Model Identification from Many Candidates

Mark L. Taper

ABSTRACT

Model identification is a necessary component of modern science. Model misspecification is a major, if not the dominant, source of error in the quantification of most scientific evidence. Hypothesis tests have become the de facto standard for evidence in the bulk of scientific work. Consequently, because hypothesis tests require a single null and a single alternative hypothesis there has been a very strong tendency to restrict the number of models considered in an analysis to two. I discuss the information criteria approach to model identification. The information criteria approach can be thought of as an extension of the likelihood ratio approach to the case of multiple alternatives. However, it has been claimed that information criteria are "confused" by too many alternative models and that selection should occur among a limited set of models. I demonstrate that the information criteria approach can be extended to large sets of models. There is a tradeoff between in the amount of model detail that can be accurately captured and the number of models that can be considered. This tradeoff can be incorporated in modifications of the parameter penalty term.

HYPOTHESES, MODELS, AND SCIENCE

The hypothesis concept plays an important role in science. The classic scientific method (Popper, 1959) continually reiterates a cycle of hypothesis

This paper was first presented in August 1998 at a symposium titled “Scientific Evidence” hosted by the Ecological Statistics section of the Ecological Society of America. I thank Subhash Lele, Jay Rotella, Prasanta Bandyopadhyay, and Brad Shepard for commenting on drafts of this manuscript.
formulation, hypothesis testing, and hypothesis refinement. A hypothesis is a concrete description of the way the universe might be or operate. The statement that the world is flat is interesting as a scientific hypothesis only if it can be coupled with some alternative, which may be nothing more complex than that the world is not flat. A hypothesis is a construct of the human mind that organizes our understanding of the universe. To quantitatively assess a scientific hypothesis, it must be translated into a model. This paper refers to structural models and to specific models. Structural models are mathematical formulations of a hypothesis, including parameterization for error and noise. Specific models are structural models with specific parameter values (see Mayo, 2004 [chapter 4, this volume], and Miller and Frost, 2004 [chapter 8, this volume]).

For example, consider the scientific hypothesis that a population exhibits density-dependent growth. A structural model representing this hypothesis might be the Ricker dynamics equation for change in population size:

$$\ln \left( \frac{N_{t+1}}{N_t} \right) = a + bN_t + \varepsilon_t,$$

where $N_t$ is population size at time $t$, $a$ is the intrinsic or maximum growth rate, $b$ is an intraspecific competition parameter providing density dependence to the model, and $\varepsilon_t$ is a time-dependent stochastic noise term. A specific model for this hypothesis and structural model might be

$$\ln \left( \frac{N_t}{N_0} \right) = 0.75 - 0.003N_t + \varepsilon_t,$$

where $\varepsilon_t$ is normally distributed with mean 0 and variance 0.01.

THE NATURE OF EVIDENCE AND MODEL IDENTIFICATION

As human beings, we can fruitfully think about the world only in terms of constructs or models. The alternative of viewing the world in full detail is frighteningly paralytic if not impossible (Borges, 1967). As scientists, we are interested in the validity of our constructs. How do we know that our models are appropriate? This is an important question. Model misspecification is a major, if not the dominant, source of error in the quantification of most scientific analysis (Chatfield, 1995). As scientists, we are trained to seek evidence supporting or refuting our ideas. But what is evidence? We can think of evidence as information contained in data about the validity of our models. Unfortunately, when we ask “Is this model true?” we have begun a failed quest. The answer is always no! No model represents the real world exactly. It is better to ask, “Does this model do a better job for me than that one?” What you can do is compare models by asking questions, such as whether this hypothesis or model is more similar to underlying truth than that model. Evidence compares models, not only to one another, but also to an
unknown truth. In other words, we should consider evidence to be some measure for the models under consideration of the relative distance to the truth. Further, whatever evidence is, it should be measured on a continuous scale.

How can one choose an evidence measure? Unfortunately, it is unlikely that there is a single best evidence measure. Evidence measures should be adapted to the problems at hand. Consider the schematic map in figure 15.1 of towns near my home. Is Bozeman closer to Big Timber than it is to Gardiner? The answer depends on whether you are flying or driving. Insisting on "distance" always being geographic distance is not as useful as a more thoughtful consideration of the problem.

Royall (1997; 2004 [chapter 5, this volume]) proposes the likelihood ratio as the fundamental measure of the evidence for one model over another (see also Forster and Sober, 2004 [chapter 6, this volume]; Edwards, 1992). The likelihood ratio is expressed as \( P(X = x|A)/P(X = x|B) \). Here, \( P(X = x|A) \) and \( P(X = x|B) \) represent the probability of a data set \( X \) being equal to the observed data set \( x \) assuming models \( A \) and \( B \) respectively. Instead of assuming the likelihood ratio as a primitive postulate, Subhash Lele develops (2004 [chapter 7, this volume]) quantitative requirements for distance measures\(^1\) to be effective evidence measures. It is easy to see by expanding equation 1 that the likelihood ratio represents a relative distance measure and is thus a special case of Lele's evidence functions. Here, \( P(X = x|T) \) is the true probability of data set \( x \) occurring.

\[
\frac{P(X = x|A)}{P(X = x|B)} = \frac{P(X = x|A)}{P(X = x|T)} \frac{P(X = x|T)}{P(X = x|B)} \tag{1}
\]

1. Technically, these are discrepancy measures, as they are not required to be symmetric.
As Lele demonstrates, the likelihood ratio is in some sense an optimal evidence measure for comparing simple single-parameter models. However, interpretation of the likelihood ratio as evidence in more complex situations can become problematic. For example, with nested linear models, likelihood increases monotonically with the number of parameters. Consequently, the probability of misleading evidence can be made arbitrarily close to 1 by adding spurious covariates. This indicates that the number of parameters in the models being compared needs to be taken into consideration in an effective evidence measure (see Forster and Sober, 2004 [chapter 6, this volume]).

One measure that compares an approximating distribution (model) to a “true distribution” is the Kullback-Leibler distance. It is the expected log-likelihood ratio of the approximating model to the “true model.”

\[
K L = \int f(x) \cdot \log \left( \frac{f(x)}{g(x)} \right) dx,
\]

where \(f(x)\) is the true probability of \(x\) and \(g(x)\) is the probability of \(x\) under the model.

The KL distance itself cannot be an evidence measure because it draws no information from data. In addition, it requires that “truth” be known. Fortunately, the KL distance can be estimated from data (up to an unknown constant) by various information criteria (Forster and Sober, 2004 [chapter 6, this volume]). Burnham and Anderson (1998) provide a very readable introduction to the field. There are a variety of approximating forms; the AIC (Akaike, 1973), the AICc (Hurvich and Tsai, 1989), the SIC (Schwarz, 1978), the CAIC (Bozdogan, 1987), AIC_hq (Hannan and Quinn, 1979), and others (see table 15.1). All take the form of constant \(-2 \ln(L) + f(k, n)\), where \(L\) is the likelihood of a model given data, \(k\) is the number of parameters in the model, and \(n\) is the number of observations.

A single information criterion (IC) value is meaningless because the unknown constant cannot be evaluated. Nevertheless, information criteria are very useful for comparing models. A model is selected as the best model in a suite of models if it has the minimum value of the information criterion being utilized. The magnitudes of differences in the information criterion values, \(\Delta\text{IC}\), are used as measures of the strength of evidence for one model over another. We are relieved of the need to evaluate the unknown constant because the differencing of IC values eliminates it.
TABLE 15.1 Common information criteria. Minimum total discrepancy forms are designed to minimize prediction error, while order consistent forms attempt to correctly identify the underlying order of the data generating model.

<table>
<thead>
<tr>
<th>Class</th>
<th>Name</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum total discrepancy forms</td>
<td>AIC</td>
<td>$-2 \ln(L) + 2k$</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>$-2 \ln(L) + 2k \left( \frac{n}{n - k - 1} \right)$</td>
</tr>
<tr>
<td>order consistent forms</td>
<td>SIC</td>
<td>$-2 \ln(L) + \ln(n)k$</td>
</tr>
<tr>
<td></td>
<td>CAIC</td>
<td>$-2 \ln(L) + (\ln(n) + 1)k$</td>
</tr>
<tr>
<td></td>
<td>IC_{nq}</td>
<td>$-2 \ln(L) + c \ln(\ln(n))k, c &gt; 1$</td>
</tr>
</tbody>
</table>

A $\Delta$IC measure is related to the likelihood ratio measure supported by Royall (2004 [chapter 5, this volume]) by the fact that $\Delta$IC are penalized log-likelihood ratios ($\Delta$IC = $-2 \ln(L_1/L_2) + \Delta f(k, n)$) (see also Forster and Sober, 2004 [chapter 6]) with the penalty depending on both the order of the models and (except for the AIC) the number of observations. The $\Delta$IC between two models reduces to minus twice the log-likelihood ratio when the number of model parameters is the same.

