2 What is the problem of simplicity?

Elliott Sober

Scientists sometimes choose between rival hypotheses on the basis of their simplicity. Non-scientists do the same thing; this is no surprise, given that the methods used in science often reflect patterns of reasoning that are at work in everyday life. When people choose the simpler of two theories, this 'choosing' can mean different things. The simpler theory may be chosen because it is aesthetically more pleasing, because it is easier to understand or remember, or because it is easier to test. However, when philosophers talk about the 'problem of simplicity', they are usually thinking about another sort of choosing. The idea is that choosing the simpler theory means regarding it as *more plausible* than its more complex rival.

Philosophers often describe the role of simplicity in hypothesis choice by talking about the problem of curve-fitting. Consider the following experiment. You put a sealed pot on a stove. The pot has a thermometer attached to it as well as a device that measures how much pressure the gas inside exerts on the walls of the pot. You then heat the pot to various temperatures and observe how much pressure there is in the pot. Each temperature reading with its associated pressure reading can be represented as a point in the coordinate system depicted below (figure 2.1). The problem is to decide what the *general* relationship is between temperature and pressure for this system, given the data. Each hypothesis about this general relationship takes the form of a line. Which line is most plausible, given the observations you have made?

One consideration that scientists take into account when they face curve-fitting problems is goodness-of-fit. A curve that comes close to the observed data fits the data better than a curve that is more distant. If goodness-of-fit were the only consideration relevant to the curve-fitting problem, scientists would always choose curves that pass exactly through the data points. They do not do this (and even if they did, the question would remain of how they choose among the infinity of curves that fit the

I am grateful to Malcolm Forster for useful discussion.

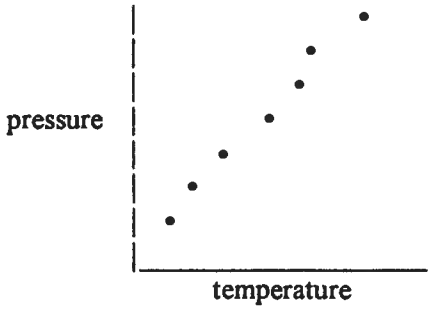


Figure 2.1 Pressure/temperature

data perfectly). Another consideration apparently influences their decisions, and this is simplicity. Extremely bumpy curves are often thought to be complex, whereas smoother curves are often thought to be simpler. Scientists sometimes reject an extremely bumpy curve that fits the data perfectly and accept a smoother curve that fits the data a little less well. Apparently, scientists care about both goodness-of-fit *and* simplicity; both considerations influence how they choose curves in the light of data.

The fact that scientists use simplicity considerations to help them choose among competing hypotheses gives rise to a philosophical problem. In fact, the problem of simplicity decomposes into three parts. The task is to show how simplicity should be *measured*, how it can be *justified*, and how it should be traded off.

1 Measuring simplicity

To strive for simplicity in one's theories means that one aims to minimize something. The minimization might involve a semantic feature of what a set of sentences says or a syntactic feature of the sentences themselves. An example of the first, semantic, understanding of simplicity would be the idea that simpler theories postulate fewer causes, or fewer changes in the characteristics of the objects in a domain of inquiry. An example of the second, syntactic, understanding of simplicity would be the idea that simpler theories take fewer symbols to express, or are expressible in terms of a smaller vocabulary.

If simplicity involves minimization, why should there be a problem in determining how simple a theory is? The problem is to figure out what to count. Theory X may seem simpler than theory Y when the counting is done one way, but the opposite conclusion may be reached if the counting is done differently. Consider, for example, the longstanding dispute in

psychology and the social sciences about psychological egoism. Egoism claims that all of our ultimate motives are focused on some benefit to self. The rival of egoism is motivational pluralism, which allows that some of our ultimate motives are egoistic, but maintains that other ultimate motives are focused on the welfare of others. Motivational pluralism says that people have both egoistic and altruistic ultimate desires. It may seem, at first glance, that egoism is simpler than pluralism, since egoism postulates just one type of ultimate desire, whereas pluralism postulates two (Batson, 1991). This may be why social scientists often favour the egoism hypothesis in spite of the fact that both egoism and pluralism are consistent with observed behaviour.

Complications arise when one considers other implications that egoism and pluralism have. Egoism says that people want to help others only because they think that helping will benefit themselves. Egoism therefore attributes to people a certain *causal belief*. Motivational pluralism postulates no such belief about the selfish benefits that helping will produce. If the simplicity of the two theories were calculated by counting such causal beliefs, the conclusion would be drawn that motivational pluralism is simpler. Egoism postulates fewer types of ultimate desire, but it postulates a larger number of causal beliefs (Sober and Wilson, 1998). What is gained on the one hand is lost on the other.

