2.2 Score Function and Fisher Information

The MLE of θ is obtained by maximising the (relative) likelihood function,

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta\in\Theta} L(\theta) = \arg\max_{\theta\in\Theta} \tilde{L}(\theta).$$

For numerical reasons, it is often easier to maximise the log-likelihood $l(\theta) = \log L(\theta)$ or the relative log-likelihood $\tilde{l}(\theta) = l(\theta) - l(\hat{\theta}_{\mathrm{ML}})$ (cf. Sect. 2.1), which yields the same result since

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta\in\Theta} l(\theta) = \arg\max_{\theta\in\Theta} \tilde{l}(\theta).$$

However, the log-likelihood function $l(\theta)$ has much larger importance, besides simplifying the computation of the MLE. Especially, its first and second derivatives are important and have their own names, which are introduced in the following. For simplicity, we assume that θ is a scalar.

Definition 2.6 (Score function) The first derivative of the log-likelihood function

$$S(\theta) = \frac{dl(\theta)}{d\theta}$$

is called the *score function*. ♦

Computation of the MLE is typically done by solving the *score equation* $S(\theta) = 0$.

The second derivative, the curvature, of the log-likelihood function is also of central importance and has its own name.

Definition 2.7 (Fisher information) The negative second derivative of the log-likelihood function

$$I(\theta) = -\frac{d^2 l(\theta)}{d\theta^2} = -\frac{dS(\theta)}{d\theta}$$

is called the *Fisher information*. The value of the Fisher information at the MLE $\hat{\theta}_{\mathrm{ML}}$, i.e. $I(\hat{\theta}_{\mathrm{ML}})$, is the *observed Fisher information*. ♦

Note that the MLE $\hat{\theta}_{\mathrm{ML}}$ is a function of the observed data, which explains the terminology "observed" Fisher information for $I(\hat{\theta}_{\mathrm{ML}})$.

Example 2.9 (Normal model) Suppose we have realisations $x_{1:n}$ of a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with unknown mean μ and known vari-

ance σ^2. The log-likelihood kernel and score function are then

$$l(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \text{and}$$

$$S(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu),$$

respectively. The solution of the score equation $S(\mu) = 0$ is the MLE $\hat{\mu}_{\mathrm{ML}} = \bar{x}$. Taking another derivative gives the Fisher information

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

which does not depend on μ and so is equal to the observed Fisher information $I(\hat{\mu}_{\mathrm{ML}})$, no matter what the actual value of $\hat{\mu}_{\mathrm{ML}}$ is.

Suppose we switch the roles of the two parameters and treat μ as known and σ^2 as unknown. We now obtain

$$\hat{\sigma}^2_{\mathrm{ML}} = \sum_{i=1}^{n}(x_i - \mu)^2/n$$

with Fisher information

$$I\left(\sigma^2\right) = \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2\sigma^4}.$$

The Fisher information of σ^2 now really depends on its argument σ^2. The observed Fisher information turns out to be

$$I\left(\hat{\sigma}^2_{\mathrm{ML}}\right) = \frac{n}{2\hat{\sigma}^4_{\mathrm{ML}}}.$$

■

It is instructive at this stage to adopt a *frequentist* point of view and to consider the MLE $\hat{\mu}_{\mathrm{ML}} = \bar{x}$ from Example 2.9 as a random variable, i.e. $\hat{\mu}_{\mathrm{ML}} = \bar{X}$ is now a function of the random sample $X_{1:n}$. We can then easily compute $\mathrm{Var}(\hat{\mu}_{\mathrm{ML}}) = \mathrm{Var}(\bar{X}) = \sigma^2/n$ and note that

$$\mathrm{Var}(\hat{\mu}_{\mathrm{ML}}) = \frac{1}{I(\hat{\mu}_{\mathrm{ML}})}$$

holds. In Sect. 4.2.3 we will see that this equality is approximately valid for other statistical models. Indeed, under certain regularity conditions, the variance $\mathrm{Var}(\hat{\theta}_{\mathrm{ML}})$ of the MLE turns out to be approximately equal to the inverse observed Fisher information $1/I(\hat{\theta}_{\mathrm{ML}})$, and the accuracy of this approximation improves with increasing sample size n. Example 2.9 is a special case, where this equality holds exactly for any sample size.

