Data analysis

The maximum entropy method

from John Skilling

MAXIMUM entropy is a remarkably powerful and general technique of data analysis. It can be used to sharpen up out-of-focus photographs (as in Fig. 1), to make maps of the radio sky, to generate images from medical scanners, to calculate spectra, to reconstruct electryn densities in a crystal and even to determine the positions of fuel rods in a nuclear reactor cooled with liquid sodium. It claims to be applicable whenever one needs to estimate a single vector of proportions $p = (p_1, p_2, \ldots, p_M)$ from seriously incomplete data, which could be fitted equally well by many different vectors. It claims to ignore what the proportions represent, which is why it can be applied to so many different inference problems. Maximum entropy always selects the simplest possible result, containing the bare minimum of structure needed to fit the data. Spurious detail is reduced as far as possible, which doubtless accounts for the practical success of the method.

The principle is always the same: one maximizes the entropy $S = -\Sigma_i p_i \log p_i$ subject to whatever constraints are imposed by the data. The theoretical basis of the method has often been less clear than the practical results however. Now papers by Shore and Johnson (*IEEE Trans. Inform. Theory* IT-26, 26; 1980 and IT-29, 942; 1983) and by Tikochinsky, Tishby and Levine (*Phys. Rev. Lett.* 52, 1357; 1984) are making the justification for the maximum entropy method both clear and compelling.

Historically, the use of maximum entropy in data analysis has often been justified by supposing that the proportions were quantized, so that $p_i = n_i/N$. If N quanta were placed 'at random' into M available locations, the relative frequency of sets of occupation numbers n would be degeneracy $N!/\Pi n_i!$, which is equivalent to $\exp(NS)$, at least if N is large. Thus maximum entropy would be the same as maximum degeneracy, which could in turn be identified with maximum probability. Partly because of the dubious status of N and the quantization, this derivation has never been entirely convincing to scientists at large. An alternative justification has been to eschew quantization and identify S directly with (minus) the Shannon information content of p, considered as a vector of probabilities. One can then talk about minimizing the configurational information of p; but once again the rationale has not proved universally convincing.

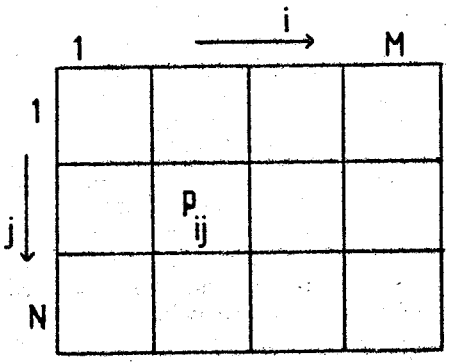

Fig. 2 For definition of i, j, M, N and p_{ij}, see text.

The new analyses reach a deeper level that bypasses the earlier derivations, and show that maximum entropy is the only consistent selection procedure. Let us start with Tikochinsky *et al.*, and with a diagram (Fig. 2), in which p_{ij} represents the probability of outcome i when performing experiment j. Tikochinsky *et al.* consider 'reproducible' experiments in which p_{ij} is independent of j, and in which incomplete data are acquired in the form of linear combinations $\langle A \rangle = \Sigma_i A_i p_i$. Because the data are incomplete (fewer than M combinations), the values of $\langle A \rangle$ do not fully determine the values of p. But if we must select just one single result, how should we go about it? Suppose the experiment is repeated N times. We are at liberty to treat each experiment independently. In this case, our selection algorithm (whatever it might be) will choose the same proportions p_i for each experiment j, and these could be combined to give an overall distribution P_i for the outcomes i after the N repetitions. We are also at liberty to treat the whole as one large experiment, in which case our selection algorithm will immediately produce an overall distribution Q_i. Tikochinsky *et al.* point out that we must surely require $P_i \equiv Q_i$ if data from reproducible experiments are to be combined consistently.

Repeating or averaging an experiment does not change its outcome, and it should not change the selected result either. In ten short equations, Tikochinsky *et al.* proceed to prove that the consistency requirements is satisfied if, and only if, the selection algorithm is maximum entropy. In a brilliantly simple way, entropy emerges as an elementary consequence of consistent reasoning.

