Introduction to Bayesian statistics

Constantino Hevia
UTDT

July 4, 2024

Classical statistics

- **Probability:** limit of the relative frequency of an event.

- **Relative frequency:** Perform a random experiment n times. Let x be the number of times that the event A occurs. The ratio x/n is the **relative frequency** of the event A in the n experiments.

- **Limit of the relative frequency of an event:** Perform the experiment many times ($n \to \infty$) and let the probability of A be

$$\Pr(A) = \lim_{n \to \infty} \frac{x}{n}$$

Classical statistics

- There is a **true** vector or unknown parameters $\boldsymbol{\theta_0} \in \Theta \subset R^k$ that parameterizes the probability distributions of events A.

- Let the events A be different realizations of possible data sets $\boldsymbol{x}$.

- The data set $\boldsymbol{x}$ is the realization of a random vector $\boldsymbol{X}$.

- Observed data $\boldsymbol{x}$ are interpreted as a particular draw from the probability function $p(\boldsymbol{x}|\boldsymbol{\theta_0})$, which depends on the true parameter vector $\boldsymbol{\theta_0}$.

- Viewed as a function of $\boldsymbol{\theta}$, $p(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function.

Classical statistics

- **Estimation:** Choose the vector of parameters that maximizes the likelihood function (i.e. maximize the probability of observing $\boldsymbol{x}$ across possible values of $\boldsymbol{\theta}$.)

- Uncertainty regarding the parameter estimate $\hat{\boldsymbol{\theta}}$ arises from the fact that the observed data represents one of many possible draws that could have been obtained from the probability function.
    - Randomness in the data drives randomness in the parameter estimates

Bayesian statistics

- **Probability:** degree of beliefs of a researcher in an event.
- Parameter $\boldsymbol{\theta}$ is a random variable. Data $\boldsymbol{x}$ is fixed and nonrandom.
- **Objective of Bayesian analysis:** make probabilistic statements regarding the parameter $\boldsymbol{\theta}$ conditional on the data $\boldsymbol{x}$.
- Probabilistic interpretation of the parameter allows us to incorporate extraneous (a priori) views regarding the parameters. This is done through the specification of a prior distribution over the possible parameters $\boldsymbol{\theta} \in \Theta$.

Bayesian statistics

- Bayes' rule and the likelihood function yield a posterior distribution of the parameters conditional on the observed data.

- Means or modes of the posterior distribution provide point estimates of the parameters.

- Uncertainty of the estimate conveyed using posterior standard deviations or other measures using the posterior distribution

- Credibility sets

Kernel of a distribution

- Let $p(\boldsymbol{x})$ be a probability density function of the random vector $\boldsymbol{X}$.
- The **kernel** of $p(\boldsymbol{x})$ is a function $k(\boldsymbol{x})$ that is **proportional** to $p(\boldsymbol{x})$, in which factors that are not function of the random variables $\boldsymbol{x}$ are omitted.
- Omitted factors may be functions of the parameters that describe the density function $p(\boldsymbol{x})$ but should never include the $\boldsymbol{x}$ variables.
- **Example:** Consider the density function of a normal random variable

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma^2}}.$$

  The associated kernel is

$$k(x) = e^{-\frac{(x-\mu)}{2\sigma^2}} \propto p(x|\mu, \sigma^2).$$

- Kernels do not integrate to 1.

Prior and posterior distributions

- Let $\boldsymbol{\theta}$ be an unknown parameter vector of interest.
- Prior distribution: we assume that $\boldsymbol{\theta}$ is random with some distribution $\pi(\boldsymbol{\theta})$.
  - Captures our previous uncertainty or views regarding the parameters.
- Likelihood function: There is a random vector $\boldsymbol{X}$ with PDF $p(\boldsymbol{x}|\boldsymbol{\theta})$.
- The joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{X}$ is $p(\boldsymbol{\theta}, \boldsymbol{x}) = \pi(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})$.
- Integrating over $\boldsymbol{\theta}$ gives the marginal density of $\boldsymbol{X}$

$$p(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Prior and posterior distributions: Bayes' Law

- By definition of conditional probability,

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

- Bayes's law

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\int_{\theta} p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{1}$$

- $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior density of $\boldsymbol{\theta}$ conditional on having observed the data $\boldsymbol{x}$.

Prior and posterior distributions

- In equation (1), $p(\boldsymbol{x})$ is a normalization constant so that $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ integrates to 1.
  - Often, it is difficult to compute $p(\boldsymbol{x})$.
- The relevant information of the posterior distribution is contained in $p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.
- The **kernel** of the posterior distribution is the product of the likelihood function and the prior distribution:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

- The posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is the object of interest. It tells us how our prior information is updated after observing the data $\boldsymbol{x}$.

Bayesian analysis

- The mode of the posterior is the maximum a posteriori (MAP) estimator:

$$\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\boldsymbol{x}).$$

- Posterior moments may be of interest: $E\left[h(\boldsymbol{\theta})|\boldsymbol{X}=\boldsymbol{x}\right] = \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}$.

- Need to draw samples from $\pi(\boldsymbol{\theta}|\boldsymbol{x})$:
  - Conceptually, very simple.
  - Operationally, it can be difficult. In most cases we only know a kernel of $\pi(\boldsymbol{\theta}|\boldsymbol{x})$.
  - Much of Bayesian statistics is about how to draw samples from the kernel of $\pi(\boldsymbol{\theta}|\boldsymbol{x})$.

Bayesian analysis

- **Objective**: Sample from $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ knowing only a kernel:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- Sometimes we recognize the form of the posterior by inspecting $p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.
  - **Conjugate priors**: given a prior $\pi(\boldsymbol{\theta})$ and the likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$, the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is in the same family of distributions as the prior $\pi(\boldsymbol{\theta})$. Used in Bayesian VARs.

- Often, $p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is a kernel of an unknown distribution.

- In such cases, we use approximate inference techniques, such as Markov Chain Monte Carlo (MCMC) methods.

