MCMC and Gibbs Sampling

Alexandra Posekany

WS 2020

Revision of Bayesian Inference

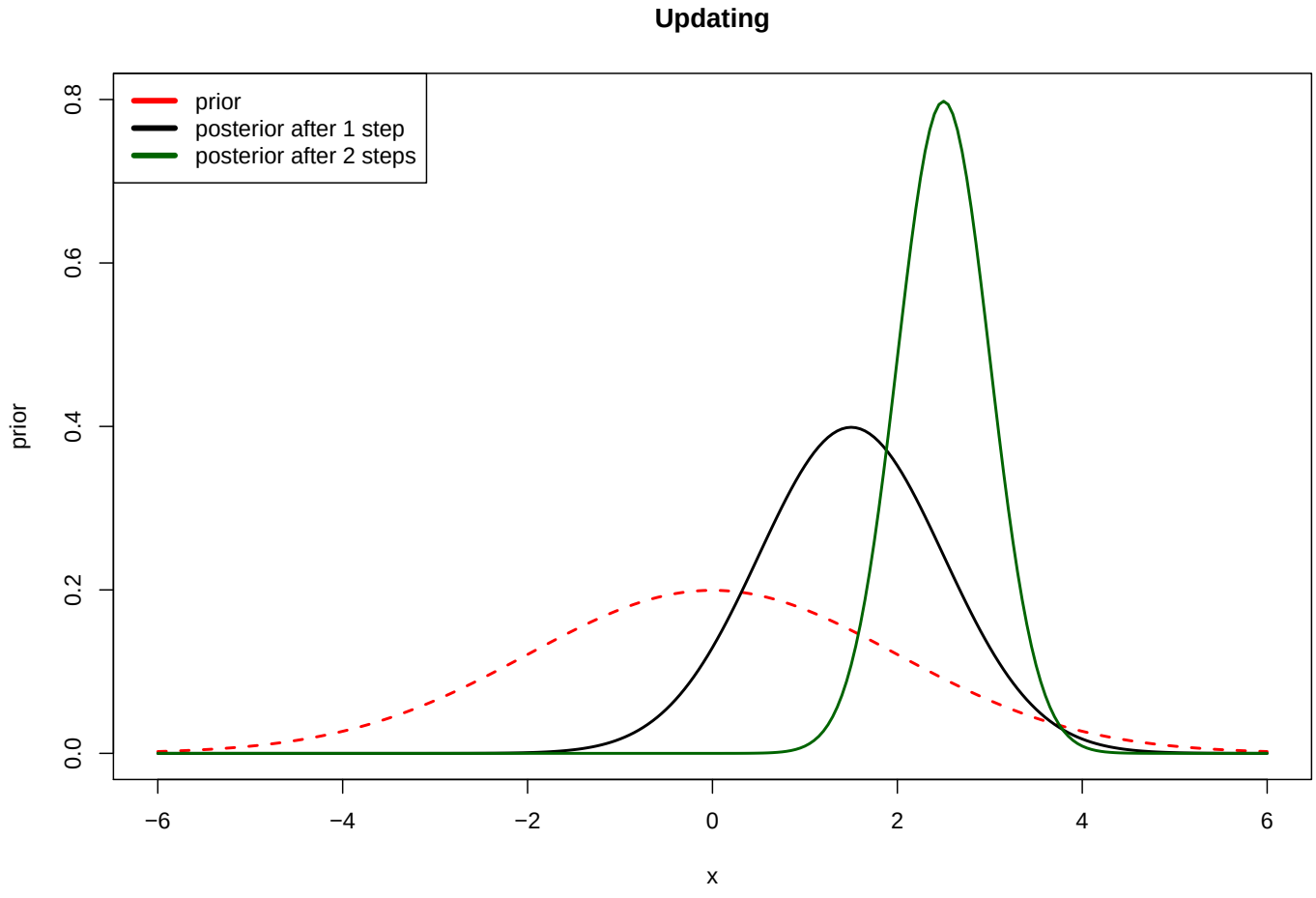
Bayes' Theorem

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_\Theta \pi(\theta)f(x|\theta)d\theta}.$$

The denominator contains the **marginal distribution**, $m(x) = \int_\Theta \pi(\theta)f(x|\theta)d\theta$ which has the normalisation purpose of assuring that $\pi(\theta|x)$ is a probability distribution.
The quintessential parts of the Bayesian model are:

- the **prior distribution** $\pi(\theta)$ which expresses the uncertainty about a model parameter θ from parameter space Θ;
- the **Likelihoodfunction** $f(x|\theta)$ which transforms the information contained in the data to the model structure and evaluates its 'fittingness',
- and the resulting **posterior distribution** $\pi(\theta|x)$.

Updating


Updating
prior
posterior after 1 step
posterior after 2 steps
prior
0.0
0.2
0.4
0.6
0.8
−6
−4
−2
0
2
4
6
x

Natural conjugate priors

- Assume that the *a priori* distribution $p(\boldsymbol{\theta})$ belongs to a class of parametric distributions $\mathcal{F}$.

  (can be either an **Informative** or **uninformative** prior)

- One says that the *a priori* distribution $p(\boldsymbol{\theta})$ is conjugate with respect to the likelihood $L(\theta|\mathbf{y})$, if the *a posteriori* distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ also belongs to $\mathcal{F}$, i.e.,

$$p(\boldsymbol{\theta}) \in \mathcal{F} \Rightarrow p(\boldsymbol{\theta} \mid \mathbf{y}) \in \mathcal{F}.$$

  When this happens, the family $\mathcal{F}$ is said a **closed** family under sampling of $\mathcal{M}$.

- The computation of the *posterior* distribution is easier when we use *prior* distributions that are conjugate with respect to the likelihood.

- In fancier models, conjugate priorsfacilitate Gibbs sampling, which is the easiest Bayesian computational algorithm.

- Until the 1990s, Bayesian inference was almost all done with conjugate *priors*.

Natural conjugate priors

- Below we have a list of natural conjugate *priors* given a specific likelihood:

Likelihood	*priori*	*posteriori*
Binomial	Beta	Beta
Negative binomial	Beta	Beta
Poisson	Gamma	Gamma
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Normal (μ unknown)	Normal	Normal
Normal (σ^2 unknown)	Gamma (or IG)	Gamma (or IG)
Normal (μ, σ^2 unknown)	Normal/Gamma (or IG)	Normal/Gamma (or IG)
Multinomial	Dirichlet	Dirichlet

Note that in the case of the Normal likelihood with σ^2 unknown, one refers to the Normal-Inverse Gamma model when working with σ^2 and to the Normal-Gamma model when working with the precision $\tau = 1/\sigma^2$.

