PROBABILITY AND STATISTICS A

Bayesian Statistical Inference and Computation

Teruo Nakatsuma

Spring Semester 2023

Faculty of Economics, Keio University

Aims Of This Course

1. Learn basic principles of Bayesian learning.
2. Learn how to conduct statistical inference (point estimation, interval estimation hypothesis testing) in the Bayesian way.
3. Learn basic principles of Markov chain Monte Carlo (MCMC) methods.
4. Hands-on practice of Python and PyMC.

Reading List i

1. Introduction to Bayesian statistics
   - Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC.
     Download the book here.
   - Hoff, P.D. (2010). *A First Course in Bayesian Statistical Methods*, Springer.
     The Japanese translation is available.
   - Johnson, A.A., Ott, M.Q. and Dogucu, M. (2022). *Bayes Rules!: An Introduction to Applied Bayesian Modeling*, Chapman & Hall/CRC.

Reading List ii

- Robert, C.P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed., Springer.

2. Advanced topics in Bayesian statistics
   - Chan, J., Koop, G., Poirier, D.J. and Tobias, J.L. (2020). *Bayesian Econometric Methods*, 2nd ed., Cambridge University Press.
   - Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press.

Reading List iii

- Prado, R. and West, M. (2021). *Time Series: Modeling, Computation, and Inference*, 2nd ed., Chapman & Hall/CRC.
- Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed., Springer.
- Rossi, P.E., Allenby, G.E. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley.

3. Classics

- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley.

Population, Sample, Parameter

1. In statistics, the population is any subject (not necessarily a group) which a researcher tries to analyze.
2. The sample is a collection of data related to the population. In a typical situation, the sample is assumed to be randomly and independently extracted from the population.
3. The parameter represents a property of the population to be analyzed. The parameter is unknown to the researcher.
4. The goal of statistics is to obtain useful insights about the parameter of the population with the sample extracted from it.

Population Distribution

We regard the population as a probability distribution and call it the population distribution. Then we can interpret the sample as a set of random variables following the population distribution, and the parameters as variables which determine the "shape" of the population distribution. Let $D = (x_1, \ldots, x_n)$ denote the sample, and $\boldsymbol{\theta}$ denote the parameter of the population distribution. To indicate that the shape of the population distribution depends on $\boldsymbol{\theta}$, the population p.m.f. or p.d.f. is denoted by $p(x_i|\boldsymbol{\theta})$ where each x_i $(i = 1, \ldots, n)$ is called an observation and supposed to be a realized value of the random variable following the population distribution. n is often called the sample size.

Example: Bernoulli Distribution

Consider a random variable X_i $(i = 1, \ldots, n)$ such that

$$X_i = \begin{cases} 1, & \text{Head is obtained;} \\ 0, & \text{Tail is obtained,} \end{cases}$$

and suppose $\mathbf{P}(X_i = 1) = \theta$ and $\mathbf{P}(X_i = 0) = 1 - \theta$. Then X_i follows the Bernoulli distribution with the p.m.f.:

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i = 0, 1.$$

Likelihood

Suppose the sample $D = (x_1, \ldots, x_n)$ are taken from a population distribution where the parameter $\boldsymbol{\theta}$ is a set of unknown parameters. The joint p.m.f. or the joint p.d.f. of D is denoted by

$$p(D|\boldsymbol{\theta}) = p(x_1, \ldots, x_n|\boldsymbol{\theta}).$$

In particular, if observations are independent of each other,

$$p(D|\boldsymbol{\theta}) = p(x_1|\boldsymbol{\theta}) \times \cdots \times p(x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}).$$

When we regard $p(D|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, it is called the likelihood or likelihood function.

Example: Bernoulli Distribution

Suppose $D = (x_1, \ldots, x_n)$ is independently generated from the same Bernoulli distribution. Then

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y(1-\theta)^{n-y},$$

where $y = \sum_{i=1}^{n} x_i$.

Suppose we have $(x_1, x_2, x_3, x_4, x_5) = (1, 0, 1, 1, 1)$. The value of $p(D|\theta)$ depends on the value of θ.

θ	0.1000	0.2000	0.3000	0.4000	0.5000
$p(D\|\theta)$	0.0001	0.0013	0.0057	0.0154	0.0312
θ	0.6000	0.7000	0.8000	0.9000	
$p(D\|\theta)$	0.0518	0.0720	0.0819	0.0656	

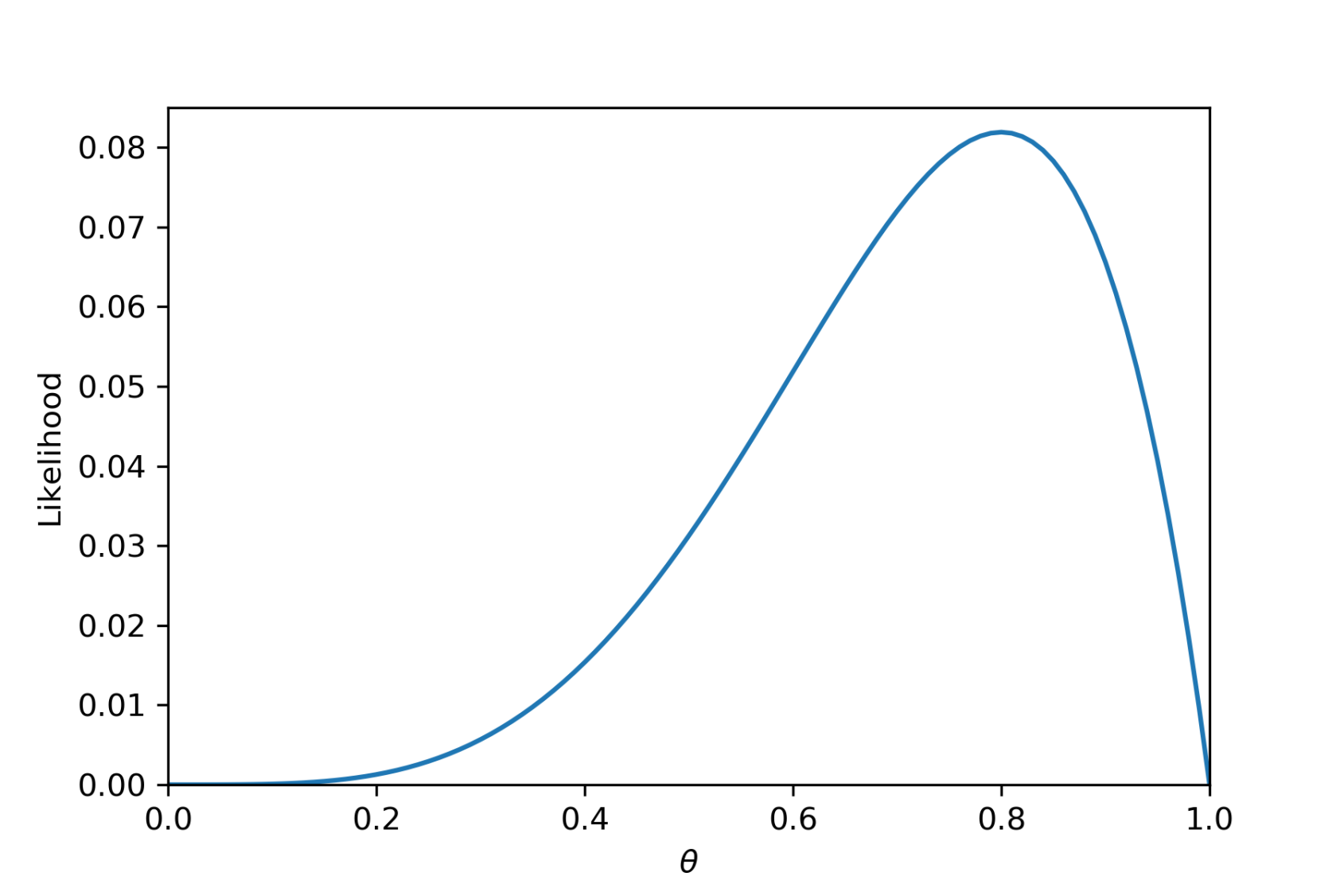

Figure 1: The likelihood of $\boldsymbol{\theta}$ in the Bernoulli distribution

Interpretation Of The Likelihood

Given the sample D, the likelihood $p(D|\boldsymbol{\theta})$ is regarded as a kind of "plausibility" of a specific value of $\boldsymbol{\theta}$.

For example, the likelihood of $\boldsymbol{\theta} = \mathbf{0.9}$ is **0.0656** while that of $\boldsymbol{\theta} = \mathbf{0.4}$ in the previous example is **0.0154**. We may say that **0.9** is about 4 times more plausible than **0.4** as the true value of $\boldsymbol{\theta}$.

To make comparison between two competing values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, we introduce the likelihood ratio:

$$\text{likelihood ratio} = \frac{p(D|\boldsymbol{\theta}_0)}{p(D|\boldsymbol{\theta}_1)}.$$

Prior Knowledge On Parameters

In practice, researchers often have information on unknown parameters before they start analysis. For example,

- $\boldsymbol{\theta}$ must take a value between 0 and 1 because it is probability;
- in case of tossing a coin, $\boldsymbol{\theta}$ is supposed to be 50% if the coin is fair.

In Bayesian statistics, we construct a distribution of unknown parameters that reflect our prior knowledge on their true values. This is call the prior distribution. Let $p(\boldsymbol{\theta})$ denote the prior distribution.

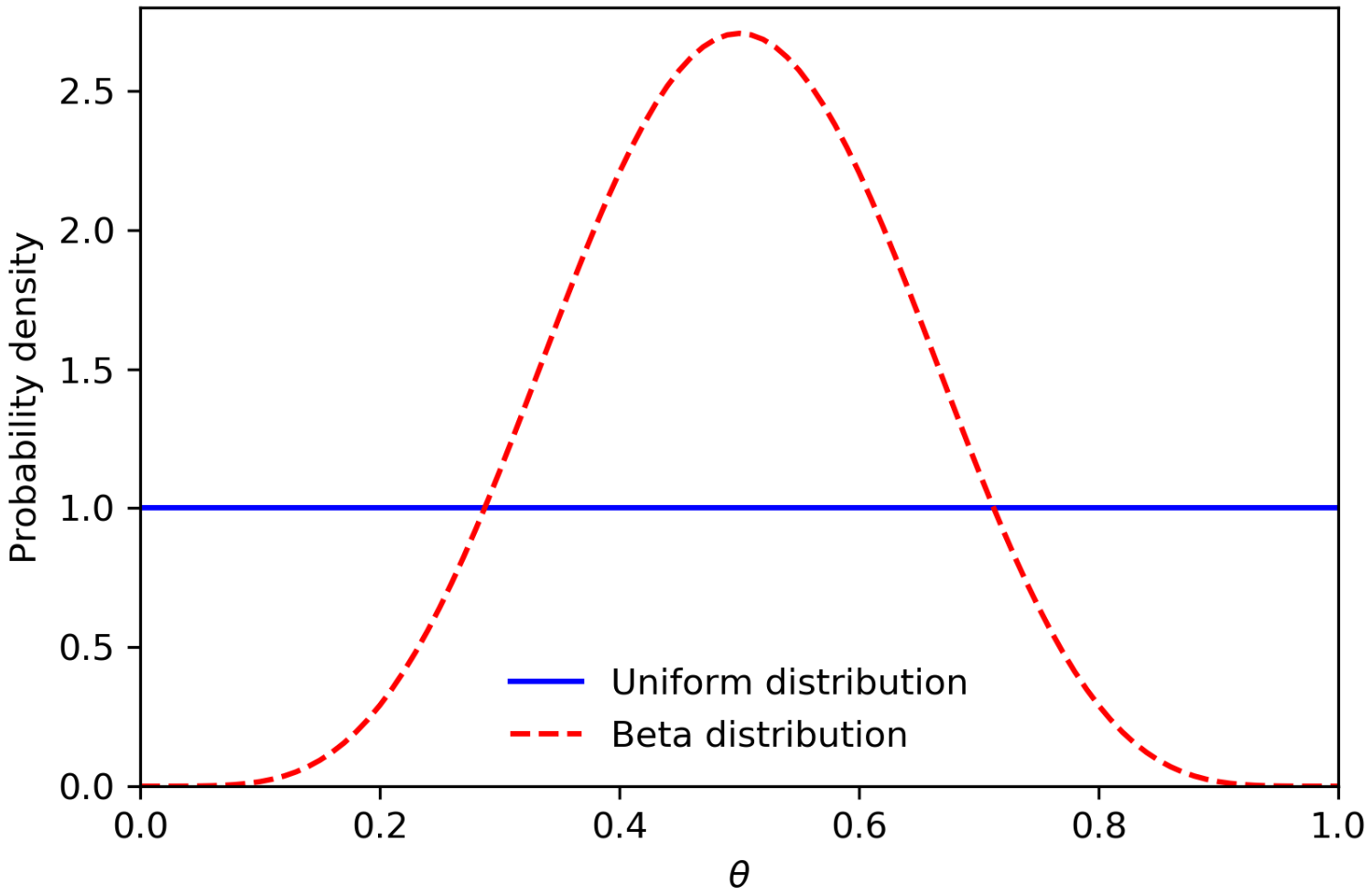

Figure 2: Prior distributions of $\boldsymbol{\theta}$ in the Bernoulli distribution

The uniform distribution $\mathrm{Uniform}(a, b)$ is

$$p(x|a,b) = \begin{cases} \frac{1}{b-a}, & (a \leqq x \leqq b); \\ 0, & (\text{otherwise}). \end{cases}$$

In the above figure, we set $a = 0$ and $b = 1$.

