# Chapter 8: Estimation and Model Fitting (Rice 8.6–8.9)

Donghyun Ko

May 27, 2026

## Contents

# 1 BayesianParameter Estimation

## 1.1 Bayesian Paradigm

Probability distributions are often used to model uncertainty in unknown quantities. Suppose we observe data

$$X = (X_1, X_2, \ldots, X_n)$$

generated from a probability model $f(x|\theta)$, where $\theta$ is an unknown parameter $\theta$. The goal of statistical inference is to use the observed data to learn about the unknown parameter.

**(Frequentist Paradigm)** In the frequentist framework, the parameter $\theta$ is assumed to be an unknown but fixed constant. Randomness is believed to arise only from the data. Inference about $\theta$ is performed using the sampling distribution of estimators such as

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n).$$

Our goal is to use DATA to estimate $\theta$ and/or draw inference about it. Examples include

- Maximum likelihood estimation
- Method of moments
- Confidence intervals.

**(Bayesian Paradigm)** In Bayesian statistics, the parameter itself is treated as a random variable. This paradigm assume that the $\theta$ arises as a realization of a random quantity (call it $\Theta$); and as such, some values are more "believable" than others. Let $\Theta$ denote the random variable corresponding to the parameter. Before observing the data, our uncertainty about $\Theta$ is described by a probability distribution called the **prior distribution**

$$f_\Theta(\theta).$$

Our goal is to update the knowledge about this quantity ($\Theta$), when the new information, DATA, is received. After observing the data, the prior distribution is updated using Bayes' theorem to produce the **posterior distribution**

$$f_{\Theta|X}(\theta|x).$$

The posterior distribution represents the updated belief about the parameter after observing the data. This is called Bayesian paradigm.

**Distinction between $\theta$ and $\Theta$ in Bayesian Statistics**

**1. Parameter value ($\theta$).**

- $\theta$ denotes a specific numerical value of the parameter. For example, $\theta$ could be 0.3, 2.1, or any other value in the parameter space.
- In the data-generating mechanism, the true parameter value might be $\theta = 0.3$, but this value is unknown to us.
- The likelihood function is typically written as $f(x|\theta)$, which is viewed as a function of $\theta$ for the observed data $x$.

**2. Random parameter ($\Theta$).**

- $\Theta$ denotes a **random variable** representing the uncertain parameter. In Bayesian statistics, uncertainty about the parameter is modeled probabilistically by treating the parameter as a random variable.
- The prior distribution of the parameter is written as $f_\Theta(\theta)$. Here, the symbol $\theta$ represents a possible value that the random variable $\Theta$ may take.

**3. Bayesian model structure.** The Bayesian framework can be summarized as

$$\Theta \sim f_\Theta(\theta) \quad \text{(prior belief about the parameter)}$$

$$X|\Theta = \theta \sim f(x|\theta) \quad \text{(data-generating model)}$$

After observing data $X = x$, the posterior distribution becomes

$$\Theta|X = x \sim f_{\Theta|X}(\theta|x).$$

Even in Bayesian statistics, the real-world data are generated from some fixed but unknown value $\theta^*$. However, since this value is unknown, our knowledge about it is modeled using the random variable $\Theta$. In the notation of $f_\Theta(\theta)$, the symbol $\theta$ represents a possible value of $\Theta$ and should not be confused with the unknown true parameter $\theta^*$. In short,

- $\Theta$ represents the **uncertainty about the true parameter**.
- $\theta$ represents **a particular candidate value** of that parameter.

## 2 Prior, Likelihood, and Posterior

Bayesian inference combines three fundamental components: the **prior distribution**, the **likelihood function**, and the **posterior distribution**. These components describe how information about an unknown parameter is updated after observing data.

**(Prior distribution)** Before observing the data, our uncertainty about the parameter is represented by the **prior distribution**

$$f_\Theta(\theta).$$

This distribution (Distribution of $\Theta$) reflects the knowledge (or belief) about the parameter prior to observing the data, called 'prior distribution'.

**(Likelihood function)** Suppose the observed data are

$$X = (X_1, \ldots, X_n).$$

This DATA corresponds to a realization of $\Theta$ equal to $\theta$. The model describing the data generating mechanism is written as

$$f_{X|\Theta}(x|\theta).$$

This function is called the **likelihood function**. It represents the probability (or density) of observing the data given a particular parameter value (a chosen value from the parameter space).

**(Posterior distribution)** After observing the data, our knowledge about the parameter (i.e., the distribution of $\Theta$) is updated by $(\Theta|X = x) \equiv (\Theta|DATA)$ through Bayes' theorem. The resulting distribution is called the **posterior distribution**

$$f_{\Theta|X}(\theta|x).$$

Bayes' theorem states that

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)}$$

where

$$f_X(x) = \int f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta$$

is the **marginal likelihood**. Since the denominator does not depend on $\theta$, we often write

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta)f_{\Theta}(\theta).$$

Thus, the posterior density can be interpreted as

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

Equivalently,

$$f_{\Theta|X}(\theta|x) = \frac{f(x_1, \ldots, x_n|\theta)f_{\Theta}(\theta)}{\int f(x_1, \ldots, x_n|\theta)f_{\Theta}(\theta)d\theta}.$$

The denominator serves as a normalizing constant ensuring the posterior density integrates to 1.

