Entropy, uncertainty and information of continuous random variables

Zdeněk Fabián
Institute of Computer Science AS CR

Energy Days 2023

Continuous random variables

Random variable X on an open interval $\mathcal{X} \subseteq \mathbb{R}$

Support $\mathcal{X}$, distribution F, density $f(x)$

(Euclidean) moments $EX^k = \int_{\mathcal{X}} x^k f(x)\, dx$

Mean EX, variance $E(X - EX)^2$

Observed data $\mathbb{X}_n = (x_1, ..., x_n)$ ... realizations of $(X_1, ..., X_n) \sim F_\theta \in \{\mathcal{F}_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$

Information in information theory

Information in $(x - dx < X < x + dx)$ is $I_F(x) = -\log f(x)dx$

The mean information/uncertainty of event X is the differential entropy

$$h(X) = EI_F = \int_{\mathcal{X}} I_F(x) f(x)\, dx$$

However, it is not a right analogue of the entropy of discrete random variables (it can be negative if a density has a pronounced central peak)

The differences between differential entropies are taken to indicate differences in uncertainty: relative entropy

$$D_{KL}(F, H) = -\int_{\mathcal{X}} \log \frac{h(x)}{f(x)} f(x)\, dx$$

Information and uncertainty in statistics

Rather unclear:

i/ Fisher information of F_θ

$$FI(\theta) = E_\theta U_\theta^2 = E_\theta \left(\frac{\partial}{\partial\theta} \log f(x;\theta) \right)^2$$

is not related to distributions, but to parameters of parametric families

ii/ As uncertainty of random variables is usually taken the variance. However, it may not exist

Cauchy $\quad EX^2 = \int_{-\infty}^{\infty} \frac{x^2}{(1+x^2)}\, dx$

Inference functions

i/ Probability theory works with ‘pure’ X

ii/ In statistics are often used, instead of $x_1, ..., x_n$ ... samples $(\phi(x_1), ..., \phi(x_n))$, where ϕ is a suitable inference function:

i/ Classical statistics: Fisher scores $U_\theta(x)$. For $\theta \in \Theta^m$ vector-valued functions

ii/ Robust statistics: (Huber’s) scores. They are scalar-valued and bounded (suppressing outliers), but not related with the assumed model F_θ

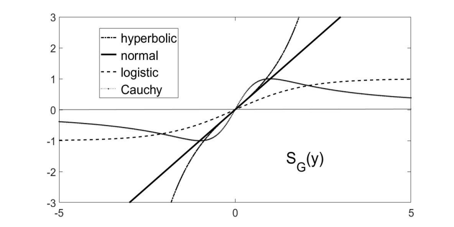
Score function in the narrow sense

Since for G on $\mathbb{R}$

$$\frac{\partial}{\partial\mu}\log g(y-\mu) = -\frac{1}{g(y-\mu)}\frac{d}{dy}g(y-\mu) \equiv S_G(y-\mu)$$

for $\mu = 0$ is **score function** as characteristic of G itself

$$S_G(y) = -\frac{g'(y)}{g(y)}$$



Properties of $S_G(Y)$

Score moments $ES_G^k(Y;\theta)$ are finite

Typical value is the mode $y^* : S_G(y;\theta) = 0$: **score mean**

Measure of variability $Var_S Y = 1/ES_G^2(\theta)$ **score variance**

CLT for iid Y according to arbitrary G

$$\bar{S}_G = \frac{1}{n}\sum_{i=1}^{n}(S_G(Y_i)$$

For $n \to \infty$ $\sqrt{n}\bar{S}_G \xrightarrow{\mathcal{D}} \mathcal{N}(0, ES_G^2)$

Score function of parametric distributions

$$S_G(y;\theta) = -\frac{1}{g(y;\theta)}\frac{d}{dy}g(y;\theta)$$

a scalar-valued function even if θ is a vector

However, they are used neither in probability nor in statistics: it does not work for distributions on $\mathcal{X} \neq \mathcal{R}$:

$$f(x) = 1 \text{ on } (0,1): \qquad -\frac{f'(x)}{f(x)} = 0$$

$$f(x) = e^{-x} \text{ on } (0,\infty): \qquad -\frac{f'(x)}{f(x)} = 1$$

Transformation-based score

Fabián (2001): X with distribution F on $\mathcal{X} \neq \mathcal{R}$ is transformed Y on $\mathbb{R}$ with “prototype” G