The beta distribution $\mathrm{Beta}(\alpha, \beta)$ is

$$p(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ 0 \leqq x \leqq 1.$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

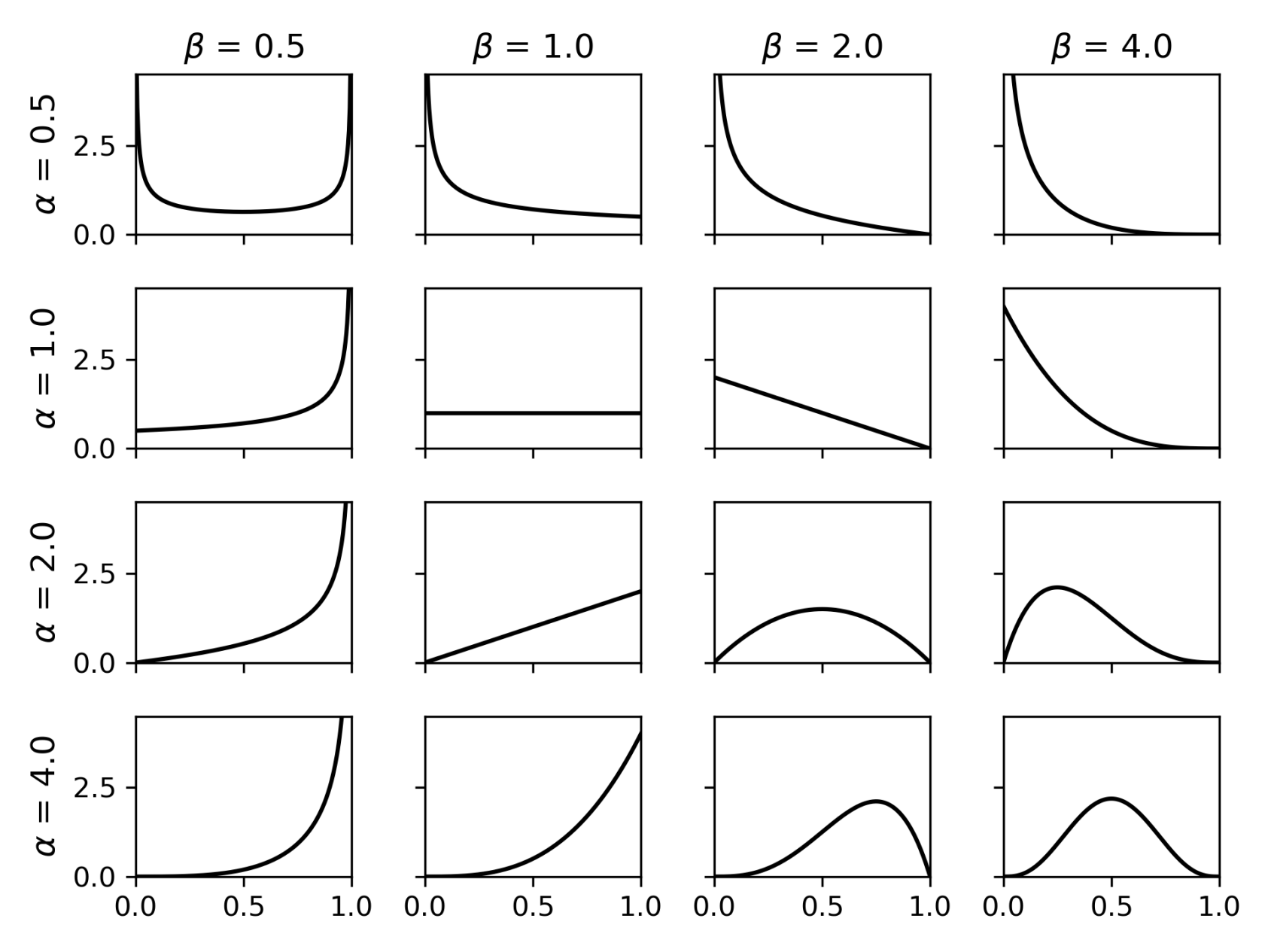

Figure 3: Beta distributions with various (α, β)

Bayes' Theorem And Posterior Distribution

Suppose $p(\boldsymbol{\theta}, D)$ is the joint distribution of $\boldsymbol{\theta}$ and D. Using the definition of the conditional distribution, we have

$$p(\boldsymbol{\theta}, D) = p(\boldsymbol{\theta}|D)p(D) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Arranging the middle and the right-hand side of the above equation, we have

Bayes' Theorem

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}.$$

- The above formula is called Bayes' theorem.
- $p(\boldsymbol{\theta}|D)$ is called the posterior distribution.
- $p(D)$ is called the normalizing constant.

Marginal Likelihood

By marginalizing the joint distribution $p(\theta, D)$, we have

$$p(D) = \int_{\Theta} p(\theta, D) d\theta = \int_{\Theta} p(D|\theta) p(\theta) d\theta.$$

Thus $p(D)$ is interpreted as "averaged likelihood" in terms of the prior $p(\theta)$. In this sense, $p(D)$ is called the marginal likelihood.

Then Bayes' theorem is rewritten as

Bayes' Theorem (Alternative Form)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)d\theta} \propto p(D|\theta)p(\theta).$$

We can ignore $p(D)$ since it does not depend on θ.

Example: Bernoulli Distribution

Suppose the prior distribution is Beta(α_0, β_0).

The posterior distribution of θ is given by

$$
\begin{aligned}
p(\theta|D) &\propto p(D|\theta)p(\theta) \\
&\propto \theta^y(1-\theta)^{n-y} \times \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\
&\propto \theta^{y+\alpha_0-1}(1-\theta)^{n-y+\beta_0-1} \\
&\propto \theta^{\alpha_\star-1}(1-\theta)^{\beta_\star-1}, \\
\alpha_\star = y + \alpha_0, &\quad \beta_\star = n - y + \beta_0.
\end{aligned}
$$

This is the beta distribution Beta$(\alpha_\star, \beta_\star)$.

Bayesian Learning

Bayes' theorem is rearranged as

$$\frac{p(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta})} = \frac{p(D|\boldsymbol{\theta})}{p(D)}.$$

Therefore

$$\frac{p(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta})} \gtreqless 1 \quad \text{if and only if} \quad \frac{p(D|\boldsymbol{\theta})}{p(D)} \gtreqless 1.$$

Since

$$\begin{cases} \dfrac{p(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta})} > 1, & \text{plausibility of } \boldsymbol{\theta} \text{ is increased;} \\ \dfrac{p(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta})} < 1, & \text{plausibility of } \boldsymbol{\theta} \text{ is decreased,} \end{cases}$$

The plausibility of $\boldsymbol{\theta}$ depends on whether $p(D|\boldsymbol{\theta})$ is higher than $p(D)$.

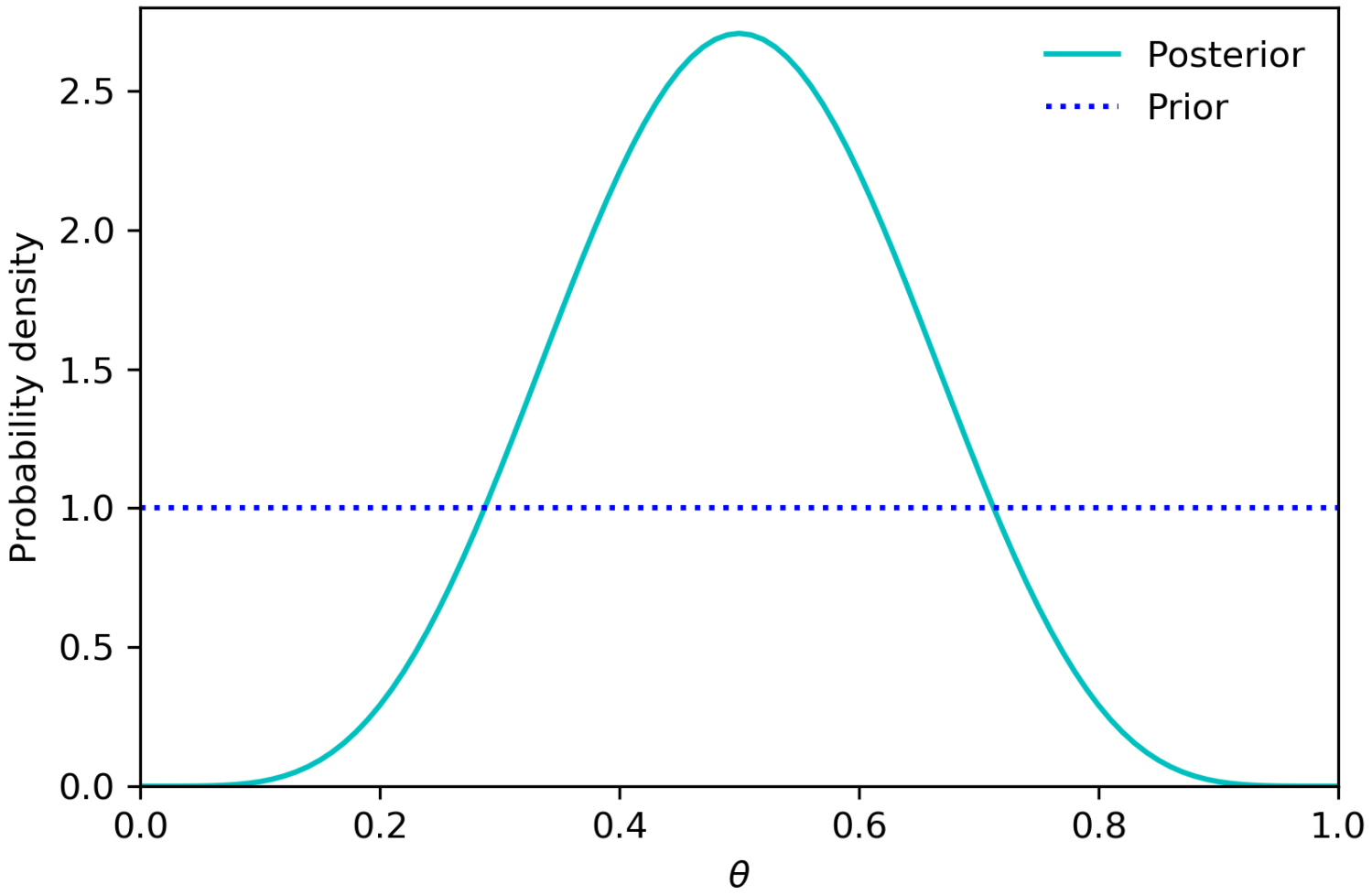

Figure 4: Posterior distributions of the probability of success

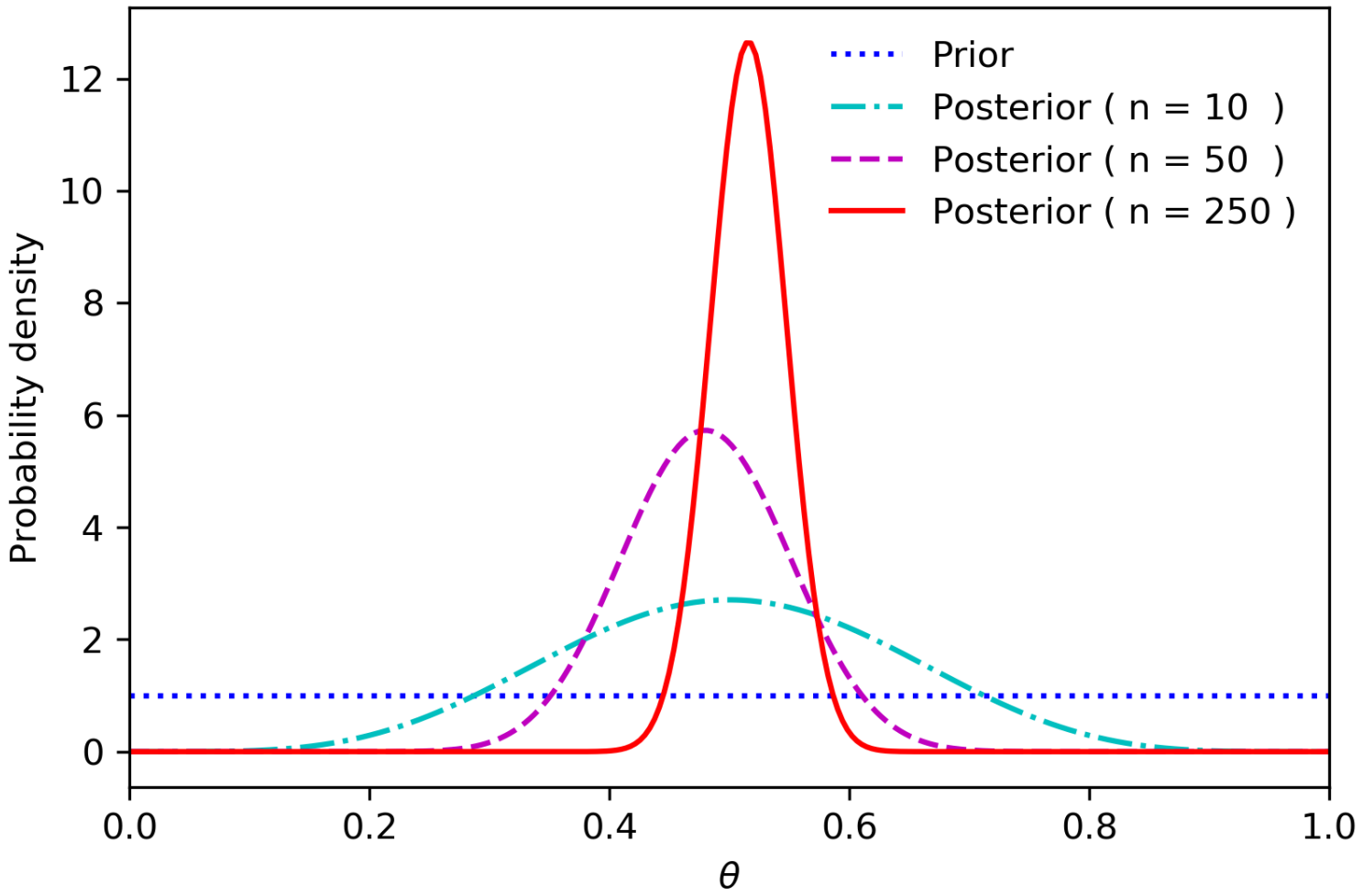

Figure 5: Sequential updating of the posterior distribution

Bayesian Inference With The Posterior Distribution

The posterior distribution $p(\boldsymbol{\theta}|D)$ embodies all available information about unknown parameters, $\boldsymbol{\theta}$. When the number of parameters to be analyzed is relatively small, displaying graphs of all (marginal) posterior distributions may be sufficient to convey useful insights on the parameters to readers.

However, when we need to analyze many parameters, it is impractical and pointless to show all graphs on the parameters in an article or report. In practice, we calculate and report several "summary statistics" that show us key characteristics of the posterior distribution. We call them the posterior statistics.

Point Estimation

On many occasions, we need to report one particular value of the parameter we regard as the most plausible guess. This type of value is called an estimate and a procedure to obtain an estimate is called point estimation.

In Bayesian statistics, an estimate of the parameter is defined as a value that minimizes the expected loss.

$$\begin{aligned}\delta_\star &= \arg\min_\delta \mathrm{E}_\theta[L(\theta,\delta)|D] \\ &= \arg\min_\delta \int_\Theta L(\theta,\delta)p(\theta|D)d\theta,\end{aligned}$$

where L is the loss function and Θ is a set of all possible values of θ (parameter space). In case of the Bernoulli probability, $\Theta = \{\theta : 0 \leqq \theta \leqq 1\}$.

