Key Concepts in Model Selection:
Performance and Generalizability

Malcolm R. Forster
University of Wisconsin, Madison
July 8, 1998

ABSTRACT: What is model selection? What are the goals of model selection? What are the methods of model selection, and how do they work? Which methods perform better than others, and in what circumstances? These questions rest on a number of key concepts in a relatively underdeveloped field. The aim of this essay is to explain some background concepts, highlight some of the results in this special issue, and to add my own.

The standard methods of model selection include classical hypothesis testing, maximum likelihood, Bayes method, minimum description length, cross-validation and Akaike's information criterion. They all provide an implementation of Occam's razor, in which parsimony or simplicity is balanced against goodness-of-fit. These methods primarily take account of the sampling errors in parameter estimation, although their relative success at this task depends on the circumstances. However, the aim of model selection should also include the ability of a model to generalize to predictions in a different domain. Errors of extrapolation, or generalization, are different from errors of parameter estimation. So, it seems that simplicity and parsimony may be an additional factor in managing these errors, in which case the standard methods of model selection are incomplete implementations of Occam's razor.

1 What Is Model Selection?

There was a time when philosophers of science were primarily interested in what Kuhn (1970) called revolutionary science; for example, in the evolution of physics from Newtonian mechanics to relativity and quantum mechanics. In contrast, the everyday nuts and bolts of science, which Kuhn called normal science, was deemed to be less interesting from a methodological point of view.

Any philosophy of science that ignores normal science is incomplete for two important reasons. First, the grand theories, when they do exist, are not compared directly to empirical observations—the comparison is mediated by models. For example, in the third book of the *Principia*, Newton made severely idealized assumptions about the number and nature of gravitational bodies in the solar system to derive approximate equations that could be compared to the known facts about planetary motions. In philosophy of science these assumptions are usually called *auxiliary assumptions*. I will refer to the equations derived from the theory using a set of auxiliary assumptions as a *model* of the theory. Newton's science was normal science par excellence, and its continuing success beyond Newton marked a grand epoch in modern physics, whose success is still not complete to this day. In other words, to understand the success of Newton's theory requires us to understand the great success of modeling.

The second reason why normal science should not be overlooked is that not every model in science is deduced from a background theory. This is especially true in the biological and psychological sciences in which modeling often proceeds "from the bottom up" with minimal guidance from a background theory. To focus exclusively on the dynamics of background theories, as opposed to models, is to ignore a large body of science. Cartwright (1983) has also pointed out that a lot of "bottom up" modeling takes place in physics in any case.

But is modeling interesting from a methodological point of view? Are there any methodological problems here that pose a challenge to science? This special issue on model selection bears witness to the fact that our current understanding of these issues is rudimentary. One point is that even under the auspices of a mature theory, there may be many competing ways a modeling a phenomenon (Cartwright, 1983). The theory does not determine which auxiliary assumptions should be used. So, there needs to be some procedure for *comparing* models. At first, it may appear that model comparison is unproblematic. Surely, one just looks at which model best fits the data. Let me call this naïve empiricism (whenever a philosopher of science attaches the adjective "naïve," it is because it is wrong). The problem with naïve empiricism is that a model can always be made to fit the data better by adding complications. But is that good science? It was surely a part of Newton's success that he managed to obtain good fit the facts in an elegant and *parsimonious* manner. The accepted standard of good normal science to fit the facts as well as possible *with the simplest possible models*. This amendment of naïve empiricism suffers from a different malady—it is notoriously vague.

Newton's first rule of hypothesizing states that we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. It is widely considered to be Newton's version of Ockham's razor (also spelt 'Occam'), named after William of Ockham (1285 - 1347/49), which states that "entities are not to be multiplied beyond necessity." Unfortunately, such rules like this are as vague as they are informal. How do we judge what is "sufficient to explain the appearances"? Presumably "sufficient" refers to the degree to which a model fits the data, but there is a problem here because if adding a "cause" is to add an adjustable parameter. The problem is that adding a parameter will almost always improve fit to some degree. The proof of this assertion is as follows: Consider the best fitting case without the extra parameter . This is equivalent to adding the extra parameter and setting its value to zero. Therefore, one can do at least as well with the parameter as without it. Now, by adjusting its value to something different from zero, one can almost always fit the data slightly better due to random noise fluctuations present in any data set.

So if adding a parameter always improves fit, then presumably its addition counts as "beyond necessity" if the additional fit is too small. But, how do we judge that? How do we make this tradeoff between addition of parameters and gain in fit? And what is the rationale behind such methods in any case? What is gained by the tradeoff, and why? These are the kinds of problems that are addressed in a new branch of statistics called model selection.

My main area of expertise is in physics, and I have learnt from mathematical psychologists that this picture of model selection is too narrow. In particular, the choice between two models need not involve the choice between adding or not adding an adjustable parameter. In a mature science in which the construction of new models is strongly constrained by a background theory, it is common that a new model is obtained from the old one by adding extra terms and extra parameters. This means that the competing models are nested in the sense that the first model is a special case of the second model with the extra parameters set to zero. On the other hand, Cutting (this volume) describes an example in which the competing models are not nested. The problem is to model the response of a subject who is asked to rate the depth of a stimulus on a scale from 1 to 99. The stimuli contained 3 square panels laid out in depth with one behind another (Figure 1). The experimenters manipulated four different sources of depth information; relative size (S), height in visual field (H), occlusion (O), and motion perspective among the panels (P), so that they could be independently present or absent in any particular stimulus. That means that there were 16 variations of the stimulus in Figure 1 shown to each subject. A variable *D* coded for the depth response, scaled to a number between 0 and 1, of a subject to a stimulus. The problem is to ex-

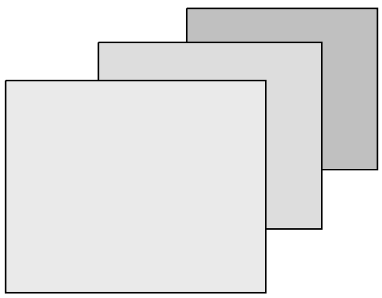

Figure 1: A stimulus like the ones used in Bruno and Cutting's experiment.

press D as a function of the presence or absence of each of the four visual cues. To write this mathematically, we need to introduce four independent variables x_S, x_H, x_O, and x_P to code for the presence or absence of the depth cues S, H, O, and P, respectively. In particular, $x_S = 1$ if S is present, and $x_S = 0$ if S is absent. The other three variables also take on the values 1 or 0 depending on whether the corresponding cue is present or absent. The problem is to choose a function that expresses D in terms of x_S, x_H, x_O, and x_P.

How is depth information combined? Bruno and Cutting (1988) proposed that the cues is combined additively. They modeled a subject's response by a linear function of the form:

$$D = S\,x_S + H\,x_H + O\,x_O + P\,x_P + B,$$

where S, H, O, and P are adjustable parameters that measure the influence that a cue has in determining the final depth judgment. B is a constant that takes account of the possibility that a subject will come up with a non-zero estimate of depth even in the absence of all the cues. Note that the parameters are constant for any particular individual, but are allowed to vary from one subject to the next. This model is known in the literature as the linear integration model (LIM).

On the other hand, Massaro (1988) proposed a non-linear model, which he called to as the fuzzy-logical model of perception (FLMP). Applied to kind of data described in Bruno and Cutting (1988), it proposed an entirely different functional form for the dependence of D on x_S, x_H, x_O, and x_P. Rather than writing it out in full, I will describe one instance of the equation: When all four cues are present,

$$D = \frac{SHOPB}{SHOPB + (1-S)(1-H)(1-O)(1-P)(1-B)}.$$

When one of the cues is missing, say S, then the parameter S is replaced by (1-S) throughout the formula. So, for example, if all cues are absent, the equation is:

$$D = \frac{(1-S)(1-H)(1-O)(1-P)B}{(1-S)(1-H)(1-O)(1-P)B + SHOP(1-B)}.$$

Although I have used the same letters to write the parameters in FLMP, there is no sense in which they are the same as the parameters in LIM. The meanings of the parameters are determined by the model in which they are embedded, and the models are different. There is no sense in which LIM is a special case of FLMP or FLMP is a special case of LIM. Nor can one model be obtained from the other by the addition or subtraction of adjustable parameters. The pair of models is non-nested. If there is a theory of model selection, it should apply to the comparison of nested and non-nested models alike.

Having thought about it, I now see that it would have useful applications in the methodology of physics as well. If we wanted to compare models across scientific revolutions, for example to compare quantum model of radiation with a classical model, then there is also no sense in which one model is nested in the other. And I note here that such comparisons are not merely of historical interest. There are many instances in which Newtonian models are still favored over relativistic or quantum mechanical alternatives. It is true that they are favored, in part, for pragmatic reasons, but it is still important to know how close is the approximation to reality, or whether there is any unambiguous answer to such a question. A general theory of model selection may increase our understanding of these issues.

In this paper, I will begin with a general characterization of what models are, and a brief survey of some of the methods of model selection covered in this volume (section 2). Almost every experimental scientist is familiar with Neyman-Pearson hypothesis testing, and it may be surprising to learn that classical hypothesis testing succeeds in trading off simplicity and fit. It is an implementation of Occam's razor. One shortfall of classical testing is that it does not generalize to the case of non-nested models.

In section 3, I will delve deeper into the foundational issues underlying model selection. It is important to understand which goals are achievable and which are not. Too often methods are designed on the basis of wishful thinking. Who would disagree that it would be nice to have a method that selected the truth, if that were possible. But is it achievable? And if it is not, then what is the next best thing? The answers are not always exactly as one would wish.

Section 4 describes some of my own theoretical investigation into the performance of some of the model selection criteria described in section 2. The research shows that the performance of different methods depend crucially on hidden features of the context in which they are applied. More specifically, for any two methods, *A* and *B*, there exists a context in which *A* performs better than *B* and there exists a context in which *B* performs better than *A*. Since simulations are usually tied to a specific (hypothetical) context, it is always possible to find a simulation in which your favorite method performs better than another. Beware of simulations!

