UiO : Matematisk institutt
Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2022
Markov Chain Monte Carlo part 3

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

McMC

- Specific chains:
  - Random walk chain
  - Independent chain
  - Gibbs sampler
- Tricks to customize sampling
  - Augmentation
  - Block update
  - Reparametrize
  - Hybrid
  - Griddy-Gibbs

McMC

- Want to sample a target distribution **f(x)**
- We construct a Markov chain which converges towards the target distribution
- Markov chain Monte Carlo
  - Transition density P(**y**|**x**) with f(**x**) as limiting distribution / stationary distribution

$$\lim_{n\to\infty} g_0 P^n = p_{\mathrm{Lim}} = f \qquad f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x}$$

- Today: Convergence,
  - Example of failures
  - How to check

Requirement for convergence

- Markov chain:
  - is Irreducible: you can visit all of parameter space
  - is Aperiodic : you do not go in loop
  - Is Recurrent : you will always return to a set
  - Has the correct stationary distribution

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x}$$

Detailed balance:

$$f(\boldsymbol{y}) P(\boldsymbol{x}|\boldsymbol{y}) = f(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x})$$

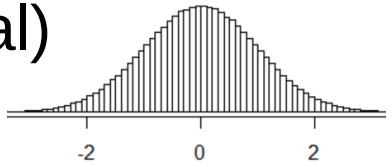
Sufficient for stationary distribution

No guarantee for the other three

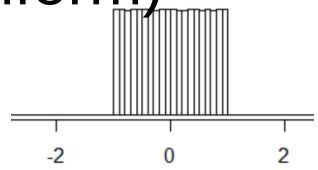
Error in independence sampler

Example 1: Independence sampler:

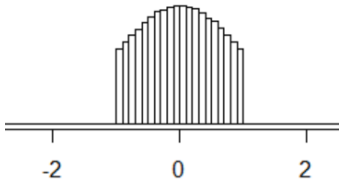
- Target: $f(x) = \phi(x; 0, 1^2)$ (standard normal)



- Proposal: $g(x) = 0.5 \quad \text{for } -1 < x \leq 1$ (uniform)



- Result: $p_L(x) = \frac{\phi(x;0,1^2)}{\Phi(1)-\Phi(-1)}$ for $-1 < x \leq 1$ (truncated)



- Your proposal does not allow you to visit outside the interval: $-1 < x \leq 1$ **irreducible** fail

Gibbs sampler failure, not irreducible

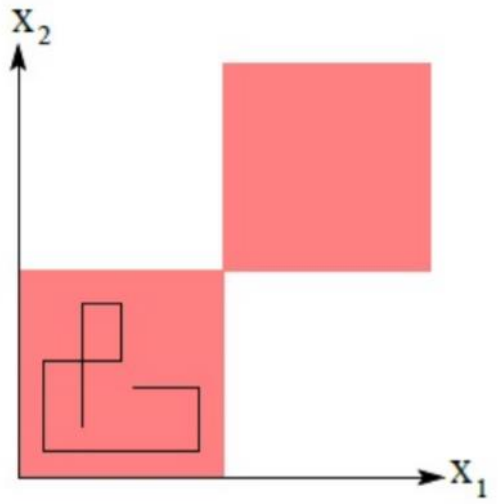
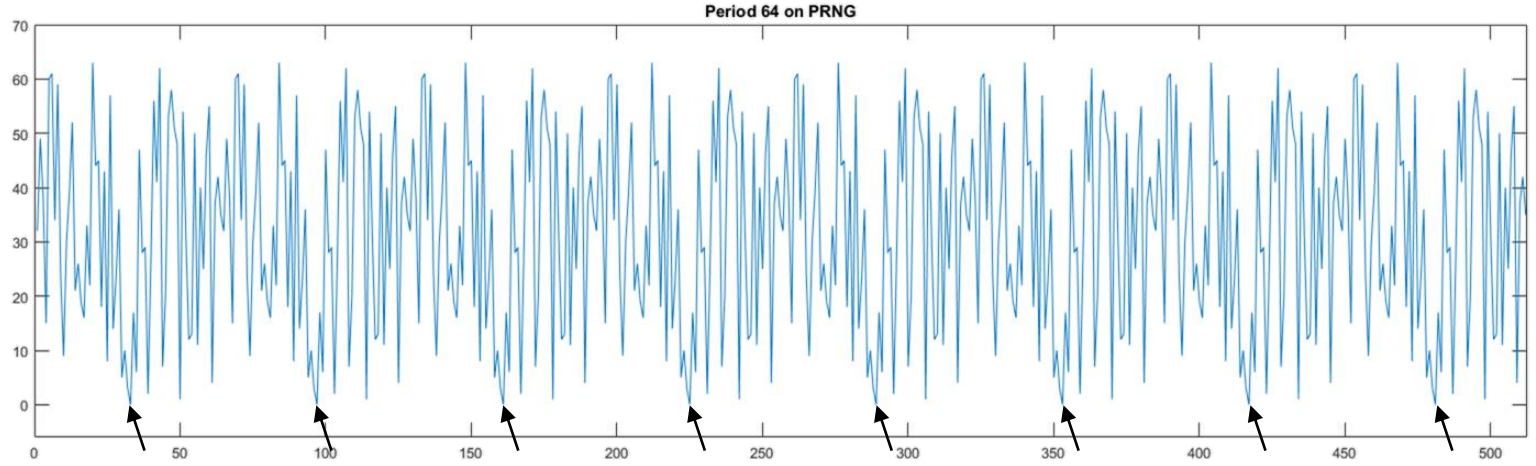


Figure 27.5 (Taken from Barber's *Bayesian Reasoning and Machine Learning*): A two dimensional distribution for which Gibbs sampling fails. The distribution has mass only in the shaded quadrants. Gibbs sampling proceeds from the l^{th} sample state (x_1^l, x_2^l) and then sampling from $p(x_2|x_1^l)$, which we write (x_1^{l+1}, x_2^{l+1}) where $x_1^{l+1} = x_1^l$. One then continues with a sample from $p(x_1|x_2 = x_2^{l+1})$, etc. If we start in the lower left quadrant and proceed this way, the upper right region is never explored.

