Introduction to Bayesian Methods - 4

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Prior distributions

The choice of prior distribution is an important aspect of Bayesian inference

- prior distributions are one of the main targets of frequentists: how much do posteriors differ when we choose different priors?
- there are two main “objective” methods for the choice of priors (MaxEnt and Jeffreys')
- here we discuss

1. The quest for "objective" priors
2. Review of the Cramer-Rao bound and related concepts
3. Information-theoretic concepts in statistics
4. Jeffreys' method
5. Reference priors
6. The Maximum Entropy Method

Random variable transformations and prior distributions

$$p_x(x)dx = p_x\left[x(y)\right]\left|\frac{dx}{dy}\right|dy = p_y(y)dy$$

$$\Rightarrow \quad p_y(y) = p_x\left[x(y)\right]\left|\frac{dx}{dy}\right|$$

- In general, if the first pdf is uniform, the other one is not. This means that choosing a uniform distribution as the "least informative" distribution is not enough, *unless we specify which variate should be uniformly distributed*.

- How can we "objectively" choose a prior distribution???

Review of the Cramer-Rao bound - proof of the Bartlett identities

- pdf normalization

$$\int_{\{x\}} p(x,\theta)dx = 1$$

- derivation of normalization formula

$$\frac{\partial}{\partial\theta}\int_{\{x\}} p(x,\theta)dx = \int_{\{x\}} \frac{\partial p(x,\theta)}{\partial\theta}dx = 0$$

- further manipulation of the previous result

$$\begin{aligned} 0 &= \int_{\{x\}} \frac{\partial p(x,\theta)}{\partial\theta}dx = \int_{\{x\}} \frac{1}{p(x,\theta)}\frac{\partial p(x,\theta)}{\partial\theta}p(x,\theta)dx \\ &= \int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial\theta}p(x,\theta)dx \\ &= \mathbb{E}\left[\frac{\partial \ln p(x,\theta)}{\partial\theta}\right] \end{aligned}$$

First Bartlett identity

Review of the Cramer-Rao bound - proof of the Bartlett identities (ctd.)

- derivation of the first identity

$$\int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx = 0$$

- further manipulation of the previous result

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta} \int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx = \int_{\{x\}} \left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} p(x,\theta) + \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} \right] dx \\
&= \int_{\{x\}} \left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} + \frac{1}{p(x,\theta)} \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} \right] p(x,\theta) dx \\
&= \int_{\{x\}} \left\{ \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} + \left[\frac{\partial \ln p(x,\theta)}{\partial \theta} \right]^2 \right\} p(x,\theta) dx \\
&= \mathbb{E}\left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} \right] + \mathbb{E}\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta} \right)^2 \right]
\end{aligned}$$

Second Bartlett identity

The Cramér-Rao bound and the Fisher information

- using both identities $$\mathrm{var}\left[\frac{\partial \ln L(x,\theta)}{\partial\theta}\right] = \mathbb{E}\left[\left(\frac{\partial \ln L(x,\theta)}{\partial\theta}\right)^2\right] = -\mathbb{E}\frac{\partial^2 \ln L(x,\theta)}{\partial\theta^2}$$

- expectation value of ML estimator $$\mathbb{E}[\hat{\theta}(D)] = \int_{\{x\}} \hat{\theta}(x) L(x,\theta) dx = \theta + b_n$$

- derivative of the previous expression

$$\frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}(D)] = 1 + \frac{\partial b_n}{\partial\theta} = \int_{\{x\}} \hat{\theta}(x)\frac{\partial L(x,\theta)}{\partial\theta} dx = \int_{\{x\}} \hat{\theta}(x)\frac{\partial \ln L(x,\theta)}{\partial\theta} L(x,\theta) dx$$

$$= \mathbb{E}\left[\hat{\theta}(D)\frac{\partial \ln L(D,\theta)}{\partial\theta}\right] = \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial\theta}\right] + \mathbb{E}\left[\hat{\theta}(D)\right]\mathbb{E}\left[\frac{\partial \ln L(D,\theta)}{\partial\theta}\right]$$

$$= \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial\theta}\right]$$

The Cramér-Rao bound and the Fisher information (ctd.)