**Example: Bayesian Analysis of a Coin Flipping Experiment**

Consider a coin flipping experiment where a coin is flipped $n = 10$ times. Let

$$X = \text{“number of heads”}.$$

Assume that

$$X|\theta \sim Binomial(n, \theta),$$

where $\theta$ is the probability of obtaining a head. The likelihood function is therefore

$$f(x|\theta) = \binom{10}{x}\theta^x(1-\theta)^{10-x}.$$

Suppose that the parameter $\theta$ can take one of two possible values:

$$\theta = 0.5 \quad \text{(fair coin), or}$$

$$\theta = 0.9 \quad \text{(strongly biased coin)}$$

with prior probabilities

$$P(\Theta = 0.5) = 0.2, \qquad P(\Theta = 0.9) = 0.8.$$

We want to compute the posterior probabilities

$$P(\Theta = 0.5|X = x), \qquad P(\Theta = 0.9|X = x).$$

Using Bayes' theorem,

$$P(\Theta = \theta_i | X = x) = \frac{P(X = x|\Theta = \theta_i)P(\Theta = \theta_i)}{\sum_j P(X = x|\Theta = \theta_j)P(\Theta = \theta_j)}.$$

For $\theta = 0.5$,

$$P(X = x|0.5) = \binom{10}{x}(0.5)^{10}.$$

For $\theta = 0.9$,

$$P(X = x|0.9) = \binom{10}{x}(0.9)^x(0.1)^{10-x}.$$

Thus,

$$P(\Theta = 0.5|X = x) = \frac{0.2\binom{10}{x}(0.5)^{10}}{0.2\binom{10}{x}(0.5)^{10} + 0.8\binom{10}{x}(0.9)^x(0.1)^{10-x}},$$

$$P(\Theta = 0.9|X = x) = \frac{0.8\binom{10}{x}(0.9)^x(0.1)^{10-x}}{0.2\binom{10}{x}(0.5)^{10} + 0.8\binom{10}{x}(0.9)^x(0.1)^{10-x}}.$$

**Interpretation**

- If $x = 1$, the observed data are much more consistent with a fair coin, so the posterior probability of $\theta = 0.5$ becomes large.
- If $x = 5$, the likelihood supports $\theta = 0.5$, but the prior favors $\theta = 0.9$, so the posterior distribution reflects a compromise between prior belief and observed data.
- If $x = 9$, the observed data strongly support $\theta = 0.9$, and the posterior probability of $\theta = 0.9$ becomes dominant.

This example illustrates how Bayesian inference combines prior information and observed data.

Example: Continuous Prior and Posterior Derivation

Consider again the coin flipping experiment where $X|\theta \sim Binomial(n, \theta)$. Suppose the prior distribution for the parameter is

$$\Theta \sim Beta(\alpha, \beta)$$

with density

$$f_\Theta(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \qquad 0 < \theta < 1.$$

**Step 1: Write the likelihood function.** Since $X|\theta \sim Binomial(n, \theta)$, the likelihood of observing $X = x$ is

$$f_{X|\Theta}(x|\theta) = P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

**Step 2: Apply Bayes' theorem.** The posterior density is

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta)f_\Theta(\theta)d\theta}.$$

Since the denominator does not depend on $\theta$, we work with the proportional form

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) f_{\Theta}(\theta).$$

This is the general form of the posterior of $\Theta|DATA$.

**Step 3: Substitute the likelihood and prior.** Substituting the expressions for the likelihood and the prior,

$$f_{\Theta|X}(\theta|x) \propto \binom{n}{x} \theta^x (1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}.$$

The terms $\binom{n}{x}$ and $B(\alpha,\beta)$ do not depend on $\theta$, so they can be absorbed into the proportionality constant. Therefore,

$$f_{\Theta|X}(\theta|x) \propto \theta^x (1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Using properties of exponents,

$$f_{\Theta|X}(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

**Step 4: Identify the posterior distribution.** The kernel above matches the form of a Beta distribution

$$Beta(a,b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

Hence, the posterior distribution of $\Theta|DATA$ is

$$\Theta|X \sim Beta(\alpha + x, \beta + n - x).$$

**Step 5: Bayesian point estimate.** In Bayesian inference, a common choice for a point estimator of the parameter is either the **posterior mean** or the **posterior mode** of the posterior distribution. A natural choice is the posterior mean, which minimizes the posterior expected squared error loss. Thus, a reasonable Bayesian point estimator for $\theta$ is

$$\hat{\theta}_B = E(\Theta|DATA) = \frac{\alpha + x}{\alpha + \beta + n}.$$

This estimator combines the prior information $(\alpha, \beta)$ and the observed data $(x, n)$, reflecting the Bayesian principle of updating prior belief using observed data.

