Fisher Information

The *Fisher information* is a way of quantifying the information that an particular observation, or data point, x contains about an unknown parameter θ.

The *score* is defined as the gradient of the log-likelihood function w.r.t. a model parameter,

$$\mathcal{S}(\theta) = \frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}. \tag{1}$$

The score is a function of the model parameter; at a particular value of the parameter, the score measures the sensitivity of the log-likelihood to changes in the parameter.

If there are multiple parameters in the model then the score is the gradient vector.

If the score is evaluated at the true value of the parameter θ, then its expectation value is zero. (Assuming the likelihood is sufficiently well behaved, e.g. all its partial derivatives exist.) The expectation value is to be understood as being taken over many different experiments; i.e. the expectation over the data x.

$$\begin{aligned}
\mathrm{E}_x\left[\mathcal{S}(\theta)\middle|\theta = \text{true value}\right] &= \int \mathrm{d}x\ S(\theta)\mathcal{L}(x|\theta) && (2)\\
&= \int \mathrm{d}x\ \frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}\mathcal{L}(x|\theta) && (3)\\
&= \int \mathrm{d}x\ \frac{1}{\mathcal{L}(x|\theta)}\frac{\partial \mathcal{L}(x|\theta)}{\partial \theta}\mathcal{L}(x|\theta) && (4)\\
&= \int \mathrm{d}x\ \frac{\partial \mathcal{L}(x|\theta)}{\partial \theta} && (5)\\
&= \frac{\partial}{\partial \theta}\int \mathrm{d}x\ \mathcal{L}(x|\theta) && (6)\\
&= \frac{\partial}{\partial \theta}1 && (7)\\
&= 0 && (8)
\end{aligned}$$

Where on the penultimate line we have used the fact that the likelihood is a (normalised) probability distribution for the data x.

The Fisher information, $\mathcal{I}(\theta)$, is defined to be variance of the score (evaluated at the true value of the parameter θ). Because $\mathcal{I} \geq 0$, a higher value of the variance implies that the absolute value of the score is larger on average; roughly speaking, this means the likelihood

function is more curved and therefore carries more information about the model parameter.

$$\mathcal{I}(\theta) = \mathrm{Var}_x\big(\mathcal{S}(\theta)\big) \tag{9}$$

$$= \mathrm{E}_x\left[\mathcal{S}(\theta)^2\middle|\theta = \text{true value}\right] - \mathrm{E}_x\left[\mathcal{S}(\theta)\middle|\theta = \text{true value}\right]^2 \tag{10}$$

$$= \int \mathrm{d}x\ \mathcal{L}(x|\theta)\mathcal{S}(\theta)^2. \tag{11}$$

Note the score and Fisher information are both dimensionfull quantities; we have $[\mathcal{S}(\theta)] = [\theta]^{-1}$ and $[\mathcal{I}(\theta)] = [\theta]^{-2}$.

The Fisher information can also be expressed in terms of the second derivative of the log-likelihood w.r.t. the parameter θ (if this derivative exists).

$$\mathcal{I}(\theta) = -\mathrm{E}_x\left[\frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial\theta^2}\middle|\theta = \text{true value}\right] \tag{12}$$

$$= -\int \mathrm{d}x\ \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial\theta^2}\mathcal{L}(x|\theta) \tag{13}$$

If there are multiple parameters in the model, i.e. θ becomes θ_μ with $\mu = 1, 2, \ldots, n$, then the Fisher information becomes a square $n \times n$ matrix, called the *Fisher information matrix*;

$$\mathcal{I}_{\mu\nu}(\theta) = -\mathrm{E}_x\left[\frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial\theta_\mu \partial\theta_n u}\middle|\theta = \text{true value}\right] \tag{14}$$

$$= -\int \mathrm{d}x\ \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial\theta^2}\mathcal{L}(x|\theta) \tag{15}$$

Exercise 0.1: Fisher information
Starting from Eq. 12, show that this is equivalent to the definition of the Fisher information in Eq. 11.

The Fisher information depends on the choice of parametrisation of the model. If ϕ is an alternative parameter related to original parameter by a (smooth) function $\theta(\phi)$, then the

$$\mathcal{I}(\phi) = \mathcal{I}(\theta)\left(\frac{\partial\theta}{\partial\phi}\right)^2, \tag{16}$$

where $\mathcal{I}(\theta)$ and $\mathcal{I}(\theta)$ are the Fisher informations for the two parameters respectively.

The Fisher information appears in the (expectation of the) Taylor expansion of the log-likelihood about the true value of the model parameter, θ'.

$$\log \mathcal{L}(x|\theta) = \mathcal{L}(x|\theta') + (\theta - \theta')\frac{\partial \log \mathcal{L}(x|\theta)}{\partial \theta}\Big|_{\theta=\theta'} + \frac{1}{2}(\theta - \theta')^2 \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial^2 \theta}\Big|_{\theta=\theta'} \dots \quad (17)$$

$$= \text{const} + (\theta - \theta')\mathcal{S}(\theta) + \frac{1}{2}(\theta - \theta')^2 \frac{\partial^2 \log \mathcal{L}(x|\theta)}{\partial^2 \theta}\Big|_{\theta=\theta'} \dots \quad (18)$$

The actual shape of the likelihood depends on the realisation of the data, x, in a particular experiment. Taking the expectation over the data gives

$$\mathrm{E}_x \left[\log \mathcal{L}(x|\theta)\right] = \text{const} - \frac{1}{2}(\theta - \theta')^2 \mathcal{I}(\theta') \dots \quad (19)$$

where we have used the fact that the expectation of the score vanishes (Eq. 8) and the definition of the Fisher information in Eq. 12.

If there are multiple parameters then the expansion in Eq. 19 becomes

$$\mathrm{E}_x \left[\log \mathcal{L}(x|\theta)\right] = \text{const} - \frac{1}{2}\sum_{\mu=1}^{n}\sum_{\nu=1}^{n}(\theta_\mu - \theta'_\mu)(\theta_\nu - \theta'_\nu)\mathcal{I}_{\mu\nu}(\theta') \dots \quad (20)$$