5 The new science of simplicity[1]

Malcolm R. Forster

1 The problem

No matter how often billiard balls have moved when struck in the past, the next billiard ball *may not* move when struck. For philosophers, this 'theoretical' possibility of being wrong raises a problem about how to *justify* our theories and models of the world and their predictions. This is the *problem of induction*. In *practice*, nobody denies that the next billiard ball *will* move when struck, so many scientists see no practical problem. But in recent times, scientists have been presented with competing methods for comparing hypotheses or models (classical hypothesis testing, BIC, AIC, cross validation, and so on) which do not yield the same predictions. Here there is a problem.

Model selection involves a tradeoff between simplicity and fit for reasons that are now fairly well understood (see Forster and Sober, 1994, for an elementary exposition). However, there are many ways of making this tradeoff, and this chapter will analyze the conditions under which one method will perform better than another. The main conclusions of the analysis are that (1) there is no method that is better than all the others under all conditions, even when some reasonable background assumptions are made, and (2) for *any* methods A and B, there are circumstances in which A is better than B, and there are other circumstance in which B will do better than A. Every method is fraught with some risk even in well behaved situations in which nature is "uniform." Scientists will do well to understand the risks.

It is easy to be persuaded by the wrong reasons. If there is always a situation in which method A performs worse than method B, then there is a computer simulation that will display this weakness. But if the analysis of this article is correct, then there is always a situation in which any

[1] My thanks go to the participants of the conference for a stimulating exchange of ideas, and to Martin Barrett, Branden Fitelson, Mike Kruse, Elliott Sober and Grace Wahba for helpful discussions on material that appeared in previous versions of this paper. I am also grateful to the Vilas Foundation, the Graduate School, and sabbatical support from the University of Wisconsin-Madison.

A. Zellner, H. A. Keuzenkamp, and M. McAleer (eds.) (2001): *Simplicity, Inference and Modelling*, Cambridge University Press, pp. 83-119.

method A will do worse. To be swayed by a single simulation to put all your money on the assumption that the examples of interest to you are the same in all relevant respects. One needs to understand what is relevant and what is not.

Another spurious argument is the (frequently cited) claim that AIC is inconsistent—that AIC does not converge in the limit of large samples to what it is trying to estimate. That depends on what AIC is trying to estimate. Akaike (1973) designed AIC to estimate the expected log-likelihood, or equivalently, Kullback-Leibler discrepancy, or predictive accuracy (Forster and Sober, 1994). In section 7, I show that AIC is consistent in estimating this quantity. Whether it is the most efficient method is a separate question. I suspect that no method has a universally valid claim to that title. The bottom line is that the comparison of methods has no easy solution, and one should not be swayed by hasty conclusions.

The way to avoid hasty conclusions is to analyze the problem in three steps:

(1) The specification of a *goal*. What goal can be reached or achieved?
(2) The specification of a *means* to the goal. What is the *criterion*, or method?
(3) *An explanation* of how a criterion works in achieving the goal.

This chapter is an exercise in applying this three-step methodology to the problem of model selection.

The chapter is organized as follows. Section 2 introduces scientific inference and its goals, while section 3 argues that standard model selection procedures lack a clear foundation in even the *easiest* of examples. This motivates the need for a deeper analysis, and section 4 describes a framework in which the goal of predictive accuracy is precisely defined. The definition of predictive accuracy is completely general and assumption free, in contrast to section 5 which develops the framework using a 'normality assumption' about the distribution of parameter estimates.[2] Even though the assumption is not universal, it is surprisingly general and far reaching. No statistician will deny that this is a very important case, and it serves as concrete illustration of how a science of simplicity should be developed. Section 6 compares the performance of various methods for optimizing the goal of predictive accuracy when the normal-

[2] 'Normality' refers to the bell-shaped normal distribution, which plays a central role in statistics. Physicists, and others, refer to the same distribution as Gaussian, after Carl Friedrich Gauss (1777 - 1855), who used it to derive the method of least squares from the principle of maximum likelihood.

ity assumption holds approximately, and explains the limitations in each method. The clear and precise definition of the goal is enough to defend AIC against the very common, but spurious, charge that it is inconsistent. I discuss this in section 7. Section 8 summarizes the main conclusions.

2 Preliminaries

A model is a set of equations, or functions, with one or more adjustable parameters. For example, suppose LIN is the family of linear functions of a dependent variable y on a single independent variable x, $\{y = a_0 + a_1x + u \mid a_0 \in \mathbb{R}, a_1 \in \mathbb{R}\}$, where $\mathbb{R}$ is the set of real numbers and u is an error term that has a specified probability distribution. The error distribution may be characterized by adjustable parameters of its own, such as a variance, although it is always assumed to have zero mean. Note that there can be more than one dependent variable, and they can each depend on several independent variables, which may depend on each other (as in causal modelling). The family LIN is characterized by two adjustable parameters, while PAR is a family of parabolic functions $\{y = a_0 + a_1x + a_2x^2 + u \mid a_0 \in \mathbb{R}, a_1 \in \mathbb{R}, a_2 \in \mathbb{R}\}$, characterized by at least three adjus-table parameters.

The distinction between variables and adjustable parameters is sometimes confusing since the adjustable parameters are variables in a sense. The difference is that x and y vary within the context of each member of the family, while the parameters only vary from one member to the next. The empirical data specify pairs of (x, y) values, which do not include parameter values. Parameters are introduced *theoretically* for the purpose of distinguishing competing hypotheses within each model.

A typical inferential problem is that of deciding, given a set of seen data (a set of number *pairs*, where the first number is a measured x-value, and the second number is a measured y-value), whether to use LIN or PAR is better for the purpose of predicting new data (a set of unseen (x, y) pairs). Since LIN and PAR are competing models, the problem is a problem of *model selection*. After the model is selected, then standard statistical methods are used to estimate the parameter values to yield a *single* functional relation between x and y, which can be used to predict y-values for novel x-values. The second step is fairly well understood. Model selection is the more intriguing part of the process although model selection is usually based on the properties of the estimated parameter values.

The philosophical problem is to understand exactly how scientists should compare models. Neither the problem, nor its proposed solutions, are limited to curve-fitting problems. That is why econometricians or

physicists, or anyone interested in prediction, should be interested in how to trade off fit with simplicity, or its close cousin, unification. For example, we may compare the solutions of Newton's equations with the solutions of Einstein's mechanics applied to the same physical system or set of systems. Here we would be comparing one huge nexus of *interconnected* models with another huge nexus where the interconnections amongst the parts follow a different pattern. Einstein's solution of the problem of explaining the slow precession of the planet Mercury's orbit around the sun depends on the speed of light, which connects that precession phe-nomenon to quite disparate electromagnetic phenomena. There is wide consensus that Einsteinian physics would come out on top because it fits the data at least as well as the Newtonian equations, and sometimes better, without fudging the result by introducing *new parameters* (the speed of light was already in use, though not in explaining planetary motions). It seems that the overall number of parameters is relevant here. These vague intuitions have swayed physicists for millennia. But physicists have not formalized them, nor explained them, nor understood them, even in very simple cases.

Recent research in statistics has lead to a number numerically precise criteria for model selection. There is classical Neyman-Pearson hypothesis testing, the Bayesian BIC criterion (Schwarz 1978), the minimization of description length (MDL) criterion (Rissanen 1978, 1987; Wallace and Freeman 1987), Akaike's information criterion (AIC) (Akaike 1973, 1974, 1977, 1985; see also Sakamoto *et al* 1986, and Forster and Sober 1994) , and various methods of cross validation (e.g., Turney 1994, Xiang and Wahba 1996). In a few short years we have gone from informal intuition to an embarrassment of riches. The problem is to find some way of critically evaluating competing methods of scientific inference. I call this the 'new science of simplicity' because I believe that this problem should be treated as a scientific problem: to understand when and why model selection criteria succeed or fail, we should model the process of model selection itself. There is no simple and no universal model of model selection, for the success of a selection method depends greatly on the circumstances, and to understand the complexities, we have to model the situation in which the model selection takes place. For philosophers of science, this is like making assumptions about the uniformity of nature in order understand how induction works. The problem is the same: How can we make assumptions that don't simply assume what we want to prove? For example, it would not be enlighten-ing to try to understand why inductive methods favor Einstein's physics over Newton's if we have to assume that Einstein's theory is true in order to model the inferential process. Fortunately, the new work on simplicity

makes use of weaker assumptions. An example of such an assumption is the 'normality assumption'. It simply places constraints on how the estimated values of parameters are distributed around their true values without placing any constraints on the true values themselves.

This is why it is so important not to confuse what I am calling the normality assumption, which is about the distribution of repeated parameter estimates, with an assumption about the normality of error distributions. For example, in the case of a binary event like coin tossing, in which a random variable[3] takes on the values 0 and 1, there is no sense in which the deviation of this random variable from the mean is normal. The error distribution is discrete, whereas the normal distribution is continuous. However, the distribution of the sample mean, which estimates the propensity of the coin to land heads, is approximately normal. A normality assumption about errors is stronger and more restrictive than an assumption of normality for the repeated parameter estimates. It is the less restrictive assumption that is used in what follows.[4]

It is true that models of model selection are a little different from standard scientific models. Scientific models are descriptive, while models of model selection are what I will call weakly normative.[5] For example, models of planetary motion describe or purport to describe planets. But models of model selection relate a model selection criterion to a goal. The goal might be predictive accuracy, empirical adequacy, truth, probable truth, or approximate truth. But whatever the goal, the project is to understand the *relationship* between the methods of scientific inference and the goal. Of this list, predictive accuracy is the one epistemic goal (minimizing description length is a non-epistemic goal) whose relationship with simplicity is reasonably well understood thanks to recent work in mathematical statistics. So, predictive accuracy is the goal considered in this paper.

Bayesianism is the dominant approach to scientific inference in North America today, but what does it take as the goal of inference? Fundamentally, Bayesianism is a theory of decision making, and can consider *any* goal. It then defines the *method* of deciding between two competing models as the maximization of the expected payoff with

---

[3] A random variable is a variable whose possible values are assigned a probability.

[4] Kiessepä (1997) shows that a normality assumption for the error distribution is not always sufficient to ensure normality of the parameter estimators. However, Cramér (1946), especially chapters 32 and 33, explains how the conditions are met asymptotically for large sample sizes in a very general class of cases.

