20

Model comparison

Entities are not to be multiplied without necessity.

William of Ockham, c 1330

We have seen in some detail how to conduct inferences – test hypotheses, estimate parameters, predict future observations – within the context of a preassigned model, representing some working hypothesis about the phenomenon being observed. But a scientist must also be concerned with a bigger problem: how to decide between different models when both seem able to account for the facts. Indeed, the progress of science requires comparison of different conceivable models; a false premise built into a model that is never questioned cannot be removed by any amount of new data.

Stated very broadly, the problem is hardly new; some 650 years ago the Franciscan Monk William of Ockham perceived the logical error in the mind projection fallacy.[1] This led him to teach that some religious issues might be settled by reason, but others only by faith. He removed the latter from his discourse, and concentrated on the areas where reason might be applied – just as Bayesians seek to do today when we discard orthodox mind projecting mythology (such as assertions of limiting frequencies in experiments that have never been performed), and concentrate on the things that are meaningful in the real world. His propositions 'amenable only to faith' correspond roughly to what we should call non-Aristotelian propositions. His famous epigram quoted above, generally called 'Ockham's razor', represents a good start on the principles of reasoning that he needed, and that we still need today. But it was also so subtle that only through modern Bayesian analysis has it been well understood.

Of course, from our present vantage point it is clear that this is really the same problem as that of compound hypothesis testing, considered already in Chapter 4. Here we need only generalize that treatment and work out further details. Then we are able to see conventional significance tests simply as model comparison in which we choose the best alternative in a prescribed *class* of alternative hypotheses.

[1] Ockham's position, stated in the language of his time, was that 'Reality exists solely in individual things, and universals are merely abstract signs.' Translated into 20th century language: the abstract creations of the mind are not realities in the external world. Unfortunately for him, some of the cherished 'realities' of contemporary orthodox theology were just the things to which he denied reality; so this got him into trouble with the Establishment. Evidently, Ockham was a forerunner of modern Bayesians, to whom all this sounds very familiar.

But some extra care is needed. As long as we work within a single model, normalization constants tend to cancel out and so need not be introduced at all in most cases. But when two different models appear in a single equation, the normalization constants do not cancel out, and it is imperative that all probabilities be correctly normalized.

20.1 Formulation of the problem

To see why the normalization constants no longer cancel, recall first what Bayes' theorem tells us about parameter estimation. A model M contains various parameters denoted collectively by θ. Given data D and prior information I, to estimate its parameters we first apply Bayes' theorem:

$$p(\theta|DMI) = p(\theta|MI)\frac{p(D|\theta MI)}{p(D|MI)}, \tag{20.1}$$

in which the presence of M on the right-hand side signifies that we are assuming the correctness of model M. The denominator serves as the normalizing constant:

$$p(D|MI) = \int \mathrm{d}\theta \; p(D\theta|MI) = \int \mathrm{d}\theta \; p(D|\theta MI)p(\theta|MI), \tag{20.2}$$

which we see is the prior expectation of the likelihood $L(\theta) = p(D|\theta MI)$; i.e. its expectation over our prior probability distribution $p(\theta|MI)$ for the parameters.

Now we move up to a higher level problem: to judge, in the light of the prior information and data, which of a given set of different models $\{M_1, \ldots, M_r\}$ is most likely to be the correct one. Bayes' theorem gives the posterior probability for the jth model as

$$p(M_j|DI) = p(M_j|I)\frac{p(D|M_jI)}{p(D|I)}, \qquad 1 \le j \le r. \tag{20.3}$$

But we may eliminate the denominator $p(D|I)$ by calculating odds ratios as we did in Chapter 4. The posterior odds ratio for model M_j over M_k is

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)}\frac{p(D|M_jI)}{p(D|M_kI)}, \tag{20.4}$$

and we see that the same probability $p(D|M_jI)$ that appears in the single-model parameter estimation problem (20.1) only as a normalizing constant, now appears as the fundamental quantity determining the status of model M_j relative to any other.[2] The exact measure of what the data have to tell us about this is always the prior expectation of its likelihood function, over the prior probability $p(\theta_j|M_jI)$ for whatever parameters θ_j may be in that

[2] This logical structure is more general even than the Bayesian formalism; it persists in the pure maximum entropy formalism, where in statistical mechanics the relative probability P_j/P_k of two different phases, such as liquid and solid, is the ratio of their partition functions Z_j/Z_k, which are the normalization constants for the sub-problems of prediction within one phase. In Bayesian analysis, the data are indifferent between two models when their normalization constants become equal; in statistical mechanics the temperature of a phase transition is the one at which the two partition functions become equal. In Bayesian analysis we shall usually prefer to express (20.4) in log-odds form; in chemical thermodynamics it has been customary for a century to state the condition of indifference between phases as equality of the 'free energies' $F_j \propto \log(Z_j)$. This illustrates the basic unity of Bayesian and maximum entropy reasoning, in spite of their superficial differences arising from the different kind of information being processed.

model (they are generally different for different models). Probabilities must be correctly normalized here, otherwise we are violating our basic rules and the odds ratio in (20.4) is arbitrary.

