British Journal for the Philosophy of Science **45** (1994). 1–35

How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions

MALCOLM FORSTER AND ELLIOTT SOBER*

ABSTRACT

Traditional analyses of the curve fitting problem maintain that the data do not indicate what form the fitted curve should take. Rather, this issue is said to be settled by prior probabilities, by simplicity, or by a background theory. In this paper, we describe a result due to Akaike [1973], which shows how the data can underwrite an inference concerning the curve's form based on an estimate of how predictively accurate it will be. We argue that this approach throws light on the theoretical virtues of parsimoniousness, unification, and non *ad hocness*, on the dispute about Bayesianism, and on empiricism and scientific realism.

I INTRODUCTION

Curve fitting is a two-step process. First one selects a family of curves (or the form that the fitted curve must take). Then one finds the curve in that family

* Both of us gratefully acknowledge support from the Graduate School at the University of Wisconsin-Madison, and NSF grant DIR-8822278 (M. F.) and NSF grant SBE-9212294 (E. S.). Special thanks go to A. W. F. Edwards, William Harper, Martin Leckey, Brian Skyrms, and especially Peter Turney for helpful comments on an earlier draft.

(or the curve of the required form) that most accurately fits the data. These two steps are universally supposed to answer to different standards. The second step requires some measure of goodness-of-fit. The first is the context in which simplicity is said to play a role. Intrinsic to this two-step picture is the idea that these different standards can come into conflict. Maximizing simplicity usually requires sacrifice in goodness-of-fit. And perfect goodness-of-fit can usually be achieved only by selecting a complex curve.

This view of the curve fitting problem engenders two puzzles. The first concerns the nature and justification of simplicity. What makes one curve simpler than another and why should the simplicity of a curve have any relevance to our opinions about which curves are true? The second concerns the relation of simplicity and goodness-of-fit. When these two *desiderata* conflict, how is a trade-off to be effected? A host of serious and inventive philosophical proposals notwithstanding, both these questions remain unanswered.

If it could be shown that a single criterion for selecting a curve gives due weight to both simplicity and goodness-of-fit, then the two problems mentioned above for traditional analyses of the curve fitting problem would fall into place. It would become clear why simplicity matters (and how it should be measured). In addition, simplicity and goodness-of-fit would be rendered commensurable by representing each in a common currency. In what follows we describe a result in statistics, stemming from the work of Akaike [1973], [1974], which provides this sort of unified treatment of the problem, in which simplicity and goodness-of-fit are both shown to contribute to a curve's expected accuracy in making predictions.[1]

2 AKAIKE WITHOUT TEARS

In this section, we present the basic concepts that are needed to formulate the curve-fitting problem and to solve it. To begin with, we need to describe the kinds of *hypotheses* that curves represent and the relationship of those curves to the *data* we have available. A 'deterministic' curve is a line in the X/Y plane; it represents a function, which maps values of X (the independent variable) onto unique values of Y (the dependent variable).[2] For example, Figure 1 depicts two such curves; each says that Y is a *linear* function of X. Each of these curves may

[1] There is a growing technical literature on the subject. Linhart & Zucchini [1986] surveys the earlier work of statisticians. Researchers in computer science have used the concept of 'shortest data descriptions' to warrant the trade-off between simplicity and goodness of fit. See Rissanen [1978], [1989], or more recently, Wallace and Freeman [1992]. While there are criteria in the literature that are quantitatively different from Akaike's, there is a measure of agreement in the way they define simplicity and goodness-of-fit. We have focused on Akaike's seminal work because he motivates his criterion in a *general* and *perspicuous* manner.

[2] The idea that there is just one independent variable is a simplifying assumption adopted for ease of exposition. The results we will describe generalize to any number of independent variables.

be obtained by fixing the values of the parameters α_0 and α_1 in the following equation:

$$Y = \alpha_0 + \alpha_1 X .$$

The two curves in Figure 1 are equally simple, we might say, because each is a straight line and each is obtained from a functional form in which there are just two adjustable parameters. These two curves belong to a *family* of curves—namely, the set of all straight lines. We will be talking about both *specific* curves and *families* of curves in what follows, so it will be important to keep the distinction between them in mind. In fact, it will turn out that there is no need to define the simplicity of a specific curve; all that is needed is the notion of the simplicity of a *family* of curves, and this Akaike's approach provides.

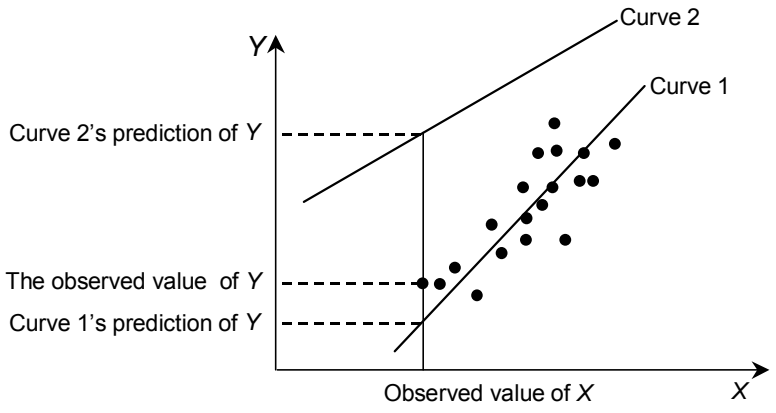

FIGURE 1

Suppose the true specific curve *determined* the outcomes of the observations we make. Then, if Curve 1 were true, the set of data points we obtain would have to fall on a straight line (i.e., on the straight line depicted by Curve 1 itself). But we will suppose that the observation process involves *error*. Even if Curve 1 were true, it is nonetheless quite possible that the data we obtain will not fall exactly on that curve. It may be impossible to say when any particular data point will fall above or below the true curve - only that it should 'tend' to be close. To represent this possibility of error, we associate a probability distribution with each curve. This distribution tells us how probable it is that the *Y*-value we observe for a given *X*-value will be 'close' to the curve. The most probable outcome is to obtain a *Y*-value that falls exactly on the true curve. Locations that are further off the curve have lower probabilities (symmetrically above and below) of being what we observe.

To make this idea concrete, suppose that we are interested in plotting the location of a planet as it moves across the sky. In this case, the *X*-axis represents time and the *Y*-axis represents location. The true curve is the actual, unique trajectory of the planet. But our observation of the planet's motion is subject to error. Even if Curve 1 in Figure 1 describes the planet's true trajectory, it

nonetheless is possible that we should obtain data that fail to fall exactly on that curve.

So there are two factors that influence the observations we make. There is the planet's actual trajectory; and there is the process of observation, which is subject to error. If the planet's trajectory is a straight line, we can combine these two influences into a single expression:

(LIN) $$Y = \alpha_0 + \alpha_1 X + \sigma U.$$

The last addend represents the influence of error. Here, of course, Y doesn't represent the planet's *actual* location, but represents its *apparent* location.[3]

Now consider the data points depicted in Figure 1. If Curve 1 were true, it is possible that we should obtain the data before us. But the same is true of Curve 2; if it were true, it also could have generated the data at hand. Although this is a similarity between the two curves, there nonetheless is a difference: the probability of obtaining the data, if Curve 1 is true, exceeds the probability of obtaining the data, if Curve 2 were true: p(Data/Curve 1) > p(Data/Curve 2).[4] Statisticians use the technical term *likelihood* to describe this difference; they would say that Curve 1 is more likely than Curve 2, given the data displayed. It is important to note that the likelihood of a hypothesis is not the same thing as its probability; don't confuse p(Data/Curve 1) with p(Curve 1/Data).

In a sense, Curve 1 fits the data better than Curve 2 does. The standard way to measure this goodness-of-fit is by a curve's *sum of squares* (SOS). As depicted in Figure 1, we compute the difference between the Y-value of a data point and the Y-value on the curve directly above or below it. We square this difference and then sum the same squared differences for each data point. Curve 1 has a lower SOS value than Curve 2, relative to the data in Figure 1. Comparing SOS values is a way to compare likelihoods. Notice that if we were to increase the number of data points, the SOS values for both curves would almost certainly go up.[5]

We can use the concept of SOS to reformulate the curve-fitting problem. Given a set of data, how are we to decide which curve is most plausible? If minimizing the SOS value were our sole criterion, we would almost always prefer bumpier curves over smoother ones. Even though Curve 1 is rather close to the data depicted in Figure 1, we could draw a more complex curve that

[3] Alternatively, the error term can be given a physical, instead of an epistemological, interpretation, if one wishes to represent the idea that nature itself is stochastic. In that case, Y would represent the planet's '*mean*' position. This difference in interpretation will not affect our subsequent discussion.

[4] When random variables are continuous, the likelihood is defined in terms of probability *densities* rather than probabilities. A lower case p is a probability density, while the upper case P is reserved for probabilities.

[5] The SOS value for a curve can't go down as the data set is enlarged; it would stay the same, if, improbably enough, the new data points fell exactly on the curve. Also note that a curve's likelihood will decline as the data set is enlarged, even if the new points fall exactly on the curve.

passes exactly through those data points. The practice of science is to not do this. Even though a hypothesis with more adjustable parameters would fit the data better, scientists seem to be willing to sacrifice goodness-of-fit if there is a compensating gain in simplicity. The problem is to understand the rationale behind this behavior. Aesthetics to one side, the fundamental issue is to understand what simplicity has to do with truth.

The universal reaction to this problem among philosophers has been to think that the only thing the data tell you about the problem at hand is given by the SOS values. The universal refrain is that 'if we proceed just on the basis of the data, we will choose a curve that passes exactly through the data points.' This interpretation means that giving weight to simplicity involves an *extra*empirical consideration. We thereby permit considerations to influence us *other* than the data at hand. Giving weight to simplicity thus seems to embody a kind of *rationalism*; a consistent empiricist must always opt for bumpy curves over smooth ones.

The elementary framework developed so far allows us to show that this standard reaction is misguided. Let us suppose that the curve in Figure 2 is true. Now consider the data that this true curve will generate. Since we assume that observation is subject to error, it is overwhelmingly probable that the data we obtain will not fall exactly on that true curve. An example of such a data set, obtained from the true curve, also is depicted in Figure 2. Now suppose we draw a curve that passes exactly through those data points. Since the data points do not fall exactly on the true curve, such a best-fitting curve will be *false*. If we think of the true curve as the 'signal' and the deviation from the true curve generated by errors of observation as 'noise,' then fitting the data perfectly involves confusing the noise with the signal. It is overwhelmingly probable that any curve that fits the data perfectly is false.

Of course, this negative remark does not provide a recipe for disentangling signal from noise. We know that any curve with perfect fit is probably false, but this does not tell us which curve we should regard as true. What we would like is a method for separating the 'trends' in the data from the random deviations from those trends generated by error. A solution to the curve fitting problem will provide a method of this sort.

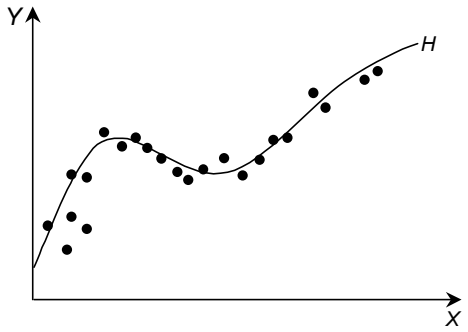

FIGURE 2

To explain Akaike's proposal, we need to introduce a precise definition of how successful a curve is in identifying the trend behind the data. In addition to talking about a curve's distance from a particular data set, we need a way to measure a curve's distance from the true curve. A constraint on this new concept is already before us: a curve that is maximally close to the data (because it passes exactly through all the data points) is probably not going to be maximally close to the truth. Closeness to the truth is different from closeness to the data. How should the concept of closeness to the truth be defined?

