# The Fisher-Rao Metric

**S.J. Maybank** CMath FIMA School of Computer Science and Information Systems, Birkbeck College, Malet Street, London, WC1E 7HX, UK.

**Abstract.**

*The information in a mass of data is summarised by fitting to the data a probability density function (pdf) chosen from a parameterised family of pdfs. Distances between pairs of pdfs are calculated using the Fisher-Rao metric, which provides a measure of the accuracy with which parameter values can be estimated from the data. Applications of the Fisher-Rao metric include the calculation of Bayesian priors, machine learning and image processing.*

## Introduction

**One of the tasks** of the statistician is to summarise the information in a mass of data by calculating the values of a relatively small number of parameters. In many cases these parameters specify a probability density function (pdf) which is said to be a model for the data. A fundamental problem arises because the same pdf can be specified by many different sets of parameters. For example, a Gaussian pdf can be specified by the mean, $\mu$, and the standard deviation, $\sigma$, or alternatively, it can be specified by the scaled mean, $\tilde{\mu} = \mu\sigma^{-1}$ and the precision $\tau = \sigma^{-2}$. If the parameters $\mu, \sigma$ are known to within a certain accuracy, then to what accuracy are the parameters $\tilde{\mu}, \tau$ known? More generally, suppose that there are two candidate Gaussian pdfs for modelling the data, one with parameters $(\mu_1, \sigma_1)$ and the other with parameters $(\mu_2, \sigma_2)$. Are the two Gaussian pdfs so similar that they are in effect the same model for the data, or do they differ significantly? Should the answer to this question be based on $(\mu_1, \sigma_1)$, $(\mu_2, \sigma_2)$ or on $(\tilde{\mu}_1, \tau_1)$, $(\tilde{\mu}_2, \tau_2)$? A moment's thought shows that the underlying objects of interest are the pdfs, and that the parameters serve only to specify particular pdfs. It follows that any meaningful comparison of the two pdfs should be independent of the choice of parameters. Does such a way of comparing pdfs exist?

Two pioneers of statistics, Ronald Fisher and Calyampudi Rao, found a way of calculating a distance between two pdfs. Fisher regarded the data as a random sample from a hypothetical space of possible data, noted that the parameter values calculated from the sampled data differ from the true parameter values and obtained an expression for the inverse covariance of these differences. The covariance measures the accuracy of the calculated parameters. Rao showed that the inverse covariance, later known as Fisher information, defines a Riemannian metric on the parameter space. This Fisher-Rao metric gives the correct distance between pdfs. If the parameterisation is changed then the description of the Fisher-Rao metric changes but the calculated distance between any two given pdfs remains the same.

## Maximum likelihood and Fisher information

In [3] Fisher divided the problems associated with the summarisation of data into three types. The first type of problem is to find a parameterised family of pdfs which is likely to contain a model for the data. To solve this problem, Fisher recommended the prior knowledge of the practical statistician. The second type is to calculate from the data the parameter values of the pdf that best models the data. The third type is to obtain a pdf for the *parameter values*, in order to assess how near the calculated values are to the true but unknown values. The investigation of problems of the second and third types led to the discovery of the Fisher-Rao metric on the parameter space.

In order to solve the second type of problem, it is necessary to find a general principle on which to base parameter estimation. Fisher rejected the Bayesian approach to parameter estimation, because there was no clear way of specifying a prior density on the parameter space [3]. Instead, he removed parameter estimation from out of the realm of probability theory by defining a new fundamental quantity, the likelihood, and advocating the maximum likelihood estimation of parameter values. Maximum likelihood estimation is defined as follows. Suppose that the pdf for data $x \in \mathbb{R}^k$ is $f(x|\theta)$, depending on a parameter vector $\theta \in \mathbb{R}^m$, and suppose that $n$ independent samples $x_1, \ldots, x_n$ from $f(x|\theta)$ are available. The joint pdf for the $n$ samples is $\Pi_{i=1}^n f(x_i|\theta)$, and the maximum likelihood estimate $\theta_1$ of $\theta$ is

$$\theta_1 = \arg\max \theta \mapsto \prod_{i=1}^n f(x_i|\theta).$$

It turned out that Bayesian parameter estimation was not so easily dismissed. Maximum likelihood estimation led to the Fisher-Rao metric, which in turn led to the definition of a prior density suitable for Bayesian estimation.

