

Model Comparison Methods

In J. Myung and Mark A. Pitt
Department of Psychology
Ohio State University
Columbus, Ohio 43210-1222
{[myung.1](mailto:myung.1@osu.edu), [pitt.2](mailto:pitt.2@osu.edu)}@osu.edu

March 10, 2003

To appear in L. Brand and M. L. Johnson (eds.), *Numerical Computer Methods, Part D*
(A volume of *Methods in Enzymology*)

Introduction

The question of how one should choose among competing explanations (models) of observed data is at the core of science. Model comparison is ubiquitous and arises, for example, when a toxicologist must decide between two dose-response models or when a biochemist needs to determine which of a set of enzyme-kinetics models best accounts for observed data.

Over the decades, a number of criteria that are thought to be important for model comparison have been proposed (e.g., Jacobs & Grainger, 1994). They include (a) *falsifiability* (Popper, 1959): whether there exist potential observations that are incompatible with the model; (b) *explanatory adequacy*: whether the theoretical account of the model helps to make sense of observed data but also established findings; (c) *interpretability*: whether the components of the model, especially its parameters, are understandable and are linked to known processes; (d) *faithfulness*: whether the model's ability to capture the underlying regularities comes from the theoretical principles the model purports to implement, not from the incidental choices made in its computational instantiation; (e) *goodness of fit*: whether the model fits the observed data sufficiently well; (f) *complexity* or *simplicity*; whether the model's description of observed data is achieved in the simplest possible manner; and (g) *generalizability*: whether the model provides a good prediction of future observations.

Although each one of these seven criteria is important in its own way, modern statistical approaches to model comparison consider only the last three (goodness of fit, complexity, generalizability), largely because they lend themselves to quantification. The other four criteria have yet to be formalized and it is not clear how some even could be or should be (e.g., interpretability).

The purpose of this chapter is to provide a tutorial on state-of-the-art statistical model comparison methods. We walk the reader through the reasoning underlying their development so that how and why a method performs as it does can be understood. The chapter is written for researchers who are interested in computational modeling but are primarily involved in empirical work. We begin by discussing the statistical foundations of model comparison.

Statistical Foundations of Model Comparison

Notation and Definition

Statistically speaking, the data vector $\mathbf{y} = (y_1, \dots, y_m)$ is a random sample from an unknown population, which represents the underlying regularity that we wish to model. The goal of modeling is to identify the model that generated the data. This is not in general possible because information in the data sample itself is frequently insufficient to narrow the choices down to a single model. To complicate matters even more, data are inevitably confounded by random noise, whether it is sampling error, imprecision of the measurement instrument, or the inherent unreliability of the data collection procedure. It is usually the case that multiple models could have generated a single data sample.

In the field of engineering, this situation is referred to as an ill-posed problem because any solution (i.e., model) is not unique given what the data tell us about the underlying regularity. The statistician's way of dealing with ill-posedness is to use all of the information available (i.e., knowledge about the model and the data together) to make a best guess as to which model most likely generated the data. Stripped bare, model selection is an inference game, with selection methods differing in their rules of play.

Formally, a model is defined as a parametric family of probability distributions. Each distribution is indexed by the model's parameter vector $\mathbf{w} = (w_1, \dots, w_k)$ and corresponds to a population. The

probability (density) function, denoted by $f(\mathbf{y}|\mathbf{w})$, specifies the probability of observing data \mathbf{y} given the parameter \mathbf{w} . Given the fixed data vector \mathbf{y} , $f(\mathbf{y}|\mathbf{w})$ becomes a function of \mathbf{w} and is called the *likelihood function* denoted by $L(\mathbf{w})$. For example, the likelihood function for binomial data is given by

$$L(\mathbf{w}) = \prod_{i=1}^k \binom{n_i}{y_i} w_i^{y_i} (1 - w_i)^{n_i - y_i} \quad (1)$$

where k is the number of conditions in an experiment, n_i is the sample size or the number of Bernoulli trials (e.g., dichotomous observations, such as success or failure of a medical treatment, made n_i times), and w_i , as an unknown parameter, represents the probability of success on each trial, and y_i ($= 0, 1, \dots, n_i$) is the number of actually observed successes.

