Lectures 9 & 10: Combining Kernels, Convergence Diagnostics

Nick Whiteley

Choosing a good proposal distribution

- Ideally: Markov chain with small correlation $\rho(\mathbf{X}^{(t-1)}, \mathbf{X}^{(t)})$ between subsequent values.
  $\rightsquigarrow$ fast exploration of the support of the target f.
- Two sources for this correlation:
  - the correlation between the current state $\mathbf{X}^{(t-1)}$ and the newly proposed value $\mathbf{X} \sim q(\cdot|\mathbf{X}^{(t-1)})$
    (can be reduced using a proposal with high variance)
  - the correlation introduced by retaining a value $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$ because the newly generated value $\mathbf{X}$ has been rejected
    (can be reduced using a proposal with small variance)
- Trade-off for finding the ideal compromise between:
  - fast exploration of the space (good mixing behaviour)
  - obtaining a large probability of acceptance
- For multivariate distributions: covariance of proposal should reflect the covariance structure of the target.

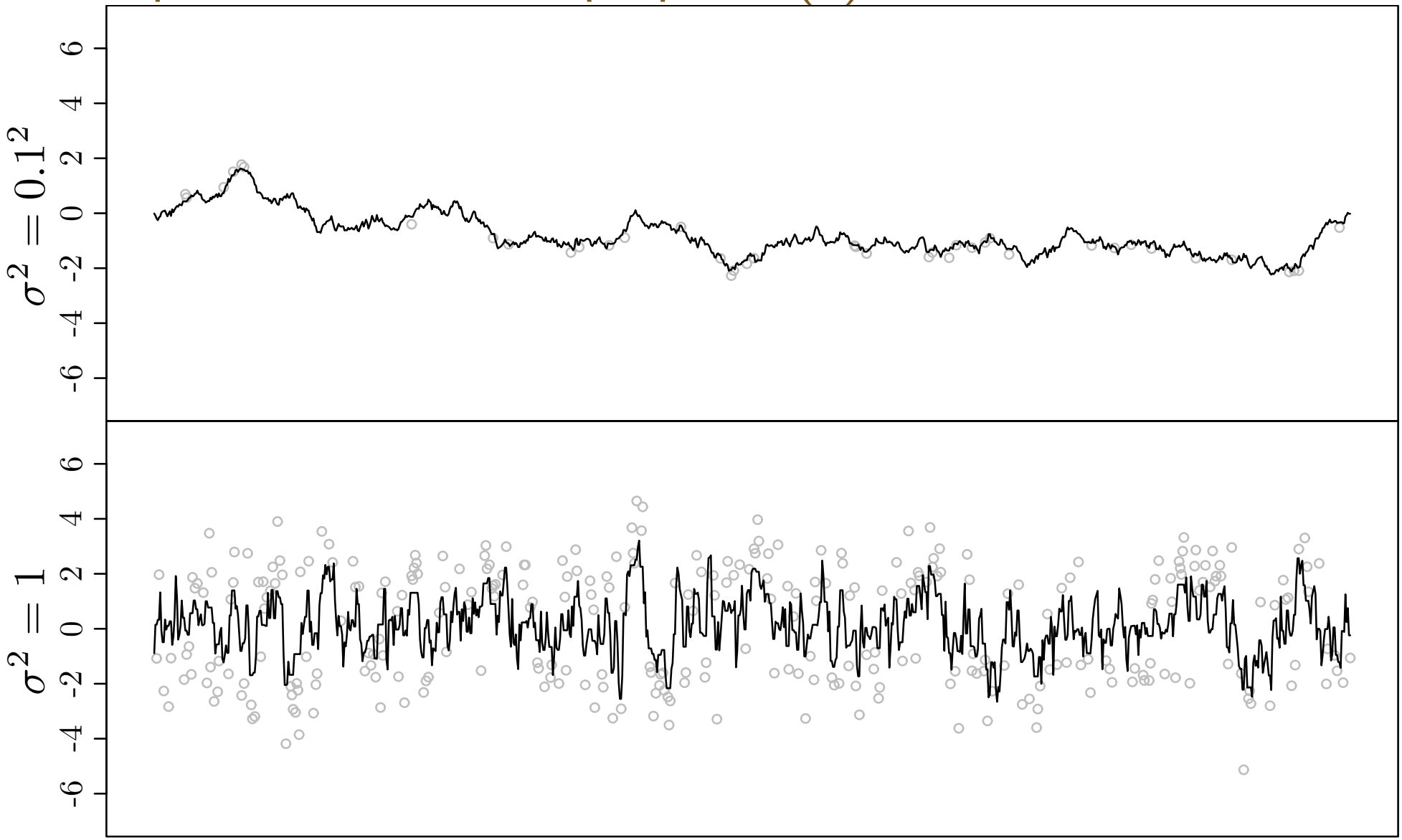
Example 5.3: Choice of proposal (1)

- Target distribution, we want to sample from: $\mathsf{N}(0,1)$ (i.e. $f(\cdot) = \phi_{(0,1)}(\cdot)$)
- We want to use a random walk Metropolis algorithm with

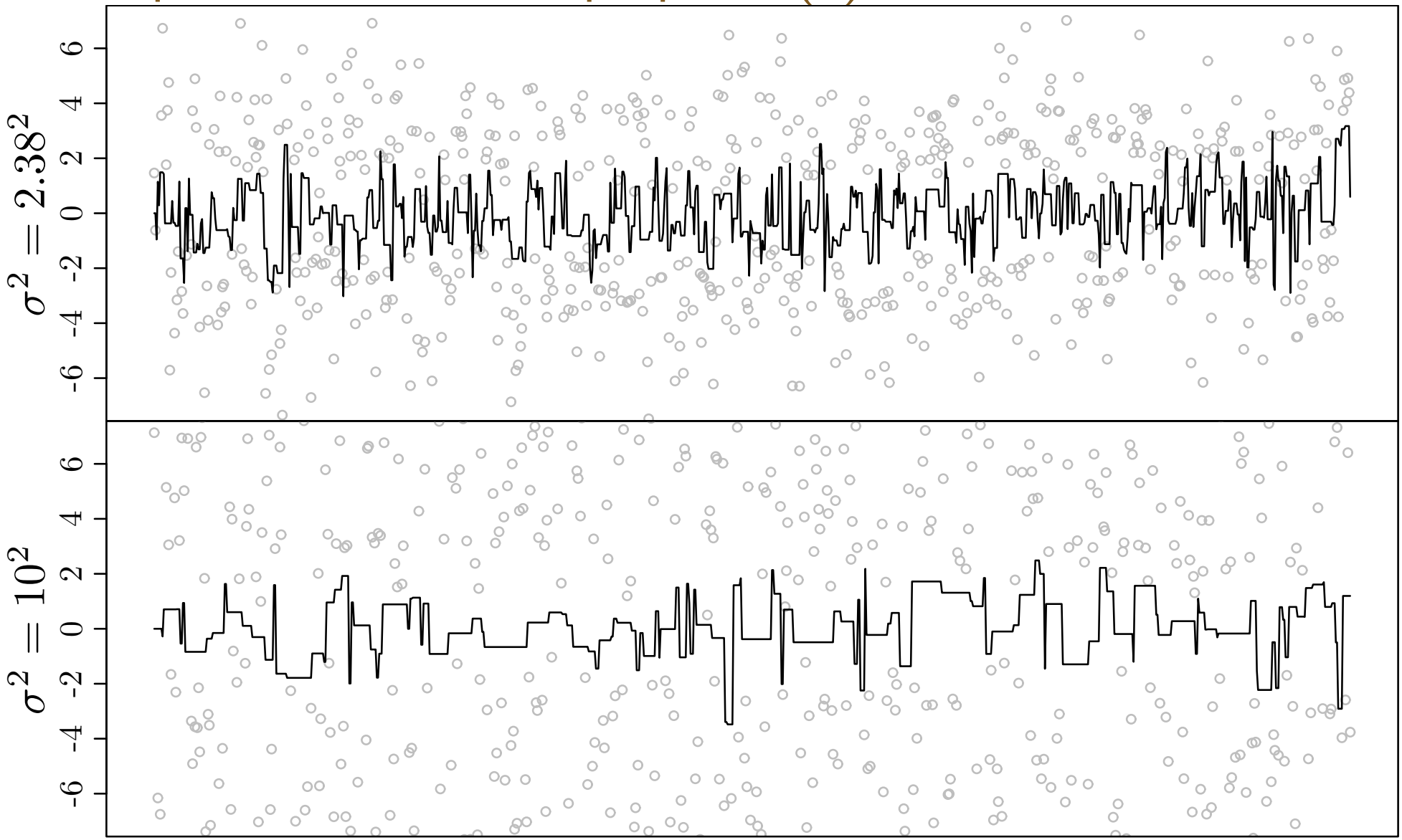
$$\varepsilon \sim \mathsf{N}(0, \sigma^2)$$

- What is the optimal choice of σ^2?
- We consider four choices $\sigma^2 = 0.1^2, 1, 2.38^2, 10^2$.

