T'aurais pas une entropie?

by jfb & co

Abstract

Where we show that it is possible to derive new entropies yielding a particular specified maximum entropy distribution. There are (probably) many errors –I hope not fundamental but is is possible; (certainly many) approximations, typos, maths and language mistakes.Suggestions and improvements will be much appreciated.

1. Maximum entropy distributions

Let f be a probability distribution defined with respect to a general measure μ on a set $\mathcal{X}$ and $S[f] = -\int_{\mathcal{X}} f(x) \log f(x) \mathrm{d}\mu(x)$ be the Shannon entropy of f. Subject to n moment constraints such as $\mathbb{E}[T_i(x)] = t_i, i = 1, \ldots, n$ and to normalization, it is well known that the maximum entropy distribution lies within the exponential family

$$f_X(x) = \exp\left(\sum_{i=1}^{n} \lambda_i T_i(x) + \lambda_0\right).$$

In order to recover known probability distributions (that must belong to the exponential family), it is then sufficient to specify a set of functions T_i, i.e., a function $T : \mathbb{R} \mapsto \mathbb{R}^n$ where n is the number of moment constraints. This has been used by many authors. For instance, the gamma distribution can be viewed as a maximum entropy distribution if one knows the moments $\mathbb{E}[X]$ and $\mathbb{E}[\log(X)]$. In order to find maximum entropy distributions with simpler constraints or distributions outside of the exponential family, it is possible to consider other entropies, which is discussed below. This problem find interests in goodness-and-fit tests based on maximum entropy principle.

2. Maximum (h, ϕ)-entropy distributions

2.1. Definition and maximum (h, ϕ)-entropy solution

Definition 1. *Let $\phi : \Omega \subset \mathbb{R}_+ \mapsto \mathbb{R}$ be a stricly convex differentiable function defined on a closed convex set Ω. Then, if f is a probability distribution defined with respect to a general measure $\mu(x)$ on a set $\mathcal{X}$,*

$$H_\phi[f] = -\int_{\mathcal{X}} \phi(f(x)) \mathrm{d}\mu(x) \tag{1}$$

is the ϕ-entropy of f.

Since $\phi(x)$ is convex, then the entropy functional $H_\phi[f]$ is concave. Also note that the composition of a concave function with a nondecreasing concave function preserves concavity, and that composition of a convex function with a nonincreasing convex function yields a concave functional.

Definition 2. *With the same assumption in definition* **??**,

$$H_{h,\phi}[f] = h\left(-\int_{\mathcal{X}} \phi(f(x))\mathrm{d}\mu(x)\right) \tag{2}$$

is called (h,ϕ)*-entropy of* f*, where*

- *either* ϕ *is convex and* h *concave nondecreasing*
- *or* ϕ *is concave and* h *convex nonincreasing*

These (h,ϕ)-entropies have been studied in [**?** **?**] for instance. In these works neither concavity (resp. convexity) of h, nor the differentiability of ϕ are imposed.

A useful related quantity to these entropies is the Bregman divergence associated with convex function ϕ:

Definition 3. *With the same assumption in definition* **??**, *the Bregman divergence associated with* ϕ *defined on a closed convex set* Ω, *is given by*

$$D_\phi(x_1, x_2) = \phi(x_1) - \phi(x_2) - \phi'(x_2)\,(x_1 - x_2)\,. \tag{3}$$

A direct consequence of the strict convexity of ϕ is the nonnegativity of the Bregman divergence: $D_\phi(x_1, x_2) \geq 0$ with equality if and only if $x_1 = x_2$.

Consider the problem of maximizing entropy (**??**) subject to constraints on some moments $\mathbb{E}\left[T(X)\right]$ where the normalization constraint is now included in T (namely $T_0(x) = 1$ and $t_0 = 1$). Without loss of generality, we consider in the sequel that ϕ is convex. Since h is nondecreasing, it is enough to look for the maximum of the ϕ-entropy (**??**),

$$\begin{cases} \max_f & -\displaystyle\int \phi(f(x))\mathrm{d}\mu(x) \\ \text{s.t.} & \mathbb{E}\left[T(X)\right] = t \end{cases} \tag{4}$$

Proposition 1. *The probability distribution* f_X *solution of the maximum entropy problem* (**??**) *satisfies the equation*

$$\phi'\big(f_X(x;t)\big) = \lambda^t\, T(x). \tag{5}$$

where vector λ *is such that* $\mathbb{E}[T(X)] = t$.

