

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

In press in B, Everitt & D. Howel (eds.), *Encyclopedia of Behavioral Statistics*. Wiley.

Mar 11, 2004

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 [4,6] and is defined in the following manner: Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample, and let $f(\mathbf{X}|\boldsymbol{\theta})$ 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 $I_n(\boldsymbol{\theta})$ 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,

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

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

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

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, $I_n(\boldsymbol{\theta}) = nI_1(\boldsymbol{\theta})$, 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 $\boldsymbol{\theta}$ is changed into another parameter $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)$ with $\eta_i = g_i(\boldsymbol{\theta})$ where each g_i is one-to-one so its inverse $g_i^{-1}(\boldsymbol{\eta}) = \theta_i$ exists. The Fisher information $I_n^*(\boldsymbol{\eta})$ for the new parameterization is obtained using the chain rule [5] as $I_n^*(\boldsymbol{\eta}) =$

$J(\boldsymbol{\eta})^T I_n(\boldsymbol{\theta}(\boldsymbol{\eta})) J(\boldsymbol{\eta})$, where $J(\boldsymbol{\eta})$ is the Jacobian matrix with elements $J(\boldsymbol{\eta})_{ij} = \partial g_i^{-1}(\boldsymbol{\eta})/\partial \eta_j$ ($i, j = 1, \dots, k$), and $\boldsymbol{\theta}(\boldsymbol{\eta}) = (g_1^{-1}(\boldsymbol{\eta}), \dots, g_k^{-1}(\boldsymbol{\eta}))$. In the rest of this article we discuss various applications of the information matrix in statistics.

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

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

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 Cramer-Rao lower bound. In particular, if $T(\mathbf{X})$ is an unbiased estimator for $\boldsymbol{\theta}$, then the numerator becomes 1, and the lower bound is simply $1/I_n(\boldsymbol{\theta})$. Note that this explains why $I_n(\boldsymbol{\theta})$ is called the ‘‘information’’ matrix: The larger the the value of $I_n(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Let the statistic $W(\mathbf{X})$ be an estimator for some function $g(\boldsymbol{\theta})$. Then the Cramer-Rao inequality states that $\text{Var}(W(\mathbf{X})) \geq \boldsymbol{\gamma}(\boldsymbol{\theta})^T I_n(\boldsymbol{\theta})^{-1} \boldsymbol{\gamma}(\boldsymbol{\theta})$ where $\boldsymbol{\gamma}(\boldsymbol{\theta})$ is a $k \times 1$ column vector with elements $\boldsymbol{\gamma}(\boldsymbol{\theta})_i = \partial g(\boldsymbol{\theta})/\partial \theta_i$.

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 $\hat{\boldsymbol{\theta}}_{ML}$ 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, $\hat{\boldsymbol{\theta}}_{ML} \rightarrow N(\boldsymbol{\theta}, I_1(\boldsymbol{\theta})^{-1}/n)$ 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, $\pi_J(\boldsymbol{\theta}) \propto \sqrt{|I_1(\boldsymbol{\theta})|}$ 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] Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.
- [6] Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, UK: Cambridge University Press.