

Key Concepts in Model Selection: Performance and Generalizability

Malcolm R. Forster

University of Wisconsin, Madison

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 paper is to explain some background concepts, to 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. © 2000 Academic Press

1. WHAT IS MODEL SELECTION?

William of Ockham (1285–1347/49) will always be remembered for his famous postulations of Ockham's razor (also spelled "Occam"), which states that "entities are not to be multiplied beyond necessity." In a similar vein, Sir Isaac Newton's first rule of hypothesizing instructs us that we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. While they

This paper is derived from a presentation at the Methods of Model Selection symposium at Indiana University in August 1997, hosted by Professor Rich Shiffrin. To all those involved in that symposium and this special issue, thank you for making it a success. As for this printed contribution, I thank my co-editors, In Jae Myung and Michael Browne, as well as co-contributors Hamparsum Bozdogan and Jerry Busemeyer, for their valuable comments on an earlier draft. Correspondence and reprint requests should be addressed to Department of Philosophy, University of Wisconsin, 600 North Park Street, 5185 Helen C. White Hall, Madison, WI 53706.

are wonderfully appealing, such rules are as vague as they are informal. How instructive are they? What does Newton mean by “sufficient to explain the appearances”? How closely does a set of equations need to fit the data in order to explain the appearances? If we increase the number of causes by adding an extra term to the equations, how do we know when this is unjustified by Newton’s criterion?

The problem is that extra terms add extra adjustable parameters, and these will *invariably* improve fit to *some* degree (since we can always obtain the same fit as before by putting the new adjustable parameters equal to zero, and if the nonzero parameters can fit some of the noise in the data, then the fit will be better). Does an extra term added to an equation count as “beyond necessity” if and only if the gain in fit is too small? If so, what counts as too small? How do we make this tradeoff between the addition of new parameters and the gain in fit? And what is the *rationale* behind such tradeoffs? What is gained by the tradeoff and why? These are the practical kinds of problems mathematical psychologists face everyday. They are issues that are addressed in a more theoretical way in the branch of mathematical statistics called model selection. Perhaps each group of researchers has something to learn from the other? That is the *raison d’être* of this special issue in the *Journal of Mathematical Psychology*.

In physics, one may think of a *model* as a set of equations containing adjustable parameters that is applied to a concrete physical system. When the construction of new models is strongly constrained by a background theory, it is common that a new model is the *extension* of an old one obtained by merely adding extra terms and extra parameters. In contrast, Cutting (2000) 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 three square panels laid out in depth with one behind another (Fig. 1). The experimenters manipulated four different sources of depth information,

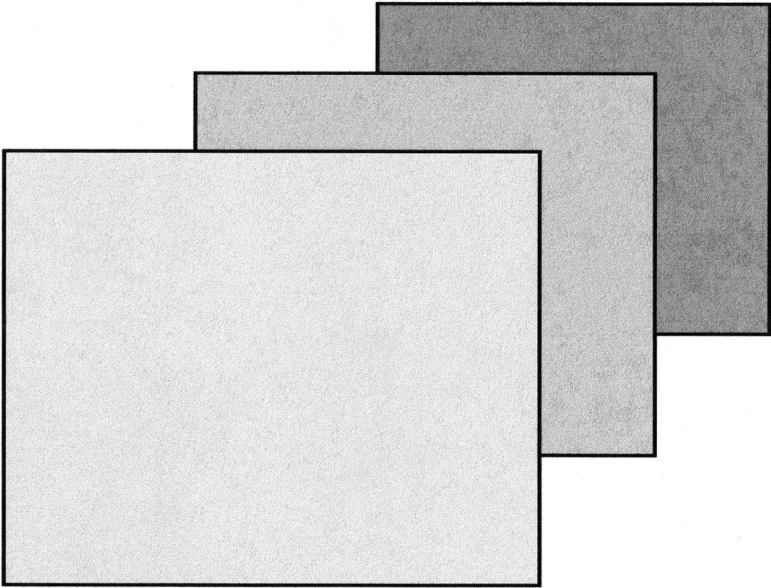


FIG. 1. A stimulus like the ones used in Bruno and Cutting’s experiment.

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 Fig. 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 express 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 are combined additively. They modeled a subject's response by a linear function of the form

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

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 nonzero 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 nonlinear model, which he called the fuzzy-logical model of perception (FLMP). Applied to the 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,

$$D = SHOPB / [SHOPB + (1 - S)(1 - H)(1 - O)(1 - P)(1 - B)].$$

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:

$$D = (1 - S)(1 - H)(1 - O)(1 - P) B / [(1 - S)(1 - H)(1 - O)(1 - P) B + SHOP(1 - B)].$$

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 nonnested models alike.

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 issue (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.

In Section 3, I will describe the model selection problem in terms of a tradeoff between the variance of parameter estimation and the initial bias of the competing models. This section is about what model selection criteria *ought to do*, and not about how well they do it. Section 4 describes some results of my own theoretical investigation into the *performance* of some of the model selection criteria described in Section 2. One result is that the performance of different methods depends 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*.

The context dependence of performance leads to an important lesson about simulations. Because 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, because they can be tailored to support any position.

Section 5 expands the scope of the paper by considering the performance of criteria with respect to *novel predictions*. This extends the concept of performance of the previous section to include errors of extrapolation or generalizability. Model selection criteria are standardly designed to maximize the fit with data that are resampled from the *same* generating distribution. When the new data are sampled from a different area or domain of the distribution, then, in principle, anything can happen, so one might think that nothing can be done about extrapolation errors that might arise. If this were correct, then science as we know it would be impossible. So, the issue arises: How well do the standard model selection methods (Section 2) perform when the aim of model selection is to generalize to new cases? Busemeyer and Wang (2000) argue that *all* the standard methods do worse than their alternative, which they call the *generalization criterion methodology*. According to some computer simulations described in Section 5, they are right in at least some cases. As far as I am aware, this is an open area of research. It is an advantage of the framework described here that it allows for the formulation of such questions.

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 error distribution. 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 as an equation for the *mean* value of the dependent variable.