MCMC and Bayesian Modeling(2017), Martin Haugh Columbia University (under resources on course page)

When does a MCMC fail periodicity?

- Rare in continuous chains, avoided by construction
- PRNG with a short period may cause a periodicity failure
  - $x_{n+1} = (ax_n + c) \bmod m$

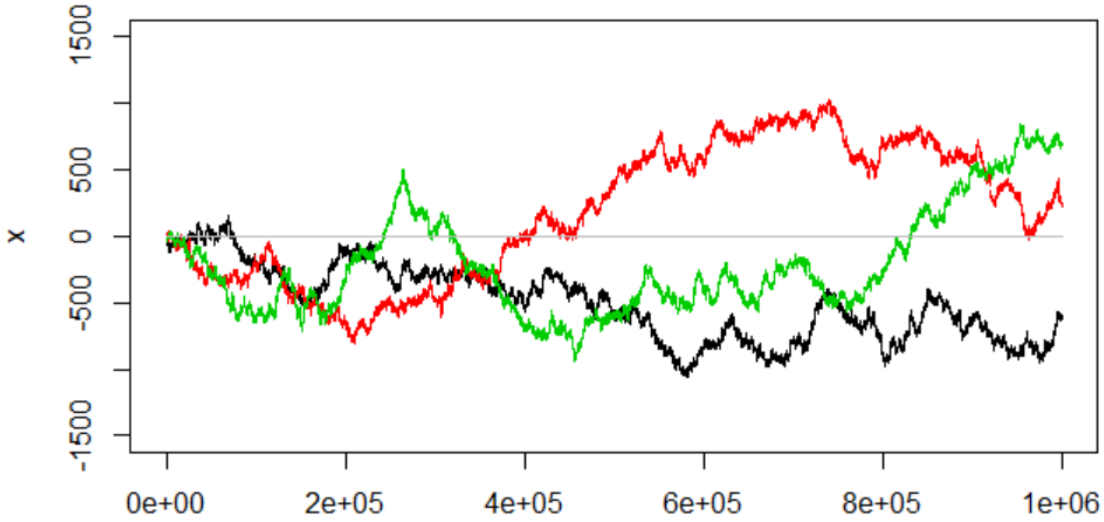


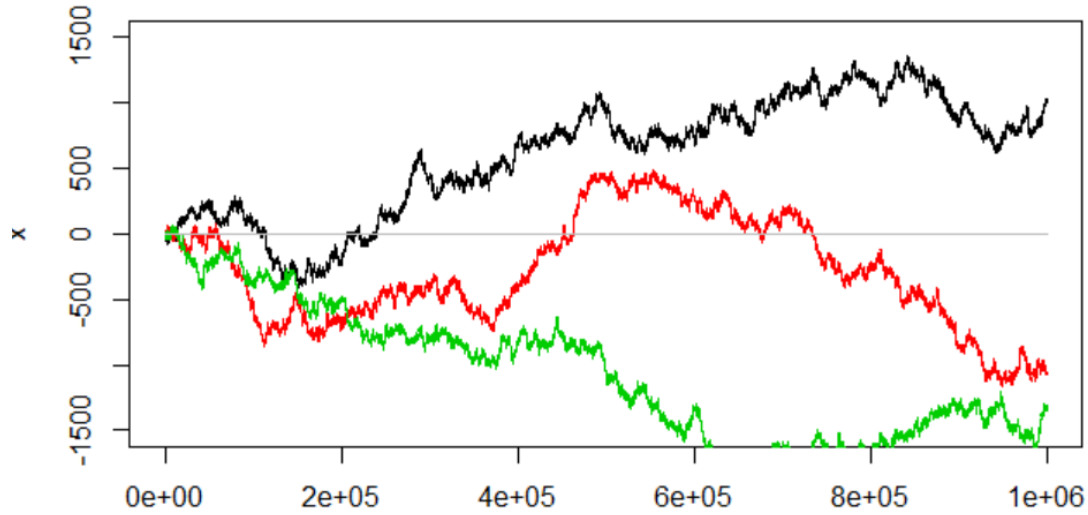
=> Use Mersenne Twister (or another modern PRNG)

Example recurrent fail : improper prior

- $f(\boldsymbol{x}) \propto 1, \boldsymbol{x} = (x_1, x_2, x_3) \in R^3$
- Random walk
- $p(\boldsymbol{x}^*|\boldsymbol{x}) = \phi(x_1^*; x_1, 1) \cdot \phi(x_2^*; x_2, 1) \cdot \phi(x_3^*; x_3, 1)$
- Irreducible? (possible to reach any point with a finite number of steps)
  - Yes, there is a positive probability for any set of non-zero measure in one step.
- Aperiodic?
  - Yes, any non zero set can be reached at any time
- Detailed balance?
  - Yes we have $p(\boldsymbol{x}^*|\boldsymbol{x})f(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{x}^*)f(\boldsymbol{x}^*)$
- So what could go wrong??
  - The chain is not recurrent

Example random walk in R^3





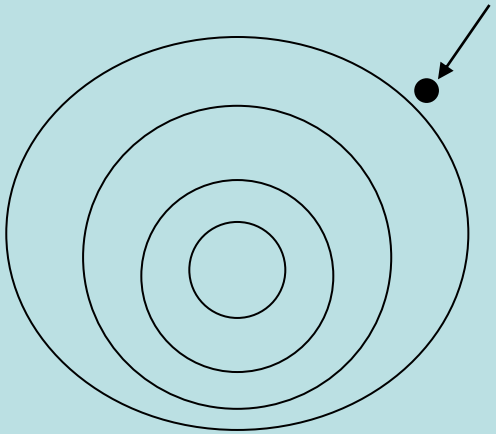
If you get sample paths like these, you might have a recurrence issue

