A Computational Theory of Subjective Probability

Phil Maguire, Philippe Moser, Rebecca Maguire and Mark Keane
NUI Maynooth - National College of Ireland

Introductory experiments

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

- Linda is a bank teller
- Linda is a bank teller and is active in the feminist movement

Introductory experiments

Can you spot the 2 Euromillions draws from the 5 sequences below ?

8	10	22	29	47
4	6	8	41	48
10	12	12	20	40
3	22	25	32	39
9	18	27	30	35

Introduction

- The mathematical concept of probability, originally formulated to describe the highly constrained environment of games of chance.

- The default assumption that probability theory provides the only logical way for people to think about likelihood.

- Tverksy and Kahneman applied probability theory to real-world situations and observed consistent deviations from the mathematical theory

The problem (😉 □)

- Is there a serious flaw in human reasoning ?

  or do

- Consistent deviations between human reasoning and a simplified, artificial mathematical theory are far more likely to reflect deficiencies in the theory than they are to reflect sub-optimality in how people think about likelihood ?

The intuition (😉 □)

- Classical probability theory only applies to cases involving a definitive probability measure function, while models of reality always involve uncertainty.
- The signal people rely on to diagnose discrepancies between their model and the real world is randomness deficiency.
- When people speak intuitively about likelihood and probability, it is the concept of representational updating which is relevant to them.

The solution (😉 □)

- Extend probability theory to situations involving an uncertain probability measure function
- The optimal model which can be derived from a set of observations is the one which maximizes the compression of that dataset, yielding the Minimum Description Length (MDL)
- => shifting the focus from an underdetermined probability measure function to the immutable mechanism of representational updating.

Main points

- Introduction to MDL Principle
    - The fundamental idea
    - Kolmogorov complexity and ideal MDL
    - MDL and model selection
- Subjective information and probability
- Experiments and discussion
- Proof that the conjunction effect is not a fallacy
- Conclusion

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- 00010001000100010001 . . . 0001000100010001000100010001
- 01110100110100100110 . . . 1010111010111011000101100010
- 00011000001010100000 . . . 0010001000010000001000110000

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- for i = 1 to 2500; print ”0001”; next; halt
- 01110100110100100110 . . . 1010111010111011000101100010
- 00011000000101010000 0 . . . 0010001000001000000001000110000

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- for i = 1 to 2500; print ”0001”; next; halt
- print ”011101001101000010101010...1010111010111011000101100010”; halt
- 00011000001010100000 . . . 0010001000010000001000110000

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- for i = 1 to 2500; print ”0001”; next; halt
- print ”01110100110100001010101...10101110101110110001011000010”; halt
- can be compressed to some length αn, with $0 < \alpha < 1$

MDL : Fundamental idea

Learning as Data Compression

- π
- Physics Data
- Natural Language
- ...

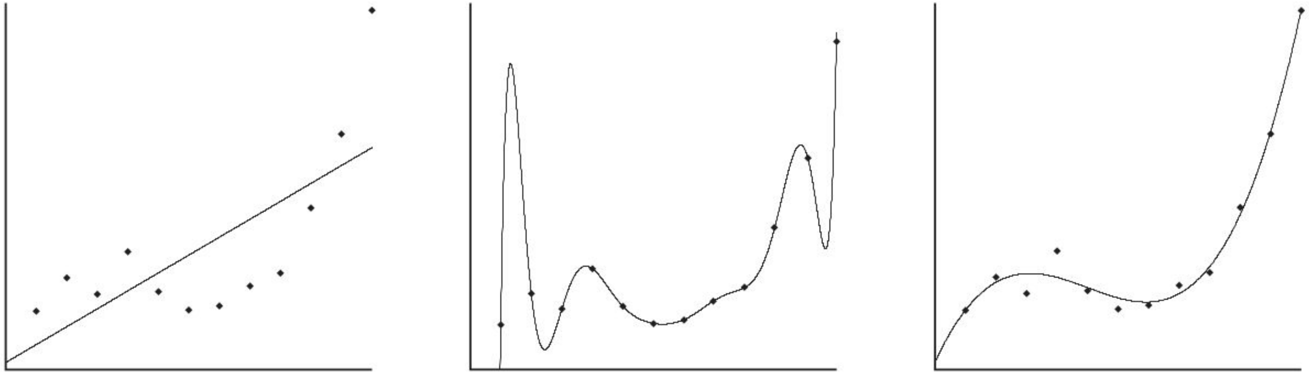
Kolmogorov Complexity and Ideal MDL

- KG of a sequence as the length of the shortest program that prints the sequence and then halts.
- Caveats:
  - Uncomputability
  - Arbitrariness/dependence on syntax
- Workaround : scale down the approach so that it does become applicable.