Sampling from a distribution

Sampling from an arbitrary distribution $f(\boldsymbol{x})$

- Let $f(\boldsymbol{x})$ be a target density function with cumulative distribution function $F(\boldsymbol{x})$.

- **Direct i.i.d. sampling** if $f(\boldsymbol{x})$ is easy to sample from: Normal, Gamma, Beta, etc.

  - **The Inverse Transform:** If X has a cumulative distribution function $F(x)$, then the random variable $U = F(X)$ has a uniform distribution $U \sim \mathbb{U}[0, 1]$.

  - **Implication:** To generate a random draw from $X \sim F$, we can generate a uniform draw $U \sim \mathbb{U}[0, 1]$ and then make the transformation $x = F^{-1}(u)$.

- What if we know the function $f(\boldsymbol{x})$ but don't know how to sample from it?

  - For example, if cannot easily compute the inverse CDF $F^{-1}(u)$.

- What if we only know a kernel $k(\boldsymbol{x})$ of $f(\boldsymbol{x})$?

Sampling from $f(x)$ if direct sampling not possible

- Want to sample from $f(x) = k(x)/C$, where $C = \int k(x)dx$ is the normalizing constant.
- Monte Carlo methods (Ulam, Von Neumann)
  - Acceptance-rejection sampling
  - Sampling importance resampling (SIR)
- Markov Chain Monte Carlo methods
  - Metropolis-Hastings
  - Gibbs Sampler

Acceptance-Rejection sampling

- There is an auxiliary density $g(x)$ that is easy to sample from.
- Need one condition:
- **Assumption:** There is a number $M < \infty$ such that $Mg(x) \geq k(x)$ for all x.
  1. Therefore, the domain of $g(x)$ must be the same as the domain of $k(x)$
  2. Also, and $Mg(x)$ is always **above** $k(x)$, so that

$$0 \leq \frac{k(x)}{Mg(x)} \leq 1.$$

- $Mg(x)$ is called an **envelope** of $k(x)$
- **Important:** the domain of $g(x)$ includes the domain of $f(x)$.

Acceptance-Rejection sampling

- **Algorithm:**
  1. Choose a proposal density $g(x)$ and an M so that $Mg(x) \geq k(x)$ for all x.
  2. Draw x from $g(x)$.
  3. Accept the draw x with probability $\frac{k(x)}{Mg(x)}$:
     - Draw u from $U(0,1)$
     - Accept x if $u \leq \frac{k(x)}{Mg(x)}$
  4. Repeat steps 2 and 3 until n draws are accepted.

Acceptance-Rejection sampling: why does it work?

- Probability of $X = x$ given that we have accepted the draw (A):

$$\Pr(x|A) = \frac{\Pr(A|x)g(x)}{\Pr(A)} = \frac{\frac{k(x)}{Mg(x)}g(x)}{\Pr(A)} = \frac{\frac{k(x)}{M}}{\Pr(A)}.$$

  The unconditional probability of accepting a draw is:

$$\Pr(A) = \int_x g(x)\frac{k(x)}{Mg(x)}dx = \frac{\int_x k(x)dx}{M} = \frac{C}{M}.$$

  Therefore,

$$\Pr(x|A) = \frac{\frac{k(x)}{M}}{\frac{C}{M}} = \frac{k(x)}{C} = f(x).$$

- $\Pr(x|A)$ is exactly the target distribution $f(x)$!

Efficiency of Acceptance-Rejection algorithm

- Probability of acceptance: $\Pr(A) = \frac{C}{M}$.
- En estimate of $\Pr(A)$ is

$$\widehat{\Pr}(A) = \frac{\text{Number of accepted draws}}{\text{Total number of draws}}$$

- Therefore, the normalizing constant can be estimated as

$$\widehat{C} = M\widehat{\Pr}(A)$$

Efficiency of Acceptance-Rejection algorithm

- Let I be the number of draws until first acceptance.
  - I has a geometric distribution.
- Expected number of draws until first acceptance is $E[I] = \frac{M}{C}$. Therefore:
  1. Would like to take M as small as possible
  2. If $g(x) = f(x)$, choose $M = C$ and $E[I] = 1$.
  3. $E[I]$ can be large if $g(x)$ is very different from $f(x)$.
  4. Would like proposal distributions that are close to $f(x)$.
- Can work well in low dimensions but can be very inefficient in high dimensions.

Acceptance-Rejection: example

- Suppose we want to sample from a mixture of normals:

$$p(x) \sim \eta N(\mu_1, \sigma_1) + (1 - \eta) N(\mu_2, \sigma_2)$$

- The kernel of the target density is

$$k(x) = \eta e^{\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + (1 - \eta) e^{\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}.$$

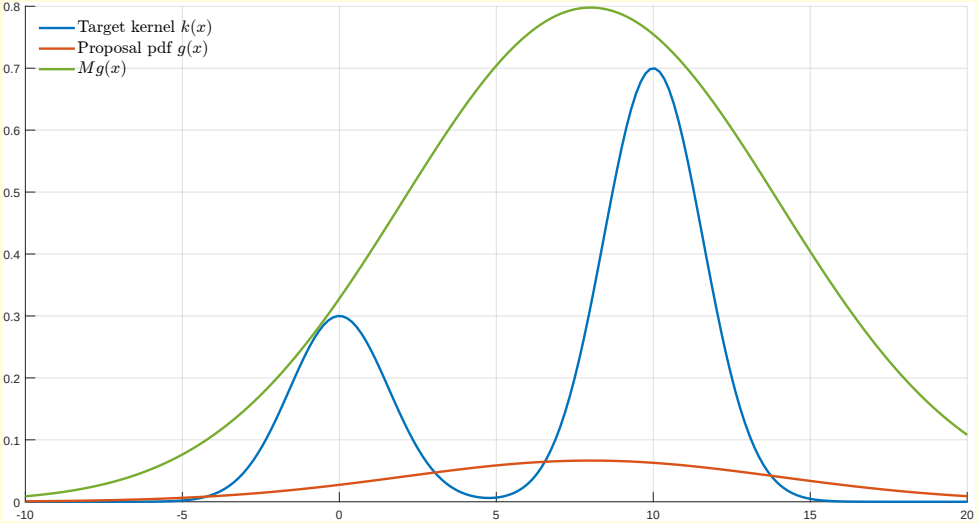
  Assume $\eta = 0.3$, $\mu_1 = 0$, $\sigma_1^2 = 2.5$, $\mu_2 = 10$, $\sigma_2^2 = 2.5$.