Let us suppose that Curve 1 in Figure 1 is true. We want a way to measure how close Curve 2 is to this true curve. Curve 1 has generated the data set displayed in the figure, and we can use the SOS measure to describe how close Curve 2 is to these data points. The idea is to define the distance of Curve 2 from Curve 1 in terms of the *average* distance of Curve 2 from the data *generated by* Curve 1. So, imagine that Curve 1 generates new data sets, and each time we measure the distance of Curve 2 from the generated data set. We repeat this procedure indefinitely, and we compute the *average* distance that Curve 2 has with respect to data sets generated by the true Curve 1. Remember that this average is computed over the space of *possible* data sets, rather than actual data sets.[6] This allows us to define distance from the truth as follows:

Distance from the true curve (T) of curve C $=^{\text{df}}$
Average[SOS of *C*, relative to data set *D* generated by *T*] −
Average[SOS of *T*, relative to data set *D* generated by *T*].

First, note that the distance from the true curve is relative to the process of data generation; it depends on the method of generating the array of *X*-values whose associated *Y*-values the curves are asked to predict.[7] Second, note that the true curve, *T*, is the curve that is closest to the truth (its distance from the truth is 0) according to this definition. However, the average SOS value of the true curve *T*, relative to the data sets that *T* generates, is *non*zero. This is because of the role of error; on average, even the true curve won't fit the data perfectly.

We now define the concept of distance from the truth for *families* of curves. The above definition defines what it means for Curve 2 to be a certain distance from the true curve. But what would it mean to describe how close to the true curve the family of straight lines (LIN) is? Here's the idea: Let's think of *two* data sets, D_1 and D_2, each generated by the true curve *T*. First, we find the specific curve within the family that fits D_1 best. Then we compute the SOS of that curve relative to the second data set D_2. Imagine carrying out this procedure

[6] Statisticians mark this distinction by using the term '*expected* value' rather than '*average* value.' We have chosen not to do this because the psychological connotations of the word 'expected' may mislead some readers.

[7] The *X*-arrays for the predicted data do not have to be the same as the *X*-array for the actual data, but both must be generated by the *same* stochastic process.

again and again for different pairs of data sets. The average SOS obtained in this way is the family's distance from the truth:

Distance from the true curve (T) of family F $=^{\text{df}}$
Average[SOS of $L_1(F)$, relative to data set D_2 generated by T] −
Average[SOS of T, relative to data set D_2 generated by T].

Here $L_1(F)$ is the best fitting ('likeliest') member of the family F, relative to data set D_1.[8]

Our definition of a family's distance from the truth is intended to measure how accurate the predictions will be that the best fitting curve in a family generates. Consider the family of straight lines (LIN) and the data displayed in Figure 1. How close is the family (LIN) to the truth? We can imagine finding the straight line that best fits the data at hand. The question we'd like to answer is how accurately *that particular straight line* will predict new data. The average distance from the truth of best fitting curves selected from that family is the distance of the family from the truth:

Distance from the true curve (T) of family F =
Average[Distance of best fitting curves in F from the truth T].

Our interest in the distance of families from the truth stems from this equality. Families are of interest because they are *instruments of prediction*; they make predictions by providing us with a specific curve–viz. the curve in the family that best fits the data.[9]

If the true curve is in fact a straight line, (LIN) will of course be very close to the truth (though the distance will be nonzero).[10] But if the truth is highly nonlinear, (LIN) will perform poorly as a device for predicting new data from old data. Let us move to a more complicated family of curves and ask the same questions. Consider (PAR), the family of parabolic equations:

(PAR) $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \sigma U.$$

Specific parabolas will be ∪-shaped or ∩-shaped curves. Notice that (LIN) is a subset of (PAR). If the true specific curve is in (LIN), it also will be in (PAR). However, the converse relation does not hold.

So if (LIN) is true, so is (PAR) (but not conversely). This may lead one to

[8] The definition of distance from the truth of a specific curve C is a special case of the definition for a family of curves F. A family is a set of curves; when a family contains just one curve, its best fitting member is just that curve itself.

[9] In the kinds of example we consider, there will be a unique curve in a family that fits the data best when the number of data points exceeds the number of adjustable parameters.

[10] A family can be literally true (by including the true curve) and still have a non-zero distance from the truth because other curves in the family (including $L(F)$) will be closer than the true curve to the *actual* data.

expect that PAR must be at least as close to the truth as (LIN) is. However, *this is not so*! Let's suppose that the true curve is, in fact, a straight line. This will generate sets of data points that mostly fail to fall on a straight line. Fitting a straight line to one set of data points will provide more accurate predictions about new data than will fitting a parabolic curve to that set. To be sure, for each data set, the best fitting parabola will be *closer to the data* than the best fitting straight line. But this leaves open how well these two curves will predict *new* data. (LIN) will be closer to the truth (in the sense defined) than (PAR) is, if the truth is a straight line.

Curves that fit a given data set perfectly will usually be false; they will perform poorly when they are asked to make predictions about *new* data sets. Perfectly fitting curves are said to '*over*fit' the data. This fact about specific curves is reflected in our definition of what it means for a family to be close to the truth. If (LIN) is closer to the truth than (PAR) is, then a straight line hypothesis fitted to one data set will do a better job of predicting new data than a parabolic curve fitted to the same data, at least on average. In this case, the more complex family is disadvantaged by the greater tendency of its best fitting case, L(PAR), to overfit the data.

The definitions just given of closeness to the truth do not show how that quantity is epistemologically accessible. To apply these definitions and compute how close to the truth a curve C (or a family F) is, one must know what the truth (T) is. Nonetheless we can use the concept of closeness to the truth to reformulate the curve-fitting problem and to provide it with a solution.

All families with at least one free parameter are able to reduce their least SOS by fitting to *random fluctuations* in the data. This is true of low dimensional families as well, though to a lesser degree. For example, the data in Figure 1 were generated by a straight line, but random fluctuations in the data enable a parabola to fit it better than *any* straight line. This shows that the phenomenon of overfitting is ubiquitous.[11] Thus, there are two reasons why the least SOS goes down as we move from lower to higher dimensional families: (a) Larger families generally contain curves closer to the truth than smaller families. (b) *Overfitting*: The higher the number of adjustable parameters, the more prone the family is to fit to noise in the data. Our promised reformulation of the curve fitting problem is this: We want to favour larger families if the least SOS goes down because of factor (a), but not if its decline is largely due to (b). If only we could correct the SOS value for overfitting, then the *corrected* SOS value would be an unbiased indication of what we are interested in—viz. the distance from the true curve.

[11] This is the same overfitting problem that plagues general purpose learning devices like neural networks. Moody [1992] and Murata *et al*. [1992] are working on generalizing the Akaike framework to apply to artificial neural networks. See Forster [1992b] for further details. It is interesting that there is such a fundamental connection between neural learning and the philosophy of science (Churchland [1989]).

At this point, we will simply state *Akaike's theorem*, without attempting to work through the mathematical argument that establishes its correctness. (See the Appendix A for a non-technical explanation of the assumptions needed, and Appendix B for the proof of the theorem in a special case. The most thorough, and accessible, technical treatment is found in Sakamoto *et al*.[1986].) Akaike [1973] discovered a way of estimating the size of the overfitting factor. The procedure is fallible, of course, but it has the mathematical property of providing an *unbiased* estimate[12] of the comparative distances of different families from the truth under favourable conditions (see Appendix A). The amazing thing about Akaike's result is that it renders closeness to the truth epistemologically accessible; the estimate turns on facts that we can readily ascertain from the family itself and from the single data set we have before us:

$$\text{Estimated[(Distance from the truth of family } F) = \text{SOS}[L(F)] + 2k\,\sigma^2 + \text{Constant.}$$

$L(F)$ is the member of the family that fits the data best, k is the number of adjustable parameters that the family contains, and σ^2 is the variance (degree of spread) of the distribution of errors around the true curve. The last term on the right hand side is common to all families, and so it drops out in comparative judgments.

The first term on the right hand side, SOS[$L(F)$], is what we have been calling the *least* SOS for the family. It represents what empiricists have traditionally taken to exhaust the testimony of evidence. The second term corrects for the average degree of overfitting for the family. Since overfitting has the effect of reducing the SOS, any correction should be positive. That this correction is proportional to k, the number of adjustable parameters,[13] reflects the intuition that overfitting will increase as we include more curves that are able to mould themselves to noise in the data. That the expected degree of overfitting also is proportional to σ^2 is plausible as well - the bigger the error deviations from the true curve, the greater the potential for misleading fluctuations in the data. Also note that if there is no error ($\sigma^2 = 0$), then the estimate for the distance from the truth reduces to the least SOS value. The postulation of error is essential if

[12] 'Unbiased' means that its *average* performance will center on the true value of the quantity being estimated. Note that an unbiased estimator can have a wide or narrow *variance*, which measures how much the estimate 'bounces around' on average. Unbiasedness is only one desideratum for 'good' estimators.

[13] In our running example, (LIN) contains two adjustable parameters and (PAR) contains three. The number of adjustable parameters is not a merely linguistic feature of the way a family is represented. For example, $Y = \alpha + \beta X + \gamma X$ is one way of representing (LIN), but k is still 2, because there is a *reparameterization* (*viz*. $\alpha' = \alpha$, $\beta' = (\beta + \gamma)$, and $\gamma' = (\beta - \gamma)$) such that $Y = \alpha' + \beta' X$. In contrast, the dimension of the family $Y = \alpha + \beta X + \gamma Z$ is 3 because there is no such reparameterization.

simplicity (as measured by k) is to be relevant to our estimates concerning what is true.[14]

We will use the term 'predictive accuracy' to describe how close to the truth a curve or family is. 'Accuracy' is a synonym for 'closeness to the truth', while the term 'predictive' serves to remind the reader that the concept is relativized to the process by which the true curve generates new data. Instead of using SOS as a measure of distance, we use the log of the likelihood to measure closeness to the data (the greater the log-likelihood, the smaller the distance from the data). Thus, we define the predictive accuracy of a curve C, denoted by A(curve C), as the average log-likelihood of C per datum. The predictive accuracy of a family F is the average predictive accuracy of its best fitting curves.[15] This leads to a more general statement of Akaike's Theorem, since the log-likelihood applies to cases, like coin tossing examples, in which the SOS value is not defined. Recalling the connection between the low SOS value of a specific curve and its high likelihood, the general statement of Akaike's theorem is as follows:

Akaike's Theorem: Estimated[A(family F)] = $(1/N)$ [log-likelihood($L(F)$) $- k$],

where N is the number of data points.[16] We no longer need to assume that the error variance, σ^2, is known, for the error variance may be treated as another adjustable parameter.[17]

[14] We regard the total absence of error as radically implausible. Even if nature were completely deterministic, there still would be *observational* errors. And even then, there still would be lawless deviations from *any* 'curve' that limits itself to an impoverished stock of independent variables. For example, it may be that the temperature at a particular place and time is determined. A curve that *truly* captures the dependence of temperature on the time of day and time of year will not predict the temperature *exactly* because there are other relevant factors. The data will *behave* as randomly *as if* the world were indeterministic. From an *epistemological* point of view, this is all that matters. Forster [1988b] and Harper [1989] examine the role of this third kind of error (arising from the action of other variables) in the 'exact' science of astronomy.

