Inverse Problems in the Bayesian Framework

Daniela Calvetti
Case Western Reserve University
Cleveland, Ohio

Raleigh, NC, July 2016

Bayes' Formula

Stochastic model: Two random variables $X \in \mathbb{R}^n$, $B \in \mathbb{R}^m$, where

- B is the observed quantity,
- X is the quantity of primary interest.

Goal: Find the *posterior probability density*,

$\pi_X(x \mid b) =$ density of X, given the observation $B = b$, $b = b_{\text{observed}}$.

Bayes' Formula

Prior information: Given the prior density $\pi_X(x)$, encoding our prior information – or prior belief – about the possible vaues of X.
Likelihood: Assuming that $X = x$, what would the forward model predict for the value distribution of B? Encode this information in $\pi_B(b \mid x)$.
Evidence: With prior and likelihood, compute

$$\pi_B(b) = \int_{\mathbb{R}^n} \pi_{XB}(x, b)dx = \int_{\mathbb{R}^n} \pi_B(b \mid x)\pi_X(x)dx.$$

Bayes' formula for probability densities

$$\pi_X(x \mid b) = \frac{\pi_B(b \mid x)\pi_X(x)}{\pi_B(b)}.$$

Linear Inverse Problems

Consider the problem of estimating x from

$$b = \mathrm{A}x + e, \quad \mathrm{A} \in \mathbb{R}^{m \times n}.$$

Stochastic extension: Write

$$B = \mathrm{A}X + E,$$

and assume the noise and prior model

$$X \sim \mathcal{N}(0, \mathrm{D}), \quad E \sim \mathcal{N}(0, \mathrm{C}).$$

Usually it is assumed that X and E are independent. In particular,

$$\mathrm{E}\{XE^{\mathsf{T}}\} = \mathrm{E}\{X\}\mathrm{E}\{E\}^{\mathsf{T}} = 0.$$

Linear Inverse Problems

However, this is not necessary, and we may have

$$\mathrm{E}\{XE^{\mathsf{T}}\} = \mathrm{R} \in \mathbb{R}^{n \times m}.$$

Define a new random variable

$$Z = \begin{bmatrix} X \\ B \end{bmatrix} \in \mathbb{R}^{n+m}.$$

Covariance matrix of Z:

$$\begin{aligned} ZZ^{\mathsf{T}} &= \begin{bmatrix} X \\ B \end{bmatrix} \begin{bmatrix} X^{\mathsf{T}} & B^{\mathsf{T}} \end{bmatrix} \\ &= \begin{bmatrix} XX^{\mathsf{T}} & XB^{\mathsf{T}} \\ BX^{\mathsf{T}} & BB^{\mathsf{T}} \end{bmatrix}. \end{aligned}$$

Compute the expectation of this matrix.

Linear Inverse Problems

Expectations:

$$\mathrm{E}\{XX^{\mathsf{T}}\} = \mathrm{D},$$

$$\begin{aligned}\mathrm{E}\{XB^{\mathsf{T}}\} &= \mathrm{E}\{X(\mathrm{A}X+E)^{\mathsf{T}}\} = \mathrm{E}\{XX^{\mathsf{T}}\mathrm{A}^{\mathsf{T}} + XE^{\mathsf{T}}\} \\ &= \mathrm{E}\{XX^{\mathsf{T}}\}\mathrm{A}^{\mathsf{T}} + \mathrm{E}\{XE^{\mathsf{T}}\} = \mathrm{D}\mathrm{A}^{\mathsf{T}} + \mathrm{R}.\end{aligned}$$

Furthermore,

$$\mathrm{E}\{BX^{\mathsf{T}}\} = \mathrm{E}\{XB^{\mathsf{T}}\}^{\mathsf{T}} = \mathrm{A}\mathrm{D} + \mathrm{R}^{\mathsf{T}},$$

Linear Inverse Problems

and, finally,

$$\begin{aligned} \mathrm{E}\{BB^{\mathsf{T}}\} &= \mathrm{E}\{(\mathrm{A}X+E)(\mathrm{A}X+E)^{\mathsf{T}}\} \\ &= \mathrm{E}\{\mathrm{A}XX^{\mathsf{T}}\mathrm{A}^{\mathsf{T}}+EX^{\mathsf{T}}\mathrm{A}^{\mathsf{T}}+\mathrm{A}XE^{\mathsf{T}}+EE^{\mathsf{T}}\} \\ &= \mathrm{A}\mathrm{D}\mathrm{A}^{\mathsf{T}}+\mathrm{R}^{\mathsf{T}}\mathrm{A}^{\mathsf{T}}+\mathrm{A}\mathrm{R}+\mathrm{C}. \end{aligned}$$

Conclusion:

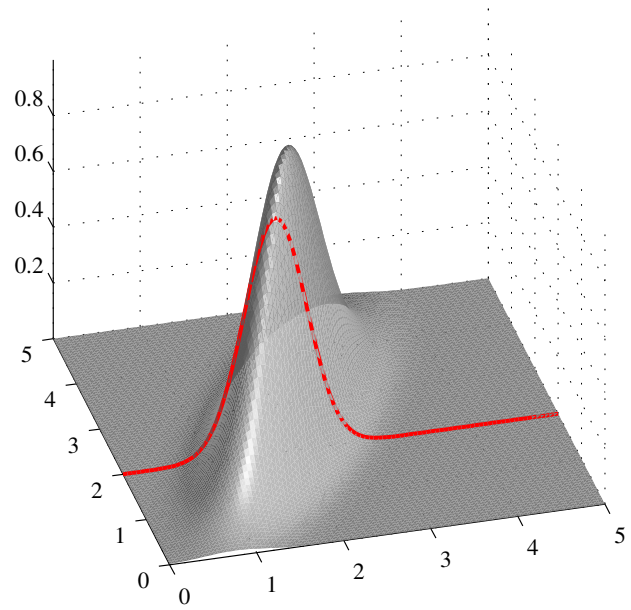
$$\mathrm{Cov}(Z)=\begin{bmatrix} \mathrm{D} & \mathrm{D}\mathrm{A}^{\mathsf{T}}+\mathrm{R} \\ \mathrm{A}\mathrm{D}+\mathrm{R}^{\mathsf{T}} & \mathrm{A}\mathrm{D}\mathrm{A}^{\mathsf{T}}+\mathrm{R}^{\mathsf{T}}\mathrm{A}^{\mathsf{T}}+\mathrm{A}\mathrm{R}+\mathrm{C} \end{bmatrix}=\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}.$$

Schur Complements and Conditioning

Given a Gaussian random variable

$$Z = \begin{bmatrix} X \\ B \end{bmatrix} \in \mathbb{R}^{n+m}$$

with covariance $\Gamma \in \mathbb{R}^{(m+n)\times(m+n)}$, what is the probability density of X, given $B = b$?



Schur Complements and Conditioning

Assume that $X \sim \mathcal{N}(0, \Gamma)$, where $\Gamma \in \mathbb{R}^{n \times n}$ is a given SPD matrix. Partitioning of X,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^k \\ \in \mathbb{R}^{n-k} \end{matrix} .$$

Question: *Assume that $X_2 = x_2$ is observed. What is the conditional probability density of X_1,*

$$\pi_{X_1}(x_1 \mid x_2) = ?$$

Schur Complements and Conditioning

Write

$$\pi_X(x) = \pi_{X_1,X_2}(x_1, x_2).$$

Bayes' formula: The distribution of unknown part x_1 provided that x_2 is known, is

$$\pi_{X_1}(x_1 \mid x_2) \propto \pi_{X_1,X_2}(x_1, x_2), \quad x_2 = x_{2,\text{observed}}.$$

In terms of the Gaussian density,

$$\pi_{X_1,X_2}(x_1, x_2) \propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\Gamma^{-1}x\right). \tag{1}$$

Schur Complements and Conditioning

Partitioning of the covariance matrix:

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \in \mathbb{R}^{n \times n}, \tag{2}$$

where

$$\Gamma_{11} \in \mathbb{R}^{k \times k}, \quad \Gamma_{22} \in \mathbb{R}^{(n-k) \times (n-k)}, \quad k < n,$$

and

$$\Gamma_{12} = \Gamma_{21}^{\mathsf{T}} \in \mathbb{R}^{k \times (n-k)}.$$

Schur Complements and Conditioning

Precision matrix $B = \Gamma^{-1}$.
Partition B:

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{3}$$

Quadratic form $x^T B x$ appearing in the exponential:

$$Bx = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} B_{11}x_1 + B_{12}x_2 \\ B_{21}x_1 + B_{22}x_2 \end{bmatrix},$$

Schur Complements and Conditioning

$$
\begin{aligned}
x^{\mathsf{T}} B x &= \begin{bmatrix} x_1^{\mathsf{T}} & x_2^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} B_{11} x_1 + B_{12} x_2 \\ B_{21} x_1 + B_{22} x_2 \end{bmatrix} \\
&= x_1^{\mathsf{T}} \left(B_{11} x_1 + B_{12} x_2 \right) + x_2^{\mathsf{T}} \left(B_{21} x_1 + B_{22} x_2 \right) \\
&= x_1^{\mathsf{T}} B_{11} x_1 + 2 x_1^{\mathsf{T}} B_{12} x_2 + x_2^{\mathsf{T}} B_{22} x_2 \\
&= \left(x_1 + B_{11}^{-1} B_{12} x_2 \right)^{\mathsf{T}} B_{11} \left(x_1 + B_{11}^{-1} B_{12} x_2 \right) \\
&\qquad + \underbrace{x_2^{\mathsf{T}} \left(B_{22} - B_{21} B_{11}^{-1} B_{12} \right) x_2}_{\text{independent of } x_1} .
\end{aligned}
$$

Schur Complements and Conditioning

From this key equation for conditional densities it follows that

$$\pi_{X_1}(x_1 \mid x_2) \propto \exp\left(-\frac{1}{2}\left(x_1 + \mathsf{B}_{11}^{-1}\mathsf{B}_{12}x_2\right)^{\mathsf{T}}\mathsf{B}_{11}\left(x_1 + \mathsf{B}_{11}^{-1}\mathsf{B}_{12}x_2\right)\right).$$

Thus the conditional density is Gaussian, with mean

$$\boxed{\bar{x}_1 = -\mathsf{B}_{11}^{-1}\mathsf{B}_{12}x_2,}$$

and covariance matrix

$$\boxed{\mathsf{C} = \mathsf{B}_{11}^{-1}.}$$

Question: *How to express these formulas in terms of* Γ*?*

Schur Complements and Conditioning

Consider a partitioned SPD matrix $\Gamma \in \mathbb{R}^{n\times n}$.
For any $v \in \mathbb{R}^k$, $x \neq 0$

$$v^T \Gamma_{11} v = \begin{bmatrix} v^T & 0 \end{bmatrix} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \begin{bmatrix} v^T \\ 0 \end{bmatrix} > 0,$$

showing the positive definiteness of Γ_{11}.
The same holds for Γ_{22}.
In particular, Γ_{11} and Γ_{22} are invertible.

