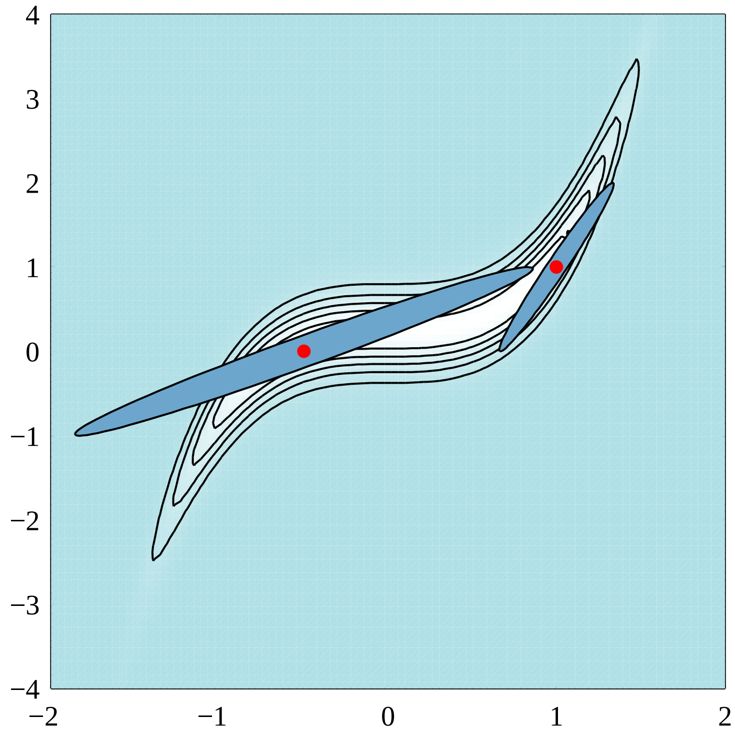
Adaptive Metropolis-Hastings (AM) algorithm

Adaptation: As the sampling proceeds, the proposal distribution is updated to conform with the underlying density.



Adaptation strategies

Analytic adaptation: Calculate a local Gaussian approximation of the probability density. Update every Mth step.

May be computationally heavy

May be difficult to find analytically and requires numerical differentiation

Sampling-based adaptation: Use the already calculated sample history to determine the new proposal distribution.

Sampling-based adaptation

Design an algorithm along the following guidelines.

- Start a usual MH sampling at a point of your choice using a white noise proposal distribution.
- After possibly removing a burn-in sequence from the beginning, calculate the empirical covariance of the sample points obtained thus far.
- Use the empirical covariance to sample new points.
- Update the covariance every Mth sample point.

ADAPTED RANDOM WALK MH ALGORITHM

1. Initialize $k = 0$, $C_k = \gamma^2 I$.
2. Generate a sample sequence of length M,

$$x_{kM+1}, x_{kM+2}, \ldots, x_{(k+1)M},$$

   using the random walk proposal

$$x_{\text{prop}} = x_{\text{curr}} + w, \quad w \sim \mathcal{N}(0, C_k)$$

3. Update

$$C_k \to C_{k+1} = \text{cov}(x_1, x_2, \ldots, x_{(k+1)M}) + \varepsilon I.$$

4. Increase $k \to k+1$ and continue from 2 until desired sample size is reached.

Questions, answers

- Is this a Markov Chain method? The update depends on *all* of the sample history via the covariance matrix!

  True! But one can show that little by little, the process forgets the past, and is *asymptotically* Markovian and therefore ergodic.

- What is that little εI doing there?

  It has two functions: Practical function is to avoid the possibility that all sample points become collinear. Theoretical function is to make sure that the ergodicity works.

- How do we update the covariance in practice?

  Let's see...

Covariance updating, stably

Divide the sample in blocks of length M:

$$\underbrace{x_1, x_2, \ldots, x_M}_{M}, \underbrace{x_{M+1}, x_{M+2}, \ldots, x_{2M}}_{M}, x_{2M+1}, \ldots$$

Average and covariance over subsamples:

$$\widehat{x}_k = \frac{1}{M} \sum_{j=(k-1)M+1}^{kM} x_j,$$

$$\widehat{C}_k = \frac{1}{M} \sum_{j=(k-1)M+1}^{kM} (x_j - \widehat{x}_k)(x_j - \widehat{x}_k).$$

Denote the *cumulative* mean and covariance by

$$\overline{x}_{kM} = \frac{1}{kM} \sum_{j=1}^{kM} x_j,$$

$$\overline{C}_{kM} = \frac{1}{kM} \sum_{j=1}^{kM} (x_j - \overline{x}_{kM})(x_j - \overline{x}_{kM}).$$

Problem: *Find a numerically stable way of updating*

$$\overline{x}_{kM} \to \overline{x}_{(k+1)M},$$

$$\overline{C}_{kM} \to \overline{C}_{(k+1)M}.$$

Updating the mean

We have

$$
\begin{aligned}
\overline{x}_{(k+1)M} &= \frac{1}{(k+1)M} \sum_{j=1}^{(k+1)M} x_j \\
&= \frac{k}{k+1} \frac{1}{kM} \sum_{j=1}^{kM} x_j + \frac{1}{(k+1)M} \sum_{j=kM+1}^{(k+1)M} x_j \\
&= \frac{k}{k+1} \overline{x}_{kM} + \frac{1}{k+1} \widehat{x}_{k+1}.
\end{aligned}
$$

We need some auxiliary results.