[5] A strongly normative statement is one which says we *should* or we *ought* to do such and such. A weakly normative statement is one that says we should do such and such *in order to optimize a given goal*, without implying that it is a goal we should optimize.

respect to that goal. The simplest idea is that the payoff of scientific theories lies in their truth. With that in mind, it is simplest to assign a payoff of 1 to a true model and 0 to a false model. Let me refer to this kind of Bayesian philosophy of science as *classical* Bayesianism, or *standard* Bayesianism.[6] Consider a choice between model A and model B. Is the expected payoff in selecting A greater than the expected payoff in selecting B? The answer is given in terms of their probabilities. If Pr(A) is the probability that A is true, and Pr(B) be the probability that B is true, then the expected payoff for A is, by definition, Pr(A) times the payoff if it's true plus the Pr(not-A) times the payoff if it's false. The second term disappears, so the expected payoff for believing A is Pr(A). Likewise, the expected payoff for believing B is Pr(B). The expected payoff for believing A is greater than the expected payoff for believing B if and only if Pr(A) is greater than Pr(B). This leads to the principle that we should choose the theory that has the greatest probability, which is exactly the idea behind the model selection criterion derived by Schwarz (1978), called BIC.

Whatever the goal, a scientific approach to model selection is usefully divided into 3 parts:

1. The specification of a *goal*. What goal can be reached or achieved in model selection? Approximate truth is too vague. Probable truth is also too vague unless you tell me what the probability is of. Truth is too vague for the same reason. Are we aiming for the truth of a theory, a model, or a more precise hypothesis?
2. The specification of a *criterion*, or a *means* to the goal. This is where simplicity will enter the picture. What kind of simplicity is involved and exactly how it is to be used in combination with other kinds of information, like fit?
3. *An explanation* of how the criterion works in achieving the goal. For example, Bayesians explain the criterion by deducing it from specific assumptions about prior probability distributions. The Akaike explanation makes no such assumptions about prior probabilities, but instead, makes assumptions about the probabilistic behavior of parameter estimates. The style of the explanation is different in each case, and is a further ingredient in what I am calling the framework.

---

[6] The classical Bayesian approach is currently dominant in the philosophy of science. See Earman (1992) for a survey of this tradition, and Forster (1995) for a critical overview. For alternative 'Akaike' solutions to standard problems in the philosophy of science, see Forster and Sober (1994). For an 'Akaike' treatment of the ravens paradox, see Forster (1994). For an Akaike solution to the problem of variety of evidence, see Kruse (1997).

It should be clear from this brief summary that the difference between the Bayesian and Akaike modeling of model selection marks a profound difference between statistical *frameworks*. What I have to say about the modeling of model selection goes to the very heart of statistical practice and its foundations. Anyone interested in induction agrees that, in some sense, truth is the *ultimate* goal of inference, but they disagree about how to measure *partial success* in achieving that goal. Classical Bayesians do not tackle the problem of defining partial success. They talk of the *probability* that a hypothesis is true, but most Bayesians deny that such probabilities are objective, in which case they do not define partial success in an objective way. There is no sense in which one Bayesian scientist is closer to the truth than another if neither actually reaches the true model.

The same criticism applies to decision-theoretic Bayesians as well. These are Bayesians who treat model selection as a decision problem, whose aim is to maximize a goal, or utility (Young, 1987), or minimize a loss or discrepancy (Linhart and Zucchini, 1986). They are free to specify any goal whatsoever, and so they are free to consider predictive accuracy as a goal. But, again, the expectation is a *subjective* expectation defined in terms of a subjective probability distribution. Typically, these Bayesians do not evaluate the *success* of their method with respect to the degree of predictive accuracy *actually achieved*. They could, but then they would be evaluating their method within the Akaike framework.

Nor do Bayesians consider the *objective* relationship between the method (the maximization of *subjectively* expected utilities) and the goal (the utilities). That is, they do not consider step (3), above. At present, it appears to be an article of faith that there is nothing better than the Bayesian method, and they provide no explanation of this fact (if it is a fact). And even if they did, I fear that it would depend on a *subjective* measure of partial success. That is why the Akaike approach is fundamental to the problem of comparing methods of model selection.

The Akaike framework defines the success of inference by how close the selected hypothesis is to the true hypothesis, where the closeness is measured by the Kullback-Leibler distance (Kullback and Leibler 1951). This distance can also be conceptualized as a measure of the accuracy of predictions in a certain domain. It is an objective measure of partial success, and like truth, we do not know its value. That is why predictive accuracy plays the role of a goal of inference, and not a means or method of inference. The issue of how well any method achieves the goal is itself a matter of *scientific* investigation. We need to develop models of model selection.

The vagueness of the notion of simplicity has always been a major worry for philosophers. Interestingly, all three methods already men-

tioned, the MDL criterion, BIC, and AIC, define simplicity in exactly the same way—as the paucity of adjustable parameters, or more exactly, the dimension of a family of functions (when the two differ, then it is the dimension that is meant, for it does not depend on how the family is described; see Forster, 1999). So, the definition of simplicity is not a source of major disagreement.

In fact, I am surprised that there is *any* disagreement amongst these schools of thought at all! After all, each criterion was designed to pursue an entirely different goal, so each criterion might be the best one for achieving its goal. The MDL criterion may be the best for minimizing description length, the BIC criterion the best for maximizing probability, and the AIC criterion the best at maximizing predictive accuracy. The point is that the claims are *logically independent*. The truth of one does not entail the falsity of the others. There is no reason why scientists should not value all three goals and pursue each one of them separately, for none of the goals are wrong-headed.

Nevertheless, researchers do tend to think that the approaches are competing solutions to the same problem. Perhaps it is because they think that it is impossible to achieve one goal without achieving the others? Hence, there is only one problem of induction and they talk of *the* problem of scientific inference. If there is only one problem, then the Akaike formulation is a precise formulation of the problem, for it provides a definition of partial success with respect to the ultimate goal of truth. For that reason, I will compare all model selection criteria within the Akaike framework.

3 A milieu of methods and an easy example

Here is a very simple example of a statistics problem. Suppose that a die has a probability θ^* of an odd number of dots landing up, which does not change over time, and each toss is independent of every other toss. This fact is not known. The two competing models are M_1 and M_2. Both models get everything right except that they disagree on the probability of an odd number of dots landing up, denoted by θ.

M_1 asserts that $\theta = ½$. This model specifies an exact probability for all events. If M_1 is a family of hypotheses, then there is only one hypothesis in the family. M_1 has no *adjustable* parameters. This is a common source of confusion, since it does mention a parameter; namely θ. But θ is given a value, and is therefore adjust*ed*, and not adjustable. M_2, on the other hand, is uncommitted about the value of θ. θ is now an adjustable parameter, so M_2 is more complex than M_1 in one sense of 'complex'. Also note that M_1 is *nested* in M_2, since all the hypotheses in M_1 also appear in

M_2. The problem is to use the observed data to estimate the probability of future events. There is no precise prediction involved, but we think of it as a prediction problem of a more general kind. The problem of induction applies to this kind of problem.

In classical statistics, there are two steps in the "solution" of this problem. The first step is to test M_1 against M_2. This is the process that I am calling model selection. The second step is to estimate the value of any adjustable parameters in the winning model by choosing the best fitting hypothesis in the family that best fits the seen data. This picks out a single hypothesis which can be used for the prediction or explanation of unseen data. While different statistical paradigms have different definitions of 'best fit', those differences usually make little difference, and I will ignore them here. I will assume that everyone measures fit by the likelihood (or log-likelihood). The naïve empirical method that ignores simplicity and goes by fit alone is called the method of maximum likelihood (ML). In the case of M_1 the maximum likelihood hypothesis has to be $\theta = ½$, since there are no others that can do better. In the case of M_2 there is a well known result that tells us that the maximum likelihood hypothesis is $\theta = \hat{\theta}$, where $\hat{\theta}$ is the relative frequency of heads-up in the observed data. Note that the second step is essential, since M_2 by itself does not specify the value of its adjustable parameter, and cannot be used to make probabilistic assertions about future data.

Here is how classical Neyman-Pearson hypothesis testing works. The simpler of two models is the null hypothesis, in this case M_1 (see figure 5.1). The decision to accept the null hypothesis or reject the null hypothesis (and therefore accept M_2) depends on how probable the data would be if the null hypothesis were true. If the data are improbable given the null hypothesis, then reject the null hypothesis, otherwise accept it. The degree of improbability is determined by the size or the level of significance of the test. A size of 5% is fairly standard ($p < .05$), which means that the null hypothesis is rejected if the observed data is a member of a class of possible data sets that collectively has a probability of 5% given the null hypothesis. The observed relative frequencies that would be lead to such a rejection are those that fall under the shaded area in figure 5.1. The value of the relative frequency shown in Figure 1 lies in that region, so that the null hypothesis is accepted in that case.

Notice that the hypothesis $\theta = \hat{\theta}$ in M_2 fits the observed facts better than the null hypothesis, yet the null hypothesis is still accepted. *Therefore classical model selection trades off fit for simplicity*, provided that the simpler hypothesis is chosen as the null hypothesis.

There are a number of peculiar features of the classical method of model selection. First, there is nothing to prevent the more complex

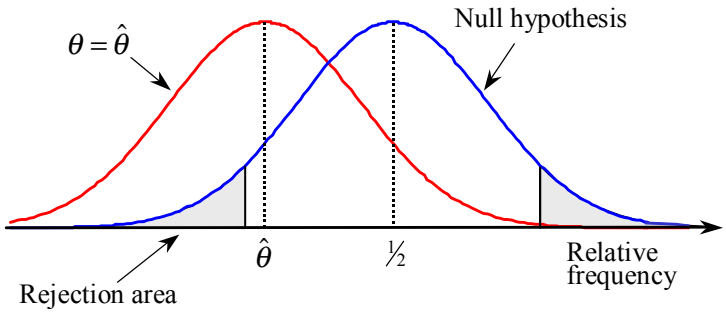

Figure 5.1: Classical Neyman-Pearson hypothesis testing.

model being chosen as the null hypothesis, and there is no reason against this practice except to say that it is not common practice. Nor is there any reason for choosing a 5% level of significance other than common practice. Finally, it is odd that the same tradeoff would be made even if M_2 had many more adjustable parameters than M_1. There is no obvious method for adjusting the size of the test to take account of these features of the context. Neyman-Pearson methods do not appear to have the kind of rationale demanded by the three steps described in the introduction.