[15] This average is computed as follows: Take a data set D_1 generated by the true curve T, and note the *predictive accuracy* of the best curve $L_1(F)$ in F relative to D_1. Imagine that this procedure is repeated with new data sets D_2, D_3, ..., each time noting the predictive values of the curves $L_2(F)$, $L_3(F)$, ... Now take the average of all these values.

[16] The factor $(1/N)$ arises from the fact that we prefer to define accuracy as the average *per datum* log-likelihood, so that the accuracy of a hypothesis does not change when we consider the prediction of data sets of different sizes.

[17] When σ^2 is treated as unknown, a curve (*by itself*) no longer confers a *probability* on the data. Literally speaking, a curve is a *family* of probability distributions—one for each numerical value of σ^2. From now on we will understand a 'curve' to be associated with some specific numerical value of σ^2. Also note that Akaike's estimate of predictive accuracy of a family of 'curves' in which σ^2 is a free parameter is related to the least SOS value for the family by a different formula (Sakamoto *et al.* [1986], p.170):

$$\text{Estimate}[A(\text{Family F}')] = -(1/2)\log[\text{SOS}(B(F))/N] - k'/N + \text{constant},$$

where F' is the higher dimensional family obtained from F by making σ^2 adjustable. Here, SOS($B(F)$) is the least SOS for the original family F, and k' is the dimension of the final family. For LIN and PAR, $k' = k + 1$.

This theorem, we believe, provides a solution to the curve-fitting problem. It explains why fitting the data at hand is *not* the only consideration that should affect our judgment about what is true. The quantity k is also relevant; it represents the bearing of simplicity. A family F with a large number of adjustable parameters will have a best member $L(F)$ whose likelihood is high; however, such a family will also have a high value for k. Symmetrically, a simpler family will have a lower likelihood associated with its best case, but will have a low value for k. Akaike's theorem shows the relevance of goodness-of-fit *and* simplicity to our estimate of what is true. But of equal importance, it states a precise rate-of-exchange between these two conflicting considerations; it shows how the one quantity should be traded off against the other. We emphasize that Akaike's theorem solves the curve-fitting problem without attributing simplicity to specific curves; the quantity k, in the first instance, is a property of families.[18]

A special case of Akaike's result is worth considering. Suppose one has a set of data that falls fairly evenly around a straight line. In this case the best fitting straight line will be very close to the best fitting parabola. So L(LIN) and L(PAR) will have almost the same SOS values. In this circumstance, Akaike's theorem says that the family with the smaller number of adjustable parameters is the one we should estimate to be closer to the truth. A simpler family is preferable if it fits the data about as well as a more complex family. Akaike's theorem describes how much of an improvement in goodness-of-fit the move to a more complicated family must provide for it to make sense to prefer the more complex family. A slight improvement in goodness-of-fit will not be enough to justify the move to a more complex family. The improvement must be large enough to overcome the penalty for complexity (represented by k).

Another feature of Akaike's theorem is that the relative weight we give to simplicity declines as the number of data points increases. Suppose that there is a slight parabolic bend in the data, reflected in the fact that the SOS value of L(PAR) is slightly lower than the SOS value of L(LIN). Recall that the absolute value of these quantities depends on the number of data points. With a large amount of data our estimate of how close a family is to the truth will be determined largely by goodness-of-fit and only slightly by simplicity. But with smaller amounts of data, simplicity plays a more determining role. Only when a nonlinear trend in the data is 'statistically significant' should that regularity be taken seriously. This is an intuitively plausible idea that Akaike's result explains.

3 UNIFICATION AS A SCIENTIFIC GOAL

It is not at all standard to think that the curve fitting problem is related intimately

[18] Thus, the problems of defining the simplicity of curves described by Priest [1976] do not undermine Akaike's proposal.

to the problem of explaining why unified theories are preferable to disunified ones. The former problem usually is associated with 'inductive' inference, the latter with 'inference to the best explanation.' We are inclined to doubt that there really are such fundamentally different kinds of nondeductive inference (Forster [1986], [1988a], [1988b]; Sober [1988b], [1990a], [1990b]).[19] In any case, Akaike's approach to curve fitting provides a ready characterization of the circumstances in which a unified model is preferable to two disunified models that cover the same domain.[20]

It is always a substantive scientific question whether two data sets should be encompassed by a single theory or different theories should be constructed for each. Should celestial and terrestrial motion be given a unified treatment or do the two sets of phenomena obey different laws? In retrospect, it may seem obvious that these two kinds of motion should receive the same theoretical treatment. But this is the wisdom of hindsight; individual phenomena do not have written on their sleeves the other phenomena with which they should be coalesced.

Traditional approaches to this problem make the allure of unification something of a mystery.[21] Given two data sets D_1 and D_1, a unified model M_u

[19] William Whewell [1840] described the process of curve fitting as a special case of a process of conceptualization called the 'colligation of facts' (Forster [1988b]). He then referred to the process that leads to the unification of disparate curve fitting solutions as the 'consilience of inductions.' On our view, both of these processes are seen as aspects of a single kind of inferential procedure. Bogen and Woodward [1988] argue that the inferential relationship of observation to theory has two parts: of observation to phenomena and of phenomena to theory. Again, it is not clear to us that these relationships are fundamentally different in kind.

[20] We will follow statistical practice and reserve the term 'model' for a *family* of hypotheses, in which each hypothesis includes a specific statement about the distribution of errors (so that likelihoods are well defined). A model leaves the values of some parameters unspecified. In applying the term to astronomy, we need only assume that some assumption about the *form* of the error distribution is included (e. g. that the distribution is Gaussian, as was assumed in Gauss's own application of the method of least squares to astronomy—see Porter [1986]). The variance of the distribution may be left as an adjustable parameter. The important point to notice is that distinguishing models from curves, or from abstract 'theories', is now critical to the philosophy of science, since Akaike's framework only provides a way of defining the simplicity of models.

[21] Friedman [1983], like some of the authors he cites (p. 242), describes unification as the process of reducing the number of independent theoretical assumptions. Of course, a model that assumes principles A, B, and C is made more probable if these assumptions are whittled down to just A and B. However, as Friedman realizes, head counting will not deliver this verdict when the postulates of one model fail to be a subset of the postulates of the other.

Friedman suggests (*e.g.*, pp. 259-60) that a unified model receives more 'boosts' in confirmation than a model of narrower scope. If model M_u covers domains D_1 and D_2, whereas model M_1 covers only domain D_1, then M_u can receive a confirmational boost from both data sets, whereas M_1 can receive a boost only from D_1. Two points need to be made about this proposal. First, although M_u receives two boosts whereas M_1 receives only one, the conjunction M_1 and M_2 receives two boosts as well. Here M_2 is a model that aims to explain only the data in D_2. The conjunction M_1&M_2 is a *dis*unified model. If one wishes to explain the virtues of unification, one should compare M_u with this conjunction, not M_u with M_1. The second point is that 'boosts' in probability are *increases* in probability, not the absolute values

might be constructed that seeks to explain them both. Alternatively, a disunified pair of models M_1 and M_2 also might be constructed, each theory addressing a different part of the total data. If M_1 fits D_1 at least as well as M_u does, and if M_2 fits D_2 at least as well as M_u does, what reason could there be to prefer M_u over the conjunction of M_1 and M_2? The temptation is to answer this question by invoking some consideration that lies outside of what the evidence says. One might appeal to the allegedly irreducible scientific goal of unification or to the connection of unification with simplicity.

The problem posed by the question of goodness-of-fit is a real one, since the combined data set D_1 and D_2 often will be more heterogeneous than either subpart is on its own. This engenders a conflict between unification and goodness-of-fit; a unified theory that encompasses both data sets will fit the data less well than a conjunction of two separate theories, each tailor-made to fit only a single data set. However, just as in the curve fitting problem, this conflict can be resolved. Once again, the key is to correct for the fact that disunified theories are more inclined to overfit the data than their unified counterparts are.

For example, consider the two data sets represented in Figure 3 and the following three models:

(M_u) The X and Y values in D_1 and D_2 are related by the function

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \sigma U.$$

(M_1) The X and Y values in D_1 are related by the function

$$Y = \beta_0 + \beta_1 X + \sigma U.$$

(M_2) The X and Y values in D_2 are related by the function

$$Y = \gamma_0 + \gamma_1 X + \sigma U.$$

Since each data set is close to collinear, M_1 will be more likely than M_u with respect to D_1 and M_2 will be more likely than M_u with respect to D_2. However, what happens when we use Akaike's Theorem to compare M_u with the conjunction M_1 and M_2, relative to the combined data? Notice that M_u has four free parameters, whereas the conjunction M_1 and M_2 has five. If its assumptions apply (see Appendix A), Akaike's Theorem entails that M_u may be more

---

thus attained. The fact that M_u receives two boosts while M_1 receives only one is quite consistent with M_u's remaining less probable than M_1. Friedman (pp. 143-4) recognizes this problem. His solution is to argue that deriving M_1 from a unified theory M_u renders M_1 more plausible than it would be if M_1 were not so derivable. We note that this claim, even if it could be sustained, does not show why M_u is more plausible than M_1 and M_2, where the unified model and its disunified competitor are *incompatible*. In addition, the fact that M_1 is more plausible in one scenario than it is in another does not bear on the question of how plausible M_u is.

In addition to these specific problems with Friedman's proposal, we also wish to note that its basic motivation is contrary to what we learn from Akaike's framework. Friedman seeks to connect unification with paucity of assumptions; as we will see in what follows, unified models impose *more* constraints than their disunified counterparts.

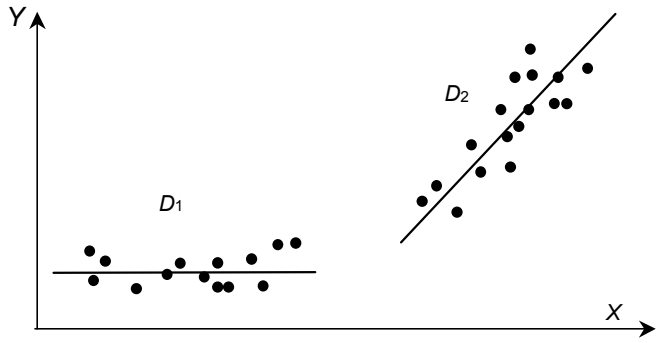

FIGURE 3

predictively accurate even though its best case is less likely than the best case of M_1 and M_2. The best fitting case of the disunified theory would have to have a log-likelihood at least 1 unit greater than the best fitting case of the unified model if the disunified model were to be judged predictively superior. This is not true for the data in Figure 3. We conclude that estimated accuracy explains why a unified model is (sometimes) preferable to its disunified competitor. At least for cases that can be analyzed in the way just described, it is gratuitous to invoke 'unification' as a *sui generis* constraint on theorizing.

