SHARPENING OCKHAM'S RAZOR
ON A BAYESIAN STROP

by

William H. Jefferys and James O. Berger
University of Texas at Austin — Purdue University

Technical Report #91-44C

Department of Statistics
Purdue University

August 1991

◇

Sharpening Ockham's Razor on a Bayesian Strop
(Key terms: Bayes' theorem; Ockham's razor)

William H. Jefferys
Department of Astronomy
University of Texas at Austin
Austin, TX 78712

James O. Berger
Department of Statistics
Purdue University
West Lafayette, IN 47907

Ockham's Razor, an established principle used every day in science, has deep connections with Bayesian reasoning, which traces directly back to Sir Harold Jeffreys' pioneering work on statistics during the 1920s and 30s. In this paper we shall explore the connection between Ockham's Razor and Bayesian statistics, and present an objective quantification of Ockham's Razor.

"Pluralitas non est ponenda sine necessitate."
—William of Ockham

Introduction.

Ockham's razor, that is, the principle that an explanation of the facts should be no more complicated than necessary, is an accepted principle in science. Over the years it has proven to be an effective tool for weeding out unprofitable hypotheses, and scientists use it every day, even when they do not cite it explicitly. See Thorburn (1918) for a history of the principle.

Ockham's razor is usually thought of as an heuristic, that experience has shown to be an effective tool. It is less widely known that under some circumstances, it can also be regarded as a *consequence* of deeper principles. This fact is implicit in Harold Jeffreys' book on probability (1939), and has more recently been emphasized by Good (1968, 1977), Jaynes (1979), Smith and Spiegelhalter (1980), Gull (1988), Loredo (1989), and MacKay (1991).

jeffreys (1939) considered the problem of fitting observed data to an empirical function. For a falling body, he considers the law

$$s = a + ut + \frac{1}{2}gt^2, \tag{1}$$

where a, u and g are adjustable parameters. So far this is only a standard problem in estimation theory. However, there are infinitely many possible laws that can represent the data set. For example, Jeffreys considers alternative laws of the form

$$s = a + ut + \frac{1}{2}gt^2 + a_3t^3 + ... + a_nt^n, \tag{2}$$

where n is greater than the number of observations and all coefficients are adjustable. Given such a law, there are infinitely many choices of the parameters that will exactly

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

fit the data, and the question is, why do we prefer (1) over (2)? The easy answer, given by Ockham's razor, is that we ought to prefer (1) to (2), provided that (1) adequately represents the observed data, on the grounds that (2) is unnecessarily complicated. On the other hand, (2) can actually represent the observed data points better than (1), since for large enough n it can be arranged to pass exactly through each data point (Figure 1). So there must be something other than the ability to fit the data that leads us to prefer the simpler law to the more complex.

Prior probabilities.

Jeffreys (1939; p. 47) suggested that the reason that we favor the simpler law is that it has higher prior probability than the more complex law (see *Box 1: Bayes' Theorem* for an explanation of the terminology). This is certainly a reasonable idea. Scientists know from experience that Ockham's razor works, and for them to reflect this experience in choosing their prior probabilities so that they favor the simpler hypothesis would be in accord with that experience. In fact, even though scientists do not usually think in terms of prior probabilities when considering a hypothesis, when they consider simple hypotheses before complex ones, they in effect are doing just this. This approach is also consistent with the tentative and step-by-step nature of science, whereby a hypothesis is taken as a working hypothesis, and altered and refined as new data become available.

Wrinch and Jeffreys proposed codifying this as a rule that would automatically give higher prior probability to laws that have fewer parameters (Wrinch and Jeffreys 1921; Jeffreys 1939). This approach would lead us to try simpler laws first, only moving on to more complicated laws as we find that the simple ones are not adequate to represent the data. So this approach provides a sort of rationalized "Ockham's Razor."

As appealing as Jeffreys' appeal to prior probabilities is, it seems that this answer may beg the question. Jeffreys lays out no clear rule for assigning prior probabilities to the various hypotheses, as he himself points out (Jeffreys 1939; p. 49). But he also points out a connection with significance tests. He shows that if one makes reasonable assumptions about the probability model, the simpler law can have greater *posterior* probability on the data than the more complex law, even if the prior probability does not favor the simple law. It is this fact that might lead one to conclude that (1) is a better choice than (2). Jeffreys never stated in so many words that this result was a form of Ockham's razor, although it seems likely that he was aware of it; the first to point out the connection explicitly appears to have been Jaynes (1979), and independently Smith and Spiegelhalter (1980), who called it an "automatic Ockham's razor," automatic in the sense that it does not depend on the prior probabilities of the hypotheses. This razor is not completely automatic, however, in that it does depend on probabilistic modeling of the effect of the complex law on the data. In Berger and Jefferys (1991) it is observed that even this input can often be avoided, leading to an objective quantification of Ockham's razor. This objective razor will be described after reviewing and illustrating the fundamental relationship between Ockham's razor and Bayesian analysis.