$\eta(x) : \mathcal{X} \to \mathbb{R} : \quad X = \eta^{-1}(Y),\, F(x) = G(\eta(x))$

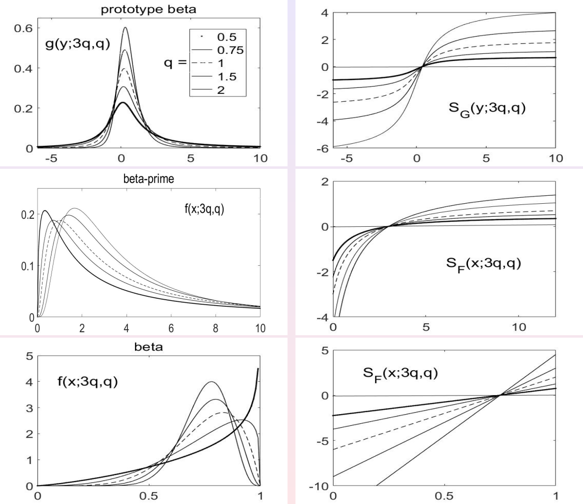
with density

$$f(x) = g(\eta(x))\eta'(x)$$

and the “score function” (**t-score**) on $\mathcal{X}$ is

$$T_F(x) = S_G(\eta(x))$$

Densities and t-scores of the “beta set”



Score mean and score variance

$$T_F(x^*) = S_G(\eta(x^*)) = S_G(y_G^*) = 0$$

so that the **score mean**, typical value od F,

$$x_F^* = \eta^{-1}(y_G^*)$$

is the projection of the mode of the prototype G into $\mathcal{X}$

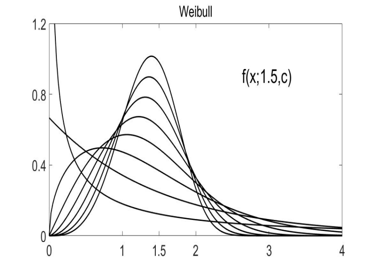
Score variance, a measure of variability of F is

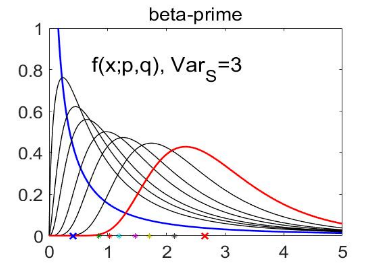
$$Var_S X = \frac{ET_F^2}{T_F'(x^*)]^2}$$

where $T_F'(x) = dT_F(x)/dx|_{x=x^*}$

Densities

of the Weibull with the same score mean and of the beta-prime with the same score variance





Information and uncertainty functions

Only mean information/uncertainty of cont. r.v. is considered

We construct $u_F(x)$ and $i_F(x)$. They should have intuitively the following properties:

i/ X from F with the concentrated mass has high Ei_F and low Eu_F and vice versa.

ii/ The courses of $i_F(x)$ and $u_F(x)$ of a concrete F should be to some extent "parallel": an observation around the score mean is a likely event carrying a little amount of information and a little amount of uncertainty as well. A reasonable properties are $i_F(x^*) = u_F(x^*) = 0$

iii/ An increase of $i_F(x)$ has in effect a decrease of $u_F(x)$

Definition: Standard distributions

Let F be a continuous distribution with t-score $T_F(x)$. Define the **information function** of F as

$$i_F(x) = T_F^2(x)$$

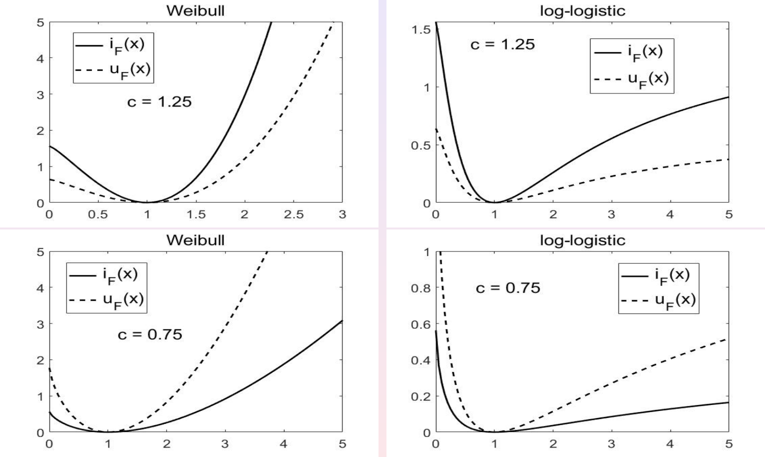
and the **uncertainty function** of F as