One can get a feel for the interpretation of $\Delta$IC values by translating them to equivalent likelihood ratios and $P$-values. For this exercise, I am comparing models with a one-parameter difference, and using a two-sided test, as might happen in the construction of a confidence interval. A $\Delta$AIC of 2 is equivalent to a likelihood ratio of 7.4 and a realized $P$-value of .091, and a $\Delta$AIC of 3 is equivalent to a likelihood ratio of 12.2 and a realized $P$-value of .051.

COMPONENTS OF ERROR IN THE APPROXIMATION OF TRUTH

Linhart and Zucchini (1986), Bozdogan (1987), Forster (2000), and others have described two kinds of discrepancies in approximating truth with models. These are (1) the discrepancy due to approximation and (2) the discrepancy due to estimation. Even if estimation were not a problem, there is a limit as to how close a particular approximating model can get to the truth. This limitation yields discrepancies of the first type. Discrepancies of the second type are a result of errors in parameter estimation that cause the discrepancy to be greater than the minimum possible for a given model. The total discrepancy between truth and an identified and estimated model is of course due to both kinds of error.
Much of the controversy surrounding information criteria–based model identification stems from the relative importance placed on different kinds of error. Using simulations, Hooten (1995) has investigated the behavior of a suite of information criteria under a variety of model identification performance criteria. He found that no criterion is superior for all purposes.

Information criteria can be divided into two broad classes. Minimum total discrepancy (MTD) criteria such as the AIC and the AIC$_c$ explicitly try to minimize the total distance between truth and an identified and fitted model. Given this intent, it is not surprising that model identification using AIC has been shown to be asymptotically equivalent to model identification using jackknife cross-validation (Stone, 1977). On the other hand, the SIC, AIC$_{hq}$, CAIC, and other Order-Consistent (OC) criteria seek to estimate the “true model order.” These criteria focus on minimizing the discrepancy due to approximation.

Practice, simulation, and theory all indicate that the AIC “overfits” (Bhan­sali and Downham, 1977; Bozdogan, 1987; Hooten, 1995). Of course, there are objections to the use of the OC criteria as well. These have been reviewed forcefully by Burnham and Anderson (1998). Some of these objections are: (1) truth has infinite order, and thus there is no “true order” to identify; (2) in the same vein, if truth has infinite order, then overfitting is impossible; (3) the OC IC do not estimate the K-L distance; and (4) the OC IC underfit. Burnham and Anderson (1998) have also objected to the use of the AIC rather than the AIC$_c$ to represent the MTD criteria, as the AIC$_c$ includes an important sample size correction.

These criticisms are not as damning to the use of OC criteria as they might at first appear, because they can be countered. (1) Truth may have an infinite order, but one’s set of approximating models does not. There is in the set a model that could (ignoring estimation error) minimize the K-L distance between the approximating model and truth. (2) Overfitting is possible and potentially deleterious. I will discuss this at greater length shortly. (3) The OC IC do estimate the K-L distance, but there is a sense in which these estimates are not optimal. (4) Whether a selected model is considered underfit or overfit depends in part on the intended use of the model.

What is overfitting/underfitting? This is not a trivial question. It is not fruitful to define an underfitted model as one with fewer parameters than truth. If truth has infinite order, then all models are underfit. One could consider a model overfit if it contained more parameters than needed to minimize some criterion. This terminology may be useful if one is committed to a particular criterion but only confuses the issue when comparing criteria. In this paper, I do not describe a model as being overfit; instead, I enumer-
ate both the numbers of "overfitting errors" and "underfitting errors" that occur in the model. Overfitting error occurs if a variable not in truth is included in the approximation. Underfitting error occurs if a variable in truth is left out of the approximation. Both errors may occur in a single model.

MODELING GOALS AND SELECTION CRITERIA

Cox (1990) distinguishes three broad categories for models based on the purpose for which models are used. Lehmann (1990) creates a similar categorization. Cox classifies models as (1) empirical models, which are designed primarily for predictions; (2) mechanistic models, whose purpose is to explain observation in terms of processes; or (3) descriptive models, which are used to discover patterns that may suggest the application of other models, either empirical or mechanistic.

All of these can appear in either structural or specific forms. For empirical models, minimizing prediction errors is the goal, and so either the AIC or the AICc is the appropriate selection criterion. For mechanistic and descriptive models, OC criteria are probably more appropriate than MTD criteria. For example, in Zeng et al. (1998) our interest was in identifying structural models for population growth in natural time series, and the SIC was our selection criterion. On the other hand, in Kramer et al. (2001), a predictive model of forest blowdown was desired, and we utilized the AIC in model building.

However, in modern science models may often serve multiple purposes, so the appropriate technique may not be clear. A researcher planning to use a model for multiple purposes may choose to select the model using a criterion appropriate for the primary purpose or may reselect models for each use with the appropriate criterion.

FEW- VS. MANY-CANDIDATE MODELS

Another central tension in the model identification approach is between use of a large or small set of candidate models (Burnham and Anderson, 1998). Several arguments militate for using a small set of models: (1) the proportion of model misidentifications increases as the number of potential models increases (Hooten, 1995); and (2) in large candidate sets, there may not be a model with an intermediate number of parameters for which the AIC is minimized (Tong, 1990).
There are two ways that IC selection may be confused by many models. The first may be termed "selection indifference." If there are a number of models in the candidate set that are similar to the best structural model (the model that would be closest to truth if parameters were known), then sampling variability may lead to the selection of a model other than the best model. Further, if there are only subtle differences in the fits of candidate models, the best model may not be selected because of the parameter penalty. The selection of any model other than the true best structural model is generally considered a mistake when quantifying model identification procedures. However, these mistakes are not very pernicious because the selected models are good approximations (Hooten, 1995).

The second type of confusion may result from what is termed "model selection bias" (Zucchini, 2000). If a researcher applies a large number of models to a given data set, then by chance one may fit well without necessarily describing the underlying generative processes, or having strong predictive power (low prediction error sum of squares, or "PRESS," values). This is a general problem in inference, and applies to hypothesis testing–based inference, as well as model identification–based inference. The prohibition against the post hoc testing of observed patterns in data and Bonferroni adjustments of critical values for multiple comparisons are frequentist devices that have been developed to circumvent this problem in the hypothesis testing framework. Burnham and Anderson (1998) also warn against the perils of "data dredging" within the model identification context.

Some arguments for the use of many models are equally compelling as arguments for the use of compact set of models. For model selection by information criteria to work well, one needs to have a "good model" in the candidate set. The use of large number of models increases the chances that at least one of the candidates will be close to truth. In a sense, using a large number of models allows the data to speak.

One's attitude about the size of the candidate model set is influenced by how one feels the information criterion approach extends classical frequentist statistics. As discussed earlier, ΔIC can be related to hypothesis tests, and thus the information criteria approach can be thought of as an extension of the hypothesis test to the case of multiple alternative models. On the other hand, information criteria have been considered the "likelihood of the model" (Burnham and Anderson, 1998). Akaike (1973) felt that the AIC was an extension of the maximum likelihood principle, while years earlier Fisher (1936) speculated that eventually the functional form of truth might be derivable from the data.

If a researcher conceives of IC model identification as an extension of
hypothesis testing to more than a single null and alternative model, then a small set of candidate models will probably seem appropriate. On the other hand, researchers who view the information criterion value as the "likelihood of a model" will naturally gravitate to large sets of models. Maximum likelihood estimates for parameters are found not by comparing likelihoods for a few selected parameter values but by comparing likelihoods for all possible parameter values. This analogy indicates all potential models should be investigated.