Proof. The maximization problem being concave, the solution exists and is unique. Equation **??** results directly from the classical Lagrange multipliers technique.

An alternative derivation of the result consists in checking that the distribution (**??**) is effectively a maximum entropy distribution, by showing that $H_\phi[f] > H_\phi[g]$ for all probability distributions with a given (fixed) moment $\mathbb{E}\left[T(X)\right]$. To this end, consider the functional Bregman divergence acting on functions defined on a comon domain $\mathcal{X}$:

$$D_\phi(f_1, f_2) = \int_{\mathcal{X}} \phi(f_1(x))\mathrm{d}\mu(x) - \int_{\mathcal{X}} \phi(f_2(x))\mathrm{d}\mu(x) - \int_{\mathcal{X}} \phi'(f_2(x))\,(f_1(x) - f_2(x))\,\mathrm{d}\mu(x).$$

From the nonnegativity of the Bregman divergence this functional divergence is nonnegative as well, and zero if and only if $f_1 = f_2$ almost everywhere. Define by

$$C_t = \{f : \mathcal{X} \mapsto \mathbb{R}_+ :\ \mathbb{E}\left[T(X)\right] = t\}$$

the set of all probability distributions defined on $\mathcal{X}$ with given moments t. Consider now $f_X \in C_t$ such that $\phi'(f_X(x)) = \lambda^t T(x)$ and any given function $f \in C_t$. Then

$$\begin{aligned} D_\phi(f, f_X) &= \int_{\mathcal{X}} \phi(f(x))\mathrm{d}\mu(x) - \int_{\mathcal{X}} \phi(f_X(x))\mathrm{d}\mu(x) - \int_{\mathcal{X}} \phi'(f_X(x))\,(f(x) - f_X(x))\,\mathrm{d}\mu(x) \\ &= -H_\phi[f] + H_\phi[f_X] - \int_{\mathcal{X}} \lambda^t\,T(x)\,(f(x) - f_X(x))\,\mathrm{d}\mu(x) \\ &= H_\phi[f_X] - H_\phi[f] \end{aligned}$$

where we used the fact that f and f_X have the same moments $\mathbb{E}\left[T(X)\right] = t$. By nonegativity of the Bregman functional divergence, we finally get that

$$H_\phi[f_X] \geq H_\phi[f]$$

for all pdf f with the same moments t than f_X, with equality if and only if $f = f_X$ almost everywhere. In other words, this shows that f_X, solution of (**??**), realizes the minimum of $H_\phi[f]$ over C_t. □

2.2. Defining new entropy functionals

Given an entropy functional, we thus obtain a maximum entropy distribution. There exists numerous (h, ϕ)-entropies in the literature. However a few of them lead to explicit forms for the maximum entropy distribution. Therefore, it is of high interest to look for the entropies that lead to a specified distribution as a maximum entropy solution. As pointed out previously, this find interests in goodness-and-fit tests based in entropies: it seems convenient to realize such tests using the entropy such that the distribution tested corresponds the its maximum entropy.

Since we will look for the function ϕ for a given probability distribution $f_X(x)$ we also see that the corresponding λ parameters can be included in the definition of the function.

Let us recall some implicit properties of $\phi(x)$.

- $\phi'(x)$ is defined on a domain included on $f_X(\mathcal{X})$;
- From the strict convexity property of ϕ, necessarily ϕ' is increasing.

The identification of a function $\phi(x)$ such that a given $f_X(x)$ is the associated maximum entropy distribution amounts to solve (**??**), that is

1. choose $T(x)$,
2. find ϕ' satisfying $\lambda^t T(x) = \phi'(f_X(x))$
3. integrate the result to get $\phi(y) = \int \phi'(y)\mathrm{d}y + c$, where c is an integration constant. The entropy being defined by $H_\phi[f] = -\int_{\mathcal{X}} \phi(f(x))\mathrm{d}\mu(x)$, the constant c will usually be zero.
4. Parameters λ may be choosen case by case in order to simplify the expression of ϕ.

