Fisher information

In mathematical statistics, the **Fisher information** (sometimes simply called **information**[1]) is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X. Formally, it is the variance of the score, or the expected value of the observed information.

The role of the Fisher information in the asymptotic theory of maximum-likelihood estimation was emphasized by the statistician Ronald Fisher (following some initial results by Francis Ysidro Edgeworth). The Fisher information matrix is used to calculate the covariance matrices associated with maximum-likelihood estimates. It can also be used in the formulation of test statistics, such as the Wald test.

In Bayesian statistics, the Fisher information plays a role in the derivation of non-informative prior distributions according to Jeffreys' rule.[2] It also appears as the large-sample covariance of the posterior distribution, provided that the prior is sufficiently smooth (a result known as Bernstein–von Mises theorem, which was anticipated by Laplace for exponential families).[3]

Statistical systems of a scientific nature (physical, biological, etc.) whose likelihood functions obey shift invariance have been shown to obey maximum Fisher information.[4] The level of the maximum depends upon the nature of the system constraints.

Definition

The Fisher information is a way of measuring the amount of information that an observable random variable $\boldsymbol{X}$ carries about an unknown parameter $\boldsymbol{\theta}$ upon which the probability of $\boldsymbol{X}$ depends. Let $\boldsymbol{f(X;\theta)}$ be the probability density function (or probability mass function) for $\boldsymbol{X}$ conditioned on the value of $\boldsymbol{\theta}$. It describes the probability that we observe a given outcome of $\boldsymbol{X}$, *given* a known value of $\boldsymbol{\theta}$. If $\boldsymbol{f}$ is sharply peaked with respect to changes in $\boldsymbol{\theta}$, it is easy to indicate the "correct" value of $\boldsymbol{\theta}$ from the data, or equivalently, that the data $\boldsymbol{X}$ provides a lot of information about the parameter $\boldsymbol{\theta}$. If $\boldsymbol{f}$ is flat and spread-out, then it would take many samples of $\boldsymbol{X}$ to estimate the actual "true" value of $\boldsymbol{\theta}$ that *would* be obtained using the entire population being sampled. This suggests studying some kind of variance with respect to $\boldsymbol{\theta}$.

Formally, the partial derivative with respect to $\boldsymbol{\theta}$ of the natural logarithm of the likelihood function is called the *score*. Under certain regularity conditions, if $\boldsymbol{\theta}$ is the true parameter (i.e. $\boldsymbol{X}$ is actually distributed as $\boldsymbol{f(X;\theta)}$), it can be shown that the expected value (the first moment) of the score, evaluated at the true parameter value $\boldsymbol{\theta}$, is 0:[5]

$$\begin{aligned}\mathbf{E}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\middle|\theta\right] &= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx \\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(x;\theta)\,dx \\ &= \frac{\partial}{\partial\theta}1 \\ &= 0.\end{aligned}$$

The **Fisher information** is defined to be the variance of the score:[6]

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \middle| \theta\right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx,$$

Note that $0 \leq \mathcal{I}(\theta)$. A random variable carrying high Fisher information implies that the absolute value of the score is often high. The Fisher information is not a function of a particular observation, as the random variable *X* has been averaged out.

If $\log f(x;\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as[7]

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) \middle| \theta\right],$$

since

$$\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right)^2 = \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2$$

and

$$\mathrm{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} \middle| \theta\right] = \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} f(x;\theta)\,dx = 0.$$

Thus, the Fisher information may be seen as the curvature of the support curve (the graph of the log-likelihood). Near the maximum likelihood estimate, low Fisher information therefore indicates that the maximum appears "blunt", that is, the maximum is shallow and there are many nearby values with a similar log-likelihood. Conversely, high Fisher information indicates that the maximum is sharp.

Regularity conditions

The regularity conditions are as follows:[8]

1. The partial derivative of $f(X;\theta)$ with respect to θ exists almost everywhere. (It can fail to exist on a null set, as long as this set does not depend on θ.)
2. The integral of $f(X;\theta)$ can be differentiated under the integral sign with respect to θ.
3. The support of $f(X;\theta)$ does not depend on θ.

If θ is a vector then the regularity conditions must hold for every component of θ. It is easy to find an example of a density that does not satisfy the regularity conditions: The density of a Uniform(0, θ) variable fails to satisfy conditions 1 and 3. In this case, even though the Fisher information can be computed from the definition, it will not have the properties it is typically assumed to have.

In terms of likelihood

Because the likelihood of θ given X is always proportional to the probability $f(X;\theta)$, their logarithms necessarily differ by a constant that is independent of θ, and the derivatives of these logarithms with respect to θ are necessarily equal. Thus one can substitute in a log-likelihood $l(\theta; X)$ instead of $\log f(X;\theta)$ in the

definitions of Fisher Information.

