The Fisher Information in Nonlinear Experimental Design

Emanuel Winterfors

Université Pierre et Marie Curie
Laboratoire Jacques-Louis Lions
75252 Paris Cedex 05, FRANCE

winterfors@ann.jussieu.fr

Introduction

The Fisher Information has been used extensively throughout the development of the theory of optimal experimental design, for linear normal models. The theory has also been extended to nonlinear models (Bayesian Experimental Design) through local linearization of the model-data parameter relationship.

There are however problems with using the Fisher information in this way. First, the Fisher Information is (unlike e.g. the Kullback-Leibler distance) dependent on choice of coordinates, as well as the choice of measure with respect to which the probability densities are defined, so experiment optimality is not universal but dependent on parameterization. Secondly, Bayesian Experiment Design theory is based on the (often unstated) assumption that all posterior distributions are unimodal or approximately normal. This is generally not the case for nonlinear models and it is all but trivial to know if this assumption is fulfilled.

The following analysis present a coordinate-independent measure for experiment quality that do not assume anything but differentiability about posterior distributions

The Fisher Information with respect to Translation parameters (FIT)

Given a space Θ with elements $\theta \in \Theta$, equipped with a measure $d\Theta$ and a probability density $p(\theta)$ so that $\int_{\theta\in\Theta} p(\theta)d\Theta = 1$, Cover et al. defined the Fisher Information with respect to a Translation parameter

$$J_T\left[\Theta\right] = \int_{\theta\in\Theta} p(\theta)\left(\frac{\partial}{\partial\theta}\log\left(p(\theta)\right)\right)^2 d\Theta \quad , \tag{1}$$

since it is equal to the classical Fisher Information of a conditional probability density $p(\theta\,|\,y)$

$$J\left[\Theta\,|\,y\right] = \int_{\theta\in\Theta} p(\theta\,|\,y)\left(\frac{\partial}{\partial y}\log\left(p(\theta\,|\,y)\right)\right)^2 d\Theta \tag{2}$$

when the conditional parameter $y=\theta$, corresponding to a translation of the probability density $p(\theta\,|\,y)$.

$J_T\left[\Theta\right]$ as defined in equation (1) is however dependent on the choice of coordinates θ of Θ. In order to get around that problem, one can introduce a coordinate independent metric $\mathbf{g}(\theta)$ on Θ, represented by a metric tensor with components $g_{ij}(\theta)$, with an inverse $\mathbf{g}^{-1}(\theta)$ with components $g^{ij}(\theta)$. This permits a coordinate-independent, scalar definition of FIT

$$\boxed{J\left[\Theta\right] = \int_{\theta\in\Theta} p(\theta)\left(\frac{\partial}{\partial\theta_i}\log\left(p(\theta)\right)\right)g^{ij}(\theta)\left(\frac{\partial}{\partial\theta_j}\log\left(p(\theta)\right)\right)d\Theta \quad , \qquad (3)}$$

where summation convention is used over coordinate indices *i* and *j*.

Interpretation of the FIT

The FIT as defined in equation (3) may be thought of as a squared surface-to-volume ratio of the probability distribution represented by $p(\theta)$. A probability distribution that is confined to a small volume will have a smaller value of the FIT than one which is spread over a larger volume. However, a probability density which is stretched out into a thin "sheet" or divided into numerous small "dots" will also have a high FIT (see figure 1 below).

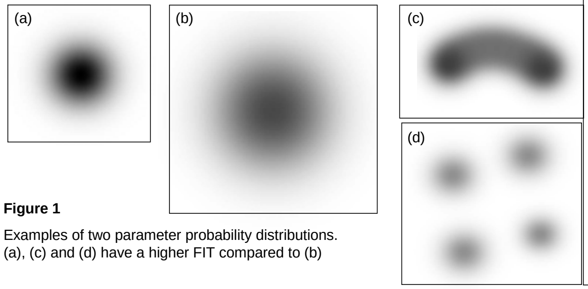

Figure 1

Examples of two parameter probability distributions.
(a), (c) and (d) have a higher FIT compared to (b)

The concept of Fisher Information with respect to a Translation parameter (referred to as FIT) is investigated and refined in order to obtain independence of choice of coordinates. Novel results are presented relating the expected gain in FIT (prior to post experiment) to the Bayesian classical Fisher Information, equivalent to the average sensitivity of the experiment.

