Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMSN50)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 12
MCMC for Bayesian computation II
Feb 27, 2020

Plan of today's lecture

1. Last time: Stochastic modeling and Bayesian inference
2. Example: Change point detection
3. Interlude: Mixing of MCMC samplers
4. Example: Korsbetningen

We are here $\longrightarrow \bullet$

The frequentist approach

The frequentist approach to statistics is characterized as follows.

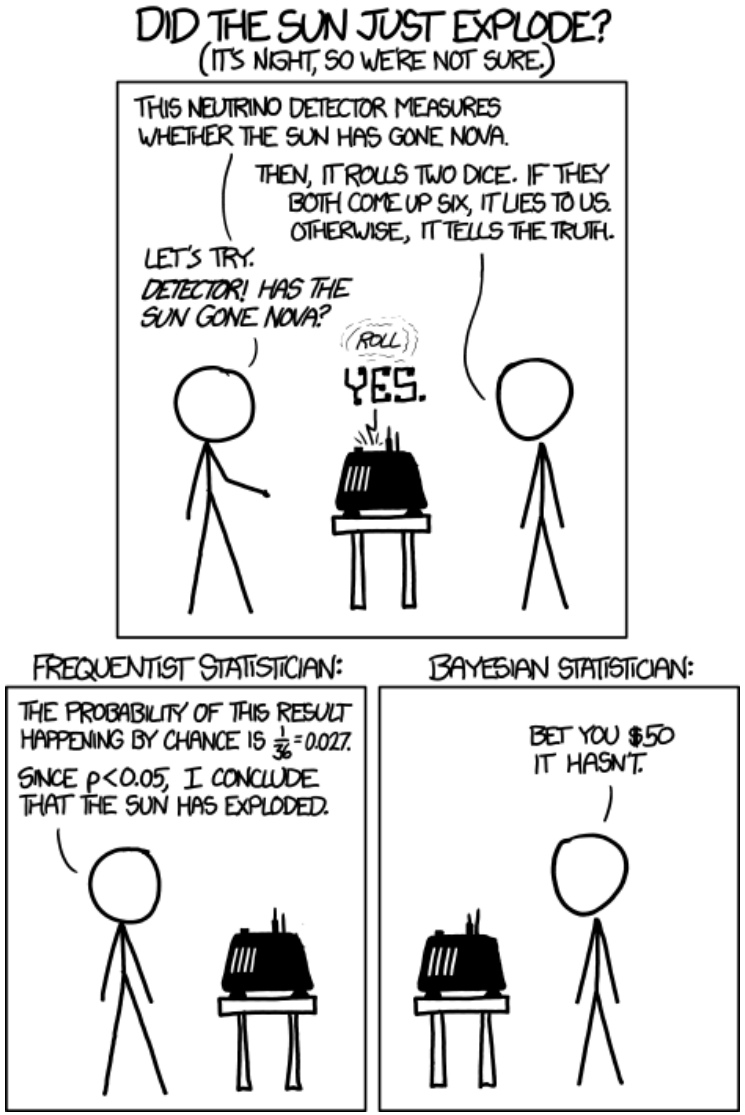
- Data y is viewed as an observation of a random variable Y with distribution $\mathbb{P}_0$ which most often is assumed to be a member of an exponential family
$$\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}.$$
- Estimates $\widehat{\theta}(y)$ are realizations of random variables.
- Confidence intervals are calculated to cover the true value in, say, 95% of the cases.
- Hypothesis testing is done by rejecting a hypothesis $\mathcal{H}_0$ if $\mathbb{P}(\text{data}\|\mathcal{H}_0)$ is small.

The Bayesian approach

Briefly, the Bayesian approach to statistics is as follows.

- The parameter θ is viewed as a random variable, and inference is based completely on the posterior distribution $f(\theta|y)$.
- It is possible to incorporate prior information in terms of the prior distribution $f(\theta)$.
- A 95% credible or posterior probability interval contains θ with a probability of 95% given the observations.
- Hypothesis tests are done by studying $\mathbb{P}(\mathcal{H}_0||\text{data})$.

