Information and Entropy of Continuous Random Variables

Fabián, Zdeněk
1996

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Information and Entropy of Continuous Random Variables

Zdeněk Fabián

Technical report No. V–694

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+422) 6605 3210 fax: (+422) 8585789
e-mail: zdenek@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Information and Entropy of Continuous Random Variables

Zdeněk Fabián

Technical report No. V–694

Abstract

Mean value of the square of a generalized score function is shown to be interpretable as an information associated with a continuous random variable. This information is in particular cases equal to the Fisher information of the corresponding distribution.

Keywords

Fisher information, entropy, information function

I. INTRODUCTION

It is still an open question what a quantity could be taken as a measure of the average amount of information associated with a continuous random variable X with density p. It is well known that it cannot be the natural candidate, the Shannon's differential entropy

$$h_S(X) = E_p(-\log p) = \int -\log p(x)\ p(x)\ dx,$$

since it can be negative. We show that such a quantity could be a suitably modified Fisher information.

Let $T \subset R$, where R denotes the real line, be an open interval with the σ-field $\mathcal{B}_T$ of its Borel subsets and let $\Theta \subset R^m$ be an open set. Consider the usual parametric model

$$\mathcal{P}_T = \{T, \mathcal{B}_T, p(u|\theta) : u \in T, \theta \in \Theta\}$$

with densities regular in the Cramer-Rao sense. A simple particular case is the location model $\{R, \mathcal{B}_R, p(x-\mu), x, \mu \in R\}$, where the location parameter μ represents a shift along the x-axis.

The Fisher information is usually defined with respect to parameters of $\mathcal{P}_T$. Recall that the Fisher information matrix $(g_{jk}(\theta))_1^m$ is given by

$$g_{jk}(\theta) = E_p s_j s_k, \qquad j,k = 1, ..., m,$$

where

$$s_j(u|\theta) = \frac{\partial \log p(u|\theta)}{\partial \theta_j} \tag{0.1}$$

is the *likelihood score* for the parameter θ_j.

The concept of the Fisher information of a distribution is much less frequent. It is defined (e.g., [2], pp.494) as

$$I(X) = E_p s^2 = \int_R s^2(x) p(x)\ dx, \tag{0.2}$$

where s is the *score function* of the distribution p, given by

$$s(x) = -\frac{p'(x)}{p(x)}. \tag{0.3}$$

It is easy to see that in the location model we have

$$I(X) = g_{11}(\mu)|_{\mu=0}. \tag{0.4}$$

Consider the function $s^2 : R \to [0, \infty)$. In the case of an unimodal distribution on $(R, \mathcal{B}_R)$, it attains its minimum value at the least informative point $x = 0$ of the distribution. By (0.4), its mean value $E_p s^2$ has the meaning of an information. It seems that the value $s^2(x)$ could represent the relative information contained in $x \in R$ (relative to other points $x' \in R$), under the assumption that the true density is p.

However, in cases of distributions whose parametric space does not contain the location parameter, (0.2) is different from any of the diagonal terms of the Fisher matrix. There is no reason in such cases to interpret $E_p s^2$ as an information. It concerns all random variables taking values in $T \neq R$.

A suitable generalization of the score function for cases $T \neq R$ has been given in [3]. We show that the mean value of the square of this generalized score function is proportional to a certain diagonal term of the Fisher matrix, so that it can be taken as an information contained in a continuous random variable in a general regular case.

II. GENERALIZED SCORE FUNCTION

Let Π_T be the set of all absolutely continuous distributions on $(T, \mathcal{B}_T)$ with densities continuously differentiable a.e.

Definition 1: Let U_T be a random variable with density $p \in \Pi_T$. Let a random variable U_R be given by the relation $U_R = \varphi^{-1}(U_T)$ where $\varphi : R \mapsto T$ is sufficiently smooth and strictly increasing. A real-valued function $q : T \mapsto R$, given by

$$q(u) = \frac{1}{p(u)} \frac{d}{du}(-L(u)p(u)), \tag{0.5}$$

where

$$L^{-1}(u) = \frac{d\varphi^{-1}(u)}{du}, \tag{0.6}$$

will be called a generalized score function (GSF) of U_T.

Random variable U_R and its distribution will be called *original*, and $U_T = \varphi(U_R)$ and its distribution $\varphi-$*related*. The concept of the GSF obviously depends on the choice of the mapping φ. We selected it in the simplest possible way for the three principally different intervals:

(i) In the case of $T = R$ we set $U_R = X$, and, naturally, $\varphi(X) = X$. Then $L(x) = 1$ and the GSF of random variable X is the usual score function (0.3).

