Computation IV

Statistics 220

Spring 2005

Copyright ©2005 by Mark E. Irwin

Metropolis - Hastings Example

Pump Example:

$$\begin{aligned} s_i|\lambda_i &\overset{ind}{\sim} Poisson(\lambda_i t_i) \\ \lambda_i|\mu,\sigma^2 &\overset{iid}{\sim} LogN(\mu,\sigma^2) \\ \mu &\sim N(\delta,\tau^2) \\ \sigma^2 &\sim \text{Inv}-\chi^2(\nu,\gamma) \end{aligned}$$

Can perform Gibbs on μ and σ^2 easily, but not on λ, due the non-conjugacy of the Poisson and log Normal distributions.

Step $i, i = 1, \ldots, 10$ (M-H):

Sample λ_i from $\lambda_i|s, \mu, \sigma^2$ with proposal $\lambda_i^* \sim logN(\lambda_i, \theta^2)$ (Multiplicative random walk)

$$
\begin{aligned}
r_i &= \frac{(\lambda_i^* t_i)^{s_i} e^{-\lambda_i^* t_i} \frac{1}{\lambda_i^* \sigma} \phi\left(\frac{\log \lambda_i^* - \mu}{\sigma}\right)}{(\lambda_i t_i)^{s_i} e^{-\lambda_i t_i} \frac{1}{\lambda_i \sigma} \phi\left(\frac{\log \lambda_i - \mu}{\sigma}\right)} \times \frac{\frac{1}{\lambda_i \theta} \phi\left(\frac{\log \lambda_i - \lambda_i^*}{\theta}\right)}{\frac{1}{\lambda_i^* \theta} \phi\left(\frac{\log \lambda_i^* - \lambda_i}{\theta}\right)} \\
&= \left(\frac{\lambda_i^*}{\lambda_i}\right)^{s_i} e^{-(\lambda_i^* - \lambda_i) t_i} \frac{\phi\left(\frac{\log \lambda_i^* - \mu}{\sigma}\right)}{\phi\left(\frac{\log \lambda_i - \mu}{\sigma}\right)}
\end{aligned}
$$

Step 11 (Gibbs): Sample μ from $\mu|\lambda, \sigma^2, \delta, \tau^2 \sim N(\text{mean}, \text{var})$ where

$$\begin{aligned} \text{mean} &= \text{var}\left(\frac{1}{\sigma^2}\sum \log \lambda_i + \frac{\delta}{\tau^2}\right) \\ \text{var} &= \left(\frac{n}{\sigma^2} + \frac{\delta}{\tau^2}\right)^{-1} \end{aligned}$$

Step 12 (Gibbs): Sample σ^2 from

$$\sigma^2|\lambda, \mu, \nu, \gamma \sim \text{Inv}{-}\chi^2\left(\nu + n, \gamma + \sum(\log \lambda_i - \mu)^2\right)$$

Parameters for run:

- Burn-in: 1000
- Imputations: 100,000
- $\delta = -50$
- $\tau^2 = 100$
- $\nu = 2$
- $\gamma = 100$
- $\theta = 0.1$

Starting values:

- $\lambda_i = l_i$
- $\mu = \frac{1}{10} \sum \log l_i$
- $\sigma^2 = \frac{1}{9} \sum \log(l_i - \mu)^2$

Other sampler options:

1. Combine steps 1 - 10 into a single draw.

$$r = \prod_{i=1}^{10} r_i = \prod_{i=1}^{10} \left(\frac{\lambda_i^*}{\lambda_i}\right)^{s_i} e^{-(\lambda_i^* - \lambda_i)t_i} \frac{\phi\left(\frac{\log \lambda_i^* - \mu}{\sigma}\right)}{\phi\left(\frac{\log \lambda_i - \mu}{\sigma}\right)}$$

   With this option all λ's change or none do. In the sampler used, whether each λ_i changes is independent of the other λ's.

   The option used is probably preferable, as it should lead to better mixing of the chain.

2. Combine sampling λ, μ, and σ^2 into a single M-H step. Probably suboptimal as the proposal distribution won't be a great match for the joint posterior distribution of λ, μ, and σ^2.