Whatever one's view, large candidate model sets arise quite reasonably and frequently in ecological problems. The problem of very large candidate sets was recently brought to my attention when I was asked to aid in the analysis of a project designed to identify the influence of abiotic and biotic factors on abundance of stream-resident westslope cutthroat trout *Oncorhyncus clarki lewisi* in Montana streams (Shepard and Taper, 1998). The study estimated cutthroat and brook trout (*Salvelinus fontinalis*) densities at seventy-one sites for three years. At all sites, physical covariates such as elevation, slope, aspect, latitude, longitude, stream size, and bed type were measured. Also measured were several anthropic factors such as the levels of grazing, mining, logging, and road use in the drainage. Dimension reduction with principal components analysis reduced thirty-eight variables to eight factors. Nevertheless, after including interactions and quadratic terms, both of which are reasonable given the ecology of these fish, we were left with seventy-two predictor variables, which could be combined into a staggeringly large number of credible candidate models.

Two alternative approaches seem reasonable in dealing with large numbers of alternative models. The first is to increase the threshold for discerning between models. Subhash Lele is currently investigating how the probability of misleading evidence for a given threshold depends on the number of alternative models. A difficulty with this approach is that the researcher is likely to end up with a very large number of alternative models that are statistically indistinguishable. A second tactic is to increase the penalty imposed for complexity as a function of the number of alternative models. Below, I demonstrate the utility of this approach with a method that determines the appropriate parameter penalty empirically from the data.

Once the complexity penalty depends on the number of models in the candidate set, then the tension between small and large candidate model sets resolves itself into a tradeoff. An increased parameter penalty clearly reduces the complexity of selected models. If a researcher considers only a few models, the penalty for complexity need not be so severe and models of greater complexity can be identified. Why should a researcher give up the
ability to identify complexity? If one has considerable knowledge of the system, complex models can be selected from a small candidate set. On the other hand, a researcher uncertain of the fundamental workings of the system can avoid missing important effects by considering large candidate model sets. Thus, a researcher using only a few models retains the ability to detect model fine structure, while a researcher using many models reduces the probability of missing important effects.

A CLARIFYING EXAMPLE

My intention is to present an example complex enough to circumvent the criticisms directed at previous simulation tests of IC-based model selection, yet simple enough to be readily understandable. Some of the criticisms of simulation tests of IC model identification are (Burnham and Anderson, 1998) (1) that the models considered are too simple to represent the real complexities of nature; (2) that the true model is included in the candidate model set; and (3) that nature has tapering effects, but most simulation models do not.

With these criticisms in mind, I simulate a multiple regression problem not unlike the real data-analytic problems described above. The data for this exercise were constructed as

$$Y = \mu + \sum_{i=1}^{10} a_i \cdot X_i + \sum_{j=1}^{1000} b_j \cdot Z_j,$$

where $Y$ is a response vector of 3000 observations. The parameters $a_1$ and $b_1$ are set to 0.7 and 0.5 respectively. Subsequent parameters are given by the recursions; $a_{i+1} = a_i \cdot 0.7$ and $b_{j+1} = b_j \cdot 0.5$. The $X_i$ are 10 known covariate vectors. The $Z_j$ are 1000 covariate vectors treated as unknown in the analysis. Also included in the analysis are 30 $W$ vectors of spurious covariates that have nothing to do with the construction of the data. Thus, data construction was deterministic and not stochastic, truth has near infinite order, there are tapering effects, and full truth is not a candidate because most covariates are unknown.

The unknown portion was constructed deterministically for heuristic purposes. The sum of the 1000 independent $Z$ vectors is indistinguishable from a normal random vector. The ignorance of the observer may reduce a complex reality to a simple set of models. Thus, the early simple simulation studies disdained by Burnham and Anderson may be interesting after all.
The data were analyzed using all subsets regression of order less than or equal to 20 covariates plus an intercept term. Model selection and parameter estimation were undertaken on data subsets of 1500, 150, and 50 observations. Information criteria were compared on the basis of underfit and overfit errors and on prediction accuracy of the selected model. Predictions were made for the values of all remaining observations using the covariates included in each model. The accuracy of predictions is reported as the prediction $R^2$ ($PR^2 = \text{variance of predicted values divided by total variance of the validation set}$). The performance of the AIC, AICc, SIC, CAIC, and SIC(x) were compared.

SIC(x) is an ad hoc criterion created for this exercise, with adjustable parameter penalty falling into Bozdogan's class of OC criteria (Bozdogan, 1987). I define the SIC(x) as:

$$SIC(x) = -2 \ln(L) + (\ln(n) + x)k.$$  \hspace{1cm} (4)

SIC(0) is identical to SIC and SIC(1) is identical to CAIC. The SIC(x) was constructed to add flexibility in the complexity penalty to the commonly used SIC and CAIC. In retrospect, the $IC_{bq}$ of Hannan and Quinn (1979) could have been used as effectively because one can always choose a $c$ that equates the $IC_{bq}$ with the SIC(x).

This example has vast data-dredging potential, with $1.141 \times 10^{14}$ candidate models being considered. Unfortunately, as indicated above, one can't dismiss candidate sets this large as unrealistic.

The results are summarized in tables 15.2 through 15.4. With large amounts of data (table 15.2), all criteria investigated identified models with essentially identical prediction accuracy. Each of the minimum total discrepancy criteria included all true covariates and therefore made no underfit errors. However, use of the MTD criteria did lead to a number of overfitting errors.
TABLE 15.3 Results with training set of 150 observations.

<table>
<thead>
<tr>
<th>Selected by:</th>
<th>Order</th>
<th>Underfit errors</th>
<th>Overfit errors</th>
<th>PR² (%)</th>
<th>Variable list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimum</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>X1–X10</td>
</tr>
<tr>
<td>AIC</td>
<td>&gt;22</td>
<td>2</td>
<td>12</td>
<td>67</td>
<td>X1–X7, X10, 12W</td>
</tr>
<tr>
<td>AICc</td>
<td>16</td>
<td>2</td>
<td>6</td>
<td>74</td>
<td>X1–X7, X10, 6W</td>
</tr>
<tr>
<td>SIC</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>74</td>
<td>X1–X7, 1W</td>
</tr>
<tr>
<td>CAIC</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>74</td>
<td>X1–X7</td>
</tr>
<tr>
<td>SIC(2)</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>74</td>
<td>X1–X7</td>
</tr>
</tbody>
</table>

TABLE 15.4 Results with training set of 50 observations.

<table>
<thead>
<tr>
<th>Selected by:</th>
<th>Order</th>
<th>Underfit error</th>
<th>Overfit error</th>
<th>PR² (%)</th>
<th>Variable list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimum</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>X1–X10</td>
</tr>
<tr>
<td>AIC</td>
<td>&gt;22</td>
<td>2</td>
<td>12</td>
<td>17</td>
<td>X1–X7, 12W</td>
</tr>
<tr>
<td>AICc</td>
<td>&gt;22</td>
<td>2</td>
<td>12</td>
<td>17</td>
<td>X1–X8, 12W</td>
</tr>
</tbody>
</table>

Intervening rows have been omitted.

| SIC(3)       | >22   | 2              | 12            | 17      | X1–X8, 12W    |
| SIC(4)       | 5     | 7              | 0             | 67      | X1–X3         |

errors with the AIC making 9 overfit errors and the AICc making 5 overfit errors. In contrast, the OC SIC and CAIC both included 9 true covariates and no spurious covariates in their selected model, leading to one underfit error and no overfit errors.

When the training set is reduced to 150 observations (table 15.3), all selected models have a reduced accuracy. The models selected by the AIC have somewhat lower prediction accuracy than do models selected by OC criteria. However, the AICc, with its small sample size correction, selects a model with a prediction accuracy just as good as the models selected by the OC criterion. Both the AIC and the AICc make large numbers of overfit errors while the OC criteria make few, if any. The AIC selects a model of the maximum order allowed.

The most striking results can be seen in the small-sample case (table 15.4). Here, both the MTD and the standard OC criteria lead to massively overfit models with very low predictive ability until the adjustable parameter penalty is substantially raised. Both the AIC and the AICc selected models with
the maximum order allowed. With the SIC(4) selected model there are no overfitting errors, and predictive ability is only modestly lower than in the medium and large data size cases despite a large number of underfitting errors.

There may be a substantial cost in the use of small model sets. If an important covariate is missed there can be a dramatic drop in predictive as well as explanatory power. This can have a larger impact than the effect of underfitting errors forced by large suites of candidate models. In this example if covariate $X_1$ is eliminated from consideration, no identification technique can identify a model with predictive power greater than 35%.