Samples of any size

The value X can represent a single sample drawn from a single distribution or can represent a collection of samples drawn from a collection of distributions. If there are n samples and the corresponding n distributions are statistically independent then the Fisher information will necessarily be the sum of the single-sample Fisher information values, one for each single sample from its distribution. In particular, if the n distributions are independent and identically distributed then the Fisher information will necessarily be n times the Fisher information of a single sample from the common distribution.

Informal derivation of the Cramér–Rao bound

The Cramér–Rao bound[9][10] states that the inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ. H.L. Van Trees (1968) and B. Roy Frieden (2004) provide the following method of deriving the Cramér–Rao bound, a result which describes use of the Fisher information.

Informally, we begin by considering an unbiased estimator $\hat{\theta}(X)$. Mathematically, "unbiased" means that

$$\mathbf{E}\left[\hat{\theta}(X) - \theta \middle| \theta\right] = \int \left(\hat{\theta}(x) - \theta\right) f(x;\theta)\, dx = 0 \text{ regardless of the value of } \theta.$$

This expression is zero independent of θ, so its partial derivative with respect to θ must also be zero. By the product rule, this partial derivative is also equal to

$$0 = \frac{\partial}{\partial\theta} \int \left(\hat{\theta}(x) - \theta\right) f(x;\theta)\, dx = \int \left(\hat{\theta}(x) - \theta\right) \frac{\partial f}{\partial\theta}\, dx - \int f\, dx.$$

For each θ, the likelihood function is a probability density function, and therefore $\int f\, dx = 1$. By using the chain rule on the partial derivative of $\log f$ and then dividing and multiplying by $f(x;\theta)$, one can verify that

$$\frac{\partial f}{\partial\theta} = f \frac{\partial \log f}{\partial\theta}.$$

Using these two facts in the above, we get

$$\int \left(\hat{\theta} - \theta\right) f \frac{\partial \log f}{\partial\theta}\, dx = 1.$$

Factoring the integrand gives

$$\int \left(\left(\hat{\theta} - \theta\right)\sqrt{f}\right)\left(\sqrt{f}\frac{\partial \log f}{\partial\theta}\right) dx = 1.$$

Squaring the expression in the integral, the Cauchy–Schwarz inequality yields

$$1 = \left(\int \left[\left(\hat{\theta} - \theta\right)\sqrt{f}\right]\cdot\left[\sqrt{f}\frac{\partial \log f}{\partial\theta}\right] dx\right)^2 \le \left[\int \left(\hat{\theta} - \theta\right)^2 f\, dx\right]\cdot\left[\int \left(\frac{\partial \log f}{\partial\theta}\right)^2 f\, dx\right].$$

The second bracketed factor is defined to be the Fisher Information, while the first bracketed factor is the expected mean-squared error of the estimator $\hat{\theta}$. By rearranging, the inequality tells us that

$$\operatorname{Var}\left(\hat{\theta}\right) \geq \frac{1}{\mathcal{I}(\theta)}.$$

In other words, the precision to which we can estimate θ is fundamentally limited by the Fisher information of the likelihood function.

Single-parameter Bernoulli experiment

A Bernoulli trial is a random variable with two possible outcomes, "success" and "failure", with success having a probability of θ. The outcome can be thought of as determined by a coin toss, with the probability of heads being θ and the probability of tails being $1 - \theta$.

Let X be a Bernoulli trial. The Fisher information contained in X may be calculated to be

$$\begin{aligned}
\mathcal{I}(\theta) &= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\middle|\theta\right] \\
&= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\left(X\log\theta + (1-X)\log(1-\theta)\right)\middle|\theta\right] \\
&= \operatorname{E}\left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}\middle|\theta\right] \\
&= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\
&= \frac{1}{\theta(1-\theta)}.
\end{aligned}$$

Because Fisher information is additive, the Fisher information contained in n independent Bernoulli trials is therefore

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}.$$

This is the reciprocal of the variance of the mean number of successes in n Bernoulli trials, so in this case, the Cramér–Rao bound is an equality.

Matrix form

When there are N parameters, so that θ is an $N \times 1$ vector $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2 & \dots & \theta_N\end{bmatrix}^{\mathsf{T}}$, then the Fisher information takes the form of an $N \times N$ matrix. This matrix is called the **Fisher information matrix** (FIM) and has typical element

$$\left[\mathcal{I}(\theta)\right]_{i,j} = \operatorname{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\middle|\theta\right].$$

The FIM is a $N \times N$ positive semidefinite matrix. If it is positive definite, then it defines a Riemannian metric on the N-dimensional parameter space. The topic information geometry uses this to connect Fisher information to differential geometry, and in that context, this metric is known as the Fisher information metric.