Examples Of Loss Functions

loss function	$L(\theta, \delta)$	point estimate
quadratic loss	$(\theta - \delta)^2$	posterior mean
absolute loss	$\lvert\theta - \delta\rvert$	posterior median
0–1 loss	$1 - 1_\theta(\delta)$	posterior mode

where $1_q(\delta)$ is the indicator function such that

$$1_\theta(\delta) = \begin{cases} 1, & (\delta = \theta), \\ 0, & (\delta \neq \theta). \end{cases}$$

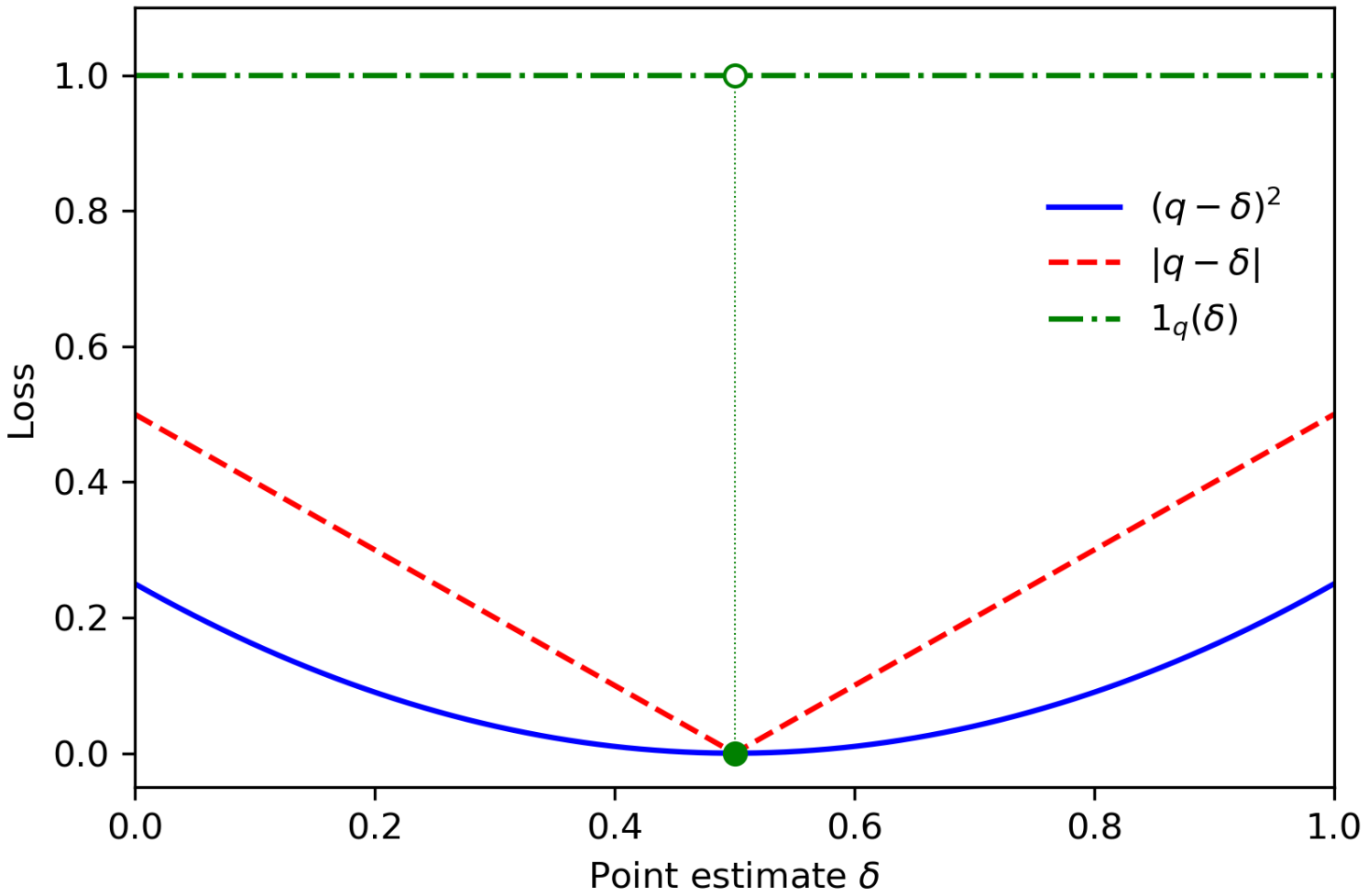

Figure 6: Examples of loss functions

Mean, Median, Mode

The mean of the distribution is the weighted average of all possible values $\boldsymbol{\theta}$ may take, i.e.,

$$\mathbf{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|D] = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}.$$

The median of the distribution is a point that divides the distribution in half, i.e.,

$$\mathbf{P}(\boldsymbol{\theta} \leqq \mathbf{Median}_{\boldsymbol{\theta}}|D) = 50\%.$$

The mode of the distribution is the highest point of the density, i.e.,

$$\mathbf{Mode}_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D).$$

Proof: Point Estimation with the Quadratic Loss

$$
\begin{aligned}
&\mathbf{E}_\theta[L(\theta,\delta)|D] = \mathbf{E}_\theta[(\theta-\delta)^2|D] \\
&= \mathbf{E}_\theta\left[(\theta - \mathbf{E}_\theta[\theta|D] + \mathbf{E}_\theta[\theta|D] - \delta)^2|D\right] \\
&= \mathbf{E}_\theta\Big[(\theta - \mathbf{E}_\theta[\theta|D])^2 + 2(\theta - \mathbf{E}_\theta[\theta|D])(\mathbf{E}_\theta[\theta|D] - \delta) \\
&\quad + (\mathbf{E}_\theta[\theta|D] - \delta)^2|D\Big] \\
&= \mathbf{E}_\theta\left[(\theta - \mathbf{E}_\theta[\theta|D])^2|D\right] + (\mathbf{E}_\theta[\theta|D] - \delta)^2 .
\end{aligned}
$$

The first term in the last equation is the posterior variance of $\boldsymbol{\theta}$, which must be positive. Thus the expected loss is minimized if the second term in the last equation is zero.

Proof: Point Estimation with the Absolute Loss i

$$
\begin{aligned}
\mathbf{E}_\theta[L(\theta,\delta)|D] &= \int_{-\infty}^{\infty} |\theta-\delta|\theta p(\theta|D)d\theta \\
&= \int_{-\infty}^{\delta} (\delta-\theta)p(\theta|D)d\theta + \int_{\delta}^{\infty} (\theta-\delta)p(\theta|D)d\theta \\
&= \delta P(\delta|D) - \int_{-\infty}^{\delta} \theta p(\theta|D)d\theta \\
&\quad + \int_{\delta}^{\infty} \theta p(\theta|D)d\theta - \delta[1-P(\delta|D)] \\
&= 2\delta P(\delta|D) - \delta - 2\int_{-\infty}^{\delta} \theta p(\theta|D)d\theta + \int_{-\infty}^{\infty} \theta p(\theta|D)d\theta.
\end{aligned}
$$

Proof: Point Estimation with the Absolute Loss ii

By applying the integration by part, we have

$$\begin{aligned}
\mathbf{E}_\theta[L(\theta,\delta)|D] &= 2\delta P(\delta|D) - \delta \\
&- 2\left\{\theta P(\theta|D)\Big|_{-\infty}^{\delta} - \int_{-\infty}^{\delta} P(\theta|D)d\theta\right\} + \mathbf{E}_\theta[\theta|D] \\
&= 2\int_{-\infty}^{\delta} P(\theta|D)d\theta - \delta + \mathbf{E}_\theta[\theta|D].
\end{aligned}$$

The first-order condition for minimization is

$$\nabla_\delta \mathbf{E}_\theta[L(\theta,\delta)|D] = 2P(\delta_\star|D) - 1 = 0.$$

Thus we have $P(\delta_\star|D) = \frac{1}{2}$, i.e., $\delta_\star$ must be $\mathbf{Median}_\theta$.

Proof: Point Estimation with the 0-1 Loss

For any $\epsilon > 0$, define

$$L_\epsilon(\theta, \delta) = 1 - 1_{[\theta-\epsilon, \theta+\epsilon]}(\delta) = \begin{cases} 0, & (\theta - \epsilon \leqq \delta \leqq \theta + \epsilon), \\ 1, & (\delta < \theta - \epsilon,\ \theta + \epsilon < \delta). \end{cases}$$

Note that $\lim_{\epsilon \to 0} L_\epsilon(\theta, \delta) = 1 - 1_\theta(\delta)$. Then we have

$$\mathbf{E}_\theta[L_\epsilon(\theta, \delta)|D] = \int_{-\infty}^{\infty} L_\epsilon(\theta, \delta) p(\theta|D) d\theta = 1 - \int_{\delta-\epsilon}^{\delta+\epsilon} p(\theta|D) d\theta,$$

which is maximized for δ that maximizes

$$\int_{\delta-\epsilon}^{\delta+\epsilon} p(\theta|D) d\theta = \mathbf{P}(\delta - \epsilon \leqq \theta \leqq \delta + \epsilon|D).$$

For infinitesimally small ϵ, such δ must be $\mathbf{Mode}_\theta$.

Remarks On Point Estimation

1. A point estimate (mean, median, mode) is merely a representative point in the posterior distribution. So it is by no means the true value of the parameter in any sense.
2. The mode is not necessarily located in the center of the posterior distribution. You must be careful about using the mode as a point estimate.
3. A Bayesian point estimate is not a random variable.
4. An estimator in the frequentist approach, on the other hand, is supposed to be a random variable.
5. Some researchers use the formula of a Bayesian point estimate as a frequentist estimator. This is often called a Bayes estimator.

Posterior Probability

The probability that the true value of $\boldsymbol{\theta}$ is within a region in the parameter space, $S_0 \subset \boldsymbol{\Theta}$, is given by

$$\mathbf{P}(\boldsymbol{\theta} \in S_0 | D) = \int_{S_0} p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}.$$

Such a probability is often called the posterior probability.

When the region is an interval, $S_0 = \{\boldsymbol{\theta} : \ a \leqq \boldsymbol{\theta} \leqq b\}$, we have

$$\mathbf{P}(a \leqq \boldsymbol{\theta} \leqq b | D) = \int_a^b p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}.$$

Credible Interval (CI)

It is tempting to state that the true value of the parameter must be within an interval with very high posterior probability (say 95%). However, there exist infinitely many intervals with 95% probability because the posterior distribution of the parameter is continuous. Thus we need extra conditions to pin down a unique interval with high posterior probability.

The credible interval of $\boldsymbol{\theta}$ is an interval $[a_c, b_c]$ such that

1. $\mathbf{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c$,
2. $\mathbf{P}(\theta < a_c | D) = \frac{c}{2}$ and $\mathbf{P}(\theta > b_c | D) = \frac{c}{2}$.

Set $c = 0.05$ for the 95% CI.

Highest Posterior Density Interval (HPDI)

The highest posterior density interval of $\boldsymbol{\theta}$ is an interval $[a_c, b_c]$ such that

1. $\mathbf{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c$,
2. for any pair $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ such that $\boldsymbol{\theta} \in [a_c, b_c]$ and $\boldsymbol{\theta}' \notin [a_c, b_c]$, $p(\boldsymbol{\theta}|D) > p(\boldsymbol{\theta}'|D)$ must hold.

In particular, if the distribution is unimodal (it has the unique mode), the HPDI must satisfy

$$\mathbf{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c,$$
$$p(a_c|D) = p(b_c|D).$$

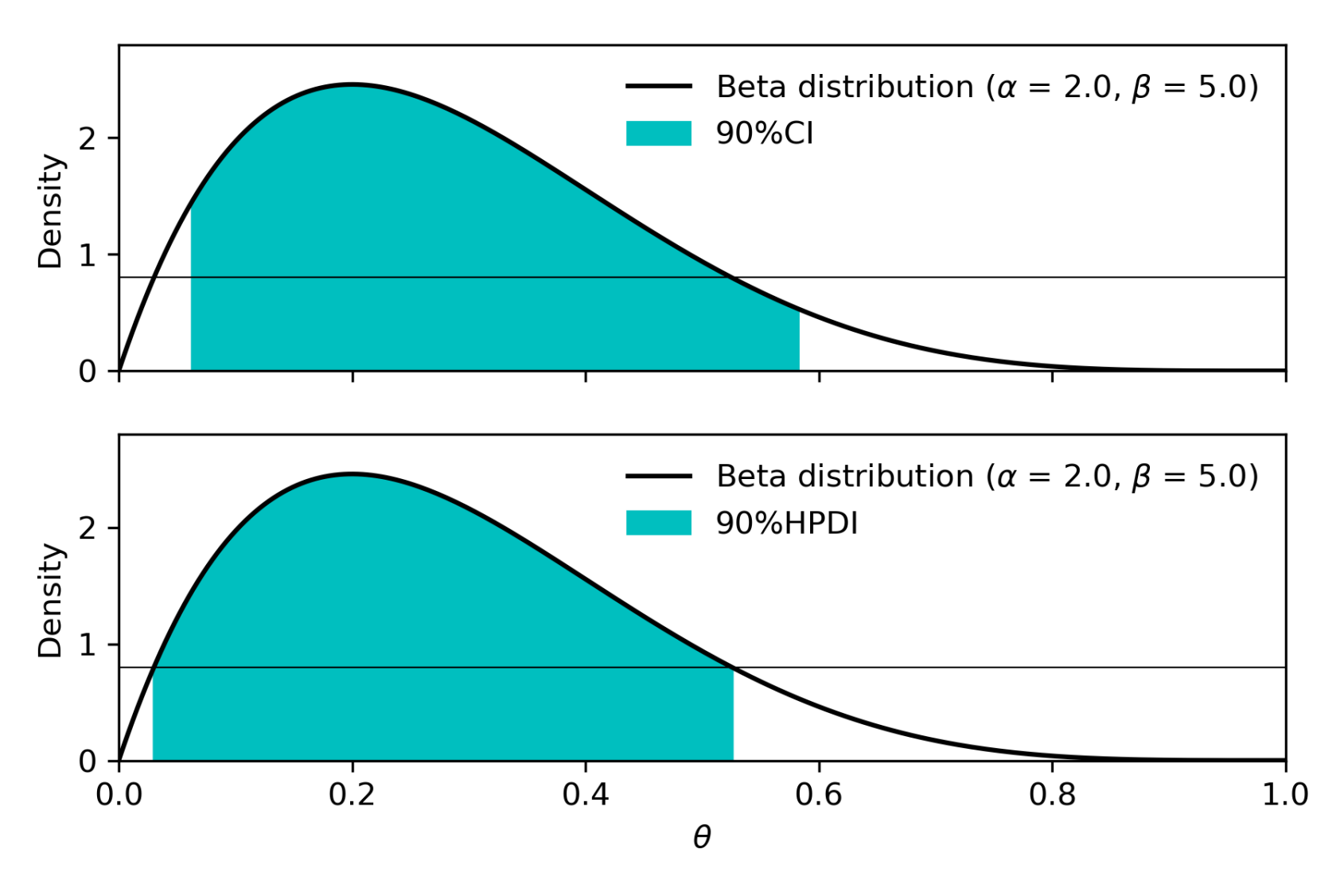

Figure 7: Comparison between CI and HPDI

Remarks On Interval Estimation

1. Both ends of CI or HPDI are not random.
2. The confidence interval in the frequentist approach, on the other hand, is a randomly shifting interval.
3. In Bayesian statistics, the posterior probability is your degree of belief. So you can say "I am 95% certain that the true value of $\boldsymbol{\theta}$ is located in the 95% HPDI."
4. The 95% confidence interval may capture the true value of the parameter with probability 95%. This probability "95%" is interpreted as the frequency with which the confidence interval will succeed in capturing the true value.

