Independent Component Analysis

(Latent Variable Model Ⅲ)

Classical multivariate statistics is largely based on the elegant theory and vast applications of multivariate normal distribution. The ubiquitous Central Limit Theorem asserts that, under mild conditions, sums of random variables are of normal distributions asymptotically.

Yet non-normal data abound, some are known to be far away from normal distributions. In this section, we consider random vector observations generated from non-normal distributions, conventionally called non-Gaussian data. In particular, we consider the method of independent component analysis, which aims at discovering non-Gaussian signal sources based on their mixture output.

1 Non-Gaussian Independent Component Analysis

Motivation of Independent Component Analysis

Often measurement data cannot be taken in isolation. In addition to random variations, measurements themselves may have been taken as a mixture of several sources, such as the recording of sounds in a noisy environment. As another example, in Electroencephalography (EEG), each electrode records a mixture of electric activities of the brain of the subject.

Independent Component Analysis (ICA) aims to identify and separate independent sources, especially sources with output of non-Gaussian distributions. ICA is also called blind signal separation in the signal procession context.

Technically, ICA is a type of latent variable model and can be viewed as one of the many variations of principal component analysis.

Comparison with Factor Analysis

Independent component analysis is a type of latent variable model, thus have many properties analogous to another latent variable model we studies earlier — the orthogonal factor model. Below we review their related properties briefly.

Review of FA

Recall that in factor analysis, we studied the orthogonal factor model, which aims at identifying latent, unobservable variables called common factors. The classical orthogonal factor model formulates a p-variate random vector $\boldsymbol{X}$ as a linear model of m underlying factors, $m \le p$.

$$\boldsymbol{X} = \boldsymbol{\mu} + L\boldsymbol{F} + \boldsymbol{\varepsilon}$$

Its expanded form is

$$\begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \ell_{11} & \cdots & \ell_{1m} \\ \vdots & \vdots & \vdots \\ \ell_{i1} & \cdots & \ell_{im} \\ \vdots & \vdots & \vdots \\ \ell_{p1} & \cdots & \ell_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_p \end{bmatrix}$$

where

- $\boldsymbol{X}$ is the original, observable p-variate random variables with mean $\boldsymbol{\mu}$.
- $F_k,\ k = 1, \cdots, m$ are model-assumed, unobservable random variables called common factors.
- Typically $m < p$ to accomplish dimension reduction.
- The factor vector $\boldsymbol{F} = (F_1, \cdots, F_m)'$ is usually required to have uncorrelated, components, thus commonly normalized to have mean vector and covariance matrix as

$$\mathbb{E}(\boldsymbol{F}) = \boldsymbol{0}_m, \qquad Cov(\boldsymbol{F}) = \mathbb{E}(\boldsymbol{F}\boldsymbol{F}') = \boldsymbol{I}_m.$$

- The covariance matrix of $\boldsymbol{X}$

$$\Sigma = \boldsymbol{L}\boldsymbol{L}' + \Psi$$

where Ψ is diagonal. Analyzing and interpreting the covariance structure is a main focus of factor models.

- Non-uniqueness of the factors

The common factors and their loadings are unique only up to orthogonal linear transformations.

For any orthogonal linear transformation of the m common factors, the transformation can be represented by an $m \times m$ orthogonal matrix $\boldsymbol{T}$, with $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}_m$. Thus

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\varepsilon}$$

Therefore any orthogonal transformation of the common factors corresponds to an equivalent factor model with a new set of loadings and rotated common factors,

$$\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{T}, \qquad \boldsymbol{F}^* = \boldsymbol{T}'\boldsymbol{F},$$

and the same covariance matrix,

$$\Sigma = \boldsymbol{L}^*\boldsymbol{L}^{*\prime} + \Psi = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{L}' + \Psi = \boldsymbol{L}\boldsymbol{L}' + \Psi$$

ICA vs FA

The model formulation of Independent Component Analysis is similar to that of the orthogonal factor models. Both can be viewed as unsupervised learning. However there are important differences.

- ICA looks for independent components.