MDL and model selection

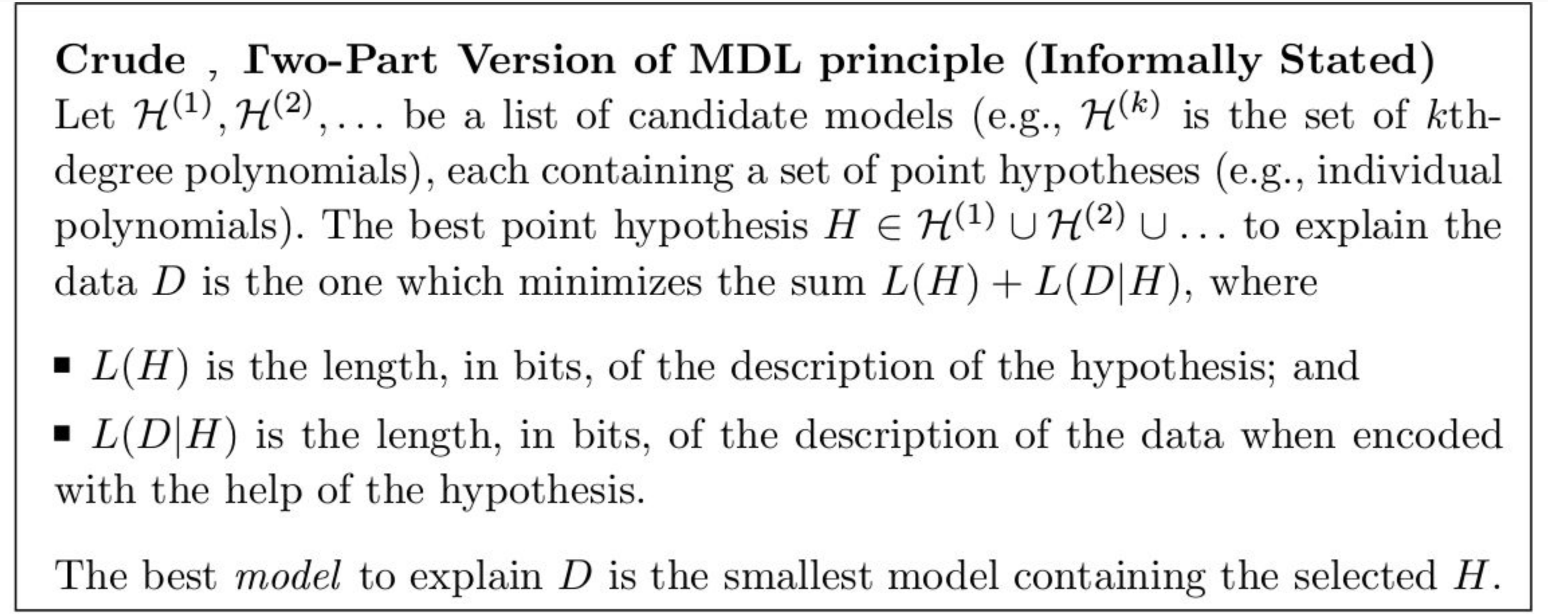
- The goal of statistical inference may be cast as trying to find regularity in the data.
- For a given set of hypotheses H and data set D, we should try to find the hypothesis or combination of hypotheses in H that compresses D most.

MDL and model selection



A simple, complex and tradeoff (third-degree) polynomial.

MDL and model selection



Crude , Γwo-Part Version of MDL principle (Informally Stated)
Let $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots$ be a list of candidate models (e.g., $\mathcal{H}^{(k)}$ is the set of kth-degree polynomials), each containing a set of point hypotheses (e.g., individual polynomials). The best point hypothesis $H \in \mathcal{H}^{(1)} \cup \mathcal{H}^{(2)} \cup \ldots$ to explain the data D is the one which minimizes the sum $L(H) + L(D|H)$, where

- $L(H)$ is the length, in bits, of the description of the hypothesis; and
- $L(D|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis.

The best *model* to explain D is the smallest model containing the selected H.

Peter Grünwald, Center of research in computer science and mathematics, NL

Subjective information and probability

Given a computable probability density function p, there are some “type of strings” we expect to be output, whereas some others are surprising.