One key advantage of the maximum likelihood estimate is that it behaves correctly under reparameterisation of the parameter space. If $\psi$ is a new parameterisation and $\psi_1$ is the maximum likelihood estimate of $\psi$, then $\psi_1 = \psi(\theta_1)$. Another advantage is that the problems of the third type can be tackled. Fisher showed that if $m = 1$ and $\theta_1$ has a Gaussian distribution with expected value $\theta$ and standard deviation $\sigma(\theta)$, then

$$\sigma(\theta)^{-2} = -n\int_{\mathbb{R}^k} \left( \frac{\partial^2}{\partial\theta^2} \ln f(x|\theta) \right) f(x|\theta)\, dx. \qquad (1)$$

The right hand side of (1) is the Fisher information. The definition of Fisher information is readily extended to the case $m > 1$, resulting in the definition of the Fisher information matrix,

$$I(\theta)_{ij} = -n\int_{\mathbb{R}^k} \left( \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f(x|\theta) \right) f(x|\theta)\, dx, \quad 1 \le i, j \le m. \qquad (2)$$

The conditions under which Fisher obtained (1) can be relaxed if $n$ is large. As $n \to \infty$, and under suitable conditions on the parameterised family, $\theta \mapsto f(x|\theta)$, of pdfs, the distribution of $\theta_1$ tends to a Gaussian distribution with expected value $\theta$ and covariance $I(\theta)^{-1}$.

## Distances in the parameter space

Suppose that $\theta$, $\tilde{\theta}$ are two parameter values in $\mathbb{R}$ and $\theta_1$, $\tilde{\theta}_1$ are maximum likelihood estimates of $\theta$, $\tilde{\theta}$. Thus, $\theta_1$ is calculated using samples drawn from $f(x|\theta)$ and $\tilde{\theta}_1$ is calculated using samples from $f(x|\tilde{\theta})$. The accuracy of the two estimates $\theta_1$, $\tilde{\theta}_1$ is the same if

$$\frac{|\theta_1 - \theta|}{\sigma(\theta)} = \frac{|\tilde{\theta}_1 - \tilde{\theta}|}{\sigma(\tilde{\theta})}.$$

This suggests that the correct measurement of the distance between $\theta_1$ and $\theta$ is $|\theta_1 - \theta|/\sigma(\theta)$, rather than $|\theta_1 - \theta|$. Similarly, in the multi-parameter case, $m \geq 1$, the correct measure of the squared distance between $\theta_1$ and $\theta$ is

$$(\theta_1 - \theta)^{\top} I(\theta)(\theta_1 - \theta). \tag{3}$$

Equation (3) suggests that $I(\theta)$ can be used to define a metric on the parameter space, but does $I(\theta)$ define a Riemannian metric? A key property of a Riemannian metric is that the distance between two points remains unchanged when the parameterisation is changed. To make an analogy, one can use the Mercator projection or the Albers projection for a map of the UK, but the distance between London and Birmingham is 122 miles in both cases. In order to investigate the effects of changes in the parameterisation, it is necessary to make a distinction between points in the parameter space and the parameter vectors used to describe them. Let $a_1$, $a_2$ be two nearby points in the parameter space with corresponding parameter values $\theta_1, \theta_2$. The squared distance $d(a_1, a_2)^2$ between $a_1$ and $a_2$ is given to leading order by

$$d(a_1, a_2)^2 = (\theta_2 - \theta_1)^{\top} I(\theta_1)(\theta_2 - \theta_1). \tag{4}$$

Let $\psi$ be a new choice of parameterisation. The information matrix, $I(\psi)$, can be calculated using (2) with $\theta$ replaced by $\psi$. Let $\psi_1, \psi_2$ be the new parameter values for $a_1$ and $a_2$. The squared distance between $a_1$ and $a_2$ is given to leading order by

$$(\psi_2 - \psi_1)^{\top} I(\psi_1)(\psi_2 - \psi_1). \tag{5}$$

If $I(\theta)$ is to define a Riemannian metric then (5) must be equal to $d(a_1, a_2)^2$, at least to leading order, and for this to happen, it is necessary that

$$I(\theta) = \left(\frac{\partial \psi}{\partial \theta}\right) I(\psi) \left(\frac{\partial \psi}{\partial \theta}\right)^{\top} \equiv \sum_{u,\upsilon=1}^{m} \frac{\partial \psi_u}{\partial \theta_i} I(\psi)_{u\upsilon} \frac{\partial \psi_\upsilon}{\partial \theta_j} \tag{6}$$

A short calculation shows that (6) follows from (2). The fact that $I(\theta)$ defines a Riemannian metric on the parameter space was first noted by Rao in [8].