- we use Schwartz's inequality for covariance $[\mathrm{cov}(x,y)]^2 \leq \sigma_x^2 \sigma_y^2$
- we apply the inequality to the previous result $1 + \frac{\partial b_n}{\partial \theta} = \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]$ and find

$$\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2 = \left\{\mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]\right\}^2 \leq \mathrm{var}[\hat{\theta}(D)]\mathrm{var}\left[\frac{\partial \ln L(D,\theta)}{\partial \theta}\right]$$

- rearranging terms, we obtain the **Cramér-Rao bound**

$$\mathrm{var}[\hat{\theta}(D)] \geq \frac{\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2}{\mathrm{var}\left[\frac{\partial \ln L(D,\theta)}{\partial \theta}\right]} = \frac{\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2}{\mathbb{E}\left[\left(\frac{\partial \ln L(D,\theta)}{\partial \theta}\right)^2\right]} = \frac{\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2}{-\mathbb{E}\frac{\partial^2 \ln L(D,\theta)}{\partial \theta^2}}$$

Definition of Fisher Information. A very concentrated pdf is very informative. Therefore, the smaller the variance, the greater the "information".

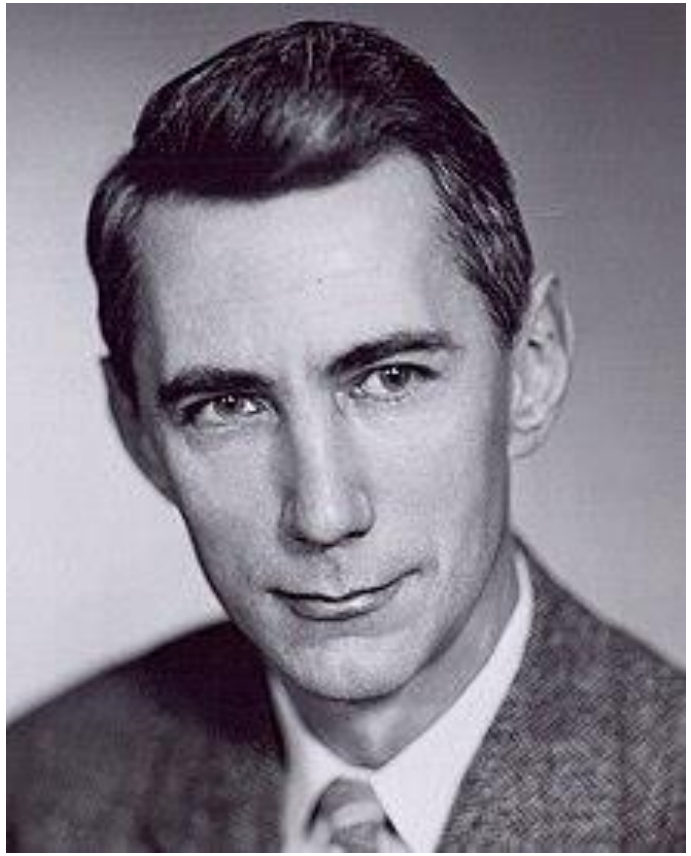
Thus, from the (unbiased, consistent) Cramér-Rao bound

$$\operatorname{var}[\hat{\theta}(D)] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln L(D,\theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbb{E}\frac{\partial^2 \ln L(D,\theta_0)}{\partial \theta_0^2}}$$

one is led to the Fisher Information

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2}$$

Information theoretic concepts in statistics



Claude Shannon
1916–2001

After graduating from Michigan and MIT, Shannon joined the AT&T Bell Telephone laboratories in 1941.

His paper ‘A Mathematical Theory of Communication’ published in the Bell System Technical Journal in 1948 laid the foundations for modern information theory. This paper introduced the word ‘bit’, and his concept that information could be sent as a stream of 1s and 0s paved the way for the communications revolution.

It is said that von Neumann recommended to Shannon that he use the term entropy, not only because of its similarity to the quantity used in physics, but also because “nobody knows what entropy really is, so in any discussion you will always have an advantage”.

