Fisher Information and Jeffreys Prior

Elvan Ceyhan

10/4/2024

In Bayesian statistics and parametric inference, **Fisher information** and **Jeffreys prior** are two closely related concepts that play an essential role in parameter estimation and model formulation. Below is an explanation of these topics, incorporating examples to clarify their applications.

1 Fisher Information

The **Fisher information** is a key concept in both frequentist and Bayesian statistics. It quantifies how much information an observable random variable X provides about an unknown parameter θ. Fisher information is fundamental to understanding the behavior of estimators and establishing the lower bound for their variance.

1.1 Definition

Fisher information $I(\theta)$ for a parameter θ is defined as:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\right)^2\right]$$

Alternatively, under certain regularity conditions, it can also be written as:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(X;\theta)\right]$$

where $f_X(X;\theta)$ is the likelihood function for the observed data X.

1.2 Intuition

Fisher information measures the *sensitivity* of the likelihood function to changes in the parameter θ. If small changes in θ lead to large changes in the likelihood, the parameter is well-identified, and the Fisher information is large. Conversely, if the likelihood is relatively flat with respect to θ, the Fisher information is small, meaning the data provides little information about the parameter.

1.3 Cramér-Rao Bound

Fisher information plays a critical role in the **Cramér-Rao inequality**, which establishes a lower bound on the variance of any unbiased estimator $\hat{\theta}$ for θ:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

This inequality implies that the Fisher information sets a limit on the efficiency of estimators: the more information the data contains about θ, the lower the variance of an unbiased estimator.

1.4 Example: Fisher Information for a Normal Distribution

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a normal distribution $X \sim N(\mu, \sigma^2)$. We will compute the Fisher information for both the mean μ and the variance σ^2.

1.5 Example 1: Normal Distribution (Mean Estimation)

Consider a normal distribution $X \sim N(\mu, \sigma^2)$, where μ is the unknown mean, and the variance σ^2 is known. We are interested in estimating the mean μ using Jeffreys prior, which is derived from the Fisher information.

1.5.1 Likelihood Function and Log-Likelihood

For a single observation X, the probability density function (PDF) is:

$$f_X(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The log-likelihood function for n independent and identically distributed (i.i.d.) observations $X_1, X_2, \ldots, X_n$ is:

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

1.5.2 First Derivative: The Score Function

The score function is the derivative of the log-likelihood function with respect to μ. It measures the sensitivity of the log-likelihood to changes in μ:

$$\frac{\partial}{\partial\mu}\log L(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

The score function indicates how changes in μ affect the likelihood. For example, when μ is close to the sample mean, the score will be close to zero, indicating that the likelihood function is maximized near this point.

1.5.3 Second Derivative: Information About μ

To calculate the Fisher information, we take the second derivative of the log-likelihood function with respect to μ. The second derivative of the log-likelihood is:

$$\frac{\partial^2}{\partial\mu^2}\log L(\mu) = -\frac{n}{\sigma^2}$$

This derivative is constant with respect to μ, meaning that the curvature of the log-likelihood function is the same for all values of μ. This uniform curvature implies that the amount of information about μ is independent of its specific value.

1.5.4 Fisher Information

The Fisher information is the negative expected value of the second derivative of the log-likelihood:

$$I(\mu) = -\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\log L(\mu)\right]$$

In this case, since the second derivative is constant and does not depend on the data, the Fisher information simplifies to:

$$I(\mu) = \frac{n}{\sigma^2}$$

For a single observation (i.e., $n = 1$), the Fisher information is:

$$I(\mu) = \frac{1}{\sigma^2}$$

This shows that the Fisher information depends inversely on the known variance σ^2: the smaller σ^2 is, the more information the data provides about μ, and vice versa. A larger variance indicates greater uncertainty in the data, reducing the amount of information about the mean.

1.5.5 Fisher Information for σ^2

We are interested in estimating the variance σ^2 of a normal distribution $X \sim N(\mu, \sigma^2)$, where μ is known, and n i.i.d. observations $X_1, X_2, \ldots, X_n$ are available.

Log-Likelihood Function for σ^2: The likelihood function for n independent and identically distributed (i.i.d.) observations is:

$$L(\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

The log-likelihood function simplifies to:

$$\log L(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

The first term, $-\frac{n}{2}\log(2\pi\sigma^2)$, accounts for the normalization of the normal distribution, while the second term represents the contribution of the data to the likelihood function.

First Derivative with Respect to σ^2: Taking the first derivative of the log-likelihood function with respect to σ^2, we get the score function, which measures how sensitive the log-likelihood is to changes in σ^2:

$$\frac{\partial}{\partial\sigma^2}\log L(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2$$

The first term, $-\frac{n}{2\sigma^2}$, comes from differentiating the logarithmic normalization term, and the second term, $\frac{1}{2\sigma^4}\sum(X_i - \mu)^2$, comes from differentiating the quadratic term involving the data.

Second Derivative with Respect to σ^2: Next, we calculate the second derivative of the log-likelihood function, which will allow us to compute the Fisher information:

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log L(\sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i - \mu)^2$$

This second derivative captures the curvature of the log-likelihood function with respect to σ^2. A large curvature implies that small changes in σ^2 significantly affect the likelihood, providing more information about the parameter.