Plagiarism and authorship.

We will regard a hypothesis H_0 as *simpler* than an alternative H_1 if it makes sharper predictions about what data will be observed. Usually this will be manifested in the fact that the more complex hypothesis can accommodate a larger set of potential observations than the simpler one. This in turn means that the simpler hypothesis is more readily falsified by arbitrary data. For example, suppose a friend (who has a reputation as a prankster) offers to flip a coin to decide who will perform a little chore: heads he wins, tails he loses. Knowing our friend's reputation, we might well be concerned that he would use trickery (like a two-headed coin) to win the toss. The hypothesis H_0 that the coin has two heads is, under this understanding, a simpler one than the hypothesis H_1 that the coin is fair. For if we toss the coin N times, then H_0 will be falsified if even a single tail comes up, whereas H_1 would be *consistent* with (i.e., not falsified by) any sequence of heads and tails.

Suppose, before the coin is flipped, we believed that the hypotheses H_0 and H_1 were equally likely. Suppose, despite our fears and for the sake of friendship, we agree to the coin-flip anyway. The coin is tossed, and it indeed comes up heads. Our belief in the two hypotheses will change as a result of this information, and (by Bayes' theorem) the posterior probability that we assign to H_0 will now be twice that which we assign to H_1 (Figure 2). However, the evidence that our friend is trying to fool us is not very strong at this point, perhaps not strong enough to challenge him for a close look at the coin. However, if this sort of thing happened to us on five separate occasions in a row, we would be rather inclined to think that our friend was playing a joke on us (Figure 3). Even though both hypotheses are consistent with the data, the simpler one is now considerably more credible.

This notion of simplicity is characterized by the fact that the simpler hypothesis divides the set of observable outcomes $\{D_k\}$ into a small set that has a relatively high probability of being observed and a large set that has a relatively small probability of being observed; whereas the more complex hypothesis tends to spread the prior probability more evenly amongst all the outcomes. This might be because there are adjustable parameters in the more complex hypothesis that allows it to accommodate a larger range of data (i.e., it has more "knobs" to adjust); or, as in the coin-tossing example, because one of the hypotheses restricts the possible outcomes more than another, as seen in Figure 3.

For example, in the days before computers, when mathematical tables were king, the compiler of a table had to contend with possible copyright infringement. Copyright law protects only the expression of an idea, not the idea itself. Since the numbers in a mathematical table do not depend upon who computes them, this poses a dilemma for the compiler. How could one demonstrate to the satisfaction of a court that a table had been copied, instead of calculated *de novo*? Clearly, a method was needed to prove beyond a reasonable doubt that plagiarism had taken place. To solve this problem, compilers frequently took advantage of the fact that numbers ending in the digit 5 can be rounded either up or down without significantly affecting the result of a calculation. By rounding such numbers randomly, for example by deciding whether to round up or down by the toss of a coin, instead of using the grade-school rule "round a number ending in 5 to the nearest even number," the compiler could embed a secret code in the table that identified the table as his work, while not significantly affecting the accuracy of the results that people would calculate using the table. Anyone who published a table with the same pattern of roundings could be prosecuted for copyright violation, since the probability of obtaining the same pattern by chance was very small.

For example, suppose that we had published a table of sines containing 1000 entries. Of these, about 100 (one in ten) would have originally ended in the digit 5 and would have been rounded either up or down at random. Two independent compilers of a table of sines would be very unlikely to come up with the same pattern of rounding, since there are $N = 2^{-100} \approx 10^{-30}$ different ways to round the 5s in the table.

As authors of a table of sines, suppose that we learned that a newly published table by another author had the same rounding pattern as our own. Let H_0 be the hypothesis that the second table was plagiarized from the first, and H_1 be the hypothesis that the second table was independently generated and just happened to have the same pattern of roundings. On the data D that the rounding patterns are identical, we can calculate that $P(D \mid H_0) = 1$ and $P(D \mid H_1) = 2^{-100} \approx 10^{-30}$. Assuming equal *prior* probabilities for the two hypotheses, one finds from the formulas in *Box 1* that the *posterior* probability that plagiarism took place differs only negligibly from 1.