FACT 1: If

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j,$$

the covariance can be written as

$$\begin{aligned} C &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x})^{\mathrm{T}} \\ &= \frac{1}{n}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} - \underbrace{\frac{1}{n}\sum_{j=1}^{n} x_j}_{=\overline{x}}\,\overline{x}^{\mathrm{T}} - \overline{x}\,\underbrace{\frac{1}{n}\sum_{j=1}^{n} x_j^{\mathrm{T}}}_{=\overline{x}^{\mathrm{T}}} + \overline{x}\overline{x}^{\mathrm{T}} \\ &= \frac{1}{n}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} - \overline{x}\overline{x}^{\mathrm{T}}. \end{aligned}$$

FACT 2: For any $\widetilde{x}$, the *non-centered covariance* is

$$\begin{aligned}
\widetilde{C} &= \frac{1}{n}\sum_{j=1}^{n}(x_j-\widetilde{x})(x_j-\widetilde{x})^{\mathrm{T}} \\
&= \frac{1}{n}\sum_{j=1}^{n}x_jx_j^{\mathrm{T}} - \underbrace{\frac{1}{n}\sum_{j=1}^{n}x_j}_{=\overline{x}}\,\widetilde{x}^{\mathrm{T}} - \widetilde{x}\,\underbrace{\frac{1}{n}\sum_{j=1}^{n}x_j^{\mathrm{T}}}_{=\overline{x}^{\mathrm{T}}} + \widetilde{x}\widetilde{x}^{\mathrm{T}} \\
&= \frac{1}{n}\sum_{j=1}^{n}x_jx_j^{\mathrm{T}} - \overline{x}\overline{x}^{\mathrm{T}} + \left(\overline{x}\overline{x}^{\mathrm{T}} - \overline{x}\widetilde{x}^{\mathrm{T}} - \widetilde{x}\overline{x}^{\mathrm{T}} + \widetilde{x}\widetilde{x}^{\mathrm{T}}\right) \\
&= C + (\overline{x}-\widetilde{x})(\overline{x}-\widetilde{x})^{\mathrm{T}}.
\end{aligned}$$

With these results, we have

$$
\begin{aligned}
C_{(k+1)M} &= \frac{1}{(k+1)M} \sum_{j=1}^{(k+1)M} (x_j - \overline{x}_{(k+1)M})(x_j - x_{(k+1)M})^{\mathrm{T}} \\
&= \frac{1}{(k+1)M} \left(\sum_{j=1}^{kM} + \sum_{kM+1}^{(k+1)M} \right) (x_j - \overline{x}_{(k+1)M})(x_j - x_{(k+1)M})^{\mathrm{T}}.
\end{aligned}
$$

Both terms are, up to a multiplicative factor, non-centered covariances.

First term:

$$
\begin{aligned}
&\frac{1}{(k+1)M}\sum_{j=1}^{kM}(x_j-\overline{x}_{(k+1)M})(x_j-\overline{x}_{(k+1)M})^{\mathrm{T}} \\
&\quad=\frac{k}{k+1}\frac{1}{kM}\sum_{j=1}^{kM}(x_j-\overline{x}_{kM})(x_j-\overline{x}_{kM})^{\mathrm{T}} \\
&\qquad\quad+\frac{k}{k+1}\left(\overline{x}_{kM}-\overline{x}_{(k+1)M}\right)\left(\overline{x}_{kM}-\overline{x}_{(k+1)M}\right)^{\mathrm{T}} \\
&\quad=\frac{k}{k+1}C_{kN}+\frac{k}{k+1}\left(\overline{x}_{kM}-\overline{x}_{(k+1)M}\right)\left(\overline{x}_{kM}-\overline{x}_{(k+1)M}\right)^{\mathrm{T}}.
\end{aligned}
$$

Substituting the updating formula:

$$
\begin{aligned}
\overline{x}_{kM} - \overline{x}_{(k+1)M} &= \overline{x}_{kM} - \frac{k}{k+1}\overline{x}_{kM} - \frac{1}{k+1}\widehat{x}_{k+1} \\
&= \frac{1}{k+1}\big(\overline{x}_{kM} - \widehat{x}_{k+1}\big),
\end{aligned}
$$

so we have

$$
\frac{1}{(k+1)M}\sum_{j=1}^{kM}(x_j - \overline{x}_{(k+1)M})(x_j - x_{(k+1)M})^{\mathrm{T}}
$$

$$
= \frac{k}{k+1}C_{kN} + \frac{k}{(k+1)^3}\big(\overline{x}_{kM} - \widehat{x}_{k+1}\big)\big(\overline{x}_{kM} - \widehat{x}_{k+1}\big)^{\mathrm{T}}.
$$

Second term:

$$\frac{1}{(k+1)M}\sum_{j=kM+1}^{(k+1)M}(x_j-\overline{x}_{(k+1)M})(x_j-x_{(k+1)M})^{\mathrm{T}}$$

$$\frac{1}{k+1}\frac{1}{M}\sum_{j=kM+1}^{(k+1)M}(x_j-\widehat{x}_{k+1})(x_j-\widehat{x}_{k+1})^{\mathrm{T}}$$

$$+\frac{1}{k+1}(\widehat{x}_{k+1}-\overline{x}_{(k+1)M})(\widehat{x}_{k+1}-\overline{x}_{(k+1)M})^{\mathrm{T}}$$

$$\frac{1}{k+1}\widehat{C}_{k+1}+\frac{1}{k+1}(\widehat{x}_{k+1}-\overline{x}_{(k+1)M})(\widehat{x}_{k+1}-\overline{x}_{(k+1)M})^{\mathrm{T}}.$$

Again, substituting the recursion formula gives

$$
\begin{aligned}
\widehat{x}_{k+1} - \overline{x}_{(k+1)M} &= \widehat{x}_{k+1} - \frac{k}{k+1}\overline{x}_{kM} - \frac{1}{k+1}\widehat{x}_{k+1} \\
&= \frac{k}{k+1}\big(\widehat{x}_{k+1} - \overline{x}_{kM}\big),
\end{aligned}
$$

and therefore

$$
\frac{1}{(k+1)M}\sum_{j=kM+1}^{(k+1)M}(x_j - \overline{x}_{(k+1)M})(x_j - x_{(k+1)M})^{\mathrm{T}}
$$

$$
= \frac{1}{k+1}\widehat{C}_{k+1} + \frac{k^2}{(k+1)^3}\big(\overline{x}_{kM} - \widehat{x}_{k+1}\big)\big(\overline{x}_{kM} - \widehat{x}_{k+1}\big)^{\mathrm{T}}.
$$

Putting the pieces together gives

$$
\begin{aligned}
C_{(k+1)M} &= \frac{k}{k+1} C_{kN} + \frac{k}{(k+1)^3} \left(\overline{x}_{kM} - \widehat{x}_{k+1}\right)\left(\overline{x}_{kM} - \widehat{x}_{k+1}\right)^{\mathrm{T}} \\
&+ \frac{1}{k+1} \widehat{C}_{k+1} + \frac{k^2}{(k+1)^3} \left(\overline{x}_{kM} - \widehat{x}_{k+1}\right)\left(\overline{x}_{kM} - \widehat{x}_{k+1}\right)^{\mathrm{T}} \\
&= \frac{k}{k+1} C_{kN} + \frac{1}{k+1} \widehat{C}_{k+1} + \frac{k}{(k+1)^2} \left(\overline{x}_{kM} - \widehat{x}_{k+1}\right)\left(\overline{x}_{kM} - \widehat{x}_{k+1}\right)^{\mathrm{T}},
\end{aligned}
$$

which is the desired updating formula.

