Information & Correlation

Jilles Vreeken

11 June 2014 (TADA)

UNIVERSITÄT DES SAARLANDES
M2CI CLUSTER OF EXCELLENCE
mpi max planck institut informatik

Questions of the day

What is information?

How can we measure correlation?

and what do talking drums
have to do with this?

Bits and Pieces

What is

- information
- a bit
- entropy
- information theory
- compression
- ...

Information Theory

Branch of science concerned
with measuring **information**

Field founded by **Claude Shannon** in 1948,
'A Mathematical Theory of Communication'

Information Theory is essentially about
uncertainty in *communication*:

not **what** you say, but what you **could** say

The Big Insight

Communication is a series
of *discrete* messages

each message reduces the uncertainty
of the recipient about
a) the series and *b*) that message

by how much?
that is the amount of information

Uncertainty

Shannon showed that uncertainty can be quantified, linking *physical* entropy to *messages*

Shannon defined
the *entropy* of a discrete random variable X as

$$H(X) = -\sum_{i} P(x_i) \log P(x_i)$$

Optimal prefix-codes

Shannon showed that uncertainty can be quantified, linking *physical* entropy to *messages*

A side-result of Shannon entropy is that

$$-\log_2 P(x_i)$$

gives the *length in bits* of the **optimal prefix code** for a message x_i

What is a prefix code?

Prefix(-free) code:

a code C where **no** code word $c \in C$
is the prefix of another $d \in C$ with $c \neq d$

Essentially, a prefix code defines a **tree**,
where each code corresponds to a path
from the root to a leaf in a decision tree

What’s a bit?