The reason for this is that H_0 makes a precise prediction about what will be seen, and is inconsistent with almost all possible data, whereas H_1 is consistent with any data that might be observed. In essence, H_1 tries to "have its cake and eat it too," hedging its bets by trying to accommodate all possible data. In contrast, H_0 boldly risks everything on a single possibility. As a result, when that single possibility turns out to be true, H_0 is rewarded for the greater risk it takes by being given a very high posterior probability compared to H_1, even though H_1 is also *consistent* with the data that we observe and cannot be *falsified* by the simple fact that the two tables agreed perfectly.

This simple example does not take into account other factors, for example, the possibility that a plagiarizer might make errors when copying the table. In such a case, not only would a table containing the exact pattern of roundings as the first be suspect, but also any table containing a pattern of roundings that differed only slightly from the first. For example, the plagiarizer might have calculated some of the numbers in the table himself, gotten tired, and then copied the rest, or he may have simply been incompetent and made mistakes. In either case, the basic ideas of the above calculation are not changed much. For example, suppose that the plagiarizer copied the table, but with a probability $0 < p \ll 1$ of making a mistake. Then the probability that the copied table would agree exactly with the original table is $(1-p)^{100}$, the probability that they would differ in exactly one place is $C_1^{100} p(1-p)^{99}$, the probability that they would differ in exactly two places is $C_2^{100} p^2 (1-p)^{98}$, and so on. (Here, C_n^m is the binomial coefficient for the number of ways we can choose n objects from a set of m objects.) If D is the observation that the two tables differ on exactly n out of the 100 digits, then the conditional probability of observing D given H_0 is no longer 1 but is given by $P(D \mid H_0) = C_n^{100} p^n (1-p)^{100-n}$. Similarly, we have that $P(D \mid H_1) = C_n^{100} 2^{-100}$. As long as $P(D \mid H_0)$ is much greater than $P(D \mid H_1)$, the probability that plagiarism took place remains close to one.

It is now routine for authors of directories, maps, mailing lists, and similar compilations deliberately to introduce innocuous but erroneous entries into the material. When plagiarism or other unauthorized use of the material takes place, the presence of these errors in the copied material proves beyond a reasonable doubt that copyright violation has occurred, and gives the compiler very strong evidence should the case should go to trial.

At McGill University, chemistry professors David Harpp and James Hogan have used a similar idea to detect cheating on multiple-choice tests. They wrote a computer program to compare the answers given by each pair of students in the class and look for a near-match between correct and incorrect answers. Of course, as teachers they hope and expect students to know the subject material, so that conclusions about cheating cannot be drawn from a student's correct answers. But if two students make the same *errors*, the evidence in favor of cheating as against accidental duplication of wrong answers can be quite significant. When two papers with similar patterns of wrong answers are found, Harpp and Hogan's program then checks the wrong answers to see if there is also a matching pattern amongst the incorrect responses. If this also looks suspicious, they then check where the two students were sitting. According to Harpp (1991), "In over fifteen exams I monitored, every suspect pair sat within one seat of each other." Most people, after considering such evidence, would agree that there is a very high probability that cheating had taken place. The analysis of the data in this case is more complicated than in our first example, however, because different questions will be answered incorrectly with differing frequencies, and because the various incorrect responses (distractors) for each question will similarly draw different numbers of responses. But there are practical solutions for these complications.

In evolutionary biology, molecular biologists have found striking evidence in support of the principle of descent with modification within the genetic messages carried by all organisms. One example is pseudogenes (Max 1986; Watson *et. al.* 1988). Pseudogenes can arise in one of several ways, when information from functional genes is duplicated with errors, such as deletion of critical codes required for gene expression, that render them nonfunctional. If a pseudogene arises in an organism, then it will be passed on to its progeny, even though the gene itself has lost its function. Thus, if we look at different species (such as humans and chimpanzees), and find that identical or nearly identical pseudogenes appear in each, this constitutes very powerful evidence in favor of the hypothesis that the two different species have a common ancestor. Just as with cheating on multiple-choice tests, or plagiarism of compiled materials, it is the verbatim or near-verbatim repetition of a mistake (in this case, the functionless pseudogene containing hundreds of "letters" of the genetic alphabet) that gives the hypothesis of copying—in evolutionary terms, descent from a common ancestor—a high posterior probability, when compared with hypotheses that are not limited by the severe constraints imposed by the principle of descent with modification.

The motion of Mercury's perihelion.

Ever since Leverrier's work in the early 19th century, astronomers were aware of a serious problem with the theory of Mercury's motion. Newtonian theory, which had been extraordinarily successful in accounting for most of the motions in the solar system, had run up against a small discrepancy in the motion of Mercury that it could not explain easily. After all of the perturbing effects of the planets had been taken into account, there remained an unexplained residual motion of Mercury's perihelion (the point in its orbit where the planet was closest to the Sun) in the amount of approximately 43 seconds of arc per century.