I have heard only one reply to this charge. The reply is that classical statistics aims to minimize the probability of rejecting the null hypothesis when it is true (i.e. minimize type I error), and minimize the probability of accepting the null hypothesis when it is false (i.e. minimize type II error), and it does this successfully. I doubt that this is the only aim of the procedure because I think that working scientists are also interested in predictive accuracy, and it is not obvious that classical testing brings us closer to that goal. And, in any case, the two parts to the goal stated above are incompatible. To minimize type I error, we should choose the size of the test to be 0%. But that will maximize the type II error. At the other extreme, one could minimize Type II errors by choosing a 100% significance level, but that would maximize the Type I error. The actual practice is a tradeoff between these two extremes. Classical statisticians need to specify a third goal if the tradeoff is to be principled.

Another objection to the Neyman-Pearson rationale for hypothesis testing is that it fails to address the problem when both models are false. For then I would have thought that any choice is in error, so trading off Type I and Type II errors, which are conditional on one or other of the models being true, is an irrelevant consideration. In other words, there is no criterion of partial success. Note that these are criticisms of the *rationale* behind the method, and not the methods themselves.

In order to explain the AIC and BIC model selection *methods* in this example, it is sufficient to think of them as classical Neyman-Pearson tests, with some special peculiarities. In particular, AIC chooses a greater rejection area (about 15.7%), while BIC recommends a smaller rejection area, which further diminishes as the number of data increases. This is the situation when the competing models differ by one adjustable parameter, as is the case in our example. Figure 5.2 plots the critical point (the point defining the boundary of the rejection area) as a function of the number of coin tosses. Notice that as the number of tosses increases, a smaller deviation of the proportion of heads up from the null result of 50% will succeed in rejecting the null hypothesis, although BIC requires are greater deviation in all cases. Therefore BIC gives greater weight to simplicity in the sense that it requires that there be stronger evidence against the hypothesis before the simpler null hypothesis is rejected.

When the models differ by a dimensions greater than one (such as would be the case if we were to compare LIN with a family of 10-degree polynomials), the size of the rejection areas decrease. This is significantly different from classical Neyman-Pearson testing, which makes no adjustment.

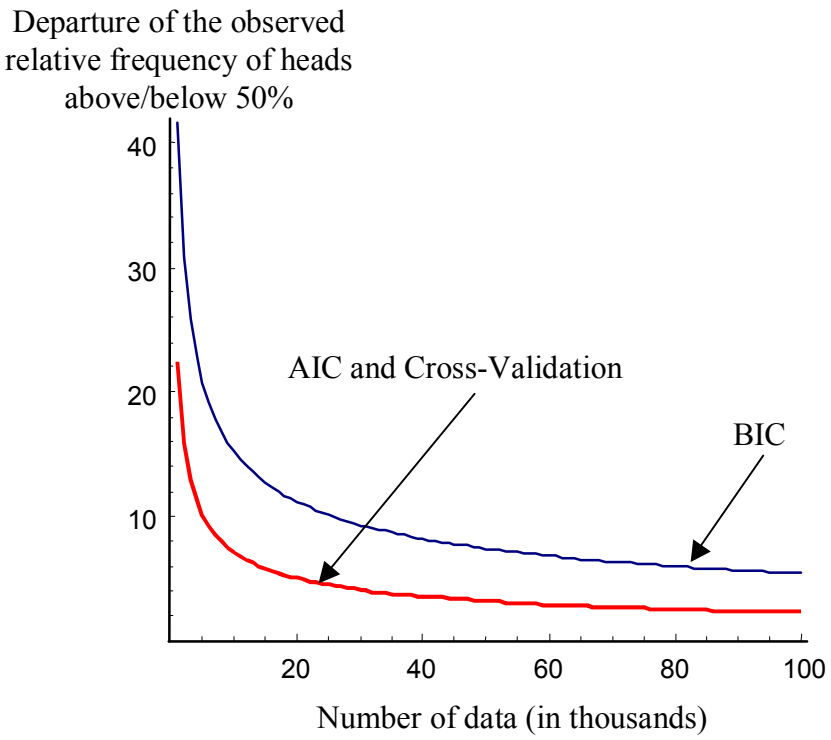

Figure 5.2: The critical point at which the null hypothesis is rejected in cross-validation, BIC, and AIC. Classical hypothesis testing would be between BIC and AIC.

Bayesians have responded to the conceptual difficulties facing classical statisticians by bringing in the prior probabilities of the competing hypotheses and their likelihoods. The posterior probability of a model is proportional to the product of the prior probability and the likelihood of the model. Therefore, the Bayesian method of comparing posterior probabilities appears to address the problem. Certainly, this approach does make a decision that depends on both of the competing models, but is it the best policy for comparing the predictive accuracy of competing models?

Perhaps Bayesians could argue like this: Truth is connected to predictive accuracy in the sense that there is no hypothesis that can be more predictively accurate than a true hypothesis, so to maximize the expected predictive accuracy of a model, we should maximize its probability. However, this argument is flawed. First, the premise is false. It is true that for a *maximally specific* hypothesis—one that gives precise values to all parameters—no hypothesis can be more accurate than the true hypothesis. However, this statement does not extend to models, which assert only that one of its hypotheses is true—models are very large disjunctions. Therefore, the predictive accuracy of a *model* is either undefined, or it depends either on the probabilistic weights given to its members, or it is identified with the predictive accuracy of the maximum likelihood hypothesis (if 'point' estimation is used). In either case, if the predictive accuracy is well defined, then the predictive accuracy of a true model will be less than the predictive accuracy of the true hypothesis. It also follows that the predictive accuracy of a false model can be higher than the predictive accuracy of a true model.

Second, even if the premise were true, the conclusion does not follow. Maximizing the probability of truth does not always maximize the expected predictive accuracy. To show this, suppose I predict the reading (plus or minus a second) on an atomic clock using my watch, which is 3 seconds fast. My predictive accuracy (suitably defined) is pretty good, but the probability that my prediction is true is zero. Contrast that to someone who makes the same prediction on the basis of a stopped clock. The probability of their prediction being true is higher than mine, yet their predictive accuracy is lousy.

Another incongruity of this Bayesian approach arises in the case of nested models, like the ones we are considering. As an independent example, consider a curve fitting example in which the model of all linear functions, LIN, is nested in the model of all parabolic functions, PAR, since all the members of LIN are contained in PAR. This can be seen by examining the equations: If the coefficient of the squared term in the equation for PAR is zero, then the equation reduces to the equation

for a straight line. Logically speaking, this nested relationship means that LIN logically entails PAR, in the sense that it is impossible for LIN to be true and PAR false. It is now a consequence of the axioms of probability that the LIN can never be more probable than PAR, and this is true for all probabilities, prior or posterior (Popper 1959, chapter 7). So, the Bayesian idea that we should select the model with the highest posterior probability leads to the conclusion that we should never choose LIN over PAR. In fact, we should never choose PAR over CUBE, where CUBE is the family of third degree polynomials, and so on. But if we are interested in predictive accuracy, there will be occasions on which we should choose LIN over PAR. Therefore, the Bayesian principle cannot serve the goal of predictive accuracy in this case.

Of course, Bayesians can simply refuse to consider this case. They might consider LIN versus PAR^-, where PAR^- is PAR minus LIN. Then the models are not nested, and the Bayesian criterion could lead to the choice of LIN over PAR^-. But it is puzzling why this difference should make a difference if we are interested in predictive accuracy, since the presence or absence of LIN nested in PAR makes no difference to any prediction, and *ipso facto*, no difference to the *accuracy* of any predictions. The failure of Bayesian principles to yield the same answer in both cases is a clear demonstration that their methods are not designed to maximize predictive accuracy. If they succeed in achieving this goal, then it is a lucky accident.

The goals of probable truth and predictive accuracy are clearly different, and it seems that predictive accuracy is the one that scientists care about most. Whenever parameter values are replaced by point estimates, there is zero chance of that specific value being the true one, yet scientists are not perturbed by this. Economists don't care whether their predictions of tomorrow's stock prices are *exactly* right; being close would still produce huge profits. Physicists don't care whether their current estimate of the speed of light is *exactly* true, so long as it has a high degree of accuracy. Biologists are not concerned if they fail to predict the exact corn yield of a new strain, so long as they are approximately right. If the probability of truth were something they cared about, then point estimation would be a puzzling practice. But if predictive accuracy is what scientists value, then their methodology makes sense.

This does not work as a criticism of *all* Bayesians. Decision-theoretic Bayesians could take predictive accuracy as their utility, and derive a criterion to maximize the expected predictive accuracy. This decision-theoretic approach is discussed in Young (1987), for example. However, the classical Bayesian approach is the most influential amongst

scientists, perhaps because it has led to the useable BIC criterion which appears to implement Occam's razor.[7]

A decision-theoretic Bayesianism that takes predictive accuracy as its utility still requires the use of prior probability distributions over propositions about the predictive accuracies of hypotheses. If we had such prior knowledge, then the Bayesian approach would make sense. But we don't. Another way of stating the criticism is that there are infinitely many Bayesian theories, and there is no way of deciding amongst them, besides using computer simulations, testing their success on real predictions, and mathematically analyzing the various criteria under a variety of assumptions. But this is just to revert to the Akaike approach, and one might wonder whether Bayesianism is anything more than the background machinery for generating criteria.

A counter-consideration is that Bayesian decision theory allows us to incorporate background information in decision-making. Certainly, when such information is available, it should be used. But Bayesians do not have a monopoly on background knowledge. It is not even true that the AIC criterion takes no account of background information, since it can be applied more globally when there is data relevant to the hypothesis that falls outside of the prediction problem at hand. For example, a model of stock market movement may take global economic parameters into account, and this may be done by considering a broader base of economic data. AIC requires that the relevance be built explicitly into the model, whereas Bayesians allow it to be represented in the prior probabilities. I believe that the background information is better built into the model, where it is publicly displayed and subjected to debate.

Cross-validation is a method widely used in learning algorithms in neural networks and in machine learning (e.g., Turney 1994). It is an interesting method because it appears to make no assumptions at all. The idea is that a curve is fitted to a subset of the observed data—often the whole data minus one data point. Such a subset of data is called a *calibrating data set*. The predictive accuracy of the fitted model is tested against the data point or points left out, which may be averaged over all possible calibrating data sets. Note that this method cannot be applied to a single specific curve, since the average fit for each data point in the set is

[7] The earliest reference to this idea I know is Rosenkrantz (1977), except he does not derive the BIC approximation, which was derived by Schwarz (1978). MacKay (1995) discusses the same version of Occam's razor in apparent ignorance of previous work. Cheeseman (1990) also discusses the classical Bayesian approach with even less sophistication and even fewer references.

just the fit with respect to the total data set, which reduces to the naïve empiricism of ML.