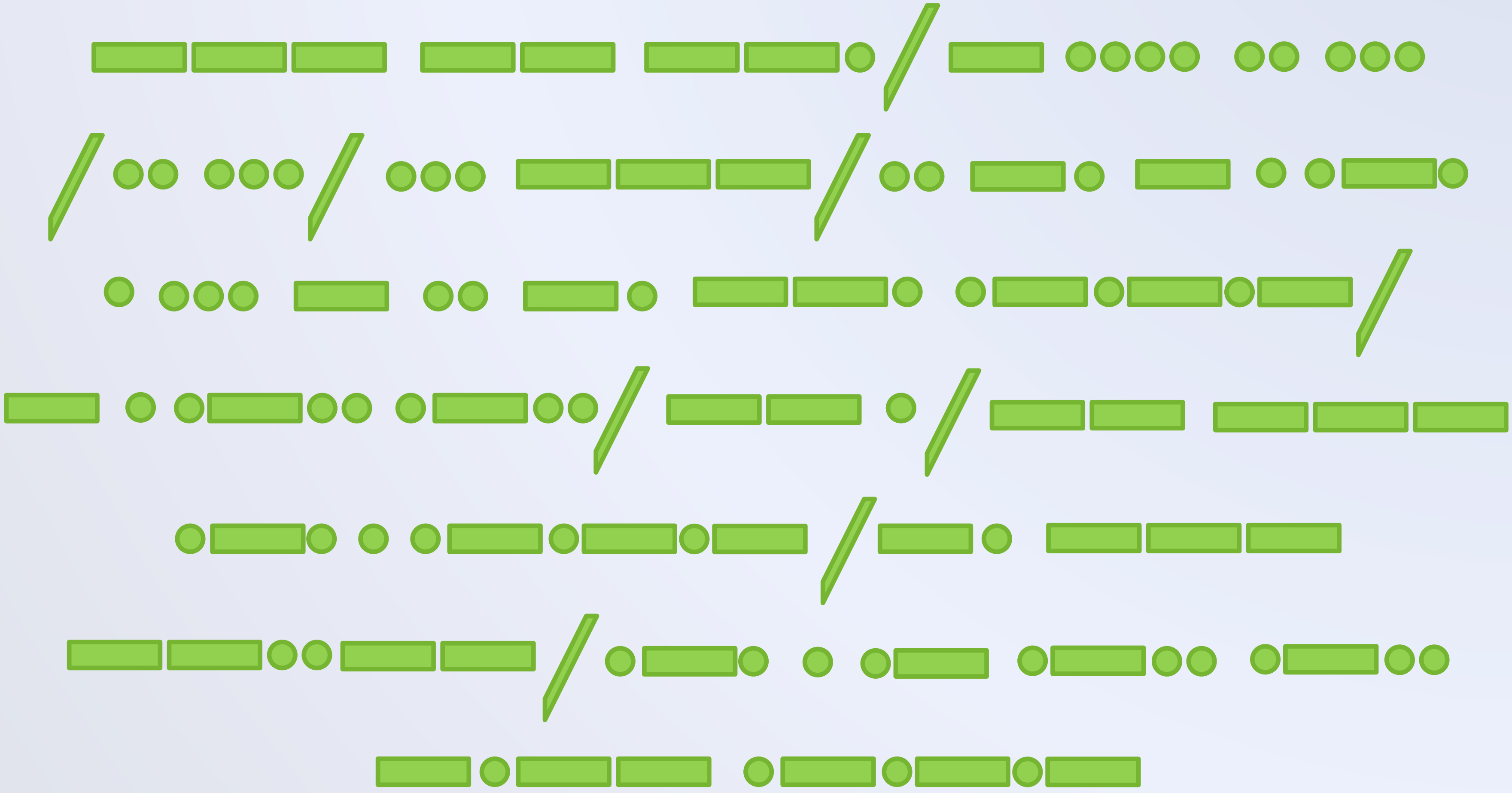
Binary digit

- **smallest** and **most fundamental** piece of information
- *yes* or *no*
- invented by Claude Shannon in 1948
- name by John Tukey

Bits have been in use for a long-long time, though

- Punch cards (1725, 1804)
- Morse code (1844)
- African ‘talking drums’

Morse code



Natural language

Punishes ‘bad’ redundancy:
often-used words are *shorter*

Rewards useful redundancy:
cotxent alolws mishaireng/raeding

African Talking Drums have used this for
efficient, fast, long-distance communication

mimic vocalized sounds: tonal language

very reliable means of communication

Measuring bits

How much information carries a given string?

how many bits?

Say we have a binary string of 10000 ‘messages’

1) 00010001000100010001...00010001000100010001000100010001

2) 01110100110100100110...101011101011101100010110001011011100

3) 00011000001010100000...001000100001000000100011000000100110

4) 0000000000000000000000000000100000000000000000000...0000000

obviously, they are 10000 bits long.

But, are they *worth* those 10000 bits?

So, how *many* bits?

Depends on the encoding!

What is the best encoding?

- one that takes the entropy of the data into account
- things that occur often should get short code
- things that occur seldom should get long code

An encoding matching Shannon Entropy is optimal

Tell us! How many bits? *Please?*

In our simplest example we have

$$P(1) = 1/100000$$
$$P(0) = 99999/100000$$

$$|code_1| = -\log(1/100000) = 16.61$$
$$|code_0| = -\log(99999/100000) = 0.0000144$$

So, knowing P our string contains

$1 * 16.61 + 99999 * 0.0000144 = 18.049$ bits

of information

Optimal....

Shannon lets us calculate optimal code lengths

- what about actual codes? 0.0000144 bits?
- Shannon and Fano invented a near-optimal encoding in 1948, within one bit of the optimal, but not lowest expected

Fano gave students an option: regular exam, or invent a better encoding

- David Huffman didn't like exams; invented Huffman-codes (1952)
- optimal for symbol-by-symbol encoding with fixed probs.

(arithmetic coding is overall optimal, Rissanen 1976)

Optimality

To encode optimally, we need optimal probabilities

What happens if we don't?

Kullback-Leibler divergence, $D(p \,||\, q)$,
measures bits we 'waste' when
we use p while q is the 'true' distribution

$$D(p \parallel q) = \sum_i \log\left(\frac{p(i)}{q(i)}\right) p(i)$$

Multivariate Entropy

So far we've been thinking about a single sequence of messages

How does entropy work for multivariate data?

Simple!

Conditional Entropy

Entropy, for when we, like, know stuff

$$H(X|Y) = \sum_{x \in \mathrm{X}} p(x) H(Y|X = x)$$

When is this useful?

