STAT-510 **2017 Fall**

Lecture 4: Fisher Information

Instructor: Xiaohui Chen   Scribe: Yutong Dai   Last Modified: 2018-03-14

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications, and they may contain factual and/or typographic errors.*

4.1 Score

In statistics, the score indicates how sensitive a likelihood function $L(X;\theta)$ is to its parameter θ.

Definition 4.1 The score for θ is the gradient of the log-lilelihood with respect to θ, i.e.

$$S(\theta) = \frac{\partial}{\partial \theta} \log f(X;\theta).$$

Remark Note that $S(\theta)$ is a function of θ and the observation X, so that, in general, it is not a statistic.

Propeties:

Mean: $E[S(\theta)|\theta] = 0$.

Proof:

$$\begin{aligned}
E[S(\theta)|\theta] &= \int f(x;\theta) \frac{\partial}{\partial \theta} \log f(x;\theta) dx \\
&= \int \frac{\partial}{\partial \theta} f(x;\theta) dx \\
&\overset{(1)}{=} \frac{\partial}{\partial \theta} \int f(x;\theta) dx \\
&= 0
\end{aligned}$$

In (1), we assume some regularity conditions. ■

Variance: $\mathbf{Var} S(\theta) = E([\frac{\partial}{\partial \theta} \log f(x;\theta)]^2|\theta)$.

Remark The variance of score is also known as Fisher Information and denoted as $\mathcal{I}(\theta)$.

4.2 Fisher Information

The Fisher information is a way of measuring the amount of information that an observable random sample X carries about an unknown parameter θ upon which the probability of X depends. Let $f(X;\theta)$ be the probability density function (or probability mass function) for X. This is also the likelihood function for θ. It describes the probability that we observe a given sample X, given a known value of θ.

If f is sharply peaked with respect to changes in θ, it is easy to indicate the "correct" value of θ from the data, or equivalently, that the data X provides a lot of information about the parameter θ. If the likelihood f is flat and spread-out, then it would take many samples like X to estimate the actual "true" value of θ that would be obtained using the entire population being sampled. This suggests studying some kind of variance with respect to θ.

Definition 4.2 (singel parameter case) The fisher information is defined as

$$\mathcal{I}(\theta) = E([\frac{\partial}{\partial\theta}\log f(X;\theta)]^2|\theta) = \int[\frac{\partial}{\partial\theta}\log f(x;\theta)]^2 f(x;\theta)dx.$$

Remark

1. The Fisher information is not a function of a particular observation, as the random variable X has been averaged out.
2. A random variable carrying high Fisher information implies that the absolute value of the score is often high.

Claim 4.3 If $\log f(x;\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions then the Fisher information may also be written as

$$\mathcal{I}(\theta) = -E[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)|\theta]$$

Proof:

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right)^2 \\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\end{aligned}$$

Since

$$E\left[\left.\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right|\theta\right] = \int\frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx = 0,$$

taking expectation on both sides of the beginnig equation concludes the proof. ■

Suppose $X_1,...,X_n \overset{iid}{\sim} f(x;\theta)$. Then the loglikelihood function is $\ell(X;\theta) = \sum_{i=1}^n \log f(X_i;\theta)$. Set deriviate of $\ell(X;\theta)$ with respect to θ, we obtain the score:

$$S(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i;\theta).$$

Since $\mathcal{I}(\theta) = E([\frac{\partial}{\partial\theta}\log f(X;\theta)]^2|\theta) = \mathbf{Var}S(\theta)$, the Fisher Information actually represents the expectation of the second-order derivative of likelihood function - curvature. The larger the curvature, the spiky the loglikelihood function is, hence more information contained.

Proposition 4.4 (additive property) Fisher Information is additive for independent observations. That is if $X_1,...,X_n \overset{iid}{\sim} f(x;\theta)$ and $\mathcal{I}_{X_1}(\theta) = E([\frac{\partial}{\partial\theta}\log f(X_1;\theta)]^2|\theta)$, then $\mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta)$.