Theories have many implications. Which of those implications should be considered when one calculates how simple a theory is? This problem is not solved merely by stipulating a metric. One must show why simplicity should be measured in one way rather than another. This task inevitably affects the second and third parts of the problem of simplicity – the problem of justifying simplicity and of trading it off.

Another example of the measurement problem is suggested by Nelson Goodman's (1965) 'new riddle of induction'. Consider the following two hypotheses:

> (H1) All emeralds are green.
> (H2) All emeralds are green until the year 2050; after that, they are blue.

(H1) seems to be the simpler hypothesis. It says that emeralds do not change colour, whereas (H2) says that they do. Here simplicity involves minimizing the number of postulated changes. If the emeralds we have observed to date are all green, then (H1) and (H2) fit the observations equally well. If simplicity is a reason to prefer one hypothesis over another, then perhaps we are entitled to choose (H1) on grounds of its greater simplicity.

Problems arise when we consider the predicate 'grue'. Let us say that an object is grue at a time if and only if the object is green and the time is before the year 2050, or the object is blue and the time is after the year 2050. Symmetrically, we will say that an object is 'bleen' at a time if and only if it is blue and the time is before the year 2050 or it is green and the time is after the year 2050. Just as blue and green are colours, we can say that grue and bleen are 'grulers'. We now can reformulate (H1) and (H2) in terms of the vocabulary just introduced:

> (H1′) All emeralds are grue until the year 2050; after that, they are bleen.
> (H2′) All emeralds are grue.

(H1′) and (H1) are logically equivalent; so are (H2′) and (H2). Which of (H1′) and (H2′) is simpler? (H2′) says that emeralds do not change gruler; (H1′) says they do. So if simplicity is calculated by seeing which theory minimizes change in gruler, (H2′) is simpler. This means that it is not an adequate solution to the measurement problem to say that simplicity should be measured by seeing how many changes a theory postulates. The proposal is incomplete. Is it change in colour or change in gruler that matters?

A similar question may be raised about the proposal that the simplicity of a proposition should be measured by considering the length of the shortest sentence that expresses it (Rissanen, 1978, 1989). In a language that contains colour predicates, but not gruler predicates, the proposition that all emeralds are green can be encoded more briefly than the proposition that all emeralds are grue. In a language that contains gruler predicates, but not colour predicates, the reverse is true. And in a language that contains both types of predicate, the two propositions have the same minimum description length. Stipulating which language should be used resolves this ambiguity, but a further question needs to be answered. Why should we adopt one language, rather than another, as the representational system within which simplicity is measured?

Syntactic measures of simplicity inevitably encounter this sort of problem. Since a proposition can be encoded in many different ways, depending on the language one adopts, measuring simplicity in terms of code features will fail to be linguistically invariant. This will not be true of measures of simplicity that are purely semantic, since semantic measures focus on features of what a proposition says, not on how the proposition is expressed in any particular language. Any measure of simplicity in terms of probability will be semantic in this sense. The conditional probability Pr(Q|P) describes a relationship between the two propositions Q

and P; once values for such quantities are specified, they remain the same regardless of what language one uses to encode the propositions. Another way to put this point is that if two sentences are logically equivalent, then they must have the same probability. (H1) and (H1′) must be equiprobable, as must (H2) and (H2′).

These comments about the difference between syntactic and semantic measures of simplicity may sound odd, given the fact that the minimum description length proposal and the Bayesian approach of Schwarz (1978) are asymptotically equivalent (Rissanen, 1978, 1989). In fact, this type of result does not contradict the point just made about the difference between syntactic and semantic proposals. Suppose a probability model is adopted for a set of propositions, and that the simplicity of a proposition is then identified with some probabilistic property that it possesses. This semantic measure of simplicity does not rule out the possibility that some suitable language might be constructed in which the simplicity of a proposition in the set is identified with the length of the shortest sentence in that language that expresses the proposition. It is possible that the semantic and syntactic definitions might induce precisely the same ordering of propositions. This would be no more surprising than the fact that the relation of logical entailment can be defined both syntactically and semantically in propositional logic, and the two definitions are equivalent.

One reason the problem of measuring simplicity is difficult is that intuitive descriptions of what simplicity means lead to opposite measurement proposals. It is often said that simpler theories make fewer assumptions. It also is often said that simpler theories have fewer adjustable parameters. These intuitive statements come into conflict when we examine families of models that are nested. Consider, for example, the following two hypotheses; the first says that x and y are linearly related, the second says that their relation is parabolic:

(LIN) $\quad y = a + bx$

(PAR) $\quad y = a + bx + cx^2$

(LIN) has two adjustable parameters (a and b); (PAR) has three (a, b, and c). Once values for these parameters are specified, a unique straight line and a unique parabola are obtained. Notice that (LIN) is simpler than (PAR) if simplicity is calculated by counting adjustable parameters. However, if simplicity involves paucity of assumptions, the opposite conclusion follows. (LIN) is equivalent to the conjunction of (PAR) and the further assumption that $c = 0$. Since (LIN) is a special case of (PAR), (LIN) says more, not less.