Clearly, it seemed as if something had been overlooked. It was known that physical mechanisms existed that might explain the discrepancy. One that seemed particularly appealing in the light of recent experience was the possibility that another planet might exist, closer to the Sun than Mercury. The reason that this idea was so appealing was that Leverrier himself, along with the English astronomer Adams, had recently (in 1846) met with brilliant success by predicting that a previously unknown planet was responsible for the known discrepancies in the motion of Uranus; not only did Leverrier and Adams hypothesize that such a planet existed, but they also suggested where it might be found, and indeed, when J.G. Galle looked for it, the planet Neptune was discovered in the predicted place. It certainly seemed possible that a similar phenomenon might explain the anomaly in Mercury's motion.

Indeed, a number of astronomers duly set out to find the new planet, dubbed 'Vulcan' in anticipation of its discovery, and some sightings were announced. However, the sightings could not be confirmed, and over time interest in the Vulcan hypothesis waned.

Other mechanisms that might explain the anomaly were also proposed. It was suggested that rings of material around the Sun could, if massive enough, produce the observed effect; or, the Sun itself might be slightly oblate, due to its rotation on its axis; or, finally, the law of gravity itself might not be exactly right. For example, the great American astronomer Simon Newcomb (1895) proposed that the exponent in Newton's law of gravity might not be exactly 2, but instead might be $2+\epsilon$, although other modifications to the law of gravity were also possible.

All these hypotheses had one characteristic in common: they possessed parameters that could be adjusted to agree with whatever data on the motion of Mercury existed. In modern parliance, we would call this a 'fudge factor.' For example, the Vulcan hypothesis had the mass and orbit of the putative planet; the ring hypothesis had the mass and location of the ring of material; the solar oblateness hypothesis had the unknown amount of the oblateness; and all the hypotheses that modified Newton's law of gravity had an adjustable parameter (like Newcomb's ϵ) that could be chosen at will.

Not all the hypotheses were equally probable, however (Roseveare 1982). As we stated above, sightings of 'Vulcan' were never confirmed, for example. As time went on, the hypothesis of matter rings of sufficient density became less and less likely (Jeffreys 1921) although some still believed in them (Poor 1921). A solar oblateness of sufficient size probably would have been detectable with 19th century techniques. However, the hypothesis that Newton's law of gravity needed an arbitrary adjustment to fit the data could not be ruled out.

What happened historically is well known. In 1915, Einstein announced his theory of general relativity, one of the consequences of which was that there should be an excess advance in the perihelion motion of the planets, that turned out to be largest in the case of Mercury. After some confusion (Roseveare 1982: pp. 154-159) it soon became clear that the amount of the advance predicted by general relativity was very close to the unexplained discrepancy in Mercury's motion. The amazing thing was that the predicted value, which is $42.98''$/century using modern values (Nobili & Wills 1986) was *not* some kind of fudge factor, but instead was an inevitable consequence of Einstein's theory!

As is well known, Einstein's theory made two other major predictions in addition to Mercury's perihelion motion (gravitational bending of light, and the slowing down of clocks in a gravitational field). There has been a lively debate over the years as to how important each has been in convincing scientists that general relativity was the correct theory of gravity (Brush 1989). In this paper we will not go into this argument, but will instead try to put ourselves into the mindset of a Bayesian observer in the early 1920s, who is trying to weigh the evidence of Mercury's motion.

An interesting pair of papers was published in 1921 (Poor, 1921; Jeffreys, 1921). Poor was an astronomer at Columbia University, who had not been convinced that general relativity was correct and still clung to the matter ring theory. Unfortunately, he also made some serious errors in his assessment of the evidence as regards the other inner planets. Jeffreys, in response, argued persuasively that the ring theory was not viable because sufficient matter did not exist. This paper was published before Jeffreys made his major contributions to probability theory, and he does not, ironically, make the Bayesian argument that we have outlined above. So we will make for Jeffreys the argument that he might have made had he returned to this question some years later.

Poor, in his Table I, gives the figure $a = +41.6'' \pm 1.4''$ for the centennial anomalous motion of Mercury. The uncertainty is certainly a probable error, so the standard error would be $\sigma = 2.0''$. In his Table II Poor gives $\alpha_E = 42.9''$ as the amount predicted by Einstein's theory, which is very close to the modern value.