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 $y = \theta x + u$, where *u* is an error term

governed by a fixed error distribution, the corresponding model M is a set of densities $\{f(x, y; \theta) \mid -\infty < \theta < \infty\}$.¹ If the error distribution itself contains an adjustable parameter, say a variance σ^2 , then it is one of the adjustable parameters characterizing the model $\{f(x, y; \theta, \sigma^2) \mid -\infty < \theta < \infty, 0 \leq \sigma^2\}$. However, parameters like σ^2 , which I will call *distributional parameters*, are less important because they do not usually characterize the structural properties of the system under study and will commonly vary from one application to the next. On the other hand, structural parameters, like θ , will tend to have greater predictive value. However, their measurement is determined by standard statistical methods, like the method of least squares, and so they are statistical parameters in that sense.

These same basic ideas generalize to the sets of equations with many dependent variables (vector equations) and many independent variables, to linear or nonlinear functional relationships, and to cases where different equations share common parameters. In physics, for example, the parameters typically play a key role in the background theory, like the half-lives of radioactive isotopes, electrical resistance, elasticity, mass, electric charge, the speed of light, initial velocities, and the fundamental physical constants. Others are less central, like the constants of integration.

A probability density represents a hypothesis which makes assertions about the statistical laws or regularities in the world. We say that the hypothesis is true if the probabilities it asserts to hold actually do hold in the world. Of course, it is never the complete truth. In this volume, the true density is alternatively called the *environmental density* (Golden, 2000) or the *operating model* (Zucchini, 2000) 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 and particular numerical assignments to the set of parameters of the model, at least on the condition that the model is identifiable (Bamber and van Santen, 2000). 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 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, forthcoming). In this essay, I shall also 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

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.

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 on 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 an intermediate step in the process.

Since the selection of a predictive density from a model involves the assignment of numerical values to all adjustable parameters, 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 set up, 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). In practice, fit is often defined in terms of the sum of squared deviations, which avoids the need for an explicit error distribution. As was first proven by Gauss (1777–1855), a ML estimate will give the same numerical result when a Gaussian error distribution is explicitly assumed, so ML is general enough to include the method of least squares as a special case. As a result, we shall treat model selection as choosing from amongst a set of maximum likelihood hypotheses selected from each of the competing models.

Not everyone in the field uses the terminology in the same way. Most importantly, Zucchini (2000) and Grünwald (2000) 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, their goal 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 authors of one view may mistakenly cite the advocates of the other view in their support.

2.1. The method of maximum likelihood. 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 (Forster, 1986). 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 the 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 of nested models, in which we decide whether to add or omit a single parameter θ . Omitting the parameter is effectively the same as setting it equal to

zero, so the choice is between the hypotheses $\theta = 0$ and $\theta \neq 0$. 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 \mathbf{x} be a variable ranging over possible data sets. The hypothesis $\theta = 0$ specifies a single density $f(\mathbf{x}; \theta = 0)$. Let $\hat{\theta}$ be the maximum likelihood estimate for θ . Then $\hat{\theta}$ is a function of \mathbf{x} whose probability distribution is determined by $\theta = 0$ to be $g(\hat{\theta}; \theta = 0)$. If $\theta = 0$ is chosen as the null hypothesis, then we may set up a 5% critical region, or rejection set, such that if $\hat{\theta}$ is sufficiently close to 0, then the null hypothesis is not rejected ($p < 0.05$, two tailed). Note that the null hypothesis would *always* be rejected if one went by fit alone because $\theta = \hat{\theta}$ is the best fitting hypothesis in the model $\theta \neq 0$, and it fits the observed facts better than $\theta = 0$. 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? One problem is that method does not extend straightforwardly to nonnested hypotheses. 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 their methods may provide an adequate implementation of Occam's razor 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, 1977, 1985). I will describe the K-L distance in greater detail later. Let \mathbf{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 $\log f(\mathbf{x}; \theta = \hat{\theta})/n - (k/n)$ is -2 times the K-L discrepancy under asymptotic conditions (which requires that the probability distribution of $\hat{\theta}$ is normal—the density $f(\mathbf{x}, \theta)$ need not be normal). This is Akaike's theorem (see Bozdogan, 2000, 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 estimated K-L discrepancy, which amounts to the maximization of $\log f(\mathbf{x}; \theta = \hat{\theta})/n - (k/n)$. The first term measures fit *per datum*, 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 (see Subsection 2.1). Note that the complexity penalty tends to zero for large data sets, so that complexity becomes relatively less important in the large sample limit.

What is the difference between AIC and classical hypothesis testing? One advantage is that AIC applies to nested and nonnested models. All that is required for the comparison of models is their maximum likelihood values, their k values, and the number of data. There is no need to say which model is the null hypothesis. The

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

other difference lies in their rationale. Akaike effectively trades off type I and type II errors in a principled way 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 $\hat{\theta}$ is a sufficient statistic for θ , which by definition means that $f(\mathbf{x}; \theta) = g(\hat{\theta}; \theta) h(\mathbf{x})$. Then AIC will reject the hypothesis $\theta = 0$ if and only if $\log g(\hat{\theta}; \theta = \hat{\theta}) - 1 > \log g(\hat{\theta}; \theta = 0)$. If $g(\hat{\theta}; \theta)$ is normal in the way required by Akaike's theorem, then $g(\hat{\theta}; \theta) = (2\pi\sigma^2)^{-1/2} \exp[-(\hat{\theta} - \theta)^2/2\sigma^2]$. Therefore, $\theta = 0$ will be rejected when $\hat{\theta}$ is far enough from 0 so that $(\hat{\theta} - 0)^2 > 2\sigma^2$. That is, $\theta = 0$ is rejected when $\hat{\theta}$ 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% ($p < 0.1573$, 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 as 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 examples, then the *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 is exactly correct? The most plausible view is that all models are idealizations of reality, 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 $\theta = 0$ is a good approximation, then an instance of the competing model $\theta \neq 0$ 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 $\theta = 0$ is a better approximation than $\theta = \hat{\theta}$. In that case, rejecting the null hypothesis would be a mistake. But it is probably not a big mistake. When the null hypothesis is approximately true, then $\hat{\theta}$ is probably close to 0.

Sometimes the rejection of an approximately true null hypothesis 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 task of a decision theory which takes social

utilities into account, and AIC is not trying to compete with such a theory. It is only designed to (fallibly) estimate the predictive accuracy of competing hypotheses. Therefore, 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 cause, which is the more socially serious part of the inference. AIC is neutral on that score.

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 $\theta = \hat{\theta}$ over $\theta = 0$, the only thing that is (provisionally) asserted is that the true value of θ is closer to $\hat{\theta}$ than it is to 0. This is a weaker assertion than concluding that the true value of θ is equal to $\hat{\theta}$. I am only defending AIC when it is *properly* understood and properly applied.

