Lecture 22: Information Theory

Reading: Sections 8.5, 14.3

GU4241/GR5241 Statistical Machine Learning
Text

Linxi Liu
April 11, 2017

Measuring Information

Information content of a random variable

We consider a random variable X with distribution P.

- P expresses what we know *before* we observe X.
- How much information do we *gain* by observing X?

That is: By information content of X, we mean the difference in information between knowing P and knowing both P and $X = x$.

To reiterate

For the definition of information, it is useful to think of...

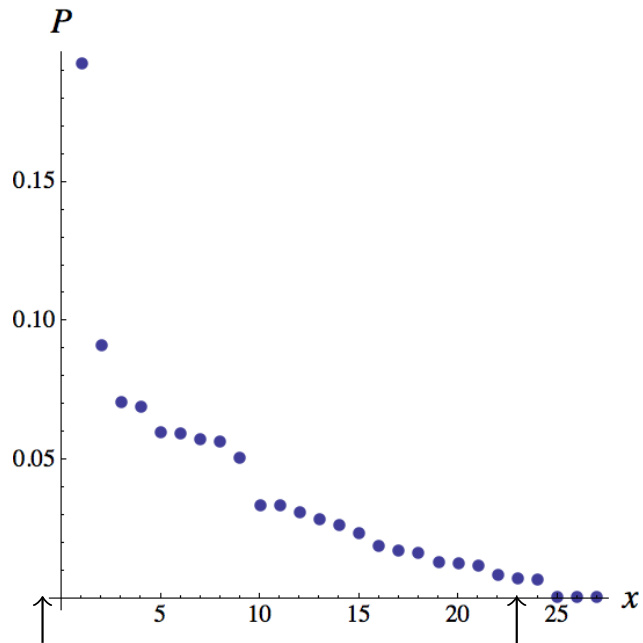
- ...the distribution P as what we *expect to happen*.
- ...the sample outcome $X = x$ as what *actually happens*.

Information

Heuristic motivation

Suppose we sample $X = x$ from a distribution P.

- If $P(x)$ is large: Small surprise; we have not gained much additional information.
- If $P(x)$ is small: We have gained more information.



X=1:
not very surprising,
low information gain

X=23:
unexpected, high
information gain

Information

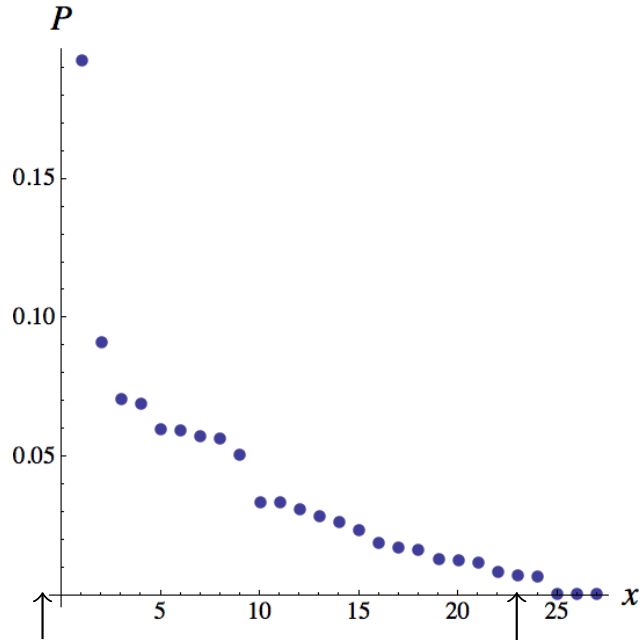
Conclusions

- The information in $X = x$ increases with $\frac{1}{P(x)}$.
- Intuitively, the information gain in two unrelated observations should be additive, so $\frac{1}{P(x)}$ itself is not a useful measure of information.

Definition

The **information** in observing $X = x$ under P is

$$J_P(x) := \log \frac{1}{P(x)} = -\log P(x) \; .$$



X=1:
not very surprising,
low information gain

X=23:
unexpected, high
information gain

Shannon's Entropy

Discrete random variables

In information theory, we have to distinguish between discrete and continuous random variables. If X is a RV with values in a space $\mathbf{X}$, we call X **discrete** if $\mathbf{X}$ has a finite or at most countably infinite number of elements.