Under certain regularity conditions, the Fisher information matrix may also be written as

$$\left[\mathcal{I}(\theta)\right]_{i,j} = -\operatorname{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(X;\theta)\,\middle|\,\theta\right].$$

The result is interesting in several ways:

- It can be derived as the Hessian of the relative entropy.
- It can be used as a Riemannian metric for defining Fisher-Rao geometry when it is positive-definite.[11]
- It can be understood as a metric induced from the Euclidean metric, after appropriate change of variable.
- In its complex-valued form, it is the Fubini–Study metric.
- It is the key part of the proof of Wilks' theorem, which allows confidence region estimates for maximum likelihood estimation (for those conditions for which it applies) without needing the Likelihood Principle.
- In cases where the analytical calculations of the FIM above are difficult, it is possible to form an average of easy Monte Carlo estimates of the Hessian of the negative log-likelihood function as an estimate of the FIM.[12][13][14] The estimates may be based on values of the negative log-likelihood function or the gradient of the negative log-likelihood function; no analytical calculation of the Hessian of the negative log-likelihood function is needed.

Orthogonal parameters

We say that two parameters θ_i and θ_j are orthogonal if the element of the ith row and jth column of the Fisher information matrix is zero. Orthogonal parameters are easy to deal with in the sense that their maximum likelihood estimates are independent and can be calculated separately. When dealing with research problems, it is very common for the researcher to invest some time searching for an orthogonal parametrization of the densities involved in the problem.

Singular statistical model

If the Fisher information matrix is positive definite for all θ, then the corresponding statistical model is said to be *regular*; otherwise, the statistical model is said to be *singular*.[15] Examples of singular statistical models include the following: normal mixtures, binomial mixtures, multinomial mixtures, Bayesian networks, neural networks, radial basis functions, hidden Markov models, stochastic context-free grammars, reduced rank regressions, Boltzmann machines.

In machine learning, if a statistical model is devised so that it extracts hidden structure from a random phenomenon, then it naturally becomes singular.[16]

Multivariate normal distribution

The FIM for a *N*-variate multivariate normal distribution, $X \sim N\left(\mu(\theta),\, \Sigma(\theta)\right)$ has a special form. Let the *K*-dimensional vector of parameters be $\theta = \begin{bmatrix} \theta_1 & \dots & \theta_K \end{bmatrix}^\mathsf{T}$ and the vector of random normal variables be $X = \begin{bmatrix} X_1 & \dots & X_N \end{bmatrix}^\mathsf{T}$. Assume that the mean values of these random variables are $\mu(\theta) = \begin{bmatrix} \mu_1(\theta) & \dots & \mu_N(\theta) \end{bmatrix}^\mathsf{T}$, and let $\Sigma(\theta)$ be the covariance matrix. Then, for $1 \le m,\, n \le K$, the (*m*, *n*) entry of the FIM is:[17]

$$\mathcal{I}_{m,n} = \frac{\partial \mu^\mathsf{T}}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n} + \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_m} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_n} \right),$$

where $(\cdot)^\mathsf{T}$ denotes the transpose of a vector, $\operatorname{tr}(\cdot)$ denotes the trace of a square matrix, and:

$$\frac{\partial \mu}{\partial \theta_m} = \begin{bmatrix} \dfrac{\partial \mu_1}{\partial \theta_m} & \dfrac{\partial \mu_2}{\partial \theta_m} & \cdots & \dfrac{\partial \mu_N}{\partial \theta_m} \end{bmatrix}^\mathsf{T};$$

$$\frac{\partial \Sigma}{\partial \theta_m} = \begin{bmatrix} \dfrac{\partial \Sigma_{1,1}}{\partial \theta_m} & \dfrac{\partial \Sigma_{1,2}}{\partial \theta_m} & \cdots & \dfrac{\partial \Sigma_{1,N}}{\partial \theta_m} \\ \dfrac{\partial \Sigma_{2,1}}{\partial \theta_m} & \dfrac{\partial \Sigma_{2,2}}{\partial \theta_m} & \cdots & \dfrac{\partial \Sigma_{2,N}}{\partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial \Sigma_{N,1}}{\partial \theta_m} & \dfrac{\partial \Sigma_{N,2}}{\partial \theta_m} & \cdots & \dfrac{\partial \Sigma_{N,N}}{\partial \theta_m} \end{bmatrix}.$$

Note that a special, but very common, case is the one where $\Sigma(\theta) = \Sigma$, a constant. Then

$$\mathcal{I}_{m,n} = \frac{\partial \mu^\mathsf{T}}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n}.$$

In this case the Fisher information matrix may be identified with the coefficient matrix of the normal equations of least squares estimation theory.

