

Introduction to Bayesian Inference

Aleksandar Dogandžić

2017-03-21

Contents

Basic Concepts 2

Applying Bayes' rule by using \propto 2

Conjugate Priors 4

Laplace's problem 5

Bayes' problem 6

Conjugate versus Non-conjugate Priors (Binomial Example) 6

Advantages and disadvantages of conjugate priors 8

Summarizing Posterior Distributions 8

Binomial example 9

Numerical summaries of the posterior distribution 11

Prior versus posterior distribution 12

Credible sets 12

Sufficiency and Bayesian Models 16

DC Level Estimation in AWGN with Known Variance 16

Single observation 16

Multiple observations 18

Proper versus Improper Priors 20

Gaussian Distribution with Unknown Variance and Known Mean 20

Estimating the variance of a Gaussian distribution with known mean 22

READING: §10 in the textbook and (Gelman et al. 2014, §2.1–2.6).

✱ SIMPLIFIED notation. When there is no potential for confusion, we use

$$f(x | \theta)$$

for probability density functions (pdfs) instead of the more cumbersome

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

Similarly, for probability mass functions (pmfs), we use $p(x | \theta)$ instead of the more cumbersome $p_{X|\Theta}(x | \theta)$.

Basic Concepts

WE observe the random variable X and wish to find the true state of nature Θ . We need the following ingredients:

1. $f_{X|\Theta}(x|\theta)$ or $p_{X|\Theta}(x|\theta)$ denote the data model, likelihood;
2. $f_{\Theta}(\theta)$ or $f_{\Theta}(\theta)$ is the prior distribution of θ (epistemic probability), i.e., our knowledge about the true state of nature.

In the Bayesian approach, we assign a prior distribution on parameter Θ . Here, Θ is often *not* really random, but the epistemic argument justifies the use of a probability distribution. We apply the Bayes' rule and base our inference on the posterior distribution of Θ :

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{\int \underbrace{f_{X,\Theta}(x,\vartheta) d\vartheta}_{\text{does not depend on } \theta}} \propto f_{X,\Theta}(x,\theta) \quad (1)$$

and, therefore,

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

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



NOTE:

- $f_{\Theta|X}(\theta|x)$ is also an epistemic probability.
- It is important to master the use of \propto .

Make sure that you understand the following:

$$\begin{aligned} f(\theta|x_1, x_2) &\propto f(\theta, x_1, x_2) \\ &\propto f(\theta, x_1|x_2) \\ &\propto f(\theta, x_2|x_1). \end{aligned}$$

Applying Bayes' rule by using \propto

CONSIDER

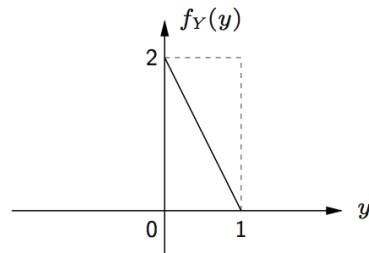
$$f_{X,Y}(x,y) = \begin{cases} 2 & x, y \geq 0, x+y \leq 1 \\ 0, & \text{otherwise} \end{cases}.$$

Now, computing $f_{X|Y}(x|y)$ can be done two ways.

✱ HARDER way.

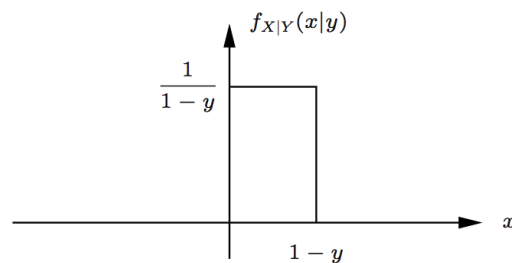
To find $f_Y(y)$, we use the law of total probability

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx \\ &= \begin{cases} \int_0^{(1-y)} 2 dx & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 2(1-y) & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$



$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \begin{cases} \frac{1}{1-y} & 0 \leq y < 1, 0 \leq x \leq 1-y \\ 0 & \text{otherwise} \end{cases}$$

In other words, $X | \{Y = y\} \sim U[0, 1 - y]$



✱ EASIER way. Employ the \propto machinery:

$$f_{X|Y}(x|y) \propto f_{X,Y}(x, y) = \begin{cases} \text{flat in } x, & 0 \leq x \leq 1-y \\ 0, & \text{otherwise} \end{cases}$$

where $0 \leq y \leq 1$.

☞ NOTE. $f_{X|Y}(x|y)$ constant over the range $0 \leq x \leq 1-y$ and zero elsewhere implies that $f_{X|Y}(x|y)$ is a uniform random variable with pdf

$$f_{X|Y}(x|y) = U(x|0, 1-y)$$

i.e.,

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{1-y}, & 0 \leq x \leq 1-y \\ 0, & \text{otherwise} \end{cases} \quad 0 \leq y \leq 1$$

which follows by looking up the table of distributions.

☞ No need for integration, we obtain the normalizing constant from the table of distributions!

Conjugate Priors

If \mathbb{F} is a class of measurement models and \mathbb{P} a class of prior distributions, then \mathbb{P} is *conjugate* for \mathbb{F} if $f_{\Theta}(\theta) \in \mathbb{P}$ and $f_{X|\Theta}(x|\theta) \in \mathbb{F}$ implies $f_{\Theta|X}(\theta|x) \in \mathbb{P}$.

☞ CONJUGATE priors allow us to find analytically tractable posteriors.

* IMPORTANT special case: If \mathbb{F} is the exponential family of distributions, then we have natural conjugate priors. Consider the exponential family

$$f_{X|\Theta}(x|\theta) = h(x)q(\theta) \exp[\eta^T(\theta)T(x)]. \quad (2)$$

Earlier, we used $q(\theta) = \exp[-B(\theta)]$ when defining the exponential family.

For conditionally independent, identically distributed (i.i.d.) $\mathbf{x} = (x[n])_{n=0}^{N-1}$ given $\Theta = \theta$, the likelihood function is also a member of the exponential family:

$$f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \left[\prod_{n=0}^{N-1} h(x[n]) \right] q^N(\theta) \exp[\eta^T(\theta)T(\mathbf{x})] \quad q^N(\theta) = [q(\theta)]^N.$$

with a natural sufficient statistic

$$T(\mathbf{x}) = \sum_{n=0}^{N-1} T(x[n]).$$

Consider the following family \mathbb{P} of prior pdfs/pmfs:

$$f_{\Theta}(\theta) \propto q^{\xi}(\theta) \exp[\eta^T(\theta)\mathbf{v}]. \quad (3)$$

Then, the posterior pdf/pmf is

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto q^{N+\xi}(\theta) \exp\{\eta^T(\theta)[T(\mathbf{x}) + \mathbf{v}]\}$$

and hence $f_{\Theta}(\theta)$ is indeed the conjugate prior for $f_{X|\Theta}(x|\theta)$.