2 Justifying simplicity

To justify simplicity is to show why it should be taken into account in judging how plausible a theory is. Ideally, this task would be accomplished by giving a general theory about what makes one hypothesis more plausible than another, and then explaining how simplicity contributes to plausibility.

I am using 'plausibility' as an informal and neutral term. There are a number of philosophical theories about what this concept means. Bayesians suggest that one theory is more plausible than another when the first is more probable. Popperians say that one theory is more plausible than another if it is better 'corroborated'; this will be true if both are consistent with the observations and the first is more falsifiable (Popper, 1959). Likelihoodists understand plausibility in terms of evidential support, which they define in terms of the probability that a hypothesis confers on the observations (Edwards, 1984). The Akaike framework maintains that one theory is more plausible than another when it has a higher expected degree of predictive accuracy; here predictive accuracy is a technical term, defined in the context of a specific type of inference problem in which one selects a specific hypothesis from a family of hypotheses, based on the data at hand, and uses that hypothesis to predict new data (Akaike, 1973; Sakamoto et al., 1986; Forster and Sober, 1994).

Although most of these approaches attempt to provide a 'global' theory of inference that applies to all hypotheses, no matter what their logical form and no matter what their subject matter, it is important to realize that these approaches can be understood as 'local' proposals as well. Perhaps Bayesianism works well in some contexts, but not in others; the same may be true for the other frameworks. It may turn out that inference obeys different rules in different settings, and that simplicity has to be justified in different ways in different contexts. Even though theories of inference have traditionally been global, it is possible that only local analyses can be constructed (Sober, 1988, 1994). Maybe there is no such thing as *the* justification of simplicity.

It may seem pessimistic to suggest that there might be no single unified justification of simplicity in scientific inference. An even more pessimistic suggestion is that the project of justification – whether global or local – is itself impossible. Perhaps simplicity can't be justified in terms of something else. If this is right, then there are two options. Either simplicity should be rejected as an irrelevant consideration in hypothesis evaluation, or it needs to be regarded as an ultimate consideration – something that is valuable for its own sake, not because it contributes to some broader aim.

Just as the question 'why be rational?' may have no non-circular answer, the same may be true of the question 'why should simplicity be considered in evaluating the plausibility of hypotheses?'

Even if local justifications of simplicity are developed for certain types of hypotheses in certain sorts of evidential contexts, the possibility remains that simplicity may not be justifiable in other settings. For example, I have argued, following Reichenbach (1938), that the justification for using simplicity to choose between theories that make different predictions does not furnish a reason for choosing between theories that are predictively equivalent (Sober, 1996). Does this mean that simplicity has some quite different rationale in the latter type of problem, or that it has no justification at all?

Another feature of the problem of justification is worth noting. If simplicity is used to justify choosing theory X over theory Y, then it must be shown that the greater simplicity of theory X makes it more plausible. What are the assumptions on which this demonstration will rest? One possibility is that purely *a priori* considerations will suffice; perhaps logic and mathematics are enough to show why simplicity contributes to plausibility in the context at hand. Although this style of justification might be desirable, it is reasonable to expect that the justification of simplicity cannot, in general, be so minimal in its empirical assumptions. Typically, substantive empirical assumptions are needed to explain why, in a particular inference problem, the simplicity of a hypothesis contributes to its plausibility (Sober, 1988). I'll illustrate this point by discussing a simple example later in this chapter.

If justifying simplicity means showing why simpler theories are more plausible, then Rissanen's idea that simplicity should be measured by minimum description length cannot stand on its own. Minimum description length is an intuitive representation of what simplicity means and an impressive body of mathematical results has been developed around this thought. However, the problem remains of saying what minimum description length has to do with plausibility. It is here that the asymptotic equivalence with Schwarz's (1978) Bayesian proposal may be important.

3 Trading off simplicity

The task of justifying simplicity involves showing that simplicity 'breaks ties'. That is, if two theories are equally good in all other respects, then the simpler of the two should be regarded as more plausible. However, theories are almost never equally good in all other respects. For example, in the curve-fitting problem discussed at the beginning of this chapter,

simplicity influences our judgments about plausibility, but so does goodness-of-fit. It almost never happens that two curves that we wish to test fit the data equally well. This means that a justification of simplicity, if it is to be relevant to scientific practice, can't merely show that the simpler of two theories is more plausible, all else being equal. Additionally, one needs to establish how sacrifices in simplicity are to be 'traded off' against gains in the other factors that affect a theory's plausibility. If curve X is simpler than curve Y, but curve Y fits the data better than curve X, how are these two pieces of information to be combined into an overall judgment about the plausibility of the two curves?