Example: proof of concept

Sampling a Gaussian in $\mathbb{R}^2$.

$$\pi(x) \propto \exp\left(-\frac{1}{2}(x-b)^{\mathrm{T}}\Gamma^{-1}(x-b)\right),$$

where

$$b = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \Gamma = UDU^{\mathrm{T}},$$

$$U = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \theta = \frac{\pi}{3},$$

$$D = \mathrm{diag}(1, 0.1).$$

```
% Defining the underlying distribution: a 2D Gaussian

 th = pi/3;
 U = [cos(th),-sin(th);sin(th),cos(th)];
 d = [1,0.1];
 D = diag(d);
 Gamma = U*D*U';
 b = [2;2];
 invGamma = inv(Gamma);
```

```
% Initializing

 x0 = [3;1];    % Initial sampling point
 step = 0.02;  % Initial step: no prior tuning for MH
 tiny = 1e-6;
 nsample = 150000;
 M = 100;       % Adaptation period
```

Observe: the step size is way too small for non-adaptive MH. The point here is to demonstrate that the adaptive method requires no tuning, i.e., you can start with sub-optimal proposal.

```
% Sampling without adaptation

 SampleNA = zeros(2,nsample);
 SampleNA(:,1) = x0;
 x = x0;
 lpdf = -0.5*(x-b)'*invGamma*(x-b);
 accrate = 0;
```

```
for j = 2:nsample
   % Draw the propsal
   xprop = x + step*randn(2,1);
   lpdfprop = -0.5*(xprop-b)'*invGamma*(xprop-b);
   % Check for acceptance
   if lpdfprop - lpdf >log(rand)
       %accept
       x = xprop;
       lpdf = lpdfprop;
       accrate = accrate + 1;
   end
   SampleNA(:,j) = x;
end
rel_accrate = 100*accrate/nsample
```

Updating

$$x_{\text{prop}} = x_{\text{curr}} + s, \quad s \sim \mathcal{N}(0, C),$$

that is,

$$\pi(s) \propto \exp\left(-\frac{1}{2} s^{\mathrm{T}} C^{-1} s\right).$$

Write the Cholesky decomposition,

$$C = R^{\mathrm{T}} R,$$

so

$$C^{-1} = R^{-1} R^{-\mathrm{T}}.$$

This means that

$$\pi(s) \propto \exp\left(-\frac{1}{2}\|R^{-\mathrm{T}}s\|^2\right),$$

or that

$$R^{-\mathrm{T}}s = w \sim \mathcal{N}(0, I).$$

Hence, the updating procedure is

$$x_{\text{prop}} = x_{\text{curr}} + R^{\mathrm{T}}w, \quad w \sim \mathcal{N}(0, I).$$

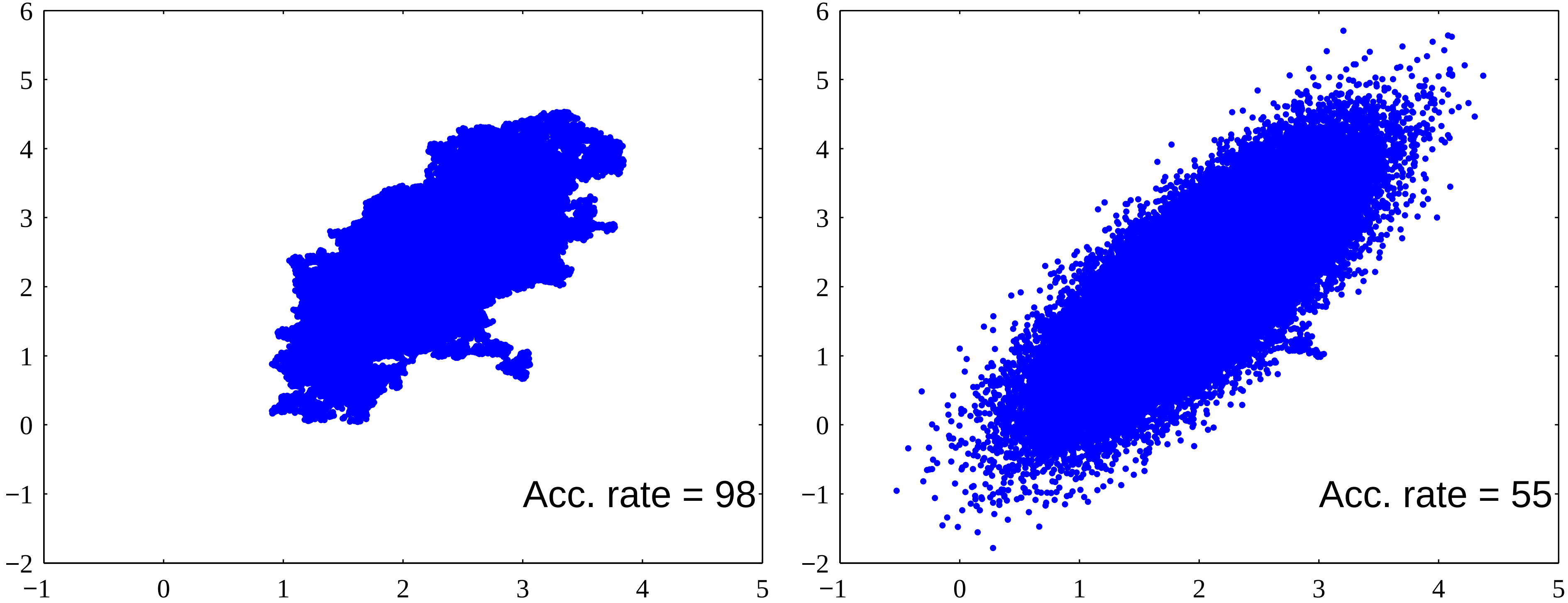
```
% Sampling with adaptation

 SampleA = zeros(2,nsample);
 SampleA(:,1) = x0;
 x = x0;
 lpdf = -0.5*(x-b)'*invGamma*(x-b);
 C = step^2*eye(2);
 mean = zeros(2,1);
 R = step*eye(2);
 accrate = 0;
 tempSample = [x];
 k = 0;
```