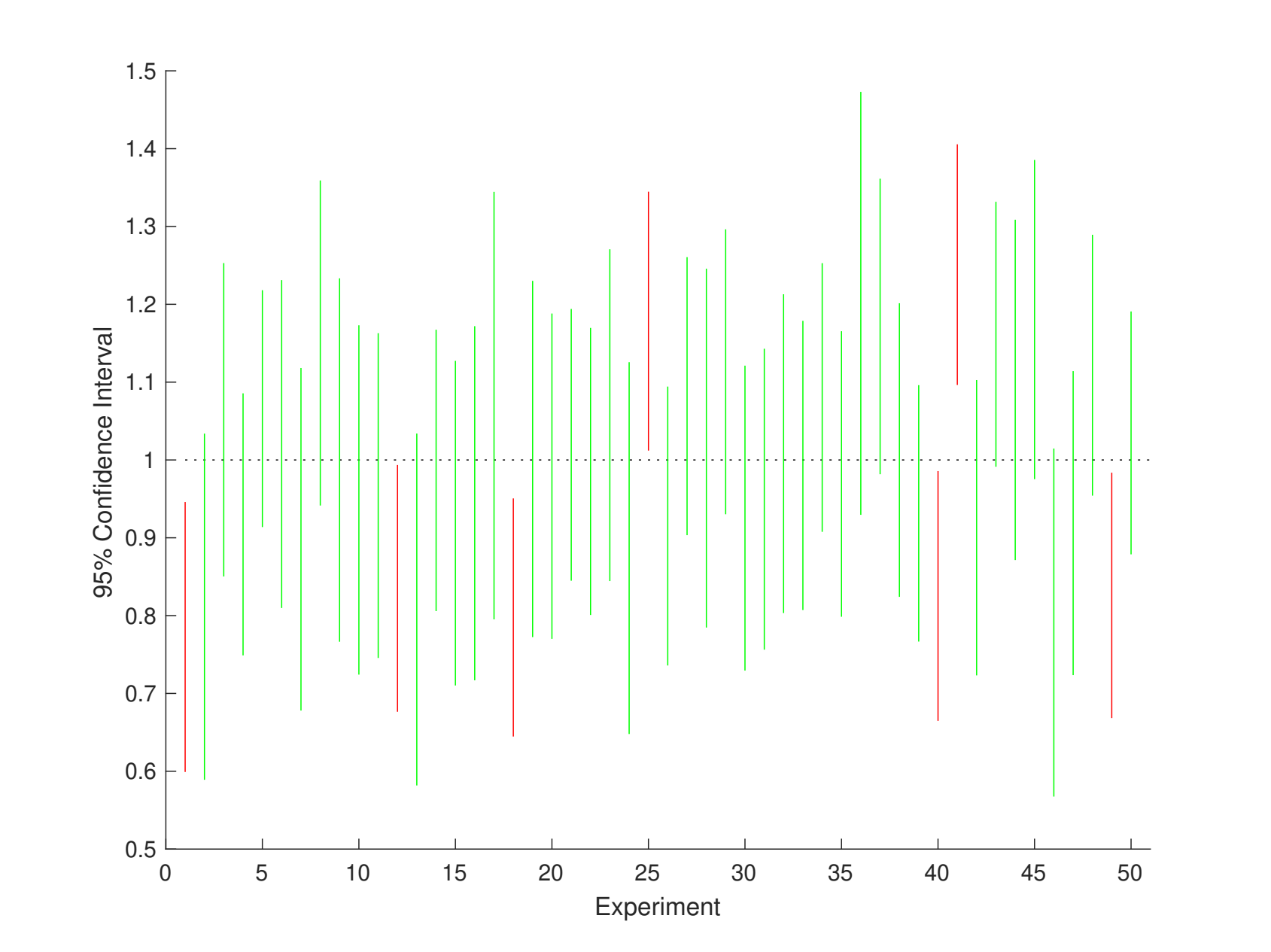

Figure 8: Confidence interval as a randomly shifting interval

Hypotheses On Parameters

In statistics, either Bayesian or frequentist, a hypothesis on the parameters is a region or interval where the true value of the parameter is supposed to be located. For example,

- $\{\boldsymbol{\theta} : \ 0.5 \leqq \boldsymbol{\theta} \leqq 1\}$,
- $\boldsymbol{\theta} = 0.5$,
- $\boldsymbol{\theta} \neq 0.5$
  $\Leftrightarrow \ \{\boldsymbol{\theta} : \ 0 \leqq \ \boldsymbol{\theta} \ < 0.5\} \cup \{\boldsymbol{\theta} : \ 0.5 < \boldsymbol{\theta} \ \leq \ 1\}$.

In general, a hypothesis H_i under which the true value of $\boldsymbol{\theta}$ is located in a region $S_i \subset \boldsymbol{\Theta}$ is expressed as

$$H_i : \ \boldsymbol{\theta} \in S_i, \quad i = 0, 1, 2, \ldots$$

Hypothesis Testing

In Bayesian statistics, plausibility of a hypothesis is measured by the posterior probability that the true value of $\boldsymbol{\theta}$ is located in S_i, that is,

$$\mathbf{P}(H_i|D) = \mathbf{P}(\boldsymbol{\theta} \in S_i|D) = \int_{S_i} p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}.$$

Competing hypotheses can be compared by using the posterior odds ratio:

$$\text{Posterior odds ratio} = \frac{\mathbf{P}(H_i|D)}{\mathbf{P}(H_j|D)}, \quad i \neq j.$$

Unlike the frequentist approach, Bayesian hypothesis testing does not involves the level of significance, the power and that dreaded P-value!

Bayes Factor

One catch of the posterior odds ratio is that it is affected by the prior information. If the prior information is biased in favor of one hypothesis, the posterior odds ratio is also biased for that hypothesis.

To control the impact of the prior information, the Bayes factor is often used. It is defined as

$$\text{Bayes factor} = \mathbf{B}_{ij} = \frac{\mathbf{P}(H_i|D)}{\mathbf{P}(H_j|D)} \div \frac{\mathbf{P}(H_i)}{\mathbf{P}(H_j)},$$

where $\mathbf{P}(H_i) = \int_{S_i} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and $\mathbf{P}(H_i)/\mathbf{P}(H_j)$ is called the prior odds ratio. Note that the Bayes factor is equivalent to the posterior odds ratio if the prior odds ratio is one.

Scale Of Bayes Factor By Jeffreys (1961)

We compare H_i against H_j ($i \neq j$). We suppose H_i is the hypothesis we keep unless we have no strong evidence against it. H_j, on the other hand, is the hypothesis we want to check whether it is supported by the evidence.

Rank	Bayes factor $\mathbf{B}_{ij}$	Support for H_j
0	$0 < \log_{10}(\mathbf{B}_{ij})$	Rejected
1	$-\frac{1}{2} < \log_{10}(\mathbf{B}_{ij}) < 0$	Barely worth mentioning
2	$-1 < \log_{10}(\mathbf{B}_{ij}) < -\frac{1}{2}$	Substantial
3	$-\frac{3}{2} < \log_{10}(\mathbf{B}_{ij}) < -1$	Strong
4	$-2 < \log_{10}(\mathbf{B}_{ij}) < -\frac{3}{2}$	Very strong
5	$\log_{10}(\mathbf{B}_{ij}) < -2$	Decisive

Two-Sided Test

On some occasions, we need to check whether the true value of $\boldsymbol{\theta}$ is exactly equal to a particular value, say $\mathbf{0.5}$ ($\boldsymbol{\theta}$ must be $\mathbf{0.5}$ if the coin is fair). For this purpose, we need to compare $H_0: \ \boldsymbol{\theta} = 0.5$ against $H_1: \ \boldsymbol{\theta} \neq 0.5$.

As a general setup, we consider

$$\begin{cases} H_0: & \boldsymbol{\theta} = \boldsymbol{\theta}_0, \\ H_1: & \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \end{cases}$$

For these hypothesis, however, it is meaningless to construct the Bayes factor because

$$\mathbf{P}(\boldsymbol{\theta} = \boldsymbol{\theta}_0) = \mathbf{P}(\boldsymbol{\theta} = \boldsymbol{\theta}_0 | D) = 0,$$

and prior and posterior odds ratio are identical to zero.

Spike-And-Slab Prior

To avoid this problem, we introduce a spike-and-slab prior:

$$p(\theta) = p_0\delta(\theta - \theta_0) + (1 - p_0)f(\theta), \quad 0 < p_0 < 1,$$

where $f(\cdot)$ is a continuous distribution of θ and $\delta(\cdot)$ is the Dirac delta function such that

- for any continuous function $g(x)$, $\int_{-\infty}^{\infty} g(x)\delta(x)dx = g(0)$;
- $\int_{-\infty}^{\infty} \delta(x)dx = 1$;
- $\delta(x) = 0$ only if $x \neq 0$.

Savage-Dickey Density Ratio

With the spike-and-slab prior, the prior odds ratio is $\frac{p_0}{1-p_0}$ and the posterior odds ratio is

$$\text{Posterior odds ratio} = \frac{p_0 f(\theta_0|D)}{(1-p_0)f(\theta_0)},$$

where $f(\theta|D)$ is the posterior distribution when $\theta \neq \theta_0$, i.e.,

$$f(\theta|D) = \frac{p(D|\theta)f(\theta)}{\int_\Theta p(D|\theta)f(\theta)d\theta}.$$

Then the Bayes factor is given by

$$\mathbf{B}_{01} = \frac{f(\theta_0|D)}{f(\theta_0)},$$

which is called the Savage-Dickey density ratio (SDDR).

Proof: SDDR i

The posterior distribution with the spike-and-slab prior is derived as follows.

$$
\begin{aligned}
p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)d\theta} \\
&= \frac{p(D|\theta)\left\{p_0\delta(\theta-\theta_0)+(1-p_0)f(\theta)\right\}}{\int_{\Theta} p(D|\theta)\left\{p_0\delta(\theta-\theta_0)+(1-p_0)f(\theta)\right\}d\theta} \\
&= \mathcal{K}^{-1}\left\{p_0 p(D|\theta)\delta(\theta-\theta_0)+(1-p_0)p(D|\theta)f(\theta)\right\},
\end{aligned}
$$

where $\mathcal{K} = p_0 p(D|\theta_0) + (1-p_0)\int_{\Theta} p(D|\theta)f(\theta)d\theta$.

Proof: SDDR ii

Since

$$\mathbf{P}(\theta = \theta_0|D) = \mathcal{K}^{-1} p_0 p(D|\theta_0),$$
$$\mathbf{P}(\theta \neq \theta_0|D) = \mathcal{K}^{-1}(1 - p_0) \int_{\Theta} p(D|\theta) f(\theta) d\theta,$$

the posterior odds ratio is

$$\frac{\mathbf{P}(\theta = \theta_0|D)}{\mathbf{P}(\theta \neq \theta_0|D)} = \frac{p_0}{1 - p_0} \times \frac{p(D|\theta_0)}{\int_{\Theta} p(D|\theta) f(\theta) d\theta} = \frac{p_0}{1 - p_0} \times \frac{f(\theta_0|D)}{f(\theta_0)}.$$

Therefore the SDDR is obtained as

$$\mathbf{B}_{01} = \frac{\mathbf{P}(\theta = \theta_0|D)}{\mathbf{P}(\theta \neq \theta_0|D)} \div \frac{\mathbf{P}(\theta = \theta_0)}{\mathbf{P}(\theta \neq \theta_0)} = \frac{f(\theta_0|D)}{f(\theta_0)}.$$

Predictive Distribution i

Let $\tilde{x}$ denote an unrealized/future value of the population distribution $p(x|\theta)$. Since it is a random variable, we can consider the joint distribution of $\tilde{x}$ and the previous data $D = (x_1, \ldots, x_n)$:

$$p(\tilde{x}, x_1, \ldots, x_n) = p(\tilde{x}, D),$$

Then, from the definition of the conditional probability, we have

$$p(\tilde{x}, D) = p(\tilde{x}|D)p(D) \quad \Rightarrow \quad p(\tilde{x}|D) = \frac{p(\tilde{x}, D)}{p(D)}.$$

Predictive Distribution ii

Furthermore, both $p(D)$ and $p(\tilde{x}, D)$ are regarded as the marginal likelihood given D and $(\tilde{x}, D)$ respectively, that is,

$$p(D) = \int_{\Theta} p(D|\theta)p(\theta)d\theta,$$
$$p(\tilde{x}, D) = \int_{\Theta} p(\tilde{x}, D|\theta)p(\theta)d\theta.$$

In sum, we have

$$p(\tilde{x}|D) = \frac{\int_{\Theta} p(\tilde{x}, D|\theta)p(\theta)d\theta}{\int_{\Theta} p(D|\theta)p(\theta)d\theta}.$$

Predictive Distribution iii

This is called the predictive distribution of $\tilde{x}$. In particular, if $\tilde{x}$ and D are independent, we have

$$p(\tilde{x}, D|\theta) = p(\tilde{x}|\theta)p(D|\theta).$$

Thus the predictive distribution of $\tilde{x}$ is rearranged as

$$\begin{aligned} p(\tilde{x}|D) &= \frac{\int_\Theta p(\tilde{x}|\theta)p(D|\theta)p(\theta)d\theta}{\int_\Theta p(D|\theta)p(\theta)d\theta} \\ &= \int_\Theta p(\tilde{x}|\theta)\frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta}d\theta \\ &= \int_\Theta p(\tilde{x}|\theta)p(\theta|D)d\theta. \end{aligned}$$

Predictive Distribution (Bernoulli Model) i

Let us derive the predictive distribution for the Bernoulli distribution.

$$
\begin{aligned}
p(\tilde{x}|D) &= \int_{\Theta} p(\tilde{x}|\theta)p(\theta|D)d\theta \\
&= \int_0^1 \theta^{\tilde{x}}(1-\theta)^{1-\tilde{x}} \frac{\theta^{\alpha_\star - 1}(1-\theta)^{\beta_\star - 1}}{B(\alpha_\star, \beta_\star)} d\theta \\
&= \frac{\int_0^1 \theta^{\tilde{x}+\alpha_\star - 1}(1-\theta)^{\beta_\star - \tilde{x}} d\theta}{B(\alpha_\star, \beta_\star)} \\
&= \frac{B(\alpha_\star + \tilde{x}, \beta_\star - \tilde{x} + 1)}{B(\alpha_\star, \beta_\star)}.
\end{aligned}
$$

Predictive Distribution (Bernoulli Model) ii

Using

$$B(\alpha+1,\beta) = \frac{\alpha}{\alpha+\beta}B(\alpha,\beta),$$
$$B(\alpha,\beta+1) = \frac{\beta}{\alpha+\beta}B(\alpha,\beta),$$

we have

$$p(\tilde{x}=1|D) = \frac{B(\alpha_\star+1,\beta_\star)}{B(\alpha_\star,\beta_\star)} = \frac{\alpha_\star}{\alpha_\star+\beta_\star},$$
$$p(\tilde{x}=0|D) = \frac{B(\alpha_\star,\beta_\star+1)}{B(\alpha_\star,\beta_\star)} = \frac{\beta_\star}{\alpha_\star+\beta_\star}.$$

Predictive Distribution (Bernoulli Model) iii

Finally

$$p(\tilde{x}|D) = \left(\frac{\alpha_\star}{\alpha_\star + \beta_\star}\right)^{\tilde{x}} \left(\frac{\beta_\star}{\alpha_\star + \beta_\star}\right)^{1-\tilde{x}}.$$

This is the Bernoulli distribution with $\theta = \frac{\alpha_\star}{\alpha_\star + \beta_\star}$.

Poisson Distribution

The p.m.f. of a Poisson distribution is

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \ \lambda > 0,$$
$$\mathbf{E}[X] = \mathbf{Var}[X] = \lambda.$$

The Poisson distribution is used to model the number of occurrences in a fixed time interval of rare events such as

- traffic accidents
- crimes
- arrival of customers

Gamma Distribution

We use a gamma distribution

$$\boldsymbol{\lambda} \sim \mathbf{Gamma}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0),$$

as the prior of $\boldsymbol{\lambda}$ in the Poisson distribution.

The p.d.f. of a gamma distribution $\mathbf{Gamma}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is given by

$$p(x|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^{\boldsymbol{\alpha}}}{\boldsymbol{\Gamma}(\boldsymbol{\alpha})} x^{\boldsymbol{\alpha}-1} e^{-\boldsymbol{\beta} x} \quad x > 0,\ \boldsymbol{\alpha} > 0,\ \boldsymbol{\beta} > 0,$$

where $\boldsymbol{\Gamma}(\cdot)$ is the gamma function:

$$\boldsymbol{\Gamma}(x) = \int_0^{\infty} z^{x-1} e^{-z} dz.$$

Likelihood

The likelihood of λ given $D = (x_1, \ldots, x_n)$ is

$$
\begin{aligned}
p(D|\lambda) &= \prod_{i=1}^{n} p(x_i|\lambda) \\
&= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}, \\
&= \frac{\lambda^{y} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}, \quad y = \sum_{i=1}^{n} x_i.
\end{aligned}
$$

Posterior Distribution

Applying Bayes' theorem, we have

$$
\begin{aligned}
p(\lambda|D) &\propto p(D|\lambda)p(\lambda) \\
&\propto \lambda^{y} e^{-n\lambda} \times \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \\
&\propto \lambda^{y+\alpha_0-1} e^{-(n+\beta_0)\lambda} \\
&\propto \lambda^{\alpha_\star - 1} e^{-\beta_\star \lambda} \\
\alpha_\star &= y + \alpha_0, \quad \beta_\star = n + \beta_0.
\end{aligned}
$$

This is the gamma distribution $\mathbf{Gamma}(\alpha_\star, \beta_\star)$.