Schur Complements and Conditioning

To calculate the inverse of Γ, we solve the equation

$$\Gamma x = y$$

in block form.
By partitioning,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^k \\ \in \mathbb{R}^{n-k} \end{matrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^k \\ \in \mathbb{R}^{n-k} \end{matrix}.$$

we have

$$\begin{aligned} \Gamma_{11} x_1 + \Gamma_{12} x_2 &= y_1, \\ \Gamma_{21} x_1 + \Gamma_{22} x_2 &= y_2. \end{aligned}$$

Schur Complements and Conditioning

Eliminate x_2 from the second equation,

$$x_2 = \Gamma_{22}^{-1}\big(y_2 - \Gamma_{21}x_1\big),$$

substitute back into the first equation:

$$\Gamma_{11}x_1 + \Gamma_{12}\Gamma_{22}^{-1}\big(y_2 - \Gamma_{21}x_1\big) = y_1,$$

and by rearranging the terms,

$$\big(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\big)x_1 = y_1 - \Gamma_{12}\Gamma_{22}^{-1}y_2.$$

Define the *Schur complement* of Γ_{22}:

$$\widetilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$$

Schur Complements and Conditioning

It can be shown that $\widetilde{\Gamma}_{22}$ must be invertible, and therefore

$$x_1 = \widetilde{\Gamma}_{22}^{-1} y_1 - \widetilde{\Gamma}_{22}^{-1} \Gamma_{12} \Gamma_{22}^{-1} y_2.$$

Similarly, interchanging the roles of x_1 and x_2,

$$x_2 = \widetilde{\Gamma}_{11}^{-1} y_2 - \widetilde{\Gamma}_{11}^{-1} \Gamma_{21} \Gamma_{11}^{-1} y_1,$$

where

$$\widetilde{\Gamma}_{11} = \Gamma_{22} - \Gamma_{21} \Gamma_{11}^{-1} \Gamma_{12}$$

is the Schur complement of Γ_{11}.

In matrix form:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \widetilde{\Gamma}_{22}^{-1} & -\widetilde{\Gamma}_{22}^{-1} \Gamma_{12} \Gamma_{22}^{-1} \\ -\widetilde{\Gamma}_{11}^{-1} \Gamma_{21} \Gamma_{11}^{-1} & \widetilde{\Gamma}_{11}^{-1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Schur Complements and Conditioning

Conclusion:

$$\Gamma^{-1} = \begin{bmatrix} \widetilde{\Gamma}_{22}^{-1} & -\widetilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ -\widetilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & \widetilde{\Gamma}_{11}^{-1} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Thus *The conditional density* $\pi_{X_1}(x_1 \mid x_2)$ *is a Gaussian*

$$\pi_{X_1}(x_1 \mid x_2) \sim \mathcal{N}(\overline{x}_1, C),$$

where

$$\boxed{\overline{x}_1 = -B_{11}^{-1}B_{12}x_2 = \Gamma_{12}\Gamma_{22}^{-1}x_2,}$$

and

$$\boxed{C = B_{11}^{-1} = \widetilde{\Gamma}_{22}.}$$

Linear Inverse Problems

For simplicity, let us assume that $R = 0$.
Posterior density $\pi_X(x \mid b)$ is a Gaussian density, with mean

$$\boxed{\overline{x} = \Gamma_{12}\Gamma_{22}^{-1} b = DA^T\left(ADA^T + C\right)^{-1} b,}$$

and covariance

$$\boxed{\Phi = \widetilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21} = D - DA^T\left(ADA^T + C\right)^{-1} AD.}$$

Example: Numerical differentiation

$$f(t) = \int_0^t g(\tau)d\tau + \text{noise}.$$

Discretization:

$$b = \mathrm{A}x + e,$$

where

$$\mathrm{A} = \frac{1}{n}\begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & & & 1 \end{bmatrix}.$$

Stochastic extension

Stochastic model

$$B = \mathsf{A}X + E.$$

Model for noise: Independent components,

$$E_j \sim \mathcal{N}(0, \sigma^2), \quad 1 \leq j \leq n.$$

Probability density:

$$\pi_E(e) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|e\|^2\right).$$

Likelihood:

$$\pi_B(b \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}\|b - \mathsf{A}x\|^2\right).$$

Autoregressive Prior Models

Discrete model,

$$x_j = g(t_j), \quad t_j = \frac{j}{n}, \quad 0 \leq j \leq n,$$

Consider two possible prior models:

1. We know that $x_0 = 0$, and believe that the absolute value of the slope of g is bounded by some $m_1 > 0$.
2. We know that $x_0 = x_n = 0$ and believe that the curvature of g is bounded by some $m_2 > 0$.

Autoregressive models

1. Slope:

$$g'(t_j) \approx \frac{x_j - x_{j-1}}{h}, \quad h = \frac{1}{n},$$

Prior information: We believe that

$$|x_j - x_{j-1}| \leq h\, m_1 \text{ with some uncertainty.}$$

2. Curvature:

$$g'(t_j) \approx \frac{x_{j-1} - 2x_j + x_{j+1}}{h^2}.$$

Prior information: We believe that

$$|x_{j-1} - 2x_j + x_{j+1}| \leq h^2 m_2 \text{ with some uncertainty.}$$

Autoregressive model

In both cases, we assume that x_j is a realization of a random variable X_j.
Boundary conditions:

1. $X_0 = 0$ with certainty. Probabilistic model for X_j, $1 \leq j \leq n$.
2. $X_0 = X_n = 0$ with certainty. Probabilistic model for X_j, $1 \leq j \leq n-1$.

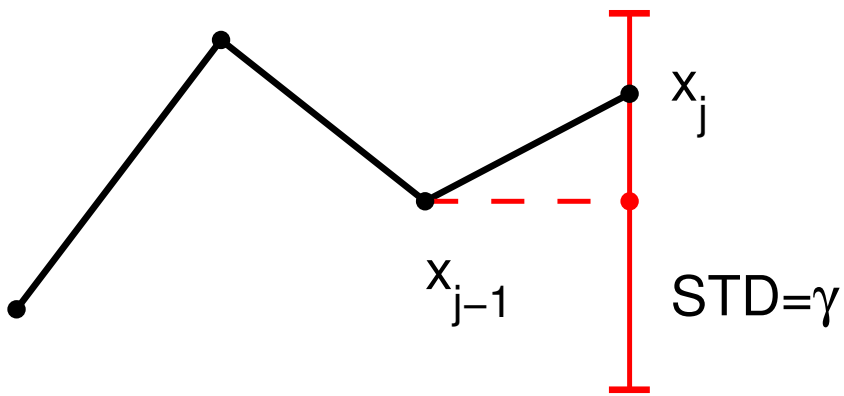
Autoregressive prior models

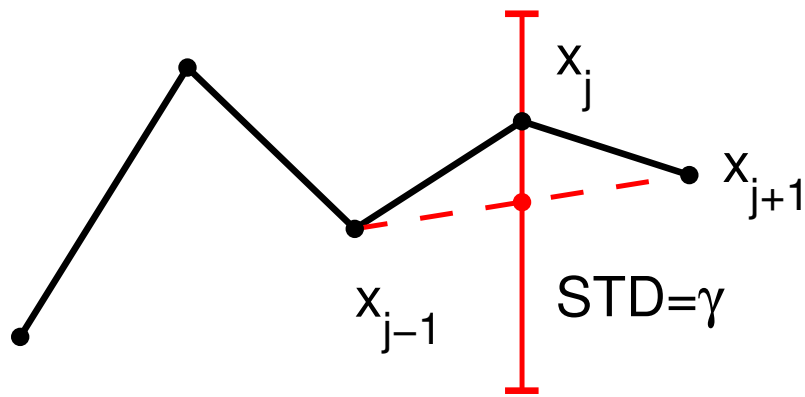
1. First order prior:

$$X_j = X_{j-1} + \gamma W_j, \quad W_j \sim \mathcal{N}(0,1), \quad \gamma = h\, m_1.$$

2. Second order prior:

$$X_j = \frac{1}{2}(X_{j-1} + X_{j+1}) + \gamma W_j, \quad W_j \sim \mathcal{N}(0,1), \quad \gamma = \frac{1}{2} h^2\, m_2.$$





Matrix form: first order model

System of equations:

$$\begin{aligned} X_1 = X_1 - X_0 &= \gamma W_1 \\ X_2 - X_1 &= \gamma W_2 \\ \vdots &\quad \vdots \\ X_n - X_{n-1} &= \gamma W_n \end{aligned}$$

$$\mathsf{L}_1 = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}.$$

$$\mathsf{L}_1 X = \gamma W, \quad W \sim \mathcal{N}(0, \mathsf{I}_n),$$

Matrix form: second order model

System of equations:

$$\begin{aligned} X_2 - 2X_1 = X_2 - 2X_1 + X_0 &= \gamma W_1 \\ X_3 - 2X_2 + X_1 &= \gamma W_2 \\ \vdots &\quad \vdots \\ -2X_{n-1} - X_{n-2} = X_n - 2X_{n-1} + X_{n-2} &= \gamma W_{n-1} \end{aligned}$$

$$\mathsf{L}_2 = \begin{bmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \\ & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-1} \end{bmatrix},$$

$$\mathsf{L}_2 X = \gamma W, \quad W \sim \mathcal{N}(0, \mathsf{I}_{n-1}),$$

Prior density

Given a model

$$\mathsf{L}X = \gamma W, \quad W \sim \mathcal{N}(0, \mathsf{I}),$$

that is,

$$\pi_W(w) \propto \exp\left(-\frac{1}{2}\|w\|^2\right),$$

we conclude that

$$\pi_X(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|\mathsf{L}x\|^2\right) = \exp\left(-\frac{1}{2}x^\mathsf{T}\left[\frac{1}{\gamma^2}\mathsf{L}^\mathsf{T}\mathsf{L}\right]x\right).$$

The inverse of the covariance matrix $=$ *precision matrix* is

$$\mathsf{D}^{-1} = \frac{1}{\gamma^2}\mathsf{L}^\mathsf{T}\mathsf{L}.$$

Testing a Prior

Question: *Given a covariance* D*, how can we check if the prior corresponds to our expectations?*