Intuitively, the model favored by the data is the one that assigns the highest probability to the observed data, and therefore ‘explains the data’ best. This is just a repetition, at a higher level, of the likelihood principle for parameter estimation within a model.

But how can an Ockham principle emerge from this? The first difficulty is that the principle has never been stated in exact, well-defined terms. Later writers have tried, almost universally, to interpret our opening quotation as saying that the criterion of choice is the ‘simplicity’ of the competing models, although it is not clear that Ockham himself used that term. Perhaps we come closer to the notion of simplicity if we restate our opening quotation as: ‘Do not introduce details that do not contribute to the quality of your inferences.’ But centuries of discussion by philosophers brought no appreciable clarification of what is meant by ‘simplicity’.[3] We think that concentration of attention exclusively on that undefined term has prevented understanding of the real point, which is merely that a model with unspecified parameters is a composite hypothesis, not a simple one; and it requires the kind of analysis given in Chapter 4 for composite hypotheses. Then some new features appear, arising from the different internal structures of the parameter spaces for the models considered.

20.2 The fair judge and the cruel realist

Now consider under what conditions we want this model comparison to take place. There are two possible positions. (1) We might adopt the posture of the scrupulously fair judge, who insists that fairness in comparing models requires that each is delivering the best performance of which it is capable, by giving each the best possible prior probability for its parameters (similarly, in Olympic games we would consider it unfair to judge two athletes by their performances when one of them is sick or injured; the fair judge wants to compare them when both are doing their absolute best). (2) We might consider it necessary to be cruel realists and judge each model taking into account the prior information we actually have pertaining to it; that is, we penalize a model if we do not have the best possible prior information about its parameters, although that is not really a fault of the model itself.

It develops that the Ockham factors express the position of the cruel realist; they are just the factors that convert the scrupulously fair comparison of the model itself – irrespective of the prior probability we are able to give it at the moment – into the comparisons of the cruel realist who insists on taking into account what is actually possible here and now. An athlete who is sick or injured merits our sympathy, but we cannot use him in the ‘big game’ tomorrow; likewise, a potentially superior model can be unusable if our prior information places its parameters far from their maximum-likelihood values. When real results are at stake, we are obliged to be cruel realists.

[3] For a time, the notion of simplicity was given up for dead, because of the seeming impossibility of defining it. The tedious details are recounted by Rosenkrantz (1977).

20.2.1 Parameters known in advance

To see this, suppose first that there is no internal parameter space; the parameters of a model are known exactly ($\theta = \theta'$) in advance. Then the model becomes, in effect, a simple hypothesis rather than a composite or compound one, and the simple form of Bayes' theorem applies. This amounts to assigning a prior $p(\theta_j|M_j I) = \delta(\theta_j - \theta'_j)$, whereupon (20.2) reduces to

$$p(D|M_j I) = p(D|\theta'_j M_j I) = L_j(\theta'_j), \tag{20.5}$$

just the likelihood of θ'_j within the jth model. Evidently, the scrupulously fair judge will note that this is a maximum if θ'_j happens to be equal to the maximum-likelihood estimate $\hat{\theta}_j$ for that model and the data. Then his posterior odds ratio (20.4) reduces to

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)} \frac{(L_j)_{\max}}{(L_k)_{\max}}. \tag{20.6}$$

But this extreme case, although fair in the aforementioned sense, may be very unrealistic; usually, the parameters are unknown, and in the problems 'amenable to reason', where useful inferences are possible, our prior information concerning the parameters must be good enough to allow useful inferences.

We have seen already in previous chapters that, if we have a reasonable amount of data, most models will give such sharply peaked likelihood functions that the prior is relatively unimportant for inferences about the *parameters*. But it is still important for inferences about the *models*, so Ockham factors defined by the priors remain important in model comparison. The simple biological problems studied by R. A. Fisher are generally of this type.

When prior information is important even for inferences about the parameters – whether from a loose model or sparse data – Ockham factors make a crucially important difference in our model comparisons. In the more complex problems studied by Harold Jeffreys and faced by modern scientists and economists, we ignore these factors at our peril.