Given a data sample, a model's descriptive adequacy is assessed by finding parameter values of the model that best fit the data in some defined sense. This procedure, called parameter estimation, is carried out by seeking the parameter vector \mathbf{w}^* that maximizes the likelihood function $L(\mathbf{w})$ given the observed data vector \mathbf{y} – a procedure known as *maximum likelihood estimation*. The resulting maximized likelihood (ML) value, $L(\mathbf{w}^*)$, defines a measure of the model's *goodness of fit*, which represents a model's ability to fit a particular set of observed data.

Other examples of goodness-of-fit measures include the minimized sum of squared errors (SSE) between a model's predictions and observations, the proportion variance accounted for or otherwise known as the coefficient of determination r^2 , and the mean squared error (MSE) defined as the square root of SSE divided by the number of observations. Among these, ML is a standard measure of goodness of fit, most widely used in statistics, and all of the model comparison methods discussed in the present chapter were developed using ML. In the rest of the chapter, goodness of fit will refer to maximized likelihood.

A Good Fit can be Insufficient and Misleading

Models are often compared based on their goodness of fit. That is, among a set of models under comparison, the scientist chooses the model that provides the best fit (i.e., highest ML value) to the observed data. The justification for this choice may be that the model best fitting the data is the one that does a better job than its rivals of capturing the underlying regularity. Although intuitive, such reasoning can be unfounded because a model can produce a good fit for reasons that have nothing to do with its ability to approximate the regularity of interest, as will be described below.

Selecting among models using a goodness-of-fit measure would make sense if data were free of noise. In reality, however, data are not “pure” reflections of the population of interest, as mentioned above. Noisy data make the task of inferring the underlying model difficult because what one is fitting is unclear: Is it the regularity, which we care about, or the noise, which we do not? Put another way, goodness of fit can be decomposed conceptually into two separate terms as follows:

$$\text{Goodness of fit} = \text{Fit to regularity} + \text{Fit to noise} \quad (2)$$

We are interested only in the first of these, fit to regularity, but any goodness-of-fit measure contains a contribution from the model's ability to fit random error as well as its ability to approximate the underlying regularity. The problem is that both quantities are unknown because when fitting a data set, we obtain the overall value of their sum, that is, a single goodness of fit. This is why a good fit can be misleading. In the worst case, a good fit can be achieved by a model that is extremely good at fitting noise yet a poor approximation of the regularity being modeled. In the next section we describe how this state of affairs can come about. The remainder of the chapter then focuses on how to correct it.

Model Complexity and Why it Matters

It turns out that a model's ability to fit random noise is closely correlated with the *complexity* of the model. Intuitively, complexity (or flexibility) refers to the property of a model that enables it to fit diverse patterns of data. For example, a model with many parameters is more complex than a model with few parameters. Also, two models with the same number of parameters but different forms of the model equation (e.g., $y = w_1x + w_2$ and $y = w_1x^{w_2}$) can differ in their complexity (Myung, 2000; Pitt, Myung

& Zhang, 2002). Generally speaking, the more complex the model, the more easily it can absorb random noise, thus increasing its fit to the data without necessarily increasing its fit to the regularity. That is, too much complexity is what causes a model to fit noise. The relationship between goodness of fit and complexity is illustrated in the three small graphs in Figure 1. The data are the dots and the lines the models. The model on the left is least complex; that on the right is most complex. Complexity is what enables the model in the lower right graph to fit the data better than the less complex models in the left and middle graphs. In fact one can always improve goodness of fit by increasing model complexity, such as adding more parameters. This is portrayed by the top curve in the large graph. An implication of this spurious phenomenon is that an overly complex model can provide a better fit than a simpler model even if the latter generated the data. That is, the more complex model will over-fit the data.

