BAYESIAN INDUCTIVE INFERENCE AND MAXIMUM ENTROPY

Stephen F. Gull
Mullard Radio Astronomy Observatory
Cavendish Laboratory
Madingley Road
Cambridge CB3 0HE, United Kingdom

ABSTRACT

The principles of Bayesian reasoning are reviewed and applied to problems of inference from data sampled from Poisson, Gaussian and Cauchy distributions. Probability distributions (priors and likelihoods) are assigned in appropriate hypothesis spaces using the Maximum Entropy Principle, and then manipulated via Bayes' Theorem. Bayesian hypothesis testing requires careful consideration of the prior ranges of any parameters involved, and this leads to a quantitive statement of Occam's Razor. As an example of this general principle we offer a solution to an important problem in regression analysis; determining the optimal number of parameters to use when fitting graphical data with a set of basis functions.

INTRODUCTION

At the Calgary meeting two years ago Ed Jaynes gave a tutorial introduction (Jaynes 1986) that provides the historical and philosophical background to the principles of Bayesian inference. Like that paper, which the reader is strongly encouraged to study, the aim here is to provide a tutorial guide to Bayesian methods. A short resumé of the basic principles is presented, but the emphasis of this paper is more technical, showing the application of the method to a selection of problems that are solved in detail. We then take a glimpse of the Frontiers of the subject, where (following Jeffreys) a quantitive statement of Occam's Razor is offered. Finally, we turn to the problem of curve-fitting, a state-of-the-art example of Bayesian methods.

THE GROUND RULES

I want to distinguish clearly three stages that together make up my Bayesian view of probability theory and statistics. I believe that all three stages are essential to the process of inductive reasoning.

G. J. Erickson and C. R. Smith (eds.),
Maximum-Entropy and Bayesian Methods in Science and Engineering (Vol. 1), 53–74.
© 1988 by Kluwer Academic Publishers.

1) Bayes' Theorem.
In its simplest form this elementary theorem relates the probabilities of two events or hypotheses A and B. It states that the joint probability distribution function (p.d.f.) of A and B can be expressed in terms of the marginal and conditional distributions:

$$pr(A,B) = pr(A)\ pr(B|A) = pr(B)\ pr(A|B).$$

Bayes' theorem is merely a re-arrangement of this decomposition, which itself follows from the requirement of consistency for the manipulation of probabilities (Cox 1946). Of course, anyone can prove this theorem, but people who believe it and use it are called Bayesians. However, before anyone, even Bayesians, can use it, the joint p.d.f. has to be assigned. Because Bayes' theorem is simply a rule for manipulating probabilities, it cannot by itself help us to assign them in the first place, and for that we have to look elsewhere.

2) Maximum Entropy.
The Maximum Entropy principle (MaxEnt) is a variational principle for the assignment of probabilities under certain types of constraint called Testable Informatation. These constaints are ones that refer to the probability distribution directly: e.g. for a discrete p.d.f. $\{p_i\}$, the ensemble average of a quantity r $\langle r \rangle = \sum_i r_i p_i$ constitutes testable information. MaxEnt states that the probabilities are given by maximising the Entropy

$$S = -\sum_i p_i \log p_i/m_i \text{ under the constraints } \sum_i p_i = 1 \text{ and } \langle r \rangle \text{ given,}$$

where $\{m_i\}$ is a suitable measure over the space of possibilities (hypothesis space). The MaxEnt rule can be justified as the only consistent variational principle for the assignment of probability distributions (Shore & Johnson 1980, Gull & Skilling 1984, Skilling 1988). It can also be derived in a multitude of other ways (Jaynes 1986). In the simplest case there is no additional information other than normalisation: MaxEnt then gives equal probabilities to all events, in accordance with Laplace's "principle of indifference". In fact, I believe that MaxEnt is the only logical method we have for the assignment of probabilities - but it is so powerful that it may be all we need. Of course, MaxEnt is rule for assigning probabilities once the hypothesis space has been defined; to choose the hypothesis space we have again to look elsewhere.

3) Choosing the hypothesis space
The real art is to choose an appropriate "space of possibilities", and to date we have no systematic way of

generating it. Transformation group arguments can often help us (Jaynes 1968) in problems involving physical quantities; the appropriate measure space is often uniform (location parameters) or uniform in the logarithm (scale parameters). MaxEnt will then assign a uniform "prior" probability distribution over this space. However, in many problems one has no guarantee that our choice is right in any final sense, and this feeling of ambiguity has led to much soul-searching. I feel (along with Jaynes, 1986) that our aims should be different. We should not seek a "final truth" in our hypothesis space, but use our common sense to capture enough structure of the real problem being solved so that we can make useful predictions. If the predictions are useful, then that is an indication the the hypothesis space is good enough for now, without prejudice to the possibility of revising it later. If the predictions are not good, this is not a disaster, for we then have learnt that the hypotheses have to be reformulated and the ways in which our predictions are wrong may help us to do this. In any case we simply have nothing to lose by choosing an interim hypothesis space and proceeding with the calculation.

Of course, not everyone sees it that way, but once you are used to the process there is nothing more painful than the sight of grown men being psychologically unable to make a simple Bayesian calculation just because they might be wrong. They could agonise forever about the hypothesis space or prior, but unless they make that calculation they will never know!

WHY I AM A BAYESIAN

I am ashamed to have to admit that, when I was a physics student, I thought that the lectures on probability theory and statistics were an unnecessary distraction from "real physics". Whatever my motives at the time, the result was that I had an open mind when confronted some years later by Bayesian statistics. Whilst observing the radio sky (see example 3) I met Geoff Daniell in a pub to discuss the analysis of the data. In the course of that evening he proved Bayes' theorem to my satisfaction by drawing on a beer-mat a circle that had two lines across it, and gave me a few examples. The following is the first example I did for myself when I returned to the telescope.

Poisson distribution - the radioactive solid

Suppose there is a sample of a radioactive solid that produces, on average, α decays per second. You have observed N decays in T seconds: what is α? The Likelihood or sampling distribution for the Poisson process is well-known as a limiting form of Binomial distribution:

$$pr(N|\alpha,T) = (\alpha T)^N \exp{-\alpha T} / N!.$$

This distribution can also be derived by MaxEnt (see for example Skilling & Gull 1984) with a constraint on $\langle N \rangle = \alpha T$, using as a measure the form $Q^N/N!$ for the number ways of distributing N objects in a large number Q of cells.

A Bayesian analysis should start with the joint p.d.f.:

$$pr(\alpha,N) = pr(\alpha)\, pr(N|\alpha) \quad (T \text{ is always known}).$$

To complete the assignment of this joint distribution we have to specify the hypothesis space sufficiently to determine the prior distribution $pr(\alpha)$. We can do this by noting that α is a scale parameter: if we were totally ignorant of the amount of radioactivity, we would be just as ignorant if there was twice (or half) as much. This leads to a unform prior in $\log\alpha$, but to be quite complete we should specify some limits $[\alpha_{min}, \alpha_{max}]$ so that the prior can be normalised. Let us define a "sensible" range of: α_{min} = Hubble's constant / Avagadro's number (1 decay per gram molecule in the age of the Universe) and $\alpha_{max} = 10^{-4}$ of a lethal radiation level (or you can find another experimenter).

We now use Bayes' theorem by writing the joint p.d.f. in its alternative form:

$$pr(\alpha,N) = pr(N)\, pr(\alpha|N).$$

Renormalising, we get the posterior distrubuton for α:

$$pr(\log\alpha|N) = (\alpha T)^N \exp{-\alpha T} / (N-1)!.$$

The re-normalisation is possible over an infinite range of α if $T>0$ and $N>0$. This is entirely reasonable: if $N=0$ then we don't yet know it is radioactive, and if $T=0$ we haven't started looking.

This is a very simple, but highly instructive example. Figure 1 shows the Likelihood and the posterior distribution for the case $T = N = 5$, $\alpha = 1$. It is the same function of the two variables (α,N), but plotted on different axes, so that there is a remarkable switch of meaning. The Likelihood gives the probability of different numbers of decays for a constant value of α, whereas the posterior gives the relative probability of different parameter values for the single value of N that was actually observed. These are completely different concepts and it is only through Bayes' theorem that there is any relationship between them.