20.2.2 Parameters unknown

Let a model M have parameters $\theta \equiv \{\theta_1, \dots, \theta_m\}$. Then, comparing (20.4) and (20.6), we write

$$p(D|MI) = L_{\max} W, \tag{20.7}$$

and this defines the Ockham factor W; it is just the amount by which the model M is penalized by our nonoptimal prior information. Written explicitly,

$$W \equiv \int \mathrm{d}\theta \, \frac{L(\theta)}{L_{\max}} p(\theta|MI). \tag{20.8}$$

If, as in Fisher's problems, the data are much more informative about θ than the prior information, then the likelihood function is sharply peaked and we could define a 'high-likelihood region' Ω' as the smallest subregion of the whole parameter space Ω that contains a specified amount (say, 95%) of the integrated likelihood. Then, most of the contribution to the integral (20.8) would come from the region Ω'. Better, the arbitrary number 0.95 can be done away with by defining first the volume $V(\Omega')$ by the condition that the integrated likelihood is just

$$\int \mathrm{d}\theta\, L(\theta) = L_{\max} V(\Omega'). \tag{20.9}$$

Then Ω' is defined as the region of volume $V(\Omega')$ that contains the maximum possible amount of integrated likelihood; that is, within Ω' the likelihood is everywhere greater than some threshold value L_0 that is reached on the boundary of Ω'.

If the prior density $p(\theta|MI)$ is so broad that it is essentially constant over this high-likelihood region Ω' surrounding the maximum-likelihood point, (20.8) reduces to

$$W \simeq V(\Omega')p(\hat{\theta}|MI), \tag{20.10}$$

so in this case the Ockham factor is essentially just the *amount of prior probability* contained in the high-likelihood region Ω' picked out by the data.

In any case, our fundamental model comparison rule (20.4) becomes

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)} \frac{(L_j)_{\max}}{(L_k)_{\max}} \frac{W_j}{W_k}, \tag{20.11}$$

in which we see revealed, by comparison with (20.6), the net Ockham factor (W_j/W_k) arising from the internal parameter spaces of the models. In (20.11), the likelihood factor depends only on the data and the models. If two different models achieve the same likelihoods $(L_j)_{\max}$, then they are potentially capable of accounting for the data equally well, and in orthodox theory it would seem that we have no basis for choice between them. Yet Bayes' theorem tells us that there is an another quality to be considered: the prior information, which is ignored by orthodox theory, may still give strong grounds for preference of one model over the other. Indeed, the Ockham factor in (20.11) may be so strong that it reverses the likelihood judgment in (20.6).

20.3 But where is the idea of simplicity?

The relation (20.11) has much meaning that unaided intuition could not (or at least, did not) see. If the data are highly informative compared with the prior information, then the relative merit of two models is determined by two factors:

(1) how high a likelihood can be attained on their respective parameter spaces Ω_j, Ω_k;
(2) how much prior probability is concentrated in their respective high-likelihood regions Ω'_j, Ω'_k?

But neither of these seems concerned with the intuitive notion of simplicity (which seems for most of us to refer to the number of different assumptions that are made – for example, the number of different parameters that are introduced – in defining a model).

To understand this, let us ask: ‘How do we all decide these things intuitively?’ Having observed some facts, what is the real criterion that leads us to prefer one explanation of them over another? Suppose that two explanations, A and B, could account for some proven historical facts equally well. But A makes four assumptions, each of which seems to us already highly plausible, while B makes only two assumptions, but they seem strained, farfetched, and highly unlikely to be true. Every historian finds himself in situations like this, and he does not hesitate to opt for explanation A, although B is intuitively simpler. Thus our intuition asks, fundamentally, not how *simple* the hypotheses are, but rather how *plausible* they are.

Of course, there is a loose connection between plausibility and simplicity, because the more complicated a set of possible hypotheses, the larger the manifold of conceivable alternatives to some particular hypothesis, and so the smaller must be the prior probability of any particular hypothesis in the set.

Now we see why ‘simplicity’ could never be given a satisfactory definition (that is, a definition that accounted in a satisfactory way for these inferences); it was a poorly chosen word, directing one’s attention away from an essential component of the inference. But from centuries of unquestioned acceptance, the idea of ‘simplicity’ became implanted with such an unshakeable mindset that some workers, even after applying Bayes’ theorem where the contrary fact stares you in the face, continued doggedly trying to interpret the Bayesian analysis in terms of simplicity.[4]

Generations of writers opined vaguely that ‘simple hypotheses are more plausible’ without giving any logical reason for it. We suggest that this should be turned around: we should say rather that ‘more plausible hypotheses tend to be simpler’. An hypothesis that we consider simpler is one that has fewer equally plausible alternatives.