Perhaps your target distribution is not a proper distribution
[not easy to tell upfront]

If you safeguard yourself against zero density regions by setting a minimum density value.
[you get into trouble]

Reccurent fail

- Since we often work with log density a probability of zero causes problems. A quick fix could be to allow the probability to be slightly positive everywhere.
  This is not a good solution
  - Having a small probability for everything gives problems ☹
    => Mc fail to be recurrent



If you get
out here and dimension
is larger than 2 chances
are that you will never
return to «central part»

Example where we it is easy to overlook detailed balance (and it matters)

- Target: $f(x) = 0.5 \quad \text{for } -1 < x \leq 1$ (uniform)
- Proposal: $g(x^*|x) = \phi(x^*; x, \sigma(x)^2)$

$$\sigma(x) = \max(1 - |x|, 0.1)$$

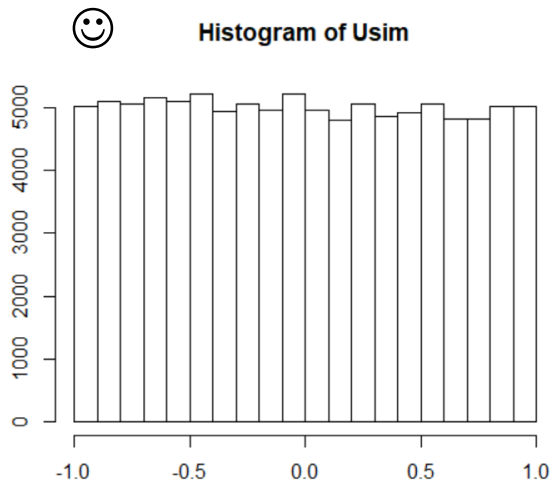
Want to avoid many proposals outside the interval

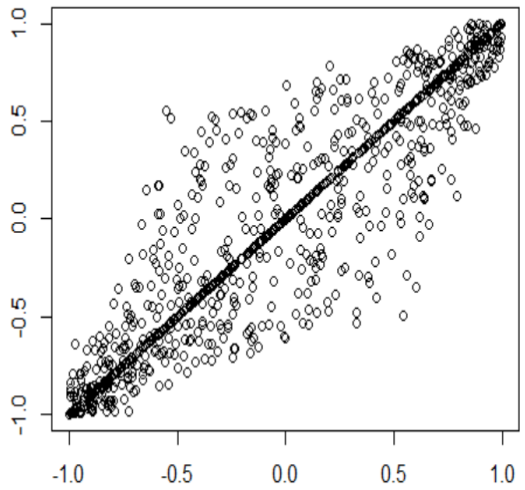
- MH-Ratio:

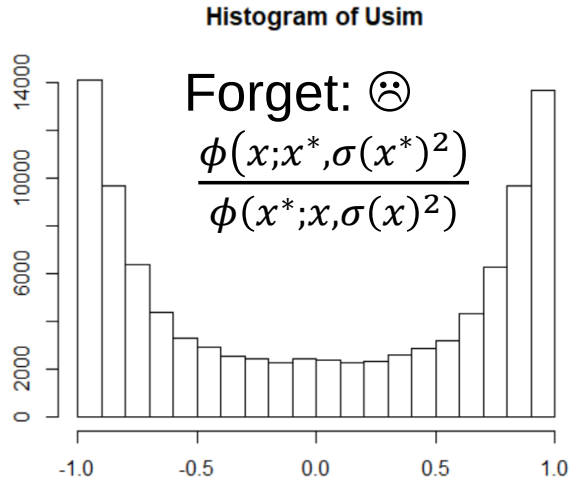
$$R(x^*|x) = \frac{f(x^*)\phi(x; x^*, \sigma(x^*)^2)}{f(x)\phi(x^*; x, \sigma(x)^2)}$$

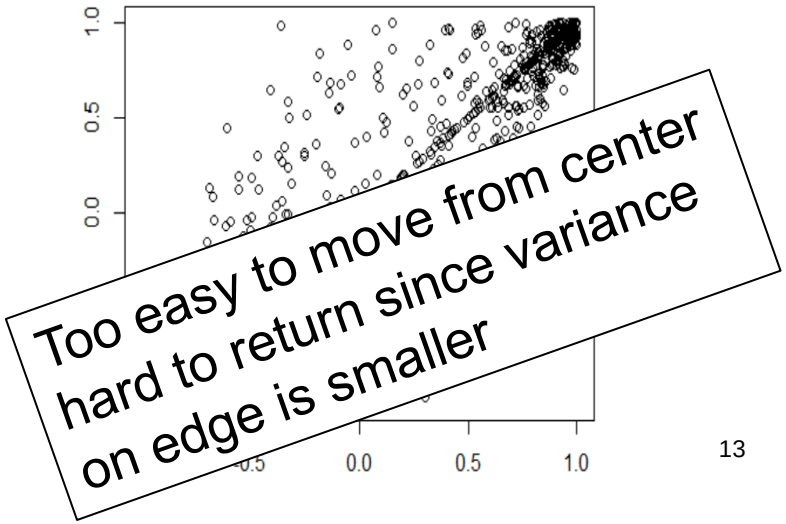
Classic mistake - forget: $\frac{\phi(x;x^*,\sigma(x^*)^2)}{\phi(x^*;x,\sigma(x)^2)}$

Results with and without error:



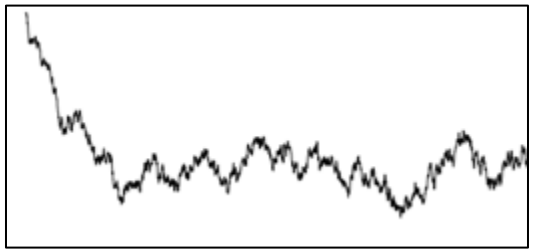




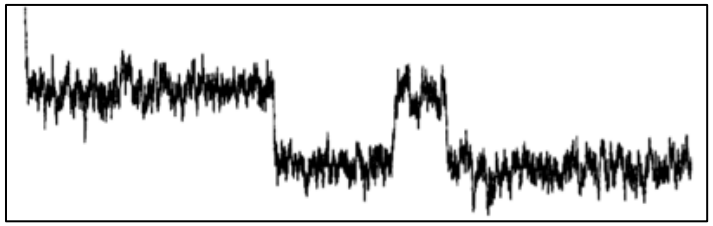


Convergence in practice

- Burn in
  - remove bias due to a bad start
- Mixing
  - Effective number of samples
- Visual
  - sample path
  - cumsum diagnostics
  - Be aware of apparent convergence
- How many chains?
  - at least two in «new territory»
- Diagnostics
  - Gelman-Rubin
- Practical
  - Monte Carlo variance less than 5%
- Acceptance rate
  - Independence sampler high
  - Random walk not too high



Need a check of all model parameters! (and important functions)



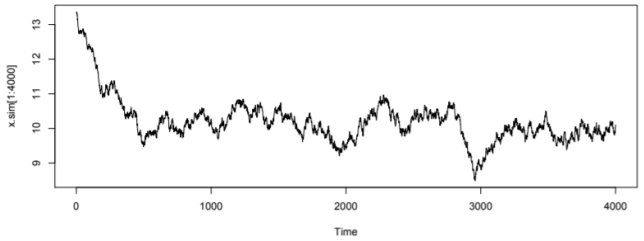
Convergence issues of MCMC

- Theoretical properties:

$$\mathbf{X}^{(t)} \xrightarrow{\mathcal{D}} f(\mathbf{x})$$

$$\hat{\theta}_1 = \frac{1}{L}\sum_{t=1}^{L} h(\mathbf{X}^{(t)}) \to E^f[h(\mathbf{X})]$$

as $t \to \infty$



D is denoted «burn in»

- Note: We also have

$$\hat{\theta}_2 = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)}) \to E^f[h(\mathbf{X})]$$

  - Advantage: Remove those variables with distribution very different from $f(\mathbf{x})$
  - Disadvantage: Need more samples
- Question: How to specify D and L?
  - D: Large enough so that $\mathbf{X}^{(t)} \approx f(\mathbf{x})$ for $t > D$ (bias small)
  - L: Large enough so that $\text{Var}[\hat{\theta}_2]$ is small enough

Effective sample size for MCMC

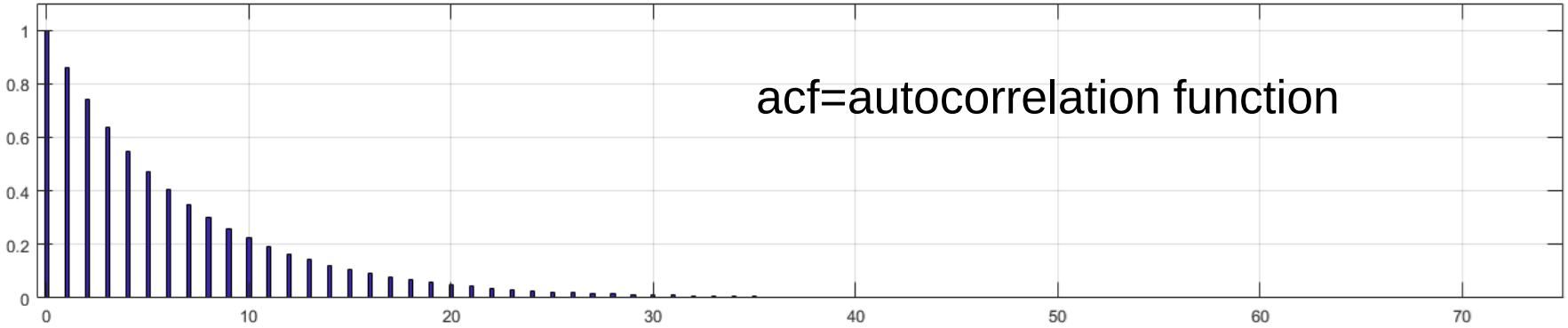
- For $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:

$$\text{Var}[\hat{\theta}] = \frac{\sigma_h^2}{L}\left[1 + 2\sum_{k=1}^{L-1}\frac{L-k}{L}\rho(k)\right] \overset{L\to\infty}{\to} \frac{\sigma_h^2}{L}[1 + 2\sum_{k=1}^{\infty}\rho(k)]$$

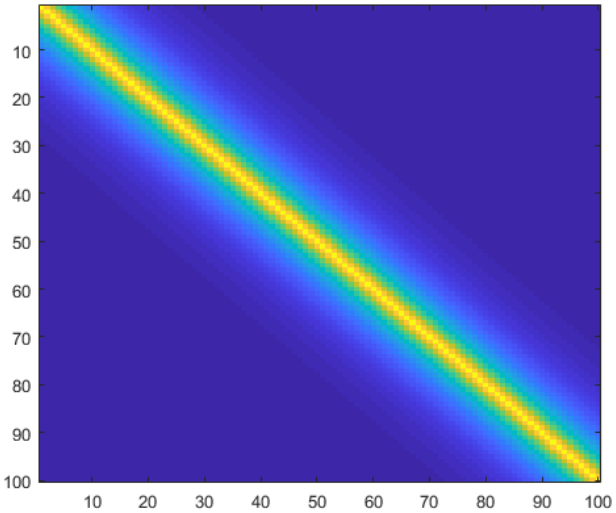
- If independent samples:

$$\text{Var}[\hat{\theta}] = \frac{\sigma_h^2}{L}$$

- Effective sample size: $\frac{L}{1+2\sum_{k=1}^{\infty}\rho(k)}$
- Use empirical estimates $\hat{\rho}(k)$
- Usual to truncate the summation when $\hat{\rho}(k) < 0.1$.