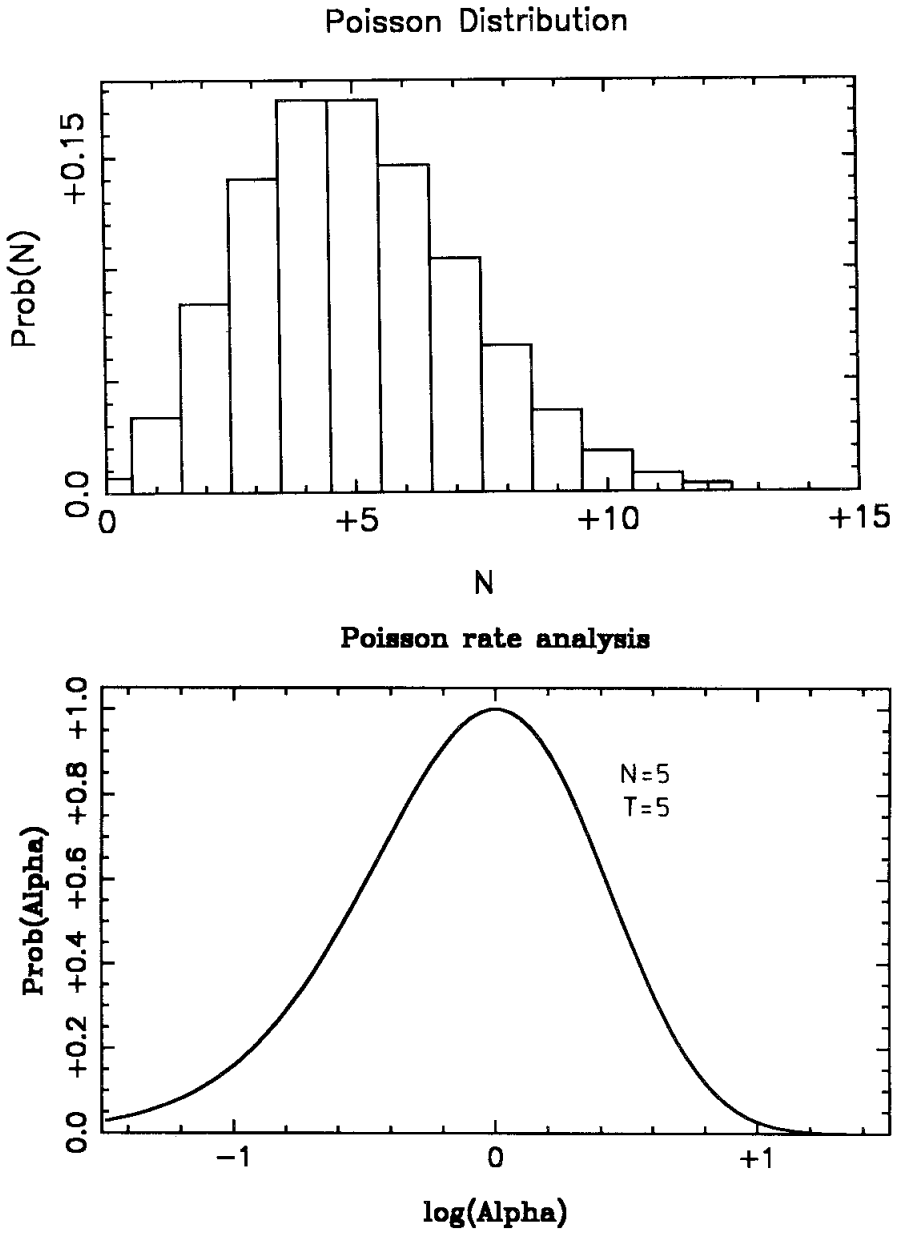

Figure 1. Likelihood and posterior probability distribution for a Poisson process with $\alpha=1$, $N=T=5$.

I think we should not lose sight of this; the Bayesian rationale given above is, of course, entirely consistent with the "Maximum Likelihood" method - in fact it provides a justification for that method. But when we use the ML method a natural misunderstanding arises by the word-play inherent in the very name "Maximum Likelihood" - it makes you think that the answer you get is the "most likely" one. Not so: you get the parameter for which the observed datum had the greatest Likelihood. It is only by confronting Bayes' theorem that one can see that this is indeed (under many circumstances) the "most likely".

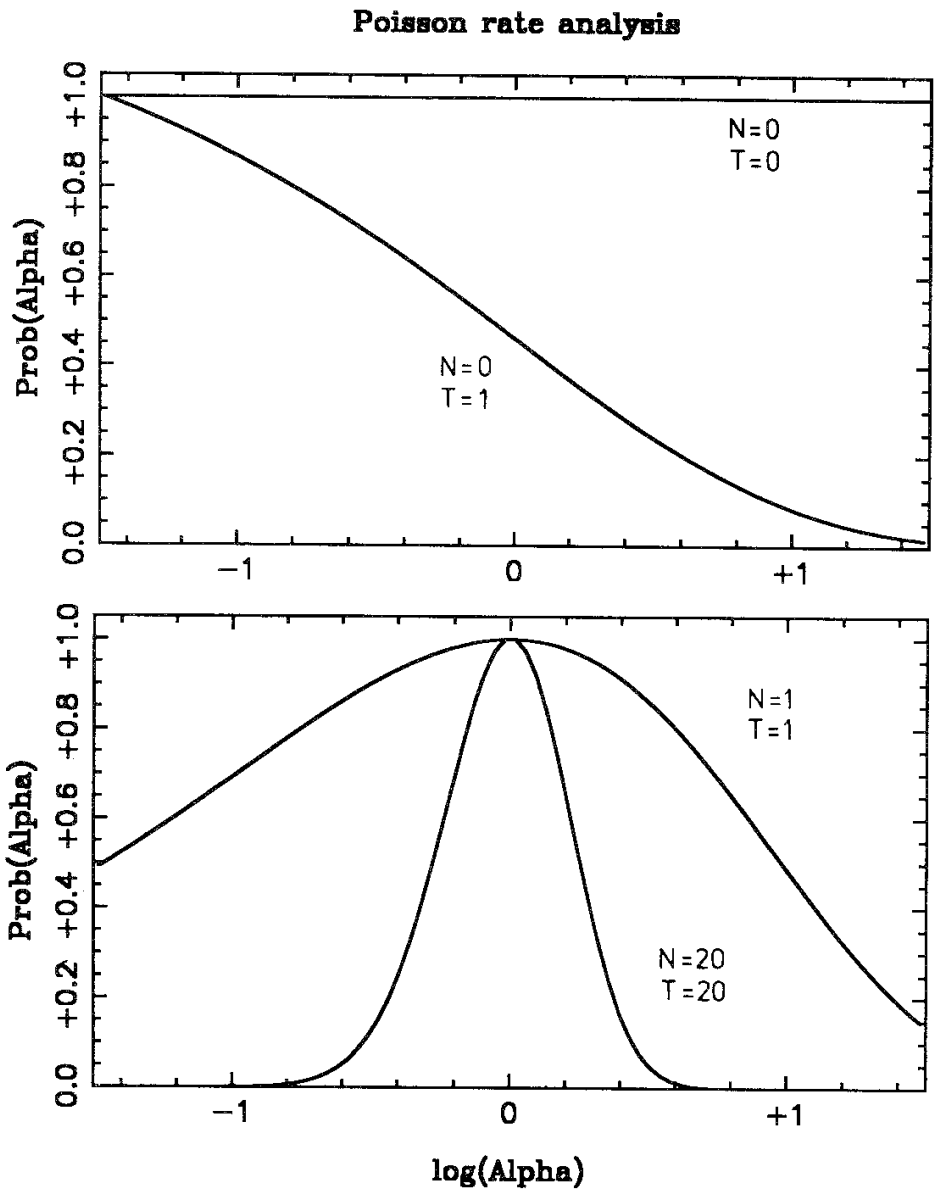

Figure 2. Evolution of the posterior p.d.f. of the rate parameter α of a Poisson process as more data become available.

Additional features

1) Figure 2 shows some results generated by a Poisson process on my PC. At T=0 the pr(logα) is uniform, and as T>0, but N remains 0, high values of α becomes less likely, and the p.d.f. can be allowed to extend over an infinite upper interval. When the first decay occurs the lower limit can also be extended to zero.

