UiO : Matematisk institutt

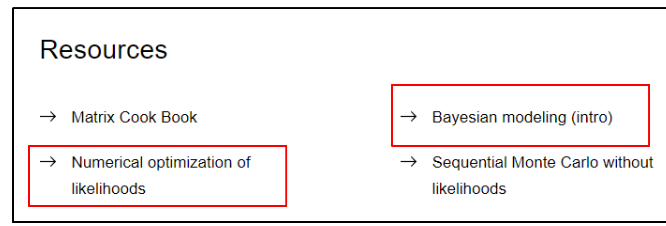
Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2021
Markov Chain Monte Carlo, code examples

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

Summary of talk with reference group

- Syllabus vs prior knowledge
  - «need to read alot outside the course»
  - Topics for background
    - Likelihood and maximization of such (1.3-1.4+)
    - Bayesian statistics (1.5 + note)
    - Markov Chain (1.7 + note)
    - Sufficient statistics [in class]
  - Give heads up for what we need to prepare



- Level of difficulty / work load
  - Course is fast track covering much material
  - Use more than 1/3 of a week on the course

Assignments

- Weekly
  - Useful (in particular for compulsory)
  - Labor intensive
  - Much work / learn a lot
  - Walk through of theory fine
  - Show more code in exercise
- Compulsory
  - Write about delivery on web
  - Project is theoretical/academic e.g. Q4
  - Everything is not clear, the Q& A helps

Lectures

- Improvement of visual aids
  - Videos of algorithms online
  - Better visualization of concepts
  - Suggest You tube videos
- Go through code in examples
- Too much in class
  - Busy slides
  - Much to absorb
  - We need 15 minutes break
  - You often run over time. Use hard stop at 45min
- Questions
  - Repeat and answer it on record
  - Hard to formulate questions with limited time

Adjustments

- Visual aids
  - Feedback taken, I'll see what I can do
  - You tube videos – if you suggest on padlet I can comment on relevance. (I put some out there)
- Code more visible in lecture
  - Go through code for McMC today
  - Go through code for SMC april 15th prior to guest lecture about computational statistics for covid-19
  - Keep this in mind for remaining lectures
- Questions, hope I still get some
- Duration of class 2x45+45

How to work in STK 4051/9051

- Before lecture
  - Read book / note
- After lecture
  - Read book / note [if you did not do it before]
  - go through R-code example
  - do exercises
- After exercise
  - do exercises [if you did not do it before] wrt code, go through R-code provided, make sure that you understand
- Always possible
  - Send mail with questions to me
  - Talk to me - use padlet

Big payoff when doing the compulsory exercise (and for life)

Online study group

- To be arranged
- Details to follow on web

Exam 2021

- Examination
  - See course webpage
- Home examination.
  - **Disclosure of exam assignment:** June 7 at 9:00 AM
  - **Submission deadline:** June 7 at 1:00 PM
- Examination system**:**
  - Inspera – see guides for digital exams
- Previous exams in course:
  - 2019 4 hour written
  - 2020 7(2) days home exam

Relation between formulas and code

- In STK4051/9051 the link between the formulas and code is important
- Derivations of formulas are expected
- The quality of code is less important if the code works. [This is not a course in programming]
- The readability/clarity of the code is important to get credit for effort if the code fail
  [Then I can see what type of error you have done]

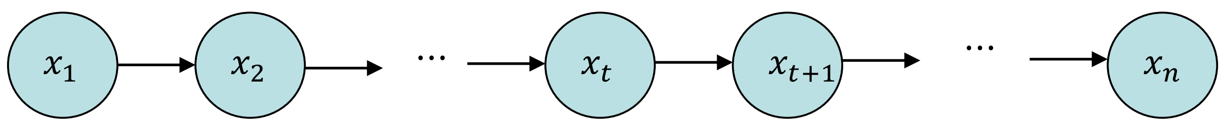
Metropolis Hastings

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

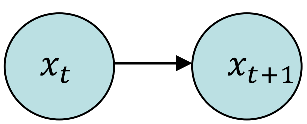
Markov chain

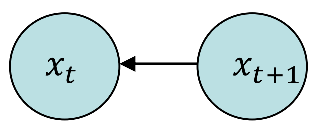
$$P(\boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{x})$$

- Important distributions
  - $f(\boldsymbol{x})$ target distribution (of current value $\boldsymbol{x}$)
  - $P(\boldsymbol{y}|\boldsymbol{x})$ distribution of next value $\boldsymbol{y}$ given current $\boldsymbol{x}$
  - $f(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})$ joint distribution: lag one "scatter"
- Detailed balance: $f(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y})$