The trade-off problem shows why the problem of measuring simplicity involves more than saying which of two theories is simpler. Additionally, one must be able to say *how much* simpler one theory is than another, and the constraints on this answer go beyond merely preserving judgments of comparative simplicity. Consider, for example, the dispute between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Schwarz, 1978). Both approaches can be thought of as equating the complexity of a family of curves with k – the number of adjustable parameters it contains. However, AIC and BIC disagree on the weight that simplicity deserves, when compared with the other factor that is said to affect a family of curve's plausibility. For a data set of fixed size, the two proposals may be formulated as follows:

$$\text{The AIC estimate of family F's plausibility} = \text{log-likelihood}\{L[F]\} - k$$

$$\text{The BIC estimate of family F's plausibility} = \text{log-likelihood}\{L[F]\} - k\log(n)/2.$$

L(F) is the likeliest curve in the family – i.e., the one that confers the highest probability on the data at hand. For large data sets, BIC places more emphasis on simplicity than AIC does.

Regardless of how the dispute between AIC and BIC is resolved, it is important to recognize that both approaches not only provide a way of measuring simplicity, but do so in a way that is commensurable with the measurement of other properties of hypotheses that affect their plausibility. This is an essential feature of any adequate solution to the problem of simplicity. It is not enough to be told that the complexity of a family of curves is measured by the number of adjustable parameters it contains. This information is useless until one is also told how simplicity and goodness-of-fit together affect a hypothesis' plausibility.

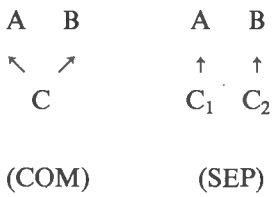
Several philosophers have asserted, without providing much of a supporting argument, that the trade-off problem has no objective solution.

For example, Kuhn (1977) claimed that scientists differ in how much importance they assign to one virtue of a theory as opposed to another, and that this difference is just a matter of taste. One scientist may think that the most important demand on a theory is that it should make accurate predictions; another may hold that the first duty of a theory is that it be elegant and general. Kuhn does not offer much of an argument for this claim; it merely constitutes his *impression* of what would be left open by any compelling and reasonably complete set of epistemological standards. Of course, it is not in dispute that scientists have different temperaments. But from the point of view of normative theory, it is far from obvious that no uniquely best trade-off exists between simplicity and the other factors that affect a theory's plausibility. Once again, this cannot be settled in advance of examining particular proposals – the AIC, the BIC and others.

4 A simple example

These various points about the problem of simplicity can be illustrated by considering a simple example. Suppose Al and Beth are students in a philosophy seminar, which is held in a seminar room in Madison whose wall of windows looks out on a lake. Students in philosophy seminars sometimes find that their attention wanders; it is possible to concentrate on the Eternal Forms for only so long. This, apparently, is what happens during one of the seminar sessions. At a certain point in the afternoon, Al and Beth both find themselves thinking 'there is a sailboat with a red sail on the lake'.

There are two hypotheses I want to consider, each of which might explain this matching of Al's and Beth's mental states. The common cause explanation (COM) says that Al and Beth were both looking at the same thing on the lake. The separate cause explanation (SEP) says that they were looking at different things. The two hypotheses might be represented as follows:



(COM) (SEP)

The (COM) hypothesis seems to be simpler than the (SEP) hypothesis. After all, one is a smaller number than two. The question I now want to

explore is why this difference in simplicity should make a difference when we decide how to evaluate (COM) and (SEP). Why is simplicity relevant?

I propose to identify a set of assumptions that suffices to guarantee that (COM) is more likely than (SEP). As noted before, I am using likelihood in the technical sense introduced by R. A. Fisher (1925). To say that (COM) is likelier than (SEP) means that (COM) confers a higher probability on the observations than (SEP) does. First, let's introduce some abbreviations for various propositions:

A = Al believes that there is a sailboat on the lake with a red sail.
B = Beth believes that there is a sailboat on the lake with a red sail.
C = The object that Al and Beth both looked at was a sailboat on the lake with a red sail.
$-$C = The object that Al and Beth both looked at was not a sailboat on the lake with a red sail.

Notice that C and $-$C are both part of the (COM) explanation – both say that Al and Beth were looking at the same object; C and its negation differ over what the characteristics of that common cause were. Next, let's introduce abbreviations for some relevant probabilities:

$\Pr(A \mid C) = a$ $\quad$ $\Pr(A \mid -C) = m$
$\Pr(B \mid C) = b$ $\quad$ $\Pr(B \mid -C) = n$
$\Pr(C) = c$

These allow us to represent the likelihood of the (COM) explanation as follows:

$$\Pr(A\&B|COM) = cab + (1 - c)mn.$$

This way of expressing the probability of the two observations, given the existence of a common cause, assumes that the belief states of Al and Beth were conditionally independent, given the state of the object they were both looking at. This is not an implausible assumption about vision – Al's probability of having a certain belief, given the state of the object that he and Beth are examining, is not affected by what Beth comes to believe. Or, at least, this is plausible, if we assume that Al and Beth don't have a conversation as they look out at the lake.