Symmetric decomposition of the precision matrix (let $\gamma = 1$ for simplicity):

$$\mathsf{D}^{-1} = \mathsf{L}^{\mathsf{T}}\mathsf{L}.$$

We know that

$$W = \mathsf{L}X \sim \mathcal{N}(0, \mathsf{I}).$$

Sampling of X:

1. Draw a realization $w \sim \mathcal{N}(0, \mathsf{I}_n)$
2. Set $x = \mathsf{L}^{-1}w$.

Random draws from priors

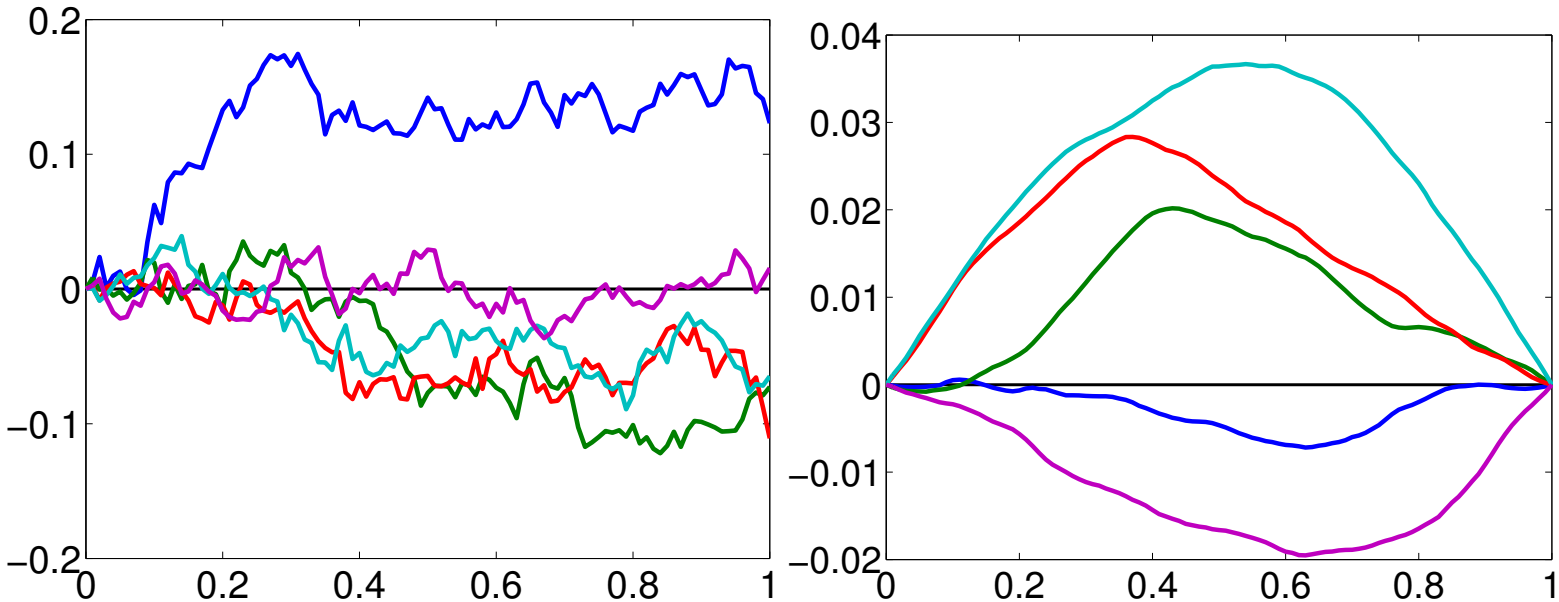
Generate m draws from the prior using the Matlab command `randn`.

```
n = 100;          % number of discretization intervals
t = (0:1/n:1);
m = 5;           % number of draws

% First order model. Boundary condition X_0 = 0

L1    = diag(ones(1,n),0) - diag(ones(1,n-1),-1);
gamma = 1/n;  % m_1 = 1
W     = gamma*randn(n,m);
X     = L1\W;
```

Plots of the random draws



Belief envelopes

Diagonal elements of the posterior covariance:

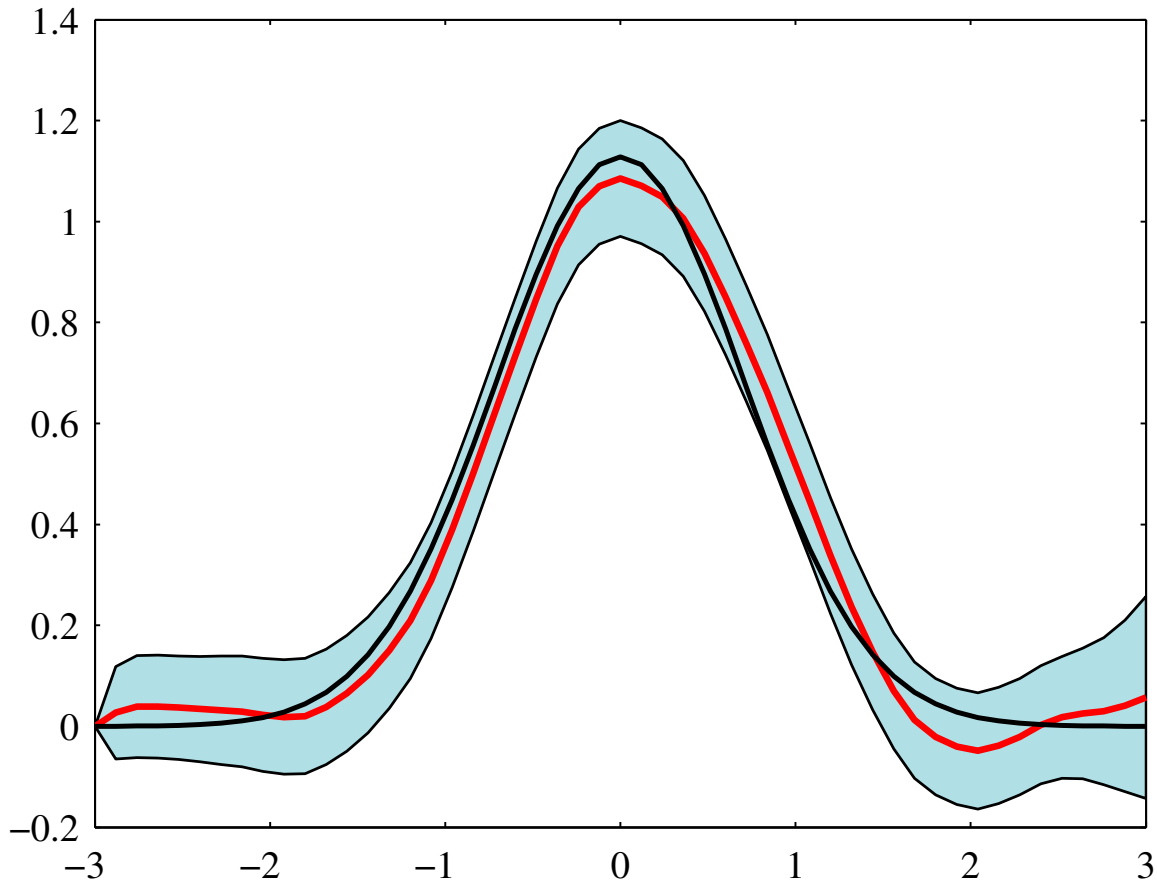
$$\Gamma_{jj} = \eta_j^2 = \text{posterior variance of } X_j.$$

Posterior belief:

$$\overline{x}_j - 2\eta_j < X_j < \overline{x}_j + 2\eta_j$$

with posterior probability $\approx 95\%$.

Bayesian solution



Mean solution and 2 STD belief envelope

Linear Inverse Problems

An alternative (but equivalent) formula for the posterior:

Prior,

$$\pi_X(x) \propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\mathsf{D}^{-1}x\right),$$

and likelihood,

$$\pi_B(b \mid x) \propto \exp\left(-\frac{1}{2}(b-\mathsf{A}x)^{\mathsf{T}}\mathsf{C}^{-1}(b-\mathsf{A}x)\right).$$

Posterior density:

$$\pi_X(x \mid b) \propto \pi_X(x)\pi_B(b \mid x) \propto \exp\left(-\frac{1}{2}Q(x)\right),$$

Linear Inverse Problems

Quadratic term in the exponential:

$$Q(x) = (b - \mathrm{A}x)^{\mathsf{T}} \mathrm{C}^{-1} (b - \mathrm{A}x) + x^{\mathsf{T}} \mathrm{D}^{-1} x$$

Collect the terms of the same order in x together:

$$Q(x) = x^{\mathsf{T}} \underbrace{\left(\mathrm{A}^{\mathsf{T}} \mathrm{C}^{-1} \mathrm{A} + \mathrm{D}^{-1}\right)}_{=\mathrm{M}} x - 2x^{\mathsf{T}} \mathrm{A}^{\mathsf{T}} \mathrm{C}^{-1} b + b^{\mathsf{T}} \mathrm{C}^{-1} b.$$

Complete the square:

$$Q(x) = (x^{\mathsf{T}} - \mathrm{M}^{-1} \mathrm{A}^{\mathsf{T}} \mathrm{C}^{-1} b)^{\mathsf{T}} \mathrm{M} (x^{\mathsf{T}} - \mathrm{M}^{-1} \mathrm{A}^{\mathsf{T}} \mathrm{C}^{-1} b) + \ldots$$

Linear Inverse Problems

Conclusion: The posterior mean and covariance have alternative expressions,

$$\boxed{\overline{x} = \left(\mathsf{A}^{\mathsf{T}}\mathsf{C}^{-1}\mathsf{A} + \mathsf{D}^{-1}\right)^{-1}\mathsf{A}^{\mathsf{T}}\mathsf{C}^{-1}b}$$

and

$$\boxed{\Phi = \left(\mathsf{A}^{\mathsf{T}}\mathsf{C}^{-1}\mathsf{A} + \mathsf{D}^{-1}\right)^{-1}.}$$

The formula for $\overline{x}$ is also known as *Wiener filtered solution*.

