Accept-Reject Algorithm

Let $\pi(x)$ be out *Target density*, i.e. the density we want to sample from.

Accept-Reject Algorithm

Choose initial value $x^{(0)}$.
For $t = 1, 2, \ldots, T$

1. Generate **Proposal**: $y \sim q(x^{(t-1)}, y)$.
2. Accept proposal with probability: $a(x^{(t-1)}, y)$
   otherwise reject it.
3. If **accepting**: $x^{(t)} = y$
4. If **rejecting**: $x^{(t)} = x^{(t-1)}$

This algorithm generate a realisation of a time homogeneous Markov chain.

How do we choose $q(x, y)$ and $a(x, y)$ so that the unique invariant distribution of the resulting Markov chain is given by $\pi(x)$?

The Metropolis-Hastings algorithm

How to choose $q(x, y)$ and $a(x, y)$?

One choice leads to the Metropolis-Hasting algorithm. The user specifies a proposal kernel $q(x, y)$. The algorithm then "automatically" chooses the correct acceptance probability.

Metropolis-Hastings algorithm

- Choose any proposal kernel $q(x, y)$
- Define the *Hastings ratio*

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)},$$

  where $H(x, y) = \infty$ if $\pi(x)a(x, y) = 0$.
- The acceptance probability is

$$a(x, y) = \min\{1, H(x, y)\}.$$

The Metropolis algorithm

A special case of the MH-algorithm is when the proposal kernel is symmetric:

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

In this case the Hastings-ratio simplifies to

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \frac{\pi(y)}{\pi(x)}.$$

Example: The most common example, is when the proposal is normal distributed with x as the mean value, and τ_p as the precision:

$$q(x, y) = \sqrt{\frac{\tau_p}{2\pi}} \exp\left(-\frac{1}{2}\tau_p(y - x)^2\right).$$

Clearly, $q(x, y) = q(y, x)$.

Burn-in

- Generate $X^{(0)} \sim \pi_0(x)$, an initial distribution, typically different from $\pi(x)$.
- Create irreducible Markov chain $X^{(0)}, X^{(1)}, X^{(2)}, \ldots$ with $\pi(x)$ as invariant distribution.
- For small values of t the distribution of $X^{(t)}$ can be quite different from $\pi(x)$.
- As a consequence, the sample mean

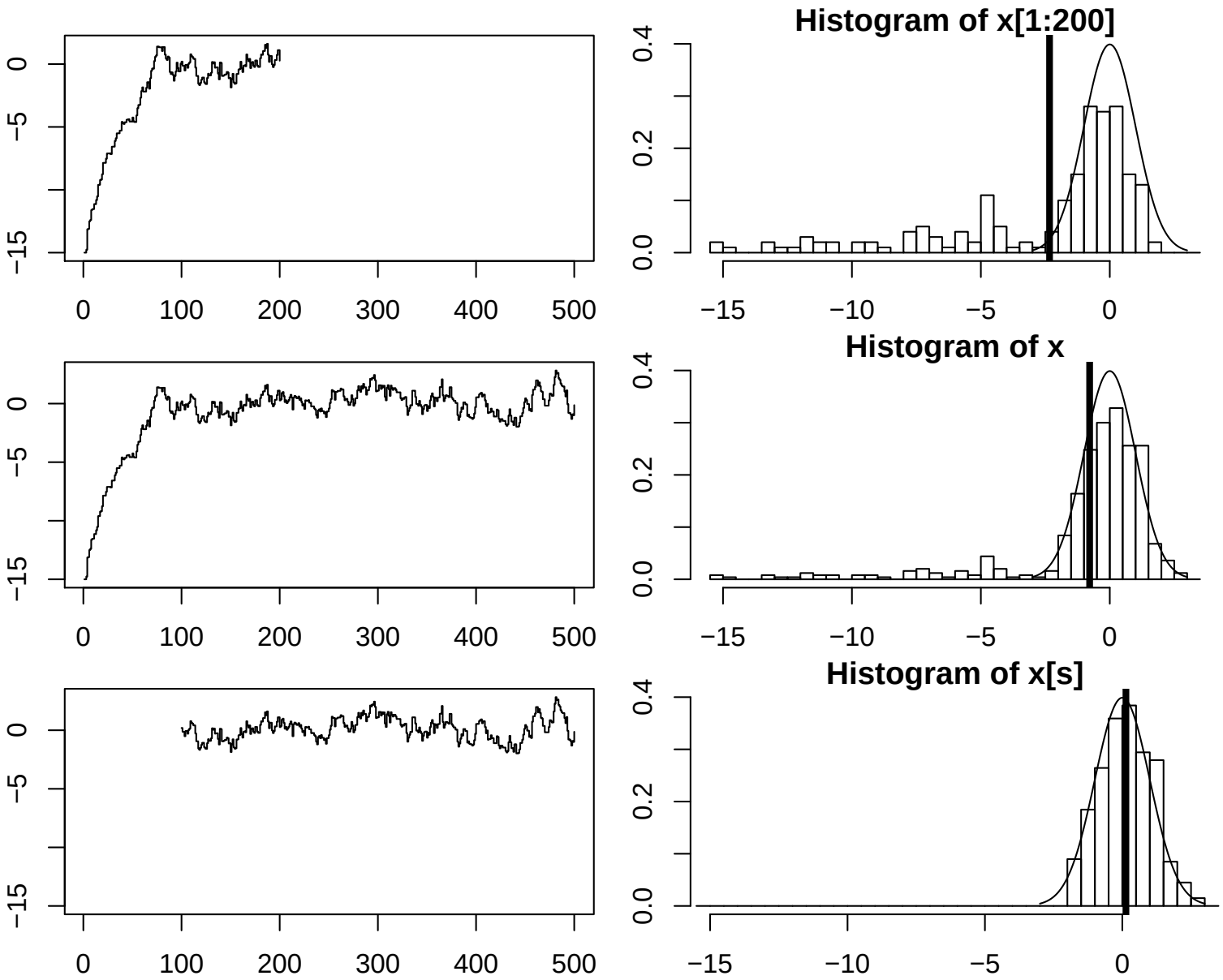
$$\frac{1}{T}\sum_{t=1}^{T} X^{(t)}$$

  is biased, i.e. $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right] \neq \mu$.
- Instead consider

$$\frac{1}{T}\sum_{t=1}^{T} X^{(m+t)},$$

  where m is the length of the **burn-in**

The effect of burn-in



Variance of the sample mean: IID Case

Assume we have independent samples $X^{(1)}, X^{(2)}, \ldots, X^{(T)}$ from $\pi(x)$.

Assume $E[X^{(t)}] = \mu$ and $Var[X^{(t)}] = \sigma^2$.

The **sample mean** is

$$\frac{1}{T}\sum_{t=1}^{T} X^{(t)}$$

For the sample mean we have the following results.

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right] = \mu$$

$$\mathbb{V}\mathrm{ar}\left[\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right] = \frac{1}{T}\sigma^2$$

$$T \cdot \mathbb{V}\mathrm{ar}\left[\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right] = \sigma^2.$$

Variance of the sample mean: Markov Chain Case

Assume $X^{(1)}, X^{(2)}, X^{(3)}, \ldots$ is an irreducible Markov chain with invariant distribution with density $\pi(x)$.

Further, assume that $X^{(1)} \sim \pi(x)$ which implies that $X^{(t)} \sim \pi(X)$ for all $t = 2, 3, 4, \ldots$, which in turn implies that $\mathbb{E}[X^{(t)}] = \mu$ and $\mathbb{V}\mathrm{ar}[X^{(t)}] = \sigma^2$.

