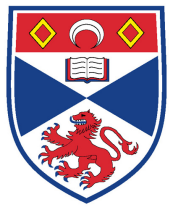
MT4531/5731: (Advanced) Bayesian Inference

Bayesian Computation (An Introduction to Markov chain Monte Carlo)

Nicolo Margaritella

School of Mathematics and Statistics, University of St Andrews

–



Outline

Outline

Markov chains

- Markov chain (MC): a stochastic sequence of numbers where each value depends *only* on the last.
  1. Choose an arbitrary start value θ^0 (starting state). Let $n = 0$.
  2. Generate a new value $\theta^{n+1} \sim \mathcal{K}(\theta^n, \theta^{n+1}) \quad (\equiv \mathcal{K}(\theta^{n+1}|\theta^n))$
  3. θ^{n+1} is now the 'current state' of the chain.
  4. Set $n = n + 1$. Go to 2 (until you have "enough" samples).
- $\mathcal{K}$ is the "transition kernel" for the chain.

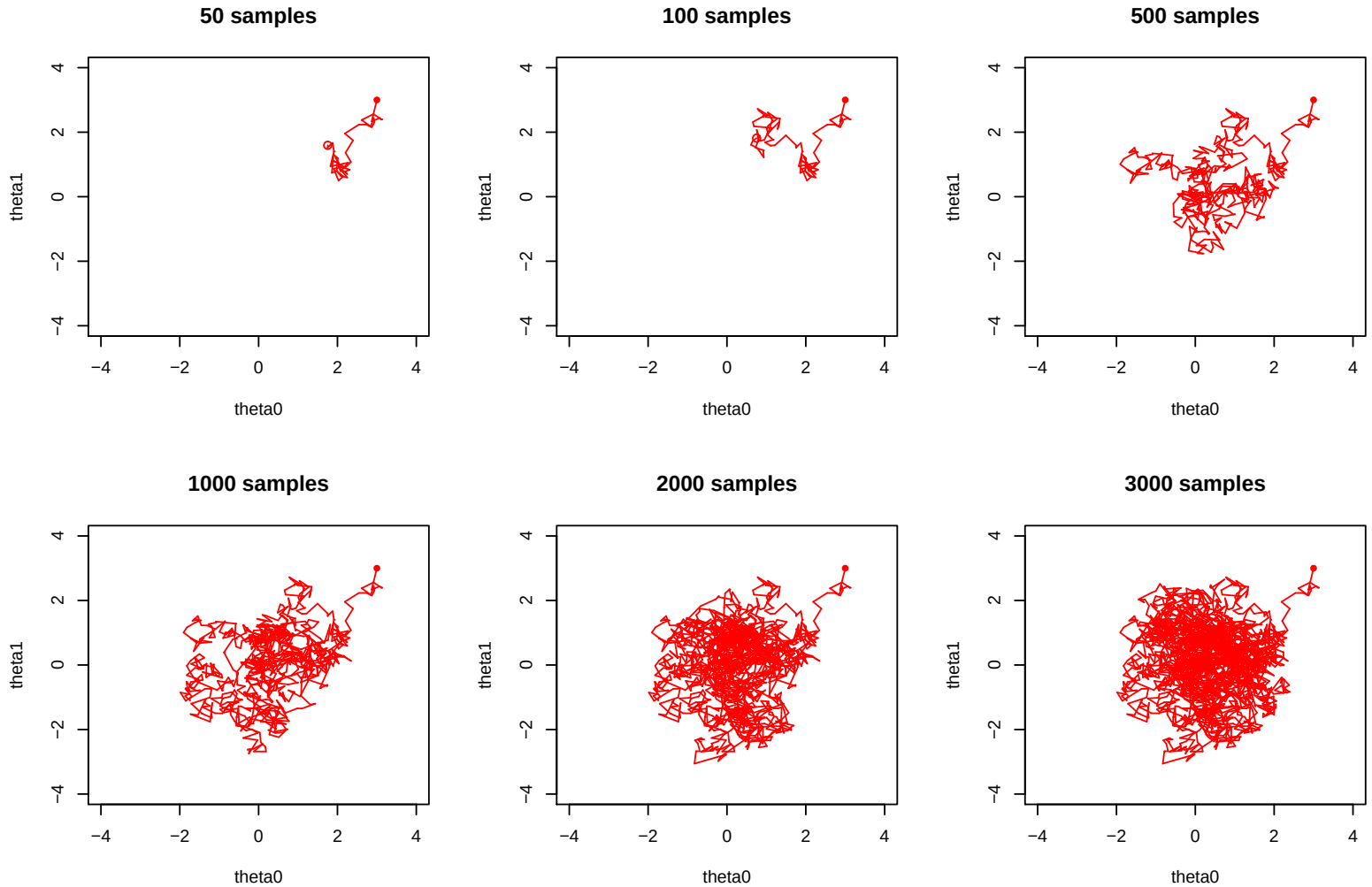
Markov chains

- Under certain conditions (that the chain is *aperiodic* and *irreducible*) the chain converges to a *stationary distribution*.
  1. A MC is irreducible if any state can be reached within finite time irrespectively of the present state.
  2. A MC is aperiodic if for any state, the chain can return to that state after a number of transitions that is a multiple of 1, and can also be 1.
- The stationary distribution is independent of the starting point.