Corresponding covariance matrix: $\boldsymbol{\Sigma}$



$$\mathrm{Var}\left(\frac{1}{L}\sum_{i=1}^{L} x_i\right) = \frac{1}{L^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1} = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\sigma_{ij} = \frac{\sigma^2}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\rho_{ij}$$

$$\sum_{i=1}^{L}\sum_{j=1}^{L}\rho_{ij} = L + 2\sum_{h=1}^{L-1}(L-h)\,\rho(h) \approx L + 2L\sum_{h=1}^{R}\rho(h)$$

$$\mathrm{Var}\left(\frac{1}{L}\sum_{i=1}^{L} x_i\right) = \frac{\sigma^2}{L}\left(1 + 2\sum_{h=1}^{R}\rho(h)\right)$$

$$1 + 2\sum_{h=1}^{R}\rho(h) = 2\left(\sum_{h=0}^{R}\rho(h)\right) - 1$$

Mixing

- For $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:

$$\mathrm{Var}[\hat{\theta}] = \frac{1}{L^2}\left[\sum_{t=D+1}^{D+L} \mathrm{Var}[h(\mathbf{X}^{(t)})] + 2\sum_{s=D+1}^{D+L-1}\sum_{t=s+1}^{D+L} \mathrm{Cov}[h(\mathbf{X}^{(s)}), h(\mathbf{X}^{(t)})]\right]$$

Assume D large, so "converged":

$$\mathrm{Var}[h(\mathbf{X}^{(t)})] \approx \sigma_h^2, \quad \mathrm{Cov}[h(\mathbf{X}^{(s)}), h(\mathbf{X}^{(t)})] \approx \sigma_h^2 \rho(t-s)$$

gives

$$\mathrm{Var}[\hat{\theta}] \approx \frac{1}{L^2}\left[\sum_{t=D+1}^{D+L} \sigma_h^2 + 2\sum_{s=D+1}^{D+L-1}\sum_{t=s+1}^{D+L} \sigma_h^2 \rho(t-s)\right]$$

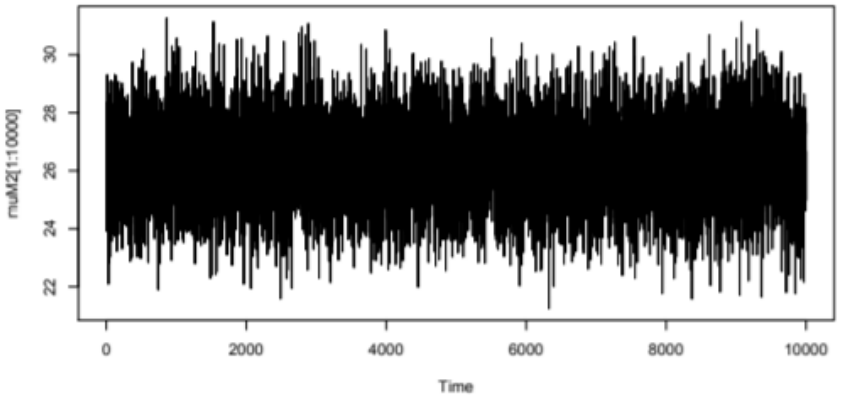
$$= \frac{\sigma_h^2}{L}\left[1 + 2\sum_{k=1}^{L-1} \frac{L-k}{L}\rho(k)\right]$$

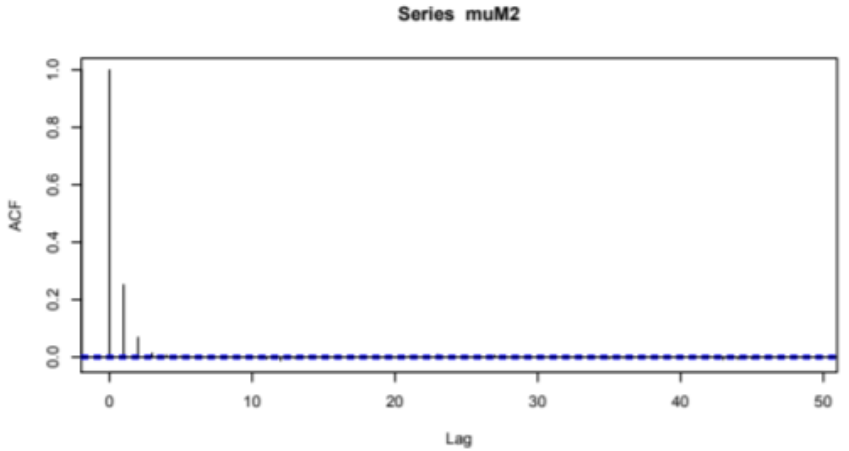
- Good mixing: $\rho(k)$ decreases fast with k!