The final section expands the scope of the essay by considering *novel predictions*. Model selection criteria are standardly designed to maximize the accuracy of re-sampled data from the same generating distribution. Busemeyer and Wang (this volume) present an interesting argument designed to show that the standard methods tend to select hypotheses that generalize poorly to data of a different kind. Their "generalization criterion methodology" is designed to address the problem. It is similar to the methodological distinction William Whewell (circa 1840) made between the prediction of data of the same kind (colligation of facts) and the prediction of data of a different kind (consilience of inductions). I will examine Busemeyer and Wang's argument in section 5.

Caution: Model selection is highly complicated phenomenon and it is very difficult to obtain many completely general results. It is a scientific problem in my view, and we should study it by modeling the process of model selection. In that spirit, I will make many simplifying assumptions along the way. The results are not be completely general, but it is the right place to begin.

2 The Many Methods of Model Selection

What is a model? From a theoretical point of view, there is an advantage to viewing models as sets of probabilistic, or statistical, hypotheses. In physics, this is done by associating models with an assumption about the probability of observational errors. In the depth perception example, depth judgments are naturally stochastic because depth judgments may differ even when visual cues are exactly the same. In other words, in each case, what appears to be a deterministic equation is interpreted an equation for the *mean* value of a variable as a function of the independent variables, or predictor variables. Technically speaking, the hypotheses are represented as probability densities, and models are sets of probability densities.

For example, in a simple curve-fitting situation in which a variable *y* is represented as a function of an independent variable *x*, as in $y = \ \ x + u$, where *u* is an error term governed by a fixed error distribution, the corresponding model *M* is a set of densities $f\ \ x, y;\ \ |\ \ \ .$[1] If the error distribution itself contains an adjustable parameter, say a variance $\ ^2$, then it is added to the list of parameters characterizing the set of probability densities: $f\ \ x, y;\ \ ,\ \ ^2\ |\ \ \ ,0\ \ \ ^2\ .$ The point is that the parameters like $\ ^2$, which are normally associated with statistical models, are mixed in with parameters like , which are not normally thought of as statistical parameters. Within the framework of model selection, there is no principled distinction between them.

This idea is easily generalized beyond the context of simple curve-fitting examples to cover any kind of scientific modeling in the quantitative sciences. For example, in physics, the parameters may include quantities like the half lives of radioactive isotopes, electrical resistance, elasticity, mass, electric charge, the speed of light, initial velocities, constants of integration, as well as error

[1] Strictly speaking, the specification of the error distribution only determines the probability distribution of *y* conditional on *x*. The missing element is a marginal probability density for *x*, which is supplied by a specification of the experiment design. This fact will play a role in the last section of this paper. See, also, Forster (1999) for further discussion of this fact.

variances. When the same parameter, such as the mass of the sun, reappears in many equations for different phenomena, then there is a sense in which all the equations are a part of a single model. The holistic character of models complicates the theory in ways that are presently unexplored. But the framework itself is quite general.

A density represents a hypothesis about objective probabilities in the world. A density is true if the probabilities it asserts to hold actually hold in the world. It may not be the complete truth. In this volume, the true density is alternatively called the environmental density (Golden, this volume) or the operating model (Zucchini, this volume) or the data generating density. A model is true if and only if one of the densities it contains is true.

There is a one-to-one correspondence between the members of a model the set and particular numerical assignments to the set of parameters of the model. For instance, if LIN is a model that asserts a linear functional dependence of y on x, then each member hypothesis is a particular line in the x-y plane, which is uniquely characterized by its slope and its point of intersection with the y-axis. A model may be represented as a set of points in a parameter space, and each member of the model is a point in that space. For that reason, I sometimes refer to a density as a *point hypothesis*. Point hypotheses not only specify the probability (density) of the observed data, which is known as the likelihood of the hypothesis relative to the observed data (not to be confused with the probability of the hypothesis given the data). But point hypotheses also specify the probability of any *possible* set of data within its domain. For that reason, I have also referred to point hypotheses as *predictive hypotheses* (Forster, 1999). In this essay, I shall sometimes call them predictive *densities*.

A model (in our usage) is a kind of hypothesis, but it is far weaker in that it only asserts that one of its densities is the true density. Logically speaking, it is the (infinite) disjunction of all the point hypotheses contained in it. The laws of probability imply that the likelihood of a model is the average of the likelihoods of its members, but this is undefined unless there is a prior probability distribution over the members of the model. Likewise, a model does not confer a well defined probability of new data, and is therefore ineffective for the purpose of anticipating the future and determining our expectations of new facts. For me, it is a fundamental axiom that the purpose of model selection is to mold our predictive expectations, and so I am forced to the ironic conclusion that model selection is not *fundamentally* about choosing a model. For me, the aim is to select a particular density from a model. Choosing a model is merely a necessary part of the process.

Since the selection of a predictive density from a model involves the assignment of numerical values to all adjustable parameters, the model selection includes parameter estimation. For those readers less familiar with statistical methods, the essential idea is that parameter estimates are determined by the member of a model that best fits the observed data (recall the one-to-one correspondence). Given the way that the framework is setup, it is convenient to assume that 'fit' is measured by likelihood, so that the method of parameter estimation is the method of maximum likelihood (ML). Therefore, the aim of model selection is to choose from amongst the maximum likelihood hypotheses of all competing models.

Not everyone in the field uses the terminology in the same way. Most importantly, Zucchini (this volume) and Grünwald (this volume) refer to a predictive density as a "model", and what I call a model as a family of "models". While it is easy to translate, I suspect that it may have encouraged some misunderstandings. When these authors talk about model selection, they contend that the goal of process is to select approximately true predictive densities, which I think is right. However an alternative (incompatible) view is that model selection aims at the selection of approximately true *models*. The terminological muddle may lead advocates of the alternative view to mistakenly cite these authors in their support.

2.1 The Method of Maximum Likelihood (ML). Maximum likelihood is principally a method of parameter estimation, but it extends straightforwardly to model selection. The rule is to choose the best of the best. That is, out of the maximum likelihood hypotheses in the competing models, select the one that has the greatest likelihood; or equivalently, the greatest log-likelihood. If one thinks of log-likelihood as a measure of fit, then ML is the implementation of naïve empiricism. It is the antithesis of Occam's razor. In fact, in the case of nested models, it can never favor anything less than most complex of all the competing models.

2.2 Classical Hypothesis Testing (N-P). This is the classical methodology of statistics made famous by Neyman and Pearson in joint publications between 1933 and 1938. It can be applied to many problems in model selection. Consider the comparison nested models, in which we decide whether to add or omit a single parameter . Omitting the parameter is effectively the same as setting it to zero, so the choice is between the hypotheses 0 and 0. In the general case, the competing models will share other adjustable parameters, but let us suppose that these are set at their previously estimated values.[2] Classical hypothesis testing does apply to a situation like this. Let $\boldsymbol{x}$ be a variable ranging over possible data sets. The hypothesis 0 specifies a particular density f x; 0. Let be the maximum likelihood estimate for . Then is a function of $\boldsymbol{x}$ whose probability distribution is determined by 0 to be g ; 0. If 0 is chosen as the null hypothesis, then we may set up a 5% critical region, or rejection set, such that if is sufficiently close to 0, then the null hypothesis is not rejected ($p < .05$, two tailed). It is an important fact that the null hypothesis would always be rejected if one went by fit alone. For is the best fitting hypothesis in the model 0, and it fits the observed facts better than 0. Therefore, when a classical test fails to reject the null hypothesis, it is favoring the simpler hypothesis *in spite of its poorer fit*. *Classical hypothesis testing succeeds in trading off goodness-of-fit for simplicity.* The only requirement is that the null hypothesis is chosen to be the simpler of the two models.

So why is there a need for new methods of model selection? The first problem is that hypothesis testing does not work when the models are non-nested. For instance, in the depth perception example of section 1, there is no reason to choose LIM rather than FLMP as the null hypothesis, and the results will be entirely different in each case. Classical methods are not based on any deep analyses or insights into the problem of model selection, even though its methods may be adequate in some problems.

2.3 AIC. The goal of Akaike's Information Criterion (AIC) is to minimize the Kullback-Leibler (*K-L*) distance of the selected density from the true density (Akaike 1973, 1974, 1985). Let $\boldsymbol{x}$ represent the total n observed data, and k the number of adjustable parameters whose variation make a difference to that discrepancy (the degrees of freedom). Then the expected value of $\log f$ x; $/n$ k/n is 2 times the *K-L* discrepancy under asymptotic conditions (which requires that the probability distribution of is normal—the density f x, need not be normal). This is Akaike's theorem (see Bozdogan (this volume) for further details), which provides a principled way of trading off simplicity and fit. The AIC rule is to select the predictive density that has the lowest expected *K-L* discrepancy, which amounts to the maximization of $\log f$ x; $/n$ k/n. The first term measures fit, while the second term penalizes complex models. Note that AIC, without the second term, would be the same as ML. The AIC rule is a modification of naïve empiricism.

What is the difference between AIC and classical hypothesis testing? One advantage is that AIC applies to nest and non-nested models. All that is required for the comparison of models is their maximum likelihood values, their k values, and the

---

[2] This is an idealization in the sense that, in general, the estimation of the new parameter is not entirely independent of the estimation of the other parameters. Later I will generalize the discussion to take this into account.

number of data. There is no need to say which model is the null hypothesis. The other difference is the rationale behind them. Akaike makes no reference to a tradeoff between type I or type II errors.

Apart from this, they are remarkably similar in examples to which they both apply. It is instructive to reconsider the example of nested hypotheses. Suppose that is a sufficient statistic for , which by definition means that f x; g ; h x .

Then AIC will reject the hypothesis 0 if and only log g ; 1 log g ; 0 . If g ; is normal in the way required by Akaike's theorem, then g ; 2 ^{2} $^{1/2}$ exp ^{2}/2 ^{2} .