The expected value of the sample mean is (again)

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right] = \mu.$$

So the expected value of the sample mean is unaffected by the shift from IID sample to Markov chain.

Variance of the sample mean: Markov Chain Case

- Regarding the variance we have

$$T \cdot \mathbb{V}\mathrm{ar}\left[\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right] \to \sigma^2\left(1 + 2\sum_{i=1}^{\infty}\rho_i\right),$$

  where

$$\rho_i = \mathrm{Corr}(X^{(t)}, X^{(t+i)}) = \frac{\mathbb{E}\left[(X^{(t)} - \mu)(X^{(t+i)} - \mu)\right]}{\sigma^2}$$

  is the lag-i auto-correlation.
- We call $\sigma^2\left(1 + 2\sum_{i=1}^{\infty}\rho_i\right)$ the *asymptotic variance*.
- Trying to get $\tau = 1 + 2\sum_{i=1}^{\infty}\rho_i$ to be as small as possible seems like a good idea.
- If we just want to estimate μ this is a brilliant idea.

Tuning

Assume the proposal kernel is

$$q(x, y) = \sqrt{\frac{\tau_p}{2\pi}} \exp\left(-\frac{1}{2}\tau_p(y - x)^2\right).$$

Now, τ_p is an "algorithm parameter" that we need to choose.

What is a good choice of τ_p? This is an example of *tuning* an algorithm.

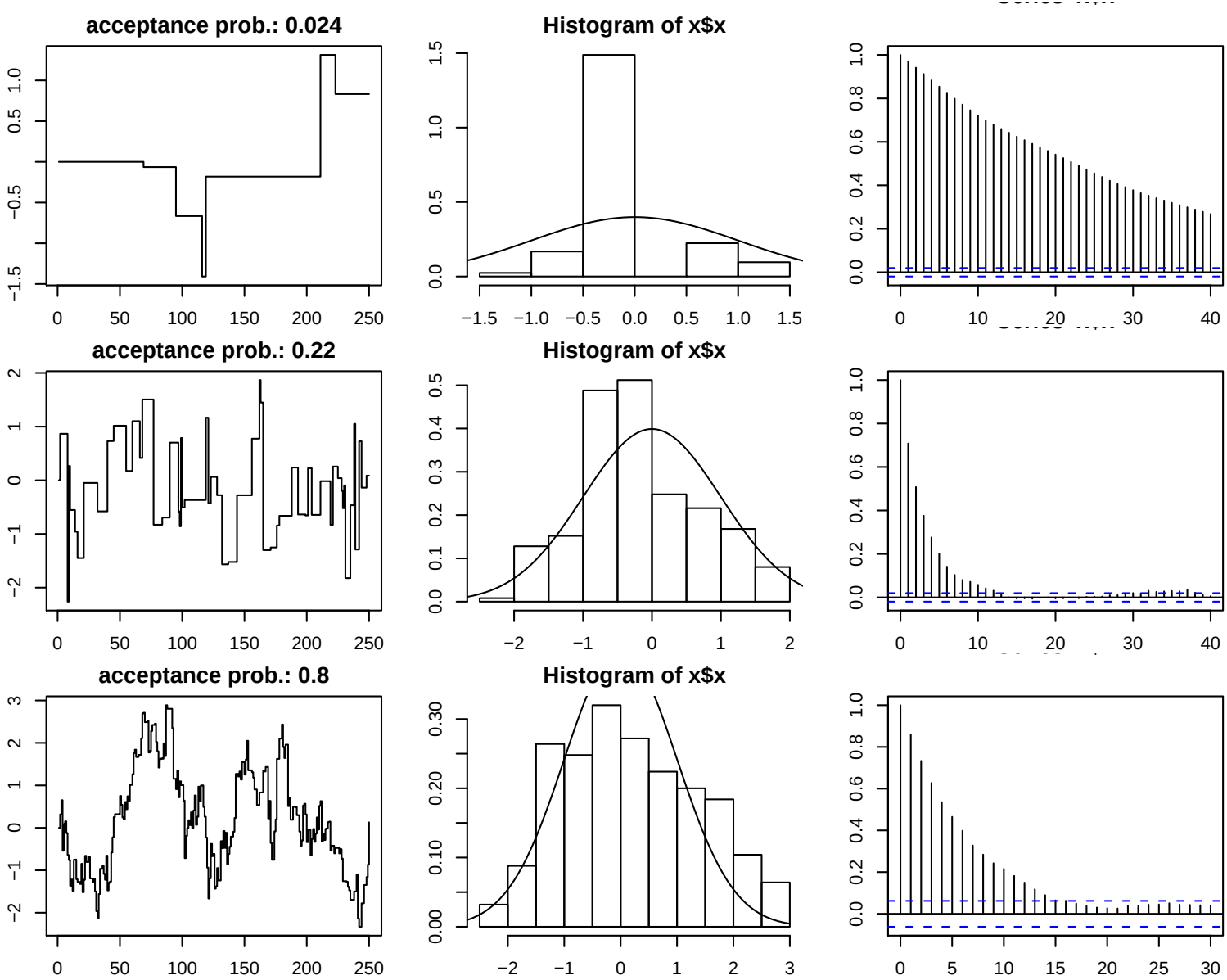
Example: Assume target density is normal

$$\pi(x) = \sqrt{\frac{\tau}{2\pi}} \exp(-\frac{1}{2}x^2)$$

The optimal choice (in terms of reducing the asymptotic variance) is so that the acceptance probability is around 40%.

If $\pi(x_1, x_2, \ldots, x_k)$ is multivariate normal, the optimal choice of τ_p corresponds to an acceptance probability of 0.234.