**(Bayesian inference based on the posterior)** In Bayesian statistics, estimation and inference about the unknown parameter are based entirely on the **posterior distribution** $f_{\Theta|X}(\theta|x)$. Recall that the posterior density can be expressed as

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) f_{\Theta}(\theta),$$

that is, the posterior is proportional to the likelihood multiplied by the prior.

**Bayesian trick.** In many problems, it is not necessary to compute the normalizing constant in Bayes' formula explicitly. Instead, we examine the right-hand side

$$f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)$$

and identify the family of distributions that matches the terms involving $\theta$. The missing proportionality constant can then be determined so that the resulting function integrates to one:

$$\int f_{\Theta|X}(\theta|x)d\theta = 1.$$

This technique is frequently used when deriving posterior distributions. Once the posterior distribution is obtained, several types of statistical summaries can be computed:

- **Point estimation.** A Bayesian point estimator for the parameter is commonly taken to be the **posterior mean** or the **posterior mode** (also called the MAP estimator):

$$\hat{\theta}_{Bayes} = E(\Theta|X) \quad \text{or} \quad \hat{\theta}_{MAP} = \arg\max_{\theta} f_{\Theta|X}(\theta|x).$$

- **Variability of the estimator.** The uncertainty associated with the estimator is measured using the **posterior variance** or the **posterior standard deviation**

$$Var(\Theta|X), \qquad SD(\Theta|X).$$

  These quantities describe how concentrated the posterior distribution is around the estimated parameter value.

- **Credible intervals.** Uncertainty in the parameter estimate is often expressed using **credible intervals**, which are intervals derived from the posterior distribution. Two common types of credible intervals are:

  - **Percentile (basic) credible interval**. This interval is constructed using the lower and upper quantiles of the posterior distribution. For example, the $5^{th}$ and $95^{th}$ percentiles produce a 90% credible interval.
  - **Highest Posterior Density (HPD) interval**. This interval contains the values of $\theta$ with the highest posterior density and has the smallest possible length among all intervals with the same probability content.

  **Interpretation of credible intervals.** Suppose $[L, U]$ is a 90% credible interval for the parameter $\Theta$ after observing the data. Then,

$$P(L \leq \Theta \leq U \mid DATA) = 0.9.$$

  This means that, given the observed data, there is a 90% probability that the parameter lies within the interval $[L, U]$.

**(Choosing the prior distribution)** An important step in Bayesian analysis is selecting an appropriate prior distribution for the parameter. Two general approaches are commonly used.

- **Orthodox (subjective) Bayesian approach**. The prior distribution $f_{\Theta}(\theta)$ is chosen to reflect genuine prior beliefs about the parameter. These beliefs may be based on expert knowledge, previous experiments, or scientific reasoning.

- **Objective Bayesian approach**. The prior distribution is chosen to be as non-informative as possible so that the data dominate the inference. Such priors are often called **noninformative** or **weakly informative** priors.

In practice, priors are usually selected by considering

- the plausible range of the parameter,
- available prior scientific knowledge,
- mathematical convenience (for example, the use of conjugate priors).

In many applications, conjugate priors are preferred because they lead to posterior distributions with closed-form expressions.

## 2.1 Conjugate Priors

A prior distribution is called a **conjugate prior** for a likelihood function if the posterior distribution belongs to the same family as the prior distribution. In Bayesian inference, the posterior density is given by

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x|u)f_\Theta(u)\,du} \propto f_{X|\Theta}(x|\theta)f_\Theta(\theta).$$

A conjugate prior is useful because after multiplying the likelihood and the prior, the posterior can often be recognized immediately as a familiar distribution, without explicitly evaluating the normalizing constant. Thus, conjugate priors are used mainly for **mathematical convenience**.

**Common examples.**

- If the likelihood comes from a Poisson distribution with unknown rate parameter $\lambda$, then a Gamma distribution is a conjugate prior for $\lambda$.
- If the likelihood comes from a Bernoulli distribution with unknown success probability $p$, then a Beta distribution is a conjugate prior for $p$.
- If the likelihood comes from a Normal distribution with **known variance**, $N(\mu, \sigma_0^2)$, then a Normal distribution is a conjugate prior for $\mu$.

**Why is this helpful?** If the posterior belongs to a known family, then we can immediately obtain:

- the posterior distribution,
- the posterior mean,
- the posterior variance,
- credible intervals.

## 2.2 Improper Priors

Definition: Improper Prior

A prior is called **improper** if its "density" does not have a finite integral, that is,

$$\int f_\Theta(\theta)\,d\theta = \infty.$$

Such a function is not a valid probability density. However, improper priors are still accepted in Bayesian analysis as long as the resulting posterior distribution is proper, meaning that the posterior density integrates to 1.