Typically factor analysis assumes the the data are of normal distributions. Then the assumed covariance structure in the orthogonal factor model implies mutual **independence** of the common factors F_i's. This independence assumption is important in using maximum likelihood to obtain parameter estimations of the factor model for given data. Without the normality assumption, the common factors F_i's are only assumed to be pairwise **uncorrelated**, which is not as strong as independence.

As the name implies, independent component analysis based on the assumption that the latent components are independent.

In general, correlation is a second-moment property, while independence is related to all orders of moments.

- ICA cares about the estimation of the components.

In Factor analysis, often the primary interest is to estimate the factor loadings ℓ_{ij}. The estimation of the latent variables $\boldsymbol{F}$ is only of secondary interest. While in independent component analysis, the exact formulation of the components is important.

Structures of Independent Component Analysis

In independent component analysis, the observed multivariate data are considered as mixture or linear combinations of independent, latent, non-Gaussian component variables, also called sources.

Noiseless, equal-dimension ICA recovery

Consider the simplest, noiseless case. The model assumption is that the obserbable p-vector $\boldsymbol{X}$ is a linear transformation of an un-observable, p-vector $\boldsymbol{S}$ of **independent components**.

$$\boldsymbol{X}_{p\times 1} = A_{p\times p}\boldsymbol{S}_{p\times 1}$$

The objective is to recover the components of $\boldsymbol{S}$.

- The components of $\boldsymbol{S} = \begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$ are assumed to be statistically independent. Therefore the joint density $f(\boldsymbol{s})$ of random vector $\boldsymbol{S}$ can be written as the product of the marginal densities $f_i(s_i)$ of the univariate components S_i.
$$f(\boldsymbol{s}) = \prod_{i=1}^{p} f_i(s_i)$$
- The components S_i's are assumed to be **non-Gaussian**. More precisely, among all p components ($p \geq 2$), at most one Gaussian component is allowed in this basic independent component analysis model.
  Otherwise, the components are not identifiable.
- If all components are know to be or close to Gaussian distributions, then ICA is not appropriate.
  In the approximately normal case, PCA or its variations should be used in the place of ICA.
- The **mixing matrix** A is assumed invertible, thus **unmixing matrix** A^{-1} exists.
- The goal is to recover the independent resources or "signals" $S_1, \cdots, S_p$.
- Ambiguity in Independent Component Analysis
  Like the factor analysis models, independent component analysis models lack of identifiability.

  Below are some identifiability issues and efforts to fix them.
  - The covariance of $\boldsymbol{S}$ can not be uniquely determined.
    Solutions: Impose $Cov(\boldsymbol{S}) = \boldsymbol{I}_p$.
    The determination of variance of each S_i is left to the local information of the specific application.
  - The signs of the components S_i's can not be determined.
    Sometimes it is okay, when the application picks the desired signs.
    Sometimes the sign issue is problematic.
  - The order of S_i can not be determined.
    The change of orders of components can be formulated as the action of a permutation matrix. For any permutation matrix P, (recall $P^{-1} = P'$)
    $$\boldsymbol{X} = A\boldsymbol{S} = (AP^{-1})(P\boldsymbol{S}) = A^*\boldsymbol{S}^*$$
    In this formulation, the covariance matrix of the independent components stays the same,
    $$Cov(\boldsymbol{S}^*) = Cov(P\boldsymbol{S}) = P'Cov(\boldsymbol{S})P = P'\boldsymbol{I}_pP = \boldsymbol{I}_p$$

ICA aims to recover a set of independent components with the above covariance matrix. Consequently the order of the component sources is not automatically recovered. Knowledge of the specific application is needed to order (as well as to scale) the components.

Ideas of ICA estimation

How do ICA models recover the non-Gaussian components S_i's?

Let $\boldsymbol{w}_1', \cdots, \boldsymbol{w}_p'$ be the row vectors of the un-mixing matrix A^{-1}. Let's re-express the to-be-recovered components S_i's as functions of the observed data $\boldsymbol{X}$.

$$\boldsymbol{S} = \begin{bmatrix} S_1 \\ \vdots \\ S_i \\ \vdots \\ S_p \end{bmatrix} = A^{-1}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{w}_1'\boldsymbol{X} \\ \vdots \\ \boldsymbol{w}_i'\boldsymbol{X} \\ \vdots \\ \boldsymbol{w}_p\boldsymbol{X} \end{bmatrix}$$

This expression shows that each of the desired independent components can be written as linear combinations of the components of the original observed $\boldsymbol{X}$, that is, a linear sum of random variables.