None of this could be comprehended at all within the confines of orthodox statistical theory, whose ideology did not allow the concept of a probability for a model or for a fixed but unknown parameter, because they were not considered ‘random variables’. Orthodoxy tried to compare models entirely in terms of their different sampling distributions, which took no note of *either* the simplicity of the model *or* the prior information! But it was unable to do even that, because then all the parameters, within a model became nuisance parameters, and that same ideology denied one any way to deal with them.[5] Thus, orthodox statistics was a total failure on this problem – it did not provide even the vocabulary in which the problem could be stated – and this held up progress for most of the 20th century.

It is remarkable that, although the point at issue is trivial mathematically, generations of mathematically competent people failed to see it because of that conceptual mindset. But once the point is seen, it seems intuitively obvious and one cannot comprehend how anyone could ever have imagined that ‘simplicity’ alone was the criterion for judging models. This just reminds us again that the human brain is an imperfect reasoning device; although it is quite good at drawing reasonable conclusions, it often fails to give a convincing

[4] Indeed, one author, for whom Ockham’s razor was *by definition* concerned with simplicity, rejected Bayesian analysis because of its failure to exhibit that error!

[5] This and other criticisms of orthodox hypothesis testing theory were made long ago by Pratt (1961).

rationale for those conclusions. For this we really do need the help of probability theory as logic.

Of course, Bayes' theorem does recognize simplicity as one component of the inference. But by what mechanism does this happen? Although Bayes' theorem always gives us the correct answer to whatever question we ask of it, it often does this in such a slick, efficient way that we are left bewildered and not quite understanding how it happened. The present problem is a good example of this, so let us try to understand the situation better intuitively.

Denote by M_n a model for which $\theta = \{\theta_1, \ldots, \theta_n\}$ is n-dimensional, ranging over a parameter space Ω_n. Now introduce a new model M_{n+1} by adding a new parameter θ_{n+1} and going to a new parameter space Ω_{n+1}, in such a way that $\theta_{n+1} = 0$ represents the old model M_n. We shall presently give an explicit calculation with this scenario, but first let us think about it in general terms.

On the subspace Ω_n the likelihood is unchanged by this change of model: $p(D|\theta M_{n+1} I) = p(D|\theta M_n I)$. But the prior probability $p(\theta|M_{n+1} I)$ must now be spread over a larger parameter space than before and will, in general, assign a lower probability to a neighborhood Ω' of a point in Ω_n than did the old model.

For a reasonably informative experiment, we expect that the likelihood will be rather strongly concentrated in small subregions $\Omega'_n \in \Omega_n$ and $\Omega'_{n+1} \in \Omega_{n+1}$. Therefore, if with M_{n+1} the maximum-likelihood point occurs at or near $\theta_{n+1} = 0$, Ω'_{n+1} will be assigned less prior probability than is Ω'_n with model M_n, and we have $p(D|M_n I) > p(D|M_{n+1} I)$; the likelihood ratio generated by the data will favor M_n over M_{n+1}. This is the Ockham phenomenon.

Thus, if the old model is already flexible enough to account well for the data, then as a general rule Bayes' theorem will, like Ockham, tell us to prefer the old model. It is intuitively simpler if by 'simpler' we mean a model that occupies a smaller volume of parameter space, and thus *restricts us to a smaller range of possible sampling distributions*. Generally, the inequality will go the other way only if the maximum-likelihood point is far from $\theta_{n+1} = 0$ (i.e. a significance test would indicate a need for the new parameter), because then the likelihood will be so much smaller on Ω'_n than on Ω'_{n+1} that it more than compensates for the lower prior probability of the latter; as noted, Ockham would not disagree.

But intuition does not tell us at all, quantitatively, how great this discrepancy in likelihoods must be in order to bring us to the point of indifference between the models. Furthermore, having seen this mechanism, it is easy to invent cases (for example, if the introduction of the new parameter is accompanied by a redistribution of prior probability on the old subspace S_n) in which Bayes' theorem may contradict Ockham because it is taking into account further circumstances undreamt of in Ockham's philosophy. So we need specific calculations to make these things quantitative.