Mutual Information and Correlation

Mutual Information

the amount of information *shared* between two variables X and Y

$$
\begin{aligned}
&I(X,Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)
\end{aligned}
$$

high ↔ correlation

low ↔ independence

Information Gain

(small aside)

Entropy and KL are used in decision trees

What is the best split in a tree?

one that results in as *homogeneous label distributions* in the sub-nodes as possible: **minimal entropy**

How do we compare over multiple options?

$$IG(T,a) = H(T) - H(T|a)$$

Low-Entropy Sets

Theory of Computation	Probability Theory 1	
No	No	1887
Yes	No	156
No	Yes	143
Yes	yes	219

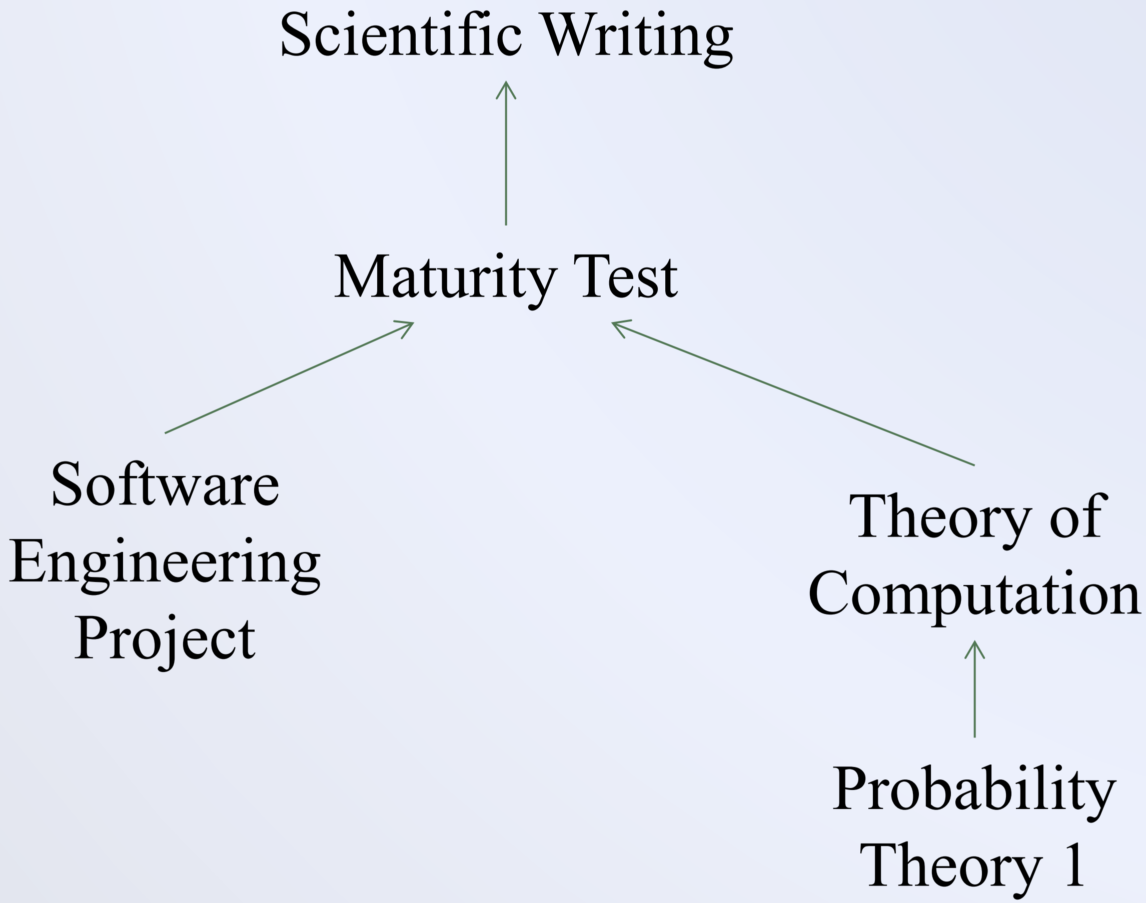
(Heikinheimo et al. 2007)

Low-Entropy Sets

Maturity Test	Software Engineering	Theory of Computation	
No	No	No	1570
Yes	No	No	79
No	Yes	No	99
Yes	Yes	No	282
No	No	Yes	28
Yes	No	Yes	164
No	Yes	Yes	13
Yes	Yes	Yes	170

(Heikinheimo et al. 2007)

Low-Entropy Trees



(Heikinheimo et al. 2007)

Entropy for Continuous-valued data

So far we only considered discrete-valued data

Lots of data is continuous-valued

(or is it)

What does this mean for entropy?

Differential Entropy

$$h(X) = -\int_{\mathbf{X}} f(x) \log f(x) dx$$

(Shannon, 1948)

Differential Entropy

How about… the entropy of Uniform(0,1/2) ?

$$-\int_0^{\frac{1}{2}} -2\log(2)\,dx = -\log(2)$$

Hm, *negative*?