Usually, non-zero sum of random variables are more Gaussian, that is, closer to a normal distribution, than the original individual variables, by the effects characterized in the Central Limit Theorem (CLT). Our goal here is to go to the opposite direction of the CLT.

To see what it means to de-Gaussian, for a fixed i, use $\boldsymbol{X} = A\boldsymbol{S}$ again in the above equation, we may write a component of the right hand side vector as

$$\text{(the } i\text{th component)} \quad Y_i = \boldsymbol{w}_i'\boldsymbol{X} = \boldsymbol{w}_i'A\boldsymbol{S} = (A'w_i)'\boldsymbol{S} = \boldsymbol{z}'\boldsymbol{S} = \sum z_jS_j, \qquad \boldsymbol{z} = A'w_i = [z_1 \ \cdots \ z_p]'$$

So Y_i can be viewed as a linear combination of the S_j's. Recall that we want to

$$\text{make} \quad Y_i = \sum z_jS_j = S_i$$

Y_i as a linear combination of the S_j's would be close to a Gaussian distribution if it is a true weighted sum of all S_j's, and would be farthest away from Gaussian if it equals to one component S_i only. Hence we can view $Y_i = \boldsymbol{w}_i'\boldsymbol{X} = S_i$ as deliberately non-Gaussian.

To recap, the idea of ICA can be stated as the following:

ICA selects $\boldsymbol{w}_i$'s so that $\boldsymbol{w}_i'\boldsymbol{X}$ is as far away from normal distribution as possible.

In other words, we aim to choose linear combinations of observed X_i's such that the linear combination is as non-Gaussian as possible. Note that this is practically going to the opposite direction of the Central Limit Theorem.

ICA data and recovered signal sources

The observed data for ICA is of the usual multivariate data form

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Often the observations $1, \cdots, n$ are ordered time points.

The relationship between the n observed p-variate data $[X_{j1}, \cdots, X_{jp}]^T, j = 1, \cdots, n$, and the recovered n data points of the p independent sources $[S_{j1}, \cdots, S_{jp}]^T, j = 1, \cdots, n$, can be better viewed in the transform of the data matrix.

$$\begin{bmatrix} X_{11} & X_{21} & & X_{j1} & & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{j2} & \cdots & X_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \cdots & & \cdots & \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ X_{1p} & X_{2p} & & X_{jp} & & X_{np} \end{bmatrix}_{p\times n} = A_{p\times p} \begin{bmatrix} S_{11} & S_{21} & & S_{j1} & & S_{n1} \\ S_{12} & S_{22} & \cdots & S_{j2} & \cdots & S_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \cdots & & \cdots & \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ S_{1p} & S_{2p} & & S_{jp} & & S_{np} \end{bmatrix}_{p\times n}$$

2 Measures of non-Gaussianity

In order to select variables as non-Gaussian as possible, we need to have measurements of non-Gaussian-ness. Below we introduce two types of such measures, one is method-of-moments based (skewness and kurtosis), one is information theory based (entropy).

2.1 Skewness and Kurtosis

First type of measures for Gausianness is based on the central moments $\mu_k = \mathbb{E}[(X-\mu)^k]$ of random variables and their normalized versions, where $\mu = \mathbb{E}(X)$. Note that $\mu_2 = \sigma^2$ is the variance of the X.

Normal density functions are symmetric with respect to the mean. A random variable is more non-Gaussian if it is more asymmetric.

Skewness of a random variable X with mean μ is defined as

$$\frac{\mu_3}{\mu_2^{3/2}} = \frac{\mathbb{E}(X-\mu)^3}{[\mathbb{E}(X-\mu)^2]^{3/2}}$$

Normal random variables are of skewness 0. Large magnitude of skewness of a random variable implies its deviation from normality, thus skewness can be used as a measure of non-Gaussianness. For example, a random variables with longer right tail has positive skewness, while longer left tail implies negative skewness.

