

# The Importance of Complexity in Model Selection

In Jae Myung

*Ohio State University*

---

Model selection should be based not solely on goodness-of-fit, but must also consider model complexity. While the goal of mathematical modeling in cognitive psychology is to select one model from a set of competing models that best captures the underlying mental process, choosing the model that best fits a particular set of data will not achieve this goal. This is because a highly complex model can provide a good fit without necessarily bearing any interpretable relationship with the underlying process. It is shown that model selection based solely on the fit to observed data will result in the choice of an unnecessarily complex model that overfits the data, and thus generalizes poorly. The effect of over-fitting must be properly offset by model selection methods. An application example of selection methods using artificial data is also presented. © 2000 Academic Press

---

## 1. INTRODUCTION

While the goal of mathematical modeling in cognitive psychology is to select one model from a set of competing models that best captures the underlying mental process, the researcher often chooses the model that best fits a particular set of data. Presumably, the justification for this procedure is that the model providing the best fit (e.g., the one that accounts for the most variance) is the one that most closely approximates the underlying mental process. In fact, this justification is unwarranted, for a highly complex model can provide a good fit without necessarily bearing any interpretable relationship with the underlying process. The purpose of this paper is to support this claim and, in so doing, demonstrate the importance of complexity in model selection. Specifically, it is shown that model selection based solely on the fit to a particular set of data will result in the choice of an unnecessarily complex model that overfits the data and thus generalizes poorly to other

This paper is based on a presentation at the symposium *Methods for Model Selection* held at Indiana University in August 1997, hosted by Professor Richard Shiffrin. A portion of the material, especially part of the simulation results in Sections 2 and 5, was previously reported in Myung and Pitt (1998). The author thanks Mark Pitt for many enlightening discussions during the preparation of this paper. Thanks are also due to Michael Browne for help with statistics-related questions, Shaobo Zhang for many extremely helpful suggestions and corrections, and finally, Krishna Tateneni, Malcolm Forster, Peter Grunwald, and two anonymous reviewers for their valuable comments. Correspondence should be directed to In Jae Myung, Department of Psychology, the Ohio State University, 1885 Neil Avenue Mall, Columbus, Ohio 43210-1222. E-mail address: [myung.1@osu.edu](mailto:myung.1@osu.edu).

data generated by the same underlying process. In order to avoid this pitfall, the over-fitting effect must be offset by selection methods.

This paper is intended to be an introduction to the topic of complexity. The issues will be explained with the aid of examples using simulated data. The paper is organized as follows: It begins with a discussion of the effects of model complexity on model fit, followed by a discussion of the relationships between complexity and generalizability. Next, seven selection methods are reviewed, which differ in their estimation of a model's generalizability. An application example of the selection methods using artificial data is also presented. Finally, the main points are summarized and the paper concludes with a final remark.

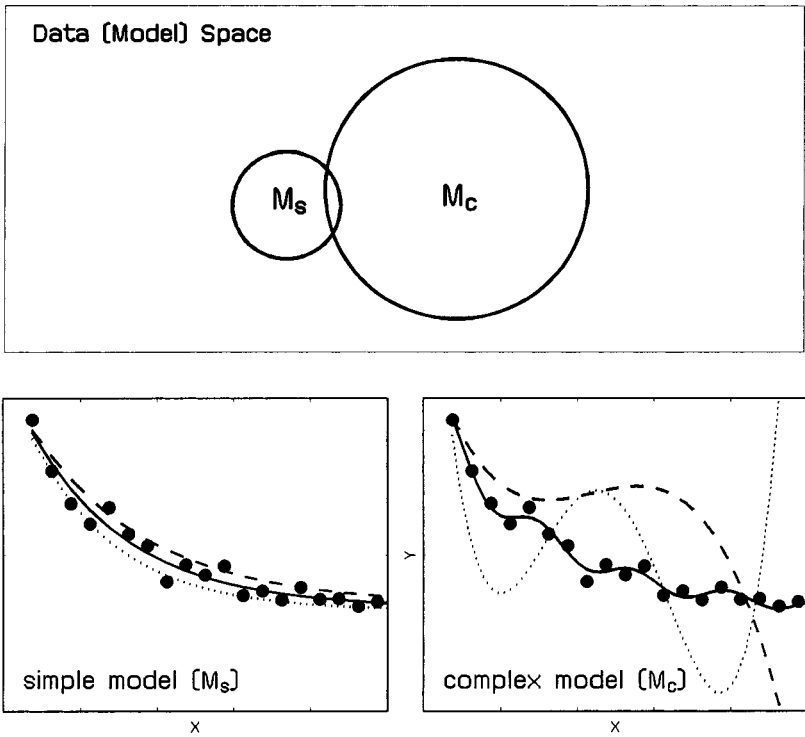
## 2. THE CONTRIBUTION OF COMPLEXITY TO MODEL BEHAVIOR

This section discusses why goodness of fit is a necessary but not sufficient condition in model selection, and why a model's complexity must also be considered.

