

Prof. H. Bensmail

Stat-664: Adv Stat Infer

Spring Semester 2006

## **Chapter 7:**

### **Priors**

**March 13, 2006**

## Maximum Entropy:

Information: see Applebaum (1996)

There are three conditions that an information measure must meet:

1. It must be positive
2. The information from observing two events jointly must be at least as large as that from observation of any of the single elementary events.
3. For independent events  $E_1$  and  $E_2$

$$I(E_1 \cap E_2) = I(E_1) + I(E_2)$$

From information theory, the function satisfying the three conditions given above must have the form

$$I(E) = -K \log_a [p(E)]$$

where  $K$  and  $a$  are positive constants.

Since  $0 \leq P(E) \leq 1$ , it follows that this information measure is positive, as  $K$  is positive, meets (1)

If the event is certain,  $I(E) = 0$  so no information is gained from knowing that the event took place.

If the event is impossible,  $p(E) = 0$  and  $I(E) = \infty$

This indicates the possibility of obtaining information from events that do not occur.

(2)

$$\begin{aligned} I(E_1 \cap E_2) &= -K \log_a [p(E_1 \cap E_2)] \\ &= -K \log_a [p(E_1)] - K \log_a [p(E_2)] \\ &= I(E_1) + I(E_2) \end{aligned}$$

then (2) and (3) are satisfied

When  $K = 1$  and  $a = 2$ , are the standards choice and the units of  $I$  in this case is called: "bits"

## Entropy for discrete Distribution

From Box and Tiao (1973); Zellner (1971); Bernardo (1979); Berger and Bernardo (1992): In Bayesian analysis, one wishes to find and use a prior distribution conveying "as little information as possible"

↓

Requires finding a distribution that minimizes some information measure.

Suppose

$$X \sim \text{Discrete}(p_1, \dots, p_K)$$

which means

$$p(X = x_i) = p_i$$

$I(X)$  is random

$$\begin{aligned} H(p_1, \dots, p_K) &= E[I(X)] \\ &= E\{-\log[p(X = x_i)]\} \\ &= -\sum p_i \log(p_i) \end{aligned} \tag{1}$$

this is called the entropy of a distribution.

**Example: sampling of genes**

Let allele A have frequency  $p$  in some population, so that all other possible alleles appear with probability  $(1 - p)$ .

Suppose that  $n$  alleles are drawn at random and  $x$  are of A form.

The process is binomial and the entropy of the distribution of the random variable  $X$  is

$$H(p) = -\sum \left[ \log \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!} \right] \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!}$$

if  $n = 1$  this reduces to the entropy of a Bernoulli distribution

$$H(p) = -p \log(p) - (1-p) \log(1-p)$$

$$\frac{dH(p)}{dp} = -\log(p) + \log(1-p)$$

when  $\frac{dH(p)}{dp} = 0$ , this gives  $p = 1/2$  as the gene frequency giving maximum entropy, with the maximized entropy being equal to 1 bit when expressed in a  $\log_2$  base.

**Conclusion:**

The gene frequency distribution producing maximum entropy is that corresponding to the situation where allele A has the same freq as that of all the other alleles combined.

In general for  $K$  states, the entropy (1) has the following gradient:

$$-\log(p_i) + \log(1 - p_1 - p_2 - \dots - p_{K-1})$$

which gives

$$H\left(\frac{1}{K}, \dots, \frac{1}{K}\right) = -\sum \frac{1}{K} \log\left(\frac{1}{K}\right) = \log(K)$$

**Entropy of a joint distribution and conditional distribution:**

Suppose that  $(X, Y)$  has a joint dist

$$(X, Y) \sim p_{xy}$$

$(x = 1, \dots, m; j = 1, \dots, n)$

The entropy of joint distribution

$$\begin{aligned}
 H(p_{11}, p_{12}, \dots, p_{mn}) &= - \sum_{x=1}^m \sum_{y=1}^n p_{xy} \log(p_{xy}) \\
 &= - \sum_{x=1}^m \sum_{y=1}^n p_x p_{y|x} \log(p_x p_{y|x}) \\
 &= \dots \\
 &= - \sum_{x=1}^m p_x \log(p_x) \sum_{y=1}^n p_y - \sum_{y=1}^n p_y \log(p_y) \sum_{x=1}^m p_x \\
 &= H(\mathbf{p}_x) + H(\mathbf{p}_y)
 \end{aligned}$$

where  $p_y$  is the vector of probabilities of the distribution of  $Y$ .

