Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA
Spring, 2006

1. Entropy

The *entropy* of a probability distribution on a finite set $\mathcal{X}$ (say, expressed through a probability mass function $p(x)$) was introduced by Shannon (1938) and further developed in (Shannon 1948), defined as:

$$H(p) \equiv -\sum_{x\in\mathcal{X}} p(x)\log\big(p(x)\big), \tag{1}$$

the expectation of the negative logarithm of the p.m.f. itself. This is a measure of the "randomness" of the distribution— for a given set $\mathcal{X}$, it attains a minimum at any point mass (where $p(x_0)=1$ for some $x_0\in\mathcal{X}$ and $p(x)=0$ for all $x\neq x_0$), and achieves its maximum value of $H(p^*)=\log(n)$ at the equiprobable distribution assigning probability $p^*(x)\equiv 1/n$ to each of the $n=|\mathcal{X}|$ points of $\mathcal{X}$ (you can prove this using Lagrange multipliers, as below, or using the calculus of variations).

Three properties of $H(p)$ are:

1. $H(p)\geq 0$, and $H(p)=0$ if and only if $p(x_0)=1$ for some $x_0\in\mathcal{X}$;
2. If X and Y are independent discrete random variables, then their marginal and joint distributions satisfy

   $$H(p_{XY}) = H(p_X) + H(p_Y)$$

3. If we extend p from $\mathcal{X}$ to $\mathcal{X}\cup\mathcal{Y}$ by $(\forall y\in\mathcal{Y})\ \{p(y)=0\}$, then $H(p)$ is unchanged.

In fact these characterize H uniquely— Khinchine (1957) showed that any function of discrete probability distributions satisfying these three properties is necessarily $H(p)$, up to an arbitrary scale factor.

For continuous (and more general) distributions it is more problematic to define *entropy* in a useful way; the tempting definition

$$H(f) \equiv -\int_{\mathcal{X}} f(x) \log \big(f(x)\big)\, dx, \tag{2}$$

is no longer invariant under changes in variables, as is Equation (1); is no longer necessarily nonnegative or finite; and also depends critically on the arbitrary choice of a dominating measure on $\mathcal{X}$ (here taken to be Lebesgue measure). Nevertheless it is occasionally useful (we will use it below); a few simple examples include

$\mathsf{Un}(a,b)$	$f(x) = \frac{1}{b-a},\ a < x < b$	$H(f) = \log(b-a)$
$\mathsf{No}(\mu,\Sigma)$	$f(x) = \frac{1}{\sqrt{\det 2\pi\Sigma}}\, e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$	$H(f) = \frac{1}{2}\log\det\big[2\pi e\Sigma\big]$
$\mathsf{Ex}(\lambda)$	$f(x) = \lambda e^{-\lambda x},\ x > 0$	$H(f) = 1 - \log\lambda$

To solve the constrained optimization problem

$$\begin{array}{lrcll} \textbf{Maximize:} & H(p) & = & -\sum_{x\in\mathcal{X}} p(x)\, \log\big(p(x)\big) & \\ \textbf{Subject to:} & c_j & = & \sum_{x\in\mathcal{X}} p(x)\, g_j(x), & j = 1\cdots k \end{array}$$

begin by seeking critical points of the Lagrangian,

$$\mathcal{L}(p,\lambda) \quad = \quad H(p) + \textstyle\sum_{j=0}^{k} \lambda_j \big(c_j - \sum_{x\in\mathcal{X}} g_j(x)\big),$$

where we constrain the probabilities $p(x)$ to sum to one by introducing a $k{+}1$'st constraint $g_0(x) \equiv 1 = c_0$. The solution, if one exists, is of the form

$$p(x) = e^{\sum_{j=0}^{k} \lambda_j g_j(x)},$$

where the $k+1$ numbers $\{\lambda_j\}$ are determined by the $k+1$ constraints.