Tikhonov Regularization Revisited

Consider the linear model

$$b = \mathsf{A}x + e, \quad e \sim \mathcal{N}(0, \mathsf{C}).$$

Likelihood:

$$\pi_B(b \mid x) \propto \exp\left(-\frac{1}{2}(b - \mathsf{A}x)^{\mathsf{T}}\mathsf{C}^{-1}(b - \mathsf{A}x)\right).$$

Assume a Gaussian prior:

$$X \sim \mathcal{N}(0, \mathsf{D}),$$

or, in terms of densities,

$$\pi_X(x) \propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\mathsf{D}^{-1}x\right).$$

Tikhonov Regularization Revisited

Bayes' formula:

$$\begin{aligned} p_X(x \mid b) &\propto p_X(x) p_B(b \mid x) \\ &= \exp\left(-\frac{1}{2} x^{\mathsf{T}} \mathsf{D}^{-1} x - \frac{1}{2} (b - \mathsf{A}x)^{\mathsf{T}} \mathsf{C}^{-1} (b - \mathsf{A}x) \right). \end{aligned}$$

By writing $\mathsf{D}^{-1} = \mathsf{L}^{\mathsf{T}}\mathsf{L}$, the negative of the exponent is

$$H(x) = \frac{1}{2} \left(\|b - \mathsf{A}x\|_{\mathsf{C}}^2 + \|\mathsf{L}x\|^2 \right),$$

a Tikhonov functional.

Tikhonov Regularization Revisited

Therefore,

$$x_{\mathrm{MAP}} = \operatorname{argmax}\{\pi_X(x \mid b)\} = \operatorname{argmin}\{\|b - \mathsf{A}x\|_{\mathsf{C}}^2 + \|\mathsf{L}x\|^2\}.$$

The Tikhonov regularization parameter is absorbed in the prior covariance matrix as well as the noise covariance matrix.

Non-Gaussian models

The Gaussian models are insufficient when

- the forward model is non-linear,
- the prior is non-Gaussian,
- the noise is non-additive,
- the noise is non-Gaussian.

Monte Carlo Integration

Assume that a probability density π_X is given in $\mathbb{R}^n$.
Problem: *Estimate numerically an integral of type*

$$\mathrm{E}\{f(X)\} = \int_{\mathbb{R}^n} f(x)\pi_X(x)dx = ?$$

- Expectation of X: $f(x) = x$.
- Covariance of X: $f(x) = (x - \overline{x})(x - \overline{x})^{\mathsf{T}}$.

Monte Carlo Integration

Difficulties with numerical integration using quadrature methods:

- We may not know the support of μ_X (Support: The set in which the function is not vanishing). Where should we put our quadrature points?
- If n is large, an integration grid becomes huge: K points/direction means K^n grid points.

Try Monte Carlo integration!

Monte Carlo Integration

Example: Given a two-dimensional set $\Omega \subset \mathbb{R}^2$. Estimate the area of Ω.
Raindrop integration: Assume that $\Omega \subset Q = [0, a] \times [0, b]$.
Draw points from uniform density over Q:

$$\{x^1, x^2, \ldots, x^N\}, \quad x^j \sim \text{Uniform}(Q).$$

Estimate of the area $|\Omega|$:

$$\frac{|\Omega|}{|Q|} = \frac{|\Omega|}{ab} \approx \frac{\#\text{ of points } x^j \in \Omega}{N},$$

solve for $|\Omega|$.

Monte Carlo Integration

The approximation corresponds to Monte Carlo integral

$$\frac{|\Omega|}{|Q|} = \frac{1}{|Q|}\int_Q \chi_\Omega(x)dx \approx \frac{1}{N}\sum_{j=1}^{N}\chi_\Omega(x^j),$$

where

$$\chi_\Omega(x) = \begin{cases} 1 & \text{if } x \in \Omega \\ 0 & \text{if} x \notin \Omega \end{cases}$$

and $1/N$ is the equal weight that every point x^j has.

$$\frac{1}{|Q|}\chi_Q(x) = \text{uniform density over } Q$$

Monte Carlo Integration

Generalize: Given a probability density π_X, write

$$\int_{\mathbb{R}^n} f(x)\pi_X(x)dx \approx \frac{1}{N}\sum_{j=1}^{N} f(x^j),$$

where

$$\{x^1, x^2, \ldots, x^N\}$$

is drawn independently from the probability distribution π_X.

Problem: *How does one draw from a probability density in* $\mathbb{R}^n$*?*

Sampling and Markov chains: Random walk

Random walk is a process of moving around by taking random steps.
Most elementary random walk:

1. Start at a point of your choice $x_0 \in \mathbb{R}^n$.
2. Draw a random vector $w_1 \sim \mathcal{N}(0, I)$ and set $x_1 = x_0 + \sigma w_1$.
3. Repeat the process: Set $x_{k+1} = x_k + \sigma w_{k+1}$, $w_{k+1} \sim \mathcal{N}(0, I)$.

Sampling and Markov chains: Random walk

In terms of random variables:

$$X_{k+1} = X_k + \sigma W_{k+1}, \quad W_{k+1} \sim \mathcal{N}(0, \mathsf{I}_n).$$

The conditional density of X_{k+1}, given $X_k = x_k$ is

$$\pi(x_{k+1} \mid x_k) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_k - x_{k+1}\|^2\right) = q_k(x_k, x_{k+1}).$$

The function q_k is called the *kth transition kernel*.

Sampling and Markov chains: Random walk

- Since
$$q_0 = q_1 = q_2 = \dots,$$
i.e., the step is always equally distributed, we call the random walk *time invariant*. (k = time).
- *The chain*
$$\{X_k,\ k = 0, 1, \cdots\}$$
of random variables, is called *discrete time stochastic process*.
- Particular feature: the probability distribution X_k *depends of the past only through the previous member* X_{k-1}:
$$\pi(x_{k+1} \mid x_0, x_1, \dots, x_k) = \pi(x_{k+1} \mid x_k).$$
A stochastic process having this property is called a **Markov chain**.

Sampling and Markov chains: Random walk

Given:

- an arbitrary transition kernel q,
- a random variable X with probability density $\pi_X(x) = p(x)$,

generate a new random variable Y by using the kernel $q(x, y)$, that is,

$$\pi(y \mid x) = q(x, y).$$

Question: *What is the probability density of this new variable?* The answer is found by **marginalization**,

$$\pi_Y(y) = \int \pi(y \mid x)\pi_X(x)dx = \int q(x, y)p(x)dx.$$

Sampling and Markov chains: Random walk

If the probability density of the new variable is equal to the one of the old one, that is,

$$\int q(x, y)p(x)dx = p(y),$$

p is called an *invariant density* of the transition kernel q.
The classical problem in the theory of Markov chains is:
Given a transition kernel, find the corresponding invariant density.

Invariant density and sampling

Recall the *sampling problem*:

Given a probability density $p = p(x)$, generate a sample that is distributed according to it.

If we had a transition kernel q with invariant density p, generating such sample from $p(x)$ would be easy:

- Start with some x_0;
- draw x_1 from $q(x_0, x_1)$;
- In general, given x_k, draw x_{k+1} from $q(x_k, x_{k+1})$.

Rephrasing the sampling problem:
Given a probability density p, find a kernel q such that p is its invariant density.

Metropolis–Hastings algorithm

Given a transition density

$$y \mapsto K(x, y), \quad x \in \mathbb{R}^n \text{ current point},$$

consider a Markov process: if $x \in \mathbb{R}^n$ is the current point, we have two possibilities:

1. Stay at x with probability $r(x)$, $0 \leq r(x) < 1$,
2. Move by using a transition kernel $K(x, y)$.

Let x and y be realizations of random variables X, Y.

$$\pi_X(x) = p(x),$$

and y is generated according to the algorithm above.
Question: What is the probability density of Y?

Let

- $\mathcal{A}$ be the event that we opt for moving from x,
- $\neg\mathcal{A}$ be the event of staying put.

The probability of $Y \in B \subset \mathbb{R}^n$ *assuming a move* is

$$\mathrm{P}\{Y \in B \mid X = x, \mathcal{A}\} = \int_B K(x, y)dy.$$

The kernel K is scaled so that

$$\begin{aligned} \mathrm{P}\{X = x, \mathcal{A}\} &= \mathrm{P}\{Y \in \mathbb{R}^n \mid X = x, \mathcal{A}\} \\ &= \int_{\mathbb{R}^n} K(x, y)dy = 1 - r(x). \end{aligned} \tag{4}$$

On the other hand, if we stay put, $Y \in B$ happens only if $X \in B$,

$$\mathrm{P}\{Y \in B \mid X = x, \neg\mathcal{A}\} = r(x)\chi_B(x) = \begin{cases} r(x), & \text{if } x \in B, \\ 0, & \text{if } x \notin B \end{cases},$$

where χ_B is the characteristic function of B.

Hence, **the total probability of arriving from x to B** is

$$\begin{aligned}
&\mathrm{P}\{Y \in B \mid X = x\} \\
&\quad = \mathrm{P}\{Y \in B \mid X = x, \mathcal{A}\} + \mathrm{P}\{Y \in B \mid X = x, \neg\mathcal{A}\} \\
&\quad = \int_B K(x, y)dy + r(x)\chi_B(x).
\end{aligned}$$

Marginalize over x and calculate the probability of $Y \in B$

$$
\begin{aligned}
\mathrm{P}\{Y \in B\} &= \int \mathrm{P}\{Y \in B \mid X = x\} p(x) dx \\
&= \int p(x) \left(\int_B K(x,y) dy \right) dx + \int \chi_B(x) r(x) p(x) dx \\
&= \int_B \left(\int p(x) K(x,y) dx \right) dy + \int_B r(x) p(x) dx \\
&= \int_B \left(\int p(x) K(x,y) dx + r(y) p(y) \right) dy .
\end{aligned}
$$

Since

$$\mathrm{P}\{Y \in B\} = \int_B \pi(y)dy,$$

we must have

$$\pi_Y(y) = \int p(x)K(x,y)dx + r(y)p(y).$$

Our goal is then to find a kernel K such that $\pi_Y(y) = p(y)$, that is

$$p(y) = \int p(x)K(x,y)dx + r(y)p(y),$$

or, equivalently,

$$(1 - r(y))p(y) = \int p(x)K(x,y)dx.$$

Substituting (4) in this formula, with the roles of x and y interchanged, we obtain

$$\int p(y)K(y,x)dx = \int p(x)K(x,y)dx$$

This equation is called the *balance equation*. This holds, in particular, if the integrands are equal,

$$p(y)K(y,x) = p(x)K(x,y).$$

The latter equation is known as *detailed balance equation*.

Metropolis-Hastings algorithm

- Start by a selecting a *proposal distribution*, or *candidate generating kernel* $q(x, y)$;
- The kernel should be chosen so that generating a Markov chain with it is easy.
- A Gaussian kernel is a popular choice.