We now need to develop a similar representation of the likelihood of the (SEP) hypothesis. First, we need some new propositions:

C_1 = The object that Al was looking at, but Beth was not, was a sailboat on the lake with a red sail.

$-C_1$ = The object that Al was looking at, but Beth was not, was not a sailboat on the lake with a red sail.

C_2 = The object that Beth was looking at, but Al was not, was a sailboat on the lake with a red sail.

$-C_2$ = The object that Beth was looking at, but Al was not, was not a sailboat on the lake with a red sail.

Notice that C_1 and its negation are propositions that are specific to the (SEP) hypothesis ; they agree that Al and Beth were looking at different things, but disagree about the characteristics of one of those objects. The same holds, of course, for C_2 and its negation.

We now need to introduce some symbols for the probabilities that will figure in our representation of the likelihood of the (SEP) hypothesis. We do this by borrowing letters from our representation of the (COM) hypothesis:

$$\Pr(A \mid C_1) = a \qquad \Pr(A \mid -C_1) = m$$
$$\Pr(B \mid C_2) = b \qquad \Pr(B \mid -C_2) = n$$
$$\Pr(C_1) = \Pr(C_2) = c$$

We now can represent the likelihood of the (SEP) hypothesis as follows:

$$\Pr(A\&B|SEP) = [ca + (1 - c)m][cb + (1 - c)n].$$

Our use of a, b, m, n and c involves a set of empirical assumptions. The fact that the letter 'a' appears in the likelihood expression for (COM) and the likelihood expression for (SEP) means that the probability of Al's believing what he does, given the state of what he is looking at, isn't affected by whether Beth is looking at that object, or at something else. The fact that 'c' occurs in both expressions means that the probability that an object on the lake is a sailboat with a red sail isn't affected by whether two people look at it, or only one does. These assumptions seem reasonable for the case at hand.

Now that we have characterized the likelihoods of the (COM) and the (SEP) hypotheses, we can derive a result that describes which hypothesis will be more likely:

(1) If $0 < c < 1$, then $\Pr(A\&B \mid COM) > \Pr(A\&B \mid SEP)$ if and only if $(a - m)(b - n) > 0$.

Assuming that c is neither 0 nor 1, the common cause hypothesis is more likely precisely when a and m differ, b and n differ, and a and m differ in the same direction that b and n do. This will be true if Al and Beth are, in a certain broad sense, similar causal systems. Suppose that Al's probability of believing that there is a sailboat on the lake with a red sail is

increased if he in fact is looking at such an object. If the same is true of Beth, then the (COM) hypothesis has the higher likelihood. The common cause hypothesis also would be more likely if Al and Beth were both 'counter-suggestible' in the same way. If Al's probability of believing that a red-sailed sailboat is present were lowered by a red-sailed sailboat's being present, and if the same were true of Beth, then the (COM) hypothesis would be more likely in this circumstance as well. Assuming that (COM) is simpler than (SEP), the assumptions just described entail that the simpler hypothesis is more likely.

Proposition (1) provides a criterion for determining which hypothesis has the higher likelihood. For a Bayesian, this result provides only half of the needed analysis. The ultimate goal is to compare the posterior probabilities of the two hypotheses. So far, we have analysed only one of the factors that influence this result. A fuller specification of what needs to be examined is provided by the following criterion, which follows from Bayes' Theorem:

(2) $\Pr(\text{COM} \mid \text{A\&B}) > \Pr(\text{SEP} \mid \text{A\&B})$ if and only if
$\Pr(\text{A\&B} \mid \text{COM})\Pr(\text{COM}) > \Pr(\text{A\&B} \mid \text{SEP})\Pr(\text{SEP})$.

Likelihood influences which posterior probability is greater, but so do the prior probabilities. If simplicity is relevant to this problem because of its connection with likelihood, the question remains of how simplicity should be traded off against the other factor that affects a hypothesis' overall plausibility.