Definition

Let X be a discrete random variable with distribution P. The *expected* information in a draw from P,

$$\mathbb{H}[X] := \mathbb{E}_P[J_P(X)]$$

is called the **Shannon entropy** of X, or the **entropy** for short.

Remarks

- Note that
$$\mathbb{E}[J_P(X)] = -\mathbb{E}_P[\log P(X)] = -\sum_{x \in \mathbf{X}} P(x) \log P(x)$$
- The entropy measures the information gained when sampling from P.
- We can interchangeably regard $\mathbb{H}$ as a property of X or of P, and we equivalently write $\mathbb{H}(P)$ for $\mathbb{H}[X]$.

Basic Properties

1. The entropy is non-negative:

$$\mathbb{H}[X] \geq 0$$

2. $\mathbb{H}(P) = 0$ means there is no uncertainty in P:

$$\mathbb{H}(P) = 0 \qquad \Leftrightarrow \qquad P(x_0) = 1 \text{ for some } x_0 \in \mathbf{X} \, .$$

3. If $\mathbf{X}$ is finite with d elements, the distribution with the largest entropy is the uniform distribution U_d, with

$$\mathbb{H}(U_d) = \log d$$

Alternative Derivation

Axiomatic description

Suppose we define *some* measure $\mathcal{H}[X]$ of information in X. Regardless of the definition, we can postulate a number of properties (axioms) that a meaningful measure should satisfy.

Additivity

- If two RVs X and Y are independent, their information content should be disjoint.
- Hence, $\mathcal{H}$ should be additive:

$$X \perp\!\!\!\perp Y \quad \Rightarrow \quad \mathcal{H}[X,Y] = \mathcal{H}[X] + \mathcal{H}[Y]$$

- More generally: We should be able to "remove the joint information" in X and Y from Y by conditioning.
- This is what we require as our first axiom:

(**Axiom I**) $\qquad\qquad \mathcal{H}[X,Y] = \mathcal{H}[X] + \mathcal{H}[Y|X]$

Axiomatic Derivation

Continuity

- We can alternatively regard $\mathcal{H}[X]$ as a function $\mathcal{H}(P)$ of the distribution of X.
- If we make a small change to P, then $\mathcal{H}(P)$ should not "jump". That is:

(**Axiom II**) $\mathcal{H}(P)$ should be continuous as a function of P.

Monotonicity

- Suppose we consider in particular the uniform distribution $P = U_d$ on d outcomes.
- If we increase d, the uncertainty in U_d increases; hence, the information gained by sampling should be higher for $d+1$ than for d:

(**Axiom III**) $$\mathcal{H}(U_d) < \mathcal{H}(U_{d+1})$$

Axiomatic Derivation

Theorem

If a real-valued function $\mathcal{H}$ on $\mathbf{X}$ satisfies Axioms I–III, then

$$\mathcal{H}(P) = c \cdot \mathbb{H}(P) \qquad \text{for all } P \,,$$

for some constant $c \in \mathbb{R}_+$. (The constant is the same for all P.)

In other words

If any information measure satisfies our requirements, it is precisely the entropy, up to a choice of scale.

Shannon's Entropy

How meaningful are the axioms?

- Over the years, about a dozen different axioms for information measures have been proposed.
- It can be shown that basically any meaningful combination of two or three of these axioms leads to the same result (i.e. determines the entropy up to scaling).

One might argue that this makes the entropy a much more fundamental quantity than most quantities used in statistics (variance etc).

Historical note

- The notion of entropy was first conceived in physics. The first precise definition was given by Boltzmann in the 1870s.
- The information-theoretic entropy was introduced in the paper

  Claude Shannon: "A mathematical theory of communication", 1948.

  This paper introduced most of the quantities we discuss here, created the field of information theory, and proved almost all of its fundamental results.

Example: Coding

Suppose we would like to compress a text document (lossless compression).

Huffman Coding

Here is a simple but efficient coding scheme:

1. Given a text, determine the frequency with which each word occurs.
2. Assign short code words to words that occur often, long code words to words that are rare.

This idea (with a specific algorithm for finding determining the code words) is called **Huffman coding**. If all we are allowed to do is to replace text words by code words, this compression method is optimal.