Another special case occurs when the mean and covariance depend on two different vector parameters, say, β and θ. This is especially popular in the analysis of spatial data, which often uses a linear model with correlated residuals. In this case,[18]

$$\mathcal{I}(\beta, \theta) = \operatorname{diag}(\mathcal{I}(\beta), \mathcal{I}(\theta))$$

where

$$\mathcal{I}(\beta)_{m,n} = \frac{\partial \mu^{\mathsf{T}}}{\partial \beta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \beta_n},$$

$$\mathcal{I}(\theta)_{m,n} = \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_m} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_n} \right)$$

Properties

Chain rule

Similar to the entropy or mutual information, the Fisher information also possesses a **chain rule** decomposition. In particular, if *X* and *Y* are jointly distributed random variables, it follows that:[19]

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_{Y|X}(\theta),$$

where $\mathcal{I}_{Y|X}(\theta) = \mathbf{E}_X\left[\mathcal{I}_{Y|X=x}(\theta)\right]$ and $\mathcal{I}_{Y|X=x}(\theta)$ is the Fisher information of *Y* relative to θ calculated with respect to the conditional density of *Y* given a specific value *X* = *x*.

As a special case, if the two random variables are independent, the information yielded by the two random variables is the sum of the information from each random variable separately:

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta).$$

Consequently, the information in a random sample of *n* independent and identically distributed observations is *n* times the information in a sample of size 1.

F-divergence

Given a convex function $f : [0, \infty) \to (-\infty, \infty]$ that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t \to 0^+} f(t)$, (which could be infinite), it defines an f-divergence D_f. Then if f is strictly convex at 1, then locally at $\theta \in \Theta$, the Fisher information matrix is a metric, in the sense that[20]

$$(\delta\theta)^T I(\theta)(\delta\theta) = \frac{1}{f''(1)} D_f(P_{\theta+\delta\theta} \| P_\theta)$$

where P_θ is the distribution parametrized by θ. That is, it's the distribution with pdf $f(x;\theta)$.

In this form, it is clear that the Fisher information matrix is a Riemannian metric, and varies correctly under a change of variables. (see section on Reparametrization)

Sufficient statistic

The information provided by a sufficient statistic is the same as that of the sample *X*. This may be seen by using Neyman's factorization criterion for a sufficient statistic. If *T*(*X*) is sufficient for *θ*, then

$$f(X;\theta) = g(T(X),\theta)h(X)$$

for some functions *g* and *h*. The independence of *h*(*X*) from *θ* implies

$$\frac{\partial}{\partial\theta}\log[f(X;\theta)] = \frac{\partial}{\partial\theta}\log[g(T(X);\theta)],$$

and the equality of information then follows from the definition of Fisher information. More generally, if $T = t(X)$ is a statistic, then

$$\mathcal{I}_T(\theta) \leq \mathcal{I}_X(\theta)$$

with equality if and only if T is a sufficient statistic.[21]

Reparametrization

The Fisher information depends on the parametrization of the problem. If θ and η are two scalar parametrizations of an estimation problem, and θ is a continuously differentiable function of η, then

$$\mathcal{I}_\eta(\eta) = \mathcal{I}_\theta(\theta(\eta))\left(\frac{d\theta}{d\eta}\right)^2$$

where $\mathcal{I}_\eta$ and $\mathcal{I}_\theta$ are the Fisher information measures of η and θ, respectively.[22]

In the vector case, suppose $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are k-vectors which parametrize an estimation problem, and suppose that $\boldsymbol{\theta}$ is a continuously differentiable function of $\boldsymbol{\eta}$, then,[23]

$$\mathcal{I}_\eta(\boldsymbol{\eta}) = \boldsymbol{J}^{\mathsf{T}}\mathcal{I}_\theta(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{J}$$

where the (i, j)th element of the $k \times k$ Jacobian matrix $\boldsymbol{J}$ is defined by

$$J_{ij} = \frac{\partial\theta_i}{\partial\eta_j},$$

and where $\boldsymbol{J}^{\mathsf{T}}$ is the matrix transpose of $\boldsymbol{J}$.

In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher–Rao metric) for the manifold of thermodynamic states, and can be used as an information-geometric complexity measure for a classification of phase transitions, e.g., the scalar curvature of the thermodynamic metric tensor diverges at (and only at) a phase transition point.[24]

In the thermodynamic context, the Fisher information matrix is directly related to the rate of change in the corresponding order parameters.[25] In particular, such relations identify second-order phase transitions via divergences of individual elements of the Fisher information matrix.

Isoperimetric inequality

The Fisher information matrix plays a role in an inequality like the isoperimetric inequality.[26] Of all probability distributions with a given entropy, the one whose Fisher information matrix has the smallest trace is the Gaussian distribution. This is like how, of all bounded sets with a given volume, the sphere has the smallest surface area.