**Example: Poisson model with an improper prior.** Assume

$$X_1, \ldots, X_n \mid \Lambda = \lambda \overset{\text{IID}}{\sim} \text{Poisson}(\lambda),$$

and suppose the prior is

$$f_\Lambda(\lambda) = \lambda^{-1}, \qquad 0 < \lambda < \infty.$$

First, check whether this prior is proper:

$$\int_0^\infty \lambda^{-1}\, d\lambda = \infty.$$

So the prior is improper. Now derive the posterior. The joint likelihood is

$$f(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \left(\prod_{i=1}^n \frac{1}{x_i!}\right) e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}.$$

Ignoring constants that do not depend on $\lambda$,

$$f(x_1, \ldots, x_n|\lambda) \propto e^{-n\lambda}\lambda^{\sum x_i}.$$

Multiply by the prior:

$$f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1, \ldots, x_n) \propto e^{-n\lambda}\lambda^{\sum x_i}\lambda^{-1} = \lambda^{\sum x_i - 1}e^{-n\lambda}.$$

This is the kernel of a Gamma density. Therefore,

$$\Lambda|X_1 = x_1, \ldots, X_n = x_n \sim \text{Gamma}\left(\sum_{i=1}^n x_i,\ n\right),$$

provided $\sum x_i > 0$, so that the posterior is proper. Therefore, even though the prior is not integrable, the posterior may still be a valid probability density.

## 2.3 Beta–Binomial Model

This is the standard Bayesian model for an unknown success probability.

**Model Setup**

Suppose

$$X|\theta \sim \text{Binomial}(n, \theta),$$

where $\theta$ is the unknown probability of success. Assume the prior distribution is

$$\Theta \sim \text{Beta}(\alpha, \beta),$$

with density

$$f_\Theta(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \qquad 0 < \theta < 1.$$

**Posterior derivation.** The likelihood is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

By Bayes' rule,

$$f_{\Theta|X}(\theta|x) \propto f(x|\theta)f_\Theta(\theta).$$

Substitute the likelihood and prior:

$$f_{\Theta|X}(\theta|x) \propto \left[\binom{n}{x}\theta^x(1-\theta)^{n-x}\right]\left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\right].$$

Ignoring constants that do not depend on $\theta$,

$$f_{\Theta|X}(\theta|x) \propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Combine exponents:

$$f_{\Theta|X}(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

This is exactly the kernel of a Beta distribution. Therefore,

$$\Theta|X = x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

**Conclusion: Beta–Binomial Conjugacy**

If

$$X|\theta \sim \text{Binomial}(n,\theta), \qquad \Theta \sim \text{Beta}(\alpha,\beta),$$

then

$$\Theta|X = x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

**Posterior mean and variance.** A common Bayesian point estimator is the posterior mean:

$$\hat{\theta}_{\text{Bayes}} = E(\Theta|X = x).$$

Since a Beta$(a, b)$ distribution has mean

$$E(\Theta) = \frac{a}{a+b},$$

we obtain

$$E(\Theta|X = x) = \frac{\alpha + x}{\alpha + \beta + n}.$$

Similarly, because

$$\text{Var}(\Theta) = \frac{ab}{(a+b)^2(a+b+1)},$$

the posterior variance is

$$\text{Var}(\Theta|X = x) = \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}.$$

**Example: 50 items, 3 defectives**

Suppose 50 items are sampled from a manufacturing process and 3 are found to be defective. Let $\theta$ be the true defective proportion. Then,

$$X|\theta \sim \text{Binomial}(50, \theta), \qquad x = 3.$$

**Case A: Statistician A uses a uniform prior.** Suppose

$$\Theta \sim U(0, 1) = \text{Beta}(1, 1).$$

Then, the posterior is

$$\Theta|X = 3 \sim \text{Beta}(1 + 3, 1 + 47) = \text{Beta}(4, 48).$$

The posterior mean is

$$E(\Theta|X = 3) = \frac{4}{52} = \frac{1}{13} \approx 0.0769.$$

**Case B: Statistician B uses an informative prior.** Suppose $\Theta \sim \text{Beta}(20, 2)$.
Then, the posterior is

$$\Theta|X = 3 \sim \text{Beta}(20 + 3, 2 + 47) = \text{Beta}(23, 49).$$

The posterior mean is

$$E(\Theta|X = 3) = \frac{23}{72} \approx 0.3194.$$

**Discussion.** The sample proportion is

$$\hat{p} = \frac{3}{50} = 0.06.$$

Under the uniform prior, the posterior mean is close to the observed sample proportion. Under the informative Beta(20, 2) prior, the posterior is pulled strongly toward the prior belief. This shows that when the prior is informative, it can substantially affect the posterior, especially when the sample size is not very large.

**Credible intervals** In Bayesian inference, interval estimates are called **credible intervals**.

Interpretation of a Credible Interval

If $[L, U]$ is a 95% credible interval for $\Theta$, then

$$P(L \leq \Theta \leq U \mid \text{data}) = 0.95.$$

This means that, given observed data, the posterior probability that $\Theta$ lies between $L$ and $U$ is 0.95.