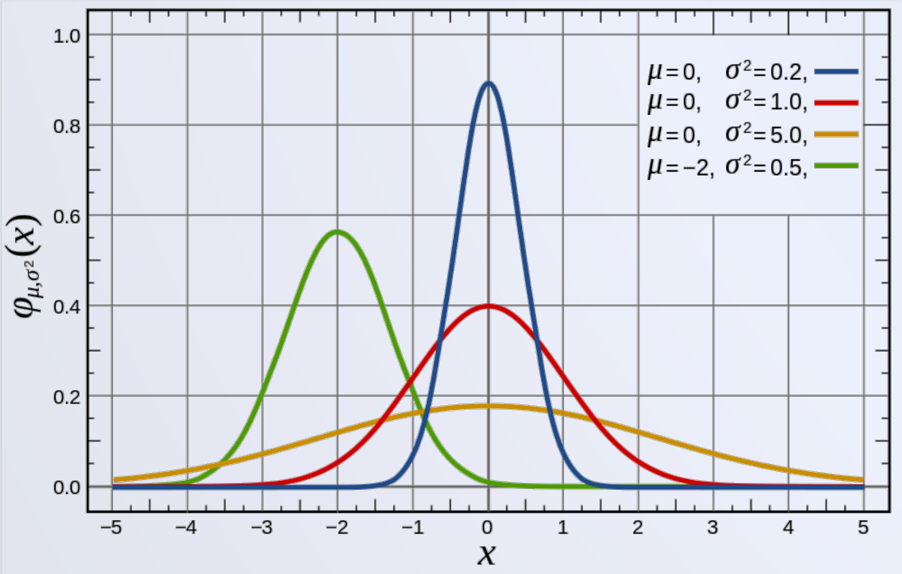
Differential Entropy

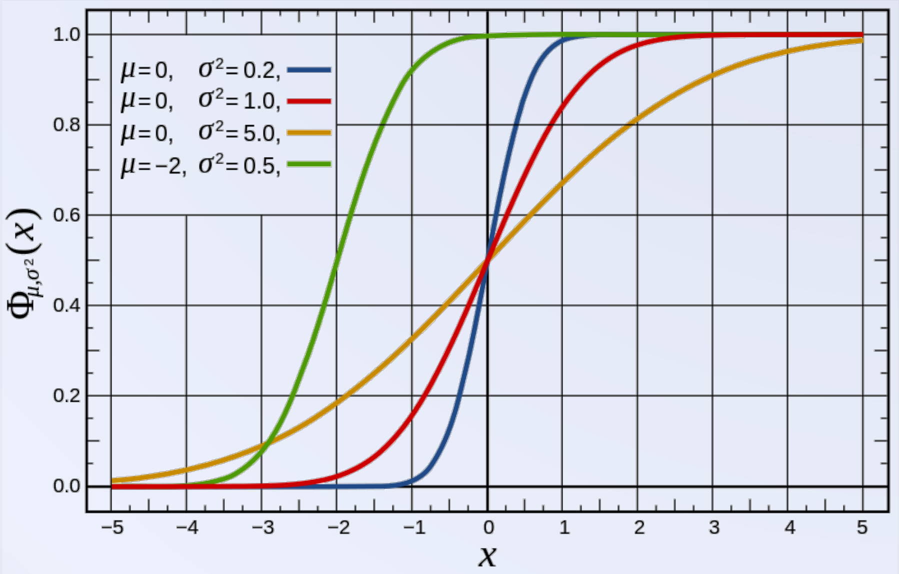
In discrete data step size 'dx' is trivial.
What is its effect here?

$$h(X) = -\int_{\mathbf{X}} f(x) \log f(x) dx$$

(Shannon, 1948)

Cumulative Distributions





Cumulative Entropy

We can define entropy for cumulative distribution functions!

$$h_{CE}(X) = -\int_{dom(X)} P(X \leq x) \log P(X \leq x) dx$$

As $0 \leq P(X \leq x) \leq 1$ we obtain $h_{CE}(X) \geq 0$ (!)

(Rao et al, 2004, 2005)

Cumulative Entropy

How do we compute it in practice?

Easy.

Let $X_1 \leq \cdots \leq X_n$ be i.i.d. random samples of continuous random variable X

$$h_{CE}(X) = -\sum_{i=1}^{n-1} (X_{i+1} - X_i) \frac{i}{n} \log \frac{i}{n}$$

(Rao et al, 2004, 2005)

Multivariate Cumulative Entropy?