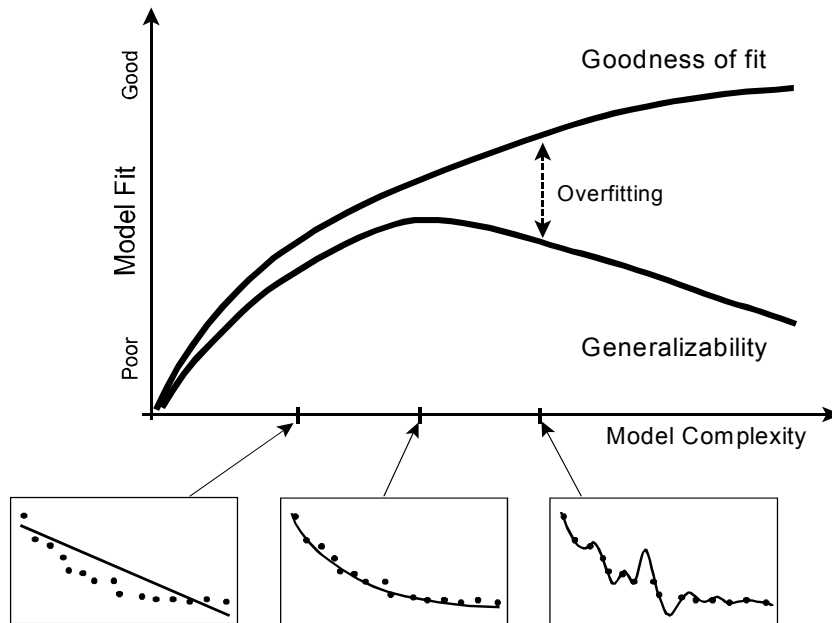


Figure 1. An illustration of the relationship between goodness of fit and generalizability as a function of model complexity. The y axis represents any fit index, where a larger value indicates a better fit (e.g., percent variance accounted for). The three smaller graphs provide a concrete example of how fit improves as complexity increases. In the left graph, the model (line) is not complex enough to match the complexity of the data (dots). The two are well matched in complexity in the middle graph, which is why this occurs at the peak of the generalizability function. In the right graph, the model is more complex than the data, capturing microvariation due to random error. Reprinted from Pitt and Myung (2002).

Over-fitting is illustrated in Table I using four dose-response models in toxicology (Sand et al, 2002). Simulated data varying in three sample sizes were generated from model M_2 , which is considered the “true” model in the sense that it generated the data. In each test, a thousand samples were generated and sampling error was added to each data set. The model M_2 , as well as three other models differing in complexity, were fitted to these data. The first row under each sample-size condition in the Table shows the percentage of samples in which the particular model fitted the data best using the goodness-of-fit measure ML. First consider the results obtained with a small sample size of $n = 20$. M_1 , with only one parameter, is clearly inadequate to capture the main trend in the data and thus never fitted the data best. It is analogous to the left-hand model in Figure 1 and is an example of under-fitting. Next consider the

results for M_3 and M_4 , which, with one extra parameter than M_2 , are more complex than M_2 . They provided a better fit much more often (62% and 23%) than the true model, M_2 (15%), even though M_2 generated the data. This is because the one extra parameter in the models enabled them to absorb sampling error above and beyond the underlying regularity. These models are analogous to the right-hand model in Figure 1. Importantly also, note that M_3 provided a better fit more often than M_4 , despite the fact that both have the same number of parameters (3). This difference in fit between the models is due to the effects of the functional form dimension of model complexity. As can be seen in the middle and lower portions of the Table, over-fitting persisted even for a relatively large sample size of $n = 100$, before it all but disappeared when sample size was increased to an unrealistic level of $n = 5000$.

TABLE I. Goodness of Fit and Generalizability of Models Differing in Complexity

Number of parameters	M_1	M_2 (true)	M_3	M_4
	1	2	3	3
Sample size: $n = 20$				
Goodness of fit	0	15	62	23
Generalizability	15	46	22	17
Sample size: $n = 100$				
Goodness of fit	0	31	56	13
Generalizability	0	64	27	9
Sample size: $n = 5000$				
Goodness of fit	0	96	4	0
Generalizability	0	98	2	0

Note: The percentage of samples in which the particular model fitted the data best. The four models are as: $M_1 : y = 1 - \exp(-0.1 - ax)$, $M_2 : y = 1/(1 + \exp(a - bx))$, $M_3 : y = c + (1 - c)(1 - \exp(-ax^b))$, $M_4 : y = c + (1 - c)/(1 + ax^{-b})$ where $a, b > 0$ and $0 < c < 1$. A thousand pairs of samples were generated from M_2 (true model) using $a = 2$ and $b = 0.2$ on the sample 10 points for $(x_1, \dots, x_{10}) = (0.001, 1, 2, 4, 7, 10, 13, 16, 20, 30)$. For each probability, a series of n ($= 20, 100, \text{ or } 5000$) independent binary outcomes (0 or 1) were generated according to the binomial probability distribution. The number of ones in the series was summed to obtain an observed count being denoted by y_i ($i = 1, \dots, 10$). This way each sample consisted of 10 observed counts. Goodness of fit was evaluated by each model's maximized likelihood and generalizability was estimated through cross-validation described in the text.