To begin with, let us define what is meant by a mathematical model. From a statistical standpoint, a mathematical model is a parametric family of probability distributions of observations. For example, a two-parameter model of visual perception may assume that the number of correct responses in a perceptual identification task follows a binomial probability density with success probability  $y(\theta_1, \theta_2)$ . The model may further assume that the success probability  $q$  is a logistic function of stimulus duration  $x$  in the form of  $y(\theta_1, \theta_2) = (1 + \theta_1 \exp(-\theta_2 \cdot x))^{-1}$ . Quite often, an equation such as this one is taken to define a mathematical model, while the underlying probability distribution is kept implicit.

The impact of complexity or flexibility on model fit can be illustrated by considering the range of data patterns (i.e., probability distributions of observations specified by the model equation) that a model can occupy in data space (top panel in Fig. 1). Data space is defined as the universe of all possible data patterns that could be observed in experimentation. Every point in this multidimensional space represents a particular data pattern, such as the shape of a frequency distribution of correct response times. All models occupy a section (or multiple sections) of data space, so it is equally appropriate to think of this space as the universe of all models under consideration.

The size of the data space occupied by a model is positively related to its complexity. A simple model ( $M_s$  in Fig. 1) will occupy a small region of space because it assumes a specific structure in the data, which will manifest itself as a relatively narrow range of similar data patterns, as illustrated in the lower-left panel of Fig. 1. When one of these patterns occurs, the model will fit the data well; otherwise it will fit poorly. Simple models are easily falsifiable, requiring a small minimum number of new points in data space to disprove the model. In contrast, a complex model ( $M_c$  in Fig. 1) will occupy a larger portion of data space. It usually is one with many parameters and a (powerful) nonlinear equation for combining parameters. Complex models do not assume a single structure in the data. Rather, this structure changes as a function of the parameter values of the model, which can be finely tuned so that the model fits a wide range of data patterns, as illustrated in the lower-right panel of Fig. 1. Overly complex models are of questionable worth because their ability to fit a wide range of data patterns can make them unfalsifiable.



**FIG. 1.** The upper panel depicts data regions occupied by two models,  $M_s$  (simple model) and  $M_c$  (complex model), and the lower panels show hypothetical fits to data (solid circle) by the models. Adapted from Fig. 1 of Myung and Pitt (1998).

There are at least two independent factors of model complexity that can significantly affect model fit, namely the number of parameters and functional form. The latter refers to the way in which the parameters are combined in the model equation. The interplay between these two factors and their effect on model fit is illustrated in the following example.

The data-fitting abilities of five models were compared relative to a standard, true model as the amount of error in the data was systematically increased. The model that generated the data was  $y = x + x^2 + N(0, \sigma^2)$  where the value of  $s$  ranged from 0 to 20 in 1-step increments. The five models, the equations of which are listed in the legend of Fig. 2, were fit to 1000 samples generated at each of the 20 error ranges. There were two conditions that differed in the number of points of the variable  $x$ . The first one contained 10 points in which  $x$  ranged from 0.1 to 4.6 in 0.5-step increments whereas the second data set contained 100 points in which  $x$  ranged from 0.10 to 5.05 in 0.05-step increments.

Models  $M_2$ ,  $M_3$ , and  $M_4$  are correctly specified in that they contain as a special case the true model that generated the data, but differ in the number of parameters and functional form.<sup>1</sup>  $M_2$  has two parameters and is the true model.  $M_3$  and  $M_4$

<sup>1</sup> In a sense, all three models,  $M_2$ ,  $M_3$ , and  $M_4$ , are true models as they all provide an exact description of the data for particular choices of parameter values. In this paper, however, by the true model we mean the simplest one of three, which is  $M_2$ .

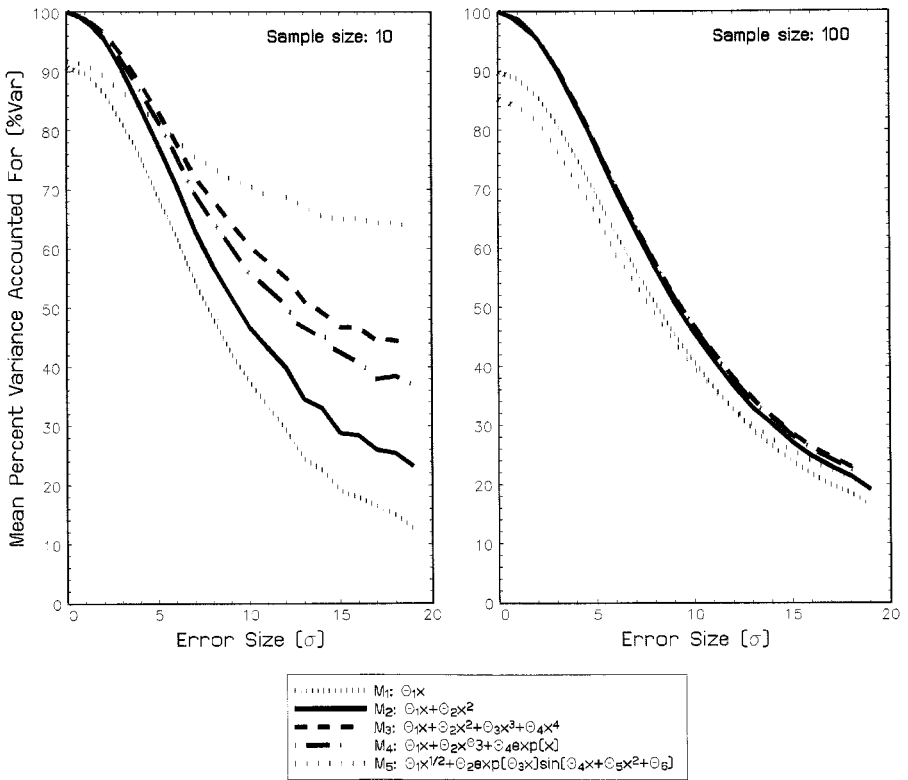


FIG. 2. Model fit as a function of error size. The left panel adapted from Fig. 2 of Myung and Pitt (1998).