* EXAMPLE. Consider the following binomial data model:

$$\begin{aligned} p_{X|\Theta}(x|\theta) &= \text{Bin}(x|N, \theta) \\ &= \binom{N}{x} \theta^x (1-\theta)^{N-x} \mathbb{1}_{(0,1)}(\theta) \\ &= \underbrace{\binom{N}{x}}_{h(x)} \underbrace{[(1-\theta) \mathbb{1}_{(0,1)}(\theta)]^N}_{q(\theta)} \exp\left[\underbrace{\ln\left(\frac{\theta}{1-\theta}\right)}_{\eta(\theta)} \underbrace{x}_{t(x)}\right] \end{aligned} \quad \text{exponential family}$$

where N is the number of trials (for example, the number of coin flips) and $\theta \in [0, 1]$ is the parameter (for example, the probability of heads).

Conjugate prior family of pdfs for θ follows by using (3):

$$\begin{aligned} f_{\Theta}(\theta) &\propto \underbrace{(1-\theta)^{N\xi} \mathbb{1}_{(0,1)}(\theta)}_{q^{\xi}(\theta)} \underbrace{\exp\left[\ln\left(\frac{\theta}{1-\theta}\right)\nu\right]}_{\exp[\eta(\theta)\nu]} \\ &= (1-\theta)^{N\xi-\nu} \theta^{\nu} \mathbb{1}_{(0,1)}(\theta). \end{aligned}$$

By looking at the table of distributions, we recognize the kernel of the beta pdf, traditionally parameterized as follows:

$$\begin{aligned} f_{\Theta}(\theta) &= \text{Beta}(\theta \mid \alpha, \beta) \\ &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \mathbb{1}_{(0,1)}(\theta) \end{aligned}$$

where $\alpha > 0$ and $\beta > 0$. We measure x coming from

$$\begin{aligned} p_{X|\Theta}(x \mid \theta) &= \text{Bin}(x \mid N, \theta) \\ &= \binom{N}{x} \theta^x (1-\theta)^{N-x} \mathbb{1}_{(0,1)}(\theta) \end{aligned}$$

and wish to infer about θ . Set the conjugate prior on θ :

$$\begin{aligned} f_{\Theta}(\theta) &= \text{Beta}(\theta \mid \alpha, \beta) \\ &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \mathbb{1}_{(0,1)}(\theta). \end{aligned}$$

Now, the posterior pdf for θ is

$$\begin{aligned} f_{\Theta|X}(\theta \mid x) &\propto p_{X|\Theta}(x \mid \theta) f_{\Theta}(\theta) \\ &\propto \theta^{x+\alpha-1} (1-\theta)^{N-x+\beta-1} \mathbb{1}_{(0,1)}(\theta) \end{aligned}$$

which is the kernel of the $\text{Beta}(x + \alpha, \beta + N - x)$ pdf; therefore,

$$f_{\Theta|X}(\theta \mid x) = \text{Beta}(\theta \mid x + \alpha, \beta + N - x).$$

Laplace's problem

LAPLACE computed posterior probabilities under a special case of this model. In particular, he considered a single observation x , the number of girls born in Paris over a time period in the 18th century, coming from

$$\{X \mid \Theta = \theta\} \sim \text{Bin}(N, \theta)$$

and set the following prior pdf:

$$\begin{aligned} f_{\Theta}(\theta) &= \text{U}(\theta \mid 0, 1) \\ &= \text{Beta}(\theta \mid 1, 1). \end{aligned}$$

Here is the measurement:

$$x = 241\,945$$

and $N = 241\,945 + 251\,527$. Laplace computed

$$\Pr_{\Theta|X}(\Theta \geq 0.5 \mid X = x) \approx 10^{-42}.$$

prob. that a newborn child is a girl

Bayes' problem

BAYES sought the probability

See (Gelman et al. 2014, §2.1).

$$\Pr_{\Theta|X}\{\Theta \in (\theta_1, \theta_2) \mid x\}.$$

His solution was based on a physical analogy of a probability space to a rectangular (e.g., a billiard) table:

- (Prior distribution) A ball W is randomly thrown (according to a uniform distribution on the table). The horizontal position of the ball on the table is θ , expressed as a fraction of the table width.
- (Likelihood) A ball O is randomly thrown N times. The value of x is the number of times O lands to the right of W.

Thus, Θ is assumed to have a (prior) uniform distribution on $[0, 1]$:

$f(\theta) = U(\theta \mid 0, 1)$. Bayes then obtained

$$\begin{aligned} \Pr_{\Theta|X}\{\Theta \in (\theta_1, \theta_2) \mid x\} &= \frac{\Pr\{\Theta \in (\theta_1, \theta_2), X = x\}}{p(x)} \\ &= \frac{\int_{\theta_1}^{\theta_2} p(x \mid \theta) f(\theta) d\theta}{p(x)} \\ &= \frac{\int_{\theta_1}^{\theta_2} \binom{N}{x} \theta^x (1 - \theta)^{N-x} d\theta}{p(x)} \end{aligned}$$

and successfully evaluated the denominator

$$\begin{aligned} p(x) &= \int_0^1 \binom{N}{x} \theta^x (1 - \theta)^{N-x} d\theta && \text{for } x = 0, 1, \dots, N \\ &= \frac{1}{N + 1} \end{aligned}$$

i.e., all possible values of x are equally likely *a priori*.

Conjugate versus Non-conjugate Priors (Binomial Example)

SUPPOSE we measure x coming from

$$p_{X|\Theta}(x \mid \theta) = \text{Bin}(x \mid N, \theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x} \mathbb{1}_{(0,1)}(\theta)$$

and wish to infer about θ .

Any density function defined on $[0, 1]$ would also be a valid prior for Θ . Bayes and Laplace used uniform $\Theta \sim U(0, 1)$. Here are some possible choices:

1. Beta(α, β), see previous example;
2. truncated Gaussian:

$$f_{\Theta}(\theta) \propto \mathcal{N}(\theta \mid \mu, \sigma^2) \mathbb{1}_{(0,1)}(\theta);$$

3. Mixture:

$$f_{\Theta}(\theta) = c \text{Beta}(\theta \mid \alpha_1, \beta_1) + (1 - c) \text{Beta}(\theta \mid \alpha_2, \beta_2)$$

where $c \in [0, 1]$.

* THE posteriors for the above priors.