However, if the method is used to compare *models* rather than particular hypotheses, then it has different properties. Each calibrating data set produces a slightly different best fitting curve in the family and there will be a penalty for large, complex, families of curves because large families will tend to produce greater variation in the curve that best fits a calibrating data set (Turney 1990). This leads to an average fit that is poorer than the fit of the curve that best fits the total data set. There is no need to explicitly define simplicity or to quantify its effects on the stability of estimation; it is taken into account implicitly rather than explicitly. It is a remarkable fact that this simple method leads to approximately the same criterion of model selection as AIC in our simple coin tossing example (see figure 5.2). It is remarkable exactly because AIC factors in simplicity explicitly while cross validation does not. But perhaps it is not so surprising once we note that they are both designed with the same goal in mind – predictive accuracy.[8] Methods of cross validation are worthy of serious attention from scientists, either as a way of complementing other criteria or as an alternative criterion. I don't know which, but I believe that the Akaike framework provides the right tools for such an investigation.

This section has surveyed the variety of inference methods that can be applied to the easiest example imaginable. Very often the methods give similar results, but the *foundations* of those methods vary greatly. Nevertheless, they should all be considered seriously. The solution is to evaluate all of them within the Akaike framework (or some natural extension of it). As you can see, this has been an argument for the Akaike *framework*, and not the Akaike criterion (AIC).

4 Predictive accuracy as a goal of model selection

How should we define predictive accuracy? First, we need to distinguish between seen and unseen data. As a *goal*, we are interested in the prediction of unseen data, rather than the data used to construct the hypothesis. The seen data is the *means* by which we can forecast how well the hypothesis will predict unseen data.

However, any particular set of data may exhibit idiosyncrasies due to random fluctuations of observational error. If we took the goal to be the

[8] I have since learned that Stone (1977) proved that AIC is equivalent to leave-one-out cross-validation asymptotically for large samples, so the result I got is to be expected because I assumed the same conditions.

prediction of a *single* set of unseen data, then the goal is too hard in the sense that particular errors are impossible to predict, and in other cases the goal may be achieved by dumb luck. It is therefore customary to define predictive accuracy differently. The idea is that a predictively accurate curve is one that is as close as possible to the *trend*, or *regularity*, *behind* the data. The technical trick used to unpack that idea is to imagine many data sets generated repeatedly by that regularity (the true curve) and define the predictive accuracy of an arbitrary hypothesis as the average fit of the curve with respect to all such data sets. In that way no particular set of errors fluctuations are given undue emphasis. In the language of probability, predictive accuracy is the expected fit of data sets generated by the true probability distribution. The expected value is therefore objectively defined. It is not the subjective expectation that would appear in a Bayesian analysis of the problem. This point is worth examining in greater detail.

Consider a curve fitting example in which y is function of x. Define the *domain of prediction* in terms of a probability distribution defined over the independent variable, $p(x)$. This distribution will define the range of x-values over which unseen data sets are sampled. There is no claim that $p(x)$ is objective in the sense of representing an objective chance, or a propensity of some kind. But it is objectively given once the domain of prediction is fixed. There are now three cases to consider:

1. There is a true conditional probability density $p^*(y/x)$, which is an objective propensity. Since $p(x)$ is objective (given the domain of prediction), the joint distribution $p(x, y)$ is objective, because it is the product of the two.
2. The probability density $p(y/x)$ is an average over the propensities $p^*(y/x,z)$, where z refers to one or more variables that affect the value of y. In this case, one needs to specify the domain of prediction more finely. One needs to specify the probability distribution $p(x, z)$. Once $p(x, z)$ is fixed, $p(x, y)$ is determined by $p^*(y/x,z)$, and is again objective.
3. The independent variable x determines a unique, error free, value of y. This is the case of noise-free data. The true curve is defined by the value of y determined by each value of x. What this means is that all points generated by the $p(x, y)$ will lie exactly on the true curve. The distribution $p(y/x)$ is a Dirac delta function (zero for all values of y except for one value, such that it integrates to 1). The probability $p(x, y)$ is still objectively determined from $p(x)$, which defines the *domain of prediction*. Moreover, $p(x, y)$ allows for a statistical treatment of parameter estimation, so it fits into the Akaike framework.

Case (3) is important for it shows how a probabilistic treatment of parameter estimation may be grounded in a probabilistic definition of the domain of prediction. There is no need to assume that nature is probabilistic. The only exception to this is when a family of curves actually contains the true curve, for in that case, there can be no curve that fits the data better than the true curve, and the estimated parameter values are always the true ones, and there will be no variation from one data set to the next. In this case, the framework will not apply. I believe that this is not a severe limitation of the framework since it is plausible to suppose that it arises very rarely. Therefore, in general, *once the domain is fixed*, the probability of sets of data generated by the true distribution in this domain is objectively determined by the true distribution.

The relativization of predictive accuracy to a domain has meaningful consequences. In many cases, a scientist is interested in predictions in a domain different from the one in which the data are sampled. For example, in time series, the observed data is sampled from the past, but the predictions pertain to the future. In the Akaike framework, the default assumption is that the domain of prediction is the same as the domain in which the data are sampled. It is imagined, in other words, that new data are re-sampled from the past. If the time series is stationary, then the past is effectively the same as the future. But in general this is not true, in which case it is an open question whether the standard model selection criteria apply (for discussion, see Forster, 2000). It is an advantage of the Akaike framework that such issues are raised explicitly.

Predictive accuracy is the expected fit of unseen data in a domain, but this definition is not precise until the notion of fit is precise. A common choice is the sum of squared deviations made famous by the method of least squares. However, squared deviations do not make sense in every example. For instance, when probabilistic hypotheses are devised to explain the relative frequency of heads in a hundred tosses by the fairness of the coin, the hypothesis does not fit the data in the sense of squared deviations. In these cases, an appropriate measure of fit is the likelihood of the hypothesis relative to the data (the probability of the data given the hypothesis).

However, does the likelihood measure apply to all cases? In order for the hypothesis to have a likelihood, we need the hypothesis to be probabilistic. In curve fitting, we do that by associating each hypothesis with an error distribution. In that way, the fit of a hypothesis with any data set is determined by the hypothesis itself, and is therefore an entirely objective feature of the hypothesis. When the error distribution is normal (Gaussian), then the log-likelihood is proportional to the sum a squared

deviations. When the error distribution is not normal, then I take the log-likelihood to be the more fundamental measure of fit.

Before we can state the goal of curve fitting, or model selection in general, we need a clear definition of the predictive accuracy of an arbitrary hypothesis. We are interested in the performance of a hypothesis in predicting data randomly generated by the true hypothesis. We have already explained that this can be measured by the expected log-likelihood of newly generated data. But we do not want this goal to depend on the number of data n because we do not really care whether the unseen data set is of size n or not. It is convenient to think of the unseen data sets as the same size as the seen data set, but it is surely not necessary. Unfortunately, the log-likelihood relative to n data increases as n increases. So, in order that the goal not depend on n we need to define the predictive accuracy of a hypothesis h as the expected *per datum* log-likelihood of h relative to data sets of size n. Under this definition, the predictive accuracy of a fixed hypothesis will be the same no matter what the value of n, at least in the special case in which the data are probabilistically independent and identically distributed.[9]

Formally, we define the predictive accuracy of an arbitrary hypothesis h as follows. Let E^* be the expected value with respect to the objective probability distribution $p^*(x, y)$, and let $Data(n)$ be an arbitrary data set of n data randomly generated by $p^*(x, y)$. Then the predictive accuracy of h, denoted by $A(h)$, is defined as

$$A(h) = \frac{1}{n} E^*\left[\log \text{likelihood}\left(Data(n)\right)\right],$$

where E^* denotes the expected value relative to the distribution $p^*(x, y)$. The goal of curve fitting, and model selection in general, is now well defined once we say what the h's are.

Models are families of hypotheses. Note that, while each member of the family has an objective likelihood, the model itself does not. Technically speaking, the likelihood of a model is an average likelihood of its members, but the average can only be defined relative to a *subjective* distribution over its members. So, the predictive accuracy of a model is undefined (except when there is only one member in the model). [10]

Model selection proceeds in two steps. The first step is to select a model, and the second step is to select a particular hypothesis from the

---

[9] For in that case, the expected log-likelihood is n times the expected log-likelihood of each datum.

[10] There are ways of defining model accuracy (Forster and Sober, 1994), but I will not do so here because it complicates the issue unnecessarily.

model. The second step is well known in statistics as the *estimation* of parameters. It can only use the seen data, and I will assume that it is the method of maximum likelihood estimation. Maximizing likelihood is the same as maximizing the log-likelihood, which selects the hypothesis that best fits the seen data. If an arbitrary member of the model is identified by a vector of parameter values, denoted by θ, then $\hat{\theta}$ denotes the member of the model that provides the best fit with the data. Each model produces a different best fitting hypothesis, so *the goal of model selection is to maximize the predictive accuracy of the best fitting cases drawn from rival models*. This is the first complete statement of the goal of model selection.

In science, competing models are often constrained by a single background theory. For example, Newton first investigated a model of the earth as a uniformly spherical ball, but found that none of the trajectories of the earth's motion derived from this assumption fit the known facts about the precession of the earth's equinoxes. He then complicated the model by allowing for the fact that the earth's globe bulges at the equator and found that the more complicated model was able to fit the equinox data. The two models are Newtonian models of the motion. However, there is no reason why Newtonian and Einsteinian models cannot compete with each other in the same way (Forster, 2000a). In fact, we may suppose that there are no background theories. All that is required is that the models share the common goal of predicting the same data.

In the model selection literature, the kind of selection problem commonly considered is where the competing models form a nested hierarchy, like the hierarchy of *k*-degree polynomials. Each model in the hierarchy has a unique dimension *k*, and the sequence of best fitting members is denoted by $\hat{\theta}_k$. The *predictive accuracy* of $\hat{\theta}_k$ is denoted by $A(\hat{\theta}_k)$. This value does not depend on the number of data, *n*. In fact, the predictive accuracy is not a property of the *seen* data at all—except in the sense that $\hat{\theta}_k$ is a function of the seen data. The aim of model selection in this context is to choose the value of *k* for which $A(\hat{\theta}_k)$ has the highest value in the hierarchy.