Metropolis-Hastings algorithm

If q satisfies the detailed balance equation, i.e.,

$$p(y)q(y,x) = p(x)q(x,y),$$

we are done, since p is an invariant density. More likely, the equality does not hold.
If

$$p(y)q(y,x) < p(x)q(x,y). \tag{5}$$

force the detailed balance equation to hold, defining K as

$$K(x,y) = \alpha(x,y)q(x,y),$$

where α is chosen so that

$$p(y)\alpha(y,x)q(y,x) = p(x)\alpha(x,y)q(x,y).$$

Metropolis-Hastings algorithm

The kernel α need not be symmetric, so let

$$\alpha(y, x) = 1.$$

Now the other factor is uniquely determined. We must have

$$\alpha(x, y) = \frac{p(y)q(y, x)}{p(x)q(x, y)} < 1.$$

Observe that if the inequality (5) goes the other way, interchange the roles of x and y, and let $\alpha(x, y) = 1$. In summary

$$K(x, y) = \alpha(x, y)q(x, y), \quad \alpha(x, y) = \min\left\{1, \frac{p(y)q(y, x)}{p(x)q(x, y)}\right\}.$$

Metropolis-Hastings algorithm

Draws in two phases:

1. Given x, draw y using the transition kernel $q(x, y)$.
2. Calculate the *acceptance ratio*,

$$\alpha(x, y) = \frac{p(y)q(y, x)}{p(x)q(x, y)}.$$

3. Flip the α–coin: draw $t \sim \text{Uniform}([0, 1])$; if $\alpha > t$, accept y, otherwise stay where you are.

Metropolis-Hastings algorithm: Random walk proposal

If $q(x, y) = q(y, x)$, the algorithm simplifies:

1. Given x, draw y using the transition kernel $q(x, y)$.
2. Calculate the *acceptance ratio*,

$$\alpha(x, y) = \frac{p(y)}{p(x)}.$$

3. Flip the α–coin: draw $t \sim \text{Uniform}([0, 1])$; if $\alpha > t$, accept y, otherwise stay where you are.

Random walk Metropolis-Hastings

1. Set sample size N. Pick initial point x^1. Set $k = 1$.
2. Propose a new point,

$$y = x^k + \delta w, \quad w \sim \mathcal{N}(0, \mathrm{I}).$$

3. Compute the acceptance ratio,

$$\alpha = \frac{\pi(y)}{\pi(x^k)}.$$

4. Flip α-coin: Draw $\xi \sim \mathrm{Uniform}([0, 1])$,
   - 4.1 If $\alpha \geq \xi$, accept: $x^{k+1} = y$,
   - 4.2 If $\alpha < \xi$, stay put: $x^{k+1} = x^k$.
5. If $k < N$, increase $k \to k + 1$ and continue from 2., else stop.

Two steps

Build the program in two steps:

1. Random walk sampling, no rejections
2. Add the rejection step

Sampling, no rejections

```
nsample = 10000;      % Sample size
x       = [0;0];      % Initial point
step    = 0.1;        % Step size of the random walk

Sample      = NaN(2,nsample);   % For memory allocation
Sample(:,1) = x;

for j = 2:N
    y = x + step*randn(2,1);

    % Accept unconditionally
    x = y;
    Sample(:,j) = x;
end
```

Add the α-coin

Write the condition

$$\frac{\pi(y)}{\pi(x)} > t$$

in logarithmic form:

$$\log \pi(y) - \log \pi(x) > \log t,$$

```
x        = [0;0];      % Initial point
step     = 0.1;        % Step size of the random walk
Sample       = NaN(2,nsample);   % For memory allocation
Sample(:,1) = x;

logpx = ...

for j = 2:N
    y = x + step*randn(2,1);
    logpy = ...
    t = rand;
    if logpy - logpx > log(t)
       % accept
       x = y;
       logpx = logpy;
    end
    Sample(:,j) = x;
end
```

- If a move is not accepted, the previous point has to be repeated:

$$\ldots, x^{k-1}, \underbrace{x^k, x^k, x^k}_{\text{rejections}}, x^{k+1}, x^{k+2}, \ldots$$

- The acceptance rate tells the relative rate of acceptances.
- Too low acceptance rate: The chain does not move
- Too high acceptance rate: The chain is essentially a random walk and learns nothing of the underlying distribution (cf. raising kids).

What is a good acceptance rate?
Rule of thumb: 15%-35%.

Example: Inverse problem in chemical kinetics

Reversible single reaction pair,

$$\mathrm{A} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{B},$$

Data: With known initial values, measure $[\mathrm{A}](t_j)$, $1 \leq j \leq n$ for

$$t_{\min} = t_1 < t_2 < \cdots < t_n = t_{\max}.$$

The noisy observation model is

$$b_j = [\mathrm{A}](t_j) + e_j, \quad e_j = \text{additive noise, noise level } = \sigma.$$

Inverse Problem: Estimate k_1 and k_{-1}.

Forward model: Mass Balance Equations

Denote

$$c_1(t) = [A](t), \quad c_2(t) = [B](t).$$

Assuming unit volume,

$$\begin{aligned} \frac{dc_1}{dt} &= -k_1 c_1 + k_{-1} c_2, \quad c_1(0) = c_{01} \\ \frac{dc_2}{dt} &= k_1 c_1 - k_{-1} c_2, \quad c_2(0) = c_{02} \end{aligned}$$

or

$$\frac{dc}{dt} = \mathrm{K} c, \quad c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix},$$

where

$$\mathrm{K} = \begin{bmatrix} -k_1 & k_{-1} \\ k_1 & -k_{-1} \end{bmatrix}.$$

Eigenvalues of K are

$$\lambda_1 = 0, \quad \lambda_2 = -k_1 - k_{-1}.$$

Time constant

$$\tau = \frac{1}{k_1 + k_{-1}}.$$

Eigenvectors:

$$v_1 = \begin{bmatrix} 1 \\ \delta \end{bmatrix}, \quad v_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \delta = \frac{k_1}{k_{-1}}.$$

Solution:

$$c = \alpha v_1 + \beta v_2 e^{-t/\tau}, \quad \alpha, \beta \in \mathbb{R}.$$

Initial conditions imply

$$\alpha = \frac{c_{01} + c_{02}}{1+\delta}, \quad \beta = \frac{\delta c_{01} - c_{02}}{1+\delta}.$$

In particular,

$$c_1(t) = f(t;k) = \frac{c_{01} + c_{02}}{1+\delta} - \frac{\delta c_{01} - c_{02}}{1+\delta} e^{-t/\tau},$$

where

$$k = \begin{bmatrix} k_1 \\ k_{-1} \end{bmatrix}.$$

Observation model: Data consists of n measurements of $c_1(t)$ corrupted by additive noise,

$$b_j = f(t_j, k) + e_j, \quad 1 \leq j \leq n.$$

Observation errors e_j mutually independent, zero mean normally distributed,

$$e_j \sim \mathcal{N}(0, \sigma^2).$$

Likelihood density

Assume that the noise is Gaussian white noise:

$$\pi_{\text{noise}}(e) \propto \exp\left(-\frac{1}{2\sigma^2}\|e\|^2\right).$$

Likelihood density is

$$\pi(b \mid k) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(b_j - f(t_j, k))^2\right).$$

Posterior density

Flat prior over an interval: We believe that

$$0 < k_1 \leq K_1, \quad 0 < k_{-1} \leq K_{-1},$$

with some reasonable upper bounds. Write

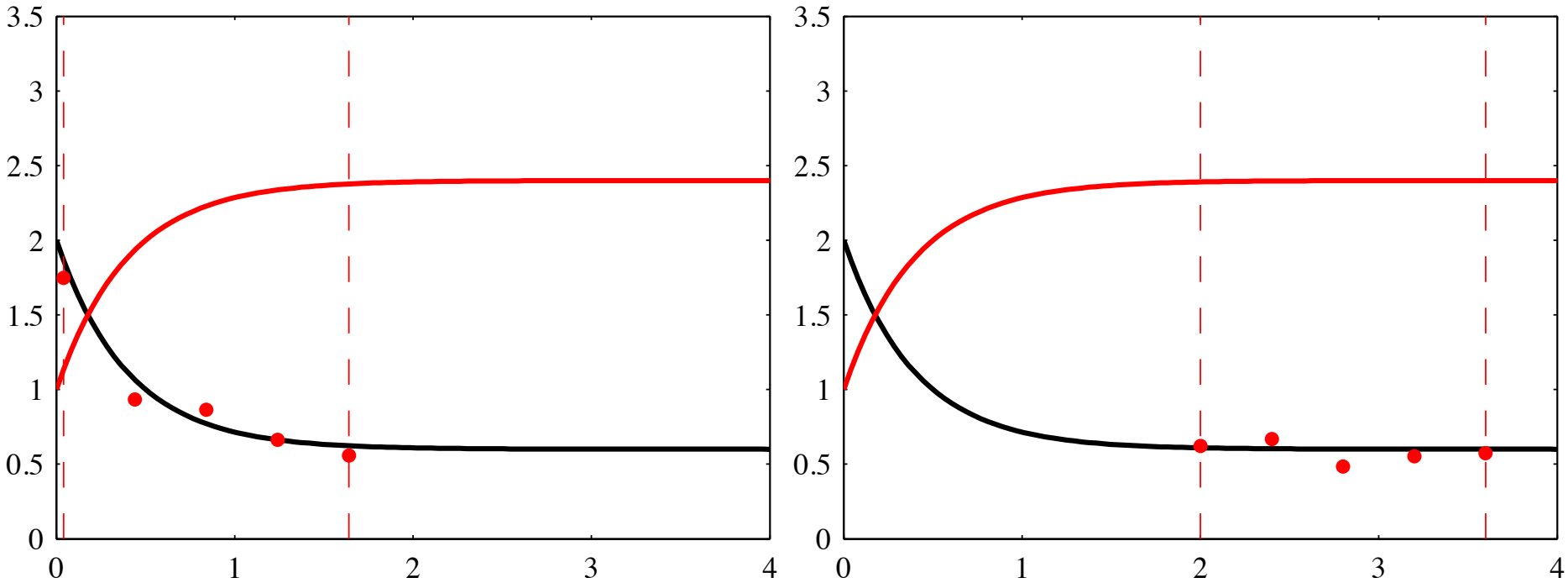
$$\pi_{\text{prior}}(k) \propto \chi_{[0,K_1]}(k_1)\chi_{[0,K_{-1}]}(k_{-1}).$$

Posterior density by Bayes' formula,

$$\pi(k \mid b) \propto \pi_{\text{prior}}(k)\pi(b \mid k).$$

Contour plots of the posterior density?

