Fisher Information 2024-06-14

jessekelighine.com
Jesse C. Chen 陳捷

Fisher Information is defined as follows:

$$\mathcal{I}_\theta \coloneqq \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right)\left(\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right)^\top \middle|\, \theta\right]$$

where $f(\cdot\,|\,\theta)$ is the likelihood function of the random variable x with parameter θ. The expectation is taken with respect to the random variable x.

In this note, we talk about what Fisher Information is, its relation to *maximum likelihood*, its conflicting intuitions, and *Cramér-Rao bound*, an inequality closely related to Fisher Information. [1]

1 What are we trying to do?

The goal is simple:

> *We want to have a way of measuring "how informative a parameter is to the likelihood function."*

If a certain parameter is very important in shaping the likelihood function, we expect our estimate of it to be more "precise." In contrast, if a parameter does not change the shape of the likelihood much, then we would expect the estimate of such parameter "poor," since likelihood functions under different true parameters look the same to us.

We will later see that *Fisher Information* achieves precisely this purpose in the context of maximum likelihood estimation.

2 Variance of Score Function

Let's first define some notation. Let $\ell(x\,|\,\theta) \coloneqq \log f(x\,|\,\theta)$ be the log-likelihood function. We use Newton's notation to denote differentiation with respect to the parameter, that is,

$$\dot{\ell}(x\,|\,\theta) \coloneqq \frac{\partial}{\partial\theta}\ell(x\,|\,\theta) = \frac{\partial}{\partial\theta}\log f(x\,|\,\theta).$$

Notice that $\dot{\ell}(x\,|\,\theta)$ has mean zero:

$$\begin{aligned}\mathbf{E}\,\dot{\ell}(x\,|\,\theta) &= \int_{\mathcal{X}}\left[\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right] f(x\,|\,\theta)dx \\ &= \int_{\mathcal{X}}\dot{f}(x\,|\,\theta)dx = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(x\,|\,\theta)dx = 0.\end{aligned}$$

[1] I this note, I shall assume the knowledge of maximum likelihood estimation and linear algebra.

Thus, we have $\mathbf{Var}\,\dot{\ell}(x \mid \theta) = \mathbf{E}\,\dot{\ell}(x \mid \theta)\dot{\ell}(x \mid \theta)^\top$, [2] which is the definition of *Fisher Information*. This observation gives us the first intuition of Fisher Information:

> **Intuition 1.** *Large Fisher Information means large variance in the score function $\dot{\ell}(x \mid \theta)$.* [a] *And a large variation in the score function means that our maximum likelihood estimate, which solves the problem*
>
> $$\arg\max_{\theta \in \Theta} \ell(x \mid \theta)$$
>
> *via first order condition*
>
> $$\dot{\ell}(x \mid \theta) = 0,$$
>
> *also has a large variance. This means that a large Fisher Information is* ***bad****.*
>
> [a] A large covariance matrix means the variance is large along every direction. You can find an intuitive explanation on my website.

3 Information Equality

Before jumping to conclusion, let's consider $\mathbf{E}\,\ddot{\ell}(x \mid \theta_0)$, the expected Hessian matrix of the log-likelihood function. It turns out $\mathbf{E}\,\ddot{\ell}(x \mid \theta_0)$ and $\mathbf{Var}\,\dot{\ell}(x \mid \theta_0)$ only differ by a negative sign:

$$\begin{aligned}
\mathbf{E}\,\ddot{\ell}(x \mid \theta) &= \int_{\mathcal{X}} \left[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log f(x \mid \theta) \right] f(x \mid \theta) dx \\
&= \int_{\mathcal{X}} \left[\frac{\ddot{f}(x \mid \theta)}{f(x \mid \theta)} - \frac{\dot{f}(x \mid \theta)\dot{f}(x \mid \theta)^\top}{f(x \mid \theta)^2} \right] f(x \mid \theta) dx \\
&= \underbrace{\frac{\partial^2}{\partial\theta^2} \int_{\mathcal{X}} f(x \mid \theta) dx}_{=0} - \int_{\mathcal{X}} \dot{\ell}(x \mid \theta)\dot{\ell}(x \mid \theta)^\top f(x \mid \theta) dx = -\,\mathbf{Var}\,\dot{\ell}(x \mid \theta)
\end{aligned}$$

This result is sometimes referred to as *information equality*. The equality states that the asymptotic variance of $\hat{\theta}$ is simply the negative of the Hessian of the log-likelihood function.

However, this means that we can also defined **Fisher Information** as $-\,\mathbf{E}\,\ddot{\ell}(x \mid \theta)$, which yields another intuition:

> **Intuition 2.** *A large Fisher information is actually* ***good****, since it is the Hessian matrix of the log-likelihood function. A larger Hessian implies a larger curvature, meaning that the log-likelihood function is more "defined" or "pointy" for our parameter, which naturally leads to a better estimation.*

This is in direct conflict with Intuition 1, how do we resolve this?

[2] We shall assume that Leibniz rule of differentiation is always satisfied for interchanging integration and differentiation.