both have four parameters but M<sub>3</sub> differs from M<sub>4</sub> in its model equation. M<sub>1</sub> and M<sub>5</sub> are misspecified models that differ in the number of parameters and also functional form.

Simulation results are shown in Fig. 2 for each model and sample size. Fit was measured as the mean percent variance accounted for (%Var) at each error size. As was expected, fit declined as error increased for all models and for both sample sizes. When there was no error in the data, models M<sub>2</sub>–M<sub>4</sub> fit the data perfectly, as they all are correctly specified. The fit by the misspecified model M<sub>1</sub> was the worst, regardless of error and sample sizes. This poor fit indicates the model does not accurately reflect the underlying process that generated the data. This outcome shows why goodness-of-fit is a necessary condition for model selection.

M<sub>2</sub>, the true model, provides a point of reference from which to evaluate M<sub>3</sub>–M<sub>5</sub>. Let us first consider the results for the small sample size (left panel of Fig. 2). As error increases, M<sub>3</sub> and M<sub>4</sub> provide better fits than M<sub>2</sub>, even though M<sub>2</sub> generated the data. The two additional parameters  $\theta_3$  and  $\theta_4$  in M<sub>3</sub> and M<sub>4</sub> enable these models to absorb a substantial amount of the error variance (up to 8–10% by error size 10), thereby improving fit. It is important to understand that this additional improvement in fit over and above that provided by M<sub>2</sub> can be due only to the model’s flexibility to capture random error, not because it more accurately approximates the model that generated the data (i.e., M<sub>2</sub>). Also note in Fig. 2 that

$M_3$  provided an even better fit than  $M_4$ . This difference in performance is due to functional form; although both models have four parameters, they are combined differently. Apparently, the extra functional complexity of  $M_3$  enables the model to absorb random error more easily than  $M_4$ . The other misspecified and over-parameterized model,  $M_5$ , fit poorly when error was small, but because of its complexity, fit the best after error size 6.

As was shown in the right panel of Fig. 2, these over-fitting effects of complexity on model fit become less dramatic as sample size increases. Regardless of sample size, however, if decisions are made based on goodness of fit alone,  $M_3$  and  $M_4$ , or even  $M_5$  especially when sample size is small, will be chosen over  $M_2$ .

In summary, the true model but also overly complex models can fit data well, which is why goodness of fit is not a sufficient condition for model selection. In other words, the over-emphasis on goodness of fit can lead to inappropriate conclusions if the role of complexity in model behavior is not taken into account (Collyer, 1985).

### 3. COMPLEXITY AND GENERALIZABILITY

Complexity affects not only model fit but also the generalizability of a model and the variability in parameter estimation. It is therefore necessary to consider them when evaluating models. A simple model will generalize better to new data sets than a complex model and thus will have a higher degree of predictive accuracy. In addition, a simple model's behavior is more tractable because parameter estimates will be more stable after repeated data fittings than those of complex models. These two points are illustrated using simulations based on artificial data.

Five (nested) linear regression models were fitted to artificial data that were generated from the true model:  $y = x + x^2 + N(\mu, \sigma^2)$ , where  $\mu = 0$  and standard deviation  $\sigma = 3$ . The five models increase in complexity in the number of parameters defined as:

$$P_i : y = \sum_{k=1}^{i+1} \theta_k x^k, \quad (i = 1, \dots, 5).$$

Note that  $P_1$  is the true model that generated the data. Five hundred pairs of samples was generated from this model as follows. Each pair of samples were independently generated using the same 10 points for  $x$ , which ranged from 0.1 to 4.6 in half-step increments. Then each of the five models was fit to the first of the pair, and least squares estimates of the parameters were obtained. The same parameters were used to make predictions for the second pair.

The results in Table 1a show how model complexity is inversely related to generalizability. For the first data set, mean square error defined as  $\sqrt{SSE/\text{sample size}}$ , where SSE stands for the sum of squares error, decreased as model complexity increased, being largest for  $P_1$  and smallest for  $P_5$ . Values in parentheses are the percentage of samples for which the corresponding model fit better than the other models. The most complex model,  $P_5$ , provided a superior fit 100% of the time.

**TABLE 1a**  
**Generalizability of Models Differing in Complexity**

	Model				
	P <sub>1</sub> (True model)	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>
First data	2.61 (0%)	2.43 (0%)	2.23 (0%)	2.03 (0%)	1.77 (100%)
Second data	3.19 (51%)	3.33 (21%)	3.47 (12%)	3.59 (8%)	3.73 (8%)

*Note.* For each data set and model, the mean square error averaged across 500 samples is shown, and percentage of samples in which the model fits better than the other models is shown in parenthesis.