Note that $\hat{\theta}_k$ will not be the predictively most accurate hypothesis in the model *k*. $\hat{\theta}_k$ fits the *seen* data the best, but it will not, in general, provide the best average fit of unseen data. The random fluctuations in any data set will lead us away from the predictively most accurate hypothesis in the family, which is denoted by θ^*_k. However, from an epistemological point of view, we don't know the hypothesis θ^*_k, so we have no choice but to select $\hat{\theta}_k$ in the second step of curve fitting. So, our goal is to maximize $A(\hat{\theta}_k)$, and not $A(\theta^*_k)$. In fact, the maximization of $A(\theta^*_k)$ would lead to the absurd result that we should select the most

complex model in the hierarchy, since $A(\theta_k^*)$ can never decrease as k increases.

While I am on the subject of "what the goal is not", let me note that getting the value of k "right" is not the goal either. It is true that in selecting a model in the hierarchy we also select of value of k. And in the special case in which $A(\theta_k^*)$ stops increasing at some point in the hierarchy, then that point in the hierarchy can be characterized in terms of a value of k, which we may denote as k^*. In other words, k^* is the smallest dimensional family in the hierarchy that contains the most predictively accurate hypothesis to occur anywhere in the hierarchy (if the true hypothesis is in the hierarchy, then k^* denotes the smallest true model). But model selection aims at selecting the best hypothesis $\hat{\theta}_k$, and this may not necessarily occur when $k = k^*$. After all, $\hat{\theta}_k$ could be closer to the optimal hypothesis when k is greater than k^* since the optimal hypothesis is also contained in those higher dimensional models. I will return to this point in section 7, where I defend AIC against the common charge that it is not statistically consistent.

5 A 'normality' assumption and the geometry of parameter space

There is a very elegant geometrical interpretation of predictive accuracy in the special case in which parameter estimates conform to a probabilistic description that I shall refer to as the *normality condition*. It is good to separate the condition from the question about what justifies the assumption. I will concentrate on its consequences and refer the inter-ested reader to Cramér (1946, chs. 32-4) for the theory behind the con-dition.

Consider the problem of predicting y from x in a specified domain of prediction. As discussed in the previous section, there is a 'true' distribution $p(x, y)$, which determines how the estimated parameter values in our models vary from one possible data set to the next. We can imagine that a large dimensional model K contains the true distribution, even though the model K is too high in the hierarchy to be considered in practice. In fact, we could define the hierarchy in such a way that it contains the true distribution, even though every model considered in practice will be false. So, let the point θ^* in the model K represent the true distribution. The maximum likelihood hypothesis in K is $\hat{\theta}_K$, which we may denote more simply by $\hat{\theta}$. There are now two separate functions over parameter space to consider. The first is the probability density for $\hat{\theta}$ over the parameter space, which we could denote by $f(\theta)$. The second is the likelihood function, $L(Data|\theta)$, which records the probability of the data given any particular point in parameter space. Both are defined over points in parameter space, but each has a very different meaning. The normality

assumption describes the nature of each function, and then connects them together.

1. The distribution $f(\theta)$ is a multivariate normal distribution centered at the point $\theta*$ with a bell-shaped distribution around that point whose spread is determined by the covariance matrix $\Sigma*$. The covariance matrix $\Sigma*$ is proportional to $1/n$, where n is the sample size (that is, the distribution is more peaked as n increases).
2. The likelihood function $L(Data|\theta)$ is *proportional* to a multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix Σ.[11] As n increases, $\log L(Data|\theta)$ increases proportionally to n, so that Σ is proportional to $1/n$.
3. Σ is equal to $\Sigma*$.

The exact truth of condition (3) is an unnecessarily strong condition, but its implications are simple and clear. Combined with (1) and (2), it implies that log-likelihoods and predictive accuracies vary according to the same metric; namely squared distances in parameter space. More precisely, there is a transformation of parameter space in which Σ is equal to I/n, where I is the identity matrix and n is the sample size. The per-datum log-likelihood of an arbitrary point θ is equal to the per-datum log-likelihood of $\hat{\theta}$ minus $\frac{1}{2}\, n|\theta-\hat{\theta}|^2$, where $|\theta-\hat{\theta}|^2$ is the square of the Euclidean distance between θ and $\hat{\theta}$ in the transformed parameter space. Moreover, the predictive accuracy of the same point θ is equal to the predictive accuracy of $\theta*$ minus $\frac{1}{2}|\theta-\theta*|^2$. Since $\hat{\theta}$ is a multivariate normal random variable distributed around $\theta*$ with covariance matrix I/n, $\sqrt{n}(\hat{\theta}-\theta*)$ is a multivariate normal random variable with mean zero and covariance matrix I. It follows that $n|\hat{\theta}-\theta*|^2$ is a chi-squared random variable with K degrees of freedom, and that $|\hat{\theta}-\theta*|^2$ is a random variable with mean K/n.

Similar conclusions apply to lower models in the hierarchy of models, assuming that they are represented as subspaces of the K-dimensional parameter space. Without loss of generality, we may suppose that the parameterization is chosen so that an arbitrary member of the model of dimension k is $(\theta_1,\theta_2,\ldots,\theta_k,0,\ldots,0)$, where the last $K-k$ parameter values are 0. The predictively most accurate member of model k, denoted θ^*_k, is the projection of $\theta*$ onto the subspace and $\hat{\theta}_k$ is the projection of $\hat{\theta}$ onto the same subspace.

[11] The likelihood function is not a probability function because it does not integrate to 1.

We may now use the normality assumption to understand the relationship between $A(\hat{\theta}_k)$ and $A(\theta^*_k)$. First note that θ^*_k is fixed, so $A(\theta^*_k)$ is a constant. On the other hand, $\hat{\theta}_k$ varies randomly around θ^*_k according to a k-variate normal distribution centered at θ^*_k. We know that $A(\theta^*_k)$ is greater than $A(\hat{\theta}_k)$, since $A(\theta^*_k)$ is the maximum by definition. Moreover, $A(\hat{\theta}_k)$ is less than $A(\theta^*_k)$ by an amount proportional to the squared distance between $\hat{\theta}_k$ and θ^*_k in the k-dimensional subspace. Therefore,

$$A\left(\hat{\theta}_k\right) = A\left(\theta^*_k\right) - \frac{\chi^2_k}{2n},$$

where χ^2_k is a chi-squared random variable of k degrees of freedom. It is a well known property of the chi-squared distribution that χ^2_k has a mean, or expected value, equal to k. That leads to the relationship between the bottom two plots in figure 5.3. Note that while $A(\theta^*_k)$ can never decrease (because the best in k +1 is at least as good as the best in k), it is also bounded above (since it can never exceed the predictive accuracy of the true hypothesis). This implies that the lower plot of $A(\hat{\theta}_k)$ as a function of k will eventually reach a maximum value and then decrease as k increases. *Hence model selection aims at a model of finite dimension*, even though the predictive accuracy $A(\theta^*_k)$ of the best hypothesis in the model will always increase as we move up the hierarchy (or, at least, it can never decrease). The distinction between $\hat{\theta}_k$ around θ^*_k is crucial to our under-standing of model selection methodology.

As an example, suppose that a Fourier series is used to approximate a function. Adding new terms in the series can improve the potential accu-

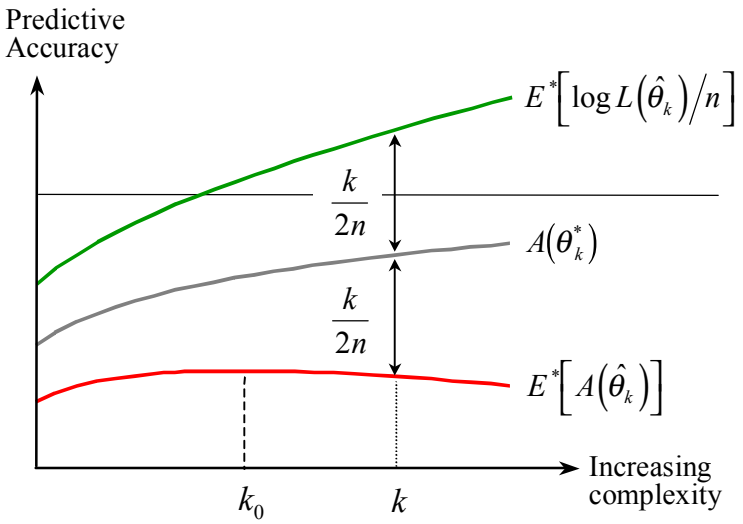

Figure 5.3 The behavior of various quantities in a nested hierarchy of models.

racy of fit indefinitely; however, the problem with overfitting is overwhelming when there are too many parameters to estimate. An historical illustration of this phenomenon is the case of ancient Ptolemaic astronomy, where adding epicycles can always improve the approximation to the planetary trajectories, yet adding epicycles beyond a certain point does not improve prediction in practice. The present framework explains this fact.

Denote the k for which $A(\hat{\theta}_k)$ is maximum as k_0. The value of k_0 depends on the estimated parameter values (on the $\hat{\theta}_k$), which depends on the actual data at hand. There will be a tendency for k_0 to increase as the number of seen data increases. This is observed in figure 5.3. The middle curve (the curve for $A(\theta^*_k)$) is entirely independent of the seen data, but the mean curve for $A(\hat{\theta}_k)$ hangs below it by a distance k/n. As n increases, it will hang closer to the middle curve, and so its maximum point will move to the right. Therefore a richer data set justifies an increase in complexity—something that is intuitively plausible on the idea that more data allow for the more accurate estimation of complex regularities. For example, a parabolic trend in a small set of data is more readily explained away as an accidental deviation from a linear regularity, while the same parabolic trend in a large number of data is not so easily dismissed.