Rejection Rates

Having some rejection can be good.

With the multiplicative random walk sampler used, if θ^2 is too small, there will be very few rejections, but the sampler will move too slowly through the space.

Increasing θ^2 will lead to better mixing, as bigger jumps can be made, though it will lead to higher rejection rates.

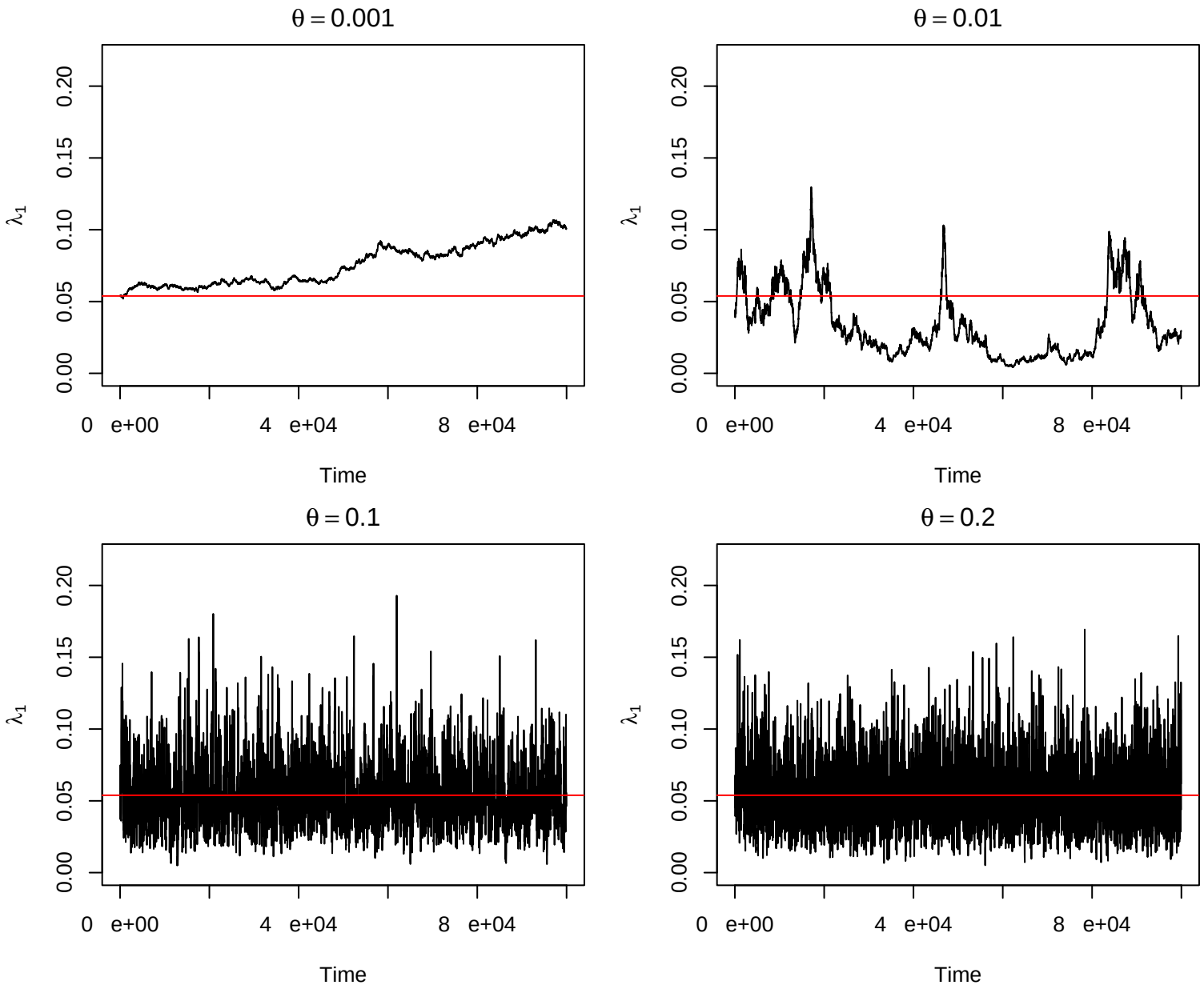
You need to find a balance between rejection rates, mixing of the chain, and coverage of the state space.