The consistency approach can be taken even further. Tikochinsky *et al.* couch their analysis in terms of probability distributions, but it often seems a little artificial to have to identify a set of physical proportions with a set of probabilities of outcomes of a random experiment. Proportions and probabilities are, of course, isomorphic, but why should we need to introduce the idea of randomness at all? And why should we need to restrict ourselves to linear data? The answer to both questions is that we need not.

Let us return to Fig. 2, but let p_{ij} now represent the proportion of the total observed quantity assigned to cell i, j of the $M \times N$ array ($\Sigma_{ij} p_{ij} = 1$): we could be considering a two-dimensional digitized picture. Suppose that the data are some arbitrary functions $U(u)$ and $V(v)$ of the marginals $u_i = \Sigma_j p_{ij}$ and $v_j = \Sigma_i p_{ij}$. We are at liberty to ignore the 'north–south' v-structure, and use our selection algorithm to choose an 'east–west' distribution u_i which fits the data U. Similarly, we can select a north–south distribution v_j which fits the data V, without reference to U. Finally, we are at liberty to use all the data, U and V together, to produce immediately an overall distribution p_{ij}. We must surely require $p_{ij} = u_i v_j$ if data from independent experiments are to be combined consistently. North–south data must not interfere with east–west structure, and vice versa. This will be the case if, and only if, the selection algorithm is maximum entropy. Again, entropy emerges as an elementary consequence of consistent reasoning.

This is the derivation of entropy which we at Cambridge now prefer (see our rigorously informal presentation in *Indirect Imaging* ed. Roberts, J.A., Cambridge University Press; 1984, a more mathematical account of which has been

Fig. 1 What maximum entropy techniques can do for car-spotters, including the police.

submitted to *Acta crystallogr.* A). We obtained our derivation from Shore and Johnson's papers, in which the notion of independent dimensions is refined to a requirement of system independence, and the whole treatment is given a proper axiomatic foundation. Shore and Johnson give four axioms — uniqueness, invariance, system independence and subset independence — which must be satisfied by any consistent selection algorithm. In careful and formal work that is remarkably similar to an axiomatic derivation of Shannon information, Shore and Johnson derive the entropy formula, generalized to $S = -\Sigma_i p_i \log(p_i/m_i)$ to allow for a possible prior model m. Although they describe p as a probability distribution, this is unnecessary; their work applies equally well to any distribution of proportions, and the technical words 'probability' and 'information' need never appear.

These ideas justify the fundamental claims made for maximum entropy in data analysis, and it is clear that we need not quantify our preference for the maximum entropy selection by anything other than the numerical value of S. It is sufficient to know that we must use maximum entropy — or lay ourselves open to the charge of inconsistency. Let's get on with it. □

John Skilling is in the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 9EW.

Palaeobotany

Early evolution of leaves

from J.B. Richardson

THE origin and evolution of leaves was a major event in land plant evolution and must have affected all other life on land. Much remains to be learnt, however, of when they first appeared and how they evolved. Fossils described on page 785 of this issue by Gensel[1] confirm what had already been suspected from compression fossils from the Gaspé Peninsula: by the end of the Lower Devonian, vascular plant evolution was quite advanced with complex forms present. The presence of divided non-laminate leaves at the end of the Lower Devonian is already documented[2] but Gensel[1] describes changes in anatomy in a lateral branch system suggestive of the early evolution of a megaphyll — a laminated leaf with a complex pattern of venation.

Did leaves arise as Zimmermann suggests in his telome theory[3], through evolutionary overtopping, planation and webbing? A telome is the most distal dichotomy of a plant. Overtopping describes what occurs if one branch of the dichotomy develops more than the other. The overtopped branches would diminish in size, eventually tranforming the equal dichotomies into a main stem with lateral branches. Planation describes the change from a three-dimensional arrangement into a single plane of neighbouring telomes. Webbing, the joining of tissues between the lateral branches of the planar telomes, would result in the familiar lamina (leaf-like structure) with dichotomous venation.