The frequentist vs the Bayesian approach ;-)

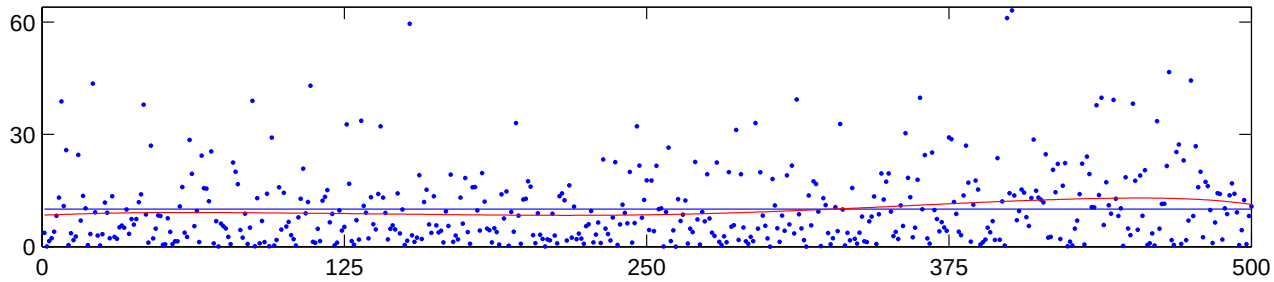


We are here $\longrightarrow \bullet$

Example: Change point detection

Consider the following model.

- We have measured the waiting times in a system and suspect that the expected waiting time changed during the monitoring period.
- The observations y_i for $i = 1, \ldots, n$ are assumed to follow exponential distributions with parameter θ_1 for $i \in \{1, \ldots, n_\mathsf{b}\}$ and parameter θ_2 for $i \in \{n_\mathsf{b} + 1, \ldots, n\}$.
- Further, we put a Gamma prior on θ_k, $\theta_k \sim \Gamma(a, b)$, with $a = 40$ and $b = 4$, and a uniform prior on n_b



Example: Change point detection (cont.)

Thus, we have unknown parameters $(\theta_1, \theta_2, n_\mathsf{b})$ and data $Y = (y_1, \ldots, y_n)$. The posterior becomes

$$
\begin{aligned}
&f(n_\mathsf{b}, \theta_1, \theta_2 | y_1, \ldots, y_n) \\
&\propto f(n_\mathsf{b}, \theta_1, \theta_2, y_1, \ldots, y_n) \\
&= f(\theta_1) f(\theta_2) f(n_\mathsf{b}) \prod_{i=1}^{n} f(y_i | n_\mathsf{b}, \theta_1, \theta_2) \\
&= \theta_1^{n_\mathsf{b}+a-1} \exp\left(-\theta_1 \left(b + \sum_{i=1}^{n_\mathsf{b}} y_i\right)\right) \\
&\qquad \times \theta_2^{n-n_\mathsf{b}+a-1} \exp\left(-\theta_2 \left(b + \sum_{i=n_\mathsf{b}+1}^{n} y_i\right)\right),
\end{aligned}
$$

Example: Change point detection (cont.)

The conditional distributions of θ_1 and θ_2 are easily calculated according to

$$\theta_1 | n_\mathsf{b}, \theta_2, y_1, \ldots, y_n \sim \Gamma\left(n_\mathsf{b} + a, b + \sum_{i=1}^{n_\mathsf{b}} y_i\right),$$

$$\theta_2 | n_\mathsf{b}, \theta_1, y_1, \ldots, y_n \sim \Gamma\left(n - n_\mathsf{b} + a, b + \sum_{i=n_\mathsf{b}+1}^{n} y_i\right).$$

The conditional distribution of n_b is however more complicated:

$$f(n_\mathsf{b} | \theta_1, \theta_2, y_1, \ldots, y_n)$$

$$\propto \theta_1^{n_\mathsf{b}} \exp\left(-\theta_1 \sum_{i=1}^{n_\mathsf{b}} y_i\right) \theta_2^{-n_\mathsf{b}} \exp\left(-\theta_2 \sum_{i=n_\mathsf{b}+1}^{n} y_i\right)$$

Example: Change point detection (cont.)