**INFORMATION CRITERION VALUE CURVES**

Determining parameter penalty appropriate for use with large candidate model sets is not a fully solved problem. Nonetheless, it seems that information criterion value curves such as figure 15.2 can be used to diagnose the degree of parameter penalty needed to locate the real information in the data. Operationally, one can increase the parameter penalty until one finds a well-defined minimum. Furthermore, all of the curves, whether or not they have a minimum, seem to carry information about the appropriate model order in the shape of the curves. All curves have a noticeable change in slope at about the same order. This process is similar to determining the

---

**Figure 15.2** Graphically determining parameter penalty. IC curves for small sample size ($n = 50$). The parameter penalty increases from the lowest curve to the highest. The curves represent, from bottom to top, the AIC, SIC, CAIC, SIC(2), SIC(3), and SIC(4) information criteria.
number of factors to retain in a principal components analysis through the inspection of screeplots (Jolliffe, 1986).

CONCLUSIONS

In this chapter, I support the definition given by Subhash Lele (2004 [chapter 7, this volume]) of evidence as a data-based relative distance measure of models to truth. Because there are a many categories of models and many uses for models, it seems reasonable to use a variety of distance measures for choosing among models in different contexts. The minimum total discrepancy criteria and the order consistent criteria are designed to answer subtly different questions, and both have domains where their use is appropriate.

Model selection bias has restricted the effective use of large candidate model sets. I have proposed an ad hoc procedure that adjusts the parameter penalty in a fashion designed to determine from the data what level of model complexity is supportable by the data given the number of alternatives in the candidate model set. This device gives researchers another tool and another tradeoff about which investigators can make reasoned decisions. The few-versus-many-models choice seems less stark now. By adjusting the parameter penalty, the researcher can reasonably use large candidate model sets and avoid missing important effects, but with a cost in that the degree of model detail identifiable has been reduced.

Commentary

Isabella Verdinelli and Larry Wasserman

Professor Taper raises the interesting question of how to choose among many models. He also suggests the possibility of using the data to choose the form of the penalty function in penalty-based methods. These are indeed important and timely problems. We would like to draw attention to some recent results in the statistical literature. We give special attention to three developments: adaptive penalties based on metric entropy (Barron, Birgé, and Massart, 1999), adaptive minimax estimation of sparse vectors (Abramovich et al., 2000), and Bayesian methods. We conclude with an ex-
ample in which we examine Professor Taper’s suggestion for choosing a data-based penalty.

ADAPTIVE PENALTIES

The question that Professor Taper raises is how we choose a model when there are many candidate models. An extreme version of this problem occurs when there are infinitely many models. Many statistical methods are versions of this problem. For example, consider fitting a polynomial of order $p$ to a data set. Now $p$ can take any value in $\{1, 2, \ldots \}$, so in choosing the order of the polynomial we are really choosing a model from an infinite list of potential models.

Suppose we have i.i.d. data $Y_1, Y_2, \ldots, Y_n$ from an unknown probability density function $f(y)$. Consider an infinite list of models $M_1, M_2, \ldots$, where each model consists of a set of density functions indexed by a parameter:

$$M_j = \{f_{\theta}; \theta_j \in \Theta_j\}.$$  

Let $\hat{\theta}_j$ be the maximum likelihood estimator of $\theta_j$ in model $M_j$. If we choose model $M_j$, then we can estimate $f_\theta$ with $\hat{f}_j \equiv f_{\hat{\theta}_j}$. As Professor Taper notes, one way to select a model is to use a penalized likelihood (or penalized least squares in the regression case). Let $L(\theta_j) = \prod_{i=1}^n f_{\theta_j}(Y_i)$ be the likelihood function for the $j$th model and let $\ell_j = \log L(\hat{\theta}_j)$. Let $j$ maximize

$$\gamma_j = \ell_j - r_j,$$

where $r_j$ is the penalty term. (If there is no maximizer, it suffices to choose a model that is sufficiently to the supremum of $\gamma_j$ over $j$. For simplicity, we assume that the supremum does occur at some value $j$.) Let $\hat{f} \equiv \hat{f}_j$ be the corresponding estimate of $f_\theta$. A reasonable goal is to choose the penalties to make $\hat{f}$ as close to $f_\theta$ as possible. Barron, Birgé, and Massart (1999) have proved the following result. For an “appropriate choice” of $r_j$, there is a constant $C > 0$, such that

$$E(d^2(\hat{f}, \hat{f})) \leq C \inf_j \left\{ \inf_{g \in M_j} D(f_\theta, g) + \frac{r_j}{n} \right\},$$

where $d(f, g) = \{\int (\sqrt{f(y)} - \sqrt{g(y)})^2 dy\}^{1/2}$ is the Hellinger distance between $f$ and $g$ and $D(f, g) = \int f(y) \log f(y)/g(y) dy$ is the Kullback-Leibler
divergence. The result requires some explanation. The "appropriate choice" of \( r_j \) depends on the "metric entropy" of the model \( M_j \) that is a measure of the complexity of the model. Essentially, it is the dimension of the model. The \( E(\cdot) \) on the left-hand side of (2) means expectation (or average), so the result bounds the average distance of the estimator from the truth. The infimum on the right-hand side of (2) occurs at the \( j \) that best balances the approximation error of the model (the first term) and the penalty (the second term). This is like the usual bias and variance tradeoff. This is only an upper bound, but it can be shown that, in many cases, this bound is close to the optimal risk; it cannot be improved. This means that penalized likelihood, with appropriate penalties, is an optimal approach. What do these penalties \( r_j \) look like? This depends on the models under consideration, and they have been worked out in special cases by Barron, Birgé, and Massart (1999) and others. A full discussion involves some technicalities that we cannot go into here. But generally, the penalties tend to be of one of two forms. In one case, the penalty is of the form \( r_j = c \delta_j \), where \( \delta_j \) is the dimension of the parameter space \( \Theta_j \) and \( c \) is a carefully chosen constant. This is similar to AIC. If there are many models with the same dimension (think of model selection in regression), then \( r_j \) will typically also involve a \( \log n \) term and will be more similar to SIC. Hence, AIC and SIC are unified from this perspective. These theoretical results are, in some cases, not yet practical, but they do give important qualitative insight. In particular, we see that the choice of penalty depends on the richness of the class of models.

**SPARSE MODELS**

Some of the most exciting breakthroughs in model selection have appeared in a series of papers by Iain Johnstone and David Donoho and their colleagues. Let us briefly mention here some results from Abramovich, Benjamini, Donoho, and Johnstone (2000), from now on referred to as ABDJ. They study the following stylized model selection problem. Let \( Y_i \sim N(\theta, 1) \) for \( i = 1, \ldots, n \) where \( n \) is large. Let \( \theta = (\theta_1, \ldots, \theta_n) \) be the vector of parameters. We will consider submodels where some of the \( \theta_j \)'s are set to zero. Note that the number of parameters is the same as the number of data points. Thus, as the sample size increases, so does the number of parameters and the number of models. This is another way of formally capturing the "many models" problem. ABDJ were inspired to study this problem because it has a deep connection with nonparametric function estimation.

Because there are so many parameters, we have no hope of estimating \( \theta \)
well unless the vector is sparse. By sparse, we mean that many of the \( \theta_i \)'s are small. This is like the regression problem in Professor Taper's chapter, where we want to estimate a regression model with many potential regression coefficients. We cannot do well if all the regression coefficients are important. We assume instead that some might be big, but many are small. The problem is to find the big (important) coefficients.

As ABDJ note, there are many ways of measuring sparseness. The simplest is to assume that \( \| \theta \|_0 \) is small where \( \| \theta \|_0 \) is the number of elements of the vector \( \theta \) that are nonzero. Another way to measure the sparseness of \( \theta \) is with the \( \ell_p \) norm defined by

\[
\| \theta \|_p = \left( \sum_{i=1}^{n} |\theta_i|^p \right)^{1/p}
\]

When \( p \) is a small positive number, this norm provides a measure of sparseness: if \( \| \theta \|_p \) is small, then \( \theta \) must be sparse.