2) The moments of the posterior distribution are easily calculated in terms of Gamma functions:

$$\langle\alpha\rangle = N/T,$$
$$\langle\alpha^2\rangle = N(N+1)/T.$$

As T and N increase, this leads to a width $\delta\alpha = N^{\frac{1}{2}}/T$ that shrinks like $T^{-\frac{1}{2}}$ as expected.

3) Another useful technique I should mention is to expand the logarithm of the p.d.f. near its maximum. A Taylor series about this point will yield an estimator and a width:

$$\log(pr(\log\alpha)) = \text{const.} - \alpha T + N\log\alpha T.$$

If we differentiate with respect to $\log\alpha$ we get an unbiased estimate:

$$\partial(\log pr)/\partial\log\alpha = -\alpha T + N \qquad \text{(zero at maximum)}$$
$$\partial^2(\log pr)/\partial\log\alpha^2 = -\alpha T.$$

This yields a maximum probability at $\alpha = N/T$ and an approximate width of $\delta\log\alpha = N^{-\frac{1}{2}}$.

These features were sufficient to convince me of the usefulness of Bayesian methods. I was, and still am, impressed by the way the beautiful result $\langle\alpha\rangle = N/T$ depends on the careful consideration of the prior for α. The next example is even more staightforward, but is still the cause of heated debate with non-Bayesians in my department.

Cauchy distribution - the lighthouse problem
(Taken from a Cambridge Part 1A examples sheet). A lighthouse is somewhere off a piece of straight coastline at position x_0 along the coast and a distance y out to sea. It emits a series of short, highly collimated flashes at random intervals and hence at random azimuths. These pulses are intercepted on the coast by photo-detectors that record only the fact that a flash has occurred, but not the azimuth from which it came. N Flashes have so far been recorded at positions $\{x_i, i=1,N\}$. Where is the lighthouse?

For any one sample the likelihood can be written in terms of the azimuthal angle θ , where $y \tan\theta = x - x_0$:

$$pr(x|x_0,y)\,dx = pr(\theta)\,d\theta = d\theta / \pi.$$

This gives the Cauchy distribution:

$$pr(x|x_0,y) = y / (\pi (y^2 + (x - x_0)^2).$$

Different pulses are independent so that the total likelihood can be written as a product:

$$pr(\{x_i\}|x_0,y) = (y/\pi)^N \prod_i (y^2 + (x_i - x_0)^2)^{-1}.$$

Use the joint p.d.f. again, taking a uniform prior probability for the position (x_0,y) as they are location

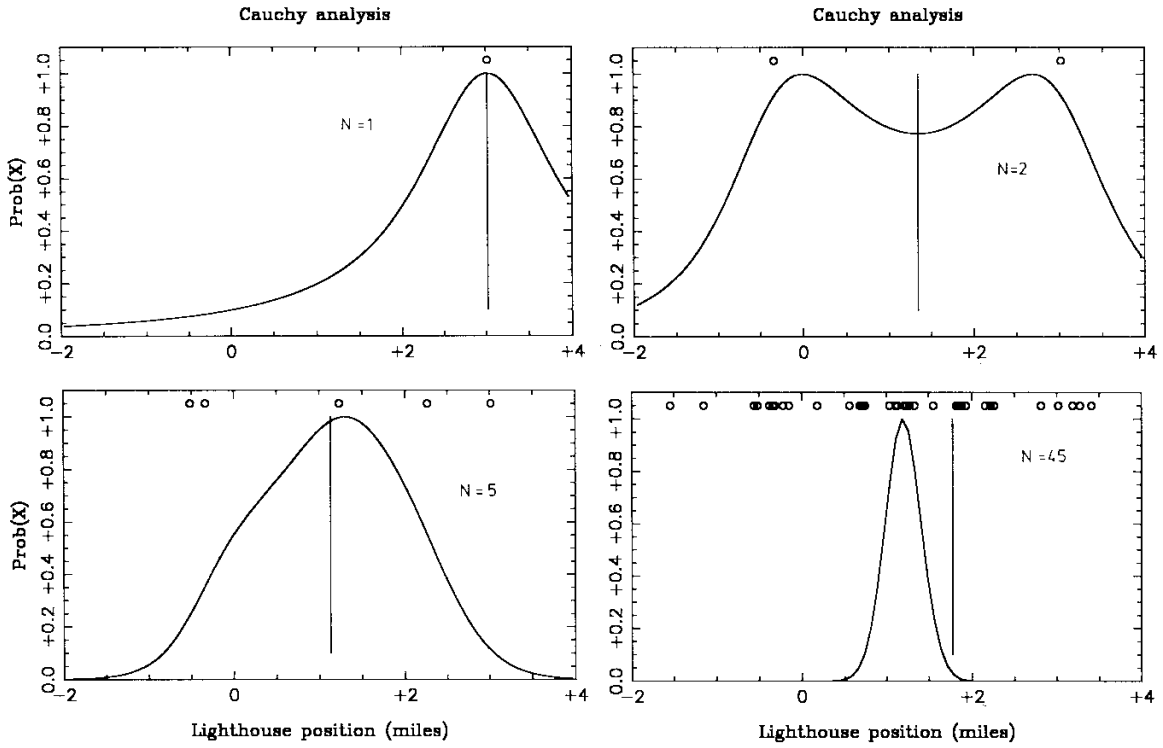

Figure 3. One-dimensional posterior p.d.f. of the lighthouse position for various data samples. Note that the distribution can be multi-modal. The vertical bar shows the postion of the sample mean. The correct position was at $x_0 = 1$.

parameters:

$$pr(\{x\},x_0,y) = pr(\{x\}|x_0,y)\ pr(x_0,y)$$
$$= pr(x_0,y|\{x\})\ pr(\{x\})$$

and obtain: $pr(x_0,y|\{x\}) \propto pr(\{x\}|x0,y)$.

This formula is illustrated by computer example for two cases:

Case 1: The lighthouse is known to be 1 mile off the coast, so that we have a one-dimensional probability distribution.

Case 2: No such restriction, so that there is two-dimensional plot.

The figures are very revealing.

1) The Cauchy distribution has very wide wings, i.e. there are many more "bad" data points than, for example in a Gaussian distribution. For the 1-dimensional case (Figure 3) this can lead to the posterior distribution being bi-modal if the first few points are sufficiently discordant.