Example 5.3: Choice of proposal (2)



Example 5.3: Choice of proposal (3)



Example 5.3: Choice of proposal (4)

	Autocorrelation $\rho(X^{(t-1)}, X^{(t)})$		Probability of acceptance $\alpha(X, X^{(t-1)})$	
	Mean	95% CI	Mean	95% CI
$\sigma^2 = 0.1^2$	0.9901	(0.9891,0.9910)	0.9694	(0.9677,0.9710)
$\sigma^2 = 1$	0.7733	(0.7676,0.7791)	0.7038	(0.7014,0.7061)
$\sigma^2 = 2.38^2$	0.6225	(0.6162,0.6289)	0.4426	(0.4401,0.4452)
$\sigma^2 = 10^2$	0.8360	(0.8303,0.8418)	0.1255	(0.1237,0.1274)

Suggests: Optimal choice is $2.38^2 > 1$.

Example 5.4: Bayesian probit model (revisited)

- So far we used: $\mathrm{Var}(\boldsymbol{\epsilon}) = 0.08 \cdot \mathbb{I}$.
- Better choice: Let $\mathrm{Var}(\boldsymbol{\epsilon})$ reflect the covariance structure
- Frequentist asymptotic theory: $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathrm{m.l.e}}) = (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$
  $\mathbf{D}$ is a suitable diagonal matrix
- Better choice: $\mathrm{Var}(\boldsymbol{\epsilon}) = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$
- Increases rate of acceptance from 13.9% to 20.0% and reduces autocorrelation:

$\boldsymbol{\Sigma} = 0.08 \cdot \mathbf{I}$	β_0	β_1	β_2	β_3
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.9496	0.9503	0.9562	0.9532
$\boldsymbol{\Sigma} = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$	β_0	β_1	β_2	β_3
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.8726	0.8765	0.8741	0.8792

(in this example $\det(0.08 \cdot \mathbb{I}) = \det(2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1})$)

5.5 Composing kernels: Mixtures and Cycles

Composing kernels: Idea

- MCMC algorithm (Gibbs sampler, Metropolis-Hastings) can be uniquely identified by the transition kernel.
- So far: only one type of update in the Metropolis-Hastings algorithm.
- Question: Can we combine different MCMC updates?
- Assume:
  - r possible MCMC updates characterised by kernels $K^{(\rho)}(\cdot, \cdot)$.
  - f is the invariant distribution of each kernel $K^{(\rho)}$.
- Two possibilities of combining the r MCMC updates:

  Cycle Perform the MCMC update in a deterministic order.

  Mixture Pick an MCMC update at random.

Cycles

Cycle of MCMC updates $K^{(1)}, \ldots, K^{(r)}$

Starting with $\mathbf{X}^{(0)}$ iterate for $t = 1, 2, \ldots$

1. Set $\boldsymbol{\xi}^{(t,0)} := \mathbf{X}^{(t-1)}$.
2. For $\rho = 1, \ldots, r$:

   Obtain $\boldsymbol{\xi}^{(t,\rho)}$ from $\boldsymbol{\xi}^{(t,\rho-1)}$ by performing an MCMC update corresponding to the kernel $K^{(\rho)}$.
3. Set $\mathbf{X}^{(t)} := \boldsymbol{\xi}^{(t,\rho)}$.

- Similar to the (systematic scan) Gibbs sampler.
- Corresponding transition kernel is

$$K^{\circ}(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) = \int \cdots \int K^{(1)}(\mathbf{x}^{(t-1)}, \boldsymbol{\xi}^{(t,1)}) K^{(2)}(\boldsymbol{\xi}^{(t,1)}, \boldsymbol{\xi}^{(t,2)}) \cdots K^{(r)}(\boldsymbol{\xi}^{(t,r-1)}, \mathbf{x}^{(t)}) \, d\boldsymbol{\xi}^{(t,r-1)} \cdots d\boldsymbol{\xi}^{(t,1)}$$

- f is the invariant distribution of K° if f is the invariant distribution of all $K^{(\rho)}$.

Mixtures

Mixture of MCMC updates $K^{(1)}, \ldots, K^{(r)}$

Starting with $\mathbf{X}^{(0)}$ iterate for $t = 1, 2, \ldots$

1. Draw ρ from $\{1, \ldots, k\}$ with probabilities $(w_1, \ldots, w_r)$.
2. Obtain $\mathbf{X}^{(t)}$ from $\mathbf{X}^{(t-1)}$ by performing an MCMC update corresponding to the kernel $K^{(\rho)}$.

- Similar to the random scan Gibbs sampler.
- Corresponding transition kernel is

$$K^{+}(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) = \sum_{\rho=1}^{r} w_\rho K^{(\rho)}(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}).$$

- f is the invariant distribution of K^{+} if f is the invariant distribution of all $K^{(\rho)}$.

Example 5.5: One-at-a-time Metropolis-Hastings: Idea

- Metropolis-Hastings algorithm 5.1 updates all components of $\mathbf{X}^{(t)}$ in a single step.
- Can we update each component $X_j^{(t)}$ separately?
  $\rightsquigarrow$ *one-at-a-time Metropolis-Hastings* algorithm.
- Can be seen as a composition of p transition kernels $K^{(1)}, \ldots, K^{(p)}$.
- Kernel $K^{(j)}$ is a Metropolis-Hastings update of $X_j^{(t)}$.
- Two possibilities of combining the kernels:
  - Cycle ("systematic scan").
  - Mixture ("random scan").

Example 5.5: One-at-a-time MH (cycle, systematic scan)

Starting with $\mathbf{X}^{(0)} = (X_1^{(0)}, \ldots, X_p^{(0)})$ iterate for $t = 1, 2, \ldots$

- Iterate for $j = 1, \ldots, p$
  - i. Draw $X_j \sim q_j(\cdot | X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j^{(t-1)}, \ldots, X_p^{(t-1)})$.
  - ii. Compute

$$\alpha_j = \min\left\{1, \frac{f(X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})}{f(X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})} \cdot \frac{q_j(X_j^{(t-1)} | X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})}{q_j(X_j | X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})}\right\}.$$

  - iii. With probability α_j set $X_j^{(t)} = X_j$, otherwise set $X_j^{(t)} = X_j^{(t-1)}$.

(corresponds to setting $\boldsymbol{\xi}^{(t,j)} = (X_1^{(t)}, \ldots, X_j^{(t)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$)

Example 5.5: One-at-a-time MH (mixture, random scan)

Starting with $\mathbf{X}^{(0)} = (X_1^{(0)}, \ldots, X_p^{(0)})$ iterate

1. Draw an index j from a distribution on $\{1, \ldots, p\}$ (e.g. uniform)
2. Draw $X_j \sim q_j(\cdot | X_1^{(t-1)}, \ldots, X_p^{(t-1)})$.
3. Compute

$$\alpha_j = \min\left\{1, \frac{f(X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_j, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})}{f(X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_j^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})} \cdot \frac{q_j(X_j^{(t-1)} | X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_j, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})}{q_j(X_j | X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_j^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})}\right\}.$$

4. With probability α_j set $X_j^{(t)} = X_j$, otherwise set $X_j^{(t)} = X_j^{(t-1)}$.
5. Set $X_\iota^{(t)} := X_\iota^{(t-1)}$ for all $\iota \neq j$.

The Gibbs sampler as a Metropolis-Hastings algorithm

Remark 5.2

The Gibbs sampler for a p-dimensional distribution is a special case of a one-at-a-time Metropolis-Hasting algorithm:

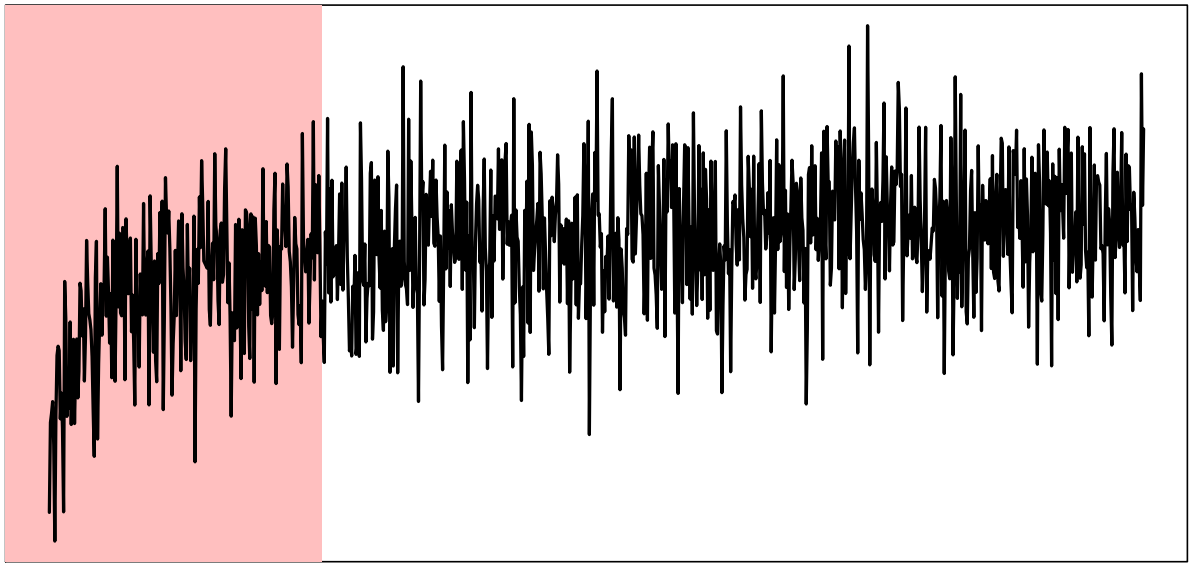
- the (systematic scan) Gibbs sampler is a cycle of p kernels,
- the random scan Gibbs sampler is a mixture of these kernels.