Let \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n) \) be an estimate of \( \theta \). Because we expect \( \theta \) to be sparse, it makes sense to set \( \hat{\theta}_i = 0 \) for some \( i \)'s. But we must decide which ones to set to 0. This, of course, is just model selection. A model \( M \) in this context is a subset of \( \{1, \ldots, n\} \), and selecting a particular model \( M \) will be interpreted as using an estimator \( \hat{\theta} \) whose elements are defined by

\[
\hat{\theta}_i = \begin{cases} 
Y_i & \text{if } i \in M \\
0 & \text{if } i \notin M
\end{cases}
\]

The question is how to select the model \( M \). We could try any of the various penalty methods mentioned in Taper. ABDJ show, instead, that a method based on false discovery rates (FDR) due to Benjamini and Hochberg (1995)—originally designed for multiple testing problems—can be adapted for this model selection problem. The procedure works as follows. Consider testing the hypothesis \( H_0 \) that \( \theta_i = 0 \). The \( P \)-value for that test is \( P_i = P(Z > |Y_i|) = 2(1 - \Phi(|Y_i|)) \) where \( Z \) has a standard normal distribution and \( \Phi \) is the standard normal cumulative distribution function. Let \( P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)} \) be the ordered \( P \)-values and let \( j \) be the largest integer such that \( P_{(j)} < \alpha j/n \). Take \( M = \{i; P_{(i)} \leq P_{(j)}\} \). This defines our model selection procedure. As noted above, this may also be thought of as a method for producing an estimate of \( \theta \) via (3), which we denote by \( \hat{\theta}(\alpha) \). We can measure how well \( \hat{\theta}(\alpha) \) estimates \( \theta \) by its minimax risk, defined by
where \( \Theta_n \) is the parameter space for \( \theta \) and \( r \) defines the type of loss function. For example, \( r = 2 \) gives the usual squared error loss.

The parameter \( \alpha \) has the following interpretation: every \( i \in M \) may be viewed as a rejection of the null hypothesis that \( \theta_i = 0 \); the expected fraction of rejections that are false (i.e., for which \( \theta_i = 0 \)) is bounded above by \( \alpha \). The fraction of false rejections is called the false discovery rate (FDR).

ABD] prove the following theorem that unifies hypothesis testing (which FDR was originally designed for) and estimation. Let \( \Theta_n = \{ \theta; n^{-1}\|\theta\|_p^p \leq a_n \} \), where \( \log^5 n/n \leq a_n \leq n^{-\delta} \) for \( \delta > 0 \). Let \( \hat{\theta}(\alpha_n) \) be the estimate based on the above procedure using level \( \alpha_n \) where \( \alpha_n \to 0 \). Then, for any \( 0 \leq p < r \leq 2 \), they proved that

\[
\sup_{\theta \in \Theta_n} E_\theta \| \theta - \hat{\theta}(\alpha_n) \|_r^r = \inf_{\theta \in \Theta_n} \sup_{\theta \in \Theta_n} E_\theta \| \theta - \hat{\theta} \|_r^r.
\]

Here, \( b_n \sim c_n \) means that \( b_n/c_n \to 1 \) as \( n \to \infty \) and the infimum on the right-hand side is over all possible estimators.

The result has the following interpretation. The right-hand side is the minimax risk that represents the best risk (expected loss). Any estimator must have a risk at least this big. The theorem says that the FDR procedure attains (asymptotically) this optimal risk. Moreover, it does so simultaneously for various values of \( p \) (which measures the sparseness) and \( r \) (which defines the loss function). This is a remarkable adaptivity result, and it suggests that the FDR procedure might be ideal for large-model selection procedures. ABDJ conjecture that the FDR procedure is similar to the penalty-based methods using a penalty of the form \( k \log(k/n) \), where \( k \) is the number of parameters in the model. This suggests (though it has not yet been proved) that such a penalty might be optimal in a wide variety of model selection problems.

**BAYESIAN METHODS**

Bayesian methods for model selection have also received much attention lately. Reviews include Kass and Raftery (1995) and Wasserman (2000). In
the Bayesian framework, one treats the model index $j$ as a parameter. One then places a prior on $j$ and for each $j$, one places a prior on the parameter $\theta_j$. Thus, the prior is of the form $\pi(j) \pi(\theta_j | j)$. It is then possible (sometimes with much computational burden) to find the model with highest posterior probability. One can also produce predictions that are obtained by averaging over the possible models with respect to their posterior probabilities.

Bayesian methods are attractive for their conceptual simplicity. However, it is important to understand the frequentist properties of these Bayesian methods. George and Foster (2000) have shown that a particular Bayesian model selection method has very good frequentist properties. ABDJ note that the George-Foster method seems to be related to the FDR method. Barron, Schervish, and Wasserman (1999), Ghosal, Ghosh, and van der Vaart (2000), Huang (2000), Shen and Wasserman (2000), and Zhao (1993, 1999) have shown that, if one places priors carefully over an infinite list of models, then the resulting posterior distribution has optimal frequentist performance.

The advantage of these Bayesian methods is that, in principle, the methods are very general. The disadvantage is that to ensure that the posterior has good frequentist properties—which strikes us as essential in complex problems—requires very carefully chosen priors. Indeed, the sharpest results so far appear to be those in Huang (2000) and the choice of priors used there was extremely delicate.

**CHOOSING PENALTIES: AN EXAMPLE**

Professor Taper suggests that one should plot the model selection scores for a variety of choices of the penalty. This is wise advice, especially given that most of the available theoretical results are large-sample results leaving the data analyst, who has only finitely many data points, with some reasonable doubt about optimality theorems. Professor Taper hints at choosing the penalty such that the resulting model score $\gamma_j$ definitively chooses a best model. This suggestion intrigued us, so we conducted a very small simulation study.

We generated $Y_i \sim N(\theta_i, 1)$, $i = 1, 2, \ldots, n$, with $n = 1,000$, and we used the estimator $\hat{\theta}$ proposed in (3). We chose the model $M$ by penalized likelihood, and we then plotted the model selection scores for the family of penalty functions

$$
\gamma_k = -2 \ln(L) + 2k \left(\frac{\ln(n)}{2}\right)^{\alpha}.
$$

(5)
Recalling that $\alpha = 0$ in (5) produces the SIC(0) score and $\alpha = 1$ gives that AIC score, we chose four values of $\alpha$: respectively 0, 0.5, 0.75, and 1. This is a slight variation of the criteria SIC($x$) suggested by Professor Taper, which would correspond to changing (5) to the form

$$\gamma_k(x) = -2 \ln(L) + 2k \left( \frac{\ln(n)}{2} + x \right)^\alpha.$$ 

Results from our simulation are shown in the four plots of Figure 15.1.1, which correspond to different levels of “sparseness” of the model’s parameters. The sparseness of the model is governed by the number of parameters
that are set to 0 in the simulation. We fixed $k (= 100, 150, 200, 300)$ of the $\theta_j$s to 5 while the remaining $(1,000 - k)$ $\theta_j$s were kept equal to 0.

All the model selection scores we considered consistently show low values in the neighborhood of the correct model order. However, note that if we choose $\alpha$ by finding the curve that sharply delineates a model, it seems that one is always led to $\alpha = 1$, and hence one would use SIC. Of course, this is only an example and without further investigation, we cannot make any definitive statement. But the example suggests that $\alpha$ may have to be chosen by some other criterion.

15.2 Commentary

Hamparsum Bozdogan

INTRODUCTION, AND A NEW INFORMATION MEASURE OF COMPLEXITY (ICOMP) CRITERION

Mark Taper raises several important issues and concerns. I wish to isolate these issues and briefly introduce what I call new generation information criteria, which I believe address these important issues data-adaptively under correct and misspecified models.

Taper's important and repeated theme throughout this chapter can be summarized as follows:

1. Tradeoff can be incorporated in modifications of the parameter penalty term that are sensitive to the size of the candidate model set. Taper proposes a model identification procedure that empirically determines an appropriate parameter penalty considering both the data and the candidate model set.
2. "Model misspecification is a major, if not the dominant, source of error in the quantification of most scientific analysis (Chatfield, 1995)."
3. "Why should a researcher give up the ability to identify complexity?"

This research was partially supported by the Scholarly Research Grant Program (SRGP) Award at the University of Tennessee during January 2000–August 2001. I extend my gratitude to Professor Mark Taper for inviting and encouraging me to make a commentary to his chapter and make a contribution to this important manuscript.
4. “Determining parameter penalty appropriate for use with large candidate model sets is not fully solved problem.”

5. “By adjusting the parameter penalty, the researcher can reasonably use large candidate model sets and avoid missing important effects, but with a cost in that the degree of model detail identifiable has been reduced.”