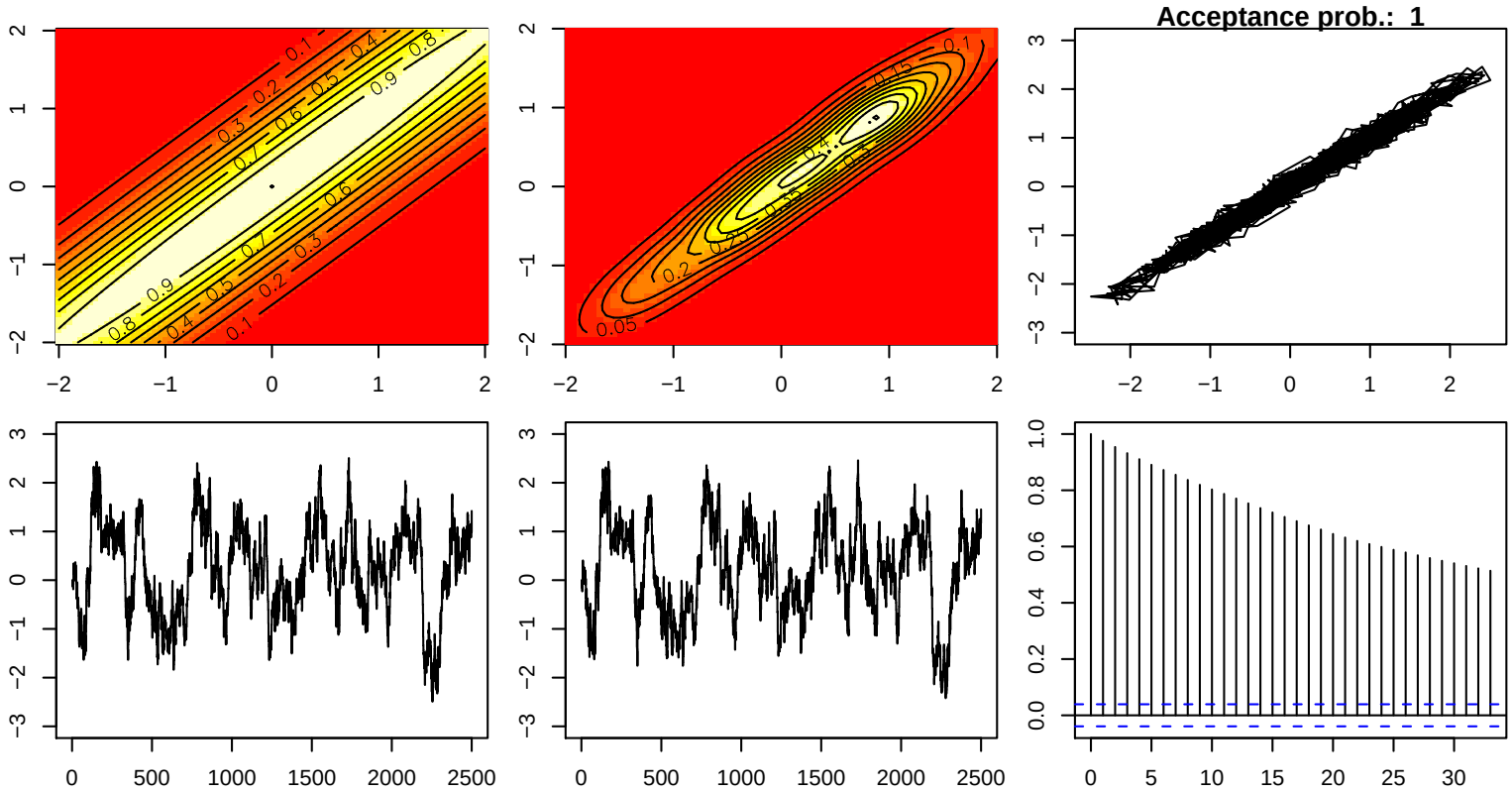
Tuning, Acceptance probability and Auto-correlation