Information-theoretic problems

Suppose we know the distribution P of words in texts. Then we can ask:

1. What is the *expected* compression rate for a random document?
2. Does our encoder achieve the optimal expected rate for P?

Example: Coding

The Source Coding Theorem (Shannon)

Suppose we are given a distribution P on words or symbols and sample a string $X^n = (X_1, \ldots, X_n)$ iid from P. Then for every $\varepsilon > 0$, there is a lossless encoder for which

$$H(P) \leq \mathbb{E}\Big[\frac{1}{n} \cdot \mathsf{length}(\mathsf{encoding}(X^n))\Big] < H(P) + \varepsilon$$

for sufficiently large n.

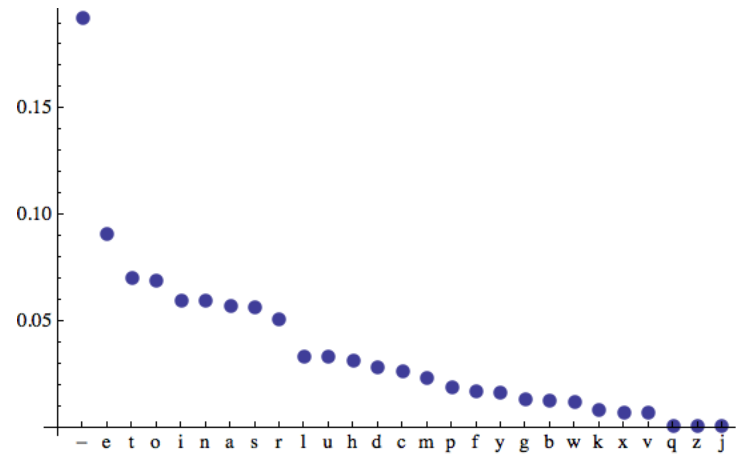
Remarks

- In other words: We can encode the sequence X^n without loss of information using $nH(P)$ bits on average.
- The entropy $H(P)$ is a lower bound for lossless compression: If an encoder achieves a better (=smaller) expectation than above, the probability that it will result in information loss approaches 1 for $n \to \infty$.

How Well Can We Compress English Text?

Character-by-character compression

- We can compress text by splitting the text into characters and assigning a code to each character.
- An empirical estimate of the distribution of characters is shown on the right. The entropy is 4.11 bit/character.
- This compression is not very effective: There are 27 characters and $2^4 < 27 \leq 2^5$, hence we can trivially encode with 5 bits/character.

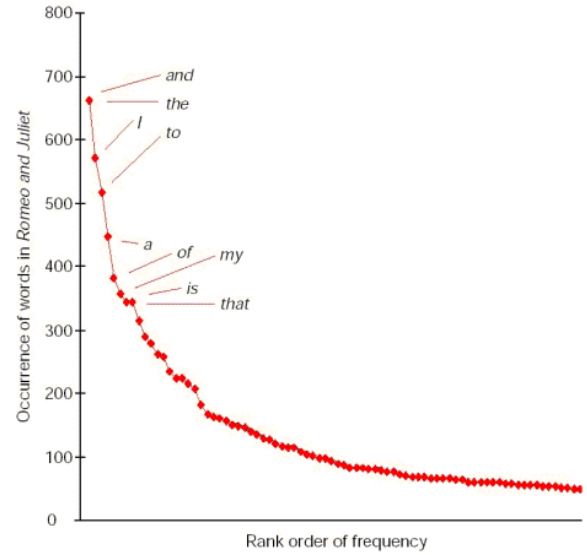


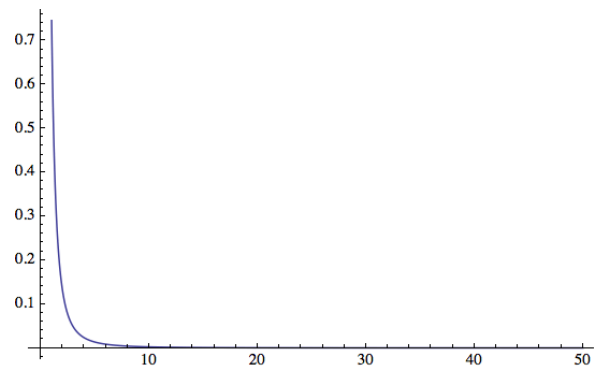
How Well Can We Compress English Text?