Therefore, 0 will be rejected when is far enough from 0 so that 0 ^{2} 2 ^{2}. That is, 0 is rejected when is more than $\sqrt{2}$ standard deviations from the mean, which corresponds to a 15.73% rejection rate when the null hypothesis is true. It is equivalent to a classical test with a rejection area of 15.73% (p < .1573, two tailed). The irony is that a 5% classical test gives greater weight to simplicity than AIC.

I anticipate that the classical community will view this is an intolerably high type I error rate. This reaction is a natural consequence of the deep differences in the way that model selection is conceptualized. In AIC's defense, first note that 15.73% is only the probability of choosing the wrong model *when the null hypothesis is true*. If the null hypothesis is not true, then the rejection the null hypothesis is not an error. So, what is the probability of the null hypothesis being true in the examples we have considered? *A priori* considerations suggest that it is *zero*. If I throw a dart onto an dart board, what is the probability that it will land exactly in the middle? When one actually looks at the models being compared, then *a priori* judgment is confirmed. What is the probability that any of the planets are spherically uniform in their distribution of mass? For that matter, what is the chance that the earth bulges at the equator in exactly the sense assumed by Newton. Or what is the chance that either of the models of depth perception considered in section 1 being exactly correct? The most plausible view is that all models are idealized to some extent, and none of them is true. But if all models are false, then the issue of type I or type II errors never arises. So, what is the point of a method designed to minimize these errors?

One response may be to say that the null hypothesis may be *approximately* true, in which case rejecting the null hypothesis does count as a mistake. Or does it? Suppose for the sake of argument that the null hypothesis is approximately true. Does it then follow that its rejection is therefore a mistake? I think not. For if 0 is a good approximation, then an instance of the competing model 0 is also a good approximation. Rejecting the null hypothesis is not necessarily a mistake because mistakes are no longer measured in black and white. Nevertheless, it may well be that 0 is a better approximation than . In that case, rejecting the null hypothesis is a mistake. But it is probably not a big mistake. When the null hypothesis is approximately true, then is probably close to 0.

Sometimes the rejection of the null hypothesis when it is approximately true can have serious social and political consequences, or it may be very expensive. The null hypothesis that there is no difference in mean intelligence amongst people of different racial backgrounds is one such example. People will be less comfortable with AIC when it threatens such theses. However, remember that AIC is not designed to take account of social consequences. This is the forte of Bayesian decision theory, and AIC is not trying to compete with it in these examples. It is only designed to (fallibly) estimate the predictive accuracy of competing hypotheses. To reject the null hypothesis is only to conclude that there is a statistical correlation between race and IQ that may extend to new cases. It suggests nothing about whether the correlation has a genetic or environmental causes.

Another important point is that the conclusions of AIC are never about the truth or falsity of a hypothesis, but about its closeness to the truth. When

AIC favors the hypothesis over 0, the only thing that is (provisionally) asserted, if anything at all, is that the true value of is closer to than it is to 0. This is a weaker assertion than concluding that the true value of is equal to . I am only defending AIC as it is *properly* understood.

Conclusions are always provisional and every selection criterion is fallible. The risks involved should be openly acknowledged and assessed. In fact, Golden (this volume) talks about "higher level" statistical tests that can be done to assess the significance of lower level decisions made by any kind of model selection criterion. He argues that model selection criteria should lead to a three-way decision—accept, reject, *or suspend judgment*. Including a third option is an excellent recommendation.

Zucchini (this volume) points out that model criteria is especially risky when a selection is made from a large number of competing models. The more models there are, the greater the chance that the winner won by luck rather than by merit. This phenomenon is known as *selection bias*. Regression towards the mean is a related phenomenon, in which very high or very low extremes are not reliably repeated. For example, most sons of very tall fathers, will tend to be closer to an average height than their fathers. Israeli flight instructors were told that they should praise their students when they performed exceptionally well, and criticize students when they performed very poorly (Kahneman and Tversky, 1973). The flight instructors (wrongly) concluded that criticism was effective in improving performance, but praise was not!

Tversky and Kahneman (1971) did a survey of mathematical psychologists, which showed that they tend to make a similar mistake. They tended to overestimated the probability that a repeated experiment will produce a statistically significant result in one experiment will be repeated in a second experiment. It appears that many people fail to grasp the meaning of their results even when they are reliable.

Browne (this volume) emphasizes that selection criteria should not be followed blindly. They are merely guides to research, not the end of research. In many ways, it would be better to think of model selection methods as tools for model *comparison* or *assessment* because the term 'selection' may suggest that something more definite has been attained. This point applies to all the methods discussed here, so it should not be used against one method at the expense of another.

2.4 Cross-Validation Techniques. Cross-validation (CV) is an entirely different technique, whose advertised goal is to evaluate the predictive accuracy of the competing models (Brown, this volume). The idea is to divide the data into two subsets; the calibration set or the training set and the test set. A model is fitted to the calibration set, to obtain $_{cal}$, which is then evaluated by its ability to fit the test set. Because the test set is not used in the construction of the hypothesis, it provides a unbiased estimate of its performance with respect to *any* new data. However, the goal is to evaluate the predictive performance of , fitted to the whole data, and this is different from $_{cal}$. The difference between the two is minimal if only one datum is left out of the training set. This advantage is offset by the unreliability of estimating predictive performance with the single datum. The solution is to repeat the procedure *n* times leaving out a different datum each time, and averaging the result. The CV rule is to select the model with the best score.

Note that leave-one-out cross-validation makes no explicit appeal to simplicity what-so-ever, so it may be surprising to learn that it is asymptotically equivalent to AIC (Stone, 1977). Perhaps it is not so surprising given that their goals are equivalent. For it turns out that to minimize the *K-L* distance between the maximum likelihood density is the same as maximizing predictive accuracy if that is defined in terms of the expected log-likelihood of new data generated by the true density (Forster and Sober, 1994).

2.5 Bayes Method. The idea behind Bayes method is that models should be compared by their posterior probabilities (Wasserman, this volume). If all models are known to be false, then they all have zero probabilities. Nevertheless it might be useful to know which is the more probable of two models *if we were to assume that at least one of them is true*. Schwarz (1978) assumed that the prior probabilities of all models were equal, and then derived an asymptotic expression for the likelihood of a model, which is sometime referred to as the Bayesian Information Criterion (B-I-C). Logically speaking, a model is a huge disjunction, which asserts that either the first density in the set is the true density, or the second density, or the third, and so on. By the probability calculus, the likelihood of a model is therefore the average likelihood of its members, where each likelihood is weighed by the prior probability of the particular density given that the model is true. The BIC rule is to favor the model with the highest value of $\log f \;\; x; \qquad /n \quad [\log n/2]k/n$. Notice that BIC is similar to AIC except that it gives greater weight to simplicity by a factor of $\log n/2$. When classical testing applies, BIC is equivalent to a classical test whose size slowly decreases as the sample size n increases.

Bayes method is one thing and BIC is another. The latter is not always an approximation of the former. Consider a competition amongst nested models in which the true density first appears in the model k^* (there may be just two models in the hierarchy, in which case k^* would be the first). If k^* is the smallest model containing the true density, then all models higher in the hierarchy are also true. Therefore, the probability that a model contains the true density can never decrease as one moves up the hierarchy. According to Bayes method, there is no model more favorable than the model at the very top of the hierarchy. Bayes method therefore chooses the most complex of all the models under consideration if they are truly nested! BIC does not have this consequence, therefore BIC is not an approximation of Bayes method in the case of truly nested models.[3]

It's not that Bayes method goes wrong here. It does exactly what it was designed to do—maximize the probability of a model's truth. Nevertheless, there is a subtly different way of setting up the problem that makes a huge difference in what Bayes method recommends. Suppose that we set up the hierarchy in such a way that each model is disjoint from all the others. This can be done by subtracting the kth model from the $k+1$ model. I will refer to such hierarchies as quasi-nested. In quasi-nested hierarchies, the models are mutually exclusive, or disjoint. This means that if k^* contains the true density, then *no other model in the hierarchy contains the true density*. It is a subtle difference, but it makes a huge difference in what Bayes method recommends. For in this case, Bayes method is approximated by BIC and it does converge asymptotically on the model k^*, and again it does exactly what it should do.

The huge difference between truly nested and quasi-nested models in Bayes method also proves that Bayes method is *not* about the *approximate* truth of models. To see this, note if $\quad 0$ is only approximately true, then there is a member of $\quad 0$ that is also approximately true because there is a member of $\quad 0$ that is infinitesimally close to $\quad 0$. Therefore, the probability that $\quad 0$ is approximately true is never greater than the probability that $\quad 0$ is approximately true. Bayesians have to maintain that models should be compared as by the probability that they contain the *exact* true density. Otherwise, Bayes method recommends the most complex model in *any* nested hierarchy.

AIC is entirely insensitive to the difference between truly nested and quasi-nested models because AIC is compares the competing maximum

[3] Forster and Sober (1994, section 7). There are other problems with Schwarz's derivation described by Wasserman (this volume), as well as a number of new Bayesian approaches to the problem. Also see Bandyopadhyay *et al* (1996) for yet another Bayesian approach. They are all built on the same conceptual foundation, which is the main subject of criticism in this essay.

likelihood hypotheses, and the candidates are the same in each case. This proves that Bayes method and AIC aim at entirely different things. The contrast is especially dramatic in a case in which we know that a simpler nested model is false, and the complex model is true. For a Bayesian, the complex model is a clear winner. Its work is already done, whereas the work of AIC has only just begun. AIC is interested in which maximum likelihood hypothesis is closest to the truth, and the truth of complex model does not imply that it yields the best maximum likelihood hypothesis (Zucchini, this volume). Ironically, AIC is not *primarily* concerned with selection of *models* at all!

Recall that AIC aims at maximum likelihood densities because models make no predictions. Bayesianism are also concerned with prediction, but they have an entirely different way of generating predictive densities. The standard approach, adopted by Schwarz (1978), is the winner-takes-all procedure in which the predictive density is the average density of all members of the winning model, where the weights are given by the posterior distribution over the members of the model, given that the model is true. Various convergence theorems show that the posterior distribution concentrates all its weight on the maximum likelihood hypothesis in the limit of large n. In that limit, the BIC rule reduces to a choice amongst the competing maximum likelihood densities, just like AIC. But the philosophy behind it is entirely different.