1.1. Kullback-Leibler

Motivated by Shannon's notion of entropy, Kullback and Leibler (1951) defined what we now know as the "Kullback-Leibler Divergence" (also called "relative entropy") between any two probability distributions on the same (not necessarily finite) set $\mathcal{X}$ by

$$K(f:g) \equiv \int_{\mathcal{X}} -f(x)\, \log\left[\frac{g(x)}{f(x)}\right]\, dx,$$

where $f(x)$ and $g(x)$ denote density functions of the two distributions; if the two distributions are discrete just replace the integral by a sum, and in fact $K(f:g)$ is well-defined for any two distributions even if they don't have densities (the Radon-Nikodym derivative $g(dx)/f(dx)$ replaces the fraction $g(x)/f(x)$). For finite sets $\mathcal{X}$, the entropy $H(f)$ introduce above can be expressed as $H(f) = n - K(f:p^*)$, where (as before) $p^*(x) \equiv 1/n$ denotes the equiprobable distribution.

The KL divergence $K(f:g)$ is a measure of the discrepancy between the two distributions, in the sense that always $K(f:g) \geq 0$ and that $K(f:g) = 0$ if and only if $f \equiv g$, but it is not a *distance* because it fails both symmetry (i.e., $K(f:g) \neq K(g:f)$ in general) and the triangle inequality (i.e., there are triplets f, g, h for which $K(f:h) \not\leq K(f:g) + K(g:h)$). It is closely related to a notion of distance, however, as we will see in Section (1.2) below. First,

Proposition 1 *Let $g > 0$ be any positive integrable function on a set $\mathcal{X}$ in Euclidean space. If the functional $K(f:g)$ attains a unique minimum over all positive measurable functions $f(x)$ at some $\hat{f}$, then $\hat{f}(x) \equiv cg(x)$ where $c = 1/\int_{\mathcal{X}} g(x)\,dx$.*

Proof. For any integrable function $h(x)$ on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} h(x)\,dx = 0$, the function

$$\psi(\epsilon) = K(\hat{f} + \epsilon h : g)$$

will exhibit a local minimum at $\epsilon = 0$. Taking derivatives, we have

$$\begin{aligned} 0 = \psi'(0) &= \frac{d}{d\epsilon}_{|\epsilon=0} \int_{\mathcal{X}} [\hat{f}(x) + \epsilon h(x)] \left\{ \log[\hat{f}(x) + \epsilon h(x)] - \log[g(x)] \right\} dx \\ &= \int_{\mathcal{X}} h(x) \log \left[\frac{\hat{f}(x)}{g(x)} \right] + \int_{\mathcal{X}} \hat{f}(x) \left[\frac{h(x)}{\hat{f}(x)} \right] dx \\ &= \int_{\mathcal{X}} h(x) \log \left[\frac{\hat{f}(x)}{g(x)} \right]. \end{aligned}$$

Since this integral must vanish for all functions $h(x)$ satisfying $\int_{\mathcal{X}} h(x)\,dx = 0$, it follows that $\hat{f}(x)/g(x)$ must some constant c; the result follows.

□

1.2. Kullback-Leibler and Fisher Information

For small $\epsilon > 0$, a second-order Taylor-series approximation of the KL divergence from $f(x \mid \theta)$ to $f(x \mid \theta + \epsilon)$ gives

$$\begin{aligned} K\big(f(x \mid \theta) : f(x \mid \theta + \epsilon)\big) &= \int_{\mathcal{X}} -f(x \mid \theta) \log \left[\frac{f(x \mid \theta + \epsilon)}{f(x \mid \theta)}\right] dx \\ &\approx \int_{\mathcal{X}} -f(x \mid \theta) \left[\epsilon \nabla \log f(x \mid \theta) + \frac{\epsilon^2}{2} \nabla^2 \log f(x \mid \theta)\right] dx \\ &= \frac{1}{2} I(\theta)\, \epsilon^2 \end{aligned}$$

or, in $q > 1$ dimensions, $\epsilon^{\intercal} I(\theta) \epsilon / 2$. This illustrates the close link between KL divergence and the "Information Metric" Riemannian distance between different distributions,

$$\begin{aligned} d_I(\theta_0, \theta_1) &= \int_{\theta_0}^{\theta_1} \sqrt{I(\theta)}\, d\theta \qquad \text{(in } q = 1 \text{ dimension)} \\ &= \inf_{\theta} \int_0^1 \sqrt{\theta_t'^{\intercal} I(\theta_t) \theta_t'}\, dt \end{aligned}$$

