Missing Data

Suppose in the model added to the unknown parameter $\theta$ we have missing data $z$. Then if we are interested about the marginal distribution of the parameter we can write it as

$$p(\theta|y) \propto \int p(y, z|\theta)p(\theta)dz$$

Data Augmentation

Given $\theta^{(t)}$ 1. Simulate $z^{(t+1)} \sim p(z|\theta^{(t)})$ 2. Simulate $\theta^{(t+1)} \sim p(\theta|z^{(t+1)})$. Each of the sequence $\theta^{(t)}$ and $z^{(t)}$ produced by this algorithm is a Markov chain with corresponding stationary distributions $\pi(\theta) = \int \pi(\theta, z)dz$ and $\pi(z) = \int \pi(\theta, z)dz$.

Example

<table>
<thead>
<tr>
<th>Number of Passages</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>139</td>
</tr>
<tr>
<td>1</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>4 or more</td>
<td>13</td>
</tr>
</tbody>
</table>

The likelihood is

$$p(\lambda|y_1, \cdots, y_5) \propto e^{-347\lambda} \lambda^{128+55x2+25x3} \times (1 - e^{-\lambda} \sum_{i=0}^{3} \lambda^i / i)$$

for $y_1 = 139, \cdots, y_5 = 13$.

For $p(\lambda) = 1/\lambda$ and $z = z_1, \cdots, z_{13}$ vector of 13 missing units larger than 4, it is possible to get the complete conditional distributions $p(\lambda|z_1, \cdots, z_{13}, y)$ and $p(z|\lambda, y)$ explicitly. (i) Simulate $z_i^{(t)} \sim Poisson(\lambda^{(t-1)})(z \geq 4); (i = 1, \cdots, 13)$. (2) Simulate $\lambda^{(t)} \sim Gam(313 + \sum_{i=1}^{13} z_i^{(t)}, 360)$.

Using the MCMC sample

After performing the MCMC, a sample vectors $\theta_1, \cdots, \theta_n$ generated from the posterior distribution $\pi$. A sample from the $i$th component of $\theta$ is given by $\theta_{1i}, \cdots, \theta_{ni}$. Marginal point or interval summaries of any real function $\phi = t(\theta)$ are estimated by their corresponding estimators based on the sample. Posterior mean of $\phi$ is estimated by $E(\phi) = \hat{\phi} = 1/n \sum_{j=1}^{n} \phi_j$
where $\phi_j = t(\theta_j)$. The posterior variance is estimated by $1/n \sum_{j=1}^{n} (\phi_j - \hat{\phi})^2$. To make the draws quasi-independent we may consider every $k$th iteration values. Also we may need to throw initial iteration values. Credibility intervals are similarly obtained by estimating the interval limits by sample quantiles.

The marginal density $\pi(\theta_i)$ can be estimated by the histogram (or smoothed version of it) of sampled values of $\theta_i$. So availability of a sample for $\theta = (\theta_1, \cdots, \theta_d)$ implies that samples for the component $\theta_j$, $j = 1, \cdots, d$ are given by the $j$th components of each sample value $\theta_i$.

Better estimators can be obtained by using conditional distributions. Recalling that $\pi(\theta_i) = \int \pi(\theta_i | \theta_{-i}) \pi(\theta_{-i} d\theta_{-i}$ a Monte Carlo estimator is given by (Rao-Blackwell estimator)

$$\pi(\theta_i) = 1/n \sum_{j=1}^{n} \pi(\theta_i | \theta_{j-i})$$

**Monitoring Convergence to the stationary distribution**

A value from the distribution of interest $\pi$ is only obtained when the number of iterations of the chain approaches infinity. In practice this is not attainable and a value obtained at a sufficiently large iteration is taken. How large the iteration should be?

(a) Graphical Methods: Draw pictures of the output of simulated chains ($t$ vs $\theta_i$), in order to detect nonstationary behaviors. A trajectory of the chain exhibiting the same qualitative behavior through iterations after a transient initial period is an indication of convergence. Similarly, the trajectory of the ergodic averages can be evaluated and plotted. Similar behavior over many successive iterations indicates convergence. Gelfand and Smith (1990), JASA, 398-409. (b) Time Series analysis: Consider a real function $\phi = t(\theta)$ and its trajectory $\phi^{(1)}, \phi^{(2)}, \cdots$ obtained from $\phi^{(j)} = t(\theta^{(j)})$, $j = 1, 2, \cdots$. This trajectory defines a time series and ergodic averages of this series can be evaluated. Assume observation of the chain for $m + n$ iterations and form averages $\bar{\phi}_1 = \frac{1}{m} \sum_{j=m+1}^{m+n} \phi^{(j)}$ and $\bar{\phi}_2 = \frac{1}{n} \sum_{j=m+n-n_a+1}^{m+n} \phi^{(j)}$. If $m$ is the length of the burn-in period, then $\bar{\phi}_1$ and $\bar{\phi}_2$ are the ergodic averages at the end and beginning of the convergence period and should behave similarly. As $n$ gets large and the ratios $n_a/n$ and $n_b/n$ remain fixed then

$$Z = \frac{\bar{\phi}_1 - \bar{\phi}_2}{\sqrt{\text{Var}(\bar{\phi}_1) + \text{Var}(\bar{\phi}_2)}}$$

goes to $N(0,1)$ in distribution.
Choose \( n_b = .1n, n_a = .5n \) and use spectral density estimators for the variance. Also can take \( t(\theta) = -2\log \pi(\theta) \) for the posterior density itself. Gewke (1992), Bayesian Statistics 4, 169-93.