Natural conjugate priors

Final important notes on conjugate *priors*

- Conjugate *priors* are convenient but may not always be flexible enough.

  When this is the case,

  - **mixtures** of conjugate *priors* can be a good alternative

- A **mixture** *prior* is given by

$$p(\theta) = \sum_{j=1}^{J} \pi_j p_j(\theta)$$

  where $p_j(\theta)$ are conjugate *priors* with distinct hyperparameters and the mixture weights π_j sum to one.

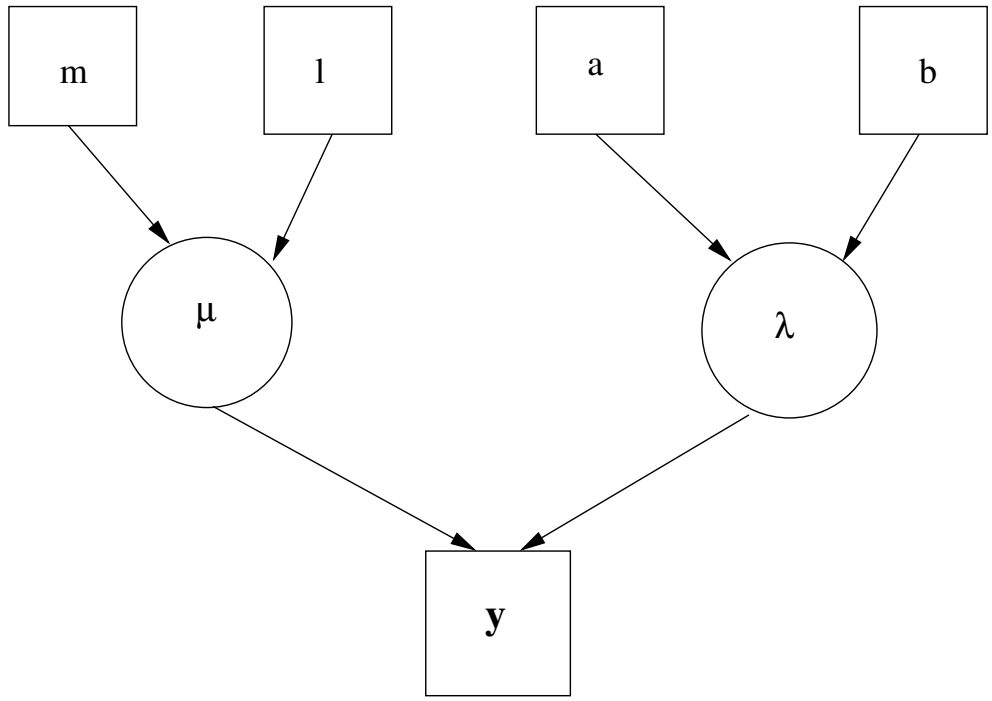
  And the best thing is, not only are they quite flexible but one also has that

  - **mixtures** of conjugate *priors* are conjugate!

Directed Acyclic Graphs

Bayesian paradigm:
consider parameters as random variables
$\rightarrow$ add prior on parameter, additional latent parameters

Directed acyclic graph (DAG): visualisation of hierarchical model



Mathematical Model

The DAG illustrates the following distributions and their parameters:

$$\begin{aligned} y|\mu,\lambda &\sim N(\mu,\lambda^{-1}) \\ \mu|m,l &\sim N(m,l^{-1}) \\ \lambda|a,b &\sim Gamma(a,b) \end{aligned}$$

based on a Likelihood function

$$f(y|\mu,\lambda) \propto \lambda^{n/2}\exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}.$$

The directions visualise the hierarchical model structure with the data and its likelihood at the centre and priors for these parameters on the next hierarchical level.

Mathematical Model

The Normal- and Gamma- distribution are natural conjugate priors for any models based on normally distributed likelihood functions, such as the equivalent of t-test for Bayesian settings based on normal distributions priors and posteriors of the mean, the linear regression model and probit regression model. All have posteriors and "full conditional" distributions of the following forms.

$$\begin{aligned} \mu|\lambda, y &\sim N(m^*, l^*) \\ m^* &= l^* \cdot (l \cdot m + \lambda \cdot n \cdot \bar{y}) \\ l^* &= l + n \cdot \lambda \end{aligned}$$

are the parameters of the normal posterior of the mean and

$$\begin{aligned} \lambda|\mu, y &\sim Gamma(a^*, b^*) \\ a^* &= a + \frac{n}{2} \\ b^* &= b + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \end{aligned}$$

of the Gamma-posterior of the precision (inverse variance) where typical for Bayesian inference the Likelihood-based pivot is **biased** by the prior information.

Posterior and Full conditionals

In the above example we used the term posterior, although the normal posterior for the mean only applies, if the precision where known, and the Gamma posterior for the precision only applies, if the mean where known. You have dealt with these **1-dimensional posterior** scenarios in your last exercises.

If both parameters where to be determined simultaneously, a **two-dimensional posterior** would have to be constructed. Its **marginal distribution** for every value of λ would be the normal posterior of the mean μ, while its **marginal distribution** for every value of μ would be the Gamma posterior of the precision λ. Thus, conditional on the value of λ the 1-dimensional normal posterior for the mean is the **marginal distribution** of this 2-dimensional posterior and therefore referred to as **full conditional posterior**. These **full conditionals** wil be the basic building stones of Gibbs samplers for MCMC simulation.

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods unite 2 principal concepts:

1. Markov Chains and
2. Monte Carlo sampling.

Revision of Monte Carlo sampling

- Monte Carlo sampling is the predominant method of Bayesian inference because it avoids asymptotic approximations and can be used in high-dimensions.
- The main idea is to approximate posterior summaries by drawing samples from the posterior distribution, and then using these samples to approximate posterior summaries of interest.
- For example, if $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$ are samples from $p(\boldsymbol{\theta} \mid \mathbf{y})$, then the mean of S samples can be used to approximate the posterior mean.
- This only provides approximations of the posterior summaries of interest.
- Many argue that this form of approximation is superior to asymptotic approximations because the Bayes CLT requires the sample size of the dataset to go to infinity and the Monte Carlo approximation requires the number of simulated values to go to infinity.
- In most cases, $S \to \infty$ is cheaper and more realistic than $n \to \infty$.
- But how to draw samples from some arbitrary distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$?