## Entropy of a continuous Distribution

The entropy of a continuous distribution is

$$H[p(\mathbf{y}|\theta)] = - \int \dots \int \log[p(\mathbf{y}|\theta)] p(\mathbf{y}|\theta) d\mathbf{y}.$$

where  $p(\mathbf{y}|\theta)$  is the density function of a random vector  $\mathbf{y}$ , indexed by  $\theta$ .

### Particular case:

Entropy is not invariant under a transformation

Suppose  $z = f(y)$  where  $z$  increases monotonically with  $y$ . Then

$$p(z|\theta) = p[f^{-1}(z)|\theta] \frac{df^{-1}(z)}{dz}$$

The entropy of the distribution of  $z$  becomes

$$\begin{aligned}
 H[p(z|\theta)] &= - \int \log \left\{ p[f^{-1}(z)|\theta] \frac{df^{-1}(z)}{dz} \right\} p[f^{-1}(z)|\theta] \frac{df^{-1}(z)}{dz} dz \\
 &= - \int \log \{ p[f^{-1}(z)|\theta] \} p[f^{-1}(z)|\theta] \frac{df^{-1}(z)}{dz} dz \\
 &\quad - \int \log \left[ \frac{df^{-1}(z)}{dz} \right] p[f^{-1}(z)|\theta] \frac{df^{-1}(z)}{dz} dz \\
 &= - \int \{ \log[p(y|\theta)] \} p(y|\theta) dy - \int \log \left[ \frac{df^{-1}(z)}{dz} \right] p(z|\theta) dz \\
 &= H[p(y|\theta)] - E_z \left\{ \log \left[ \frac{df^{-1}(z)}{dz} \right] \right\}
 \end{aligned}$$

### Entropy of a uniform distribution:

$$\begin{aligned}
H[p(y|\theta)] &= -\int_a^b \left[ \log\left(\frac{1}{b-a}\right) \right] \frac{1}{b-a} dy \\
&= \log(b-a)
\end{aligned}$$

### Entropy of a normal distribution

$$\begin{aligned}
H[p(y|\mu, \sigma^2)] &= -\int \left\{ \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]\right) \right\} \\
&\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right] dy \\
&= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy \\
&\quad + \frac{1}{2\sigma^2} \int (y-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy \\
&= \frac{1}{2} [1 + \log(2\pi\sigma^2)]
\end{aligned}$$

### Priors conveying little information

**Jeffrey's 1961:** We use prior that has little relative to the contributions of data.  
 We have seen that effect of priors dissipates as the sample sizes increases.  
 So problem: effect of prior choice on finite sample inferences

### Uniform prior:

noninformative: most widely used priors

In the absence of evidence all possibilities should have the same prior probability  
 (Bernardo and Smith 1994)

If  $\theta =$  one of  $K$  possible values, then the noninformative prior on  $\theta$  is

$$(1/K, \dots, 1/K)$$

this is also a maximum entropy distribution when all that is known is that there are  $K$  mutually exclusive and exhaustive states.

Suppose

$$\theta \sim \text{Uniform prior}$$

then if  $\lambda = f(\theta)$  where  $f$  is monotone

$$\begin{aligned}
p(\lambda) &= p[f^{-1}(\lambda)] \left| \frac{f^{-1}(\lambda)}{d\lambda} \right| \\
&\propto \left| \frac{f^{-1}(\lambda)}{d\lambda} \right|
\end{aligned}$$

if the transformation is linear, the jacobian is a constant, so it follows that the density of  $\lambda$  is uniform as well.

If the transformation is nonlinear, the density varies with  $\lambda$ , this implies that if one claims

ignorance with respect to  $\theta$ , the same cannot be said about  $\lambda$ . This is a severe inconsistency.

Example: Improper uniform prior for normality case

$$\theta \sim N(\mu_\theta, \sigma_\theta^2).$$

if  $\sigma_\theta^2$  is very small. The values are concentrated around  $\mu_\theta$ .

If  $\sigma_\theta^2$  is large, this generate a flatter dist, and in the limit, it is a uniform distribution.