Tricky.

Very tricky.

Too tricky for now.

(Nguyen et al, 2013, 2014)

Cumulative Mutual Information

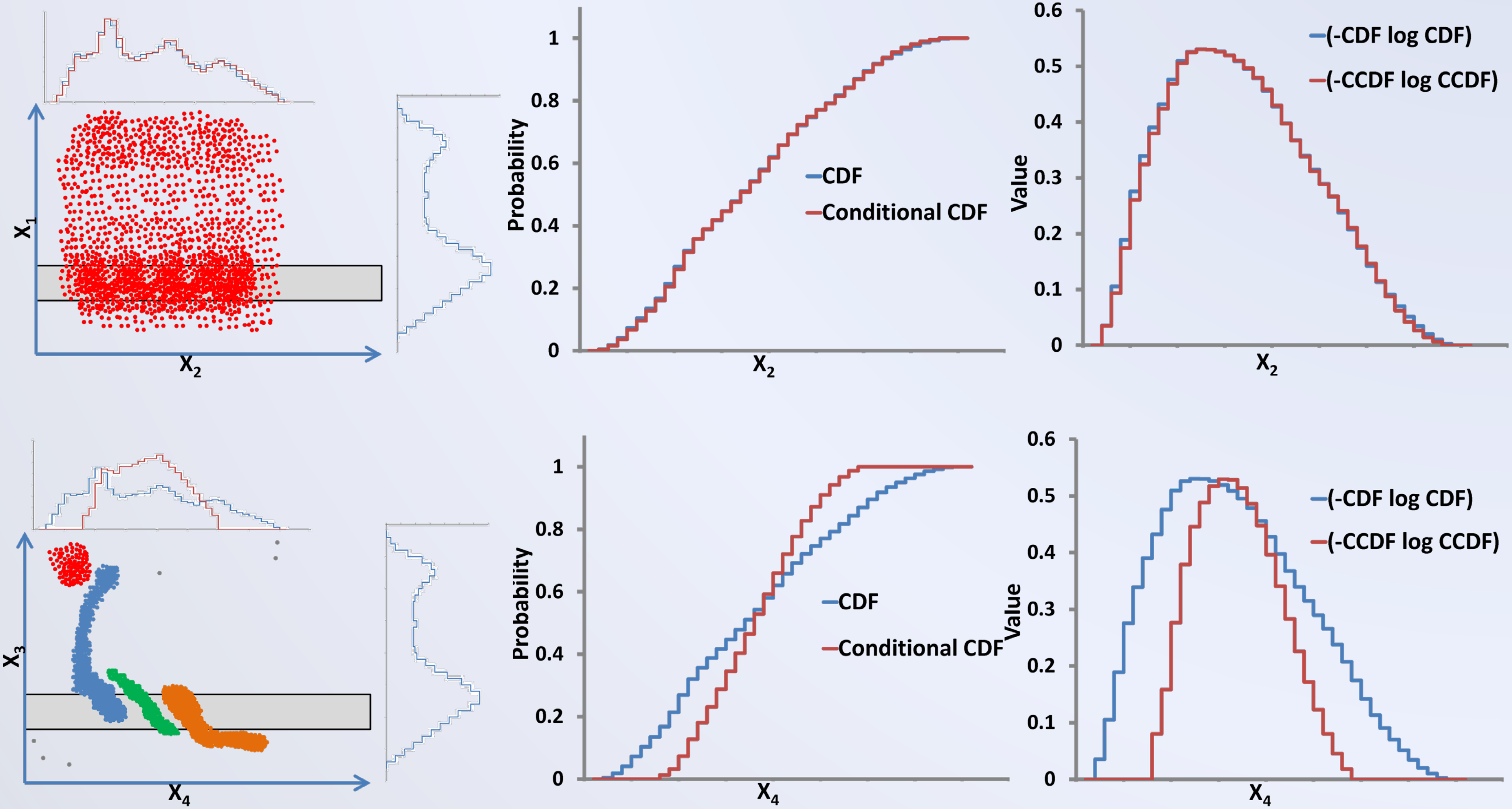
Given continuous valued data
over a set of attributes X we want to identify

$$Y \subset X$$

such that Y has high mutual information.

Can we do this with cumulative entropy?