4.3 Conditional Fisher Information

Let $X, Y \sim f_{XY}(x,y;\theta)$. The Fisher Information about θ in Y given X is defined by

$$\mathcal{I}_{Y|X}(\theta) = \int \mathcal{I}_{Y|X=x}(\theta)f(x;\theta)dx.$$

Lemma 4.5 Let $X, Y \sim f_{XY}(x, y; \theta)$. Then $\mathcal{I}_{XY}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_{Y|X}(\theta)$.

Proof:

$$\log f_{XY}(x, y; \theta) = \log f_Y(y|X = x; \theta) + \log f_X(x; \theta)$$

Set second-order derivative and take expectation on both sides, we have

$$\begin{aligned}
&- E_{XY}(\frac{\partial^2}{\partial\theta^2} \log f_{XY}(x, y; \theta)) \\
&= -E_{XY}(\frac{\partial^2}{\partial\theta^2} \log f_Y(y|X = x; \theta)) - E_{XY}(\frac{\partial^2}{\partial\theta^2} \log f_X(x; \theta)) \\
&= -[\int \frac{\partial^2}{\partial\theta^2} \log f_Y(y|X = x; \theta) f_Y(y|X = x) dy \int f_X(x) dx] - \int \frac{\partial^2}{\partial\theta^2} \log f_X(x; \theta) f_X(x) dx \int f_Y(y|X = x) dy \\
&= \mathcal{I}_X(\theta) + \mathcal{I}_{Y|X}(\theta)
\end{aligned}$$

■

Corollary 4.6

1. $\mathcal{I}_{XY}(\theta) \geq \mathcal{I}_X(\theta)$, where the equality is attained iff $\forall x, y, \quad f_Y(y|X = x; \theta)$ doesn't depend on θ.
2. If X, Y are independent, then $\mathcal{I}_{XY}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)$.

Definition 4.7 For two statistic $T_1(X), T_2(X)$. We say $T_1(X)$ is more informative than $T_2(X)$ if $I_{T_1(X)}(\theta) - I_{T_2(X)}(\theta) \succeq 0$, where $\succeq$ means positive defined. (We shall define fisher information later in matrix form.)

Theorem 4.8 (Sufficient Statistic do not lose any information) For any statistic $T(X)$, we have $\mathcal{I}_X(\theta) \geq \mathcal{I}_{T(X)}(\theta)$ and the equality holds iff $T(X)$ is a sufficient statistic.

Proof: From Lemma 4.5, we know that

$$\mathcal{I}_{(X,T(X))}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_{T(X)|X}(\theta).$$

Since

$$f_{T(X)}(y|X = x; \theta) = \begin{cases} 1; & y = T(x) \\ 0; & \text{otherwise} \end{cases},$$

$f_{T(X)}(y|X = x; \theta)$ doesen't depend on θ. From Corollary 4.6, we have $\mathcal{I}_{(X,T(X))}(\theta) = \mathcal{I}_X(\theta)$. Additionaly, $\mathcal{I}_{(X,T(X))}(\theta) = \mathcal{I}_{T(X)}(\theta) + \mathcal{I}_{X|T(X)}(\theta)$, so

$$\mathcal{I}_X(\theta) \geq \mathcal{I}_{T(X)}(\theta)$$

"$\Leftarrow$":

Suppose $T(X)$ is a sufficient statistic, then $f(x|T(X) = t; \theta)$ dose not depend on θ. From Corollary 4.6, we conclude that

$$\mathcal{I}_{(X,T(X))}(\theta) = \mathcal{I}_{T(X)}(\theta).$$

Therefore, $\mathcal{I}_X(\theta) = \mathcal{I}_{T(X)}(\theta)$.