Information theoretic concepts in statistics

- The amount of information can be viewed as a "degree of surprise" on learning the value of a variable *x*. If we are told that a highly improbable event has just occurred, we will have received more information than if we were told that some very likely event has just occurred, and if we knew that the event was certain to happen we would receive no information.

- Thus, information carried by an event (symbol) must be a function of its probability

- If two events (symbols) are independent the total information is the sum of the information carried by each of them, therefore

$$p(x)p(y) \quad \Rightarrow \quad h(x) + h(y)$$

- logarithms have this property, therefore we take

$$h(x) \propto \ln p(x)$$

- more specifically, we choose

$$h(x) = -\log_2 p(x)$$

Information theoretic concepts in statistics – 2

- definition of the **Shannon entropy: average information** carried by the events (symbols)

$$H = -\sum_{k=1}^{N} p_k \log_2 p_k = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k}$$

Example:

- just two symbols, 0 and 1, same probability

$$H = -2\left(\frac{1}{2}\log_2 \frac{1}{2}\right) = 1 \text{ bit}$$

there are 2 equal terms

average information conveyed by each symbol

the result is given in pseudounit "bits" (for natural logarithms this is "nats")

Information theoretic concepts in statistics – 2

- just two symbols, 0 and 1, probabilities ¼ and ¾ , respectively

$$H = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \approx 0.81 \text{ bit}$$

- 8 symbols, equal probabilities

$$H = -\sum_1^8 \frac{1}{8}\log_2\frac{1}{8} = \log_2 8 = 3 \text{ bit}$$

- 8 symbols, with probabilities ½, ¼, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64

$$\begin{aligned} H &= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + 4\times\frac{1}{64}\log_2\frac{1}{64}\right) \\ &= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + 4\times\frac{3}{32} = 2 \text{ bit} \end{aligned}$$

Information theoretic concepts in statistics – entropy in statistical mechanics

- consider a system where states *n* are occupied by N_n identical particles (*n, n=1, … , M*).

- the number of ways to fill these states is given by

$$\Omega = \frac{N!}{N_1!N_2!\dots N_M!}$$

- then Boltzmann's entropy is

$$H_B = k_B \ln \Omega = k_B \ln \frac{N!}{N_1!N_2!\dots N_M!} \approx k_B \left[(N \ln N - N) - \sum_i (N_i \ln N_i - N_i) \right]$$

$$= k_B \left[N \ln N - \sum_i N p_i (\ln p_i + \ln N) \right] = k_B N \sum_i p_i \ln \frac{1}{p_i}$$

Information theoretic concepts in statistics – additivity of entropy

If symbols are emitted simultaneously and independently by two sources, the joint probability distribution is

$$p(j,k) = p_1(j)p_2(k)$$

and therefore, the joint entropy is

$$\begin{aligned} H &= -\sum_{j,k} p(j,k)\log_2 p(j,k) = -\sum_{j,k} p_1(j)p_2(k)\log_2[p_1(j)p_2(k)] \\ &= -\sum_{j} p_1(j)\log_2 p_1(j) - \sum_{k} p_2(k)\log_2 p_2(k) \\ &= H_1 + H_2 \end{aligned}$$

Information theoretic concepts in statistics – the uniform distribution has maximal entropy

This is an easy result that follows using one Lagrange multiplier to keep probability normalization into account

$$H + \lambda \sum_{k=1}^{N} p_k = -\sum_{k=1}^{N} p_k \log_2 p_k + \lambda \sum_{k=1}^{N} p_k$$

$$= -\frac{1}{\ln 2} \sum_{k=1}^{N} p_k \ln p_k + \lambda \sum_{k=1}^{N} p_k$$

$$\frac{\partial}{\partial p_j}\left(H + \lambda \sum_{k=1}^{N} p_k\right) = -\frac{1}{\ln 2}(\ln p_j + 1) + \lambda = 0$$

$$p_j = \exp(\lambda \ln 2 - 1) = 1/N$$

all probabilities have the same value

Information theoretic concepts in statistics – differential entropy

The Shannon entropy cannot be extended to continuous distribution in a straightforward way. Consider a discretized version of the probability distribution:

$$P_k = \int_{k\Delta}^{(k+1)\Delta} p(x)dx = p(x_k^*)\Delta \quad , \text{where } x_k^* \in (k\Delta, (k+1)\Delta)$$

$$H = -\sum_k P_k \ln P_k$$

$$= -\sum_k p(x_k^*) \ln p(x_k^*)\Delta - \sum_k p(x_k^*)\Delta \ \ln \Delta \approx \underbrace{-\int p(x) \ln p(x)dx}_{\text{differential entropy}} - \ln \Delta$$