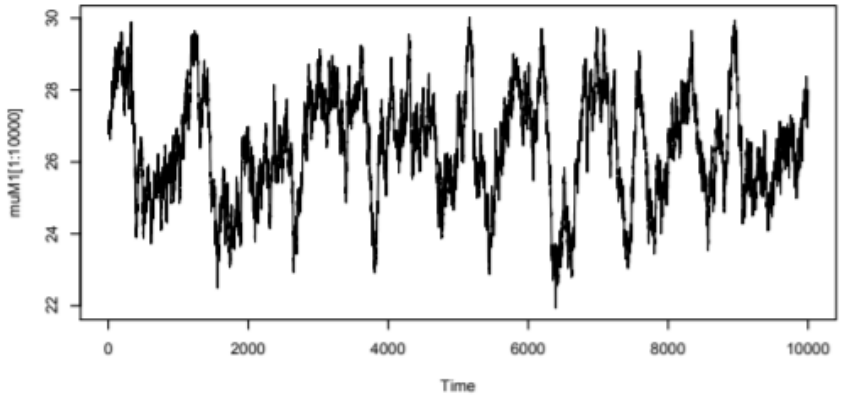
Convergence for all

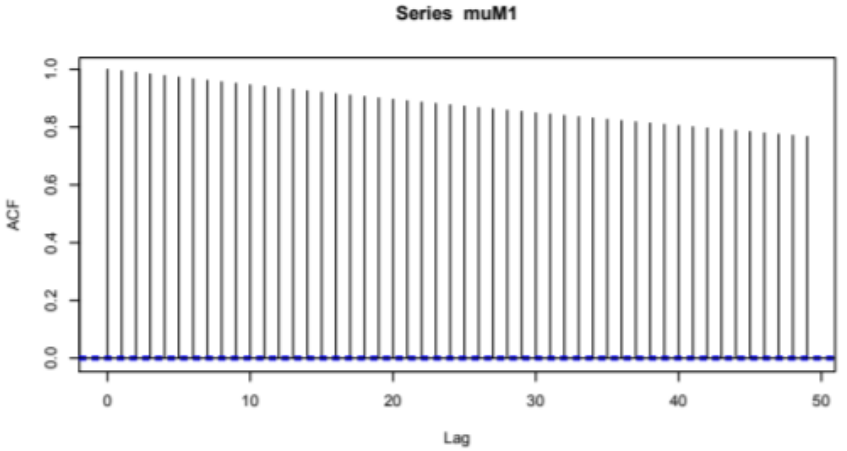
- We can have the situation that: some variables mix well while other have bad mixing
- Need to check
  - Each model parameter
  - Important functions of model parameters
- Consider also
  - Likelihood and/or posterior density (to proportionality)
- Convergence check is essential, the McMC code is not complete without it.