## New characterisations

The Fisher-Rao metric originated in the maximum likelihood approach to parameter estimation, but is it more fundamental than this origin suggests? Can the Fisher-Rao metric be derived from the axioms of probability theory or from the basic requirements for logical reasoning? How many "reasonable" measures of distance can there be on a parameter space for pdfs?

Let $D(a_1, a_2)$ be a measure of the distance between points $a_1, a_2$ corresponding to pdfs $f(x|\theta_1)$, $f(x|\theta_2)$, respectively. Two fundamental requirements for $D$ are firstly that it should be invariant under reparameterisations of the parameter space and secondly that it should be invariant under reparameterisations of the data space containing $x$. The reason for these requirements is straightforward: the parameterisations can be varied at will by the statistician. If $D(a_1, a_2)$ were not invariant, then it could be small in one context, and large in another context, even though the data and the family of pdfs are the same in both contexts. Arguments based on the value of $D(a_1, a_2)$ would then not be meaningful from the statistical point of view. Shun-ichi Amari showed that for reasonable choices of $D$, these invariance requirements ensure that

$$D(a_1, a_2) = c(\theta_2 - \theta_1)^{\top} I(\theta_1)(\theta_2 - \theta_1), \text{ to leading order.}$$

where $c$ is a constant [1]. Thus locally, $D$ reduces to a scaled version of the Fisher-Rao metric.

If the samples $x_1, \ldots, x_n$ are drawn from a finite discrete space $X_m = \{0, \ldots, m\}$, then a stronger result is possible. The parameter space for the pdfs on $X_m$ is the $m$-simplex $S_m$ of points $(p_0, \ldots, p_m)$ in $\mathbb{R}^{m+1}$ which satisfy $p_i \geq 0, 1 \leq i \leq m$, and $\sum_{i=0}^{m} p_i = 1$. A linear map $T: S_m \to \mathbb{R}^{k+1}$ is said to be a stochastic map if $T(S_m) \subseteq S_k$. A natural requirement for a stochastic map is that it should not increase the distance between any two points in $S_n$, $d(Ta_1, Ta_2) \leq d(a_1, a_2)$. Under this requirement, the effect of the map is to remove information from $a_1$, $a_2$, or at best to leave the information in $a_1$, $a_2$ unchanged. Nicholai Chentsov showed that the only metric on each $S_n$ compatible with all the stochastic maps is a scaled version of the Fisher-Rao metric [2].

## Applications

Rao introduced the Fisher-Rao metric in order to cluster points in the parameter space. Points close together under the metric can be clustered together on the grounds that the corresponding pdfs are similar. If the $f(x|\theta)$ are Gaussian pdfs with a common covariance but differing expected values, then (4) coincides with the widely used Mahanalobis distance.

In contrast with Fisher and Rao, Harold Jeffreys took a Bayesian view of statistics, and saw in the Fisher-Rao metric a way of defining a prior pdf on the parameter space [6]. The prior pdf, $p(\theta)$, is defined by

$$p(\theta)d\theta = \frac{|\det(I(\theta))|^{1/2}}{\mathrm{Vol}(S)} d\theta, \tag{7}$$

where the volume, Vol($S$), of the parameter space $S$ under the Fisher-Rao metric is

$$\mathrm{Vol}(S) = \int_S |\det(I(\theta))|^{1/2} d\theta.$$

The scale invariant prior, $\sigma^{-1}d\sigma$, for the standard deviation is a special case of (7). Once a suitable prior pdf is defined, Bayes rule can be applied to obtain the pdf for $\theta$, conditional on the measurements. A key advantage of Jeffreys' prior is that the probability assigned by $p(\theta)$ to subsets of $S$ is independent of the choice of parameterisation of $S$.

A simple algorithm for parameter estimation can be defined by sampling $S$ at a finite number of points $\theta_1, \ldots, \theta_N$ and testing each $\theta$ to see if it is compatible with the data. The volume, Vol($S$), is a measure of the complexity of the algorithm. If Vol($S$) is small, then $N$ is small, and parameter estimation is straightforward. The algorithm is particularly effective if the data are a mixture, in which some points are drawn from a single distribution $f(x|\theta)$, and others are drawn from unrelated distributions.

Recent applications of Jeffreys' prior include machine learning [5] and the detection of structures in digital images [7]. In the minimum description length approach to machine learning, the data are modelled by a pdf $f(x|\theta)$ and then compressed. The best choice of $\theta$ is the one for which the length in bits of the compressed data is a minimum. In order to measure fairly the length of the compressed data, it is necessary to include the length of the description of $\theta$. This length depends on Vol($S$) and on the

resolution chosen in $S$. If there is a large amount of data, then small differences in $\theta$ cause significant differences in the length of the compressed data. It follows that a high resolution is required, leading to an increase in the number of bits required to describe $\theta$.

