

4

What is information?

We describe various types of information inherent in probabilistic models, many of which play an important role in statistical inference.

4-1 Overview

There are several notions of information related to random variables. While quite different all of them are closely related (**Figure 4-1**).

The most fundamental and most important is the Kullback-Leibler (KL) information, or equivalently, the Boltzmann entropy. It measures the similarity of two probabilistic models. The Shannon entropy and Fisher's information matrix are further information quantities, both of which can be directly derived from Kullback-Leibler information. We also discuss mutual information as an important application of Kullback-Leibler information to quantify independence.

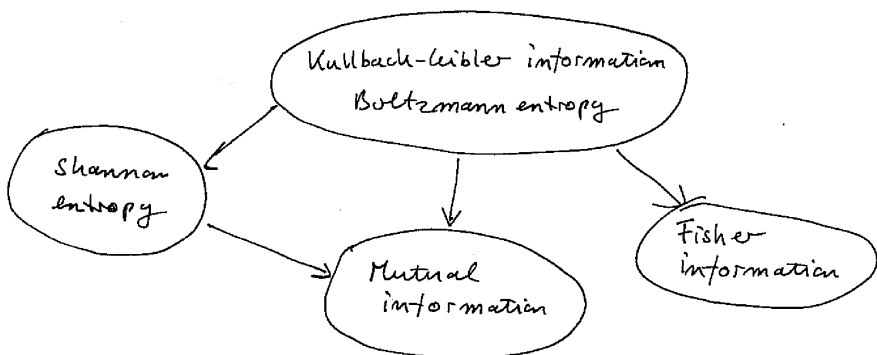


Figure 4-1: Measures of information discussed in this chapter, and their relationships.

4–2 Kullback-Leibler information

The Kullback-Leibler (KL) information is a fundamental quantity in probability and statistics that measures the similarity of two distributions. It is hard to overstate the importance of KL information: many statistical procedures for inference use KL information either directly or indirectly, including the likelihood and penalized likelihood approaches. KL information was formally introduced in the 1950s but the same quantity was already discovered almost a century earlier in the 1870s by the Ludwig Boltzmann, a physicist famous for developing the stochastic underpinnings of thermodynamics.

Suppose we have two models defined on the same sample space, one with distribution F and a second one with distribution G . Then the expectation

$$I^{\text{KL}}(F; G) := E_F \log \left(\frac{f(X)}{g(X)} \right) \quad \text{Equation 4–1}$$

is the **Kullback-Leibler information** of G with regard to F . Applying **Equation 2–2** with $h(X) = \log\{f(X)/g(X)\}$ we get to the formulas

$$I^{\text{KL}}(F; G) = \begin{cases} \sum_{i=1}^d f(x_i) \log \left(\frac{f(x_i)}{g(x_i)} \right) & \text{for a discrete variable, and} \\ \int_{-\infty}^{\infty} f(x) \log \left(\frac{f(x)}{g(x)} \right) dx & \text{for a continuous variable.} \end{cases}$$

With a negative sign attached KL information becomes the **Boltzmann entropy**

$$B(F; G) = -I^{\text{KL}}(F; G).$$

Mathematically, I^{KL} belongs to the class of f -divergences, and a further common name it is **relative entropy**.

The KL information has the following properties:

- (i) It is always non-negative: $I^{\text{KL}}(F; G) \geq 0$.
- (ii) $I^{\text{KL}} = 0$ *only* if the two models F and G are identical.
- (iii) The KL divergence is *not* symmetric: $I^{\text{KL}}(F; G) \neq I^{\text{KL}}(G; F)$.
- (iv) I^{KL} is invariant against transformations of the sample space. If instead of X the random variable $Y = h(X)$ is considered, where h is an invertible function, and the distributions G and F are transformed accordingly (cf. **Equation 2–7**), the Kullback-Leibler information remains the same. Thus, I^{KL} is a **geometric quantity** that is independent of the choice of a specific coordinate system.

The KL information is a measure of the difference between two distributions, i.e. the larger I^{KL} the further apart are the two models. However, because of the lack of symmetry in the two arguments it is not a metric. Therefore, the term KL *distance* must be avoided and the expression KL *divergence* should be used instead. A symmetric version of I^{KL} is given by the **Jeffreys information** $J(F; G) = I^{\text{KL}}(F; G) + I^{\text{KL}}(G; F)$.