These important issues bring us to the point of “overall” model complexity. In general statistical modeling and model evaluation problems, the concept of model complexity plays an important role. At the philosophical level, complexity involves notions such as connectivity patterns and the interactions of model components. Without a measure of “overall” model complexity, prediction of model behavior and assessing model quality is difficult. This requires detailed statistical analysis and computation to choose the best-fitting model among a portfolio of competing models for a given finite sample. As I have argued with Akaike on many occasions during my visit to the Institute of Statistical Mathematics (ISM) in Tokyo, Japan, as a Visiting Research Fellow during 1988, in AIC and AIC-type criteria, counting and penalizing the number of parameters in a model is necessary but by no means sufficient. Model complexity in statistics depends intrinsically on many factors other than the model dimension, such as the several forms of parameter redundancy, parameter stability, random error structure of the model, and the linearity and nonlinearity of the parameters of the model, to mention a few (see, e.g., Bozdogan, 2000, 64).

The purpose of this short note and commentary is to affirm that we need to pay attention to the important issues listed above, and to develop and present information-theoretic ideas of a measure of overall model complexity in statistical estimation to help provide new approaches relevant to statistical inference.

As is well known, based on Akaike’s (1973) original AIC, many model selection procedures that take the form of a penalized likelihood (a negative log likelihood plus a penalty term) have been proposed (Sclove, 1987). For example, for AIC, this form is given by

$$AIC(k) = -2 \log L(\hat{\theta}_k) + 2m(k),$$

where $\hat{\theta}_k$ is the maximum likelihood estimator (MLE) of the parameter vector $\theta_k$ for model $M_k$, and $L(\hat{\theta}_k)$ is the maximized likelihood function, and $m(k)$ is the number of free parameters in $M_k$. According to Akaike (1987,
the accuracy of parameter estimates is measured by the expected log likelihood of the fitted model, where the expectation is taken with respect to the true data generating process. AIC is asymptotically unbiased estimator of minus twice this expected log likelihood.

Motivated by considerations similar to those in AIC, Rissanen's (1976) final estimation criterion (FEC) as well as the analytical extensions of AIC in Bozdogan (1987), Bozdogan (1988, 1990, 1994, 2000) introduced a new generation model selection criterion called ICOMP. “I” stands for “information,” and “COMP” stands for “complexity.” ICOMP criterion is based on the complexity of an element or set of random vectors via a generalization of van Emden's (1971) entropic covariance complexity index. Using an information-theoretic interpretation, ICOMP views complexity as the discrimination information of the joint distribution of the parameter estimates against the product of their marginal distributions. Discrimination information is equal to 0 if the distributions are identical and is positive otherwise (van Emden, 1971, 25). The most general form of ICOMP is based on the estimated inverse Fisher information matrix (IFIM) of the model. For a general multivariate linear or nonlinear model defined by statistical model = signal + noise, the ICOMP criterion for model \( M_k \) is defined by

\[
\text{ICOMP}(\text{IFIM}) = -2 \log L(\hat{\theta}_k) + 2 C_1(\hat{F}^{-1}(\hat{\theta}_k)),
\]

where \( \hat{F}^{-1}(\hat{\theta}_k) \) is the estimated IFIM under model \( M_k \), and

\[
C_1(\hat{F}^{-1}) = \frac{s}{2} \log \left[ \frac{\text{tr}(\hat{F}^{-1})}{s} \right] - \frac{1}{2} \log |\hat{F}^{-1}|
\]

is the maximal covariance complexity of \( \hat{F}^{-1}(\hat{\theta}_k) \), and \( s = \text{dim}(\hat{F}^{-1}) = \text{rank}(\hat{F}^{-1}) \). The term \( C_1(\hat{F}^{-1}) \) in (3) is an upper bound to van Emden's (1971) covariance complexity index, measuring both inequality among the variances and the contribution of the covariances in \( \hat{F}^{-1}(\hat{\theta}_k) \). Large values of \( C_1 \) indicate a high interaction between the parameter estimates. \( C_1 \) takes into account the fact that as we increase the number of free parameters in a model, the accuracy of the parameter estimates decreases. As preferred according to the principle of parsimony, ICOMP chooses simpler models that provide more accurate and efficient parameter estimates over more complex, overspecified models.

The first component of ICOMP in (2) measures the lack of fit of the model, and the second component measures the complexity of the esti-
mated IFIM, which gives a scalar measure of the celebrated Cramér-Rao lower bound matrix, which takes into account the accuracy of the estimated parameters and implicitly adjusts for the number of free parameters included in the model (see, e.g., Cramér, 1946, and Rao, 1945, 1947, 1948). According to ICOMP, the best model among a set of competitors is the one that minimizes ICOMP.

Under well-known regularity conditions, the MLE \( \hat{\theta} \) is asymptotically normally distributed with covariance matrix \( \hat{F}^{-1} \). In this case, \( C_1(\hat{F}^{-1}) \) measures the Kullback-Leibler (KL) (1951) information divergence against independence of the parameter estimates (Kullback, 1968; Harris, 1978; Theil and Fiebig, 1984) and ICOMP can be viewed as an approximation to the sum of two KL discrepancies. Similar to AIC, the first measures the KL divergence between the true data generating process and the fitted model. Beyond AIC, the second measures the KL divergence against independence of the parameter estimates.

With ICOMP, complexity is viewed not as the number of parameters in the model, but as the degree of interdependence (i.e., the correlational structure among the parameter estimates). By defining complexity in this way, ICOMP provides a more judicious penalty term than AIC, Rissanen's (1978, 1986) minimum description length (MDL), Schwarz's (1978) SBC (or ), and Bozdogan's (1987) consistent AIC (CAIC). Lack of parsimony and profusion of complexity are automatically adjusted by \( C_1(\hat{F}^{-1}) \) across the competing models in a portfolio. This measures the quantum of information contained in a particular likelihood function (see Ghosh, 1988, 34). Therefore, ICOMP has the attractive feature of implicitly adjusting for the number of parameters and for the sample size, and thus controlling the risks of both insufficiently paramaterized and overparameterized models.

Similar to AIC, ICOMP can be made invariant with respect to scaling and orthonormal transformations by using the correlational form of IFIM given by

\[
\text{ICOMP(IFIM)}_R = -2 \log L(\hat{\theta}_k) + 2C_1(\hat{F}^{-1}_R(\hat{\theta}_k)),
\]

where

\[
\hat{F}^{-1}_R(\hat{\theta}) = D_F^{-1/2}\hat{F}^{-1}D_F^{-1/2},
\]

and \( D = \text{Diag}(\text{IFIM}) \).

In this way, ICOMP becomes invariant to one-to-one transformations of the parameter estimates and more. In the literature, several authors such
as McQuarie and Tsai (1998, 367) and Burnham and Anderson (1998, 69), without reviewing the impact and the applications of ICOMP to many complex modeling problems, have erroneously interpreted the contribution of this novel approach over AIC and AIC-type criteria.

ICOMP FOR MISSPECIFIED MODELS

"Model misspecification is a major, if not the dominant, source of error in the quantification of most scientific analysis (Chatfield, 1995)." In this section, we generalize ICOMP above to the case of a misspecified model and develop ICOMP under misspecification. To our knowledge, in the literature, there do not appear to be any existing criteria that handle misspecification, except the Takeuchi's (1976) information criterion (TIC), or AICr. Our approach will provide new tools for researchers to determine for themselves whether or not the probability model is misspecified as we fit and evaluate them. This is very important in practice, but is often ignored. Furthermore, we will be able to do model evaluation in the presence of skewness and/or kurtosis, especially when we carry out subset selection of best predictors, detection of outliers, and so on in linear or nonlinear regression settings. Our approach will eliminate the serious errors made in inference when the usual standard techniques are used when there is skewness and/or kurtosis present in the data and normality is always assumed.

Following White (1982), we give a simple test of information matrix equivalence to check the misspecification of a model. We construct a consistent estimator of the covariance matrix $\text{Cov}(\hat{\theta}_\xi)$ given by

$$
\text{Cov}(\hat{\theta})_{\text{Misspec}} = \hat{F}^{-1}\hat{R}\hat{F}^{-1}.
$$

(6)

When a model is misspecified, then the two forms of the Fisher information matrices, that is, the inner- (\hat{F}^{-1}) and outer-product (\hat{R}) forms, are not the same. Thus, White's (1982) covariance matrix in (6) would impose greater complexity than the inverse Fisher information matrix (IFIM).