1. For the prior Beta(α, β), the posterior is Beta($x + \alpha, \beta + N - x$), see previous example.
2. Truncated Gaussian:

$$f(\theta \mid x) \propto \mathcal{N}(\theta \mid \mu, \sigma^2) \theta^x (1 - \theta)^{N-x} \mathbb{1}_{(0,1)}(\theta). \quad (4)$$

This posterior pdf *does not* have a nice form.

3. Mixture:

$$\begin{aligned} f(\theta \mid x) &\propto [c \text{Beta}(\theta \mid \alpha_1, \beta_1) + (1 - c) \text{Beta}(\theta \mid \alpha_2, \beta_2)] \theta^x (1 - \theta)^{N-x} \mathbb{1}_{(0,1)}(\theta) \\ &\propto c \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \theta^{\alpha_1+x-1} (1 - \theta)^{\beta_1+N-x-1} \mathbb{1}_{(0,1)}(\theta) \\ &\quad + (1 - c) \frac{\Gamma(\alpha_2 + \beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} \theta^{\alpha_2+x-1} (1 - \theta)^{\beta_2+N-x-1} \mathbb{1}_{(0,1)}(\theta) \end{aligned}$$

which is a mixture of betas:

$$f(\theta \mid x) = c_1 \text{Beta}(\theta \mid \alpha_1 + x, \beta_1 + N - x) + (1 - c_1) \text{Beta}(\theta \mid \alpha_2 + x, \beta_2 + N - x)$$

where $c_1 \in [0, 1]$. HW: find c_1 .

* COMMENTS:

1. Binomial likelihood \times beta prior = beta posterior; therefore conjugate prior.
2. The prior is clearly not conjugate.
3. Binomial likelihood \times beta mixture prior = beta mixture posterior.

Advantages and disadvantages of conjugate priors

ADVANTAGES:

- Easy to deal with mathematically and computationally.
- Interpretable as additional data; e.g., a $\text{Beta}(a, b)$ prior in the binomial success problem can be thought to be equivalent to seeing a data set earlier that had a successes and b failures.

DISADVANTAGE: Can be overly restrictive, some prior beliefs cannot be described.

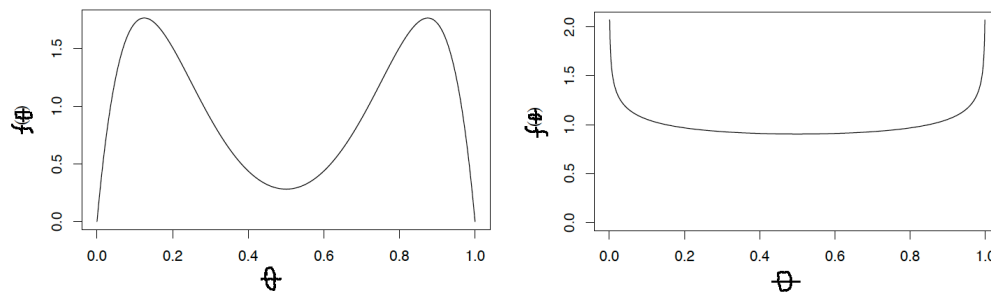


Figure 1: (Left) A non-beta prior and (right) a beta prior.

- * **EXAMPLE.** Suppose you wanted to use a beta prior with binomial data, but your prior beliefs matched Fig. 1(left). There is no beta pdf that matches the density in Fig. 1(left). About the closest you could get would be the density in Fig. 1(right), which does not get the bimodal nature right.

Whether or not a conjugate prior exists depends on the form of the likelihood function. Most cases do not have conjugate distributions.

About the only case where conjugate priors are guaranteed to exist is when the data distribution is a member of the exponential family. Exponential family includes many important distributions, such as Gaussian, binomial, multinomial, Poisson, gamma, and beta.

Summarizing Posterior Distributions

WE often need summaries of the posterior distribution.

In a univariate (one-parameter) problem, such as for the binomial success probability or for the normal mean (with known variance) a plot of the posterior density is useful.

If the prior is conjugate, the plot is easy. If the prior is non-conjugate, the plot is *not* much harder.

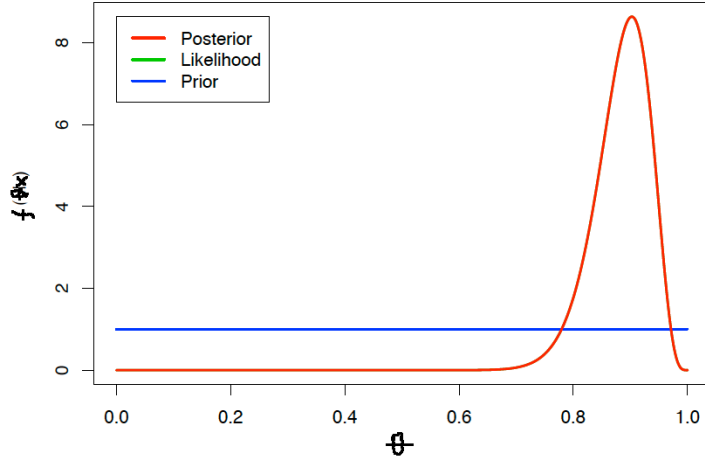


Figure 2: Prior $f_{\Theta}(\theta) = U(\theta | 0, 1) = \text{Beta}(\theta | 1, 1)$, likelihood, and posterior.

Binomial example

CONSIDER a binomial experiment with $N = 41$ trials and $x = 37$ successes.

- * CONJUGATE beta priors. For the prior $f_{\Theta}(\theta) = U(\theta | 0, 1) = \text{Beta}(\theta | 1, 1)$, the posterior

$$f_{\Theta|X}(\theta | x) = \text{Beta}(\theta | 38, 5)$$

is shown in Fig. 2. For the prior $f_{\Theta}(\theta) = \text{Beta}(\theta | 3, 3)$, the posterior

$$f_{\Theta|X}(\theta | x) = \text{Beta}(\theta | 40, 7)$$

is shown in Fig. 3.

- * TRUNCATED Gaussian prior. For the prior $f_{\Theta}(\theta) \propto \mathcal{N}(\theta | 0.5, 0.1^2) \mathbb{1}_{(0,1)}(\theta)$, the posterior computed using (4) is shown in Fig. 4.
- * BETA mixture prior. For the prior $f_{\Theta}(\theta) = 0.5 \text{Beta}(\theta | 8, 2) + 0.5 \text{Beta}(\theta | 2, 8)$, the posterior is

$$f_{\Theta|X}(\theta | x) = 0.5 \text{Beta}(\theta | 45, 6) + 0.5 \text{Beta}(\theta | 39, 12)$$

shown in Fig. 5.