where the infimum is over all differentiable paths θ from θ_0 to θ_1, and where θ_t' is the velocity vector $(d/dt)\theta_t$; evidently, for near-by $\theta_0 \approx \theta_1$,

$$2\, K\big(f(x \mid \theta_0) : f(x \mid \theta_1)\big) \approx d_I(\theta_0, \theta_1)^2,$$

and for any θ_0, θ_1,

$$d_I(\theta_0, \theta_1) = \inf_{\theta} \lim_{n \to \infty} \sum_{i=1}^{n} \sqrt{2K\left(f\big(x \mid \theta_{(i-1/n)}\big) : f\big(x \mid \theta_{i/n}\big)\right)}.$$

Roughly, KL is half the *square* of a distance measure, and small open KL "balls" are also small open Information-metric balls.

2. Prior and Posterior Distributions

A number of authors have used information-theoretic (*i.e.*, entropy-based) ideas and/or Kullback-Leibler divergence in the search for Bayesian prior distributions that are as diffuse as possible, in hope of expressing *no* prior knowledge about a parameter $\theta \in \Theta$, beginning with statistician Dennis V. Lindley (1956) and physicist Edwin T. Jaynes (1957a; 1957b). There is a thriving “Max-Ent” community of (mostly Bayesian) statisticians and economitricians, among them ISBA founder Arnold Zellner. The most widely successful approach, in my opinion, is that of Bernardo (1979) (developed more fully in Berger and Bernardo 1992), which we now sketch.

Bernardo, motivated by Lindley (1956), sought (initially) to *maximize* the KL divergence between the prior and posterior distributions, looking for an extremum π for the expected value (under the marginal distribution for X) of the KL divergence from the prior $\pi(\theta)$ to the posterior $\pi(\theta \mid X)$,

$$K(\pi : \pi_{|X}) = \int_{\mathcal{X}} \left\{ \int_{\Theta} -\pi(\theta \mid x)\ \log \left[\frac{\pi(\theta \mid x)}{\pi(\theta)} \right] d\theta \right\} m(x)\, dx,$$

where $m(x) \equiv \int_\Theta f(x \mid \theta)\, \pi(\theta)\, d\theta$ is the marginal density function for X and where $\pi(\theta \mid x) = m(x)^{-1} f(x \mid \theta) \pi(\theta)$ is the posterior (or conditional) density for θ, given x. The initial attempt hit three obstacles: the optimum often doesn’t exist; when it does, it is often a discrete distribution concentrated on a few points, obiously a poor candidate for a prior distribution expressing minimal prior information; and even if it does exist and isn’t discrete, it can be very difficult to compute. All three difficulties were overcome by the artifice of imaginging a large sample of size n, rather than a single observation, and eventually taking a limit as $n \to \infty$; thus we try to maximize

$$K(\pi : \pi_{|\mathbf{X_n}}) = \int_{\mathcal{X}^n} \left\{ \int_{\Theta} -\pi(\theta \mid \mathbf{x}_n)\ \log \left[\frac{\pi(\theta \mid \mathbf{x}_n)}{\pi(\theta)} \right] d\theta \right\} m(\mathbf{x}_n)\, d^n x, \qquad (3)$$

where $m(\mathbf{x}_n) \equiv \int_\Theta \left\{ \prod_{j=1}^n f(x_j \mid \theta) \right\} \pi(\theta)\, d\theta$. Writing

$$H(p) = \int -p(\theta) \log[p(\theta)]\, d\theta$$

for any p.d.f. $p(\theta)$, we can re-write Equation (3) as

$$\begin{aligned}
K\big(\pi : \pi_{|\mathbf{X_n}}\big) &= -H(\pi) + \int_{\mathcal{X}^n} m(\mathbf{x}_n) H\big(\pi(\theta \mid \mathbf{x}_n)\big)\, d^n x \\
&= \int_\Theta \pi(\theta) \log \pi(\theta)\, d\theta \\
&\quad + \int_{\mathcal{X}^n} \left[\int_\Theta f(\mathbf{x}_n \mid \theta) \pi(\theta)\, d\theta \right] H\big(\pi(\theta \mid \mathbf{x}_n)\big)\, d^n x \\
&= \int_\Theta -\pi(\theta) \log \left[\exp\left\{ -\int_{\mathcal{X}^n} f(\mathbf{x}_n \mid \theta) H\big(\pi(\theta \mid \mathbf{x}_n)\big)\, d^n x \right\} / \pi(\theta) \right] d\theta \qquad (4)
\end{aligned}$$