The method of model averaging (Wasserman, this volume) abandons the assumption that the winning model is true. It only supposes that *one* of the competing models is true, and derives a predictive density as a weighted average of all the densities in all the models. On the positive side, it diminishes the problem of selection bias (Zucchini, this volume), because it is not a risky kind of winner-takes-all procedure favored by other model selection techniques. On the other hand, this advantage comes at the expense of making the predictions rather imprecise. It raises an interesting theoretical question about when such a tradeoff is worthwhile.

3.5 Minimum Description Length Criteria. Within computer science circles, the best known implementation of Occam's razor is the minimum description length criteria (MDL) or the minimum message length criteria (MML) (Grünwald, this volume). The motivating idea is that the best model is one that facilitates the shortest encoding of observed data. I have little to say about these methods except to mention that one version of MDL is asymptotically equivalent to BIC. That means that the following sections provide some information about the performance of at least one of these criteria in prediction problems.

3 The Bias/Variance Dilemma

Having surveyed five different methods of model selection, it may come as no surprise that there is considerable disagreement about the goal of model selection. Yet from a theoretical point of view, it is essential to compare the criteria against a common standard. I shall adopt the Akaike point of view, in which the aim to select a predictive density that best fits the true density within a particular domain of prediction. I do not pretend that this is the only interesting goal to study, but it is the focus of this essay.

I have already made the point that focusing exclusively on models is wrong. For example, simulation results in this volume are often reported in terms of the relative frequency of choosing the "right" model and the "wrong" model. Yet in the case of nested models, this distinction is very strange. If model A contains the true density and it is nested in model B then model B also contains the true density. Both models are "correct". In the non-nested case, the distinction may be well defined. Yet it is still unfaithful to the Akaike philosophy, in which the *size* of a mistake matters. The black and white distinctions of right and wrong or true and false are vestiges of the classical and Bayesian viewpoints, and it is hard to shake off the shackles of past conceptualizations.

The task in front of us is to (i) define the goal of model selection, (ii) describe the factors that determine when the goal is met, and (iii) measure the performance of all criteria against this goal. I will begin as informally as possible, in order to explain as much as possible with as little as possible for as long as possible.

"Lack of fit" is also referred to as *discrepancy* (Zucchini, this volume). Unless I state otherwise, by "discrepancy" I mean the discrepancy between a density and the true density. There are many ways of defining the discrepancy, but suppose we have chosen a particular measure. In any model selection problem, the right choice and the wrong choice is defined by the discrepancy. Discrepancy, like truth, is something that is unseen. It defines the goal, but does not provide the means.

In the case of two nested models, where *B* is obtained from *A* by adding parameters, the overall discrepancy is usefully broken up into two components: the bias and the variance. Proposition 1 describes the bias, while Proposition 2 describes the variance. The overall discrepancy depends on the bias and the variance.

Definition 1: Let $^*_A$ be the member of model *A* that has the least discrepancy in the family, and let $^*_B$ the member of *B* with the least discrepancy. Let the discrepancy of $^*_A$ with respect to the true density is called the *model bias* of *A*, or the *approximation discrepancy* (Zucchini, this volume), or the *model error* (Kruse, 1997), or the degree of misspecification.

Proposition 1: If *A* is nested in *B*, then the model bias of *A* is greater than or equal to the model bias of *B*.

Proof: Let $^*_A$ denote the density in *A* that best approximates the true density. Then $^*_A$ is also in *B* (or there is a member of *B* that is arbitrarily close to $^*_A$ in the case of quasi-nested models). Therefore, the best approximation in *B* is no worse than the best approximation in *A*.

In summary, the advantage of adding a parameter is that the model bias will never increase. The second question is whether there is any risk of losing something. Are there any disadvantages in adding a parameter? The answer is "yes", although the question is a bit more involved.

The easiest way to think about the disadvantage of adding parameters is to imagine that both models share a common parameter , while model *B* has one extra parameter . If both models are represented in the 2-dimensional space of points , then model *A* is the line = 0, while *B* is the whole plane. Let the member of *A* with the smallest discrepancy be $^*_A$,0 , while the member of *B* with the smallest discrepancy is $^*_B$, * . Similarly, the maximum likelihood densities in *A* and *B* be $_A$,0 and $_B$, respectively. Note that in general $^*_A$ $^*_B$ and $_A$ $_B$. However, there is often a way of transforming the parameters , so that $^*_A$ $^*_B$, $_A$ $_B$, and is stochastically independent of (remember that maximum likelihood estimates are functions of the data ***x***, and are therefore random variables). Such transformations do exist asymptotically if certain regularity conditions hold (Cramér, 1946, chapters 32 and 33). If the parameters have been transformed in this way, then all relevant differences between the models are characterized solely in terms of the parameter .

Definition 2: The discrepancy difference between a maximum likelihood density and the minimum discrepancy density is called the *estimation discrepancy*.

If the estimation discrepancy of both models is the same with respect to their shared parameter, , then the difference in estimation discrepancy is determined by the parameter . However, model *A* has no estimation discrepancy in because its minimum discrepancy value is 0 and is estimated value is 0. This motivates the following proposition:

Proposition 2: If *A* is nested in *B*, then the estimation discrepancy of model *B* is greater than or equal to the estimation discrepancy of model *A*.

So, the disadvantage of adding a parameter is that it will almost always increase the estimation discrepancy. Whether adding a parameter is an advantage *overall* depends on whether the gain in model bias outweighs the increased estimation discrepancy. This is commonly referred to as the bias/variance dilemma (e.g., Geman *et al* (1992)).

When the models are not nested, the concepts of bias and variance still apply. The only difference is that there is not necessarily a tradeoff—there may be a clear cut winner that minimizes the bias and the variance at the same time.

4 How Well Do the Methods Work?

Section 2 described the methods of model selection and section 3 described the goal. In this section, I will put these two components together by asking how well the methods *succeed* in minimizing the discrepancy of the selected densities. This is a question about how well the methods actually *perform*. It is a theoretical question, but it goes one step beyond the analysis of the previous section.

Because the problem is hard, I will solve it only within a narrower set of conditions. I will consider one particular discrepancy function and I will assume that the conditions guaranteed to hold asymptotically in sample size under certain regularity conditions (Cramér, 1946, chapters 32 and 33) actually hold for finite data (at least to a sufficiently good approximation). In what follows, I will simply refer to these as "asymptotic conditions" without *necessarily* implying that the sample size is large. Sakamoto *et al* (1986) is an excellent introduction to this area of mathematical statistics.

Let $\boldsymbol{x}$ be a shorthand notation for n data, $x_1, x_2, \ , x_n$, and let $f_{\ ,\ }(\mathrm{x})$ be the density in a family B associated with the parameter values $\ ,\ $. Suppose that $f^*(\mathrm{x})$ is the true density, which is not necessarily in the family. Note that $\boldsymbol{x}$ is being treated as a random vector governed by the true density. The actual data, x_0, provides one particular value of $\boldsymbol{x}$. Then the Kullback-Leibler discrepancy is defined as:[4]

$$_{K\text{-}L}(f_{\ ,\ }) \quad \frac{1}{n}\int f^*(\mathrm{x})\log f^*(\mathrm{x})\,d\mathrm{x} \quad \frac{1}{n}\int f^*(\mathrm{x})\log f_{\ ,\ }(\mathrm{x})\,d\mathrm{x}$$

A well known theorem (Kullback and Leibler, 1951) shows that $_{K\text{-}L}(f_{\ ,\ }) \quad 0$, where the equality holds if and only if $f_{\ ,\ }(\mathrm{x}) \quad f^*(\mathrm{x})$ for all $\boldsymbol{x}$, except for a subset of $\boldsymbol{x}$ values of measure zero ("almost everywhere" in language of measure theory) . That is, any density with zero discrepancy must agree with the true density on almost all probabilities.

Now introduce the maximum likelihood estimates for $\ ,\ $ to be and respectively. Note that they are functions of $\boldsymbol{x}$, so that and are also random variables governed by a true density function $g^*(\ ,\ )$, which is derivable from the true density $f^*(\mathrm{x})$. I will also assume that there exist values * and * of the parameters and such that $g_{\ ^*,\ ^*}(\ ,\ )$ is the true distribution of the maximum likelihood statistics and . Note that this does *not* imply that model B contains the true density. I am only assuming that there is a density in B that determines the true density for and , which is a much weaker assertion.

Asymptotic conditions further imply that $g_{\ ,\ }(\ ,\ )$ is a multivariate normal distribution whose first moments (means) are and , respectively, and whose second moments (covariance matrix) do not depend on or . This is all we need in order to ensure that there exists a transformation of the parameters in which $g_{\ ,\ }(\ ,\ ) =$ $g(\ )\,g(\ )$, where g and g are normal distributions with means and , respectively, and variances equal to $1/n$.

The standard parameterization of a family of densities will not have these nice properties. However, the existence of a parameterization with these properties is all that is relevant to the theoretical

[4] This differs from the formula in Zucchini (this volume) by a constant term and a factor of $1/n$, which make no practical difference because the constant term drops out in any comparison. I have also taken it for granted that all discrepancies are discrepancies with the true density.

study of this section. In everything that follows, I will assume that and are the transformed parameters. In the Appendix, I explain how the *K-L* discrepancy reduces to:

$$2 \quad f_{\,,} \qquad {}^{*\,2} \qquad {}^{*\,2} \quad const.$$

In other words, the *K-L* discrepancy of any density , in model *B*, is determined by the sum of the squared distances of the parameters from their optimal values, modulo a constant. This formula applies to *any* member of *B* and to any member of *A*, since *A* is nested in *B*. It is easily generalized for any number of adjustable parameters.