We can obtain the posterior plots easily:

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

where

$$\begin{aligned} f_X(x) &= \int f_{\Theta}(\theta) f_{X|\Theta}(x | \theta) d\theta \\ &\approx \sum_{i=1}^m f_{\Theta}(\theta_0 + i \Delta\theta) f_{X|\Theta}(x | \theta_0 + i \Delta\theta) \Delta\theta = c \end{aligned}$$

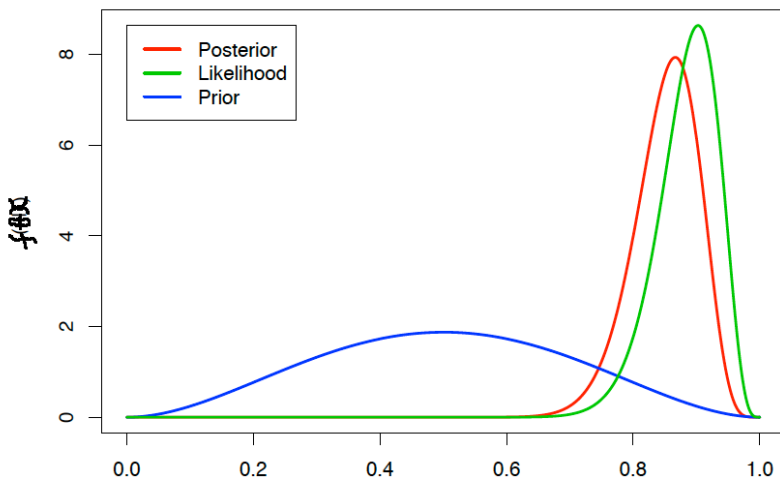


Figure 3: Prior $f_{\Theta}(\theta) = \text{Beta}(\theta | 3, 3)$, likelihood, and posterior.

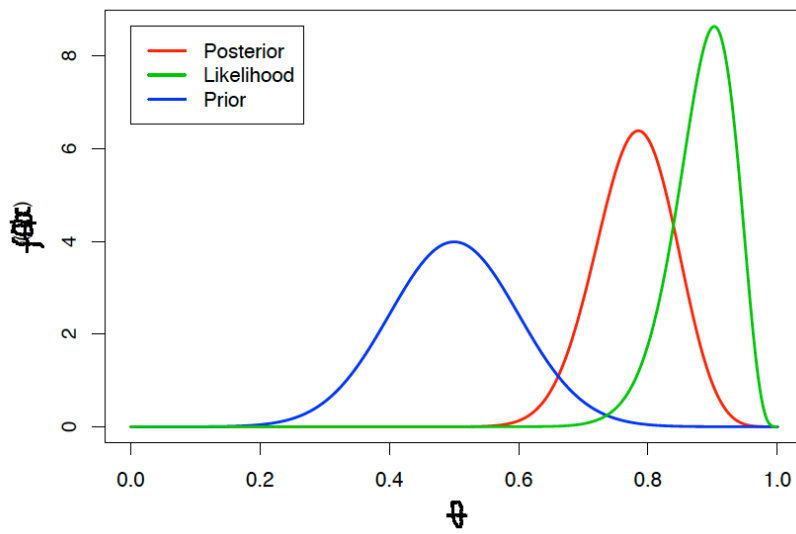


Figure 4: Prior $f_{\Theta}(\theta) \propto \mathcal{N}(\theta | 0.5, 0.1^2) \mathbb{1}_{(0,1)}(\theta)$, likelihood, and posterior.

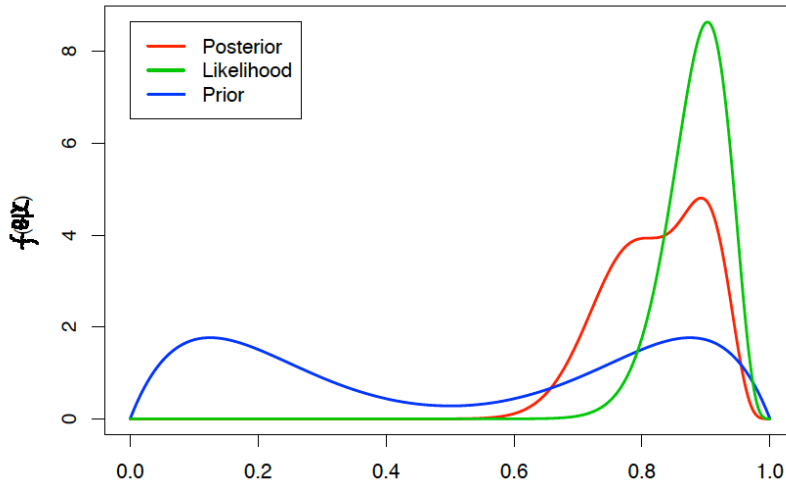


Figure 5: Prior

$f_{\Theta}(\theta) = 0.5 \text{Beta}(\theta | 8, 2) + 0.5 \text{Beta}(\theta | 2, 8)$
likelihood, and posterior.

i.e., calculate the unnormalized density at m equally spaced points on the interval $[\theta_0, \theta_1]$, where $\Delta\theta = (\theta_1 - \theta_0)/m$. Then

$$f(\theta | x) \approx \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{c}.$$

Note that this approximation will work well when θ is small and m is large. What is classified as a small $\Delta\theta$ and a large m depends on how smooth $f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)$ is.

Other numerical quadrature methods (e.g., Simpson's Rule) could be used to calculate the normalization constant of the posterior.

This adjustment is only needed if we wish compare this posterior distribution to another distribution (e.g., the prior). To see the shape of the posterior, plot $f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)$ as a function of θ . The scaling adjustment just relabels the y axis on the plot.

Numerical summaries of the posterior distribution

WHEN summarizing a posterior distribution we are usually interested in two quantities, location and spread.

* MEASURES of location. Three common measures of location are used:

- posterior mean: $E_{\Theta|X}(\Theta | \mathbf{x})$,
- posterior median $\text{median}(\theta | \mathbf{x})$ defined by

$$\int_{-\infty}^{\text{median}(\theta | \mathbf{x})} f_{\Theta|X}(\theta | \mathbf{x}) d\theta = \int_{\text{median}(\theta | \mathbf{x})}^{+\infty} f_{\Theta|X}(\theta | \mathbf{x}) d\theta,$$

- posterior mode: $\arg \max_{\theta} f_{\Theta|X}(\theta | \mathbf{x})$.

Apart from being common summaries of location, these choices can also be justified in the decision theory framework, which we will discuss later.