Information theoretic concepts in statistics – relative entropy

Considering two sets of symbols (same number of symbols), we can consider the relative information carried by each symbol in one set with respect to the corresponding one in the other set

$$(-\log_2 p_k) - (-\log_2 g_k) = -\log_2 \frac{p_k}{g_k}$$

Then, the average difference of the information carried by the p_k's with respect to the reference set (the **relative entropy**) is

$$H_R = -\sum_k p_k \log_2 \frac{p_k}{g_k}$$

This extends without problems to continuous distributions

$$H_R = -\int p(x) \log_2 \frac{p(x)}{g(x)} dx$$

Information theoretic concepts in statistics – the Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is a simple redefinition of the relative entropy

$$H_R = -\int p(x) \log_2 \frac{p(x)}{g(x)} dx \quad \Rightarrow \quad D_{\mathrm{KL}}(p||g) = \int p(x) \ln \frac{p(x)}{g(x)} dx$$

- Natural logs instead of logs base 2
- Change of sign
- NOT symmetrical with respect to the p,g exchange
- Has several interesting properties

Information theoretic concepts in statistics – Jensen's inequality

Consider a convex function, then

$$f(a+(b-a)t) \leq f(a) + [f(b) - f(a)]t$$

$$\Rightarrow \quad f[a(1-t)+bt] \leq f(a)(1-t) + f(b)t$$

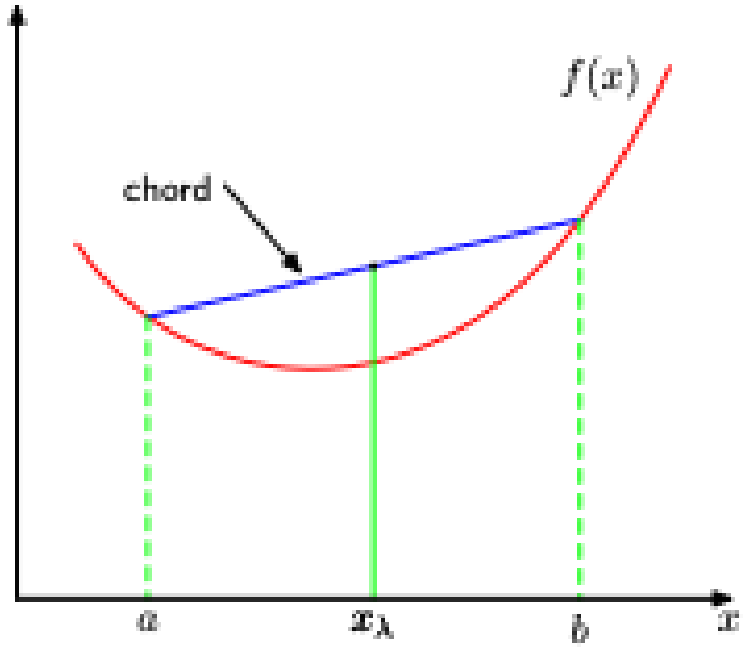
which can be written as

$$f[at_1 + bt_2] \leq f(a)t_1 + f(b)t_2 \quad \text{with } t_1 + t_2 = 1$$

Now, we conjecture the extension

$$f\left(\sum_{k=1}^{n} x_k t_k\right) \leq \sum_{k=1}^{n} f(x_k)t_k \quad \text{with } \sum_{k=1}^{n} t_k = 1$$

and prove the inequality by induction.