Word-by-word compression

- The distribution of words in languages is highly concentrated on a few common words.

  (Upper plot: Ranked word occurrences in *Romeo and Juliet*.)
- If we rank words in English by frequency of occurrence, the occurrence distribution is well-approximated by a Zipf distribution with parameter between 1.5 and 2 (lower plot).
- Due to concentration, these distributions have relatively low entropy.
- Consequence: If we split into words instead or characters, we can achieve much better compression rates.
- Common compression algorithms (e.g. Lempel-Ziv) split into substrings which are not necessarily words.





Kullback-Leibler Divergence

Comparing distributions

We can use the notion of information to compare one distribution to another.

Heuristic motivation

Suppose we wish to compare two distributions P and Q on $\mathbf{X}$.

- The entropy $\mathbb{H}[Q] = \mathbb{E}_Q[J_Q(X)]$ measures how much information gain (in terms of Q) we can *expect* from a random sample from Q.
- Now ask instead: How much information gain in terms of Q can we expect from a random sample drawn from P? We compute: $\mathbb{E}_P[J_Q(X)]$.
- A measure of difference between P and Q should vanish if $Q = P$. Since $P = Q$ means $\mathbb{E}_P[J_Q(X)] = \mathbb{H}(P)$, which is usually not 0, we have to normalize by subtracting $\mathbb{H}(P)$.

Kullback-Leibler Divergence

Definition

The function

$$D_{\mathsf{KL}}(P\|Q) := \mathbb{E}_P[J_Q(X)] - \mathbb{H}(P)$$

is called the **Kullback-Leibler divergence** or the **relative entropy** of P and Q.

Basic Properties

Equivalent forms

$$D_{\mathsf{KL}}[P\|Q] = \mathbb{E}_P[J_Q(X) - J_P(X)] = \sum_{x\in\mathbf{X}} P(x)\log\frac{P(x)}{Q(x)}$$

Positive definiteness

$$D_{\mathsf{KL}}[P\|Q] \geq 0 \qquad \text{and} \qquad D_{\mathsf{KL}}[P\|Q] = 0 \Leftrightarrow P = Q\ .$$

The KL divergence is not a metric

Intuitively, D_{KL} can be used like a distance measure between distributions, however:

- It is *not* symmetric: $D_{\mathsf{KL}}[P\|Q] \neq D_{\mathsf{KL}}[Q\|P]$ in general.
- It does *not* satisfy a triangle inequality.

Convexity

A very useful property of $\mathbb{H}$ and D_{KL} is convexity:

- $\mathbb{H}(P)$ is concave as a function of P.
- $D_{\mathsf{KL}}[P\|Q]$ is convex in the pair (P, Q).

Conditioning

- How can we compute the entropy of Y conditional on X?
- For a fixed value $X = x$, we can simply compute $\mathbb{H}$ from the conditional probability $P(Y|X = x)$ as

$$\mathbb{H}[Y|X = x] = -\sum_{y \in \mathbf{X}} P(y|x) \log P(y|x) \ .$$

- To make the definition independent of x, we take the expectation

$$\mathbb{H}[Y|X] := \mathbb{E}_{P(x)}[\mathbb{H}[Y|X = x]] \ .$$

  This is called the **conditional entropy** of Y given X.
- A few lines of arithmetic show:

$$\mathbb{H}[Y|X] = -\sum_{x,y \in \mathbf{X}} P(x, y) \log P(y|x)$$

Mutual Information

Heuristic Motivation

- Another question we can ask about a pair X, Y of random variables is: How much information do they share?
- In other words: How much does observing X tell us about Y?
- If X and Y contain no shared information, they are independent, and their joint distribution is $P(x, y) = P(x)P(y)$.
- Idea: Compare the actual joint distribution to the independent case using KL divergence.

We first define the mutual information in a different way, but will then see that the idea above indeed applies.

Definition

The function

$$I[X, Y] := \mathbb{H}[X] - \mathbb{H}[X|Y] = \mathbb{H}[Y] - \mathbb{H}[Y|X]$$

is called the **mutual information** of X and Y.