(ii) In the case of $T = (0, \infty)$ we set $U_T = Z$ and $Z = \varphi(X) = e^X$. Then $L(z) = z$ and $X = \varphi^{-1}(Z) = \ln Z$. This choice is certainly in the spirit of statistics. Positive data are often logarithmically transformed, and some pairs of distributions in current use defined on Borel subsets of $T = (0, \infty)$ and R are often considered to be logarithmically related. The GSF of Z is given by the explicit formula

$$q(z) = -1 - z\frac{p'(z)}{p(z)} = -1 - zs(z). \tag{0.7}$$

(iii) Let $a, b \in R$. In the case of $T = (a, b)$ we set $U_T = W$. It might seem that there are many possible transformations $\varphi_{ab} : R \to (a, b)$ and that the concept of GSF on a finite interval (a, b) (or $[a, b]$) should be ambiguous. In fact, this is not the case. To be consistent with (ii), we require

$$\lim_{\substack{a \to 0 \\ b \to \infty}} \varphi_{ab}^{-1}(w) = \ln w. \tag{0.8}$$

A general mapping satisfying (0.8) is

$$\varphi_{ab}^{-1}(w) = \ln \frac{(b-c)(w-a)}{(c-a)(b-w)} \tag{0.9}$$

where $a < c < b$ and, incidentally, $c = c(a,b)$ with $\lim\limits_{\substack{a\to 0\\ b\to\infty}} c(a,b) = 1$. The Jacobian of the transformation (0.9) and the corresponding GSF do not depend on c. Indeed,

$$L^{-1}(w) = \frac{d}{dw}(\varphi_{ab}^{-1}(w)) = \frac{b-a}{(w-a)(b-w)}.$$

Thus, the general transformation (0.9) provides a unique GSF on (a,b) in the form

$$q(w) = (b-a)^{-1}[-(b+a) + 2w - (w-a)(b-w)p'(w)/p(w)], \tag{0.10}$$

which reduces on $T = (0,1)$ into

$$q(w) = -1 + 2w - w(1-w)s(w).$$

We mention some properties of the GSF.

Proposition 1: Let U_R and $U_T = \varphi(U_R)$ be random variables with densities p_R, p and GSFs q_R, q, respectively. Then,

$$p(u) = p_R(\varphi^{-1}(u))L^{-1}(u) \tag{0.11}$$

$$q(u) = q_R(\varphi^{-1}(u)). \tag{0.12}$$

Proof. The relation between distribution functions F and F_R of U_T and U_R, respectively, is $F(u) = F_R(\varphi^{-1}(u))$, so that

$$p(u) = dF(u)/du = dF_R(x)/dx \cdot dx/du = p_R(x) \cdot d\varphi^{-1}(u)/du.$$

By (0.5),

$$q(u) = \frac{L(u)}{p_R(x)}\frac{d}{du}(-p_R(x)) = -\frac{L(u)}{p_R(x)}\frac{dp_R(x)}{dx}L^{-1}(u) = q_R(x).$$

■

The GSF of U_T is thus the transformed score function of the original random variable U_R. Further, we have $E_p q = 0$. Indeed, letting $c_1 = \inf\{u : u \in T\}$, $c_2 = \sup\{u : u \in T\}$ and using (0.5) and (0.11),

$$E_p q = \int_{c_1}^{c_2} q(u)p(u)\,du = -L(u)p(u)|_{c_1}^{c_2} = p_R(x)|_{-\infty}^{\infty} = 0.$$

The GSF can be bounded or semi-bounded. It follows immediately from (0.12) that if the GSF of an original distribution is unbounded, bounded, or semibounded, the GSF of a φ-related distribution must be unbounded, bounded, or semibounded, respectively. Now consider distributions on $(R, \mathcal{B}_R)$. If $p_R \sim e^{-x^2}$, the corresponding

$q_R = O(x)$; whereas if $p_R \sim e^{-|x|}$, then $q_R = O(1)$. Thus, a bounded GSF indicates a slow decay of the corresponding density to zero, which characterizes heavy-tailed distributions.

The concept of GSF can easily be generalized for the case of a parametric family $\mathcal{P}_T$.