Remind that ϕ' must be increasing, thus, necessarily, $\lambda^t\,T(x)$ and $f_X(x)$ must have the same sense of variation. Moreover, at a first glance, eq. (**??**) requires f_X and $\lambda^t T(x)$ sharing the same symmetries. Namely, if for two different values $x_1, x_2 \in \mathcal{X}$ the distribution satisfies $f_X(x_1) = f_X(x_2)$ then $\lambda^t\,T(x_1) = \lambda^t\,T(x_2)$ (this does mean that $T(x_1)$ and $T(x_2)$ must be equals). Thus, eq. (**??**) rewrites

$$\phi'(y) = \lambda^t T(f_X^{-1}(y)) \tag{6}$$

where f_X^{-1} can be multivalued; in such a situation, ϕ' remains well defined. Note again that for given T and f_X, the solution is not unique due to parameters λ, which can be chosen finely so as to simplify

the expression of ϕ'. Eq. (**??**) can then be integrated, at least formaly, to achieve H_ϕ (and thus any $H_{h,\phi}$ entropy with nondecreasing h).

For instance, for one moment constraint, if λ_1 is negative, then

- for $T_1(x) = x$, $f_X(x)$ must be decreasing,
- for $T_1(x) = x^2$ or $T_1(x) = |x|$, $f_X(x)$ must be unimodal with a maximum at zero.

3. ϕ-escort, ϕ-Fisher information and generalized Cramér-Rao inequality

4. Some examples

4.1. Normal distribution and second-order moment

For a normal distribution, and second order moment constraint

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{x^2}{2\,\sigma^2}\right) \qquad \text{and} \qquad \lambda^t T(x) = \lambda_0 + \lambda_1 x^2$$

We begin by computing the inverse of $y = f_X(x)$ where $x \in \mathbb{R}_+$ for instance, which gives

$$\phi'(y) = \left(\lambda_0 - \sigma^2 \log(2\pi\sigma^2)\,\lambda_1\right) - 2\sigma^2\lambda_1 \log y$$

The judicious choice

$$\lambda_0 = 1 - \log(\sqrt{2\pi}\sigma) \qquad \text{and} \qquad \lambda_1 = -\frac{1}{2\,\sigma^2}$$

leads to function

$$\phi(y) = y \log y$$

that gives nothing more than the Shannon entropy as expected.

4.2. q-Normal distribution and second-order moment

For q-normal distribution, also known as Tsallis distributions, and a second order moment constraint,

$$f_X(x) = C_q\Big(1 - (q-1)\beta x^2\Big)_+^{\frac{1}{(q-1)}} \qquad \text{and} \qquad \lambda^t T(x) = \lambda_0 + \lambda_1 x^2$$

where $q > 0$ and $x_+ = \max(x, 0)$ we get

$$\phi'(y) = \left(\lambda_0 + \frac{\lambda_1}{(q-1)\beta}\right) - \frac{\lambda_1\, y^{q-1}}{C_q^{q-1}(q-1)\beta}$$

In this case, a judicious choise of parameters is

$$\lambda_0 = \frac{q\, C_q^{q-1}}{(q-1)\beta} \qquad \text{and} \qquad \lambda_1 = -q\, C_q^{q-1}\beta$$

that yields to

$$\phi(y) = \frac{y^q}{q-1}.$$

and an associated entropy can be

$$H_{h,\phi}[f] = \frac{1}{1-q}\left(\int_{\mathcal{X}} f(x)^q \mathrm{d}\mu(x) - 1\right),$$

which is nothing but Tsallis entropy.

4.3. q-exponential distribution and first-order moment

4.4. Hyperbolic secant distribution and first-order moment

Let us consider some specific cases.