The proposal q_j corresponding to the j-th kernel consists of drawing $X_j^{(t)} \sim f_{X_j|X_{-j}}$.
The corresponding probability of acceptance is uniformly equal to 1.

7 Convergence diagnostics

Practical considerations: Burn-in period

- Theory (ergodic theorems) allows for the use of the entire chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots)$.
- However distribution of $(\mathbf{X}^{(t)})$ for small t might still be far from the stationary distribution f.
- Can be beneficial to discard the first iterations $\mathbf{X}^{(t)}$, $t = 1, \ldots, T_0$ (*burn-in period*).
- Optimal T_0 depends on mixing properties of the chain.



Practical considerations: Thinning (1)

- MCMC methods typically yield positively correlated chain: $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ large for small τ.
- Idea: build a subchain by only keeping every m-th value: Consider a Markov chain $(\mathbf{Y}^{(t)})_{t=1,\ldots,\lfloor T/m \rfloor}$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$ instead of $(\mathbf{X}^{(t)})_{t=1,\ldots,T}$ (*thinning*).
- $(\mathbf{Y}^{(t)})_t$ exhibits less autocorrelation than $(\mathbf{X}^{(t)})_t$, i.e.

$$\rho(\mathbf{Y}^{(t)}, \mathbf{Y}^{(t+\tau)}) = \rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+m \cdot \tau)}) < \rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)}),$$

  if the correlation $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ decreases monotonically in τ.
- Price we have to pay: length of $(\mathbf{Y}^{(t)})_{t=1,\ldots,\lfloor T/m \rfloor}$ is only $(1/m)$-th of the length of $(\mathbf{X}^{(t)})_{t=1,\ldots,T}$.

Practical considerations: Thinning (2)

- If $\mathbf{X}^{(t)} \sim f$ and corresponding variances exist,

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right) \leq \mathrm{Var}\left(\frac{1}{\lfloor T/m \rfloor}\sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)})\right),$$

  i.e. thinning cannot be justified when objective is estimating $\mathbb{E}_f(h(\mathbf{X}))$.
- Thinning can be a useful concept
  - if computer has insufficient memory.
  - for convergence diagnostics: $(\mathbf{Y}^{(t)})_{t=1,\ldots,\lfloor T/m \rfloor}$ is closer to an i.i.d. sample than $(\mathbf{X}^{(t)})_{t=1,\ldots,T}$.

The need for convergence diagnostics

- Theory we have studied guarantees (under certain conditions) the convergence of the Markov chain $\mathbf{X}^{(t)}$ to the desired distribution.
- This does not imply that a *finite* sample from such a chain yields a good approximation to the target distribution.
- Validity of the approximation must be confirmed in practise.
- Convergence diagnostics help answering this question.
- Convergence diagnostics are *not* perfect and should be treated with a good amount of scepticism.

Different diagnostic tasks

Convergence to the target distribution Does $\mathbf{X}^{(t)}$ yield a sample from the target distribution?

- Has $(\mathbf{X}^{(t)})_t$ reached a stationary regime?
- Does $(\mathbf{X}^{(t)})_t$ cover the support of the target distribution?

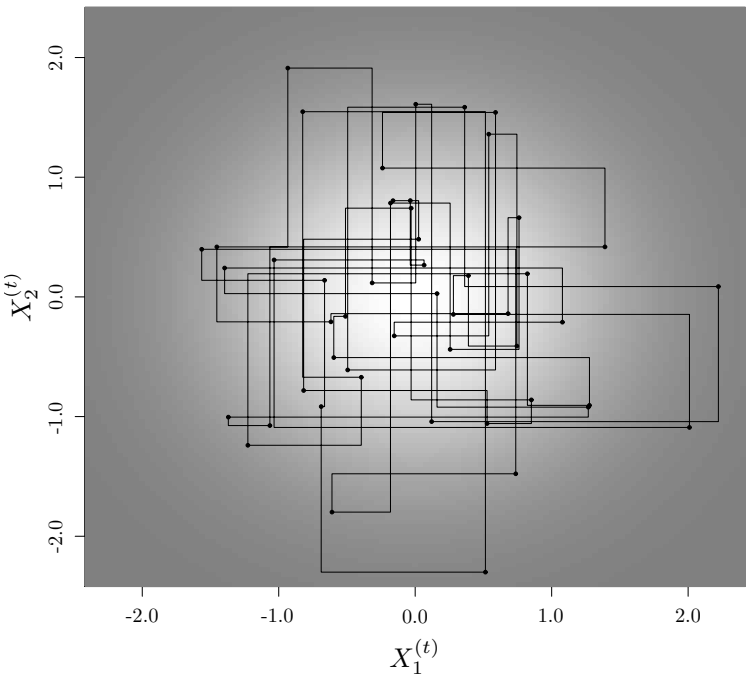
Convergence of the averages Does $\sum_{t=1}^{T} h(\mathbf{X}^{(t)})/T$ provide a good approximation to the expectation $\mathbb{E}_f(h(\mathbf{X}))$ under the target distribution?

Comparison to i.i.d. sampling How much information is contained in the sample from the Markov chain compared to i.i.d. sampling?

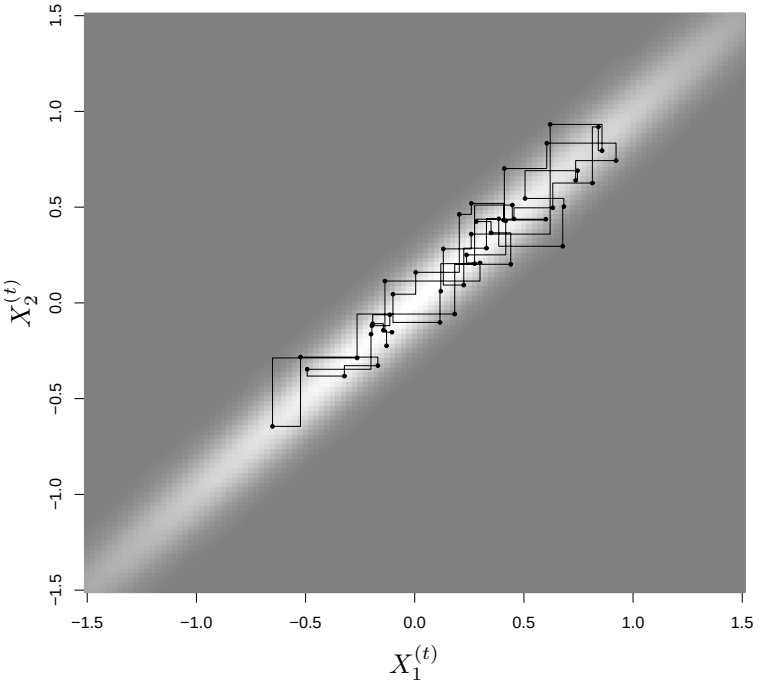
Pathological example 1: potentially slowly mixing

Gibbs sampler from a bivariate Gaussian with correlation $\rho(X_1, X_2)$

$$\rho(X_1, X_2) = 0.3$$



$$\rho(X_1, X_2) = 0.99$$



For correlations $\rho(X_1, X_2)$ close to ± 1 the chain can be poorly mixing.

Pathological example 2: no central limit theorem

The following MCMC algorithm has the Beta$(\alpha, 1)$ distribution as stationary distribution:

Starting with any $X^{(0)}$ iterate for $t = 1, 2, \ldots$

1. With probability $1 - X^{(t-1)}$, set $X^{(t)} = X^{(t-1)}$.
2. Otherwise draw $X^{(t)} \sim \text{Beta}(\alpha + 1, 1)$.