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 (Browne, 2000). 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 $\hat{\theta}_{cal}$, 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 an unbiased estimate of its performance with respect to *any* new data. However, the goal is to evaluate the predictive performance of $\hat{\theta}$, fitted to the whole data, and this is different from $\hat{\theta}_{cal}$. 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 whatsoever, 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 minimizing 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. Bayes' method says that models should be compared by their posterior probabilities (Wasserman, 2000). 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 (BIC). 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 $\log f(\mathbf{x}; \theta = \hat{\theta})/n - [\log(n)/2] k/n$. Note that BIC is similar to AIC except that it

gives greater weight to simplicity by a factor of $\log(n)/2$. Where classical hypothesis 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. To understand why this is so, it is important to make a distinction. I will say that model A is *truly* nested in model B when all the densities in A are also in B . For example, if A is represented by the equation $y = \beta_1 \cdot x$ and B by the equation $y = \beta_0 + \beta_1 \cdot x$, then A is a special case of B in which $\beta_0 = 0$. On the other hand, if we exclude the members of A from membership in B by explicitly specifying that $\beta_0 \neq 0$ in the definition of B , then we say that A is *quasi*-nested in B . It may seem that removing a set of measure zero from B should make no practical difference at all. But that is exactly why it is puzzling to learn that it makes a huge difference for Bayes' method.

Consider a competition between two truly nested models, A and B . If the true density is in A then the true density is in B . Therefore, the probability that A contains the true density, that is, the probability of A , can never be greater than the probability of B (Popper, 1959). It does not matter whether the probabilities are prior probabilities or whether they are posterior probabilities. Therefore, according to Bayes' method, it is *impossible* to favor the simpler model over the more complex model, and there is no implementation of Occam's razor *if the models are truly nested*! BIC does not have this consequence; therefore BIC is not an approximation of Bayes' method in this case.³

If Bayes' method is to be an implementation of Occam's razor, as Bayesians claim (Wasserman, 2000), then it is charitable to assume that they are referring only to the choice amongst quasi-nested models. This maneuver succeeds in restoring consistency to their claims. Nevertheless, it does not resolve the puzzle about why there *should* be any difference between truly nested and quasi-nested models. In the other methods of model selection, such as AIC or cross-validation, there is no difference between these two cases.

In the argument above, I stated that the probability of A is the probability that A is true. What if we understood it to be the probability that A is *approximately* true? To see that this cannot be right, note that if $\theta = 0$ is only approximately true, then there is a member of $\theta \neq 0$ that is also approximately true because there is a member of $\theta \neq 0$ that is infinitesimally close to $\theta = 0$. Therefore, the probability that $\theta = 0$ is approximately true is never greater than the probability that $\theta \neq 0$ is approximately true. This argument applies to truly nested and quasi-nested models, so Bayesians have to maintain that the probability of A is the probability that A contain the *exact* true density.

It should be clear from this discussion that Bayes' method and AIC optimize entirely different things. The contrast is especially dramatic in any 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 more probably true (obviously), and its choice is

³ See Forster and Sober (1994, Section 7). There are other problems with Schwarz's derivation described by Wasserman (2000), 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.

already made. In contrast, the work of AIC has only just begun. AIC is interested in which maximum likelihood hypothesis within each model is closest to the truth, and the truth of a complex model does not imply that it yields the best point hypothesis (Zucchini, 2000). AIC is not *primarily* concerned with selection of *models* at all!

Bayes' method does not select a point density from the winning model, although it does yield a predictive density; namely 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, conditional on the assumption 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 . Therefore, in that limit, the Bayesian predictive density converges to the maximum likelihood density, which is the same as AIC's choice (Wasserman, 2000). It is a simplification I will make when I compare the performance of the BIC and AIC in Section 4.

2.6. 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, 2000). 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.

I should also mention Bozdogan's ICOMP criterion (Bozdogan, 2000) in this subsection because it shares some of the same emphasis on the *descriptions* of models (see also footnote 4).

2.7. Limitations of all methods above. There are some caveats that apply equally to all methods. Regression toward the mean is a 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 tended to overestimate the probability that statistically significant results in one experiment will be repeated in a second experiment. Both of these are correctable mistakes.

However, *selection bias* is a related phenomenon which is not so clearly correctable. Zucchini (2000) points out that model criteria are especially risky when a selection is made from a large number of competing models. The random fluctuations in the data will increase the scores of some models more than others. The more models there are, the greater the chance that the winner won by luck rather than by merit.

The Bayesian method of model averaging (Wasserman, 2000) 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, 2000), 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.

Golden (2000) 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 the third option is an excellent recommendation.

Browne (2000) 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.

2.8. Generalization test methodology. The final, and perhaps the most serious, limitation of these methods is identified by Busemeyer and Wang (2000). If we think of model selection as using data sampled from one distribution in order to predict data sampled from another, then it is clear that all the methods listed above assume that the two distributions are the same. Yet in practice, predictions are sampled from a distribution that is an extension of the data generating distribution. This is most obvious in the case of time series models, where the goal is (usually) to *extrapolate* or *generalize* beyond the range of the observed data, but it also arises more generally. In such cases, errors of estimation arise not only from small sample fluctuations, but also from the failure of the sampled data to properly *represent* the domain of prediction. Standard methods of model selection are not designed to manage this second source of error.

Busemeyer and Wang (2000) introduce a method that they call the generalization test methodology as an attempt to address this problem. In brief, it is very similar to cross-validation, except that the calibration set and the test set (Subsection 2.4) are judiciously chosen to test for extrapolation properties. This is such an important issue that I have devoted Section 5 of this article to that topic.

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 point of view of Zucchini (2000) in which the aim of model selection is to select a predictive density that *best fits* the true density within a particular domain, where fit is defined in terms of a *discrepancy* function. The most important feature of discrepancy functions is that they define a *distance* (but not necessarily a metric) from the true density or the closeness to the true density. AIC adopts the K-L measure as its discrepancy function. I shall often use this as a concrete illustration, although the thrust of this paper is more general.

I have already made the point that focusing exclusively on models is a mistake. 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 nonnested case, the distinction may be well defined. Yet it is still unfaithful to the discrepancy idea, 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 conceptualizations.

The task of model selection 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 referred to as *discrepancy* (Zucchini, 2000). 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 are 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 is a combination of on the bias and the variance.