- The proposal density is

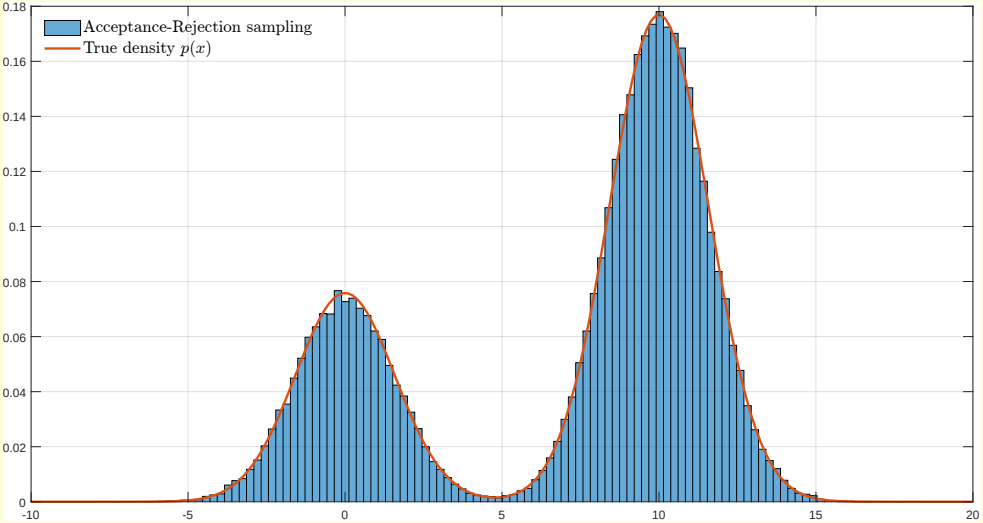
$$g(x) \sim N(\mu_3, \sigma_3)$$

  and set $\mu_3 = 8$ and $\sigma_3 = 6$.

Acceptance-Rejection: kernel, proposal, and envelope



Acceptance-Rejection sampling



Sampling-Importance Resampling (SIR)

- Weighted bootstrap procedure that does not require the finite bound M to exist.
- $f(x) = k(x)/C$ is the target distribution but we can only evaluate the kernel $k(x)$.
- $g(x)$: proposal density, easy to sample from. Domain of $g(x)$ includes domain of $f(x)$.
- **Algorithm:**
  1. Draw n samples from $g(x)$: $\{x_1, x_2, \ldots, x_n\}$.
  2. Compute importance weights $\omega_i = k(x_i)/g(x_i)$ for $i = 1, 2, \ldots, n$ and then:

$$q_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}.$$

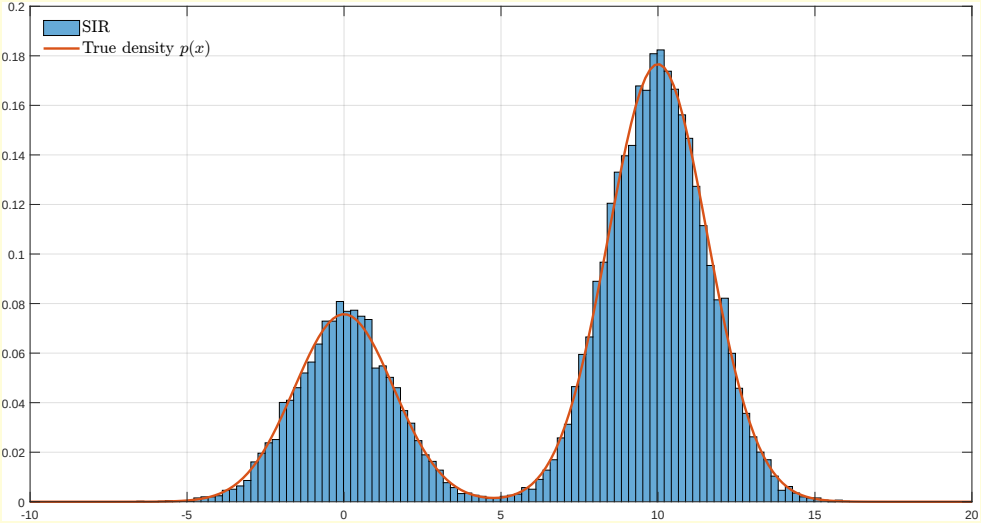
  3. Draw x^* from the discrete distribution $\{x_1, x_2, \ldots, x_n\}$ with mass q_i on x_i.
  4. x^* is approximately distributed according to $f(x)$, with the approximation improving as n increases.

Sampling-Importance Resampling: *CDF* of x^*

$$
\begin{aligned}
\Pr(x^* \le a) &= \sum_{i=1}^{n} q_i \mathbb{1}(x_i \le a) = \frac{\frac{1}{n}\sum_{i=1}^{n} \omega_i \mathbb{1}(x_i \le a)}{\frac{1}{n}\sum_{i=1}^{n} \omega_i} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{k(x_i)}{g(x_i)} \mathbb{1}(x_i \le a)}{\frac{1}{n}\sum_{i=1}^{n} \frac{k(x_i)}{g(x_i)}} \xrightarrow{n\to\infty} \frac{E_g\left[\frac{k(x)}{g(x)} \mathbb{1}(x \le a)\right]}{E_g\left[\frac{k(x)}{g(x)}\right]} \\
&= \frac{\int_{-\infty}^{\infty} \frac{k(x)}{g(x)} \mathbb{1}(x \le a) g(x) dx}{\int_{-\infty}^{\infty} \frac{k(x)}{g(x)} g(x) dx} = \frac{\int_{-\infty}^{\infty} k(x) \mathbb{1}(x \le a) dx}{\int_{-\infty}^{\infty} k(x) dx} \\
&= \frac{\int_{-\infty}^{a} k(x) dx}{C} = \int_{-\infty}^{a} f(x) dx \\
&= F(a)
\end{aligned}
$$

As $n \to \infty$, drawing from the discrete distribution is equivalent to drawing from $f(x)$.