A convex function *f*(*x*) is one for which every chord (shown in blue) lies on or above the function (shown in red)

Information theoretic concepts in statistics – Jensen's inequality – 2

If

$$f\left(\sum_{k=1}^{n+1} x_k t_k\right) \leq \sum_{k=1}^{n+1} f(x_k) t_k \quad \text{with} \quad \sum_{k=1}^{n+1} t_k = 1$$

then isolate the (n+1)-th parameter

$$\sum_{k=1}^{n} t_k = 1 - t_{n+1}, \quad \text{so that} \quad \sum_{k=1}^{n} \frac{t_k}{1 - t_{n+1}} = 1$$

and rearrange the l.h.s. of the inequality

$$\begin{aligned} f\left(\sum_{k=1}^{n+1} x_k t_k\right) &= f\left((1 - t_{n+1}) \sum_{k=1}^{n} x_k \frac{t_k}{1 - t_{n+1}} + x_{n+1} t_{n+1}\right) \\ &\leq f\left(\sum_{k=1}^{n} x_k \frac{t_k}{1 - t_{n+1}}\right)(1 - t_{n+1}) + f(x_{n+1}) t_{n+1} \\ &\leq \sum_{k=1}^{n} f(x_k) \frac{t_k}{1 - t_{n+1}}(1 - t_{n+1}) + f(x_{n+1}) t_{n+1} = \sum_{k=1}^{n+1} f(x_k) t_k \end{aligned}$$

Information theoretic concepts in statistics – the Kullback-Leibler divergence - 2

Jensen's inequality can be restated in a simple way if the t's are mapped into probabilities

$$f\left(\sum_{k=1}^{n} x_k p_k\right) \leq \sum_{k=1}^{n} f(x_k) p_k \quad \Rightarrow \quad f\left[\mathbb{E}(x)\right] \leq \mathbb{E}[f(x)]$$

Equivalently

$$f\left(\int x p(x) dx\right) \leq \int f(x) p(x) dx$$

Now we can apply the inequality to the KL divergence (the -log function is convex) and find

$$D_{KL}(p||g) = -\int p(x) \ln \frac{g(x)}{p(x)} dx \geq -\ln\left(\int g(x) dx\right) = 0$$

Information theoretic concepts in statistics – The KL divergence is a quasi-metric (however a local version of the KL divergence is the Fisher information, which is a true metric)

The KL divergence can be used to measure the "distance" between two distributions.

Example: the KL divergence

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

for the distributions

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

$$q(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$D_{\mathrm{KL}}(p||q) = \frac{\mu^2}{2\sigma^2}$$

Now consider a family of parametric distributions and evaluate the KL divergence between two close elements of the family

$$D_{\mathrm{KL}}\left[p(x,\theta)||p(x,\theta+\epsilon)\right] = \int_{-\infty}^{+\infty} p(x,\theta)\ln\frac{p(x,\theta)}{p(x,\theta+\epsilon)}dx$$
$$= \mathbb{E}\left[\ln p(x,\theta) - \ln p(x,\theta+\epsilon)\right]$$

Since

$$\ln p(x,\theta+\epsilon) \approx \ln p(x,\theta) + \frac{\partial \ln p(x,\theta)}{\partial\theta}\epsilon + \frac{1}{2}\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\epsilon^2$$

we find, using the first Bartlett identity,

$$D_{\mathrm{KL}}\left[p(x,\theta)||p(x,\theta+\epsilon)\right] = -\mathbb{E}\left(\frac{\partial \ln p(x,\theta)}{\partial\theta}\epsilon + \frac{1}{2}\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\epsilon^2\right)$$
$$= -\frac{1}{2}\mathbb{E}\left[\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

i.e., locally the KL divergence is just the Fisher information

Information theoretic concepts in statistics – The KL divergence can be transformed into a true distance between pdf's

- Jeffreys' distance

$$D_{\mathrm{J}}(p||q) = \frac{1}{2}D_{\mathrm{KL}}(p||q) + \frac{1}{2}D_{\mathrm{KL}}(q||p)$$

- Jensen-Shannon distance

$$D_{\mathrm{JS}}(p||q) = \frac{1}{2}D_{\mathrm{KL}}\left(p||\frac{p+q}{2}\right) + \frac{1}{2}D_{\mathrm{KL}}\left(q||\frac{p+q}{2}\right)$$

Information theoretic concepts in statistics – Using the KL divergence

Suppose that data is being generated from an unknown distribution $p(x)$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(x|\boldsymbol{\theta})$, governed by a set of adjustable parameters $\boldsymbol{\theta}$, for example, a multivariate Gaussian.