Definition 2: Let U_R be a random variable with density $p_R(x|\gamma) \in \Pi_R$. Let $p(u|\theta)$ be the density of random variable $U_T = \varphi(U_R)$. The GSF of U_T is defined by

$$q(u|\theta) = \frac{1}{p(u|\theta)} \frac{d}{du}(-L(u)p(u|\theta)), \tag{0.13}$$

where $L(u)$ is given by (0.6).

Let $\Theta_R = R \times \Theta_{m-1}$ where $\Theta_{m-1} \subset R^{m-1}$ is an open convex set. Consider the density of U_R in the form $p(x - \mu|\alpha)$ where $\gamma_1 = \mu \in R$ is the location parameter and $\alpha = (\gamma_2, ..., \gamma_m) \in \Theta_{m-1}$. We call the parameter $\nu = \varphi(\mu) \in T$ of the density $p(u|\theta) = p(u|\nu, \alpha)$ of random variable $U_T = \varphi(U_R)$ the *transformed location parameter.*

Proposition 2: The GSF of a random variable with distribution $p(u|\nu, \alpha)$, where ν is the transformed location parameter, is given by

$$q(u|\nu, \alpha) = L(\nu)s_1(u|\nu, \alpha),$$

where s_1 is the likelihood score for ν.

Proof: Denote $y = \varphi^{-1}(u) - \varphi^{-1}(\nu)$. As in (0.11), we have $p(u|\theta) = L^{-1}(u)p_R(y|\alpha)$. Using (0.1), (0.6), (0.13) and (0.12)

$$s_1(u|\nu, \alpha) = \frac{1}{p(u|\nu, \alpha)} \frac{dp(u|\nu, \alpha)}{d\nu} = \frac{L(u)}{p_R(y|\alpha)} \frac{d(L^{-1}(u)p_R(y|\alpha))}{dy} \frac{dy}{d\nu}$$

$$= -\frac{p'_R(y|\alpha)}{p_R(y|\alpha)} L^{-1}(\nu) = q_R(y|\alpha)L^{-1}(\nu) \ = L^{-1}(\nu)q(u|\nu, \alpha).$$

■

Thus, if the vector of parameters of a distribution contains the transformed location parameter, the GSF is proportional to the likelihood score for this parameter.

III. INFORMATION OF A CONTINUOUS DISTRIBUTION

Definition 3: Let q be the GSF of a random variable U_T. A function

$$i_q(u) = q^2(u),$$

will be called the information function of U_T.

Proposition 3: Let $T \neq R$. The solution of the equation $i_q(u) = 0$, if unique, appears to be the least informative point of the distribution p.

Proof: By (0.11), density $p(u)$ appears to be a product of two terms. The term $L^{-1}(u)$ is common to all distributions on a given $(T, \mathcal{B}_T)$, so that it does not carry any information about the random variable U_T. All the information contained in U_T is thus condensed into the term $p_R(\varphi^{-1}(u))$. The maximum of p_R exists and defines the least informative point u^* of the distribution p. By (0.11) and (0.5), $(d/du)p_R(\varphi^{-1}(u)) = (d/du)(L(u)p(u)) = -q(u)p(u)$, so that u^* is the solution of the equation $q(u^*) = 0$. ■

Information function of a parametric distribution $p(u|\theta)$ is, obviously, $i_q(u|\theta) = q^2(u|\theta)$.

Definition 4: A value

$$I_q(U_T|\theta) = \int_T q^2(u|\theta)\, p(u|\theta) du$$

will be called the q-information of random variable U_T.

Clearly, $I_q(U_T|\theta)$ is non-negative and finite for all Cramer-Rao regular distributions. In a model with the transformed location parameter ν, we have $I_q(U_T) = \lim_{\nu \to \varphi(0)} I_q(U_T|\nu)$. Moreover, by Proposition 2, $I_q(U_T|\nu) = L^2(\nu) g_{11}(\nu)$, where $g_{11}(\nu)$ is the Fisher information about ν. The mean value of i_q is thus proportional to a quantity which is known to be an information measure.

This conclusion, together with Proposition 3 (which obviously holds true also in the case of a parametric distribution), is the basis of our belief that $i_q(u)$ can be interpreted as a relative information contained in $u \in T$ provided that the true distribution is p, and the q-information as an information of a distribution p.

IV. EXAMPLES

Both the differential entropy and q-information are values (possibly dependent on parameters) characterizing a continuous random variable. We show by means of some examples that the latter has a reasonable meaning.