The asymmetry in the arguments of I^{KL} is not coincidental — the two models F and G play very different roles. In particular, the expectation in **Equation 4–1** is with respect only to F , and not to G . In typical application of I^{KL} the model F is considered to be the true model and G the approximative model. In this setting the KL divergence measures, loosely speaking, the *information lost* by using candidate model G rather than the correct model F . Note that the minus sign in the definition of the Boltzmann entropy explicitly accounts for the fact that there is always a loss when using a model different from the true model. An alternative, probabilistic interpretation of the KL divergence is given in the next section where **Equation 4–1** is derived from a combinatorial point of view.

Example 4–1: KL divergence between two normal distributions

Consider two univariate normal models F_0 and F with means μ_0 and μ and variances σ_0^2 and σ^2 . After some calculation we get for the KL information

$$I^{\text{KL}}(F_0; F) = \frac{1}{2} \left\{ \frac{(\mu - \mu_0)^2}{\sigma^2} + \frac{\sigma_0^2}{\sigma^2} - \log \left(\frac{\sigma_0^2}{\sigma^2} \right) - 1 \right\}. \quad \text{Equation 4–2}$$

This expression is invariant against changes of scale and location — but not against other more general coordinate transformations. This not in contradiction with the geometric property of I^{KL} because general changes of units will lead outside the normal family, and thus the above formula is not appropriate.

For multivariate normal F_0 and F of dimension d with means μ_0 and μ and covariances Σ_0 and Σ the KL information is given by

$$I^{\text{KL}}(F_0; F) = \frac{1}{2} \left\{ (\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) + \text{tr} \left(\Sigma^{-1} \Sigma_0 \right) - \log \det \left(\Sigma^{-1} \Sigma_0 \right) - d \right\},$$

which for $d = 1$ reduces to **Equation 4–2**. For identical covariances ($\Sigma = \Sigma_0$) the KL information becomes symmetric and proportional to $(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0)$, which is known as the **squared Mahalanobis distance**.

Example 4–2: KL divergence between two exponential distributions

The KL divergence from an exponential distribution with mean μ_0 to an approximating exponential with mean μ has a similar form as in the normal example. With density $f(x|\mu) = \frac{1}{\mu} e^{-x/\mu}$ we get

$$I^{\text{KL}}(F_0; F) = \frac{\mu_0}{\mu} - \log \left(\frac{\mu_0}{\mu} \right) - 1.$$

Example 4-3: Pearson's chi-squared statistic as approximation to KL divergence

Applying a Taylor series expansion of p around q we can approximate

$$p \log \frac{p}{q} \approx (p - q) + \frac{(p - q)^2}{2q}.$$

With this in mind, twice the KL divergence between two discrete distributions P and Q with dimension d and probability masses p_i and q_i is

$$G^2 = 2 I^{\text{KL}}(P; Q) = 2 \sum_{i=1}^d p_i \log \left(\frac{p_i}{q_i} \right) \approx \sum_{i=1}^d \frac{(p_i - q_i)^2}{q_i} = \chi^2.$$

Thus, Pearson's chi-squared goodness of fit statistic between an observed P and an expected Q may be viewed as approximation to (twice) the KL divergence between P and Q . The quantity $G^2 = 2 I^{\text{KL}}(P; Q)$ is also known as the **G-squared statistic**, or likelihood ratio chi-squared statistic.