$$u_F(x) = \frac{i_F(x)}{[T_F'(x^*)]^2}$$

where $T_F'(x^*) = dT_F(x)/dx|_{x=x^*}$

According to the definition, **mean information of a continuous distribution $Ei_F = ET_F^2$ is the Fisher information for the score mean and mean uncertainty Eu_F is the score variance**

Information and uncertainty functions



Mean information and differential entropy

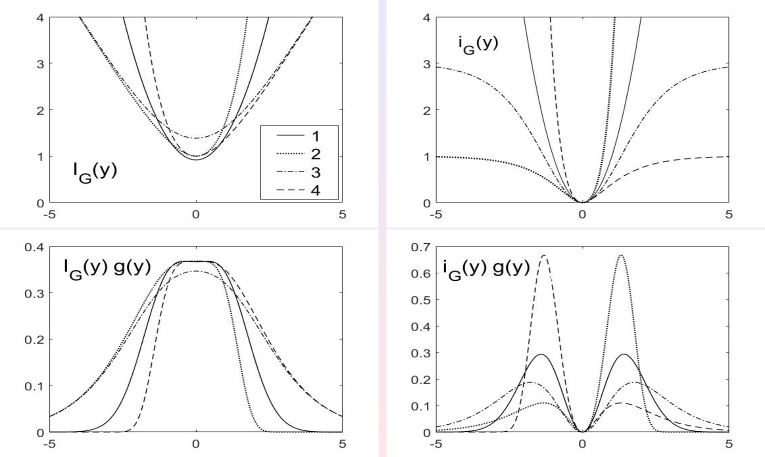
The “information function” $I_G(y) = -\log g(y)$ of information theory is compared with $i_G(y)$

Information functions “weighted” by densities on bottom panels

Distributions:

1 normal, 2 Gumbel, 3 logistic, 4 extreme value

Creating diff. entropy and mean uncertainty



Parametric distributions

For F_θ we easily generalize the t-score as

$$S_F(x;\theta) = S_G(\eta(x))\eta'(x^*)$$

where x^* is the score mean x^*

Define the information function of F as

$$i_F(x;\theta) = S_F^2(x;\theta)$$

Let $\theta = (x^*, \delta)$. The uncertainty function of F we define by

$$u_F(x;\theta) = \frac{i_F(x;\theta)}{[ES_F']^2}$$

where $S_F' = \frac{\partial S_F(x;x^*,\delta)}{\partial x^*}$

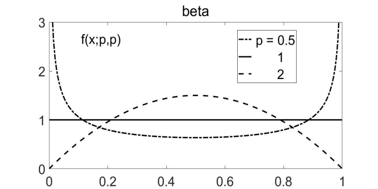
beta distribution

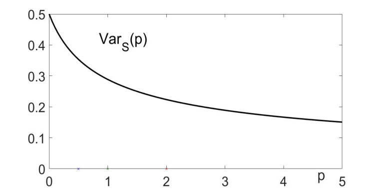
$$f(x;p,q) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, \; T_F(x;p,q) = (p+q)x - p$$

has linear t-score and its mean uncertainty is its variance

$$\omega^2 = Var_S X = \frac{pq}{(p+q+1)(p+q)^2}$$

beta(p,p) with $p = 0.5, 1, 2$, and $\sqrt{Var_S}$ of *beta*(p,p) as a function of p $\lim_{p\to 0}$ beta$(p,p) = (\frac{1}{2}, \frac{1}{2})$





Information and uncertainty in a data sample

Let $\mathbb{X}_n$ be a random sample from $F(x; x^*)$. Information contained in $\mathbb{X}_n$ is related to the score mean of F. The average

$$\frac{1}{n}\sum_{i=1}^{n} S_F^2(x_i; x^*)$$

converges as $n \to \infty$ to a finite value $ES_F^2(x^*)$. A new observation increases the precision of the estimate of x^* only, reducing the uncertainty (variance if asymptotically normal) of the estimate