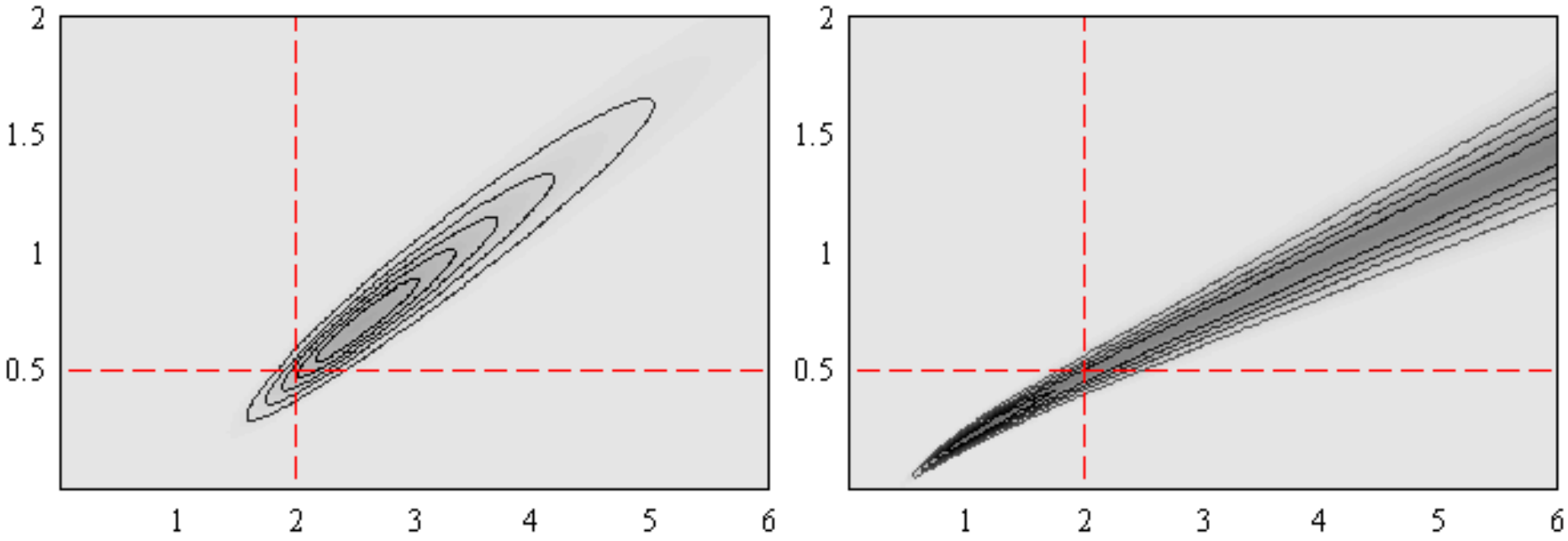
Data



$K_1 = 6$, $K_{-1} = 2$, five uniformly distributed measurements of $c_1(t)$ over the interval $[t_{\min}, t_{\max}]$,

$t_{\min} = 0.1\,\tau, t_{\max} = 4.1\,\tau$ (left) , $\quad t_{\min} = 5\,\tau, \quad t_{\max} = 9\,\tau$ (right)

Posterior densities



$K_1 = 6$, $K_{-1} = 2$, five uniformly distributed measurements of $c_1(t)$ over the interval $[t_{\min}, t_{\max}]$,

$t_{\min} = 0.1\,\tau$, $t_{\max} = 4.1\,\tau$ (left) , $\quad t_{\min} = 5\,\tau$, $\quad t_{\max} = 9\,\tau$ (right)

The hair cross indicates the value used for data generation.

MCMC exploration

Generate the data: Define

$$t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

where

$$b_j = \frac{c_{01} + c_{02}}{1 + \delta_{\text{true}}} - \frac{\delta_{\text{true}} c_{01} - c_{02}}{1 + \delta_{\text{true}}} e^{-t/\tau_{\text{true}}} + e_j.$$

$$\delta_{\text{true}} = \frac{k_{1,\text{true}}}{k_{-1,\text{true}}}, \quad \tau_{\text{true}} = \frac{1}{k_{1,\text{true}} + k_{-1,\text{true}}}.$$

Random walk Metropolis-Hastings

Start with the transient measurements.
White noise proposal,

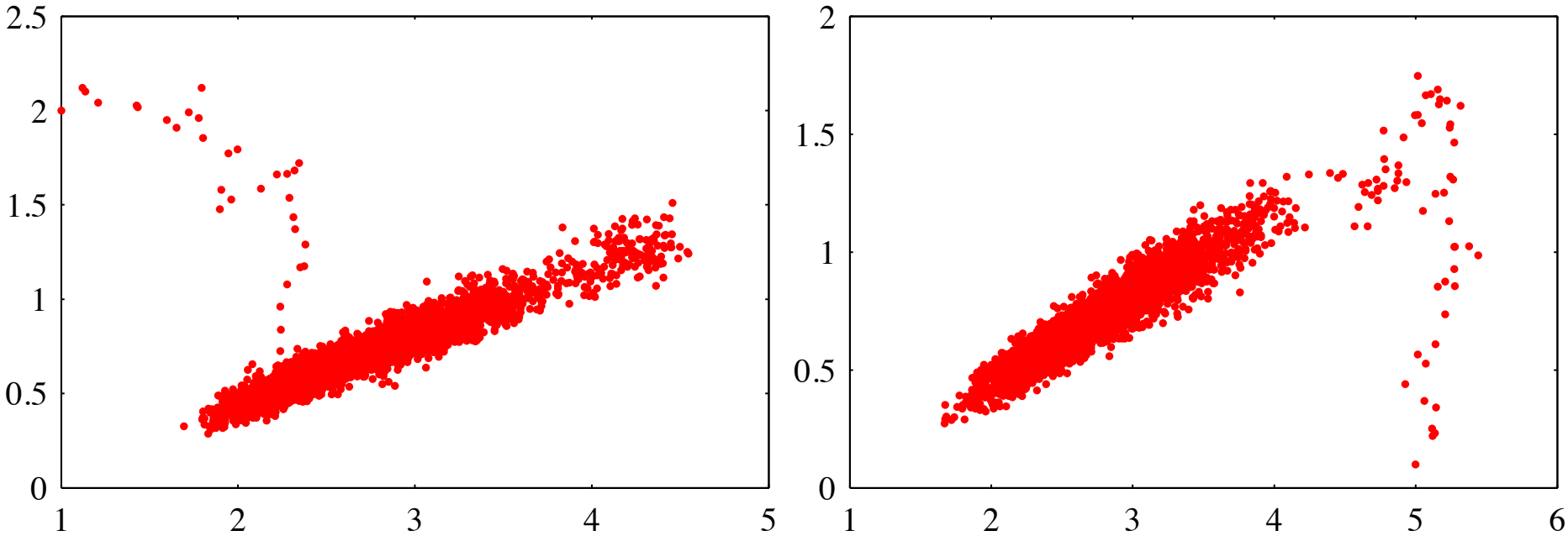
$$k_{\text{prop}} = k + \delta w, \quad w \sim \mathcal{N}(0, I).$$

Choose first $\delta = 0.1$, different initial points

$$k_0 = (1, 2) \text{ or } k_0 = (5, 0.1).$$

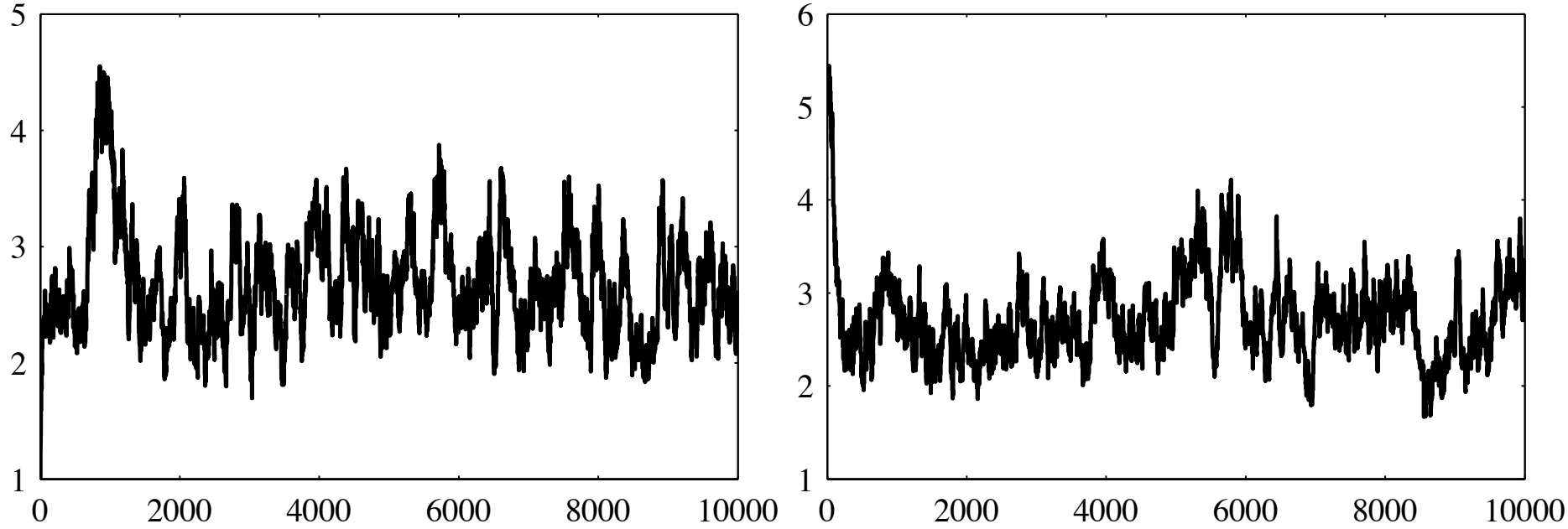
Acceptance rates with these values are of the order 45%.

