Statistics and Probability Letters 112 (2016) 5–9

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bayesian information in an experiment and the Fisher information distance

Stephen G. Walker

Department of Mathematics, University of Texas at Austin, TX, USA

ARTICLE INFO

Article history:
Received 19 June 2015
Received in revised form 18 January 2016
Accepted 18 January 2016
Available online 3 February 2016

Keywords:
Shannon entropy
Information
Bayesian design

ABSTRACT

There are two forms of Fisher information; for the parameter of a model and for the information in a density model. These two forms are shown to be fundamentally connected through a measure of gain in information from a Bayesian experiment.

1. Introduction

For the statistical model $f(x|\theta)$, with $x \in \mathbb{R}$ and $\theta \in \Theta \subset \mathbb{R}$, we assume that $f(x|\theta)$ is differentiable with respect to both x and θ. To start we have θ as a 1 dimensional parameter, which is extended to higher dimensions later in the paper. The amount of information about θ in a single observation experiment is given by the Fisher information;

$$I(\theta) = \int \left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 f(x|\theta)\, \mathrm{d}x. \tag{1}$$

It has a number of applications, most notably in large sample asymptotic studies; see, for example, Lehmann and Casella (1998).

On the other hand, a well known measure of the information carried by a density function is the Shannon entropy, which in the continuous case is often referred to as the differential Shannon entropy. This is defined as

$$H(Y) = -\int f_Y(y)\, \log f_Y(y)\, \mathrm{d}y, \tag{2}$$

where the random variable Y has density function f_Y. See Shannon (1948). This was originally defined in the discrete case; i.e. Y is a discrete random variable. Some of the key properties, such as positivity, are lost in the differential case.

Nevertheless, the Shannon entropy was adopted as a measure of the information in a prior and posterior distribution; see Lindley (1956). To elaborate, suppose $p(\theta)$ denotes a prior distribution for model $f(x|\theta)$ and $p(\theta|x)$ is the posterior distribution. The idea is that the gain in information by observing outcome X is given by

$$\Delta(X) = H(\Theta) - H(\Theta|X) \tag{3}$$

E-mail address: S.G.Walker@math.utexas.edu.

http://dx.doi.org/10.1016/j.spl.2016.01.014

where

$$H(\Theta) = -\int p(\theta)\,\log p(\theta)\,\mathrm{d}\theta \quad \text{and} \quad H(\Theta|X=x) = -\int p(\theta|x)\,\log p(\theta|x)\,\mathrm{d}\theta.$$

The expected gain in information by the experiment is therefore given by $\Delta = \int \Delta(x)\,m(x)\,\mathrm{d}x$ where $m(x) = \int f(x|\theta)\,p(\theta)\,\mathrm{d}\theta$. One can easily find simplified forms for Δ, which can also be represented as

$$\Delta = \int D\big(p(\cdot|x), p(\cdot)\big)\,m(x)\,\mathrm{d}x$$

where D denotes the Kullback–Leibler divergence;

$$D(p_1, p_2) = \int p_1(x)\,\log\frac{p_1(x)}{p_2(x)}\,\mathrm{d}x.$$

Applications in the Bayesian design of experiments can be found in Sebastiani and Wynn (1997).

If $I(\theta)$ represents the information about θ in a single observation experiment and $p(\theta)$ is the prior, then we ask why the expected information in the experiment is not simply

$$\widetilde{\Delta} = \int I(\theta)\,p(\theta)\,\mathrm{d}\theta, \tag{4}$$

rather than Δ. We will discuss this, showing that (4) arises by using the Fisher information distance, $F(p_1, p_2)$ (to be defined later), i.e.

$$\widetilde{\Delta} = \int F(p(\cdot|x), p(\cdot))\,m(x)\,\mathrm{d}x,$$

rather than the Kullback–Leibler divergence, in Section 3. However, for now, we note that $\int I(\theta)\,p(\theta)\,\mathrm{d}\theta$ does appear in a Bayesian type Cramér–Rao bound; see Gill and Levit (1995).

In Section 2 we show that an alternative recognized measure of information of a density, $J(Y)$, also known as the Fisher information, is such that if the gain in information given observation X is

$$\Delta^*(X) = J(\Theta|X) - J(\Theta),$$

then the expected gain,

$$\Delta^* = \int \Delta^*(x)\,m(x)\,\mathrm{d}x,$$

coincides with $\widetilde{\Delta}$.

2. Fisher information for a density function

Widely used in the functional analysis community, the Fisher information for a density f_Y is given by

$$J(Y) = \int \frac{f_Y'(y)^2}{f_Y(y)}\,\mathrm{d}y. \tag{5}$$

See, for example, Bobkov et al. (2014). This can also be represented as

$$J(Y) = \int f_Y(y)\left(\frac{\partial}{\partial y}\log f_Y(y)\right)^2\mathrm{d}y$$

and of course $J(Y)$ will coincide with (1) when $f_Y(y) = f(y-\theta)$; i.e. a location model.

Now (5) is not so well known in the statistics literature; though there are some references. For example, Brown (1982) used it to prove the central limit theorem via a Cramér–Rao inequality, and elaborating on this paper, Johnson and Barron (2004), also use (5), as well as the Fisher information distance, to be defined later in this paper. As well, Papaioannou and Ferentinos (2005) discuss properties of (5) as a direct comparison with the properties of (1).

Theorem 1. *It is that* $\Delta^* \equiv \widetilde{\Delta}$.

Proof. Now

$$J(\Theta|X=x) = \int \frac{\left(\frac{\partial}{\partial\theta}p(\theta|x)\right)^2}{p(\theta|x)}\,\mathrm{d}\theta$$