These simulation results should highlight the dangers of selecting models using only a goodness-of-fit measure. The potential exists for choosing the wrong model, which makes goodness-of-fit risky to use as a method of model selection.

Generalizability: A Solution that Embodies Occam's Razor

Returning to Eq. (2), a model's ability to fit the regularity in the data, represented by the first term on the right-hand side of the equation, defines its *generalizability*. Generalizability, or predictive accuracy, refers to how well a model predicts the statistical properties of future, as yet unseen, samples from a replication of the experiment that generated the current data sample. That is, when a model is evaluated against its competitors, the goal should be not to assess how much better it fits a single data sample, but how well it captures the process that generated the data. This is achieved when generalizability, not goodness of fit, is the goal of model selection.

Statistically speaking, generalizability is defined in terms of a *discrepancy function* that measures the degree of approximation or dissimilarity between two probability distributions (Linhart & Zucchini, 1986). Specifically, a discrepancy between two distributions, f and g , is any well-behaved function, $D(f, g)$, that satisfies $D(f, g) \geq D(f, f) = 0$ for $f \neq g$. The larger the value of $D(f, g)$, the less one probability distribution approximates the other. The following equation gives a formal definition of generalizability of a given model f_M :

$$\text{Generalizability} := E_T [D(f_T, f_M)] = \int D[f_T, f_M(\mathbf{w}^*(\mathbf{y}))] f_T(\mathbf{y}) d\mathbf{y} \quad (3)$$

In the equation, f_T represent the probability distribution from which all data are generated (i.e., true model), and $f_M(\mathbf{w}^*(\mathbf{y}))$, as a member of the model family under investigation, is the best-fitting probability distribution given a data vector \mathbf{y} , and thus represents a goodness of fit measure. According to the above equation, generalizability is a mean discrepancy between the true model and the model of interest, averaged across all data that could possibly be observed under the true model.

Whereas goodness of fit is monotonically related to complexity, the relationship between generalizability and complexity is not so straightforward. An illustration of this is shown in Figure 1. Generalizability increases positively with complexity only up to the point where the model is optimally complex to capture the regularities in the data. Any additional complexity beyond this point will cause generalizability to diminish as the model begins to over-fit the data by absorbing random noise. Shown in the middle inset is the model with the highest generalizability. It captures the main trends in the data, but little if any of the random fluctuations of the points that in all likelihood are due to sampling error. In short, a model must be sufficiently complex to capture the underlying regularity yet not too complex to cause it to over-fit the data and thus lose generalizability.

The improvements gained by using generalizability in model selection are shown in Table I. Generalizability was assessed using two independent data samples. Each model was fitted to the first sample to obtain the best-fitting parameter values. The resulting fit defines the model's goodness of fit, as shown in the first row under each sample-size condition. Next, with the best-fitting parameters of each model fixed, the models were fitted to the second data sample. The quality of this second fit is a measure of generalizability known as Cross Validation.¹

The second row under each sample-size condition in the Table shows the percentage of samples in which the particular model generalized the best. Comparison with the goodness-of-fit row reveals that an overly complex model (e.g., M_3) generalizes poorly (22% for $n = 20$ and $n = 27\%$ for $n = 100$) whereas the simpler, true model (e.g., M_2) generalizes better (46% for $n = 20$ and $n = 64\%$ for $n = 100$). This example demonstrates that the cost of a model's superior fit to a particular data sample can result in a loss of generalizability when fitted to new data samples generated by that same process. This is because the model's fit to the first sample was far too good, absorbing random error in the data, not just the regularity.

To reiterate, in model comparison, one would like to choose the model, among a set of competing models, that maximizes generalizability. Unfortunately, generalizability is not directly observable because noise always distorts the regularity of interest. Generalizability, therefore, must be *estimated* from a data sample. It is achieved by weighting a model's goodness of fit relative to its complexity. An overly complex model is penalized to the extent that the extra complexity merely helps the model fit the random error, as illustrated in the right inset in Figure 1. Generalizability is viewed as a formal implementation of Occam's razor: Goodness of fit is traded off with complexity. A number of generalizability measures that implement this basic principle have been developed and are discussed in the next section.