Continuous State Space Markov Chains

Markov Chains

- **Markov Chain**

  A sequence of random variables $\theta_0, \theta_1, \theta_2, \ldots$ defined in a discrete state space $\{1, 2, \ldots, N\}$ is said a **Markov chain** if

  $$\Pr(\theta_{n+1} = j \mid \theta_n = i, \theta_{n-1} = i_{n-1}, \ldots, \theta_0 = i_0) = \Pr(\theta_{n+1} = j \mid \theta_n = i), \quad \forall n \geq 0.$$

  Basically, the future is independent from the past given the present.

  - $p_{ij} = \Pr(\theta_{n+1} = j \mid \theta_n = i)$ is said the transition probability from state i to state j and

    $$\mathrm{P} = (p_{ij})_{i,j=1,\ldots,N}$$

    is the $N \times N$ **transition matrix** of the chain.

(see **[3]** for more detail)

Transition Kernel

Transition kernel

A transition kernel is a function $\mathcal{K}$ defined on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ defining the probability of choosing a value x_{n+1} from set A of

$$P[X_{n+1} \in A | X_n = x_n] \quad = \quad \int_A \mathcal{K}(x_n, dx)$$

such that

1. $\mathcal{K}(x, .)$ is a **probability measure** $\forall x \in \mathcal{X}$: , i.e. for every fixed value x of the state space $\mathcal{X}$ $\mathcal{K}(x, .)$ is a function operating on the Borel sets which assigns a probability (depending on x) to every set of $\mathcal{B}(\mathcal{X})$;
2. $\mathcal{K}(., A)$ is a **measurable function** $\forall A \in \mathcal{B}(\mathcal{X})$:, i.e. for every fixed set A the function $\mathcal{K}(., A)$ operates on the state space $\mathcal{X}$ and is measurable.

Overview of Markov chain properties

explore whole space

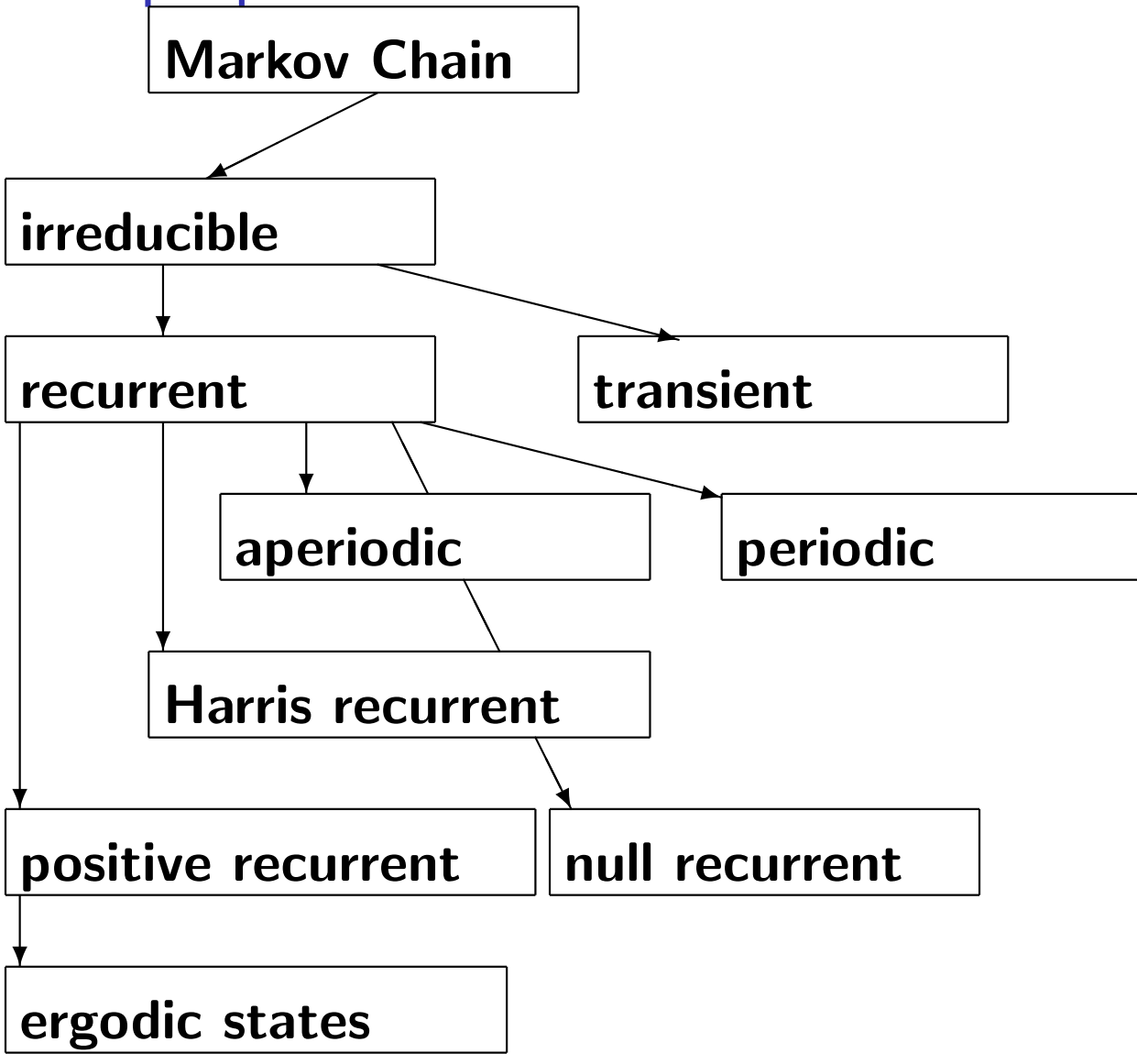
reach every set often enough

move in cycles

finite stopping times

existence of invariant measure

limiting behaviour



Detailed balance conditions

Detailed Balance Condition

Definition: A Markov chain with transition kernel $\mathcal{K}(.,.)$ satisfies the detailed balance condition if there exists a function f satisfying

$$\mathcal{K}(y,x)f(y) = \mathcal{K}(x,y)f(x) \quad \forall(x,y)$$

Detailed Balance Condition, reversibility

If a Markov chain with transition kernel $\mathcal{K}$ satisfies the detailed balance condition with π a probability density function, the following statements hold true:

1. The density π is the invariant density of the chain.
2. The chain is reversible.

Markov Chain Monte Carlo

Basic MCMC samplers:

- The **Metropolis-Hastings** sampler: most universal sampling scheme
- The **Gibbs** sampler: most commonly used, simple to understand and straightforward to calculate and implement
- The **Reversible Jump** sampler and the **Birth and Death** sampler: deal with varying parameter sizes
- **Hybrid** sampler: combines at least 2 of the sampling approaches

Markov Chain Monte Carlo

Idea: to obtain samples from a distribution without this distribution being explicitly available

Aim: constructing an ergodic Markov chain with stationary distribution ξ in order to acquire samples from that distribution.

plug sampled values into the Monte Carlo integration

in a Bayesian frame work: posterior distributions often analytically intractible for complex, e.g. hierarchical Bayesian models

Metropolis-Hastings sampler

Rejection sampling (revisited)

- Let X be a continuous random variable with p.d.f. f from which it is not easy (or possible) to sample from

- Assume we find a density function $g(x)$ from which we can easily simulate from and such that

$$\exists M > 0 : f(x) \leq Mg(x), \qquad \forall x$$

The **acceptance-rejection method** proceeds as follows

Step 1 `Generate a candidate observation` x_c `from` g

Step 2 `Compute the probability of accepting` x_c `as`

$$\alpha = \frac{1}{M} \frac{f(x_c)}{g(x_c)}$$

Step 3 `Generate an observation` u `from a` $U(0,1)$ `distribution`

Step 3 `If` $u \leq \alpha$ `set` $x = x_c$`. Otherwise go back to step 1.`

Step 4 `Repeat previous steps until you reach the desired sample size`

Link between Rejection sampling and Metropolis-hastings sampling

- Let θ be a continuous random variable with posterior distribution $p(\theta|\mathbf{y})$ from which it is not easy (or possible) to sample from

- Assume we find a density $g(\theta)$ (**envelope function**) from which we can easily simulate from and such that

$$\exists M > 0 : p(\theta|\mathbf{y}) \leq Mg(\theta), \qquad \forall \theta$$

(In this case one says that $p(\theta|\mathbf{y})$ resides in the envelope)

The **acceptance-rejection method** proceeds as follows

Step 1 `Generate a candidate observation` θ_c `from` g

Step 2 `Compute the probability of accepting` θ_c `as`

$$\alpha = \frac{1}{M} \frac{p(\theta_c|\mathbf{y})}{g(\theta_c)}$$

Step 3 `Generate an observation` u `from a` $U(0,1)$ `distribution`

Step 3 `If` $u \leq \alpha$ `set` $\theta = \theta_c$`. Otherwise go back to step 1.`

Step 4 `Repeat previous steps until you reach the desired sample size`

Metropolis-Hastings sampler

Aim: drawing $(x^{(t)})$ such that $(x^{(t)})$, $t = 0, 1, \ldots$ are a Markov chain with stationary distribution being the objective **target density** ξ

Approach: auxiliary conditional distribution, **proposal density** $q(.|.)$ of a proposed value given the 'old' value.

good proposal

- easy to simulate from or
- symmetric (i.e.$q(x|y) = q(y|x)$) so that it cancels out in the acceptance probability

Metropolis-Hastings sampler

- For $t = 0$: take starting value x_0
- $t > 0$:
  1. generate proposal $Y_t \sim q(y|x^{(t-1)})$
  2. Either

     move to proposed value Y_t with $\alpha(x^{(t-1)}, Y_t)$ or

     stay at old value $x^{(t-1)}$ with $1 - \alpha(x^{(t-1)}, Y_t)$

     where $\alpha(x, y) = \min\left\{\frac{\xi(y)}{\xi(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$ is the acceptance probability.

     The transition kernel of the Metropolis-Hastings sampler is

$$\mathcal{K}(x, y) = \alpha(x, y)q(y|x) + (1 - \int \alpha(x, y)q(y|x)dy)\delta_x(y)$$

MH stationary distribution

The generic Metropolis-Hastings algorithm is well-defined for any target and proposal distribution, but certain regularity conditions are important for ξ to be the limiting distribution:

- $\cup_{x \in supp_\xi} supp_{q(.|x)} \supset supp_\xi$, the minimal necessary condition for ξ to be the limiting distribution of the chain.
- let $supp_\xi$ be connected, not necessary, but very helpful for applications and important for irreducibility and existence of a single stationary distribution
- Tail behaviour: for efficient MH sampler the tails of the proposal should be heavier than the tails of the target distribution

Example: Metropolis-Hastings for student's t distribution

Consider a non-central student's t-distribution model with known degrees of freedom ν and scale 1.

$$\begin{aligned} X &\sim t_\nu(\theta, 1) \\ f(x, \theta) &\propto (\nu + (x-\theta)^2)^{-\frac{\nu+1}{2}} \end{aligned}$$

We choose a flat prior for θ: $\pi(\theta) \propto 1$, and the proposal distribution is standard normal $N(0,1)$. Given 1 sample of x, $\theta^{(t-1)}$ and the proposal ζ drawn from $N(0,1)$ the acceptance probability for run $t \geq 1$ would be:

$$\alpha(\theta^{(t-1)}, \zeta) = \left(\frac{\nu + (x-\zeta)^2}{\nu + (x-\theta^{(t-1)})^2} \right)^{-\frac{\nu+1}{2}} \frac{\exp\left(-\frac{1}{2}(\theta^{(t-1)})^2\right)}{\exp\left(-\frac{1}{2}\zeta^2\right)}$$

for any proposed value of θ within the parameter's support.

MH stationary distribution

Detailed balance condition

Let $(X^{(t)})$ be the chain produced by the Metropolis-Hastings algorithm. For every conditional distribution $q(.|.)$ whose support includes the support of the target distribution $\xi(.)$, the following two statements hold:

1. the kernel of the chain satisfies the detailed balance condition with ξ.
2. ξ is a stationary distribution of the chain.