The proof involves taking a multivariate random variable X with density function f and adding a location parameter to form a family of densities $\{f(x-\theta) \mid \theta \in \mathbb{R}^n\}$. Then, by analogy with the Minkowski–Steiner formula, the "surface area" of X is defined to be

$$S(X) = \lim_{\varepsilon \to 0} \frac{e^{H(X+Z_\varepsilon)} - e^{H(X)}}{\varepsilon}$$

where Z_ε is a Gaussian variable with covariance matrix εI. The name "surface area" is apt because the entropy power $e^{H(X)}$ is the volume of the "effective support set,"[27] so $S(X)$ is the "derivative" of the volume of the effective support set, much like the Minkowski-Steiner formula. The remainder of the proof uses the entropy power inequality, which is like the Brunn–Minkowski inequality. The trace of the Fisher information matrix is found to be a factor of $S(X)$.

Applications

Optimal design of experiments

Fisher information is widely used in optimal experimental design. Because of the reciprocity of estimator-variance and Fisher information, *minimizing* the *variance* corresponds to *maximizing* the *information*.

When the linear (or linearized) statistical model has several parameters, the mean of the parameter estimator is a vector and its variance is a matrix. The inverse of the variance matrix is called the "information matrix". Because the variance of the estimator of a parameter vector is a matrix, the problem of "minimizing the variance" is complicated. Using statistical theory, statisticians compress the information-matrix using real-valued summary statistics; being real-valued functions, these "information criteria" can be maximized.

Traditionally, statisticians have evaluated estimators and designs by considering some summary statistic of the covariance matrix (of an unbiased estimator), usually with positive real values (like the determinant or matrix trace). Working with positive real numbers brings several advantages: If the estimator of a single parameter has a positive variance, then the variance and the Fisher information are both positive real numbers; hence they are members of the convex cone of nonnegative real numbers (whose nonzero members have reciprocals in this same cone).

For several parameters, the covariance matrices and information matrices are elements of the convex cone of nonnegative-definite symmetric matrices in a partially ordered vector space, under the Loewner (Löwner) order. This cone is closed under matrix addition and inversion, as well as under the multiplication of positive real numbers and matrices. An exposition of matrix theory and Loewner order appears in Pukelsheim.[28]

The traditional optimality criteria are the information matrix's invariants, in the sense of invariant theory; algebraically, the traditional optimality criteria are functionals of the eigenvalues of the (Fisher) information matrix (see optimal design).

Jeffreys prior in Bayesian statistics

In Bayesian statistics, the Fisher information is used to calculate the Jeffreys prior, which is a standard, non-informative prior for continuous distribution parameters.[29]

Computational neuroscience

The Fisher information has been used to find bounds on the accuracy of neural codes. In that case, X is typically the joint responses of many neurons representing a low dimensional variable θ (such as a stimulus parameter). In particular the role of correlations in the noise of the neural responses has been studied.[30]

Derivation of physical laws

Fisher information plays a central role in a controversial principle put forward by Frieden as the basis of physical laws, a claim that has been disputed.[31]

Machine learning

The Fisher information is used in machine learning techniques such as elastic weight consolidation,[32] which reduces catastrophic forgetting in artificial neural networks.

Fisher information can be used as an alternative to the Hessian of the loss function in second-order gradient descent network training.[33]

Relation to relative entropy

Fisher information is related to relative entropy.[34] The relative entropy, or Kullback–Leibler divergence, between two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ can be written as

$$KL(p:q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Now, consider a family of probability distributions $f(x;\theta)$ parametrized by $\theta \in \Theta$. Then the Kullback–Leibler divergence, between two distributions in the family can be written as

$$D(\theta, \theta') = KL(p(\cdot\,;\theta) : p(\cdot\,;\theta')) = \int f(x;\theta) \log \frac{f(x;\theta)}{f(x;\theta')}\, dx.$$

If $\boldsymbol{\theta}$ is fixed, then the relative entropy between two distributions of the same family is minimized at $\boldsymbol{\theta' = \theta}$. For $\boldsymbol{\theta'}$ close to $\boldsymbol{\theta}$, one may expand the previous expression in a series up to second order:

$$D(\theta, \theta') = \frac{1}{2} (\theta' - \theta)^{\mathsf{T}} \left(\frac{\partial^2}{\partial \theta'_i\, \partial \theta'_j} D(\theta, \theta') \right)_{\theta' = \theta} (\theta' - \theta) + o\left((\theta' - \theta)^2\right)$$

But the second order derivative can be written as

$$\left(\frac{\partial^2}{\partial \theta'_i\, \partial \theta'_j} D(\theta, \theta') \right)_{\theta' = \theta} = -\int f(x;\theta) \left(\frac{\partial^2}{\partial \theta'_i\, \partial \theta'_j} \log(f(x;\theta')) \right)_{\theta' = \theta} dx = [\mathcal{I}(\theta)]_{i,j}.$$

Thus the Fisher information represents the curvature of the relative entropy of a conditional distribution with respect to its parameters.