Example 1: $T = R$. Here GSF is the score function, and $I_q(X)$ equals the Fisher information of distribution (0.2). Densities, information functions and Fisher information of some distributions are given in Table 1.

While $i_S(x) = -\log p(x)$ is unbounded for all distributions, $i_q(x)$ can be unbounded, semi-bounded or bounded depending on the type of the distribution. We judge that the boundedness of the information function of heavy-tailed distributions has a good reason. An occurrence of outlier values in samples from heavy-tailed distributions only slightly influences estimates, contrary to disastrous effects that arise in similar situations in cases of sharply-peaked distributions with unbounded i_q (see e.g. [4]).

The q-information is high for sharply-peaked and low for heavy-tailed distributions. A sample from a sharply-peaked distribution is thus carrying, on average, more q-information about the distribution than a sample from a heavy-tailed distribution.

Perhaps the mean Fisher uncertainty of a continuous random variable (*q-entropy*, say) could be expressed by the reciprocal value of the q-information,

$$h_q(U_T|\theta) = I_q(U_T|\theta)^{-1}. \tag{0.14}$$

Example 2: Related distributions on $T = (0, \infty)$. Consider random variables X_j with distributions given in Example 1. Densities and information functions of φ-related random variables $Z_j = e^{X_j}$ are given in Table 2. The q-information of Z_j and X_j are equal.

Example 3: Non-symmetric exponential family. Let $T = (0, \infty)$. Consider a family of distributions with densities in the form

$$p(z|\nu, \beta, \lambda) = \frac{\beta\lambda^\lambda}{\Gamma(\lambda)z}\left(\frac{z}{\nu}\right)^{\lambda\beta} e^{-\lambda\left(\frac{z}{\nu}\right)^\beta}, \qquad \nu, \beta, \lambda > 0 \tag{0.15}$$

where Γ is the gamma function. Some members of the family are, for instance, the following distributions:

Distribution	ν	β	λ
exponential	ν	1	1
Rayleigh	$\gamma\sqrt{2}$	2	1
Weibull	ν	β	1
Erlang	n/β	1	n
gamma	$c\lambda$	1	λ
chi-squared	n/β^2	1	$n/2$

The differential entropy of the family is given by

$$h_S(\nu, \beta, \lambda) = -\log\beta - \lambda\log\lambda + \log\Gamma(\lambda) + \log\nu + (\lambda - 1/\beta)(\log\lambda - \psi(\lambda)) + \lambda.$$

The formula is very cumbersome, since there are logarithms of norming factors involved in it.

The GSF of the family $q(z|\nu, \beta, \lambda) = \lambda\beta((z/\nu)^\beta - 1)$ is of semi-bounded type. The q-information is given by an extremely simple expression

$$I_q(\nu, \beta, \lambda) = \lambda\beta^2.$$

Notice that I_q is independent of the transformed location parameter ν. The Fisher information about the parameter ν is, by Proposition 2, $g_{11}(\nu) = \lambda\beta^2/\nu^2$. The family original to (0.15) has densities

$$p(x|\mu, \sigma, \lambda) = \frac{\lambda^\lambda}{\sigma\Gamma(\lambda)} e^{\lambda(x-\mu)/\sigma} e^{-\lambda e^{(x-\mu)/\sigma}},$$

where $\mu = \log\nu$ and $\sigma = \beta^{-1}$, and the q-information $I_q(\mu, \sigma, \lambda) = \lambda/\sigma^2$.

Example 4: Uniform distribution, $T = [0, b]$. The differential entropy is $h_S(b) = \log b$. Its value and even the sign depends on b. The GSF of the uniformly distributed

random variable is, by (0.10), $q(w|b) = 2w/b - 1$. Hence, $I_q(b) = b^{-3}\int_0^b (2w-b)^2\, dw = 1/3$ independently of the length of the interval.

Considering a discrete distribution as a sampled version of a continuous one, the GSF of a discrete random variable taking on values in $T = (0, \infty)$ or $T = (a, b)$ can be defined by the means of formulas (0.7) and (0.10), respectively, after replacing the derivatives of the density by differences (e.g., $p'(x_i) \approx p(x_i) - p(x_{i-1})$). In the case of the discrete uniform distribution with density $p(k|n) = 1/(n+1), k = 0, ..., n$, the influence function is, by (0.10), $q(k|n) = 2k/n - 1$, and the q-information $I_q(n) = \sum_{k=0}^{n}(2k/n-1)^2/(n+1) = 2(2n+1)/3n - 1$, with $\lim_{n\to\infty} I_q(n) = 1/3$.