Note that the standard error (1.77) for this model is substantially lower than the expected standard error ( $\sigma = 3.00$ ) for the true model with the correct parameter values. The difference, 1.23 ( $= 3.00 - 1.77$ ), represents the extent of over-fitting. When the generalizability of the models was assessed by fitting the five models to the second data set with the parameter values obtained from the first data set, the more complex models performed more poorly than simple models. Not only was the standard error highest for P<sub>5</sub>, but it also yielded the best fit least often (8%). Less complex models (e.g., P<sub>2</sub>) yielded better fits, with the true model (P<sub>1</sub>) having the smallest error and fitting the best most often.

The influence of complexity on parameter estimation is shown in Table 1b, which contains the mean standard deviation of the two parameter estimates ( $\theta_1$  and  $\theta_2$ ) in each model, averaged over the 500 samples of the first data set. Note how variation is directly related to complexity, being far greater for complex models than for simple ones. The distributions are so wide for complex models that they suggest the parameter values themselves carry little information, making them theoretically meaningless. A complex model succeeds in fitting the observed data very well despite having large errors for each parameter because the errors conspire to cancel each other out for the special purpose of fitting the observed data. But when it comes to generalizing to new data, the conspiracy falls apart.<sup>2</sup>

In summary, one outcome that the preceding simulations demonstrate is an inverse relationship between complexity and generalizability. A model with many parameters (complex) will fit an observed data set better than a model with few parameters (simple), even if the latter generated the data. The complex model, however, generalizes poorly to new data sets precisely because it overfits the first data set by absorbing random error. It is unlikely that an overly complex model will accurately reflect the mental processes responsible for generating the data. An implication of this observation for model selection is that a better-fitting complex model is in general not always preferable. This trade-off between goodness of fit and complexity embodies the principle of Occam's razor (William of Occam, ca. 1290–1349), which states that "entities should not be multiplied beyond necessity." Selection methods in essence implement this principle in one way or another.

<sup>2</sup> This interpretation of a complex model in terms of parameter variability is provided by Malcolm Forster. The author thanks him for the insight.

**TABLE 1b**  
**Variability of Parameter Estimates in Models Differing in Complexity**

	Model				
	P <sub>1</sub> (True model)	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>
Parameter $\theta_1$	1.39	3.41	7.09	12.98	20.53
Parameter $\theta_2$	0.37	2.19	8.05	22.71	51.20

*Note.* For each model, the mean standard deviation of two parameter estimates averaged across 500 samples is shown.

#### 4. MODEL SELECTION AS COMPLEXITY ADJUSTMENT

A central theme of model selection is that to avoid choosing unnecessarily complex models, a model should be selected based on its generalizability, rather than its goodness of fit. This goal is realized by defining a selection criterion that makes an appropriate adjustment to its goodness of fit by taking into account the contribution from model complexity. There are at least seven different selection methods that are currently in use. They differ from one another in terms of how such adjustments are made to best estimate a model's generalizability. Here I briefly describe each method. For details, the reader is directed to more comprehensive reviews, many of which are in this issue.

*AIC, BIC, and RMSD.* Several selection methods have been proposed that adjust for variation in the number of parameters among models, essentially penalizing models with additional parameters. They include the Akaike information criterion (AIC; Akaike, 1973; Bozdogan, 2000), the Bayesian information criterion (BIC; Schwarz, 1978; Wasserman, 2000), and the root mean squared deviation (RMSD):<sup>3</sup>

$$AIC_k := -2 \log(ML_k) + 2p_k$$

$$BIC_k := -2 \log(ML_k) + p_k \log(n)$$

$$RMSD_k := \sqrt{SSE_k / (N - p_k)}.$$

In the above equations,  $\log(ML_k)$  is the natural logarithm of the maximized likelihood function (ML) for model  $k$ ,  $p_k$  is the number of free parameters in the model,  $n$  is the number of independent (scalar) observations that contributes to the likelihood (see, e.g., Raftery, 1993, p. 166),  $N$  is the number of data points being

<sup>3</sup> Although the RMSD is often used as a heuristic criterion in selecting among mathematical models of cognition (see, e.g., Friedman *et al.* 1995) and therefore included in the current discussion, to the author's knowledge, no statistical justification exists as to what the criterion attempts to estimate. On the other hand, the root mean square error of approximation (RMSEA, Steiger, 1990; Browne & Cudeck, 1992), which is similar in form to the RMSD and has been employed to date only in the analysis of moment structures, has a well-justified interpretation. Future work of clarifying the statistical meaning of the RMSD and its relationships to other criteria such as the RMSEA is clearly needed.

fitted, and finally,  $SSE_k$  is the sum of squares error for model  $k$ .<sup>4</sup> Note that the term  $(N - p_k)$  represents the degrees of freedom. For normal distributed errors, the first term of the AIC and BIC,  $-2 \log(ML_k)$ , can be replaced by  $\{n \cdot \log(SSE_k) + \text{constant}\}$ . These model selection methods prescribe that the model that minimizes a given criterion should be chosen.