- * MEASURES of spread. The common choices for measuring spread of the posterior distribution are the posterior variance $\text{var}_{\Theta|X}(\Theta | \mathbf{x})$ and standard deviation $\sqrt{\text{var}_{\Theta|X}(\Theta | \mathbf{x})}$. Other possibilities, though less common, are posterior interquartile range and mean absolute deviation: $E_{\Theta|X}[|\Theta - E_{\Theta|X}(\Theta | \mathbf{x})| | \mathbf{x}]$.

Prior versus posterior distribution

THE posterior distribution can be seen as a compromise between the prior and the data, which can be seen from the two well-known relationships reviewed in handout revprob:

$$E_{\Theta}(\Theta) = E_X[E_{\Theta|X}(\Theta | X)] \quad (4a)$$

$$\text{var}_{\Theta}(\Theta) = E_X[\text{var}_{\Theta|X}(\Theta | X)] + \text{var}_X[E_{\Theta|X}(\Theta | X)] \quad (4b)$$

where (4a) and (4b) follow from (10) and (11) in handout revprob.

Here,

- (4a) states that our prior mean is the average of all possible posterior means (averaged over all possible data sets).
- (4b) states that the posterior variance is, on average, smaller than the prior variance: The size of the difference depends on the variability of the posterior means.

Credible sets

WE can make interval inferences based on the posterior distributions and construct *credible sets*, also known as *Bayesian confidence intervals*.

Consider a subset A of the parameter space for θ . Then, A is a 100c % credible set for θ if

$$\Pr_{\Theta|X}(\theta \in A | \mathbf{x}) = c.$$

The most common approach to credible intervals (scalar credible sets) are *central credible intervals*. A central credible interval $[\tau_L, \tau_U]$ satisfies

$$\frac{1-c}{2} = \int_{-\infty}^{\tau_L} f_{\Theta|X}(\theta | \mathbf{x}) d\theta, \quad \frac{1-c}{2} = \int_{\tau_U}^{\infty} f_{\Theta|X}(\theta | \mathbf{x}) d\theta$$

where τ_L and τ_U are the $0.5(1-c)$ and $1-0.5(1-c)$ quantiles of the posterior pdf, see Fig. 6. An alternative is the 100c % highest posterior

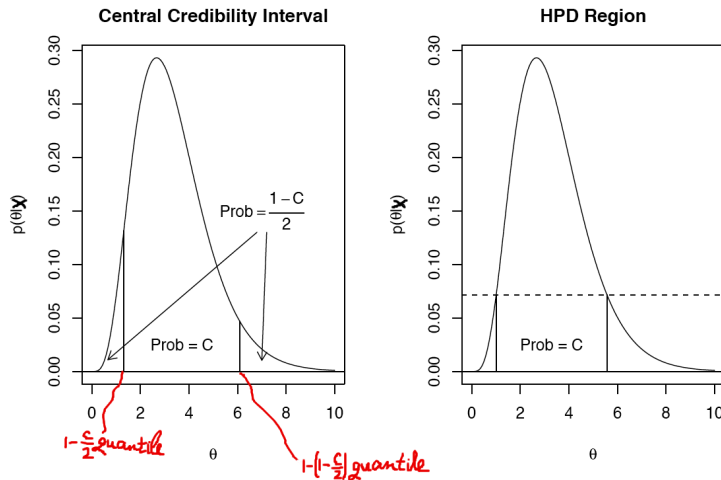


Figure 6: Central interval: (1.313, 6.102) with length 4.789; HPD interval: (1.006, 5.571) with length 4.565.

density (HPD) region, which is defined as the smallest region of the parameter space with probability c .

The central interval is usually easier to determine than the HPD interval because it only involves finding quantiles of the posterior distribution.

To find the HPD region, we should be able to determine

- the superlevel set $A_q = \{\theta \mid f_{\Theta|X}(\theta | \mathbf{x}) \geq q\}$ and
- the corresponding probability $\Pr_{\Theta|X}(\Theta \in A_q | \mathbf{x})$.

The HPD region is then given by the q for which

$$\Pr_{\Theta|X}(\Theta \in A_q | \mathbf{x}) = c.$$

In the example graph, it was not too bad since the posterior is unimodal. However, if the posterior is multimodal, A_q may be a bunch of disjoint sets, see, e.g., Fig. 7.

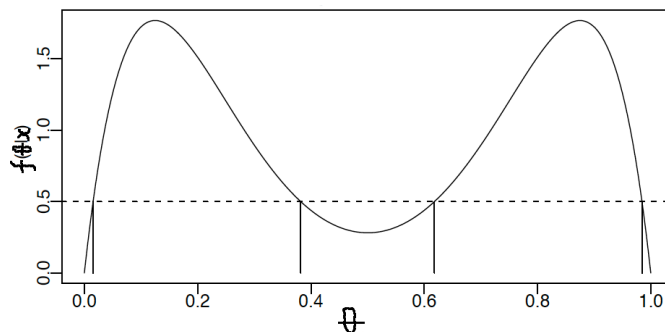


Figure 7: Disjoint HPD interval.

Handling multiple modes is difficult, particularly in large-scale problems where we may not be able to identify all the modes and find

probabilities $\Pr_{\Theta|X}(\theta \in A_q | \mathbf{x}) = c$. Hence, HPD regions are rarely used.

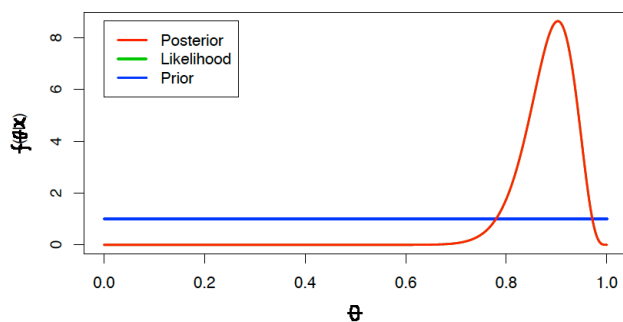


Figure 8: The 95 % central credible set for the example in Fig. 2.

Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	???	0.289	(0.025, 0.975)
Posterior	0.884	0.902	0.048	(0.774, 0.960)

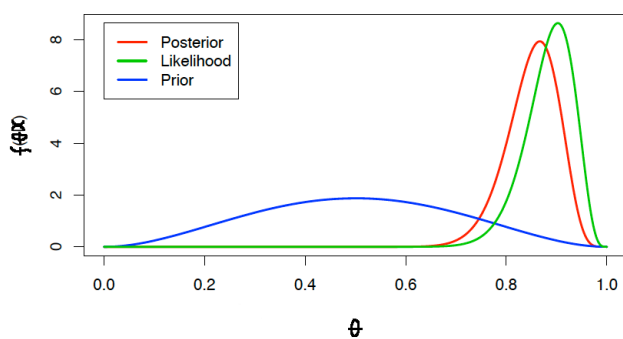


Figure 9: The 95 % central credible set for the example in Fig. 3.

Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	0.5	0.189	(0.147, 0.853)
Posterior	0.851	0.867	0.051	(0.737, 0.937)

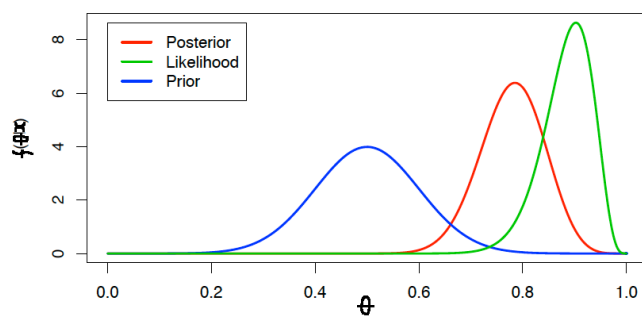


Figure 10: The 95 % central credible set for the example in Fig. 4.

Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	0.5	0.1	(0.304, 0.696)
Posterior	???	???	???	(???, ???)

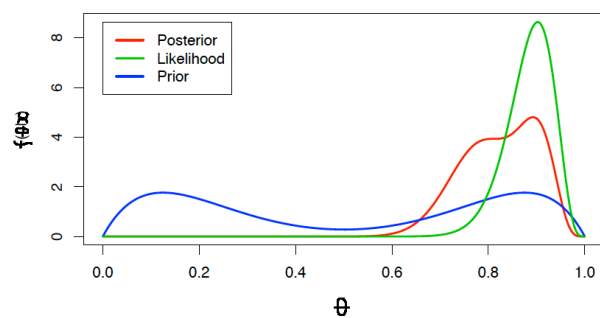


Figure 11: The 95 % central credible set for the example in Fig. 5.

Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	??? & ???	???	(???, ???)
Posterior	0.823	???	???	(???, ???)

Sufficiency and Bayesian Models

WE discuss sufficiency for Bayesian models.

Theorem 1 (Kolmogorov). If a statistic $T(\mathbf{X})$ is sufficient for a parameter θ , then, for any prior $f_{\Theta}(\theta)$,

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

 **THEOREM 1** implies that, given $T(X)$, X and Θ are independent.

DC Level Estimation in AWGN with Known Variance

Reading: (Gelman et al. 2014, §2.5)

Single observation

CHOOSE the measurement model:

$$f_{X|\Theta}(x | \theta) = \mathcal{N}(x | \theta, \sigma^2)$$

where we assume that σ^2 is known. Hence, the likelihood for one measurement is

$$f_{X|\Theta}(x | \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x - \theta)^2\right]. \quad (5)$$

Now,

$$\begin{aligned} f_{X|\Theta}(x | \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\underbrace{\theta^2 - 2x\theta + x^2}_{\text{quadratic in } \theta})\right] \\ &= h(x) \exp\left(-\frac{1}{2\sigma^2}\theta^2 + \frac{x\theta}{\sigma^2}\right) \end{aligned}$$

which fits the exponential-family form in (2) with

$$\begin{aligned} q(\theta) &= \exp\left(-\frac{1}{2\sigma^2}\theta^2\right) \\ t(x) &= x \\ \eta(\theta) &= \frac{\theta}{\sigma^2}. \end{aligned}$$

The conjugate prior pdf for θ has the following form (see (3)):

$$f_{\Theta}(\theta) \propto \underbrace{\exp\left(-\frac{\xi}{2\sigma^2}\theta^2\right)}_{[q(\theta)]^\xi} \underbrace{\exp\left(\frac{\theta}{\sigma^2}v\right)}_{\exp[\eta(\theta)v]}$$

which can be reparameterized as

$$f_{\Theta}(\theta) \propto \exp\left[-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right].$$

complete the squares

Therefore, the conjugate prior pdf for the likelihood function in (5) is

$$f_{\Theta}(\theta) = \mathcal{N}(\theta \mid \mu_0, \tau_0^2)$$

where μ_0 and τ_0^2 are *known hyperparameters*.¹ We now compute the posterior pdf by collecting the terms that contain θ and θ^2 :

¹ We can continue and assign a prior joint pdf for the hyperparameters, which would lead to a hierarchical Bayesian model.

$$\begin{aligned} f_{\Theta|X}(\theta \mid x) &\propto f_{X|\Theta}(x \mid \theta) f_{\Theta}(\theta) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{x^2 - 2x\theta + \theta^2}{\sigma^2} + \frac{\theta^2 - 2\mu_0\theta + \mu_0^2}{\tau_0^2}\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}\right)\theta^2 + \left(\frac{x}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right)\theta\right] \end{aligned}$$


which implies that $f_{\Theta|X}(\theta \mid x)$ is a Gaussian pdf with mean μ_1 and variance τ_1^2 , where

$$\frac{1}{\tau_1^2} = \frac{1}{\sigma^2} + \frac{1}{\tau_0^2} \quad (6a)$$

$$\frac{\mu_1}{\tau_1^2} = \frac{x}{\sigma^2} + \frac{\mu_0}{\tau_0^2}. \quad (6b)$$

see the kernel of the Gaussian pdf in (6)

We will generalize the above expressions to multiple measurements.

 RECALL that $z \sim \mathcal{N}(\mu, \sigma^2)$ implies

$$f_Z(z) \propto \exp\left(-0.5\frac{1}{\sigma^2}z^2 + \frac{1}{\sigma^2}\mu z\right).$$

* COMMENTS:

- The posterior mean, obtained by solving (6), is a weighted average of the observation and the prior mean:

$$\begin{aligned} \mu_1 &= \mu_1(x) = \frac{(1/\sigma^2)x + (1/\tau_0^2)\mu_0}{(1/\sigma^2) + (1/\tau_0^2)} \\ &= \frac{\text{likelihood precision } x + \text{prior precision } \mu_0}{\text{likelihood precision} + \text{prior precision}}. \end{aligned}$$

- We will show that the posterior mean is the Bayesian minimum mean-square error (MMSE) estimate of θ .
- Here, the weights are given by the *precisions*² $1/\sigma^2$ and $1/\tau_0^2$.

² The inverse of variance is called precision.