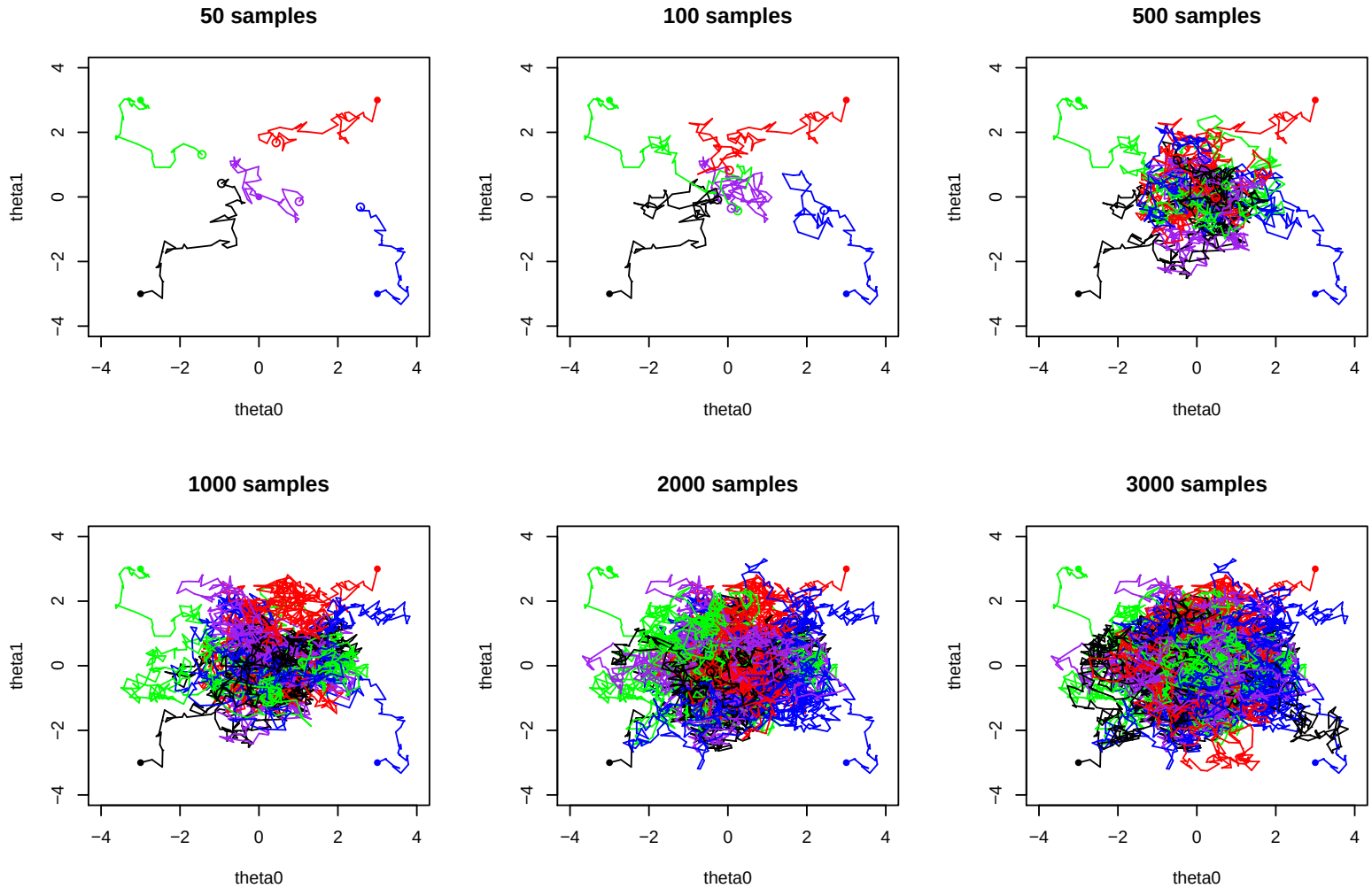
Markov chain Monte Carlo

- Choose a transition kernel such that the stationary distribution is $\pi(\boldsymbol{\theta}|\boldsymbol{x})$
- Start the chain at some value, and run until it has converged to the stationary distribution
- Collect subsequent samples, and use them for inference (see lecture on Monte Carlo integration)