Thus, we sample the posterior of $(\theta_1, \theta_2, n_\mathsf{b})$ using a Gibbs sampler with a MH step for n_b, yielding a hybrid sampler. The MH step goes as follows.

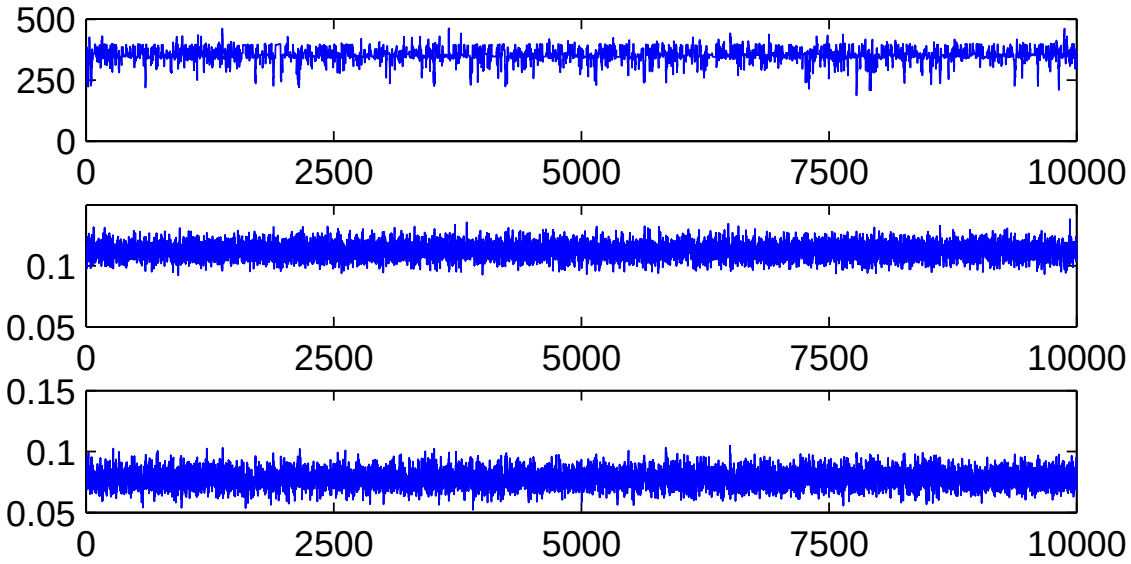
Given n_b, we propose a candidate n_b^* uniformly on the integers $\{n_\mathsf{b} - R, \ldots, n_\mathsf{b}, \ldots, n_\mathsf{b} + R\}$, for some R. This forms a symmetric proposal on $\{1, \ldots, n\}$.

The acceptance probability for the MH step thus becomes

$$\alpha(n_\mathsf{b}, n_\mathsf{b}^*) = 1 \wedge \frac{\theta_1^{n_\mathsf{b}^*} \theta_2^{-n_\mathsf{b}^*} \exp(-\theta_1 \sum_{i=1}^{n_\mathsf{b}^*} y_i) \exp(-\theta_2 \sum_{i=n_\mathsf{b}^*+1}^{n} y_i)}{\theta_1^{n_\mathsf{b}} \theta_2^{-n_\mathsf{b}} \exp(-\theta_1 \sum_{i=1}^{n_\mathsf{b}} y_i) \exp(-\theta_2 \sum_{i=n_\mathsf{b}+1}^{n} y_i)}.$$