20.4 An example: linear response models

Now we give a simple analysis that illustrates the above conclusions and allows us to calculate definite numerical values for the likelihood and Ockham factors. We have a common scenario: a data set $D \equiv \{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of measured values of (x, y) in n pairs of observations. We may think of x as the 'cause' and y as the 'effect', although this is not required. For the general relations below, the 'independent variables' x_i need not be uniformly spaced or even monotonic increasing in the index i. From these data and any prior information we have, we are to decide between two conceivable models for the process generating the data. For model M_1 the responses are, but for irregular measurement errors e_i, linear in the cause:

$$M_1 : \quad y_i = \alpha x_i + e_i, \qquad 1 \leq i \leq n, \tag{20.12}$$

while for model M_2 there is also a quadratic term:

$$M_2 : \quad y_i = \alpha x_i + \beta x_i^2 + e_i, \tag{20.13}$$

which represents, if β is negative, an incipient saturation or stabilizing effect (if β is positive, an incipient instability). We may think, for concreteness, of x_i as the dose of some medicine given to the ith patient, y_i as the resulting increase in blood pressure. Then we are trying to decide whether the response to this medicine is linear or quadratic in the dosage. But this mathematical model applies equally well to many different scenarios.[6] Whichever model is correct, we assume that the x_i are measured with negligible error, but the errors of measurement of y_i are supposed to be the same for either model, so we assign a joint sampling distribution to them:

$$p(e_1 \cdots e_n|I) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{e_i^2}{2\sigma^2}\right\} = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{w}{2}\sum_i e_i^2\right\} \tag{20.14}$$

where $w \equiv 1/\sigma^2$ is the 'weight' parameter, more convenient in calculations than σ^2. This simple scenario has the merit that all calculations can be done exactly with pencil and paper, so that the final result can be subjected to arbitrary extreme conditions and will remain correct, and we can see which limiting operations are and are not well-behaved.

20.4.1 Digression: the old sermon still another time

Again, we belabor the meaning of this, as discussed in Chapter 7. In orthodox statistics, a sampling distribution is always referred to as if it represented an 'objectively real' fact, the frequency distribution of the errors. But we doubt whether anybody has ever seen a real problem in which one had prior knowledge of any such frequency distribution, or indeed prior knowledge that any limiting frequency distribution exists.

[6] For example, x_i might be the amount of ozone in the air in the ith year, y_i the average temperature in that year. Or, x_i may be the amount of some food additive ingested by the ith Canadian rat, y_i the amount of cancer tissue that rat developed. Or, x_i may be the amount of acid rain falling on Northern Germany in the ith year, y_i the number of pine trees that died in that year; and so on. In other words, we are now in the realm of what were called 'linear response models' in the Preface, and the results of these calculations have a direct bearing on many currently controversial health and environmental issues. Of course, most real problems will require more sophisticated models than we are considering now, but having seen this simple calculation it will be clear how to generalize it in many different ways.

How could one ever acquire information about the long-run results of an experiment that has never been performed? That is part of the mind projecting mythology that we discard.

We recognize, then, that assigning this sampling distribution is only a means of describing our own *prior state of knowledge* about the measurement errors. The parameter σ indicates the general magnitude of the errors that we expect. The prior information I might, for example, be the variability observed in past examples of such data; or in a physics experiment it might not be the result of any observations, but rather obtained from the principles of statistical mechanics, indicating the level of Nyquist noise for the known temperature of the apparatus.

In particular, the absence of correlations in (20.14) is not an assertion that no correlations exist in the real data; it is only a recognition that we have no prior knowledge of such correlations, and therefore to suppose correlations of either sign is as likely to hurt as to help the accuracy of our inferences. In one sense, by being noncommittal about it, we are only being honest and frankly acknowledging our ignorance. But in another sense, we are taking the safest, most conservative, course; using a sampling distribution which will yield reasonable results *whether or not correlations actually exist*. But if we knew of any such correlations, we would be able to make still better inferences (although not much better) by use of a sampling distribution which contains them.

The reason for this is that correlations in a sampling distribution tell the robot that some regions of the vector sample space are more likely than others, even though they have the same mean-square error magnitudes $\overline{e^2}$; then some details of the data that it would otherwise have to dismiss as noise can be recognized as providing further evidence about systematic effects in the model.

We return to the problem. The sampling distribution for model M_1 is

$$M_1: \qquad p(D|\alpha M_1) = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{nw}{2} Q_1(\alpha)\right\} \tag{20.15}$$

with the quadratic form

$$Q_1(\alpha) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha x_i)^2 = \overline{y^2} - 2\alpha\overline{xy} + \alpha^2\overline{x^2}, \tag{20.16}$$

where the bars denote averages. The maximum-likelihood estimate of α is then found from $\partial Q_1/\partial\alpha = 0$, or,

$$\alpha = \hat{\alpha} \equiv \frac{\overline{xy}}{\overline{x}}, \tag{20.17}$$

which in this case is also called the ‘ordinary least squares’ estimate. Supposing the weight w is known; the likelihood (20.15) for model M_1 is then

$$L_1(\alpha) = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{nw}{2}\left[\overline{y^2} + \overline{x^2}(\alpha - \hat{\alpha})^2 - \overline{x^2}\hat{\alpha}^2\right]\right\} \tag{20.18}$$