Example: Single chain



Example: 5 chains



Questions, questions

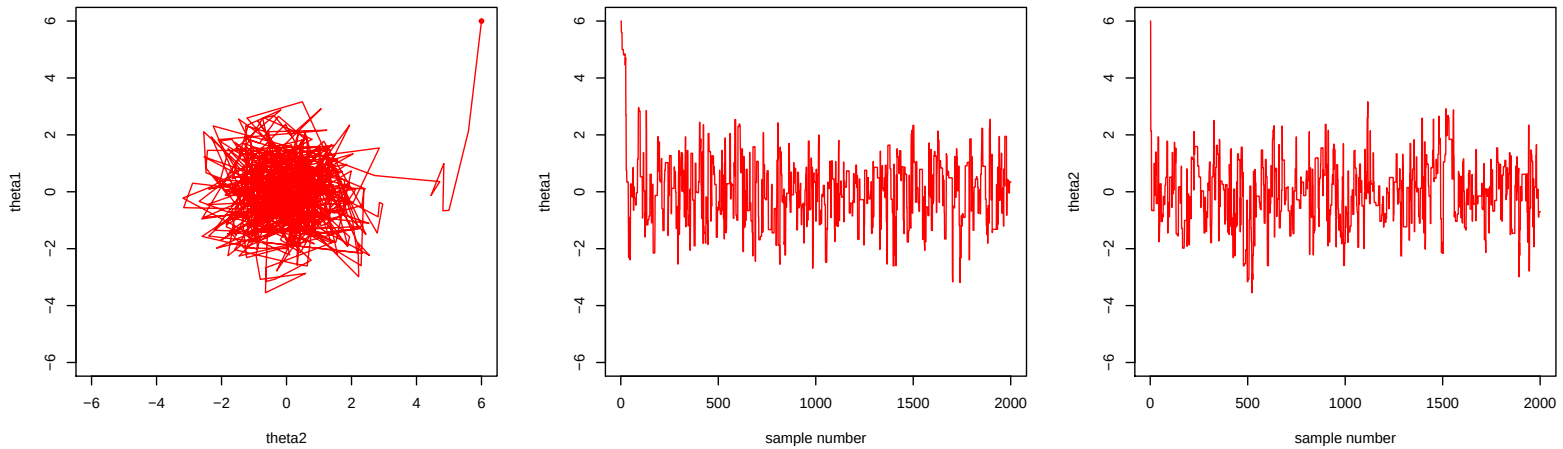
- How do we choose a transition kernel to get the correct stationary distribution?
  - Many techniques available – some will be introduced in subsequent lectures.
  - Bottom line: can be surprisingly simple, even for very complex models.
- How long do we need to run the chain before it converges and we can start using the samples?
- How many samples do we need for accurate inference?

Outline

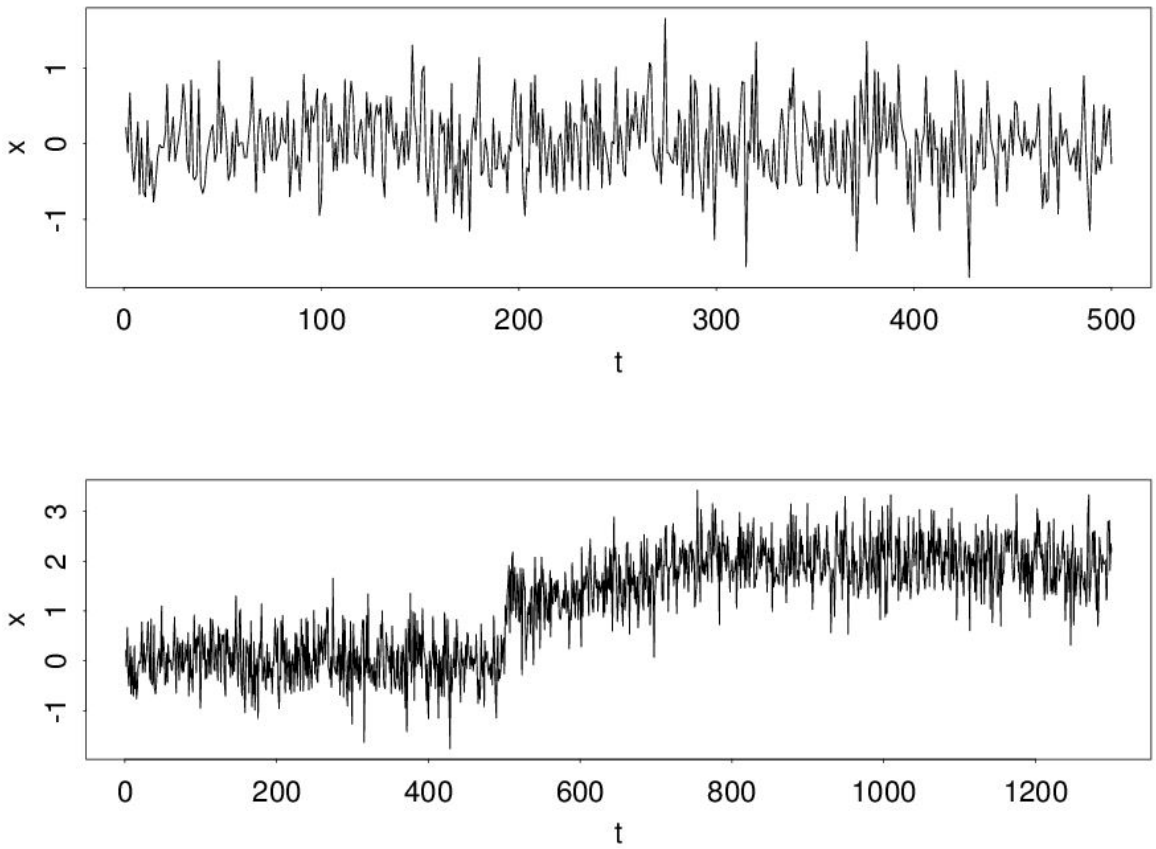
Convergence and burn-in

- We discard samples taken before the chain has converged
- The phase before convergence is called "burn-in"
- How do we determine when burn-in has finished?