Note that each of these three criteria consists of two terms, the first term representing lack of fit and the second term representing model complexity (i.e.,  $2p_k$  for AIC and  $p_k \log(n)$  for BIC) or simplicity (i.e.,  $(N - p_k)$  for RMSD). These two terms are combined additively in the AIC and BIC or multiplicatively in the RMSD. In both cases, model selection is carried out by trading off lack-of-fit against complexity. A complex model with many parameters, having a large value in the complexity term, will not be selected unless its fit is good enough to justify the extra complexity.

The number of parameters is the only dimension of complexity that these methods consider. As we have seen in the earlier simulation, functional form can also significantly affect model fit and therefore needs to be taken into account in model selection. The selection methods described below are sensitive to functional form as well as the number of parameters.

*Minimum description length (MDL).* This measure of complexity is defined for a given model as the number of binary digits required to unambiguously describe the observed data with the help of the model (Rissanen, 1987; Grunwald, 2000). It is rooted in algorithmic coding theory (Kolmogorov, 1968) and is obtained as a large sample approximation of stochastic complexity (Rissanen, 1983, 1986). The  $MDL_k$  for model  $k$  for large  $n$  is given by

$$MDL_k := -\log(ML_k) + \frac{1}{2} \log |H_k(\hat{\theta})|$$

In the equation,  $\hat{\theta}$  is the ML estimate and  $H_k(\hat{\theta})$  is the Hessian matrix of the minus log likelihood of data  $D$  defined as  $-\nabla_{\theta}^2 \log p(D | M_k(\hat{\theta}))$ , which is often referred to as the observed Fisher information matrix. The model with a minimum MDL should be chosen.

The second term, containing the Hessian matrix, can be interpreted as a complexity measure. Two remarks concerning this measure are in order. First, the MDL is sensitive to both dimensions of model complexity as its value depends upon the functional form of the model as well as the number of the parameters. As sample size increases, however, sensitivity to functional form will gradually diminish as  $|H_k(\hat{\theta})| \rightarrow \alpha n^{p_k}$  as  $n \rightarrow \infty$  where  $\alpha$  is a scalar constant (see, e.g., Raftery, 1993). In this case, the MDL is simply reduced to one-half of the BIC. Second, values of

<sup>4</sup> To understand the difference between  $N$  and  $n$ , first note that the BIC is derived under the assumption that the likelihood function can be expressed as a multiplication of  $n$  normal or approximately normal probability densities (Schwartz, 1978, p. 462). Accordingly,  $n$  will often be equal to the data size  $N$  or the sample size  $s$  depending upon the problem in question. For example, in a simple regression model of  $y_i = \theta x_i + e_i$  where  $x_i$ 's are values of an independent variable and  $e_i$  is independent and normally distributed with zero mean and a fixed variance for  $i = 1, \dots, N$ ,  $n$  is equal to the data size  $N$ . On the other hand, in a case in which  $y_i$ 's ( $i = 1, \dots, s$ ) are independent, identically distributed observations according to a parametrized probability density,  $n$  will be equal to the sample size  $s$ .

the complexity measure may depend upon the data through the parameter estimate  $\hat{\theta}$ , though this is not always the case. For example, for a multivariate linear regression model  $\mathbf{Y} = \mathbf{X}\theta + \varepsilon$  where  $\varepsilon$  is normal distributed with a null mean vector and a known covariance matrix  $\Sigma_k$ , the Hessian matrix is related to the inverse variance-covariance matrix (i.e.,  $H_k(\hat{\theta}) = \mathbf{X}'\Sigma_k^{-1}\mathbf{X}$ ) and hence is independent of the data. An interesting implication of this observation for normal probability models is that from a MDL point of view, the smaller the variance of a normal probability model, the more complex the model. In other words, a model that is highly peaked around  $\hat{\theta}$  (i.e., finely tuned to fit observed data) is more complex than the model that fits the data for a wide range of parameter values, in accord with the intuitions about complex models discussed earlier (Section 2; see also Myung & Pitt, 1997, for a related discussion).

*Information-theoretic measure of complexity (ICOMP).* Bozdogan (1990 and 2000; Bozdogan & Barse, 1997; Bozdogan & Haughton, 1997) recently developed this measure of model complexity. The ICOMP is defined as an entropic measure of statistical dependence between the parameter estimates and is given by

$$ICOMP_k := -\log(ML_k) + \frac{p_k}{2} \log \left[ \frac{\text{trace}(\Omega_k(\hat{\theta}))}{p_k} \right] - \frac{1}{2} \log |\Omega_k(\hat{\theta})|,$$

where  $\Omega_k(\hat{\theta})$  is the covariance matrix of the parameter estimates for model  $k$ . The model that minimizes the ICOMP should be selected.

The second and third terms of the ICOMP together represents a measure of model complexity, whose value is a monotonically increasing function of the number of parameters (Bozdogan & Haughton, 1997, Appendix A).  $\Omega_k(\hat{\theta})$  is often approximated by the inverse Hessian matrix as  $\Omega_k(\hat{\theta}) \approx H_k^{-1}(\hat{\theta})$  (Efron & Hinkley, 1978). In this approximation, the ICOMP becomes identical to the MDL, except for the second term.

