The Importance of Being Ignorant

Using Entropy for Interpretation and Inference

R Nityananda

In many real life situations, we have to draw conclusions from data which are not complete and have been affected by measurement errors. Such problems have been addressed from the time of Bayes and Laplace (late 1700's) using concepts which parallel Boltzmann's use of entropy in thermal physics. The idea is to assign probabilities to different possible conclusions from a given set of data. A critical – and sometimes controversial – input is a 'prior probability', which represents our knowledge before any data are given or taken! This body of ideas is introduced in this article with simple examples.

Rajaram Nityananda works in areas of physics and astronomy involving optical and statistical concepts. After increasing entropy at the Raman Research Institute in Bangalore for more than two decades, he moved recently to the National Centre for Radio Astrophysics of the Tata Institute of Fundamental Research at Pune.

From the earliest times, thinkers have recognised two distinct ways of learning about the world we live in. Our educational system gives prominence to the first one – 'deduction'. The best example is of course Euclid's construction of geometry from a few innocent looking axioms. In the world of fiction, Sherlock Holmes claimed to 'deduce' what had really happened in a crime from a few clues. But in reality, what most of us (Sherlock Holmes included) practise, should be called 'induction'. Logicians have given this name to drawing conclusions from observations or experiments by a rather different process. To start with, we have a large number of possible hypotheses to choose from. Observations and experimental data are used to narrow down the possibilities. The word 'hypothesis' is being used in a rather simple sense here. For example, if we are trying to determine the elliptical orbit of an asteroid, the 'hypothesis' is just a set of numbers giving the plane of the orbit, the size and shape and orientation of the ellipse in this plane,

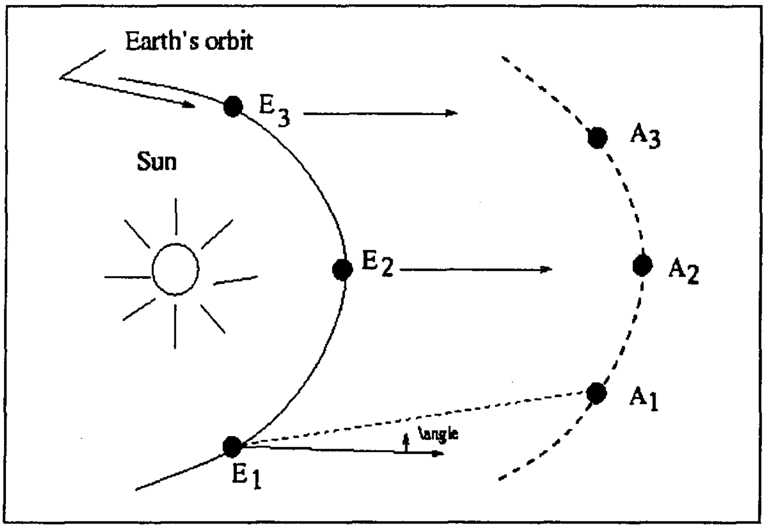

Figure 1. The (unknown) orbit of the asteroid is shown by dashed lines. At three different times t_1, t_2, t_3, observations give the three directions (but not distances) of E_1A_1, E_2A_2, and E_3A_3. The earth's orbit $E_1E_2E_3$ is assumed known.

and where the asteroid sits on the ellipse at a given time. We do not directly measure these numbers but rather the angular position in the sky at different times, as seen from the earth which is itself a moving platform. The situation is illustrated in *Figure* 1.

Gauss faced precisely this problem of orbit determination in the year 1801. A few observations of Ceres, the very first asteroid discovered, were available. He invented the so called 'method of least squares' to choose the best orbit consistent with the measurements available. We now explain how his method fits in with our earlier general discussion. To simplify matters, we will assume, as in *Figure* 1, that the two orbits, of earth and asteroid, lie in a plane. We show in *Figure* 2 two kinds of graphs. One, made up of individual points, gives the observations. The continuous curves, give the predictions of different possible orbits (i.e., hypotheses).