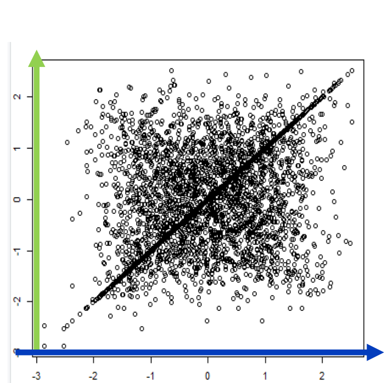


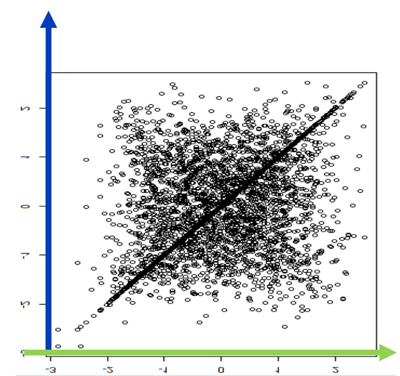
$$f(\boldsymbol{x_t})f(\boldsymbol{x_{t+1}}|\boldsymbol{x_t}) = f(\boldsymbol{x_t}|\boldsymbol{x_{t+1}})\, f(\boldsymbol{x_{t+1}})$$





$$f(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}) = f(\boldsymbol{x_{t+1}}, \boldsymbol{x_t})$$
$$f(\boldsymbol{a}, \boldsymbol{b}) = f(\boldsymbol{b}, \boldsymbol{a})$$





Metropolis-Hastings algorithm

- $P(\mathbf{y}|\mathbf{x})$ defined through an algorithm:
  1. Sample a candidate value $\mathbf{X}^*$ from a proposal distribution $g(\cdot|\mathbf{x})$.
  2. Compute the Metropolis-Hastings ratio
  $$R(\mathbf{x},\mathbf{X}^*) = \frac{f(\mathbf{X}^*)g(\mathbf{x}|\mathbf{X}^*)}{f(\mathbf{x})g(\mathbf{X}^*|\mathbf{x})}$$
  3. Put
  $$\mathbf{Y} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{1, R(\mathbf{x},\mathbf{X}^*)\} \\ \mathbf{x} & \text{otherwise} \end{cases}$$
- For $\mathbf{y} \neq \mathbf{x}$:

$$P(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}|\mathbf{x}) \min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}\right\}$$

Proposal distribution → $g(\mathbf{y}|\mathbf{x})$

Acceptance probability → $\min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}\right\}$

Examples

- Target ditribution $f(\mathbf{x})$ (given)
- Proposal distribution $g(\cdot|\mathbf{x})$ (to be invented)
- Acceptance rate (to be computed) $\min\left\{1, \underbrace{\frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}}_{\text{M-H ratio}}\right\}$
- Specific chains:
  - Independence ex: $g(\boldsymbol{y}|\boldsymbol{x}) = \phi(\boldsymbol{y})$
  - Random walk ex: $g(\boldsymbol{y}|\boldsymbol{x}) = \phi(\boldsymbol{y} - \boldsymbol{x})$
  - Gibbs sampler ex: $g(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{p} f\left(y_j \middle| \boldsymbol{x}_{-j}\right)\delta(\boldsymbol{y}_{-j} = \boldsymbol{x}_{-j})$

this really is $g(\boldsymbol{y}, j|\boldsymbol{x})$

$P(\text{changing index } j)$

Convergence?

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

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

Effective sample size for MCMC

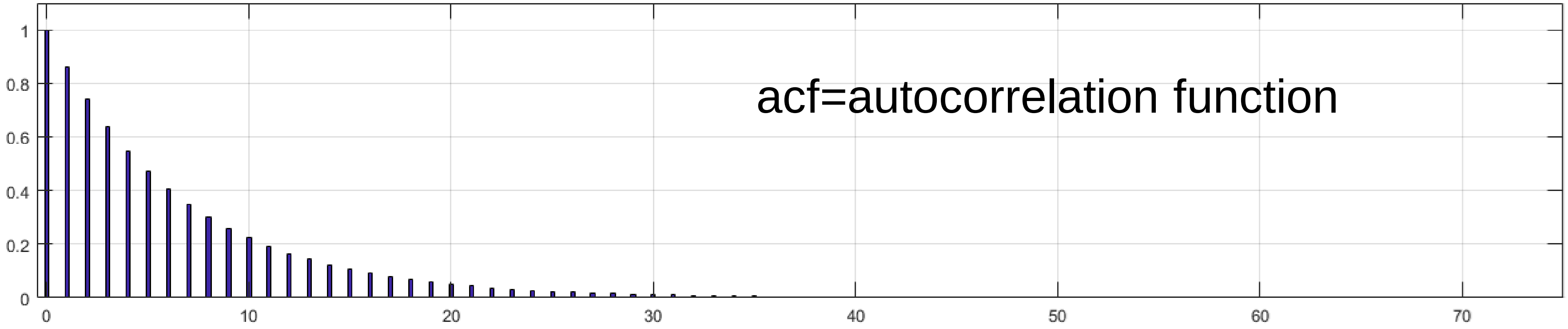
- For $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:

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

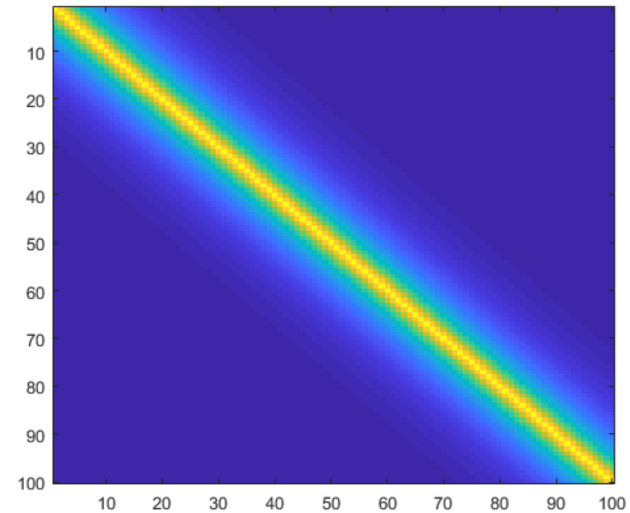
- If independent samples:

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

- 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: $\boldsymbol{\Sigma}$



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

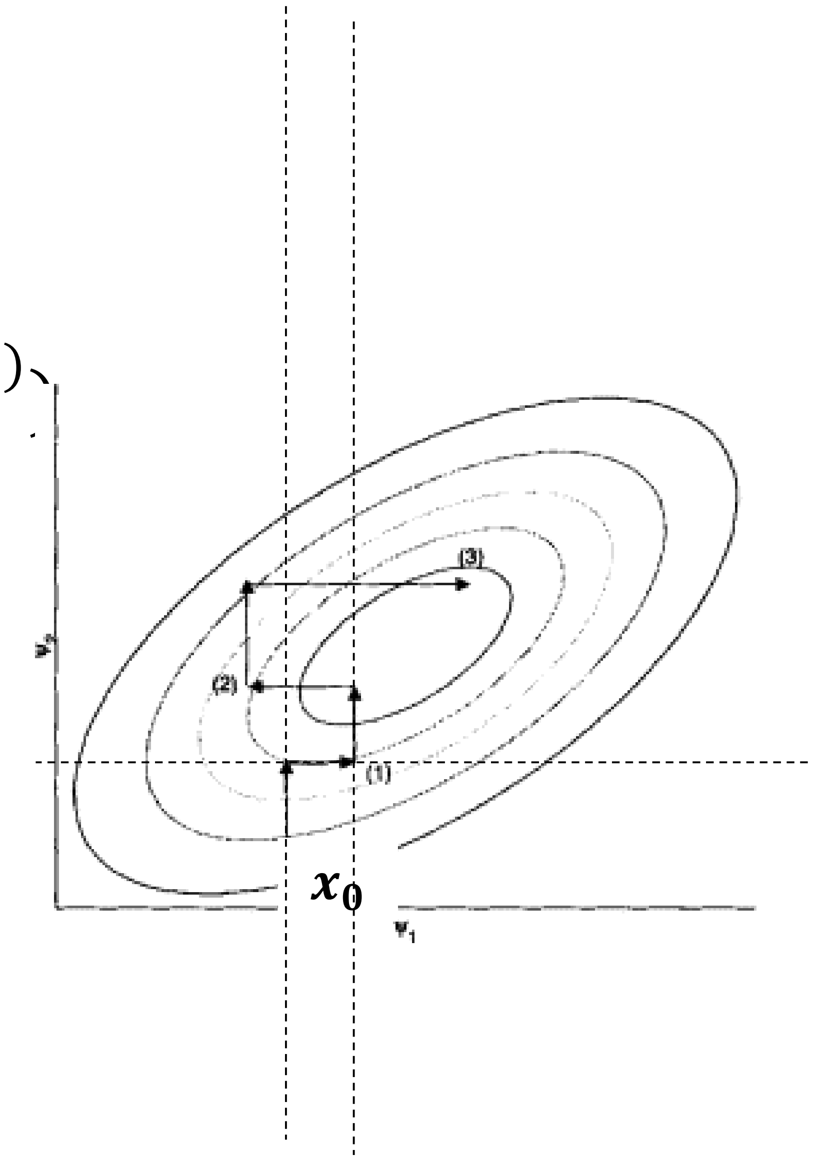
$$\sum_{i=1}^{L}\sum_{j=1}^{L}\rho_{ij} = L + 2\sum_{h=1}^{L-1}(L-h)\,\rho(h) \approx L + 2L\sum_{h=1}^{R}\rho(h)$$