One way to determine $\boldsymbol{\theta}$ is to minimize the Kullback-Leibler divergence between $p(x)$ and $q(x|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We cannot do this directly because we don't know $p(x)$. Suppose, however, that we have observed a finite set of training points x_n, for $n = 1, \ldots, N$, drawn from $p(x)$ (an **empirical distribution**). Then the expectation with respect to $p(x)$ can be approximated by a finite sum over these points

$$D_{KL}(p||q) = \mathbb{E}_p\left[\ln \frac{p(x)}{q(x|\boldsymbol{\theta})}\right] \approx \frac{1}{N}\sum_{n=1}^{N}\left[-\ln q(x_n|\boldsymbol{\theta}) + \ln p(x_n)\right]$$

likelihood

parameter-independent term

From this equation we see that we can obtain an approximate distribution by minimizing the KL divergence, i.e., by maximizing the likelihood.

Information theoretic concepts in statistics – Mutual information

We can use the KL divergence to measure the degree of statistical dependence between pairs of variates, by measuring the distance between $p(\mathbf{x},\mathbf{y})$ and $p(\mathbf{x})p(\mathbf{y})$

$$I(\mathbf{x},\mathbf{y}) = D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})] = \int p(\mathbf{x},\mathbf{y}) \ln \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x}d\mathbf{y}$$

This quantity is called the **mutual information**.

We also find,

$$\begin{aligned} I(\mathbf{x},\mathbf{y}) &= -\int p(\mathbf{x},\mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})} d\mathbf{x}d\mathbf{y} \\ &= -\int p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})} d\mathbf{x}d\mathbf{y} \\ &= -\int p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \left[\ln p(\mathbf{x}) - \ln p(\mathbf{x}|\mathbf{y})\right] d\mathbf{x}d\mathbf{y} \\ &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} + \int p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x}d\mathbf{y} \\ &= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] \\ &= H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}] \end{aligned}$$

differential entropy

conditional entropy

we can view the mutual information as the reduction in the uncertainty about **x** by virtue of being told the value of **y** (or vice versa).

Jeffreys' priors

An invariant form for the prior probability in estimation problems

By Harold Jeffreys, F.R.S.

(*Received* 23 *November* 1945)

It is shown that a certain differential form depending on the values of the parameters in a law of chance is invariant for all transformations of the parameters when the law is differen-

Jeffreys' priors – the KL divergence is invariant with respect to generic random variable transformations.

From the definition of KL divergence, and from the transformation formula for pdf's we find

$$\int_{-\infty}^{+\infty} p_y(y) \ln\left(\frac{p_y(y)}{q_y(y)}\right) dy = \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)\left|\frac{dx}{dy}\right|}{q_x(x)\left|\frac{dx}{dy}\right|}\right) dx$$

$$= \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)}{q_x(x)}\right) dx$$

In this case, our random variables are the parameter estimates, therefore the KL divergence is invariant with respect to parameter (random variable transformations), **therefore the associated Fisher Information from the local expansion of the KL divergence is also invariant with respect to parameter transformations**.

From the equation that relates KL divergence and Fisher Information, we find a corresponding pdf as follows. Equation

$$D_{\mathrm{KL}}\left[p(x|\theta)||p(x|\theta+\epsilon)\right] = -\frac{1}{2}\mathbb{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

means that the KL divergence depends quadratically on small changes of the expansion parameter and that the KL divergence remains constant if the term on the r.h.s. remains constant.

Dimensionally, the Fisher information is quadratic with respect to a pdf, therefore we take its square root to define a pdf, i.e.,

$$f(\theta) \sim \sqrt{I(\theta)}$$

This must be normalized to obtain a pdf that is invariant with respect to parameter transformations.