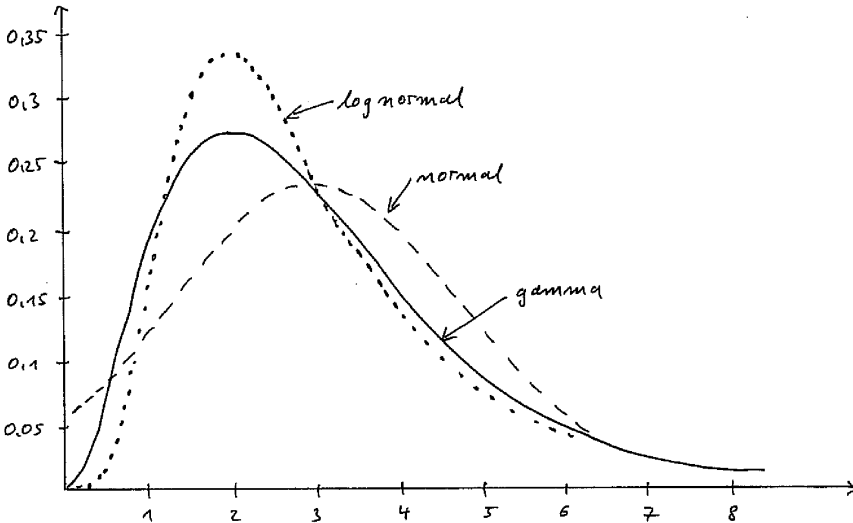
Example 4-4: Approximation of a gamma distribution

Figure 4-2: A gamma density and two approximating models (log-normal and normal).

We would like to approximate a gamma distribution with shape parameter $\alpha = 3$ and scale parameter $\beta = 1$ by either a log-normal distribution or a normal distribution with matching mean (3) and variance (3) — see **Figure 4-2**. Which of the two candidates approximates the gamma model best?

For the answer we compute the KL information between the gamma and the two approximating models by numerical integration and obtain $I^{\text{KL}} = 0.0594$ for the matching

log-normal and 0.1207 for the normal distribution. This confirms what is evident already by visual inspection: the KL number to the log-normal model is smaller and thus it is a better approximation of the gamma model than the normal distribution.

4–3 Boltzmann's derivation and probabilistic interpretation

The definition of the Kullback-Leibler divergence in **Equation 4–1**) as an expectation of log-densities can be derived and better understood by a simple argument which goes back Ludwig Boltzmann.

We consider a simple urn experiment where we randomly allocate n elements to m different compartments (in the original setting these correspond to gas particles and energy levels). The assignment of a single element to one of the urns is governed by a given distribution G , with probability mass function $g_i = g(x_i)$. The probability to see a specific allocation n_1, \dots, n_m with $n = \sum_{i=1}^m n_i$ is given by multinomial probability

$$\Psi = \Pr(n_1, \dots, n_m | g_1, \dots, g_m) = \frac{n!}{\prod_{i=1}^m n_i!} \prod_{i=1}^m g_i^{n_i}.$$

If we instead of probability look at log-probability

$$\log \Psi = \log(n!) - \sum_{i=1}^m \log(n_i!) + \sum_{i=1}^m n_i \log(g_i).$$

For large n and n_i the Sterling approximation $\log(n!) \approx n \log(n) - n$ simplifies the terms containing factorial functions, so that we get

$$\begin{aligned} \log \Psi &\stackrel{\text{large sample}}{=} - \sum_{i=1}^m n_i \log\left(\frac{n_i}{n}\right) + \sum_{i=1}^m n_i \log(g_i) \\ &= -n \sum_{i=1}^m \frac{n_i}{n} \log\left(\frac{n_i/n}{g_i}\right). \end{aligned}$$

For large n the observed frequencies n_i/n approach a probability mass function $f_i = f(x_i)$, so that finally we get the asymptotic identity

$$\frac{1}{n} \log \Psi \stackrel{\text{large sample}}{=} B(F; G) = -I^{\text{KL}}(F; G). \quad \text{Equation 4–3}$$

In essence, the KL information is the negative log probability per data point to observe the correct model F under the specified model G . This is really quite a remarkable result, in particular as it exhibits a clear Bayesian flavor: whenever we use I^{KL} we implicitly evaluate the probability of the true model!

4–4 Expected Fisher information

The **expected Fisher information matrix** I^{Fisher} , which plays a very important role in likelihood and Bayes theory, is a local version of KL information.