Fisher Information for σ^2: The Fisher information is the negative of the expected value of the second derivative of the log-likelihood function. Since the expectation of $\sum_{i=1}^{n}(X_i - \mu)^2$ is $n\sigma^2$ (the sum of squared deviations for normal random variables), we can simplify the expression for Fisher information:

$$I(\sigma^2) = -\mathbb{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\log L(\sigma^2)\right]$$

Substituting the expected value of the sum of squared deviations, we get:

$$I(\sigma^2) = \frac{n}{2\sigma^4}$$

This shows that the Fisher information for σ^2 is inversely proportional to σ^4, which means that as the variance σ^2 increases, the amount of information provided by the data about σ^2 decreases. Conversely, smaller variances provide more information, as the curvature of the likelihood function becomes sharper.

Interpretation: The Fisher information for σ^2 quantifies how much the observed data contributes to estimating the variance parameter. Since the Fisher information decreases as σ^2 increases, this indicates that the parameter becomes harder to estimate accurately as the variance grows larger. Intuitively, if the data are spread out more (higher variance), it becomes more challenging to pinpoint the true value of σ^2.

2 Jeffreys Prior

Jeffreys prior is a commonly used *non-informative prior* in Bayesian statistics, designed to be invariant under reparameterization. This means that if you change the parameterization of your model, Jeffreys prior remains consistent with the transformed parameter. It is particularly useful when there is little prior knowledge about the parameter, and you wish to adopt a neutral, objective prior.

2.1 Definition

Jeffreys prior for a parameter θ is defined as:

$$\pi_J(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta)$ is the **Fisher information** for the parameter θ.

Because the Fisher information measures how much information the data provides about θ, Jeffreys prior guarantees that the prior is proportional to the square root of this information, resulting in a prior that reflects the structure of the problem but avoids imposing subjective biases.

2.2 Invariance under Reparameterization

One of the most appealing properties of Jeffreys prior is its invariance under transformations of the parameter. If $\theta' = g(\theta)$, the Jeffreys prior in the new parameterization remains proportional to the Fisher information:

$$\pi_J(\theta') = \pi_J(\theta) \left| \frac{d\theta}{d\theta'} \right|$$

This guarantees that the prior does not depend on how the parameter is expressed, making it a robust choice for non-informative priors.

2.2.1 Jeffreys Prior for μ

Consider a normal distribution $X \sim N(\mu, \sigma^2)$, where μ is the unknown mean, and the variance σ^2 is known. Recall in this case, the Fisher information is:

$$I(\mu) = \frac{n}{\sigma^2}$$

Jeffreys prior is derived from the Fisher information and is proportional to the square root of the Fisher information:

$$\pi_J(\mu) \propto \sqrt{I(\mu)}$$

Since the Fisher information is constant with respect to μ, Jeffreys prior for μ simplifies to:

$$\pi_J(\mu) \propto \sqrt{\frac{1}{\sigma^2}} = \frac{1}{\sigma}$$

This is a constant prior with respect to μ (since σ^2 is known), meaning:

$$\pi_J(\mu) \propto 1$$

Thus, Jeffreys prior is *flat* or *non-informative*, reflecting that all values of μ are equally likely before observing any data. This is appropriate when we have no prior information about μ. The prior is uninformative because μ is a *location parameter*, meaning that the data's location (or central tendency) shifts without affecting its spread or variability.

2.2.2 Interpretation

This flat prior expresses complete ignorance about the parameter μ, treating every value as equally plausible before any data is observed. After data is collected, the posterior distribution for μ will be driven entirely by the likelihood. The fact that the prior is flat for μ makes Jeffreys prior an attractive choice when performing Bayesian inference for location parameters, as it avoids introducing subjective biases into the analysis.

Moreover, the fact that Jeffreys prior is proportional to the square root of the Fisher information ensures that it is invariant under reparameterization. This invariance property makes Jeffreys prior a natural and robust choice for non-informative priors in Bayesian analysis.

2.3 Example 2: Normal Distribution (Variance Estimation)

Now, suppose we want to estimate the variance σ^2 in the normal distribution $X \sim N(\mu, \sigma^2)$, where the mean μ is known. Recall that the Fisher information for σ^2 is:

$$I(\sigma^2) = \frac{1}{2\sigma^4}$$

Thus, Jeffreys prior for σ^2 is:

$$\pi_J(\sigma^2) \propto \sqrt{I(\sigma^2)} = \frac{1}{\sigma^2}$$

This prior is an *improper prior* because it does not integrate to 1 over the entire domain, but it is still commonly used in practice, especially when combined with data through the likelihood in Bayesian inference.

3 Summary

- **Fisher Information**: Measures the amount of information that an observable random variable provides about a parameter. It is used to establish the Cramér-Rao bound, providing a lower bound on the variance of unbiased estimators.

- **Jeffreys Prior**: A non-informative prior proportional to the square root of the Fisher information, which is invariant under reparameterization and useful in Bayesian inference.

These concepts are foundational in Bayesian and frequentist inference, with Fisher information quantifying the precision of parameter estimates, and Jeffreys prior ensuring objectivity in Bayesian analysis.