Two common types are:

- **Quantile-based credible interval**
- **Highest posterior density (HPD) interval**

For the Beta–Binomial model, if

$$\Theta | X = x \sim \text{Beta}(\alpha + x, \beta + n - x),$$

then a 95% quantile-based credible interval is

$$\left[q_{0.025},\, q_{0.975}\right],$$

where $q_{0.025}$ and $q_{0.975}$ are the 2.5th and 97.5th percentiles of the posterior Beta distribution. In R, these can be computed by

$$\texttt{qbeta}(\texttt{0.025}, \alpha + \texttt{x}, \beta + \texttt{n} - \texttt{x}), \quad \text{and} \quad \texttt{qbeta}(\texttt{0.975}, \alpha + \texttt{x}, \beta + \texttt{n} - \texttt{x}).$$

## 2.4 Normal Likelihood with Known Variance and Normal Prior

Now consider a Normal model with unknown mean and known variance. It is convenient to use the **precision**

$$\xi = \frac{1}{\sigma^2}$$

instead of the variance.

**Model Setup**

Suppose

$$X_1, \ldots, X_n \overset{\text{IID}}{\sim} N(\theta, \sigma_0^2),$$

where $\sigma_0^2$ is known, so equivalently the precision $\xi_0 = 1/\sigma_0^2$ is known. Assume the prior

$$\Theta \sim N(\theta_{\text{prior}}, 1/\xi_{\text{prior}}),$$

where $\theta_{\text{prior}}$ and $\xi_{\text{prior}}$ are known constants.

### Posterior derivation

The likelihood is

$$f(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \left(\frac{\xi_0}{2\pi}\right)^{1/2} \exp\left[-\frac{\xi_0}{2}(x_i - \theta)^2\right].$$

Ignoring constants not depending on $\theta$,

$$f(x_1, \ldots, x_n | \theta) \propto \exp\left[-\frac{\xi_0}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right].$$

The prior density is

$$f_\Theta(\theta) \propto \exp\left[-\frac{\xi_{\text{prior}}}{2}(\theta - \theta_{\text{prior}})^2\right].$$

Therefore,

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \exp\left[-\frac{\xi_0}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right] \exp\left[-\frac{\xi_{\text{prior}}}{2}(\theta - \theta_{\text{prior}})^2\right].$$

Combine exponents:

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left\{\xi_0 \sum_{i=1}^{n}(x_i-\theta)^2 + \xi_{\text{prior}}(\theta-\theta_{\text{prior}})^2\right\}\right].$$

Expand the quadratic terms:

$$\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_{i=1}^{n} x_i^2 - 2\theta\sum_{i=1}^{n} x_i + n\theta^2,$$

and

$$(\theta-\theta_{\text{prior}})^2 = \theta^2 - 2\theta\theta_{\text{prior}} + \theta_{\text{prior}}^2.$$

Substituting and collecting terms involving $\theta$ gives

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left\{(n\xi_0+\xi_{\text{prior}})\theta^2 - 2\left(\xi_0\sum_{i=1}^{n} x_i + \xi_{\text{prior}}\theta_{\text{prior}}\right)\theta + C\right\}\right],$$

where $C$ does not depend on $\theta$. Now complete the square. Define

$$\xi_{\text{post}} = n\xi_0 + \xi_{\text{prior}},$$

and

$$\theta_{\text{post}} = \frac{\xi_0\sum_{i=1}^{n} x_i + \xi_{\text{prior}}\theta_{\text{prior}}}{n\xi_0+\xi_{\text{prior}}}.$$

Then,

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \exp\left[-\frac{\xi_{\text{post}}}{2}(\theta-\theta_{\text{post}})^2\right],$$

which is the kernel of a Normal density. Therefore,

$$\Theta|\mathbf{X} \sim N\left(\theta_{\text{post}}, \frac{1}{\xi_{\text{post}}}\right).$$

That is,

$$\Theta|\mathbf{X} \sim N\left(\frac{\xi_0\sum_{i=1}^{n} x_i + \xi_{\text{prior}}\theta_{\text{prior}}}{n\xi_0+\xi_{\text{prior}}}, \frac{1}{n\xi_0+\xi_{\text{prior}}}\right).$$

The posterior mean is a weighted average of the sample information and the prior mean, with weights determined by precisions. A larger precision means greater certainty.

**95% Bayes basic interval and HPD interval**

Since the posterior is Normal and symmetric,

$$\Theta|\mathbf{X} \sim N\left(\theta_{\text{post}}, \frac{1}{\xi_{\text{post}}}\right),$$

a 95% Bayes basic interval is

$$\theta_{\text{post}} \pm 1.96\sqrt{\frac{1}{\xi_{\text{post}}}}.$$

Because the posterior distribution is symmetric and unimodal, the 95% HPD interval is the same:

$$\left[\theta_{\text{post}} - \frac{1.96}{\sqrt{\xi_{\text{post}}}},\ \theta_{\text{post}} + \frac{1.96}{\sqrt{\xi_{\text{post}}}}\right].$$

So in this case, the Bayes basic interval and the HPD interval are **not different**.