DEFINITION 1. Let θ_A^* be the member of model A that has the least discrepancy in the family, and let θ_B^* be the member of B with the least discrepancy. Let the discrepancy of θ_A^* with respect to the true density be called the *model bias* of A , the *approximation discrepancy* (Zucchini, 2000), 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 θ_A^* denote the density in A that best approximates the true density. Then θ_A^* is also in B (or there is a member of B that is arbitrarily close to θ_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 φ , while model B has one extra parameter θ . If both models are represented in the 2-dimensional space of points (φ, θ) then model A is the line $\theta=0$, while B is the whole plane. Let the member of A with the smallest discrepancy be $(\varphi_A^*, 0)$, while the member of B with the smallest discrepancy is (φ_B^*, θ^*) . Similarly, the maximum likelihood densities in

A and B are $(\hat{\varphi}_A, 0)$ and $(\hat{\varphi}_B, \hat{\theta})$, respectively. Note that in general $\varphi_A^* \neq \varphi_B^*$ and $\hat{\varphi}_A \neq \hat{\varphi}_B$. However, there is often a way of transforming the parameters (φ, θ) so that $\varphi_A^* = \varphi_B^*$, $\hat{\varphi}_A = \hat{\varphi}_B = \hat{\varphi}$, and $\hat{\theta}$ is stochastically independent of $\hat{\varphi}$ (remember that maximum likelihood estimates are functions of the data \mathbf{x} and are therefore random variables). Such transformations do exist asymptotically if certain regularity conditions hold (Cramér, 1946, chap. 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 θ .⁴

DEFINITION 2. The discrepancy difference between a maximum likelihood density and the minimum discrepancy density is called the *estimation discrepancy*. It is alternatively called the *estimation variance* or just the *variance*.

If the estimation discrepancy of both models is the same with respect to their shared parameter, φ , then the difference in estimation discrepancy is determined by the parameter θ . However, model A has no estimation discrepancy in θ because its minimum discrepancy value and its estimated value are both 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 . Therefore, the disadvantage of adding a parameter is that it will almost always increase the estimation discrepancy.*

DEFINITION 3 Let the *overall discrepancy* be the discrepancy between the true density and the probability density that is selected from a model by the best fit criterion.

The overall discrepancy is what model selection aims to minimize, but it is composed of the model bias, or the approximation discrepancy, and the estimation discrepancy. Proposition 1 tells us that the model bias decreases with the number of parameters, but Proposition 2 tells us that the estimation error increases with the number of parameters. 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 nonnested, the concepts of bias and variance still apply. The situation is illustrated in Fig. 2, which shows a 3-dimensional parameter space. Each point in that space assigns a particular set of numerical values to the three parameters and hence picks out a unique point hypothesis, or probability density. A one-parameter model, like A , might be the family of points on a line, while a 2-parameter family, like B , is the family of points on a plane. A is nested in B if and only if the line lies in the plane. In Fig. 2, B is not nested in A . I have also shown a special case in which the true density is in the same 3-dimensional space. Note that it is not in A or B , so A and B are both false. Under the same conditions mentioned in preparation for Proposition 2, a data set is also represented by a

⁴ Bozdogan (2000) has a different view about the relevance of such transformations. My assumption is that such transformations are merely a matter of mathematical convenience because it makes no difference to the discrepancy or the estimated discrepancy, since these are invariant under such transformations (Forster, 1999). However, Bozdogan's ICOMP criterion is not invariant under all transformations and is therefore built upon a different philosophical foundation.

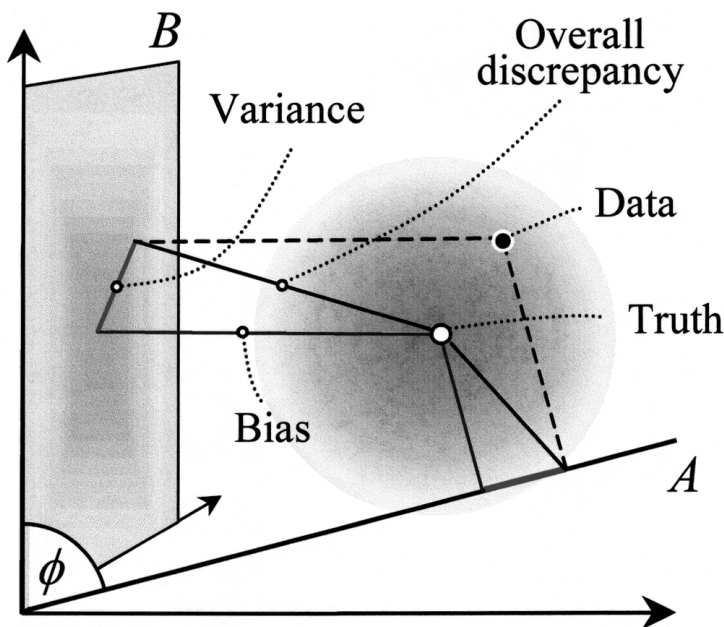


FIG. 2. Nonnested models. In parameter space, each point represents a point hypothesis. A one-parameter model like A is represented by a line, while a two-parameter model, like B , is represented by a plane in Bruno and Cutting's experiment.

point in parameter space; namely by the best fitting density in the space. The best fitting cases in A and B are then the orthogonal projections of that point onto those subspaces. Likewise, the least discrepancy cases are the orthogonal projections of the true hypothesis onto the same subspaces (see Sakamoto *et al.*, 1986; Bozdogan, 1987). The discrepancy between any two point hypotheses is the *square* of the distances between them in parameter space. As a consequence of Pythagoras's theorem, the overall discrepancy is equal to the sum of the approximation discrepancy (bias) and estimation discrepancy (variance).

In the nonnested case, there *need* be no bias–variance tradeoff—one model may have a clear-cut advantage on both counts. However, that does not mean that there is no model selection problem. In all cases, the problem is that bias and variance are unknown quantities. So how do we tell whether one model selection criterion is better than another at optimizing the overall discrepancy? Simulations may help, provided that they examine the full range of possible contexts.

4. HOW WELL DO THE METHODS WORK?

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

Consider a model selection problem between two models, A and B , with respect to an observed data set \mathbf{x}_0 generated by a true probability density, which one may

label θ^* . What I am about to say does not depend on whether the models are nested or nonnested. In such a context, there will be some fact of the matter about which of the two maximum likelihood densities has the least discrepancy (with θ^*) and by how much.