Natural Conjugate Prior

Some of you already noticed the following observations.

- In case of the Bernoulli distribution, if we use a beta distribution as the prior for the probability of success, the posterior is also a beta distribution.
- In case of the Poisson distribution, if we use a gamma distribution as the prior for the mean, the posterior is also a gamma distribution.

A class of prior distributions that make the posterior distribution belong to the same family of the prior distribution is called the natural conjugate prior or conjugate prior.

Predictive Distribution

The predictive distribution for the Poisson distribution is

$$p(\tilde{x}|D) = \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1} \left(\frac{\beta_\star}{\beta_\star + 1}\right)^{\alpha_\star} \left(\frac{1}{\beta_\star + 1}\right)^{\tilde{x}},$$

$$\tilde{x} = 0, 1, 2, \ldots$$

This is a negative-binomial distribution.

Proof i

Following the definition of the predictive distribution, we have

$$
\begin{aligned}
p(\tilde{x}|D) &= \int_0^\infty \frac{\lambda^{\tilde{x}} e^{-\lambda}}{\tilde{x}!} \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \lambda^{\alpha_\star - 1} e^{-\beta_\star \lambda} d\lambda \\
&= \frac{\beta_\star^{\alpha_\star}}{\tilde{x}!\Gamma(\alpha_\star)} \int_0^\infty \lambda^{\tilde{x}+\alpha_\star - 1} e^{-(\beta_\star + 1)\lambda} d\lambda \\
&= \frac{\Gamma(\tilde{x} + \alpha_\star)}{\tilde{x}!\Gamma(\alpha_\star)} \frac{\beta_\star^{\alpha_\star}}{(\beta_\star + 1)^{\tilde{x}+\alpha_\star}}.
\end{aligned}
$$

Proof ii

Since $\Gamma(n) = (n-1)!$, if $\alpha_\star$ is an integer, we obtain

$$\frac{\Gamma(\tilde{x}+\alpha_\star)}{\tilde{x}!\Gamma(\alpha_\star)} = \frac{(\tilde{x}+\alpha_\star-1)!}{\tilde{x}!(\alpha_\star-1)!} = \binom{\tilde{x}+\alpha_\star-1}{\alpha_\star-1}.$$

Therefore

$$\begin{aligned} p(\tilde{x}|D) &= \binom{\tilde{x}+\alpha_\star-1}{\alpha_\star-1}\frac{\beta_\star^{\alpha_\star}}{(\beta_\star+1)^{\tilde{x}+\alpha_\star}} \\ &= \binom{\tilde{x}+\alpha_\star-1}{\alpha_\star-1}\left(\frac{\beta_\star}{\beta_\star+1}\right)^{\alpha_\star}\left(\frac{1}{\beta_\star+1}\right)^{\tilde{x}}. \end{aligned}$$

Exponential Distribution

The p.d.f. of an exponential distribution $\mathbf{Exp}(\lambda)$ is

$$p(x|\lambda) = \lambda e^{-\lambda x},$$
$$0 < x < \infty, \quad \lambda > 0,$$
$$\mathbf{E}[X] = \frac{1}{\lambda}, \quad \mathbf{Var}[X] = \frac{1}{\lambda^2}.$$

The exponential distribution is used to model the length of duration.

Posterior Distribution i

Suppose we have the i.i.d. sample $D = (x_1, \ldots, x_n)$ from $\mathbf{Exp}(\lambda)$.

The likelihood is

$$p(D|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

We use the gamma distribution $\mathbf{Gamma}(\alpha_0, \beta_0)$ as the prior of λ.

Posterior Distribution ii

The posterior distribution of $\boldsymbol{\lambda}$ is derived as

$$
\begin{aligned}
p(\lambda|D) &\propto p(D|\lambda)p(\lambda) \\
&\propto \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \times \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \\
&\propto \lambda^{n+\alpha_0-1} e^{-\left(\sum_{i=1}^n x_i + \beta_0\right)\lambda} \\
&\propto \lambda^{\alpha_\star - 1} e^{-\beta_\star \lambda},
\end{aligned}
$$

$$
\alpha_\star = n + \alpha_0, \quad \beta_\star = \sum_{i=1}^n x_i + \beta_0,
$$

which is the gamma distribution $\mathbf{Gamma}(\alpha_\star, \beta_\star)$.

Predictive Distribution

$$
\begin{aligned}
&p(\tilde{x}|D) \\
&= \int_0^\infty p(\tilde{x}|\lambda)p(\lambda|D)d\lambda \\
&= \int_0^\infty \lambda e^{-\lambda\tilde{x}} \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \lambda^{\alpha_\star - 1} e^{-\beta_\star \lambda} d\lambda \\
&= \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \int_0^\infty \lambda^{\alpha_\star} e^{-(\tilde{x}+\beta_\star)\lambda} d\lambda = \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \frac{\Gamma(\alpha_\star + 1)}{(\tilde{x} + \beta_\star)^{\alpha_\star + 1}} \\
&= \frac{\alpha_\star \beta_\star^{\alpha_\star}}{(\tilde{x} + \beta_\star)^{\alpha_\star + 1}},
\end{aligned}
$$

which is a Pareto distribution.

Alternative Parameterization i

In many textbooks, the exponential distribution is defined As

$$p(x|\boldsymbol{\theta}) = \frac{1}{\boldsymbol{\theta}} e^{\frac{1}{\theta}},$$
$$0 < x < \infty, \quad \boldsymbol{\theta} > 0,$$
$$\mathbf{E}[x] = \boldsymbol{\theta}, \quad \mathbf{Var}[X] = \boldsymbol{\theta}^2.$$

Obviously, reparameterization $\boldsymbol{\lambda} = s(\boldsymbol{\theta}) = 1/\boldsymbol{\theta}$ leads to the aforementioned expression of the exponential distribution.

Alternative Parameterization ii

We can derive the posterior distribution of $\boldsymbol{\theta}$ by applying the change-of-variable formula. The Jacobian is

$$|\nabla s(\theta)| = \frac{1}{\theta^2}.$$

Thus the posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned} p(\theta|D) &= \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} s(\theta)^{\alpha^*-1} \exp\left[-\beta^* s(\theta)\right] |\nabla s(\lambda)| \\ &= \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \theta^{-(\alpha^*+1)} \exp\left(-\frac{\beta^*}{\theta}\right), \end{aligned}$$

which is called the inverse Gamma distribution.

Alternative Parameterization iii

In general, the p.d.f. of the inverse Gamma distribution $\mathbf{Inv.Gamma}(\alpha, \beta)$ is given by

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right), \; x > 0, \; \alpha > 0, \; \beta > 0.$$

Note that

$$X \sim \mathbf{Gamma}(\alpha, \beta) \quad \Rightarrow \quad \frac{1}{X} \sim \mathbf{Inv.Gamma}(\alpha, \beta).$$

Normal Distribution

The p.d.f. of a normal distribution $\mathbf{Normal}(\mu, \sigma^2)$ is

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

$$-\infty < x < \infty, \ -\infty < \mu < \infty, \ \sigma^2 > 0,$$

$$\mathbf{E}[X] = \mu, \quad \mathbf{Var}[X] = \sigma^2.$$

- The normal distribution is the mainstay of statistics.
- Many economic data are supposed to follow a normal distribution, though it is not the case for financial data.
- Many sophisticated statistical models are built upon the normal distribution.
- The normal distribution is often a limit of the other distribution — the central limit theorem.

Normal-Inverse-Gamma Distribution i

The natural conjugate prior for (μ, σ^2) is

$$\mu|\sigma^2 \sim \mathbf{Normal}\left(\mu_0, \frac{\sigma^2}{n_0}\right), \ \sigma^2 \sim \mathbf{Inv.Gamma}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right),$$

which is called the normal-inverse-gamma distribution:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2),$$

$$p(\mu|\sigma^2) = \sqrt{\frac{n_0}{2\pi\sigma^2}} \exp\left[-\frac{n_0(\mu - \mu_0)^2}{2\sigma^2}\right],$$

$$p(\sigma^2) = \frac{\left(\frac{\lambda_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\lambda_0}{2\sigma^2}\right).$$

Likelihood i

The likelihood of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is

$$
\begin{aligned}
p(D|\mu, \sigma^2) &= \prod_{i=1}^{n} p(x_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathbf{exp}\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \mathbf{exp}\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right].
\end{aligned}
$$

Likelihood ii

Since

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}\left\{(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu)+(\bar{x}-\mu)^2\right\}\\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2,\end{aligned}$$

the likelihood is rewritten as

$$p(D|\mu,\sigma^2) \propto (\sigma^2)^{-\frac{n}{2}}\,\mathbf{exp}\left[-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right].$$

Posterior Distribution Of (μ, σ^2)

Applying Bayes' theorem, we have

$$
\begin{aligned}
&p(\mu, \sigma^2 | D) \\
&\propto p(D|\mu, \sigma^2) p(\mu|\sigma^2) p(\sigma^2) \\
&\propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right] \\
&\quad \times (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{n_0(\mu - \mu_0)^2}{2\sigma^2}\right] \times (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left[-\frac{\lambda_0}{2\sigma^2}\right] \\
&\propto (\sigma^2)^{-\frac{n+\nu_0+3}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right.\right. \\
&\qquad\qquad \left.\left. + n_0(\mu - \mu_0)^2 + \lambda_0\right\}\right].
\end{aligned}
$$

Completing The Square

"Completing the square" is referred to as a transformation of a quadratic functions:

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}.$$

By completing the square, we have

$$\begin{aligned}
&n(\bar{x} - \mu)^2 + n_0(\mu - \mu_0)^2 \\
&= (n + n_0)\mu^2 - 2(n\bar{x} + n_0\mu_0)\mu + n\bar{x}^2 + n_0\mu_0^2 \\
&= (n + n_0)\left(\mu - \frac{n\bar{x} + n_0\mu_0}{n + n_0}\right)^2 + \frac{nn_0}{n + n_0}(\mu_0 - \bar{x})^2.
\end{aligned}$$

Joint Posterior Distribution i

Therefore the joint posterior distribution of (μ, σ^2) is

$$p(\mu, \sigma^2 | D) \propto (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{n_*(\mu - \mu_*)^2}{2\sigma^2}\right] \times (\sigma^2)^{-\left(\frac{\nu_*}{2}+1\right)} \exp\left(-\frac{\lambda_*}{2\sigma^2}\right),$$

where

$$\mu_* = \frac{n\bar{x} + n_0\mu_0}{n + n_0}, \quad n_* = n + n_0, \quad \nu_* = n + \nu_0,$$

$$\lambda_* = \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{nn_0}{n + n_0}(\mu_0 - \bar{x})^2 + \lambda_0.$$

Joint Posterior Distribution ii

This is also a normal-inverse-gamma distribution.

$$\mu|\sigma^2, D \sim \mathbf{Normal}\left(\mu_*, \frac{\sigma^2}{n_*}\right),$$
$$\sigma^2|D \sim \mathbf{Inv.Gamma}\left(\frac{\nu_*}{2}, \frac{\lambda_*}{2}\right).$$

Integrating Out Nuisance Parameters

To proceed to posterior inference on the mean $\boldsymbol{\mu}$, we need the marginal posterior distribution $\boldsymbol{p(\mu|D)}$. This is obtained by integrating out the other parameter $\boldsymbol{\sigma^2}$ as

$$\begin{aligned} p(\mu|D) &= \int_0^\infty p(\mu, \sigma^2|D) d\sigma^2 \\ &= \int_0^\infty p(\mu|\sigma^2, D) p(\sigma^2|D) d\sigma^2. \end{aligned}$$

A parameter that is not the primary subject of our analysis is called the nuisance parameter.

Marginal Posterior Distribution Of μ i

The marginal posterior distribution of $\boldsymbol{\mu}$ is derived as

$$
\begin{aligned}
p(\mu|D) &= \int_0^\infty p(\mu|\sigma^2, D)p(\sigma^2|D)d\sigma^2 \\
&= \sqrt{\frac{n_*}{2\pi}} \frac{\left(\frac{\lambda_*}{2}\right)^{\frac{\nu_*}{2}}}{\Gamma\left(\frac{\nu_*}{2}\right)} \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_*+1}{2}+1\right)} \exp\left[-\frac{n_*(\mu-\mu_*)^2+\lambda_*}{2\sigma^2}\right] d\sigma^2 \\
&= \frac{\sqrt{n_*}\lambda_*^{\frac{\nu_*}{2}} 2^{-\frac{\nu_*+1}{2}} \Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)\sqrt{\pi}} \left[\frac{n_*(\mu-\mu_*)^2+\lambda_*}{2}\right]^{-\frac{\nu_*+1}{2}} \\
&= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)} \sqrt{\frac{n_*}{\pi\lambda_*}} \left[1+\frac{n_*(\mu-\mu_*)^2}{\lambda_*}\right]^{-\frac{\nu_*+1}{2}}.
\end{aligned}
$$

Marginal Posterior Distribution Of μ ii

The integral is evaluated with the following formula:

$$\int_0^\infty x^{-(\alpha+1)} e^{-\frac{\beta}{x}} dx = \beta^{-\alpha}\Gamma(\alpha).$$

The marginal posterior distribution $p(\mu|D)$ is a (Student's) t-distribution:

$$\mu|D \sim \mathrm{t}\left(\nu_*, \mu_*, \tau_*^2\right), \quad \tau_*^2 = \frac{\lambda_*}{\nu_* n_*},$$

In general, the p.d.f. of the t-distribution $\mathrm{t}(\nu, \mu, \sigma^2)$ is

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left[1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right]^{-\frac{\nu+1}{2}}.$$

Student's t-Distribution

The t-distribution in introductory statistics is defined as

$$T = \frac{Z}{\sqrt{V/\nu}},\ Z \sim \mathbf{Normal}(0,1),\ V \sim \chi^2(\nu),\ Z \perp V,$$

where $\chi^2(\nu)$ represents the χ^2-distribution with degree of freedom ν and "$\perp$" implies independence. In our notation, $T \sim \mathbf{t}(\nu, 0, 1)$.