This is what Zimmerman proposed but how is the theory testable? One of the restraints on evolutionary hypothesis is the geological record. Do these morphological features appear in the geological succession in the order demanded by the theory? The answer is both yes and no. Simple dichotomously-branched axes have been recorded in the mid-Silurian (Wenlock)[4], while by the late Silurian, rhyniopsids, with a similar appearance, were common but not diverse. Overtopping was present in the early Devonian (Lower Gedinnian)[5]. At that time there was considerable diversity of the rhyniopsids, and zosterophylls, distinguishable from the rhyniopsids on the basis of lateral sporangia, were present[6]. Later, during the Siegenian, the zosterophylls diversified[7] and the trimerophytes, such as *Dawsonites*, appeared. Later still, in the Lower Emsian of Belgium, a similar flora along with several *Psilophyton* spp. was present[8] and work by Andrews indicates a high diversity of plants at that time (see *inter alia* ref.9). Trimerophytes in the Emsian exemplify overtopping and reduction of laterals; as Stewart[10] states "Trimerophytopsida exhibits almost every branching pattern to be found in megaphyllous vascular plants". Thus, as re-emphasized by Gensel's paper, Zimmermann's leaf-forming processes, apart from planation and possibly webbing, had happened by the late Lower Devonian and much of the geological record supports the telome theory on the origin of megaphylls.

Zimmermann derived a great variety of leaves and leaf-like structures, including microphyllous 'leaves' of the lycopsida, from a rhyniopsid-like ancestor. Bower's enation theory[11] is an alternative proposal for deriving microphylls — small leaves with a single central vein. In the telome theory, the microphylls arise by reduction of bunches (trusses) of telomes. Bower's theory is that bumps (enations) emerged from the surface of the stems. As these enations extended (in the evolutionary sense) they developed vascular tissue, first at the base of the enation (as in *Asteroxylon*) and then throughout the enation, as in *Baragwanathia*, which thus became a true microphyll. According to the telome theory this series would be reversed: reduced vascularized microphylls would precede unvascularized outgrowths. Which

100 years ago

SOME forty years ago Dr. Joule raised the question whether a body that is magnetised ndergoes any changes in its temperature; but the question has not yet received a definite solution, the rise of temperature which accompanies magnetisation being ascribed by some to induction currents, and not directly to magnetism. While recognising the influence of the former, Mr. Borgman has tried to show that there is also a change of temperature due to the magnetism and demagnetisation, and that the amount of heat thus disengaged is proportionate to the squares of the temporary magnetism. M. Bachmetieff, having made, at the University of Zurich, an extensive series of experiments, the first part of which is now published in the *Journal of the Russian Chemical Society* (vol. xvi. fasc. 3), arrives at the conclusion that magnetism, *by itself*, produces variations of temperature in magnetised bodies, and that this "magnetic heat" is equal to the product of the magnetic moment by the magnetising force multiplied by a constant; it increases also, within a certain limit, with the frequency of the interruptions of the magnetising current, and increases still more when the direction of the current is alternately changed. Its amount is not equal throughout the length of an iron cylinder, reaching its maximum about its middle and decreasing towards its ends. Its cause must be searched for in purely mechanical forces.

RATHER a strange occurence came recently before my notice, and thinking perhaps you might care to insert it in your columns, I send you the facts of the circumstance. A few days since, towards evening, I killed a snake just close behind my house; it measured about a yard and a half in length, was one of the most deadly of the numerous kinds of snakes found in Java, and bears the name of "Oelar belang." On examining it later I found what I thought to be the tail of another small snake protruding from its mouth, but on pulling it out I was greatly surprised to discover that it was really a snake of the same species, and of almost the same length. There was certainly not more than three inches' difference in the length of the two snakes, and at the time I killed the outside snake only about an inch and a half or two inches of the tail of the one he had swallowed protruded from his mouth. The natives here say that the two snakes must have been fighting, the victor afterwards swallowing his opponent. I should be pleased to know whether such an instance has ever before been brought before your notice, or whether it is really an uncommon case.
Soemedang, Java.