Gauss invented the so called 'method of least squares' to choose the best orbit consistent with the measurements available.

Our first reaction is that it needs only four numbers to specify the orbit in the plane. These could be the x and y coordinates of the asteroid, and the x and y components of its velocity, at a given time (January 1, 1801, for example!). Four measurements ought to be enough, and we should be able to *deduce* the orbit without guesswork.

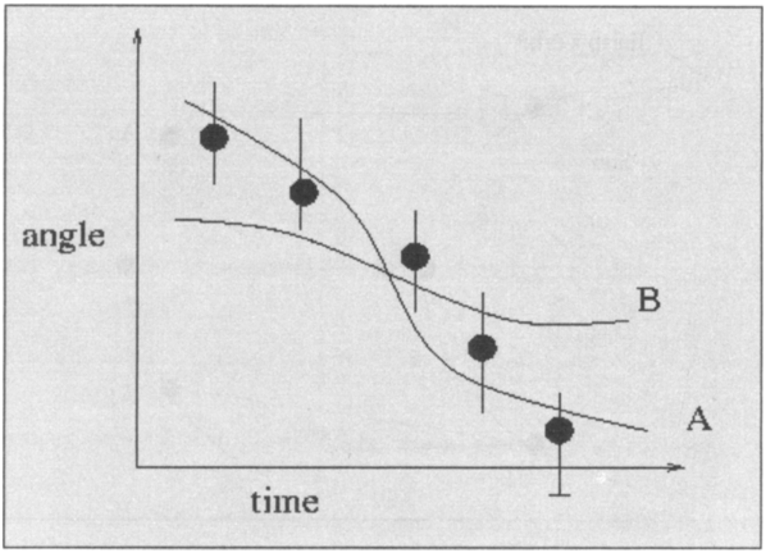

Figure 2. The points show the actual observations. The continuous curve A shows what we might regard as the best orbit. B is another orbit at a greater distance than A. The vertical lines through the observed points represent errors of measurement.

But Gauss, although the prince of mathematicians, also knew the real world better. Measurements are never exact, and the points would not lie exactly on the predicted curve even if we knew the orbit! We can state this in another way. For each measurement, we can draw a vertical bar which represents the possible range in which the true value (of the angle) could lie. Each point has now become 'fuzzy' or 'blurred' in the vertical direction (The measurement along the x-axis, viz time, is usually very accurate and we do not worry about its errors here.)

Now we can readily see that there is a corresponding fuzziness or uncertainty in the curve drawn though the points. We have moved from deduction to induction. Other names for this process are 'inversion' (going back from the data to the hypothesis) and 'statistical inference.'

Measurements are never exact, and the points would not lie exactly on the predicted curve even if we knew the orbit!

Going back to *Figure* 2, why do we choose the curve A rather than the curve B? An experimenter would say that 'the deviations of curve A from the measurements are consistent with the error bars, while curve B lies well outside the error bars.

Now let us try and be more quantitative. Each error bar is really not a line with sharp limits. Larger errors are less probable, but not impossible. In fact, Gauss himself, building on the work of de Moivre and Laplace, proposed that the probability for the error to be x falls off proportionally to exp $\left(\frac{-x^2}{2\sigma^2}\right)$. This is the bell shaped graph sketched in *Figure* A. *Box* 1 gives a few more details about this remarkable, widespread distribution which we all call gaussian. The basic message of *Box* 1 is that the error is itself the sum of many smaller contributions each of which may not have a gaussian dis-

Box 1. The Gaussian Distribution

A coin is tossed eight times. What is the most probable number of heads? Four of course. Why is eight heads less probable than four? Because there is only one way to get eight heads, HHHHHHHH. But there are ${}^8C_4$=70 ways to get four heads, since we now have freedom to choose any four of the eight tosses to show heads. The full table of numbers is