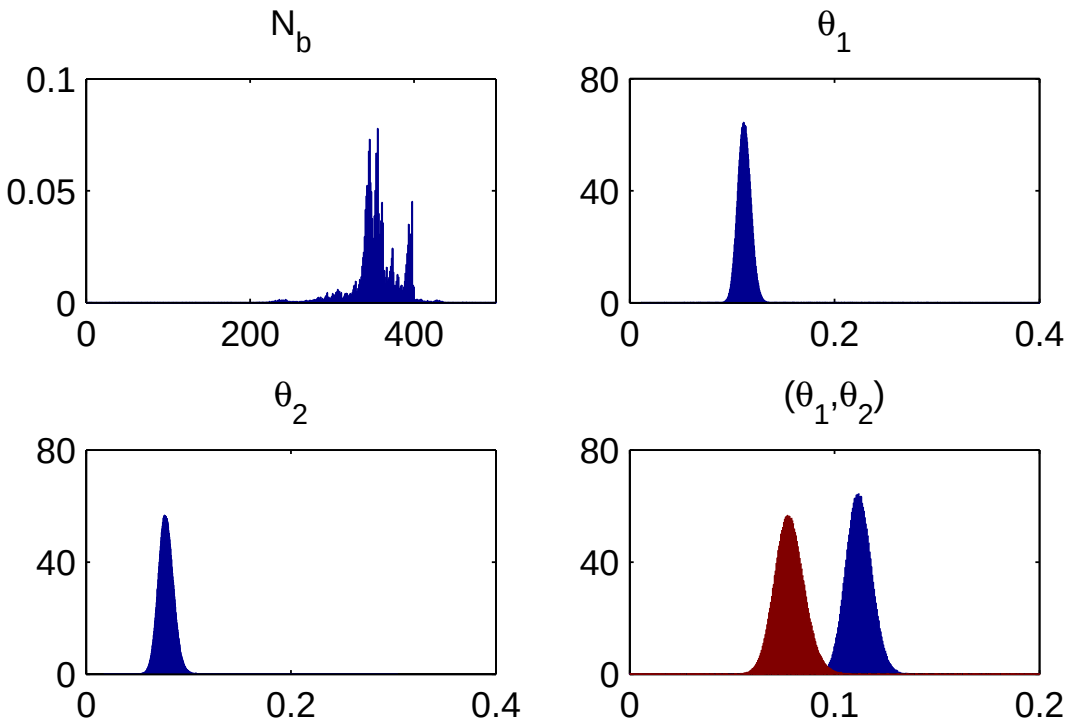
Example: Change point detection (cont.)

Running this Gibbs sampler with $R = 75$ gives an acceptance rate of 33%.



Example: Change point detection (cont.)

The resulting histograms of the parameters are as follows:



Selecting priors

The posterior is computed via Bayes's formula

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta')f(\theta')\,\mathrm{d}\theta'} \propto f(y|\theta)f(\theta).$$

In Bayesian modeling there is always an interplay between the prior and the data:

- The posterior is drawn away from the data towards the prior. How far depends on the strength of the prior.
- However, enough data will most likely overwhelm the prior.

Two common prior-types are

- conjugate priors.
- improper (flat) priors.

Conjugate priors

Conjugate priors

- are such that the prior and the posterior belong to the same distribution class for a given likelihood.
- allow theoretical calculations and Gibbs sampling.
- are sometimes criticized since we select priors for ease of calculation. However, often the parameters are flexible enough to allow for a reasonable prior.
- may be hard to derive for complex models.

Conjugate priors

Conjugate priors for θ for some common likelihoods. All parameters except θ are assumed fixed and known and data $(y_1, \ldots, y_n)$ is assumed to be conditionally independent given θ.

Likelihood	*Prior*	*Posterior*
$\text{Bin}(n, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + y, \beta + n - y)$
$\text{Ge}(\theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + n, \beta + \sum_{i=1}^n y_i - n)$
$\text{NegBin}(n, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + n, \beta + y - n)$
$\text{Gamma}(k, \theta)$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + nk, \beta + \sum_{i=1}^n y_i)$
$\text{Po}(\theta)$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + \sum_{i=1}^n y_i, \beta + n)$
$\mathcal{N}(\mu, \theta^{-1})$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(m, s^2)$	$\mathcal{N}\left(\frac{m/s^2 + n\bar{y}/\sigma^2}{1/s^2 + n/\sigma^2}, \frac{1}{1/s^2 + n/\sigma^2}\right)$

Improper priors

Improper, or flat, priors are used when prior information is deficient.

For instance, if $\theta \in \mathbb{R}$, $f(\theta) \propto 1$ is an improper prior since it is not integrable; however, we allow this as long as the posterior is a well-defined density.