- As the likelihood precision $1/\sigma^2$ increases, we have

$$\mu_1(x) \rightarrow x.$$

- As the prior precision $1/\tau_0^2$ increases, we have

$$\mu_1(x) \rightarrow \mu_0.$$

- The posterior mean is the measurement x *shifted*³ toward the prior mean:

$$\mu_1(x) = x - \frac{\sigma^2}{\sigma^2 + \tau_0^2}(x - \mu_0)$$

or the prior mean adjusted toward the measurement x :

$$\mu_1(x) = \mu_0 + \frac{\tau_0^2}{\sigma^2 + \tau_0^2}(x - \mu_0).$$

³ When the prior mean is zero, the posterior mean is the measurement x *shrunk*, i.e., has reduced magnitude.

- Posterior precision is the sum of the prior and likelihood precisions:

$$\frac{1}{\tau_1^2} = \frac{\sigma^2 + \tau_0^2}{\sigma^2 \tau_0^2} = \frac{1}{\sigma^2} + \frac{1}{\tau_0^2}.$$

Multiple observations

EXAMPLE 10.2 in the textbook. Consider now N conditionally i.i.d. observations $(X[n])_{n=0}^{N-1}$ given θ :

$$\begin{aligned} f_{\Theta|X}(\theta | \mathbf{x}) &\propto f_{\Theta}(\theta) f_{X|\Theta}(\mathbf{x} | \theta) \\ &\propto \exp\left[-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right] \prod_{n=0}^{N-1} \exp\left[-\frac{1}{2\sigma^2}(x[n] - \theta)^2\right] \end{aligned}$$

where $\mathbf{x} = (x[n])_{n=0}^{N-1}$. This posterior pdf depends on \mathbf{x} only through the sample mean

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

i.e., \bar{x} is the sufficient statistic for θ in this model. Note that

$$\{\bar{X} | \Theta = \theta\} \sim \mathcal{N}(\theta, \sigma^2/N).$$

new likelihood using sufficiency

By invoking sufficiency, we reduce our problem to the single-observation case, where \bar{x} is our equivalent single observation. Hence,

$$\begin{aligned} f_{\Theta|X}(\theta | \mathbf{x}) &= f_{\Theta|\bar{X}}(\theta | \bar{x}) \\ &\propto f_{\Theta}(\theta) f_{\bar{X}|\Theta}(\bar{x} | \theta) \\ &= \mathcal{N}(\theta | \mu_N(\bar{x}), \tau_N^2) \end{aligned} \tag{7a}$$

sufficiency

with

$$\mu_N(\bar{x}) = \frac{(N/\sigma^2)\bar{x} + (1/\tau_0^2)\mu_0}{N/\sigma^2 + 1/\tau_0^2}, \quad \frac{1}{\tau_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau_0^2}. \quad (7b)$$

👉 COMMENTS:

- If N is large, the influence of the prior pdf disappears and the posterior pdf effectively depends only on \bar{x} and σ^2 .
- If $\tau_0^2 = \sigma^2$, the prior has the same weight as adding one more observation with value μ_0 .
- When $\tau_0^2 \nearrow +\infty$ with N fixed or $N \nearrow +\infty$ with τ_0 fixed, we have

$$f_{\Theta|\bar{X}}(\theta | \bar{x}) \rightarrow \mathcal{N}(\theta | \bar{x}, \sigma^2/N) \quad (8)$$

which is a good general approximation whenever our prior knowledge about θ is vague or the number of observations N is large. In this scenario, the influence of the prior disappears. Furthermore, $\tau_0^2 \nearrow +\infty$ corresponds to

$$f_{\Theta}(\theta) \propto 1 \quad (9)$$

and leads to the posterior pdf proportional to the likelihood:

$$\begin{aligned} f_{\Theta|\mathbf{X}}(\theta | \mathbf{x}) &= f_{\Theta|\bar{X}}(\theta | \bar{x}) \\ &\propto \underbrace{f_{\Theta}(\theta)}_{\propto 1, \text{ see (9)}} \underbrace{f_{\bar{X}|\Theta}(\bar{x} | \theta)}_{\text{likelihood}} \\ &\propto \underbrace{f_{\bar{X}|\Theta}(\bar{x} | \theta)}_{\text{likelihood}}. \end{aligned}$$

sufficiency

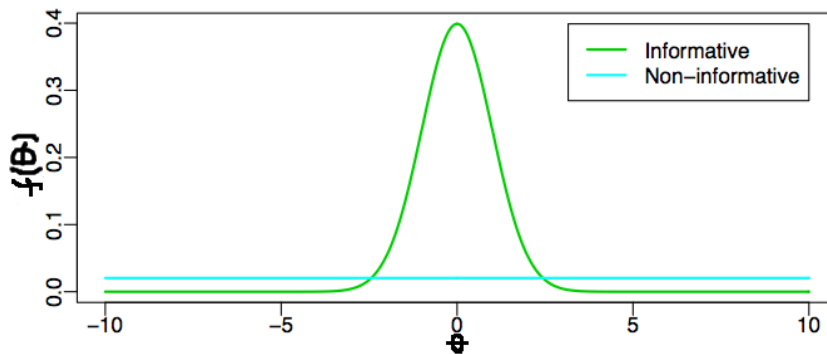


Figure 12: Informative and noninformative priors.

Since

$$\int_{-\infty}^{+\infty} 1 = +\infty$$

the prior (9) is improper and does not describe a valid probability density. However, we can still use it because the posterior pdf in (8) is proper, see Fig. 13.

See handout noninfpriors for more on noninformative priors.

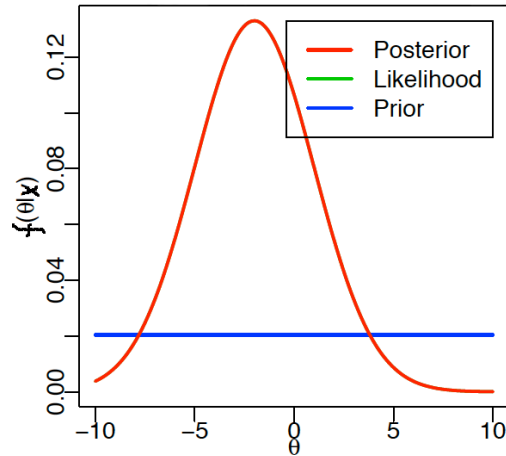


Figure 13: For the noninformative prior in (9), the posterior pdf is proportional to the likelihood function.

Proper versus Improper Priors

A prior $f_{\Theta}(\theta)$ is called *proper* if it is a valid probability distribution:

$$f_{\Theta}(\theta) \geq 0 \quad \forall \theta, \quad \int_{\text{sp}_{\Theta}} f_{\Theta}(\theta) d\theta = 1. \quad (10)$$

A prior $f_{\Theta}(\theta)$ is called *improper* if

$$f_{\Theta}(\theta) \geq 0 \quad \forall \theta, \quad \int f_{\Theta}(\theta) d\theta = +\infty. \quad (11)$$

If a prior is proper, so is the posterior

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

If a prior is improper, the posterior may or may not be proper. For many common problems, popular improper noninformative priors⁴ lead to proper posteriors, assuming enough data have been collected. But, this has to be checked.