No. of heads	0	1	2	3	4	5	6	7	8
No. of cases	1	8	28	56	70	56	28	8	1

and they are plotted in *Figure* A

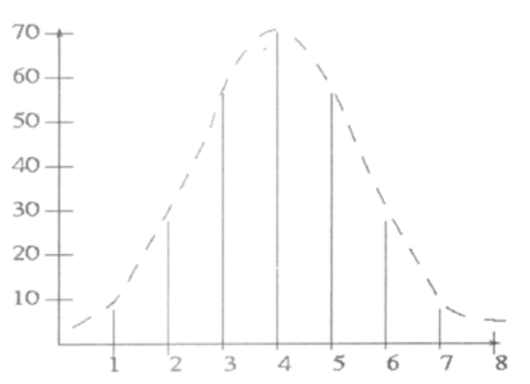

Figure A.

To get the probability, we have to divide by 256. We have also superposed a bell shaped curve. This is how the probability for n heads behaves when the number of tosses is very large (of course, we have to relabel the axes if we have 158 tosses instead of 8!). This is the famous gaussian distribution. Its mathematical form is

$$P(x) = A \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

A is a constant of proportionality.

$x = m$ is the peak of the curve and also the average value of x. σ^2 is a constant which is called 'variance'. It measures the average of the square of the deviation of x from m.

tribution. But the sum does approach this law in many cases. We can think of the height of the gaussian as measuring the number of ways that a given error could be built up from the underlying individual contributions ("errorlets'?). The logarithm of this number is, therefore, proportional to

$$\log(\exp\left(-x^2\sigma_i/2\sigma_i^2\right)) = \text{constant } - \frac{x_i^2}{2\sigma_i^2}.$$

σ_i^2 is the average of the square of x_i. Why do we take the logarithm? This is a convenient thing to do when we want to multiply numbers! Come back to our original problem of determining the best orbit (*Figure* 2). When we guess a given curve, A or B, we are automatically attributing the deviations of the points from the curve to experimental error. So we should be asking ourselves – 'What is the probability that the errors took the values that we are suggesting?' This probability is obtained by multiplying gaussian functions for the individual errors at each measured point. We now want to maximise the joint probability, i.e., the product of probabilities. So we maximise the logarithm, which is

$$\log(\text{Probability of errors}) =$$

$$\text{const} + \text{another const}\left(-\sum_i \frac{x_i^2}{2\sigma_i^2}\right).$$

In the simple case where all the $\sigma_i's$ are equal, this means we have to *minimise* the sum of the squares of all the errors (because of the negative sign in front of it). This is the famous method of least squares, and it is eminently sensible. It prevents us from doing silly things like drawing the theoretical graph well away from the points. It ensures that errors have both signs. But let us remember that least squares is not sacred or perfect. It is only as good as the assumptions that went into it. When the errors do not have a gaussian distribution, or

But let us remember that least squares is not sacred or perfect. It is only as good as the assumptions that went into it.

when we have some physical limits which restrict our orbit, we can, and must do better. Our example was really meant to introduce a broader framework for hypothesis testing.

This broader framework came even before Gauss. It is attributed to Bayes and Laplace, who worked in the late 1700's.[1] The basic ('Bayesic?') idea is to use a simple theorem of conditional probability due to Bayes (*Box* 2). We need it in the form.

We should warn the reader that there are many other approaches to statistical inference. The Bayesean approach of this article uses concepts closest to entropy.

Box 2. Bayes' Theorem for Conditional Probabilities

One way of understanding this theorem is via *Figure* B in which points stand for events and areas stand for probabilities.