in which we could discard any factor independent of α, but that will disappear presently of its own accord, in (20.23). If we were using this to estimate α from the data alone, our result would be

$$(\alpha)_{\text{est}} = \hat{\alpha} \pm \frac{1}{\sqrt{nw\overline{x^2}}} = \hat{\alpha} \pm \frac{1}{\sqrt{n}} \frac{\sigma}{x_{\text{rms}}} = \hat{\alpha} \pm \delta\alpha, \tag{20.19}$$

where $x_{\text{rms}} = \sqrt{\overline{x^2}}$ is the root-mean-square value of the x_i. Thus the volume (in this case, width) of the high-likelihood region Ω' may be taken as roughly $V(\Omega') = 2(\delta\alpha)$.

Now, using (20.17), the 'global' sampling distribution for model M_1 in (20.3) contains two factors:

$$p(D|M_1 I) = \int \mathrm{d}\alpha \; p(D|\alpha M_1) p(\alpha|M_1 I) = L_{\max}(M_1) W_1, \tag{20.20}$$

where

$$L_{\max}(M_1) = L_1(\hat{\alpha}). \tag{20.21}$$

The Ockham factor for model M_1 is therefore

$$W_1 = \int \mathrm{d}\alpha \; \frac{L_1(\alpha)}{L_1(\hat{\alpha})} p(\alpha|M_1 I), \tag{20.22}$$

and we find for the likelihood ratio

$$\frac{L_1(\alpha)}{L_1(\hat{\alpha})} = \exp\left[-\frac{nw\overline{x^2}}{2} (\alpha - \hat{\alpha})^2 \right]. \tag{20.23}$$

This makes it evident that $W_1 \leq 1$, since the likelihood ratio cannot exceed unity and the prior is normalized.

Now we must assign a prior for α. Usually, we will have some reason, such as previous experience with such problems, for guessing a value of the general order of magnitude of some quantity α_0, but we are not at all confident of the accuracy of that guess, except to think that $|\alpha - \alpha_0|$ cannot be enormously large (else there would be such a catastrophe that we would not be concerned with this problem); but we would seldom have any more specific prior information about it. We can indicate this by assigning the normalized prior density

$$p(\alpha|M_1 I) = \sqrt{\frac{w_0}{2\pi}} \exp\left\{ -\frac{w_0}{2} (\alpha - \alpha_0)^2 \right\}, \tag{20.24}$$

which says that we think it unlikely that $|\alpha - \alpha_0|$ is much greater than $\sigma_0 = 1/\sqrt{w_0}$. From both the central limit theorem as discussed in Chapter 7 and the maximum entropy principle as discussed in Chapter 11, this Gaussian functional form of prior is preferred in principle over all others as representing the actual state of knowledge that we have in virtually all real problems. Then it is fortunate that this form also enables us to do the integration (20.22) exactly, with the result

$$W_1 = \sqrt{\frac{w_0}{nw\overline{x^2} + w_0}} \exp\left\{-\frac{nw\overline{x^2}w_0}{2(nw\overline{x^2} + w_0)}(\hat{\alpha} - \alpha_0)^2\right\}. \tag{20.25}$$

Rewriting this in terms of the half-width $\delta\alpha = 1/\sqrt{nw\overline{x^2}}$ of the high-likelihood region and the half-width $\sigma_0 = 1/\sqrt{w_0}$ of the prior for α, it becomes

$$W_1 = \frac{1}{\sqrt{1 + (\sigma_0/\delta\alpha)^2}} \exp\left\{-\frac{(\hat{\alpha} - \alpha_0)^2}{2\sigma_0^2}\right\}. \tag{20.26}$$

This has several limiting forms. If the prior estimate α_0 is exactly equal to the ordinary least squares estimate $\hat{\alpha}$, it reduces to

$$W_1 = \frac{1}{\sqrt{1 + (\sigma_0/\delta\alpha)^2}}. \tag{20.27}$$

Then, if $\sigma_0 \gg \delta\alpha$, we have

$$W_1 \simeq \frac{\delta\alpha}{\sigma_0}, \tag{20.28}$$

which is indeed just the amount of prior probability contained in the high-likelihood region. In this case, the Ockham factor is the ratio by which the parameter space is contracted by the information in the data, which expresses how much the vagueness of our prior information deteriorates the performance of model M_1, by placing prior probability outside its high-likelihood region. If the prior estimate α_0 differs from the ordinary least squares estimate $\hat{\alpha}$ by less than σ_0, this remains approximately correct.