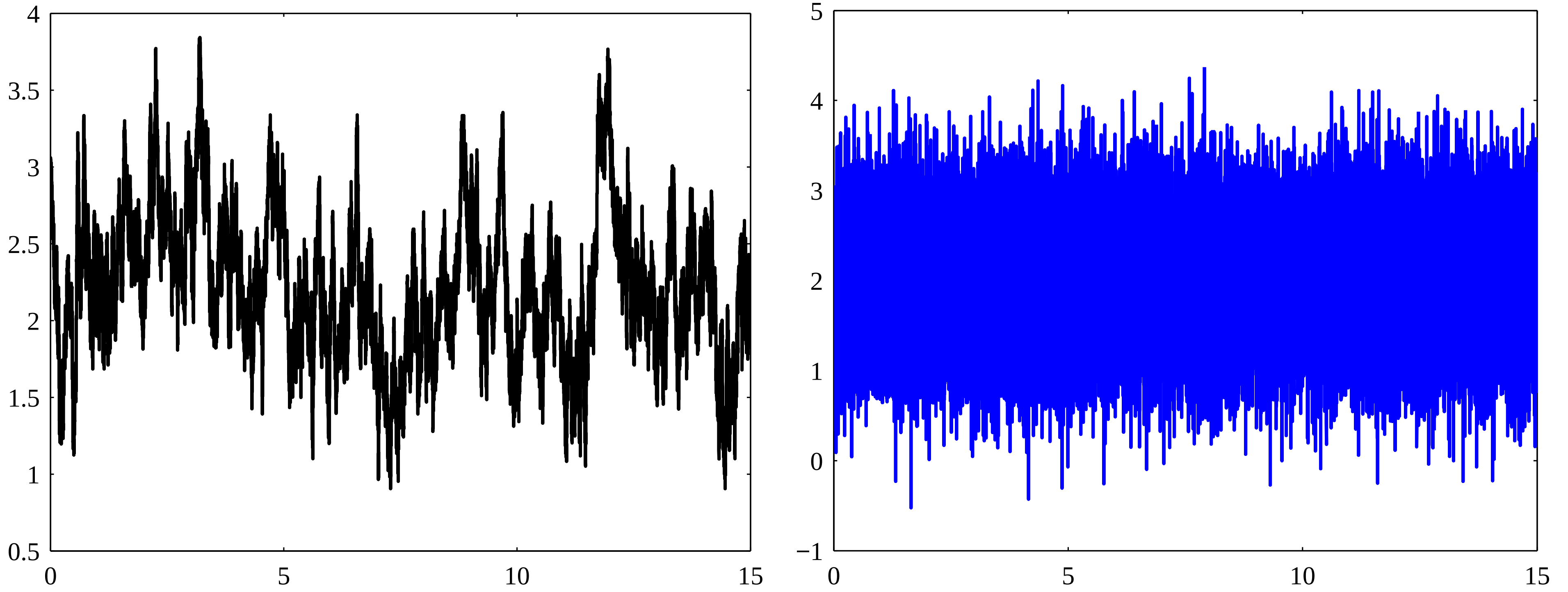
```
for j =2:nsample
   % Draw the proposal
   xprop = x + R'*randn(2,1);
   lpdfprop = -0.5*(xprop-b)'*invGamma*(xprop-b);
   % Check for acceptance
   if lpdfprop - lpdf >log(rand)
       %accept
       x = xprop;
       lpdf = lpdfprop;
       accrate = accrate + 1;
   end
   SampleA(:,j) = x;
   tempSample = [tempSample x];
```

```
    if mod(j,M) == 0
        % Update the proposal distribution
        xk = 1/M*sum(tempSample')';
        aux = tempSample - xk*ones(1,M);
        Ck = 1/M*aux*aux';
        mean = k/(k+1)*mean + 1/(k+1)*xk;
        C = k/(k+1)*C + 1/(k+1)*Ck + k/(k+1)^2*(mean-xk)*(mean-xk)';
        R = chol(C);
        k = k+1;
        tempSample = [];
    end
end rel_accrateA = 100*accrate/nsample
```

Scatter plots



Sample histories



The $p\%$ probability region:

$$\Gamma = UDU^{\mathrm{T}} \Rightarrow \Gamma^{-1} = UD^{-1}U^{\mathrm{T}},$$

$$\begin{aligned} \pi(x) &\propto \exp\left(-\frac{1}{2}(x-b)^{\mathrm{T}}\Gamma^{-1}(x-b)\right) \\ &= \exp\left(-\frac{1}{2}\|D^{-1/2}U^{\mathrm{T}}(x-b)\|^2\right), \end{aligned}$$

so

$$W = D^{-1/2}U^{\mathrm{T}}(X-b) \sim \mathcal{N}(0,I).$$

The $p\%$ probability region for W is a disc D_α of radius α:

$$\begin{aligned} \mathrm{P}\{W \in D_\alpha\} &= \frac{1}{2\pi} \int_{D_\alpha} \exp\left(-\frac{1}{2}\|w\|^2\right) dw \\ &= \frac{1}{2\pi} \int_0^\alpha \int_0^{2\pi} e^{-r^2/2} d\theta r dr \\ &= 1 - e^{-\alpha^2/2} = \frac{p}{100}, \end{aligned}$$

