Fisher Information

2024-04-07
陳 捷 Jesse C. Chen
jessekelighine.com

The *Fisher Information* is given by the following:

$$\mathcal{I}_\theta := \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right)^2\,\middle|\,\theta\right] \tag{1}$$

where $f(\cdot)$ is the conditioned likelihood function (density function) of random variable x. The expectation is taken with respect to the random variable x.

1 Idea

The idea is to measure "how much information about a parameter θ can we get from observing a random variable x." By information we mean "how much can θ affect the distribution of x." In the extreme cases, if the distribution of x does not depend of θ, then no information about θ is gained from observing x; if the distribution of x is highly dependent of the parameter θ, then a lot of information about θ is gained from observing x, since different θ's would imply very differnet distributions of x.

2 Explanation of the Formula

To measure this, we use formula (1). Essentially, we measure how much the density function will change if we nudge the parameter θ by a little bit. We use $\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)$ since we want to measure the "percentage of change" due to θ. However, we cannot simple calculate the expectation of $\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)$ since under some reularity conditions we have

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

That is, the expected "percentage of change" is zero. Thus, we measure the *variance* of "percentage of change" to guage how much θ affects the distribution of x.

Let $\ell(x\,|\,\theta) := \log f(x\,|\,\theta)$ and dot denote differentiation with respect to the parameter, i.e., $\dot{\ell}(x\,|\,\theta) := \frac{\partial}{\partial\theta}\ell(x\,|\,\theta)$. From the derivation above, we have that $\mathbf{E}\,\dot{\ell}(x\,|\,\theta) = 0$; also, from the definition of Fisher informtion, we have $\mathbf{Var}\,\dot{\ell}(x\,|\,\theta) = \mathcal{I}_\theta$.

3 Maximum Likelihood Estimation

If we estimate the parameter θ using MLE (using $\hat{\theta}_{\text{MLE}}$), then it can be shown that

$$\hat{\theta}_{\text{MLE}} \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_\theta^{-1}\right), \tag{2}$$

that is, the bigger the Fisher information, the better the estimation.

Proof. Let $x = \{x_1, ..., x_n\}$ be iid samples from $f(x \mid \theta_0)$ where θ_0 is the true parameter. Let $\ell_x(\theta)$ denote $\sum_{i=1}^n \ell_{x_i}(\theta)$ be the likelihood function.

We know that $\dot{\ell}_x(\theta) \sim (0, n\mathcal{I}_\theta)$ since x_i are iid samples from $f(x \mid \theta_0)$. We want to find the expectation of $\ddot{\ell}_x(\theta)$. Consider the second partial derivative of $\ell_x(\theta)$ with respect to θ:

$$\ddot{\ell}_x(\theta) := \frac{\partial^2}{\partial\theta^2}\ell_x(\theta) = \frac{\ddot{f}(x \mid \theta)}{f(x \mid \theta)} - \left(\frac{\dot{f}(x \mid \theta)}{f(x \mid \theta)}\right)^2$$

Take expectation of both sides and we have

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

where $\int_{\mathcal{X}} \ddot{f}(x \mid \theta)dx = 0$ is obtained by interchanging the integral and differentiation. Hence, $-\mathbf{E}\,\ddot{\ell}_x(\theta) = n\mathcal{I}_\theta$.

Since $\hat{\theta}_{\text{MLE}}$ is an MLE estimator, it satisfies $\dot{\ell}_x(\hat{\theta}_{\text{MLE}}) = 0$ by first order condition. By mean value theorem, we have

$$\underbrace{\dot{\ell}_x(\hat{\theta}_{\text{MLE}})}_{=0} - \dot{\ell}_x(\theta_0) = \ddot{\ell}_x(\bar{\theta})(\hat{\theta}_{\text{MLE}} - \theta_0)$$