If in (20.27), $\sigma_0 \to 0$, we have $W_1 \to 1$, the maximum possible value; if the prior information already told us exactly the ordinary least squares estimate from the data, with zero error tolerance, model W_1 is not penalized at all. But in all other cases there is some penalty. For example, if $|\alpha_0 - \hat{\alpha}| \gg \sigma_0$, then the evidence of the data strongly contradicts the prior information, and the model is severely penalized.

For model M_2 the sampling distribution is still given by (20.15), but now with the quadratic form

$$Q_2(\alpha, \beta) \equiv \frac{1}{n}\sum(y_i - \alpha x_i - \beta x_i^2)^2 = \overline{y^2} + \alpha^2\overline{x^2} + \beta^2\overline{x^4} - 2\alpha\overline{xy} - 2\beta\overline{x^2y} + 2\alpha\beta\overline{x^3}, \tag{20.29}$$

and the maximum-likelihood estimates $(\hat{\alpha}, \hat{\beta})$ are now the roots of the simultaneous equations $\partial Q_2/\partial\alpha = 0, \quad \partial Q_2/\partial\beta = 0$, or

$$\begin{aligned} \overline{x^2}\hat{\alpha} + \overline{x^3}\hat{\beta} &= \overline{xy} \\ \overline{x^3}\hat{\alpha} + \overline{x^4}\hat{\beta} &= \overline{x^2y}, \end{aligned} \tag{20.30}$$

of which the solution is

$$\hat{\alpha} = \frac{(\overline{x^4})(\overline{xy}) - (\overline{x^3})(\overline{x^2y})}{(\overline{x^2})(\overline{x^4}) - (\overline{x^3})^2}, \qquad \hat{\beta} = \frac{(\overline{x^2})(\overline{x^2y}) - (\overline{x^3})(\overline{xy})}{(\overline{x^2}(\overline{x^4}) - (\overline{x^3})^2}, \tag{20.31}$$

and we note that, as $\overline{x^3} \to 0$, these relax into estimates

$$\hat{\alpha} \to \frac{\overline{xy}}{\overline{x^2}}, \qquad \hat{\beta} \to \frac{\overline{x^2y}}{\overline{x^4}}, \tag{20.32}$$

where $\hat{\alpha}$ is the ordinary least squares estimate found using model M_1 (20.17). Now, as in (20.22), the Ockham factor for model M_2 is

$$W_2 = \int \mathrm{d}\alpha \int \mathrm{d}\beta \, \frac{L_2(\alpha, \beta)}{L_2(\hat{\alpha}, \hat{\beta})} p(\alpha\beta|M_2 I), \tag{20.33}$$

and, after some rather tedious algebra, we find that the likelihood ratio just constructs a familiar quadratic form:

$$\frac{L_2(\alpha, \beta)}{L_2(\hat{\alpha}, \hat{\beta})} = \exp\left\{-\frac{nw}{2} Q(\alpha, \beta)\right\}, \tag{20.34}$$

where

$$\begin{aligned} Q(\alpha, \beta) &\equiv Q_2(\alpha, \beta) - Q_2(\hat{\alpha}, \hat{\beta}) \\ &= \overline{x^2}(\alpha - \hat{\alpha})^2 + 2\overline{x^3}(\alpha - \hat{\alpha})(\beta - \hat{\beta}) + \overline{x^4}(\beta - \hat{\beta})^2. \end{aligned} \tag{20.35}$$

Now we assign a joint prior

$$p(\alpha\beta|M_2 I) = \sqrt{\frac{w_0}{2\pi}} \exp\left\{-\frac{w_0}{2}(\alpha - \alpha_0)^2\right\} \sqrt{\frac{w_1}{2\pi}} \exp\left\{-\frac{w_1}{2}(\beta - \beta_0)^2\right\} \tag{20.36}$$

in which w_0, a_0 are the same as in (20.24), so that the marginal prior for α is the same in the two models (otherwise we would be changing two different circumstances instead of one in going from M_1 to M_2, which would make the results very hard to interpret):

$$p(\alpha|M_1 I) = p(\alpha|M_2 I). \tag{20.37}$$

The Ockham factor for model M_2 is then

$$W_2 = \frac{\sqrt{w_0 w_1}}{2\pi} \int \mathrm{d}\alpha \int \mathrm{d}\beta \exp\left\{-\frac{1}{2}\left[nw\, Q(\alpha, \beta) + w_0(\alpha - \alpha_0)^2 + w_1(\beta - \beta_0)^2\right]\right\}, \tag{20.38}$$

and again this integration can be carried out exactly, with the result

$$W_2 = \sqrt{\frac{w_0 w_1}{(w_0 + nw\overline{x^2})(w_1 + nw\overline{x^4})}}\ \exp\{x\}. \tag{20.39}$$

Editor's Exercise 20.1. As written, the denominator in (20.39) is correct only if the condition $\overline{x^3} \to 0$ is used. Using this simplifying assumption, derive W_2 and define x.