which is equivalent to

$$\alpha = \sqrt{2 \log \frac{100}{100 - p}}.$$

Given a sample $\{x_1, x_2, \ldots, x_N\}$, we

- calculate the sample $\{x_1, x_2, \ldots, x_N\}$,

$$w_j = D^{-1/2} U^{\mathrm{T}} (x_j - b),$$

- calculate the relative amount of these sample points are within the disc D_α,

$$r_p(N) = \frac{100}{N} \#\{w_j \mid \|w_j\| < \alpha\}.$$

When N grows, we should have

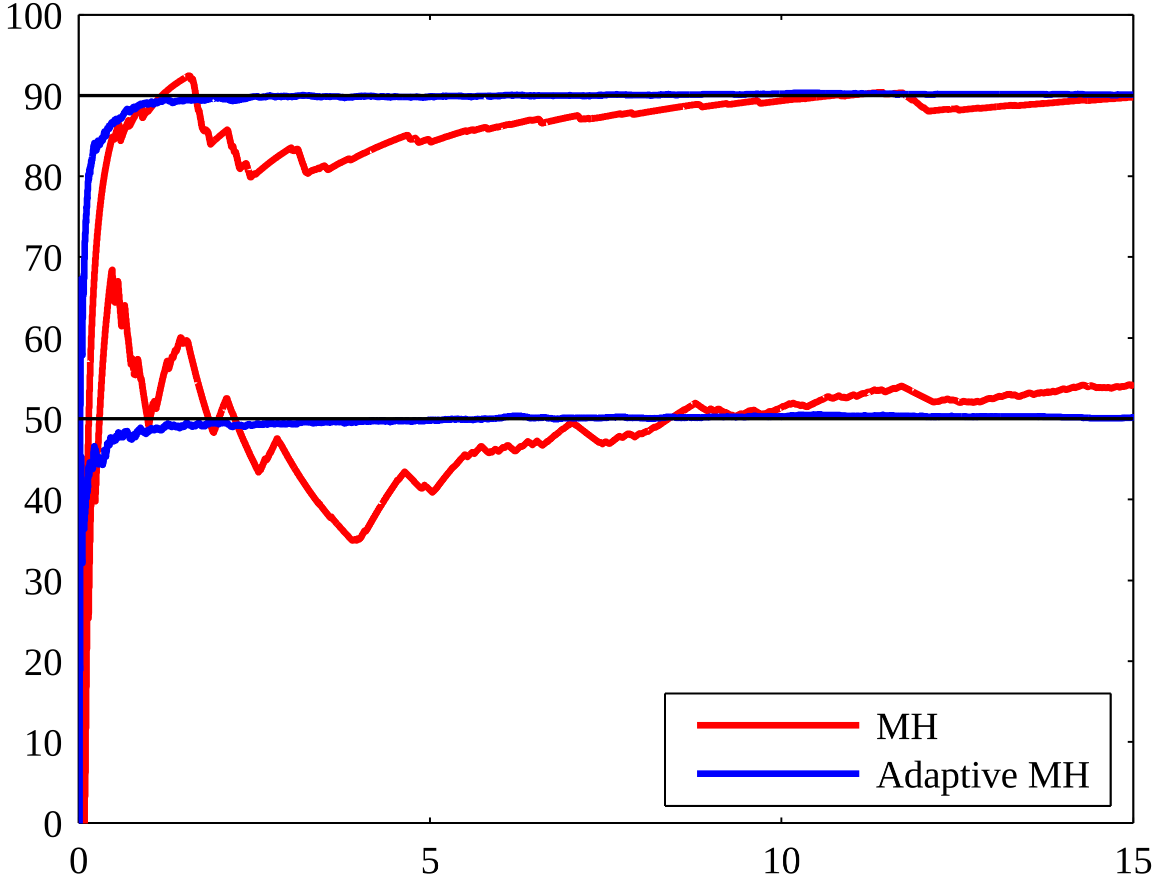
$$r_p(N) \to p.$$

Program

```
% Number of points within a p percent ellipse

 W = diag(1 ./sqrt(d))*U'*(SampleNA-b*ones(1,nsample));
 normW = sqrt(sum(W.^2));
 p = [90,50];

 for j = 1:2
    alpha = sqrt(2*log(100/(100-p(j))));
    xinside = (normW<alpha);
    reln = 100*cumsum(xinside)./[1:nsample];
end
```


100
90
80
70
60
50
40
30
20
10
0
0
5
10
15
MH
Adaptive MH

Example: inverse problems in chemical engineering

Recall the reversible chemical reactions

$$A \rightleftarrows B,$$

with reaction rates k_1 and k_2, respectively.

Concentrations C_A and C_B satisfy

$$\begin{aligned} \frac{dC_A}{dt} &= -k_1 C_A + k_2 C_B \\ \frac{dC_B}{dt} &= k_1 C_A - k_2 C_B, \end{aligned}$$

with initial data

$$C_A(0) = C_{A,0}, \quad C_B(0) = C_{b,0}.$$

Inverse problem

Assume that we know the initial concentrations.

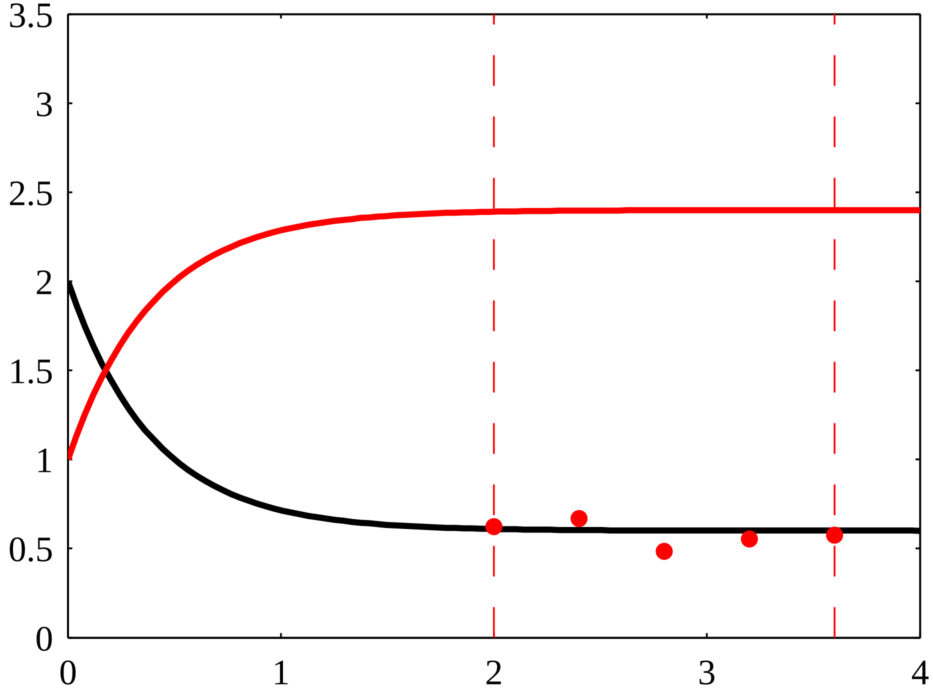
Data: For $0 < t_1 < t_2 \cdots < t_n$, measure $C_A(t_j)$, $1 \leq j \leq n$.