Useful Relationships

Conditioning reduces entropy

$$\mathbb{H}[X, Y] = \mathbb{H}[Y|X] + \mathbb{H}[X]$$

Mutual information as a Kullback-Leibler divergence

$$I[X, Y] = D_{\mathsf{KL}}[P(x, y) \| P(x)P(y)] = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

Note: This compares $P(x, y)$ to the case where X, Y are independent (which means $P(x, y) = P(x)P(y)$).

Mutual information characterizes independence

$$I[X, Y] = 0 \qquad \Leftrightarrow \qquad X \perp\!\!\!\perp Y$$

The Continuous Case

If the sample space $\mathbf{X}$ is uncountable (e.g. $\mathbf{X} = \mathbb{R}$), instead of P and Q we consider densities p and q, we have to substitute integrals for sums.

Differential entropy

$$\mathbb{H}[X] := -\int_{\mathbf{X}} p(x) \log p(x) dx$$

Since p is a density, we can have $\log p(x) > 0$, and $\mathbb{H}[X]$ can be negative. To distinguish it from the entropy, $\mathbb{H}[X]$ is called the **differential entropy**.

KL divergence and mutual information

D_{KL} and I are defined analagously to the discrete case:

$$D_{\mathsf{KL}}(p\|q) := \int_{\mathbf{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

$$I[X, Y] := \int_{\mathbf{X}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx$$

Properties in the Continuous Case

Differential entropy

- Since p is a density, we can have $\log p(x) > 0$, and $\mathbb{H}[X]$ can be negative.
- The term differential entropy is used to distinguish it from the entropy.

KL divergence

The KL divergence for densities still satisfies

$$D_{\mathsf{KL}}(p\|q) \geq 0 \qquad \text{and} \qquad D_{\mathsf{KL}}(p\|q) = 0 \quad \Leftrightarrow \quad p = q \; .$$

As a consequence, the mutual information still satisfies

$$I[X, Y] \geq 0 \qquad \text{and} \qquad I[X, Y] = 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \; .$$

KL Divergence and Maximum Likelihood

Idea

Suppose we observe data $x_1, \ldots, x_n$ and assume a model $\mathcal{P} = \{p(x|\theta) | \theta \in \mathcal{T}\}$. We could fit the model using the KL divergence as a cost measure:

$$\hat{\theta} := \arg\min_{\theta \in \mathcal{T}} D_{\mathsf{KL}}(\mathbb{F}_n | p(x|\theta))$$

Computation

$$\begin{aligned}
\hat{\theta} &= \arg\min_{\theta \in \mathcal{T}} D_{\mathsf{KL}}(\mathbb{F}_n | p(x|\theta)) = \arg\min_{\theta \in \mathcal{T}} \Big(\int_{\mathbf{X}} \mathbb{F}_n(x) \log \frac{\mathbb{F}(x)}{p(x|\theta)} dx \Big) \\
&= \arg\min_{\theta \in \mathcal{T}} \Big(\int_{\mathbf{X}} \mathbb{F}_n(x) \log \mathbb{F}_n(x) dx - \int_{\mathbf{X}} \mathbb{F}_n(x) \log p(x|\theta) dx \Big) \\
&= \arg\max_{\theta \in \mathcal{T}} \Big(\int_{\mathbf{X}} \mathbb{F}_n(x) \log p(x|\theta) dx \Big) = \arg\max_{\theta \in \mathcal{T}} \Big(\int_{\mathbf{X}} \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x) \log p(x|\theta) dx \Big) \\
&= \arg\max_{\theta \in \mathcal{T}} \Big(\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta) \Big) = \hat{\theta}_{\mathsf{MLE}}
\end{aligned}$$

Minimizing KL divergence between $\mathbb{F}_n$ and the model is equivalent to maximum likelihood estimation!

Maximum Entropy Methods

The maximum entropy principle

Suppose we have to choose a model distribution from a given set $\mathcal{P}$ of admissible distributions. The **maximum entropy principle** says: Always choose the distribution

$$P = \arg\max_{Q \in \mathcal{P}} \mathbb{H}(Q)$$

with the highest entropy in $\mathcal{P}$. P is called the **maximum entropy distribution**, which is sometimes abbreviated to 'MaxEnt distribution'.

Rationale

- When choosing a model distribution, we should try to avoid illicit assumptions.
- Higher entropy $\leftrightarrow$ higher uncertainty $\leftrightarrow$ fewer assumptions.

This idea was introduced by the physicist E. T. Jaynes, who championed it as a general modeling approach.