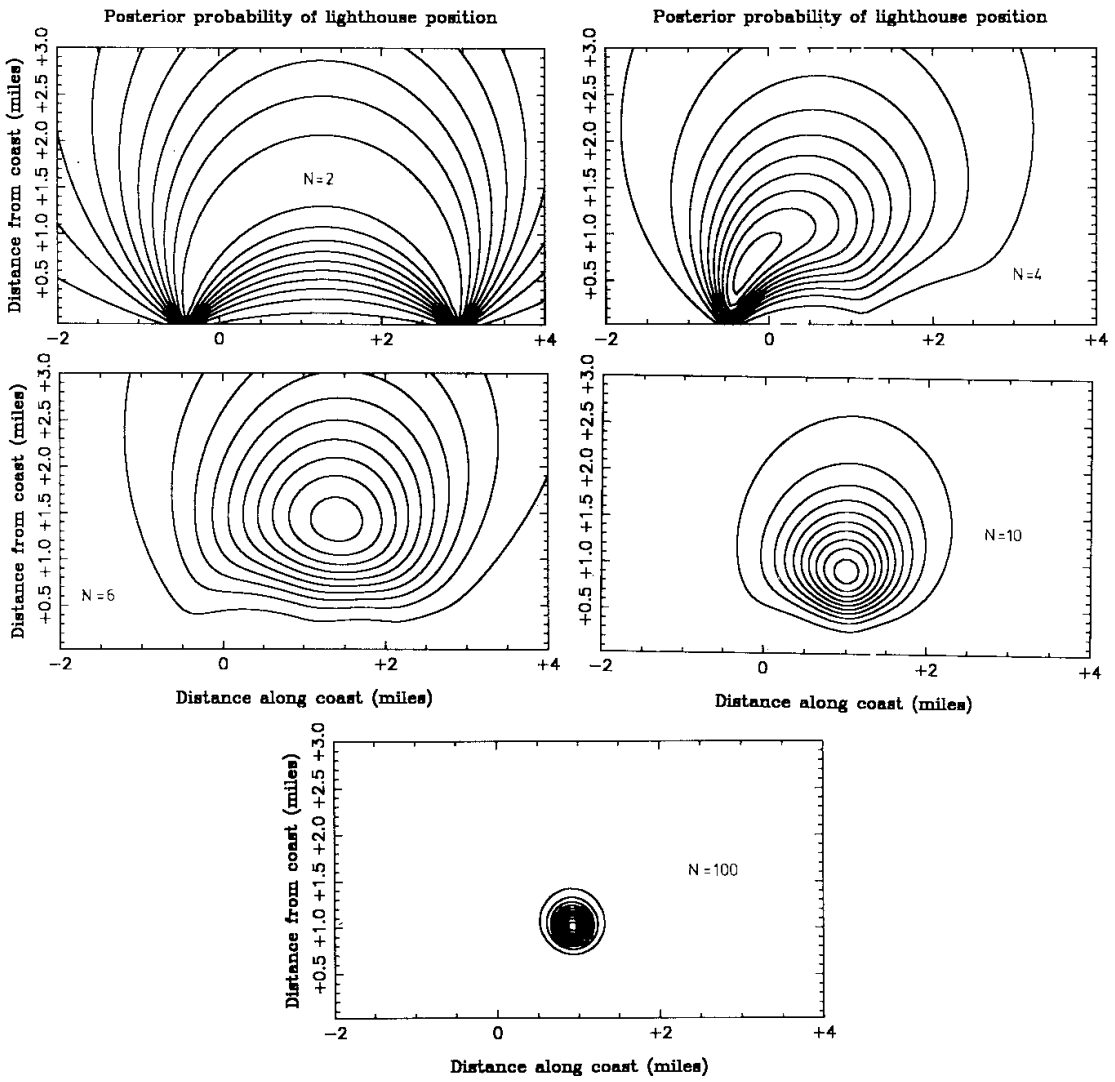

Figure 4. Two-dimensional plot of lighthouse position as function of x_0 and y. The correct position was at x_0=y=1.

2) Nevertheless the bulk of "good" data eventually overwhelm the bad and the allowed range of (x_0,y) shrinks (Figure 4).

3) The sample mean $\Sigma x/N$ is not a good statistic for this problem, and does not approach the value of x_0 any more closely as N increases. For an excellent discussion of this see Jaynes (1976).

Gaussian distribution

Meanwhile, back at the telescope, I was observing a patch of sky repeatedly in an attempt to detect a putative "hole" in the temperature of the Cosmic Microwave Background Radiation. The depth of this hole is about 0.5mK, and individual 1 minute measurements had a variance of about 10mK (and cost about $3 each). (Patience has now been rewarded with 3 results of $\approx 10\sigma$ after 10 years (Birkinshaw et al. 1985)). Suppose that we model the data collection as

a Gaussian process with mean μ and standard deviation σ. You have N samples $\{x_i\}$: what are μ and σ? This simple problem is worth solving here because it illustrates quite a few of the mathematical subtleties that will appear later in the section on Bayesian curve-fitting.

The single-sample likelihood for the Gaussian distribution can be derived by MaxEnt, using constraints on the first two moments of pr(x): $\langle x \rangle = \mu$ and $\langle (x - \mu)^2 \rangle = \sigma^2$, and a uniform measure m(x). The MaxEnt likelihood for multiple samples $\{x_i\}$ is then independent:

$$\mathrm{pr}(\{x_i\}|\mu\sigma) = (2\pi\sigma^2)^{-N/2} \exp -\sum_i (x_i-\mu)^2/2\sigma^2.$$

To manipulate this expression it is best to re-write the exponential as: $-(1/2\sigma^2)\ [N\mu^2 - 2\mu\Sigma x_i + \Sigma x_i^2]$.

Now complete the square, defining the sample mean and variance $\bar{x} = \Sigma x/N$ and $V = \Sigma x^2 - N\bar{x}^2$:

$$-(1/2\sigma^2)\ [N(\mu - \bar{x})^2 + V].$$

We are now ready for Bayes' theorem using the joint p.d.f. again:

$$\mathrm{pr}(\{x_i\},\mu\sigma) = \mathrm{pr}(\{x_i\}|\mu\sigma)\ \mathrm{pr}(\mu\sigma) = \mathrm{pr}(\{x_i\})\ \mathrm{pr}(\mu\sigma|\{x_i\}).$$

The prior for μ and σ has been much discussed: σ is a scale parameter and should have a uniform prior in $\log\sigma$; μ is a location parameter and should have a uniform prior. At the time that this talk was presented I followed conventional wisdom that this implied a uniform prior $\mathrm{pr}(\mu,\log\sigma)$. But if we start by allocating the prior in $\log\sigma$ over some range $[\sigma_{min}, \sigma_{max}]$ then we clearly have to assign the range in μ <u>given the knowledge of σ</u>. Perhaps the range in μ should be proportional to σ, which would lead rather surprisingly to $\mathrm{pr}(\mu,\log\sigma) \propto 1/\sigma^2$. Yoel Tikochinsky has another argument based on a transformation group that yields the same result. Although I will now assume $\mathrm{pr}(\mu,\log\sigma) = \text{constant}$, I think that Yoel has a good point and that there still seems to be some life in this old argument!

Write the posterior distribution:

$$\mathrm{pr}(\mu,\log\sigma|\{x_i\}) \propto \sigma^{-N} \exp -[N(\mu-\bar{x})^2+V]/2\sigma^2.$$

The marginal distributions are interesting: the distribution for μ is a "Student-t" with N-1 degrees of freedom.

$$\mathrm{pr}(\mu|\{x_i\}) = \int d\log\sigma\ \mathrm{pr}(\mu,\log\sigma|x\{_i\})$$

$$\propto [N(\mu-\bar{x})^2 + V]^{-N/2}.$$

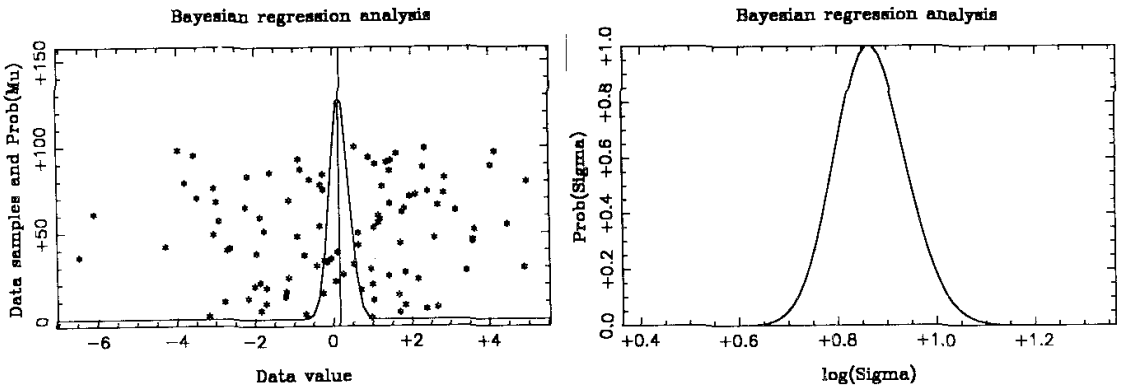

Figure 5. Marginal posterior distributions of mean and standard deviation for a Gaussian distribution. There were 100 samples with $\mu=0$ and $\sigma=2$.

Marginalising the other way we find:

$$pr(\log\sigma|\{x_i\}) \propto \sigma^{-N+1} \exp -V/2\sigma^2,$$

and defining $X=V/\sigma^2$:

$$pr(X|\{x_i\}) \propto X^{(N-3)/2} \exp -X/2.$$

In more conventional language this says that V/σ^2 is distributed like χ^2 with N-1 degrees of freedom. A good estimator is therefore $\sigma^2 \approx V/(N-1)$. This can also be seen by differentiating $\log(pr(\log\sigma))$ with respect to $\log\sigma$. Some results are plotted as Figure 5.

Note that in these examples I have treated scale parameters systematically by taking logarithms. This is good practice, because the prior is uniform in the logarithm, corresponding to the suggestion that we use log graph paper to plot the distribution. If we insist on plotting the parameter itself, then then prior is $1/\sigma$, for example. This looks a bit mysterious, even to a practising Bayesian like myself. But we don't need to confuse - take logarithms.