Let's consider a family of distributions F_{θ} indexed by the parameter vector $\theta = (\theta_1, \dots, \theta_d)^T$ of dimension d . For a specific θ we compare the distribution F_{θ} with its neighbor $F_{\theta+\varepsilon}$ obtained by shifting the parameter vector θ by some small amount ε . The KL divergence between F_{θ} and $F_{\theta+\varepsilon}$ is

$$I^{\text{KL}}(F_{\theta}; F_{\theta+\varepsilon}) = E_{F_{\theta}} \left\{ \log f(X|\theta) - \log f(X|\theta + \varepsilon) \right\}.$$

By quadratic approximation of the last term in the brackets as $\log f(X|\theta + \varepsilon) \approx \log f(X|\theta) + \nabla \log f(X|\theta)^T \varepsilon + \frac{1}{2} \varepsilon^T \nabla^T \nabla \log f(X|\theta) \varepsilon$ and taking into account that the expectation of $\nabla \log f(X|\theta)$ vanishes (see Problems) we get

$$\begin{aligned} I^{\text{KL}}(F_{\theta}; F_{\theta+\varepsilon}) &\approx E_{F_{\theta}} \left\{ -\frac{1}{2} \varepsilon^T \nabla^T \nabla \log f(X|\theta) \varepsilon \right\} \\ &= \frac{1}{2} \varepsilon^T I^{\text{Fisher}}(\theta) \varepsilon \end{aligned} \quad \text{Equation 4–4}$$

with the $d \times d$ matrix $I^{\text{Fisher}}(\theta)$ given by

$$I^{\text{Fisher}}(\theta) = -E_{F_{\theta}} \left\{ \nabla^T \nabla \log f(X|\theta) \right\}. \quad \text{Equation 4–5}$$

The expected Fisher information $I^{\text{Fisher}}(\theta)$ is, unlike KL divergence, symmetric and provides a local metric in the space of distributions indexed by the parameter vector θ . In a later chapter we will discuss further geometrical properties of the Fisher information, e.g., it is formally a **metric tensor** describing the manifold of models.

As a local version of KL information I^{Fisher} is invariant against a change of coordinates in the sample space. However, if the index θ is changed to $\phi(\theta)$ then expected Fisher information transforms according to

$$I_{\phi}^{\text{Fisher}}(\phi) = J_{\theta}(\phi)^T I_{\theta}^{\text{Fisher}}(\theta(\phi)) J_{\theta}(\phi), \quad \text{Equation 4–6}$$

where $J_{\theta}(\phi)$ is the Jacobi matrix containing the partial derivatives of $\theta(\phi)$.

Example 4–5: Expected Fisher information of the normal model

We consider a normal distribution with mean μ and variance σ^2 . The expected Fisher information matrix may be directly computed via **Equation 4–5** but this is a bit tedious. A much faster route consists of reusing the result from **Example 4–1**. Using **Equation 4–2**

we compute the KL divergence of $N(\mu, \sigma^2)$ from $N(\mu + \varepsilon_1, \sigma^2 + \varepsilon_2)$. Approximating the logarithm $\log(1+x) \approx x - x^2/2$ we obtain

$$I^{\text{KL}}(N(\mu + \varepsilon_1, \sigma^2 + \varepsilon_2); N(\mu, \sigma^2)) \approx \frac{1}{2} \left\{ \frac{\varepsilon_1^2}{\sigma^2} + \frac{\varepsilon_2^2}{2\sigma^4} \right\}$$

so that by comparison with **Equation 4–4** we finally get

$$\mathbf{I}^{\text{Fisher}}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

This result is quite interesting: for the location parameter μ we get $I^{\text{Fisher}}(\mu) = \frac{1}{\sigma^2}$, i.e. the expected Fisher information is a constant independent of μ , and this constant equals the inverse variance, i.e. the precision. In contrast, for the parameter σ^2 the Fisher information $I^{\text{Fisher}}(\sigma^2) = \frac{1}{2\sigma^4}$ does depend on the parameter σ^2 .

Example 4–6: Expected Fisher information for the standard deviation

In the previous example, we computed the expected Fisher information of the variance, $I^{\text{Fisher}}(\sigma^2)$. What happens if we reparameterize and consider instead of σ^2 the standard deviation σ as our primary parameter?

From **Equation 4–2** we get $I^{\text{KL}}(N(\mu, (\sigma + \varepsilon)^2); N(\mu, \sigma^2)) \approx \frac{1}{2}\varepsilon(\frac{2}{\sigma^2})\varepsilon$ which implies $I^{\text{Fisher}}(\sigma) = \frac{2}{\sigma^2}$. Note that $I^{\text{Fisher}}(\sigma)$ differs from $I^{\text{Fisher}}(\sigma^2) = \frac{1}{2\sigma^4}$.