Assuming normality of the measurement error, we can write for the conditional probability of the data a on the hypothesis that the true value is α, the density

$$P(a \mid \alpha) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-\alpha)^2}{2\sigma^2}\right).$$

Taking $\alpha = \alpha_E = 42.9''$ under Einstein's theory, E, yields for the conditional probability density of the data

$$P(a|E) = \frac{1}{\sqrt{2\pi}\sigma} \ \exp\left(-\frac{(a-42.9)^2}{2\sigma^2}\right).$$

Obtaining the conditional probability density of the data under a "fudged Newtonian" theory, F, is more complicated. If we could determine a *prior density* $P(\alpha|F)$ for the true value of α under F, then the conditional density of the data under F would be

$$P(a|F) = \int P(a|\alpha)P(\alpha|F)d\alpha. \tag{3}$$

But what should we choose for $P(\alpha|F)$?

Note, first, that $P(\alpha|F)$ must not depend in any way upon the data, but should reflect other information about α that is available through consideration of F. Since Newtonian theory was already well established at the time, and large deviations from Newtonian theory ($\alpha = 0$) are less believable than small ones, it would be natural to choose $P(\alpha|F)$ to be decreasing as α moves away from zero. Also, choosing $P(\alpha|F)$ to be symmetric is natural, at least for those theories in which, *a priori*, the effect on α could have been either positive or negative. A simple density that reflects these assumptions is the normal density with mean zero and standard deviation τ,

$$P(\alpha|F) = \frac{1}{\sqrt{2\pi}\tau} \ \exp\left\{-\frac{\alpha^2}{2\tau^2}\right\}.$$

It remains only to choose the standard deviation τ, which can be interpreted as the probable magnitude of the effect of α that would result under F. In the case at hand, the only data that can rule out moderate values of τ, say several times larger than $43''$ per century, would be observations of the inner planets, primarily of Venus. Unfortunately, the eccentricity of Venus' orbit is quite small, and so the perihelion motion of Venus, much smaller than that of Mercury to begin with, is difficult to observe accurately. For Earth and Mars the observations are much more accurate, but the effect is much smaller, comparable to the standard errors. However, the data on the inner planets can rule out a very large deviation from Newtonian theory, say $100''$ per century or so for Mercury. A value for τ of, say, $50''$ per century for Mercury's orbit seems reasonable and in agreement with the data on the orbits of the other inner planets. Since these are just rough calculations in any case, let us therefore take $\tau = 50''$.

Using equation (3), the conditional density of the data under F can now be computed to be

$$P(a|F) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \ \exp\left(-\frac{a^2}{2(\tau^2+\sigma^2)}\right).$$

Recalling that $a = 41.6$, $\sigma = 2.0$, and $\tau = 50$, we can then estimate the degree to which the data favor E over F by the Bayes factor (see *Box 1*)

$$B = \frac{P(a|E)}{P(a|F)} = \sqrt{1+\overline{\tau}^2} \ \exp\left(-\frac{D_E^2}{2}\right) \ \exp\left(\frac{D_F^2}{2(1+\overline{\tau}^2)}\right), \tag{4}$$

where $D_E = (a-\alpha_E)/\sigma = -0.65$, $D_F = a/\sigma = 20.8$ and $\overline{\tau} = \tau/\sigma = 25.0$. The Bayes factor is to be interpreted as the odds provided by the data for E compared with F; in particular, when B is greater than 1, the data favor E, and when B is less than 1 they favor F (Figure 4).

Plugging in the numbers, we obtain $B = 28.6$, which is moderately strong evidence in favor of the Einstein hypothesis. Indeed, as shown in Figure 5, the lowest that the odds ratio can be is $B = 27.76$, no matter what value of τ we adopt. Ironically, the data that Poor himself provides in his paper *against* general relativity favor the Einstein hypothesis over the "fudged Newtonian" hypothesis. If Poor had made a Bayesian calculation, he might not have come to the conclusion that he did!

Discussion of the relativity example.

Equation (4) contains several factors. The exponential factors reflect the "fit" of the data to the hypotheses. For example, D_E measures how far the data are from the Einstein value—in this case, less than one standard deviation, so that the corresponding exponential factor is near one. Similarly, the data are consistent with the "fudged Newtonian" hypothesis, resulting in the second exponential factor also being near one.

In this example, nearly all of B is due to the factor $\sqrt{1+\tau^2}$, which measures the ratio between the spread of the prior distribution for the "fudged Newtonian" hypothesis to that of the data. Since this spread is relatively large, it means that the "fudged Newtonian" hypothesis has to waste a considerable amount of prior probability on hypothetical values of α that are never observed. Gull (1988) calls this factor the "Ockham factor."