For example, in a regression model, there are a number of ways a researcher can misspecify the model. Some of these are discussed in Godfrey (1988, 100). The most common misspecification occurs:

1. when the functional form of the model is not correctly specified
2. when there are "near-linear" dependencies among the predictor vari-
Model Identification from Many Candidates

ables, which is known as the problem of multicollinearity in regression models

3. when there are high skewness and kurtosis in the variables, which causes nonnormality of the random error or disturbances

4. when there is autocorrelation and heteroscedasticity

Therefore, under misspecification, ICOMP is defined by

$$\text{ICOMP(IFIM)}_{\text{Misspec}} = -2 \log L(\hat{\theta}) + 2C_{1}(\widehat{\text{Cov}(\hat{\theta})}_{\text{Misspec}}).$$  \hspace{1cm} (7)

When we assume that the true distribution does not belong to the specified parametric family of pdfs, that is, if the parameter vector $\theta$ of the distribution is unknown and is estimated by the maximum likelihood method, then it is no longer true that the average of the maximized log likelihood converges to the expected value of the parameterized log likelihood. That is,

$$\frac{1}{n} \log L(x|\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\hat{\theta}) \rightarrow E_{x}[\log f(X|\hat{\theta})].$$  \hspace{1cm} (8)

In this case, the bias, $b$, between the average of the maximized log likelihood and the expected maximized log likelihood is given by

$$b = E_{G}\left[\frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\hat{\theta}) - \int_{R} \log f(x|\hat{\theta}) dG(x)\right]$$

$$= \frac{1}{n} \text{tr}(F^{-1}R) + O(n^{-1}).$$  \hspace{1cm} (9)

We note that $\text{tr}(\hat{F}^{-1}\hat{R}^{-1})$ is the well-known Lagrange-multiplier test statistic. See, for example, Takeuchi (1976), Hosking (1980), and Shibata (1989). Thus, we have the generalized Akaike's (1973) information criterion (GAIC) defined by

$$\text{GAIC} = -2 \log L(\hat{\theta}) + 2\text{tr}(\hat{F}^{-1}\hat{R}^{-1}).$$  \hspace{1cm} (10)

GAIC is the same as Takeuchi's (1976) information criterion, discussed above. For more details, see Konishi and Kitagawa (1996) and Konishi (1999).

The basic idea underlying the information matrix equivalence test (IMET) is that it relies on the equality between the two forms of the Fisher infor-
mation matrices. These are useful to check the misspecification of a model. So if \( \hat{F} = \hat{R} \), then the bias reduces to

\[
b = \frac{1}{n} k + O(n^{-2})
\] (11)

which gives AIC as a special case. In this respect, when the true model that is entertained is not in the model set considered, which is often the case in practice, AIC will confuse identification of the best-fitting model and it will not penalize the presence of skewness and kurtosis.

**ICOMP AS A BAYESIAN CRITERION IN MAXIMIZING A POSTERIOR EXPECTED UTILITY**

From Bozdogan and Haughton (1998), we define ICOMP as a Bayesian criterion close to maximizing a posterior expected utility given by

\[
\text{ICOMP}(\text{IFIM})_B = -2 \log L(\hat{\theta}_M) + k + 2C_1(\hat{F}^{-1}(\hat{\theta}_M)).
\] (12)

**OTHER NEW GENERATION MODEL SELECTION CRITERIA**

More recently, Bozdogan and Ueno (2000) extended Bozdogan's (1987) CAICF both in Bayesian and in frequentist frameworks. Their approach is based on a different estimation of the expectation of the quadratic variation in AIC. The result obtained includes both Akaike's approach and that of Schwarz (1978) and Rissanen (1978) as special cases. Further, it includes the additional term \( \text{tr}(\hat{F}^{-1}\hat{R}^{-1}) \). It generalizes to the case of model selection under misspecification. This new criterion is given by

\[
\text{CAICF}_{\text{Gen}} = -2 \log L(\hat{\theta}) + k \log(n) + \log|\hat{F}(\hat{\theta})| + 2 \text{tr}(\hat{F}^{-1}(\hat{\theta})\hat{R}(\hat{\theta})).
\] (13)

Further, we give an approximation to (13) that corrects the bias for small as well as large sample sizes:

\[
\text{CAICF}_c = -2 \log L(\hat{\theta}) + k \log(n) + \log|\hat{F}(\hat{\theta})| + 2 \left( \frac{nk}{n-k-2} \right).
\] (14)
Indeed as \( n \to \infty \), the term \( 2n/(n - k - 2) \) goes to 2, and CAICF reduces to CAICF; see Bozdogan (1987).

The Bayesian model selection (BMS) criterion of Bozdogan and Ueno (2000) is given by

\[
BMS = -2 \log L(\hat{\theta}) - 2 \log \pi(\hat{\theta}) + k \log(n) - k \log(2\pi) \\
+ \log |\hat{F}(\hat{\theta})| + tr\{F^{-1}(\hat{\theta})R(\hat{\theta})\}.
\]  

(15)

Dropping the constant term \(-k \log(2\pi)\) in (15), because it will not affect the comparisons of the models, BMS further simplifies to

\[
BMS = -2 \log L(\hat{\theta}) - 2 \log \pi(\hat{\theta}) + k \log(n) + \log |\hat{F}(\hat{\theta})| \\
+ tr\{F^{-1}(\hat{\theta})R(\hat{\theta})\}.
\]

(16)

For a constant prior \( \pi(\hat{\theta}) \), we have

\[
BMS = -2 \log L(\hat{\theta}) + k \log(n) + \log |\hat{F}(\hat{\theta})| + tr\{F^{-1}(\hat{\theta})R(\hat{\theta})\}.
\]

(17)

The minimum message length (MML) criterion of Wallace and Freeman (1987) is defined by

\[
MML_{87} = -2 \log L(\hat{\theta}) + k + \log |\hat{F}(\hat{\theta})| - 2 \log \pi(\hat{\theta}) + k \log(\kappa_k).
\]

(18)

In (18), \( \pi(\hat{\theta}) \) is the estimated prior distribution and \( \kappa_k \) is the quantification constant (Baxter, 1996).

**OBSERVATIONS**

The increase of \( |\hat{F}(\hat{\theta})| \) with the complexity of the models is not so obvious, since it depends on the type of models that are being entertained and fitted. When the observed data are i.i.d., we can write

\[
|\hat{F}(\hat{\theta})_k| = O(n^{1/2}).
\]

(19)

If we take the first two terms of \( \text{CAICF}_{\text{Gen}} \) or \( \text{CAICF}_c \), we can deduce the naive MDL of Rissanen (1978) and the SBC of Schwarz (1978):

\[
\text{MDL}/\text{SBC} = -2 \log L(\hat{\theta}_k) + k \log(n).
\]

(20)
We carried out a subset selection of variables in multiple regression analysis based on the following simulation protocol. Let

\[ X_1 = 10 + \varepsilon_1, \]
\[ X_2 = 10 + 0.3\varepsilon_1 + \alpha\varepsilon_2, \text{ where } \alpha = \sqrt{1 - 0.3^2} = \sqrt{0.91} = 0.9539 \]
\[ X_3 = 10 + 0.3\varepsilon + 0.5604\varepsilon + 0.8282\varepsilon_3, \]
\[ X_4 = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon_4, \]
\[ X_5 = -5 + 0.5x_1 + x_2 + 0.5\varepsilon_5, \]

where \( \varepsilon_i \) are independent and identically distributed, according to \( N(0, \sigma^2) \), for \( i = 1, 2, \ldots, n \) observations, and also \( \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5 \approx N(0, \sigma^2 = 1) \). The response variable is generated from the true model:

\[ y_{(nx1)} = [1, X_{or}]_{(nx4)}\beta_{(4x1)} + \varepsilon_{(nx1)}, \]

with

\[ X_{or} = [x_1, x_2, x_3], \]

where

\[ \beta_{4x1} = \begin{bmatrix} -8 \\ 1 \\ 0.5 \\ 0.3 \end{bmatrix} \]

\[ \varepsilon_{(100x1)} \sim N(0, \sigma^2I), \sigma^2 = 1. \]

Note that the generation of \( y \) does not contain variables \( x_4 \) and \( x_5 \), but \( x_4 \) and \( x_5 \) are implicitly related to the other variables. The multicollinearity is controlled by \( \alpha \), and indeed the generated model is misspecified. As the sample size gets larger, the correlation matrix of the predictors depicts more of the relationship among the variables, and thus makes it difficult for the model selection criteria to pick the true model that generated the data. Under this setup, we carried out subset selection of variables using \( y \), and \( X = [x_1, x_2, x_3, x_4, x_5] \). So there are \( 2^5 = 32 \) possible subset models including the constant term (constant model), and \( 2^5 - 1 = 31 \) possible subset no-constant models. In this case, we expect to pick \( X = [x_1, x_2, x_3] \) subset variable model as the best-fitting model over the saturated full model and other subset models,
and study the relative performance of the information criteria. All the computations are carried out using a new cutting-edge open architecture and self-learning computational toolboxes using informational model selection procedures in MATLAB developed by this author. These modules are available from the author. We now summarize our results in carrying out the subset selection of variables to determine the best predictors in 100 replications of the simulation experiment for different sample sizes.