For some problems, a rejection rate of 50% is fine and I've seen reports for large problems using normal random walk proposals the rejection rates of 75% are optimal.

The book recommends rejections rates around 80% (when altering a vector of parameters) and 60% (when altering one parameter at a time).

Rejection rates under different random walk standard deviations

Parameter	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.1$	$\theta = 0.2$
λ_1	0.00093	0.00947	0.07067	0.13899
λ_2	0.00005	0.00373	0.03015	0.05986
λ_3	0.00008	0.00713	0.06990	0.13774
λ_4	0.00032	0.01225	0.11783	0.22687
λ_5	0.00017	0.00376	0.05352	0.10601
λ_6	0.00097	0.01217	0.13792	0.26114
λ_7	0.00017	0.00346	0.03036	0.05523
λ_8	0.00007	0.00231	0.02806	0.05822
λ_9	0.00046	0.00633	0.06073	0.12077
λ_{10}	0.00057	0.01332	0.14533	0.27805
μ	0.00000	0.00000	0.00000	0.00000
σ^2	0.00000	0.00000	0.00000	0.00000


θ = 0.001
θ = 0.01
θ = 0.1
θ = 0.2
λ1
0.00
0.05
0.10
0.15
0.20
0 e+00
4 e+04
8 e+04
Time

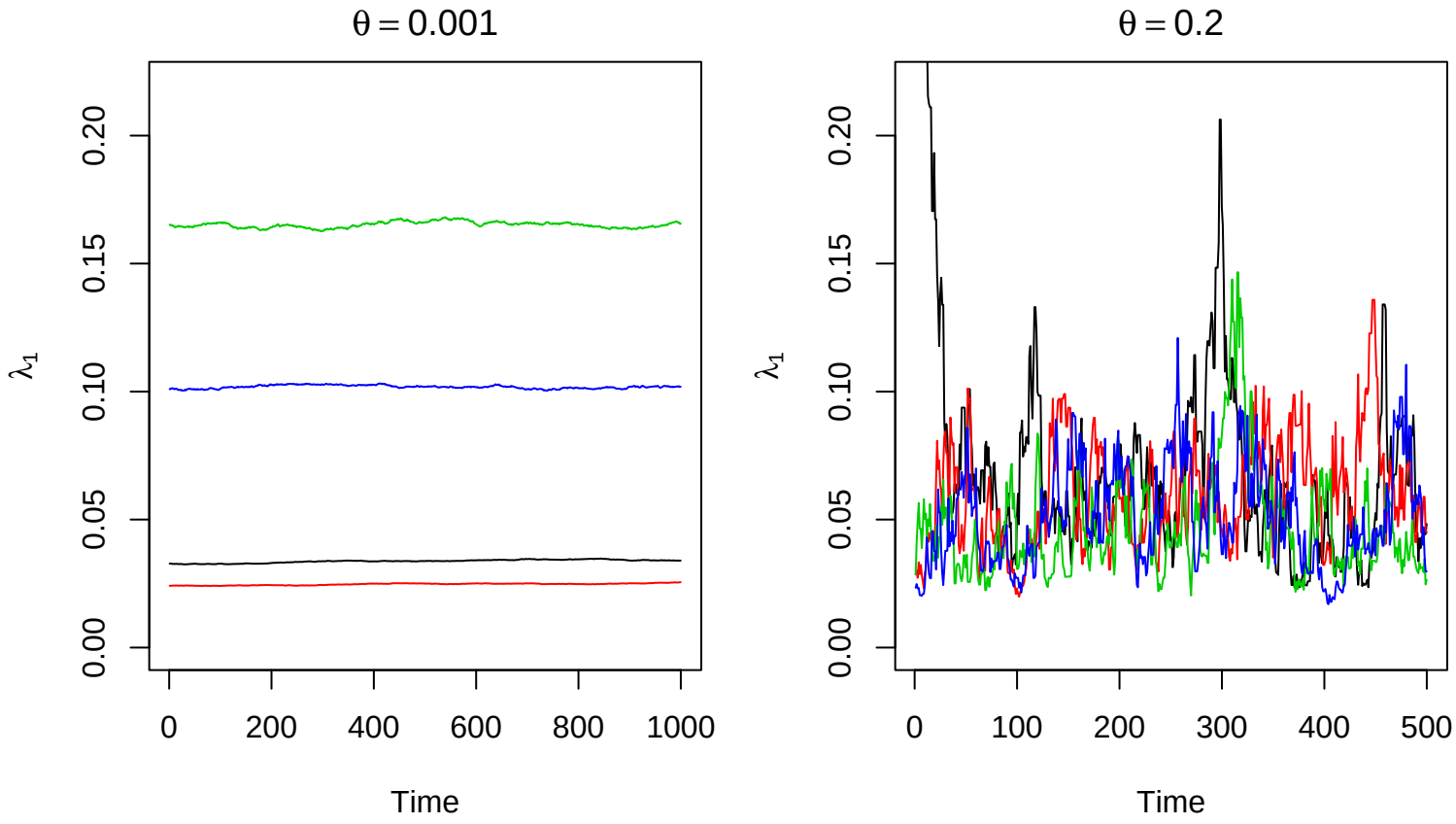
Parameter	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.1$	$\theta = 0.2$	Gibbs
λ_1	0.0755	0.0335	0.0538	0.0531	0.0534
λ_2	0.0559	0.1089	0.0606	0.0646	0.0665
λ_3	0.0775	0.0718	0.0804	0.0801	0.0796
λ_4	0.1124	0.1188	0.1122	0.1111	0.1111
λ_5	0.6817	0.5501	0.5651	0.5565	0.5603
λ_6	0.5066	0.5982	0.6020	0.6033	0.6019
λ_7	0.6452	0.1460	0.9281	0.8702	0.8889
λ_8	1.0675	1.0572	0.8023	0.8949	0.8902
λ_9	2.5088	1.5289	1.8545	1.7993	1.8553
λ_{10}	2.0121	2.1340	2.0880	2.0876	2.0856
μ	-2.2718	-2.6757	-2.5723	-2.5381	-2.5405
σ^2	26.6847	27.1960	27.3542	27.1282	27.2422