It is this Ockham factor that allows us to choose between the two hypotheses. The reason for the Ockham factor is simply that the "fudged Newtonian" hypothesis has an additional degree of freedom that allows it to accommodate a much larger range of hypothetical data than does the Einstein hypothesis. This means that it must spread its risk over a larger parameter space in order not to miss the region supported by the data. In this sense, it is a less simple theory than the Einstein hypothesis. The Einstein hypothesis makes a sharp prediction about Mercury's perihelion motion, which depends only on the known values of the constant of gravity and the speed of light. If we do not measure a value for the perihelion motion that is close to the predicted value, we can reject the Einstein theory. This is not possible for the "fudged Newtonian" hypothesis.

An objective Ockham's razor.

The above analysis required choice of $P(\alpha|F)$, the prior density of α under the fudged Newtonian hypothesis. It was argued that this density should be symmetric about $\alpha = 0$ and decreasing in $|\alpha|$; the particular choice of a normal $N(0, 50^2)$ density was then made. This last step was rather arbitrary and, in general, can be difficult. It is thus of considerable interest that a quantification of Ockham's razor can be given that is *independent* of the particular choice of $P(\alpha|F)$. Indeed, in Berger and Jefferys (1991), the following is established: *for any $P(\alpha|F)$ that is symmetric and decreasing in $|\alpha|$,*

$$B = \frac{P(a|E)}{P(a|F)} \geq \sqrt{\frac{2}{\pi}} \left[|D_F| + \sqrt{2\ln(|D_F|+1.2)}\right] \exp\left\{-\frac{D_E^2}{2}\right\}. \tag{5}$$

For the Mercury example, the right hand side of (5) is 15.04, so that the evidence in favor of Einstein is at least 15 to 1 for any reasonable $P(\alpha|F)$.

Note that equation (5) actually defines an objective Ockham's razor for the following generic situation: we observe data $a \sim N(\alpha, \sigma^2)$, σ^2 given, and wish to compare the "simple" model $H_0\colon \alpha = \alpha_E$ with the "complex" model, H_1, which has α freely varying with $\alpha = 0$ being the previous "standard." Then (5) provides an objective lower bound for the odds of H_0 to H_1 provided by the data. Of course, as with any lower bound, (5) is useless if the right hand side happens to be small; there is then no recourse but to attempt specification of $P(\alpha|H_1)$. Often, however, the bound will be large enough so that one can confidently choose H_0 without having to specify $P(\alpha|H_1)$.

Parsimonious model selection.

In choosing between two or more models, one of which is (essentially) true, the Bayesian ideas that we have discussed (see also *Box 1*) provide excellent mechanisms for selection. A somewhat different situation that is frequently encountered, however, is that of fitting an empirical model to data—a model that is not "true," but that will be used for prediction of the phenomenon under study. This can arise either when the true model is unknown or when the true model is too complex to be computationally useful.

Selecting among possible empirical models in this setting involves quite different considerations. Accuracy of future predictions is, of course, a major concern, but simplicity of the model for computational or interpretational reasons is also highly relevant. This latter factor can lead to a "parsimonious Ockham's razor," which chooses the simpler model for such practical reasons, not because it is "true." Although Bayesian analysis can also be crucial in understanding model selection in this scenario, the relevant Bayesian techniques are considerably different and discussion would take us too far afield.

Conclusions.

Ockham's razor, far from being merely an *ad hoc* principle, can under many practical situations in science be justified as a consequence of Bayesian inference. Bayesian analysis can shed new light on what the notion of the "simplest" hypothesis consistent with the data actually means. We have seen three different ways in which Ockham's razor can be interpreted in Bayesian terms: in the choice of the prior probabilities of hypotheses, using scientific experience to judge that simpler hypotheses are more likely to be correct; as a consequence of the fact that a hypothesis with fewer adjustable parameters will automatically have an enhanced posterior probability, due to the fact that the predictions it makes are sharp; and in the choice of parsimonious empirical models. All these are in agreement with our intuitive notion of what makes a theory powerful and believable.

Acknowledgements.

The authors wish to thank their colleagues who have commented on drafts of this paper: Raynor L. Duncombe, David M. Hillis, and Myra Samuels. This research was supported by the National Aeronautics and Space Administration, under NASA Contract NAS5–29285, and by the National Science Foundation, Grant DMS–8923071

References

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer–Verlag, New York.

Berger, James O. and Berry, D. (1988). "Statistical analysis and the illusion of objectivity." *American Scientist* **76**, 159–165.

Berger, James O. and Jefferys, William H. (1991). "Minimal Bayesian testing of precise hypotheses, model selection, and Ockham's razor." Technical Report, Purdue University.