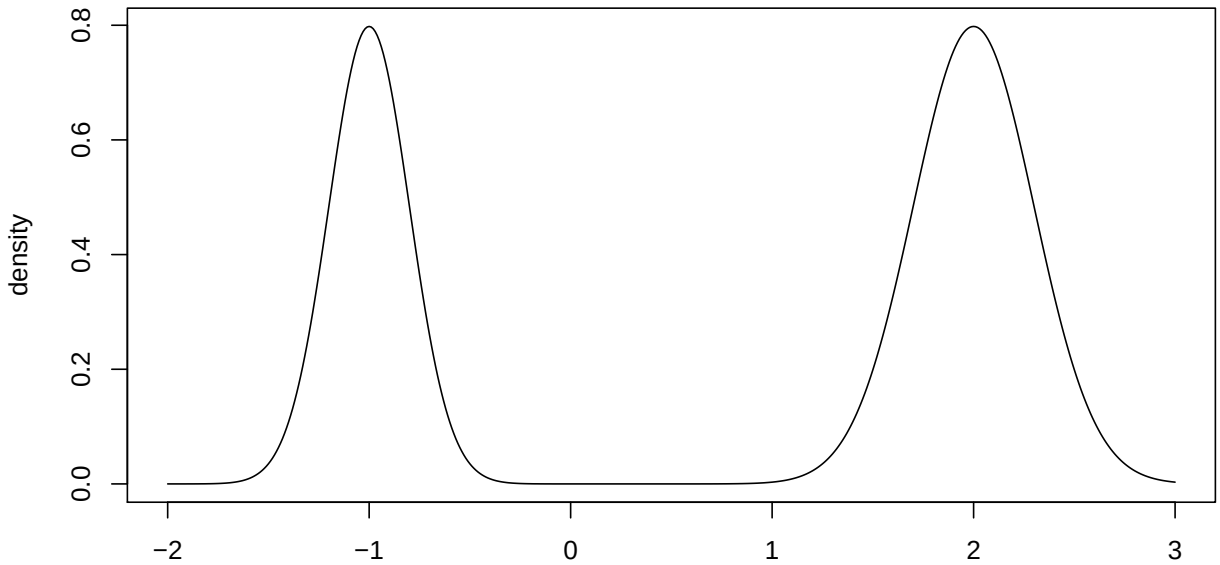
Markov chain converges very slowly (no central limit theorem applies).

Pathological example 3: nearly reducible chain

Metropolis-Hastings sample from a mixture of two well-separated Gaussians, i.e. the target is

$$f(x) = 0.4 \cdot \phi_{(-1,0.2^2)}(x) + 0.6 \cdot \phi_{(2,0.3^2)}(x)$$

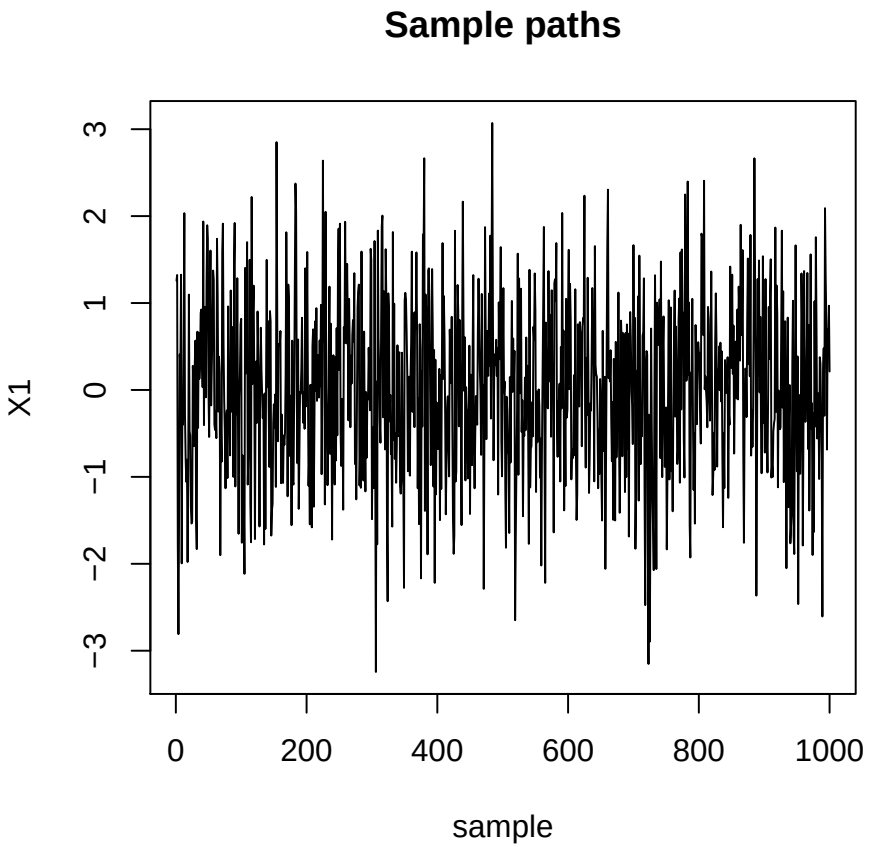
If the variance of the proposal is too small, the chain cannot move from one population to the other.

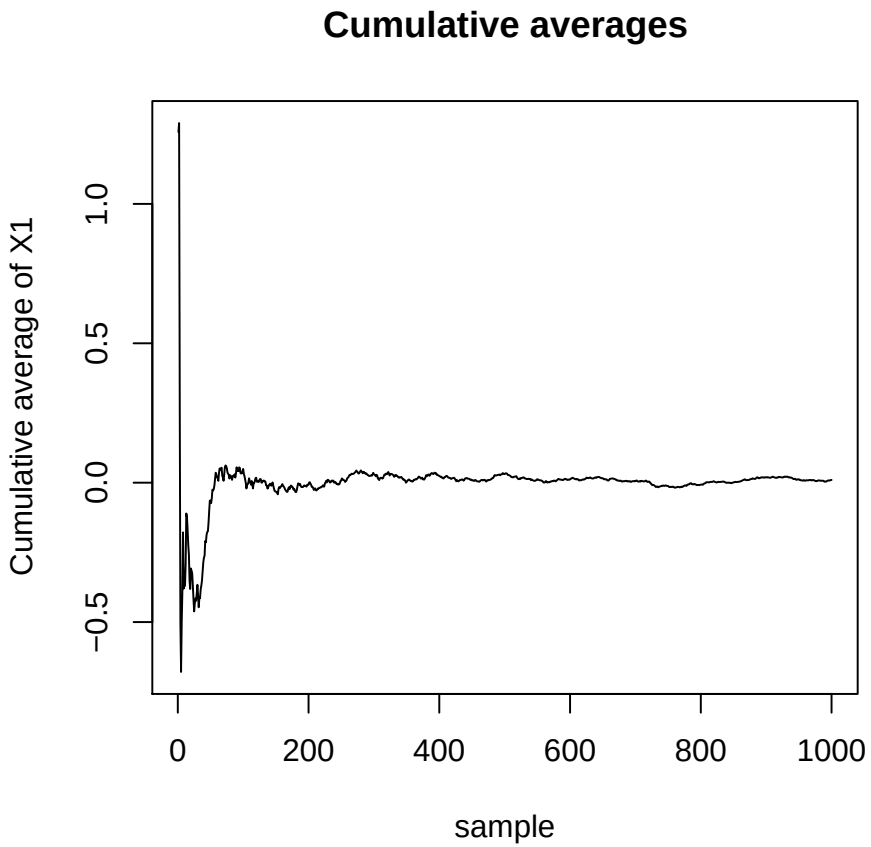


Basic plots

- Plot the sample paths $(X_j^{(t)})_t$.
  should be oscillating very fast and show very little structure.
- Plot the cumulative averages $(\sum_{\tau=1}^{t} X_j^{(\tau)}/t)_t$.
  should be converging to a value.
- Alternatively plot CUSUM $(\bar{X}_j - \sum_{\tau=1}^{t} X_j^{(\tau)}/t)_t$ with $\bar{X}_j = \sum_{\tau=1}^{T} X_j^{(\tau)}/T$.
  should be converging to 0.
- Only very obvious problems visible in these plots.
- Difficult to assess multivariate distributions from univariate projections.

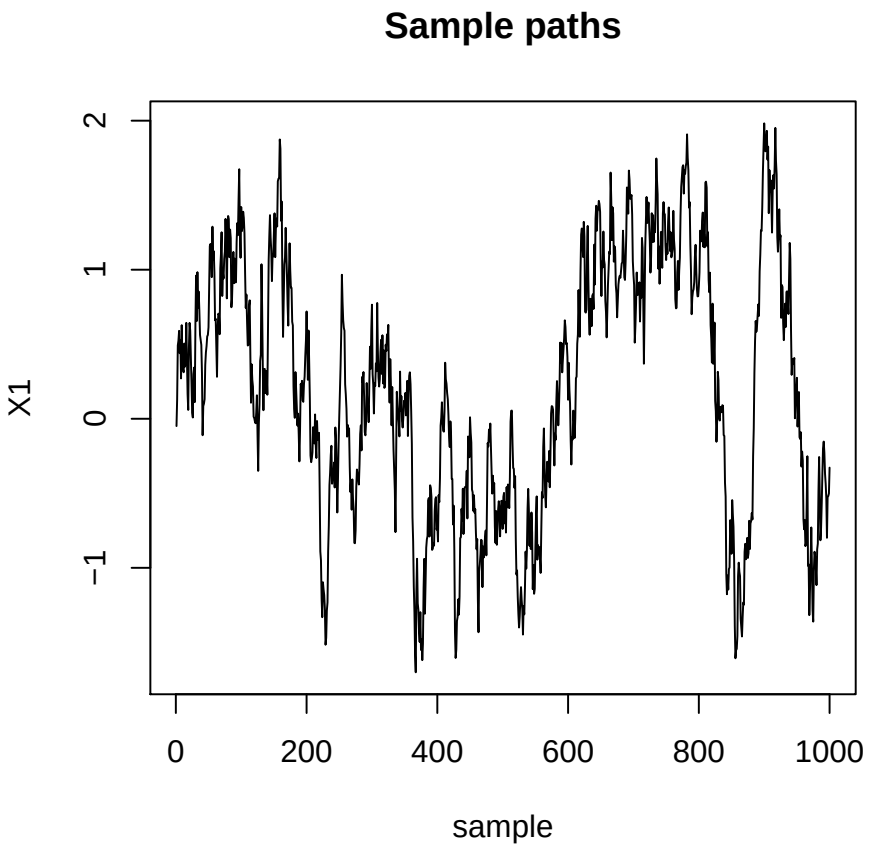
Basic plots for pathological example 1 ($\rho(X_1, X_2) = 0.3$)