SIR example


SIR
True density $p(x)$
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-10
-5
0
5
10
15
20

Markov Chain Monte Carlo (MCMC)

- Sample from $f(x) = k(x)/C$, where $C = \int k(x)dx$ is the normalizing constant.
- Create a Markov chain whose invariant distribution is equal to $f(x)$.
  - Simulate the Markov chain until stationarity is achieved.
  - A draw from the stationary distribution of the Markov chain is a draw from $f(x)$.
- Very general procedure that can be used in most settings.

Markov Chain Monte Carlo (MCMC)

- A stochastic process $\{X_1, X_2, \ldots, X_t\}$ on a discrete space Ω is a first order **Markov Chain** if

$$\Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_1 = x_1) = \Pr(X_t = x_t | X_{t-1} = x_{t-1}).$$

- A Markov chain can be defined by its transition Kernel.
  - If Ω is **discrete**, the transition kernel is a transition matrix with elements

$$P(y|x) = \Pr(X_t = y | X_{t-1} = x), \quad x, y \in \Omega.$$

  - If the space Ω is **continuous**, the kernel is a conditional density $Q(x, x')$ such that

$$\Pr(X \in A | x) = \int_A Q(x, y) dy.$$

  x is the current state and $X \in A$ is the set of values that the state can take tomorrow.

Markov Chain Monte Carlo (MCMC)

- A stationary distribution of a Markov chain is a distribution $\pi(x)$ on Ω such that

$$\pi(y) = \sum_{x \in \Omega} P(y|x)\pi(x).$$

  If the space Ω is continuous, the stationary distribution satisfies

$$\pi(y) = \int_{\Omega} Q(x, y)\pi(x)dx.$$

- Assume that the Markov chain is satisfies conditions such that the stationary distribution exists and is unique.
  - Nerd alert!: a sufficient condition is that the Markov chain is irreducible and aperiodic.

Markov Chain Monte Carlo (MCMC)

Definition: A Markov chain is said to be **time reversible** if there exists a probability measure π on Ω such that

$$P(x|y)\pi(y) = P(y|x)\pi(x) \tag{2}$$

- Note that if $\pi(\cdot)$ satisfies (2), then it is the invariant distribution of the Markov chain:

$$\sum_x P(y|x)\pi(x) \overset{eq.(2)}{=} \sum_x P(x|y)\pi(y) = \pi(y).$$

- Condition (2) implies that, in the long run, the chain moves from x to y at the same rate as it moves from y to x.

- Time reversibility is the key element the MCMC algorithms.

- There are equivalent results for the continuous state case.

Metropolis-Hastings algorithm

- Want to sample from $f(x) = k(x)/C$, with $C = \int k(x)dx$.

- We will construct a reversible Markov chain as follows.
  1. Choose M, set $j = 1$, and initialize the algorithm with a value x_1.
  2. Draw y from some Markov transition kernel $Q(y|x_j)$.
  3. Set $x_{j+1} = y$ with probability $\alpha(y|x_j)$, where

$$\alpha(y|x_j) = \min\left\{\frac{k(y)}{k(x_j)}\frac{Q(x_j|y)}{Q(y|x_j)}, 1\right\}$$

     Otherwise set $x_{j+1} = x_j$. In particular:
     - Draw u from $U(0,1)$. If $u \leq \alpha(y|x_j)$ then $x_{j+1} = y$. Otherwise, $x_{j+1} = x_j$.
  4. If $j \leq M$, set $j \rightsquigarrow j+1$ and go to 2.

Claim: The resulting Markov chain is reversible with stationary distribution $f(x)$.

Metropolis-Hastings algorithm

- Only need to evaluate $k(x)$. But note that $\frac{k(y)}{k(x_j)} = \frac{f(y)}{f(x_j)}$. The constant C is not needed in the algorithm.

- If the candidate draw y is rejected, the current state x_j becomes the next value in the sequence. Note difference with acceptance-rejection sampling.

- The transition kernel of the Markov chain is

$$P(y|x) = \begin{cases} \alpha(y|x)Q(y|x) & y \neq x \\ 1 - \sum_{u \neq x} \alpha(u|x)Q(u|x) & y = x \end{cases}$$

- Metropolis-Hasting algorithm is defined by the transition Kernel $Q(y|x)$.
  - Alternatives?

Metropolis-Hastings algorithm

Proof of Claim: We check that $p(x)$ satisfies the reversibility condition (2). The probability of transitioning from x to y according to the algorithm is $P(y|x) = \alpha(y|x)Q(y|x)$. Then:

$$\begin{aligned}
P(y|x)f(x) =& \alpha(y|x)Q(y|x)f(x) \\
=& \min\left\{\frac{f(y)}{f(x)}\frac{Q(x|y)}{Q(y|x)}, 1\right\} Q(y|x)f(x) \\
=& \min\left\{Q(x|y)f(y), Q(y|x)f(x)\right\} \\
=& \min\left\{1, \frac{f(x)}{f(y)}\frac{Q(y|x)}{Q(x|y)}f(x)\right\} Q(x|y)f(y) \\
=& \alpha(x|y)Q(x|y)f(y) \\
=& P(x|y)f(y)
\end{aligned}$$

Therefore, $f(x)$ is the stationary distribution of the Metropolis-Hastings Markov chain.

Metropolis-Hastings: choice of $Q(y|x)$

- Random Walk Metropolis-Hastings (RWMC)
- Popular choice for $Q(y|x)$ is a random walk:

$$y = x + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \Sigma). \tag{3}$$

- The random walk satisfies the properties of a good transition kernel.
- How do we choose Σ?
  - A good choice is the Hessian of the distribution of interest. But sometimes it is not available or too difficult to compute.