Information distance of distributions

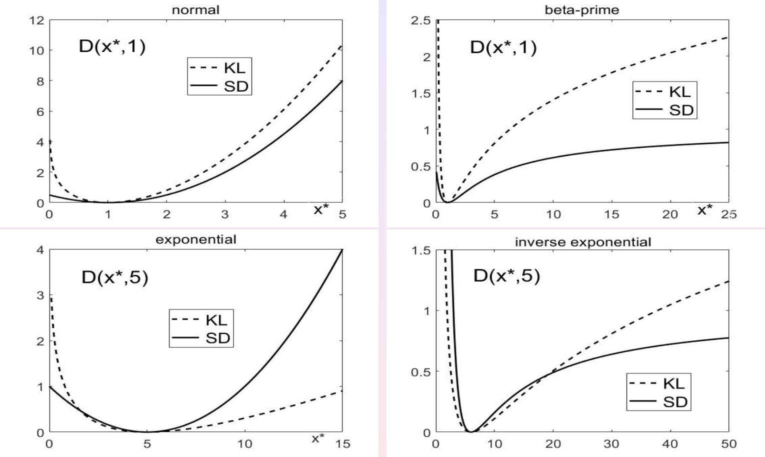
Relative entropy (Kullback-Leibler distance (KL)

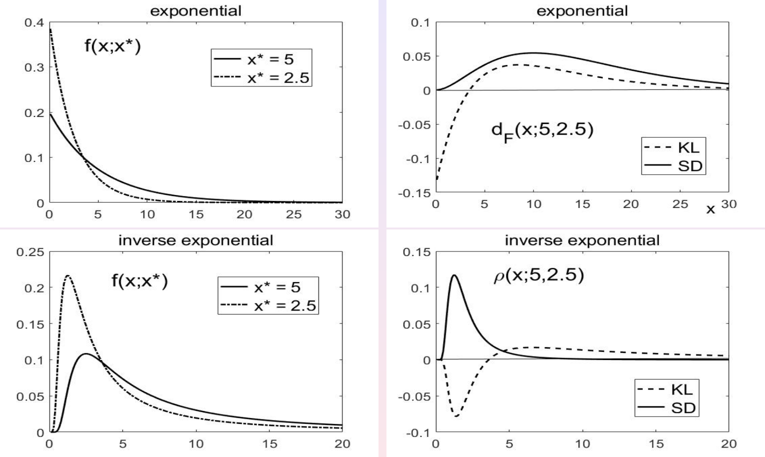
$$D_{KL}(F,H) = -\int_{\mathcal{X}} \log \frac{h(x)}{f(x)} f(x)\, dx$$

Information or score distance (SD)

$$D_{SD}(F,H) = \frac{1}{2ES_F^2} \int_{\mathcal{X}} [S_F(x) - S_H(x)]^2 f(x)\, dx$$

Let $D_j = \int_{\mathcal{X}} \rho_j(x)\, dx$


normal
D(x*,1)
KL
SD
x*
beta-prime
D(x*,1)
KL
SD
x*
exponential
D(x*,5)
KL
SD
x*
inverse exponential
D(x*,5)
KL
SD


exponential
f(x;x*)
x* = 5
x* = 2.5
0.4
0.3
0.2
0.1
0
0 5 10 15 20 25 30
exponential
d_F(x;5,2.5)
KL
SD
0.1
0.05
0
-0.05
-0.1
-0.15
0 5 10 15 20 25 x 30
inverse exponential
f(x;x*)
x* = 5
x* = 2.5
0.25
0.2
0.15
0.1
0.05
0
0 5 10 15 20
inverse exponential
ρ(x;5,2.5)
KL
SD
0.15
0.1
0.05
0
-0.05
-0.1
0 5 10 15 20

Fabián, Z. (2001). Induced cores and their use in robust parametric estimation. *Communs. Stat. Theory Meth.*

Fabián, Z. and Vajda, I. (2003). Core functions and core divergences of regular distributions. Kybernetika **39**, 29-42.

Fabián, Z. (2016). Score function of distribution and revival of the moment method. *Communs. Stat. Theory Meth.*

Fabián, Z. (2021). Mean, mode or median? The score mean. *Communs. Stat. Theory Meth.*

Fabián, Z. (2023). A measure of variability within parametric families of continuous distributions. *Communs. Stat. Theory Meth.*