Inference and Assessing Convergence

With iterative sampling schemes, it may not be clear that we are getting the correct answer from our sampler. Two problems are

1. Has the chain run long enough to get into the stationary distribution and adequately cover the sample space.

   For example, in the pump example, for the smaller random walk standard deviations, θ, the chains had not covered the sample space well. For $\theta = 0.001$, clearly the chain had problems. However when $\theta = 0.2$, things are much better.



The starting values for λ_i are $l_i * logN(0,1)$. For λ_1, $l_1 = 0.053$ so the starting values are drawn from a positively skewed distribution with mean $= 0.087$, median $= 0.053$, and standard deviation $= 0.115$.

2. Within sample auto-correlation.

   Assume that $\mathrm{Corr}(X_t, X_{t+j}) = \rho_j$. Then

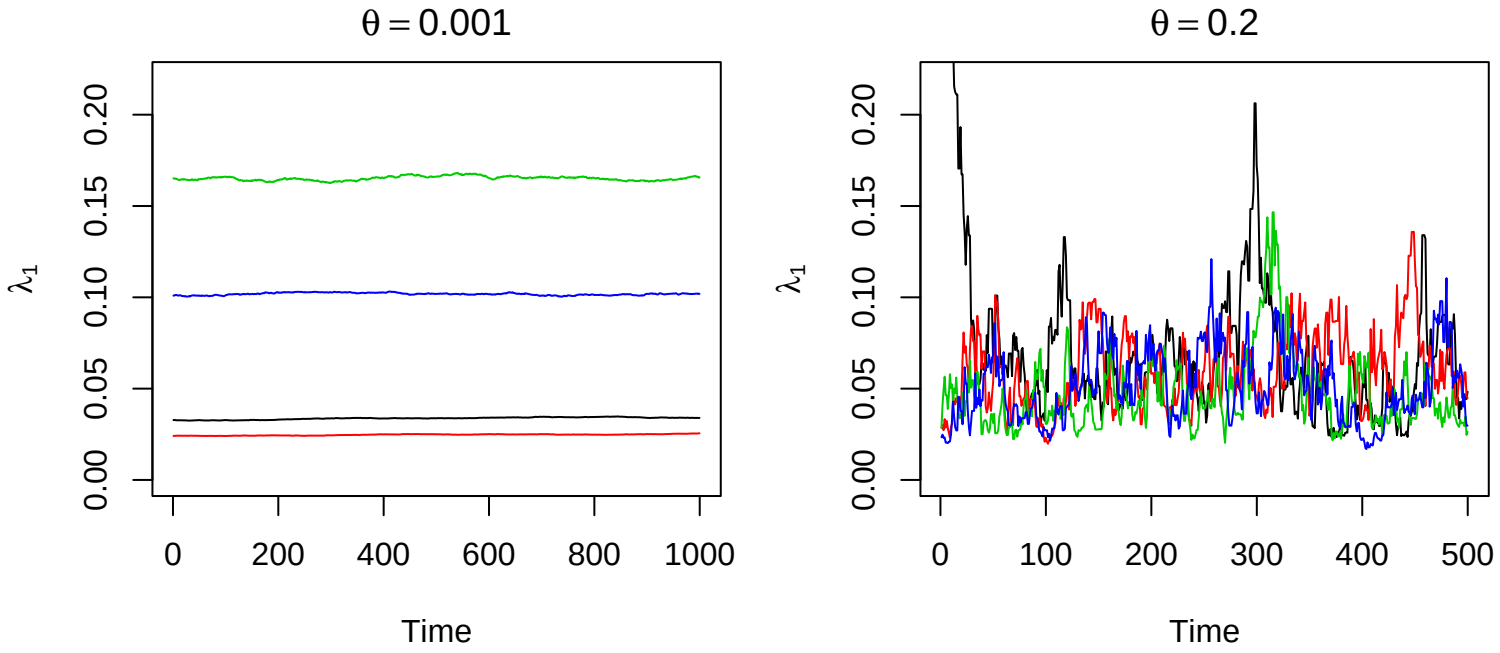
$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}\left(1 + \sum_{j=1}^{n-1} \frac{n-j}{n}\rho_j\right)$$

   As MCMC samplers tend to have positive autocorrelations, estimating the posterior mean of a distribution by the average of the sampler will usually be less efficient than an independent sampler. So the effective sample size is often less than we used in the simulation.

   So to get the same level of accuracy, we need to simulate more realizations than with an independence sampler. If space is a concern, thin the sequence, keeping every kth realization from the chain.

Monitoring Convergence

To monitor convergence of sampler, we can compare multiple chains to see if they are acting similarly. If they are, it suggests that the chains have converged to there stationary distributions.



While graphs are useful, we also want numeric summaries to compare the $m \geq 2$ chains.

Analysis of convergence is often done on scalar quantities, either individual parameters or functions of parameters.

The idea is to see if the variability between the chains is similar to the variability within the chains.

Assume there are $m \geq 2$ chains with n values kept from each chain after removing the burn in part and thinning the remaining values. Let $\psi_{ij}; i = 1, \ldots, n, j = 1, \ldots, m$ be the scalar of interest.

Let B and W be the between and within sequence variances of ψ_{ij}

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})^2 \qquad W = \frac{1}{m} \sum_{j=1}^{m} s_j^2$$

where

$$\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} \qquad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j} \qquad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{.j})^2$$

The marginal posterior variance $\mathrm{Var}(\psi|y)$ can be estimated by

$$\widehat{\mathrm{Var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

This quantity will overestimate $\mathrm{Var}(\psi|y)$ if the starting points are overdispersed, but is unbiased if the draws are from the stationary distribution or in the limit as $n \to \infty$. This estimate is analogous to the variance estimate in a cluster sample.