A calculus-of-variations argument shows that, under suitable regularity conditions, any expression (like Equation (4)) of the form $\int \pi(\theta)\log[f(\theta)/\pi(\theta)]\,d\theta$ is maximized by a solution of the form $\pi(\theta) \propto f(\theta)$; thus the optimal $\pi(\theta)$ will be of the form

$$\pi(\theta) \propto \exp\left\{ \int_{\mathcal{X}^n} -f(\mathbf{x}_n \mid \theta) H\big(\pi(\theta \mid \mathbf{x}_n)\big)\, d^n x \right\} \tag{5}$$

From Bayes' theorem and the Bayesian CLT we know that (also under suitable regularity)

$$\begin{aligned} \pi(\theta \mid \mathbf{x}_n) &= m(\mathbf{x}_n)^{-1} f(\mathbf{x}_n \mid \theta)\,\pi(\theta) \\ &\approx \sqrt{\det\Big(\frac{n}{2\pi} I(\theta)\Big)}\;\exp\Big\{-\frac{n}{2}(\theta-\hat{\theta})' I(\theta)(\theta-\hat{\theta})\Big\} \end{aligned}$$

and hence that

$$H\big(\pi(\theta \mid \mathbf{x}_n)\big) \approx \frac{q}{2}\log(2\pi e/n) - \frac{1}{2}\log\det\big(I(\theta)\big),$$

since the normal p.d.f. $p \sim \mathsf{No}(\mu, \Sigma)$ in q dimensions has entropy $H(p) = (1/2)\log\det(2\pi e \Sigma)$, so we have

$$\begin{aligned} \pi(\theta) &\propto \exp\left\{ \int_{\mathcal{X}^n} -f(\mathbf{x}_n \mid \theta) H\big(\pi(\theta \mid \mathbf{x}_n)\big)\, d^n x \right\} \\ &\approx \exp\left\{ -\frac{q}{2}\log(2\pi e/n) + \frac{1}{2}\log\det\big(I(\theta)\big) \right\} \\ &\propto \sqrt{\det I(\theta)}, \end{aligned} \tag{6}$$

giving Jeffreys' prior a new motivation.

Bernardo and Berger do not recommend this choice for parameters of dimension $q > 1$; instead they have an elaborate argument for how to build what they call Reference Priors, one dimension at a time, for the goal of estimating "parameters of interest" in the presence of "nuisance parameters". Take a look at the references below for more details.

References

Berger, J. O. and Bernardo, J. M. (1992), "On the development of the referrence prior method," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 35–49.

Bernardo, J. M. (1979), “Reference posterior distributions for Bayesian inference (with discussion),” *Journal of the Royal Statistical Society, Ser.* B *(Statistical Methodology)*, 41, 113–147.

Jaynes, E. T. (1957a), “Information Theory and Statistical Mechanics,” *Physical Review*, 106, 620–630.

Jaynes, E. T. (1957b), “Information Theory and Statistical Mechanics II,” *Physical Review*, 108, 171–190.

Khinchine, A. Y. (1957), *Mathematical Foundations of Information Theory*, New York, NY: Dover.

Kullback, S. and Leibler, R. A. (1951), “On Information and sufficiency,” *Annals of Mathematical Statistics*, 22, 79–86.

Lindley, D. V. (1956), “On a measure of the information provided by an experiment,” *Annals of Statistics*, 27, 986–1005.

Shannon, C. E. (1938), “A Symbolic Analysis of Relay and Switching Circuits,” *Transactions of the American Institute of Electrical Engineers*, 57, 713–723.

Shannon, C. E. (1948), “A Mathematical Theory of Communication,” *Bell System Technical Journal*, 27, 379–423 and 623–656.