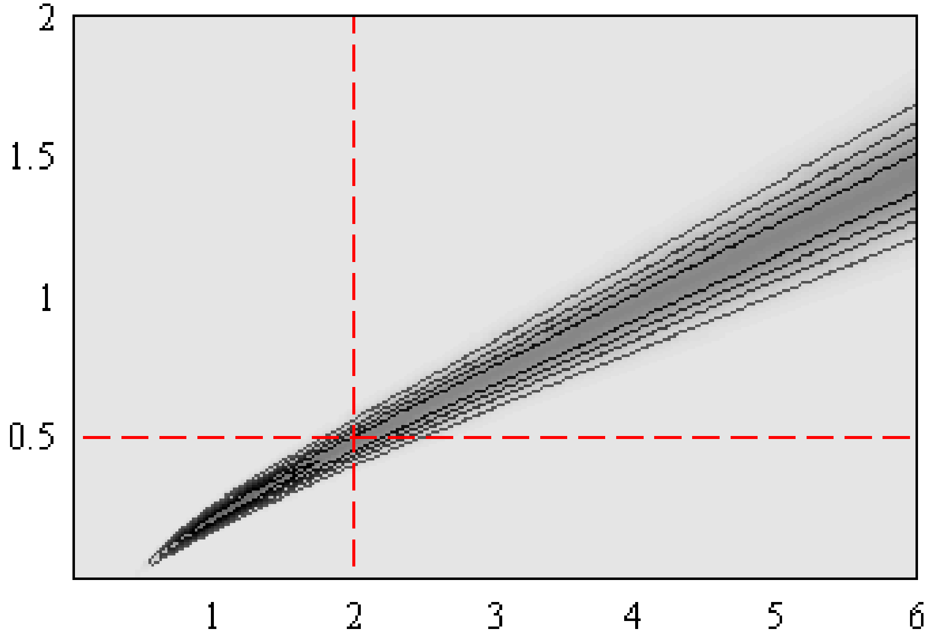
Estimate k_1 and k_2.

Noisy observations:

$$b_j = C_A(t_j) + e_j, \quad e_j \sim \mathcal{N}(0, \sigma^2).$$

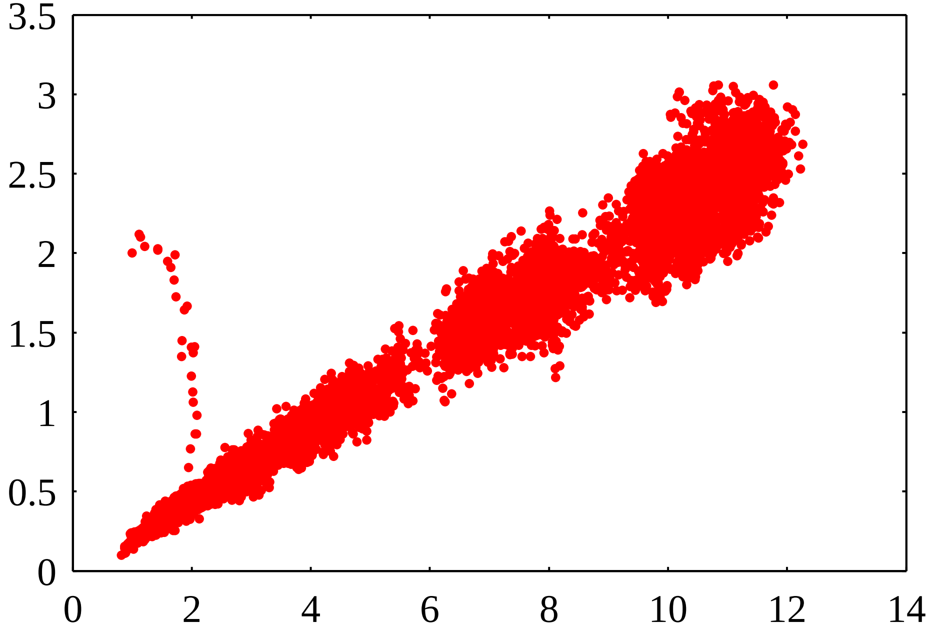
Steady state data

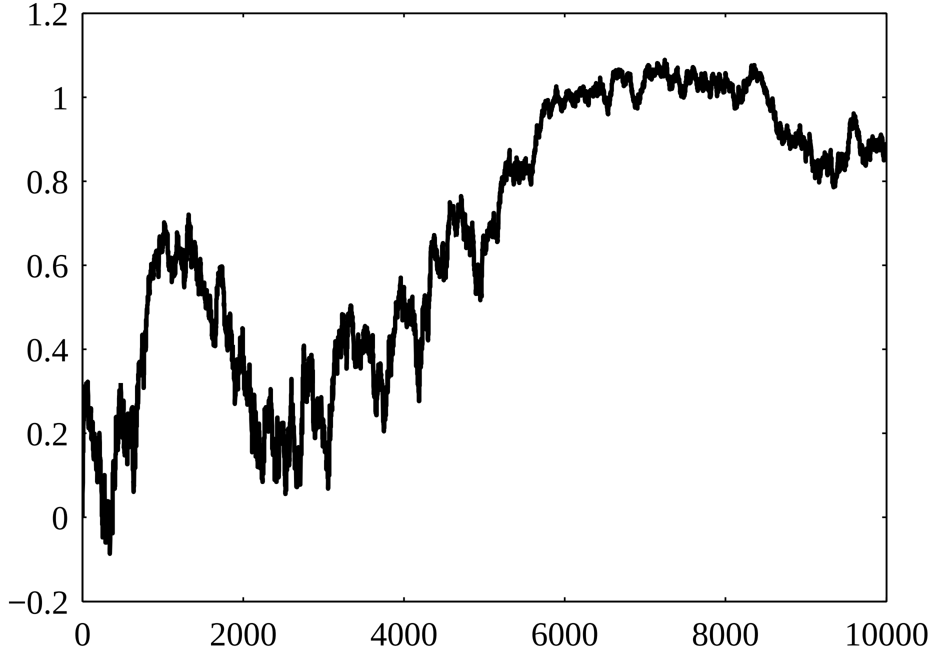




$$C_{A,0} = 2, \quad C_{B,0} = 1, \quad k_1 = 2, \quad k_2 = 0.5, \quad \sigma = 0.2$$