The history of astronomy provides one of the earliest examples of the problem at hand. In Ptolemy's geocentric astronomy, the relative motion of the earth and the sun is independently replicated within the model for each planet, thereby unnecessarily adding to the number of adjustable parameters in his system. Copernicus's major innovation was to decompose the apparent motion of the planets into their individual motions around the sun together with a *common* sun-earth component, thereby reducing the number of adjustable parameters. At the end of the non-technical exposition of his programme in *De Revolutionibus*, Copernicus repeatedly traces the weakness of Ptolemy's astronomy back to its failure to impose any *principled* constraints on the separate planetary models.

In a now famous passage, Kuhn ([1957], p.181) claims that the unification or 'harmony' of Copernicus' system appeals to an 'aesthetic sense, and that alone'. Many philosophers of science have resisted Kuhn's analysis, but none has made a convincing reply. We present the maximization of estimated predictive accuracy as the rationale for accepting the Copernican model over its Ptolemaic rival. For example, if each additional epicycle is characterized by 4 adjustable parameters, then the likelihood of the best basic Ptolemaic model, with just twelve circles, would have to be e^{20} (or more than 485 million) times the likelihood of its Copernican counterpart with just seven circles for the evidence to favour the Ptolemaic proposal.[22] Yet it is generally agreed that these basic

[22] If the log-likelihood is penalized by subtracting k, then the likelihood is penalized by multiplying it by a 'decay factor' e^{-k}.

models had about the same degree of fit with the data known at the time. The advantage of the Copernican model can hardly be characterized as merely aesthetic; it is observation, not *a prioristic* preference, that drives our choice of theory in this instance.[23]

4. CAUSAL MODELING

Newton's first Rule of Reasoning in Philosophy in *Principia* was that 'we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.' Here Newton gives voice to a version of Ockham's razor -- explanations that postulate fewer causes should be preferred over explanations that postulate more. Although this injunction is often thought to be quite separate from the criterion of evidential support, some everyday applications of the rule can be given a simple representation in Akaike's framework.

The entries in the following table represent the probabilities that an event C has, given the four combinations of the putative causes A and B:

$$P(C / -)$$

	A	$-A$
B	$w + a + b + i$	$w + b$
$-B$	$w + a$	w

Next we define a characteristic function χ_A:

$\chi_A = 1$ if A occurs

$\chi_A = 0$ if A does not occur.

Ditto for the definition of χ_B.

We now can formulate three hypotheses about the probability that C has in these four possible circumstances:

(INT) $P(C / \chi_A = x_A , \chi_B = x_B) = w + ax_A + bx_B + ix_Ax_B$

(ADD) $P(C / \chi_A = x_A , \chi_B = x_B) = w + ax_A + bx_B$

(SING) $P(C / \chi_A = x_A , \chi_B = x_B) = w + ax_A$.

(SING) says that only a single cause (namely A) makes a difference in whether C occurs. (ADD) says that two causes play a role and that their relationship is additive. (INT) says that there are two causes whose contributions are interactive (*i.e.*, nonlinear or nonadditive). The hypotheses are listed in order of increasing parsimoniousness—one cause is simpler than two, and an additive

[23] Forster (1988b) and Harper (1989) argue that the subsequent impact of Kepler and Newton may be understood in the same terms.

model with two causes is simpler than an interactive model for those two causes.

As in the curve fitting problem, it is standard to understand causal modeling as a problem with two parts. First one selects a hypothesis about the form the causal relationship is to take; then one finds the best hypothesis of that form by estimating parameter values. Rather than solving the first problem by appeal to simplicity, our approach shows how estimated predictive accuracy can be brought to bear from the beginning. Suppose one has a large and equal number of observations for each of the four treatment cells. Let the empirical frequencies of C in those four cells be:

$P(C / -)$

	A	$-A$
B	0.5	0.2
$-B$	0.5	0.2

The three hypotheses now have the same best case, namely one in which $w = 0.2$, $a = 0.3$, $b = 0$, and $i = 0$. Recall that the estimated predictive accuracy of each model is $1/N$ times its maximum log-likelihood minus k/N. This means that when one model is a special case of another and they have the same best case, the model of lower dimensionality has greater estimated predictive accuracy. It follows that (SING) has greater estimated predictive accuracy than (ADD) and (ADD) has greater estimated predictive accuracy than (INT). For the data just given, predictive accuracy explains why it is vain to postulate more causes when fewer suffice.[24] And as in our discussion of unification, it is possible to adjust the data set so as to provide a rationale for favouring a hypothesis of greater complexity.

5 THE PROBLEM OF *AD HOCNESS*

The bugbear of *ad hoc* hypotheses has traditionally been raised within the framework of a hypothetico-deductive philosophy of science. Predictions can be deduced from theories only with the help of auxiliary hypotheses. On this view, we test a theory by observing whether its predictions are true. However, the Quine-Duhem thesis states that the core theory may always be shielded from refutation by making after-the-fact adjustments in the auxiliary hypotheses, so that correct predictions are deduced. The classic example of this is Ptolemaic astronomy, where the model may always be amended in the face of potential refutation by adding another circle—so much so that the expression ‘adding

[24] In this example, it isn’t just that fewer causes are preferable to more; in addition, we have shown why an additive model for two causes is preferable to an interactive model of those two causes. Counting causes is a special case of the more general consideration of dimensionality. Forster [1988b] argues that Newton was sensitive to this wider conception.

epicycles to epicycles' has become synonymous with 'ad hocness'. Although we reject the hypothetico-deductive picture of science, we do accept the usual conclusion that there is an important distinction to be drawn between reasonable revision and *ad hoc* evasion.

Philosophers of science have recognized that protection of the core theories by *post hoc* revision is not always bad. The example usually cited is Leverrier's postulation of Neptune's existence to protect Newtonian mechanics from the anomalous wiggles in Uranus' orbit. The problem is to understand the epistemological grounds for distinguishing good from bad revisions of auxiliary hypotheses (which Lakatos [1970] refers to as the *protective belt*). As is customary, we reserve the term 'ad hoc' for revisions of the bad kind, but reject the *ad hominem* or historicist construal of the term. *Ad hocness*, if it is relevant to questions of evidence, has nothing to do with the motives of the person advocating the hypothesis, or with historical sequences of theories and their evidence.[25]

Lakatos [1970] notes, with approval, that Leverrier's amendment of the prior Newtonian planetary model produced *novel predictions*; he introduces the derogatory term 'degenerating' for research programmes that fail to do this. But there are at least two problems with this approach. Musgrave [1974] warns that a careless reading of the term 'novel' may tempt us into a view of confirmation in which historical contingencies are given undue emphasis. The second defect in Lakatos's idea is that it fails to distinguish estimated predictive *success* from predictive *power*. It is obvious that predictive power is important, for without it there can be no predictive success. But predictive power is not enough to indicate that model revisions are of the good kind. For example, the continued addition of epicycles in Ptolemy's astronomy is not degenerate *in Lakatos's sense*. Each addition leads to novel predictions about the future positions of the planets. What we need is a measure of the predictive success that these additions can be expected to bring, and this is what Akaike's idea of estimated predictive accuracy provides.

Our proposal is that a research programme is *degenerative* just in case loss in simplicity is not compensated by a sufficient gain in fit with data. Of course, the fit will always improve, but the improvement may not be enough to increase the estimated predictive value.

Established research programmes often achieve considerable predictive success, so why do some researchers put their money on an undeveloped programme? First note that on our proposal there is no impediment for new programmes to take over the predictive successes of old ones. There is no 'problem of old evidence' (Glymour [1980], Eells [1985]), since estimated

[25] We do not rule out the possibility that historical or psychological circumstances may sometimes be a reliable *indication* of *ad hocness*. Our only point is that these circumstances do not *make* a theory *ad hoc*, anymore than a barometer makes it rain.

predictive accuracy does not depend on the historical sequence of discovery. But further, it is perfectly understandable that researchers may decide where to invest their energy by formulating a judgment about *projected* predictive success, and the degree to which current programmes are degenerating is thus a relevant consideration.[26]

6 THE SUB-FAMILY PROBLEM

While this explication of Lakatos' notion is a point in favour of our approach, there is another type of ad hocness that is a threat to Akaike's programme. A literal reading of Akaike's Theorem is that we should use the best fitting curve from the family with the highest estimated predictive value. However, for any such family, it is possible to construct an *ad hoc* family of curves with the same best fitting curve, with yet higher estimated predictive accuracy: Fix one or more of the adjustable parameters at their maximum likelihood values. Each subfamily, so constructed, will have the same best case. At the end of the procedure, we obtain a zero dimensional family whose *only* member is the best fitting curve of the original family. The Akaike estimate of the predictive accuracy of this singleton family is just the log-likelihood of the curve. If this is allowed, then we are pushed back towards selecting complicated curves that fit the data exactly. We call this *the sub-family problem*.[27]

Our resolution of this problem returns us to an idea described in Section 2: If a curve fits the data so well that it looks 'too good to be true', then it probably is. In order to spell this out, we now describe a theorem (stronger than Akaike's) that characterizes the behaviour of the *error* in estimating the predictive accuracy of families. The *error* of the estimated predictive accuracy of family F, or Error[Estimated($A(F)$)], is defined as the difference between Akaike's estimate of the predictive accuracy of family F and the true predictive accuracy of that family. Notice that the true predictive accuracy is *constant*—it does not depend on which hypothetical data set generated by the truth happens to be the actual data set. On the other hand, the *estimated* predictive accuracy of F does depend on the actual data—it is what statisticians call a random variable. So Error[Estimated($A(F)$)] also depends on the data, and the following theorem describes this dependence by decomposing it into the sum of three errors:[28]

[26] The Akaike approach also finesses the problem of 'Kuhn loss': Superceding theories do not always carry over all the successes of their predecessor. For example, Cartesian vortex theory 'explains' why all planets revolve around the sun in the same direction, whereas Newton's theory dismisses this as a mere coincidence. Within Akaike's framework, the losses are weighed against the gains in the common currency of likelihoods.

[27] The reader should not be misled into thinking that the subfamily problem is a problem for Akaike's criterion alone; it arises for any proposal that measures simplicity by the paucity of parameters.

[28] The result we are about to describe is close to, but not identical with, equation (4.55) in Sakamoto *et al*. ([1986], p.77). Similar formulae were originally proven in Akaike [1973]. See Forster [1992a] for further explanation.

The Error Theorem:

Error[Estimated($A(F)$)] =
Residual Fitting Error + Common Error + Sub-family Error.

It is important to remember that these errors are not errors of prediction - they are errors in the estimation of predictive accuracy. This is why the Error Theorem might be called a 'meta-theorem' - it is a theorem about the 'meaning' of Akaike's Theorem. However, it rests on the same assumptions as Akaike's Theorem (see Appendix A).

Akaike's Theorem states that the average of Error[Estimated($A(F)$)] over all possible data sets generated by the truth is zero, which is to say Akaike's estimate of predictive accuracy is *statistically unbiased*.[29] 'Statistically unbiased' means that its *average* performance will center on the true value of the quantity being estimated; it is a minimal requirement for 'good' estimators. Akaike's estimate conforms to this standard, but sometimes fails to meet another desideratum, which we will refer to as *epistemic unbiasedness*. We shall now explain the distinction in terms of an example.