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

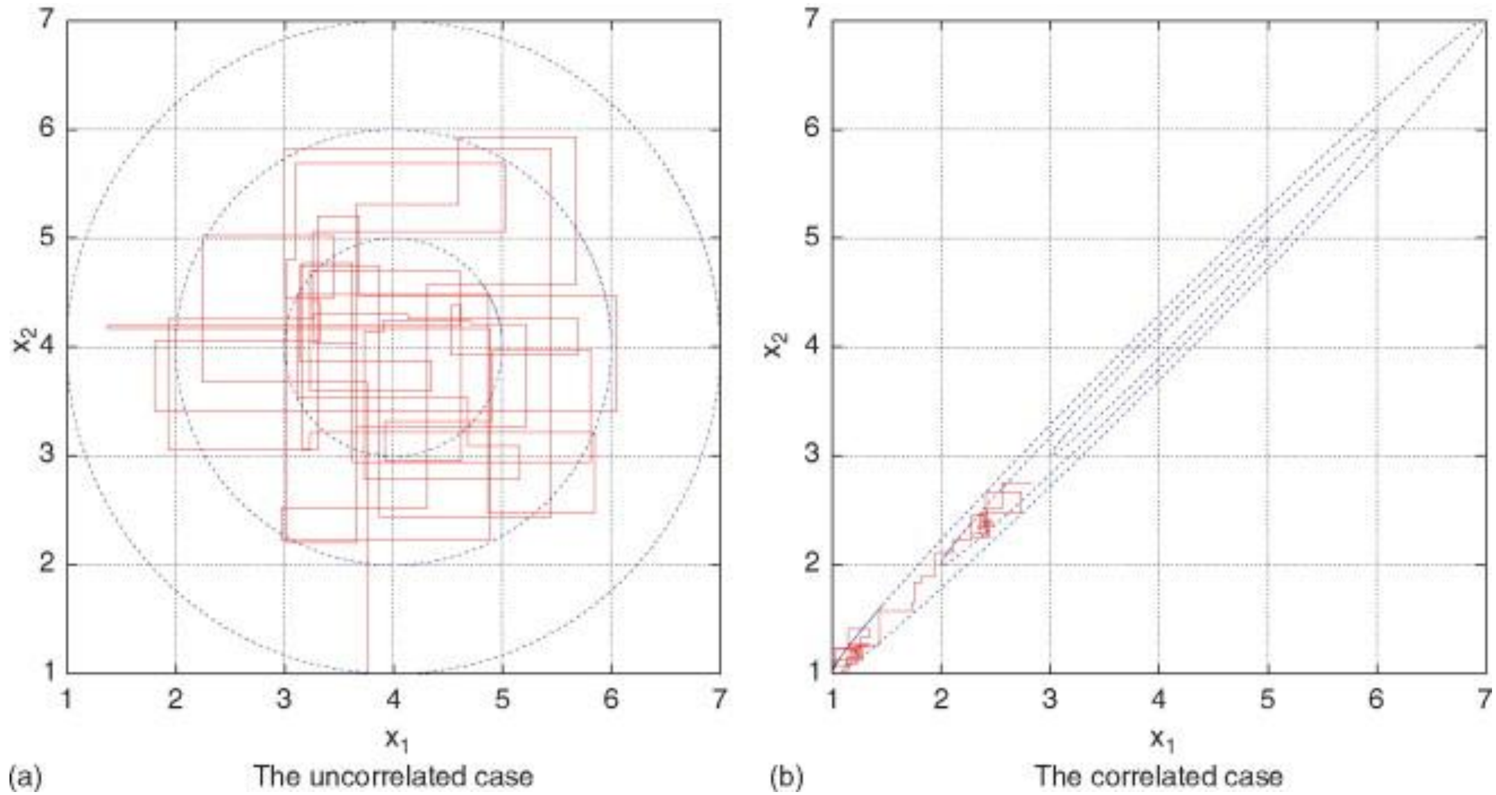
$$1 + 2\sum_{h=1}^{R}\rho(h) = 2\left(\sum_{h=0}^{R}\rho(h)\right) - 1$$

Gibbs sampler (2D)

1. Initiate $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)})$
2. sample
   I. $f(x_1^{(t+1)} | x_2^{(t)})$
   II. $f(x_2^{(t+1)} | x_1^{(t+1)})$
3. Goto 2



Gibbs sampler for the Bivariate normal distribution: (you will have this on exercise)