The horizontally striped region A represents all cases or trials in which some event a occured. The vertically striped region B similarly stands for all instances of b. The intersection C of these two regions is cross hatched and represents cases where both a & b occured. We can now say

Area of $C = p(a,b)$ = joint probability of a and b
Area of $A = p(a)$ = probability of a
Area of $B = p(b)$ = probability of b

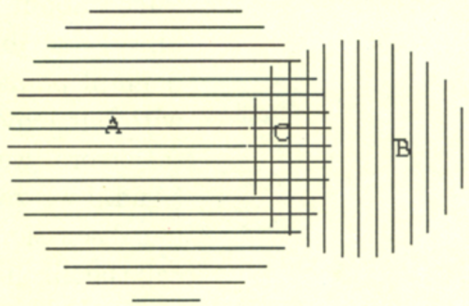

Figure B. Venn diagram illustrating Bayes' theorem.

Conditional probability of a given that b has occured $= p(a\,|b)$

$$= \frac{\text{area of } C}{\text{area of } A} = \frac{p(a,b)}{p(a)}.$$

Hence, $p(a,b)=p(a \mid b)\,.\,p(a)$. Similarly, $p(a,b) = p(b \mid a)\,p(b)$.
Hence, equating these two,

$$p(a \mid b) = \frac{p\,(b \mid a)\,p(a)}{p(b)}.$$

Since the left side is a function of a for fixed b, we can treat the denominator as a constant, as we have in the main text.

Probability of H (given D) states the goal of all experimental science, viz., we are given data, and we try to assign probabilities to different hypotheses based on this data.

Probability of H(given D) $\propto$ Probability of D (given H) multiplied by Probability of H (not given anything.)

Our choice of notation is deliberate. H stands for hypothesis, D stands for data. The left hand side states the goal of all experimental science, viz., we are given data, and we try to assign probabilities to different hypotheses based on this data. That is what the notation P(H| D) means. The right hand side of our equation tells us how we are to achieve this goal. It has two factors. The first one is the conditional probability P(D|H). In words, given a hypothesis (orbit in our earlier example) what is the probability that the given data could arise (e.g., angle measurements of the asteroid)? We have already talked about this when we multiplied gaussian (probability) distributions for the errors at the various experimental points. In general, if we know how to predict with our hypothesis and we understand our experimental errors, we should have no difficulty with P (D|H). (And if we don't the first priority is to do so!). Our earlier discussion stopped at P(D|H) – which statisticians call the 'likelihood function' when regarded as a function of H – for fixed D. Of course, it is an honest probability distribution for D, when H is fixed.

But the rules of probability tell us that this is not enough. We have to face up squarely to the second factor on the right side P(H). This is the *unconditional* probability that a particular hypothesis H is true. Since this has nothing to do with the data, it is called the 'prior' distribution. Perhaps the philosophy of Kant shaped this terminology. He believed that some things like space and time had to be given to us 'a priori', right at the beginning. We already had some kind of prior distribution in mind in our orbit problem. We only drew curves like A or B which were based on Newton's laws of motion and gravitation, and did not try others. Using P(H) to reject what we know to be impossible even before the observations are taken, is a good idea. But P(H) also as-signs different weights to two hypotheses which are both possible to start with. This seems like introducing the experimenters prejudice into the interpretation of data! Hot debates continue on this point. The ghost of the prior has haunted Bayesian statistical inference from its birth. Laplace himself coined a 'principle of insufficient reason'. It was a way of making the prior a constant or flat function so as to be as even handed as possible. This is similar in spirit to our accepting 1/2 and 1/6 as the probabilities for coins and dice. But when we come to a continuous variable q, going from zero to one, do we say it has equal probabilities to be less than or greater than $\frac{1}{2}$? There is a trap here pointed out by Laplace's countryman Bertrand. Why not apply the same (insufficient!) reasoning to q^2, which goes from zero to one? We would then conclude the q would be less than 0.707 with probability $\frac{1}{2}$. Clearly one needs further input to decide on a prior in cases like this.