Metropolis-Hastings: choice of $Q(y|x)$

- Independent MH: another popular choice.
- Just make $Q(y|x) = Q(y)$ independent of x.
- Similarity with acceptance sampling.
- Acceptance probability is now

$$\alpha(y|x) = \min\left\{\frac{k(y)}{k(x)}\frac{Q(x)}{Q(y)}, 1\right\}$$

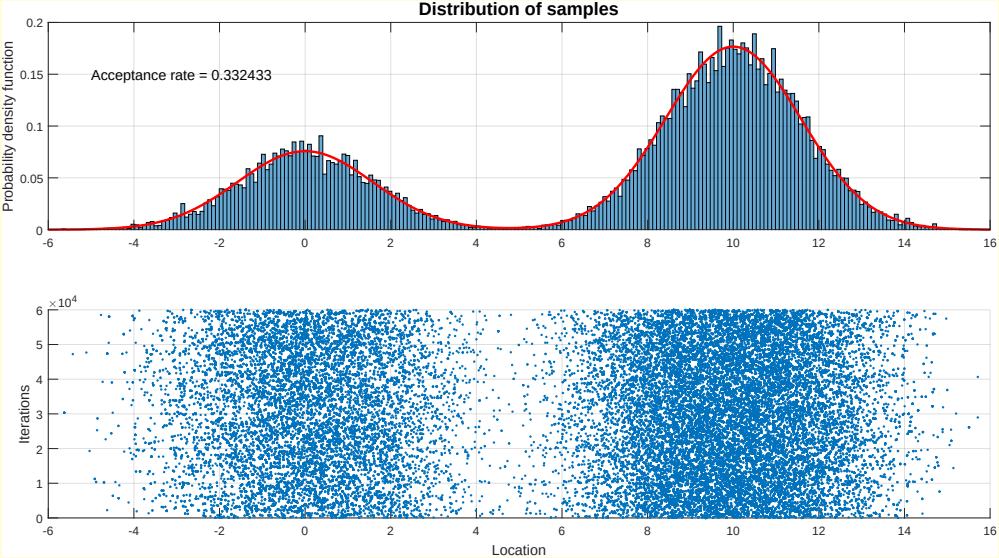
Acceptance rate

- When the candidate draw is rejected, we keep the current value in the chain
- Choosing an appropriate acceptance rate is important for the performance of the algorithm.
- Roberts, Gelman and Gilks, (1997): Acceptance rate should be about:
  - 45% for 1 dimensional problems
  - 26% for 6 dimensions
  - 23% in the limit
- Most users of MH choose an acceptance rate between 0.2 and 0.4.

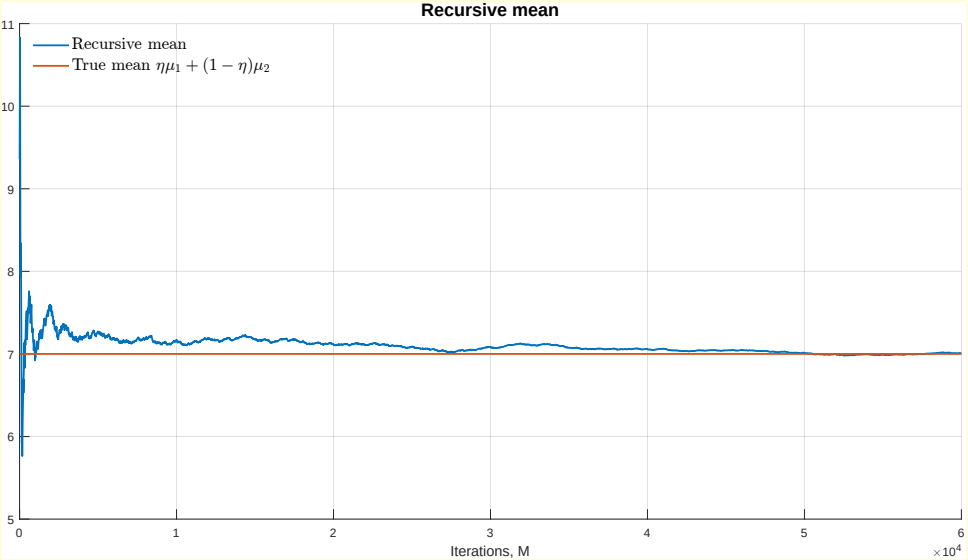
Convergence diagnostics

- To use MCMC we must:
  1. Ensure that the Markov Chain has converged to its stationary distribution
  2. Only use samples generated after stationarity has been reached.
- Methods to assess convergence:
  1. Plot variables of interest as a function of iteration number.
  2. Plot histograms in different blocks of the chain.
  3. Plot recursive means, rolling means, etc.
  4. Chapter 12 of Robert and Casella, “Monte Carlo Statistical Methods”, (2004) contains several procedures.