1. For a normal distribution, $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ and $T(x) = x^2$, we begin by computing the inverse $y = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, which gives $-\frac{1}{2}x^2 - \log\sqrt{2\pi} = \log(y)$. Choosing $\lambda = -\frac{1}{2}$, $\mu = -\log\sqrt{2\pi}$ and integrating, we obtain

$$\phi(y) = y \log y - y$$

2. For a Tsallis q-exponential, $f_X(x) = C_q \left(1 - (q-1)\beta x\right)_+^{\frac{1}{(q-1)}}$, $x \geq 0$, and $T(x) = x$. We simply have $C_q^{q-1}\left(1 - (q-1)\beta x\right) = y^{q-1}$. With $\lambda = qC_q^{q-1}\beta$ and $\mu = qC_q^{q-1}/(1-q)$, this yields

$$\phi(y) = \frac{y^q}{1-q}.$$

Taking $\mu = \left(qC_q^{q-1} + 1\right)/(1-q)$ gives

$$\phi(y) = \frac{y^q - y}{1-q},$$

and an associated entropy can be

$$H_\phi[f] = \frac{1}{1-q}\left(\int f(x)^q \mathrm{d}\mu(x) - 1\right),$$

which is nothing but Tsallis entropy.[1]

3. The same entropy functional can readily be obtained for the so-called q-Gaussian, or Student-t and -r distributions $f_X(x) = C_q \left(1 - (q-1)\beta x^2\right)_+^{\frac{1}{(q-1)}}$. It suffices to follow the very same steps as above with $T(x) = x^2$.
4. Let $f_X(x)$ be the hyperbolic secant distribution, with density

$$f_X(x) = \frac{1}{2}\mathrm{sech}(\frac{\pi}{2}x) = \frac{1}{2}\cosh^{-1}(\frac{\pi}{2}x).$$

Obviously, $\frac{\pi}{2}x = \cosh(2y) = \phi'(y)$ with $T(x) = x$, $\lambda = \frac{\pi}{2}$, and

$$\phi(y) = \sinh(2y).$$

So doing, we obtain an hyperbolic sine entropy with the hyperbolic secant distribution as the associated maximum entropy distribution.

Of course, the preceeding derivations require that (**??**) is effectively solvable. Furthermore, one has also to choose or design a specific $T(x)$ statistic, as well as the parameters λ and μ. In the examples above, we used $T(x) = x$ and $T(x) = x^2$. Particular choices such as $T(x) = x^2$ or $T(x) = |x|$ obviously lead to symmetrical densities. The case of nonsymmetrical unimodal densities seems to be much more involved. For instance, if we take $T(x) = x$, then the resolution of (**??**) amounts to compute the inverse relation of $f_X(x)$, which is is multi-valued. We will deal now with this special case.

[1] Of course, we can also take the first $\phi(y) = \frac{y^q}{1-q}$, integrate and add any constant, since adding a constant do not modify the actual value af the minimizer (or maximizer if we consider concave entropies).

4.5. Entropies for unimodal nonsymmetric distributions

Assume, without loss of generality that the mode is $x = 0$. Let $T(x) = x$, and $\lambda = -1$, $\mu = 0$. In such case, we have to find ϕ satisfying $x = -\phi'(f_X(x)) = -\phi'(y)$. We see that ϕ' is minus the inverse relation of $y = f_X(x)$. But $f_X(x)$ is not injective and to each y correspond a positive and a negative value of x. Hence we have two partial inverses, say ϕ'_+ and ϕ'_- such that $\phi'^{-1}_+(-x) = f_X(x)$ for $x \geq 0$ and $\phi'^{-1}_-(-x) = f_X(x)$ for $x \leq 0$. We observed above that if $f_X(x)$ is non increasing, that is assumed here for $x \geq 0$, then $\phi''_+ \geq 0$ and $\phi_+$is convex. Then, our proposal is to use the functional ϕ_+ for defining a ϕ-entropy

$$H_\phi[f_X] = \int \phi_+(f_X(x))\,\mathrm{d}\mu x$$

associated with a specific nonsymmetric probability distribution. In this setting, it is understood that the maximum entropy distribution $f_X(x) = \phi'^{-1}(-x)$ will have to be computed as $\phi'^{-1}_+(-x) = f_X(x)$ for $x \geq 0$ and $\phi'^{-1}_-(-x) = f_X(x)$ for $x \leq 0$. Of course, this does not forbid to model one-sided probability distribution, provided that the constraint is included in the formulation of the maximum entropy problem.