Trace plots

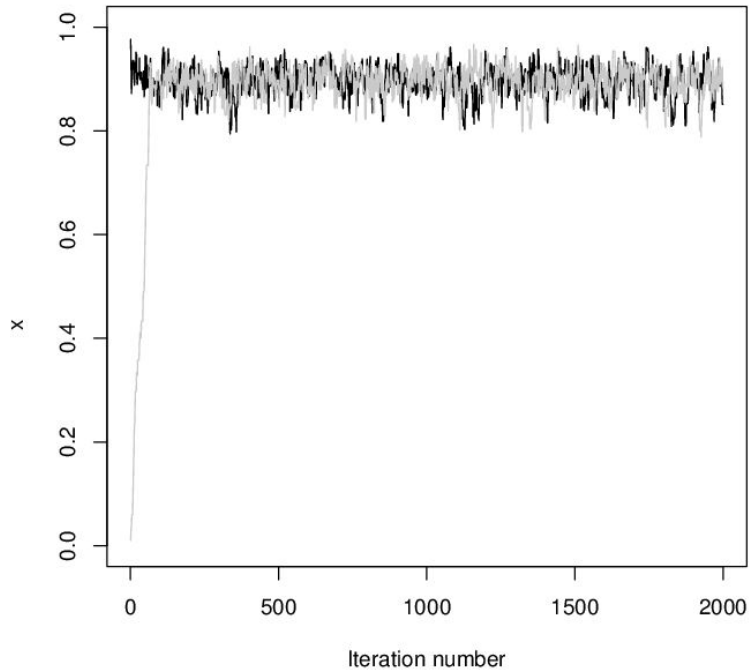


... but take care!



Visual method

- Need at least 2 chains for this.
- Start at diverse ("overdispersed") start points – or random.
- The time until the chains come together is called *mixing time*.



Brooks-Gelman-Rubin (BGR) method

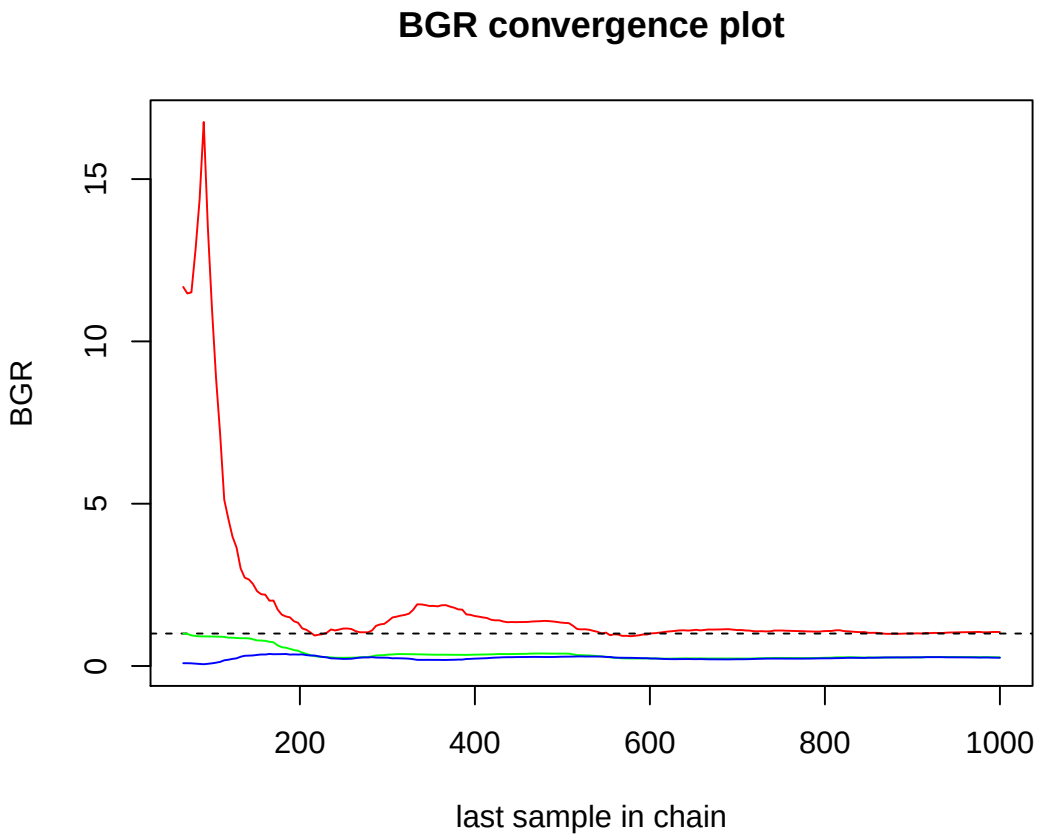
- Various flavours – all based on ANOVA-like ideas
- Need multiple chains
- Given chains of length $2n$, discard 1st n; for the remaining n calculate

$$\hat{R} = \frac{\text{width of 80\% credible interval of pooled chains}}{\text{mean of width of 80\% credible interval of individual chains}}$$