↓

improper prior (because the integrale is not finite)

However, the posterior is proper:

if  $y_i \sim N(\mu, \sigma^2)$  ( $i = 1, \dots, n$ ) to be iid with known variance.

If  $p(\mu) \propto \text{constant}$ . This generates an improper distribution, unless finite boundaries are assigned to the values of  $\mu$ .

But the posterior:

$$p(\mu|y_1, \dots, y_n) \propto \exp\left[-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right]$$

which integrates to  $\sqrt{2\pi\sigma^2/n}$ .

## Other vague priors

Although the uniform distribution is probably the most widely employed "vague" prior, other distributions that supposedly convey vague prior knowledge have been suggested.

**Example:** probability of success  $\theta$  in a Binomial distribution

$$\theta \sim Be(a, b)$$

is a prior and  $x$  is the number of successes after  $n$  independent trials  
when  $a = b = 0$ , this generates an improper prior

however:

$$p(\theta|n, x, a = 0, b = 0) \propto \theta^{x-1}(1 - \theta)^{n-x-1}$$

which is

$$Be(x, n - x)$$

Beta is proper when  $x > 0$  and  $(n - x) > 0$  so when  $x = 0$  or  $n = x$  then it is improper.

so an improper prior can lead to an improper posterior.

For example, when  $\theta$  is small, *post* will not be proper unless  $n$  is large.

## A single parameter: Jeffrey's prior

Jeffrey (1961) proposed a class of improper priors for representing the ignorance. He suggested that they are noninformative because:

1. If a parameter  $\theta$  can take any value in  $-\infty, +\infty$ , it should be uniformly distributed
2. If it takes values between 0 and  $+\infty$  then its log should have a uniform distribution

1-

$$p(-\infty < \theta < +\infty) = \int_{-\infty}^{+\infty} d\theta = \infty$$

so choosing any pair of interval is indeterminate.

If  $p(a < \theta < b)$  is determinate, where  $a$  and  $b$  are large, then this satisfy Jeffery's rule.

2- if  $\log(\theta) \sim \text{Uniform}$  then  $\theta \propto 1/\theta$  then

$$\int_0^{+\infty} \frac{1}{\theta} d\theta = \infty, \int_0^a \frac{1}{\theta} d\theta = \infty, \int_a^{\infty} \frac{1}{\theta} d\theta = \infty$$

which represents ignorance since ratio between the last two is indeterminate.

### Jeffrey's Prior:

Jeffery's prior: Invariance requirement

It means that:

if  $p(\theta)$  is the prior of  $\theta$  and  $\lambda = f(\theta)$ ,  $f$  is monotone, then

$$p(\theta)d\theta = p(\lambda)d\lambda$$

Jeffrey's idea for a prior density of  $\theta$ , is the square root of Fisher's information measure

$$\begin{aligned} p(\theta) &= \sqrt{I(\theta)} = \sqrt{E\left[\frac{dl(\theta|y)}{d\theta}\right]^2} \\ &= \sqrt{-E\left[\frac{d^2l(\theta|y)}{(d\theta)^2}\right]} \end{aligned}$$

where  $l(\theta|y)$  is the log likelihood function.

$$\begin{aligned} p(\lambda) &\propto \sqrt{E\left[\frac{dl(\theta|y)}{d\theta}\right]^2} \frac{d\theta}{d\lambda} \\ &\propto \sqrt{E\left[\frac{dl(f^{-1}(\lambda)|y)}{d\theta} \frac{d\theta}{d\lambda}\right]^2} \\ &\propto \sqrt{E\left[\frac{dl(f^{-1}(\lambda)|y)}{d\lambda}\right]^2} = \sqrt{I(\lambda)} \end{aligned}$$

since the likelihood is invariant under a change in parameterization.