However W should be an underestimate of $\mathrm{Var}(\psi|y)$ as for finite n, each chain will not have covered the whole sample space. However as $n \to \infty$ W is a consistent estimate of $\mathrm{Var}(\psi|y)$.

Thus we can monitor convergence by comparing $\widehat{\mathrm{Var}}^{+}(\psi|y)$ with W. If they are similar, it suggests that the chains have converged. On comparison that can be done is by

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\psi|y)}{W}}$$

which declines to 1 as $n \to \infty$.

Thus if $\hat{R}$ is large, it suggests that the chains have not been run long enough.

What is a large $\hat{R}$ depends on the problem of interest, but a rule of thumb is the $\hat{R}$ should be below 1.1. If this is the case, then the mn draws can be combined into a single sample for inference purposes.

(Note: I've also seen $\hat{R} < 1.2$ as a rule of thumb.)

Parameter	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.1$	$\theta = 0.2$	Gibbs
λ_1	88.2941	4.7877	1.0439	1.0193	1.0010
λ_2	80.9220	4.8551	1.2036	1.0095	1.0024
λ_3	31.3008	5.6364	1.0248	1.0126	1.0017
λ_4	75.6527	1.7486	1.0062	1.0037	1.0010
λ_5	65.2930	3.8716	1.0852	1.0086	1.0009
λ_6	31.9666	2.1821	1.0217	1.0000	1.0010
λ_7	36.9225	6.4352	1.1559	1.0190	1.0020
λ_8	102.3836	5.9662	1.1599	1.1416	1.0025
λ_9	32.8548	2.6505	1.0022	1.0207	1.0021
λ_{10}	116.9852	3.3047	1.0191	1.0035	1.0008
μ	1.0027	1.0164	1.0047	1.0024	1.0015
σ^2	1.0030	1.0007	1.0020	1.0003	1.0018

Note that a small $\hat{R}$ doesn't necessarily imply that the chain has convergence. If you don't have your starting points dispersed enough, all the chains might miss an important part of the sample space. This can occur when the distribution you are sampling from is multimodal and you don't have a chain starting from one of the modes.

Also as $\hat{R}$ is based on means and variances, it works better with quantities that are approximately normal. Thus looking at transformed variables can be better for examining convergence. For example, log transform positive random variables and logit transform variables on (0,1).

Effective Sample Size

Using W and B we can compute a quantifying the effective sample size of our sampler, i.e. the sample size for an independence sampler with the same variance.

If our sample were truly independent, then B would be an unbiased estimate of $\mathrm{Var}^+(\psi|y)$. However due to the autocorrelation in an MCMC sample, B has a positive bias in estimating $\mathrm{Var}^+(\psi|y)$.

Thus one estimate of the effective sample size is

$$n_{\text{eff}} = mn\frac{\widehat{\mathrm{Var}}^+(\psi|y)}{B}$$

Actually it is better to report $\min(n_{\text{eff}}, mn)$, as superefficient estimation is extremely rare.

Parameter	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.1$	$\theta = 0.2$	Gibbs
λ_1	4.0005	4.18	47.95	103.78	4000
λ_2	4.0006	4.17	12.88	201.70	3900
λ_3	4.0040	4.12	81.97	155.54	1600
λ_4	4.0006	5.94	298.71	476.49	4000
λ_5	4.0009	4.28	26.36	220.30	810
λ_6	4.0039	5.06	92.88	3666.42	4000
λ_7	4.0029	4.09	15.84	105.42	1100
λ_8	4.0003	4.11	15.52	17.12	2700
λ_9	4.0037	4.66	719.28	97.01	3300
λ_{10}	4.0002	4.40	104.84	497.14	3100
μ	623.1476	120.93	386.60	671.59	1500
σ^2	559.2017	1653.44	772.37	2668.19	1300

(The Gibbs values are from `WinBUGS` and are rounded to 2 significant digits.)

Note that if m is small, then B will have large sampling variability.

Thus n_{eff} is only a crude estimate of the effective sample size.

It is possible to determine more precise measures based on time-series analysis of the chains.