An application of Jeffreys' prior to line detection in an image is described [7]. The image is assumed to be a disc, scaled to have unit radius. Cartesian coordinates are chosen with an origin at the centre of the disc. The parameter vector for a line $l$ is $(\rho(l), \phi(l))$ where $(\rho(l)\cos(\phi(l)), \rho(l)\sin(\phi(l)))$are the Cartesian coordinates of the point on $l$ closest to the origin. Each line, $l$, has an associated pdf, $f(x|\rho(l), \phi(l))$, for the measured point $x$ in the open unit disk. The parameter space for the pdfs $f(x|\rho(l), \phi(l))$ is $S = (0,1)\times[0, 2\pi)$. The parameterisation $(\rho(l), \phi(l))$ fails for lines through the origin, therefore these lines are omitted from $S$. The Fisher-Rao metric on $S$ is closely approximated by

$$\frac{1}{\sigma^2}\begin{pmatrix}1 & 0\\ 0 & (1-\rho^2)/3\end{pmatrix}, \tag{8}$$

where $\sigma$ is the standard deviation of the error in measuring a coordinate of a point in the unit disk.

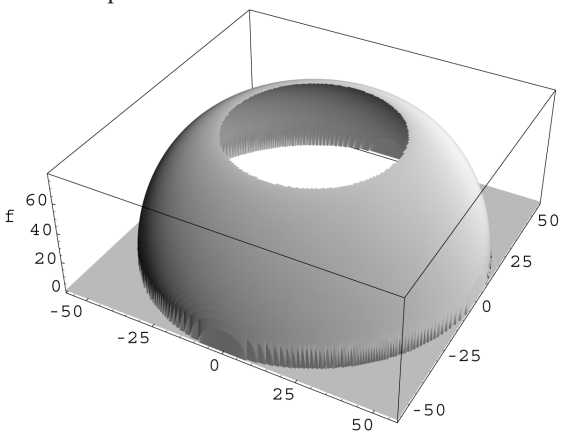


Figure 1: Parameter space for lines embedded as a surface $F$ in $\mathbb{R}^3$. The flat shaded area is not part of $F$

The set $S = (0,1)\times[0, 2\pi)$ is a part of the Euclidean plane, but the Euclidean metric does not coincide with the metric defined on $S$ by (8). However, $S$ can be embedded as a surface $F$ in $\mathbb{R}^3$ such that the metric defined on $F$ by (8) coincides with the metric defined on $F$ by the Euclidean metric in $\mathbb{R}^3$. The surface $F$ is shown in Figure 1 for $\sigma = 10^{-2}$. It is a surface of revolution because (8) is independent of the angular coordinate $\phi$. Lines passing near to the origin are represented by points towards the top of $F$ and lines far from the origin are represented by points towards the bottom of $F$.

In the line detection algorithm, $S$ is sampled at points $(\rho(l_i), \phi(l_i)), 1 \le i \le N$. A line $l_i$ is detected if a significant number of the measurements are within a distance $O(\sigma)$ of $l_i$. The measurements are points $x(i), 1 \le i \le n$, in the image at which there is a large change in the pixel values over a short distance in the image.

## Conclusion

The invariance of the Fisher-Rao metric under reparameterisations of the data space and the parameter space ensures its importance in the theory and the applications of statistics. It has a central role to play whenever it is necessary to estimate parameters using measurements subject to unknown errors. More ambitiously, Roy Frieden argues in [4] that the fundamental equations of physics can be derived from a variational principle based on the Fisher information.❑


REFERENCES

1 Amari, S.-I. (1984) *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, vol. 28. Springer-Verlag.

2 Chentsov, N. N. (1972) *Statistical Decision Rules and Optimal Inference* (in Russian), Nauka, Moscow. English translation (1982), AMS.

3 Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London* Series A, **222**, 309-368.

4 Frieden, B.R. (2001) Physics from Fisher information. *Mathematics Today*, **37**, 115-119.

5 P.D. Grünwald (2007 *The Minimum Description Length Principle*, MIT Press.

6 Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, Series A, **186**, No. 1007, pp. 453-461

7 Maybank, S.J. (2004) Detection of image structures using Fisher information and the Rao metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 1579-1589.

8 Rao, C.R. (1945) Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, **37**, 81-89.



Email: sjmaybank@dcs.bbk.ac.uk