In the same manner, $T \sim \mathbf{t}(\nu, \mu, \sigma^2)$ is defined as

$$T = \mu + \frac{\sigma Z}{\sqrt{V/\nu}},\ Z \sim \mathbf{Normal}(0,1),\ V \sim \chi^2(\nu),\ Z \perp V.$$

Predictive Distribution i

The predictive distribution for the normal distribution is

$$\tilde{x}|D \sim \mathrm{t}(\nu_*, \mu_*, \tau_*^2(1+n_*)).$$

Proof: We must evaluate

$$p(\tilde{x}|D) = \int_0^\infty \underbrace{\int_{-\infty}^\infty p(\tilde{x}|\mu,\sigma^2)p(\mu|\sigma^2,D)d\mu}_{g(\sigma^2)} p(\sigma^2|D)d\sigma^2.$$

Predictive Distribution ii

First we derive the closed form of the integral:

$$g(\sigma^2) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\tilde{x}-\mu)^2}{2\sigma^2}\right] \sqrt{\frac{n_*}{2\pi\sigma^2}} \exp\left[-\frac{n_*(\mu-\mu_*)^2}{2\sigma^2}\right] d\mu$$

$$= \frac{\sqrt{n_*}}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left[-\frac{(\tilde{x}-\mu)^2 + n_*(\mu-\mu_*)^2}{2\sigma^2}\right] d\mu.$$

Note that, by completing the square, we have

$$\begin{aligned}
&(\tilde{x}-\mu)^2 + n_*(\mu-\mu_*)^2 \\
&= (1+n_*)\mu^2 - 2(\tilde{x}+n_*\mu_*)\mu + \tilde{x}^2 + n_*\mu_*^2 \\
&= (1+n_*)\left(\mu - \frac{\tilde{x}+n_*\mu_*}{1+n_*}\right)^2 + \frac{n_*}{1+n_*}(\tilde{x}-\mu_*)^2.
\end{aligned}$$

Predictive Distribution iii

Thus

$$
\begin{aligned}
g(\sigma^2) &= \sqrt{\frac{n_*}{2\pi\sigma^2(1+n_*)}}\,\mathbf{exp}\left[-\frac{n_*(\tilde{x}-\mu_*)^2}{2\sigma^2(1+n_*)}\right] \\
&\quad \times \underbrace{\int_{-\infty}^{\infty}\sqrt{\frac{1+n_*}{2\pi\sigma^2}}\,\mathbf{exp}\left[-\frac{1+n_*}{2\sigma^2}\left(\mu-\frac{\tilde{x}+n_*\mu_*}{1+n_*}\right)^2\right]d\mu}_{1} \\
&= \sqrt{\frac{n_*}{2\pi\sigma^2(1+n_*)}}\,\mathbf{exp}\left[-\frac{n_*(\tilde{x}-\mu_*)^2}{2\sigma^2(1+n_*)}\right].
\end{aligned}
$$

Predictive Distribution iv

Substituting this result for $g(\sigma^2)$, we have

$$
\begin{aligned}
p(\tilde{x}|D) &= \sqrt{\frac{n_*}{2\pi(1+n_*)}}\frac{\left(\frac{\lambda_*}{2}\right)^{\frac{\nu_*}{2}}}{\Gamma\left(\frac{\nu_*}{2}\right)} \\
&\quad \times \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_*+1}{2}+1\right)} \exp\left[-\frac{\frac{n_*}{1+n_*}(\tilde{x}-\mu_*)^2+\lambda_*}{2\sigma^2}\right] d\sigma^2 \\
&= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)}\sqrt{\frac{n_*}{\pi\lambda_*(1+n_*)}}\left[1+\frac{n_*(\tilde{x}-\mu_*)^2}{\lambda_*(1+n_*)}\right]^{-\frac{\nu_*+1}{2}} \\
&= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)\sqrt{\pi\nu_*\tau_*^2(1+n_*)}}\left[1+\frac{(\tilde{x}-\mu_*)^2}{\nu_*\tau_*^2(1+n_*)}\right]^{-\frac{\nu_*+1}{2}},
\end{aligned}
$$

where $\tau_*^2 = \frac{\lambda_*}{\nu_* n_*}$. ■

Regression Model i

Suppose the conditional expectation of y_i is a linear combination of k independent variables $(x_{1i}, \ldots, x_{ki})$, i.e.,

$$\mathbf{E}[y_i|x_i] = \mathbf{E}[y_i|x_{1i}, \ldots, x_{ki}] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}.$$

In practice, we almost always put the constant term β_0 in the model. By adding the error term, we have a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i, \ \epsilon_i \sim \mathbf{Normal}(0, \sigma^2).$$

This is sometimes called the multiple regression model.

Regression Model ii

Now define

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Then the regression model is summarized as

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathbf{Normal}(0, \sigma^2 \mathbf{I}).$$

The unknown parameters in the regression model are $\theta = (\beta, \sigma^2)$.

Prior Distribution

We assume the following prior distribution:

$$p(\theta) = p(\beta|\sigma^2)p(\sigma^2),$$

$$p(\beta|\sigma^2) = (2\pi\sigma^2)^{-\frac{k+1}{2}}|A_0|^{\frac{1}{2}}\,\mathbf{exp}\left[-\frac{1}{2\sigma^2}(\beta-\beta_0)'A_0(\beta-\beta_0)\right],$$

$$p(\sigma^2) = \frac{\left(\frac{\lambda_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}(\sigma^2)^{-(\frac{\nu_0}{2}+1)}\,\mathbf{exp}\left[-\frac{\lambda_0}{2\sigma^2}\right],$$

which implies that the conditional prior of $\boldsymbol{\beta}$ given $\boldsymbol{\sigma}^2$ is a multivariate normal distribution $\mathbf{Normal}\left(\boldsymbol{\beta}_0, \boldsymbol{\sigma}^2 A_0^{-1}\right)$ while the prior of $\boldsymbol{\sigma}^2$ is $\mathbf{Inv.Gamma}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right)$.

Posterior Distribution i

The likelihood is

$$p(y|X,\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right].$$

By applying Bayes' theorem, we have

$$\begin{aligned} p(\theta|D) &\propto p(y|X,\theta)p(\theta) \\ &\propto (\sigma^2)^{-\left(\frac{\nu_0+n+k+1}{2}+1\right)} \exp\left[-\frac{Q}{2\sigma^2}\right], \\ Q &= (y - X\beta)'(y - X\beta) + (\beta - \beta_0)'A_0(\beta - \beta_0) + \lambda_0. \end{aligned}$$

Posterior Distribution ii

By completing the square, we have

$$
\begin{aligned}
&(y - X\beta)'(y - X\beta) + (\beta - \beta_0)'A_0(\beta - \beta_0) \\
&\quad = (\beta - \beta_\star)'A_\star(\beta - \beta_\star) + (y - \hat{y})'(y - \hat{y}) \\
&\qquad + (\beta_0 - \hat{\beta})'C_\star(\beta_0 - \hat{\beta}),
\end{aligned}
$$

where

$$
\begin{aligned}
&\hat{\beta} = (X'X)^{-1}X'y, \quad \beta_\star = (X'X + A_0)^{-1}(X'y + A_0\beta_0), \\
&A_\star = X'X + A_0, \quad \hat{y} = X\hat{\beta}, \quad C_\star = ((X'X)^{-1} + A_0^{-1})^{-1}.
\end{aligned}
$$

Posterior Distribution iii

Therefore, the posterior distribution of $\boldsymbol{\theta}$ is obtained as

$$p(\theta|D) \propto (\sigma^2)^{-\frac{k+1}{2}} \exp\left[-\frac{1}{2\sigma^2}(\beta - \beta_\star)' A_\star (\beta - \beta_\star)\right]$$
$$\times (\sigma^2)^{-(\frac{\nu_\star}{2}+1)} \exp\left[-\frac{\lambda_\star}{2\sigma^2}\right],$$

where

$$\nu_\star = n + \nu_0,$$
$$\lambda_\star = (y - \hat{y})'(y - \hat{y}) + (\beta_0 - \hat{\beta})' C_\star (\beta_0 - \hat{\beta}) + \lambda_0.$$

Posterior Distribution iv

This posterior distribution is equivalent to

$$p(\theta|D) = p(\beta|\sigma^2, D)p(\sigma^2|D),$$

$$p(\beta|\sigma^2, D) = (2\pi\sigma^2)^{-\frac{k+1}{2}}|A_\star|^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(\beta-\beta_\star)'A_\star(\beta-\beta_\star)\right],$$

$$p(\sigma^2|D) = \frac{\left(\frac{\lambda_\star}{2}\right)^{\frac{\nu_\star}{2}}}{\Gamma\left(\frac{\nu_\star}{2}\right)}(\sigma^2)^{-\left(\frac{\nu_\star}{2}+1\right)} \exp\left[-\frac{\lambda_\star}{2\sigma^2}\right],$$

that is,

$$\beta|\sigma^2, D \sim \mathbf{Normal}\left(\beta_\star, \sigma^2 A_\star^{-1}\right),\ \sigma^2|D \sim \mathbf{Inv.Gamma}\left(\frac{\nu_\star}{2}, \frac{\lambda_\star}{2}\right).$$

Proof Of Completing The Square i

Note that

$$
\begin{aligned}
&(y - X\beta)'(y - X\beta) \\
&= (y - X\hat{\beta} + X\hat{\beta} - X\beta)'(y - X\hat{\beta} + X\hat{\beta} - X\beta) \\
&= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\
&\quad - 2\underbrace{(y - X\hat{\beta})'X}_{=0}(\beta - \hat{\beta}) \\
&= (y - \hat{y})'(y - \hat{y}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}).
\end{aligned}
$$

Proof Of Completing The Square ii

Then

$$
\begin{aligned}
&(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) + (\beta - \beta_0)' A_0 (\beta - \beta_0) \\
&= \beta' (X'X + A_0)\beta - 2(X'X\hat{\beta} + A_0\beta_0)'\beta + \hat{\beta}' X' X \hat{\beta} + \beta_0' A_0 \beta_0 \\
&= (\beta - (X'X + A_0)^{-1}(X'X\hat{\beta} + A_0\beta_0))' (X'X + A_0)(\beta - \beta_\star) \\
&\quad - (X'X\hat{\beta} + A_0\beta_0)' (X'X + A_0)^{-1} (X'X\hat{\beta} + A_0\beta_0) \\
&\quad + \hat{\beta}' X' X \hat{\beta} + \beta_0' A_0 \beta_0 \\
&= (\beta - \beta_\star)' A_\star (\beta - \beta_\star) + \hat{\beta}' (X'X - X'X(X'X + A_0)^{-1} X'X)\hat{\beta} \\
&\quad + \beta_0' (A_0 - A_0 (X'X + A_0)^{-1} A_0)\beta_0 - 2\hat{\beta}' X'X(X'X + A_0)^{-1} A_0 \beta_0.
\end{aligned}
$$

Proof Of Completing The Square iii

Since $(A+B)^{-1} = A^{-1} - A^{-1}(B^{-1}+A^{-1})^{-1}A^{-1}$, we have

$$\begin{aligned} C_\star = ((X'X)^{-1} + A_0^{-1})^{-1} &= X'X - X'X(X'X + A_0)^{-1}X'X \\ &= A_0 - A_0(X'X + A_0)^{-1}A_0. \end{aligned}$$

Furthermore

$$\begin{aligned} X'X(X'X + A_0)^{-1}A_0 &= (A_0^{-1}(X'X + A_0)(X'X)^{-1})^{-1} \\ &= ((X'X)^{-1} + A_0^{-1})^{-1} = C_\star. \end{aligned}$$

In sum, we have

$$\begin{aligned} &(y - X\beta)'(y - X\beta) + (\beta - \beta_0)'A_0(\beta - \beta_0) \\ &= (\beta - \beta_\star)'A_\star(\beta - \beta_\star) + (y - \hat{y})'(y - \hat{y}) + (\beta_0 - \hat{\beta})'C_\star(\beta_0 - \hat{\beta}). \blacksquare \end{aligned}$$

Marginal Posterior Distribution i

The marginal posterior distribution of regression coefficients $\boldsymbol{\beta}$ is obtained by integrating out $\boldsymbol{\sigma}^2$ as

$$p(\boldsymbol{\beta}|D) = \int_0^\infty p(\boldsymbol{\beta}|\sigma^2, D)p(\sigma^2|D)d\sigma^2.$$

$$\int_0^\infty (\sigma^2)^{-\left(\frac{\nu_\star+k+1}{2}+1\right)} \mathbf{exp}\left[-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)'A_\star(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)+\lambda_\star}{2\sigma^2}\right] d\sigma^2$$

$$= \left[\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)'A_\star(\boldsymbol{\beta}-\boldsymbol{\beta}_\star)+\lambda_\star}{2}\right]^{-\frac{\nu_\star+k+1}{2}} \boldsymbol{\Gamma}\left(\frac{\nu_\star+k+1}{2}\right).$$

Marginal Posterior Distribution ii

Thus $p(\beta|D)$ is derived as

$$
\begin{aligned}
p(\beta|D) &= \int_0^\infty p(\beta|\sigma^2, D)p(\sigma^2|D)d\sigma^2 \\
&= \frac{\Gamma\left(\frac{\nu_\star+k+1}{2}\right)}{\Gamma\left(\frac{\nu_\star}{2}\right)\pi^{\frac{k+1}{2}}}|\lambda_\star A_\star^{-1}|^{-\frac{1}{2}} \\
&\quad \times \left[1+\frac{1}{\lambda_\star}(\beta-\beta_\star)'A_\star(\beta-\beta_\star)\right]^{-\frac{\nu_\star+k+1}{2}},
\end{aligned}
$$

which is the multivariate t distribution

$$
\mathrm{t}(\beta_\star, \lambda_\star A_\star^{-1}, \nu_\star),\ \mathrm{E}[\beta] = \beta_\star,\ \mathrm{Var}[\beta] = \frac{\lambda_\star}{\nu_\star - 2}A_\star^{-1}.
$$

Marginal Posterior Distribution iii

Let us derive the posterior distribution of $\boldsymbol{\beta}_j$ $(j = 1, \ldots, k+1)$. First of all, let $H = \frac{\lambda_\star}{\nu_\star} A_\star^{-1}$ and consider the following partition of $\boldsymbol{\beta}_\star$ and H:

$$\boldsymbol{\beta}_\star = \begin{bmatrix} \boldsymbol{\beta}_{\star j} \\ \boldsymbol{\beta}_{\star \cdot} \end{bmatrix}, \quad H = \begin{bmatrix} h_{jj} & h'_{j\cdot} \\ h_{j\cdot} & H_{\cdot\cdot} \end{bmatrix}.$$

Using the property of the multivariate t distribution, we have

$$\left. \frac{\boldsymbol{\beta}_j - \boldsymbol{\beta}_{\star j}}{\sqrt{h_{jj}}} \right| D \sim \mathrm{t}(\nu_\star).$$

Bayesian Inference On Regression Coefficients i

- **Point estimation:** The point estimate of β_j is $\beta_{\star j}$ for the {quadratic, absolute, 0-1} loss since the t distribution is symmetric around $\beta_{\star j}$
- **Interval estimation:** Consider $t_{\frac{p}{2}}$ such that

$$\Pr\left\{\frac{\beta_j - \beta_{\star j}}{\sqrt{h_{jj}}} > t_{\frac{p}{2}}\right\} = \frac{p}{2}.$$

Then the $100(1-p)\%$ {credible, HPD} interval is given by

$$\beta_{\star j} - t_{\frac{p}{2}}\sqrt{h_{jj}} \leq \beta_j \leq \beta_{\star j} + t_{\frac{p}{2}}\sqrt{h_{jj}}.$$

Bayesian Inference On Regression Coefficients ii

- Test of significance:

$$H_0: \ \beta_j = 0; \quad H_1: \ \beta_j \neq 0.$$

can be tested with SDDR:

$$\text{SDDR} = \frac{f_t(0|\beta_{\star j}, h_{jj}, \nu_\star)}{f_t(0|\beta_{0j}, \lambda_0 a_{jj}/\nu_0, \nu_0)},$$

where $f_t(\cdot)$ is the p.d.f. of the t distribution.