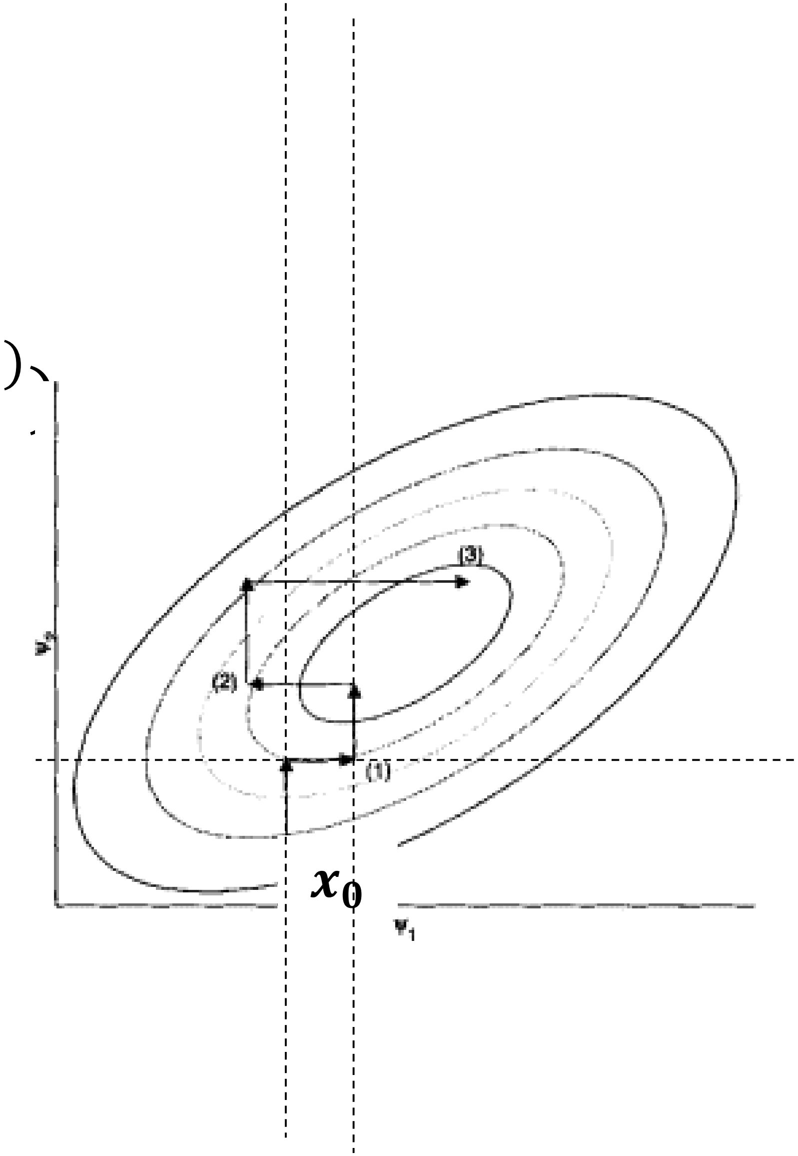


Will work well

Bad mixing

Gibbs sampler

1. Initiate $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)})$
2. sample
   I. $f(x_1^{(t+1)} | x_2^{(t)})$
   II. $f(x_2^{(t+1)} | x_1^{(t+1)})$
3. Goto 2



Main challenge:

Compute $f(x_1^{(t+1)} | x_2^{(t)})$, …

Capture recapture

- Estimate number of pups in a fur seal colony

- Aim: Estimate population size, N, of a species
- Procedure:
  - At time t_1: Catch $c_1 = m_1$ individuals, each with probability α_1.
    Mark and release
  - At time $t_i, i > 1$: Catch c_i individuals, each with probability α_i.
    Count number of newly caught individuals, m_i, mark the unmarked and release all

- Likelihood:

$$L(N,\boldsymbol{\alpha}|\mathbf{c},\mathbf{m}) \propto \binom{N}{\sum_{k=1}^{I} m_k} \prod_{i=1}^{I} \alpha_i^{c_i}(1-\alpha_j)^{N-c_i}$$

- Prior:

$$f(N) \propto 1$$

$$f(\alpha_i|\theta_1,\theta_2) \sim \text{Beta}(\theta_1,\theta_2)$$

$$E(\alpha_i|\theta_1,\theta_2) = \frac{\theta_1}{\theta_1+\theta_2}$$

- Can derive ($r = \sum_{k=1}^{I} m_k$):

$$N|\boldsymbol{\alpha},\mathbf{c},\mathbf{m} \sim r + \texttt{NegBinom}(r+1, 1-\prod_{i=1}^{I}(1-\alpha_i))$$

$$\alpha_i|N,\boldsymbol{\alpha}_{-i},\mathbf{c},\mathbf{m} \sim \text{Beta}(c_i+\theta_1, N-c_i+\theta_2)$$

- `Example_7_6.R`

Capture-recapture - extended approach

- Assume now a prior $f(\theta_1, \theta_2) \propto \exp\{-(\theta_1 + \theta_2)/1000\}$
- Conditional distributions:

$$N|\cdot \sim r + \texttt{NegBinom}(r + 1, 1 - \prod_{i=1}^{I}(1 - \alpha_i))$$

$$\alpha_i|\cdot \sim \text{Beta}(c_i + \theta_1, N - c_i + \theta_2)$$

$$(\theta_1, \theta_2)|\cdot \sim \underbrace{k \left[\frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}\right]^I \prod_{i=1}^{I} \alpha_i^{\theta_1}(1 - \alpha_i)^{\theta_2} \exp\left\{-\frac{\theta_1 + \theta_2}{1000}\right\}}_{\text{Sample using M.H}}$$