First, consider a standard example of a statistically unbiased estimate: the measurement of the mass of an object. For this measurement, the deviation from the true mass value is determined by a symmetrical error distribution centred on the true mass value, so that it is just as probable that the measured value is below the true value as it is above the true value. The measured value of mass is a statistically unbiased estimate of the true mass. But now suppose that we modify this estimate by adding +10 or −10 depending on whether a fair coin lands heads or tails, respectively. Supposed that the measured value of mass was 7 *kg*, and the fair coin lands heads. Then the new estimate is 17 *kg*. Surprisingly, this new estimate is also a statistically unbiased estimate of the true mass! The reason is that in an imagined series of repeated instances, the +10 will be subtracted as often as it is added, so that the value of the *average* value of the modified estimate will still be equal to the true mass value. However, we know that the modified estimate is an *over*estimate in this instance, because we know that the coin landed heads. If the coin had landed tails, then the estimate would have been −3 *kg*, and would have been known to be an *under*estimate. In either case, we say that the modified estimate is *epistemically biased*. In sum, the unmodified measurement value is a statistically and epistemically unbiased estimate of the mass, while the modified estimate is statistically unbiased, but epistemically biased. Other things being equal, we prefer an estimate that is epistemically unbiased.

With this distinction in hand, the Error Theorem is able to explain the limitations of Akaike's method. Here is a brief overview of our analysis:

[29] *Statistical* unbiasedness is really a property of the formula for obtaining the estimate, rather than the particular value of the estimator.

First, the common error is the same for *all* families (hence its name); it cancels out when we make comparisons, and has no effect on model selection. It will not be mentioned again. Second, the Residual Fitting Error is statistically and epistemically unbiased. But the Subfamily Error has a peculiar property. It is statistically unbiased (as is required by Akaike's Theorem); however, it is not always free of epistemic bias. Sometimes Akaike's estimate displays an epistemic bias, and this bias is highlighted by the subfamily problem. A careful analysis of the Subfamily Error will reveal the source and nature of the problem.

We begin by filling in some background. One of the assumptions of these theorems is that there is some complex K-dimensional family of hypotheses (curves) that includes the true hypothesis, and that every family F that we may wish to consider is a subfamily of this superfamily (which we will call K). Every hypothesis under consideration may be represented as a point in the *parameter space* of K. This space may be treated as a K-dimensional vector space. So, if we imagine that our coordinate frame is centered on the Truth (where else?), then various hypotheses may be located in different directions, as shown in Figure 4. The two vectors shown are particularly important because the subfamily error is equal to the dot product, or scalar product, of these two vectors. The first vector is the one to $L(K)$, the best fitting curve in K. Clearly this vector will move around when we consider different data sets generated by the truth. In fact, its tip falls just as probably on one point as on any other on the circle shown, although its length will vary as well. The other vector is fixed. It is the vector from the truth, T, to the hypothesis in the family F that is closest to T (viz. the most predictively accurate hypothesis in F). Now, the dot product is the product of the lengths of these two vectors times the cosine of the angle between them. The cosine factor is +1 if the vectors are parallel, 0 if they are orthogonal, −1 if they are anti-parallel, and in between for in between angles.

The Akaike estimate for a low dimensional family whose best fitting case is close to the data (and such families are the dangerous 'pretenders', for they 'unfairly' combine high log-likelihoods with small penalties for complexity)

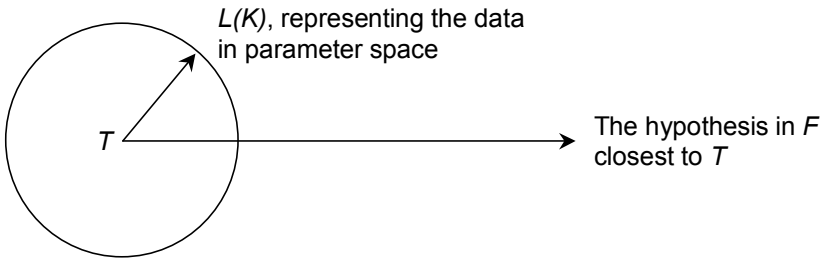

FIGURE 4

exhibits an epistemic bias, as we now explain. The most predictively accurate hypothesis in such small families will also be close to the data, and therefore close to $L(K)$. The danger is that the tips of the two vectors (whose dot product is equal to the subfamily error) will be close together. Then the cosine factor is close to +1 and the subfamily error is large *and positive*. To illustrate this aspect of the relationship of Akaike's Theorem and the Error Theorem, consider the following example. Suppose we have a very large data set that exhibits strong linearity. We wish to estimate the predictive accuracies of L(LIN) and L(POLY-n), where POLY-n is the family of n-degree polynomials with n parameters free, and $L(F)$ is obtained by using the data to single out the best fitting curve in family F.[30] We may apply Akaike's Theorem to (LIN) and (POLY-n) *directly*, or we can apply it to the singleton families containing just L(LIN) and L(POLY-n), respectively. The surprising fact - that the *ad hoc* Akaike estimate for L(POLY-n) is surely an *over*estimate of the predictive accuracy of L(POLY-n) - may have been anticipated from the fact that unreliable *ad hoc* comparisons of L(POLY-n) and L(LIN) will always favour L(POLY-n), because it is always closer to the data. In sum, both the direct and the *ad hoc* method of accuracy estimation are statistically unbiased (as required by Akaike's Theorem), but the *ad hoc* application of Akaike's method yields an estimate that we *know* is too high. The *ad hoc* application yields an estimate that is *epistemically biased*.[31]

We have now unpacked our slogan about a curve's looking 'too good to be true' to provide deeper insights into the source and solution of the subfamily problem: The Akaike estimates of the predictive accuracy of $L(F)$ obtained by viewing $L(F)$ as the best fitting case in the *ad hoc* hierarchy of subfamilies of F tend to be *too high*. Indeed, that is exactly what we observe—the Akaike estimate of $L(F)$ increases steadily as we move down the hierarchy towards the singleton subfamily. In sum: We have good reason not to trust the Akaike accuracy estimates for *ad hoc* subfamilies constructed by fixing adjustable parameters at their maximum likelihood values. We emphasize that this has nothing to do with *when* subfamilies are constructed, or *who* constructs them.

Our analysis of the Error Theorem has been brief and necessarily incomplete. Much more research is needed on the *management of errors* in Akaike's method of model selection. Our aim has been to give the reader a taste for the heuristic power of Akaike's framework in addressing such foundational questions. We close by pointing out that the resolution we have sketched depends (like Akaike's Theorem) on the existence of prediction errors, for otherwise the vector

[30] Remember (from Section 2) that we are interested in estimating the predictive accuracy of a family only because it also provides an estimate of the predictive accuracy of its best fitting curve.

[31] Although the estimate is known to be too high, given the data at hand, the Akaike estimate of the predictive accuracy of that same singleton family relative to *other* data sets generated by the true 'curve' will be too low. On average, of course, the estimate will be centred on the true value.

to $L(F)$ would be 0 and there would no subfamily errors for *any* family.

7. THE BEARING ON BAYESIANISM

The fundamental principle behind Akaike's method is that we should aim to select hypotheses that have the greatest predictive accuracy. Since the truth has the maximum possible predictive accuracy and accuracy is a measure of 'closeness', Akaike's recipe aims to move us *towards* the truth. In contrast, the central thesis of the kind of Bayesianism we will criticize here is that hypotheses should be compared as to their *probability* of truth.[32]

In this section, we examine the possibility that Akaike's method might be recast in a Bayesian framework. Since our argument is many-faceted, we provide a brief summary here. We criticize two different Bayesian proposals that promise to yield a solution to the curve fitting problem. The first Bayesian strategy is to focus on families—show that the best families by Akaike's standards are the most probable families, and then give a Bayesian justification for selecting the best fitting case. The second approach is to bypass families, and show how the most accurate individual hypotheses end up with higher posterior probabilities. After criticizing these suggestions, we end the section by suggesting that Bayesian methods may be useful for assessing the risks in applying Akaike's criterion.

The key element of any Bayesian approach is the use of Bayes' Theorem, which says that the probability of any hypothesis H given any data is proportional to its prior probability times its likelihood: $p(H/\text{Data}) \propto p(H) \times p(\text{Data}/H)$. However, it is an unalterable fact about probabilities that (PAR) is more probable than (LIN), *relative to any data you care to describe*. No matter what the likelihoods are, there is no assignment of priors consistent with probability theory that can alter the fact that $p(\text{PAR/Data}) \geq p(\text{LIN/Data})$. The reason is that (LIN) is a special case of (PAR). How, then, can Bayesians explain the fact that scientists sometimes prefer (LIN) over (PAR)?[33]

Bayesians might propose to address this problem as follows. Instead of (LIN)

[32] The problems we will enumerate for Bayesianism in what follows apply with equal force to what might be called *incremental Bayesianism*. This doctrine has no interest in assigning absolute values to prior and posterior probabilities, but seeks only to make sense of differences or ratios that obtain between these quantities. If H_1 and H_2 are both confirmed by the data, both $P(H_1/\text{Data})/P(H_1)$ and $P(H_2/\text{Data})/P(H_2)$ are greater than unity. To compare these ratios to find out which hypothesis received the larger boost, we need to evaluate the likelihood ratio $P(\text{Data}/H_1)/P(\text{Data}/H_2)$. When the hypotheses are single curves, the better fitting hypothesis automatically receives the higher boost. When the hypotheses are families, evaluating this ratio leads to the problems we will describe in connection with Bayesian approaches to defining the likelihood of families.

[33] One might seek to evade this conclusion by saying that (LIN) and (PAR) are embedded in different theoretical contexts, that this difference gives rise to differences in meaning between their respective theoretical parameters, and that it follows from this that (PAR) is *not* entailed by (LIN). Although we are prepared to grant that this might be plausible in certain special cases, we doubt that this is an adequate response *in general*.

and (PAR), let us consider (LIN) and (PAR*), where (PAR*) is some subset of (PAR) from which (LIN) has been removed. Since (LIN) and (PAR*) are disjoint, nothing prevents us from ordering their prior probabilities as we see fit. In response, we note that this *ad hoc* maneuver does not address the problem of comparing (LIN) versus (PAR), but *merely changes the subject*. In addition, it remains to be seen how Bayesians can justify an ordering of priors for the hypotheses thus constructed and how they are able to make sense of the idea that families of curves (as opposed to *single* curves) possess well defined likelihoods.

Rosenkrantz [1977] and Schwarz [1978] independently argued for a proposal of the first kind—ignoring the problems of logical entailment, they seek to compare the *likelihoods* of *families* of curves.[34] So consider some family of curves F with dimension k. The idea is to define the *average* likelihood of the family in terms of some prior weighting of the members of the family, $p(\text{Curve}/F)$.[35]

If $p(\text{Curve}/F)$ is *strictly* informationless, then it is easy to see that $p(\text{Data}/F) = 0$. Almost every curve in the family will be very far from the data. This means that if we accord equal weight to every curve in F, the average likelihood of F will be zero. What if we let $p(\text{Curve}/F)$ be 'almost' informationless? This means that we divide the curves in the family into two subsets -- within one subset (which includes curves close to the data points), we let the weights be equal and nonzero; outside this volume, we let the weights be zero. We illustrate this proposal by returning to the examples of (LIN) and (PAR), where the error variance σ^2 is known. For (LIN), we specify a volume V_1 of parameter values for α_0 and α_1 within which the likelihoods are non-negligible. For PAR, we specify a volume V_2 of parameter values for β_0, β_1, and β_3 with the same characteristic. If we let boldface $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ range over curves in (LIN) and (PAR) respectively, the average likelihoods of those families then may be expressed approximately as follows:

$$p(\text{Data/LIN}) = (1/V_1) \int \cdots \int p(\text{Data}/\boldsymbol{\alpha},\text{LIN})\, d\boldsymbol{\alpha}$$

$$p(\text{Data/PAR}) = (1/V_2) \int \cdots \int p(\text{Data}/\boldsymbol{\beta},\text{PAR})\, d\boldsymbol{\beta},$$

where the integration is restricted to the subsets of curves with non-zero weights. Note that as larger and larger volumes are taken into account, the average likelihoods approach zero (as the weighting become more strictly informationless).