A bivariate case example

Target: $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$

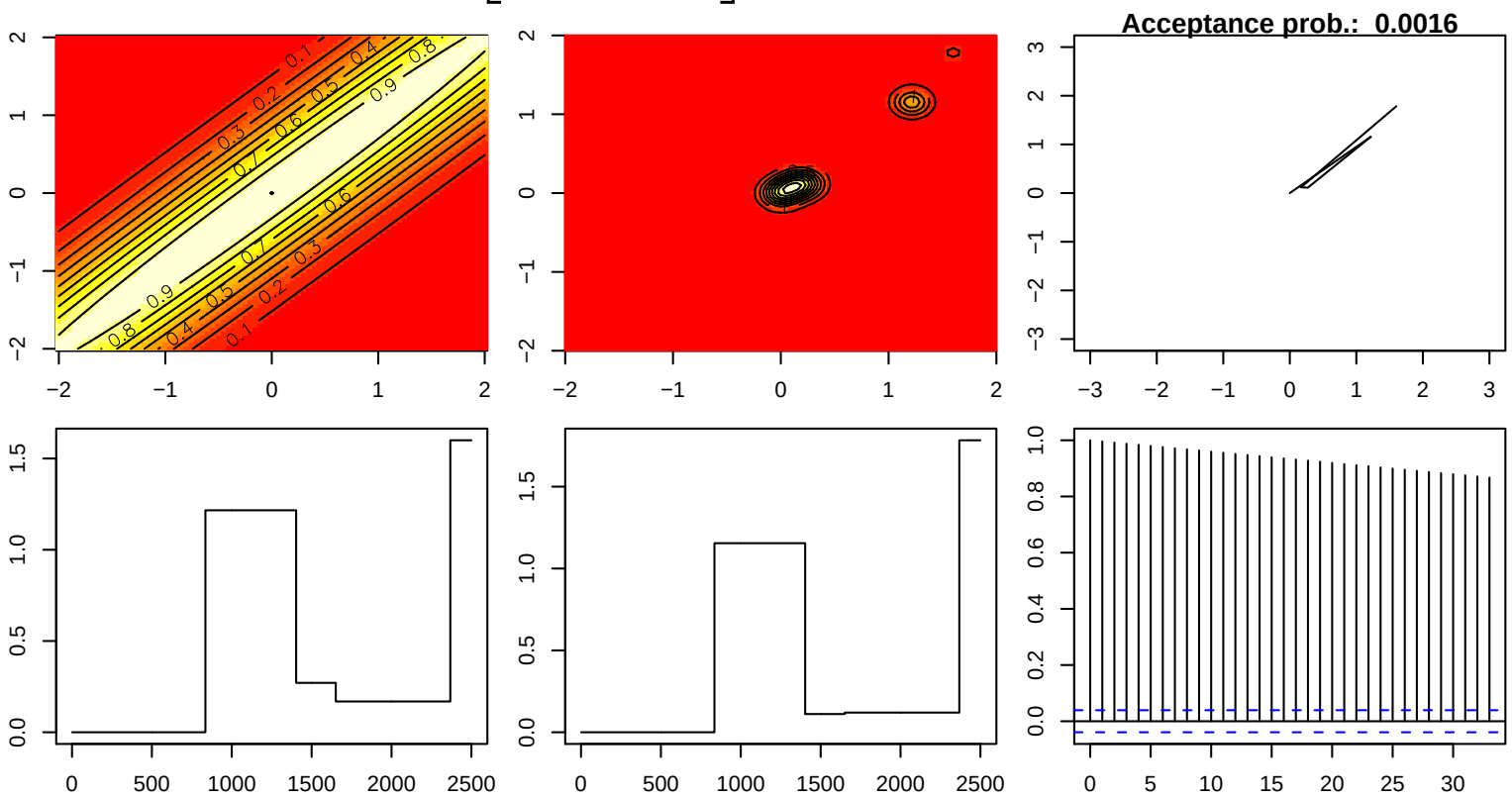
Gibbs sampling



A bivariate case example

Target: $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$

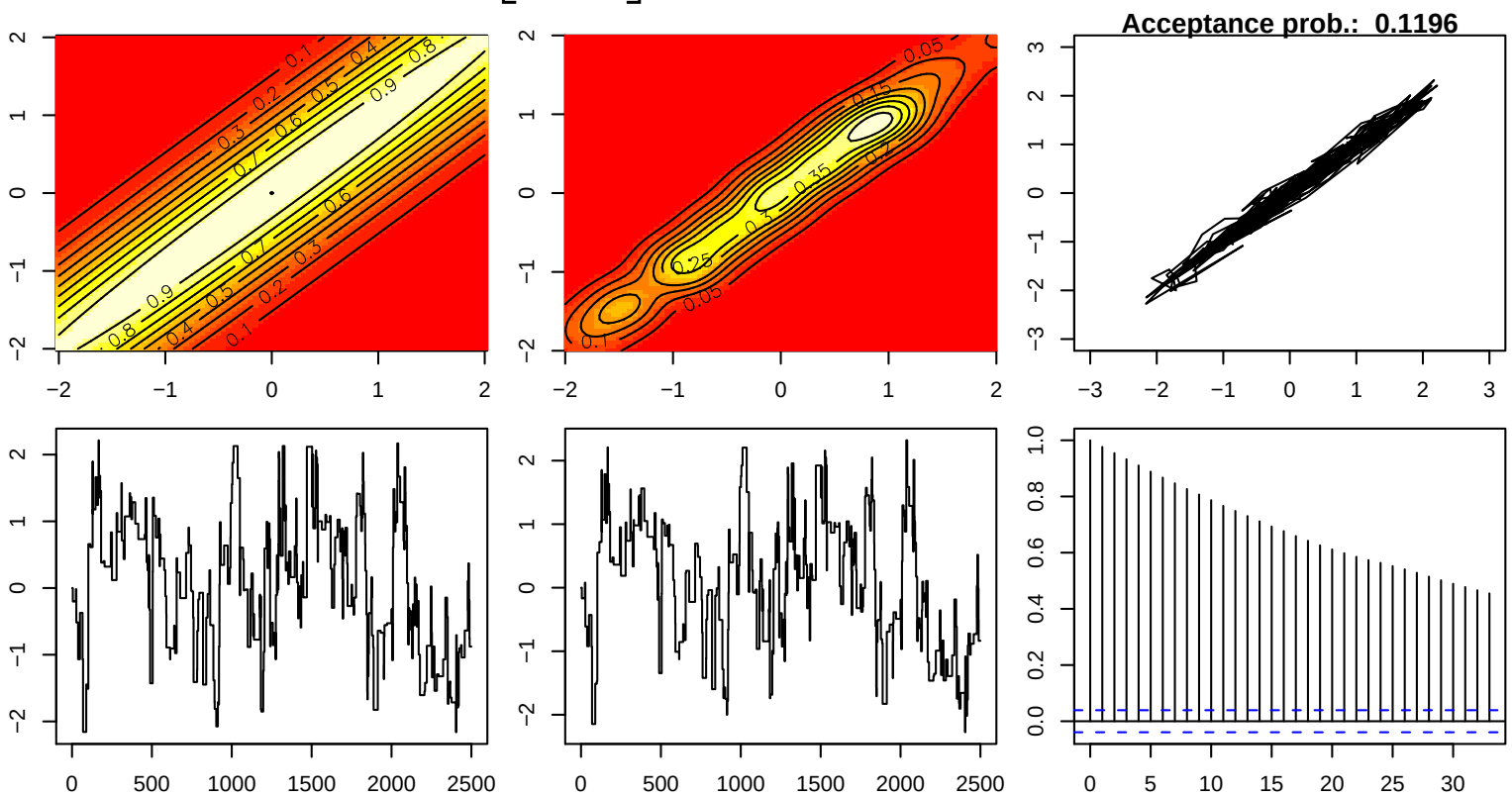
Proposal covariance $\begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$



A bivariate case example

Target: $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$

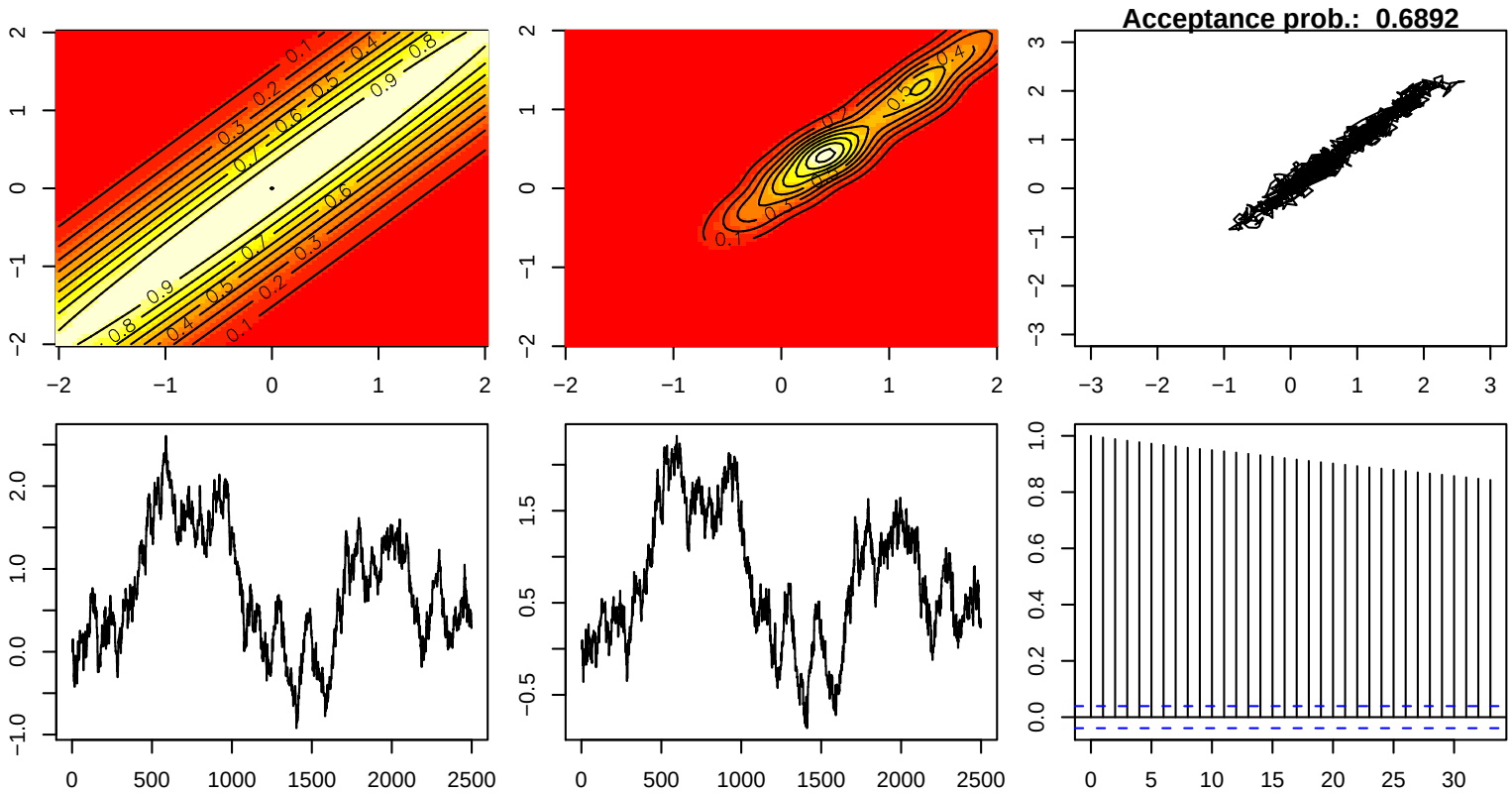
Proposal covariance: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