Derivation Of SDDR i

The marginal prior distribution of $\boldsymbol{\beta}_j$ is derived as

$$\mathrm{t}\left(\beta_{j0}, \frac{\lambda_0 a_{jj}}{\nu_0}, \nu_0\right).$$

Let $f_t(\cdot|\beta_{0j}, \lambda_0 a_{jj}/\nu_0, \nu_0)$ denote the p.d.f. of the above t distribution. Consider a conditional spike-and-slab prior:

$$p(\beta_j|\beta_{-j}, \sigma^2) = q_0\delta(\beta_j) + (1 - q_0)f(\beta_j|\beta_{-j}, \sigma^2),$$

where $\boldsymbol{\beta}_{-j}$ is the set of regression coefficients except for $\boldsymbol{\beta}_j$ and $f(\beta_j|\beta_{-j}, \sigma^2)$ is the p.d.f. of the conditional distribution of $\boldsymbol{\beta}_j$ in $\mathbf{Normal}(\boldsymbol{\beta}_0, \sigma^2 A_0^{-1})$ given $\boldsymbol{\beta}_{-j}$.

Derivation Of SDDR ii

Then we have

$$
\begin{aligned}
&p(\beta_j|D) \\
&= \frac{\iint p(D|\beta,\sigma^2)p(\beta_j|\beta_{-j},\sigma^2)p(\beta_{-j}|\sigma^2)p(\sigma^2)d\beta_{-j}d\sigma^2}{\iiint p(D|\beta,\sigma^2)p(\beta_j|\beta_{-j},\sigma^2)p(\beta_{-j}|\sigma^2)p(\sigma^2)d\beta_j d\beta_{-j}d\sigma^2} \\
&= \frac{q_0 m(D|\beta_j=0)\delta(\beta_j) + (1-q_0)m(D|\beta_j)f(\beta_j)}{q_0 m(D|\beta_j=0) + (1-q_0)m(D)},
\end{aligned}
$$

$$
f(\beta_j) = \iint f(\beta_j|\beta_{-j},\sigma^2)p(\beta_{-j}|\sigma^2)p(\sigma^2)d\beta_{-j}d\sigma^2,
$$

$$
m(D) = \iint p(D|\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2)d\beta d\sigma^2,
$$

$$
m(D|\beta_j) = \iint p(D|\beta,\sigma^2)p(\beta_{-j}|\sigma^2)p(\sigma^2)d\beta_{-j}d\sigma^2.
$$

Derivation Of SDDR iii

Since

$$\mathbf{Pr}\{\beta_j = 0|D\} \propto q_0 m(D|\beta_j = 0),$$
$$\mathbf{Pr}\{\beta_j \neq 0|D\} \propto (1 - q_0)m(D),$$

the Bayes factor is derived as

$$\begin{aligned} B_{01}(D) &= \frac{\mathbf{Pr}\{\beta_j = 0|D\}}{\mathbf{Pr}\{\beta_j \neq 0|D\}} \div \frac{\mathbf{Pr}\{\beta_j = 0\}}{\mathbf{Pr}\{\beta_j \neq 0\}} \\ &= \frac{q_0 m(D|\beta_j = 0)}{(1 - q_0)m(D)} \div \frac{q_0}{1 - q_0} = \frac{m(D|\beta_j = 0)}{m(D)}. \end{aligned}$$

Derivation Of SDDR iv

Furthermore, by using the following relationship

$$f(\beta_j = 0|D) = \frac{m(D|\beta_j = 0)f(\beta_j = 0)}{m(D)},$$

we have

$$B_{01}(D) = \frac{m(D|\beta_j = 0)}{m(D)} = \frac{f(\beta_j = 0|D)}{f(\beta_j = 0)}.$$

Since $f(\cdot)$ is the p.d.f. of the t distribution, we obtain the SDDR. ■

Marginal Likelihood

The marginal likelihood of the regression model is given by

$$m(D) = \int p(y|X,\theta)p(\theta)d\theta = \frac{p(y|X,\theta)p(\theta)}{p(\theta|D)},$$

where $D = (y, X)$. Therefore we obtain the marginal likelihood as

$$m(D) = \frac{\Gamma\left(\frac{\nu_\star}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\pi^{\frac{n}{2}}}\sqrt{\frac{\lambda_0^{\nu_0}|A_0|}{\lambda_\star^{\nu_\star}|A_\star|}}.$$

Predictive Distribution Of Regression i

Suppose we have p sets of independent variables and we want to predict the corresponding p dependent variables:

$$\tilde{X} = \begin{bmatrix} 1 & \tilde{x}_{11} & \cdots & \tilde{x}_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{x}_{1p} & \cdots & \tilde{x}_{kp} \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_p \end{bmatrix}.$$

Then we may consider an auxiliary regression model:

$$\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathbf{Normal}(0, \sigma^2 \mathbf{I}),$$

Predictive Distribution Of Regression ii

which implies that $\tilde{y}$ follows $\mathbf{Normal}(\tilde{X}\beta, \sigma^2\mathbf{I})$. So the p.d.f. of $\tilde{y}$ is

$$p(\tilde{y}|\tilde{X}, \theta) = (2\pi\sigma^2)^{-\frac{p}{2}} \mathbf{exp}\left[-\frac{1}{2\sigma^2}(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta)\right].$$

The predictive distribution of $\tilde{y}$ is defined as

$$\begin{aligned} p(\tilde{y}|D) &= \int p(\tilde{y}|\tilde{X}, \theta)p(\theta|D)d\theta \\ &= \int \left\{\int p(\tilde{y}|\tilde{X}, \theta)p(\beta|\sigma^2, D)d\beta\right\} p(\sigma^2|D)d\sigma^2. \end{aligned}$$

Predictive Distribution Of Regression iii

This is the multivariate t distribution

$$\mathrm{t}(\tilde{X}\beta_\star, \lambda_\star \Omega_\star, \nu_\star),$$

where

$$\Omega_\star = \mathrm{I} + \tilde{X} A_\star^{-1} \tilde{X}', \quad \bar{\beta}_\star = (\tilde{X}'\tilde{X} + A_\star)^{-1}(\tilde{X}'\tilde{y} + A_\star \beta_\star).$$

$$\mathrm{E}[\tilde{y}|D] = \tilde{X}\beta_\star = \tilde{X}\mathrm{E}[\beta|D],$$

$$\mathrm{Var}[\tilde{y}|D] = \frac{\lambda_\star}{\nu_\star - 2}\Omega_\star = \mathrm{E}[\sigma^2|D]\mathrm{I} + \tilde{X}\mathrm{Var}[\beta|D]\tilde{X}',$$

$$\mathrm{E}[\sigma^2|D] = \frac{\lambda_\star}{\nu_\star - 2}, \quad \mathrm{Var}[\beta|D] = \frac{\lambda_\star}{\nu_\star - 2} A_\star^{-1}.$$

Derivation Of The Predictive Distribution i

First, we complete the square again.

$$
\begin{aligned}
&(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) + (\beta - \beta_\star)' A_\star (\beta - \beta_\star) \\
&= \beta'(\tilde{X}'\tilde{X} + A_\star)\beta - 2(\tilde{X}'\tilde{y} + A_\star\beta_\star)'\beta + \tilde{y}'\tilde{y} + \beta_\star' A_\star \beta_\star \\
&= (\beta - (\tilde{X}'\tilde{X} + A_\star)^{-1}(\tilde{X}'\tilde{y} + A_\star\beta_\star))'(\tilde{X}'\tilde{X} + A_\star)(\beta - \bar{\beta}_\star) \\
&\quad - (\tilde{X}'\tilde{y} + A_\star\beta_\star)'(\tilde{X}'\tilde{X} + A_\star)^{-1}(\tilde{X}'\tilde{y} + A_\star\beta_\star) + \tilde{y}'\tilde{y} + \beta_\star' A_\star \beta_\star \\
&= (\beta - \bar{\beta}_\star)'(\tilde{X}'\tilde{X} + A_\star)(\beta - \bar{\beta}_\star) \\
&\quad + \tilde{y}'(\mathbf{I} - \tilde{X}(A_\star + \tilde{X}'\tilde{X})^{-1}\tilde{X}')\tilde{y} \\
&\quad + \beta_\star'(A_\star - A_\star(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star)\beta_\star \\
&\quad - 2\tilde{y}'\tilde{X}(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star\beta_\star.
\end{aligned}
$$

Derivation Of The Predictive Distribution ii

To complete the square of the last three terms, we use the following formulae:

$$\mathbf{I} - \tilde{X}(A_\star + \tilde{X}'\tilde{X})^{-1}\tilde{X}' = (\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1},$$
$$(A_\star + \tilde{X}'\tilde{X})^{-1} = A_\star^{-1} - A_\star^{-1}\tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}A_\star^{-1}.$$

By multiplying both sides of the second equation by $A_\star$ from the left and the right, we have

$$\begin{aligned} A_\star(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star &= A_\star(A_\star^{-1} - A_\star^{-1}\tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}A_\star^{-1})A_\star \\ &= A_\star - \tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}. \end{aligned}$$

Derivation Of The Predictive Distribution iii

Thus

$$A_\star - A_\star(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star = \tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}.$$

In the same manner, we have

$$\begin{aligned}
&\tilde{X}(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star \\
&= \tilde{X}(A_\star^{-1} - A_\star^{-1}\tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}A_\star^{-1})A_\star \\
&= \tilde{X} - \tilde{X}A_\star^{-1}\tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X} \\
&= \tilde{X} - (\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}' - \mathbf{I})(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X} \\
&= \tilde{X} - \tilde{X} + (\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X} \\
&= (\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}.
\end{aligned}$$

Derivation Of The Predictive Distribution iv

Therefore

$$
\begin{aligned}
&\tilde{y}'(I - \tilde{X}(A_\star + \tilde{X}'\tilde{X})^{-1}\tilde{X}')\tilde{y} + \beta_\star'(A_\star - A_\star(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star)\beta_\star \\
&\quad - 2\tilde{y}'\tilde{X}(A_\star + \tilde{X}'\tilde{X})^{-1}A_\star\beta_\star \\
&= \tilde{y}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{y} + \beta_\star'\tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}\beta_\star \\
&\quad - 2\tilde{y}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}\beta_\star \\
&= \tilde{y}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{y} + \beta_\star'\tilde{X}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}\beta_\star \\
&\quad - 2\tilde{y}'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}\tilde{X}\beta_\star \\
&= (\tilde{y} - \tilde{X}\beta_\star)'(\mathbf{I} + \tilde{X}A_\star^{-1}\tilde{X}')^{-1}(\tilde{y} - \tilde{X}\beta_\star).
\end{aligned}
$$

Derivation Of The Predictive Distribution v

Finally, we have

$$
\begin{aligned}
&p(\tilde{y}|\tilde{X},\theta)p(\beta|\sigma^2,D) \\
&= (2\pi\sigma^2)^{-\frac{p+k+1}{2}}|A_\star|^{\frac{1}{2}}\,\mathbf{exp}\left[-\frac{1}{2\sigma^2}(\tilde{y}-\tilde{X}\beta_\star)'\mathbf{\Omega}_\star^{-1}(\tilde{y}-\tilde{X}\beta_\star)\right] \\
&\quad\times\mathbf{exp}\left[-\frac{1}{2\sigma^2}(\beta-\bar{\beta}_\star)'(\tilde{X}'\tilde{X}+A_\star)(\beta-\bar{\beta}_\star)\right].
\end{aligned}
$$

Since

$$
\begin{aligned}
&\int \mathbf{exp}\left[-\frac{1}{2\sigma^2}(\beta-\bar{\beta}_\star)'(\tilde{X}'\tilde{X}+A_\star)(\beta-\bar{\beta}_\star)\right]d\beta \\
&\quad= (2\pi)^{\frac{k+1}{2}}|\sigma^{-2}(\tilde{X}'\tilde{X}+A_\star)|^{-\frac{1}{2}},
\end{aligned}
$$

Derivation Of The Predictive Distribution vi

we have

$$\int p(\tilde{y}|\tilde{X},\theta)p(\beta|\sigma^2,D)d\beta$$
$$= (2\pi\sigma^2)^{-\frac{p}{2}}|\mathbf{\Omega}_\star|^{-\frac{1}{2}}\,\mathbf{exp}\left[-\frac{1}{2\sigma^2}(\tilde{y}-\tilde{X}\beta_\star)'\mathbf{\Omega}_\star^{-1}(\tilde{y}-\tilde{X}\beta_\star)\right].$$

Here we utilize $|A_\star^{-1}||\tilde{X}'\tilde{X}+A_\star| = |\mathbf{I}+\tilde{X}A_\star^{-1}\tilde{X}'| = |\mathbf{\Omega}_\star|$. The above expression is the p.d.f. of $\mathbf{Normal}(\tilde{X}\beta_\star,\sigma^2\mathbf{\Omega}_\star)$.

Derivation Of The Predictive Distribution vii

Next, we evaluate the integral with respect to $\boldsymbol{\sigma}^2$. Then we have

$$
\begin{aligned}
p(\tilde{y}|D) &= \int \left\{ \int p(\tilde{y}|\tilde{X}, \theta) p(\beta|\sigma^2, D) d\beta \right\} p(\sigma^2|D) d\sigma^2 \\
&= \int (2\pi\sigma^2)^{-\frac{p}{2}} |\boldsymbol{\Omega}_\star|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (\tilde{y} - \tilde{X}\beta_\star)' \boldsymbol{\Omega}_\star^{-1} (\tilde{y} - \tilde{X}\beta_\star) \right] \\
&\quad \times \frac{\left(\frac{\lambda_\star}{2}\right)^{\frac{\nu_\star}{2}}}{\Gamma\left(\frac{\nu_\star}{2}\right)} (\sigma^2)^{-\left(\frac{\nu_\star}{2}+1\right)} \exp\left[-\frac{\lambda_\star}{2\sigma^2} \right] d\sigma^2 \\
&= \frac{\Gamma\left(\frac{\nu_\star+p}{2}\right)}{\Gamma\left(\frac{\nu_\star}{2}\right) \pi^{\frac{p}{2}}} |\lambda_\star \boldsymbol{\Omega}_\star|^{-\frac{1}{2}} \left[1 + \frac{1}{\lambda_\star} (\tilde{y} - \tilde{X}\beta_\star)' \boldsymbol{\Omega}_\star^{-1} (\tilde{y} - \tilde{X}\beta_\star) \right]^{-\frac{\nu_\star+p}{2}} .
\end{aligned}
$$

Derivation Of The Predictive Distribution viii

For this derivation, we use

$$\int (\sigma^2)^{-\left(\frac{\nu_\star+p}{2}+1\right)} \mathbf{exp}\left[-\frac{1}{2\sigma^2}\left\{\lambda_\star + (\tilde{y}-\tilde{X}\beta_\star)'\boldsymbol{\Omega}_\star^{-1}(\tilde{y}-\tilde{X}\beta_\star)\right\}\right] d\sigma^2$$

$$= \left\{\lambda_\star + (\tilde{y}-\tilde{X}\beta_\star)'\boldsymbol{\Omega}_\star^{-1}(\tilde{y}-\tilde{X}\beta_\star)\right\}^{-\frac{\nu_\star+p}{2}} \boldsymbol{\Gamma}\left(\frac{\nu_\star+p}{2}\right).$$

This is the p.d.f. of $\mathbf{t}(\tilde{X}\beta_\star, \lambda_\star\boldsymbol{\Omega}_\star, \nu_\star)$. ■

Summary Of Bayesian Analysis

- θ — $m \times 1$ vector of unknown parameters
- D — data
- $p(x|\theta)$ — population distribution
- $p(\theta)$ — prior distribution
- $p(D|\theta)$ — likelihood
- $p(\theta|D)$ — posterior distribution

Bayes' Theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\mathbb{R}^m} p(D|\theta)p(\theta)d\theta}.$$

Difficulties In Bayesian Analysis

1. In Bayesian analysis, we are required to evaluate multiple integrals to proceed statistical inference.
2. For example, we need $\int_{\mathbb{R}^m} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ to derive the exact expression of the posterior distribution, but it is not available except for a limited number of examples (e.g., natural conjugate prior).
3. It is the case for the following quantities:
   - Marginal distribution: $p(\boldsymbol{\theta}_j|D) = \int_{\mathbb{R}^{m-1}} p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}_{-j}$.
   - Mean: $\mathbf{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}_j|D] = \int_{-\infty}^{\infty} \boldsymbol{\theta}_j p(\boldsymbol{\theta}_j|D)d\boldsymbol{\theta}_j$.
   - Median: $\int_{-\infty}^{\text{Median}_{\boldsymbol{\theta}_j}} p(\boldsymbol{\theta}_j|D)d\boldsymbol{\theta}_j = \frac{1}{2}$.
   - Probability: $\mathbf{P}(a \leqq \boldsymbol{\theta}_j \leqq b|D) = \int_a^b p(\boldsymbol{\theta}_j|D)d\boldsymbol{\theta}_j$.
   - Predictive dist.: $p(\tilde{x}|D) = \int_{\mathbb{R}^m} p(\tilde{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$.