Brush, S. (1989). "Prediction and theory evaluation: The case of light bending." *Science* **246**, *1124-1129.* See also the responses to this article in *Science,* **248**, *422-423.*

Edwards, W., Lindman, H. and Savage, L.J. (1963). "Bayesian statistical inference for psychological research." *Psychological Review* **70**, 193–242.

Good, I.J. (1968). "Corroboration, explanation, evolving probability, simplicity, and a sharpened razor." *British J. Philosophy of Science* **19**, 123–143.

Good, I.J. (1977). "Explicativity: a mathematical theory of explanation with statistical applications." *Proceedings of the Royal Society A* **354**, 303–330.

Gull, S. (1988). "Bayesian inductive inference and maximum entropy. In G.J. Erickson and C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol 1), 53-74.* Dordrecht: Kluwer Academic Publishers.

Harpp, David (1991). Quoted in "Big Prof is Watching You," *Discover,* April 1991 issue, *pp. 12-13.*

Jaynes, E.T. (1979). "Inference, method, and decision: Towards a Bayesian philosophy of science." *Journal of the American Statistical Association* **74**, *740-41.*

Jeffreys, H. (1921). "Secular perturbations of the inner planets." *Science* **54**, *248*.

Jeffreys, H. (1939). *Theory of Probability, Third Edition*. Oxford: Clarendon Press.

Laplace, P.S. (1812). *Théorie Analytique des Probabilities*. Courcier, Paris.

Lindley, D.V. (1957). "A statistical paradox." *Biometrika* **44**, *187-192*.

Loredo, T.J. (1990). "From Laplace to Supernova 1987A: Bayesian inference in astrophysics." In P. Fougere (ed.), *Maximum Entropy and Bayesian Methods, 81-142*. Dordrecht: Kluwer Academic Publishers.

MacKay, David J.C. (1991). "Bayesian Interpolation." Submitted to*Neural Computation*.

Max, Edward E. (1986). "Plagiarized errors and molecular genetics: Another argument in the evolution-creation controversy." *Creation/Evolution* **XIX** *34-46*.

Newcomb, Simon (1895). *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. Washington: Government Printing Office. *pp. 109-122*.

Nobili, A.M. & Will, C.M. (1986). "The real value of Mercury's perihelion advance." *Nature* **320**, *39-41*.

Poor, C.L. (1921). "The motions of the planets and the relativity theory." *Science* **54**, *30-34*.

Roseveare, N.T (1982). *Mercury's Perihelion from Le Verrier to Einstein*. Oxford: Clarendon Press.

Smith, A.F.M. and Spiegelhalter, D.J. (1980). "Bayes factors and choice criteria for linear models." *J. Royal Statist. Soc. B* **42**, *213-220*.

Thorburn, W.M. (1918). "The myth of Occam's razor." *Mind* **27**, 345–353.

Watson, James D., Hopkins, Nancy H., Roberts, Jeffrey W., Steitz, Joan A. and Weiner, Alan M. (1988). *Molecular Biology of the Gene, 4th Edition*. Menlo Park: Benjamin/Cummings Publishing Company. *pp. 649-663*.

Wrinch, D. and Jeffreys, H. (1921). "On certain fundamental principles of scientific inquiry." *Phil. Mag.* **42**, *369-390*.

Box 1: Bayes' Theorem

The foundation of this paper is Bayes' theorem, proved by the Rev. Thomas Bayes (1763). At its core, Bayes' theorem represents a way—Bayesians would argue the most consistent way—of incorporating new data into our understanding of the world. Let $H_1, H_2, \ldots, H_n$ be n mutually exclusive and exhaustive hypotheses, and let $P(H_i \mid I)$ represent our personal probability that the hypothesis H_i is true, given all the relevant prior information I that is available to us. Let D represent some new piece of data that comes to our attention. Then Bayes' theorem tells us that we should update our personal probabilities according to the rule

$$P(H_i \mid D\&I) = \frac{P(D \mid H_i\&I)P(H_i \mid I)}{P(D \mid I)},$$

where $P(D \mid H_i\&I)$ is the probability that we would observe D, given that H_i is true and assuming the information I, and $P(H_i \mid D\&I)$ is our updated personal probability that H_i is true, given both the old information I and the new information D. The denominator is the total probability of observing the data, summed over all hypotheses:

$$P(D \mid I) = \sum_j P(D \mid H_j\&I)P(H_j \mid I)$$

When considering only two hypotheses, it is often more convenient to work with the *odds ratio* Ω, which is just the ratio of the two probabilities. Thus, given H_0 and H_1, we find that the *odds* $\Omega_{0/1}(D\&I)$ in favor of H_0 as against H_1 after considering both D and I are given by

$$\begin{aligned}\Omega_{0/1}(D\&I) &= \frac{P(H_0 \mid D\&I)}{P(H_1 \mid D\&I)} \\ &= \frac{P(D \mid H_0\&I)}{P(D \mid H_1\&I)}\frac{P(H_0 \mid I)}{P(H_1 \mid I)} \\ &= B_{0/1}(D\&I)\Omega_{0/1}(I)\end{aligned}$$