A bivariate case example

Target: $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$

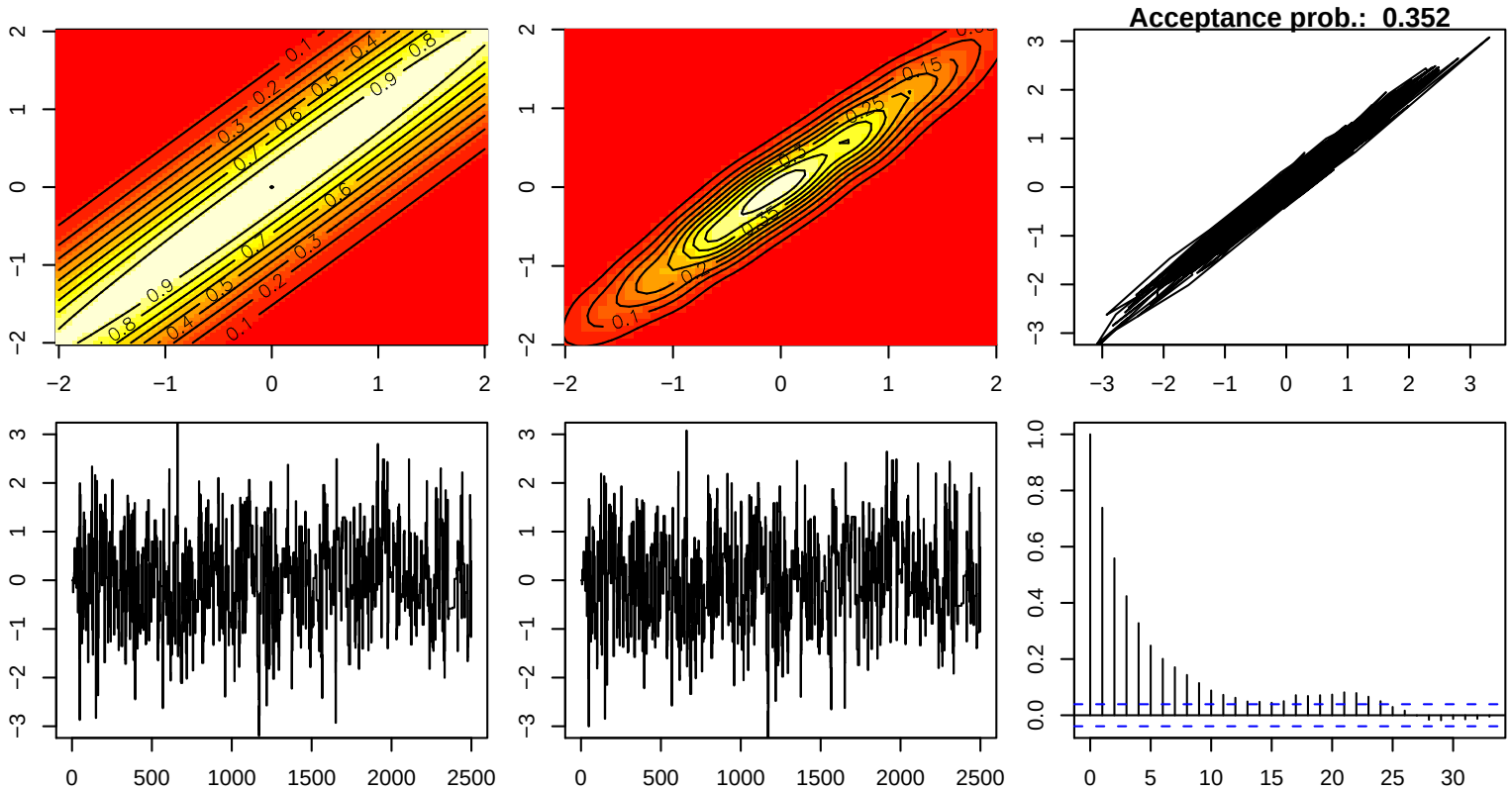
Proposal covariance: $\begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$



A bivariate case example

Target: $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$

Proposal covariance: $\frac{2.38^2}{2} \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$



Optimum proposal

Assume target is a d-dimensional normal:

$$\pi(\mathbf{x}) = \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and the proposal is normal:

$$q(\mathbf{x}, \mathbf{y}) = \mathcal{N}_d(\mathbf{x}, \boldsymbol{\Sigma}_q)$$

Then the optimum choice of proposal variance $\boldsymbol{\Sigma}_q$ is

$$\boldsymbol{\Sigma}_q = \frac{2.38^2}{d}\boldsymbol{\Sigma}$$

Catch: $\boldsymbol{\Sigma}$ is unknown.

Solutions: Pilot run or *adaptive MCMC*.

Reminder: The Gibbs sampler

Aim: We want to sample $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ from a pdf/pf $\pi(\boldsymbol{\theta})$.

Assume $\theta_i \in \Omega_i \subseteq \mathbf{R}^{d_i}$. Then, $\theta \in \Omega_1 \times \Omega_2 \times \cdots \times \Omega_k \subseteq \mathbf{R}^{d_1+d_2+\cdots+d_k}$

We can now (under some conditions) generate an *approximate* sample from $\pi(\boldsymbol{\theta})$ as follows:

Gibbs Sampler

- Choose initial value $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$.
- For $t = 1, 2, \ldots, T$
  - For $i = 1, 2, \ldots, k$
    1. Generate $\theta_i^{(t)} \sim \pi(\theta_i | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_k^{(t-1)})$

Question: What if we cannot generate samples from one or more of the full conditional distributions?

Solution: Use a Metropolis-Hastings update instead!

Metropolis within Gibbs (MwG)

Gibbs Sampler

- Choose initial value $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$.
- For $t = 1, 2, \ldots, T$
  - For $i = 1, 2, \ldots, k$
    1. Generate proposal $\theta_i' \sim q(\theta_i' | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_i^{(t-1)}, \ldots, \theta_k^{(t-1)})$
    2. Calculate Hastings ratio

$$H(\theta_i^{(t-1)}, \theta_i') = \frac{\pi(\theta_i' | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_k^{(t-1)})}{\pi(\theta_i^{(t-1)} | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_k^{(t-1)})} \times \frac{q(\theta_i^{(t-1)} | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_i', \ldots, \theta_k^{(t-1)})}{q(\theta_i' | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_i^{(t-1)}, \ldots, \theta_k^{(t-1)})}$$

    3. With probability

$$\min\left\{1, H(\theta_i^{(t-1)}, \theta_i')\right\}$$

    set $\theta_i^{(t)} = \theta_i'$ (accept) otherwise set $\theta_i^{(t)} = \theta_i^{(t-1)}$ (reject).

Metropolis within Gibbs: Comments

- Notice that each of the k component updates have $\pi(\boldsymbol{\theta})$ as their invariant distribution.
- Hence the MwG algorithm has $\pi(\boldsymbol{\theta})$ as its invariant distribution.
- Irreducibility is *not* automatically fulfilled.

- **Special case**: Assume that $q(\theta_i|\theta_{-i})$ is given by the full conditional:

$$q(\theta_i'|\theta_1^{(t)},\ldots,\theta_i^{(t)},\theta_{i+1}^{(t-1)},\ldots,\theta_k^{(t-1)}) = \pi(\theta_i'|\theta_1^{(t)},\ldots,\theta_{i-1}^{(t)},\theta_{i+1}^{(t-1)},\ldots,\theta_k^{(t-1)}).$$

  Then $H(\theta_i^{(t-1)},\theta_i') = 1$, hence all proposals are accepted.
- In fact, this is exactly the usual Gibbs sampler!

Prior predictions

Predicting future observations *without* data.

Notation: Let $\tilde{x}$ denote a prediction.