*Cross-validation (CV).* The idea behind cross-validation (CV; Stone, 1974; Browne, 2000) is that a model should be selected based on its ability to capture the behavior of unseen or future observations from the same underlying process. In CV, the data are divided into two nonoverlapping samples, calibration and validation. The calibration sample is used to estimate parameters of the model. These values are then fixed in the model and its ability to fit the validation sample is measured, from which a cross-validation index is obtained. Formally, the CV can be expressed as

$$CV_k := F_{k, \text{Validation}}(\hat{\theta}_{\text{Calibration}}),$$

where  $\hat{\theta}_{\text{Calibration}}$  stands for the parameter values estimated from the calibration sample. In the equation,  $F_k$  is a lack-of-fit or goodness-of-fit measure (e.g., SSE or %Var). The model that best fits the validation sample is chosen. When more than two samples are available for analysis, performance can be improved by averaging CV over multiple validation samples. Alternatively, a combination of multiple calibration samples and one validation sample or a similar variation may be used for the same purpose.

*Bayesian model selection (BMS)*. The crux of BMS (Kass & Raftery, 1995; Wasserman, 2000; Myung & Pitt, 1997) is the Bayes factor, which is defined as a ratio of two marginal likelihoods, each corresponding to a model being tested. The marginal likelihood is the probability of the data *averaged* over the entire range of parameter values, and its logarithmic value defines the BMS as follows,

$$BMS_k := \log p(D | M_k) = \log \int p(D | M_k(\theta)) \pi_k(\theta) d\theta,$$

where  $\pi_k(\theta)$  is the prior density of the parameter  $\theta$  for model  $M_k$ . The model with the largest log marginal likelihood is selected.

Under suitable assumptions, the marginal likelihood can be expressed as

$$p(D | M_k) \approx \frac{ML_k}{|H_k(\hat{\theta}) + G_k(\hat{\theta})|^{1/2} [(2\pi)^{p_k/2} \pi_k(\hat{\theta})]},$$

where  $H_k(\hat{\theta})$  is the same Hessian matrix defined earlier, and  $G_k(\hat{\theta})$  is the Hessian matrix of the minus log prior defined as  $G_k(\hat{\theta}) = -\nabla^2 \log \pi_k(\hat{\theta})$ . From this simplified expression of the marginal likelihood, one can intuitively interpret the denominator term as a Bayesian complexity measure. Note again the appearance of the Hessian matrix in BMS, as in MDL and ICOMP. The above expression reveals how Bayesian model selection works: Maximization of the marginal likelihood is accomplished by pitting maximization of goodness of fit against minimization of model complexity. Incidentally, the BIC can be shown to be a large sample approximation of the simplified expression under flat priors, that is,  $BIC_k \approx -2 \log p(D | M_k)$  (e.g., Wasserman, 2000).

To recap, complexity adjustment is behind each of these seven methods. They differ only in how complexity is measured and how it is combined with a goodness of fit measure to yield an overall selection criterion. Another way to conceptualize what these selection methods do is in terms of averaging to avoid overly complex models. It is grounded explicitly or implicitly in the reasoning behind each method. That is, the seven methods can be grouped under two different approaches. In the first approach, which may be termed the *generalization-based* approach, a model is evaluated in terms of its ability to fit not only observed data, but also unseen (e.g., future) data from the same process. The AIC, ICOMP, RMSD, and CV belong to this approach. For instance, the AIC is derived as a large sample approximation of the expected Kullback–Leibler information distance between the probability density of the true model and that of the fitted model, the expectation taken over all possible observations under the true probability density. Similar notions have motivated the other three criteria. In the second, *explanation-based* approach, a model is evaluated in terms of the expected likelihood of the presently observed data alone over all possible parameter values.<sup>5</sup> BMS, BIC, and MDL are three

<sup>5</sup> Hanna (1969) made a similar distinction between generalization and explanation in model selection. Among various information-theoretic measures proposed in his paper, the *expected descriptive power* in Eq. (1.2) (p. 301) over possible outcomes reflects generalizability. Another measure called the *coefficient of predictive power* (p. 297), which is obtained by averaging over different parameter's values, reflects the explanatory power of the model.

examples of this approach. Parameter averaging in BMS is obvious in the marginal likelihood and in BIC as the latter is an approximation of the former. For the MDL, its precursor, the stochastic complexity, is defined in terms of certain marginal likelihoods (Rissanen, 1983).

Despite their different theoretical origins, an important commonality is shared by the two approaches. Model selection is based on not just its *maximized* performance (e.g., maximum likelihood) but instead on an *averaged* performance, that is, an average across all potential (present and future) data patterns in the generalization approach, or an average over all possible parameter values in the explanation-based approach.

No matter what complexity adjustment or averaging is performed in the selection methods, the ultimate question that we, as modelers, might ask is whether any of these methods will indeed succeed in identifying the true model that generated the data or at least the closest approximation to the true model in some defined sense. Here we mention two asymptotic results known for some of the methods under restricted assumptions. First, Nishii (1984) showed that for comparing among nonnested normal linear regression models one of which is the true model, BIC is consistent in the sense that the probability of selecting the true model approaches one as sample size goes to infinity whereas AIC and ICOMP are both inconsistent. On the other hand, if the true model is not among those considered in the normal linear regression analysis, Bozdogan and Haughton (1997) showed that all three criteria (AIC, BIC, ICOMP) perform perfectly in selecting the model that is the closest approximation to the true model in the Kullback–Leibler sense. Second, Browne and Cudeck (1992) showed that in the context of covariance structure modeling, both the AIC and an estimate of the expected cross-validation index will produce the same ranking among a set of competing models. In the following section, we demonstrate the application of the seven methods using artificial data and compare their performance under different sample sizes.