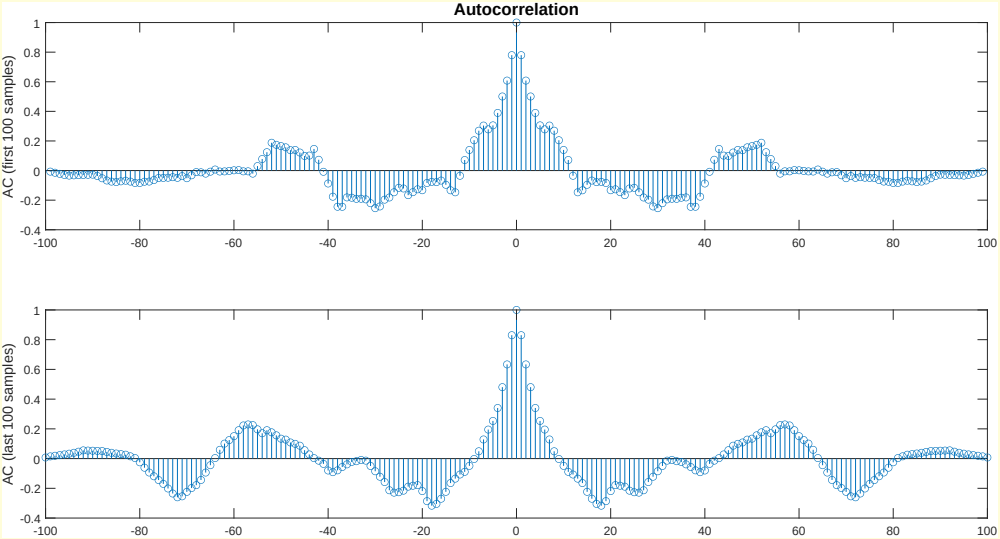
RWMH sampling



RWMH example: recursive average



RWMH example: autocorrelations



Gibbs Sampler

- Let $p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_k)$, where $x \in \mathbb{R}^k$, denote the target density.
- In some cases, we cannot draw from $p(\boldsymbol{x})$ but we can draw from conditional distributions $p(\boldsymbol{y}|\boldsymbol{z})$ and $p(\boldsymbol{z}|\boldsymbol{y})$ for some partition $\boldsymbol{x} = [\boldsymbol{y}, \boldsymbol{z}]$.
  - The partition could be in more than two blocks: $\boldsymbol{x} = \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}\right]$, $n \leq k$.
- In this case, we can construct a Markov chain whose invariant distribution is $p(\boldsymbol{x})$.
- What partition to use? That will depend on the particular case.
- The Gibbs sampler is useful to perform Bayesian analysis of the linear regression model and for Bayesian VARs.

Gibbs Sampler

- **Algorithm:** Choose N and arbitrary starting values $[\boldsymbol{y}_0, \boldsymbol{z}_0]$. Set $j = 1$. Then do the following:

  1. Draw $\boldsymbol{y}_j$ from $p(\boldsymbol{y}|\boldsymbol{z}_{j-1})$.
  2. Draw $\boldsymbol{z}_j$ from $p(\boldsymbol{z}|\boldsymbol{y}_j)$.
  3. Store $[\boldsymbol{y}_j, \boldsymbol{z}_j]$, set $j \rightsquigarrow j+1$ and go to step 1 while $j < N$.

- This algorithm defines a transition mechanism $[\boldsymbol{y}_{j-1}, \boldsymbol{z}_{j-1}] \to [\boldsymbol{y}_j, \boldsymbol{z}_j]$ which is the realization of a Markov chain.

- Of course, can change the order in which we draw the different blocks.

Gibbs Sampler

- Not only is $\{\boldsymbol{y}_j, \boldsymbol{z}_j\}$ a Markov chain, but also each subsequence $\{\boldsymbol{y}_j\}$ and $\{\boldsymbol{z}_j\}$.
- For example, the chain $\{\boldsymbol{y}_j\}$ has transition density

$$P(\boldsymbol{y}_j|\boldsymbol{y}_{j-1}) = \int p(\boldsymbol{y}_j|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{y}_{j-1})d\boldsymbol{z}$$

  which depends only on $\boldsymbol{y}_{j-1}$.
- **Theorem:** the invariant distribution of the Gibbs sampler is $p(\boldsymbol{x})$.
- **Claim:** The Gibbs-Sampler is a particular Metropolis-Hastings algorithm.

Bayesian linear regression

Bayesian linear regression

- Consider the linear regression model

$$y_i = x_i'\beta + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2) \text{ and } i = 1, 2, \ldots, n,$$

where y_i is the dependent variable and $x_i, \beta \in R^k$ (there are k regressors).

- Stacking the observations we get

$$Y = X\beta + \epsilon; \qquad \epsilon \sim N(0, \sigma^2 I_n),$$

where Y is $n \times 1$, X is $n \times k$, β is $k \times 1$, and ϵ is $n \times 1$.

- The likelihood function is

$$p(Y|\beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} \mid \sigma^2 I_n \mid^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right] \tag{4}$$

(It also depends on X but let's keep it implicit.)

Bayesian linear regression

- Note: The FOC with respect to β implies that MLE is equal to OLS:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{5}$$

- The unknowns are β and σ.
- Usually, we use the priors $\beta \sim N(\beta_0, \Sigma_0)$ and $\sigma^2 \sim IGa(a, b)$ (i.e. the prior of the variance is an inverse Gamma distribution with parameters a, b).
- We want to find the posterior distribution $p(\beta, \sigma^2|Y)$.
- But easier to find $p(\beta|\sigma^2, Y)$ and $p(\sigma^2|\beta, Y)$.

Bayesian linear regression

- $p(\beta|\sigma^2, Y)$ is a Normal distribution and $p(\sigma^2|\beta, Y)$ is an IGa distribution.
- How can use the conditional densities $p(\beta|\sigma^2, Y)$ and $p(\sigma^2|\beta, Y)$ to sample from $p(\beta, \sigma^2|Y)$?
- It is convenient to define $\tau \equiv \sigma^{-2}$ as the *precision* of ϵ_i.
- The prior for τ is a Gamma distribution, $\tau \sim Ga(a, b)$.
- In this case, the likelihood function is

$$p(Y|\beta, \tau) = (2\pi)^{-\frac{n}{2}} \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}(Y - X\beta)'(Y - X\beta)\right] \tag{6}$$

Bayesian linear regression with known τ

- Add prior beliefs about β in the form of a distribution $p(\beta)$.

$$p(\beta) \sim N(\beta_0, \Sigma_0).$$

  where β_0 is $k \times 1$ and Σ_0 is $k \times k$.

- Here we take $\tau = \sigma^{-2}$ as a known parameter.