How should we understand the prior probability of the hypotheses in question? How probable was it that Al and Beth looked at the same object on the lake? There is no way to answer this question *a priori.* However, if we adopt the following model of the process of looking, an answer can be obtained. Let us suppose that Al and Beth independently select an object at random from the ones on the lake to look at. This would not be true if some objects were far more interesting to look at than others, but let's assume that Al and Beth are as described. Then, if there are r objects on the lake, the probability that Al and Beth are looking at the same object is $r(1/r)(1/r) = 1/r$. This means that the prior probabilities of the two hypotheses are $\Pr(\text{COM}) = 1/r$ and $\Pr(\text{SEP}) = (r - 1)/r$. For large r, the common cause hypothesis is vastly less probable *a priori.* If the common cause hypothesis is to have the higher posterior probability, then it must overcome this deficiency by having a very superior likelihood. Rewriting (2) and substituting the two prior probabilities just obtained, the result is:

(3) $Pr(COM \mid A\&B) > Pr(SEP \mid A\&B)$ if and only if $Pr(A\&B \mid COM)/Pr(A\&B \mid SEP) > (r - 1)$.

Proposition (1) says that the (COM) hypothesis will have a higher likelihood than the (SEP) hypothesis in a wide range of circumstances. Now proposition (3) tells us how much greater this likelihood needs to be, if the (COM) hypothesis is to be more probable, given the evidence at hand. Of course, just as empirical assumptions entered into our derivation of proposition (1), the same is true for the reasoning behind proposition (3).

This example, I think, illustrates the main ingredients that are needed for the problem of simplicity to be solved in a given type of inference problem. The common cause explanation is simpler than the separate cause explanation of why Al and Beth have the same thought. This should affect our judgment about which explanation is more plausible only to the extent that simplicity can be connected to some independently acceptable conception of what plausibility means. This task is accomplished by showing that the common cause hypothesis has the higher likelihood, on the assumption that likelihood is relevant when we assess the plausibility of competing hypotheses. Empirical assumptions that are fairly reasonable entail that the simpler hypothesis is more likely in this problem. However, there apparently is more to a hypothesis' plausibility than its likelihood. Prior probability also enters, and the common cause hypothesis, though it is more likely, has a lower prior probability. We therefore need a principle that tells us how simplicity (in the guise of likelihood) should be traded off against prior probability. Further empirical assumptions allow this rate of exchange to be calculated. Simpler theories aren't always more probable, given the evidence at hand. Proposition (3) allows us to determine when this is true for the problem under analysis.

In this example, it is interesting that the relevance of simplicity to inference does not depend on the assumption that 'nature is simple' or that it usually is so. It is more likely, in Fisher's sense, that Al and Beth were looking at the same thing, rather than at different things, given that their beliefs matched. This is wholly independent of the question of whether people usually look at the same or at different things. The simpler hypothesis is more *likely*; this does not entail that the simpler hypothesis is more *probable*. In fact, it is not unreasonable to suspect that the very reverse may be true, in terms of the prior probabilities involved. If I tell you that Al and Beth are both looking at the lake, but don't tell you what they are thinking, is it more probable that they are looking at the same object on the lake, or at different objects? Unless there is a single enormously attention-grabbing object there (for example,

a giant fire-breathing sea serpent rising from the lake's centre), it is reasonable to think that the latter hypothesis has the higher prior probability.

I am not suggesting that the analysis of this example provides a satisfactory model for how the entire problem of simplicity should be addressed. However, besides illuminating the general character of 'the problem of simplicity', it also applies to an interesting type of scientific inference. When evolutionists seek to reconstruct the phylogenetic relationships that connect different species to each other, they generally prefer hypotheses that allow them to view similarities as homologies. For example, consider the fact that numerous mammalian species have hair. One could explain this by saying that the shared characteristic is a homology inherited from a common ancestor, or by saying that it is a homoplasy – that it evolved independently in separate lineages. Common sense and scientific practice both embody a preference for the common cause explanation. This preference has a likelihood justification, if the processes occurring in lineages obey a principle that I have called the backwards inequality (Sober, 1988):

$$\Pr(\text{Descendant has hair} \mid \text{Ancestor has hair}) > \Pr(\text{Descendant has hair} \mid \text{Ancestor lacks hair}).$$

This inequality formally resembles the constraint that $a > m$ and $b > n$ discussed in connection with Al and Beth. The backwards inequality is independent of the question of whether stasis is more probable than change – that is, of whether the following forwards inequality obtains:

$$\Pr(\text{Descendant has hair} \mid \text{Ancestor has hair}) > \Pr(\text{Descendant lacks hair} \mid \text{Ancestor has hair}).$$

The forwards inequality is a contingent matter that may be true for some traits in some lineages and false for others. The backwards inequality, however, is much more robust; it follows from the idea that evolution can be modelled as a Markov process in which there is a small constant probability of change per small interval of time (Sober, 1988).

5 Model selection

The example about Al and Beth makes Bayesianism look like a very promising approach to the problem of simplicity. Bayesianism says that the relative plausibility of hypotheses in the light of evidence is to be assessed by calculating likelihoods and prior probabilities. In the problem just analysed, fairly plausible empirical assumptions can be used to

make sense of both these quantities, and one of them seems to explain why simplicity should influence our judgment of a hypothesis' plausibility. If simplicity could always be understood in this format, the problem of simplicity would be solved and its solution would count as a major triumph for Bayesianism.