BAYESIAN HYPOTHESIS TESTING

The story of Mr. A and Mr. B

Why do we prefer theories with only a few parameters? The principle proposed by William of Occam - that there is more intrinsic merit in simpler theories - is universally accepted by scientists. But why? The following argument, due to Harold Jeffreys (1939, Chapter 5) explains that a simple theory can become more probable than a complicated one when confronted with data.

Suppose we have two competing theories to explain the data

D. The theory proposed by Mr. B has a parameter λ, which has to be known before the data can be predicted, but Mr. A's theory has none, and predicts the data directly. An example that occurred in physics some time ago was the Brans-Dicke scalar field theory that included a ratio ω of the strength of scalar and tensor fields (Brans & Dicke 1961). If there was no scalar component ($\omega = 0$) then the theory reduced to Einstein's General Relativity. The data in question were the classical tests of G.R., along with some new measurements of solar oblateness.

Write the likelihoods $pr(D|A)$ and $pr(D|B,\lambda)$. There is presumably a value of λ that fits the data best - call it λ_0. Let us suppose for the sake of illustration that for our particular case the likelihood is a Gaussian with width $\Delta\lambda$. Also, we must suppose that Mr. B's extra parameter allows him to fit the data better than Mr. A's inflexible one, which may or may not be a special case of Mr. B's. (The Brans-Dicke camp might say G.R. was just a special case of their theory, but the other side might retort that no such parameter existed!)

The question to be asked is then: how much bigger should $pr(D|B,\lambda_0)$ be than $pr(D|A)$ for Mr. B's theory to be preferred? We need a <u>quantitative</u> statement of Occam's Razor. Let us try to calculate the relative probabilities of Mr. A and Mr. B's theories in the light of the data.

$$\frac{pr(A|D)}{pr(B|D)} = \frac{pr(A|D)}{\int d\lambda\ pr(B,\lambda|D)} = \frac{pr(A)\ pr(D|A)}{\int d\lambda\ pr(B,\lambda)\ pr(D|B,\lambda)}$$

$$\frac{pr(A)}{pr(B)} \times \frac{pr(D|A)}{\int d\lambda\ pr(\lambda|B)\ pr(D|B,\lambda)}$$

The difficult term is the prior for the parameter λ in Mr. B's theory. Let us take it as uniform in some range $[\lambda_{min}, \lambda_{max}]$ specified by Mr. B. Then, using the assumption of a Gaussian likelihood we find:

$$\frac{pr(A|D)}{pr(B|D)} = \frac{pr(A)}{pr(B)} \times \frac{pr(D|A)}{pr(D|B,\lambda_0)} \times \frac{(\lambda_{max} - \lambda_{min})}{(2\pi)^{\frac{1}{2}}\ \Delta\lambda}.$$

The first term in this product is a prior prejudice in favour of Mr. A or Mr. B that has nothing to do with the theory being tested. It might be taken as unity, or might even reflect their past performances. The second term is the best-case likelihood ratio, that is expected to favour Mr.

B. The third term is the "Occam factor" we are looking for and is due to the posterior collapse of Mr. B's hypothesis space. If A and B were equally probable to start with, then Mr. B has to spread his share of probability over a bigger space from λ_{min} to λ_{max}. When the data are given, many of these possible parameter values perish, and only the range $\Delta\lambda$ survive.

This analysis is the same as that given by Jeffreys, he then says that there are difficulties, which indeed there are, because λ_{min} and λ_{max} are left in an unsatisfactorily ambiguous state: what stops us taking an infinite range? That gives an infinite penalty for the parameter, which is just as bad as having no penalty at all. We have to be fair to both Mr. A and Mr. B. However, when stated in the abstract as here, I think that this ambiguity is inevitable - there can be no panacea to solve all such problems. On the other hand we can certainly make progress for many <u>specific</u> problems, when our prior information, whilst still vague, is not actually zero.

A further note of interest is that the decomposition of the posterior probability is precisely the same (if you take the logarithm) as that given by Peter Cheesman and others in their "minimum message-length" approach.

<u>BAYESIAN CURVE-FITTING</u>

Suppose that you are given a graph consisting of N pairs of {x,y} values, and that the values of the ordinate $\{y_i\}$ are subject to a constant, but unknown, amount of noise σ. The task is to fit a set of M parameters $\{a_j\}$ so that the $\{y_i\}$ can be adequately represented in terms of a set of <u>basis</u> functions $\{f_j(x)\}$:

$$y(x_i) \approx \hat{y}(x_i) = \sum_{j=1}^{M} a_j f_j(x_i),$$

or: $\hat{\mathbf{y}} = \mathbf{f}.\mathbf{a}$, where f is an (N x M) matrix. The functions {f} might, for example, be a set of polynomials, or a Fourier series.

I must emphasise that this model problem is one where we suppose that measurement noise σ is added to an exact underlying relation $\hat{\mathbf{y}} = \mathbf{f}.\mathbf{a}$ and that the $\{a_j\}$ are unknown, but <u>with no intrinsic variation from sample to sample</u> of $\{y_i\}$. Another scenario for this sort of problem is the case where there is very little measurement noise, but the data $\{y_i\}$ relate to individual objects that have a spread of $\{a_j\}$ values. An example of this latter type of problem is the colour-luminosity relation for main-sequence stars (the Hertzprung-Russell diagram). Although this other case is

interesting, it is not the problem addressed here.

The ultimate goal of our analysis is to answer the question of how many parameters M we should use. However, we start with the relatively straightfoward task of determining the parameters $\{a_j\}$ when M and σ are known in advance. Write the joint p.d.f. as:

$$pr(\{y\},\{a\}|\sigma,M) = pr(\{a\}|\sigma,M)\ pr(\{y\}|\{a\},\sigma,M).$$

(In all of what follows the values of $\{x_i\}$ and N are known as well, but will be omitted to avoid cluttering the conditioning statements.) For the moment take the prior $pr(\{a\})$ as uniform over some large hypervolume $\delta^M a$. The likelihood (which can be derived by MaxEnt) can be taken as an independent Gaussian:

$$pr(\{y\}|\{a\},\sigma,M) = (2\pi\sigma^2)^{-N/2} \exp -\sum_i (y_i - \sum_j f_{ij}a_j)^2/2\sigma^2.$$

We can make life easier by a little rearrangement of the exponential; write it as $-V/2\sigma^2$, where:

$$V = \Sigma(\mathbf{y} - \mathbf{fa})^2 = V_0 - 2\ \mathbf{a}^t.\mathbf{B} + \mathbf{a}^t.\mathbf{A}.\mathbf{a},$$

$$V_0 = \sum_i y_i^2,\ B_j = \sum_i y_i f_j(x_i),\ A_{jl} = \sum_i f_j(x_i) f_l(x_i) = \mathbf{f}^t\mathbf{f}.$$

The (MxM) **A** matrix tells us about the structure of the space spanned by the {f}; it is strictly positive definite if the {f} are linearly independent. If the basis functions are dependent or M>N then we don't really deserve to solve the problem. A further definition we will need is the Maximum Likelihood estimator $\hat{\mathbf{a}}$, which is the solution of the equation: $\mathbf{A}.\hat{\mathbf{a}} = \mathbf{B}$. We then have:

$$V = V(M) + (\mathbf{a} - \hat{\mathbf{a}})^t.\mathbf{A}.(\mathbf{a} - \hat{\mathbf{a}}),$$

$$V(M) = V_0 - \hat{\mathbf{a}}^t.\mathbf{B}\ .$$

The minimum of V, namely V(M), occurs at $\mathbf{a} = \hat{\mathbf{a}}$.

We now use Bayes' Theorem to obtain the posterior distribution:

$$pr(\{a\}|\{y\},\sigma,M) \propto \exp -V(M)/2\sigma^2\ \exp -(\mathbf{a}-\hat{\mathbf{a}})^t\mathbf{A}(\mathbf{a}-\hat{\mathbf{a}})/2\sigma^2.$$

This is a multivariate Gaussian distribution for the variables {a}, with best estimator $\langle \mathbf{a} \rangle = \hat{\mathbf{a}}$ and covariance:

$$\langle \delta a_j \delta a_l \rangle = \int d^M a\ \delta a_j \delta a_l\ pr(\{a\}|\{y\}) = \sigma^2\ [\mathbf{A}^{-1}]_{jl},$$

where $\delta\mathbf{a} = \mathbf{a} - \hat{\mathbf{a}}$.