How are these two likelihoods to be compared? The volume V_1 has two dimensions in parameter space; the volume V_2 has three. Although Rosenkrantz

[34] They ignore the entailment problem by comparing only the likelihoods of families; they bracket the Bayesian comparison of *posterior probabilities*.

[35] Here, the 'average likelihood' is an average over the members of a family of curves, and the Data are fixed. In contrast, the 'average log-likelihoods' we discussed in previous sections were averages of the log-likelihood of a *single* curve with respect to many (hypothetical) data sets.

[1977] and Schwarz [1978] do not formulate their analysis in terms of the volumes V_1 and V_2, their proposal is equivalent to setting $V_1 = V_2$. This is one way to render commensurable the volumes of different dimensionality that appear in the likelihood expressions.[36]

The trouble is that the proposal is not *invariant under reparameterization*. Consider the following pair of equations:

(LIN) $$Y = \alpha_0 + \alpha_1 X + \sigma U$$

(LIN′) $$Y = (\alpha_0')/3 + (\alpha_1'/2) X + \sigma U.$$

These equations define exactly the same family of straight lines. Yet, the proposal entails that the latter has 6 times the average likelihood of the former.[37]

Let us now turn to another strategy that Bayesians might pursue in finding a solution to the weighting problem. This is to let $p(\boldsymbol{\alpha}/\text{LIN})$ be equal to some informative probability $p(\boldsymbol{\alpha}/\text{LIN},E_0)$. Here the weighting scheme is a posterior probability, constructed on the basis of some evidence E_0 that was acquired before the Data. The difficulty with this proposal is that it only pushes the problem back a step. One still has to make sense of the average likelihood $p(E_0/\text{LIN})$. This requires us to evaluate quantities of the form $p(\boldsymbol{\alpha}/\text{LIN})$. Eventually, this must lead the Bayesian back to the quest for informationless (or almost informationless) priors, which we have discussed already.[38] In light of these considerations, we think it is highly questionable that this first Bayesian

[36] The ad hocness of any such assumption is noted by Aitkin [1991], who refers his readers to Lindley [1957].

[37] The reader can most easily grasp this result by considering the problem of integrating a function $f(x)$, where $f(x) = 1$ between the limits 0 and 1, and $f(x) = 0$ elsewhere. Clearly,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 .$$

Yet if we transform coordinates such that $x' = 6x$, while equating $g(x')$ and $f(x)$ for corresponding values of x and x', we obtain

$$\int_{-\infty}^{\infty} g(x')\,dx' = 6 .$$

[38] Nevertheless, Schwarz [1978] has pressed ahead and derived an interesting asymptotic expression for the average likelihood (with the V term omitted). Under conditions similar to those for Akaike's Theorem,

$$\text{Log(Average Likelihood of } F) = \log p(\text{Data}/L(F)) - (\log N)\, k/2 + \text{ other terms},$$

where $L(F)$ is the maximum likelihood hypothesis in F, N is the number of data, and k is the dimension of F. The 'other terms' are negligible for large N. The resulting recipe for model selection is often referred to as the Bayesian Information Criterion, or BIC for short. We will not evaluate the criterion here. But we deny that it is securely grounded in the Bayesian framework, for the reasons we have given. In that regard, it is interesting to note that the same criterion has been independently derived from quite different principles by Akaike [1977] and Rissanen [1978], [1989].

approach—in which *families* of curves are the objects of investigation—can provide a satisfactory treatment of the curve fitting problem.[39]

So let us consider a Bayesian who compares the probabilities of *particular* curves. The problem here is that there seems to be no principled way for estimated predictive accuracies to affect the estimated probability of their truth. For such a Bayesian is bound by Bayes' Theorem, which says that the posterior probability of such a particular hypothesis is proportional to the prior probability times the likelihood relative to the *total* evidence:

$$p(\text{Curve/Data}) = p(\text{Curve})\, p(\text{Data/Curve})\, /p(\text{Data})\ .$$

The likelihood term, p(Data/Curve), simply measures the goodness-of-fit, so the only vehicle for including any estimate of the predictive value of the curve is in the prior probability, p(Curve). In order to replicate the Akaike result, we would need

$$p(\text{Curve}) = p(\text{Data})\, e^{-k}\ ,$$

where p(Data) is merely a normalization factor. But we do not see how a Bayesian can justify assigning *priors* in accordance with this scheme.

The problem is not avoided by adopting a subjectivist approach that eschews the need for objective justification. The problem is deeper than that. The trouble is that a *particular* curve, as opposed to a family of curves, cannot be assigned a value of *k on a priori grounds*. After all, any curve is a member of *many families* of different dimensions. While this problem for Akaike arises in the guise of the subfamily problem, the proposed solution was to distrust subfamilies that have a special relationship *with the data*. However, no comparable solution is available to the Bayesians because the determination of k must be made independently of the data. Thus, Bayesians must find an entirely different kind of solution to the subfamily problem,

[39] However, Aitkin [1991] has a different 'average likelihood' proposal, which allegedly solves the curve fitting problem. He computes the average by weighing each curve in the family by its *posterio*r probability p(Curve/Data), given *all* the available data. A theorem based on the same assumptions as Akaike's Theorem shows that:

$$\text{Log(Aitkin Average Likelihood of } F) = \text{ Log-likelihood}(L(F)) - (\text{k}/2)\log 2\ .$$

Since log2 is less than 1 (the logarithms are to base e), Aitkin imposes less than 1/2 of Akaike's penalty for complexity. This is already an uncomfortable consequence because the Error Theorem shows that (PAR) will be chosen over (LIN) by Aitkin's criterion more often than not *even when (LIN) is true*. But the real problem is that the criterion is just 'pulled out of a hat.' What will families of greater average posterior likelihood provide for us? Will they tend to bring us closer to the truth, or give us more accurate predictions, or what? Aitkin provides no answers to these questions. Given that Aitkin's proposal does not have more fundamental principles to fall back on, how does he cope with the subfamily problem? There is no analogue to the Error Theorem for Aitkin because there is no sense in which average likelihood is in error if it is not estimating anything. Also see the commentaries immediately following Aitkin's paper, including one by Akaike.

and we fail to see how this can be done.[40]

Our diagnosis of the problem is that Bayesianism is unable to capture the proper significance of considering *families* of curves. We work with families because they deliver the most reliable estimates of the predictive accuracy of a few curves; namely their best fitting cases. There is no reason to suspect that such an enterprise can be construed as maximizing the probability that these best fitting cases are true. Why should we be interested in the *probability* of these curves' being true, when it is intuitively clear that no curve fitting procedure will ever deliver curves that are *exactly* true? If we have to live with false hypotheses, then it may be wise to lower our sights, and aim at hypotheses that have the highest possible *predictive accuracy*. Thus, the brand of Bayesianism most popular amongst philosophers is founded on too narrow a conception of the scientific enterprise.[41]

Having said all that, we do not draw the rash conclusion that Bayesian methodology is irrelevant to Akaike's new predictive paradigm. There are many Bayesian solutions to practical statistical problems. However, Akaike's reconceptualization of statistics does recommend that the *foundations* of Bayesian statistics require rethinking.[42] A positive suggestion may be that Bayesian methods can help determine the probability that one hypothesis is more predictively accurate than another. In that way, Bayesian methods might be usefully brought to bear on the problem of assessing the *reliability* of estimated accuracies, for that appears to be an important and open area of research.

8. EMPIRICISM AND REALISM

One virtue of our approach is that it makes clear what the simplicity of a curve has to do with the reasons one might have for believing it. Popper [1959] argued that simpler curves are more falsifiable; Sober [1975] suggested that simpler curves are more informative. These proposals, and others like them,[43] make it

[40] In this respect, we think it is instructive to consider the recent attempt by Jefferys and Berger [1992] to provide a Bayesian rationale for Ockham's razor. We criticize their proposal in Sober and Forster [1992].

[41] It is easy to construct examples which show that maximizing probability of truth is different from maximizing closeness to the truth. A common example is the use of averages to estimate a discrete number, say the number of children in an American family. An estimate of 1.9 children has less probability of being true in any case than an estimate of 2, but may be predictively more accurate nevertheless.

[42] Akaike [1985] shows how the rule of Bayesian conditionalization, as a method of updating probabilities, may be understood in terms of maximizing expected predictive accuracy.

[43] Turney [1990] demonstrates that simpler families of curves are more *stable*. Roughly, the instability of a family of curves, relative to the data, is the expected 'distance' (measured by the SOS) of a new best fitting curve from the old best fitting curve when the data are perturbed in accordance with the *known* error distribution. Turney's measure of instability takes one step

difficult to say why one ought to believe simpler curves rather than their more complex competitors. In contrast, the analysis we have proposed greatly simplifies the task of justification. When a simpler curve is more plausible than its more complex alternatives, this is because it has a higher estimated predictive accuracy.

We believe that our account of curve fitting is good news for empiric*ism*, although it does not accord with what has been said by many empiric*ists*. The idea that some *sui generis* criterion of simplicity is relevant to judging the plausibility of hypotheses is deeply inimical to empiricism. For empiricism, hypothesis evaluation should be driven by data, not by *a priori* assumptions about what a 'good' hypothesis should be like. Empiricists often take this point to heart and conclude that simplicity is a merely pragmatic virtue, one having to do with the usefulness of hypotheses, but not with their plausibility (*cf. e.g.*, Van Fraassen [1980], pp. 87-89). The embarrassing thing about this dismissal of simplicity is that it applies not just to highly theoretical hypotheses, but to quite mundane empirical generalizations of the sort that figure in some curve fitting problems. In these contexts, skepticism about simplicity threatens to lead the empiricist down the garden path to skepticism about induction (Sober [1990a]). Empiricists therefore should welcome the idea that curve fitting does not require a *sui generis* criterion of simplicity. This does not show that some form of radical empiricism is true. Rather, we draw the more modest conclusion that *the data tell you more than you may have thought*.[44]

Although our goal has been to show how the simplicity of a curve can reflect important facts about its predictive accuracy, we do not claim that all uses of simplicity and parsimony in science reduce to purely evidential considerations. We do not deny that scientists often have pragmatic reasons for using simpler curves instead of more complex ones. However, we would insist that these pragmatic considerations not be confused with evidential ones. Monolithic theories about simplicity and parsimony—which claim that these considerations are *never* evidential or that they are *never* merely pragmatic—should be replaced by a more pluralistic approach. At least in the context of the curve fitting problem, Akaike's technical result provides a benchmark that identifies the degree to which simplicity has evidential significance. Any further weight accorded to simplicity, we suspect, derives from pragmatic considerations.