So far, we have just touched the fringes of entropy concept, when we looked at the logarithm of the number of ways that a given error could occur. But we are now prepared for the basic problem which faced Boltzmann when he investigated the theory of gases in the latter half of the nineteenth century. There is detailed discussion in the article by Bhattacharjee in this issue, and we only give the bare minimum needed for this story. Boltzmann would take the total energy and total volume of a gas as given – these correspond to the data set D. Let us think of the detailed position (x) and velocity (v) information of all the molecules as our hypothesis H. Boltzmann (and his great American contemporary, Gibbs) divided the space of x and v into cells of equal volume, measured by the product $dxdv$. Notice that he singled out x rather than x^3, v rather than v^5. This prior was based on his analysis of the dynamics of collisions between molecules. The rest is history. He was able to deduce Maxwell's probability distribution law for the

In modern quantum language, one can state Boltzmann's prior in a physically appealing way. Every single energy level of the whole system gets equal probability to start with.

molecular velocity components v_x, v_y, v_z. [2] Even better, he was able to show how collisions would produce such a distribution even if it was not present to start with. These results were in full agreement with experiments, both earlier and later. In modern quantum language, one can state Boltzmann's prior in a physically appealing way. Every single energy level of the whole system gets equal probability to start with. While Boltzmann chose the volume in $x - v$ space based on classical collisions, today we know that this is equivalent to counting energy states in quantum theory.

We now move forward about half a century to 1948. Stimulated by rapid advances in electronics, one of the best telephone systems in the world was established in the United States. Many of the new developments came from the Bell Telephone Laboratories (Bell Labs for short) and were published in the *Bell System Technical Journal.* Claude E Shannon, a young researcher at Bell, contributed two papers on the Mathematical Theory of Communication. His deep insight was to introduce a quantitative measure of the amount of information being communicated. After all, this information is what we really pay the telephone company for! If you receive a message from someone in the English language, you already know the approximate fraction of E's, T's, A's, etc. Let us say there are a hundred letters in a telegram. There is a large number, W, of possible English messages with a hundred letters.[3] You open the telegram and find out which one of the W is your message. Shannon proposed that the information gained be measured by $S = \log_2 W$. The reason for taking the logarithm is the same as earlier. Two successive telegrams (on unrelated subjects!) would correspond to $W_1 \times W_2$ possible messages. Shannon's measure ensures that the information (and perhaps your telegram bill!) is additive, i.e., $S = \log_2 W_1 + \log_2 W_2 = S_1 + S_2$.

This is related to Boltzmann's entropy. He would call

[2] Interestingly, this is a product of three gaussian distributions for v_x, v_y, v_z.

[3] If all the letters were equally likely to occur, this number would be $W = 26^{100}$.

S a measure of your ignorance before you opened the telegram, rather than your enlightenment after you opened it. But it is sensible to take the two quantities as equivalent.

Why choose 2 as the base of logarithms? The simplest situations are when the message simply says which of two (equiprobable) options was realised. When the nurse steps out of the maternity ward and tells the anxious father 'Its a girl', $W=2$, and $S=1$. This is called one bit of information. Everyone in this computer age knows that 'bit' is short for 'binary digit', something which takes 2 values, zero and one.

Shannon's concept of information took the world by storm. There was tremendous enthusiasm to apply it to every field. An indignant journal editor even wrote the following lines – "We will no longer consider papers with titles like information theory, photosynthesis, and religion"!

We have presented Shannon's work in conjunction with the ideas of Bayes and Boltzmann. This attempt at complete synthesis actually came a few years after Shannon, in the influential work of the physicist Edwin T Jaynes. He and his followers have explored the application of 'maximum entropy' (as they call this approach) to a variety of practical problems. Both Shannon and Jaynes

