Fisher information is a key concept in the theory of statistical inference [4,6] and is defined in the following manner: Let \( X = (X_1, ..., X_n) \) be a random sample, and let \( f(X|\theta) \) denote the probability density function for some model of the data, which has parameter vector \( \theta = (\theta_1, \ldots, \theta_k) \). Then the Fisher information matrix \( I_n(\theta) \) of sample size \( n \) is given by the \( k \times k \) symmetric matrix whose \( ij \)-th element is given by the covariance between first partial derivatives of the log-likelihood,

\[
I_n(\theta)_{i,j} = \text{Cov} \left[ \frac{\partial \ln f(X|\theta)}{\partial \theta_i}, \frac{\partial \ln f(X|\theta)}{\partial \theta_j} \right]. \tag{1}
\]

An alternative, but equivalent, definition for the Fisher information matrix is based on the expected values of the second partial derivatives, and is given by

\[
I_n(\theta)_{i,j} = -E \left[ \frac{\partial^2 \ln f(X|\theta)}{\partial \theta_i \partial \theta_j} \right]. \tag{2}
\]

Strictly, this definition corresponds to the expected Fisher information. If no expectation is taken we obtain a data-dependent quantity that is called the observed Fisher information. As a simple example, consider a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), where \( \theta = (\mu, \sigma^2) \). The Fisher information matrix for this situation is given by

\[
I_n(\theta) = \begin{pmatrix}
\frac{n}{\sigma^2} & 0 \\
0 & \frac{n\sigma^4}{2}
\end{pmatrix}.
\]

It is worth noting two useful properties of the Fisher information matrix. Firstly, \( I_n(\theta) = nI_1(\theta) \), meaning that the expected Fisher information for a sample of \( n \) independent observations is equivalent to \( n \) times the Fisher information for a single observation. Secondly, it is dependent on the choice of parameterization. Suppose the parameter \( \theta \) is changed into another parameter \( \eta = (\eta_1, ..., \eta_k) \) with \( \eta_i = g_i(\theta) \) where each \( g_i \) is one-to-one so its inverse \( g_i^{-1}(\eta) = \theta_i \) exists. The Fisher information \( I_n^*(\eta) \) for the new parameterization is obtained using the chain rule [5] as

\[
I_n^*(\eta) = \begin{pmatrix}
\frac{\partial g_1}{\partial \eta_1} & \cdots & \frac{\partial g_1}{\partial \eta_k} \\
\vdots & \ddots & \vdots \\
\frac{\partial g_k}{\partial \eta_1} & \cdots & \frac{\partial g_k}{\partial \eta_k}
\end{pmatrix} \cdot I_n(\theta) \cdot \begin{pmatrix}
\frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_k} \\
\vdots & \ddots & \vdots \\
\frac{\partial g_k}{\partial \theta_1} & \cdots & \frac{\partial g_k}{\partial \theta_k}
\end{pmatrix}.
\]
\[ J(\eta)^T I_n(\theta(\eta)) J(\eta), \] where \( J(\eta) \) is the Jacobian matrix with elements \( J(\eta)_{ij} = \partial g^{-1}_i(\eta)/\partial \eta_j \) (\( i, j = 1, \ldots, k \)), and \( \theta(\eta) = (g^{-1}_1(\eta), \ldots, g^{-1}_k(\eta)) \). In the rest of this article we discuss various applications of the information matrix in statistics.

**The Cramer-Rao Inequality.** Let \( T(X) \) be any statistic and let \( \psi(\theta) \) be its expectation such that \( \psi(\theta) = E[T(X)] \). Under some regularity conditions, it follows that for all \( \theta \),

\[
\text{Var}(T(X)) \geq \frac{(d\psi(\theta) / d\theta)^2}{I_n(\theta)}.
\]

This is called the Cramer-Rao inequality or the information inequality, and the value of the right hand side of (3) is known as the Cramer-Rao lower bound. In particular, if \( T(X) \) is an unbiased estimator for \( \theta \), then the numerator becomes 1, and the lower bound is simply \( 1/I_n(\theta) \). Note that this explains why \( I_n(\theta) \) is called the “information” matrix: The larger the value of \( I_n(\theta) \) is, the smaller the variance becomes, and therefore, we would be more certain about the location of the unknown parameter value. The Cramer-Rao inequality generalizes to the multi-parameter case, where \( \theta = (\theta_1, \ldots, \theta_k) \). Let the statistic \( W(X) \) be an estimator for some function \( g(\theta) \). Then the Cramer-Rao inequality states that \( \text{Var}(W(X)) \geq \gamma(\theta)^T I_n(\theta)^{-1} \gamma(\theta) \) where \( \gamma(\theta) \) is a \( k \times 1 \) column vector with elements \( \gamma(\theta)_i = \partial g(\theta)/\partial \theta_i \).

**Asymptotic Theory.** The maximum likelihood estimator has many useful properties, including reparametrization-invariance, consistency, and sufficiency. Further, it follows under some regularity conditions that the sampling distribution of a maximum likelihood estimator \( \hat{\theta}_{ML} \) is asymptotically unbiased and also asymptotically normal with its variance-covariance matrix obtained from the inverse Fisher information matrix of sample size 1, that is, \( \hat{\theta}_{ML} \rightarrow N(\theta, I_1(\theta)^{-1}/n) \) as \( n \) goes to infinity.

**Bayesian Statistics.** The Fisher information also arises in Bayesian inference. The following noninformative prior, known as Jeffreys’ prior [3], is defined in terms of the Fisher information, \( \pi_j(\theta) \propto \sqrt{|I_j(\theta)|} \) where \( |I_j(\theta)| \) is the determinant of the information matrix. This prior can be useful for three reasons. First, it is reparametrization-invariant so the same prior is obtained under all reparameterizations [3]. Second, Jeffreys’ prior is a uniform density on the space of probability distributions in the sense that it assigns equal mass to each “different” distribution [1]. In comparison, the uniform prior defined as \( \pi_U(\theta) = c \) for some constant \( c \) assigns equal mass to each different value of the parameter and is not reparametrization-invariant. Third, Jeffrey’s prior is the one that maximizes the amount of information about \( \theta \), in the Kullback-Leibler sense, that the data are expected to provide [2].

**References**