Assume:

- **Data model**: $\tilde{x}|\theta \sim \pi(x|\theta)$
- **Prior**: $\pi(\theta)$

The above assumptions implies a joint distribution of data, x, and parameter, θ:

$$\pi(x, \theta) = \pi(x|\theta)\pi(\theta).$$

We are interested in predicting a future observation, i.e. the (marginal) distribution of x, i.e. when ignoring θ, i.e.

$$\tilde{x} \sim \pi(x),$$

where

$$\pi(x) = \int \pi(x|\theta)\pi(\theta)d\theta.$$

Prior prediction: Normal case, τ known

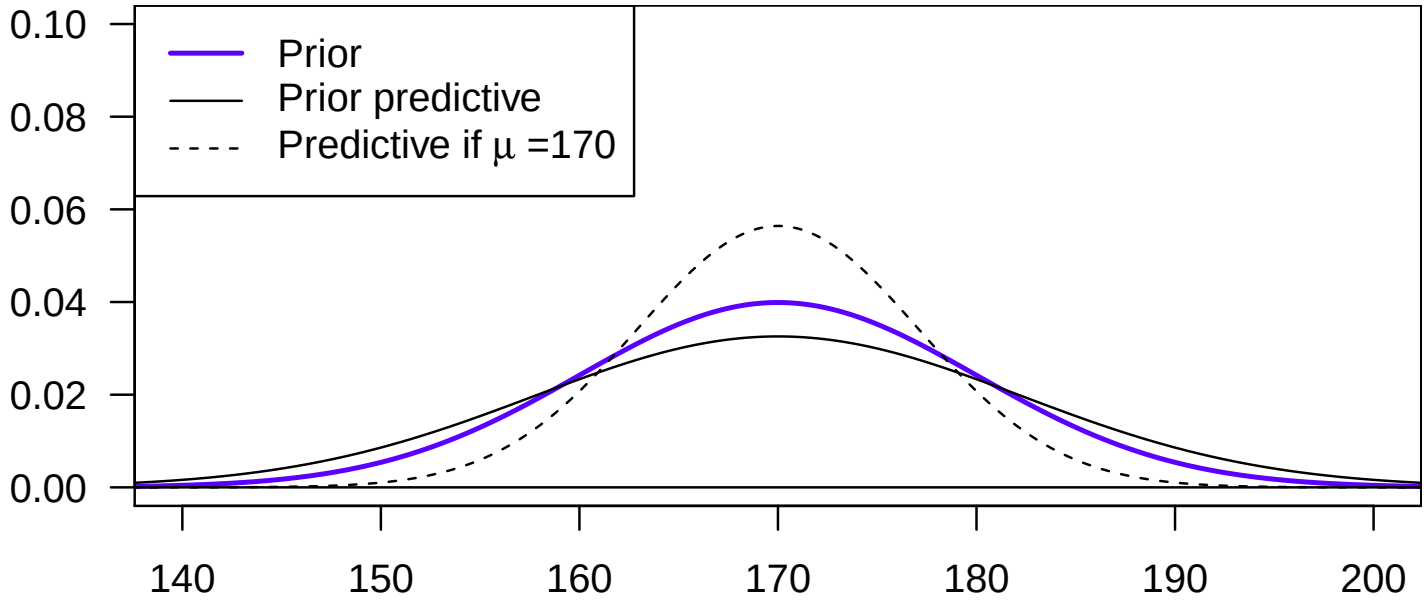
Assume:

- **Data model**: $\pi(x|\mu) \sim \mathcal{N}(\mu, \tau)$.
- **Prior**: $\pi(\mu) = \mathcal{N}(\mu_0, \tau_0)$

Prior predictive distribution

$$
\begin{aligned}
\pi(x) &= \int \pi(x|\mu)\pi(\mu)d\mu \\
&= \int \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu-\mu_0)^2\right) d\mu \\
&= \sqrt{\frac{\tau\tau_0}{\tau+\tau_0}\frac{1}{2\pi}} \exp\left(-\frac{1}{2}\frac{\tau\tau_0}{\tau+\tau_0}(x-\mu_0)^2\right) \\
&= \mathcal{N}\left(\mu_0, \frac{\tau\tau_0}{\tau+\tau_0}\right).
\end{aligned}
$$

Prior predictive distribution



Simulating the prior predictive distribution

If $\pi(x)$ is difficult to derive or not easily simulated from *directly* we can use another strategy.

Simulating the prior predictive distribution can be done as follows:

1. Generate parameter from prior: $\theta \sim \pi(\theta)$
2. Conditional on θ generate x: $\tilde{x} \sim \pi(x|\theta)$

Now $\tilde{x}$ is a sample from the prior predictive distribution.

Posterior prediction

Predicting future observation *given* data.

Assume:

- **Data model**: $x|\theta \sim \pi(x|\theta)$
- **Prior**: $\pi(\theta)$

The joint distribution of predicted data $\tilde{x}$, data x and parameter θ is

$$\begin{aligned}\pi(\tilde{x}, x, \theta) &= \pi(\tilde{x}|\theta)\pi(x|\theta)\pi(\theta) \\ &\propto \pi(\tilde{x}|\theta)\pi(\theta|x).\end{aligned}$$

Notice: Here $\pi(\tilde{x}|\theta)$ and $\pi(x|\theta)$ represent the same distribution.

The posterior predictive distribution is the (marginal) distribution of $\tilde{x}$ conditional on data x:

$$\begin{aligned}\pi(\tilde{x}|x) &= \int \pi(\tilde{x}, \theta|x)d\theta = \int \frac{\pi(\tilde{x}, \theta, x)}{\pi(x)}d\theta \propto \int \pi(\tilde{x}|\theta)\pi(x|\theta)\pi(\theta)d\theta \\ &\propto \int \pi(\tilde{x}|\theta)\pi(\theta|x)d\theta\end{aligned}$$

Posterior prediction: Normal case, τ known

Data model: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau)$.

Prior: $\pi(\mu) = \mathcal{N}(\mu_0, \sigma_0)$.

Posterior: $\pi(\mu|\mathbf{x}) = \mathcal{N}\left(\frac{n\tau\bar{x}+\tau_0\mu_0}{n\tau+\tau_0}, n\tau + \tau_0\right)$.