Unfortunately, matters are not so simple. The role of simplicity in the curve-fitting problem is hard to understand within a Bayesian framework. In this type of inference problem, two quite different kinds of proposition are discussed. There are specific curves, and there are the families of curves to which those specific curves belong. The straight line hypothesis '$y = 3x + 4$' is an example of the former; the family (LIN) – the set of all straight lines – is an example of the latter. Specific curves have no adjustable parameters; families contain at least one adjustable parameter. Let us consider how prior probabilities and likelihoods might be assigned to hypotheses of both these types.

Whether one is talking about specific curves or families of curves, it is usually quite difficult to know how to make sense of their prior probabilities. Harold Jeffreys' (1957) Bayesian proposal is instructive in this regard. Although he provided slightly different proposals in different publications, his idea basically was that the complexity of a differential equation should be measured by summing the absolute values of the integers (degrees and derivative orders) in it, along with the number of adjustable parameters it contains. Jeffreys further suggested that simpler equations should be assigned higher prior probabilities (Hesse, 1967).

A consequence of this proposal is that (LIN) is said to be more probable than (PAR). As Popper (1959) pointed out, this contradicts the axioms of probability; if one hypothesis entails another, the former cannot have the higher probability, no matter what the evidence is on which one conditionalizes. To rescue Jeffreys' idea from contradiction, it is sometimes suggested that the families under comparison be disjoint. Rather than compare (LIN) and (PAR), one should compare (LIN) and (PAR′), where (PAR′) says that $y = a + bx + cx^2$ *and* that $c \neq 0$. There is no logical barrier to stipulating that (LIN) has a higher prior probability than (PAR′).

However, an important question still needs to be addressed. *Why* should (LIN) be assigned a higher prior probability than (PAR′)? I suggest that Bayesians should feel uncomfortable with this proposal, since it says that it is more probable that $c = 0$ than that $c \neq 0$. If I told you I was about to drop a very sharp dart onto a line that is one mile long, would you think it more probable that the dart will fall exactly at the beginning of the line than that it will fall elsewhere? Although this assumption is logically consistent, it cannot be said to be very plausible. The proposal

that (LIN) be assigned a higher prior probability than (PAR′) seems to be an act of desperation, one that conflicts with the intuitions that usually motivate Bayesian reasoning.

Setting aside the difficulty of what should be said about nested models, the general problem here is that the prior probability assignments that Bayesians propose often seem to lack justification. The prior probabilities assigned to the (COM) and (SEP) hypotheses that I discussed in connection with Al and Beth were based on empirical facts about how often two people look at the same thing when they look at a lake. These assignments of priors can be rationally assessed because we are able to view each hypothesis as a possible outcome of a chance process; in this case, frequency data are relevant and in principle available. It is a familiar criticism of Bayesianism that many of the hypotheses that are of interest in science cannot be understood in this way. As a result, assigning them prior probabilities has only a subjective significance, representing someone's degrees of belief.

If prior probabilities don't offer much promise as a vehicle by which Bayesians can explain the relevance of simplicity in the problem of model selection, what about likelihoods? The idea that specific curves have likelihoods – that they confer probabilities on the data – makes perfectly good sense, once an error distribution is specified. The usual model of error entails that a curve that is near the data will have a higher likelihood than a curve that is more distant; goodness-of-fit is relevant to curve fitting because goodness-of-fit reflects likelihood. However, the fact that the likelihoods of specific curves are well defined does not help explain why simplicity is relevant to model evaluation. The reason is that a complex curve can always be found that fits the data better than a simpler curve that fits the data imperfectly. If our only goal is to maximize likelihood, there is no reason to value simplicity.

Since the likelihoods of specific curves cannot explain why simplicity is desirable, perhaps we should consider the likelihoods of families of curves. This approach requires that we ask, for example, what the probability is of obtaining the data, if (LIN) is correct? This quantity is an average over all the specific straight lines (L_1, L_2, ...) that belong to the family:

$$\Pr(\text{Data} \mid \text{LIN}) = \sum \Pr(\text{Data} \mid L_i)\Pr(L_i \mid \text{LIN}).$$

Some of the L_i's are very near the data, so the value of $\Pr(\text{Data} \mid L_i)$ for those straight lines will be large; however, many straight lines will be quite far away, and so the value of $\Pr(\text{Data} \mid L_i)$ for them will be small. The *average* straight line is further from the data than any finite

number, so Pr(Data | LIN) = 0, if we assume that $Pr(L_i \mid LIN) = Pr(L_j \mid LIN)$, for all straight lines L_i and L_j.