DEFINITION 4. If a criterion makes the correct choice, then the *performance gain* in any instance 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 expected performance gain on all possible instances generated by the true density, where the expectation is calculated relative to the true density. I will refer to the overall performance more briefly as the *performance*. Like the discrepancies themselves, there is a fact of the matter about the performance of any criterion in any particular context.

The problem is to understand how the performances of criteria depend on the context and how the performances of criteria compare in different contexts. Because the problem is hard, I have tackled it within a narrowed set of conditions. First, I consider only the Kullback–Leibler discrepancy function. Second, I assume that certain conditions that hold asymptotically in sample size under certain regularity conditions (Cramér, 1946, chap. 32 and 33) actually hold for finite data (at least to a sufficiently good approximation).⁵ I will simply refer to these as *asymptotic conditions* without *necessarily* implying that the sample size is large. In Section 5, I will investigate how well various model selection methods work under these two assumptions (Forster, forthcoming).

In the case of nested models,⁶ under the asymptotic conditions assumed here, the performance of any criterion is a function *only* of the *difference* in model bias, which I will denote by $\Delta bias$ (Forster, forthcoming). $\Delta bias$ is never negative because A can never be less biased than B in the nested case. There is a fact of matter about the difference in model bias between particular models A and B , and it does not depend on the observed data \mathbf{x}_0 , or on the sample size n . $\Delta bias$ is a property of the context—of the true density, the domain of prediction (Forster, forthcoming), and the models.

Figure 3 shows the performance of AIC and BIC plotted as a function of $\sqrt{\Delta bias}$, where $\sqrt{\Delta bias}$ is proportional to the distance between the least discrepancy cases in B and A in parameter space. Figure 3 shows the special case in which B has one more adjustable parameter than A . Instances in which B has more than one additional parameters show qualitatively similar features, except that AIC is closer to being optimal for all values of $\Delta bias$.

I have plotted the performances of AIC, NP ($p < 0.05$), and BIC, relative to the performance of the method of maximum likelihood (ML), which in the case of nested models, always selects the more complex of the two models. Remember that $\Delta bias$ is a constant for any given prediction problem, so you are located at some fixed point on the x -axis in Fig. 3. The expected value of the variances (estimation

⁵ The first half of Sakamoto *et al.* (1986) provides an introduction to this area of mathematical statistics, while the second half applies it to various kinds of model selection problems.

⁶ It makes no difference whether the models are truly nested or quasi-nested (see Subsection 2.5).

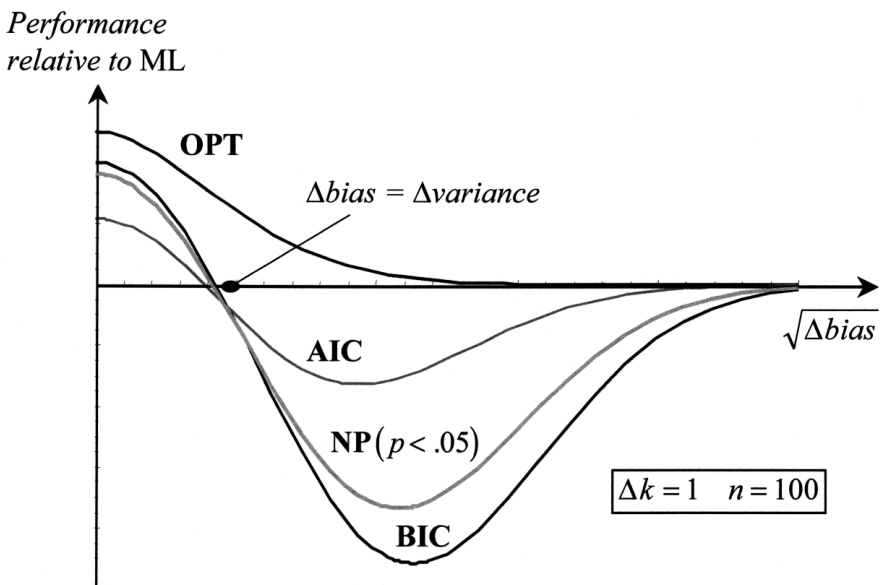


FIG. 3. Nested Models. The relative performance of AIC, BIC, and Neyman–Pearson hypothesis testing relative to ML, when one of two nested models has an extra parameter.

discrepancies) of each model is equal to k/n , where k is the number of degrees of freedom, which is usually the same as the number of adjustable parameters, and n is the number of data. Note that it depends only on k and n . The *difference* in the variance is given by $\Delta \text{variance} = \Delta k/n$. As we move from the left to the right along the x -axis, we move from a situation in which $\Delta \text{bias} < \Delta \text{variance}$, through the point at which $\Delta \text{bias} = \Delta \text{variance}$, onto the region in which $\Delta \text{bias} > \Delta \text{variance}$. Note that the relative performances depend dramatically on where you are located on this continuum. The region in which $\Delta \text{bias} < \Delta \text{variance}$ is quite different from the region in which $\Delta \text{bias} > \Delta \text{variance}$.

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

By saying that a rule is optimal at a point, I mean that a rule that chooses the predictive hypothesis with the least discrepancy *every time* would have a negligibly better performance. The performance of this optimal rule is labeled OPT in Fig. 3. Many people assume that ML is *never* optimal because it always chooses the most complex model in a nested sequence or hierarchy of models, such as all polynomials. 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 not whether a mistake is made, but the size of the mistake that matters in defining performance. A similar confusion often leads people to say that BIC performs better than AIC for large sample sizes, but this is not true if performance is defined as I have defined it here.

The slogan “AIC is inconsistent, but BIC is not” refers to the fact that AIC (as well as any classical NP 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 discrepancies of the selected predictive densities. It turns out that the *size* of the mistake made by AIC diminishes to zero in the limit of large n (although the convergence is slower than for BIC and other criteria—see Bozdogan, 1987). By “the size of the mistake” I do not mean the probability of type I error because that probability is a constant 15.73% (as I proved in Section 2). By “size” I mean “magnitude.” The *size* of the mistakes tend to zero even if their frequency does not. Since many theoreticians still maintain that there is a problem with the way that AIC behaves in the large sample limit, including those in the Akaike school (e.g., Bozdogan, 1987), one has to conclude that they are holding AIC to a different standard of performance. It is not always clear what that standard is or whether it is desirable.