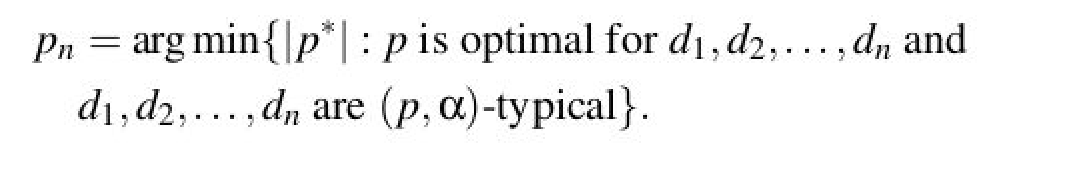
let α > 0 be a constant, called the surprise threshold, which represents the level of randomness deficiency that necessitates representational updating.

$$K(x|p^*) \geq -\log p(x) - \alpha.$$

$$K(x) = K(p) - \log p(x) \pm O(1)$$

Subjective information and probability

Suppose an observer experiences observations d_1 , d_2 , . . . generated by some source with computable probability density.



$$p_n = \arg\min\{|p^*| : p \text{ is optimal for } d_1, d_2, \ldots, d_n \text{ and } d_1, d_2, \ldots, d_n \text{ are } (p, \alpha)\text{-typical}\}.$$

Subjective information and probability

Observation d_{n+1} is α-surprising if the length of its shortest description given p is less than the number of bits a Shannon-Fano code based on p would require after subtracting the surprise level α :

$$K(d_{n+1}|p_n^*) < -\log p_n(d_{n+1}) - \alpha.$$

The subjective information of d_{n+1} is therefore : $K(p_{n+1}^*|p_n^*).$

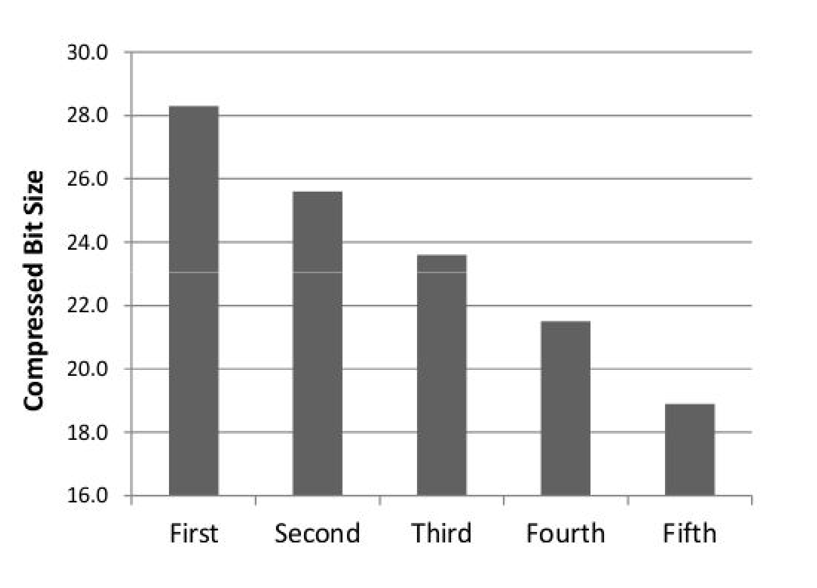
We can thus workout the subjective probability of d_{n+1} : $2^{-K(p_{n+1}^*|p_n^*)}.$

Experiments and discussion

Experiment 1 Hypothesis : people use subjective probability rather than classical probability to judge the likelihood for real-world events.

Method : Distractor sequences met the constraints of having compressed bit-sizes of between 23, 21 and and 19 bits while the average lottery length is 30.9 bits. True sequences were 28 and 26 bits long.