(c) Nonparametric Tests: Standard non-parametric tests, such as Kolmogorov-Smirnov can be applied in the stationarity assessment of a single output of the chain \( \theta^{(t)} \). When the chain is stationary \( \theta^{(t_1)} \) and \( \theta^{(t_2)} \) have the same marginal distribution for arbitrary times \( t_1 \) and \( t_2 \). Divide the MCMC samples in two parts \( (\theta^{(1)}, \ldots, \theta^{(T/2)}) \) and \( (\theta^{(T/2+1)}, \ldots, \theta^{(T)}) \).

But most of the nonparametric tests are devised and calibrated in terms of iid samples, there need to be a correction for the correlation between the \( \theta^{(t)} \)'s. Select a batch size \( G \), which means choose the subsamples from each half as \( \theta^{(G)}, \theta^{(2G)}, \ldots \) which are quasi-independent. Then create usual Kolmogorov-Smirnov statistic based on these samples.

(d) Multiple Chains: Use ANOVA techniques. The overall idea is to test whether dispersion within chains is larger than dispersion between chains. This is equivalent to the histogram of all chains being similar to all the histograms of individual chains. Considering \( m \) parallel chains and a real function \( \phi = t(\theta) \), there are \( m \) trajectories \( \{\phi_{i1}, \ldots, \phi_{in}\}, i = 1, \ldots, m \). The variance between between chains \( B \) and within chains \( W \) are given by

\[
B = \frac{n}{m-1} \sum_{i=1}^{m} (\bar{\phi}_i - \bar{\phi})^2
\]

and

\[
W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} (\phi_{ij} - \bar{\phi}_i)^2
\]

where \( \bar{\phi}_i \)= average of observation of chain \( i \), \( \bar{\phi} \) = Averages of averages. Under convergence, all these \( mn \) values are drawn from the posterior and the variance of \( \phi \) (say \( \sigma_\phi^2 \)) can be consistently estimated by \( W \), \( B \) and the weighted average \( \hat{\sigma}_\phi^2 = (1 - 1/n)W + (1/n)B \). If the chains have not yet converged, \( \hat{\sigma}_\phi^2 \) overestimate \( \sigma_\phi^2 \). (due to oversispersion of the initial values)

\( W \) will underestimate \( \sigma_\phi^2 \) because each chain will not have adequately traversed the complete state space. Potential scale reduction= \( \hat{R} = \sqrt{\hat{\sigma}_\phi^2}/W \), that is always larger than 1. As \( n \to \infty \) both \( \hat{R} \to 1 \). So convergence can be evaluated by the proximity of \( \hat{R} \) to 1. (practically value below 1.2 is fine). Gelman and Rubin (1992), Statistical Science, 457-511.

(E) Methods based on conditional distributions: If \( \theta \) can be divided in two blocks \( \theta_1 \) and \( \theta_2 \). Then \( \pi(\theta) = \pi(\theta_1|\theta_2)\pi(\theta_2) = \pi(\theta_2|\theta_1)\pi(\theta_1) \) for all \( \theta \). With Gibbs sampling
full conditional are available but need to estimate the marginals \( \hat{\pi}(\theta_i), i = 1, 2 \). Now say
\[\eta = \pi(\theta_1 | \theta_2) \hat{\pi}(\theta_2) = \pi(\theta_2 | \theta_1) \hat{\pi}(\theta_1) .\]
if the chain has converged then \( \eta \) should be near to 0.

Say \( \eta_1 = \frac{\pi(\theta_2 | \theta_1) \hat{\pi}(\theta_1)}{\pi(\theta_2 | \theta_1) \hat{\pi}(\theta_1)} \) where \( \theta^* = (\theta_1^*, \theta_2^*) \) is another value from the state space. Similarly define \( \eta_2 = \frac{\pi(\theta_1 | \theta_2) \hat{\pi}(\theta_2)}{\pi(\theta_1 | \theta_2) \hat{\pi}(\theta_2)} \). If the chain converged then \( \eta_1 \) and \( \eta_2 \) will be close and if they are close to \( \eta_3 = \pi(\theta) / \hat{\pi}(\theta^*) \) then it has converged to the right stationary distribution.

Zellner and Min (1995) JASA, 921-27. Ritter and Tanner (1992) JASA 861-868 also propose to assess convergence of the chain by looking at \( \eta_1 \) and \( \eta_2 \). They suggest evaluating the ratios at the chain values \( \theta^{(n)} \) and plotting the histograms of the ratios. As \( n \to \infty \), these histograms should become closer to a degenerate distribution at the value of 1. See Cowles and Carlin 91996), JASA, 883-904 FOR DISCUSSION OF ALL THESE METHODS.