Gibbs sampler

Gibbs sampler

Idea: use **full conditional distributions**

$$\xi_i(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p) \quad i = 1, 2, \ldots, p$$

associated with the target distribution to generate samples from target distribution, if we can sample from these distributions

Thus, unlike the MH sampler the Gibbs sampler is by definition **multidimensional**! (at least two variables are required for the conditional distributions)

Gibbs sampler

- **Gibbs sampling** was proposed in the early 1990s (Geman and Geman, 1984; Gelfand and Smith, 1990) and fundamentally changed Bayesian computing.
  - It is attractive because it can sample from high-dimensional posteriors
  - The main idea is to break the problem of sampling from the high-dimensional joint distribution into a series of samples from low-dimensional conditional distributions. (the full conditional *posteriors*)
- The algorithm is straightforward:
  - One begins by setting initial values for all parameters, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.
  - Variables are then sampled one at a time from their **full conditional distributions**

$$p(\theta_j \mid \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p, \mathbf{y}), \qquad j = 1, \ldots, d.$$

  - Rather than 1 sample from p-dimensional joint, we make p-dimensional samples.

Gibbs sampler

- Generally, given a parameter vector $\theta = (\theta_1, \ldots, \theta_d)$, the Gibbs sampler works as follows.

Step 1 Specify initial values $(\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.

Step 2 For $t = 1, \ldots, T$

2.1 Simulate $\theta_1^{(t)} \sim p(\theta_1 \mid \mathbf{y}, \theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)})$

2.2 Simulate $\theta_2^{(t)} \sim p(\theta_2 \mid \mathbf{y}, \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)})$

...

2.d Simulate $\theta_d^{(t)} \sim p(\theta_d \mid \mathbf{y}, \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{d-1}^{(t)})$

Step 3 Discard the first k observations of the chain and compute the summary statistics from the *posterior* distribution based on $(\theta_1^{(k+1)}, \ldots, \theta_d^{(k+1)}), \ldots, (\theta_1^{(T)}, \ldots, \theta_d^{(T)})$

Gibbs sampler

The transition kernel of this algorithm is therefore

$$\mathcal{K}(x^{(t+1)}|x^{(t)}) = \prod_{j=1}^{p} \xi(x_j^{(t+1)}|x_1^{(t+1)}, \ldots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \ldots, x_p^{(t)})$$

One has for t sufficiently large that

$$(\theta_1^{(t)}, \ldots, \theta_d^{(t)}) \overset{\text{approx.}}{\sim} p(\theta_1, \ldots, \theta_d \mid \mathbf{y})$$

where this chain of values is a **Markov chain**.

Gibbs sampler

- a single Gibbs transition can be viewed as a special case of a single component Metropolis-Hastings move with acceptance probability 1
- $\Rightarrow$ 2-stage Gibbs sampler inherits all properties of the Metropolis-Hastings Sampler
- not the case for the multi-stage Gibbs sampler,
  $\rightarrow$ most well-behaved example of a hybrid sampler
- as an MH step: limited choice of proposal distributions and prior knowledge of some analytical or probabilistic properties of the target distribution, as defined by the full conditionals
- impossible to vary the number of parameters

Two-stage Gibbs sampler

For the two-stage Gibbs sampler, not only $(X_1^{(t)}, X_2^{(t)})$ is a Markov chain, but also the marginals $(X_1^{(t)})$ and $(X_2^{(t)})$ are **Markov chains**, with transition kernel

$$K(x_1, x_1^*) = \int f_{X_2|X_1}(x_2|x_1) f_{X_1|X_2}(x_1^*|x_2) dx_2$$

only depending on $X_1^{(t)}$.

Two-stage Gibbs sampler

The **stationary distribution** for the marginal process is the **marginal distribution**:

$$
\begin{aligned}
f_{X_1}(x_1^*) &= \int f_{X_1|X_2}(x_1^*|x_2) f_{X_2}(x_2) dx_2 \\
&= \int f_{X_1|X_2}(x_1^*|x_2) \int f_{X_2|X_1}(x_2|x_1) f_{X_1}(x_1) dx_1 dx_2 \\
&= \int \left(\int f_{X_1|X_2}(x_1^*|x_2) f_{X_2|X_1}(x_2|x_1) dx_2 \right) f_{X_1}(x_1) dx_1 \\
&= \int K(x_1, x_1^*) f_{X_1}(x_1) dx_1
\end{aligned}
$$

Hammersley Clifford theorem (2-stage sampler)

Hammersley Clifford (2-stage)
The conditional distributions $f_{X_1|X_2}(x_1|x_2)$ and $f_{X_2|X_1}(x_2|x_1)$ contain sufficient information to produce samples of the joint distribution $f(x_1, x_2)$ associated with these full conditionals:

$$f(x_1, x_2) = \frac{f_{X_2|X_1}(x_2|x_1)}{\int \left(\frac{f_{X_2|X_1}(x_2|x_1)}{f_{X_1|X_2}(x_1|x_2)} \right) dx_2}$$

Example: Gibbs sampler

Gibbs Sampler for normal distribution with Normal- and Gamma- conjugate priors. All have posteriors and "full conditional" distributions of the following forms.

The full conditional distributions of

$$\begin{array}{rcc} \mu|\lambda, y & \sim & N(m^*, l^*) \\ & & m^* = l^* \cdot (l \cdot m + \lambda \cdot n \cdot \bar{y}) \\ & & l^* = l + n \cdot \lambda \\ \lambda|\mu, y & \sim & Gamma(a^*, b^*) \\ & & a^* = a + \frac{n}{2} \\ & & b^* = b + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \end{array}$$

are the basis for the Gibbs sampler updating

$$\begin{array}{rcl} \mu^{(t)}|y, \lambda^{(t-1)} & \sim & N(m^* = l^* \cdot (l \cdot m + \lambda^{(t-1)} \cdot n \cdot \bar{y}), l^* = l + n \cdot \lambda^{(t-1)}) \\ \lambda^{(t)}|\mu^{(t)}, y & \sim & Gamma(a^* = a + \frac{n}{2}, b^* = b + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu^{(t)})^2) \end{array}$$

Gibbs - MH connection

Connection between the Gibbs update and a single Metropolis Hastings step is the case of I containing just one single index.

In the case of the Gibbs sampler this proposal distribution is chosen to be

$$q(y_I|x_I, x_{I^C}) = \xi(y_I|x_{I^C}) \tag{1}$$

independent of x_I.

$\Rightarrow$ the acceptance probability becomes 1 independently of x and y.

Hybrid samplers

Hybrid samplers or **Metropolis-within-Gibbs** samplers are **multi-step** samplers based on the same principle as the Gibbs sampler that draws from the full posterior can be obtained by sampling from the **full conditional distributions**.

each sampling step can be

- Gibbs,
- Metropolis-Hastings,
- reversible jump
- or some other type or MCMC sampler.