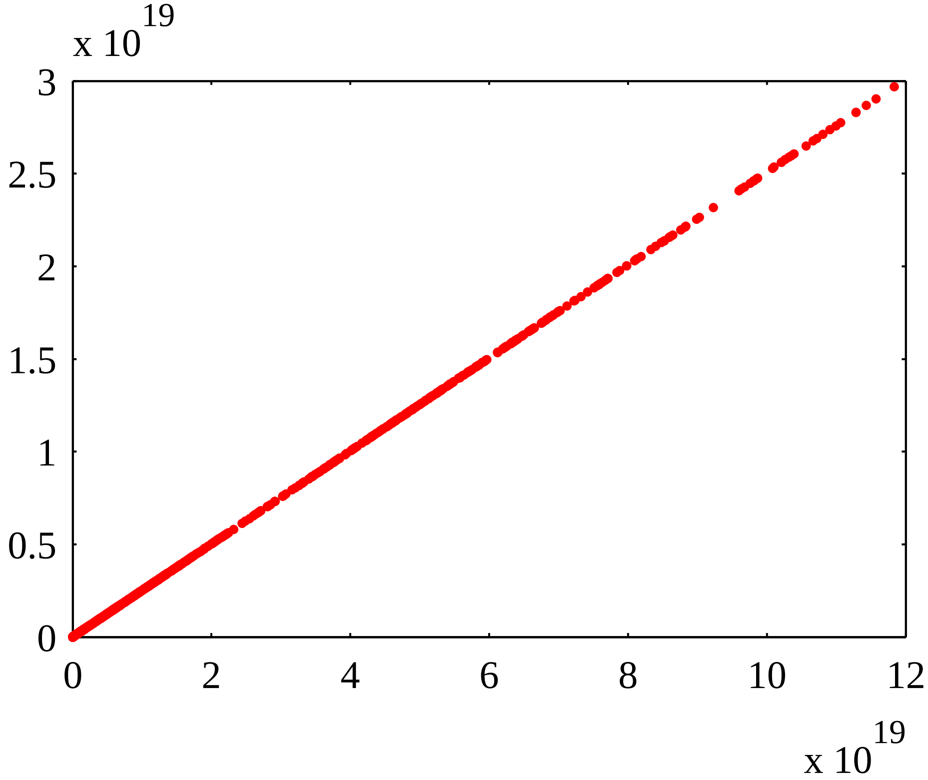
Sampling with non-adaptive MH

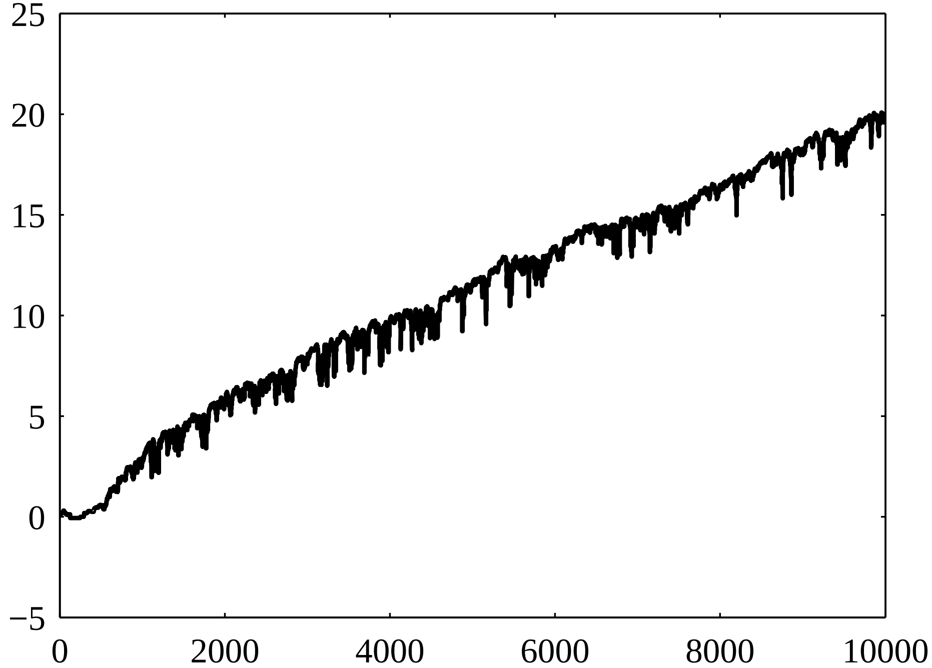




Initial point $(1, 2)$

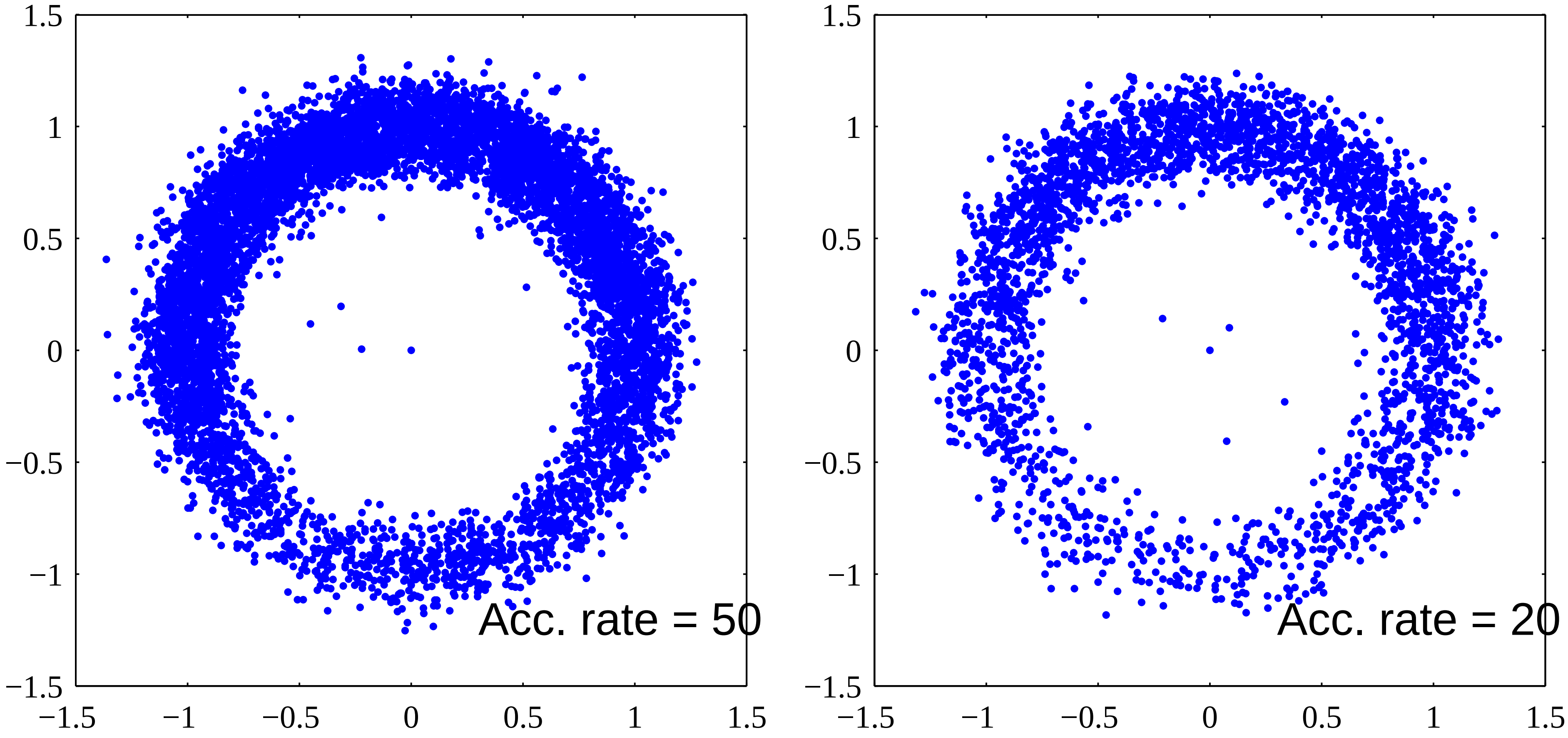
Sampling with adaptive MH



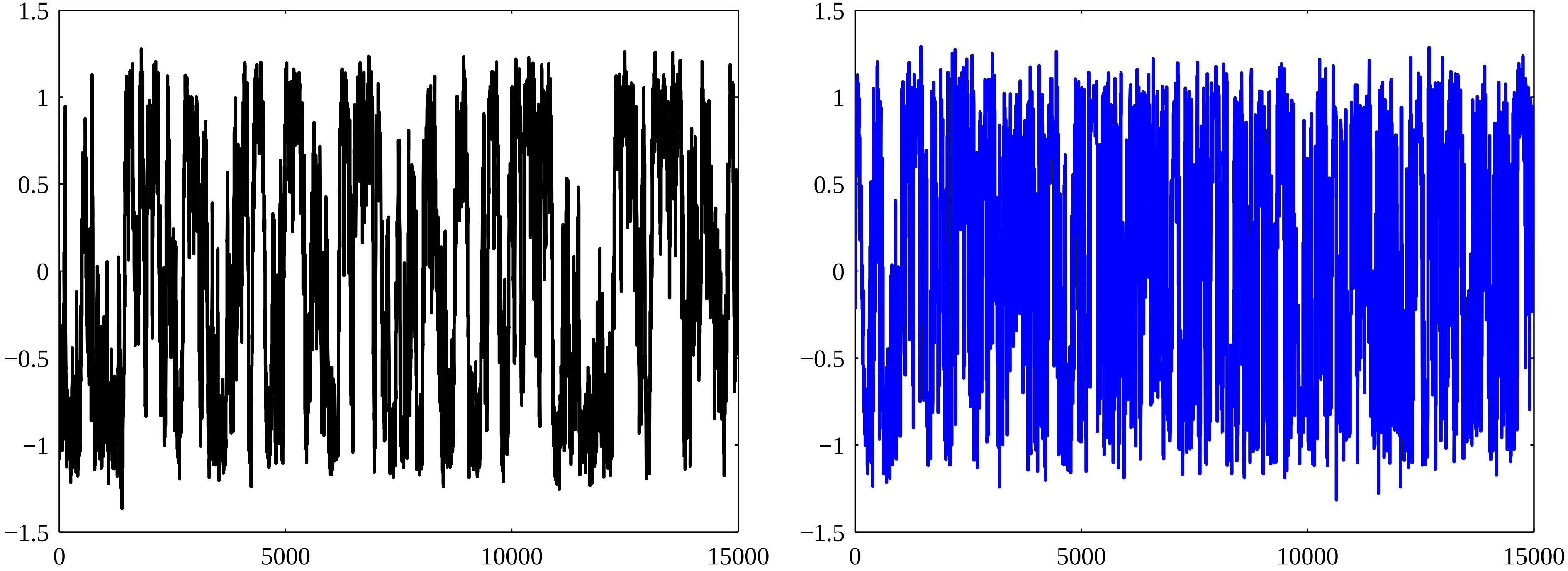


Adaptation after every 100 sample points

Example: Horseshoe distribution



Non-adaptive (left) vs. adaptive (right).



Non-adaptive (left) vs. adaptive (right).

Observations

In this example, the adaptation, as defined here, is not of great help:

The distribution is almost circular, so the asymptotic covariance is almost an identity, and we end up drawing essentially from a white noise density.

The only advantage is that the step length need no tuning.