Experiments and discussion



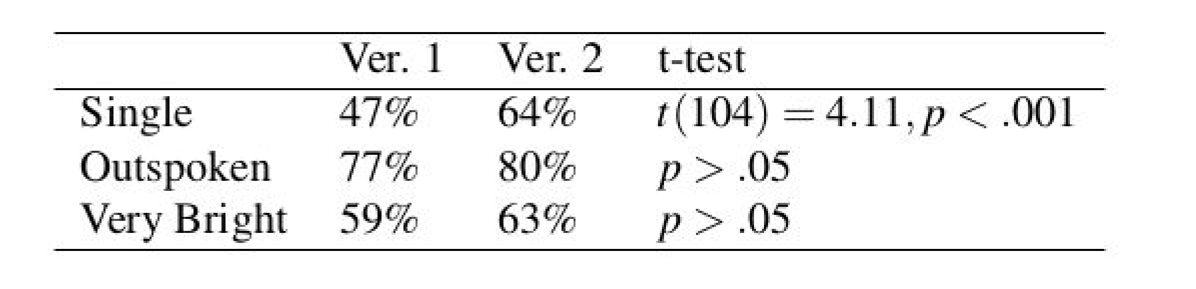
130 participants, correlation between ranking and compressed description was 0.965 with p < .001

Experiments and discussion

Experiment 2 Hypothesis : Are the people right about Linda ?

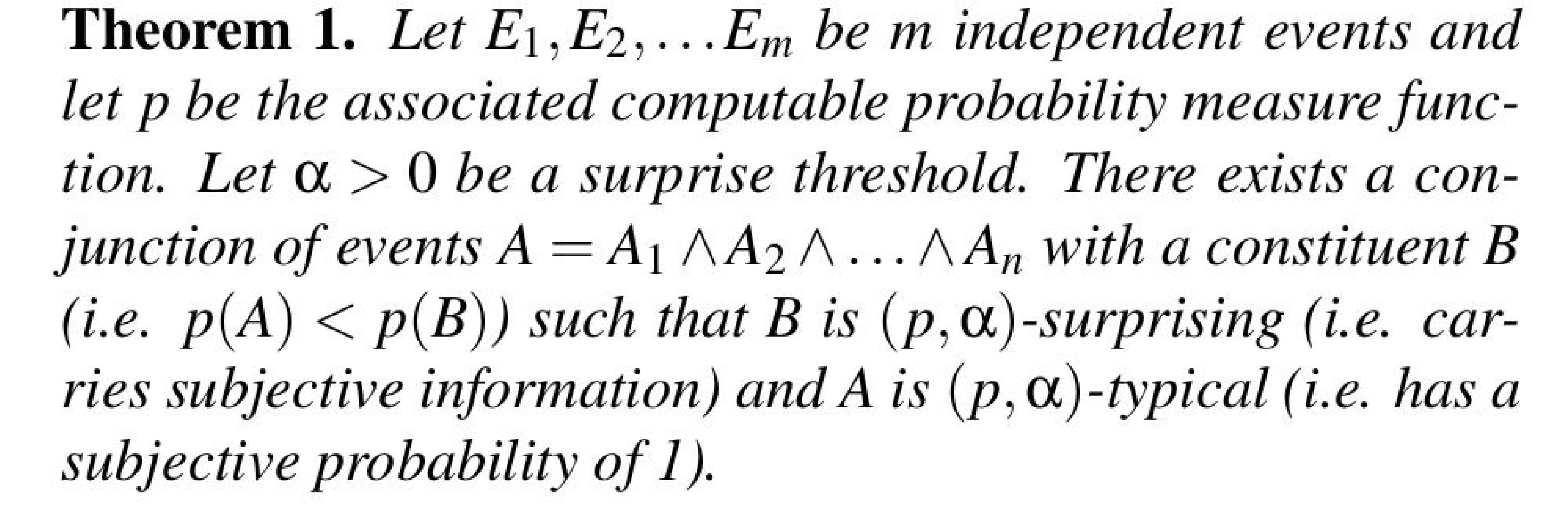
Method : Outcomes included in the description in 2 versions, removed the information that Linda is very bright, single and outspoken.

Results :



	Ver. 1	Ver. 2	t-test
Single	47%	64%	$t(104) = 4.11, p < .001$
Outspoken	77%	80%	$p > .05$
Very Bright	59%	63%	$p > .05$