Our analysis supports the idea that the simplicity of a family of curves is an

---

towards estimating the degree of overfitting, as we have characterized it. However, in our opinion, his paper does not show why stability should be relevant to the question of what to believe. We also note that Turney leaves open the justification for trade offs between simplicity and goodness-of-fit. Akaike's Theorem is more general than Turney's theorem in any case—it is not restricted to the standard curve fitting situation, and does not assume a *known* error variance.

[44] For the bearing of this thesis on traditional arguments against the existence of component forces in Newtonian physics, see Forster [1988b].

epistemic epiphenomenon.[45] Sometimes simpler curves are to be preferred over more complicated ones, but the reason for this is not that simplicity is an epistemic end-in-itself. At other times, more complex curves are to be preferred over simpler alternatives, but this is not because the irreducible demands of simplicity are overwhelmed by more weighty considerations of some other sort. Whether a simpler curve is preferable to some more complex alternative, or the reverse is true, has nothing to do with simplicity and everything to do with predictive accuracy.

Our brand of empiricism is not antithetical to the realist view that science *aims* at the truth,[46] in the same sense that archers aim at the bull's-eye even when they have no hope of hitting it. In the past, the curve fitting problem has posed a dilemma: Either accept a realist interpretation of science at the price of viewing simplicity as an irreducible and *a prioristic* sign of truth and thereby eschew empiricism, or embrace some form of anti-realism. Akaike's solution to the curve fitting problem dismantles the dilemma. It now is possible to be a realist and an empiricist at the same time.

Popper [1968] initiated a realist program that takes the 'disastrous meta-induction' (Laudan [1984]) seriously - all of our scientific theories in the past have been false, so it is likely that all of our theories in the future will also be false. Even granting this prediction of failure, it may make sense to claim that our theories *aim* at the truth if we could (1) define a measure of closeness-to-the-truth, and (2) show how theory choice could be viewed as implementing some method that would, more often than not, take us closer to the truth. Proposed solutions to the problem of defining verisimilitude have never gained wide acceptance,[47] and the second part of the programme is seldom discussed.

We have already described predictive accuracy as a measure of closeness to

[45] This thesis complements the view of parsimony developed in Sober [1988b], [1990b]. It also might be formulated in terms of the idea of *screening off*: Simplicity is correlated with plausibility, but only because simplicity also is correlated with predictive accuracy. Once the estimated predictive accuracy of a hypothesis is held fixed, its simplicity has nothing further to contribute to an assessment of its plausibility.

[46] We do not claim that this is the *only* aim of science. We agree with sociologists of science that a complete account of the *practice* of science must include an account of pragmatic and social values. Modern theories of decision making are well equipped to model scientific practice in terms of pragmatic, social, and evidential considerations, in a way that is compatible with realism (Hooker [1987]). However, we do oppose those extremists who believe that internal evidential considerations play no role in the social dynamics of science.

[47] Popper's original definition of verisimilitude was formulated in terms of the *deductive* consequences of theories; fatal flaws were detected independently by Tichý [1974] and by Miller [1974]. Tichý [1974] presents an alternative definition of his own, which Miller [1974] shows to be language dependent. Miller [1975] also argues that the intuitive notion of accuracy of prediction is also subject to the same kind of language variance. Good's [1975] reply to Miller's paper contains a brief explanation of why a probabilistic definition of accuracy, like Akaike's, is not susceptible to Miller's objection. See Forster [1992a] for further discussion.

the truth. To that extent, Akaike's approach revitalizes Popper's programme.[48] However, we suspect that those neo-Popperians who seek some grand metaphysical definition of closeness to the truth will be disappointed with a notion of predictive accuracy defined by reference to a specified domain of inquiry.[49] Nonetheless, we are convinced that any definition of verisimilitude must be limited in this way if we are primarily interested in epistemological questions. In any event, the important point is that Akaike's Theorem lays the epistemological foundation for our progress towards the truth in this domain-relative sense.

In spite of our sympathy for Popper's quest for a concept of verisimilitude, we nonetheless reject hypothetico-deductivism, on which the Popperian programme is founded.[50] The hypothetico-deductivist strategy has been to adopt an idealized model of science in which there are no probabilistic errors in the data, to use this error-free idealization to solve various philosophical problems, and then to add an account of error as an afterthought.[51] Our analysis suggests that many central problems in the philosophy of science are not decomposable in this way. Simplicity and unification are relevant to our judgments about what is truth-like only to the extent that observing and inferring are subject to error.

9. APPENDIX A: THE ASSUMPTIONS BEHIND AKAIKE'S THEOREM

There are three kinds of assumption behind the proof of Akaike's Theorem. First, there is a 'uniformity of nature' assumption that says that the true curve, whatever it is, remains the same for both the old and the new data sets considered in the definition of predictive accuracy. The second kind of assumption consists of mathematically formulated conditions that ensure the 'asymptotic normality' of the likelihood function (*viz*. the likelihood viewed as a function of parameter values). These assumptions contribute to proving various central limit theorems in mathematical statistics. The final assumption is that the sample size (the amount of data) is large enough to ensure that the likelihood function will approximate its asymptotic properties. It is the second assumption that requires the most explaining. We first say what the 'normality' assumption

[48] This perspective also is relevant to Cartwright's [1983] argument that the proliferation of mutually incompatible models in physics is a reason to reject realism. This is an embarrassment to a realist who interprets all (viable) models as true. On the other hand, our brand of realist is only interested in interpreting hypotheses as being more or less close-to-the-truth. A plurality of models is conducive to a more modest realist programme.

[49] We note in this connection that there are philosophical issues raised by the concept of prediction that are not addressed by Akaike's notion of predictive accuracy.

[50] Note that hypothetico-deductivism, as we understand it, is not rescued by the fact that *probabilistic* assertions about future data are deduced from scientific hypotheses. For hypothetico-deductivism demands that at least some of the deductive consequences of our theories are *observations*, but we do not observe probabilities.

[51] See Forster [1994] for a discussion of how this bears on Hempel's raven paradox.

is, and describe the pivotal role it has played in statistics.

The normal, or Gaussian, probability distribution is easily recognized in its one dimensional form by its characteristic bell shape. In its more general multivariate form, the normal distribution has come to play a pivotal role in experimental and theoretical statistics. In experimental statistics, error distributions (in the estimation of parameter values) are found to be approximately normal, especially for large data sets. According to Cramér ([1946], p.231), 'Such is the case, e.g., with the distributions of errors of physical and astronomical measurements, a great number of demographical and biological distributions, etc.' In fact, the assumption that measurement errors are normally distributed around a mean value is so widespread in science that it is often referred to as *the law of errors*. On the theoretical side, 'the central limit theorem affords a theoretical explanation of these empirical facts.' In a somewhat humorous tone, Cramér ([1946], p.232) sums up by quoting Lippman as saying: 'everyone believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact,' and adds that 'both parties are perfectly right, provided that their belief is not too absolute.'

Mathematically, these assumptions are difficult to state explicitly, not just because they are mathematically esoteric, but also because there are various ways in which the assumptions may be weakened (see Cramér [1946]). For this reason, mathematical statisticians almost always vaguely refer to the assumptions as 'certain regularity conditions.' They would certainly not make the brazen claim that these conditions hold for all real scientific models, and we follow their lead here. However, we do wish to say that the conditions are not unduly restrictive. There is no need to assume that the error distributions associated with the observational data are themselves approximately bell-shaped. The standard coin tossing example illustrates the point. The assumed 'error' distribution is the binomial distribution (the probability getting the high value is p, while the probability of the low value is $(1-p)$), yet the distribution for the p-estimates is asymptotically normal. The second point is that asymptotic normality is not restricted to models that are linear in their parameters. For example, suppose that the product $\alpha\beta$ occurs in one of the equations of the model. If $\hat{\alpha}$ and $\hat{\beta}$ are their maximum likelihood estimates and the values of α and β are sufficiently close to these estimates, then we may write: $\alpha\beta = (\hat{\alpha} + \Delta\alpha)(\hat{\beta} + \Delta\beta) \approx \hat{\alpha}\hat{\beta} + \hat{\alpha}\Delta\beta + \hat{\beta}\Delta\alpha$. Here, $\hat{\alpha}$ and $\hat{\beta}$ are constants, and the nonlinear product is now linear in the new, transformed, parameters $\Delta\alpha$ and $\Delta\beta$. This approximation will be valid because the region of non-negligible likelihoods becomes more narrowly concentrated around the best estimates as the sample size increases. The same argument applies to other sufficiently smooth nonlinear equations, such as $Y = \sin(\alpha X + \beta)$, and so on.

Perhaps the most restrictive assumption is that the sample size be large. This

does not mean merely that the *total* data set is large, but that there is enough data within the domain of each parameter. For example, the approximate normality of the model M_1 and M_2 in Section 3 requires that *both* of the data sets D_1 and D_2 are sufficiently large.

10 APPENDIX B: A PROOF OF A SPECIAL CASE OF AKAIKE'S THEOREM

Suppose that we are sampling from a target population of values of a random variable X (e.g. the population of possible measurements of the mass of an object) with mean μ^* (the true mass) and variance σ^2 (the error of measurement), where the true probability distribution p for the values x of the random variable X is normal, or Gaussian. That is,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu^*)^2 \right].$$

Now consider a hypothesis ('curve') that (falsely) asserts that the mean is μ. The hypothesis in question asserts that the probability distribution for measured values of X is

$$q(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2 \right].$$

Hypotheses like $q(x)$ form a *family* of hypotheses, each of which corresponds to a particular value of the parameter μ. Thus, it is notationally convenient to denote the hypothesis itself by μ. (It will be clear from the context when μ is the parameter, the parameter value, or the hypothesis in the family corresponding to a parameter value.) The simplicity of a family of hypotheses (referred to by statisticians as a *model*) is measured by the number of adjustable parameters; in this case there is only one (μ).

If we accept this family of hypotheses, the next step is to find the best fitting hypothesis, and this is the hypothesis that confers the highest probability (density) on the data (i.e. has the maximum likelihood out of all the members of the family). We denote the maximum likelihood hypothesis (which is also

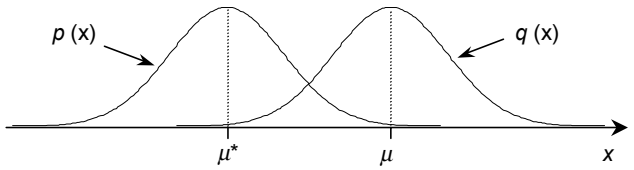

FIGURE 5

the maximum log-likelihood hypothesis) by $\hat{\mu}$. How will $\hat{\mu}$, obtained from past data, fare in the prediction of new data drawn from the same population? For any *particular* datum x, we might measure the accuracy with which it is predicted by its goodness-of-fit; viz. the log-likelihood, $\log p(x)$. But we are really interested in the 'average datum' drawn from the population, so we define the *predictive accuracy* (A for 'accuracy') of an arbitrary hypothesis μ to be:

$$A(\mu) =^{\text{df}} E^*(\log q(x)),$$

where $q(x)$ is the probability distribution in the family corresponding to the parameter value μ, and E^* is the expected value calculated with respect to the true hypothesis (μ^*). That is,

$$A(\mu) = \int_{-\infty}^{\infty} p(x) \log q(x)\, dx .$$

Note that $A(\mu)$ is the expected log-likelihood *per datum* for a data set of arbitrary size N. From the diagram, it is intuitively clear that a distribution $q(x)$ with central point μ that is far from the true value μ^* is not going to do so well in predicting data randomly sampled from the true population. By the same token, $p(x)$ is going to do the best job of fitting the data it generates. The following result gives this intuitive fact a quantitative representation:

$$A(\mu) = A(\mu^*) - \tfrac{1}{2} (\mu - \mu^*)^2/\sigma^2. \qquad (1)$$

Proof: The log of

$$\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

is clearly equal to

$$-\tfrac{1}{2} (\mu - \mu^*)^2/\sigma^2.$$

But,

$$(x-\mu)^2 = (x - \mu^* - (\mu - \mu^*))^2 = (x - \mu^*)^2 - 2(x-\mu)(\mu - \mu^*) + (\mu - \mu^*)^2.$$

When we take expectations and simplify the result follows. This completes the proof.