Diagnostics

- The convergence of the d-tuple obtained at iteration t, $(\theta_1^{(t)}, \ldots, \theta_p^{(t)})$ to a draw from a joint posterior distribution (*equilibrium distribution of the chain*) occurs under mild regular conditions that are generally satisfied for most statistical models (see, e.g., Geman and Geman, 1984, or Roberts and Smith, 1993)

- The method may need quite a few iterates to converge and/or initiate convergence. The period until it initiates convergence is called the **burn-in** period. All those iterates/samples should be discarded

  - the burn-in period can be seen by plotting the iterations *versus* the generated values (traceplots)

  - If all values are within a zone **without** showing strong periodicity and/or (especially) tendencies, then there is no evidence of lack of convergence.

Autocorrelation of the chains

- In addition, updating parameters one at a time can lead to high autocorrelation
  - The lag h autocorrelation function (ACF) for parameter θ_j is $\rho_j(h) = \text{Cor}(\theta_j^{(t)}, \theta_j^{(t-h)})$, i.e., it is the correlation between the given parameter value in the Markov chain separated by h iterations. The term $h > 1$ is usually referred to as lag.
  - An independent chain will have approximately zero autocorrelation at each lag.
  - Note that the autocorrelation function is always equal to 1 for the value $h = 0$, since $\text{Cor}(\theta_j^{(t)}, \theta_j^{(t)}) = 1$.

```
acf(timeseries)
```

Mixing Properties

- Ideally, for efficient Markov chains, there should be a fast decrease in the value of the autocorrelation function as the lag increases.
  - This would imply that there is little relationship between values of the Markov chain within a small number of iterations.
  - Conversely, poorly mixing chains will typically have a very shallow gradient in the ACF plot, with high autocorrelation values for even relatively large values of h.

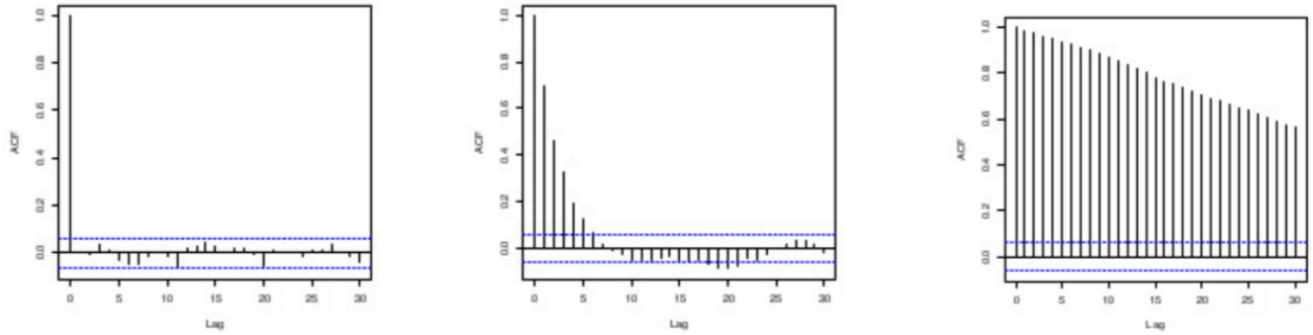


Figure 1. Sample ACF plots representing (a) ideal mixing, (b) typical good mixing, (c) poor mixing.

Thinning the chain

- A somewhat crude, yet reasonably effective, method dealing with autocorrelation is to only keep every k draws from the posterior and discard the rest; this is known as thinning the chain.

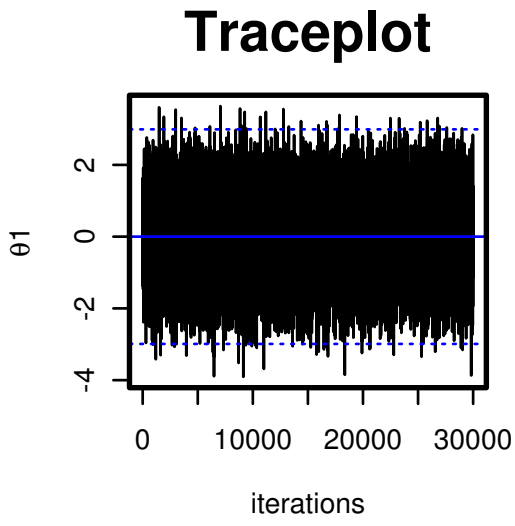
- The advantages of thinning are both simplicity and a reduction in memory usage–saving and working with large chains can be burdensome.

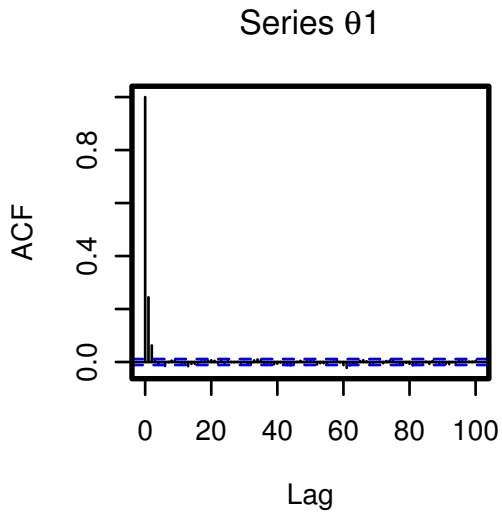
- The disadvantage is that we are clearly throwing away information; thinning can never be as efficient as using all the iterations. **Also, there is debate as to enlarging the original chain instead of thinning...**