The same result can also be directly obtained by applying the transformation rule of **Equation 4–6**. With $\theta = (\mu, \sigma^2)^T$, $\phi = (\mu, \sigma)^T$, and $\mathbf{J}_\theta(\phi) = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix}$ we get

$$\mathbf{I}^{\text{Fisher}}(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

This result generalizes to arbitrary location-scale families (see Problems): a location parameter always has constant Fisher information, and for any scale parameter σ the Fisher information is proportional to $1/\sigma^2$.

Example 4–7: Expected Fisher information in a collection of iid random variables

If we consider a set of n independent and identically distributed (iid) random variables $\mathbf{X} = (X_1, \dots, X_n)$ what is the total expected Fisher information in \mathbf{X} ?

For independent random variables the joint density $f(\mathbf{x}|\theta) = f(x_i|\theta)^n$ is the product of the individual densities. Plugged into **Equation 4–5** this results in $nI^{\text{Fisher}}(\theta)$ where $I^{\text{Fisher}}(\theta)$ is the information for θ in any of the components X_i . Thus, the total Fisher information is the sum of the Fisher information in each individual variable.

4–5 Mutual information

A further important and very useful quantity derived from KL information is the **mutual information** between two random variables.

Mutual information is a measure of statistical dependencies. Two variables X_1 and X_2 are said to be stochastically independent if the joint density $f(x_1, x_2)$ factorizes into the product of the marginal densities $g_1(x_1)$ and $g_2(x_2)$. The mutual information is the KL information from the joint density to the product density,

$$\text{MI}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) \log \left(\frac{f(x_1, x_2)}{g_1(x_1) g_2(x_2)} \right) dx_1 dx_2. \quad \text{Equation 4–7}$$

Mutual information inherits all favorable properties of KL information including transformation invariance. It is applicable to both discrete and continuous variables. Furthermore, by construction mutual information is symmetric, always non-negative and becomes zero only if the variables are independent.

Example 4–8: Mutual information between two normal distributed variables

Consider two variables X_1 and X_2 following a joint normal distribution with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 and correlation ρ . A bit of calculation shows that the mutual information between X_1 and X_2 equals

$$\text{MI}(X_1, X_2) = -\frac{1}{2} \log(1 - \rho^2).$$

Hence, for normal variables mutual information is a function of the correlation. If the correlation vanishes ($\rho = 0$) then we have $\text{MI} = 0$. Conversely, for perfect correlation ($\rho = 1$) the mutual information will be infinite. Intuitively, this makes perfect sense, as correlation is a measure of how much information is contained in one random variable for the other. However, it is also important to keep in mind that the normal model is very special — for non-normal random variables zero correlation does in general *not* imply independence (see also **Example 2–2**).

Example 4–9: Mutual information between two categorical variables

The mutual information between two discrete variables X_1 and X_2 with dimensions n_1 and n_2 is straightforward to compute. If the marginal mass functions are denoted by $\pi_i^{(1)}$ and $\pi_j^{(2)}$ with $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$, and the joint probability mass function is given by the $n_1 \times n_2$ table π_{ij} the mutual information is

$$\text{MI}(X_1, X_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \pi_{ij} \log \frac{\pi_{ij}}{\pi_i^{(1)} \pi_j^{(2)}}$$

which can be approximated using a Taylor series (see **Example 4-3**) as

$$\text{MI}(X_1, X_2) \approx \frac{1}{2} \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\pi_{ij} - \pi_i^{(1)} \pi_j^{(2)} \right)^2}{\pi_i^{(1)} \pi_j^{(2)}} = \frac{1}{2} \chi^2,$$

yielding (half) the Pearson's chi-squared statistic of independence.

4-6 Shannon entropy

Probably the most widely known information criterion is the **Shannon entropy**. For a discrete random variable with distribution F the Shannon entropy is defined by

$$H(F) = - \sum_{i=1}^m f(x_i) \log f(x_i). \quad \text{Equation 4-8}$$

This expression looks very similar to the Boltzmann entropy $B(F, G)$, and indeed for G uniform with $g(x_i) = 1/m$ there is the relationship

$$H(F) = B(F, G) + \log m.$$

Therefore, the Shannon entropy simply measures the difference of the uniform distribution and the distribution of the random variable of interest. The Shannon entropy $H(F)$ is always greater than or equal to zero, and it reaches its maximum value ($\log m$) if F is itself the uniform distribution.