# 3 Bayesian Inference and Computational Aspects

A central difficulty in Bayesian analysis is calculating the normalizing constant in the posterior density:

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{\int f_{X,\Theta}(x,\theta)\, d\theta}.$$

There are three main strategies to tackle this issue:

- **Approach 1: Recognize the posterior distribution directly from the numerator.** Since
$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) f_\Theta(\theta),$$
if the right-hand side matches a known density kernel, we can identify the posterior without evaluating the integral explicitly.
- **Approach 2: Approximate the posterior by a Normal distribution.** This is the idea of the **Laplace approximation**.
- **Approach 3: Use sampling-based methods.** If the posterior is too complicated to identify or approximate well, we can sample from it using **Markov Chain Monte Carlo (MCMC)** methods.

## 3.1 Laplace Approximation

If the posterior density cannot be identified exactly, one useful idea is to approximate it by a Normal distribution.

**Idea of Laplace Approximation**

If the posterior is sharply peaked near its maximum, then the log-posterior can be approximated by a quadratic function near that maximum. Exponentiating the quadratic form gives a Normal approximation.

Suppose

$$X_1, \ldots, X_n \overset{\text{IID}}{\sim} f(x|\theta), \qquad \Theta \sim f_\Theta(\theta).$$

Then, the posterior is

$$f_{\Theta|X}(\theta|x) \propto f_\Theta(\theta) \prod_{i=1}^{n} f(x_i|\theta).$$

Write

$$q(\theta) = \ln f_\Theta(\theta) + \ell(\theta), \quad \text{where} \quad \ell(\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta).$$

So,

$$f_{\Theta|X}(\theta|x) \propto e^{q(\theta)}.$$

Let $\hat{\theta}$ maximize $q(\theta)$. Then, $\hat{\theta}$ is the posterior mode. Using a 2nd-order Taylor expansion around $\hat{\theta}$,

$$q(\theta) \approx q(\hat{\theta}) + (\theta - \hat{\theta}) q'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 q''(\hat{\theta}).$$

Since $\hat{\theta}$ maximizes $q(\theta)$,

$$q'(\hat{\theta}) = 0.$$

Therefore,

$$q(\theta) \approx q(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 q''(\hat{\theta}).$$

Because $q''(\hat{\theta}) < 0$ at a maximum,

$$q(\theta) \approx q(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 [-q''(\hat{\theta})].$$

Exponentiating,

$$f_{\Theta|X}(\theta|x) \propto e^{q(\theta)} \approx e^{q(\hat{\theta})} \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 [-q''(\hat{\theta})]\right).$$

Ignoring the constant factor $e^{q(\hat{\theta})}$,

$$f_{\Theta|X}(\theta|x) \approx \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 [-q''(\hat{\theta})]\right),$$

which is the kernel of a Normal density. Hence,

$$\Theta|X \approx N\left(\hat{\theta}, [-q''(\hat{\theta})]^{-1}\right).$$

**Special case.** If the prior is nearly constant in the region where the likelihood is large, then the posterior is dominated by the likelihood. In that case, $\hat{\theta}$ is approximately the MLE, and

$$\Theta|X \approx N\left(\hat{\theta}_{MLE}, [-\ell''(\hat{\theta}_{MLE})]^{-1}\right).$$

**Remark.** The assumption that the likelihood dominates the prior is used only for simplicity. In general, Laplace approximation can be applied to the full function

$$q(\theta) = \ln f_{\Theta}(\theta) + \ell(\theta),$$

and then the maximizer is the posterior mode, not necessarily the MLE.

### 3.2 Markov Chain Monte Carlo (MCMC)

If neither direct identification nor Laplace approximation works well, then we can access the posterior distribution by drawing samples from it.

**MCMC Idea**

MCMC methods are used to compute numerical approximations of integrals of the form

$$\int H(\theta) f(\theta)\, d\theta = E_f[H(\Theta)].$$

If $\theta^{(1)}, \ldots, \theta^{(N)}$ are samples from $f(\theta)$, then

$$E_f[H(\Theta)] \approx \frac{1}{N} \sum_{i=1}^{N} H(\theta^{(i)}).$$

In Bayesian inference, we apply this to the posterior:

$$E_{\Theta|X}[H(\Theta)] = \int H(\theta) f_{\Theta|X}(\theta|x)\, d\theta.$$

**Why do we care about MCMC?**

- Posterior densities are often too complicated to handle analytically.
- MCMC methods allow us to generate samples from the posterior and use them to estimate expectations, variances, and credible intervals.
- These methods are especially useful in high-dimensional models.

**Basic idea.** MCMC constructs a sequence

$$\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$$

called a **Markov chain**. After running long enough, the chain reaches equilibrium, and the sampled values behave like draws from the target distribution.