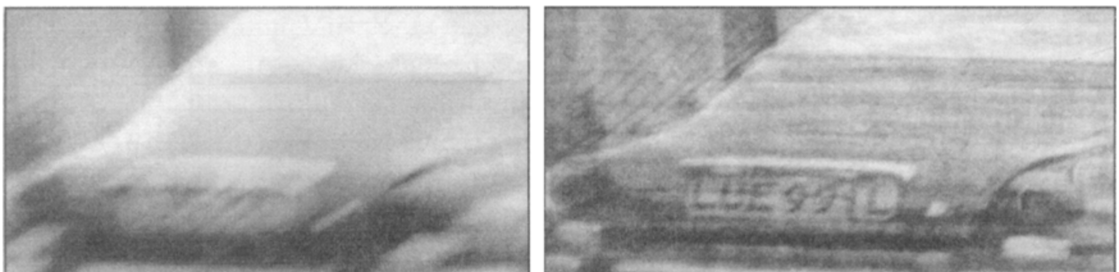

Figure 3. The photograph shows the effect of applying maximum entropy deconvolution to a motion-blurred picture. Processing by A Lehar and Maximum Entropy Data Consultants Ltd. for the UK Home Office. Our thanks to Steve Gull and his colleagues at the Mullard Radio Astronomy Observatories in Cambridge who were instrumental in developing maximum entropy methods for such problems.

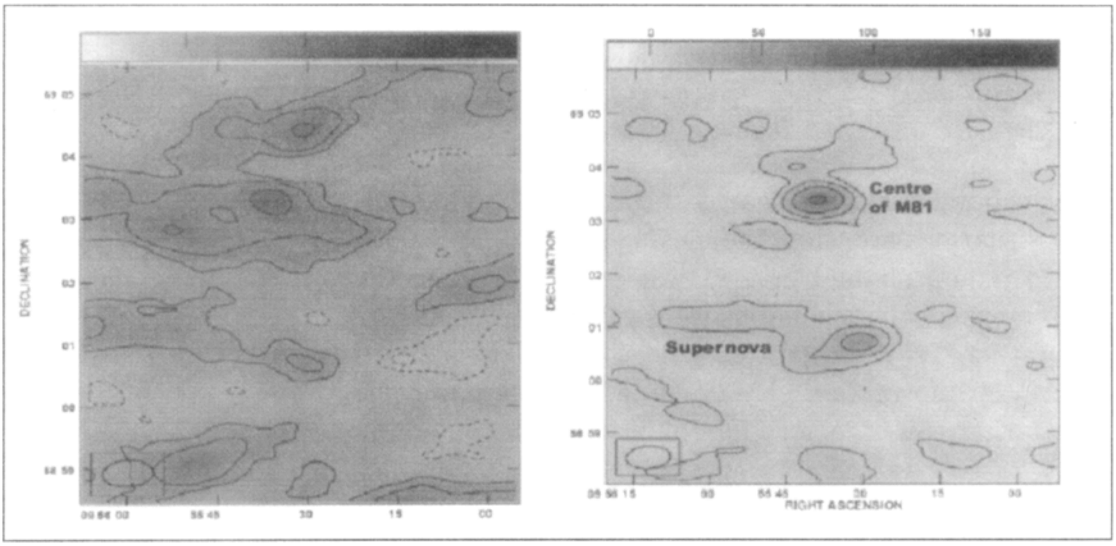

Figure 4.

died recently, living to see their ideas bear fruit over nearly half a century.

Although we cannot give details here, inversion based on maximum entropy methods is in wide use. A dramatic real life example (*Figure* 3) would be a blurred photograph of a car. In this context, the blurred photograph stands for the data D, while the reconstructed picture corresponds to H. After inversion, one is able to read the number plate clearly! An example of removing blurring in astronomy using prior information is given in Figure 4.

We must of course remember that maximum entropy is not a magic wand. The fact that we are able to read the number plate means that the information in the data (blurred photograph), plus the information in the prior, were enough to recover what we were looking for. In a given problem, there is usually a range of possible priors which would be regarded as reasonable. Most workers would regard results which are insensitive to choices in this range as genuine. When results start becoming sensitive to the prior, it is time to go out and get more data or work on a different problem.

Address for Correspondence
R Nityananda
National Centre for Radio Astrophysics
Post Bag 3, Ganeshkhind
Pune 411 007, India