The net Ockham factor in favor of M_1 over M_2 is computed from (20.27) and (20.39):

$$\frac{W_1}{W_2} = \frac{1/\sqrt{1 + (\sigma_0/\delta\alpha)^2}}{\sqrt{(w_0 w_1)/(w_0 + nw\overline{x^2})(w_1 + nw\overline{x^4})}\exp\{x\}}. \tag{20.40}$$

Editor's Exercise 20.2. Rewrite (20.40) in terms of the half-widths: $\delta\alpha = 1/\sqrt{nw\overline{x^2}}$, $\sigma_0 = 1/\sqrt{w_0}$, $\delta\beta = 1/\sqrt{nw\overline{x^4}}$, and $\sigma_1 = 1/\sqrt{w_1}$. Under what conditions will model M_2 will be favored over M_1?

20.5 Comments

Actual scientific practice does not really obey Ockham's razor, either in its previous 'simplicity' form or in our revised 'plausibility' form. As so many of us have deplored, the attractive new hypothesis or model, which accounts for the facts in such a neat, plausible way that you want to believe it at once, is usually pooh-poohed by the official Establishment in favor of some drab, complicated, uninteresting one; or, if necessary, in favor of no alternative at all. The progress of science is carried forward mostly by the few fundamental dissenting innovators, such as Copernicus, Galileo, Newton, Laplace, Darwin, Mendel, Pasteur, Boltzmann, Einstein, Wegener, Jeffreys – all of whom had to undergo this initial rejection and attack. In the cases of Galileo, Laplace, and Darwin, these attacks continued for more than a century after their deaths. This is not because their new hypotheses were faulty – quite the contrary – but because this is part of the sociology of science (and, indeed, of all scholarship). In any field, the Establishment is seldom in pursuit of the truth, because it is composed of those who sincerely believe that they are already in possession of it.

Progress is delayed also by another aspect of this. Scholars who failed to heed the teachings of William of Ockham about issues amenable to reason and issues amenable only to faith, were – and still are – doomed to a lifetime of generating nonsense. We note the most common form this nonsense has taken in the past.

20.5.1 Final causes

It seems that every discussion of scientific inference must deal, sooner or later, with the issue of belief or disbelief in final causes. Expressed views range all the way from Jacques Monod (1970) forbidding us even to mention purpose in the Universe, to the religious fundamentalist who insists that it is evil not to believe in such a purpose. We are astonished by the dogmatic, emotional intensity with which opposite views are proclaimed, by persons who do not have a shred of supporting factual evidence for their positions.

But almost everyone who has discussed this has supposed that by a 'final cause' one means some supernatural force that suspends natural law and takes over control of events (that is, alters positions and velocities of molecules in a way inconsistent with the equations of motion) in order to ensure that some desired final condition is attained. In our view, almost all past discussions have been flawed by failure to recognize that operation of a final cause does not imply controlling molecular details.

When the author of a textbook says: 'My purpose in writing this book was to . . .', he is disclosing that there was a true 'final cause' governing many activities of writer, pen, secretary, word processor, extending usually over several years. When a chemist imposes conditions on his system which forces it to have a certain volume and temperature, he is just as truly the wielder of a final cause dictating the final thermodynamic state that he wished it to have. A bricklayer and a cook are likewise engaged in the art of invoking final causes for definite purposes. But – and this is the point almost always missed – these final causes are *macroscopic*; they do not determine any particular 'molecular' details. In all cases, had those fine details been different in any one of billions of ways, the final cause would have been satisfied just as well.

The final cause may then be said to possess an entropy, indicating the number of microscopic ways in which its purpose can be realized; and the larger that entropy, the greater is the probability that it will be realized. Thus the principle of maximum entropy applies also here.

In other words, while the idea of a microscopic final cause runs counter to all the instincts of a scientist, a macroscopic final cause is a perfectly familiar and real phenomenon, which we all invoke daily. We can hardly deny the existence of purpose in the Universe when virtually everything we do is done with some definite purpose in mind. Indeed, anybody who fails to pursue some definite long-run purpose in the conduct of his life is dismissed as an idler by his colleagues. Obviously, this is just a familiar fact with no religious connotations – and no anti-religious ones. Every scientist believes in macroscopic final causes without thereby believing in supernatural contravention of the laws of physics. The wielder of the final cause is not suspending physical law; he is merely choosing the Hamiltonian with which some system evolves *according to physical law*. To fail to see this is to generate the most fantastic, mystical nonsense.