Example: a simple Gaussian Likelihood for *n* datapoints, with known variance

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$\ln L(D|\mu) \sim \sum_n \left(-\ln\sigma - \frac{(x_n-\mu)^2}{2\sigma^2}\right)$$ fixed sigma

$$I(\mu) = \mathbb{E}\left[-\frac{\partial^2 \ln L(D|\mu)}{\partial\mu^2}\right] \sim \text{constant}$$

This points to a uniform prior for μ. In general, this uniform prior is an improper prior.

Example: a simple Gaussian Likelihood for *n* datapoints, with known mean

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$I(\sigma) = \mathbb{E}\left[-\frac{\partial^2 \ln L(D|\sigma)}{\partial\sigma^2}\right] \sim \frac{1}{\sigma^2}$$

fixed mu

$$\sqrt{I(\sigma)} \sim \frac{1}{\sigma}$$

This power-law pdf is another improper prior.

Example: Poisson distribution

$$L(D|a) = \prod_n \frac{a^{k_n}}{k_n!} e^{-a}$$

$$I(a) = \mathbb{E}\left[-\frac{\partial^2 \ln L(D|a)}{\partial a^2}\right] \sim \frac{1}{a}$$

$$\sqrt{I(a)} \sim \frac{1}{\sqrt{a}}$$

This power-law pdf is yet another improper prior.

Example: binomial distribution

$$L(D|\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}$$

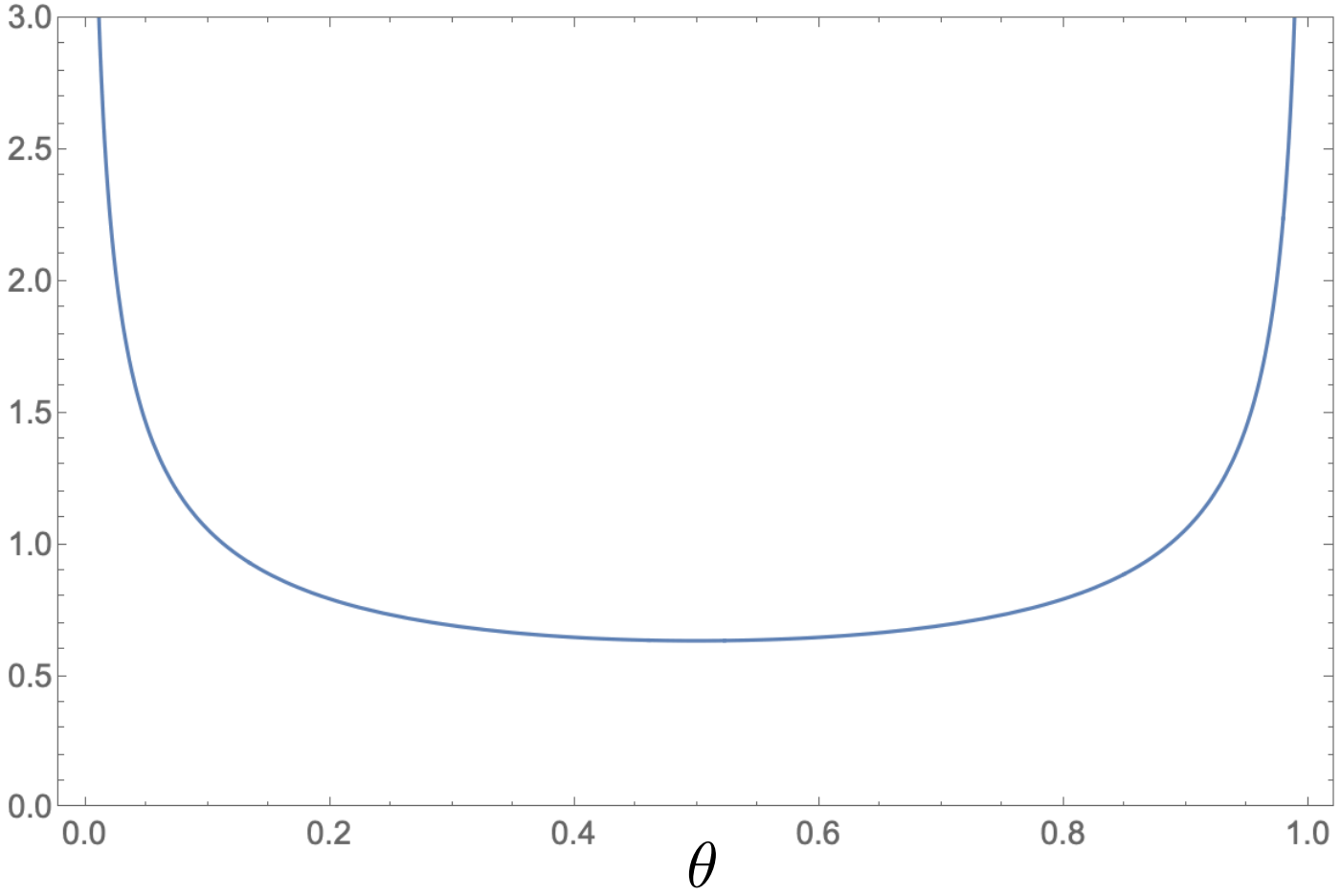
$$\ln L(D|\theta) \sim n \ln \theta + (N-n) \ln(1-\theta)$$