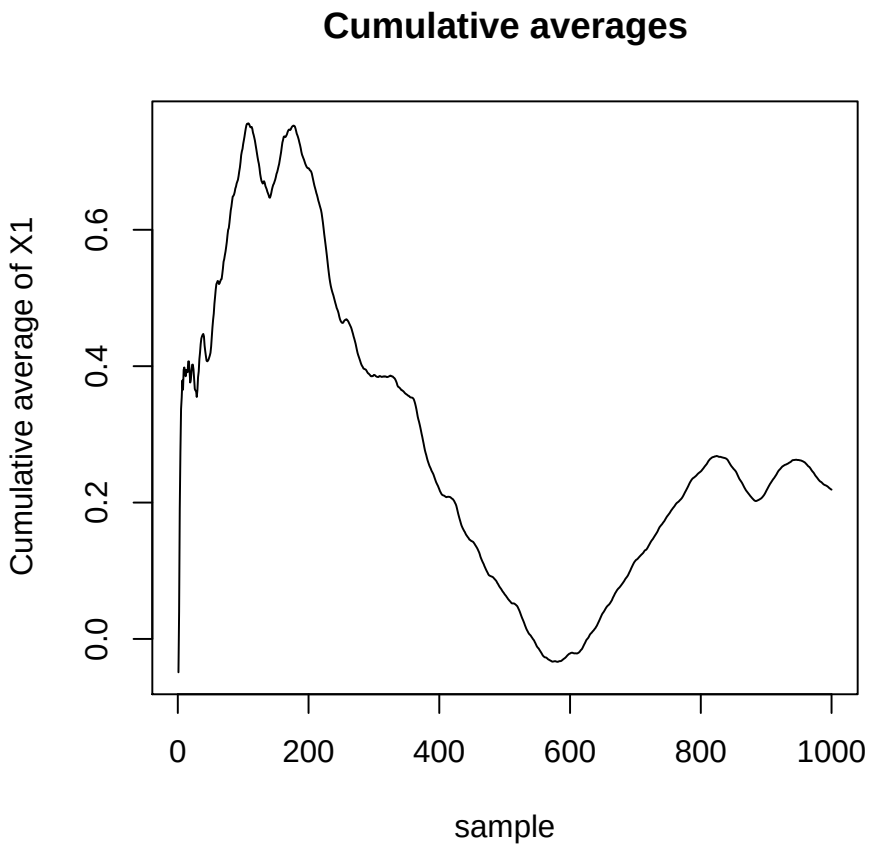




Looks OK.

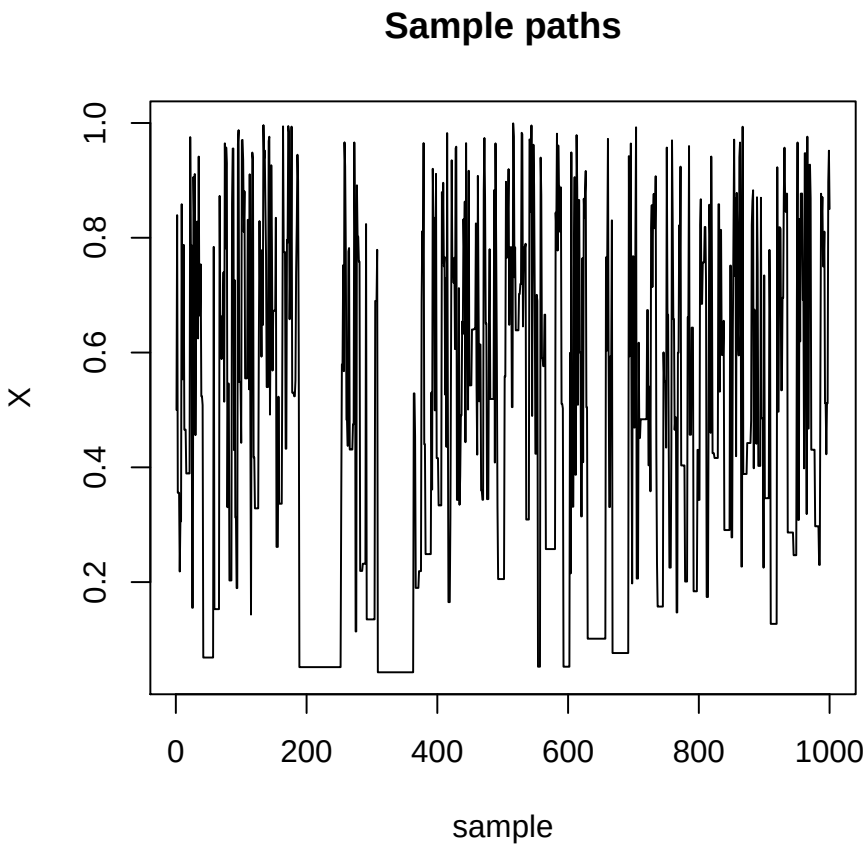
Basic plots for pathological example 1 ($\rho(X_1, X_2) = 0.99$)

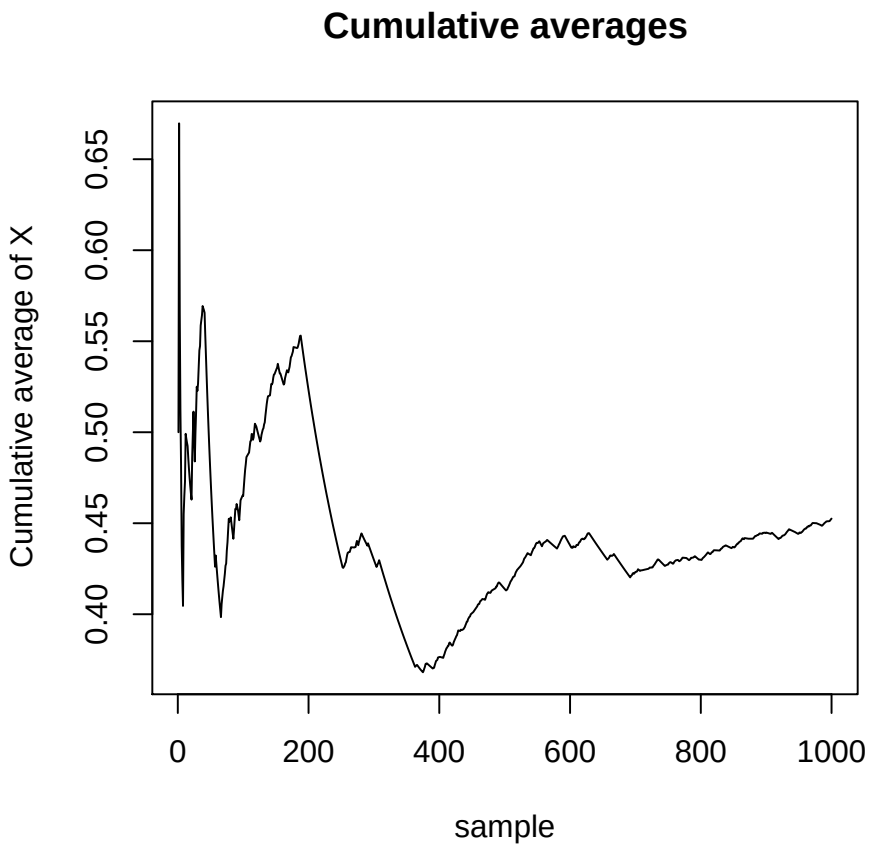




Slow mixing speed can be detected.