An immediate application of this formula provides a precise formulation of the bias/variance dilemma (section 3). Let , refer to the density $f_{\,,}$ x , and let $_0$ and $_0$ refer to the maximum likelihood values of and relative to the actual data x_0. Then the difference of the overall discrepancies between the maximum likelihood hypotheses in models *B* and *A* is:

$$_{0},0 \qquad _{0},\ _{0} \quad \frac{1}{2} \quad {}^{*}\ 0^{\,2} \qquad _{0} \quad {}^{*\,2}\ .$$

The first term on the right hand side is gain due to decreasing the model bias that results from adding the parameter, while the second term is the loss due to an increased estimation discrepancy. When the gain in bias outweighs the loss in variance, there is a gain in the overall discrepancy.

How does this analysis extend to the case of non-nested models? Here it is necessary to modify the reduced formula for *K-L* discrepancy because part of the constant term is only constant for densities in the model *B* (and *A*). The generalized formula is:

$$2 \quad , \qquad {}^{*\,2} \qquad {}^{*\,2} \quad 2 \quad {}^{*},\ {}^{*} \quad const.$$

The third term is the model bias of *B*, while the first two terms measure the estimation discrepancy. The constant term is constant for all densities. A similar formula may be derived for a second model under the same assumptions for an entirely disjoint set of parameters. Here it is important to understand that the derivation only requires that each model is correct in what it says about the distribution of the maximum likelihood estimators of its parameters. It is possible that two disjoint models are correct in what they say here even though it is impossible for both them to be correct about the density of ***x***. Again, the success of the competition is defined by the different in model bias the difference in the estimation discrepancy, even though there may be no tradeoff between them.

The choice of one model in favor of another is correct if and only it leads to a gain in overall discrepancy. A model selection criterion must make that choice based on the difference in log-likelihoods and other information such as the difference in the numbers of adjustable parameters, k, and the number of data *n*. The missing element in this framework is a representation of the difference in the log-likelihoods. This is achieved by embedding the models in a common parameter space. For example, if model *A* is a one-parameter model and model *B* is a two-parameter model, then they may be embedding in a three-dimensional parameter space (see Figure 2). *A* is a line and *B* is a plane. *A* is nested in *B* only when then angle between them, , is 0. The parameters are transformed so that the point , , , which represents the data, has a normal probability distribution centered at the point *, *, * with a variance of 1/*n* in any direction. This means that the mean *K-L*

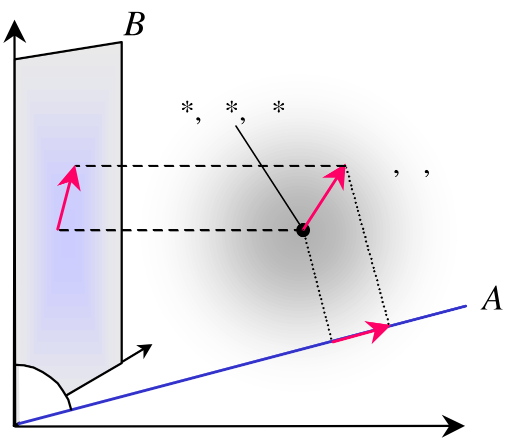

Figure 2: A three-dimensional parameter space in which model *A* is a line and model *B* is a plane.

discrepancy between $*, *, *$ and $, ,$ is proportional to $3/n$. The densities in the models that play the same role are the orthogonal projections of these points onto the respective subspaces. Therefore the mean estimation discrepancy for model *A* is proportional to $2/n$, and the mean discrepancy for model *A* is proportional to $1/n$. This, by the way, is the fact that was exploited by Akaike to correct for the effect of complexity on the estimation discrepancies in model selection.

Within this representation, the log-likelihood of all densities is less than the log-likelihood of $, ,$, which has the greatest log-likelihood of any hypothesis. The size of the difference is equal to $n/2$ times its squared distance from $, ,$. The *n* arises from the fact that the squared distances have been scaled so that they reflect the *K-L* discrepancy, which I have defined to be independent of *n*, whereas log-likelihoods do depend on *n*.

For example, in the special case in which $=$ 90 , then the maximum likelihood hypotheses are $0,0,$ and $, ,0$ respectively. Then the difference in their log-likelihoods is given by:

$$l \; , \; ,0 \quad l\; 0,0, \qquad {}^{2} \qquad {}^{2} \qquad {}^{2} \; n/2,$$

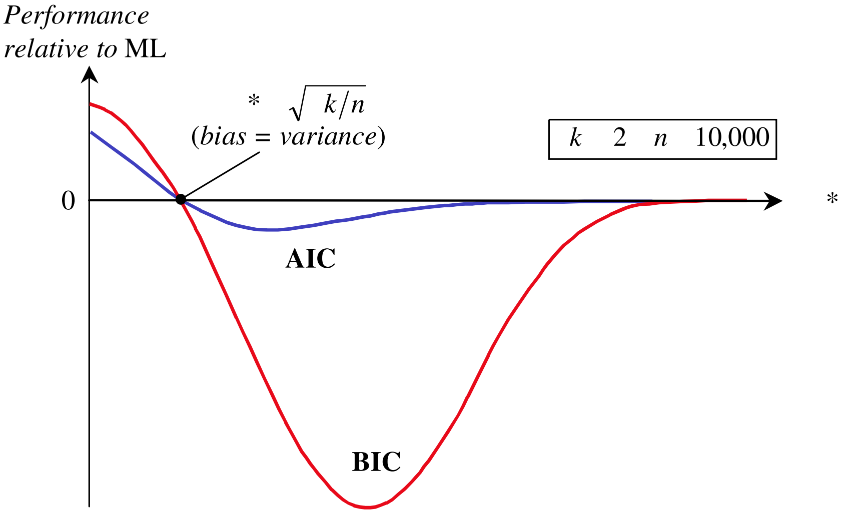

Figure 3: The performance of AIC and BIC depends on the context. Higher is better.

where *l* stands for "log-likelihood". Now we have all the ingredients needed to study the performance of any model selection criterion .

If a criterion makes the correct choice, then the *performance gain* is equal to the decrease in discrepancy resulting from that choice (in comparison with the alternative choice). If the criterion makes the wrong choice, then the performance gain is negative. The overall *performance* of a criterion is the average performance gain on all possible instances generated by the true density.

Note that in any particular model selection problem, there will be some fact of the matter about the difference in model bias between *A* and *B*. Note also, that this does not depend on the observed data x_0, or on the sample size *n*. It is a property of the prediction problem itself. Moreover, it is the only feature of the background context that is relevant to performance of a model selection criterion under the asymptotic conditions assumed here. That is, performance is a function of the *difference* in model bias, which I will denote by

$$bias \;\; \tfrac{1}{2} \qquad *^{\,2}, \text{ where } \quad * \qquad * \;\; 0 .$$

Figure 3 shows the performance of AIC and BIC plotted as a function of $*$ in the case of nested models, in which the more complex model has two additional parameters. More exactly, I have plotted the performance of AIC, or BIC, minus the performance of method of maximum likelihood (ML), which in this context always selects the more complex of the two models. Remember that $*$ is a constant for any given problem, so you are "located" at some fixed point on the *x*-axis in Figure 3. Notice that the relative performances depend dramatically on where you are "located". The region in which $* \quad \sqrt{k/n}$ is quite different from the region in which $* \quad \sqrt{k/n}$. The point dividing these two regions is the point at which the gain in bias equals the loss in variance.

ML is a good baseline to choose because, as it turns out, all methods are equivalent to ML when the difference in bias large compared to the difference in variance (* $\sqrt{k/n}$). This is shown at the right hand side of the graphs. If the complex model has a large advantage in bias then it is almost certainly reflected in the log-likelihoods, which is all the ML exploits. Thus, the performance of *all* the criteria in this region is optimal.

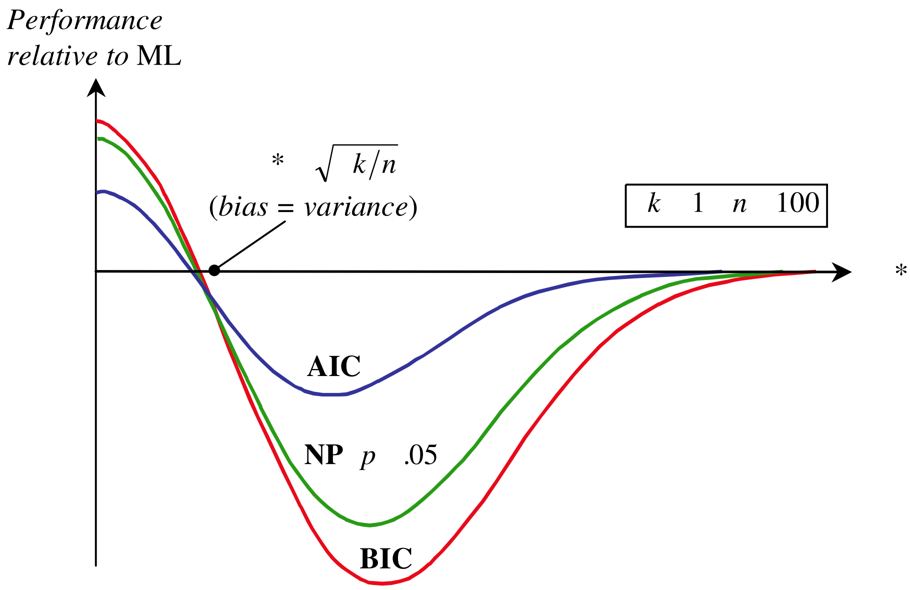

Figure 4: The relative performance of AIC, BIC, and Neyman-Pearson hypothesis testing with one extra parameter.

By saying that a rule is "optimal", I mean that a rule that chooses the predictive hypothesis with the least discrepancy *every time* would have a negligibly better performance. Many people that ML is never optimal because it always chooses the most complex model in a nested hierarchy. But choosing a complex model does not preclude it from choosing a predictive density that is close to those in the simpler models of the hierarchy. Remember that it is the discrepancy of the predictive density that matters in performance measurement. The same confusion often leads people to say that BIC performs better than AIC for large sample sizes, but this is not true if "performance" is properly defined (Forster, 1999).