Maximum Entropy Under Constraints

Maximum entropy under constraints

Suppose the set $\mathcal{P}$ of distributions is defined by a constraint. For example:

$$\mathcal{P} = \text{ all distributions on } \mathbb{R} \text{ with variance } \sigma_0^2 \ .$$

Example 1: Trivial constraint

Suppose the only constaint is that the choice of sample space, e.g. $\mathbf{X} = [0, 1]$. Then the maximum entropy distribution is the uniform distribution on $[0, 1]$.

Example 2: Given variance

If $\mathcal{P} = \{$ distributions on $\mathbb{R}$ with $\text{Var}[X] = \sigma_0^2\}$, then P is Gaussian with variance σ_0^2.

The Exponential Family Again

Expectations as constraints

Suppose $\mathbf{X} = \mathbb{R}^d$, and we formulate constraints by choosing functions $S_1, \ldots, S_m : \mathbf{X} \to \mathbb{R}$ and positing their expected values.

That is, the constrained set is

$$\mathcal{P} := \{Q \mid \mathbb{E}_Q[S_1(X)] = s_1, \ldots, \mathbb{E}_Q[S_m(X)] = s_m\} .$$

Constrained optimization problem (for the discrete case)

We add the constraints to the objective function $\mathbb{H}(Q)$ using Lagrange multipliers $\theta_1, \ldots, \theta_n$. We also include a normalization constraint with Lagrangre multiplier θ_0.

$$P = \arg\max_Q \mathbb{H}(Q) + \theta_0 \Big(1 - \sum_{x \in \mathbf{X}} Q(x)\Big) + \theta_1 \Big(s_1 - \sum_{x \in \mathbf{X}} S_1(x) Q(x)\Big) + \ldots + \theta_m \Big(s_m - \sum_{x \in \mathbf{X}} S_m(x) Q(x)\Big)$$

Exponential Family

Maximum entropy solution

The solution of the constrained optimization problem is

$$P(x) = \frac{1}{Z(\theta)} e^{\langle S(x), \boldsymbol{\theta} \rangle} ,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$.

Continuous distributions

Exponential family densities $p(x|\theta)$ for continuous random variables can similarly obtained as maximum entropy models given constraints of the form $\mathbb{E}_p[S_j(x)] = s_j$. This case requires more technicalities, due to the properties of the differential entropy.

Statistical physics

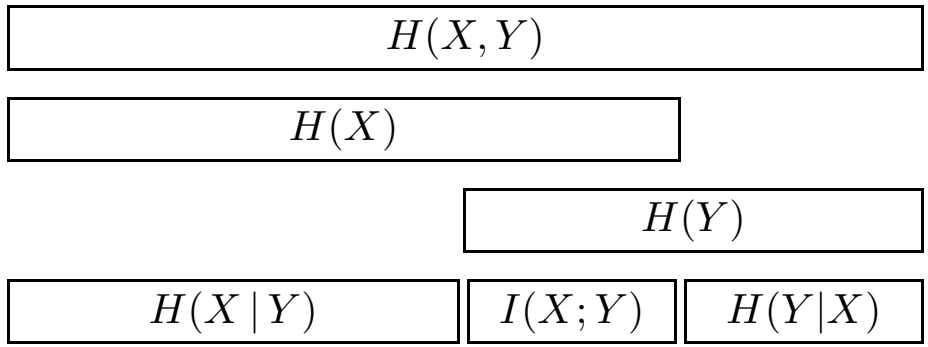
In physics, the maximum entropy distribution under given constraints is called the **Gibbs distribution**.

Summary: Information Theory and Statistics

- Maximum likelihood minimizes D_{KL} between empirical distribution and model.
- Variance, covariance and the χ^2-statistic can be regarded as first-order approximations to entropy, mutual information and KL divergence.
- Various methods can be derived by substituting information-theoretic for traditional statistical quantities.
- Example: A dimension-reduction technique called *independent component analysis* can be motivated as (roughly speaking) a PCA-like method which measures independence in terms of mutual information rather than covariance.

Summary

The various additive relationships can be summarized as follows:



Further reading

David J. C. MacKay: *Information Theory, Inference, and Learning Algorithms.*
Cambridge University Press, 2003.
Online version: See link on course homepage.