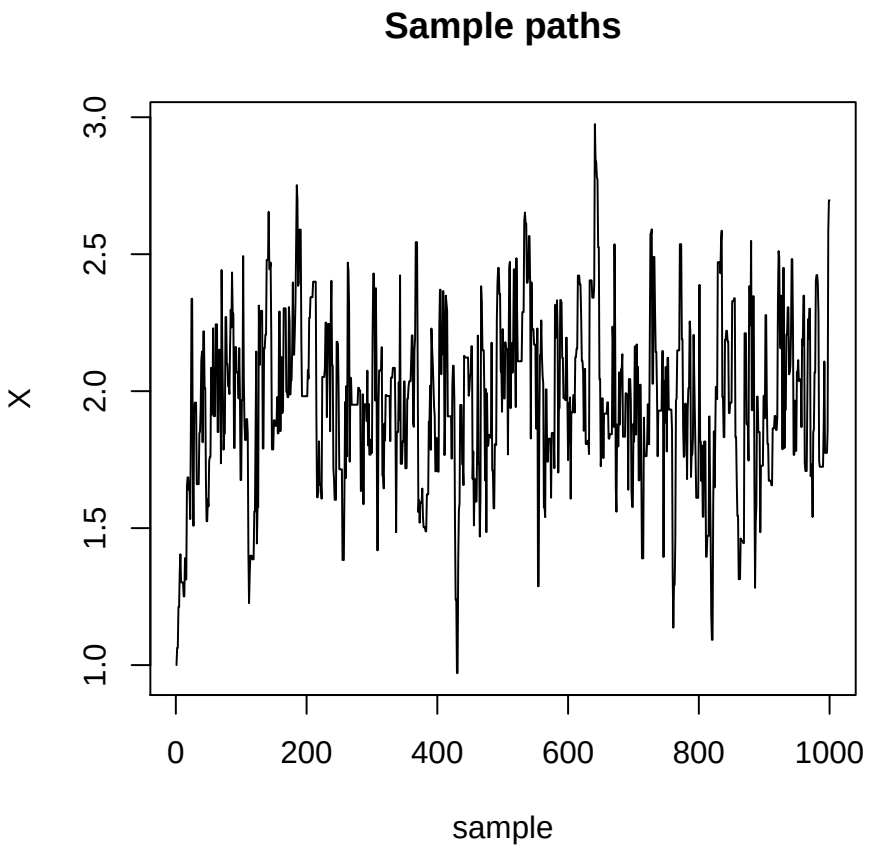
Basic plots for pathological example 2

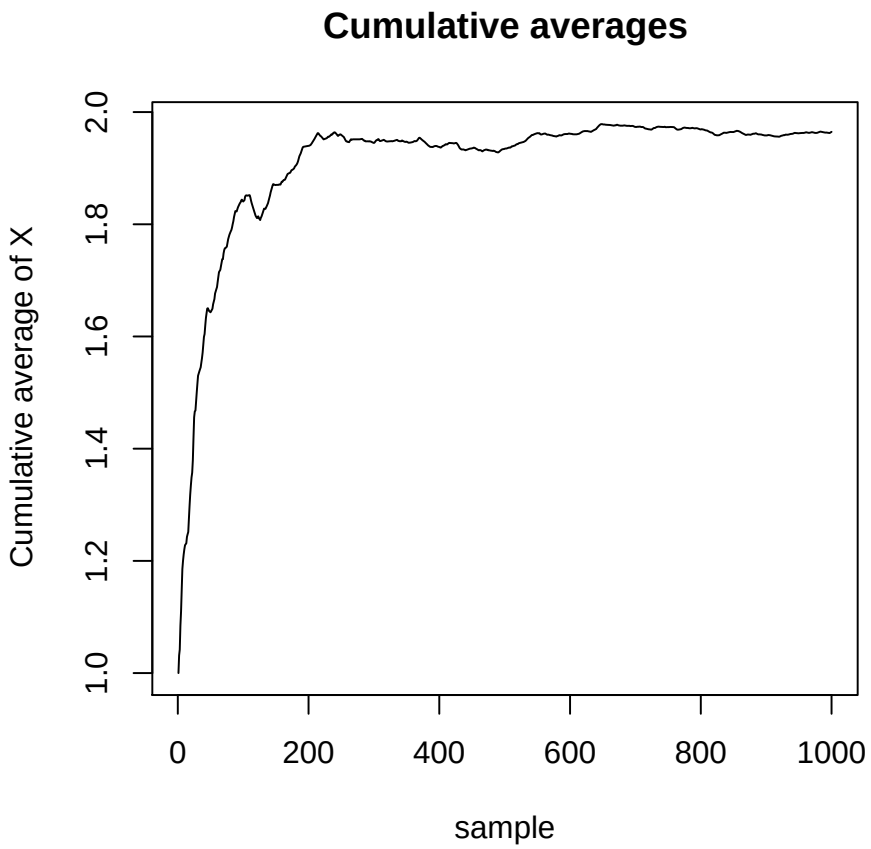




Slow convergence of the mean can be detected.

Basic plots for pathological example 3





We *cannot* detect that the sample only covers one part of the distribution.
("you've only seen where you've been")

Non-parametric tests of convergence

- Partition chain in 3 blocks:

  burn-in $(\mathbf{X}^{(t)})_{t=1,\ldots,\lfloor T/3 \rfloor}$

  first block $(\mathbf{X}^{(t)})_{t=\lfloor T/3 \rfloor+1,\ldots,2\lfloor T/3 \rfloor}$

  second block $(\mathbf{X}^{(t)})_{t=2\lfloor T/3 \rfloor+1,\ldots,T}$

- Distribution of $\mathbf{X}^{(t)}$ in both blocks should be identical.
- Idea: Use of a non-parametric test to test whether the two distributions are identical.
- Problem: Tests designed for i.i.d. samples.
  $\rightsquigarrow$ Resort to a (less correlated) thinned chain $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$.

Kolmogorov-Smirnov test

- Two i.i.d. populations: $Z_{1,1}, \ldots, Z_{1,n}$ and $Z_{2,1}, \ldots, Z_{2,n}$
- Estimate empirical CDF in each population:

$$\hat{F}_k(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty, z]}(Z_{k,i})$$

- Test statistic is the maximum difference between the two empirical CDFs:

$$K = \sup_{x \in \mathbb{R}} |\hat{F}_1(x) - \hat{F}_2(x)|$$

- For $n \to \infty$ the CDF of $\sqrt{n} \cdot K$ converges to the CDF

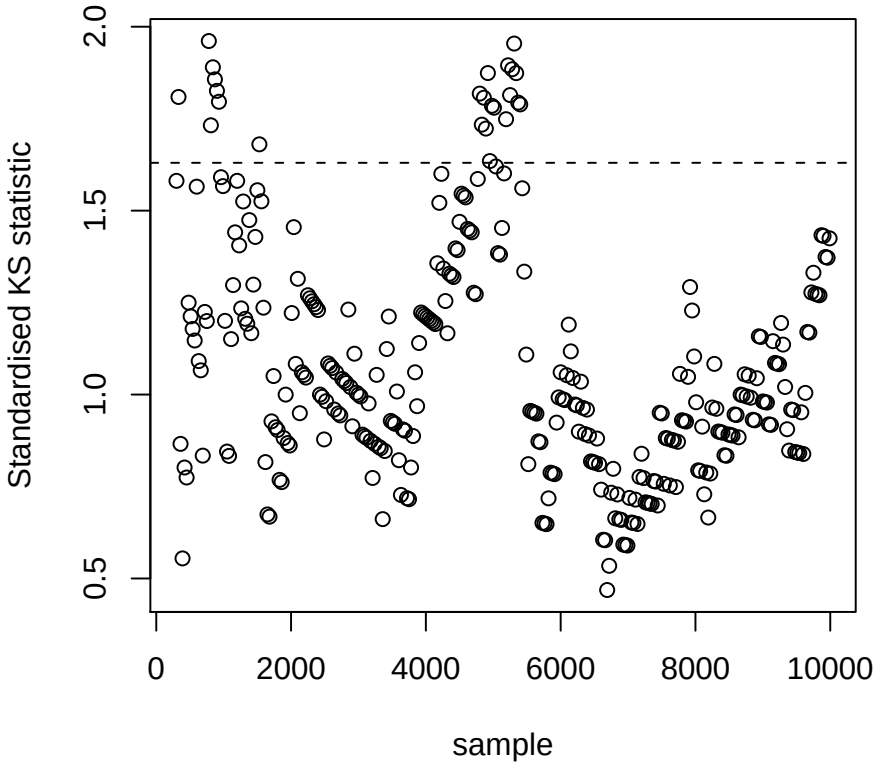
$$R(k) = 1 - \sum_{i=1}^{+\infty} (-1)^{i-1} \exp(-2i^2 k^2)$$