The slogan "AIC is inconsistent, but BIC is not" refers to the fact that AIC (and any classical test) "overshoots" the smallest true model in a nested hierarchy in the limit of large n, while BIC does not. Again there is an exclusive focus on the selection of models, with no consideration to the predictive densities. It turns out that the *size* of the mistake made by AIC diminishes to zero in the limit of large n, yet this mitigating circumstance is never mentioned (Forster, 1999). The point is that the performance of AIC has to be understood in terms of its goal of maximizing the fit between selected densities and the true density.

When the bias difference is higher (* $\sqrt{k/n}$), then the rules that give the *least* weight to simplicity perform better. In this region, ML, which gives no weight to simplicity, performs better than AIC, which performs better than BIC. While this is a general feature of these graphs (for all k and all n), the magnitude of the differences may vary. Figure 3 shows the plot when the complex model has two extra parameters (k 2), and n = 10,000. While the relative performances change when k changes, I have found that they do not perceptible change when n is changed. Of course, the absolute performance change with n (they all get better as n increases), but the relative performances look the same as shown in Figure 3. When k is increased, then the performance advantage of AIC over BIC continues to widen in the region * $\sqrt{k/n}$. I will discuss the case of one extra parameter later, for this is arguably the most common situation.

When the bias difference is lower (* $\sqrt{k/n}$) in Figure 3, then the rules like BIC, which give a higher weight to simplicity, perform at close to optimal levels. The extreme case of this occurs at the far left when * 0 . A special case of this is when model *A* contains the true density, for then both *A* and *B* have zero bias, and therefore their bias difference is zero. For example, suppose the problem is to find *y* as a linear function of a potential regressors $x_1, x_2, \quad , x_r$, when in fact *y* is not correlated with any of these variables. AIC will tend to add about 15.37% of the variables unnecessarily. This is a mistake, although the size of the mistake will diminish as *n* increases because the coefficients of the added terms will tend to zero. Nevertheless, BIC will make less of a mistake because it will add fewer unnecessary regressors.

The problem for BIC is that one does not know whether one is in such a situation. If one did, then we could do even better than BIC by adding no regressors at all. But we do not. BIC is gambling everything on the *a priori* unlikely scenario that *y* has exactly zero correlation with all of $x_1, x_2, \quad , x_r$.

Note that *region* of the graph you are in will depend on *n* for although $\quad *$ is constant, whether $\quad *$ is less than, equal to, or greater than $\sqrt{\quad k/n}$ depends on *n*. In particular, as *n* increases, the

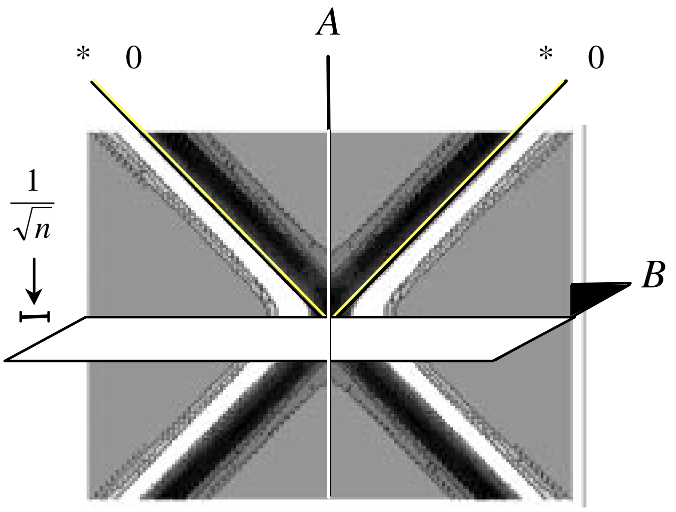

Figure 5: The relative performance of AIC versus BIC for the possible locations of the minimum discrepancy density in parameter space. The competition is between non-nested models *A* and *B*, where *B* has one more parameter than *A*. The white region is where AIC performs better, while the black region is where BIC performs better. The gray area is where they have the same performance. The standard deviation of the estimation error is shown on the left. *n* = 100 in this computation.

plots shrink in size, and the point at which the bias and the variance are the same moves to the left. So, if you imagine that you zoom in as the plots shrink, then effectively you move to the right as *n* increases. The only exception to this is when you are exactly at the far left, where $\quad * \quad 0$. But this is *a priori* so unlikely that one should not put much weight on it.

A related warning: Do not put undue weight on a single simulation result. Each simulated problem, like real problems, will have a fixed bias difference $\quad *$. For any two model selection methods, it is always possible to find a simulated problem which favors one over the other. *A single simulation is uninformative at best, and misleading at worse.*

When a pair of nested models differ by only one parameter, the situation is qualitatively the same as before (Figure 4). Quantitatively, there are some changes. The exchange between AIC and BIC is more favorable to BIC. The converse is also true. As the difference in parameters, $\quad k$, increases, the performance of AIC becomes closer and closer to the optimal, and the relative performance of BIC is less favorable.

I have also plotted the performance of a standard Neyman-Pearson hypothesis test ($p < .05$) in Figure 4. It performs between AIC and BIC because it gives a weight to simplicity that lies between AIC and BIC.

As I explained in a previous section, the bias/variance dilemma only really applies to the competition between nested models. So what about the performance of these criteria on non-nested models? The selection of models still involves considerations of bias and variance; however, the model with fewer parameters still has an expected advantage in variance, although it may have an advantage in bias as well.

The fundamental result is that AIC performs better than BIC only when the complex model has a bias advantage. BIC never performs better than AIC when the complex model has a bias advantage, except at some points near the intersection of *A* and *B*. As a piece of practical advice, if you expect a non-nested complex model to have an advantage in bias, then it is better to use AIC than BIC, especially if the difference in parameters is greater than one. If you know that the simpler model has a bias advantage, then use BIC.

The trouble with such advise is that it may not be practical. It presupposes that one has independent information about the location of minimum discrepancy density $*$. Where is this information obtained? It may come from other data that is thought to be similar in relevant respects. But if the information is about the position of $*$ then presumably it is because there is a common parameter applying to the modeling to both sets of data. In that case, the criteria should be applied globally to the composite model, in which case we still have no information independent of what the criteria applied more globally. However, it may be that the models have not been formally unified. At this pre-theoretical stage, we may have a reliable intuition about how the unification would work, in which case it would make sense to use it.

5 Unification and Generalization

At a conference connectionism for cognitive scientists held at Carleton University, May 1996, I presented a paper arguing for the importance of taking account of simplicity, unification and the consilience of inductions in solving prediction problems of any kind. The response initiated by David Rumelhart, and developed by Geoff Hinton, was that simplicity is not particularly important because it is only disappears from consideration when the sample size is sufficiently large. Indeed, this is confirmed by the results in the previous section, where AIC and BIC are equivalent to ML, which takes no account of simplicity. Yet I believe that the response is profoundly mistaken. Does that mean I have to abandon the framework and disown the results described in this paper? No, but I will need to extend the framework.

The first point to explain is that the discrepancy of a predictive density is always relative to a domain. Recall that the *K-L* discrepancy of a predictive density $f\ \ \mathrm{x}$ is

$$f \qquad \frac{1}{n}\ \ f^*\ \mathrm{x}\ \ \log f\ \ \mathrm{x}\ \ dx \quad const.,$$

where $\boldsymbol{x}$ is the set of all observable quantities under consideration and stands in for the adjustable parameters. In the depth perception example described in section 1 (see Cutting, this volume), $\boldsymbol{x}$ would involve the array of variables [x_S, x_H, x_O, x_P, D] and is the set of parameters *S*, *H*, *O*, *P*, and *B*. Exactly how the expression applies to this example is far from obvious. An actual experiment involves a sequence of stimuli followed by a measurement of the dependent variable *D*. Each stimuli is characterized by an assignment of value to the independent variables x_S, x_H, x_O, and x_P. If there are *n* stimuli presented to a subject, then the experiment is defined by a 4 by *n* matrix in which each row gives the four values of the independent variables encoding that stimulus. This matrix is the *experimental design matrix*, $\boldsymbol{X}_0$. The *response vector* $\boldsymbol{D}$ is a 1 by *n* matrix of the instances of the variable *D* in each of the *n* trials. $\boldsymbol{X}_0$ is a matrix of *numbers*, whereas $\boldsymbol{D}$ is a vector of *variables*. The probability density $f\ \ \mathrm{x}$ is a joint density $f\ \ \mathrm{X,D}$, where $\boldsymbol{X}$ is a matrix in which the numbers in X_0 are replaced by the corresponding random variables. However, the LIM and FLMP models only specify the conditional density $f\ \ \mathrm{D/X}$. In order to obtain the joint density, we need to add a probability distribution over the 16 possible values of the independent variables, $p_0\ \ \mathrm{X}$. However, if the probability $p_0\ \ \mathrm{X}$ to assign probability 1 to one to the array of actual values of $\boldsymbol{X}_0$ and 0 to all other possible values, then $p_0\ \ \mathrm{X}$ uniquely specifies the design matrix $\boldsymbol{X}_0$. The experimenter determines the design matrix, and, therefore, the experimenter determines the probability distribution $p_0\ \ \mathrm{X}$.