4.5.1. Example 1. The logistic distribution

The pdf of the logistic distribution is given by

$$f_X(x) = \frac{e^{-\frac{x}{s}}}{s\left(1+e^{-\frac{x}{s}}\right)^2}.$$

This distribution, which resembles the normal distribution but has heavier tails, has been used in many applications. By direct calculations, we obtain

$$\begin{cases} \phi'_-(y) = & s\ln\left(\frac{1}{2}\,\frac{-2\,ys+1+\sqrt{-4\,ys+1}}{ys}\right), \\ \phi'_+(y) = & s\ln\left(-\frac{1}{2}\,\frac{2\,ys-1+\sqrt{-4\,ys+1}}{ys}\right). \end{cases}$$

The associated entropy is then

$$\phi_+(y) = \frac{1}{2}\sqrt{-4\,ys+1} + ys\,\ln\left(-\frac{\sqrt{-4\,ys+1}-1}{\sqrt{-4\,ys+1}+1}\right),$$

for $y \in [0, \frac{1}{4s}]$, and where we have introduced a integration constant such that $\min_y \phi_+(y) = 0$. For $y > \frac{1}{4s}$, we extend the function and let $\phi_+(y) = +\infty$. Figure **??** gives a representation of this entropy for $s = 1$.

4.5.2. Example 2. The gamma distribution

The probability density function of the gamma distribution is given by

$$f_X(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta\,x}}{\Gamma(\alpha)}$$

We obtain

$$\phi'(y) = -\mathrm{e}^{\frac{1}{\alpha-1}\left(-W\left(-\frac{\beta\left(y\Gamma(\alpha)\beta^{-\alpha}\right)^{(\alpha-1)^{-1}}}{\alpha-1}\right)\alpha+W\left(-\frac{\beta\left(y\Gamma(\alpha)\beta^{-\alpha}\right)^{(\alpha-1)^{-1}}}{\alpha-1}\right)+\ln\left(y\Gamma(\alpha)\beta^{-\alpha}\right)\right)},$$

where W is the Lambert W multivalued 'function' defined by $z = W(z)e^{W(z)}$ (ie the inverse relation of $f(w) = we^w$). Unfortunately, in the general case, we do not have a closed form for $\phi(y)$ as the integral of $\phi'(y)$.[2] Restricting us to the case $\alpha = 2$, we have

$$\phi(y) = \frac{\left(1 - W\left(-\frac{y}{\beta}\right) + y\left(W\left(-\frac{y}{\beta}\right)\right)^2\right)}{\beta\, W\left(-\frac{y}{\beta}\right)} + \frac{\beta}{e},$$

which is convex if we choose the -1 branch of the Lambert function. An example with $\alpha = 2$ and $\beta = 3$ is given on Figure **??**.

4.5.3. Example 3. The arcsine distribution

As a last example, and though it is not a unimodal density (! but yields the same problem for inversion), let us consider the case of the arcsine distribution (see wiki). This distribution, defined for $x \in (0, 1)$, is a special case of the Beta distribution with parameters $\alpha = \beta = 1/2$. It has the following pdf:

$$f_X(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

Observe that $\min_x f_X(x) = 2/\pi$. Doing our now usual calculations, we obtain

$$\begin{cases} \phi'_-(y) = & -\frac{y\pi + \sqrt{y^2\pi^2 - 4}}{2y\pi}, \\ \phi'_+(y) = & -\frac{y\pi - \sqrt{y^2\pi^2 - 4}}{2y\pi}. \end{cases}$$

and the expression of the entropy is

$$\phi_+(y) = \frac{1}{2}\,\frac{\sqrt{y^2\pi^2 - 4}}{\pi} + \frac{1}{\pi}\arctan\left(2\,\frac{1}{\sqrt{y^2\pi^2 - 4}}\right) - \frac{1}{2}\,y,$$

for $y \geq 1/\pi$. The entropy is shown on Figure **??**.

[2] This might not be completely unacceptable. Indeed, it is really not difficult to compute numerically the values of $\phi(y)$.