It is worth dwelling for a while on these formulae, particularly that for the covariance. This result is "obvious" if you "diagonalise" the matrix **A** (in imagination only!), because the individual dimensions in the integral then separate. Another thing we can do with the formula is to use it for prediction of $\hat{y}(x)$:

$$\langle\hat{y}\rangle = \mathbf{f}(x) \cdot \hat{\mathbf{a}},$$

$$\langle\delta\hat{y}^2\rangle = \sigma^2 \sum_{jl} [\mathbf{A}^{-1}]_{jl} f_j(x) f_l(x).$$

A final note is that the practical solution of the equation for $\hat{\mathbf{a}}$ is best done by least-squares solution of $|\mathbf{B}-\mathbf{f}.\mathbf{a}|^2$, not by solving the normal equations directly. The numerical conditioning of the least-squares solution is determined by the singular values of **f**, which are the square roots of the eigenvalues of **A** itself.

The determination of σ

We now turn to the next easiest problem: M is given but the noise σ is unknown. To do this we expand the hypothesis space to include σ as a parameter, taking a uniform prior for $\log\sigma$ because σ is a scale parameter. To be definite we take this over a range $[\sigma_{min},\sigma_{max}]$, but the range will not matter if N>M. This procedure is equivalent to "forgetting" σ - the posterior distribution for σ will then single out the most likely value of σ consistent with the data. We now should be careful about all the factors of σ that appear in the normalisation of the joint p.d.f, and find:

$$\mathrm{pr}(\log\sigma,\{a\}|\{y\},M) \propto \sigma^{-N} \exp -V(M)/2\sigma^2 \exp -\delta\mathbf{a}^t\mathbf{A}\delta\mathbf{a}/2\sigma^2.$$

We now integrate this over {a} to find the marginal distribution:

$$\mathrm{pr}(\log\sigma|\{y\}) = \int d^M a\ \mathrm{pr}(\log\sigma,\mathbf{a}|\{y\})$$

$$\propto \sigma^{M-N} \exp -V(M)/2\sigma^2.$$

This means that the posterior distribution of $V(M)/\sigma^2$ is like χ^2 with N-M degrees of freedom. If we have to choose a single value of σ, then $\sigma^2 = V(M)/(N-M)$ is a pretty good guess. As might be expected, we lose a degree of freedom for each parameter estimated.

The real problem: which M to use?

Figure 6 shows some data kindly provided to me by colleagues as a blind test. It is (I was assured) a polynomial, with added noise, though not much. Also plotted is the fit for

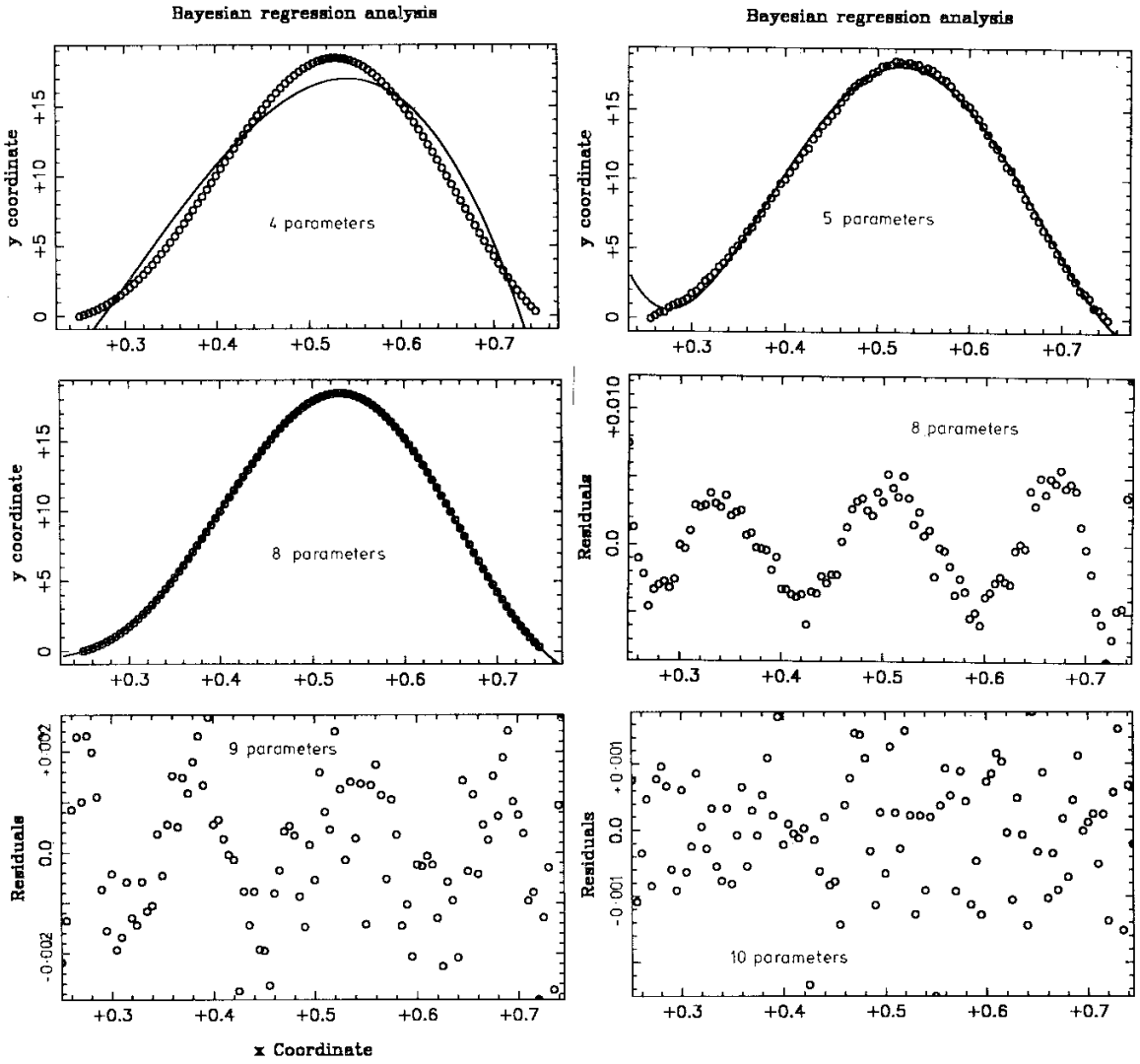

Figure 6. Best-fitting polynomial curves of order 4,5 and 8 compared with a sample of 100 {x,y} values. The residuals are shown for polynomial orders 8,9 and 10.

M=4,5,8 and the residuals for M=8,9,10. The graph of the minimum Variance V(M) against polynomial order M is shown in Figure 7. We see now the real problem: the V(M) curve decreases monotonically, quickly at first, but then more slowly. But how much decrease of V(M) must we have before it is worth adding a new parameter? We need to be fair: if we accept any decrease, then we approach the dreaded "Sure Thing" Theory (Copyright (c) E.T. Jaynes), if we are over-cautious we will miss true structure. In Bayesian terms, this problem is related to the hypervolume $\delta^M a$ associated with any M. We must complete the assignment of priors:

$$pr(\{a\},M) = pr(M)\ pr(\{a\}|M).$$

The final prior pr(M) scarcely matters and we take it as constant in $1<M<M_{max}$.