Since (1) holds for any hypothesis in the family, it surely holds for the hypothesis that best fits the past data. Thus,

$$A(\hat{\mu}) = A(\mu^*) - \tfrac{1}{2}(\mu - \mu^*)^2 \big/ \sigma^2 .$$

While interesting, this result is still epistemologically unhelpful because we don't know $A(\mu^*)$ and we don't know the value of μ^*. The second problem is surmounted in the following way. We may estimate $A(\hat{\mu})$ by the expected value of the right hand side, where the expected value is taken over the maximum likelihood estimate $\hat{\mu}$. That is,

$$\text{Estimate of } A(\hat{\mu}) = E^*\left[A(\mu) - \tfrac{1}{2}(\hat{\mu} - \mu^*)^2 \big/ \sigma^2 \right].$$

But the *central limit theorem* tells us that the expected sum of squared deviations of an estimate of μ from its true value is just σ^2/N, where N is the number of data in the sample from which the estimate is taken (the number of 'past data'). Thus, we have

$$\text{Estimate of } A(\hat{\mu}) = A(\mu^*) - \tfrac{1}{2}/N. \qquad (2)$$

The only remaining problem is to estimate $A(\mu^*)$. Again the qualitative facts are clear. If $\hat{\mu}$ is the best fitting hypothesis relative to past data, then it fits the past data better than any other hypothesis (by definition), and therefore it fits better than μ^*. Thus, if $l(\hat{\mu})$ is the log-likelihood of the best fitting hypothesis, then $l(\hat{\mu}) > l(\mu^*)$ and $E^*(l(\hat{\mu})/N) > E^*(l(\mu^*)/N) =^{\text{df}} A(\mu^*)$. The question as to *how much* greater is answered by the following result (without proof):

$$A(\mu^*) = E^*(l(\hat{\mu})/N) - \tfrac{1}{2}/N. \qquad (3)$$

If we now combine (2) and (3) we get:

$$\text{Estimate of } A(\hat{\mu}) = E^*(l(\hat{\mu}) - 1)/N.$$

Since $l(\hat{\mu}) - 1$ is clearly an unbiased estimate of $E^*(l(\hat{\mu}) - 1)$, we finally arrive at the main result, as it applies to this example:

Akaike [1973]: $\text{Estimate of } A(\hat{\mu}) = (1/N)[l(\hat{\mu}) - 1].$

That is, if we are interested in the predictive accuracy of the best fitting hypothesis from the family, we should not judge its accuracy by its goodness-of-fit, for that estimate is usually biased towards being too high. An *unbiased* estimate is obtained by using a *corrected* measure of goodness-of-fit.

The important fact is that this result generalizes (surprisingly well) to a variety of conditions, and to examples of models with many adjustable parameters. If k is the number of adjustable parameters in a model, then we may state Akaike's theorem in its general form:

Akaike [1973]: $\text{Estimate of } A(\hat{\mu}) = (1/N)[l(\hat{\mu}) - k].$

This is the formula that quantifies the trade-off between simplicity (the number of adjustable parameters) and goodness-of-fit (the maximum log-likelihood).

Department of Philosophy
University of Wisconsin, Madison 53706

REFERENCES

AITKIN, M. [1991]: 'Posterior Bayes Factors.' *Journal of the Royal Statistical Society*. *B* 1: 110-128.

AKAIKE, H. [1973]: 'Information Theory and an Extension of the Maximum Likelihood Principle.' B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*: 267-81. Budapest: Akademiai Kiado.

AKAIKE, H. [1974]: 'A New Look at the Statistical Model Identification.' *IEEE Transactions on Automatic Control*, vol. AC-19: 716-23.

AKAIKE, H. [1977]: 'One Entropy Maximization Principle.' P. R. Krishniah (ed.), *Applications of Statistics*: 27-41. Amsterdam: North-Holland.

AKAIKE, H. [1985]: 'Prediction and Entropy.' In A. C. Atkinson and S. E. Feinberg (eds.), *A Celebration of Statistics*. New York: Springer. 1-24.

BOGEN, J. & J. WOODWARD [1988]: 'Saving the Phenomena.' *The Philosophical Review*, vol. XCVII: 303-352.

CARTWRIGHT, N. [1983]: *How the Laws of Physics Lie*. Oxford: Oxford University Press.

CHURCHLAND, P. M. [1989]: *A Neuralcomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge: MIT Press.

CRAMÉR H. [1946]: *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

EELLS, E. [1985]: 'Problems of Old Evidence.' *Pacific Philosophical Quarterly*: 283-302.

FORSTER, M. [1986]: 'Unification and Scientific Realism Revisited.' In A. Fine and P. Machamer (eds.), *PSA 1986*. E. Lansing, Michigan: Philosophy of Science Association. 1: 394-405.

FORSTER, M. [1988a]: 'Confirmation of Component Causes.' *PSA 1988*. E. Lansing, Michigan: Philosophy of Science Association. 1: 3-9.

FORSTER, M. [1988b]: 'Unification, Explanation, and the Composition of Causes in Newtonian Mechanics.' *Studies in the History and Philosophy of Science*: 55-101.

FORSTER, M [1992a]: 'Progress Towards the Truth.' In preparation.

FORSTER, M. [1992b]: 'The Problem of Overfitting in Artificial Neural Networks.' In preparation.

FORSTER, M. [1994]: 'Non-Bayesian Foundations for Statistical Estimation, Prediction, and the Ravens Example.' *Erkenntnis* **40:** 357 - 376.

FRIEDMAN, M. [1983]: *Foundations of Space-Time Theories*. Princeton, NJ: Princeton University Press.

GLYMOUR, C. [1980]: *Theory and Evidence*. Princeton: Princeton University Press.

GOOD, I. J. [1985]: 'Comments on David Miller,' *Synthese* 30: 205-206.

HARPER, W. [1989]: 'Consilience and Natural Kind Reasoning.' In J. R. Brown and J. Mittelstrass (eds.) *An Intimate Relation*: 115-152. Dordrecht: Kluwer Academic Publishers.

HOOKER, C. [1987]: *A Realistic Theory of Science*. Albany: State University of New York Press.

HOWSON, C. and URBACH, P. [1989]: *Scientific Reasoning: The Bayesian Approach*. La Salle, Illinois: Open Court.

JEFFERYS, W. and J. BERGER [1992]: 'Ockham's Razor and Bayesian Analysis.' *American Scientist* 80: 64-72.

KUHN, T. [1957]: *The Copernican Revolution*. Cambridge, Mass.: Harvard University Press.

LAKATOS, I. [1970]: 'Falsificationism and the Methodology of Scientific Research Programmes' in Lakatos and Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

LAUDAN, L. [1984]: 'A Confutation of Convergent Realism.' In J. Leplin (ed.), *Scientific Realism*. Berkeley and Los Angeles: The University of California Press.

LINDLEY, D. V. [1957]: 'A Statistical Paradox.' *Biometrika* 44: 187-192.

LINHART, H. and W. ZUCCHINI [1986]: *Model Selection*. N. Y.: John Wiley & Sons.

MILLER, D. [1974]: 'Popper's Qualitative Theory of Verisimilitude,' *British Journal for Philosophy of Science* 25: 166-77.

MILLER, D. [1975]: 'The Accuracy of Predictions,' *Synthese* 30: 159-191.

MOODY, J. [1992]: 'The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems.' In J. E. Moody, S. J. Hanson and R. P. Lippmann (eds.) *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kauffmann Publishers.

MURATA, N., S. YOSHIZAWA, and S. AMARI [1992]: 'Network Information Criterion - Determining the Number of Hidden Units for an Artificial Neural Network Model.' Unpublished Manuscript, June 22 1992, University of Tokyo.

MUSGRAVE, A. [1974]: 'Logical Versus Historical Theories of Confirmation.' *British Journal for the Philosophy of Science* 25: 1-23.

POPPER, K. [1959]: *The Logic of Scientific Discovery*. London: Hutchinson.

POPPER, K. [1963]: *Conjectures and Refutations*. London: Routledge and Kegan Paul.

PRIEST, G. [1976]: 'Gruesome Simplicity.' *Philosophy of Science 43*: 432-437.

RISSANEN, J. [1978]: 'Modeling by the Shortest Data Description.' *Automatica* 14: 465-471.

RISSANEN, J. [1989]: *Stochastic Complexity in Statistical Inquiry*. Singapore: World Books.

ROSENKRANTZ, R. [1977]: *Inference, Method, and Decision*. Dordrecht: D. Reidel.

SAKAMOTO, Y., M. ISHIGURO, and G. KITAGAWA [1986]: *Akaike Information Criterion Statistics*. Dordrecht: Kluwer Academic Publishers.

SCHWARZ, G. [1978]: 'Estimating the Dimension of a Model.' *Annals of Statistics* 6: 461-5.

SOBER, E. [1975]: *Simplicity*. Oxford: Oxford University Press.

SOBER, E. [1988a]: 'Likelihood and Convergence.' *Philosophy of Science* 55: 228-37.

SOBER, E. [1988b]: *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, Mass.: MIT Press.

SOBER, E. [1990a]: 'Contrastive Empiricism.' In W. Savage (ed.), *Minnesota Studies in the Philosophy of Science: Scientific Theories*, vol. 14, Minneapolis: University of Minnesota Press, 392-412.

SOBER, E. [1990b]: 'Let's Razor Ockham's Razor.' In D. Knowles (ed.), *Explanation and Its Limits*. Royal Institute of Philosophy Supplementary Volume 27, Cambridge: Cambridge University Press, 73-94.

SOBER, E. and M. FORSTER [1992]: 'Lessons in Likelihood.' *American Scientist* 80: 212-13.

TICHÝ, P. [1974]: 'On Popper's Definitions of Verisimilitude,' *British Journal for the Philosophy of Science*. 25: 155-160.

TURNEY, PETER [1990]: 'The Curve Fitting Problem—A Solution.' *British Journal for the Philosophy of Science* 41: 509-30.

VAN FRAASSEN, B. [1980]: *The Scientific Image*. Oxford: Oxford University Press.

WALLACE, C. S. and P. R. FREEMAN [1992]: 'Single Factor Analysis by MML Estimation.' *Journal of the Royal Statistical Society B* 54: 195-209.

WHEWELL, W. [1840]: *The Philosophy of the Inductive Sciences* (1967 edition). London: Frank Cass & Co Ltd.