Kolmogorov-Smirnov test

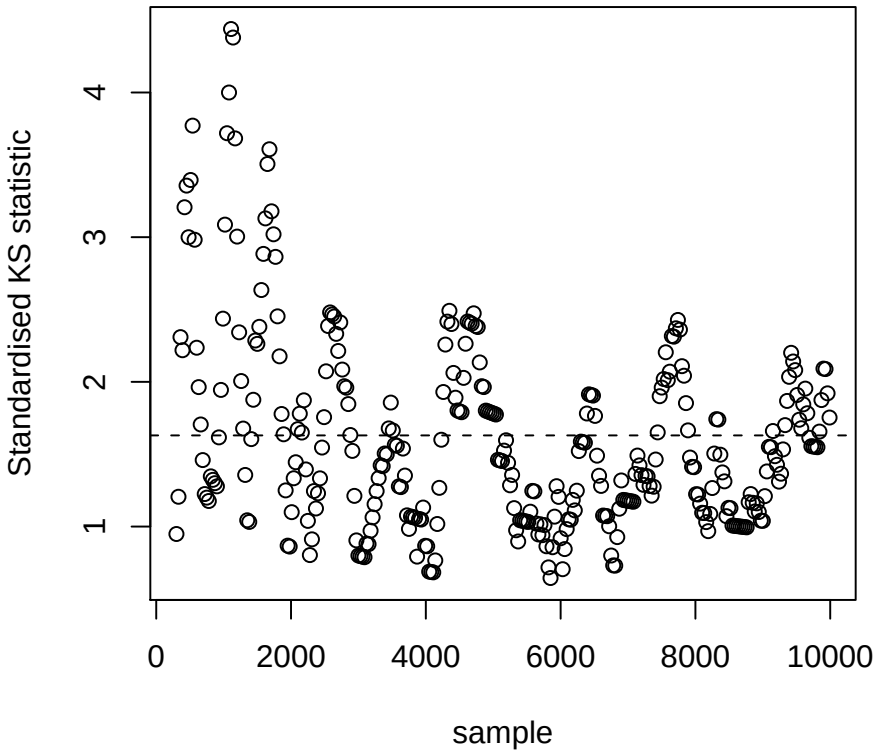
- In our case the two populations are
  thinned first block $(\mathbf{Y}^{(t)})_{t=\lfloor T/(3\cdot m)\rfloor+1,\ldots,2\lfloor T/(3\cdot m)\rfloor}$
  thinned second block $(\mathbf{X}^{(t)})_{t=2\lfloor T/(3\cdot m)\rfloor+1,\ldots,\lfloor T/m\rfloor}$
- Even the thinned chain $(\mathbf{Y}^{(t)})_t$ is autocorrelated
  $\rightsquigarrow$ test invalid from a formal point of view.
- Standardised test statistic $\sqrt{\lfloor T/(3\cdot m)\rfloor}\cdot K$ can still be used a heuristic tool.

KS test for pathological example 1

$$\rho(X_1, X_2) = 0.3$$

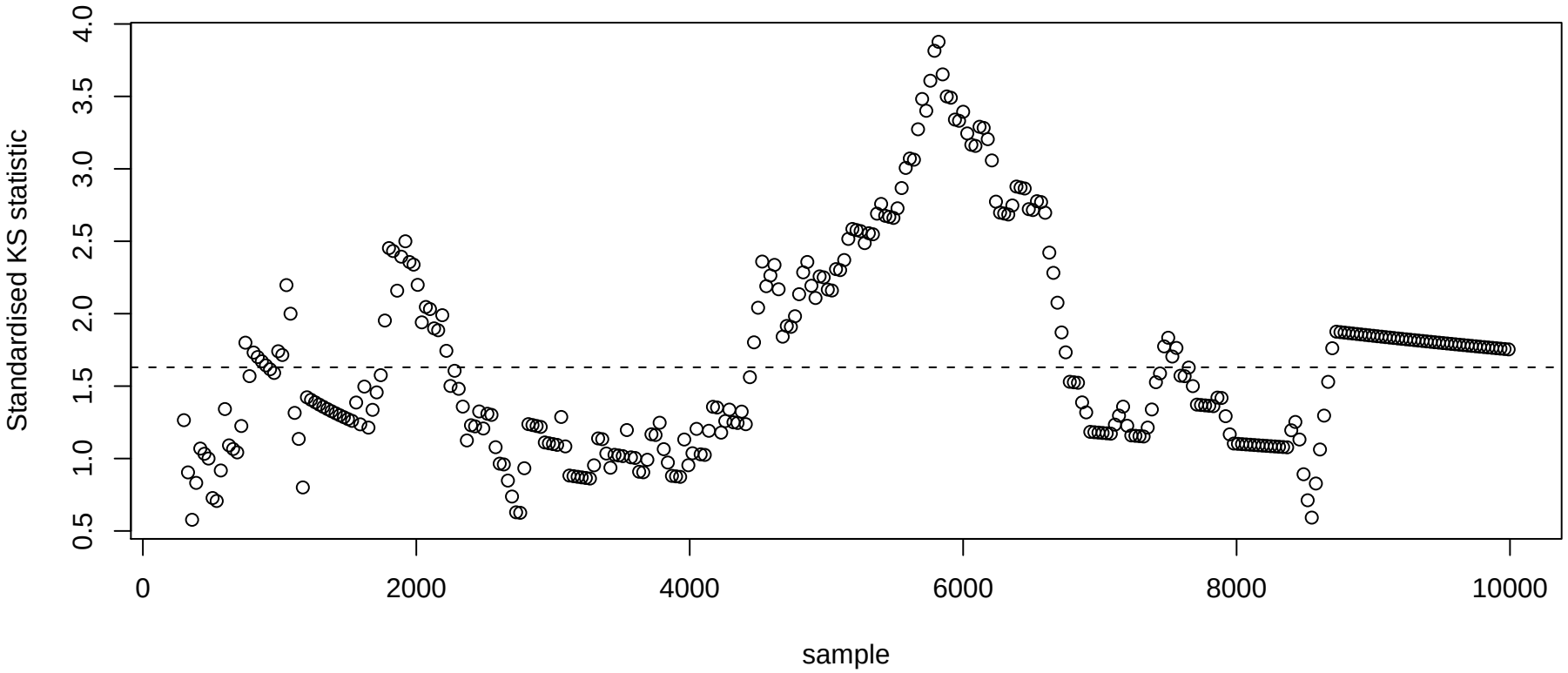


$$\rho(X_1, X_2) = 0.99$$



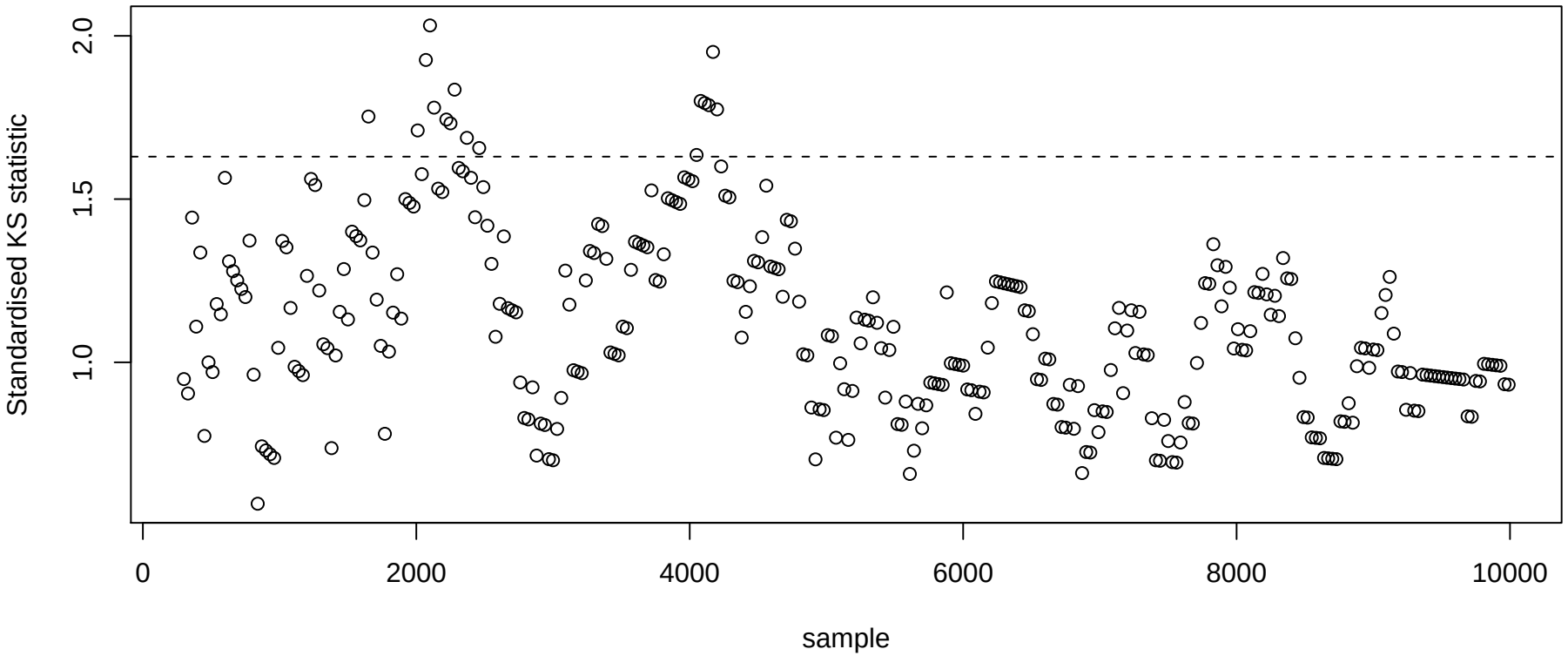
Slow mixing speed can be detected for the highly correlated chain.

KS test for pathological example 2



Problems can be detected.