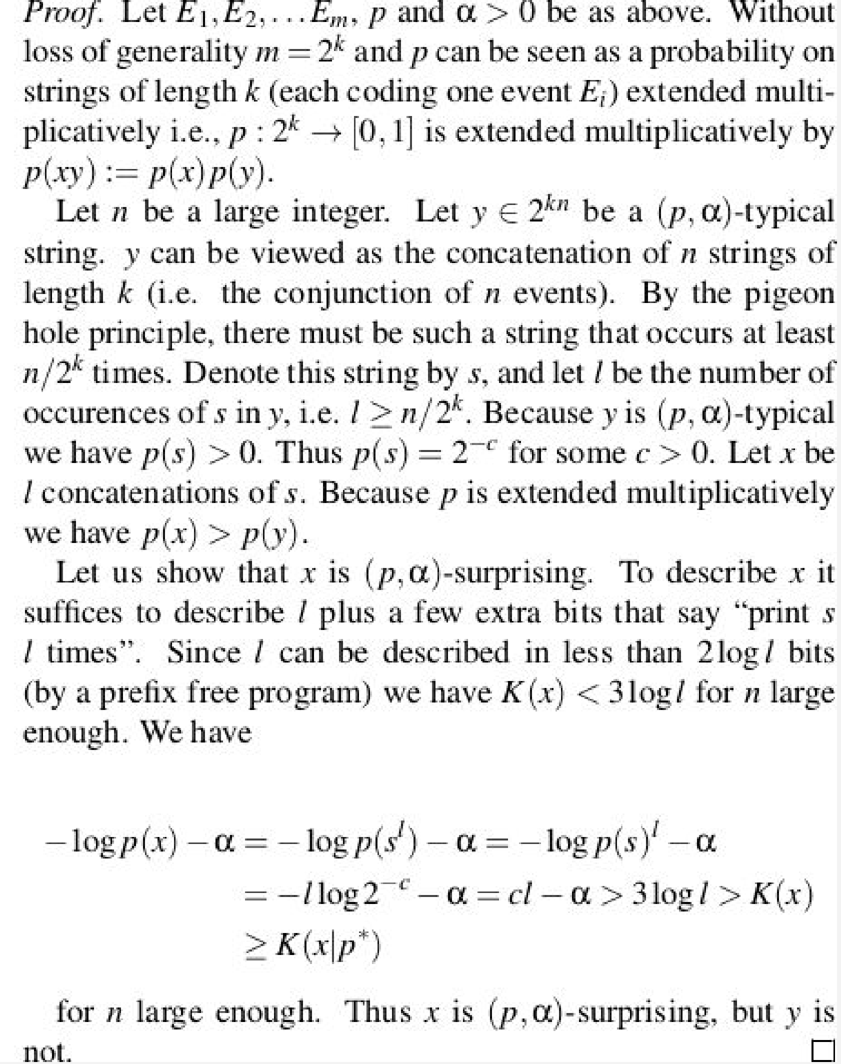
The conjunction effect is not a fallacy



Theorem 1. *Let $E_1, E_2, \ldots E_m$ be m independent events and let p be the associated computable probability measure function. Let $\alpha > 0$ be a surprise threshold. There exists a conjunction of events $A = A_1 \wedge A_2 \wedge \ldots \wedge A_n$ with a constituent B (i.e. $p(A) < p(B)$) such that B is (p, α)-surprising (i.e. carries subjective information) and A is (p, α)-typical (i.e. has a subjective probability of 1).*

Conclusion

Mathematical theories which have been developed for precision models in the exact sciences retain their validity when used to describe complex cognition in the real world.



Proof. Let $E_1, E_2, \ldots E_m$, p and $\alpha > 0$ be as above. Without loss of generality $m = 2^k$ and p can be seen as a probability on strings of length k (each coding one event E_i) extended multiplicatively i.e., $p : 2^k \to [0,1]$ is extended multiplicatively by $p(xy) := p(x)p(y)$.

Let n be a large integer. Let $y \in 2^{kn}$ be a (p,α)-typical string. y can be viewed as the concatenation of n strings of length k (i.e. the conjunction of n events). By the pigeon hole principle, there must be such a string that occurs at least $n/2^k$ times. Denote this string by s, and let l be the number of occurences of s in y, i.e. $l \geq n/2^k$. Because y is (p,α)-typical we have $p(s) > 0$. Thus $p(s) = 2^{-c}$ for some $c > 0$. Let x be l concatenations of s. Because p is extended multiplicatively we have $p(x) > p(y)$.

Let us show that x is (p,α)-surprising. To describe x it suffices to describe l plus a few extra bits that say "print s l times". Since l can be described in less than $2\log l$ bits (by a prefix free program) we have $K(x) < 3\log l$ for n large enough. We have

$$\begin{aligned} -\log p(x) - \alpha &= -\log p(s^l) - \alpha = -\log p(s)^l - \alpha \\ &= -l \log 2^{-c} - \alpha = cl - \alpha > 3\log l > K(x) \\ &\geq K(x|p^*) \end{aligned}$$

for n large enough. Thus x is (p,α)-surprising, but y is not. □