Identifying Interacting Subspaces



Multivariate Cumulative Entropy

First things first. We need

$$h_{CE}(X \mid y) = \int h_{CE}(X \mid y) p(y) dy$$

which, in practice, means

$$h_{CE}(X \mid Y) = \sum_{y \in Y} h_{CE}(X \mid y) p(y)$$

with y just some datapoints, and $p(y) = \frac{|y|}{n}$

How do we choose y?

such that $h_{CE}(X|Y)$ is **minimal**

Entrez, CMI

We cannot (realistically) calculate

$$h_{CE}(\{X_1, \dots, X_m\})$$

in one go

but…

Mutual Information has this nice factorization property…

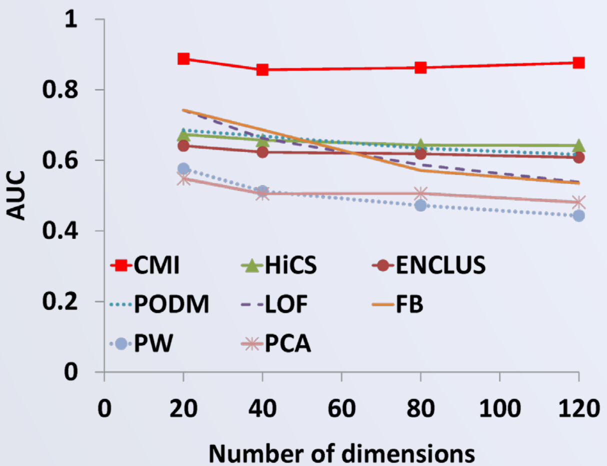
so, what we can do is

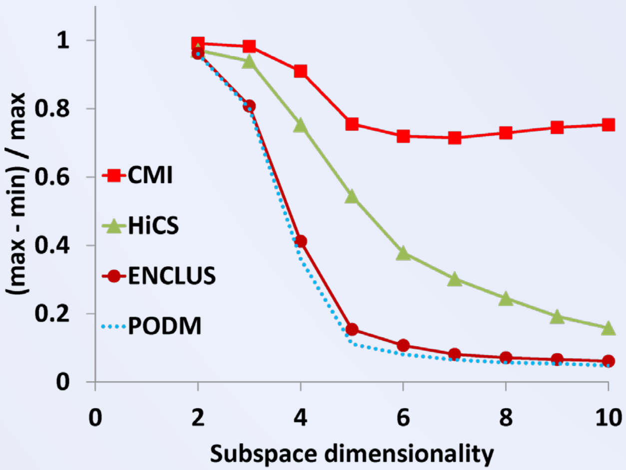
$$\sum_{i=2} h_{CE}(X_i) - \sum_{i=2} h_{CE}(X_i | X_1, \dots, X_{i-1})$$

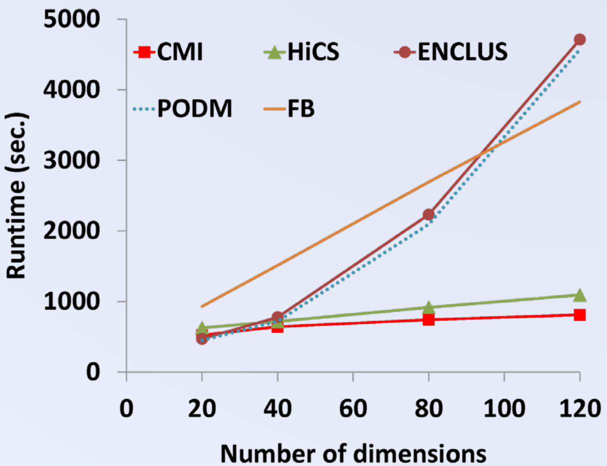
The CMI algorithm

super simple:
a priori-style

CMI in action







Conclusions

Information is about uncertainty of what you *could* say

Entropy is a core aspect of information theory

- lots of **nice** properties
- optimal prefix-code lengths, mutual information, etc

Entropy for continuous data is… more tricky

- *differential* entropy is a bit problematic
- **cumulative** distributions provide a way out,
  but are mostly unchartered territory

Thank you!

Information is about uncertainty of what you *could* say

Entropy is a core aspect of information theory

- lots of **nice** properties
- optimal prefix-code lengths, mutual information, etc

Entropy for continuous data is… more tricky

- *differential* entropy is a bit problematic
- **cumulative** distributions provide a way out,
  but are mostly unchartered territory