- `Example_7_7.R`

= Hybrid Gibbs sampler

Variable augmentation, mixture distribution

- Mixture distribution

$$Y \sim f(y) = \delta\phi(y, \mu_0, 0.5) + (1-\delta)\phi(y, \mu_1, 0.5), \quad \mu_0 = 7, \mu_1 = 10$$

- Prior $\delta \sim \text{Uniform}[0, 1]$
- Aim: Simulate $\delta \sim p(\delta | y_1, ..., y_n)$

$$p(\delta | y_1, ..., y_n) \propto \prod_{i=1}^{n} [\delta\phi(y_i, 7, 0.5) + (1-\delta)\phi(y_i, 10, 0.5)]$$

Difficult to simulate from directly

- Note, can write model for Y by

$$\begin{aligned} \Pr(Z = z) &= \delta^{1-z}(1-\delta)^z, & z &= 0, 1 \\ Y | Z = z &\sim \phi(y, \mu_z, 0.5), & \mu_0 &= 7, \mu_1 = 10 \end{aligned}$$

$$p(\delta | y_1, ..., y_n, z_1, ..., z_n) \quad \propto \text{Beta}(\delta, n - \sum_{i=1}^{n} z_i + 1, \sum_{i=1}^{n} z_i + 1)$$

$$\Pr(Z_i = z_i | \delta, y_i) \propto \delta^{1-z_i}(1-\delta)^{z_i}\phi(y_i, \mu_{z_i}, 0.5)$$

$$\propto \begin{cases} \frac{\delta\phi(y_i, \mu_0, 0.5)}{\delta\phi(y_i, \mu_0, 0.5) + (1-\delta)\phi(y_i, \mu_1, 0.5)} & z_i = 0 \\ \frac{(1-\delta)\phi(y_i, \mu_1, 0.5)}{\delta\phi(y_i, \mu_0, 0.5) + (1-\delta)\phi(y_i, \mu_1, 0.5)} & z_i = 1 \end{cases}$$

Example mixture distribution cont…

- Aim: Simulate $\delta \sim p(\delta|y_1, \dots, y_n)$
- Approach: Simulate from $p(\delta, \mathbf{Z}|y_1, \dots, y_n)$
- Gibbs sampling
  1. Initialize $\delta^{(0)}$, set $t = 0$
  2. Simulate $\mathbf{Z}^{(t+1)} \sim p(\mathbf{z}|\delta^{(t)}, \mathbf{y})$
  3. Simulate $\delta^{(t+1)} \sim p(\delta|\mathbf{z}^{(t+1)}, \mathbf{y})$
  4. Increment t and go to step 2.

$$p(\delta|y_1, \dots, y_n, z_1, \dots, z_n) \quad \propto \text{Beta}(\delta, n - \sum_{i=1}^n z_i + 1, \sum_{i=1}^n z_i + 1)$$

$$\Pr(Z_i = z_i|\delta, y_i) \propto \delta^{1-z_i}(1-\delta)^{z_i}\phi(y_i, \mu_{z_i}, 0.5)$$

$$\propto \begin{cases} \frac{\delta\phi(y_i,\mu_0,0.5)}{\delta\phi(y_i,\mu_0,0.5)+(1-\delta)\phi(y_i,\mu_1,0.5)} & z_i = 0 \\ \frac{(1-\delta)\phi(y_i,\mu_1,0.5)}{\delta\phi(y_i,\mu_0,0.5)+(1-\delta)\phi(y_i,\mu_1,0.5)} & z_i = 1 \end{cases}$$

Marov Chain for McMC need to be

- Irreducible
  Visit any state in a finite number of steps
- Aperiodic
  Not looping into a cycle
- Recurrent
  You will always return

- Satisfy the fixpoint equation

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

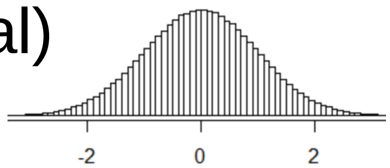
  – Sufficient: Detailed balance

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

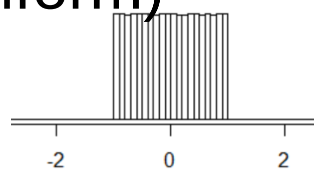
Error in independence sampler

Example 1: Independence sampler:

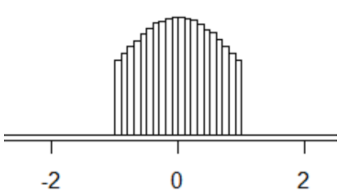
- Target: $f(x) = \phi(x; 0,1^2)$ (standard normal)