- Combining the prior belief with the likelihood function (4) and using Bayes's theorem

$$p(\beta|\tau, Y) = \frac{p(Y|\beta, \tau)p(\beta)}{p(Y|\tau)} \implies p(\beta|\tau, Y) \propto p(Y|\beta, \tau)p(\beta)$$

Bayesian linear regression with known τ

- Prior distribution $p(\beta)$:

$$p(\beta) = (2\pi)^{-\frac{k}{2}} \mid \Sigma_0 \mid^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0)\right]$$
$$\propto \exp\left[-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0)\right]$$

- Kernel of the likelihood function:

$$p(Y|\beta, \tau) \propto \exp\left[-\frac{\tau}{2}(Y - X\beta)'(Y - X\beta)\right]$$

- Combine the expressions and compute the posterior kernel:

$$p(\beta|\tau, Y) \propto \exp\left[-\frac{\tau}{2}(Y - X\beta)'(Y - X\beta)\right] \exp\left[-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0)\right]$$

Bayesian linear regression with known τ

- After some tedious algebra we get

$$p(\beta|\tau, Y) \propto \exp\left[-\frac{1}{2}(\beta - \beta_1)' \Sigma_1^{-1} (\beta - \beta_1)\right]$$

where

$$\begin{aligned}\Sigma_1 &= \left(\Sigma_0^{-1} + \tau X'X\right)^{-1} \\ \beta_1 &= \Sigma_1 \left(\Sigma_0^{-1}\beta_0 + \tau X'Y\right).\end{aligned}$$

- This is the kernel of a normal distribution, so that $\beta|\tau, Y \sim N(\beta_1, \Sigma_1)$

- **Note:** for the proposed prior, the posterior has the same distribution as the likelihood. This is called “natural conjugate prior”.

Bayesian linear regression with known τ

- But we can do more. The OLS (and MLE) estimator implies

$$\hat{\beta} = (X'X)^{-1}X'Y \implies X'Y = (X'X)\hat{\beta}.$$

- Replacing this result into the definition of β_1 gives

$$\begin{aligned}
\beta_1 =& \Sigma_1 \left(\Sigma_0^{-1}\beta_0 + \tau X'X\hat{\beta}\right) \\
\beta_1 =& \left(\Sigma_0^{-1} + \tau X'X\right)^{-1} \left(\Sigma_0^{-1}\beta_0 + \tau X'X\hat{\beta}\right) \\
\beta_1 =& W\beta_0 + (I - W)\hat{\beta}.
\end{aligned}$$

- The posterior mean is a weighted average of the prior mean and the OLS estimate.
  - If the variance of the prior is infinity ($\Sigma_0^{-1} \to 0$), the posterior mean is the OLS estimate.
  - If the data is uninformative ($\tau \to 0$) the mean of the posterior is the mean of the prior.

Bayesian linear regression with known τ: uniform prior

- Sometimes we don't have a priori reasons to prefer one parameter value over others.
- In such case, we may use uniform priors, $p(\beta) \propto 1$
- The posterior is normal $p(\beta|\tau, Y) = N(\hat{\beta}, \hat{\Sigma})$ with

$$\hat{\Sigma} = (\tau X'X)^{-1}$$
$$\hat{\beta} = (X'X)^{-1}X'Y.$$

  that is, the posterior mean is the OLS estimate and the posterior covariance matrix is the covariance matrix of the OLS coefficients.

- **Exercise 1:** prove the previous claim.

Example

- Consider the simple linear model

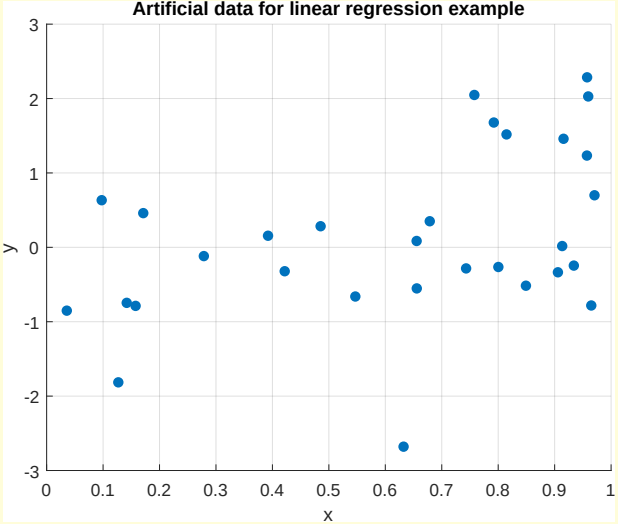
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

- Simulate data assuming $\beta_0 = -1$, $\beta_1 = 2$, $\sigma = 1$, $n = 30$, $x_i \sim \text{Uniform}(0, 1)$.

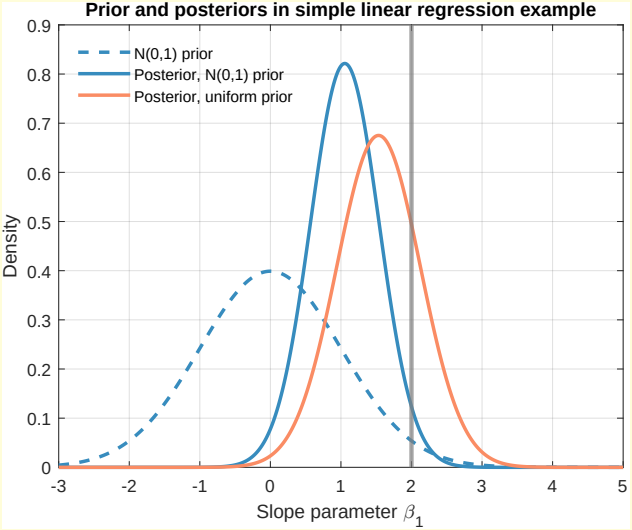
- Consider two priors:
  1. Uniform prior: $p(\beta) \propto 1$.
  2. Normal prior: $p(\beta) = N(0, I_2)$.

- Given the priors, compute the posterior for β_1 and plot.

Example


Artificial data for linear regression example
3
2
1
0
-1
-2
-3
y
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x

Example



Exercise 2: do the same plot but for the parameter β_0.