Equality of expected gain in FIT and average sensitivity

The gain in FIT of an experiment after an observation y is $J\left[\Theta\,|\,y\right]-J\left[\Theta\right]$. When designing an experiment, one may be interested in maximizing the expected gain in FIT. The expected FIT of a conditional probability distribution $p(\theta\,|\,y)$ and a marginal probability density $p(y)$ can be defined

$$J\left[\Theta\,|\,\Omega\right] = \int_{y\in\Omega} p(y)J\left[\Theta\,|\,y\right]d\Omega \quad . \tag{4}$$

This allows for defining the expected gain in FIT as $J\left[\Theta\,|\,\Omega\right]-J\left[\Theta\right]$. In order to easily compute this expected gain, it is necessary to first define a coordinate independent version of the classical Fisher information defined in equation (2)

$$J_\theta\left[\Omega\,|\,\theta\right] = g^{ij}(\theta)\int_{y\in\Omega} p(y\,|\,\theta)\left(\frac{\partial}{\partial\theta_i}\log\left(p(\theta\,|\,y)\right)\right)\left(\frac{\partial}{\partial\theta_j}\log\left(p(\theta\,|\,y)\right)\right)d\Omega \quad . \tag{5}$$

$J_\theta\left[\Omega\,|\,\theta\right]$ can be interpreted as a measure of the sensitivity of the observed parameters $y\in\Omega$ to changes in the model parameters θ. The average (or, expected) $J_\theta\left[\Omega\,|\,\Theta\right]$ is defined as in equation (4).

$$\boxed{J\left[\Theta\,|\,\Omega\right]-J\left[\Theta\right] = J_\theta\left[\Omega\,|\,\Theta\right] \quad . \qquad (6)}$$

Application to nonlinear experiment design

In a typical experiment design situation, one wishes to determine the value of some model parameter $\theta\in\Theta$ by observing some data variables $y\in\Omega$ where the probability of making a particular observation $p(y\,|\,\theta,\xi)$ given θ and experiment design ξ is known, along with a prior probability distribution $p(\theta)$. Having made an observation y, the posterior probability distribution can be calculated using Bayes' formula $p(\theta\,|\,y,\xi) = p(y\,|\,\theta,\xi)p(\theta)/p(y\,|\,\xi)$.

To estimate the expected information gain $J\left[\Theta\,|\,\Omega\right]-J\left[\Theta\right]$, one normally needs to calculate $J\left[\Theta\,|\,y\right]$ for all possible posteriors given all observations $y\in\Omega$. This is generally computationally challenging, even using efficient MCMC algorithms.

The implication of equation (6) is that it is not necessary to compute any posterior distributions, one can simply compute $J_\theta\left[\Omega\,|\,\theta\right]$ for all (or a MC sample of) points in model space $\theta\in\Theta$. This is largely feasible for a wide range of problems, even in rather high dimension.

Worth noting is that equation (6) is applicable to all differentiable models, without making any assumptions or approximations on the shape of the posterior distributions. The computed expected information gains are also independent on choice of coordinates θ and y on Θ and Ω.

Comparison to the Shannon Information

The Shannon Information is the most commonly used quality criterion in experiment design. It is usually defined

$$I\left[\Theta\right] = \int_{\theta\in\Theta} p(\theta)\log\left(p(\theta)\right)d\Theta \quad . \tag{7}$$

Defining the expected Shannon Information of a conditional probability density $I\left[\Theta\,|\,\Omega\right]$ as in equation (4), one can derive a theorem similar to equation (6) for calculating the expected gain in Shannon Information (see Lindley, 1956)

$$I\left[\Theta\,|\,\Omega\right]-I\left[\Theta\right] = I\left[\Omega\,|\,\Theta\right]-I\left[\Omega\right] \quad . \tag{8}$$

This also allows for the calculation of expected information gains without the need of computing all possible posterior distributions, but the term $-I\left[\Omega\right]$ (equal to the entropy of the Shannon Entropy of the marginal distribution in data space Ω) is not trivial to estimate. Numerical MC methods have been used (see Ryan, 2003 or van den Berg, 2003), but it remains much more challenging than computing expected gains in FIT using equation (6).