Model Comparison Methods

Measures of generalizability

The foremost goal of model comparison is achieving good generalizability by striking the right balance between two opposing pressures, goodness of fit and complexity. In this section we introduce five representative measures of generalizability that achieve this balance. They are the Akaike Information

¹ This method of estimating generalizability is known as cross-validation and is discussed in the next section. It is used in this example because of its simplicity. We have found that it performs poorly compared to more sophisticated measures like MDL and BMS.

Criterion (AIC; Akaike, 1973; Bozdogan, 2000), the Bayesian Information Criterion (BIC; Schwarz, 1978), minimum description length (MDL; Rissanen, 1983, 1996; Grunwald, 2000; Hansen & Yu, 2001), cross-validation (CV; Stone 1974; Browne, 2000), and finally, Bayesian model selection (BMS; Kass & Raftery, 1995; Wasserman, 2000). For a comprehensive treatment of these and other comparison methods, the reader is directed to a special issue of the *Journal of Mathematical Psychology* (Myung, Forster & Browne, 2000) and Burnham and Anderson (2002).

The five model comparison criteria are defined as follows:

$$\begin{aligned}
 AIC &= -2 \ln L(\mathbf{w}^*) + 2k \\
 BIC &= -2 \ln L(\mathbf{w}^*) + k \ln(n) \\
 MDL &= -\ln L(\mathbf{w}^*) + \frac{k}{2} \ln\left(\frac{n}{2\pi}\right) + \ln \int \sqrt{|I(\mathbf{w})|} d\mathbf{w} \\
 CV &= -\ln f(\mathbf{y}_{\text{val}} | \mathbf{w}^*(\mathbf{y}_{\text{cal}})) \\
 BMS &= -\ln \int L(\mathbf{w}) \pi(\mathbf{w}) d\mathbf{w}
 \end{aligned} \tag{4}$$

where $\ln L(\mathbf{w}^*)$ is the natural logarithm of the model's maximized likelihood, k is the number of parameters of the model, n is the sample size. Also in the above equation, $\pi(\mathbf{w})$ is the parameter prior density and $I(\mathbf{w})$ is the Fisher information matrix in mathematical statistics (e.g., Schervish, 1995), the hard bracket $|\cdot|$ denotes the determinant of a matrix, and finally, \mathbf{y}_{val} and \mathbf{y}_{cal} are defined later when discussing the CV criterion. These comparison methods prescribe that the model minimizing the given criterion should be preferred.

AIC, BIC and MDL

For AIC, BIC and MDL, the first term represents a lack of fit measure and the remaining terms represent a model complexity measure. Each criterion defines complexity differently. AIC considers the number of parameters (k) as the only relevant dimension of complexity whereas BIC considers sample size (n) as well. In MDL, the second and third terms together represent a complexity measure. The second term of MDL is essentially the same as that of BIC. It is the third term that is unique in MDL, and it accounts for the effects of complexity due to functional form. Functional form is reflected through the Fisher information matrix $I(\mathbf{w})$. To grasp this, note that the matrix $I(\mathbf{w})$ is defined in terms of the second derivative of the log-likelihood function $\{\ln L(\mathbf{w})\}$, the value of which depends upon the form of the model equation, for instance, whether $y = w_1x + w_2$ or $y = w_1x^{w_2}$.

Why are there such different measures of generalizability (AIC, BIC, MDL)? How do they differ from one another and in what sense? AIC was derived as a large sample approximation of the discrepancy between the true model and the fitted model in which the discrepancy is measured by the Kullback-Leiber distance (Kullback & Leibler, 1951). As such, AIC purports to select the model, among a set of candidate models, that is closest to the truth in the Kullback-Leibler sense. BIC has its origin in Bayesian statistics and seeks the model that is "most likely" to have generated the data in the Bayesian sense. BIC can be seen as a large sample approximation of a quantity related to BMS, which is discussed below. Finally, MDL originated from algorithmic coding theory in computer science. The goal of MDL is to select the model that provides the shortest description of the data in bits. The more the model is able to compress the data by extracting the regularities or patterns in it, the better the model's generalizability because these uncovered regularities can then be used to predict accurately future data. As noted earlier, unlike AIC and BIC, MDL considers the functional form of a model, and thus is designed for comparing models that differ along this dimension. Given this additional sensitivity, MDL is expected to perform more accurately than its two competitors. The price to pay for MDL's superior performance is the computational challenge in its calculation; the evaluation of the integral term generally requires use of numerical integration techniques (e.g., Gilks, Richardson & Spiegelhalter, 1996).