$$\mathbb{E}\left[-\frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right] \sim \frac{N\theta}{\theta^2} + \frac{N - N\theta}{(1-\theta)^2}$$

$$= \frac{N}{\theta} + \frac{N}{1-\theta}$$

$$= \frac{N}{\theta(1-\theta)}$$

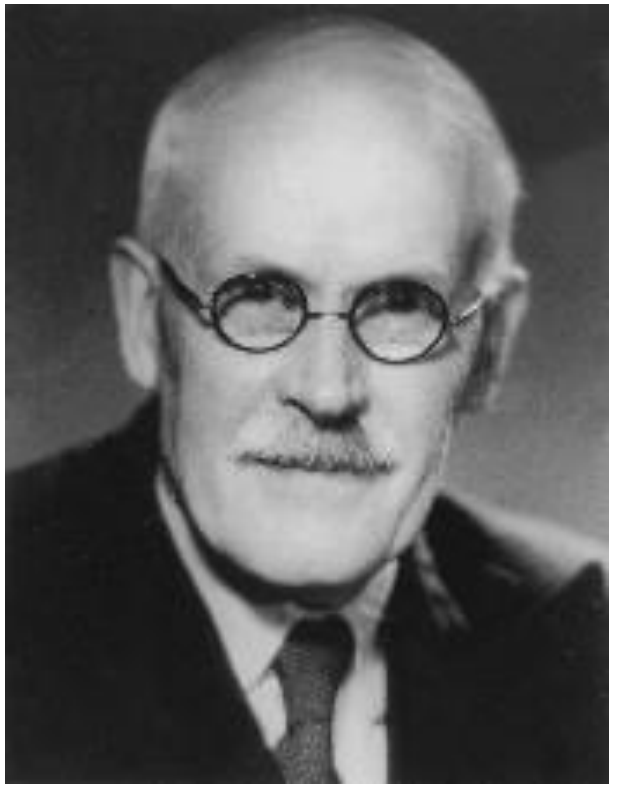
$$I(\theta) \sim \frac{1}{\theta(1-\theta)} \quad \Rightarrow \quad \sqrt{I(\theta)} \sim \frac{\theta^{1/2}(1-\theta)^{1/2}}{B(1/2,1/2)} = \text{Beta}(\theta|1/2,1/2)$$



A lesson learned from Jeffreys' priors

Jeffreys priors are tuned to the Likelihood, but doesn't this sound strange? Shouldn't the prior distribution be related to the prior information alone?

Well ... no, the Likelihood is also constructed using prior information (obviously!). So, in this approach the Likelihood and the priors are both determined using the available prior information.



Harold Jeffreys
(1891-1989)

Additional comments on Jeffreys' priors

- In general, they are NOT conjugate priors, but are limits of conjugate priors
- They work well for single parameter models, but NOT for multivariate models