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



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



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

Gibbs sampler can fail to be irreducible

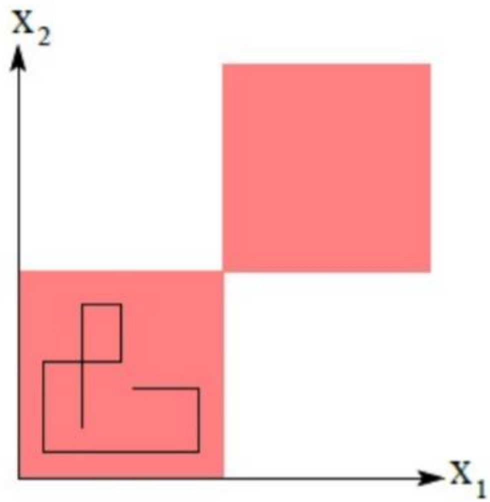


Figure 27.5 (Taken from Barber's *Bayesian Reasoning and Machine Learning*): A two dimensional distribution for which Gibbs sampling fails. The distribution has mass only in the shaded quadrants. Gibbs sampling proceeds from the l^{th} sample state (x_1^l, x_2^l) and then sampling from $p(x_2|x_1^l)$, 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)

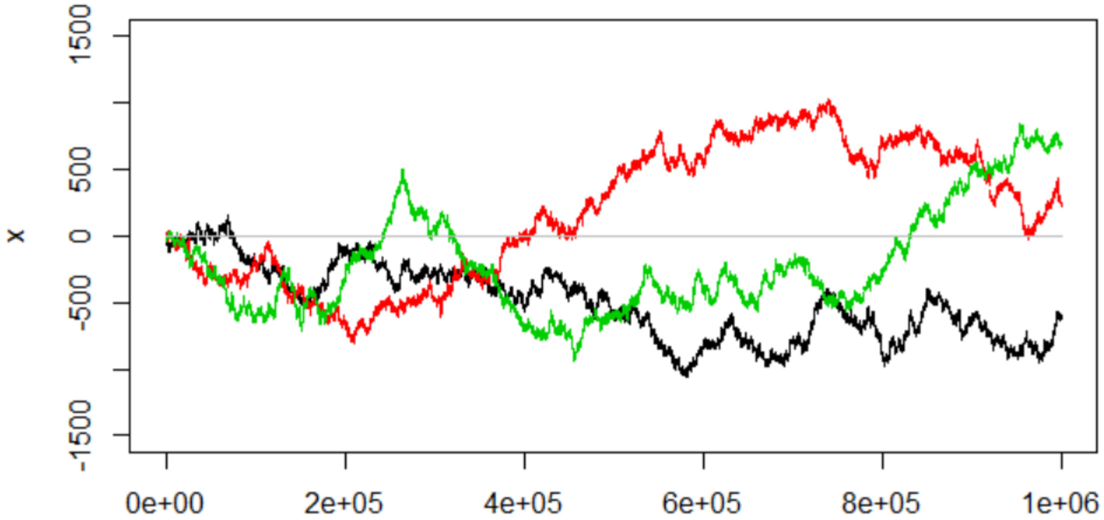
When does a MCMC fail periodicity?

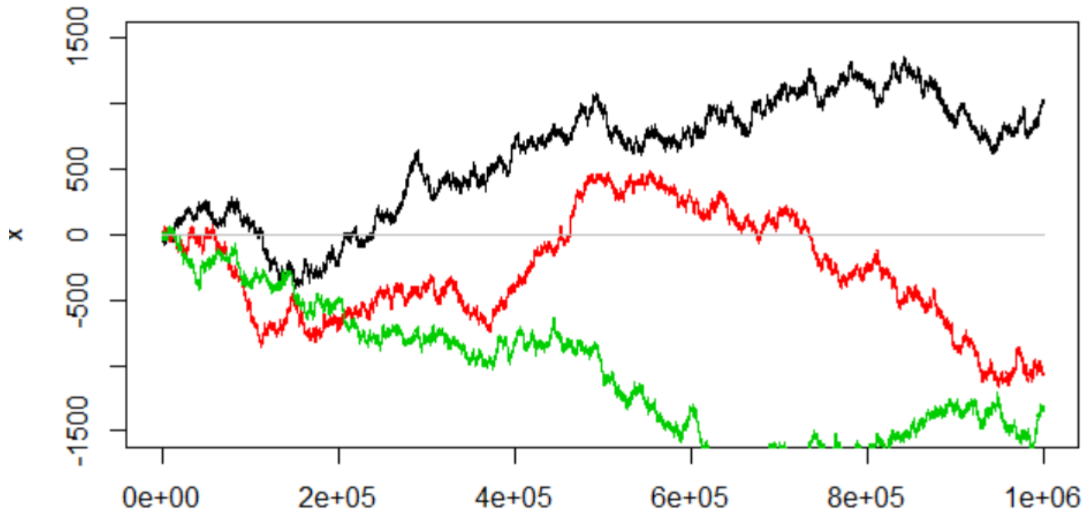
- Rare in continious chains, avoided by construction
- PRNG with a short period might give you a similar type of failure. Use Mersenne Twister