**Common algorithms.** When the posterior distribution is too complicated to analyze directly, we can approximate it by generating many draws from it using **Markov Chain Monte Carlo (MCMC)** methods. The key idea is that, even if we cannot compute the posterior density in closed form, we may still be able to generate a long sequence of draws whose long-run behavior matches the posterior distribution. Then, posterior characteristics can be approximated empirically from the simulated draws. In particular, if, after discarding burn-in, we retain

$$\theta_{K+1}, \theta_{K+2}, \ldots, \theta_N,$$

then we may estimate:

$$E(\Theta|X=x) \approx \bar{\theta} = \frac{1}{N-K} \sum_{t=K+1}^{N} \theta_t,$$

$$\mathrm{Var}(\Theta|X=x) \approx \frac{1}{N-K} \sum_{t=K+1}^{N} (\theta_t - \bar{\theta})^2,$$

$$SD(\Theta|X=x) \approx \left[ \frac{1}{N-K} \sum_{t=K+1}^{N} (\theta_t - \bar{\theta})^2 \right]^{1/2}.$$

We may also approximate:

- the posterior mode by the highest peak in the histogram of the retained draws,
- a basic 95% Bayesian credible interval by the empirical 2.5% and 97.5% quantiles of the retained draws.

Thus, MCMC converts a difficult posterior-analysis problem into a simulation problem: once we have enough draws from the posterior, many posterior summaries can be estimated numerically. The two most important algorithms introduced in these notes are **Metropolis–Hastings** and **Gibbs sampling**. Both produce a sequence of dependent draws, called a **Markov chain**, and after the chain has run long enough, its equilibrium (stationary) distribution is the target posterior distribution.

**Metropolis–Hastings algorithm**

The **Metropolis–Hastings (MH) algorithm** is used when direct sampling from the posterior distribution is difficult. Its main idea is to construct a Markov chain that converges to the desired posterior distribution as its stationary distribution. One of its biggest advantages is that we do *not* need to know the normalizing constant of the posterior density. This is extremely important in Bayesian inference, because the posterior usually has the form

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) f_{X|\Theta}(x|\theta),$$

and the denominator in Bayes' formula is often very hard to compute. In Metropolis–Hastings, this denominator cancels when we form a ratio, so we can work only with the kernel of the posterior. Suppose the current state of the Markov chain at iteration $t-1$ is $\theta_{t-1}$. The MH algorithm proceeds as follows.

1. **Initialize.** Choose a starting value $\theta_0$ and set $t = 0$.

2. **Propose a candidate.** Generate a candidate value $\theta^*$ from a proposal distribution centered at the current state $\theta_{t-1}$. This proposal distribution is often called the **jumping distribution**, denoted by
$$\theta^* \sim J(\theta_{t-1}).$$
A common choice is a Normal distribution centered at the previous state.

3. **Compute the acceptance ratio.** Calculate
$$r = \frac{f_{\Theta|X}(\theta^*|x)}{f_{\Theta|X}(\theta_{t-1}|x)}.$$
Using the posterior kernel, this becomes
$$r = \frac{f_{\Theta}(\theta^*) f_{X|\Theta}(x|\theta^*)}{f_{\Theta}(\theta_{t-1}) f_{X|\Theta}(x|\theta_{t-1})}.$$
The unknown normalizing constant cancels out, which is exactly why this method is practical.

4. **Accept or reject the candidate.**

   - If $r \geq 1$, then the proposed value has at least as much posterior support as the current value, so we accept it with probability 1 and set
   $$\theta_t = \theta^*.$$

   - If $r < 1$, then we accept the proposed value only with probability $r$. Otherwise we stay where we are:
   $$\theta_t = \begin{cases} \theta^*, & \text{with probability } r, \\ \theta_{t-1}, & \text{with probability } 1 - r. \end{cases}$$

   Equivalently, the acceptance probability is
   $$A(\theta^*, \theta_{t-1}) = \min\{r, 1\}.$$

5. **Increment and repeat.** Set $t = t + 1$ and repeat this process many times.

**Why does this algorithm work?** The rule is designed so that the chain tends to move toward regions where the posterior density is large. If the proposed point $\theta^*$ has a larger posterior density than the current point, it is always accepted. If it has a smaller posterior density, it still has some chance of being accepted. This is important: if we only moved uphill, the chain might get trapped too easily. By occasionally accepting lower-density points, the chain continues to explore the parameter space. After many iterations, the chain spends most of its time in regions of high posterior probability, and its equilibrium distribution becomes the posterior distribution itself.

**Choice of proposal distribution.** The standard deviation of the proposal distribution must be chosen carefully.

- If the proposal standard deviation is too **small**, then most proposed moves are accepted, but the chain moves only in tiny steps. As a result, exploration is slow.
- If the proposal standard deviation is too **large**, then most proposed moves are rejected, so the chain stays at the same value for many iterations and sampling becomes inefficient.
- Therefore, a good proposal distribution should balance movement and acceptance so that the chain explores efficiently while still accepting a reasonable fraction of proposals.