where $\bar{\theta}$ is an value between θ_0 and $\hat{\theta}_{\text{MLE}}$. By rearranging the terms we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) = -\left(\frac{\ddot{\ell}_x(\bar{\theta})}{n}\right)^{-1} \frac{\dot{\ell}_x(\theta_0)}{\sqrt{n}} &\xrightarrow{d} -(n\mathcal{I}_\theta)^{-1}\mathcal{N}\left(0, n\mathcal{I}_\theta\right) \\
&= \mathcal{N}\left(0, (n\mathcal{I}_\theta)^{-1}\right)
\end{aligned}$$

by Slutsky's Theorem where $(\ddot{\ell}_x(\bar{\theta})/n)^{-1} \xrightarrow{p} (n\mathcal{I}_\theta)^{-1}$ by weak law of large number and $\dot{\ell}_x(\theta_0)/\sqrt{n} \xrightarrow{d} \mathcal{N}\left(0, n\mathcal{I}_\theta\right)$ by central limit theorem. □

Notice that this results also applies to multi-dimension parameters $\vec{\theta}$ by replacing Fisher information by a Fisher information matrix and $\ddot{\ell}_x(\theta)$ by the Hessian.

4 Cramér-Rao Bounds

One reason that MLE is favourable is that it is very efficient in the sense that it has the smallest possible variance of all the estimators. Cramér-Rao Bounds show that for any unbiased estimator $\tilde{\theta}$ of the parameter θ, we have

$$\mathbf{Var}\,\tilde{\theta} \geq \frac{1}{n\mathcal{I}_\theta}.$$

However, MLE are not always unbiased (e.g. MLE for variance). But since the bias of MLE are on the order of $1/n$, the bias is relatively small compared to its variance.

Proof. Consider an unbiased estimator $\tilde{\theta} = t(x)$ of θ. We have

$$\begin{aligned}\int_{\mathcal{X}} t(x)\dot{\ell}_x(\theta)f(x\,|\,\theta)dx = \int_{\mathcal{X}} t(x)\dot{f}(x\,|\,\theta)dx &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} t(x)f(x\,|\,\theta)dx \\ &= \frac{\partial}{\partial\theta}\theta = 1.\end{aligned}$$

Since $\dot{\ell}_x(\theta)$ has expectation zero, $\int_{\mathcal{X}}[t(x)-\theta]\dot{\ell}_x(\theta)f(x\,|\,\theta)dx = 1$. By applying Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left[\int_{\mathcal{X}}[t(x)-\theta]\dot{\ell}_x(\theta)f(x\,|\,\theta)dx\right]^2 &\leq \int_{\mathcal{X}}[t(x)-\theta]^2 f(x\,|\,\theta)dx \cdot \int_{\mathcal{X}}\dot{\ell}_x(\theta)^2 f(x\,|\,\theta)dx \\ \implies 1 &\leq \mathbf{Var}\,t(x)\cdot\mathcal{I}_\theta\end{aligned}$$

□

5 Observed Fisher Information

In equation (2), we derived the asymptotic distribution of MLE and found that the variance of such distribution is the inverse of Fisher information. However, in practice, it is hard to obtain Fisher information since it requires probability calculation. Therefore, Fisher advocated of the use of *Observed Fisher Information*, denoted by $I(x)$, defined by

$$I(x) := -\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}) = \frac{\partial^2}{\partial\theta^2}\ell_x(\theta)\bigg|_{\hat{\theta}_{\text{MLE}}},$$

which can be calculated numerically given the data x. This is basically an application of the *plug-in principle*.

Apart from being easy to calculate, another benefit, argued by Fisher, of using the observed Fisher information is that it is a kind of *conditioned inference* (or in Fisher's jargon, approximate ancillary), which bridges the gap between Bayesian and frequentist inference: by using the "observed" information, we are essentially updating our belief (or our confident) in the estimator based on the observed data, which might be more relevant while making inferences. ■

References

[1] Wikipedia: https://en.wikipedia.org/wiki/Fisher_information

[2] Handbook of Econometrics. volume 4. p.2141. Whitney K. Newey & Daniel Mcfadden.

[3] Computer Age Statistical Inference. Chapter *Fisherian Inference and Maximum Likelihood Estimation*. Bradley Efron & Trevor Hastie.