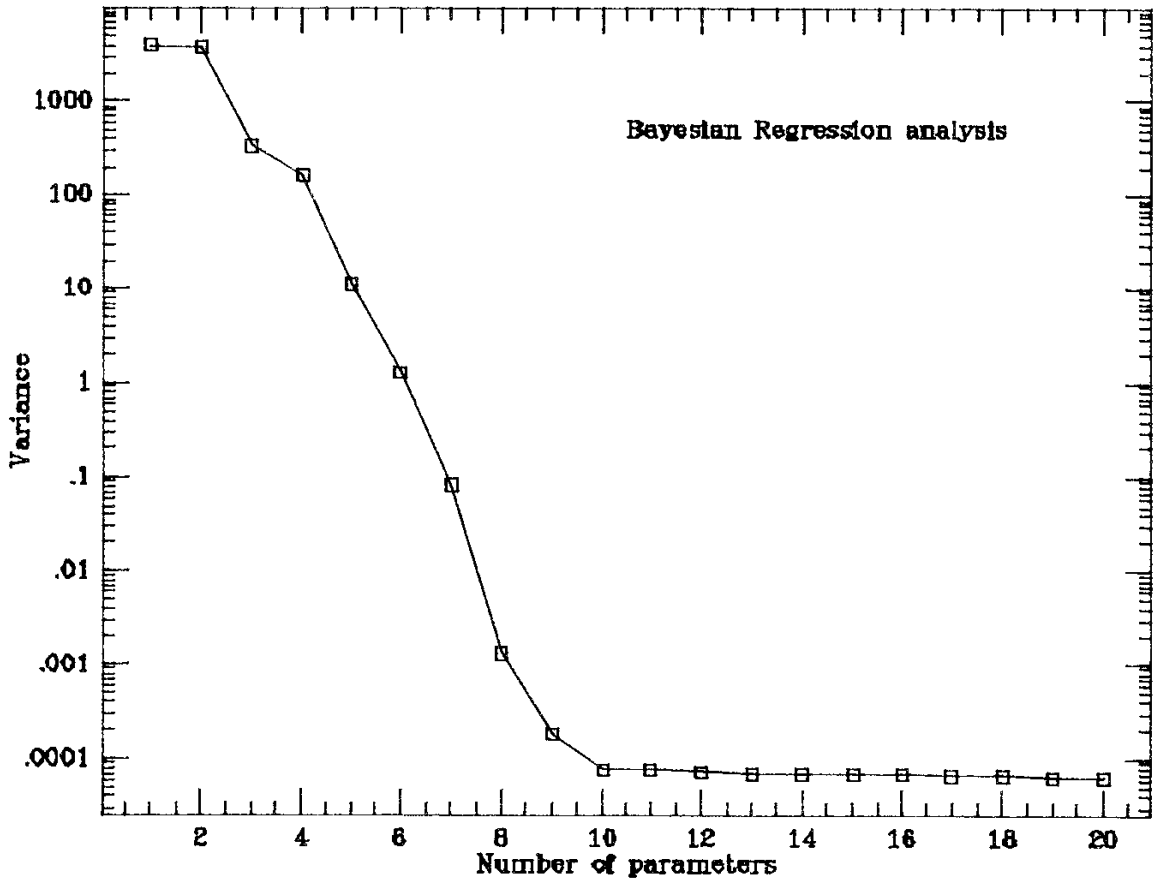

Figure 7. Minimum variance V(M) as a function of polynomial order for the example of Figure 6.

The important factor is, of course, $pr(\{a\}|M)$; we need a prior that ties together the different values of M. A possible way in which we can do this is by referring everything to the N-dimensional space of the ordinates $\{\hat{y}(x_i)\}$. Suppose that the points $\{x,\hat{y}\}$ are drawn on a piece of graph-paper, and that the $\hat{y}$-axis extends from -R to +R. We know that the ordinates will be somewhere on this graph-paper (and so are the samples $\{y_i\}$ to within the noise σ), but we want to encode this information gently, in a way that does not prejudice the shape of the curve. If we knew R, we could accomplish this rather neatly by an ensemble average constraint of the variance:

$$\langle \Sigma \hat{y}^2 \rangle = N R^2 .$$

Note that this is a statement only about the average variance, we are not constraining the variance itself. But

$$\langle \Sigma \hat{y}^2 \rangle = \int d^M a \sum_i \left(\sum_j f_j(x_i) a_j \right)^2 pr(\{a\}|R)$$

is testable information; it relates directly to $pr(\{a\})$. We can therefore derive $pr(\{a\})$ using MaxEnt, maximising

$$S = -\int da\ pr(a)\ \log(p(a)/m(a))$$

over some large measure space $m(a)$ (taken as constant because the $\{a\}$ are location parameters in a M-dimensional vector space). We use the Partition function:

$$Z(\beta) = \int da\ \exp-(\beta/2)a^t\mathbf{A}a = (2\pi/\beta)^{M/2}\ (\det \mathbf{A})^{-\frac{1}{2}},$$

where the constraint is satified by: $NR^2 = M/\beta$. This leads to the prior:

$$pr(\{a\}|M,R) = (\beta/2\pi)^{M/2}\ (\det \mathbf{A})^{\frac{1}{2}}\ \exp -(\beta/2)a^t\mathbf{A}a.$$

This prior neatly incorporates all the properties of orthogonality and normalisation of the basis functions, and relates everthing to the same N-dimensional hypervolume NR^2. In that sense we are being fair. But what should be be the size of this hypervolume? We face the same problem as before; if the hypervolume is to big, we pay too great a price for a new parameter and will miss real structure; if the hypervolume is too small we take too many parameters and have the additional disadvantage that the prior biases the answer too much. The hypervolume has to be "just right", which means $NR^2 \approx \Sigma y^2$ (of the data set). We can show this by expanding our hypothesis space yet further to include different values of R (or β). R and β are scale parameters, so we take uniform prior in logR or logβ. By doing this we essentially "forget" the size of the graph-paper (which we didn't know anyway!), yet retain the "fairness" property between the different values of M. The posterior distribution of pr(logR) will then automatically select the best hypervolume for our purposes, just as previously happened for the case of pr(logσ). We could, of course, simply integrate R out of the problem here and now, and obtain a nice-looking prior:

$$pr(a|M) \propto (a^t\mathbf{A}a)^{-M/2}.$$

The proportionality warns us that this is an improper prior, still depending on the limits of logR, with weak (logarithmic) singularities at both large and small values of a. However, this integration would be counter-productive as far as practical manipulation is concerned; we will keep the Gaussian distributions around as long as possible, because we can always integrate them exactly. The difficult functional forms are those involving R and σ, and we will delay their determination until last. For the moment we note that the consequence of the prior (at fixed β) is to change our estimate of the parameters:

$$\langle a \rangle = k\ \hat{a},$$

with $k = 1/(1 + \sigma^2\beta)$, the fractional weight of the data versus the prior. There is thus a (small) bias of the parameters towards zero (very small in the example given).

Our final formula for the posterior distribution is:

$$pr(M,\log\sigma,\log\beta|\{y_i\}) \propto$$

$$\beta^{M/2}\ \sigma^{-N}\ (\beta + 1/\sigma^2)^{-M/2}\ \exp[-V(M)/2\sigma^2 - (\beta k/2)B^t.\hat{a}].$$

This formula is the basis of the computer program that produced the figures. For each M, the maximum posterior probability was found and a numerical "steepest descents" integration performed to get the marginal distribution $pr(M|\{y\})$. However, with the caveat explained in the next section, we can proceed further analytically for the limiting case $V(M) << V_0$ (i.e. good data!). For this case can set $k = 1$ and $B^t.a \approx V_0$, and find that the integral over β and σ separates into two Gamma function integrals:

$$pr(\log\sigma,\log\beta,M) \propto \beta^{M/2}\ \exp-\beta V_0/2\ \ \sigma^{M-N}\ \exp-V(M)/2\sigma^2.$$

This implies chi-squared distributions for β and $1/\sigma^2$ and estimators:

$$\langle\sigma^2\rangle \approx V(M)/(N-M)$$

as before, and

$$\langle 1/\beta\rangle \approx V_0/M,$$

leading to $V_0 \approx NR^2$ as predicted.

Further, integrating over $\log\sigma$ and $\log\beta$ we find:

$$\log pr(M) = \text{const.} + \log(\Gamma(M/2)) + \log(\Gamma((N-M)/2)) + \ldots$$

$$\ldots + (N-M)/2\ \log\ (V_0/V(M)).$$

The error in this formula is $O(MV(M)/V_0)$, which is small for $V(M)<<V_0$.

The performance of this formula can be judged by Figure 8, which shows the posterior distribution as a function of M. It is instructive to look again at the residuals shown in Figure 6: most people agree that there is clear evidence for a new parameter at M=9, but it would be a brave man that suggested one at M=10.