Cross-validation (CV)

CV is a sampling-based method in which generalizability is estimated directly from the data without an explicit consideration of model complexity. The method works as follows. First, the observed data are split into two sub-samples of equal size. We then fit the model of interest to the first, calibration sample (\mathbf{y}_{cal}) and find the best-fitting parameter values, denoted by $\mathbf{w}^*(\mathbf{y}_{\text{cal}})$. With these values fixed, the model is fitted to the second, validation sample (\mathbf{y}_{val}). The resulting fit of the model to the validation data \mathbf{y}_{val} defines the model's generalizability estimate.

CV can easily be implemented using any computer programming language as its calculation does not require sophisticated computational techniques, in contrast to MDL. CV's ease of use is offset by the unreliability of its generalizability estimate, especially for small sample sizes. On the other hand, unlike AIC and BIC, CV takes into account the functional form effects of complexity, although the implicit nature of CV makes it difficult to discern how this is achieved.

Bayesian Model Selection (BMS)

BMS, a sharper version of BIC, is defined as the minus *marginal likelihood*. The marginal likelihood is the probability of the observed data given the model, averaged over the entire range of the parameter vector and weighted by the parameter prior density $\pi(\mathbf{w})$. As such, BMS aims to select the model with the highest mean likelihood of the data. The often cited *Bayes factor* (Kass & Raftery, 1995) is a ratio of marginal likelihoods between a pair of models being compared. As in other Bayesian methods, the prior $\pi(\mathbf{w})$ in the marginal likelihood is to be determined by utilizing available information (i.e., informative priors) or otherwise as a non-informative prior such as a uniform density (Kass & Wasserman, 1996).

BMS does not yield an explicit measure of complexity though complexity is taken into account and is hidden in the integral. It is through the integral that the functional form dimension of complexity, as well as the number of parameters and the sample size, is reflected. It is also the integral that makes it non-trivial to implement BMS. As in MDL, the integral in BMS must be evaluated numerically.

Among the five comparison methods discussed above, BMS and MDL represent state-of-the-art techniques that will generally perform more accurately across a range of modeling situations than the other three criteria (AIC, BIC, CV). On the other hand, the latter three are attractive given their ease of use, and are likely to perform adequately under certain circumstances. In particular, if the models being compared do differ in number of parameters and further, sample size is sufficiently large, then one may be able to use AIC or BIC with confidence instead of the more sophisticated BMS and MDL.

Relations to Generalized Likelihood Ratio Test

Although the generalized likelihood ratio test (GLRT) is often used to test the adequacy of a model in the context of another model, it is not an appropriate method for model comparison. In this section we briefly comment on GLRT and its relation to the model comparison methods discussed above. The generalized likelihood ratio test (Wilks, 1938; Johnson & Wichern, 1998) is a null hypothesis significance test and is based on the G^2 statistic defined as:

$$G^2 = -2 \ln \frac{ML_A}{ML_B} \quad (5)$$

In the equation ML_A/ML_B is a ratio of the maximized likelihoods of two nested models, A and B, with A being nested within B. A model is said to be nested within another if the latter yields the former as a special case. For instance, $y = w_1x$ is nested within $y = w_1x + w_2x^2$.

The sampling distribution of G^2 under the null hypothesis that the reduced model A holds follows a χ^2 -distribution with $df = k_B - k_A$ where k_B and k_A are the numbers of free parameters of models B and A, respectively. When the null hypothesis is retained (i.e., the p-value does not exceed the alpha level), the conclusion is that the reduced model A provides a sufficiently good fit to the observed data and therefore the extra parameters of the full model B unnecessary. If the null hypothesis is rejected, one concludes that model A is inadequate and the extra parameters are necessary to account for the observed data.