The relationship between $A(\hat{\theta}_k)$ and $A(\theta^*_k)$ exhibits what is commonly called the bias/variance tradeoff (Geman *et al*, 1992). Let me first explain what is meant by the terms 'bias' and 'variance'. *Model bias* is the amount that the best case in the model is less predictively accurate than the true hypothesis. By 'best case', I mean the hypothesis in the model with the highest predictive accuracy, not the best fitting case. In other words, model bias is the difference between $A(\theta^*_k)$ and the predictive accuracy of the true hypothesis. As $A(\theta^*_k)$ increases (see figure 5.3), it gets closer to the best possible value, so the model bias decreases. Of course, we do not know which hypothesis is the most predictively accurate. So, model bias is not something that models wear on their sleeves. Nevertheless, we can make some reasonable guesses about model bias. For example, the model that says that planets orbit the sun on square paths is a very biased model because the best possible square orbit is not going fit the true orbit very well. At the other extreme, any model that contains the true hypothesis has zero bias. In nested models, the bias is less for more complex hypotheses.

The variance, on the other hand, refers to the squared distance of the best fitting hypothesis $\hat{\theta}_k$ from the most predictively accurate hypothesis θ^*_k. It is governed by the chi-squared variable in the previous equation. The variance of estimated hypothesis from the best case favors simplicity.

In conclusion, complexity is good for reduction of bias, whereas simplicity reduces the tendency to overfit. The optimum model is the one that makes the best tradeoff between these two factors. The bias/variance dilemma refers to the fact that as we go up in a hierarchy of nested models, the bias decreases, but the expected variance increases. A model selection criterion *aims* at the best trade off between bias and variance, but neither bias nor variance is known, so this theoretical insight does not lead directly to any criteria. It tells us what we *aim* to do, not how to do it.

An interesting special case is where a family k_1 at some point in the hierarchy already contains the true hypothesis. In that case, there is no decrease in bias past that point. But going higher in the hierarchy leads to some loss, because the additional parameters will produce a tendency to overfit. This means that going from model k_1 to k_1+1 has no *expected* advantages in terms of predictive accuracy. So, it would be best to stop in this case. However, this fact does not lead to a criterion either, unless we know that the k_1 model is true. If we already knew that, we would need no criterion.

6 Comparing selection criteria

In this section I will compare the performance of AIC and BIC in the selection of two nested models differing by one adjustable parameter in contexts in which the normality assumption holds. While the normality condition will not hold for many examples, it is a central case in statistics because the Central Limit theorems show that it holds in a wide variety of circumstances (see Cramér 1946, chapters 32 and 33). More importantly, the arguments leveled against AIC in favor of BIC are framed in this context. So, my analysis will enable us to analyze those arguments in the next section.

The normality assumption also determines the stochastic behavior of the log-likelihood of the seen data, and we can exploit this knowledge to obtain a criterion of model selection. Let $\log L(\hat{\theta}_k)$ be the log-likelihood of $\hat{\theta}_k$ relative to the seen data. If $\hat{\theta}_k$ is a random variable, then $\log L(\hat{\theta}_k)/n$ is also a random variable. Its relationship to $A(\hat{\theta}_k)$ is also displayed in figure 5.3: $\log L(\hat{\theta}_k)/n$ is, on average, higher than $A(\hat{\theta}_k)$ by a value of k/n (modulo a constant, which doesn't matter because it cancels out when we compare models). So, an unbiased[12] estimate of the predictive accuracy

[12] An estimator of a quantity (in this case an estimator of predictive accuracy) is *unbiased* if the expected value of the estimate is equal to the quantity being estimated. This sense of 'bias' has nothing to do with model bias.

of the best fitting curve in any model is given by $\log L(\hat{\theta}_k)/n - k/n$. If we judge the predictive accuracies of competing models by this estimate, then we should choose the model with the highest value of $\log L(\hat{\theta}_k)/n - k/n$. *This is the Akaike information criterion* (AIC).

The BIC criterion (Schwarz 1978) maximizes the quantity $\log L(\hat{\theta}_k)/n - k\log[n]/2n$, giving a greater weight to simplicity by a factor of $\log[n]/2$. This factor is quite large for large *n*, and has the effect of selecting a simpler model than AIC. As we shall see, this an advantage in some cases and a disadvantage in other cases. There is an easy way of understanding why this is so. Consider two very extreme selection rules: The first I shall call the Always-Simple rule because it always selects the simpler model no matter what the data say. Philosophers will think of this rule as an extreme form a rationalism. The second rule goes to the opposite extreme and always selects the more complex model no matter what the data, which I call the Always-Complex rule. In the case of nested models, the Always-Complex rule always selects the model with the best-fitting specification and is therefore equivalent to a maximum likelihood (ML) rule. It is also a rule that philosophers might describe as a naïve form of empiricism, since it gives no weight to simplicity. BIC and AIC are between these two rules: BIC erring towards the Always-Simple side of the spectrum, while AIC is closer to the ML rule.

Consider any two nested models that differ by one adjustable parameter, and assume that normality conditions apply approximately. Note we need *not* assume that the true hypothesis is in either model (although the normality conditions are easier to satisfy when it is). The simple example in section 3 is of this type, but the results here are far more general. The only circumstance that affects the expected performance of the rules in this context is the difference in the model biases between the two models. The model bias, remember, is defined as the amount that the most predictively accurate member of the family is less predictively accurate than the true hypothesis. Under conditions of normality, the difference in model bias is proportional to the squared distance between the most accurate members of each model. In our easy example, this is proportional to $(\theta^* - ½)^2$. Note that the Always-Simple rule selects the hypothesis $\theta = ½$ and the ML rule selects the hypothesis $\theta = \hat{\theta}$, where $\hat{\theta}$ is the maximum likelihood value of the statistic (the relative frequency of 'heads up' in our example). Under the normality assumption the predictive accuracies of these hypotheses are proportional to the squared distance to θ^* in parameter space. That is,

$$A(\theta = ½) = -const.(\theta^* - ½)^2 \text{ and } A(\theta = \hat{\theta}) = -const.(\theta^* - \hat{\theta})^2.$$

Therefore, the null hypothesis $\hat{\theta} = ½$ is a better choice than the alternative $\theta = \hat{\theta}$ if and only if ½ is closer to θ^* than $\hat{\theta}$ is to θ^*. Notice that the first distance is proportional to the complex model's advantage in bias, while the expected value of the second squared distance is just the variance of the estimator $\hat{\theta}$. Therefore, the ML rule is more successful than the Always-Simple rule, on average, if and only if, the advantage in model bias outweighs the increased variance, or expected overfitting, that comes with complexity. This is the bias/variance dilemma.

A simple corollary to this result is that the two extreme rules, Always-Simple and Always-Complex, enjoy the same success (on average) if the model bias advantage exactly balances the expected loss due to variance. It is remarkable that two diametrically opposed methods can be equally successful in some circumstances. In fact, we may expect that all rules, like BIC and AIC, will perform equivalently when the bias difference is equal to the variance difference.

The situation in which the bias and variance differences are equal is a *neutral point* between two kinds of extremes—at one end of the spectrum the variance is the dominant factor, and at the other extreme, the bias difference is the overriding consideration. In the first case simplicity is the important factor, while in the second case goodness of fit is the important criterion. So, when the model bias difference is less than the expected difference in variance, we may expect BIC to perform better since it gives greater weight to simplicity. And when the model bias is greater than the variance, we may expect AIC to perform better than BIC, though neither will do better than ML.

These facts are confirmed by the results of computer computations shown in figure 5.4. In that graph, the expected gain in predictive accuracy, or what amounts to the same thing, the gain in expected predictive accuracy, is plotted against the model bias difference between the two models in question. Higher is better. The expected performance of naïve empiricist method of ML is taken as a baseline, so the gain (or loss if the gain is negative) is relative to ML. The performance is therefore computed as follows. Imagine that a data set of size *n* is randomly generated by the true distribution in a domain of prediction. The method in question then selects its hypothesis. If it is the same as the ML hypothesis, then the gain is zero. If it chooses the simpler model, then the gain will be positive if the resulting hypothesis is predictively more accurate, and negative if it is less accurate, on average. The overall performance of the method is calculated as its expected gain. The expectation is calculated by weighting each possible case by the relative frequency of its occurrence as determined by the true distribution.

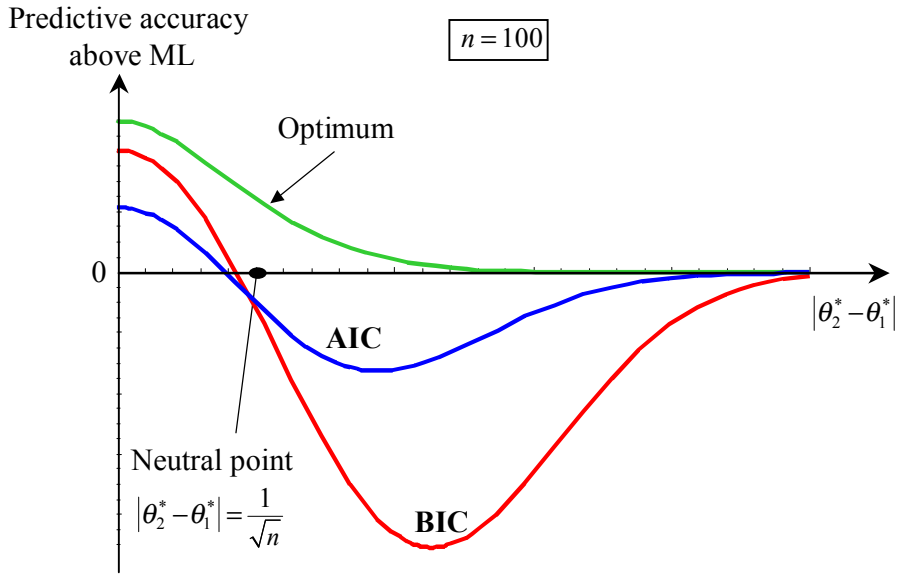

Figure 5.4 At the neutral point, the advantage of bias had by the complex model balances its disadvantage in variance, and all selection rules result in roughly the same expected predictive accuracy. In situations where the difference in model bias is smaller, methods that favor simplicity do better, like BIC, while in all other contexts, it is better to give less weight to simplicity, in which case AIC does better than BIC. The plot looks the same for a very wide variety of values of n.

The performance of any method will depend on the difference in bias between the two models. The horizontal axis is scaled according to raw (un-squared) distances in parameter space, so it is actually represents the square root of the model bias differences.[13] On the far left is the special case in which both models have the same bias. That is the point at which there is no advantage in complexity. To the right are points for which the model bias is decreased in the more complex model. For nested models, the bias factor will always favor the more complex model, although this is not always true for non-nested models.