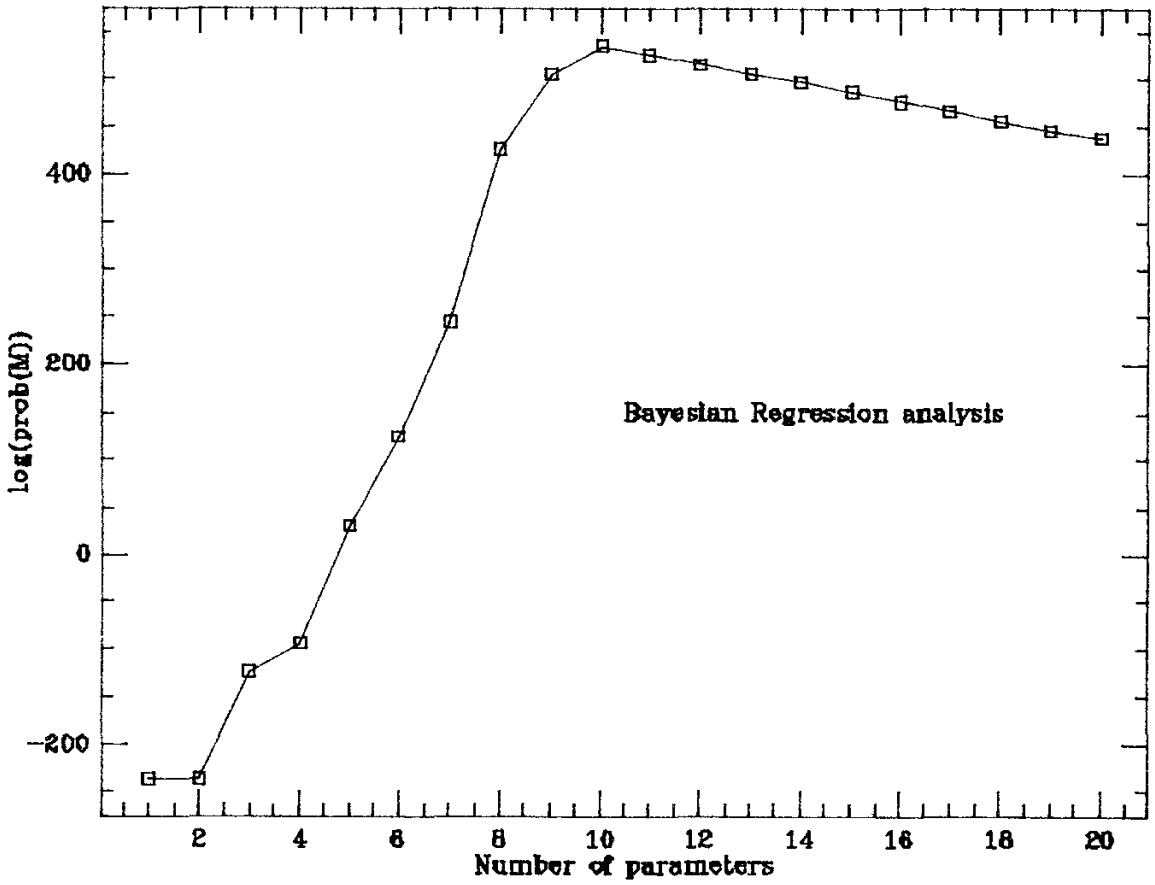

Figure 8. Posterior probability distribution of the polynomial expansion order for the example of Figure 6. There is a maximum at M = 10.

Sermon on the spike

When we perform the integral over β more carefully, by changing to $k\beta$ as a new variable, we encounter a singularity:

$$\int_0^{1/\sigma^2-\epsilon} d(k\beta)\ (k\beta)^{M/2-1}\ \exp(-k\beta B^t.\hat{a}/2)\ (1 - k\beta\sigma^2)^{-1}$$

where σ is related to the maximum allowed prior range of β: $\epsilon = 1/\sigma^2\beta_{max}$. There is, therefore, a tiny "spike" which gives a logarithmically divergent contribution to the integral. This behaviour is related to the fact that we took the range parameter β (or R) to be a scale parameter, and is again a warning that some aspect of that prior assumption remains relevant in the posterior distribution. Rather than being frightened of these spikes that occur in problems of this type, let us instead make a simple calculation to see just how relevant our prior information is, after the data have arrived. We can do this by calculating what the cutoff β_{max} would have to be in order to make a 50 per cent contribution to the integral. The main part of the integral is approximately $\Gamma(M/2)\ (2/V_0)^{M/2-1}$, and the spike involves

the value of the cutoff. Making suitable approximations we find that the fraction in the spike

$$\sim (V_0/2\sigma^2)^{M/2-1} \exp(-V_0/2\sigma^2) \log(\sigma^2\beta_{max}) / \Gamma(M/2)$$

For our example, with $V_0 = 4 \times 10^3$, $\sigma = 10^{-4}$, M=10, we find that the spike is important if:

$$\sigma^2\beta_{max} \sim \exp(\exp(2\times10^{11})),$$

where we are justified in ignoring a factor of 10^{45} as "small"! Such values of β_{max} are, of course, quite incomprehensible even to an astronomer, and indicates that our integral does indeed converge for all practical purposes.

However, let us think for a moment about the origin of this divergence. The data provide us with likelihood factors of about $\exp(-V_0/2\sigma^2)$, which are certainly large (see above), but nevertheless finite. The prior contains the range parameter β as a scale parameter, and in particular allows us to think of the limit $\beta_{max} \Longrightarrow \infty$, which corresponds to reducing the allowed range to zero. Eventually we arrive at the case where the prior is so sure that **a = 0** that it is incapable of learning from the data. If this situation is permitted without limit, then the finite likelihood factors will not be able to overwhelm the prior. The purpose of the above calculation was to show just how pig-headed one would have to be in order to ignore the data completely. In that respect it is telling us something useful. We note, finally, that there is no corresponding problem with the limit $\beta_{min} \Longrightarrow 0$.

CONCLUSIONS

We have given a selection of examples that illustrate the simplicity and power of Bayesian methods. Bayes' rule is used to manipulate probabilities in the light of experimental data; MaxEnt is used to assign probability distributions given testable information. However, it is up to us to choose a hypothesis space that is suitable for our problem, and this not only requires us to assign an appropriate measure in the space of possibilities, but to define a range of allowed values for any parameters involved.

The collapse of hypothesis space hyper-volume leads to a penalty for introducing a new parameter. This was first described by Jeffreys, but it is a very general phenomenon that deserves to be better known.

A tentative solution has been offered to the problem of

determining the optimal number of parameters in regression analysis. The essential feature of this solution is the attempt to treat all expansion orders equally, by relating their available parameter-space hyper-volumes to a common range parameter.

ACKNOWLEDGMENTS

Thanks are due to Yoel Tikochinsky and Larry Bretthorst for helpful comments, and to Gary Erickson for patience beyond the call of duty. The graph-fitting problem was motivated by a debate at the Laramie 1985 meeting between Peter Cheesman and Jorma Rissanen.

REFERENCES

Birkinshaw, M., Gull, S.F. & Hardebeck, H. (1984). Nature, 309, 34-35.

Brans, C. & Dicke, R.H. (1961). Phys. Rev., 124, 925.

Cox, R.P. (1946). Probability, Frequency and Reasonable Expectation. Am. Jour. Phys. 17, 1-13.

Gull, S.F. & Skilling, J. (1984). Maximum entropy method in image processing. IEE Proc.,131(F), 646-659.

Jaynes, E.T. (1968). Prior probabilities. Reprinted in E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics, ed. R. Rosenkrantz, 1983 Dordrecht: Reidel.

Jaynes, E.T. (1976). Confidence intervals versus Bayesian Intervals. Reprinted in E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics, ed. R. Rosenkrantz, 1983. Dordrecht: Reidel.

Jaynes, E.T. (1986). Bayesian Methods - an Introductory Tutorial. In Maximum Entropy and Bayesian Methods in Applied Statistics. ed. J.H. Justice. Cambridge University Press.

Jeffreys, H. (1939). Theory of Probability, Oxford University Press. Later editions 1948, 1961,1983.

Shore, J.E. & Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans.Info.Theory, IT-26, 26-39 and IT-29, 942-943.

Skilling, J. (1986). The Axioms of Maximum Entropy. Presented at 1986 Maximum Entropy conference, Seattle, Washington (this volume).

Skilling, J. & Gull, S.F. (1984). The entropy of an image. SIAM Amer. Math. Soc. proc. Appl. Math.,14, 167-189