In this example, we need to substitute $f\ \ \mathrm{D/X}\ \ p_0\ \ \mathrm{X}$ for $f\ \ \mathrm{x}$ and $f^*\ \ \mathrm{D/X}\ \ p_0\ \ \mathrm{X}$ for $f^*\ \ \mathrm{x}$ in the formula for the *K-L* discrepancy. Because the logarithm of a product is the sum of

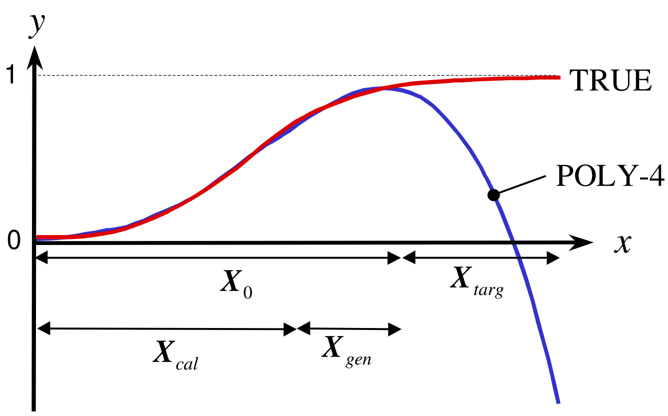

Figure 6: A 4-degree polynomial matches the true curve closely within the domain from which the data are drawn, but extrapolates poorly beyond that domain.

the logarithms, and an integration over p_0 X dX sets X X_0, the discrepancy simplifies to:

$$f \qquad f^*\; D/X_0 \;\; \log f \;\; D/X_0 \;\; dD \quad const\,.$$

Notice that the discrepancy is a function of the design matrix X_0. Let us explicitly record that fact by using the notation $_{X_0}$ f , or more simply $_0$ f . In model selection, I am assuming that the *goal* is the minimize the discrepancy of the predictive density. But why should the goal be limited to predictions *under the same experimental design as that used to collect the original data*? Scientists may be interested in predictions in other kinds of experiments, with a different design matrix X_{targ} (*targ* for "target"), in which case we would want to minimize the discrepancy $_{X_t}$ f , written $_t$ f .

How well do the standard methods of model selection optimize the accuracy of predictions within a different domain of prediction? The results of the previous section apply only to the case in which the target domain is chosen to be the same as the data domain (X_{targ} X_0). Hence, those results do not support the view that simplicity and unification disappear from consideration when the sample size is sufficiently large. All bets on that question are off. I believe that this is the problem of extrapolation raised by Busemeyer and Wang (this issue).

I shall develop the question in an easy curve-fitting example. Suppose that we want to predict values of *y* given values of *x*. The observed data are a set of (*x*, *y*) pairs collected for a range of *x*-values between 0 and 3.5 at intervals of .1. These *x*-values determine the experimental design matrix X_0. Now suppose that we are interested in predicting *y* for *x*-values between 3.5 and 5, which defines a different design matrix X_{targ}.

Figure 6 illustrates the problem with the extrapolation of a curve from one domain to another. A model consisting of 4-degree polynomials (POLY-4) fits the target curve (TRUE) extremely well within the domain X_0. However, the extrapolation of the POLY-4 curve to the target domain to X_{targ} is disastrous. In terms of the concepts developed in previous sections, the X_0 minimum discrepancy curve in POLY-4 has a very large discrepancy in the target domain X_{targ}. The problem is not that POLY-4 has no curve that fits the true curve in X_{targ}. Nor is the problem that POLY-4 has no curve capable fitting both domains at once. It may be that POLY-4 can *accommodate* new data in the target domain. The problem is that the model fails to *predict* the novel facts *in advance*.

One reaction to this phenomenon is to say: "Well, yes, there is always a risk in extrapolation. That is exactly what is shown by Hume's problem of induction, and more recently by the no-free-lunch theorems in machine learning (Wolpert, 1996). But the best we can do in the circumstances is to have faith in the uniformity of nature, and hope that fit in one domain will extend to fit in another. If the extrapolation does not work out, then there is nothing we can do about it. I believe that this response is mistaken.

Busemeyer and Wang (this volume) share my conviction. Their idea is that successful extrapolation in the past may be a useful indicator of further extrapolation, and they refer to their methodology as the *generalization criterion methodology*. The basic idea is nothing new: William Whewell (circa

1840), whose methodology Darwin claimed to have used in the *Origin of Species*, made the distinction between prediction of the same kind as prediction of a new kind and claimed that the latter is a particularly forceful and convincing kind of evidence for the truth of a theory. Whewell called this peculiar kind of evidence the *consilience of inductions* (Butts, 1989, p. 153).[5]

Such methods rest on the simple inductive argument that if extrapolation has been successful in the past, then extrapolation will be successful in the future. Of course, there is no guarantee that nature will cooperate in this regard; but for that matter, there are no guarantees for the success of predictions of any kind. The issue is not whether such an argument is fallible, but whether there are situations in which past extrapolation is useful indicator of future extrapolation, and whether this empirical information is not already exploited by the standard model selection criteria.

To investigate this issue, I extended the simulated prediction problem in Figure 6 to include six other models. Four of them were simplifications of POLY-4, whose equation is:

POLY-4: $y \quad {}_1 \quad {}_2x \quad {}_3x^2 \quad {}_4x^3 \quad {}_5x^3 \quad u$,

where u is a residue term whose probability distribution is normal with a fixed variance. The models CUBIC, PAR, and LIN are special cases of POLY-4 obtained by setting the appropriate parameters equal to zero, in the obvious way. LIN_0 is a further simplification of LIN obtained by setting ${}_1$ equal to zero. To these five models, I added two exponential models:

EXP: $y \quad \exp\left[\quad {}_1 x \quad {}_2 \quad {}^4 \right] \quad u$.

EXP^+: $y \quad {}_3 \quad \exp\left[\quad {}_1 x \quad {}_2 \quad {}^4 \right] \quad u$.

The formula for the true curve was:

TRUE: $y \quad ½ \quad ½\tanh x \quad 2$.

I assumed that there was "sufficiently large" number of seen data generated in the domain X_0 so that AIC, BIC, and ML make the same choices in every instance case. That is, I considered a situation in which Rumelhart and Hinton claim that simplicity would play no role. Assuming that the data are spread evenly over each domain at intervals of 0.1, Table 1 records the scores in three different domains shown in Figure 6 (ignore X_{cal} for the moment). Because I assumed zero variance for generating density, the scores cannot vary from one trial to the next. Moreover the scores are exactly equal to the model biases in the respective domains because there are no estimation errors.

The second column of Table 1 gives the AIC, BIC, and ML scores for the domain of the observed data. These numbers measure the how well the models fit the domain X_0 (lower numbers are better). The results are that the exponential models score better than all the polynomial models, but that within each group, the more complex models score better. This is exactly as expected, and it shows once again that simplicity has no advantage if X_0 is the target domain.

It is interesting to note that the order in the

k	Model	X_{cal}	X_0	X_0+X_{targ}
3	EXP^+	1.5	5.3	9.0
2	EXP	12.2	10.4	9.1
5	POLY-4	7.0	324.6	1,935
4	CUBIC	881	805	7,265
3	PAR	230	9,677	31,292
2	LIN	10,749	14,061	55,672
1	LIN_0	18,274	29,241	56,425

Table 1: This table shows the biases for the model in each domain. All values are squared deviations per datum mulitiplied by10^5. Lower numbers are better. The models are ordered by their biases in the data domain (second column). This is order in which they are favored by AIC, BIC, and ML, in the limit for large *n*.

5 In Forster (1988), I discuss how the consilience of inductions provides an empirical foundation for distinguishing component causes, and how it played a role in Newton's argument for universal gravitation. Forster (in preparation) argues that the same kind of evidence can provide an empirical foundation for the asymmetry of causal relationships in causal modeling.

third column of the first table is exactly the same. That is, a low relative discrepancy in the range from x 0 to x 3.5 is a success indicator of a low relative discrepancy in the range from x 0 to x 5. The fit in X_0 is a successful indicator of a model to *accommodate* the data over the extended domain. So, if that were the goal, then the standard methods would be useful in this example. But this is not the goal.

The goal is to approximate the truth in the target domain in advance of seeing any data from that domain. The success of the models at achieving this goal is indicated by their score in the second column of Table 2. Again, the exponential models are better than the polynomial models, but the order within each group is almost completely reversed. Except for CUBIC and PAR, the simpler models in each group extrapolate better. While the result in this simulation may not be general, it at least shows in one example that the standard model selection criteria are not reliable indicators of generalizability.

The third column in Table 2 orders the models in the same way, so it does not matter whether we consider the target domain to include X_0 or not.

The remaining question is the epistemological one. Is there any empirical indicator of generalizability of a model? Of course, there is always the wait-and-see method: Fit the curve in X_0 and wait and see how well it fits data from X_{targ}. But that is not the question. One wants to know whether there is any indicator of this obtainable *in advance* from X_0 alone. Future data is not accessible, and cannot be used in such a criterion. The generalization methodology is to divide X_0 into two sub-domains. Let X_{cal} is the *calibration* domain defined, say, by values of x from 0 to 2.5, while X_{gen} is the *generalization* domain defined by values of x ranging from 2.5 to 3.5. The idea is to fit the curve to X_{cal} and test its ability to predict the data in X_{gen}. This is like cross-validation except that there is only one subdivision of the data, and it chosen so that the direction of the test extrapolation is most likely to indicate the success of the wider extrapolation.

The test scores are shown in first of the three columns of Table 2. The score correctly indicates that the exponential models are better than any of the polynomial models, although the standard criteria got that much right as well. However, there are some improvements. Within the exponential models, it correctly indicates that EXP extrapolates better than EXP$^+$, whereas the standard criteria did not. And within the polynomial family of models, the ordering is better than that achieved by the standard criteria, though not perfect. It fails to predict how badly POLY-4 extrapolates. In particular, POLY-4 gets a better score than LIN_0, but its extrapolation is far worse. Also remember that when sample sizes are small, there will be an added problem with sample errors because the divided data sets have fewer data in them than the full data sets.

Nevertheless, it does appear that the generalization scores are providing us with useful empirical information that is not exploited by the standard selection criteria. Perhaps there are also situations in which the information is not only un-exploited,

k	Model	$\boldsymbol{X}_{cal}$ $\boldsymbol{X}_{gen}$	$\boldsymbol{X}_0$ $\boldsymbol{X}_{targ}$	$\boldsymbol{X}_0$ $\boldsymbol{X}_0+\boldsymbol{X}_{targ}$
2	EXP	10.4	6.1	9.1
3	EXP$^+$	29.3	10.4	10.4
1	LIN_0	2,779	3,519	1,677
2	LIN	818	8,220	2,851
4	CUBIC	5,260	12,133	3,822
3	PAR	8,683	24,830	7,953
5	POLY-4	1,486	84,795	26,608

Table 2: All scores are squared deviations per datum mulitiplied by10^5. Lower scores are better. The first column shows generalization test score, as determined from the observed data. The models are ordered by this score. The second column shows how well the best fitting density of each model extrapolates to its target domain. The models are ordered by their score in the second column. The third column shows how well the same density fits the true density over the whole domain.

but also relatively clear cut and decisive. Such information might at least supplement the standard criteria. It is also possible that the standard criteria should have an additional simplicity term, which does not decay to zero for large n.