History

The Fisher information was discussed by several early statisticians, notably F. Y. Edgeworth.[35] For example, Savage[36] says: "In it [Fisher information], he [Fisher] was to some extent anticipated (Edgeworth 1908–9 esp. 502, 507–8, 662, 677–8, 82–5 and references he [Edgeworth] cites including Pearson and Filon 1898 [. . .])." There are a number of early historical sources[37] and a number of reviews of this early work.[38][39][40]

See also

- Efficiency (statistics)
- Observed information
- Fisher information metric
- Formation matrix
- Information geometry
- Jeffreys prior
- Cramér–Rao bound
- Minimum Fisher information
- Quantum Fisher information

Other measures employed in information theory:

- Entropy (information theory)
- Kullback–Leibler divergence
- Self-information

Notes

1. Lehmann & Casella, p. 115
2. Robert, Christian (2007). "Noninformative prior distributions". *The Bayesian Choice* (2nd ed.). Springer. pp. 127–141. ISBN 978-0-387-71598-8.
3. Le Cam, Lucien (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer. pp. 618–621. ISBN 0-387-96307-3.
4. Frieden & Gatenby (2013)
5. Suba Rao. "Lectures on statistical inference" (http://www.stat.tamu.edu/~suhasini/teaching613/inference.pdf) (PDF).
6. Fisher (1922)
7. Lehmann & Casella, eq. (2.5.16), Lemma 5.3, p.116.
8. Schervish, Mark J. (1995). *Theory of Statistics* (https://www.worldcat.org/oclc/852790658). New York, NY: Springer New York. p. 111. ISBN 978-1-4612-4250-5. OCLC 852790658 (https://www.worldcat.org/oclc/852790658).
9. Cramer (1946)
10. Rao (1945)
11. Nielsen, Frank (2010). "Cramer-Rao lower bound and information geometry" (https://arxiv.org/abs/1301.3578). *Connected at Infinity II*: 18–37. arXiv:1301.3578 (https://arxiv.org/abs/1301.3578).
12. Spall, J. C. (2005). "Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings". *Journal of Computational and Graphical Statistics*. **14** (4): 889–909. doi:10.1198/106186005X78800 (https://doi.org/10.1198%2F106186005X78800). S2CID 16090098 (https://api.semanticscholar.org/CorpusID:16090098).

13. Spall, J. C. (2008), "Improved Methods for Monte Carlo Estimation of the Fisher Information Matrix," *Proceedings of the American Control Conference*, Seattle, WA, 11–13 June 2008, pp. 2395–2400. https://doi.org/10.1109/ACC.2008.4586850
14. Das, S.; Spall, J. C.; Ghanem, R. (2010). "Efficient Monte Carlo Computation of Fisher Information Matrix Using Prior Information". *Computational Statistics and Data Analysis*. **54** (2): 272–289. doi:10.1016/j.csda.2009.09.018 (https://doi.org/10.1016%2Fj.csda.2009.09.018).
15. Watanabe, S. (2008), Accardi, L.; Freudenberg, W.; Ohya, M. (eds.), "Algebraic geometrical method in singular statistical estimation", *Quantum Bio-Informatics*, World Scientific: 325–336, Bibcode:2008qbi..conf..325W (https://ui.adsabs.harvard.edu/abs/2008qbi..conf..325W), doi:10.1142/9789812793171_0024 (https://doi.org/10.1142%2F9789812793171_0024), ISBN 978-981-279-316-4.
16. Watanabe, S (2013). "A Widely Applicable Bayesian Information Criterion". *Journal of Machine Learning Research*. **14**: 867–897.
17. Malagò, Luigi; Pistone, Giovanni (2015). *Information geometry of the Gaussian distribution in view of stochastic optimization*. *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*. pp. 150–162. doi:10.1145/2725494.2725510 (https://doi.org/10.1145%2F2725494.2725510). ISBN 9781450334341. S2CID 693896 (https://api.semanticscholar.org/CorpusID:693896).
18. Mardia, K. V.; Marshall, R. J. (1984). "Maximum likelihood estimation of models for residual covariance in spatial regression". *Biometrika*. **71** (1): 135–46. doi:10.1093/biomet/71.1.135 (https://doi.org/10.1093%2Fbiomet%2F71.1.135).
19. Zamir, R. (1998). "A proof of the Fisher information inequality via a data processing argument". *IEEE Transactions on Information Theory*. **44** (3): 1246–1250. CiteSeerX 10.1.1.49.6628 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.6628). doi:10.1109/18.669301 (https://doi.org/10.1109%2F18.669301).
20. Polyanskiy, Yury (2017). "Lecture notes on information theory, chapter 29, ECE563 (UIUC)" (https://people.lids.mit.edu/yp/homepage/data/LN_stats.pdf) (PDF). *Lecture notes on information theory*. Archived (https://web.archive.org/web/20220524014051/https://people.lids.mit.edu/yp/homepage/data/LN_stats.pdf) (PDF) from the original on 2022-05-24. Retrieved 2022-05-24.
21. Schervish, Mark J. (1995). *Theory Statistics*. Springer-Verlag. p. 113.
22. Lehmann & Casella, eq. (2.5.11).
23. Lehmann & Casella, eq. (2.6.16)
24. Janke, W.; Johnston, D. A.; Kenna, R. (2004). "Information Geometry and Phase Transitions". *Physica A*. **336** (1–2): 181. arXiv:cond-mat/0401092 (https://arxiv.org/abs/cond-mat/0401092). Bibcode:2004PhyA..336..181J (https://ui.adsabs.harvard.edu/abs/2004PhyA..336..181J). doi:10.1016/j.physa.2004.01.023 (https://doi.org/10.1016%2Fj.physa.2004.01.023). S2CID 119085942 (https://api.semanticscholar.org/CorpusID:119085942).
25. Prokopenko, M.; Lizier, Joseph T.; Lizier, J. T.; Obst, O.; Wang, X. R. (2011). "Relating Fisher information to order parameters". *Physical Review E*. **84** (4): 041116. Bibcode:2011PhRvE..84d1116P (https://ui.adsabs.harvard.edu/abs/2011PhRvE..84d1116P). doi:10.1103/PhysRevE.84.041116 (https://doi.org/10.1103%2FPhysRevE.84.041116). PMID 22181096 (https://pubmed.ncbi.nlm.nih.gov/22181096). S2CID 18366894 (https://api.semanticscholar.org/CorpusID:18366894).
26. Costa, M.; Cover, T. (Nov 1984). "On the similarity of the entropy power inequality and the Brunn-Minkowski inequality" (https://ieeexplore.ieee.org/document/1056983). *IEEE Transactions on Information Theory*. **30** (6): 837–839. doi:10.1109/TIT.1984.1056983 (https://doi.org/10.1109%2FTIT.1984.1056983). ISSN 1557-9654 (https://www.worldcat.org/issn/1557-9654).
27. Cover, Thomas M. (2006). *Elements of information theory* (https://www.worldcat.org/oclc/59879802). Joy A. Thomas (2nd ed.). Hoboken, N.J.: Wiley-Interscience. p. 256. ISBN 0-471-24195-4. OCLC 59879802 (https://www.worldcat.org/oclc/59879802).