"$\Rightarrow$":

Since $\mathcal{I}_X(\theta) = \mathcal{I}_{T(X)}(\theta)$, $\mathcal{I}_{(X,T(X))}(\theta) = \mathcal{I}_X(\theta)$ and $\mathcal{I}_{(X,T(X))}(\theta) = \mathcal{I}_{T(X)}(\theta) + \mathcal{I}_{X|T(X)}(\theta)$, we know that

$$\mathcal{I}_{X|T(X)}(\theta) = 0.$$

From Corollary 4.6, we conclude that $f(x|T(X) = t; \theta)$ dose not depend on θ and hence $T(X)$ is a sufficient statistic. ■

Theorem 4.9 Suppose $W(X)$ is an ancillary statistic, then

$$\mathcal{I}_{W(X)}(\theta) = 0, \quad \forall \theta \in \Theta.$$

Claim 4.10 (Conditional Inference) Supppose $X \sim f(x;\theta)$ and there are two satistic $T_1 = T_1(X), T_2 = T_2(X)$. If T_1 is an ancillary statistic, then

$$\mathcal{I}_{T_1,T_2}(\theta) = \mathcal{I}_{T_2|T_1}(\theta), \quad \forall \theta \in \Theta.$$

From this claim, we know that if we want to know the Fisher Information of (T_1, T_2), we need to follow steps given below.

- Step1: Find the conditional pdf/pmf of T_2 given $T_1 = t_1$, denoted by $g_{T_2|T_1=t_1}(t_2;\theta)$
- Step2: Compute the conditional Fisher Information at $T_1 = t_1$

$$\mathcal{I}_{T_2|T_1=t_1} = \int [\frac{\partial}{\partial\theta} log(g_{T_2|T_1=t_1}(t_2;\theta))]^2 g_{T_2|T_1=t_1}(t_2;\theta)dt_2.$$

- Step3: Average $\mathcal{I}_{T_2|T_1=t_1}$ over all possible t_1 with weights given by the pdf/pmf of T_1, denoted by

$$\mathcal{I}_{T_2|T_1} = \int \mathcal{I}_{T_2|T_1=t_1}(\theta) f_{T_1}(t_1)dt_1.$$

Example 1: Consider $X_1, X_2 \overset{iid}{\sim} N(\theta, 1)$ and $T_1 = X_1$ (not a sufficient statistc), $T_2 = X_1 - X_2$ (an ancillary statistic). We want to find the distribution of $f_{T_1|T_2=v}(u)$. Note that

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim N_2(\begin{pmatrix} \theta \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}),$$

and the Lemma 4.11, we know that $T_1|T_2 = v \sim N(\theta/2, 1/2)$. Hence, $\mathcal{I}_{T_1|T_2=v} = \frac{1}{2\sigma^4} = 2$ and furthermore $\mathcal{I}_{T_1|T_2} = 2$. We conclude that

$$I_{T_1,T_2}(\theta) = I_{X_1,X_2} = 2.$$

The last equality holds that (T_1, T_2) and (X_1, X_2) are all sufficient staistic.

Remark This example agian shows that an ancillary statistic contains no information about the θ, once we combine it with another statistic, which has someinformation about θ, then we can have all information abouth θ.

Lemma 4.11 If

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}),$$

then

$$X_1|X_2 = x_2 \sim N(\bar{\mu}, \bar{\Sigma}),$$

where

$$\begin{cases} \bar{\mu} & = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \bar{\Sigma} & = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{cases}$$

When there are p parameters, so that θ is a $p \times 1$ vector $\theta = \left[\theta_1, \theta_2, \ldots, \theta_N\right]^{\mathrm{T}}$, then the Fisher information takes the form of an $p \times p$ matrix.

Definition 4.12 (multiple parameter case) The Fisher information matrix (FIM) matrix has typical element

$$\left[\mathcal{I}\left(\theta\right)\right]_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\middle|\,\theta\right].$$

Claim 4.13 If $\log f(x;\theta)$ is twice differentiable with respect to each θ_i, and under certain regularity conditions then the Fisher information may also be written as

$$[\mathcal{I}(\theta)]_{ij} = -E[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X;\theta)|\theta]$$

Proof:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X;\theta) &= \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(X;\theta)}{f(X;\theta)} - \frac{\frac{\partial}{\partial\theta_i}f(X;\theta)\frac{\partial}{\partial\theta_j}f(X;\theta)}{f(X;\theta)^2} \\
&= \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)
\end{aligned}$$