Example 2.10 (Binomial model) The score function of a binomial observation $X = x$ with $X \sim \text{Bin}(n, \pi)$ is

$$S(\pi) = \frac{dl(\pi)}{d\pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi}$$

and has been derived already in Example 2.1. Taking the derivative of $S(\pi)$ gives the Fisher information

$$\begin{aligned} I(\pi) &= -\frac{d^2 l(\pi)}{d\pi^2} = -\frac{dS(\pi)}{d\pi} \\ &= \frac{x}{\pi^2} + \frac{n-x}{(1-\pi)^2} \\ &= n\left\{\frac{x/n}{\pi^2} + \frac{(n-x)/n}{(1-\pi)^2}\right\}. \end{aligned}$$

Plugging in the MLE $\hat{\pi}_{\text{ML}} = x/n$, we finally obtain the observed Fisher information

$$I(\hat{\pi}_{\text{ML}}) = \frac{n}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})}.$$

This result is plausible if we take a frequentist point of view and consider the MLE as a random variable. Then

$$\text{Var}(\hat{\pi}_{\text{ML}}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(X) = \frac{1}{n^2} n\pi(1-\pi) = \frac{\pi(1-\pi)}{n},$$

so the variance of $\hat{\pi}_{\text{ML}}$ has the same form as the inverse observed Fisher information; the only difference is that the MLE $\hat{\pi}_{\text{ML}}$ is replaced by the true (and unknown) parameter π. The inverse observed Fisher information is hence an estimate of the variance of the MLE. ■

How does the observed Fisher information change if we reparametrise our statistical model? Here is the answer to this question.

Result 2.1 (Observed Fisher information after reparametrisation) *Let $I_\theta(\hat{\theta}_{\text{ML}})$ denote the observed Fisher information of a scalar parameter θ and suppose that $\phi = h(\theta)$ is a one-to-one transformation of θ. The observed Fisher information $I_\phi(\hat{\phi}_{\text{ML}})$ of ϕ is then*

$$I_\phi(\hat{\phi}_{\text{ML}}) = I_\theta(\hat{\theta}_{\text{ML}}) \left\{\frac{dh^{-1}(\hat{\phi}_{\text{ML}})}{d\phi}\right\}^2 = I_\theta(\hat{\theta}_{\text{ML}}) \left\{\frac{dh(\hat{\theta}_{\text{ML}})}{d\theta}\right\}^{-2}. \tag{2.3}$$

Proof The transformation h is assumed to be one-to-one, so $\theta = h^{-1}(\phi)$ and $l_\phi(\phi) = l_\theta\{h^{-1}(\phi)\}$. Application of the chain rule gives

$$
\begin{aligned}
S_\phi(\phi) &= \frac{dl_\phi(\phi)}{d\phi} = \frac{dl_\theta\{h^{-1}(\phi)\}}{d\phi} \\
&= \frac{dl_\theta(\theta)}{d\theta} \cdot \frac{dh^{-1}(\phi)}{d\phi} \\
&= S_\theta(\theta) \cdot \frac{dh^{-1}(\phi)}{d\phi}. \qquad (2.4)
\end{aligned}
$$

The second derivative of $l_\phi(\phi)$ can be computed using the product and chain rules:

$$
\begin{aligned}
I_\phi(\phi) &= -\frac{dS_\phi(\phi)}{d\phi} = -\frac{d}{d\phi}\left\{ S_\theta(\theta) \cdot \frac{dh^{-1}(\phi)}{d\phi} \right\} \\
&= -\frac{dS_\theta(\theta)}{d\phi} \cdot \frac{dh^{-1}(\phi)}{d\phi} - S_\theta(\theta) \cdot \frac{d^2h^{-1}(\phi)}{d\phi^2} \\
&= -\frac{dS_\theta(\theta)}{d\theta} \cdot \left\{ \frac{dh^{-1}(\phi)}{d\phi} \right\}^2 - S_\theta(\theta) \cdot \frac{d^2h^{-1}(\phi)}{d\phi^2} \\
&= I_\theta(\theta) \left\{ \frac{dh^{-1}(\phi)}{d\phi} \right\}^2 - S_\theta(\theta) \cdot \frac{d^2h^{-1}(\phi)}{d\phi^2}.
\end{aligned}
$$