4 Maximum Likelihood

Let's be precise. Suppose we have the data set $\{x_i\}_{i=1}^n \overset{\text{iid}}{\sim} f(\cdot \mid \theta_0)$. Then we denote the joint log-likelihood as

$$\ell(\{x_i\} \mid \theta) \coloneqq \sum_{i=1}^{n} \ell(x_i \mid \theta).$$

We want to derive the asymptotic distribution of the maximum likelihood estimator $\hat{\theta}$ as $n \to \infty$. By mean value theorem, we have

$$\underbrace{\dot{\ell}(\{x_i\} \mid \hat{\theta})}_{=0} - \dot{\ell}(\{x_i\} \mid \theta_0) = \ddot{\ell}(\{x_i\} \mid \bar{\theta})(\hat{\theta} - \theta_0)$$

where $\bar{\theta}$ is a value between θ_0 and $\hat{\theta}$. Supposing we already have the consistency result $\hat{\theta} \xrightarrow{p} \theta_0$, we have

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= -\left[\frac{\ddot{\ell}(\{x_i\} \mid \bar{\theta})}{n}\right]^{-1} \left(\frac{\dot{\ell}(\{x_i\} \mid \theta_0)}{\sqrt{n}}\right) \\ &\xrightarrow{d} -\mathbf{E}\left[\ddot{\ell}(x \mid \theta_0)\right]^{-1} \mathcal{N}(0, \mathbf{Var}\, \dot{\ell}(x \mid \theta_0)).\end{aligned}$$

Now we know the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ is a normal distribution centered at 0 with variance $\mathbf{E}[\ddot{\ell}(x \mid \theta_0)]^{-1} \mathbf{Var}\, \dot{\ell}(x \mid \theta_0) \mathbf{E}[\ddot{\ell}(x \mid \theta_0)]^{-1}$.

Notice this this result confirms with both of our intuitions: we have $\mathbf{Var}\, \dot{\ell}(x \mid \theta)$ in the nominator (in concord with Intuition 1), and $\mathbf{E}\, \ddot{\ell}(x \mid \theta_0)$ in the denominator (in concord with Intuition 2). Hence, it is a trade-off between these two effects we found in the two intuitions. And by *information equality*, we obtain the famous maximum likelihood result:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_\theta^{-1}).$$

Therefore, a **"larger"** Fisher Information is what we really want. This does not mean that Intuition 2 is correct and Intuition 1 is wrong, it is a result of both intuitions combined.

Note that the maximum likelihood result achieves what we set out to do in the first place. We now know that the estimation quality of $\hat{\theta}$ depends on the size of $\mathcal{I}_\theta$: the larger $\mathcal{I}_\theta$ is, the more precise the estimation, i.e., under the same number of samples n, a larger $\mathcal{I}_\theta$ yields a more precise result. [3]

[3] In some contexts (mostly in Bayesian statistics), the inverse of a covariance matrix is called the *precision matrix*. Thus, we can view Fisher Information as the precision matrix of the asymptotic distribution. I find this name particularly fitting in this context.

5 Cramér-Rao Bound

Now we know that Fisher Information tells us the precision of our estimation, but can we do better? Can we be more precise? The answer is **NO**, given that we want the estimation to be unbiased. [4] This result is called *Cramér-Rao bound*, stating that any other unbiased estimation of θ_0 must have variance larger than $\mathcal{I}_\theta^{-1}$.

Suppose $\tilde{\theta} = t(\{x_i\})$ is an unbiased estimator for θ_0, it's straightforward via some algebra to see that we have the covariance between $\dot{\ell}(\{x_i\} \mid \theta)$ and $t(\{x_i\}) - \theta$ as an identity matrix of dimension k:

$$\mathbf{E}[t(\{x_i\}) - \theta]\dot{\ell}(\{x_i\} \mid \theta)^\top = I_k,$$

where k is the length of θ_0.

A slight variation of Cauchy-Schwarz inequality states that for random vectors v and u both with mean 0, we have that $\mathbf{E}(vv^\top) - \mathbf{E}(vu^\top)\,\mathbf{E}(uu^\top)^{-1}\,\mathbf{E}(uv^\top)$ is positive definite. [5] If we plugin $v = t(\{x_i\}) - \theta$ and $u = \dot{\ell}(\{x_i\} \mid \theta)$, we have

$$\mathbf{Var}\, t(\{x_i\}) - (n\mathcal{I}_\theta)^{-1}$$

is positive definite. Therefore, $\mathbf{Var}\, t(\{x_i\})$ is larger than $(n\mathcal{I}_\theta)^{-1}$. [6]

What does it mean to be compared to $(n\mathcal{I}_\theta)^{-1}$? When n is sufficiently large, we have the approximation

$$\hat{\theta} - \theta_0 \overset{A}{\sim} \mathcal{N}(0, (n\mathcal{I}_\theta)^{-1})$$

Hence, we can see that the variance of our maximum likelihood estimator is approximately $(n\mathcal{I}_\theta)^{-1}$, which beats every other unbiased estimator.

* * *

In a nutshell:

> 1. **Maximum likelihood** is the best method in the sense that it produces the most precise estimator among all the unbiased estimators.
> 2. **Fisher Information** is a reasonable definition of "information," as it describes the best "precision" under all unbiased estimators of θ.

[4] Maximum likelihood estimation is not always unbiased, but the order of the bias is small $O(1/n)$.

[5] Note $\mathbf{E}(v + Au)(v + Au)^\top$ is positive definite. Let $A = -\,\mathbf{E}\, vu^\top (\mathbf{E}\, uu^\top)^{-1}$ and we obtain the desired result.

[6] If you are not sure what does positive definiteness have to do with the concept of "larger," you can find an intuitive explanation on my website.