Example 5: Beta distribution. The uniform distribution is the maximum differential entropy distribution on the given interval. If $T = (0, 1)$, it holds $h_S(1) = 0$ and the differential entropy of any other continuous distribution is negative. Let us consider the beta distribution with density

$$p(w|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} w^{\alpha-1}(1-w)^{\beta-1} \qquad w \in [0, 1], \alpha, \beta \geq 0$$

where B is the beta function. By (0.10), the corresponding GSF is $q(w|\alpha, \beta) = (\alpha + \beta)w - \alpha$ and the q-information is given by

$$I_q(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \int_0^1 [(\alpha+\beta)w - \alpha]^2 w^{\alpha-1}(1-w)^{\beta-1}\, dw = \frac{\alpha\beta}{\alpha+\beta+1}.$$

For illustration we give some values of the q-entropy (0.14):

β	4	2	1	0.5	0.25
$h_F(1, \beta)$	1.5	2	3	5	9

It is obvious that the uniform distribution, $p(w|1, 1)$, is not the distribution with the maximum uncertainty, measured by (0.14), on $T = (0, 1)$. We think that this result is well justified. The relation $h_F > 3$ holds in cases of small values of parameters α, β, where the density of the beta distribution is antimodal, exhibiting small relative probabilities in the central area and large ones at the ends of the interval. It can be interpreted as though the result of the observation would be more uncertain, in a Fisher sense, when an occurrence of an event is likely in two, almost separated areas, rather than in the case of equally likely events.

Example 6. Triangular distribution. Consider a triangular distribution with density $p(w|a) = 2w/a$, $0 \leq w \leq a$ and $p(w|a) = 2(1-w)/(1-a)$ when $a < w \leq 1$. The differential entropy is $h_S(X|a) = 1/2 - \log 2$ (e.g., [2]), which is independent of a. The corresponding information function is discontinuous and equals, according to (0.10), $i_q(w) = (-1+2w-w(1-w)/w))^2 = (3w-2)^2$ when $0 \leq w \leq a$ and $i_q(w|a) = (3w-1)^2$ when $a < w \leq 1$. The q-information, given by $I_q(a) = (5a^3/2 - 5a^2 + 2a + 1/2)/(1-a)$, depends on a. It illustrates the fact that the q-entropy, contrary to Shannon's entropy, is capable of encompassing the morphology of distributions.

TABLE 1

INFORMATION FUNCTION AND q-INFORMATION OF SOME DISTRIBUTIONS

	Distribution	$p(x)$	$i_q(x)$	I_q
1		$\frac{1}{2K_0(1)}e^{-\cosh x}$	$\sinh^2 x$	1.429
2	normal	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	x^2	1
3	doubly exponential	$e^x e^{-e^x}$	$(e^x-1)^2$	1
4	logistic	$e^x/(1+e^x)^2$	$\mathrm{tgh}^2(x/2)$	1/3
5	Cauchy	$\pi^{-1}(1+x^2)^{-1}$	$4x^2/(1+x^2)^2$	1/2

(K_0 is the modified Bessel function of the third kind.)

TABLE 2

INFORMATION FUNCTION OF φ-RELATED DISTRIBUTIONS

j	Distribution	$p(z)$	$i_q(z)$
1	Wald type	$\frac{1}{2K_0(1)z}e^{-\frac{1}{2}(z+1/z)}$	$\frac{1}{4}(z-1/z)^2$
2	lognormal	$\frac{1}{\sqrt{2\pi}\,z}e^{-\frac{1}{2}\log^2 z}$	$\log^2 z$
3	exponential	e^{-z}	$(z-1)^2$
4	log-logistic	$1/(z+1)^2$	$((z-1)/(z+1))^2$
5	log-Cauchy	$(\pi z)^{-1}(1+\log^2 z)^{-1}$	$4\log^2 z/(1+\log^2 z)^2$

REFERENCES

[1] R.C.Rao, *Linear Statistical Inference and Its Applications.* New York: Wiley, 1973.

[2] T.M.Cover, J.A.Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

[3] Z.Fabián, "Generalized score function and its use," in *Transactions of 12-th Prague Conference on Information Theory*, pp.67-72, 1994.

[4] F.R. Hampel, P.J. Rousseeuw, E.M. Ronchetti, W.A. Stahel, *Robust Statistic. The Approach Based on Influence Functions,* New York: Wiley, 1987.