Recall that the prior prediction (of one observation) was

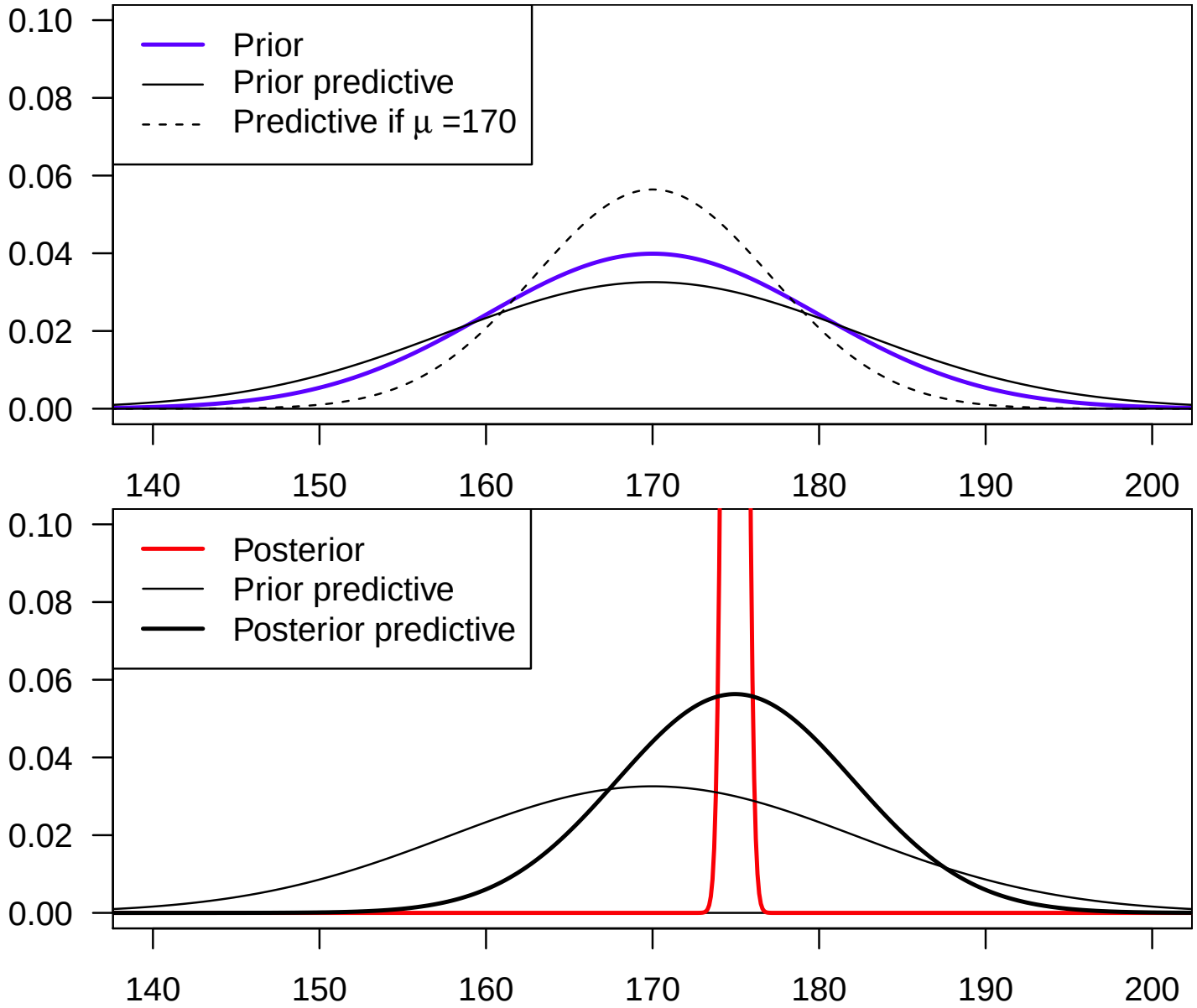
$$\tilde{x} \sim \mathcal{N}\left(\mu_0, \frac{\tau_0\tau}{\tau + \tau_0}\right)$$

Since the posterior is the "prior" for the posterior prediction we have

$$\tilde{x}|\mathbf{x} \sim \mathcal{N}\left(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}, \frac{(n\tau + \tau_0)\tau}{\tau + n\tau + \tau_0}\right)$$

When n is large we have $\tilde{x}|\mathbf{x} \overset{approx}{\sim} \mathcal{N}(\bar{x}, \tau)$.

Prior and posterior predictive distributions



Posterior prediction using a graph

Model checking

Idea: If the model is correct posterior predictions of the data should look similar to the observed data.

Difficulty: Who to choose a good measure of "similarity".

Example: We have observed a sequence of $n = 20$ zeros and ones:

$$1\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

Model: $X_1, X_2, \ldots, X_{20}$ are IID and $P(X_i = 1) = p$.

Prior: $\pi(\pi) = Be(\alpha, \beta)$.

Posterior: $\pi(\pi|\mathbf{x}) = Be(\#1 + \alpha, \#0 + \beta)$.

Model checking: We simulate posterior predictive realisations $\tilde{\mathbf{X}}^{(1)}, \tilde{\mathbf{X}}^{(2)}, \ldots, \tilde{\mathbf{X}}^{(N)}$, where

$$\tilde{\mathbf{X}}^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_2^{(n)}).$$

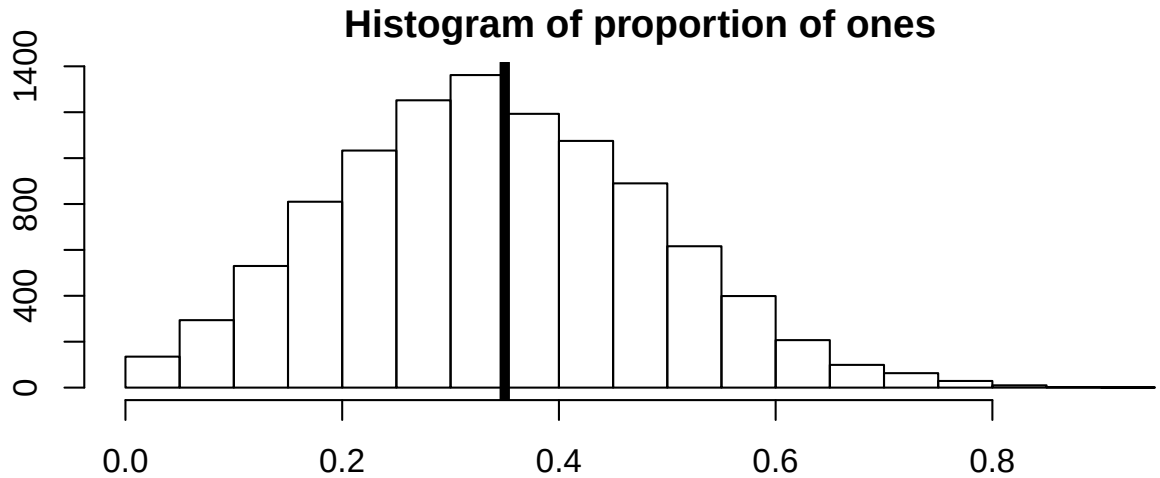
If these vectors look "similar" to the data above, the model is probably ok.

Model checking: First attempt (A failure)

Define summary function

$$s(\mathbf{x}) = \text{Number of ones in the sequence } \mathbf{x}$$

Histogram for $s(\tilde{\mathbf{x}}^{(i)})$ for $N = 10,000$ independent posterior predictions:



Clearly the observed number of ones is in no way unusual compared to the posterior predictions.

This is really expected — so we need another summary function $s(x)$.

Model checking: Second attempt (A success)

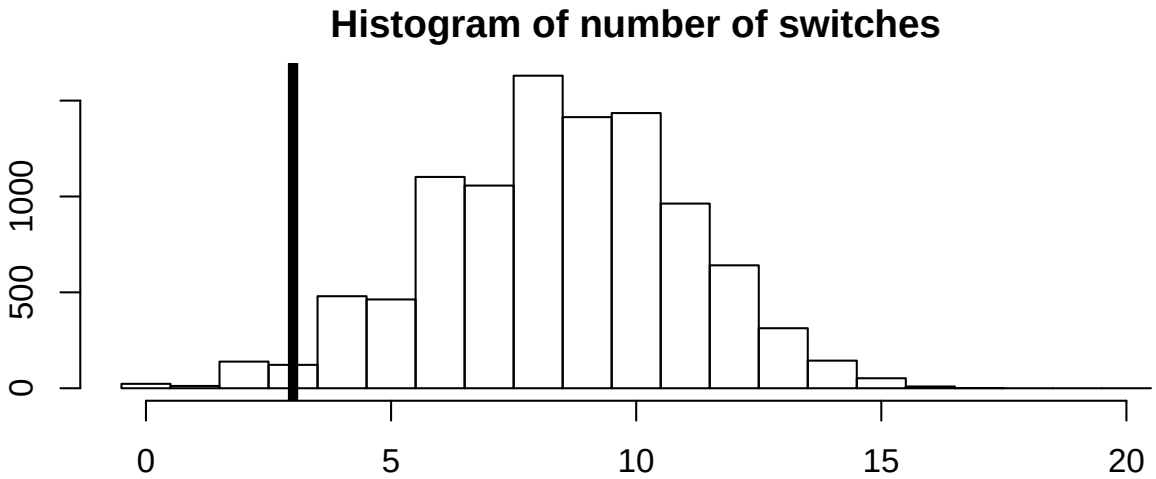
Define summary function

$$s(\mathbf{x}) = \text{Number of switches between ones and zeros in } \mathbf{x}$$