The exponential of the Shannon Entropy $\exp\left(-I\left[\Theta\right]\right)$ may be interpreted as a measure of the volume covered by a probability distribution. This implies that stretching out a probability distribution or dividing it up in several peaks as in figures 1 (a, c and d) will not affect the value of Shannon Information, as long as the volume covered is the same. An example of this can be seen in the example to the right.

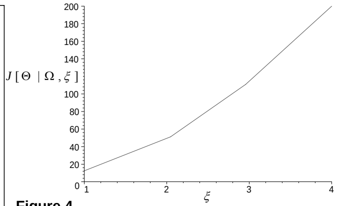

Figure 4

Expected gain in FIT as function of design parameter

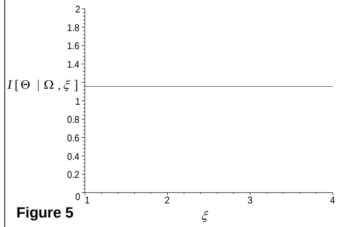

Figure 5

Expected gain in Shannon Information as function of design parameter

Example

A single model parameter $\theta\in\Theta \;=[0..\pi]$ with a homogeneous prior probability density $p(\theta)=1/\pi$ is to be estimated by observations of y whose relation to θ is described by the conditional probability density

$$p(y\,|\,\theta,\xi) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(y-\sin(\xi\theta))^2}{2\sigma^2}\right) \quad , \tag{2}$$

which describe a sine curve with Gaussian measurement noise (see figures 2 and 3). The experiment design parameter ξ is a positive integer determining the frequency of the curve.

Figures 4 and 5 shows the dependence of expected gains in Shannon Information and FIT, respectively.

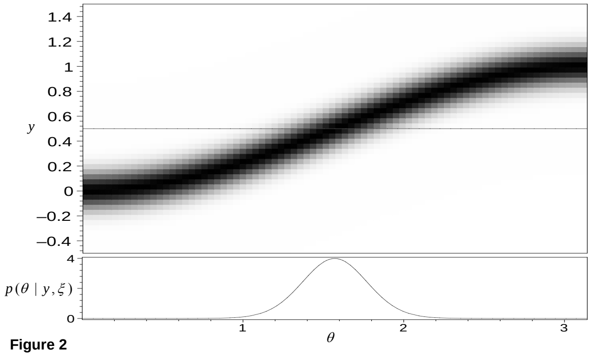

Figure 2

Joint probability density of experiment in example, for design parameter ξ=1. Posterior probability density shown for an observation *y*=0.5.

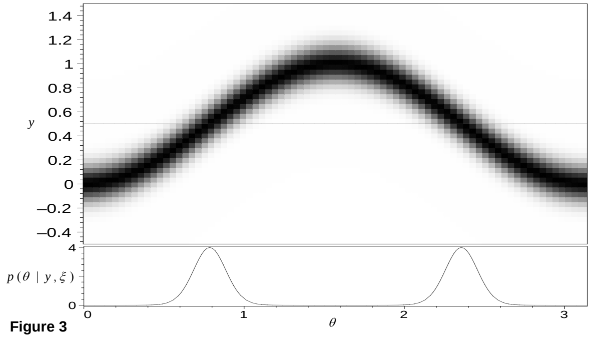

Figure 3

Joint probability density of experiment in example, for design parameter ξ=2. Posterior probability density shown for an observation *y*=0.5.

References

A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, 1991.

D. V. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005, 1956.

J. A. van den Berg, A Curtis, and J. Trampert. Bayesian, nonlinear experimental design applied to simple, geophysical examples. *Geophys. J. Int.*, 55(2):411–421, 2003.