MAP Estimation

One exception is the posterior mode:

$$\text{Mode}_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^m} p(\boldsymbol{\theta}|D).$$

Note that the posterior mode is equivalent to

$$\text{MAP}_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^m} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

which does not depend on the normalizing constant. Since the exact expressions of the prior distribution and the likelihood are known in many applications, we can solve the above optimization problem without evaluating the normalizing constant.

The posterior mode is often called MAP (maximum a posteriori) estimate.

Monte Carlo Methods i

Monte Carlo methods are widely applied numerical techniques to evaluate integrals. Suppose we need to evaluate the following expectation:

$$\mathbf{E}_{\boldsymbol{\theta}}[h(\boldsymbol{\theta})] = \int_{\mathbb{R}^m} h(\boldsymbol{\theta}) p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}.$$

Examples:

- Mean $\mathbf{E}_{\boldsymbol{\theta}}[\theta_j|D]$: $h(\boldsymbol{\theta}) = \theta_j$.
- Variance $\mathbf{Var}_{\boldsymbol{\theta}}[\theta_j|D]$: $h(\boldsymbol{\theta}) = \left(\theta_j - \mathbf{E}_{\boldsymbol{\theta}}[\theta_j|D]\right)^2$.
- Probability $\mathbf{P}(a \leqq \theta_j \leqq b|D)$: $h(\boldsymbol{\theta}) = 1_{[a,b]}(\theta_j)$.
- Predictive distribution $p(\tilde{x}|D)$: $h(\boldsymbol{\theta}) = p(\tilde{x}|\boldsymbol{\theta})$.

Monte Carlo Methods ii

Suppose we have a set of pseudo-random numbers generated from the posterior distribution $p(\boldsymbol{\theta}|D)$, $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}\}$, which is called the Monte Carlo sample, and define

$$\hat{h} = \frac{1}{T} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)}).$$

If the law of large numbers holds,

$$\hat{h} \rightsquigarrow \mathbf{E}_{\boldsymbol{\theta}}[h(\boldsymbol{\theta})] \quad \text{as} \quad T \to \infty,$$

where "$\rightsquigarrow$" means the convergence in probability.

Monte Carlo Methods iii

Therefore, when T is sufficiently large, $\mathbf{E}_{\boldsymbol{\theta}}[h(\boldsymbol{\theta})]$ can be well approximated with $\hat{h}$. This is the basic idea behind the Monte Carlo integration method.

Monte Carlo approximations of the posterior statistics are given as follows:

- Mean: $\hat{\mathbf{E}}_{\boldsymbol{\theta}}[\theta_j|D] = \frac{1}{T}\sum_{t=1}^{T}\theta_j^{(t)}$.
- Variance: $\widehat{\mathbf{Var}}_{\boldsymbol{\theta}}[\theta_j|D] = \frac{1}{T}\sum_{t=1}^{T}\left(\theta_j^{(t)} - \hat{\mathbf{E}}_{\boldsymbol{\theta}}[\theta_j|D]\right)^2$.
- Probability: $\widehat{\mathbf{P}}(a \leqq \theta_j \leqq b|D) = \frac{1}{T}\sum_{t=1}^{T}1_{[a,b]}(\theta_j^{(t)})$.
- Predictive distribution: $\hat{p}(\tilde{x}|D) = \frac{1}{T}\sum_{t=1}^{T}p(\tilde{x}|\boldsymbol{\theta}^{(t)})$.

Monte Carlo Methods iv

Note that the Monte Carlo method merely gives us an approximated value of the true posterior statistic (the true one is obtained when $T \to \infty$). These errors are measured by

Numerical errors in Monte Carlo approximation

$$\mathbf{SE}[\hat{h}] = \sqrt{\frac{1}{T(T-1)} \sum_{t=1}^{T} \left(h(\theta^{(t)}) - \hat{h} \right)^2}.$$

Basically this is equivalent to the standard error of the sample mean in statistics.

Kernel Density Estimation

The marginal posterior p.d.f. $p(\theta_j|D)$ can be evaluated with a kernel density estimation method:

$$\hat{p}(\theta_j) = \frac{1}{Tw}\sum_{t=1}^{T} K\left(\frac{\theta_j - \theta_j^{(t)}}{w}\right),$$

where $K(\cdot)$ is a kernel function. In practice, the p.d.f. of the standard normal distribution:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

is often used as $K(\cdot)$, which is called the Gaussian kernel. w is the bandwidth and determines the smoothness of the evaluated density function.

Monte Carlo Approximation Of Quantiles i

Quantile

The $100q\%$ quantile $\boldsymbol{\theta}_j^{[q]}$ is defined as

$$\mathbf{P}(\boldsymbol{\theta}_j \leqq \boldsymbol{\theta}_j^{[q]} | D) = q, \quad 0 < q < 1.$$

We can approximate the quantile $\boldsymbol{\theta}_j^{[q]}$ with

Sample Quantile

$$\hat{\boldsymbol{\theta}}_j^{[q]} = \max_{1 \leqq t \leqq T} \boldsymbol{\theta}_j^{(t)} \quad \text{s.t.} \quad \frac{1}{T} \sum_{s=1}^{T} 1_{\left(-\infty,\, \boldsymbol{\theta}_j^{(t)}\right)} \left(\boldsymbol{\theta}_j^{(s)}\right) \leqq q,$$

Monte Carlo Approximation Of Quantiles ii

For example, when $T = 10,000$, $\hat{\theta}_j^{[0.05]}$ is the 500th smallest value in the Monte Carlo sample.

- Median: $\hat{\theta}_j^{[0.5]}$.
- $100(1-q)\%$ credible interval: $\hat{\theta}_j^{[\frac{q}{2}]} \leqq \theta_j \leqq \hat{\theta}_j^{[1-\frac{q}{2}]}$.
- $100(1-q)\%$ HPDI [Chen and Shao (1999)]:

$$\hat{\theta}_j^{\left[\frac{t^*}{T}\right]} \leqq \theta_j \leqq \hat{\theta}_j^{\left[1-q+\frac{t^*}{T}\right]}, \quad 1 \leqq t \leqq qT,$$

where

$$t^* = \arg\min_{1 \leqq t \leqq qT} \left| \hat{\theta}_j^{[\frac{t}{T}]} - \hat{\theta}_j^{[1-q+\frac{t}{T}]} \right|.$$

Markov Chain Monte Carlo (MCMC)

- So far we suppose we have the Monte Carlo sample generated from the posterior distribution.
- However it is not obvious how to generate random numbers from an unknown distribution.
- For this purpose, we apply Markov chain sampling methods (e.g., Gibbs sampler, Metropolis-Hastings algorithm, Hamiltonian Monte Carlo method).
- Markov chain sampling + Monte Carlo integration = Markov chain Monte Carlo (MCMC)
- I will not cover these methods because they are so complicated. Please read books in the reading list.
- Instead we use PyMC for MCMC implementation.

Convergence Diagnostics

Unfortunately Markov chain sampling does not always work, even though the distribution of generated random numbers must converge to the posterior distribution in theory.

To make it work properly, we need careful tuning and monitoring. PyMC can automatically tune the sampling algorithm to some extent. So we discuss how to check the convergence of generated chains.

- n — the number of draws
- m — the number of chains
- θ_{ij} — draw i in chain j $(i = 1, \ldots, n,\ j = 1, \ldots, m)$
- $\{\theta_1, \ldots, \theta_T\}$ — Monte Carlo sample $(T = m \times n)$

Numerical Inefficiency

Effective sample size

$$\hat{T}_e = \frac{T}{1 + 2\sum_{s=1}^{S} \hat{\rho}_s},$$

where

$$\hat{\rho}_s = \frac{\hat{\gamma}_s}{\hat{\gamma}_0},\ \hat{\gamma}_s = \frac{1}{T}\sum_{t=s+1}^{T}(\theta_t - \bar{\theta})(\theta_{t-s} - \bar{\theta}),\ \bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t.$$

$\hat{\rho}_s$ is called the autocorrelation (of lag s). If a chain of random numbers are independent, all autocorrelations are zeros. In practice, we often observe strong and persistent positive autocorrelations in the chain, which leads to a smaller $\hat{T}_e$.

Gelman-Rubin Convergence Diagnostic

Gelman-Rubin convergence diagnostic $\hat{R}$

$$\hat{R} = \sqrt{\frac{\hat{V}}{\hat{W}}}, \quad \hat{V} = \frac{n-1}{n}\hat{W} + \frac{1}{n}\hat{B},$$

$$\hat{B} = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_{\cdot j} - \bar{\theta})^2, \quad \bar{\theta}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij},$$

$$\hat{W} = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{\cdot j})^2.$$

If the mean of each chain is equal to each other, $\frac{1}{n}\hat{B}$ will be close to zero when n is sufficiently large. Therefore $\hat{R}$ must be close to one; otherwise, the convergence is doubtful.

Non-Standard Dependent Variables

In the regression model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i,$$

we implicitly suppose the dependent variable y_i in the regression model is continuous and can be either positive or negative. In other words, y_i is supposed to be a real-valued random variable.

As long as the normality assumption is valid and the conditional expectation of y_i is expressed as a linear function:

$$\mathrm{E}[y_i|x_i] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

this assumption seems benign. In some cases, however, it is impractical to assume so.

Binary Data

Suppose y_i takes either 1 or 0 with the constant probability, i.e.,

$$y_i = \begin{cases} 1, & \text{with probability } q_i; \\ 0, & \text{with probability } 1 - q_i, \end{cases}$$

which is called a Bernoulli distribution. This type of data appears in empirical analysis on decision making (e.g., consumer's choice) or events (e.g., bankruptcy).

In this case, the conditional expectation of y_i is q_i, the conditional probability $\mathbf{P}(y_i = 1|x_i)$ itself. Since $0 \leqq q_i \leqq 1$ by definition, it is unrealistic to assume that q_i is a linear function of x_i.

Count Data

Suppose y_i takes non-negative integers, $0, 1, 2, \ldots$. This type of data is used in analyzing occurrences of rare events such as traffic accidents, crimes, mechanical failures, and so forth.

It is often assume that y_i follows a Poisson distribution:

$$\mathbf{P}(y_i = y|x_i) = \frac{e^{-\lambda_i}\lambda_i^y}{y!}, \; y = 0, 1, 2, \ldots, \; \lambda_i > 0.$$

In this case, the conditional expectation of y_i is λ_i, which must be positive. Thus the linearity assumption on the conditional expectation of y_i seems inappropriate.

Generalized Linear Model

Let $\boldsymbol{\mu}_i$ denote the conditional expectation $\mathbf{E}[y_i|x_i]$ and

$$x_i^\intercal \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

for brevity. To relax the limitation of the linearity assumption $\boldsymbol{\mu}_i = x_i^\intercal \boldsymbol{\beta}$, we introduce a transformation of the conditional expectation $g(\cdot)$:

$$g(\mu_i) = x_i^\intercal \boldsymbol{\beta} \quad \text{or} \quad \mu_i = g^{-1}(x_i^\intercal \boldsymbol{\beta}).$$

$g(\cdot)$ is called a link function and a regression-type model of the transformed conditional expectation is called a generalized linear model (GLM).

Link Functions

1. Logit link

$$\log \frac{\mu_i}{1-\mu_i} = x_i^\intercal \beta \quad \Leftrightarrow \quad \mu_i = \frac{1}{1+e^{-x_i^\intercal \beta}}.$$

2. Probit link

$$\Phi^{-1}(\mu_i) = x_i^\intercal \beta \quad \Leftrightarrow \quad \mu_i = \Phi(x_i^\intercal \beta),$$

where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$

3. Log link

$$\log \mu_i = x_i^\intercal \beta \quad \Leftrightarrow \quad \mu_i = e^{x_i^\intercal \beta}.$$

Remarks

1. Since $\boldsymbol{\mu}_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$ in the linear regression model, $g(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i$, which is called the linear link function.
2. Since both logit and probit link functions assure that $\boldsymbol{\mu}_i$ takes a value between 0 and 1, they are suitable for the binary data model.
3. When the probability p_i is transformed with the logit link function, such a GLM is called a logit model.
4. When the probability p_i is transformed with the probit link function, such a GLM is called a probit model.
5. Since the log link function assures that $\boldsymbol{\mu}_i$ is positive, it is used in the Poisson model of count data.

Binary Choice Model

A GLM of Bernoulli binary data with either logit or probit link function is often called a binary choice model, though the binary data are not necessarily related to decision making.

Examples

- Consumer's choice
  $y_i = 1$, if Consumer i owns an iPhone; 0, otherwise.
- Labor force participation
  $y_i = 1$, if Person i works; 0, otherwise.
- Bankruptcy
  $y_i = 1$, if Firm i goes bankrupt; 0, otherwise.

Likelihood in the Logit / Probit Model

Since the probability of y_i is expressed as

$$\mathbf{P}(y_i = y|x_i) = q_i^y(1-q_i)^{1-y},\ q_i = g^{-1}(x_i^\intercal \beta),\ y = 0, 1,$$

where g is either logit or probit link function. The joint probability of $D = (y_1, \ldots, y_n)$ is given by

$$p(D|\beta) = \prod_{i=1}^{n} q_i^{y_i}(1-q_i)^{1-y_i}.$$

This is the likelihood in the logit / probit model. Since this likelihood is a complicated non-linear function of $\boldsymbol{\beta}$, it is impractical to derive the closed form of the posterior distribution.

Poisson Regression Model

A GLM with Poisson count data with the log link function is called a Poisson regression model. Since the probability of y_i is expressed as

$$\mathbf{P}(y_i = y|x_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}, \ \lambda_i = e^{x_i^\top \beta}, \ y = 0, 1, 2, \ldots,$$

the likelihood is given by

$$p(D|\beta) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \propto \prod_{i=1}^{n} \lambda_i^{y_i} \cdot e^{-\sum_{i=1}^{n} \lambda_i}.$$

Since the above likelihood is a complicated non-linear function of $\boldsymbol{\beta}$, it is impractical to derive the closed form of the posterior distribution.