However, every symmetric distribution has skewness 0, Thus for symmetric non-Gaussian distributions, skewness is not informative. Other non-Gaussianness measures are needed.

Kurtosis is a measure of peakedness of the probability density function of a random variable X, defined by

$$\frac{\mu_4}{\mu_2^2} = \frac{\mathbb{E}(X-\mu)^4}{[\mathbb{E}(X-\mu)^2]^2} \quad \in [0, \infty)$$

Normal distribution has kurtosis $= 3$.

Excess Kurtosis is defined as the deviation from the normal kurtosis:

$$\kappa = \frac{\mu_4}{\mu_2^2} - 3 \quad \in [-3, \infty)$$

Large magnitude of Excess Kurtosis indicates deviation from normal distribution.

An example

Consider a case of dimension $p = 2$. S_1, S_2 are independent random variables of uniform distribution on $[-\sqrt{3}, \sqrt{3}]$.

$$f_i(s_i) = \begin{cases} \frac{1}{2\sqrt{3}}, & |s_i| \le \sqrt{3} \\ 0, & otherwise. \end{cases}$$

Let

$$\boldsymbol{X} = A\boldsymbol{S} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \boldsymbol{S}$$

Suppose that a sample of 1000 observations of $\boldsymbol{X}$ are obtained.

- What will PCA do to the data matrix X? (Maximizing component variance.)
- What will ICA do on X? (Maximizing component non-Gaussian measures such as kurtosis)
- What happens if both S_i are $\sim N(0,1)$?

(See similar cases in the in-class demo.)

2.2 Entropy

Another type of measure on Gausianness or non-Gaussianness is based on the special characteristics of normal distributions in terms of information entropy.

Entropy is a measure of randomness or unpredictability of information content. In physics, entropy implies the amount of "disorder" of a system.

Entropy (Shannon Entropy)

Entropy of a discrete random variable X with $p_i = \mathbb{P}(X = i)$ is defined as

$$H(X) = -\sum_i p_i \log p_i = \sum_i p_i \log \frac{1}{p_i}$$

The base of the logarithm is a constant a, $\log = \log_a$.
Several commonly used logarithm bases are $a = 2$, $a = e$, and $a = 10$.

As shown in the definition, the original entropy is on discrete probability distributions. Differential entropy is the generalization of Shanon entropy to continuous random variables.

Differential entropy ($a.k.a.$ continuous entropy)

Differential entropy of a continuous random variable X with probability density function $f(x)$ is defined as

$$H(X) = -\int_{\mathbb{R}} f(x) \log f(x)\, dx$$

where $0 \log 0$ is defined to 0, which is reasonable since $\lim_{x\to 0+} x \log x = 0$.

Gaussian Entropy (exercises)

- For $X \sim N(\mu, \sigma^2)$ with density
$$\phi(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
its differential entropy has the form
$$H(X) = \log(\sigma\sqrt{2\pi e})$$
where the logarithm is of base e. Notice that Gaussian differential entropy is independent of μ.
- For any continuous random variable with mean 0, variance σ^2, and probability density function f,
$$H(X) \leq \log(\sigma\sqrt{2\pi e})$$
which can be derived using
$$-\int_{\mathbb{R}} f(x) \log \phi(x) dx = \log(\sigma\sqrt{2\pi e})$$
where $\phi(x) = \phi(x; 0, \sigma^2)$.
- Therefore, among all continuous random variable with mean zero, variance σ^2, normal random variable has the largest differential entropy.

This can be stated that, among all probability distributions, normal distributions achieve the maximum entropy, which means the most randomness.

Therefore entropy, or rather, the distance of an entropy from the normal entropy (negentropy), can be used to develop measures of non-Gaussianity.