Visual inspection of chain

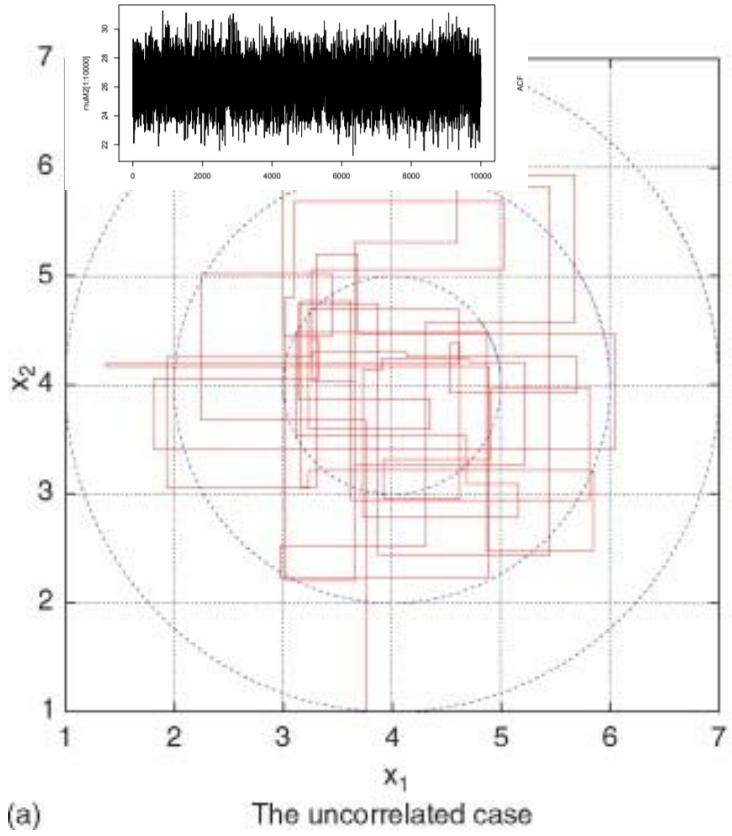


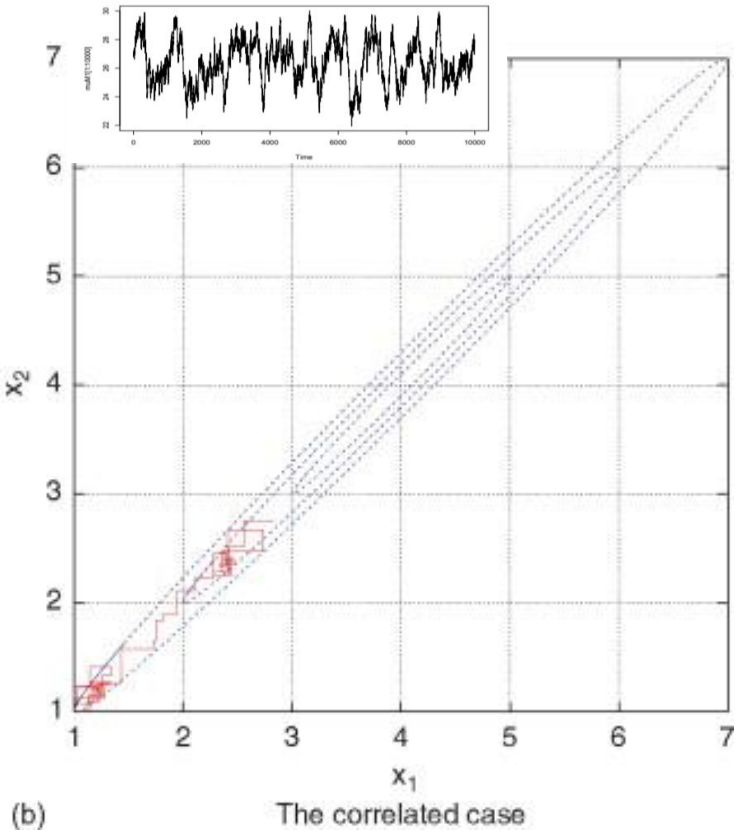






Gibbs sampler for the Bivariate normal distribution





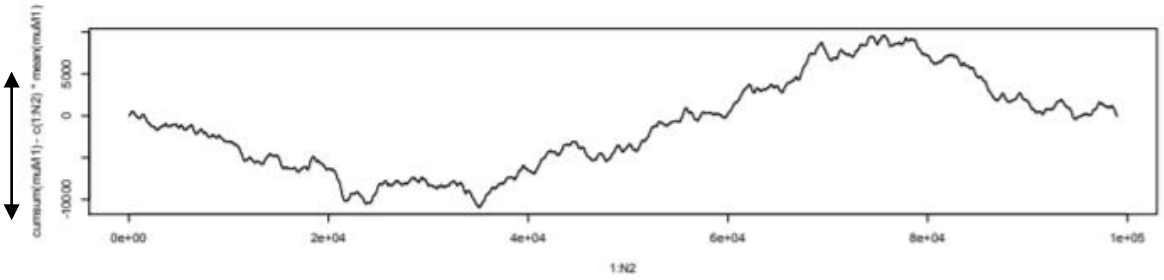
Will work well

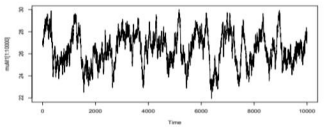
Bad mixing

How to assess convergence

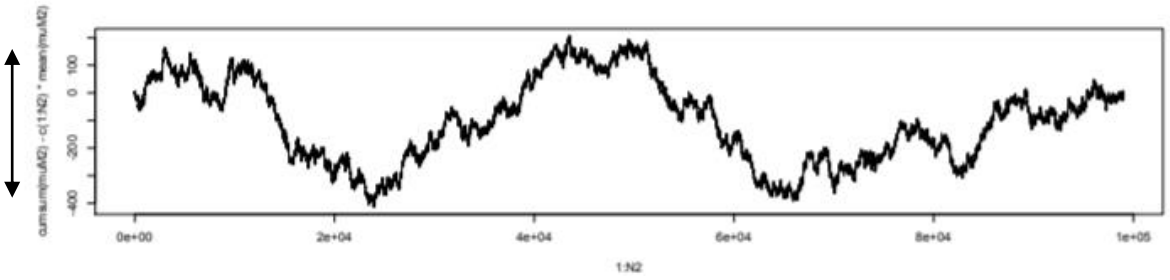
- Graphical diagnostics:
  - Sample paths:
    - Plot $h(\mathbf{X}^{(t)})$ as function of t
    - Useful with different $h(\cdot)$ functions!
  - Cusum diagnostics
    - Plot $\sum_{i=1}^{t}[h(\mathbf{X}^{(i)}) - \hat{\theta}_n]$ versus t
    - Wiggly and small excursions from 0: Indicate chain is mixing well

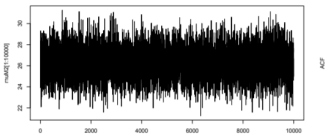
Bad





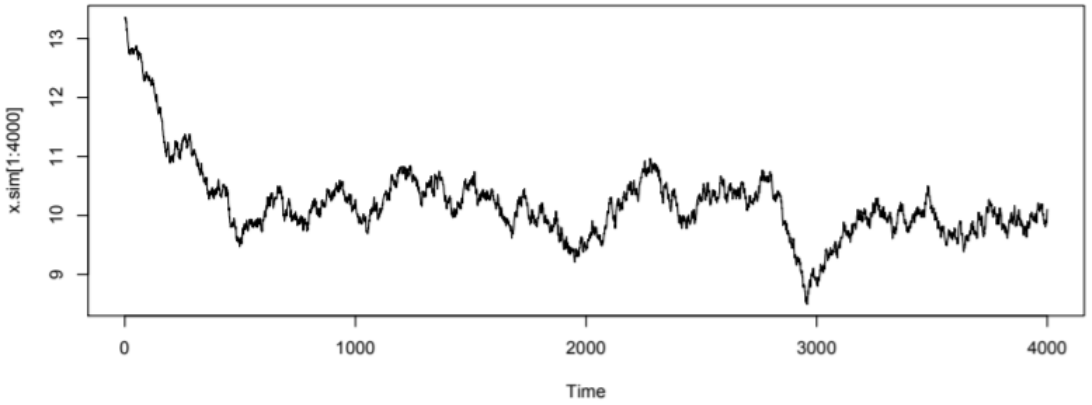
Better



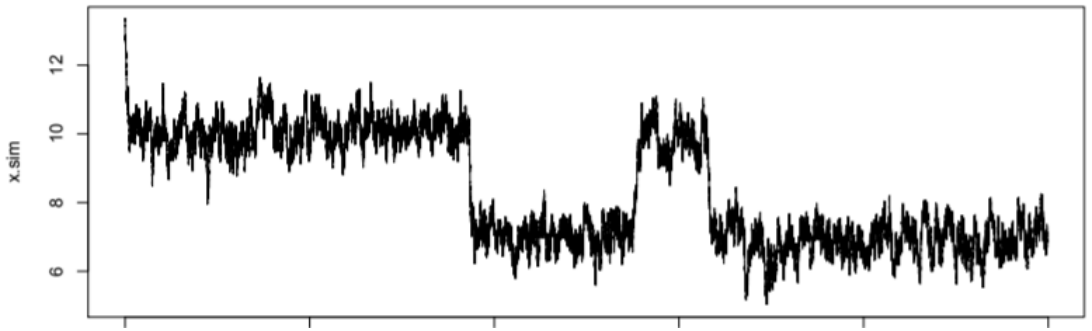


Apparent convergence

- $f(x) = 0.7 \cdot N(7, 0.5^2) + 0.3 \cdot N(10, 0.5^2)$
- Metropolis-Hastings with proposal $N(x^{(t)}, 0.05^2)$
- First 4000 samples (400 discarded)