When the bias difference is such that it balances the variance ($\Delta bias = \Delta variance$), then each model is as good as the other. It does not matter, on average, which model you choose, and therefore, it does not matter which selection criterion you use. The performance is not affected by how little or how much weight is given to simplicity. All methods are equivalent in performance, even though they may disagree on the choice of hypothesis. Even if one chooses a hypothesis randomly, or on purely *a priori* grounds (always the simplest, for example), then it does not matter in the end. In Fig. 3, this is why the performances of AIC, BIC, and ML are approximately the same at this point.

When the bias difference is higher ($\Delta bias > \Delta variance$), 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 in turn performs better than BIC.

When the bias difference is lower ($\Delta bias < \Delta variance$) in Fig. 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 $\Delta bias = 0$. 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 (although this is not the only way in which $\Delta bias = 0$ is true). For example, suppose the problem is to find y as a linear function of a potential regressors x_1, x_2, \dots, x_r , when in fact y is not correlated with any of these variables. AIC will tend to add about 15.73% 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 in which $\Delta bias = 0$. If one did, then we could do even better than BIC by adding no regressors at all. But we do not. BIC is gambling on the *a priori* unlikely scenario that y has exactly zero correlation with all of x_1, x_2, \dots, x_r .

Note that the *region* of the graph you are in will depend on n , for although $\Delta bias$ does not depend on n , whether $\Delta bias$ is less than, equal to, or greater than $\Delta k/n$ depends on n . In particular, as n increases, the 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 $\Delta bias = 0$. 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 $\Delta bias$. 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.*

As I explained in a previous section, the bias–variance *dilemma* applies only 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 results in Fig. 4 compare the performances of AIC and BIC in the case of two nonnested models (Fig. 2), such that A is orthogonal to B . Each point represents a possible location of the point hypothesis with the least discrepancy in the parameter space (Fig. 2). Label that point hypothesis θ^* . If the truth lies in the space, then it will be that hypothesis θ^* . Under the asymptotic conditions assumed in this section, the problem is the same that as if θ^* were the data generating density. The question is: If θ^* were located at a given point, which method, out of AIC and BIC, would perform best (on average) in that circumstance? If the point is in the white region, then the answer is that AIC performs better in that context. If the point is in the black

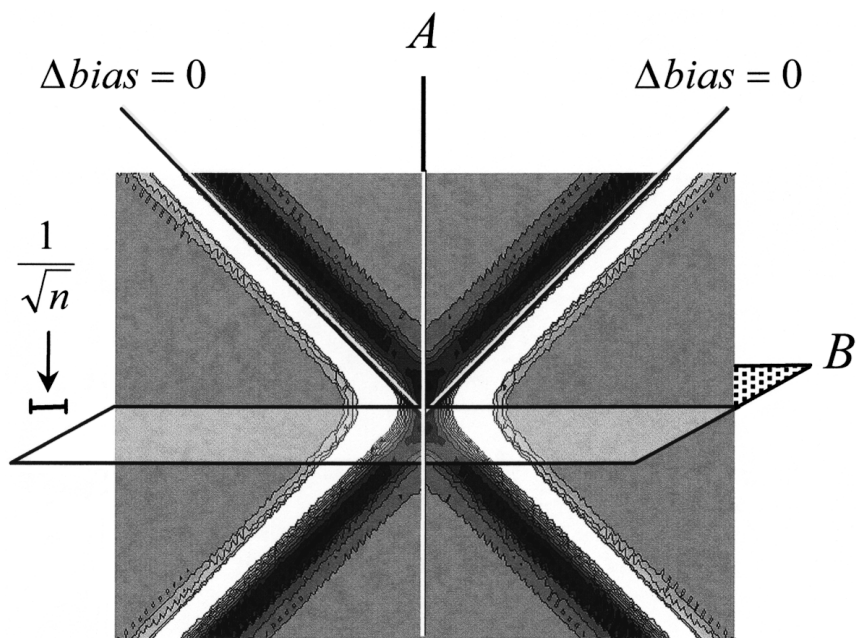


FIG. 4. Nonnested models. The relative performance of AIC versus BIC for the possible locations of the true density in parameter space. The competition is between nonnested models A and B , where B has one more parameter than A . The white region is where AIC performs better, while the black region is 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. In this computation, $n = 100$.

region, then BIC performs better in such a situation. If the point is in a neutral gray area, then their performances are the same. Inspection of Fig. 4 shows that the parameter space falls into two regions: (a) The region in which the more complex model has an advantage in bias over its simpler competitor. This is the standard situation in the case of nested models. Here AIC performs as well as or better than BIC, because AIC gives less weight to simplicity. However, there is a small exception to that when θ^* is very near the intersection of A and B . (b) The region in which the simpler model has the advantage in bias. Here BIC performs as well as or better than AIC because BIC gives more weight to simplicity. The variance appears to determine the thickness of the black or white strips.

What practical advice follows from these results? One can say that if a complex model has a bias advantage (i.e., is less biased), then it is better to use AIC than BIC. If you know that the simpler model has a bias advantage, then use BIC. The trouble with such advice is that it may not be practical, for it presupposes that one knows something about the location of θ^* . Where do we get this information? It may come from other data is thought to be similar in relevant respects. If there is some unifying connection between one data set and another, then presumably the two situations can be modeled in a global way, such that there is a common parameter applying to both sets of data. If the models have not been formally unified, then, at this pretheoretical stage, one may have a reliable intuition about how the unification would work. In that case, make use of the intuition.

The bottom line is that no single criterion can claim to work the best in all cases. Model selection criteria may be a useful guide to research, but it is not a substitute for careful thinking and common sense reasoning (Browne, 2000). The next section makes the same point in a more dramatic way.

5. UNIFICATION AND GENERALIZATION

At a conference on 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 (Butts, 1989; Forster, 1988) in solving prediction problems of any kind. A response initiated by David Rumelhart, and developed by Geoff Hinton, was that simplicity is not particularly important because it only disappears from consideration when the sample size is sufficiently large. Indeed, this is confirmed by the results in the previous section, where in that limit, AIC and BIC are equivalent to ML, which gives no weight to simplicity. Yet I continue to believe that the response is mistaken. Does that mean I have to abandon the framework and disown the results described in this paper? No, but I do need to extend the framework.