2.3 Joint, conditional, differential entropy

Definitions

For finite, discrete random variable X, its (Shannon) **entropy** is defined as

$$H(X) = \mathbb{E}[-\log P(X)] = -\sum_i p(x_i) \log p(x_i), \qquad p(x_i) = \mathbb{P}(X = x_i)$$

Entropy for more than one random variables

For two variables, their **joint entropy** is defined as

$$H(X, Y) = -\sum_{i,j} p(x_i, y_j) \log p(x_i, y_j)$$

For continuous random variables X, Y with joint density function $f(x, y)$, the joint entropy or **joint differential entropy** is defined as

$$H(X, Y) = -\int_{\mathbb{R}^2} f(x, y) \log f(x, y) dx dy$$

Conditional entropy

It is natural to define the conditional entropy of X given $Y = y$ as

$$H(X \mid Y = y) = -\sum_i p(x_i \mid y) \log p(x_i \mid y)$$

The **conditional entropy** of X given Y is

$$\begin{aligned} H(X|Y) &= \sum_j p(y_j) H(X|Y = y_j) \\ &= \sum_j p(y_j) \left(-\sum_i p(x_i|y_j) \log p(x_i|y_j) \right) \\ &= \sum_j p(y_j) \left(\sum_i \frac{p(x_i, y_j)}{p(y_j)} \log \frac{p(y_j)}{p(x_i, y_j)} \right) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \end{aligned}$$

For continuous random variable X with density function $f(x)$, the **conditional differential entropy** is analogously defined as

$$\begin{aligned} H(X|Y) &= -\int_{\mathbb{R}^2} f(x, y) \log f(x|y) dx dy \\ &= \int_{\mathbb{R}} f(y) \left(-\int_{\mathbb{R}} f(x|y) \log f(x|y) dx \right) dy \\ &= \int_{\mathbb{R}} f(y) H(X|Y = y) dy \end{aligned}$$

Properties of joint and conditional entropy

- $Y = AX$, where A is an $n \times n$ matrix, then
$$H(Y) = H(X) + \log |\det(A)|$$
- Rewrite the conditional density as
$$\begin{aligned} H(X|Y) &= \sum_{i,j} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)} + \sum_j \left(\sum_i p(x_i, y_j) \right) \log p(y_j) \\ &= -\sum_{i,j} p(x_i, y_j) \log p(x_i, y_j) + \sum_j p(y_j) \log p(y_j) \end{aligned}$$
we obtain
$$H(X|Y) = H(X, Y) - H(Y)$$
and
$$H(Y|X) = H(X, Y) - H(X)$$
- Bayes' rule
$$H(Y|X) = H(X|Y) - H(X) + H(Y)$$
- Chain rule
$$H(X_1, \cdots, X_n) = \sum_{i=1}^n H(X_i|X_{i-1}, \cdots, X_1)$$

- Symmetric relations

$$H(X|Y) + H(Y|X) = 2H(X,Y) - H(Y) - H(X)$$
$$H(X) + H(Y) = 2H(X,Y) - H(X|Y) - H(Y|X)$$

Mutual information

Mutual information of a random vector X and its with n components $(X_1, \cdots, X_n)$ is defined as

$$I(X) = I(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X)$$

$I(X)$ is the **Kullback-Leibler distance** between the distribution of X and its independence version, that is, the random vector with the same marginal distributions but independence components.

$I(X)$ is a natural measure of dependence of the components of X. The smaller $I(X)$, the more independent the components of X.

Consider the ICA setup $\boldsymbol{X} = A\boldsymbol{S}$ with orthogonal A.

$$I(\boldsymbol{S}) = \sum_{i=1}^{p} H(S_i) - H(\boldsymbol{S}) = \sum_{i=1}^{p} H(S_i) - H(\boldsymbol{X}) - \log|A^{-1}| = \sum_{i=1}^{p} H(S_i) - H(\boldsymbol{X})$$

with $\boldsymbol{S} = A^{-1}\boldsymbol{X}, S_i = \boldsymbol{w}_i'\boldsymbol{X}$.

Therefore, using the measure of mutual information, seeking maximum independence among components $S_1, \cdots, S_n$ from mixture $\boldsymbol{X} = A\boldsymbol{S}$ comes down to seeking orthogonal linear transformation that minimizes $I(\boldsymbol{S}) = I(A^{-1}\boldsymbol{X})$.