$$\begin{aligned} p(\theta|y) &\propto L(\theta|y) \sqrt{I(\theta)} \\ &\propto L(\theta|y) \sqrt{E\left[\frac{dl(\theta|y)}{d\theta}\right]^2} \end{aligned}$$

and

$$p(\lambda|y) \propto L(\lambda|y) \sqrt{E\left[\frac{dl(f^{-1}(\lambda)|y)}{d\theta}\right]^2} \frac{d\theta}{d\lambda}$$

$$\propto L(\lambda|y) \sqrt{I(\lambda)}$$

**Example1:**

Normal distr with unknown mean  $\mu$

$$l(\mu|\sigma^2, y) = c - \frac{n(\bar{y} - \mu)^2}{2\sigma^2}$$

$$\left[\frac{dl}{d\mu}\right]^2 = \left[\frac{n(\bar{y} - \mu)}{\sigma^2}\right]^2$$

and

$$E\left(\left[\frac{dl}{d\mu}\right]^2\right) = I(\mu) = \frac{n}{\sigma^2}$$

Jeffrey's prior is

$$\sqrt{\frac{n}{\sigma^2}}$$

**Example 2:**

Normal dist with unknown variance  $\sigma^2$ :

$$l(\sigma^2|\mu, y) = c - \frac{n}{2} \log(\sigma^2) - \frac{\sum(y_i - \mu)^2}{2\sigma^2}$$

$$\left[\frac{dl}{d\sigma^2}\right]^2 = -\frac{n}{2\sigma^4} + \frac{\sum(y_i - \mu)^2}{\sigma^6}$$

and

$$E\left(\left[\frac{dl}{d\sigma^2}\right]^2\right) = I(\sigma^2) = \frac{n}{2\sigma^4}$$

then the jeffrey's prior for  $\sigma^2$  is  $\sqrt{\frac{n}{2\sigma^4}} \propto \frac{1}{\sigma^2}$

**Many parameters:**

$$p(\theta) = \sqrt{|\mathbf{I}(\theta)|}$$

**Example:**

consider the linear regression model under normality

Unknown parameter is  $\beta(p \times 1)$  and  $\sigma^2$

$$I(\theta) = \begin{bmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

then

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-\frac{p+2}{2}}$$

## Maximum Entropy Prior Distribution

The problem is to find  $p(\theta)$  that maximizes the entropy

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

Solution is complicated unless it is based on discretizing the prior distribution.

One can define a partition of  $(a, b)$  such that:

$$a = \theta_0 < \theta_1 < \dots < \theta_{K-1} < \theta_K = b$$

Define the event  $E_j \in (\theta_{j-1}, \theta_j)$  for  $j = 1, \dots, K$

and the discrete  $\theta^*$  with  $K$  mutually exclusive and exhaustive states and probability distribution

$$\begin{aligned} \Pr(\theta^* = E_j) &= \int_{\theta_{j-1}}^{\theta_j} p(\theta) d\theta \\ &= F(\theta_j) - F(\theta_{j-1}) = p_j \end{aligned}$$

then

$$H(\theta^*) = - \sum_{j=1}^K p_j \log \frac{p_j}{m_j}$$

to maximize  $H(\theta^*)$  with respect to  $p_j$  we calculate

$$\frac{\partial H(\theta^*)}{\partial p_j} = -1 - \log \frac{p_j}{m_j} - \lambda$$

and

$$\begin{aligned} p_j &= m_j \exp[-(1 + \lambda)] \\ \sum p_j &= \sum m_j \exp[-(1 + \lambda)] \end{aligned}$$

which gives:  $\exp[-(1 + \lambda)] = 1$  so that  $\lambda = -1$  and  $p_j = m_j$

hence in the limit  $p(\theta) = m(\theta)$  **the uniform distribution.**

## Reference prior

Bernardo (1979)

It is defined as any positive function  $h(\theta)$  such that

$$\pi(\theta|y) \propto p(y|\theta)\pi(\theta)$$

Using a detailed development, the solution to the above is  $h(\theta)$

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

where  $y$  is one measurement

### Example:

Suppose we have  $n$  Bernoulli trials with prob of success  $\theta$

The dist of a single observation  $y$  is:

$$p(y|\theta) \propto \theta^y(1-\theta)^{1-y}$$

The information measure from a single draw is

$$\begin{aligned} & E_{y|\theta} \left\{ -\frac{d^2}{(d\theta)^2} [y \log(\theta) + (1-y) \log(1-\theta)] \right\} \\ &= E_{y|\theta} \left[ \frac{y}{\theta^2} + \frac{(1-y)}{(1-\theta)^2} \right] \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

So the reference prior is

$$\pi(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}}$$