There are several crucial differences between GLRT and the five model comparison methods. First and importantly, GLRT does not assess generalizability, which is the goal of model comparison. Rather, it is a hypothesis testing method that simply judges descriptive adequacy of a given model. (For contemporary criticisms of null hypothesis significance testing, see, e.g., Berger & Berry, 1998; Cohen,

1994). Even if the null hypothesis is retained under GLRT, the result does not necessarily indicate that model A is more likely to be correct or generalize better than model B, or vice versa. Second, GLRT requires the nestedness assumption hold of the two models being tested. In contrast, no such assumptions is required for the five comparison methods, making them much more versatile. Third, GLRT was developed in the context of linear regression models with normally-distributed error, further restricting its use. GLRT is also inappropriate for testing non-linear models and models with non-normal error. Again, no such restrictions are imposed on the preceding model comparison methods. In short, given its limited applicability and narrow interpretability, GLRT is a poor method of model comparison.

Model Comparison at Work

In this section, we present a model-recovery test to demonstrate the relative performance of two comparison methods, AIC and MDL. These two were chosen because they differ from each other in how model complexity is defined. AIC considers only the number of parameters, whereas MDL takes into account the functional form dimension as well. The maximized likelihood (ML), solely a goodness of fit measure, was included as a baseline from which the improvements in model recovery could be evaluated as the two dimensions of complexity are introduced into the selection method.

Three models, M_2 - M_4 from Table I, were compared. A thousand data sets (with sampling noise added) were generated from each model, and all three models were then fitted to each group of 1000 data sets. The selection methods were compared on their ability to recover the model that generated the data. A good method should be able to identify the true model (i.e., the one that generated the data) 100% of the time. Any deviations from perfect recovery reveal a bias in the selection method.

The simulation results are reported in Table II. The top 3x3 matrix shows model recovery performance under ML. The result in the first column of the matrix, which is essentially the same as that in the first row of the middle panel in Table I, shows that an over-complex model, M_3 , was chosen more often than the true data-generating model, M_2 (58% vs. 30%). This bias is not surprising given that model M_3 has one more parameter than model M_2 , and that a goodness of fit measure such as ML does not consider this or any other dimension of complexity. The first column in the simulation using AIC shows that when the difference in the number of parameters was taken into account, the bias was largely corrected. Now, the true model M_2 was chosen 76% of the time, much more often than the more complex model, M_3 (16%).

To see the effects of functional form in model selection, turn your attention to the second and third columns. The two models, M_3 and M_4 , have the same number of parameters (3) but differ from each other in functional form. For these two models, an asymmetric pattern of recovery performance was observed for the models under ML. Model M_4 was correctly recovered 90% of the time whereas model M_3 was recovered much less often (61%). In other words, M_4 fitted its competitor's data as well as its own data, but the reverse was not true. Essentially the same pattern of model recovery was obtained under AIC. That is, AIC also overestimated the generalizability of M_4 relative to M_3 . Because the two models have the same number of parameters, this bias must be due to a different dimension of complexity, namely functional form. Calculation of the complexity of these models (using the two right-hand terms of MDL in eq. 4) shows M_4 to indeed be more complex than M_3 (9.36 vs 8.57), with a complexity difference of 0.79. This means that to be selected under MDL, M_4 must provide a higher value of the log maximized likelihood than M_3 by at least 0.79 for it to be selected. Compared with AIC, MDL imposes a stiffer tariff to counteract M_4 's added flexibility in fitting random error. When the effects of complexity due to functional form were neutralized by using MDL, recovery generally improved, especially for the data generated from M_3 (bottom 3x3 matrix).

This example demonstrates the importance of accounting for all relevant dimensions of complexity in model comparison. However, the reader is cautioned not to over-generalize the above simulation results. They should not be regarded as indicative of how the selection methods will perform across all settings. Model comparison is an inference problem. The quality of the inference depends strongly on the characteristics of the data (e.g., sample size, experimental design, type of random error) and the models themselves (e.g., model equation, parameters, nested vs non-nested). For this reason, it is unreasonable to expect a selection method to perform perfectly all the time. Rather, like any statistical test, a comparison method is developed using an unlimited amount of data (i.e., asymptotically), meaning that the method may not work as well with a small sample of data, but should improve as sample size

increases.