KS test for pathological example 3

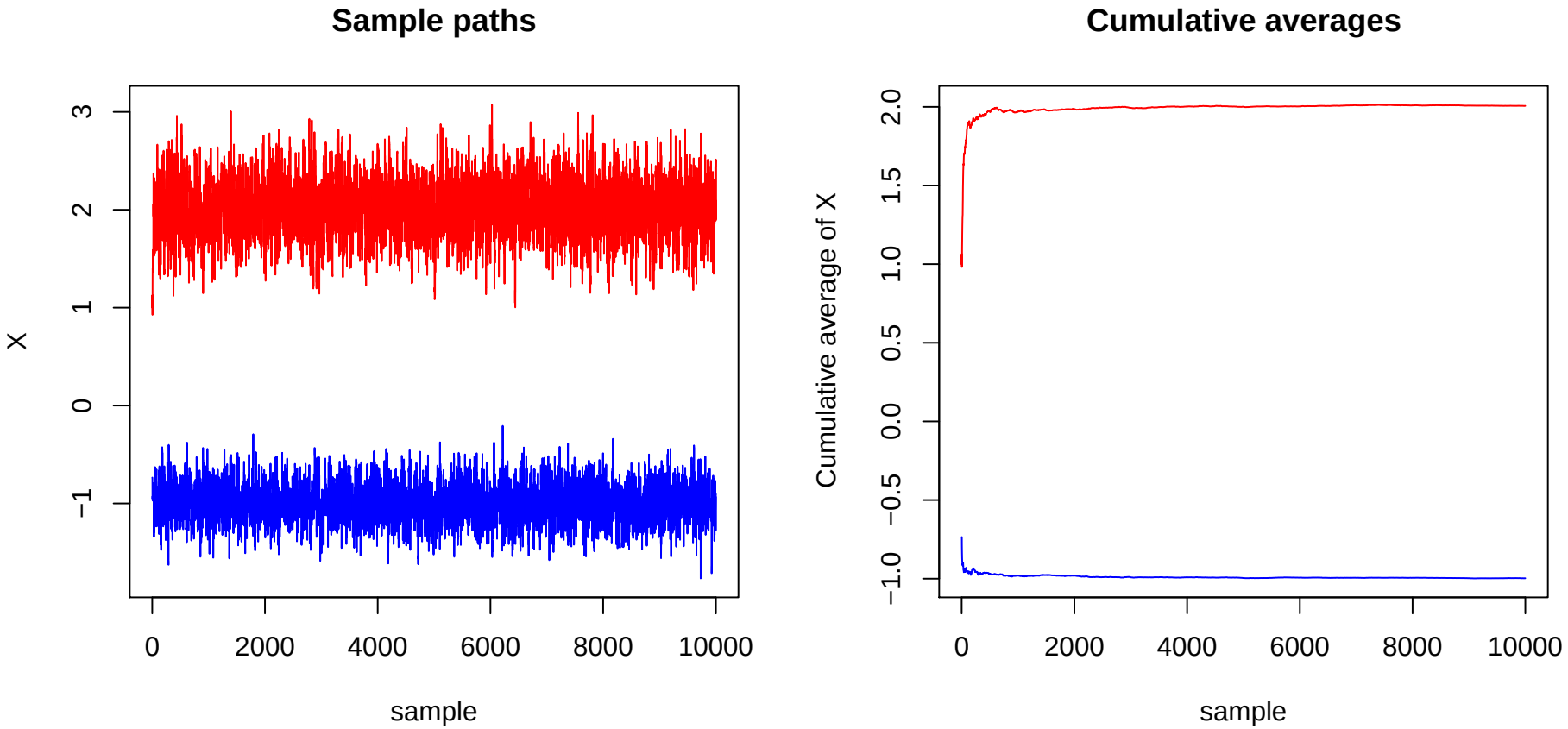


We *cannot* detect that the sample only covers one part of the distribution.
("you've only seen where you've been")

Comparing multiple chains

- Compare $L > 1$ chains $(\mathbf{X}^{(1,t)})_t, \ldots, (\mathbf{X}^{(L,t)})_t$.
- Initialised using overdispersed starting values $\mathbf{X}^{(1,0)}, \ldots, \mathbf{X}^{(L,0)}$.
- Idea: Variance and range of each chain $(\mathbf{X}^{(l,t)})_t$ should equal the range and variance of all chains pooled together.
- Compare basic plots for the different chains.
- Quantitative measure:
  - Compute distance $\delta_\alpha^{(l)}$ between α and $(1-\alpha)$ quantile of $(X_k^{(l,t)})_t$.
  - Compute distance $\delta_\alpha^{(\cdot)}$ between α and $(1-\alpha)$ quantile of the pooled data.
  - The ratio $\hat{S}_\alpha^{\text{interval}} = \dfrac{\sum_{l=1}^{L} \delta_\alpha^{(l)}/L}{\delta_\alpha^{(\cdot)}}$ should be around 1.
- Alternative: compare variance within each chain with the pooled variance estimate.
- Choosing suitable initial values $\mathbf{X}^{(1,0)}, \ldots, \mathbf{X}^{(L,0)}$ difficult in high dimensions.

Comparing multiple chains plots for pathological example 3



$$\hat{S}_{\alpha}^{\text{interval}} = 0.2703 \ll 1$$

We can detect that the sample only covers one part of the distribution (provided the chains are initialised appropriately).

Riemann sums and control variates

- Consider order statistic $X^{[1]} \leq \ldots \leq X^{[T]}$.
- Provided $(X^{[t]})_t = 1 \ldots, T$ covers the support of the target, the Riemann sum

$$\sum_{t=2}^{T} (X^{[t]} - X^{[t-1]}) f(X^{[t]})$$

converges to

$$\int f(x) dx = 1.$$

- Thus if $\sum_{t=2}^{T} (X^{[t]} - X^{[t-1]}) f(X^{[t]}) \ll 1$, the Markov chain has failed to explore all the support of the target.
- Requires that target density f is available inclusive of normalisation constants.
- Only effective in 1D.
- Riemann sums can be seen as a special case of *control variates*.

Riemann sums for pathological example 3

For the chain stuck in the population with mean 2 we obtain

$$\sum_{t=2}^{T}(X^{[t]} - X^{[t-1]})f(X^{[t]}) = 0.598 \ll 1,$$

so we can detect that we have not explored the whole distribution.

Effective sample size

- MCMC algorithms yield a positively correlated sample $(\mathbf{X}^{(t)})_{t=1,\ldots,T}$.
- MCMC sample of size T thus contains less information than an i.i.d. sample of size T.
- Question: how much less information?
- Approximate $(h(\mathbf{X}^{(t)}))_{t=1,\ldots,T}$ by an $AR(1)$ process, i.e. we assume that

$$\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho^{|\tau|}.$$

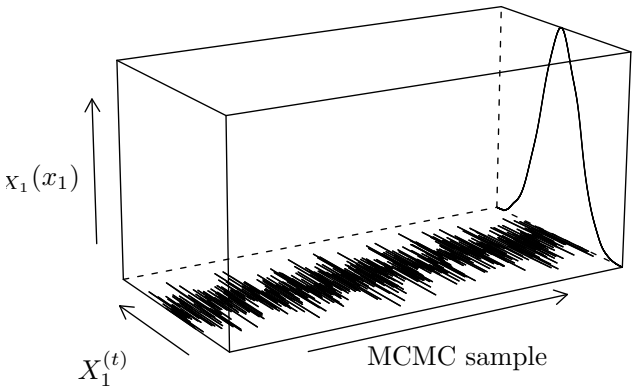
- Variance of the estimator is

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right) \approx \frac{1+\rho}{1-\rho} \cdot \frac{1}{T}\mathrm{Var}\left(h(\mathbf{X}^{(t)})\right)$$

- Same variance as an i.i.d. sample of the size $T \cdot \dfrac{1-\rho}{1+\rho}$.
- Thus define $T \cdot \dfrac{1-\rho}{1+\rho}$ as *effective sample size*.

Effective sample for pathological example 1

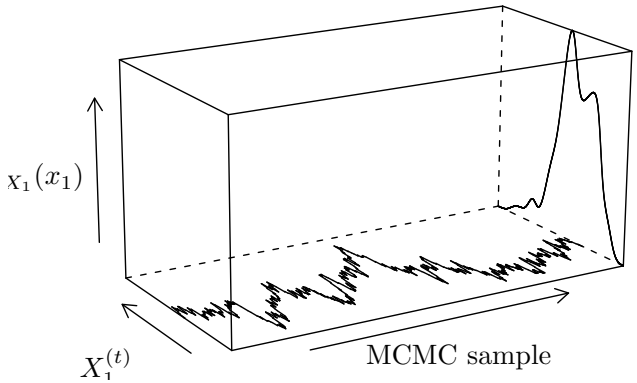
Rapidly mixing chain
$(\rho(X_1, X_2) = 0.3)$
10,000 samples



$\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.078$

ESS for estimating $\mathbb{E}_f(X_1)$ is 8,547.

Slowly mixing chain
$(\rho(X_1, X_2) = 0.99)$
10,000 samples



$\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.979$

ESS for estimating $\mathbb{E}_f(X_1)$ is 105.