2.4 NegEntropy

For convenience in computation and approximation, sometimes the negentropy measure is used.

NegEntropy, short for Negative Entropy, is a non-Gaussian-ness measure, a measure of distance to normality.

The negEntropy for a standardized random variable $X \sim (0,1)$, i.e., of mean 0 variance 1, is a comparison of its entropy to that of a standard normal $Z \sim N(0,1)$ with the same mean 0 and variance 1.

$$J(X) = H(Z) - H(X)$$

A commonly used convenient approximation (supposedly from Jones 1987) is (details and derivations omitted)

$$J(X) \approx \frac{1}{12}E(X^3)^2 + \frac{1}{48}\kappa(X)^2 \tag{1}$$

where $\kappa(X)$ is the excess kurtosis of X. Notice that

$$J(Z) = H(Z) - H(Z) = 0, \qquad for \quad Z \sim N(0,1),$$

and the approximation form in (1) has similar property,

$$\frac{1}{12}E(Z^3)^2 + \frac{1}{48}\kappa(Z)^2 = 0, \qquad for \quad Z \sim N(0,1).$$

Therefore negEntropy $J(X)$ and its approximation (1) can serve as non-Gaussian-ness measure.

3 ICA Appendix: Background on entropy*

In information theory, the quantity entropy measures a type of "information" contained in a random variable in terms of the randomness of its outcomes.

Information theory studies the quantification, storage, and communication of information.

The occurrence of a likely event is common, of high probability, high certainty or less randomness. On the other hand, unlikely events are of low probability, thus of high uncertainty or more randomness.

The basic intuition in quantifying information is that the occurrence of an unlikely event is more informative than the occurrence of a likely event.

Entropy as a measure of information uncertainty

Entropy is meant to be a measure of the amount of uncertainty or information content.

There are a few desired properties for such a measure:

- Likely events should have low information content, unlikely events should have higher information content.
- Deterministic events that are guaranteed to happen should have no information content.
- Independent events should have additive information.

Shannon entropy of a finite, discrete valued random variable quantifies the uncertainty or randomness of the outcomes of the variable. Information entropy is originally measured in *bits*. One bit is typically defined as the information entropy of a binary random variable that is 0 or 1 with equal probability, or the information that is gained when the value of such a variable becomes known.

Examples

Let the variable be the outcome of a die or a coin toss. Let k be the number of bits needed to capture the uncertainty of the variable. Notice that k bits is equivalent to 2^k distinct stages.

- Example 1 (biased coin toss).

  The die toss has $p = 1/4$ chance for the variable to have value '1', $1 - p = 3/4$ chance for the variable to have value '0'.

  The situation is viewed as to be equivalent to a fair die with four faces, only one face has value '1'. In other words, one out of four equally likely outcomes is '1'.

  In this case, we may say that we need an information-storage capacity of four distinct states to quantify the uncertainty of the outcome '1'.

  Then $k = 2$ binary bits are needed to quantify the uncertainty information of the random variable, to distinguish or pin down the uncertainty of the outcome '1'.

  For example, if the 2-bits is represented by {00, 01, 10, 11}, we may use 00 for the very outcome '1', and use $01, 10, 11$ for the other three possible outcomes of the four-face fair die.

  In this case, $p = 1/4$ leads to $k = 2$ binary bits: Because the storage capacity of **distinct states** needed is

$$2^k = 2^2 = 4 = 1/p$$

  according to the binary capacity of bits.

The number of bits needed, also called **self-information** of the outcome, is

$$k = \log_2(1/p) = 2$$

which reflects the level of information or uncertainty associated with the particular outcome '1'.

Note that physically only one bit is needed to store the two possible values of the variable in a digital device. However conceptually or logically two bits are needed to quantify the uncertainty.

- Example 2 (fair coin toss).

  Compare with the situation that the die toss has $p = 1/2$ chance for the variable to have value '1', $1-p = 1/2$ chance for the variable to have value '0'.