M. MONTOIGNY has recently published a pamphlet on the influence of the atmosphere in the apparition of colours seen in the scintillation of stars. He has previously noticed that there is a great predominance of blue in the scintillating colour when rain is approaching, and he is now so convinced of the accuracy of this forecast that it is included among others in the *Bulletin Météorologique* published by the Observatory of Brussels.

From *Nature* 30, 3 July 1884.

(within errors) to those found in the fossil record. On closer examination we found that most of the signal was coming from the 11 largest craters, those with a diameter >10 km. One would expect comets in a shower to be larger than the background comets if, for example, the comets from the inner cloud were more massive than those from the outer regions, as might be expected from an accretion model of comet formation. Of the 11 large craters, one is clearly not associated with the periodicity, and as many as two others may be chance coincidences. Thus, the evidence indicates that about 9 ± 3 of the 13 craters with precise ages originated from the periodic showers. (The uncertainty estimate includes both systematic and Poisson errors.) Stated another way, we conclude that $70 \pm 25\%$ of the craters were formed during showers.

Weissman claims that the total number of comets hitting the Earth from showers is substantially greater than the number hitting in the time between showers—this is not necessarily true in our model, and the impact data indicate that the numbers may be roughly equal. Perhaps Weissman's confusion comes from a poorly worded statement in our original paper[1] that the number of comets that arrive in a single shower "may be as much as one or two orders of magnitude greater than the number that arrive between showers". In this statement we were referring to the effects of the companion star alone, and were not including comets whose infall is triggered by random passing stars. Weissman states near the end of his letter: "there do not appear to be many random events mixed with the periodic signal"; as stated above, we believe that $30 \pm 25\%$ of the impact craters (principally those with diameters <10 km) may be random in arrival times.

Weissman says that the probability of a large comet (with a 10 km nucleus) hitting is small, and that such impacts are likely to occur every 500 Myr, a period "considerably longer than is observed". Only one such large impact has been proven in the last 250 Myr—the one at the end of the Cretaceous. The crater for this impact, if it still exists, should be 100-200 km across. (If the crater is on the sea floor it would not have been found.) All the other extinctions during this period were smaller, with associated craters 10-100 km across. Weissman claims there should be one large impact in 500 Myr; the data show one in 250 Myr. Although we do not necessarily accept Weissman's number (he does not give a derivation) we find no significant discrepancy with the data.

If the orbit of the companion star were once much closer to the Sun, as is necessary in our model, then Weissman objects that the showers would have been more intense in the past. This is true, but is not an objection. In fact, there is clear evidence that bombardment in the remote past was once much greater than in the last 250 Myr. The 'late heavy bombardment' of the Moon (when most of the lunar craters were formed) ended 3.9×10^9 yr ago, conceivably when the companion star was scattered to a larger orbit. Weissman correctly states that an ancient close-in orbit would have severely depleted the Oort comet cloud, implying that there was originally an 'immense cloud mass'. As the mass of this early cloud could still have been much less than the mass of the companion star, we can see no valid objection. If the companion star exists, then we certainly will have to change our models of the dynamics of the early Solar System. But inconsistency with the models is not the same as inconsistency with the data on which those models were based. We know of no data inconsistent with our model of a solar companion star, and we know of no alternative model consistent with the measured period and phase of the impacts.

RICHARD A. MULLER*
PIET HUT†
MARC DAVIS‡
WALTER ALVAREZ§

* *Department of Physics and Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720, USA*
† *Institute for Advanced Study, Princeton, New Jersey 08540, USA*
‡ *Department of Physics and Astronomy, University of California, Berkeley, California 94720, USA*
§ *Department of Geology and Geophysics, University of California, Berkeley, California 94720, USA*

1. Davis, M., Hut, P. & Muller, R. A. *Nature* **308**, 715-717 (1984).
2. Whitmire, D. & Jackson, A. *Nature* **308**, 713-715 (1984).
3. Alvarez, W. & Muller, R. A. *Nature* **308**, 718-720 (1984).
4. Raup, D. & Sepkoski, J. *Proc. natn. Acad. Sci. U.S.A.* **81**, 801-805 (1984).
5. Hills, J. G. *Astr. J.* **86**, 1730-1740 (1981).
6. Hut, P. *Nature* **311**, 636-638 (1984).
7. Grieve, R. *Geol. Soc. Am. spec. Pap.* **190**, 25-37 (1982).