The rest of the context is held fixed: The models differ by one adjustable parameter, the number of seen data is fixed, and normality conditions hold. Remember that the seen data set itself is not held fixed. We are interested in the expected performance averaged over all possible seen data sets of size n, where the expectation is determined by the true distribution.

The curve labeled the 'optimum rule' in figure 4 records the perfor-

[13] If it were scaled by the squared distances, then the results would look even less favorable to the BIC criterion.

mance of the following 'perfect' method of selection: of the two hypothesis, choose the one that is the most predictively accurate. Sometimes the simpler model will 'win', sometimes the more complex model will 'win'. In the cases in which the simpler model is chosen, the policy is doing the opposite from the ML method. This 'policy' does better than ML when the model bias gain is relatively small, which reflects the fact that the decreased overfitting outweighs the loss in model bias. But when the model bias advantage of complex models is large enough, the complex model is almost always doing better in spite of its greater tendency to overfit. Note that the optimum rule cannot be implemented in practice, for it supposes that we know the predictive accuracies of the hypotheses in question. Of course, we do not know this. 'Real' methods can only make use of things we know, like the number of adjustable parameters, the number of seen data, and the fit with seen data. The optimum curve is shown on the graph because it marks the absolute upper bound in performance for any real criterion.

BIC manages to meet that optimum for the special case (on the far left in Figure 4) in which both models are equally biased. In our easy example, this corresponds to the case in which the null hypothesis is actually true ($\theta^* = ½$). If we knew this were the case, then we would want to choose the null hypothesis no matter what the data are, which is to say that the Always-Simple rule is also optimum in this situation. It is hardly surprising that both these rules do better than AIC in this situation.

Nevertheless, this situation may be relevant to scientific research. Raftery (1994) argues that this situation is likely to arise in regression problems in which scientists consider many possible independent variables when few, if any, are truly relevant to the dependent variable. In an extreme case we can imagine that a set of 51 variables are all probabilistically independent. Pick one as the depend variable and consider all models that take this variable to be a linear function of some proper subset of the remaining variables. Since the coefficients of each term in the equation can be zero, all of the models contain the true hypothesis (in which all the coefficients are zero). Therefore all the models are unbiased (in fact, they are all true). That means that complex models lose by their increased tendency to overfit, and have no compensating gains in bias. For instance, in comparing two nested models in which one adds a single independent variable, AIC will incorrectly add the variable 15.7% of the time no matter how many data we collect. BIC will make this mistake less often, and the frequency of the mistake diminishes to zero as we collect more data.

While AIC is making a mistake in this situation, the mistake is not as bad as it sounds. The goal is to maximize predictive accuracy, and the severity of the mistake is measured by the loss in predictive accuracy. If

the estimated value of the coefficient of the added variable is close to zero, then the loss in predictive accuracy may be very small. Even the extreme case of adopting the maximum likelihood rule (ML), which adds all 50 variables, the loss in predictive accuracy due to overfitting is equal to 50/*n*, on average, which diminishes as *n* increases.[14] AIC will tend to add about 8 variables, instead of 50, although the loss will be more than 8/*n* because it will add the variables with the larger estimated coefficients. The plot in Figure 4 suggests that the loss is around 28/*n*. For smaller *n*, this may be quite a large loss, but notice that the loss tends to zero as *n* increases, despite that fact that the proportion of wrongly added variables does not tend to zero. That is why it is important to be clear about the goal (I will return to this point in the next section).

In the plot in figure 5.4, $n = 100$. But, surprisingly, the plots look the same for a wide variety of values I tested, from $n = 100$, and up. Again, the reason that the *relative* performance of BIC and AIC does not change much is because of the fact that the relative *cost* of each BIC mistake goes up even though the frequency of BIC mistakes diminishes for BIC. Note that the *absolute* cost, in terms of predictive accuracy, decreases to zero for both methods as *n* tends to infinity.

Before leaving the special case, it is important to emphasize that scientists do not *know* that they are in such a situation. If they did know, there would be no need for any method of model selection—just pick the simplest model. It is precisely because the context is unknown that scien-tists want to use a selection rule. So, it would be wrong to prefer BIC solely on the basis of what happens in this special case.

The raison d'être of model selection is the possibility of facing the situations represented further to the right on the *x*-axis in Figure 4. There we quickly approach the neutral point at which all 'real' methods perform approximately the same. This point occurs when the model bias difference equals the variance of the true distribution (of the parameter estimator). With the units we have chosen, this occurs at the point marked $1/\sqrt{n}$. At points of greater difference in model bias, the fortunes of BIC and AIC change dramatically, and at model bias differences corresponding to about 3 standard deviations, BIC is paying a huge price for weighing simplicity so heavily.

In the case illustrated, the competing models differ by just one adjustable parameter $(\Delta k = 1)$. In other computer computations, I have found that BIC has an even greater disadvantage on the right-hand side of the

[14] This is because the maximum likelihood hypothesis is, on average, a (squared) distance of 1/*n* from the optimum hypothesis, θ^* (see figure 5.4). (This depends on an appropriate scaling of distances in parameter space.) The loss is then multiplied for each variable.

neutral point, while its advantage over AIC on the left is less. The near optimality of BIC in one case exposes us to considerable risk in other contexts.

It is interesting to consider what happens when the number of seen data, n, increases. I have defined model bias in a way that does not depend on n, so the point on the x-axis in Figure 4 that represents the context we are in does not change as n changes. As n increases, the relative shapes of the curves do not change, but they shrink in size. That is, the heights above and below the x-axis get smaller inversely proportionally to n, *and the neutral point moves to the left*. If we imagine that the graph is magnified as it shrinks, so it appears the same size to us, then the only change is that the point on the x-axis that represents the current context moves to the right. So, what happens if we steadily increase the number of seen data over time? We start out at an initial value of n, call it n_0. Then we collect more data, and n increases. At the beginning, we are either to the left of the neutral point or we are not. If we start at the left, then BIC will be better than AIC initially. But as the data number increases, we *must* move through the region in which BIC is performing poorly. If we do not start out to the left of the neutral point, then AIC is never worse than BIC. So, no matter what happens, we are exposed to a case in which BIC is worse than AIC as the sample size increases. In the limit as n tends to infinity, all methods approximate the optimal curve. So, the risks associated with BIC appear at intermediate values of n. Analyses that look only at the behavior of the methods for asymptotically large values of n will overlook this weakness of BIC at intermediate sample sizes.

The analysis of this section has looked at the comparison of two fixed nested models. These results do not extend straightforwardly to the case of selecting models in a *hierarchy* of nested models (some remarks will address this in the next section). However, the special case considered here does substantiate my thesis that BIC pays a price for its near optimality in one special case.

7 The charge that AIC is inconsistent

It is frequently alleged that AIC is inconsistent,[15] while BIC is not, thereby suggesting that BIC performs better in the limit of large n. This allegation is repeated in many publications, and in so many con-

[15] Philosophers unfamiliar with statistical terminology should note that this does *not* refer to *logical* inconsistency. Rather, an estimator is statistically *consistent* if it converges in probability to the true value of what it is trying to estimate (the target value).

versations, that I am unable to document all of them. I will pick on just one example. Keuzenkamp and McAleer (1995, page 9) state that AIC "fails to give a consistent estimate of k," which they attribute to Rissanen (1987, page 92) and Schwarz (1978). Bozdogan (1987) takes the criticism to heart, and derives an extension of AIC that is consistent in this sense. My conclusion will be that there is no sensible charge to answer, and so there is no need to modify AIC (at least, not for this reason). An immediate corollary is that all the competing criteria are consistent in the relevant sense. In any case, even if it did turn out unfavorably for AIC, it would be wrong to place too much emphasis on what happens in the long term, when scientists are only interested in finite data.[16]

There are actually many different questions that can be asked about the consistency of AIC. The first is whether AIC is a consistent method of *maximizing* predictive accuracy in the sense of converging on the hypothesis with the greatest predictive accuracy in the large sample limit. The second is whether AIC is consistent estimator of predictive accuracy, which is a subtlety different question from the first. And the third is whether AIC converges to the smallest true *model* in a nested hierarchy of models. The answer to the first two questions will be yes, AIC is consistent in this sense while the answer to the third is no, AIC is not consistent in this sense, but this fact does not limit its ability to achieve its goal. Here are the details.

Whatever it means to 'estimate k', it is certainly not what AIC was *designed* to estimate. The goal defined by Akaike (1973) was to estimate predictive accuracy. Because Akaike is the author of this approach, the charge that AIC is inconsistent might be read by many observers as saying that AIC is an inconsistent estimate *of predictive accuracy*. I will begin by showing that this charge of inconsistency is false, and then return to the quoted charge.

Akaike's own criterion minimizes the quantity $-2(\log L(\hat{\theta}_k) - k)$, which estimates $-2nA(\hat{\theta}_k)$. But note that this is a strange thing to estimate, since it depends on the number of seen data, n. It is like estimating the *sum* of heights of n people drawn from a population. The target value would be $n\mu$, where μ is the mean height in the population. Rather, the target *should* be a feature of the population alone, namely μ. To proceed otherwise is to mix up the *means* to the goal, which *is* a function of n, and the goal itself (which is not a function of n). So, the correct procedure is to use the sample mean, $\overline{x}$, to estimate μ, and this is a consistent estimate.

[16] See Sober (1988) for a response to the inconsistency of likelihood estimation in some situations, and Forster (1995, section 3) for a critique of the Bayesian idea that priors are harmless because they are 'washed out' in the long run.

Now suppose we were to use $n\bar{x}$ to estimate $n\mu$. Then of course the estimator would be inconsistent because the error of estimation grows with increasing n. This is hardly surprising when the target value keeps growing. The correct response to this problem would be to say, as everyone does, that $\bar{x}$ is a consistent estimate of μ. Surprisingly, this is exactly the situation with respect to AIC. AIC, in Akaike's formulation, *is* an inconsistent estimate because its target value grows with n. Akaike (1973, 1974, 1977, 1985) sets up the problem in a conceptually muddled way.

The correct response to the 'problem' is to divide the estimator and target by n, so that the target does not depend on the sample size. This is exactly what I have done here, and what Forster and Sober (1994) were careful to do when they introduced the term 'predictive accuracy' to represent what the AIC criterion aimed to estimate (Akaike does not use this term). AIC does provide a consistent estimate of predictive accuracy when it is properly defined.