Perhaps there are fairly general results to be found, but as in any young science, the road ahead is uncertain. The results should not be generalized and the suggestions are merely speculative. But they are also very interesting.

6 Summary and Conclusions

Model selection is a relatively new branch of mathematical statistics. The very basic concepts in the field are difficult and complicated to understand, but the framework shows promise in its ability to formulate some traditional problems of epistemology in a rigorous way.

The standard methods of model selection, like classical hypothesis testing, maximum likelihood, Bayes method, minimum description length, cross-validation and Akaike's information criterion, are able to compensate for the errors in the estimation of model parameters. The greater the number of parameters, the greater the compensation, which means that they tradeoff fit with simplicity if the simplicity of a model is measured by the paucity of parameters. They vary in the weight they give to simplicity, and my results show that whether that is good or bad depends on unseen features of the context of the problem at hand.

Although they are not all designed for this purpose, the standard criteria all share the feature that the weight given to simplicity diminishes as the number of available data increases. This is exactly what they should do if their purpose is to compensate for estimation error, for the estimation error diminishes to zero as the sample size tends to infinity.

In Jae Myung (this volume) nicely describes the goal mathematical modeling in cognitive psychology to "identify the model that most closely approximates the underlying process." I agree. However, there is an ambiguity in such a statement when there is a tradeoff between approximating the underlying process in one experimental domain and approximating it in another. This problem arises when the goal is to generalize a model to new domains of prediction outside the domain from which the current data was sampled. In that case, it appears that simplicity and unification may play a role that is quite independent of its effect on estimation error. While the problem of generalizability is rigorously described in the framework developed here, there are few results beyond some suggestive computer simulations (including one in this paper and in Busemeyer and Wang, this volume). This is an important new area of research that is relatively untouched in the articles in this special issue.

Appendix

Let $\boldsymbol{x}$ be a shorthand notation for n data, $x_1, x_2, \quad , x_n$, and let

$$f_{\,,}\ \mathrm{x} \quad f_{\,,}\ x_1, x_2, \quad , x_n \quad f_{\,,}\ x_1\ f_{\,,}\ x_2 \quad f_{\,,}\ x_n$$

be a probability density in a family B associated with the parameterized by $\ ,\ $. Assume that $f^*\ \mathrm{x}$ is the true density, which is not necessarily in B. Note that $\boldsymbol{x}$ is being treated as a random vector whose value is generated by the true density. Now let the maximum likelihood estimates for $\ ,\ $ to be and respectively. Note that they are functions of $\boldsymbol{x}$, so that and are also random variables governed by the true density $f^*\ \mathrm{x}$. Under asymptotic conditions, and are efficient (minimum variance) estimators of and , which implies that:

$$f_{\,,}\ \mathrm{x} \quad g_{\,,} \quad , \ h \quad {}_1, \ {}_2, \quad , {}_{n\ 2}\ |J|^{\ 1},$$

where $|J|$ is the determinant of the Jacobian of the transformation from $x_1, x_2, \quad , x_n$ to $\ ,\ ,\ {}_1, \ {}_2, \quad , {}_{n\ 2}$ (see Cramér, 1946, section 22.2), and h does not depend on the parameters and . This implies that the values of and are jointly sufficient to summarize all the information that is relevant about the data $\boldsymbol{x}$ to estimate and (Cramér, 1946, section 32.4). This is also that we need to know about the data to completely characterize the *K-L* discrepancies.

I assume that there exist values * and * of the parameters and such that $g_{*,*}$, is the true distribution of the maximum likelihood statistics and . Note that this does *not* imply that model *B* contains the true density, since *h* need not be the true density for the transformation variables $_1, _2, \ldots, _{n\,2}$.

Asymptotic conditions further imply that $g_{,}$, is a multivariate normal distribution whose first moments (means) are and , respectively, and whose second moments (covariance matrix) do not depend on or , at least over the range of values relevant to the present situation (which must include the values of and that have non-negligible probability, and those within the range of 0, since that is the value assumed in model *A*). This ensures that there exists a transformation of the parameters in which $g_{,}$, = $g \; g$, where g and g are normal distributions with means and , respectively, and variances equal to 1/*n*. In everything that follows, I will assume that and are the transformed parameters.

If we now substitute the above expression for $f_{,}$ x into the definition of the *K-L* discrepancy, and simplify, we obtain:

$$_{K\text{-}L} \; f_{,} \quad \frac{1}{n} \; g_{*} \quad \log g \quad d$$
$$\frac{1}{n} \; g_{*} \quad \log g \quad d \quad const.$$

Given that g and g are $N \; ,1/n$ and $N \; ,1/n$ respectively, this expression further simplifies to:

$$_{K\text{-}L} \; f_{,} \quad \frac{1}{2} \; g_{*} \qquad {}^{2} d$$
$$\frac{1}{2} \; g_{*} \qquad {}^{2} d \quad const.$$

A useful identity to use here is:

$$^{2} \qquad * \quad * \quad ^{2}$$
$$*^{2} \; 2 \quad * \quad * \qquad *^{2}$$

Using this identity, plus the analogous one for , plus some obvious evaluations of the integrals, and given that g_{*} and g_{*} are $N \; *,1/n$ and $N \; *,1/n$, we obtain:

$$2 \; _{K\text{-}L} \; f_{,} \qquad *^{2} \qquad *^{2} \quad const.$$

A simple corollary of this result is that * and * are, in fact, the minimum discrepancy values of the parameters. They are *not* necessarily the "true values" of the parameters, since we are not assuming that any density in the family is the true density.

Author Notes

This paper is derived from a presentation at the Methods of Model Selection symposium at Indiana University in August 1997, hosted by Prof. Rich Shiffrin. To all those involved in that symposium, and the ensuing special issue in the *Journal of Mathematical Psychology*, thank you for making it such a terrific success.

References

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

Akaike, H. (1974): “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control,* vol. AC-19: 716-23.

Akaike, H. (1977): “On the Entropy Maximization Principle.” P. R. Krishniah (ed.), *Applications of Statistics:* 27-41. Amsterdam: North-Holland.

Akaike, H. (1985): “Prediction and Entropy.” In A. C. Atkinson and S. E. Fienberg (eds.), *A Celebration of Statistics.* New York: Springer. 1-24.

Bamber, D. and Jan P. H. van Santen (1999): “How to tell whether a model is testable: The example of discrete-state selective-influence models,” *Journal of Mathematical Psychology.*

Bandyopadhyay, Prasanta, and R. Boik, P. Basu (1996): “The Curve Fitting Problem: A Bayesian Approach.”

Philosophy of Science **63** (supplement), S264-S272.

Browne, M. (1999): "Cross-validation methods." *Journal of Mathematical Psychology*.

Bruno, N. and Cutting, J. E. (1988): "Minimodularity and the perception of layout." *Journal of Experimental Psychology: General*, **117**, 161-170.

Busemeyer, J. R. and Yi-Min Wang (1999): "Model comparisons and model selections based on generalization test methodology," *Journal of Mathematical Psychology*.

Butts, Robert E. (ed.) (1989). *William Whewell: Theory of Scientific Method.* Hackett Publishing Company, Indianapolis/Cambridge.

Cartwright, Nancy (1979). "Causal Laws and Effective Strategies." *Noûs* **13**: 419-37.

Cartwright, Nancy (1983): *How the Laws of Physics Lie.* Oxford: Oxford University Press.

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

Cutting, J. E. (1999): "Accuracy, scope, and flexibility of models," *Journal of Mathematical Psychology*.

Forster, M. R. (1988): "Unification, Explanation, and the Composition of Causes in Newtonian Mechanics." *Studies in the History and Philosophy of Science* **19:** 55 - 101.

Forster, M. R. (1995): "Bayes and Bust: The Problem of Simplicity for a Probabilist's Approach to Confirmation." *The British Journal for the Philosophy of Science* **46:** 399-424.

Forster, M. R. (1999): "The New Science of Simplicity" in H. A. Keuzenkamp, M. McAleer, and A. Zellner (eds.) *Simplicity*, Cambridge: University of Cambridge Press.

Forster, M. R. (in preparation): "Causation, Prediction, and Accommodation," available online at http://philosophy.wisc.edu/forster/cause1.pdf .

Forster, M. R. and E. Sober (1994): "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions." *The 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**, pp. 1-58.

Golden, R. M. (1999): "Statistical Tests for Comparing Possibly Misspecified and Non-nested Models." *Journal of Mathematical Psychology.*

Grünwald, P. (1999): "Model Selection based on Minimum Description Length." *Journal of Mathematical Psychology*.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Kruse, M. B. (1997): "Variation and the Accuracy of Predictions." *British Journal for the Philosophy of Science* **48**: 181-193.

Kuhn, Thomas (1970): *The Structure of Scientific Revolutions,* Second Edition. Chicago: University of Chicago Press.

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

Massaro, D. W. (1988): "Ambiguity and perception in experimentation," *Journal of Experimental Psychology: General*, **117**, 417-421.

Myung, In Jae (1999): "The Importance of Complexity in Model Selection." *Journal of Mathematical Psychology*.

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

Schwarz, G. (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6:** 461-5.

Tversky, A. & Kahneman, D. (1971). "Belief in the law of small numbers." In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3-20). Cambridge: Cambridge University Press.

Wolpert, David H. (1996): "The lack of a priori distinctions between learning algorithms," *Neural Computation* **8**: 1341 - 1390.

Wasserman, L. (1999): "Bayesian model selection and model averaging." *Journal of Mathematical Psychology*.

Zucchini, W. (1999): "An Introduction to Model Selection." *Journal of Mathematical Psychology*.