Bayesian linear regression with known β

- Now assume that the prior of the precision τ follows a Gamma density $\tau \sim Ga(a_0, b_0)$

$$p(\tau) = \frac{b_0^{a_0}}{\Gamma(a_0)} \tau^{a_0 - 1} \exp\left(-b_0 \tau\right).$$

- The mean and variance of τ are $E(\tau) = a_0 / b_0$ and $\text{Var}(\tau) = a_0 / b_0^2$.
- The hyperparameters a_0 and b_0 determine the shape of the distribution.
- Equivalently, the prior of the variance σ^2 is an inverse Gamma distribution.
- Combining the Gamma prior with the likelihood function (4) and using Bayes's theorem implies

$$p(\tau | \beta, Y) \propto p(Y | \beta, \tau) p(\tau)$$

Bayesian linear regression with known β

- Prior distribution $p(\tau)$:

$$p(\tau) \propto \tau^{a_0 - 1} \exp\left(-b_0 \tau\right)$$

- Kernel of the likelihood function:

$$p(Y|\beta, \tau) \propto \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right]$$

- Combine the expressions and compute the posterior kernel:

$$\begin{aligned} p(\tau|\beta, Y) \propto & \tau^{\frac{n}{2}} \tau^{a_0 - 1} \exp\left[-b_0 \tau\right] \exp\left[-\frac{\tau}{2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right] \\ \propto & \tau^{a_0 + \frac{n}{2} - 1} \exp\left[-\left(b_0 + \frac{1}{2}\left(Y - X\beta\right)'\left(Y - X\beta\right)\right)\tau\right] \end{aligned}$$

Bayesian linear regression with known β

- Note that

$$p(\tau|\beta, Y) \propto \tau^{a_0+\frac{n}{2}-1} \exp\left[-\tau\left(b_0 + \frac{1}{2}(Y - X\beta)'(Y - X\beta)\right)\right]$$

is the kernel of a Gamma distribution, so that

$$\tau|\beta Y \sim Ga(a_1, b_1)$$

with

$$\begin{aligned} a_1 &= a_0 + \frac{n}{2}, \\ b_1 &= b_0 + \frac{1}{2}(Y - X\beta)'(Y - X\beta). \end{aligned}$$

Example

- Consider the linear regression

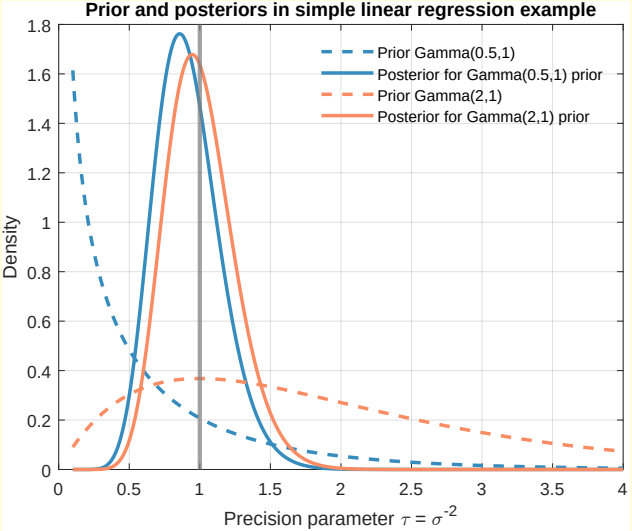
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

- Simulate data assuming $\beta_0 = -1$, $\beta_1 = 2$, $\sigma = 1$,$n = 30$, $x_i \sim \text{Uniform}(0, 1)$.

- Consider two Gamma priors for the precision $\tau = \sigma^{-2}$:
  1. $\tau \sim Ga(0.5, 1)$.
  2. $\tau \sim Ga(2, 1)$.

- Given the priors, compute the posterior for τ and plot.

Example


Prior and posteriors in simple linear regression example
Prior Gamma(0.5,1)
Posterior for Gamma(0.5,1) prior
Prior Gamma(2,1)
Posterior for Gamma(2,1) prior
Density
Precision parameter $\tau = \sigma^{-2}$
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0.5
1.5
2
2.5
3
3.5
4

Bayesian linear regression

- Linear regression: $Y = X\beta + \epsilon$ with $\epsilon \sim N(0, I_n/\tau)$.
- Unknowns: β and τ.
- Compute the likelihood $p(Y|\beta, \tau)$.
- Prior densities: $\beta \sim N(\beta_0, \Sigma_0)$ and $\tau \sim Ga(a_0, b_0)$
- We were able to compute the conditional posteriors $p(\beta|\tau, Y)$ and $p(\tau|\beta, Y)$.
- **But we want to sample from** $p(\beta, \tau|Y)$. How do we do it?

Bayesian linear regression

- Linear regression: $Y = X\beta + \epsilon$ with $\epsilon \sim N(0, I_n/\tau)$.
- Unknowns: β and τ.
- Compute the likelihood $p(Y|\beta, \tau)$.
- Prior densities: $\beta \sim N(\beta_0, \Sigma_0)$ and $\tau \sim Ga(a_0, b_0)$
- We were able to compute the conditional posteriors $p(\beta|\tau, Y)$ and $p(\tau|\beta, Y)$.
- **But we want to sample from** $p(\beta, \tau|Y)$. How do we do it?
  - Use the **Gibbs Sampler**.

Gibbs Sampler in the linear regression model

- We will construct a Markov Chain $\{\beta^j, \tau^j\}$ for $j = 1, 2, \ldots, N$ where N is a large.

- **Algorithm**: `Choose a large` N `and an arbitrary` τ^0`, and set` $j = 1$`. Then iterate on the following loop:`
  1. `Draw` β^j `from` $p(\beta|\tau^{j-1}, Y)$.
  2. `Draw` τ^j `from` $p(\tau|\beta^j, Y)$.
  3. `Store` β^j, τ^j`, set` $j = j + 1$ `and return to step 1 while` $j < N$.

- With this algorithm we compute a long sequence $\{\beta^j, \tau^j\}$ whose invariant distribution can be shown to be the posterior density $p(\beta, \tau|Y)$.

- We usually discard the first few hundreds or thousands draws to eliminate the impact of the initial arbitrary value.

Example: Gibbs sampler