In the data the number of switches is 3:

$$1\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

Histogram for $s(\tilde{\mathbf{x}}^{(i)})$ for $N = 10,000$ independent posterior predictions:

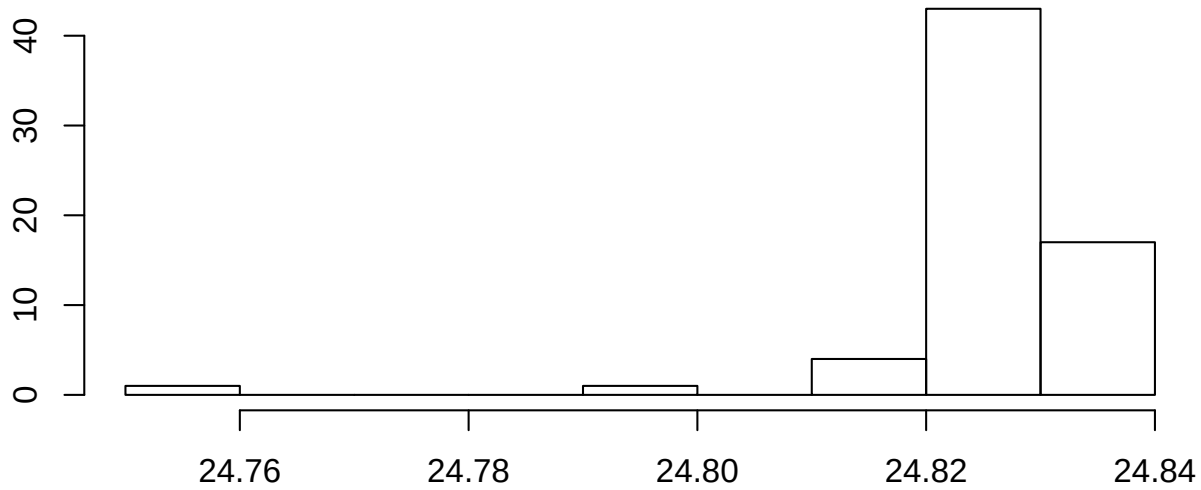


Only around 1.7% of the posterior prediction have 3 or fewer switches.

This suggests that the model assumption of independence is questionable.

Example: Speed of light

66 measurements of the time it takes light to travel 7445 meters:



Data model:

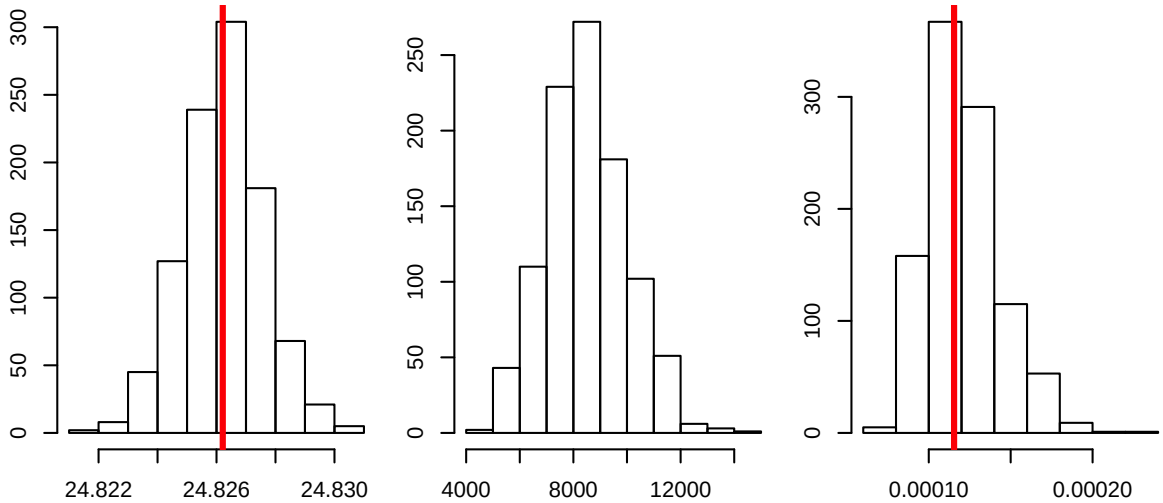
$$x_1, \ldots, x_{66} \overset{iid}{\sim} \mathcal{N}(\mu, \tau)$$

Prior:

$$\pi(\mu, \tau) = \mathcal{N}(\mu; 0, 0.001) \times Gamma(\tau; 0.001, 1000)$$

Example: Speed of light

Poestrior distribution of μ, τ and $1/\tau$:



Red lines denote sample mean and sample variance, respectively.

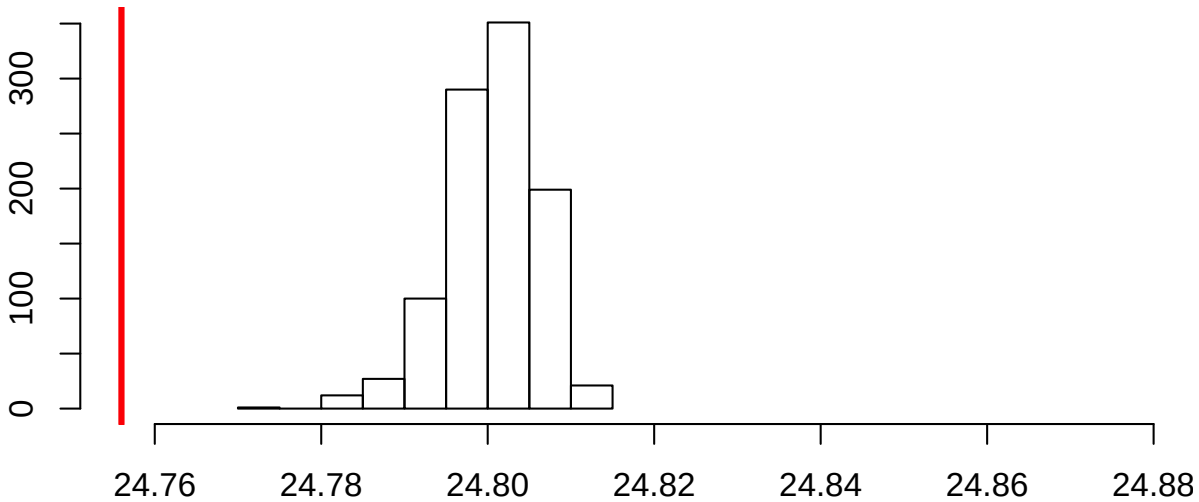
Seem reasonable.

Example: Speed of light

Data contain one very low measurement. Is this unusual?

Generate 1000 posterior predictive samples $\mathbf{x}^{(i)} = x_1^{(i)}, \ldots, x_{66}^{(i)}$, $i = 1, \ldots, 1000$

Definer $s(\mathbf{x}) = \min\{x_1, \ldots x_{66})$



Conclusion: The smallest value in the data is very unlikely under the assumed model.