WHITMIRE AND JACKSON REPLY—The contribution of shower comets to the mean cratering rate need only be comparable to the (assumed) steady-state random contribution to produce a strong signal in the Fourier spectrum of dated craters. The number of terrestrial impacts resulting from a shower of 2×10^9 comets, brighter than $H_{10} = 11$ and with perihelia ≤ 1 AU, is given by the product of (2×10^9 comets) and (4 perihelia passages per comet) and (2×10^{-9} impacts per perihelion passage), which is equal to 16 impacts. According to Weissman[1], a long period comet, brighter than $H_{10} = 11$, has a mass $\geq 5 \times 10^{14}$ g and produces a crater ≥20 km in diameter. The mean cratering rate, for craters ≥20 km, over the last 300 Myr is[2] $0.35 (\pm 0.13) \times 10^{-14}$ km^{-2} yr^{-1}, which corresponds to ~50 impacts of comets, brighter than $H_{10} = 11$, per 28 Myr. Thus the shower contribution to the mean cratering rate is less than the total rate and comparable to the mean background rate.

According to Wetherill and Shoemaker[3], a random 10-km asteroid is expected to impact on the Earth about every 50 Myr. The absence of any random extinction events[4], and the evidence that major extinctions occur over intervals of ~1 Myr, both suggest that the primary extinction mechanism may be associated with the enhanced impact rate (≥30 times background) during a shower, rather than with a single catastrophic impact, although such an event should also occur occasionally.

The stability of the companion's present orbit towards stellar perturbations and the galactic tides has recently been investigated by Hut[5] and Hills[6]. These results indicate that the orbital period should random-walk away by 10-20% over a 250-Myr interval, a result not obviously incompatible with the cratering and extinction data. The evolution of the companion from a tighter orbit[7], and the associated cratering rates, remain to be investigated in detail.

Sepkoski and Raup[4] have recently extended their original analysis and have further restricted the uncertainty in the extinction period. They find the period to be 26.2 ± 1 Myr and conclude that this is incompatible with the 33 ± 3-Myr interval between the Sun's galactic plane crossings.

DANIEL P. WHITMIRE
Department of Physics, University of Southwestern, Louisiana, P.O. Box 44210, Lafayette, Louisiana 70504, USA
ALBERT A. JACKSON IV
Computer Science Corporation, 1300 Bay Area, Houston, Texas 77058, USA

1. Weissman, P. R. *Geol. Soc. Am. spec. Pap.* **190**, 15-24 (1982).
2. Grieve, R. A. F. & Dence, M. R. *Icarus* **38**, 230-242 (1979).
3. Wetherill, G. W. & Shoemaker, E. M. *Geol. Soc. Am. spec. Pap.* **190**, 1-13 (1982).
4. Sepkoski, J. J. & Raup, D. M. *Dynamics of Extinction* (ed. Elliott, D.) (Wiley, New York, in the press).
5. Hut, P. *Nature* **311**, 638-641 (1984).
6. Hills, J. *Nature* **311**, 636-638 (1984).
7. Whitmire, D. P. & Jackson, A. A. *Nature* **308**, 713-715 (1984).

The maximum entropy method for data analysis

IN data analysis it is very satisfying when a useful technique, which has arisen in a more or less *ad hoc* manner, finds respectability as a manifestation of some formal theory. Skilling[1] invokes this argument for using the maximum entropy method for data analysis. Experience that the procedure often seems to work quite well is strengthened by reference to recent theoretical work. Skilling's conclusion is that maximum entropy is the only regularization method that should be used for a very wide range of data analysis. I suggest that this claim is unjustified on both pragmatic and theoretical grounds.

Let us confine our attention to one specific application, the enhancement of images as illustrated in Fig. 1 of Skilling's paper. What seems to me the major advantage of the maximum entropy method over many, but not all, other regularization