In this formula, $B_{0/1}$ is the *Bayes factor* of H_0 on H_1. Note that it does *not* depend on the prior probabilities of the hypotheses, which are involved only in the *prior odds* $\Omega_{0/1}(I)$. Thus the Bayes factor measures the amount by which the new data favors H_0 against H_1. This separation of the effect of the data from the prior probabilities is important for scientific communication.

Box 2: Controversies

Bayes' theorem has had a long and controversial history, which is too extensive to review here. See Berger (1985) or Edwards, Lindman, and Savage (1963) for general discussion of these controversies. After many years of eclipse, Bayesian methods have undergone a revival, particularly after the publication of the influential book, *Theory of Probability* (Jeffreys, 1939) .

There are two main points of contention between Bayesians and traditional (frequentist) statisticians. The first is philosophical: Some argue that since only one of the hypotheses H_i is true, it makes no sense to talk about the "probability" that H_j is true. This has a certain logic if one interprets probabilities as frequencies; but Bayesians also use the term "probability" to refer to the *degree of plausibility* of a hypothesis, and indeed, this is the way most working scientists use the term. Bayes' theorem thus becomes a way of calculating how the observation of new data increases the plausibility of some hypotheses at the expense of others. The Bayesian use of the word "probability of a hypothesis" is in fact the natural one for working scientists.

The second point is that there are no universally accepted ways of assigning the prior probabilities ("priors") $P(H_i \mid I)$ that Bayes' theorem requires. This means that different scientists, faced with the same data, may come to different conclusions. The traditional approach to statistics, such as the use of significance tests of hypotheses, sought to overcome this perceived problem of subjectivity.

Bayesians have several responses to this. One school (see Edwards, Lindman, and Savage, 1963, for an introduction) believes that there is nothing inherently wrong with subjectivism, and indeed, that the frequentist approach is really no more objective, although it has successfully disguised this fact (Berger and Berry, 1988). Subjectivist Bayesians point out that it is quite common for scientists to disagree about the plausibility of hypotheses, and contend that this is a natural, and indeed inescapable state of affairs. Furthermore, frequentists can give the appearance of objectivity only by answering the "wrong" questions—for instance, providing the probability of a "rejection region" given the hypothesis is true, when a scientist really wants to know the probability that the hypothesis is true given the data.

Another school (Laplace, 1812, Jeffreys, 1939) has developed methods of choosing and utilizing "objective" prior distributions for a wide class of problems. With problems for which such are available, Bayesian analysis can claim to be as objective as any statistical methods. Note, however, that for the model selection problem we have discussed, objective prior distributions cannot be utilized; thus it was necessary to specify $P(\alpha|F)$ in the Mercury perihilion problem to determine the Bayes factor, although an objective lower bound was available. See Berger and Jefferys (1991) for further discussion.

Figure Captions

Figure 1: The dots are data relating time to the distance a falling object travels. The curves are the best quadratic fit to the data (solid line), and an exact 7-th degree polynomial fit to the data (dashed line). Though the quadratic fit is not exact, it is more appealing because of its simplicity.

Figure 2: The probability of flipping zero or one head for a two-headed coin (hatched) and a fair coin (white). If tails is flipped, the hypothesis that the coin is two-headed is falsified; however, if heads is flipped, the hypothesis that the coin is two-headed is favored by odds of 2:1.

Figure 3: The probability of flipping n heads in five tries, as a function of the number of heads that have been tossed, for a two-headed coin (hatched) and a fair coin (white). Note how heavily favored is the two-headed hypothesis when we actually observe only heads.

Figure 4: The odds ratio, B, of the Einstein to the "Fudged Newton" hypotheses, as a function of τ, the "guess" as to the standard deviation of θ under the Fudged Newton hypothesis. Note that the lowest the odds can be is 27.76.

Figure 5: The probability densities of the data, a, under the Fudged Newton and Einstein hypotheses. For the actual data $a = 41.6$, the odds in favor of Einstein's model are $B = d_2/d_1 = 28.6/1$. Because the Einstein hypothesis makes a sharp prediction, it is tall and narrow in the region near its prediction, which causes it to be favored when the observed data are near the predicted value. (The scales are somewhat altered for clearer presentation.)