  Now the situation is equivalent to a fair coin with two faces, one face has value '1'. So one out of two equally likely outcomes is 1. This time $k = 1$ bit is needed to quantify the uncertainty information of the random variable, to distinguish the outcome "1".

  We only need an information-storage capacity of two distinct states to quantify the uncertainty of outcome '1'.

  Thus in this case, $p = 1/2$ leads to $k = 1$: the needed storage capacity of the two distinct states is

  $$2^k = 2^1 = 2 = 1/p$$

  The number of bits (self-information) needed for the event is

  $$k = \log_2(1/p) = 1.$$

- Example 3 (certain coin toss).

  Consider the extreme case that a die toss always lands with face 1, the probability of the event $p = 1$. There is no uncertainty. No bits are needed to quantify the uncertainty.

  In this case $p = 1, k = 0$. The number of distinct stages is

  $$2^k = 2^0 = 1 = 1/p,$$

  and the slef-information, the bits needed is (defining $\log(1/0) = 0$)

  $$k = \log_2(1/p) = 0.$$

  Note that physically we still need at least part of one bit to store the outcome '1' , the value of the variable. However conceptually no bits are needed to quantify the uncertainty, since there is no uncertainty.

- Example 4 (probability weighted average).

  Possible outcomes of a three-faced die toss are $\{A, B, C\}$ with probability $p_A = 1/2,\ p_B = 1/4,\ p_C = 1/4$ respectively. Based on the above derivation,

  - Outcome of "A" has probability $1/2$ to occur, needs $k_A = 1$ bit to quantify its information.
  - Outcome of "B" has probability $1/4$ to occur, needs $k_B = 2$ bits.
  - Outcome of "C" has probability $1/4$ to occur, needs $k_C = 2$ bits.

  From Example 1 and Example 2, we know the number of bits, or self-information, of each outcome.

  Weighted by the outcome probabilities, the overall average or expected number of bits is

  $$\bar{k} = p_A k_A + p_B k_B + p_C k_C = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = 1.5(bit)$$

Equivalently,

$$\begin{aligned}\bar{k} &= p_A \log_2(1/p_A) + p_B \log_2(1/p_B) + p_C \log_2(1/p_C) \\ &= \frac{1}{2}\log_2(1/2) + \frac{1}{4}\log_2(1/4) + \frac{1}{4_2}\log(1/4) \\ &= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = 1.5(bit)\end{aligned}$$

This is the average bit of all possible outcomes in terms of the probability distribution.

This quantity conveys the amount of uncertainty in the random event with 3 possible outcome.

- Revisit Example 1 (biased coin toss)

  In the same token, we may evaluate the average amount of uncertainty in the probability distribution in Example 1, where outcome probabilities are $p_1 = p,\ p_2 = 1 - p$. The average amount of uncertainty is

  $$\begin{aligned}p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2) &= (1/4)\log_2(4) + (3/4)\log_2(4/3) \\ &= (1/4) \times 2 + (3/4) \times 0.415 \\ &= 0.5 + 0.311 \approx 0.81(bit)\end{aligned}$$

- Revisit Example 2 (fair coin toss)

  Comparatively, the event in Example 2 should contain more uncertainty that Example 1 does.

  Indeed, for Example 2, the average number of bits is

  $$p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2) = (1/2)\log_2(2) + (1/2)\log_2(2) = 1.0(bit)$$

Remarks

- The above case can be unified by the formula

  $$\bar{k} = p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2) + \cdots + p_n \log_2(1/p_n), \qquad p_i \in (0,1), \quad p_1 + \cdots + p_n = 1. \tag{2}$$

  which leads to the original definition of base-2 entropy.
  (The definition can extend to $p_i \in [0,1]$ with $0 \log 0 = 0$.)

- Other heuristic ways to describe the average amount of information:

  - the average number of yes-no questions need to pin down the output, or

  - the average number of fair bounces (the k_i's, the bits) of a binary machine to generate the output with the given probabilities.

References

Sections 14.7 in *The Elements of Statistical Learning* Hastie, Tibshirani and Friedman.
Sections 10.1-10.4 in *Analysis of Multivariate and High-Dimensional Data* by Koch.