Since

$$E\left[\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(X;\theta)}{f(X;\theta)}\middle|\,\theta\right] = \int\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x;\theta)dx = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f(x;\theta)dx = 0,$$

taking expectation on both sides of the beginnig equation concludes the proof. ■

Example 1: Suppose $X \sim N(\mu, \sigma^2)$, and $\theta = (\mu, \sigma^2)$. Then

$$\begin{cases} \frac{\partial}{\partial\mu}\log f(X;\mu,\sigma^2) = \frac{X-\mu}{\sigma^2} \\ \frac{\partial}{\partial\sigma^2}\log f(X;\mu,\sigma^2) = -\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4}. \end{cases}$$

$\mathcal{I}_{11}(\theta) = E(\frac{X-\mu}{\sigma^2})^2 = \frac{1}{\sigma^2}$.

$$\begin{aligned}
\mathcal{I}_{22}(\theta) &= E(-\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4})^2 \\
&= \frac{1}{4\sigma^4}E[(\frac{X-\mu}{\sigma})^2 - 1]^2 \\
&= \frac{1}{4\sigma^4}E[Z^2 - 1]^2 \\
&= \frac{1}{2\sigma^4},
\end{aligned}$$

where $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. The last eqality holds as $EX^n = (n-1)!!\sigma^n$ when n is even.

$\mathcal{I}_{12}(\theta) = E[(\frac{X-\mu}{\sigma^2})(-\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4})] = \frac{1}{2\sigma^3}EZ^3 = 0.$

To sum up,

$$\mathcal{I}_X(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

Example 2: Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$a $X = (X_1, X_2, ..., X_n)$. Denote $\bar{X}_n = \frac{1}{n}\sum_i X_i \sim N(\mu, \sigma^2/n)$ then by Proposition 4.4, we know that

$$\mathcal{I}_X(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

However, as

$$\mathcal{I}_{\bar{X}_n} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

there's information lose in the $[\mathcal{I}_{\bar{X}_n}]_{22}$.

Consider $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2$, we have $V = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$. Since $\log f_V(v) = -\log[2^{(n-1)/2}\Gamma(\frac{n-1}{2})] + \frac{n-3}{2}\log v - \frac{1}{2}v$ and $T \triangleq S^2 = h(V) = \frac{\sigma^2}{n-1}V$, we know that

$$f_T(t;\sigma^2) = f_V(h^{-1}(t);\sigma^2)|\frac{\partial}{\partial t}h^{-1}(t)|.$$

So

$$\frac{\partial}{\partial \sigma^2}\log f_T(t;\sigma^2) = \frac{(n-1)t}{2\sigma^4} - \frac{n-1}{2\sigma^2}$$

$$\mathcal{I}_{S^2}(\theta) = E_{f_T(t;\sigma^2)}[\frac{(n-1)t}{2\sigma^4} - \frac{n-1}{2\sigma^2}]^2 = \frac{n-1}{2\sigma^4}.$$

Therefore,

$$\mathcal{I}_{S^2} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{n-1}{2\sigma^4} \end{pmatrix}.$$

Theorem 4.14 Suppose $X \sim f(x;\theta)$ and $Y = h(X)$, where $h(.)$ is one-to-one, diffreniable and don't depend on θ. Then $\mathcal{I}_X(\theta) = \mathcal{I}_Y(\theta)$.

Proof: For simplicity, we only focus on one dimesion case and assume $f(x;\theta)$ is a pdf. Notice in $f_Y(y;\theta) = f_X(h^{-1}(y);\theta)|\frac{\partial}{\partial y}h^{-1}(y)|, y \in$ support, the term $|\frac{\partial}{\partial y}h^{-1}(y)|$ dose not depend on θ, hence $\frac{\partial}{\partial t}\log f_Y(y;\theta) = \frac{\partial}{\partial t}\log f_X(h^{-1}(y);\theta)$.

$$\mathcal{I}_Y(\theta) = E[\frac{\partial}{\partial t}\log f_Y(y;\theta)]^2 = \mathcal{I}_X(\theta).$$

∎