One way around this problem is to use some of the data to induce a weighting among these straight lines. Straight lines close to the data are assigned higher weight; $Pr(L_i \mid LIN)$ is said to be greater than $Pr(L_j \mid LIN)$, if L_i is close to the data and L_j is far away. This is essentially the approach of Schwarz (1978). When the average likelihood Pr(Data | LIN) is computed using this system of weights, it can turn out, depending on the data, that (LIN) has a higher average likelihood than (PAR′). Schwarz assumes that the competing models have the same prior probability (in this respect he differs from Jeffreys), with the net result that (LIN) can turn out to have the higher posterior probability.

As mentioned earlier, Schwarz's BIC gives more emphasis to simplicity than the AIC does. However, it is important to notice that Schwarz's analysis of the problem focuses on estimating the average likelihood of families of curves. Akaike's theorem, on the other hand, says that the AIC is an unbiased estimator of predictive accuracy, not of average likelihood. In a sense, the AIC and the BIC provide estimates of different things; yet, they almost always are thought to be in competition. If the question of which estimator is better is to make sense, we must decide whether the average likelihood of a family or its predictive accuracy is what we want to estimate. The predictive accuracy of a family tells you how well the best-fitting member of that family can be expected to predict new data; a family's average likelihood tells you how well, on average, the different members of the family fit the data at hand. Both quantities might be of interest, but they are different.

If we fix on predictive accuracy as our goal, what is the upshot? This decision doesn't automatically mean that AIC is better than BIC. After all, BIC might be a good device for estimating predictive accuracy, even though it was derived to estimate average likelihood. In any event, a very important fact here is that Akaike proved that the AIC is an unbiased estimator of predictive accuracy; it follows that the BIC must be a biased estimator of that quantity. On the other side, it is often said that the AIC is statistically inconsistent, but that BIC is statistically consistent; Malcolm Forster's chapter in the present volume addresses this claim and shows that it rests on using an inappropriate definition of predictive accuracy. Other optimality properties need to be considered as well, if a full assessment of AIC and BIC is to be obtained.

Even though it is important to resolve this question, we should not lose sight of what the AIC and the BIC approaches have in common. The idea that the goal of curve-fitting (and model selection, generally) is to achieve predictive accuracy helps explain why simplicity is relevant in this infer-

ential context. The more adjustable parameters a family contains, the greater its risk of *over-fitting* the data, of mistaking noise for signal (Forster and Sober, 1994). Simplicity is relevant because complex families often do a bad job of predicting *new* data, though they can be made to fit the *old* data quite well. This is part of the practical experience that model builders have in different sciences. The mathematical framework constructed by Akaike and his co-workers allows one to understand *why* complexity contributes to over-fitting, and even to predict, from the data at hand, the degree to which this is apt to happen.

6 Conclusion

Simplicity considerations are relevant in many inference problems. In some, simplicity matters because simpler hypotheses have higher likelihoods; in others, simplicity matters because simpler hypotheses have higher prior probabilities; and in still others, simplicity matters because it reduces the risk of over-fitting. This 'local' approach has thrown considerable light on the longstanding philosophical problem of simplicity; we now have available a set of ideas that indicates how simplicity should be measured, justified and traded off against other desirable properties that hypotheses may possess. The progress that has been made on this problem is encouraging. If the near future resembles the recent past (a simple assumption!), this progress can be expected to continue.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado.

Batson, D. (1991). *The Altruism Question*. Hillsdale, NJ: Erlbaum.

Edwards, A. (1984). *Likelihood*. 2nd edition. Baltimore: Johns Hopkins University Press.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Forster, M. and E. Sober (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45: 1–35.

Goodman, N. (1965). *Fact, Fiction, and Forecast*. Indianapolis: Bobbs-Merrill.

Hesse, M. (1967). Simplicity. *The Encyclopedia of Philosophy*. Vol. 7, pp. 445–8.

Jeffreys, H. (1957). *Scientific Inference*. 2nd edition. Cambridge: Cambridge University Press.

Kuhn, T. (1977). Objectivity, value judgment, and theory choice. In *The Essential Tension*, pp. 320–39. Chicago: University of Chicago Press.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Reichenbach, H. (1938). *Experience and Prediction*. Chicago: University of Chicago Press.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica* 14: 465–71.

(1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Books.

Sakamoto, Y., M. Ishiguro and G. Kitagawa (1986). *Akaike Information Criterion Statistics*. Dordrecht: Kluwer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461–5.

Sober, E. (1988). *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, MA: MIT Press.

(1994). Let's razor Ockham's razor. In *From a Biological Point of View*. New York: Cambridge University Press.

(1996). Parsimony and Predictive Equivalence. *Erkenntnis* 44: 167–97.

Sober, E. and D. Wilson (1998). *Unto Others – the Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.