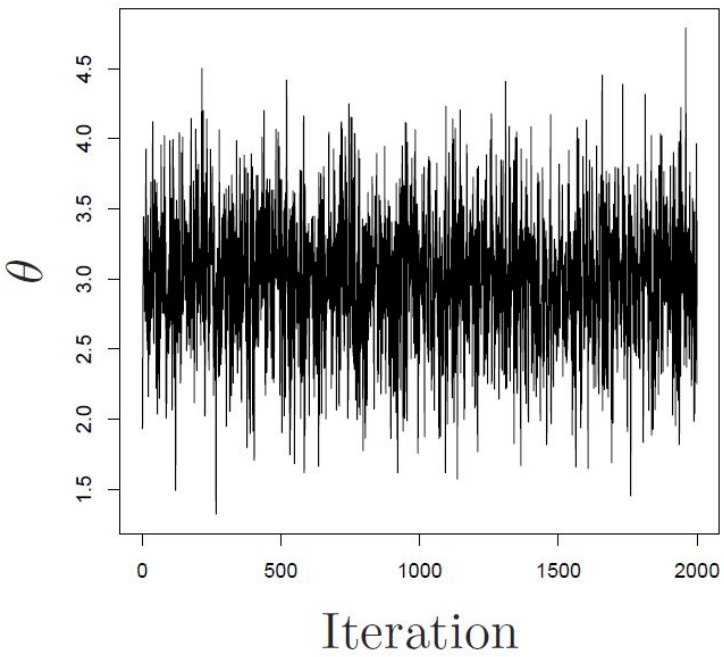
- Assume convergence when $\hat{R} \approx 1$.
  (Also need to check stability of pooled and within interval CI widths)

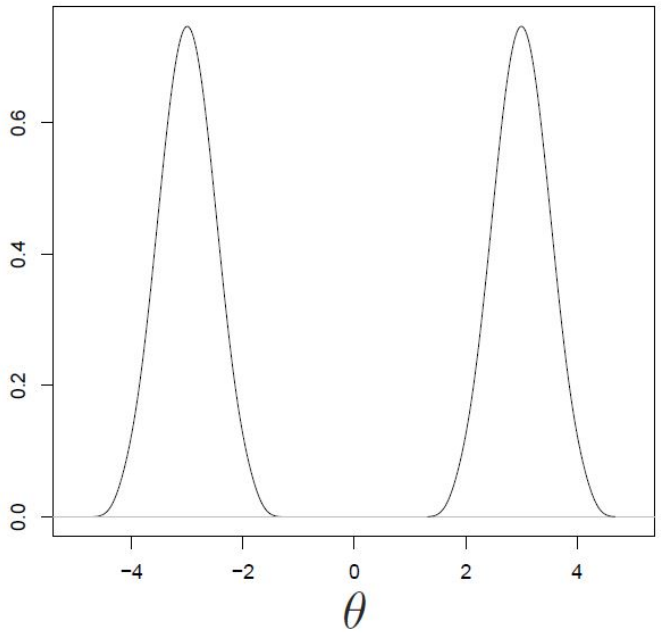
Example - BGR

red = BGR;
green = normalised width of the central 80% CI of the pooled runs;
blue = normalised average width of 80% CIs within the individual runs.



... but you never really know!





Monte Carlo Error

- (After burn-in) How many iterations (samples) do we need to sample for reliable inference?
- Recall that the sample mean $\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta^i$ has distribution

$$\overline{\theta} \sim N\left(\mathbb{E}_\pi(\theta), \frac{\sigma^2}{n}\right).$$

  where $\sqrt{\frac{\sigma^2}{n}}$ is the Monte Carlo (MC) error.
- With independent samples, we estimate σ^2 using,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2$$

- The MC error allows to calculate bounds on the error in our estimates (e.g., $\pm 2\times$MC error gives approximate 95% limits for our estimate of the mean).

Batching

- How do we calculate MC error when the samples are correlated?
- Divide the chain of samples into m batches of length T where batches are large enough that the mean is reasonably well estimated, and batch means are independent.
- Then, estimate σ^2 by

$$\hat{\sigma}^2 = \frac{T}{m-1} \sum_{i=1}^{m} (\overline{\theta}_i - \overline{\theta})^2.$$

and MC error as $\sqrt{\hat{\sigma}^2/n}$.

Effective sample size

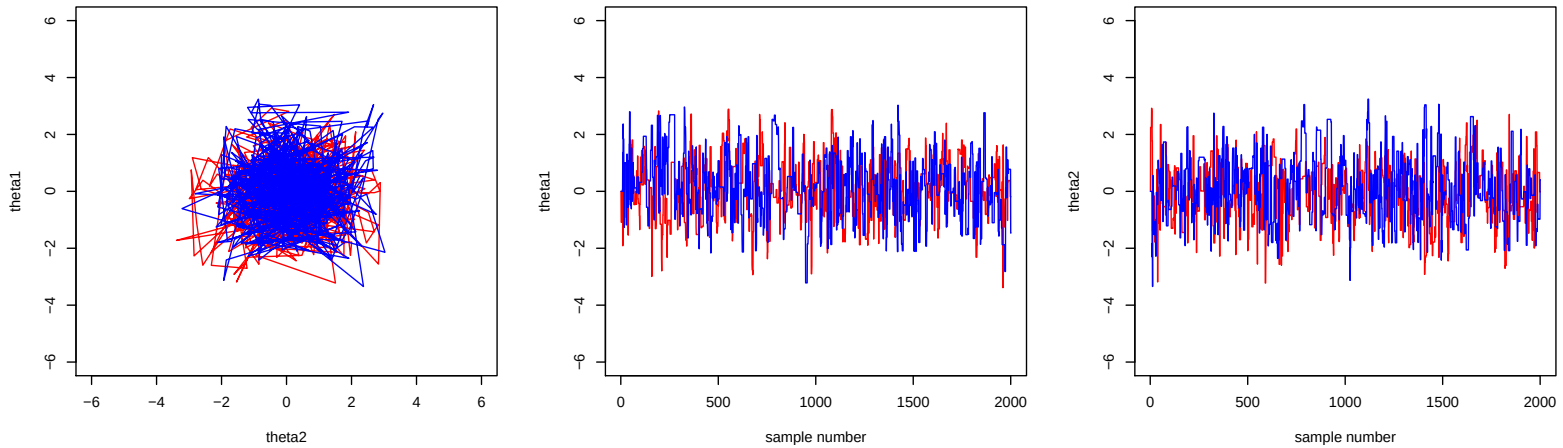
- Another method for estimating MC error relies on the concept of 'effective sample size' (ESS).
- The Effective Sample Size, M, is given by,