28. Pukelsheim, Friedrick (1993). *Optimal Design of Experiments*. New York: Wiley. ISBN 978-0-471-61971-0.
29. Bernardo, Jose M.; Smith, Adrian F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons. ISBN 978-0-471-92416-6.
30. Abbott, Larry F.; Dayan, Peter (1999). "The effect of correlated variability on the accuracy of a population code". *Neural Computation*. **11** (1): 91–101. doi:10.1162/089976699300016827 (https://doi.org/10.1162%2F089976699300016827). PMID 9950724 (https://pubmed.ncbi.nlm.nih.gov/9950724). S2CID 2958438 (https://api.semanticscholar.org/CorpusID:2958438).
31. Streater, R. F. (2007). *Lost Causes in and beyond Physics*. Springer. p. 69. ISBN 978-3-540-36581-5.
32. Kirkpatrick, James; Pascanu, Razvan; Rabinowitz, Neil; Veness, Joel; Desjardins, Guillaume; Rusu, Andrei A.; Milan, Kieran; Quan, John; Ramalho, Tiago (2017-03-28). "Overcoming catastrophic forgetting in neural networks" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5380101). *Proceedings of the National Academy of Sciences*. **114** (13): 3521–3526. doi:10.1073/pnas.1611835114 (https://doi.org/10.1073%2Fpnas.1611835114). ISSN 0027-8424 (https://www.worldcat.org/issn/0027-8424). PMC 5380101 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5380101). PMID 28292907 (https://pubmed.ncbi.nlm.nih.gov/28292907).
33. Martens, James (August 2020). "New Insights and Perspectives on the Natural Gradient Method" (https://arxiv.org/pdf/1412.1193.pdf) (PDF). *Journal of Machine Learning Research* (21). arXiv:1412.1193 (https://arxiv.org/abs/1412.1193).
34. Gourieroux & Montfort (1995), page 87 (https://books.google.com/books?id=gqI-pAP2JZ8C&pg=PA87)
35. Savage (1976)
36. Savage(1976), page 156
37. Edgeworth (September 1908, December 1908)
38. Pratt (1976)
39. Stigler (1978, 1986, 1999)
40. Hald (1998, 1999)