Evaluating $I_\phi(\phi)$ at the MLE $\phi = \hat{\phi}_{\mathrm{ML}}$ (so $\theta = \hat{\theta}_{\mathrm{ML}}$) leads to the first equation in (2.3) (note that $S_\theta(\hat{\theta}_{\mathrm{ML}}) = 0$). The second equation follows with

$$
\frac{dh^{-1}(\phi)}{d\phi} = \left\{ \frac{dh(\theta)}{d\theta} \right\}^{-1} \quad \text{for } \frac{dh(\theta)}{d\theta} \neq 0. \qquad (2.5)
$$

□

Example 2.11 (Binomial model) In Example 2.6 we saw that the MLE of the odds $\omega = \pi/(1-\pi)$ is $\hat{\omega}_{\mathrm{ML}} = x/(n-x)$. What is the corresponding observed Fisher information? First, we compute the derivative of $h(\pi) = \pi/(1-\pi)$, which is

$$
\frac{dh(\pi)}{d\pi} = \frac{1}{(1-\pi)^2}.
$$

Using the observed Fisher information of π derived in Example 2.10, we obtain

$$
\begin{aligned}
I_\omega(\hat{\omega}_{\mathrm{ML}}) &= I_\pi(\hat{\pi}_{\mathrm{ML}}) \left\{ \frac{dh(\hat{\pi}_{\mathrm{ML}})}{d\pi} \right\}^{-2} = \frac{n}{\hat{\pi}_{\mathrm{ML}}(1-\hat{\pi}_{\mathrm{ML}})} \cdot (1-\hat{\pi}_{\mathrm{ML}})^4 \\
&= n \cdot \frac{(1-\hat{\pi}_{\mathrm{ML}})^3}{\hat{\pi}_{\mathrm{ML}}} = \frac{(n-x)^3}{nx}.
\end{aligned}
$$

As a function of x for fixed n, the observed Fisher information $I_\omega(\hat{\omega}_{\text{ML}})$ is monotonically decreasing (the numerator is monotonically decreasing, and the denominator is monotonically increasing). In other words, the observed Fisher information increases with decreasing MLE $\hat{\omega}_{\text{ML}}$.

The observed Fisher information of the log odds $\phi = \log(\omega)$ can be similarly computed, and we obtain

$$I_\phi(\hat{\phi}_{\text{ML}}) = I_\omega(\hat{\omega}_{\text{ML}}) \left(\frac{1}{\hat{\omega}_{\text{ML}}} \right)^{-2} = \frac{(n-x)^3}{nx} \cdot \frac{x^2}{(n-x)^2} = \frac{x(n-x)}{n}.$$

Note that $I_\phi(\hat{\phi}_{\text{ML}})$ does not change if we redefine successes as failures and vice versa. This is also the case for the observed Fisher information $I_\pi(\hat{\pi}_{\text{ML}})$ but not for $I_\omega(\hat{\omega}_{\text{ML}})$. ■

2.3 Numerical Computation of the Maximum Likelihood Estimate

Explicit formulas for the MLE and the observed Fisher information can typically only be derived in simple models. In more complex models, numerical techniques have to be applied to compute maximum and curvature of the log-likelihood function. We first describe the application of general purpose optimisation algorithms to this setting and will discuss the Expectation-Maximisation (EM) algorithm in Sect. 2.3.2.

2.3.1 Numerical Optimisation

Application of the *Newton–Raphson algorithm* (cf. Appendix C.1.3) requires the first two derivatives of the function to be maximised, so for maximising the log-likelihood function, we need the score function and the Fisher information. Iterative application of the equation

$$\theta^{(t+1)} = \theta^{(t)} + \frac{S(\theta^{(t)})}{I(\theta^{(t)})}$$

gives after convergence (i.e. $\theta^{(t+1)} = \theta^{(t)}$) the MLE $\hat{\theta}_{\text{ML}}$. As a by-product, the observed Fisher information $I(\hat{\theta}_{\text{ML}})$ can also be extracted.

To apply the Newton–Raphson algorithm in `R`, the function `optim` can conveniently be used, see Appendix C.1.3 for details. We need to pass the log-likelihood function as an argument to `optim`. Explicitly passing the score function into `optim` typically accelerates convergence. If the derivative is not available, it can sometimes be computed symbolically using the `R` function `deriv`. Generally no derivatives need to be passed to `optim` because it can approximate them numerically. Particularly useful is the option `hessian = TRUE`, in which case `optim` will also return the negative observed Fisher information.