$$M = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

  where ρ_k is the correlation of samples k iterations apart.
- In practice an approximation is calculated (as the formula involves a sum to infinity)
- If samples are dependent, our *effective sample size* (ESS) is less than n.
- The more dependent they are, the smaller the ESS gets.
- The *autocorrelation plot* shows the correlation between successive samples.

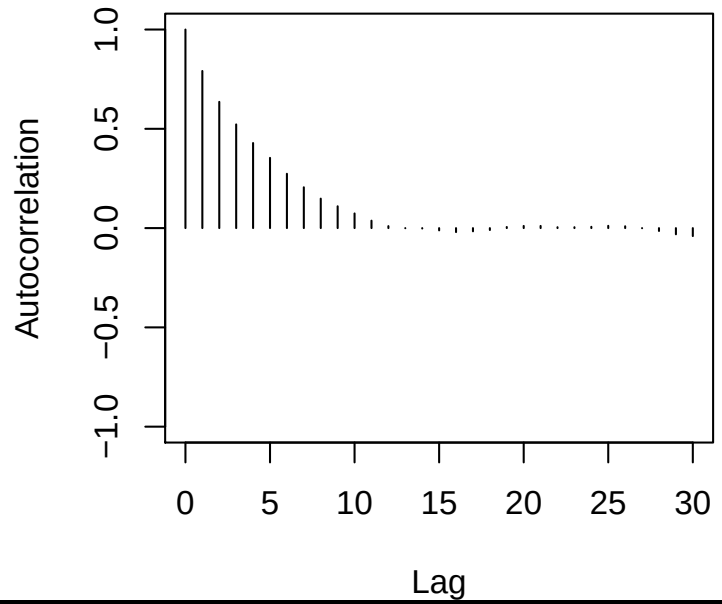
Example - moderate autocorrelation

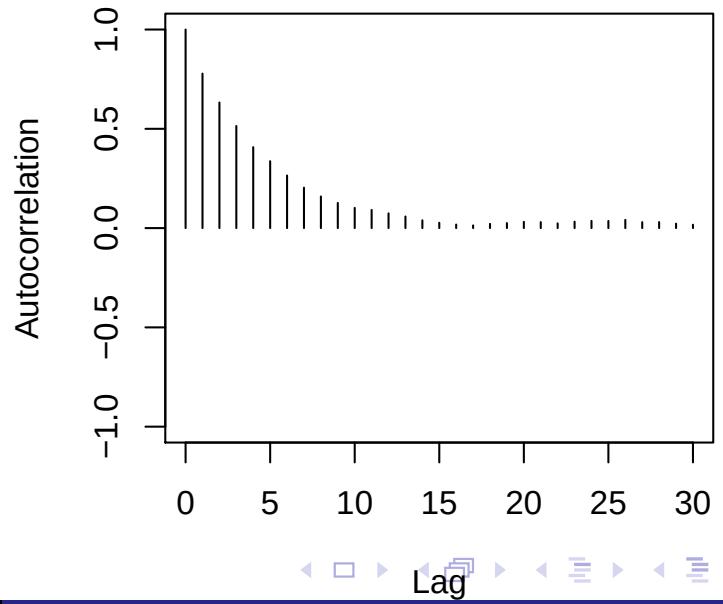
4000 samples (after burn-in) – 2000 per chain.



Autocorrelation function - moderate autocorrelation

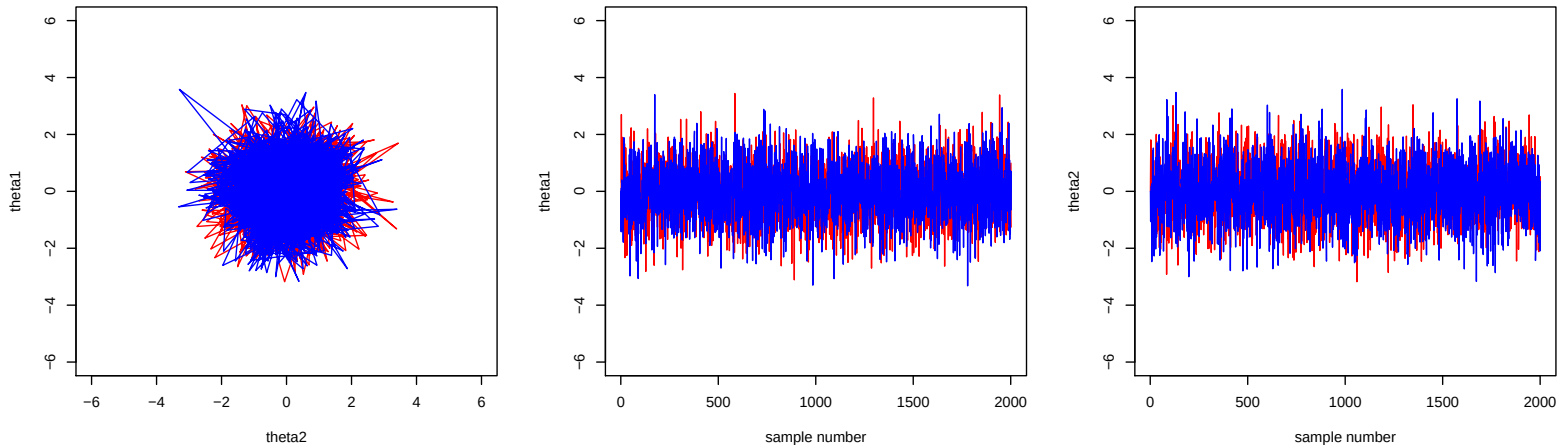
4000 samples (after burn-in). ESS $\theta_0 = 418$; ESS $\theta_1 = 416$.
In this plot, the first bar is the correlation of samples 0 iterations apart, i.e. the correlation between a sample and the same sample, which is always 1. The second bar is the correlation of samples 1 iteration apart, and so on and so forth.