Shannon entropy also has a close link to mutual information. For discrete variables mutual information can be written in terms of the Shannon entropies of the marginal and joint distributions,

$$\text{MI}(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2).$$

Thus, mutual information is the difference of the joint entropy and the sum of the two marginal entropies.

Surprisingly, it is rather difficult to generalize Shannon entropy from a discrete to a continuous random variable. Replacing the summation in **Equation 4-8** by an integral leads to **differential entropy**

$$H(F) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

Unfortunately, differential entropy is not a fully satisfying extension of Shannon entropy, as it exhibits a number of unfavorable defects. For instance, it can take on negative as well as positive values. Moreover, differential entropy is not invariant under coordinate transformations, so unlike the KL information its value is dependent on the choice of units.

These issues can be traced back to the divergence of the above term $\log m$ in the process of going from a discrete to continuous variable, or equivalently, to the problem of generalizing the discrete uniform distribution G to the continuous case. Hence, we recommend using Shannon entropy only in its original discrete domain, and otherwise to employ KL information.

Example 4–10: Negative differential entropy

A simple example to demonstrate that differential entropy behaves differently than the Shannon entropy is the uniform distribution $U(0, a)$. Its density within its support is $f(x) = \frac{1}{a}$ so that $H = - \int_0^a \frac{1}{a} \log(\frac{1}{a}) dx = \log a$. Thus, for $0 < a < 1$ the differential entropy of $U(0, a)$ is negative.

4–7 Bibliographic notes

Shannon entropy is introduced in Shannon (1948) and Kullback-Leibler information in Kullback and Leibler (1951). Akaike (1985) traces the origins of KL information back to Boltzmann's entropy and also describes its probabilistic interpretation. In Jaynes (2003), section 11.4, a similar derivation is given for the Shannon entropy. Jaynes also discusses the problem of generalizing Shannon entropy to the continuous case (section 12.3). Fisher information goes back to Fisher (1925) and its relationship to KL information is given in Kullback and Leibler (1951). The fact that expected Fisher information is a metric tensor was first noted by Rao (1945).

4–8 Problems

- 4–1 Show that KL information is either positive or zero. Hint: use Jensen's inequality (**Equation 2–8**).
- 4–2 Show that KL information and Fisher information is invariant against coordinate transformations in the sample space.
- 4–3 Verify **Example 4–1** and **Example 4–4**.
- 4–4 Write a computer program demonstrating that the Boltzmann derivation of KL information is correct.
- 4–5 Let $S_1 = \nabla \log f(X|\theta)$. Show that $E(S_1) = 0$ and $\text{Var}(S_1) = I^{\text{Fisher}}(\theta)$. Hint: use the fact that $f(x|\theta)$ is a probability density and that expectation and differentiation can be exchanged.
- 4–6 Show that the expected Fisher information $I^{\text{Fisher}}(\theta)$ is constant if θ is a location parameter, and that it is proportional to $1/\theta^2$ for a scale parameter.

- 4–7 Compute the expected Fisher information for the parameters of the Bernoulli, Binomial, Poisson, and Cauchy distributions.

In a nutshell

- There are many different concepts of information. They are all related, concern the relationships between two distributions, and quantify the information contained in random variable for another.
- The most fundamental quantity is Kullback-Leiber information, or equivalently, Boltzmann entropy. It has many favorable properties, including transformation invariance.
- The Boltzmann entropy has probabilistic interpretation as the log-probability to observe the true model under a specified generating model.
- Pearson's chi-squared statistic is an approximation to (twice) the KL divergence.
- Expected Fisher information is a local version of KL information and has a geometrical meaning as a metric in the space of models.
- Mutual information measures independence and is an important application of KL information.
- Shannon entropy can be viewed as a special case of KL information where the approximating model is the uniform distribution. It is problematic to apply for continuous random variables.

Related chapters

Likelihood, AIC, Geometry