For instance, let y be an observation from $Y \sim \mathcal{N}(\theta, 1)$, where $\theta \in \mathbb{R}$. Since we do not have any prior information concerning θ we put $f(\theta) \propto 1$ for all $\theta \in \mathbb{R}$. After this we proceed, formally, like

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta')f(\theta')\,\mathrm{d}\theta'} = \frac{\mathcal{N}(y;\theta,1)\cdot 1}{\int \mathcal{N}(y;\theta',1)\cdot 1\,\mathrm{d}\theta'}$$

$$\overset{\text{symm.}}{=} \frac{\mathcal{N}(\theta;y,1)\cdot 1}{\int \mathcal{N}(\theta';y,1)\cdot 1\,\mathrm{d}\theta'} = \mathcal{N}(\theta;y,1).$$

We are here $\longrightarrow \bullet$

Mixing of MCMC samplers

We recall that the asymptotic variance of τ_N is given by

$$\sigma^2 = r(0) + 2\sum_{\ell=1}^{\infty} r(\ell) \quad \text{with} \quad r(\ell) = \lim_{n\to\infty} \mathbb{C}(\phi(X_{n+\ell}), \phi(X_n)).$$

Some remarks:

- Consequently, in order to obtain a low variance of τ_N, the covariance function $r(\ell)$ should decrease rapidly with ℓ.
- For geometrically ergodic chains $r(\ell)$ tends to zero geometrically fast.