Scatter plots



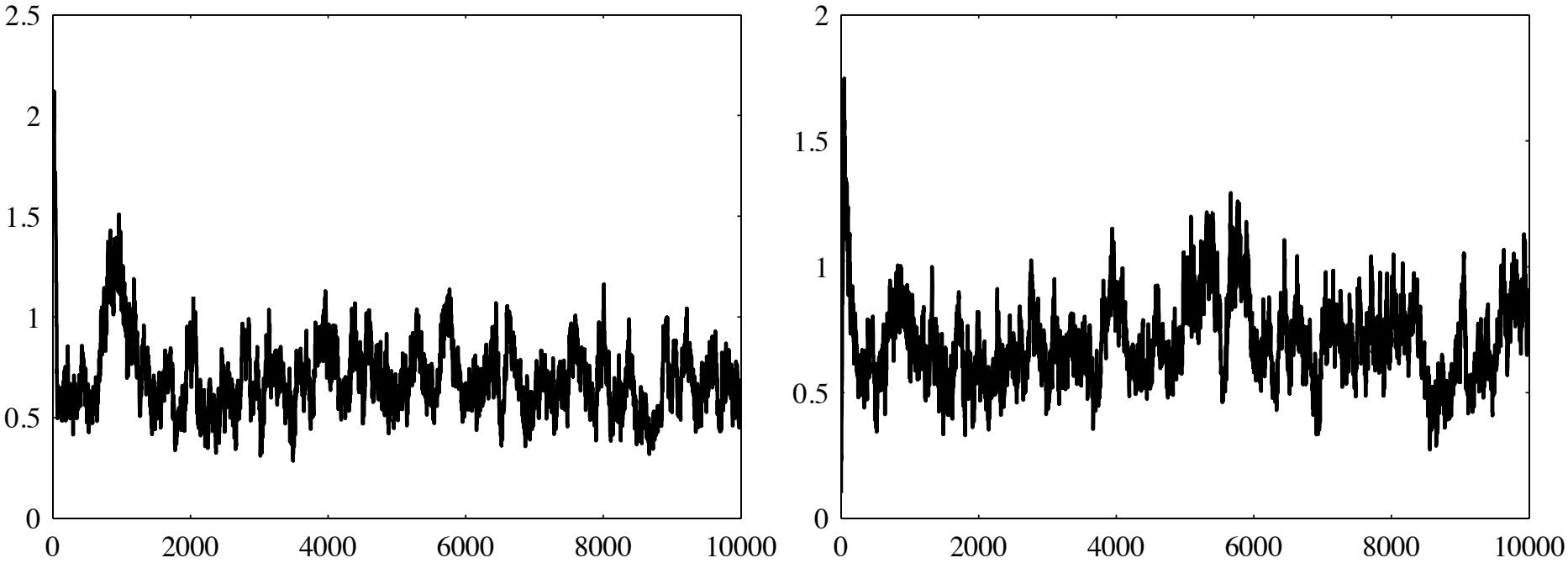
Transient measurements, $t_{\text{min}} = 0.1\,\tau$, $t_{\text{max}} = 4.1\,\tau$

Sample histories:First component



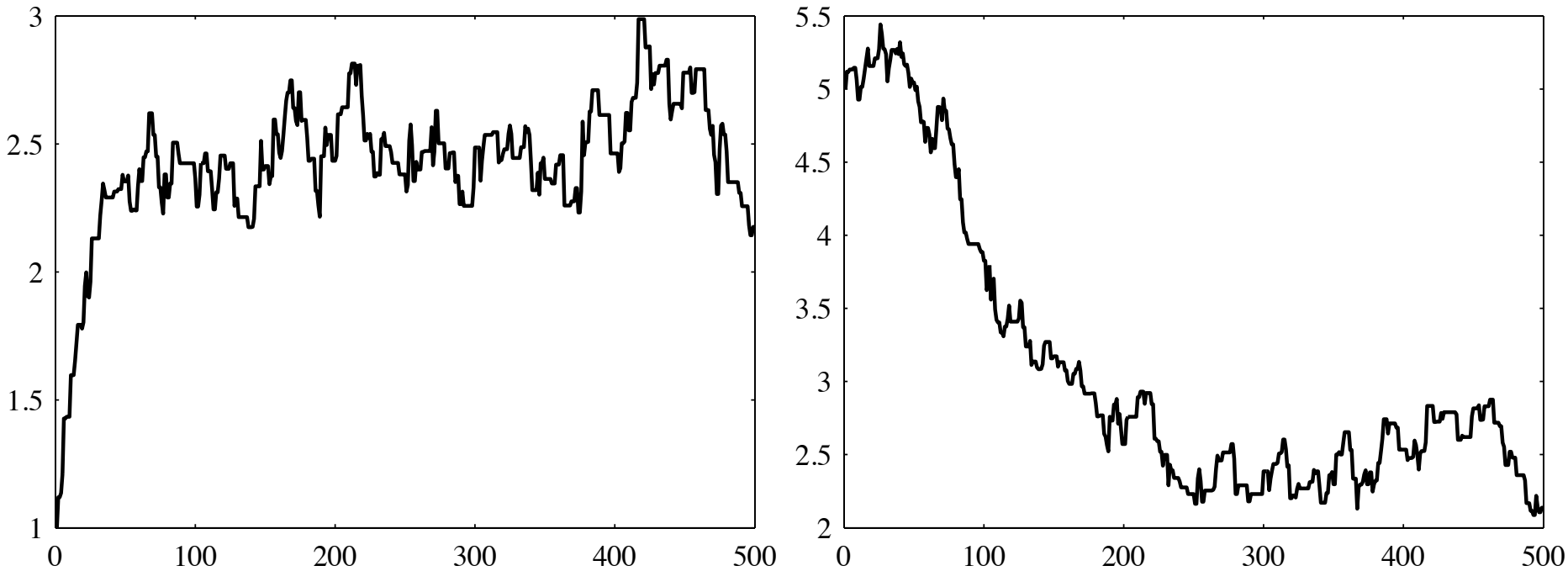
Initial value $k_1 = 1$ (left) and $k_1 = 5$ (right).

Sample histories: Second component



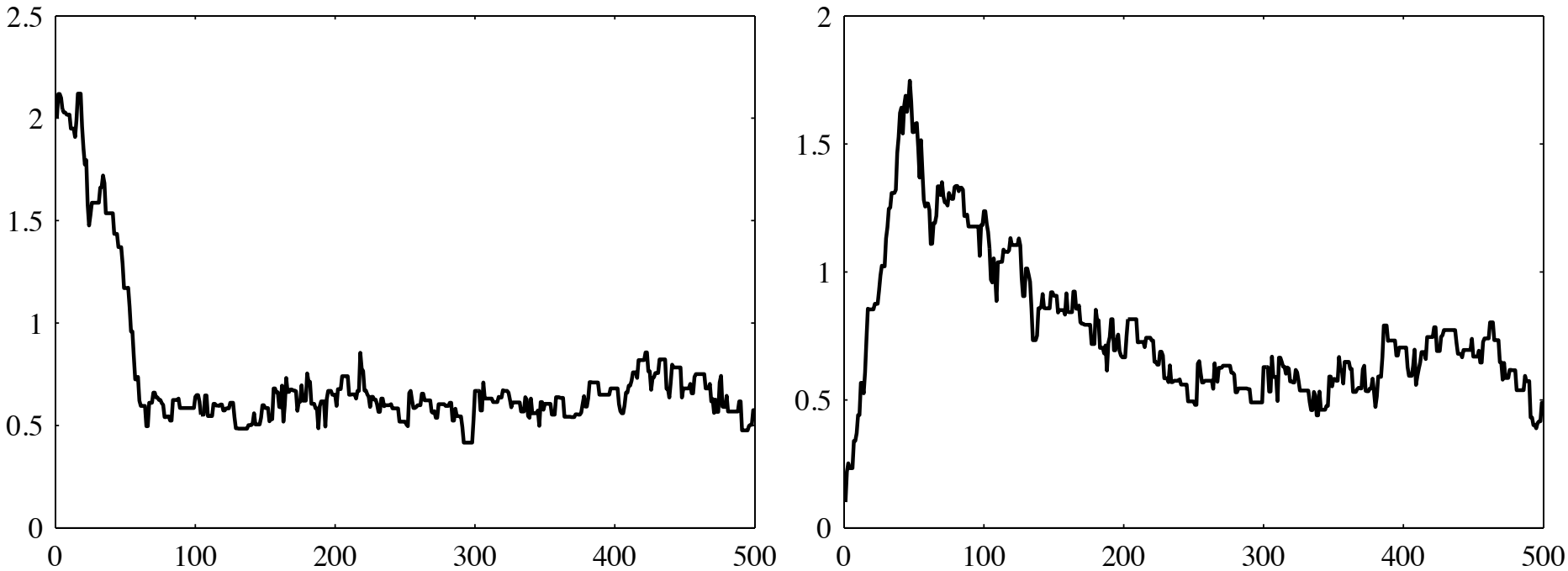
Initial value $k_2 = 2$ (left) and $k_2 = 0.2$ (right).

Burn-in: first component



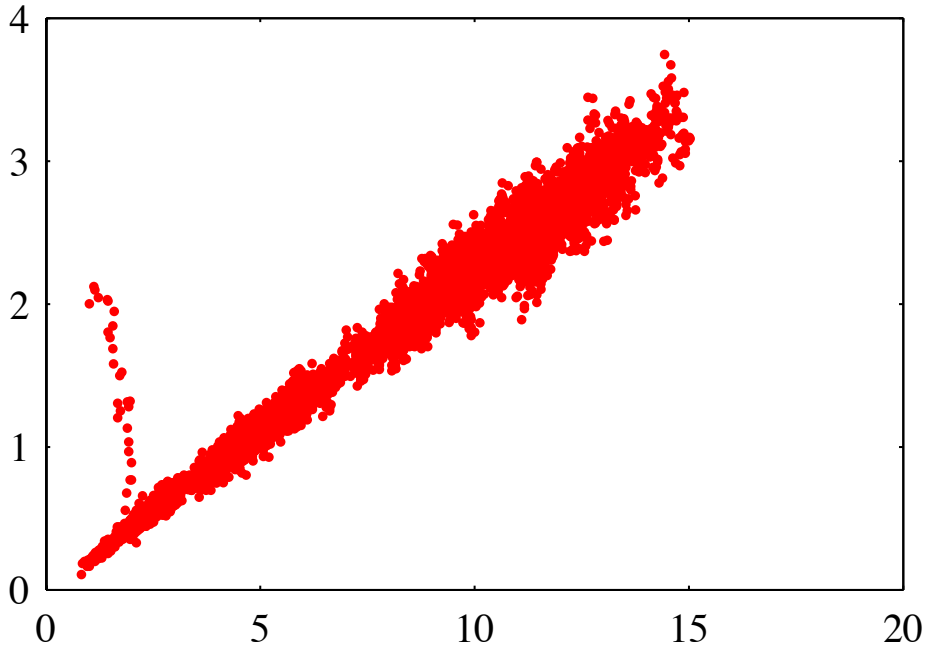
Initial value $k_1 = 1$ (left) and $k_1 = 5$ (right).

