3.3.5 Autocorrelation in the Markov chain

One common cause of concern is that the draws from a Markov chain are correlated.

- The autocorrelation does not affect how we estimate posterior means, variances, plot marginal posteriors etc based on the output from the chain.
- The autocorrelation affects the precision of the estimates and estimation of the numerical standard error of the estimates becomes an issue.

1. Skip Draws

- Use only every m^{th} draw from the chain, choosing m just large enough to make the correlation between θ_i and θ_{i+m} neglilible. Based on a sample of size R/m we can then estimate the numerical standard error of the estimate

$$\overline{g}_{R/m} = \frac{m}{R}\sum_{i=1}^{R/m} g\left(\theta^{([i-1]m+1)}\right)$$

as

$$\widehat{V}\left(\overline{g}_{R/m}\right) = \sqrt{\frac{\widehat{V}\left(g\left(\theta\right)\right)}{R/m}}.$$

This is an inefficient procedure. It can be shown that $V\left(\overline{g}_{R/m}\right) \geq V\left(\overline{g}_{R}\right)$.

2. Batch means

- Batched means. Divide the data in R/m batches, each containing m consecutive draws from the Markov chain, and calculate the batch means

$$\overline{g}_{m,j} = \frac{1}{m}\sum_{(j-1)m+1}^{jm} g\left(\theta^{(i)}\right).$$

For reasonably large m the batch means will be essentially uncorrelated and we can estimate the variance of the batch means as

$$\widehat{V}\left(\overline{g}_{m}\right) = \frac{1}{R/m-1}\sum_{j=1}^{R/m}\left(\overline{g}_{m,j}-\overline{g}_{R}\right)^{2}.$$

An estimate of the variance of $\overline{g}_R$ is then given by

$$\widehat{V}\left(\overline{g}_{R}\right) = \frac{m}{R}\widehat{V}\left(\overline{g}_{m}\right) = \frac{m}{R\left(R/m-1\right)}\sum_{j=1}^{R/m}\left(\overline{g}_{m,j}-\overline{g}_{R}\right)^{2}.$$

3. Use every draw and compute correct estimate for variance of the sample mean: (Preferred method)

- Spectral estimate. The asymptotic variance of the mean of g is

$$\sigma^2 = V_\pi(g) + 2\sum_{k=1}^{\infty} Cov_\pi\left[g\left(x^{(i)}\right), g\left(x^{(i+k)}\right)\right].$$

The variance of $\overline{g}_R$ can the be consistently estimated using standard time series techniques as

$$\widehat{V}(\overline{g}_R) = \frac{1}{R}\sum_{j=-m}^{m}\left(1 - \frac{|j|}{m+1}\right)\widehat{\gamma}_j$$

where

$$\widehat{\gamma}_j = \frac{1}{R}\sum_{i=1}^{R-j}\left[g\left(\theta^{(i)}\right) - \overline{g}_R\right]\left[g\left(\theta^{(i+j)}\right) - \overline{g}_R\right]$$

is the estimated covariance between $g\left(\theta^{(i)}\right)$ and $g\left(\theta^{(i+j)}\right)$. Consistency requires that the truncation $m = o\left(R^{1/4}\right)$.

- The autocorrelation will in general lead to a loss of efficiency compared to the case when we can generate iid draws from the posterior. As with importance sampling we can measure the loss with the relative numerical efficiency

$$RNE = \frac{\widehat{V}(g(\theta))/R}{\widehat{V}(\overline{g})}.$$

3.3.6 Assessing convergence

Although the theoretical results doesn't require that we wait until the Markov chain has converged on the stationary distribution before we start sampling from the chain convergence is an important issue in practice since we will always have a finite number of draws.

- Convergence diagnostics are based on the output from the chain and can never prove convergence.
- Has the chain converged on a stationary distribution? Has it converged on the right one?
- Is the chain mixing well or does it have a tendency to get stuck in parts of the parameter space?
  - High autocorrelation.

  - Multimodality.

- Sensitivity to starting values.
- How large should the burn in be and how many draws from the sampler do we need?
- Simple diagnostics
  - Autocorrelations, should be small.
  - Running means of functions, g, of interest. Do they converge?
  - Run several chains with different starting points and compare the results.

Formal diagnostics The litterature on MCMC diagnostics is reviewed in Cowles & Carlin (1996), Brooks & Roberts (1998) and Mengersen, Robert & Guihennec-Jouyaux (1999). Here we will give two simple diagnostics.

Geweke (1992) proposed monitoring convergence by the statistic

$$z_G = \frac{\overline{g}_a - \overline{g}_b}{\sqrt{V\left(\overline{g}_a\right) + V\left(\overline{g}_b\right)}}$$

where $g\left(\cdot\right)$ is some function of the output of the chain and

$$\overline{g}_a = \frac{1}{n_a}\sum_{i=m+1}^{m+n_a} g\left(\theta^{(i)}\right),\ \overline{g}_b = \frac{1}{n_b}\sum_{i=m+n-n_b+1}^{m+n} g\left(\theta^{(i)}\right)$$

for a chain that is run $m+n$ steps with $n > n_a+n_b$ and the distance between the estimates such that they can be taken to be uncorrelated. The variances are estimated taking account of the autocorrelation structure in the chain, for example by the spectral estimate above. If the chain has converged after m steps the distribution of z_G is well aproximated by a standard normal. Calculating z_G for a range of values of m and comparing to a standard normal will thus give an indication of the burn in needed for the chain.

Gelman & Rubin (1992) proposed running several chains started at points that are overdispersed compared to the posterior. Let $\theta_j^{(i)}$ denote the output from chain j for m chains run n steps from burn-in and define the between and within chain variances as

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\overline{g}_j - \overline{g}\right)^2,\quad W = \frac{1}{m\left(n-1\right)}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(g\left(\theta_j^{(i)}\right) - \overline{g}_j\right)^2.$$

$$\overline{g}_j = \frac{1}{n}\sum_{i=1}^{n} g\left(\theta_j^{(i)}\right),\quad \overline{g} = \frac{1}{m}\sum_{j=1}^{m}\overline{g}_j$$

Convergence failure or convergence on different stationary distributions after the selected burn-in is indicated by the between chain variation, B, being larger than the within chain variation, W. An overall estimate of the variance is given by $V = (1 - 1/n)\,W + B/n$. Gelman and Rubin suggest using $R = \sqrt{V/W}$ as a measure of convergence and accepting convergence if $R < 1.2$.