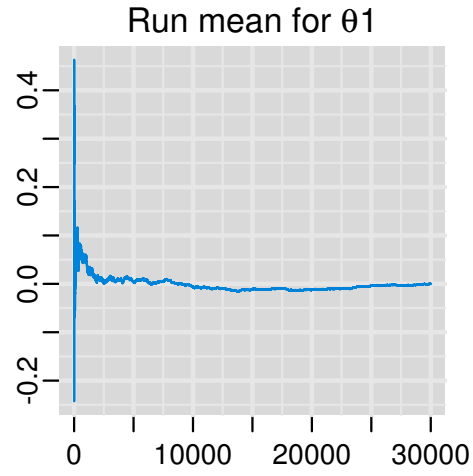
Code for Chain Diagnostics

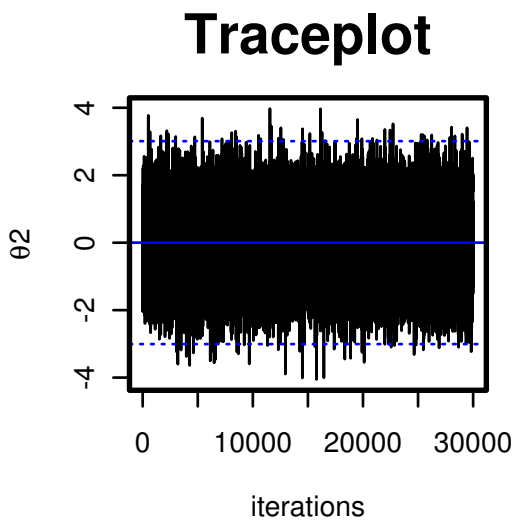
```
# trace plots
  par(mfrow=c(2,2),mar=c(5,5,4,2))
  plot(1:M,t1,type="l",ylim=c(min(t1,mean(t1)-3*sd(t1)),
      max(t1,mean(t1)+3*sd(t1))),xlab="iterations",
      ylab=expression(paste(theta,"1")),cex.main=2,main="Traceplot")
  abline(h=mean(t1),lty=1,col=4);
  abline(h=mean(t1)+3*sd(t1),lty=3,col=4)
  abline(h=mean(t1)-3*sd(t1),lty=3,col=4);  box(lwd=2)
 # ACF plots with run mean
  acf(t1,lag.max=100,main=expression(paste("Series ",theta,"1")),cex.main=2);  box(lwd=2)
  plot(1:M,t2,type="l",ylim=c(min(t2,mean(t2[-(1:200)])-3*sd(t2[-(1:200)])),
      max(t2,mean(t2[-(1:200)])+3*sd(t2[-(1:200)]))),xlab="iterations",
      ylab=expression(paste(theta,"2")),cex.main=2,main="Traceplot")
  abline(h=mean(t2[-(1:200)]),lty=1,col=4)
  abline(h=mean(t2[-(1:200)])+3*sd(t2[-(1:200)]),lty=3,col=4)
  abline(h=mean(t2[-(1:200)])-3*sd(t2[-(1:200)]),lty=3,col=4);  box(lwd=2)
  acf(t2,lag.max=100,main=expression(paste("Series ",theta,"2")),cex.main=2); box(lwd=2)
 # histrograms of marginal posteriors
  rmeanplot(t1,lwd=2,main=expression(paste("Run mean for",theta,"1")),mar=c(2,2,1.5,1)+ 0.1)
  rmeanplot(t2,lwd=2,main=expression(paste("Run mean for",theta,"2")),mar=c(2,2,1.5,1)+ 0.1)
  par(mfrow=c(1,2),mar=c(5,5,4,2))
  hist(t1,freq=F,col="grey",main="Histogram",xlab=expression(paste(theta,"1")))
  curve(dnorm(x,0,1),from=-4,4,add=T,lwd=2,col=2)
  hist(t2,freq=F,col="grey",main="Histogram",xlab=expression(paste(theta,"2")))
  curve(dnorm(x,0,1),from=-4,4,add=T,lwd=2,col=2)
```