## 5. APPLICATION EXAMPLE OF THE SELECTION METHODS

The five models used earlier ( $M_1$ – $M_5$  in Fig. 2) were used to examine how the seven selection methods vary in their sensitivity to model complexity and sample size. Each of the five models was fitted to the same artificial data at error size  $\sigma = 1$  used in Fig. 2, and their ability to recover the true model was measured. For MDL, the Hessian matrix was computed numerically, and it was also used in place of the covariance matrix for ICOMP. For  $CV_s$ , an average MSE across  $s$  independent validation samples was used as the cross-validation index. For BMS, the noninformative Jeffreys' prior of  $\pi(\theta_i) \propto 1/|\theta_i|$  on a finite interval of the parameters was used for all the parameters (Kass & Wasserman, 1996). The BMS integral was evaluated numerically using the simple Monte Carlo method. The results are summarized in Table 2.

Model recovery using MSE provides a baseline from which to examine how selection changes when complexity is incorporated into the selection method. Because selection based on MSE is not adjusted for model complexity, its performance is predictable based on the simulations presented earlier (Section 2). Regardless of

TABLE 2

Comparison of Performance among the Seven Selection Methods to the Simulated Data

	Model fitted				
	M <sub>1</sub>	M <sub>2</sub> (True model)	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>
Sample size 10					
MSE	2.58 (0%)	0.82 (0%)	0.73 (78%)	0.81 (11%)	2.30 (11%)
AIC	43.9 (0%)	22.6 (74%)	24.1 (23%)	26.4 (1%)	48.8 (2%)
BIC	44.2 (0%)	23.2 (85%)	25.3 (15%)	27.6 (0%)	50.6 (0%)
RMSD	2.72 (0%)	0.92 (59%)	0.94 (39%)	1.05 (1%)	3.64 (1%)
MDL	23.1 (0%)	13.6 (87%)	16.7 (0%)	15.8 (13%)	37.8 (0%)
ICOMP	21.0 (0%)	11.4 (100%)	19.2 (0%)	19.3 (0%)	33.3 (0%)
CV <sub>1</sub>	2.64 (0%)	1.01 (45%)	1.04 (27%)	1.05 (24%)	2.26 (4%)
CV <sub>10</sub>	2.63 (0%)	0.99 (71%)	1.06 (10%)	1.04 (18%)	2.25 (1%)
BMS	-48.4 (0%)	-20.7 (98%)	-23.6 (2%)	-25.1 (0%)	-27.4 (0%)
Sample size 100					
MSE	3.08 (0%)	0.97 (0%)	0.97 (64%)	0.99 (32%)	3.93 (4%)
AIC	687 (0%)	459 (94%)	462 (4%)	465 (1%)	731 (1%)
BIC	690 (0%)	464 (100%)	472 (0%)	476 (0%)	747 (0%)
RMSD	3.09 (0%)	0.98 (74%)	0.99 (12%)	1.01 (12%)	4.05 (2%)
MDL	346 (0%)	234 (100%)	241 (0%)	241 (0%)	379 (0%)
ICOMP	343 (0%)	230 (100%)	238 (0%)	239 (0%)	373 (0%)
CV <sub>1</sub>	3.08 (0%)	0.99 (68%)	1.00 (23%)	1.01 (5%)	3.01 (4%)
CV <sub>10</sub>	3.08 (0%)	0.99 (91%)	1.00 (3%)	1.01 (6%)	3.11 (0%)
BMS	-572 (0%)	-151 (97%)	-167 (2%)	-191 (1%)	-192 (0%)

*Note.* For each method, the arithmetic mean, averaged across 1000 samples (for BMS, 100 samples), is shown. In CV<sub>s</sub>, an average MSE across *s* independent validation samples was used as the cross-validation index. The value in parentheses is the percent of the samples in which the indicated model fit better than the other three models. The four models were as follows: M<sub>1</sub>,  $y = \theta_1 x$ ; M<sub>2</sub>,  $y = \theta_1 x + \theta_2 x^2$ ; M<sub>3</sub>,  $y = \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ ; M<sub>4</sub>,  $y = \theta_1 x + \theta_2 x^{\theta_3} + \theta_4 e^x$ ; M<sub>5</sub>,  $y = \theta_1 x^{0.5} + \theta_2 \exp(\theta_3 x) \sin(\theta_4 x + \theta_5 x^2 + \theta_6)$ .

sample size, MSE favors the three overly complex models M<sub>3</sub>–M<sub>5</sub>, particularly M<sub>3</sub>. The two simpler models, including the true model (M<sub>2</sub>), were never chosen.