Now, let us return to the earlier charge of inconsistency. When there is talk of 'estimating k' the discussion is typically being restricted to the context of a nested hierarchy of models. Here there are two cases to consider. The first is the case in which the true hypothesis appears somewhere in the hierarchy, while in the second it does not. Let me consider them in turn.

In the former case, the true hypothesis will first appear in a model of dimension k^*, *and in every model higher in the hierarchy*. When one talks of estimating k, one is treating the value of k determined by the selected model as an estimate of k^*. But why should it be desirable that k be as close as possible to k^*? In general it is not desirable. For example, consider the hierarchy of nested polynomials and suppose that the true curve is a parabola (i.e., it is in PAR). If the data is sampled from a relatively narrow region in which the curve is approximately linear (which is to say that there is not much to gain by going from LIN to PAR), then for even quite large values of n, it may be best to select LIN over PAR, and better than any other family of polynomials higher in the hierarchy. Philosophically speaking, this is the interesting case in which a false model is better than a true model. However, for sufficiently high values of n, this will change, and PAR will be the better choice (because the problem of overfitting is then far less). Again, this is an example in which asymptotic results are potentially misleading because they do not extend to intermediate data sizes.

Let us consider the case in which n is large enough to make PAR the best choice (again in that case in which the true curve is in PAR). Now AIC will eventually overshoot PAR. Asymptotically, AIC will not converge on PAR (Bozdogan 1987; Speed and Yu, 1991). This is the basis

for the quoted charge that AIC is inconsistent. But how serious are the consequences of this fact? After all, AIC does successfully converge on the true hypothesis!

One might object: "But how can it converge on the true parabola if it doesn't converge on PAR?" But the objector is forgetting that the true curve is also in all the models higher in the hierarchy because the models are nested. So, there is no need for the curve favored by AIC to be in PAR in order for it to converge to a member of PAR. The fact that I am right about this is seen independently from the fact that the maximum likelihood estimates of the parameter values converge to their true values. This implies that even ML converges on the true hypothesis, and certainly ML overshoots k^* far more than AIC!

In the second case the true hypothesis does not appear anywhere in the hierarchy of models. In this case the model bias will keep decreasing as we move up the hierarchy, and there will never be a point at which it stops decreasing. The situation is depicted in figure 5.3. For each n, there will be an optimum model k_0, and this value will keep increasing as n increases. The situation here is complicated to analyse, but one thing is clear. There is no *universally valid* theorem that shows that BIC does better than AIC. Their relative performances will depend on the model biases in the hierarchy in a complicated way.

In both cases, the optimum model moves up the hierarchy as n increases. In the first case, it reaches a maximum value k^*, and then stops. The crucial point is that in all cases, the error of AIC (as an estimate of predictive accuracy) converges to zero as n tends to infinity. So, there is no *relevant* charge of inconsistency to be leveled against AIC in any situation. In fact, there is no such charge to be leveled against any of the methods I have discussed, which is to say that asymptotic results do not succeed in differentiating any method from any other. The crucial question concerns what happens for intermediate values of n. Theoreticians should focus on the harder questions, for there are no easy knock-down arguments against one criterion or another.

8 Summary of results

The analysis has raised a number of issues: is there any universal proof of optimality, or more realistically, Is one criterion more optimal than known competitors? Or does it depend on the circumstances? What is the sense of optimality involved? I believe that the framework described in this chapter shows how to approach these questions, and has yielded some answers in special cases. The main conclusion is that the perfor-

mance of model selection criteria varies dramatically from one context to another. Here is a more detailed summary of these results:

- All model selection criteria may be measured against the common goal of maximizing predictive accuracy.
- Predictive accuracy is always relative to a specified domain of prediction, and different domains define different, and perhaps conflicting, goals.
- It is commonly claimed that AIC is inconsistent. However, *all* criteria are consistent in the sense that they converge on the optimum hypothesis for asymptotically large data sizes.
- Because all methods are consistent in the relevant sense, this asymptotic property is irrelevant to the comparison of selection methods.
- The relevant differences in the selection criteria show up for *intermediate* sized data sets, although what counts as 'intermediate' may vary from one context to the next.
- When the more complex model merely adds adjustable parameters without reducing model bias, then BIC makes a better choice than AIC, but no method does better than always choosing the simpler model in this context.
- When a more complex model does reduce bias, but just enough to balance the expected loss due to overfitting, then this is a 'neutral point' at which all methods enjoy roughly the same degree of success.
- When a more complex model reduces model bias by an amount that exceeds the expected loss due to overfitting, then AIC does quite a lot better than BIC, though ML performs better than both.

The demonstration of these results is limited to the comparison of two nested models under conditions of normality, and it supposes that the domain of prediction is the same as the sampling domain (it deals with interpolation rather than extrapolation—see Forster, 2000 for some results on extrapolation). This leaves a number of open questions. How do these results extend to hierarchies of nested models, and to non-nested models? What happens when normality conditions do not apply? What if the domain of prediction is different from the domain from which the data are sampled? While I have few answers to these questions, I have attempted to describe how such an investigation may proceed.

What are the *practical* consequences of these results? In the case investigated here, I have plotted the relative performances of model selection criteria against the biases of the models under consideration. The problem is that the model biases are generally unknown.

A sophisticated Bayesian might assign a prior probability distribution over the model biases. For example, if the model biases along the x-axis

in figure 5.4 have approximately the same weight, then the expected performance of AIC will be better than BIC. If such a prior were available, it would not only adjudicate between AIC and BIC, but it would also allow one to *design* a third criterion that is better than both. However, it is difficult to see how any such prior could be justified.

If such priors are unavailable, then it seems sensible to favor AIC over BIC, if that were the only choice.[17] After all, AIC is a better estimator of predictive accuracy than BIC, since BIC is a biased[18] estimator of predictive accuracy. When you correct for the bias in BIC you get AIC. BIC merely sacrifices bias with no known gain in efficiency or any other desirable property of estimators.

Irrespective of any practical advice available at the present time, the main conclusion of this chapter is that the Akaike *framework* is the right framework to use in the investigation of practical questions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*, Budapest, Akademiai Kiado, pp. 267-81.

(1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control,* vol. AC-19: 716-23.

(1977). On the entropy maximization principle, in P. R. Krishniah (ed.), *Applications of Statistics:* 27-41. Amsterdam, North-Holland.

(1985). Prediction and Entropy, in A. C. Atkinson and S. E. Fienberg (eds.), *A Celebration of Statistics*, pp. 1-24, New York, Springer.

Bearse, P. M., H. Bozdogan and A. Schlottman (1997). Empirical econometric modeling of food consumption using a new informational complexity approach. *Journal of Applied Econometrics*. October 1997.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52:** 345-370.

(1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics—Theory and Method* **19**: 221-278.

Bozdogan, H. and D. Haughton (forthcoming). Information complexity criteria for regression models. *Computational Statistics and Data Analysis*.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: a Practical Information-Theoretic Approach*. New York: Springer.

---

[17] Of course, they are not the only choices. For example, Bearse *et al* (1997) and Bozdogan (1990) derive alternative criteria to AIC and BIC. Burnham and Anderson (1998) provide a recent survey of variations on AIC.

[18] An estimator of a quantity, in this case an estimator of predictive accuracy, is *biased* if the expected value of the estimate is not equal to the quantity being estimated. This sense of 'bias' has nothing to do with model bias.

Cheeseman, P. (1990). On finding the most probable model. In Jeff Shrager and Pat Langley, *Computational Models of Scientific Discovery and Theory Formation*, pp.73-93. San Mateo, CA: Morgan Kaufmann Inc.

Cramér H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, The MIT Press, Cambridge.

Forster, M. R. (1994). Non-Bayesian foundations for statistical estimation, prediction, and the ravens example. *Erkenntnis* **40:** 357 - 376.

Forster, M. R. (1995). Bayes and bust: the problem of simplicity for a probabilist's approach to confirmation. *British Journal for the Philosophy of Science* **46:** 399-424.

(1999). Model selection in science: the problem of language variance. *British Journal for the Philosophy of Science* **50**: 83-102.

(2000). Key concepts in model selection: performance and generalizability, *Journal of Mathematical Psychology* **44**: 205-231.

(2000a). Hard problems in the philosophy of science: Idealisation and commensurability. In R. Nola and H. Sankey (eds) *After Popper, Kuhn, and Feyerabend*. Kluwer Academic Press, pp. 231-250.

Forster, M. R. 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.

Geman, S., E. Bienenstock and R. Doursat 1992, Neural networks and the bias/variance dilemma. *Neural Computation* **4:** 1-58.

Keuzenkamp, H. and McAleer, M. (1995). Simplicity, scientific inference and economic modeling. *The Economic Journal* **105:** 1-21.

Kiessepä, I. A. (1997). Akaike information criterion, curve-fitting, and the philosophical problem of simplicity. *British Journal for the Philosophy of Science* **48:** 21-48.

Kruse, M. (1997). Variation and the accuracy of predictions. *British Journal for the Philosophy of Science* **48**: 181-193.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22:** 79-86.

Linhart, H. and W. Zucchini (1986). *Model Selection.* New York: John Wiley & Sons.

MacKay, D. J. C. (1995). Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* **6:** 496-505.

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

Raftery, A. E. (1994). Bayesian model selection and social research. Working Paper no. 94-12, Center for Studies in Demography and Ecology, University of Washington.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica* **14:** 465-471.

(1987). Stochastic complexity and the MDL principle. *Economic Reviews* **6:** 85-102.

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

Rosenkrantz, R. D. (1977). *Inference, Method, and Decision*. Dordrecht: Reidel.

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, Elliott (1988). Likelihood and convergence. *Philosophy of Science* **55:** 228-37.

Speed, T. P. and Bin Yu (1991). Model selection and prediction: normal regression, Technical Report No. 207, Statistics Dept., University of California at Berkeley.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B* **39**: 44-47.

Turney, P. D. (1990). The curve fitting problem—a solution. *British Journal for the Philosophy of Science* **41:** 509-30.

(1994). A theory of cross-validation error. *The Journal of Theoretical and Experimental Artificial Intelligence* **6:** 361-392.

Young, A. S. (1987). On a Bayesian criterion for choosing predictive sub-models in linear regression. *Metrika* **34:** 325-339.

Wallace, C. S. and P. R. Freeman (1987). Estimation and inference by compact coding, *Journal of the Royal Statistical Society* **B 49:** 240-265.

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica* **6**: 675-692.