TABLE II. Model Recovery Performance of Model Comparison Methods

Model comparison method	Model fitted:	Data were generated from:		
		M ₂	M ₃	M ₄
ML	M ₂	30	0	2
	M ₃	58	61	8
	M ₄	12	39	90
AIC	M ₂	76	0	7
	M ₃	16	61	5
	M ₄	8	39	88
MDL	M ₂	90	0	15
	M ₃	7	92	6
	M ₄	3	8	79

Note: The percentage of samples in which the particular model fitted the data best. The three models, M₂, M₃ and M₄, are defined in Table I. A thousand samples were generated from each model using the same 10 points of $(x_1, \dots, x_{10}) = (0.001, 1, 2, 4, 7, 10, 13, 16, 20, 30)$. For each probability, a series of $n = 100$ (sample size) independent binary outcomes (0 or 1) were generated according to the binomial probability distribution. The parameter values used to generate the simulated data were as follows: $(a, b) = (2, 3)$ for M₂; $(a, b, c) = (0.3, 0.5, 0.1)$ for M₃ and $(a, b, c) = (500, 0.3, 0.1)$ for M₄. In parameter estimation and also in the calculation of MDL, the following parameter ranges were used: $0 < a < 100, 0 < b < 10$ for M₂; $0 < a < 100, 0 < b < 10, 0 < c < 1$ for M₃; $0 < a < 10,000, 0 < b < 10, 0 < c < 1$ for M₄.

Conclusion

Computational modeling is currently enjoying a heyday as more and more scientists harness the power of statistics and mathematics to develop quantitative descriptions of the phenomenon under study. Progress in this endeavor depends on there being equally sophisticated methods for comparing such models. The aim of this chapter was to introduce the reader to contemporary model selection methods. As stated early on, the problem is ill-posed because in the absence of sufficient data, solutions to the problem are non-unique. To date, the preferred strategy for solving this problem has been to maximize generalizability. It is achieved by evaluating the amount of information in the data relative to the information capacity (i.e., complexity) of the model.

When using any of these selection methods, we advise that the results be interpreted in relation to the other criteria important in model selection. It is easy to forget that AIC and MDL are just fancy statistical tools that were invented to aid the scientific process. They are not the arbiters of truth. Like any such tool, they are blind to the quality of the data and the plausibility of the models under consideration. They will be most useful when considered in the context of the other selection criteria outlined at the beginning of this chapter (e.g., interpretability, falsifiability).

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrox and F. Caski, *Second International Symposium on Information Theory* (pp. 267-281).

- Akademia Kiado, Budapest.
- Berger, J. O. & Berry, D. A. (1998). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159-165.
- Bozdogan, H. (2000). Akaike information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44, 62-91.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44, 108-132.
- Burnham, L. S., & Anderson, D. R. (2002). *Model selection and inference: A practical information-theoretic approach* (2nd edition). Springer-Verlag. New York.
- Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, 49, 997-1003.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall. New York.
- Grunwald, P. (2000). The minimum description length principle. *Journal of Mathematical Psychology*, 44, 133-152.
- Hansen, M. H. & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96, 746-774.
- Jacobs, A. M. & Grainger, J. (1994). Models of visual word recognition—sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 1311-1334.
- Johnson, R. A. & Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis* (4th edition), p. 234-235. Prentice Hall. Upper Saddle, NJ.
- Kass, R. E. & Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association*, 90, 773-795.
- Kass, R. E. & Wasserman, L. W. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343-1370.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79-86.
- Linhart, H., & Zucchini, W. (1986). *Model Selection*. New York, NY: John Wiley & Sons.
- Myung (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44, 190-204.
- Myung, I. J., Forster, M., & Browne, M. W., eds. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, 44, 1-2.
- Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6, 421-425.
- Pitt, M. A., Myung, I. J. & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109, 472-491.
- Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.
- Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11, 416-431.
- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Information Theory* 42, 40-47.
- Sand, S., Filipsson, A. F. & Victorin, K. (2002). Evaluation of the benchmark dose method for dichotomous data: Model dependence and model selection. *Regulatory Toxicology and Pharmacology*, 36, 184-197.
- Schervish, M. J. (1995). *The Theory of Statistics*. New York: Springer-Verlag.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- 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.
- Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9, 60-62.

Author Note

Both authors were supported by NIH Grant R01 MH57472