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





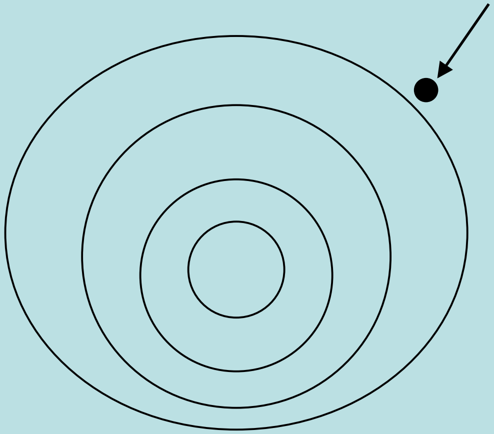
If you get sample paths like these, you might have a recurrence issue

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

If you safe guard 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

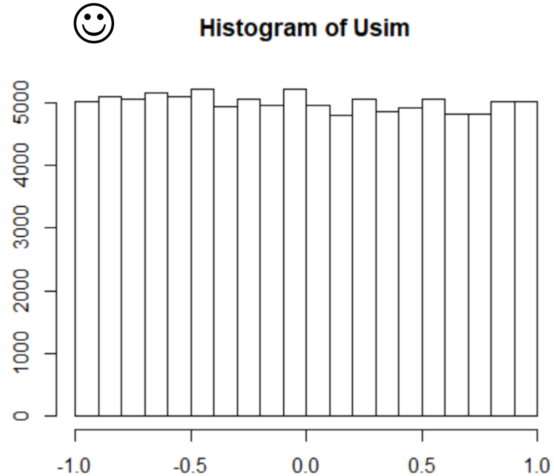
- MH-Ratio:

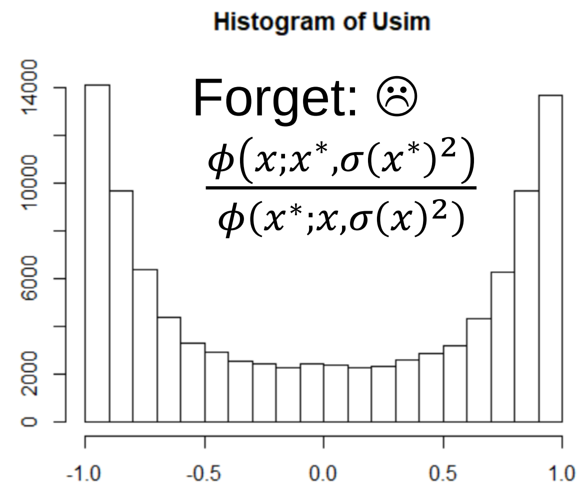
$$R(x^*|x) = \frac{f(x^*)\phi(x; x^*, \sigma(x^*)^2)}{f(x)\phi(x^*; x, \sigma(x)^2)}$$

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

Results with and without error:

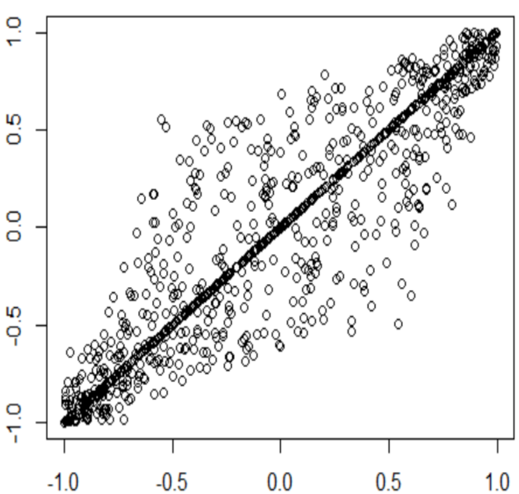
☺

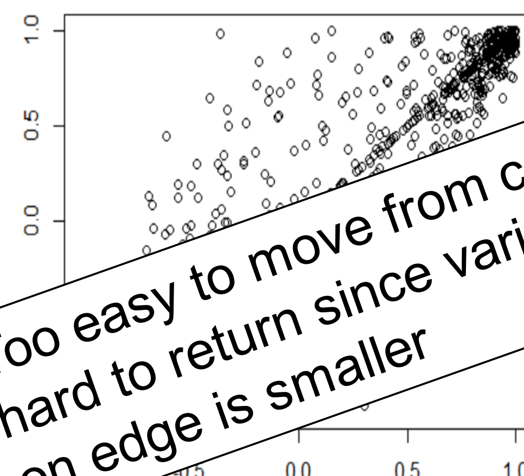




Forget: ☹

$$\frac{\phi(x;x^*,\sigma(x^*)^2)}{\phi(x^*;x,\sigma(x)^2)}$$





Too easy to move from center
hard to return since variance
on edge is smaller