Frequency of choosing the best subset regression model in 100 replications of the Monte Carlo experiment for \( n = 75 \): ICOMP(IFIM) chooses the subset \( \{0, 1, 2, 3, \ldots, \} \) 82 times; CAIC\(_{Gen} \), 70 times; BMS, 82 times; and AIC, 69 times. This indicates the confusion of AIC in misspecified and highly collinear models. For AIC, overfitting and underfitting are not well balanced. On the other hand, ICOMP balances the overfitting and underfitting better than CAIC\(_{Gen} \) and BMS in this experiment.

Frequency of choosing the best subset regression model in 100 replications of the Monte Carlo experiment for \( n = 100 \): ICOMP(IFIM) chooses the subset \( \{0, 1, 2, 3, \ldots, \} \) 91 times; CAIC\(_{Gen} \), 87 times; BMS, 92 times; and AIC, 73 times. This too indicates the confusion of AIC in misspecified and highly collinear models. In this case, AIC overfits the model more than any other criterion. Again, ICOMP balances the overfitting and underfitting better than CAIC\(_{Gen} \) and BMS.

Frequency of choosing the best subset regression model in 100 replications of the Monte Carlo experiment for \( n = 200 \): ICOMP(IFIM) chooses the subset \( \{0, 1, 2, 3, \ldots, \} \) 97 times; CAIC\(_{Gen} \), 100 times; BMS, 99 times; and AIC, 74 times. Once again AIC is confused in misspecified and highly collinear models. On the other hand, overfitting and underfitting disappear for ICOMP, CAIC\(_{Gen} \), and BMS as the sample size gets larger, which is the case (see, e.g., Bozdogan, 1987, on consistent criteria for more details).

We note that AIC’s performance has not improved beyond 70% compared to ICOMP(IFIM), CAIC\(_{Gen} \), and BMS in moderate to large samples despite the fact that the empirical power of AIC is about 84.27%. This is, again, due to the fact that counting and penalizing the number of parameters may be a necessary but not a sufficient condition to capture the entire model complexity. By measuring the degree of interaction between the parameter estimates via their measure of complexity of the covariance matrix, we are now able to determine not only which regression estimates are degraded by the presence of collinearity, but also which are not adversely affected. It is because of these problems that the tradeoff between the interaction of the parameter estimates of a model and the interaction of the residuals is very important in both linear and nonlinear models. Therefore, it is important to
take parameter interdependencies and accuracies into account in the model evaluation criterion. ICOMP(IFIM), CAIC\textsubscript{Gen}, and BMS do these for us data-adaptively. For more on this within the multiple regression context, we refer the reader to Bozdogan and Haughton (1998), where the asymptotic consistency properties of the class of ICOMP criteria are studied in detail, both when one of the models considered is the true model and when none of the models considered is the true model.

CONCLUSION

In this short note and commentary, we introduced several new generation model selection criteria that take into account misspecification of the model, data-adaptively modify the parameter penalty, determine the complexity of the model, and balance overfitting and underfitting risks more judiciously. ICOMP considers situations both with correlated and with uncorrelated model residuals by including dependence, and both linear and nonlinear model parameters. As it is formulated, ICOMP is very general in the sense that it can be easily extended to other distributional assumptions on the model structures. It requires minimal assumptions and measures the strength of the structure of a model. It represents relations such as similarity, dissimilarity, and higher-order correlations within the model. If the variables show neither similarity nor dissimilarity, then the complexity becomes 0 and one should not use ICOMP. The difference between ICOMP, AIC, and SBC/MDL is that with ICOMP we have the advantage of working with both biased and unbiased estimates of the parameters and measure the complexities of their covariances to study the robustness properties of different methods of parameter estimates. AIC and AIC-type criteria are based on MLEs, which often are biased, and they do not take into account the concept of parameter redundancy, accuracy, and the parameter interdependencies in the model fitting and selection process. In the literature (see, e.g., Li, Lewandowsky, and DeBrunner, 1996), a measure of a model's total sensitivity to all of its parameters is often defined in terms of the trace of FIM, and, in some cases, it is defined by the determinant of IFIM, called a generalized variance. Using such measures alone as performance measures has serious disadvantages to which one should pay attention. In concluding, we note that our Monte Carlo simulation of subset selection of variables in multiple regression analysis under collinearity of variables clearly demonstrates the superiority of ICOMP-class criteria to AIC in model selection, prediction, and perturbation studies. We believe the set of potentially fruitful applica-
tions of information-theoretic model selection criteria are vast in scientific investigations and in empirical work to discover the nature of scientific evidence (see, e.g., Bearse and Bozdogan, 1998). We hope that future research will continue to explore these avenues.

Finally, both robust and misspecification versions or forms of ICOMP have been developed by this author, and the results of this will be reported elsewhere.

---

### 15.3 Rejoinder

**Mark L. Taper**

I thank Drs. Verdinelli, Wasserman, and Bozdogan for their interesting and useful commentaries on my chapter.

Drs. Verdinelli and Wasserman present an example showing the breakdown of the SIC(x) in the analysis of model sets with sparse effects of uniform magnitude. As mentioned in my chapter, the SIC(x) was explicitly designed for models with tapered effects, but it is very useful to have the range of the SIC(x)'s utility clarified.

They have also brought us current on related articles in the statistical literature published since my chapter was written in 1998. Another important recent article not mentioned either in the commentary or in my chapter is Bai, Rao, and Wu (1999).

Bai, Rao, and Wu propose a class of general information criteria (GIC) that consider the data and all compared models in the construction of the penalty term. The SIC(x) that I proposed would fall into their class, with an additional restriction on the size of x relative to the data size.

Bai, Rao, and Wu compare the efficacy of a member of their class and find it superior to information criteria not taking the model set into consideration. I am gratified by this work. I do not consider the SIC(x) a definitive but only an ad hoc solution, whose primary purpose is to demonstrate the problem. I welcome further development by mathematical statisticians.

Although the effectiveness of the GIC and the SIC(x) have not been compared yet, I can point out one computational difference. The GIC requires the evaluation of all regressions in the model set, whereas the SIC(x), because it can take advantage of the "leaps and bounds" algorithm, does not. This may allow the SIC(x) to be employed in the analysis of larger model sets than the GIC.
I appreciate Dr. Bozdogan's recognition of the importance of the questions that I have raised. Further, I agree with him that a model's order (number of parameters) is only a crude index of its complexity. The information-based penalties that Bozdogan proposes should give better corrections for the parameter estimation component of the total discrepancy. Further, because sample size correction is implicit rather than explicit, ICOMP should be superior for the identification of hierarchical models and other cases where observations are not independent.

Unfortunately, the ICOMP does not consider the magnitude of the model set. However, I see no reason why a hybrid criterion might not be utilized. One could define an ICOMP(x) as:

\[-2 \log(L(\hat{\theta}_k)) + 2 \exp(x)C_1(\hat{f}^{-1}(\hat{\theta}_k)).\]

As in the SIC(x), the x in the above would be increased until a clear minimum was observed.

Dr. Bozdogan extends the ICOMP criterion in a fashion that acknowledges model misspecification. This is an exciting development, in keeping with the belief expressed in this chapter and in chapters 7 (Lele), 14 (Lindsey), and 16 (Taper and Lele), that models are always misspecified. I am gratified that my paper may have stimulated a portion of this work.

The bias correction term implicit in the ICOMP and explicit in the CAICF_gen and the CAIF_c should prove to be very useful. The structurally similar correction in the AIC_c certainly has been a major contribution. However, I have one caveat to raise about these criteria based on the empirical information matrix. The information number requires considerable data to estimate accurately. I have found that the CAICF that Dr. Bozdogan proposed in his classic 1987 paper is quite unstable if the number of observations is low, or even moderate. Even with bias corrections, the behavior of these forms may deteriorate rapidly with small sample size.

REFERENCES