- The speed at which $r(\ell)$ tends to zero is typically described using the term mixing.
  - Strong mixing = fast forgetting = rapidly decreasing $r(\ell)$.
  - Bad mixing = slow forgetting = slowly decreasing $r(\ell)$.

Why is good mixing important?

Bad choices of proposal distributions may lead to bad mixing, implying high variance and the need for a large MC sample size to ensure good estimates. This causes problems for the MCMC algorithm in the sense that it may

- fail to converge altogether or
- need a very long time to converge.

Optimal mixing

Let us focus for a while on the MH algorithm.

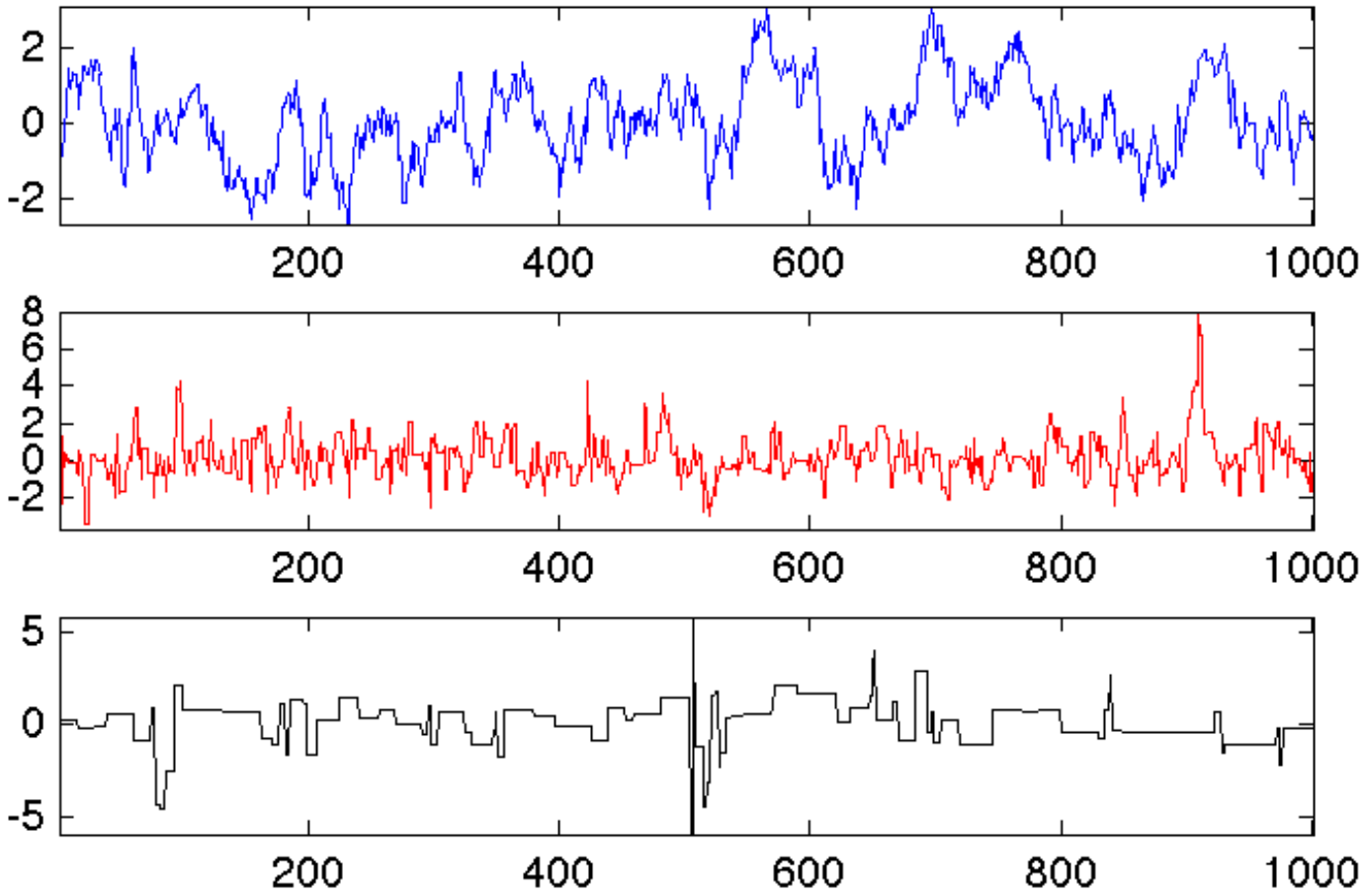
- When designing a random walk proposal, $X_k^* = X_k + \epsilon$ with $\epsilon \sim \mathcal{N}(0, s\mathbf{\Sigma})$, two things will effect the acceptance rate:
  1. How well $\mathbf{\Sigma}$ captures the dependence of the target distribution.
  2. How appropriate the scaling $s > 0$ is.
- One way to obtain a covariance matrix $\mathbf{\Sigma}$ that captures well the dependence structure of the target distribution $f(x)$ is to let