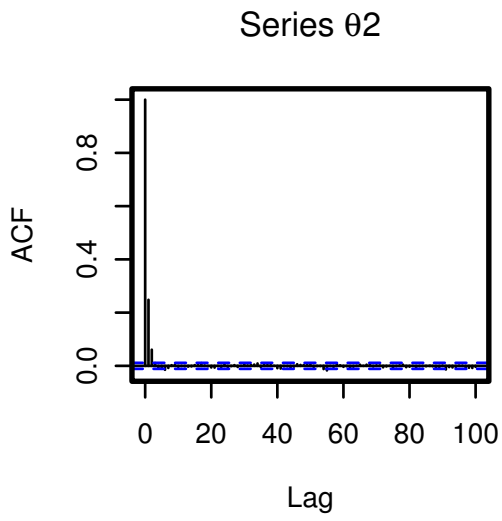
Interpreting Diagnostics

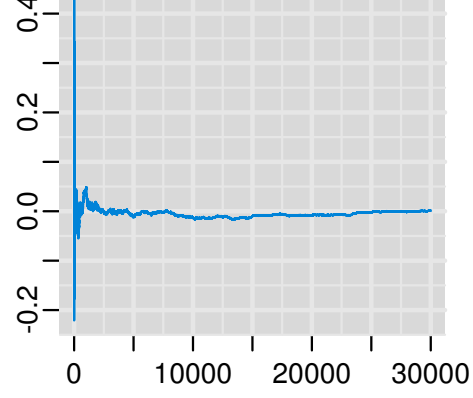






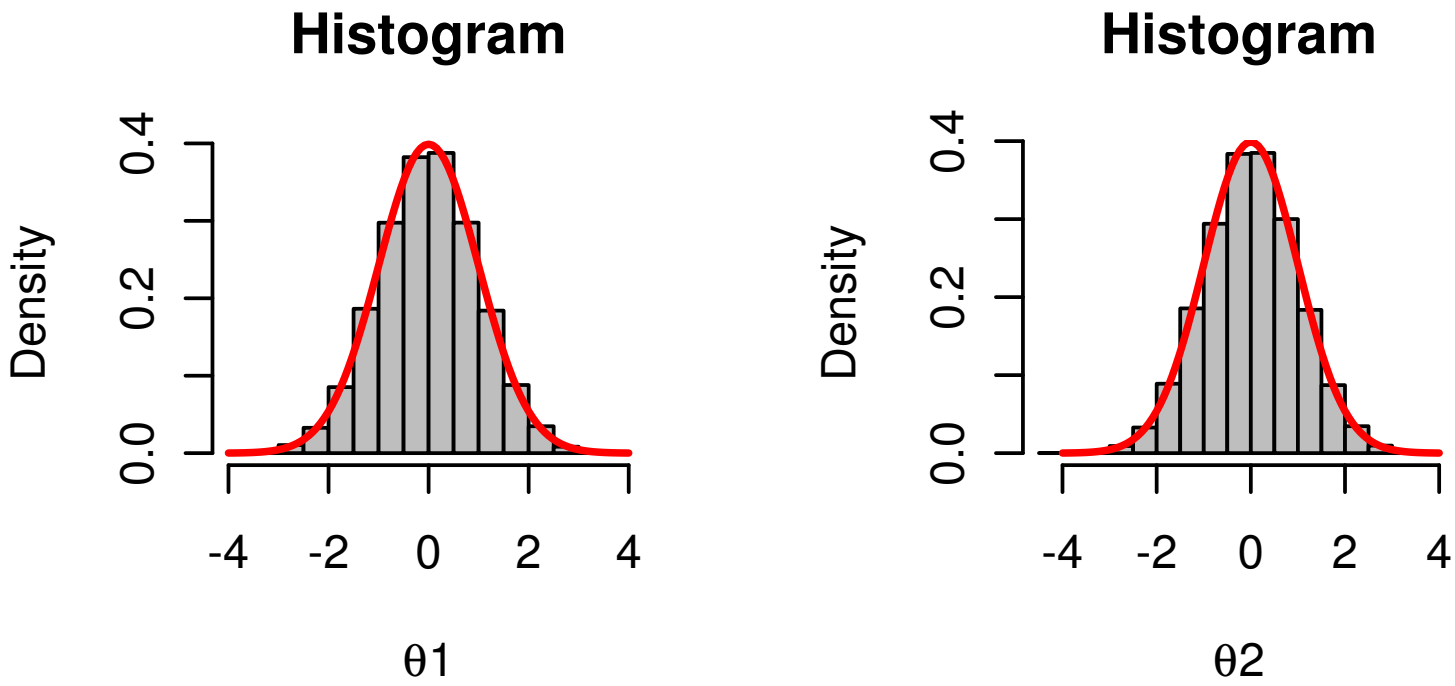




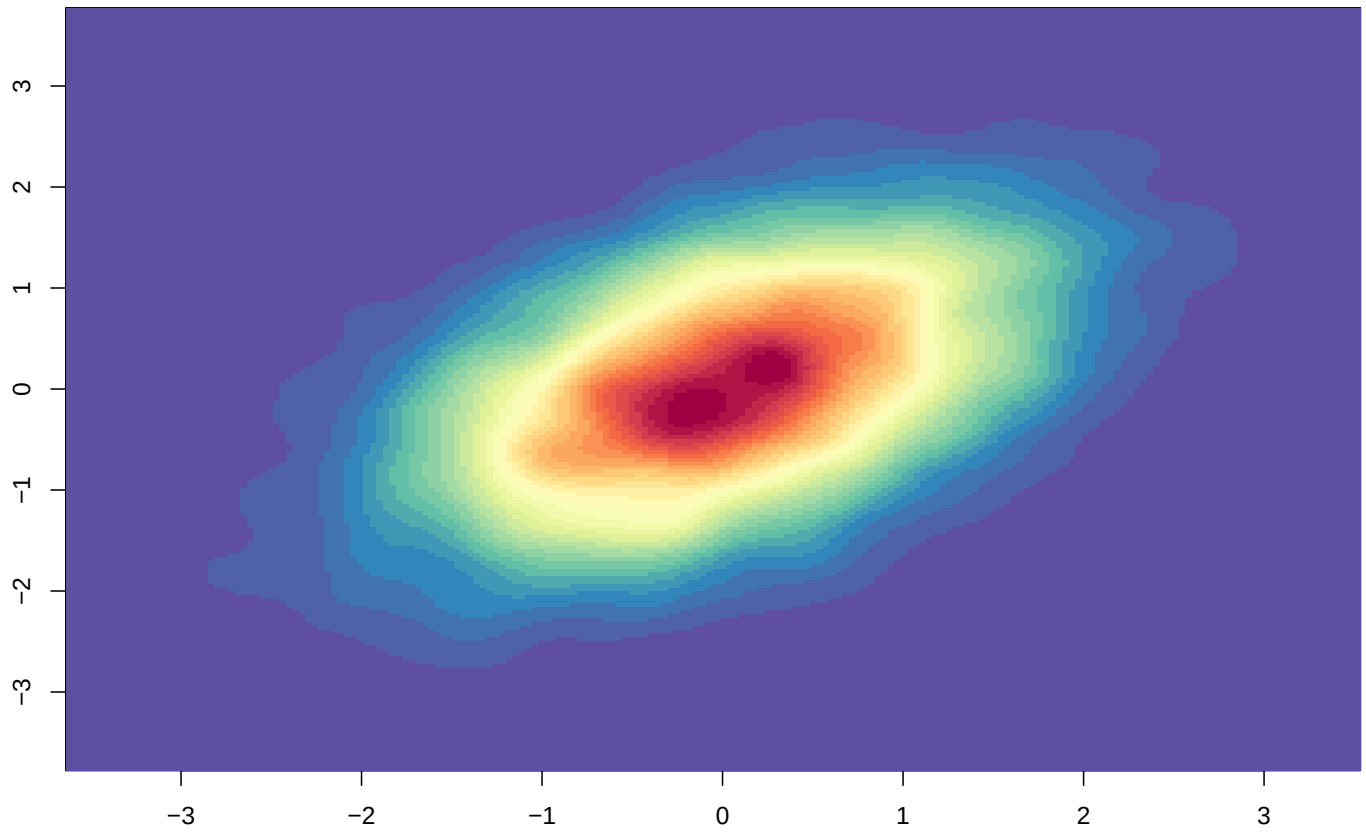


Visualising the marignal posteriors

The histograms visualise the 1-dimensional marginal posteriors. As we have seen that auto-correlation drops to 0 immediately and mixing is thus very good, we can treat these values as iid samples, even though they were generated from a Markov (=auto-correlated) process.



Visualsising the 2-dimensional posterior



Exercise

Consider the standard normal bivariate distribution with parameters $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ with p.d.f. given by

$$p(\theta_1, \theta_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\left\{ -\frac{1}{2(1-\rho^2)}(\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2) \right\}}, \text{ with } \theta_1, \theta_2 \in \mathbb{R}, \rho \in [0, 1]$$

where we also know that the marginal distributions are given by

$$p(\theta_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta_i^2}, i = 1, 2.$$

In what follows, consider $\rho = 0.5$, chain size $M = 30000$. Choose a non-informative prior-setting to initialise the algorithm.

- Implement a Gibbs sampler to sample from this distribution.
- Use the Metropolis–Hastings algorithm with block-wise update to simulate from this distribution.
- Perform chain diagnostics on the resulting chain.