⁴ e.g., Jeffreys' priors, to be discussed in `handout noninfpriors`

☞ THE posterior distribution *must* be proper.

Gaussian Distribution with Unknown Variance and Known Mean

Reading: (Gelman et al. 2014, §2.6)

DATA model:

$$f(x | \sigma^2) = \mathcal{N}(x | \mu, \sigma^2)$$

where $\sigma^2 \geq 0$ is now the parameter of interest and μ is a known constant.

* **EXAMPLE.** We now find the conjugate prior family of pdfs for σ^2 under this model. First, write $f(x | \sigma^2)$ explicitly:

$$f(x | \sigma^2) = \exp \left[\underbrace{-\frac{1}{2\sigma^2}}_{\eta(\sigma^2)} \underbrace{(x - \mu)^2}_{t(x)} \right] \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \mathbb{1}_{[0,+\infty)}(\sigma^2)}_{q(\sigma^2)} .$$

A conjugate prior family of pdfs for σ^2 follows from (3):

$$f_{\sigma^2}(\sigma^2) \propto \underbrace{(2\pi\sigma^2)^{-\xi/2} \mathbb{1}_{[0,+\infty)}(\sigma^2)}_{[q(\sigma^2)]^\xi} \underbrace{\exp\left(-\frac{1}{2\sigma^2}v\right)}_{\exp[\eta(\sigma^2)v]} .$$

How does this pdf *look like*? By looking up the table of distributions we see that it is an inverse-gamma pdf:

$$f_{\sigma^2}(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \mathbb{1}_{[0,+\infty)}(\sigma^2)$$

where α and β are known hyperparameters.⁵

For ease of interpretation, use an equivalent prior pdf: A *scaled inverted χ^2 distribution* with scale σ_0^2 and ν_0 degrees of freedom; here σ_0^2 and ν_0 are the known hyperparameters. In other words, we take the prior distribution of σ^2 to be the distribution of


$$\frac{\sigma_0^2 \nu_0}{X} \quad (12)$$

where X is a $\chi_{\nu_0}^2$ random variable (see the underlined part of the distribution handout). We use the following notation for this distribution:

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

or

$$\begin{aligned} f(\sigma^2) &= \text{Inv-}\chi^2(\sigma^2 | \nu_0, \sigma_0^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right) \mathbb{1}_{[0,+\infty)}(\sigma^2). \end{aligned}$$

 **NOTE:** From the table of distributions, we also obtain the following facts:

- the mean of $f(\sigma^2)$ is

$$\mathbb{E}(\sigma^2) = \frac{\sigma_0^2 \nu_0}{\nu_0 - 2} \quad (13)$$

and

- when ν_0 is large, the variance behaves like $(\sigma_0^2)^2/\nu_0$, implying that large ν_0 yields high precision.

⁵ Example 10.3 in the textbook uses this distribution as a prior distribution for the variance parameter.

Estimating the variance of a Gaussian distribution with known mean

FOR i.i.d. $(X[n])_{n=0}^{N-1}$ given σ^2 , the likelihood function is in the exponential-family form:

$$\begin{aligned} f(\mathbf{x} | \sigma^2) &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu)^2\right\} \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{NT(\mathbf{x})}{2\sigma^2}\right] \end{aligned}$$

where

$$T(\mathbf{x}) \triangleq \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu)^2$$

is the natural sufficient statistic. Choose the conjugate prior pdf, i.e., the scaled inverted χ^2 distribution:

$$\begin{aligned} f(\sigma^2) &= \text{Inv-}\chi^2(\sigma^2 | \nu_0, \sigma_0^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \mathbb{1}_{[0,+\infty)}(\sigma^2). \end{aligned}$$

Now,

$$\begin{aligned} f(\sigma^2 | \mathbf{x}) &\propto f(\sigma^2) f(\mathbf{x} | \sigma^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) (\sigma^2)^{-N/2} \exp\left[-\frac{NT(\mathbf{x})}{2\sigma^2}\right] \mathbb{1}_{[0,+\infty)}(\sigma^2) \\ &\propto (\sigma^2)^{-(\nu_N/2+1)} \exp\left(-\frac{\nu_N\sigma_N^2}{2\sigma^2}\right) \mathbb{1}_{[0,+\infty)}(\sigma^2) \end{aligned}$$

with

$$\nu_N = \nu_0 + N$$

and

$$\sigma_N^2 = \sigma_N^2(\mathbf{x}) = \frac{NT(\mathbf{x}) + \nu_0\sigma_0^2}{N + \nu_0}.$$

Therefore, $f(\sigma^2 | \mathbf{x})$ is also a scaled inverted χ^2 distribution. Now, the posterior mean⁶ is

⁶ and the MMSE estimate of σ^2 , to be shown later

$$\begin{aligned} E(\sigma^2 | X = \mathbf{x}) &= \frac{\sigma_N^2 \nu_N}{\nu_N - 2} \\ &= \frac{NT(\mathbf{x}) + \nu_0\sigma_0^2}{N + \nu_0 - 2} \end{aligned}$$

obtained by using (13), but now for the posterior pdf.

* COMMENTS:

- The MMSE estimate of σ^2 is a weighted average of the prior guess and a data based estimate:

$$E(\sigma^2 | X = \mathbf{x}) = \frac{NT(\mathbf{x}) + \nu_0\sigma_0^2}{N + \nu_0 - 2}$$

where the weights are obtained using the prior and sample degrees of freedom.

- Interpretation of the prior information: the chosen prior provides information equivalent to v_0 observations with average variance equal to σ_0^2 .
- As $N \nearrow +\infty$, $\sigma_N^2 \rightarrow T(\mathbf{x})$ and

$$E(\sigma^2 | \mathbf{X} = \mathbf{x}) \rightarrow T(\mathbf{x}).$$

- As $v_0 \nearrow +\infty$, $\sigma_N^2 \rightarrow \sigma_0^2$ and

$$E(\sigma^2 | \mathbf{X} = \mathbf{x}) \rightarrow \sigma_0^2.$$

Acronyms

HPD highest posterior density. 12–14

i.i.d. independent, identically distributed. 4, 18, 22

MMSE minimum mean-square error. 17, 22

pdf probability density function. 1, 3–5, 7, 12, 17–19, 21, 22

pmf probability mass function. 1, 4

References

Gelman, A., J. B. Carlin, H. S. Stern, David B. Dunson, Aki Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Taylor & Francis (cit. on pp. 1, 6, 16, 20).