Inspection of the recovery rates for the seven methods reveals how much recovery improves when model complexity is incorporated into the selection process. Let us first consider the results for the small sample size (top half of Table 2). The true model (M<sub>2</sub>) was successfully recovered at least 45% of the time, although performance differed across the methods. When AIC was the selection method, M<sub>2</sub> was correctly recovered 74% of the time, while M<sub>3</sub> was chosen 23% of the time. The three models, M<sub>1</sub>, M<sub>4</sub>, and M<sub>5</sub>, were never or rarely chosen. Inspection of the AIC values reveals why. Their larger values compared with M<sub>2</sub> and M<sub>3</sub> indicate that either the fit was too poor (for M<sub>1</sub>) or they were penalized heavily for having additional complexity even if their fits were good (for M<sub>4</sub> and M<sub>5</sub>). A similar interpretation can be drawn for BIC, RMSD, MDL, and CV. All showed a bias to select complex models. Not surprisingly, in CV<sub>10</sub>, the performance improved as the number of independent validation samples to be averaged was increased. This is

because the validation index more accurately estimates the true population value as more validation samples are used. Performance under ICOMP and BMS was most impressive. Under these criteria, the biases toward complex or simpler models exhibited by the preceding selection methods completely or virtually disappeared. Here, selection was clear cut, with the true model ( $M_2$ ) being recovered at least a full 98% of the time.

When sample size is relatively large (bottom half of Table 2), as was expected, model recovery improved significantly across most of the selection methods, even to the level of complete recovery with BIC, MDL, and ICOMP.

In summary, all seven selection methods performed reasonably well in identifying the true model, though performance varied widely in their bias toward simple or complex models as well as in their sensitivity to sample size. Before closing this section, a word of caution is in order: The above results should not be taken as representative behavior of the selection methods. The sole purpose of the simulation was to *demonstrate* their application, not to investigate their properties. Accordingly, performance of the methods in this example should not be generalized to other data and settings. Relative performance of the selection methods may vary considerably depending upon the specific set of models under consideration (e.g., nested vs. nonnested, correctly specified vs. misspecified), the characteristics of observed data, the sample size, the level of random error, etc.

## 6. SUMMARY AND CONCLUSION

The main point of this article has been that model selection should be based not solely on goodness of fit, but must also consider model complexity. Put another way, while the goal of mathematical modeling in cognitive psychology is to identify the model that most closely approximates the underlying process, choosing the model that best fits a particular set of data will not achieve this goal.

From a statistical standpoint, observed data represent a sample from an unknown population (true underlying process) so the data are tainted with sampling error. Consequently, when the data are fitted to a model, the model's performance reflects not only the population pattern but also spurious patterns due to sampling error. Such spurious patterns will be specific to the particular sample and will not repeat themselves in other samples. A complex model with many parameters, because of its extra flexibility, tends to capture these spurious patterns more easily than a simple model with few parameters. Consequently, the complex model yields a better fit to the data, not because of its ability to more accurately approximate the underlying process but rather because of its ability to capitalize on sampling error. Therefore, choosing a model based solely on its fit, without appropriately filtering out the effects due to sampling error, will result in choosing an overly complex model that generalizes poorly to other data from the same underlying process. A consequence of such practice is that the model may become more sophisticated as additional parameters or modifications of the model are introduced to account for newly found discrepancies—which may be, in fact, sampling errors—between a model's predictions and new observations, and in the process, however, the model's generalizability may be further decreased!

In conclusion, to avoid selecting a powerful model, but one with little scientific significance, model selection should not be based solely on a model's ability to fit particular sample data but instead should be based on its ability to capture the characteristics of the population (i.e., generalizability). Only after assessing its generalizability can we be confident in the correctness of the model.

## REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In B. N. Petrox and F. Caski (Eds.), *Second international symposium on information theory*, pp. 267. Akademiai Kiado, Budapest.
- Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics Theory Methods*, **19**, 221–278.
- Bozdogan, H. (2000). Akaike information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, **44**, 62–91.
- Bozdogan, H., & Bearnse, P. M. (1997). Model selection using informational complexity with applications to vector autoregressive (VAR) models. In D. Dowe (Ed.), *Information, statistics and induction in sciences (ISIS) anthology*. Berlin/New York: Springer-Verlag.
- Bozdogan, H., & Haughton, D. M. A. (1997). Information complexity criteria for regression models. Unpublished manuscript.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.
- Browne, M. W., & Cudeck, R. C. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, **21**, 230–258.
- Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics*, **38**, 476–481.
- Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457–487.
- Friedman, E., Massaro, D. W., Kitzis, S. N., & Cohen, M. M. (1995). A comparison of learning models. *Journal of Mathematical Psychology*, **39**, 164–178.
- Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133–152.
- Hanna, J. F. (1968). Some information measures for testing stochastic models. *Journal of Mathematical Psychology*, **6**, 294–311.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.
- Kolmogorov, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Transaction on Information Theory*, **14**, 662–664.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95.
- Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger and A. M. Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition*, pp. 327–355. Hillsdale, NJ: Erlbaum.
- Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, **12**, 758–765.
- Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*, pp. 163–180. Thousand Oaks, CA: Sage.

- Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416–431.
- Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, **14**, 1080–1100.
- Rissanen, J. (1987). Stochastic complexity and the MDL principle. *Econometric Reviews*, **6**, 85–102.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, **25**, 173–180.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

Received: November 10, 1997; revised: August 25, 1998