References

- Cramér, Harald (1946). *Mathematical methods of statistics*. Princeton mathematical series. Princeton: Princeton University Press. ISBN 0691080046.
- Edgeworth, F. Y. (Jun 1908). "On the Probable Errors of Frequency-Constants" (https://zenodo.org/record/1449470). *Journal of the Royal Statistical Society*. **71** (2): 381–397. doi:10.2307/2339461 (https://doi.org/10.2307%2F2339461). JSTOR 2339461 (https://www.jstor.org/stable/2339461).
- Edgeworth, F. Y. (Sep 1908). "On the Probable Errors of Frequency-Constants (Contd.)" (https://zenodo.org/record/1449468). *Journal of the Royal Statistical Society*. **71** (3): 499–512. doi:10.2307/2339293 (https://doi.org/10.2307%2F2339293). JSTOR 2339293 (https://www.jstor.org/stable/2339293).
- Edgeworth, F. Y. (Dec 1908). "On the Probable Errors of Frequency-Constants (Contd.)" (https://zenodo.org/record/1449468). *Journal of the Royal Statistical Society*. **71** (4): 651–678. doi:10.2307/2339378 (https://doi.org/10.2307%2F2339378). JSTOR 2339378 (https://www.jstor.org/stable/2339378).
- Fisher, R. A. (1922-01-01). "On the mathematical foundations of theoretical statistics" (https://doi.org/10.1098%2Frsta.1922.0009). *Philosophical Transactions of the Royal Society of London, Series A*. **222** (594–604): 309–368. Bibcode:1922RSPTA.222..309F (https://ui.adsabs.harvard.edu/abs/1922RSPTA.222..309F). doi:10.1098/rsta.1922.0009 (https://doi.org/10.1098%2Frsta.1922.0009).

- Frieden, B. R. (2004) *Science from Fisher Information: A Unification*. Cambridge Univ. Press. ISBN 0-521-00911-1.
- Frieden, B. Roy; Gatenby, Robert A. (2013). "Principle of maximum Fisher information from Hardy's axioms applied to statistical systems" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010149). *Physical Review E*. **88** (4): 042144. arXiv:1405.0007 (https://arxiv.org/abs/1405.0007). Bibcode:2013PhRvE..88d2144F (https://ui.adsabs.harvard.edu/abs/2013PhRvE..88d2144F). doi:10.1103/PhysRevE.88.042144 (https://doi.org/10.1103%2FPhysRevE.88.042144). PMC 4010149 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010149). PMID 24229152 (https://pubmed.ncbi.nlm.nih.gov/24229152).
- Hald, A. (May 1999). "On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares" (https://doi.org/10.1214%2Fss%2F1009212248). *Statistical Science*. **14** (2): 214–222. doi:10.1214/ss/1009212248 (https://doi.org/10.1214%2Fss%2F1009212248). JSTOR 2676741 (https://www.jstor.org/stable/2676741).
- Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley. ISBN 978-0-471-17912-2.
- Lehmann, E. L.; Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer. ISBN 978-0-387-98502-2.
- Le Cam, Lucien (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag. ISBN 978-0-387-96307-5.
- Pratt, John W. (May 1976). "F. Y. Edgeworth and R. A. Fisher on the Efficiency of Maximum Likelihood Estimation" (https://doi.org/10.1214%2Faos%2F1176343457). *Annals of Statistics*. **4** (3): 501–514. doi:10.1214/aos/1176343457 (https://doi.org/10.1214%2Faos%2F1176343457). JSTOR 2958222 (https://www.jstor.org/stable/2958222).
- Rao, C. Radhakrishna (1945). "Information and accuracy attainable in the estimation of statistical parameters". *Bulletin of the Calcutta Mathematical Society*. Springer Series in Statistics. **37**: 81–91. doi:10.1007/978-1-4612-0919-5_16 (https://doi.org/10.1007%2F978-1-4612-0919-5_16). ISBN 978-0-387-94037-3.
- Savage, L. J. (May 1976). "On Rereading R. A. Fisher" (https://doi.org/10.1214%2Faos%2F1176343456). *Annals of Statistics*. **4** (3): 441–500. doi:10.1214/aos/1176343456 (https://doi.org/10.1214%2Faos%2F1176343456). JSTOR 2958221 (https://www.jstor.org/stable/2958221).
- Schervish, Mark J. (1995). *Theory of Statistics*. New York: Springer. ISBN 978-0-387-94546-0.
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900* (https://archive.org/details/historyofstatist00stig). Harvard University Press. ISBN 978-0-674-40340-6.
- Stigler, S. M. (1978). "Francis Ysidro Edgeworth, Statistician". *Journal of the Royal Statistical Society, Series A*. **141** (3): 287–322. doi:10.2307/2344804 (https://doi.org/10.2307%2F2344804). JSTOR 2344804 (https://www.jstor.org/stable/2344804).
- Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press. ISBN 978-0-674-83601-3.
- Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley. ISBN 978-0-471-09517-0.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Fisher_information&oldid=1135625401"