- Full 10000 samples



Number of chains

- Assume possible to perform N iterations
  - One long chain of length N, or
  - J parallel chains, each of length N/J?
- Burnin:
  - One long chain: Only need to discard D samples
  - Parallel chains: Need to discard $J \cdot D$ samples
- Check of convergence
  - Easier with many parallel chains
- Efficiency
  - Parallel chains give more independent samples
- Computational issues
  - Possible to utilize multiple cores with parallel chains

The Gelman-Rubin diagnostic

- Motivated from analysis of variance
- Assume J chains run in parallel
- jth chain: $x_j^{(D+1)}, \ldots, x_j^{(D+L)}$ (first D discarded)
- Define

$$\bar{x}_j = \frac{1}{L} \sum_{t=D+1}^{D+L} x_j^{(t)} \qquad \bar{x}_\cdot = \frac{1}{J} \sum_{j=1}^{J} \bar{x}_j$$

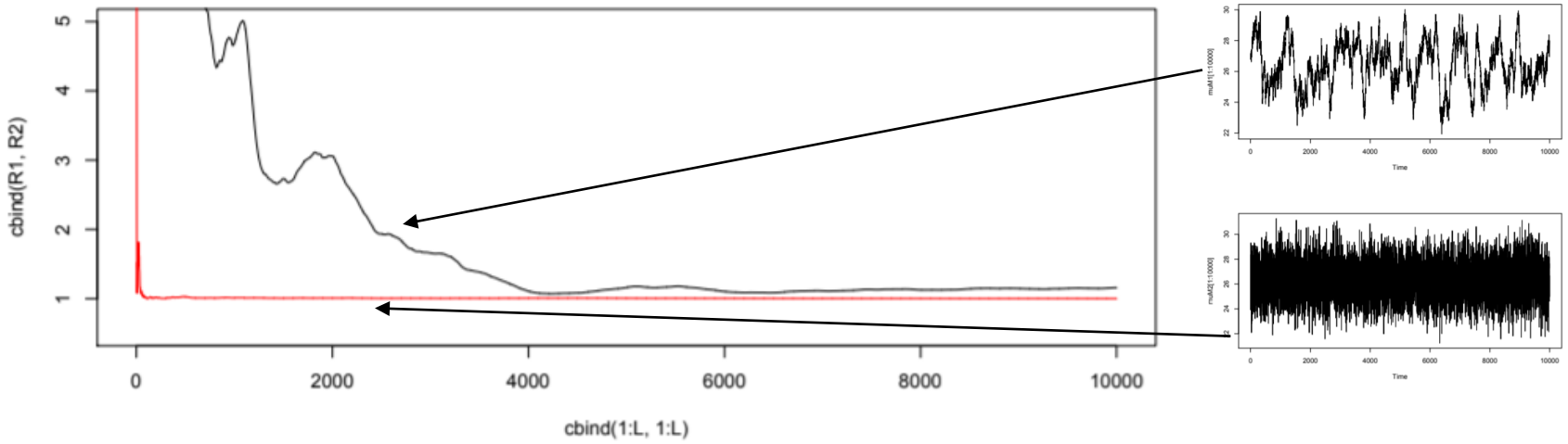
$$B = \frac{L}{J-1} \sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2$$

$$W = \frac{1}{J} \sum_{j=1}^{J} s_j^2 \qquad s_j^2 = \frac{1}{L-1} \sum_{t=D+1}^{D+L} (x_j^{(t)} - \bar{x}_j)^2$$

- If converged, both B and W estimates $\sigma^2 = \mathrm{Var}_f[X]$
- Diagnostic: $R = \frac{\frac{L-1}{L} W + \frac{1}{L} B}{W}$
- "Rule": $\sqrt{R} < 1.1$ indicate D and L are sufficient

Example: Exercise 7.8

- $D = 100, L = 1000$: $\sqrt{R_1} = 1.588$, $\sqrt{R_2} = 1.002$,
- $D = 1000, L = 1000$: $\sqrt{R_1} = 1.700$, $\sqrt{R_2} = 1.004$,
- $D = 1000, L = 10000$: $\sqrt{R_1} = 1.049$, $\sqrt{R_2} = 1.0008$



Data uncertainty and Monte Carlo uncertainty (a practical approach)

- Parameter: $\theta = E^f[h(\mathbf{X})]$
- Estimator: $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:
- Two types of uncertainty
  - Variability in $h(\mathbf{X})$: $\sigma_h^2 = \mathrm{Var}^f[h(\mathbf{X})]$
    - Estimator: $\hat{\sigma}_h^2 = \frac{1}{L}\sum_{t=D+1}^{D+L}[h(\mathbf{X}^{(t)}) - \hat{\theta}]^2$
  - MC variability in $\hat{\theta}$:
    - Estimator: Divide data into batches of size $b = \lfloor L^{1/a} \rfloor$, make estimates $\hat{\theta}$ within each batch and variance from these
- Recommendation: Specify L so that MC variability is less than 5% of variability in $h(\mathbf{X})$.

Practical argument: since the uncertainty in h(X) has a fixed uncertainty (for a give dataset), it makes little sense to estimate the mean with an accuracy several orders below.

Choices

- Gibbs sampler
  - Random or deterministic scan?
    - Deterministic scan most common
    - When high correlation, random scan can be more efficient
- Independence chain:
  - $g(\cdot) \approx f(\cdot)$
  - High acceptance rate
  - Tail properties most important
  - f /g should be bounded
- Random walk proposal
  - Tune variance so that acceptance rate is between 25% and 50%

Convergence?

- Burn in
  - remove bias due to a bad start
- Mixing
  - Effective number of samples
- Visual
  - sample path
  - cumsum diagnostics
  - Be aware of apparent convergence
- One or many chains?
  - at least two in «new territory»
- Diagnostics
  - Gelman-Rubin
- Practical
  - Monte Carlo variance less than 5%

Need a check of all model parameters! (and important functions)