The first point I need to explain is that the discrepancy of a predictive density is always relative to a domain. Consider the Kullback–Leibler discrepancy of a predictive density $f_\theta(\mathbf{x})$,

$$\Delta(f_\theta) = -\frac{1}{n} \int f^*(\mathbf{x}) \log f_\theta(\mathbf{x}) d\mathbf{x} + \text{const.},$$

where \mathbf{x} is the set of all observable quantities under consideration and θ stands in for all the adjustable parameters. In the depth perception example described in Section 1 (see Cutting, 2000), \mathbf{x} would involve the array of variables $[x_S, x_H, x_O, x_P, D]$ and θ is the set of parameters S, H, O, P , and B . Exactly how this discrepancy 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 stimulus is characterized by an assignment of values 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 $4 \times n$ matrix in which each row gives the four values of the independent variables encoding that stimulus. This matrix is the *experimental design matrix*, \mathbf{X}_0 . The *response vector* \mathbf{D} is a $1 \times n$ matrix of the instances of the variable D in each of the n trials. \mathbf{X}_0 is a matrix of *numbers*, whereas \mathbf{D} is a vector of *variables*. The probability density $f_\theta(\mathbf{x})$ is a joint density $f_\theta(\mathbf{X}, \mathbf{D})$, where \mathbf{X} is a matrix in which the numbers in \mathbf{X}_0 are replaced by the corresponding random variables. However, the LIM and FLMP models only specify the *conditional* density $f_\theta(\mathbf{D}/\mathbf{X})$. In order to obtain the joint density, we need to add a probability distribution over the 16 possible values of the independent variables, $p_0(\mathbf{X})$. Now, $p_0(\mathbf{X})$ uniquely singles out the design matrix \mathbf{X}_0 only if $p_0(\mathbf{X})$ assigns probability of 1 to \mathbf{X}_0 and 0 to all other possible values. The experimenter determines the design matrix and thereby defines the probability distribution $p_0(\mathbf{X})$, correct?

If we substitute $f_\theta(\mathbf{D}/\mathbf{X}) p_0(\mathbf{X})$ for $f_\theta(\mathbf{x})$ and $f^*(\mathbf{D}/\mathbf{X}) p_0(\mathbf{X})$ for $f^*(\mathbf{x})$ in the formula for the K-L discrepancy, the K-L discrepancy simplifies to:

$$\Delta(f_\theta) = - \int f^*(\mathbf{D}/\mathbf{X}_0) \log f_\theta(\mathbf{D}/\mathbf{X}_0) d\mathbf{D} + \text{const.}$$

Note that the resulting discrepancy is a function of the design matrix \mathbf{X}_0 . Let us explicitly record that fact by using the notation $\Delta_{\mathbf{X}_0}(f_\theta)$, or more simply $\Delta_0(f_\theta)$. In model selection, I am assuming that the *goal* is to 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 \mathbf{X}_{targ} (*targ* for “target”), in which case we would want to minimize the discrepancy $\Delta_{\mathbf{X}_{\text{targ}}}(f_\theta)$, written $\Delta_{\text{targ}}(f_\theta)$.

How well do the standard methods of model selection optimize the accuracy of predictions, or minimize the expected discrepancies, 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 ($\mathbf{X}_{\text{targ}} = \mathbf{X}_0$). 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. This is the problem of extrapolation raised by Busemeyer and Wang (2000).

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 0.1. These x -values determine the experimental design matrix \mathbf{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 \mathbf{X}_{targ} .

Figure 5 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 \mathbf{X}_0 . However, the extrapolation of the POLY-4 curve to the target domain to \mathbf{X}_{targ} is disastrous. In terms of the concepts developed in previous sections, the \mathbf{X}_0 minimum discrepancy curve in POLY-4 has a very large discrepancy in the target domain \mathbf{X}_{targ} . The problem is not that POLY-4 has no curve that fits the true curve in \mathbf{X}_{targ} . Nor is the problem that POLY-4 has no curve capable of fitting both domains at once. POLY-4 is able to *accommodate* new data in the target domain. The problem is that the model fails to *predict* the novel facts.

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). The best we can do in the circumstances, according to this point of view, 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 (2000) share this conviction (see also Kruse, 1997). 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 predictions of the same kind and predictions of a new kind and claimed that the latter are a particularly forceful and convincing kind of evidence for the truth of a theory. Whewell called this particular kind of evidence the *consilience of inductions* (Butts, 1989, p. 153).⁷

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 a 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 Fig. 5 to include six other models. Four of them were simplifications of POLY-4, whose equation is:

$$\text{POLY-4: } y = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3 + \theta_5 x^4 + u,$$

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

⁷ In Forster (1998), I discuss how the consilience of inductions provides an empirical foundation for distinguish component causes and how it played a role in Newton’s argument for universal gravitation. Forster (in preparation) argues further that the same kind of evidence can provide an empirical foundation for the asymmetry of causal relationships in causal modeling.

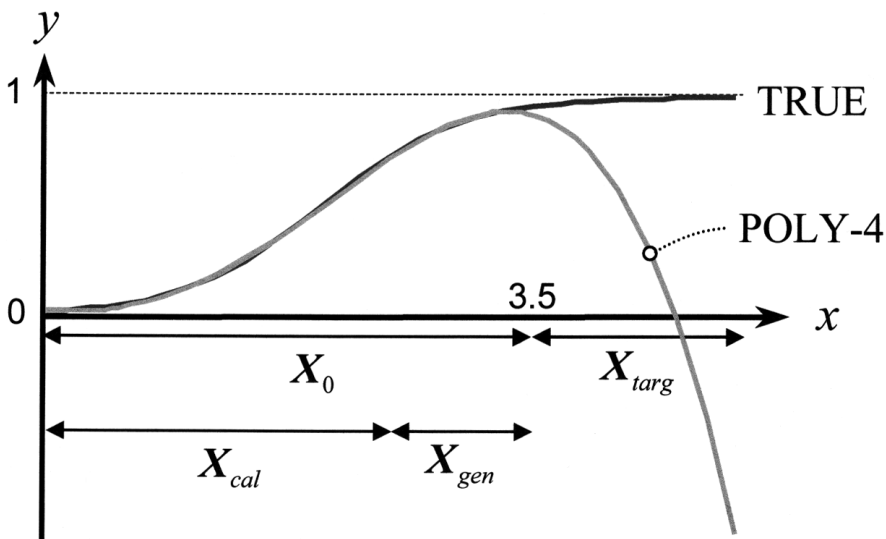


FIG. 5. A 4-degree polynomial matches the true curve closely within the domain from which the data are drawn, but extrapolates poorly beyond that domain.