**Burn-in and posterior summaries.** Once we obtain a long chain

$$\theta_1, \theta_2, \ldots, \theta_N,$$

we usually discard the first $K$ draws,

$$\theta_1, \ldots, \theta_K,$$

because they are heavily influenced by the starting value. This discarded portion is called the **burn-in**. The remaining draws

$$\theta_{K+1}, \ldots, \theta_N$$

are then treated as approximate posterior draws and used to estimate the posterior mean, posterior variance, posterior standard deviation, posterior mode, and credible intervals.

### Gibbs sampling

**Gibbs sampling** is another important MCMC method. Unlike Metropolis–Hastings, it does not rely on proposing a candidate and then accepting or rejecting it. Instead, it relies on being able to sample directly from **conditional distributions**. For this reason, Gibbs sampling is especially useful when the posterior involves several parameters and each parameter has a manageable full conditional distribution. Suppose the unknown parameter consists of two components, say $(\Theta, \Xi)$, and suppose our target is the joint posterior distribution

$$f_{\Theta,\Xi|X}(\theta, \xi|x).$$

Even if this joint distribution is difficult to sample from directly, it may still be possible to sample from the two full conditional distributions

$$f_{\Theta|\Xi,X}(\theta|\xi, x) \qquad \text{and} \qquad f_{\Xi|\Theta,X}(\xi|\theta, x).$$

The Gibbs sampling algorithm then works as follows.

1. **Initialize.** Pick a starting value for one parameter, for example $\xi_0$.

2. **Update $\Theta$.** Given the current value $\xi_{t-1}$, generate

$$\theta_t \sim f_{\Theta|\Xi,X}(\theta|\xi_{t-1}, x).$$

3. **Update $\Xi$.** Using the newly generated value $\theta_t$, generate

$$\xi_t \sim f_{\Xi|\Theta,X}(\xi|\theta_t, x).$$

4. **Repeat.** Set $t = t + 1$ and continue alternating these conditional updates many times.

This produces a sequence of pairs

$$(\theta_0, \xi_0), (\theta_1, \xi_1), (\theta_2, \xi_2), \ldots$$

which forms a Markov chain. After enough iterations, these pairs behave like draws from the joint posterior distribution of $(\Theta, \Xi)|X = x$. Then, for example, the marginal posterior distribution of $\Theta$ can be studied simply by looking at the retained $\theta_t$ values. As in MH, we typically discard an initial burn-in portion before using the sample for inference.

**Why is Gibbs sampling attractive?** Gibbs sampling is often simpler than Metropolis–Hastings when the full conditional distributions are standard distributions such as Normal or Gamma. In that case, every update step is straightforward: just draw from the appropriate conditional distribution. Also, unlike MH, there is no rejection step. Every conditional draw is automatically accepted. This makes Gibbs sampling especially appealing in multi-parameter Bayesian models.

**Example: Normal model with unknown mean and precision.** Consider a model

$$X|\Theta = \theta, \Xi = \xi \sim N(\theta, \xi^{-1}),$$

with independent priors

$$\Theta \sim N(\theta_{\text{prior}}, \xi_{\text{prior}}^{-1}), \qquad \Xi \sim \text{Gamma}(\alpha, \lambda).$$

From the joint posterior, we identify the full conditional distributions by keeping only the terms involving one parameter at a time and matching kernels. This gives

$$\Theta|\Xi, X = x \sim N(\theta_{\text{post}}, 1/\xi_{\text{post}})$$

and

$$\Xi|\Theta, X = x \sim \text{Gamma}(\cdots),$$

so the Gibbs sampler alternates:

1. sample a new $\Theta$ given the current value of $\Xi$,

2. then sample a new $\Xi$ given the newly updated value of $\Theta$.

Repeating this many times gives approximate draws from the joint posterior distribution. The notes also emphasize that we can then use the retained draws of $\Theta$ alone to form a credible interval for $\Theta$.

**Comparing the two algorithms.**

- **Metropolis–Hastings** is more general. It can be used even when direct sampling from conditional distributions is not available. However, it requires a proposal distribution and an accept/reject step.
- **Gibbs sampling** is often simpler when the full conditional distributions are known and easy to sample from. Every draw is accepted, but the method depends on our ability to sample from those conditional distributions.

**Practical interpretation for students.** You can think of these two methods as two different ways of "walking around" the posterior distribution.

- In **Metropolis–Hastings**, we propose a move and then decide whether to accept it.
- In **Gibbs sampling**, we do not propose-and-reject; instead, we repeatedly redraw each parameter from its own conditional distribution.

In both cases, after the chain has run long enough and burn-in is removed, the retained draws are used to approximate posterior summaries numerically. That is the practical goal of MCMC.

**Important practical terms.**

- **Trace plot:** plot of sampled values against iteration number.
- **Burn-in:** early iterations discarded before equilibrium is reached.
- **Posterior histogram:** histogram of retained draws, approximating the posterior density.

A trace plot should look stable, without long-term trend, once the chain has converged. After burn-in is removed, the histogram of sampled values approximates the posterior distribution. To obtain a representative sample from the posterior, the chain often needs to be run for a very long time.