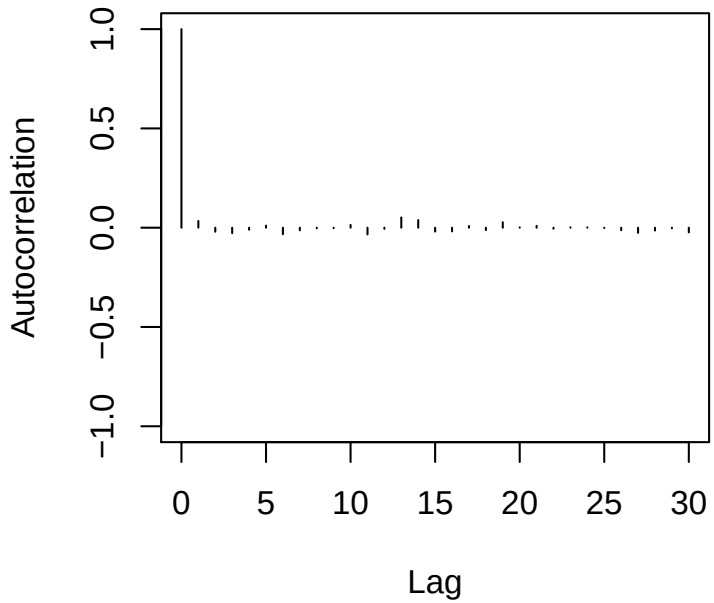
Example - no autocorrelation

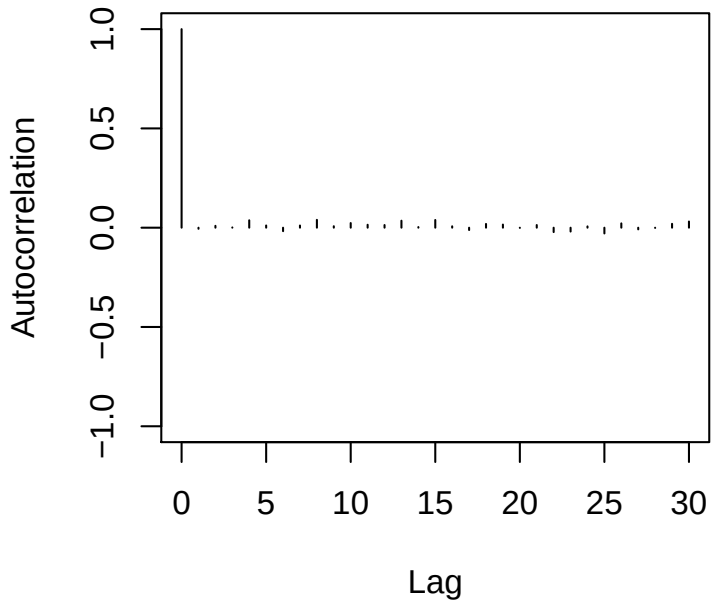
4000 samples (after burn-in) – 2000 per chain.



Autocorrelation function - no autocorrelation

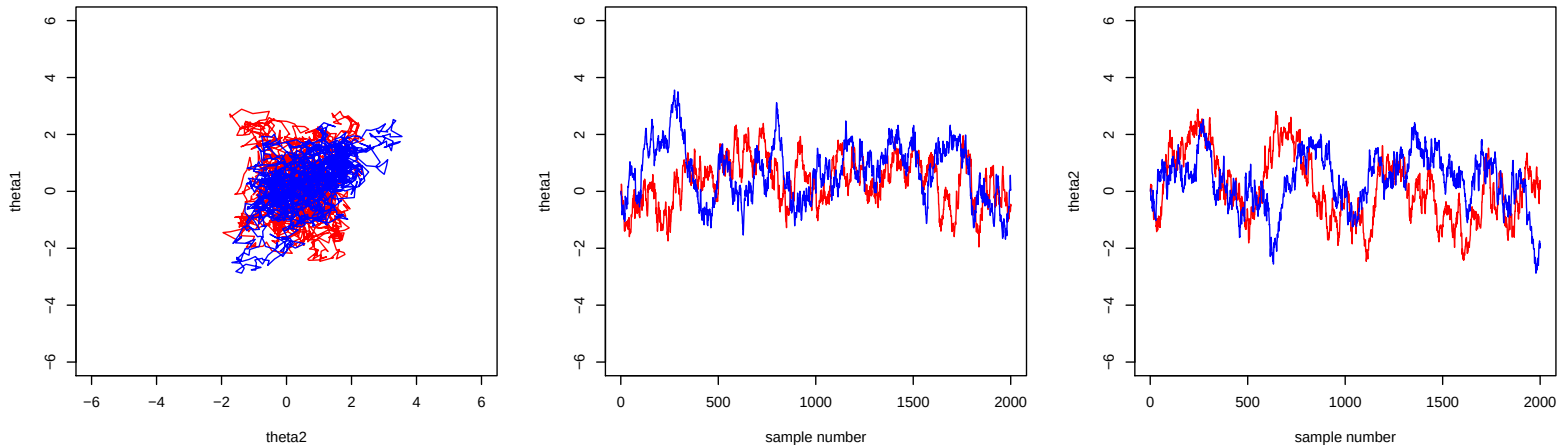
4000 samples (after burn-in). ESS $\theta_0 = 3870$; ESS $\theta_1 = 3524$.





