Information Matrix

Jay I. Myung & Daniel J. Navarro
Department of Psychology, Ohio State University
1827 Neil Avenue, Columbus OH 43210, USA.
{myung.1, navarro.20}@osu.edu

Abstract

Fisher information essentially describes the amount of information data provide about an unknown parameter. It has applications in finding the variance of an estimator, as well as in the asymptotic behavior of maximum likelihood estimates, and in Bayesian inference.

Keywords: Cramer-Rao bound, information inequality, Jeffreys' prior.

Fisher information is a key concept in the theory of statistical inference [4,7] and essentially describes the amount of information data provide about an unknown parameter. It has applications in finding the variance of an estimator, as well as in the asymptotic behavior of maximum likelihood estimates, and in Bayesian inference. To define Fisher information, let $\boldsymbol{X} = (X_1, ..., X_n)$ be a random sample, and let $f(\boldsymbol{X}|\boldsymbol{\theta})$ denote the probability density function for some model of the data, which has parameter vector $\boldsymbol{\theta} = (\theta_1, \dots \theta_k)$. 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 deriviatives of the log-likelihood,

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

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

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

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 μ and variance σ^2, where $\boldsymbol{\theta} = (\mu, \sigma^2)$. The Fisher information matrix for this situation is given by $I_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$.

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, that is, how the parameters of a model are combined in the model's equation to define the probability density function. If the parameters are changed into new parameters by describing the latter as a function of the former, then the information matrix of the revised parameters can be found analytically from the information matrix of the old parameters and the function that transforms the old parameters to the new ones [6].

The Cramer-Rao Inequality. Perhaps the most important application of the Fisher information matrix in statistics is in determining an absolute lower bound for the variance of an arbitrary unbiased estimator. Let $T(\boldsymbol{X})$ be any statistic and let $\psi(\theta)$ be its expectation such that $\psi(\theta) = E[T(\boldsymbol{X})]$. Under some regularity conditions, it follows that for all θ,

$$Var(T(\boldsymbol{X})) \geq \frac{\left(\frac{d\psi(\theta)}{d\theta}\right)^2}{I_n(\theta)}. \tag{3}$$

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 famous Cramer-Rao lower bound [5]. In particular, if $T(\boldsymbol{X})$ is an unbiased estimator for θ, 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 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. It is straightforward to generalize the Cramer-Rao inequality to the multi-parameter case [6].

Asymptotic Theory. The maximum likelihood estimator has many useful properties, including reparametrization-invariance, consistency, and sufficiency. Another remarkable property of the estimator is that it achieves the Cramer-Rao minimum variance asymptotically. That is, it follows under some regularity conditions that the sampling distribution of a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$ 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, $\hat{\boldsymbol{\theta}}_{ML} \to N(\boldsymbol{\theta}, I_1(\boldsymbol{\theta})^{-1}/n)$ as n goes to infinity.

Bayesian Statistics. Fisher information also arises in Bayesian inference. The information matrix is used to define a noninformative prior that generalizes the notion of "uniform" but also is equipped with some desirable properties. This is called Jeffreys' prior [3] defined as $\pi_J(\boldsymbol{\theta}) \propto \sqrt{|I_1(\boldsymbol{\theta})|}$ where $|I_1(\boldsymbol{\theta})|$ 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 $\pi_U(\boldsymbol{\theta}) = c$ 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 $\boldsymbol{\theta}$, in the Kullback-Leibler sense, that the data are expected to provide [2].

References

[1] Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation, 9*, 347-368.

[2] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion) *Journal of the Royal Statistical Society, Series B, 4*, 113-147.

[3] Jeffreys, H. (1961). *Theory of Probability* (3rd edition). London, UK: Oxford University Press.

[4] Lehman, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd edition). New York, NY: Springer.

[5] Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society 37*, 81-91; (Republished in S. Kotz & N. Johnson (eds.), *Breakthroughs in Statistics: 1889-1990, vol. 1*).

[6] Schervish, M. J. (1995). *Theory of Statistics.* New York: Springer.

[7] Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data.* Cambridge, UK: Cambridge University Press.