$$\mathbf{\Sigma}_{ij} = \frac{2.38}{d} \left(- \frac{\partial^2 \log f(x)}{\partial x_i \partial x_j} \bigg|_{x=x_{\text{mode}}} \right)^{-1}.$$

- Rule of thumb: A good acceptance rate is around 30% (23%–44%)!

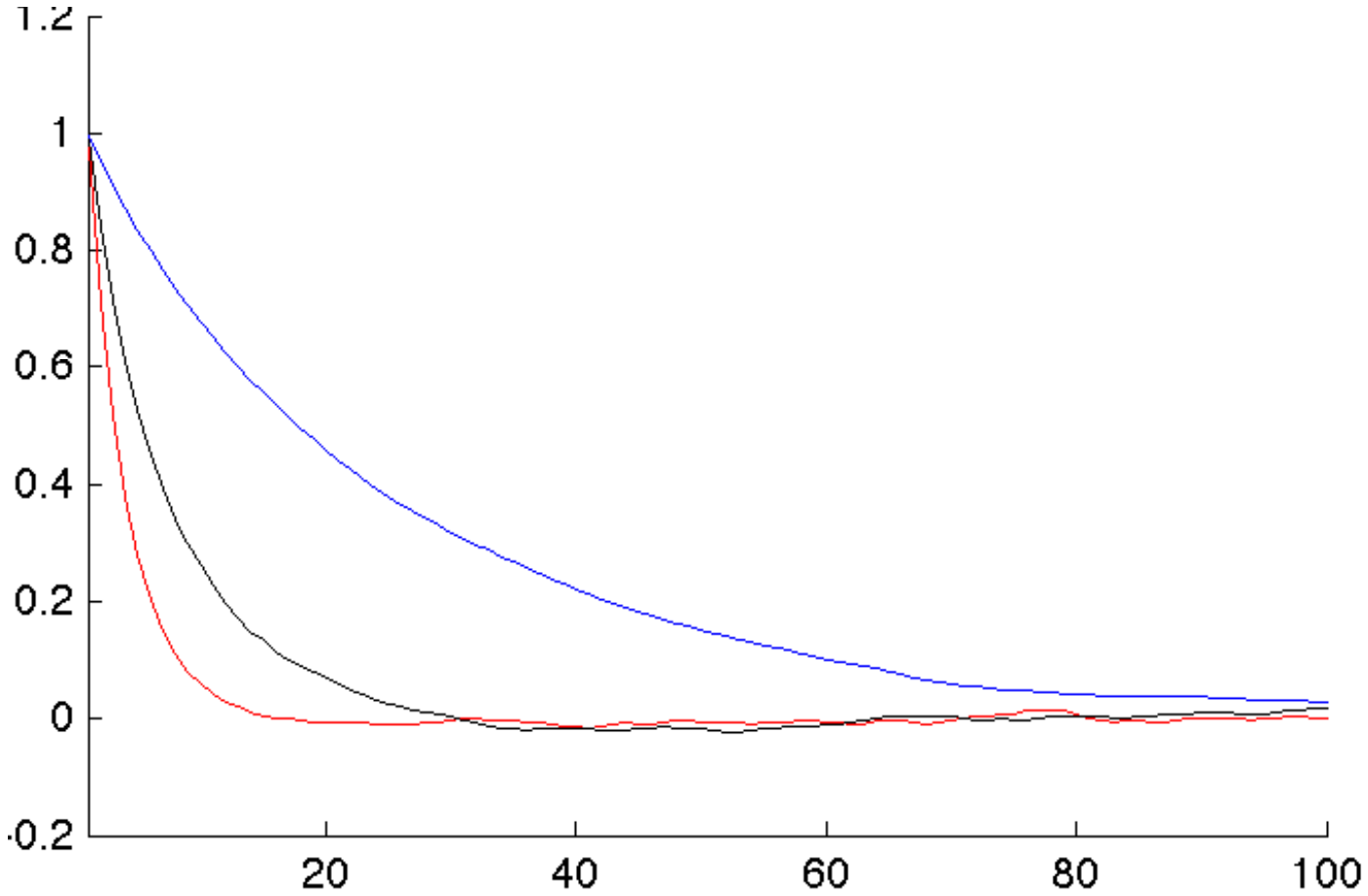
Mixing—Random walk proposal

Using symmetric normal proposal with three different values for s (small, medium, large, respectively) yields the following trajectories:



Mixing—Random walk proposal

Correlation function for the three chains:



We are here $\longrightarrow \bullet$

Example: Korsbetningen

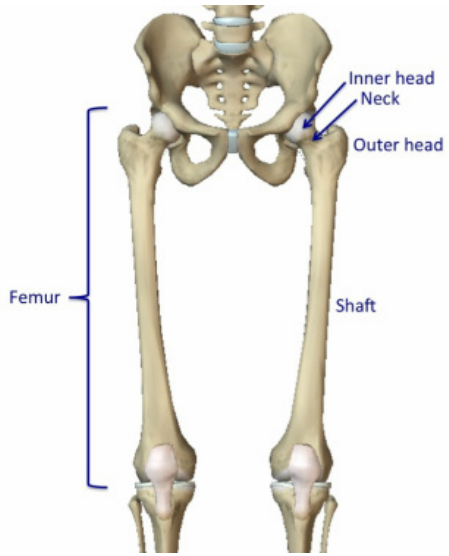
Latin inscription:
Anno Domini MCCCLXI feria III post Jacobi ante portas Visby in manibus Danorum ceciderunt Gutenses, hic sepulti, orate pro eis!

On the third day after saint Jacob, in the year of our lord 1361, the Goths fell outside the gates of Visby at the hands of the Danish. They are buried here. Pray for them!

Example: Korsbetningen—background

The background is the following.