Burn-in: Second component



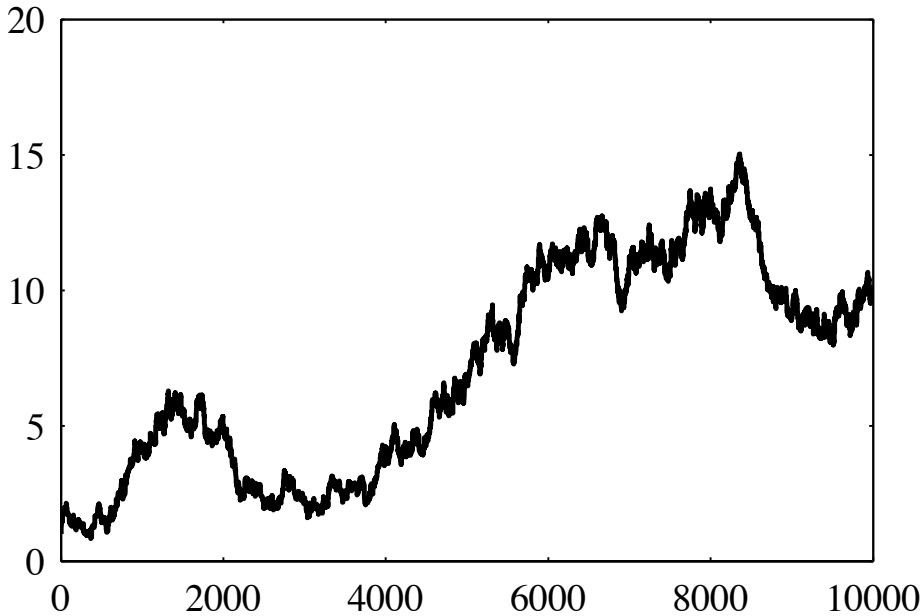
Initial value $k_2 = 2$ (left) and $k_2 = 0.2$ (right).

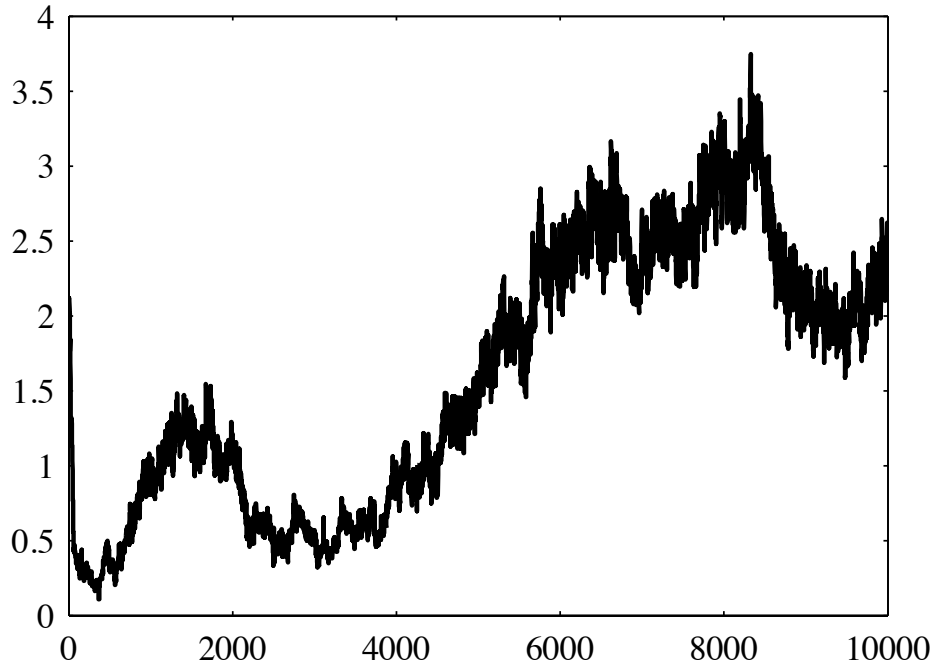
Scatter plots: Steady state measurement



$t_{\min} = 5\,\tau$, $t_{\max} = 9\,\tau$. Use the same step size as before.
Initial point $(k_1, k_2) = (1, 2)$.

Sample histories, a.k.a. fuzzy worms



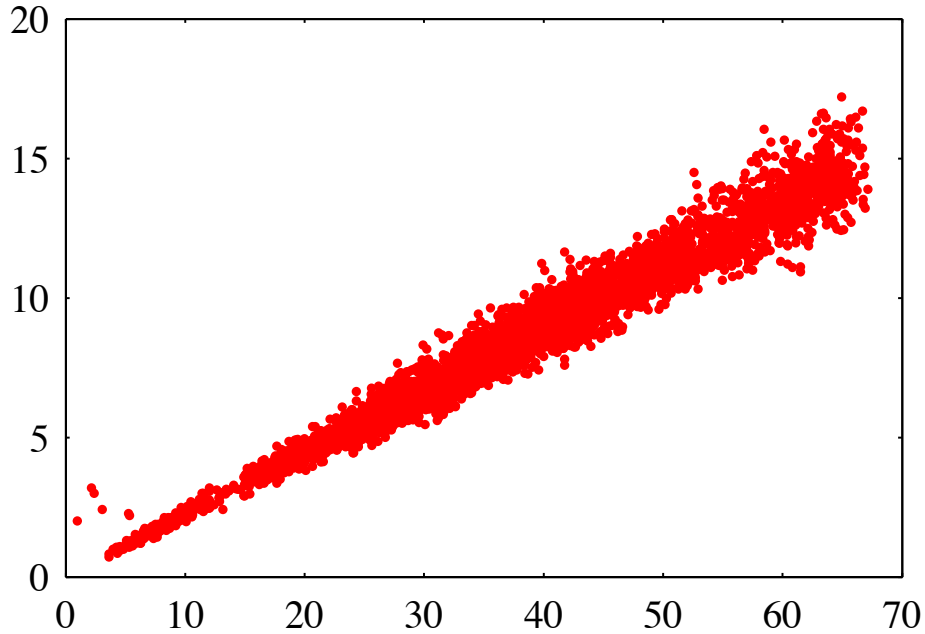


Increase the step size $0.1 \to 1$.

Initial point $(k_1, k_2) = (1, 2)$.

Acceptance remains high, about 55%

Scatter plots, steady state data

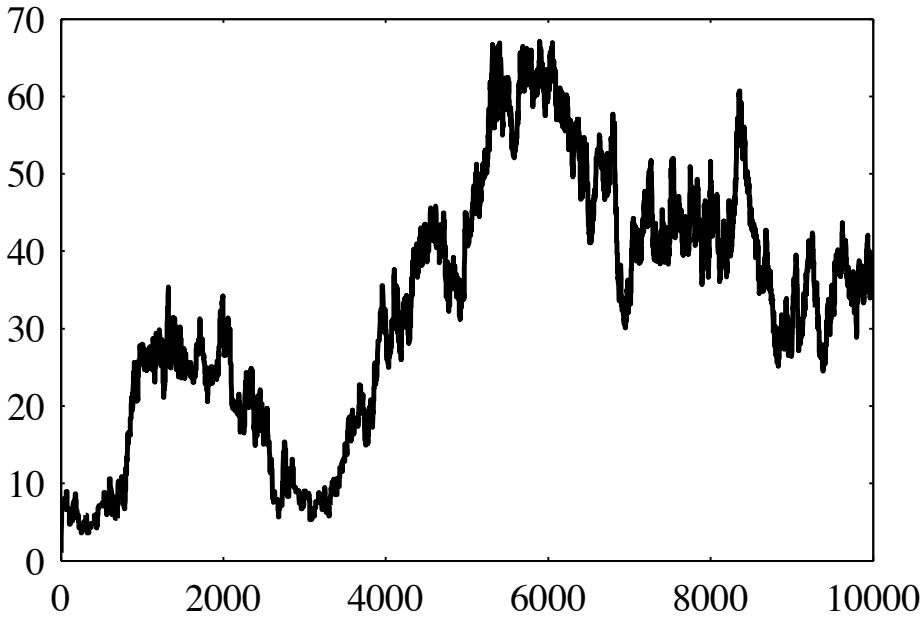


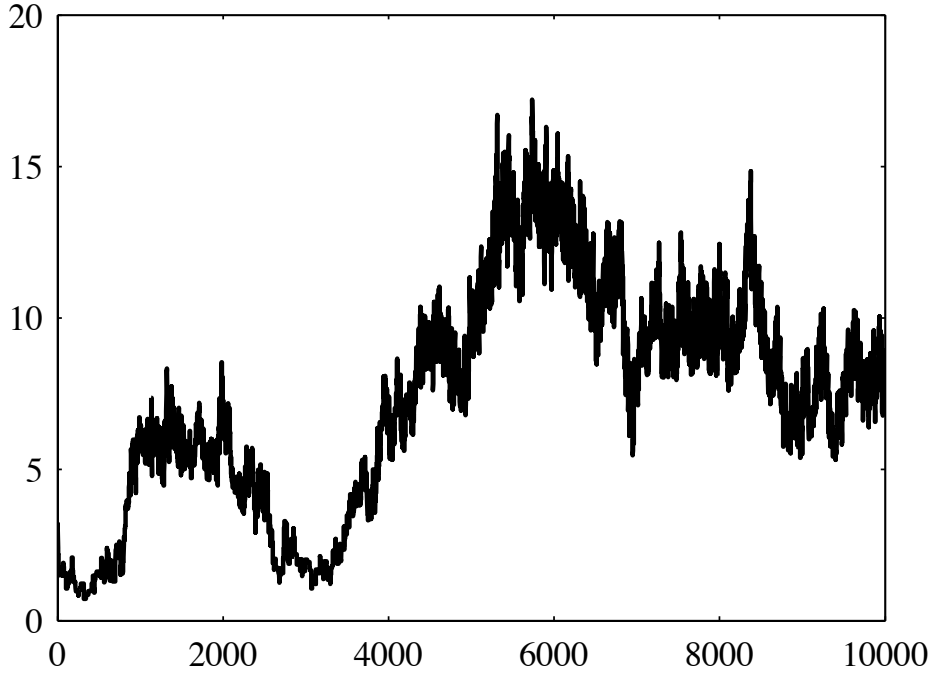
Increase the step size $0.1 \rightarrow 1$.

Initial point $(k_1, k_2) = (1, 2)$.

Acceptance remains high, about 55%

Fuzzy worms


70
60
50
40
30
20
10
0
0
2000
4000
6000
8000
10000


20
15
10
5
0
0
2000
4000
6000
8000
10000