by setting the appropriate parameters equal to zero, in the obvious way. LIN_0 is a further simplification of LIN obtained by setting θ_1 equal to zero. To these five models, I added two exponential models:

$$\text{EXP: } y = \exp[-\theta_1(x - \theta_2)^4] + u.$$

$$\text{EXP}^+: y = \theta_3 + \exp[-\theta_1(x - \theta_2)^4] + u.$$

The formula for the true curve was:

$$\text{TRUE: } y = \frac{1}{2} + \frac{1}{2} \tanh(x - 2).$$

I assumed that there was a sufficiently large number of seen data generated in the domain \mathbf{X}_0 so that AIC, BIC, and ML make the same choices in every instance. That is, I am considering a situation in which Rumelhart and Hinton claim that simplicity would play no role because there are no errors arising from small sample sizes. However, there is an important distinction to be made between sampling errors arising from small sample sizes and those arising from *unrepresentative* samples. 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 Fig. 5 (ignore the \mathbf{X}_{cal} column for the moment). Because I assumed zero variance for the 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 from the right of Table 1 gives the AIC, BIC, and ML scores for the domain of the observed data. These numbers measure how well the models fit the domain \mathbf{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 \mathbf{X}_0 is the target domain.

TABLE 1
Biases for the Model in Each Domain

k	Model	X_{cal}	X_0	$X_0 + X_{targ}$
1	LIN_0	18,274	29,241	56,425
2	LIN	10,479	14,061	55,672
3	PAR	230	9,677	31,292
4	CUBIC	881	805	7,265
5	POLY-4	7.0	324.6	1,935
2	EXP	12.2	10.4	9.1
3	EXP ⁺	1.5	5.3	9.0

It is interesting to note that the order in the rightmost column of the Table 1 is exactly the same as the second column from the right. That is, a low discrepancy in the range from $x=0$ to $x=3.5$ is a successful indicator of a low discrepancy in the range from $x=0$ to $x=5$. The fit in X_0 is successfully predicting the model's ability to *accommodate* the data over the extended domain. So, if that were the goal, then the standard methods such as AIC, BIC, CV, and ML would be useful in this example. But this is not the goal. Once again, the goal is *prediction*, not accommodation.

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 rightmost columns 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 within each group generalize 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 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

TABLE 2

k	Model	$X_{cal} \rightarrow X_{gen}$	$X_0 \rightarrow X_{targ}$	$X_0 \rightarrow X_0 + X_{targ}$
1	LIN_0	2,779	3,519	1,677
2	LIN	818	8,220	2,851
3	PAR	8,683	24,830	7,953
4	CUBIC	5,260	12,133	3,822
5	POLY-4	1,486	84,795	26,608
2	EXP	10.4	6.1	9.1
3	EXP ⁺	29.3	10.4	10.4

Note. All scores are squared deviations per datum multiplied by 10^5 . Lower scores are better. The first column shows generalization test scores, as determined from the observed data. The second column from the right shows how well the best fitting density of each model extrapolates to its target domain. The rightmost column shows how well the same density fits the true density over the whole domain.

that is not the question. One wants to know whether there is any indicator of this obtainable *in advance* from \mathbf{X}_0 alone. Future data are not accessible and cannot be used in such a criterion. The generalization methodology divides \mathbf{X}_0 into two subdomains. Let \mathbf{X}_{cal} be the *calibration* domain defined, say, by values of x from 0 to 2.5, while \mathbf{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 \mathbf{X}_{cal} and test its ability to predict the data in \mathbf{X}_{gen} . This is like cross-validation except that there is only one subdivision of the data, and it is 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 the first of the last 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 right as well. However, there are some striking improvements. Within the exponential models, it correctly indicates that EXP extrapolates better than EXP^+ , whereas the standard criteria did not. Within the polynomial family of models, the ordering is far better than that achieved by the standard criteria, though not perfect. On the minus side, it fails to predict just how badly POLY-4 extrapolates: POLY-4 gets a better score than LIN_0 , even though LIN_0 is the best at extrapolation and POLY-4 is the worst. On the plus side, the generalization method picks out LIN as the best polynomial model, and it does a pretty good job at extrapolation; far better than POLY-4, which is favored by the standard methods.

Overall, 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 unexploited, 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 no general results to be found, but as in any young science, the road ahead is uncertain. These 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 in the methodology of science 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 that 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.

According to Myung (2000), the goal of mathematical modeling in cognitive psychology is 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, 2000). This is an important new area of research.

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.
- Akaike, H. (1977). On the entropy maximization principle. In P. R. Krishniah (Ed.), *Applications of statistics* (pp. 27–41). Amsterdam: North-Holland.
- Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson and S. E. Fienberg (Eds.), *A celebration of statistics* (pp. 1–24). New York: Springer.
- Bamber, D., & van Santen, J. (2000). How to assess a model's testability and identifiability, *Journal of Mathematical Psychology*, **44**, 20–40.
- Bandyopadhyay, P., Boik, R., & Basu, P. (1996). The curve fitting problem: A Bayesian approach. *Philosophy of Science* **63** (supplement), S264–S272.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370.
- Bozdogan, H. (2000). Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology*, **44**, 62–91.
- Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.
- Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, **117**, 161–170.
- Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology, *Journal of Mathematical Psychology*, **44**, 171–189.
- Butts, R. E. (Ed.) (1989). *William Whewell: Theory of scientific method*. Indianapolis/Cambridge: Hackett.
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.
- Cutting, J. E. (2000). Accuracy, scope, and flexibility of models, *Journal of Mathematical Psychology*, **44**, 3–19.
- Forster, M. R. (1986). Statistical covariance as a measure of phylogenetic relationship, *Cladistics* **2**, 297–317.

- 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). Model selection in science: The problem of language variance. *The British Journal for the Philosophy of Science*, **50**, 83–102.
- Forster, M. R. (forthcoming). The new science of simplicity. In H. A. Keuzenkamp, M. McAleer, & A. Zellner (Eds.), *Simplicity*. Cambridge, UK: 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., & Sober, E. (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., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma, *Neural Computation* **4**, 1–58.
- Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, **44**, 153–170.
- Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133–152.
- Kahneman, D., & 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.
- Kullback, S., & Leibler, R. A. (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, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190–204.
- Popper, K. (1959), *The logic of scientific discovery*. London: Hutchinson.
- Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht: Kluwer Academic.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–465.
- 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, UK: Cambridge University Press.
- Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms, *Neural Computation* **8**, 1341–1390.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.
- Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41–61.

Received: July 13, 1998; revised: October 19, 1998