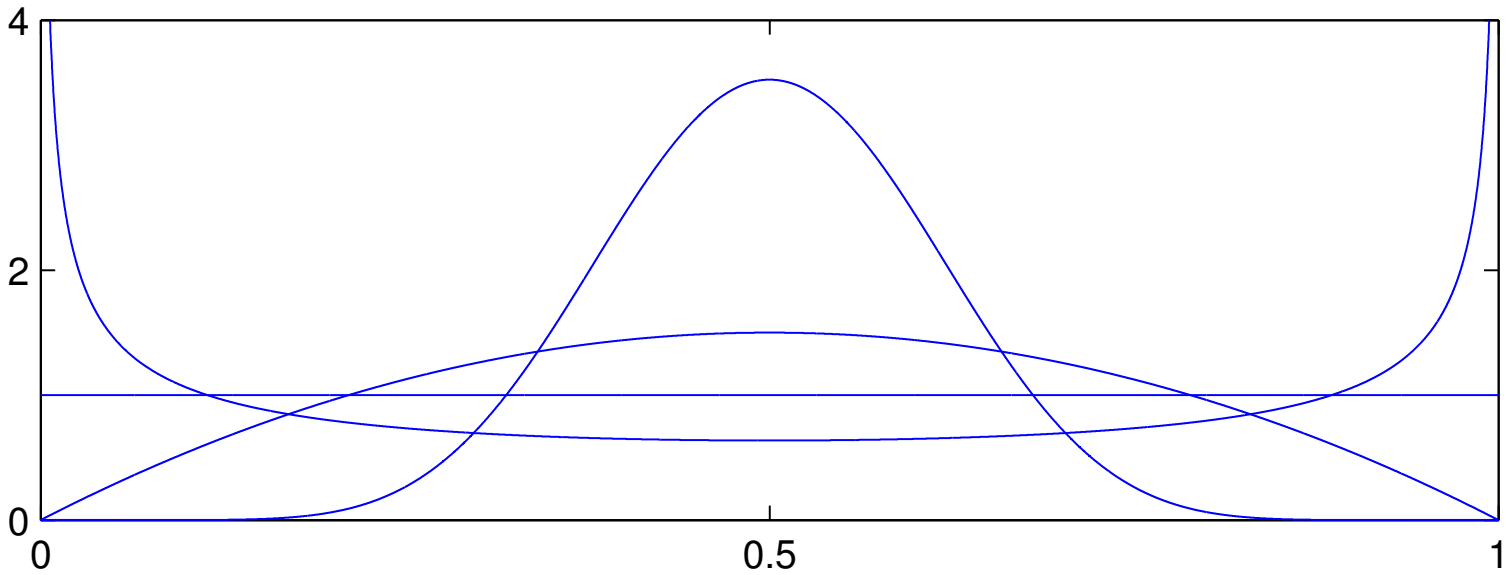
- In 1361 the Danish king Valdemar Atterdag conquered Gotland and captured the rich Hanseatic town of Visby.
- Most of the defenders were killed in the attack and are buried in a field, Korsbetningen, outside of the walls of Visby.
- In 1929–1930 the gravesite (with several graves) was excavated . In grave one e.g. a total of 493 femurs, 237 right and 256 left, were found.
- We want to estimate the number of buried bodies.
- An interesting tv-programme about the event can be found at https://www.youtube.com/watch?v=IGY2OqMXF9w



Example: Korsbetningen—model

We set up the following model.

- Assume that the numbers y_1 and y_2 of left resp. right legs are two observations from a Bin(n, p) distribution.
- Here n is the total number of people buried and p is the probability of finding a leg, left or right, of a person.
- We put a conjugate Beta(a, b)-prior on p and a $\mathcal{U}(256, 2500)$ prior on n.



Example: Korsbetningen—a hybrid MCMC

We proceed as follows:

- A standard Gibbs step for

$$p|n, y_1, y_2 \sim \text{Beta}(a + y_1 + y_2, b + 2n - (y_1 + y_2)).$$

- MH for n, with a symmetric proposal obtained by drawing, given n, a new candidate n^* among the integers $\{n - R, \ldots, n, \ldots, n + R\}$.
- The acceptance probability becomes

$$\alpha(n, n^*) = 1 \wedge \frac{(1-p)^{2n^*}(n^*!)^2(n - y_1)!(n - y_2)!}{(1-p)^{2n}(n!)^2(n^* - y_1)!(n^* - y_2)!}.$$

Example: Korsbetningen—a hybrid MCMC