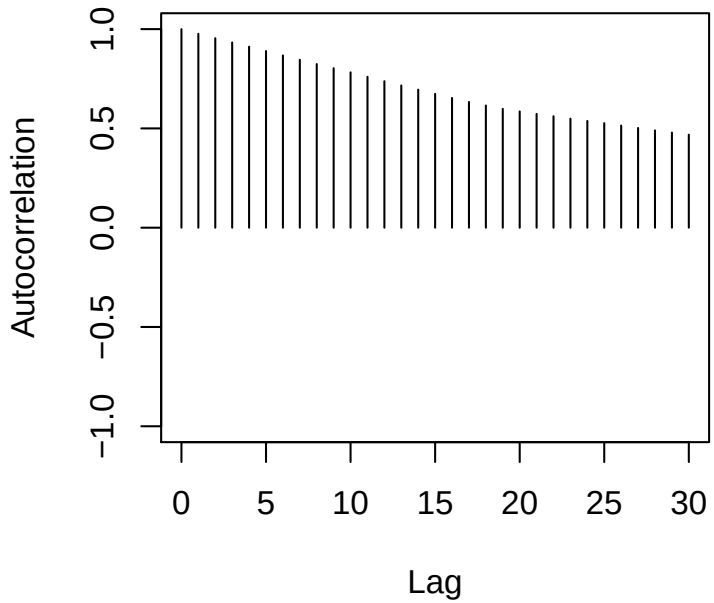
Example - high autocorrelation

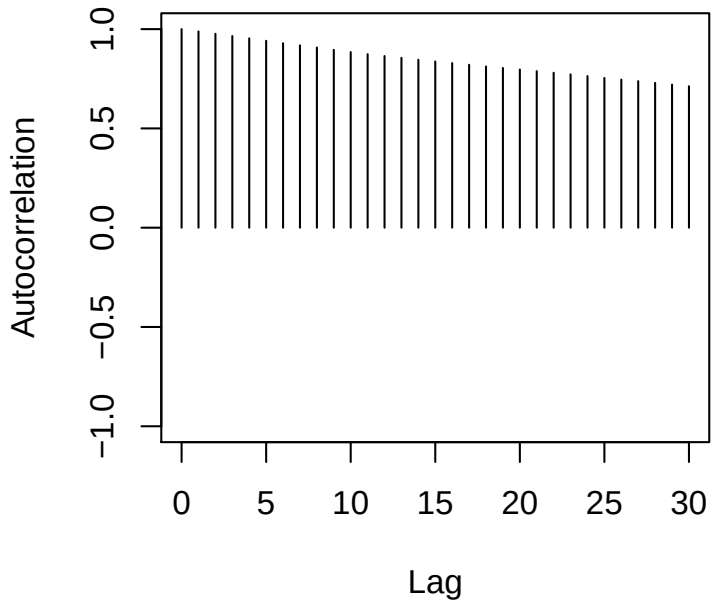
4000 samples (after burn-in) – 2000 per chain.



Autocorrelation function - high autocorrelation

4000 samples (after burn-in). ESS $\theta_0 = 39$; ESS $\theta_1 = 30$.





Effective sample size

- Once we have an estimate of M, the effective sample size, we can estimate MC error using the variance estimate,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2$$

as with independent samples, but then the MC error is estimated as $\sqrt{\hat{\sigma}^2/\hat{M}}$ rather than $\sqrt{\hat{\sigma}^2/n}$.

Thinning

- Sometimes, because samples are very dependent and so the MC chain mixes slowly, we have to collect so many samples it is hard to store and process them all.
- In this case, we can *thin* by storing only every kth.
- This reduces autocorrelation, but it also discards information.
- Better not to thin, unless you have to.

Outline

Summary

- MCMC: construct a Markov chain with $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ as the stationary distribution.
- Burn-in
  - Use multiple (overdispersed) start points; check trace plots, ACF and BGR
  - Remember we can never be completely sure the chain has converged
- Run length post-convergence
  - Check ACF for slow mixing, calculate ESS and MC error.
  - Retain enough samples to accurately estimate the parameters of interest.

Use these rules and you won't slip up!

Next lecture: BUGS

⇒ In the next lecture we will start using the NIMBLE package in RStudio to run our Bayesian analyses!

- Read Appendix B in the lecture notes for instructions to install NIMBLE on your machine.