Psychological Review
2002, Vol. 109, No. 3, 472–491

0033-295X/02/$5.00 DOI: 10.1037//0033-295X.109.3.472

Toward a Method of Selecting Among Computational Models of Cognition

Mark A. Pitt, In Jae Myung, and Shaobo Zhang
The Ohio State University

The question of how one should decide among competing explanations of data is at the heart of the scientific enterprise. Computational models of cognition are increasingly being advanced as explanations of behavior. The success of this line of inquiry depends on the development of robust methods to guide the evaluation and selection of these models. This article introduces a method of selecting among mathematical models of cognition known as *minimum description length*, which provides an intuitive and theoretically well-grounded understanding of why one model should be chosen. A central but elusive concept in model selection, complexity, can also be derived with the method. The adequacy of the method is demonstrated in 3 areas of cognitive modeling: psychophysics, information integration, and categorization.

How should one choose among competing theoretical explanations of data? This question is at the heart of the scientific enterprise, regardless of whether verbal models are being tested in an experimental setting or computational models are being evaluated in simulations. A number of criteria have been proposed to assist in this endeavor, summarized nicely by Jacobs and Grainger (1994). They include (a) plausibility (are the assumptions of the model biologically and psychologically plausible?); (b) explanatory adequacy (is the theoretical explanation reasonable and consistent with what is known?); (c) interpretability (do the model and its parts—e.g., parameters—make sense? are they understandable?); (d) descriptive adequacy (does the model provide a good description of the observed data?); (e) generalizability (does the model predict well the characteristics of data that will be observed in the future?); and (f) complexity (does the model capture the phenomenon in the least complex—i.e., simplest—possible manner?).

The relative importance of these criteria may vary with the types of models being compared. For example, verbal models are likely to be scrutinized on the first three criteria just as much as the last three to thoroughly evaluate the soundness of the models and their assumptions. Computational models, on the other hand, may have already satisfied the first three criteria to a certain level of acceptability earlier in their evolution, leaving the last three criteria to be the primary ones on which they are evaluated. This emphasis on the latter three can be seen in the development of quantitative methods designed to compare models on these criteria. These methods are the topic of this article.

In the last two decades, interest in mathematical models of cognition and other psychological processes has increased tremendously. We view this as a positive sign for the discipline, for it suggests that this method of inquiry holds considerable promise. Among other things, a mathematical instantiation of a theory provides a test bed in which researchers can examine the detailed interactions of a model's parts with a level of precision that is not possible with verbal models. Furthermore, through systematic evaluation of its behavior, an accurate assessment of a model's viability can be obtained.

The goal of modeling is to infer the structural and functional properties of a cognitive process from behavioral data that were thought to have been generated by that process. At its most basic level, then, a mathematical model is a set of assumptions about the structure and functioning of the process. The adequacy of a model is first assessed by measuring its ability to reproduce human data. If it does so reasonably well, then the next step is to compare its performance with competing models.

It is imperative that the model selection method that is used to select among competing models accurately measures how well each model approximates the mental process. Above all else, the method must be valid. Otherwise, the purpose of modeling is undermined. One runs the risk of choosing a model that in actuality is a poor approximation of the underlying process of interest, leading researchers astray. The potential severity of this problem should make it clear that sound methodology is not only integral to but also necessary for theoretical advancement. In short, model selection methods must be as sophisticated and robust as the models themselves.

In this article, we introduce a new quantitative method of model selection. It is theoretically well grounded and provides a clear

Mark A. Pitt, In Jae Myung, and Shaobo Zhang, Department of Psychology, The Ohio State University.

Portions of this work were presented at the 40th annual meeting of the Psychonomic Society, Los Angeles, California, November 18–22, 1999, and at the 31st and 32nd annual meetings of the Society for Mathematical Psychology, Nashville, Tennessee (August 6–9, 1998) and Santa Cruz, California (July 29–August 1, 1999), respectively. We thank D. Bamber, R. Golden, and Andrew Hanson for their valuable comments and attention to detail in reading earlier versions of this article.

Mark A. Pitt and In Jae Myung contributed equally to the article, so order of authorship should be viewed as arbitrary. The section Three Application Examples is based on Shaobo Zhang's doctoral dissertation submitted to the Department of Psychology at The Ohio State University. In Jae Myung and Mark A. Pitt were supported by National Institute of Mental Health Grant MH57472.

Correspondence concerning this article should be addressed to Mark A. Pitt or In Jae Myung, Department of Psychology, The Ohio State University, 1885 Neil Avenue Mall, Columbus, Ohio 43210-1222. E-mail: pitt.2@osu.edu or myung.1@osu.edu

understanding of why one model should be chosen over another. The purpose of the article is to provide a good conceptual understanding of the problem of model selection and the solution being advocated. Consequently, only the most important (and new) technical advances are discussed. A more thorough treatment of the mathematics can be found in other sources (Myung, Balasubramanian, & Pitt, 2000; Myung, Kim, & Pitt, 2000; Myung & Pitt, 1997, 1998). After introducing the problem of model selection and identifying model complexity as a key property of a model that must be considered by any selection method, we introduce an intuitive statistical tool that assists in understanding and measuring complexity. Next, we develop a quantitative measure of complexity within the mathematics of differential geometry and show how it is incorporated into a powerful model selection method known as minimum description length (MDL). Finally, application examples of MDL and the complexity measure are provided by comparing models in three areas of cognitive psychology: psychophysics, information integration, and categorization.

Generalizability Instead of Goodness of Fit

Model selection in psychology has largely been limited to a single criterion to measure the accuracy with which a set of models describes a mental process: goodness of fit (GOF). The model that fits a particular set of observed data the best (i.e., accounts for the most variance) is considered superior because it is presumed to approximate most closely the mental process that generated the data. Typical measures of GOF include the root mean squared error (*RMSE*), which is the square root of the sum of squared deviations between observed and predicted data divided by the number of data points fitted, and the maximum likelihood, which is the probability of obtaining the observed data maximized with respect to the model's parameter values. GOF as a selection criterion is attractive because it appears to measure exactly what one wants to know: How well does the model mimic human behavior? In addition, the GOF measure is easy to calculate.

GOF is a necessary and important component of model selection: Data are the only link to the underlying cognitive process, so a model's ability to describe the output from this process must be considered in model selection. However, model selection based solely on GOF can lead to erroneous results and the choice of an inferior model. Just because a model fits data well does not necessarily imply that the regularity one seeks to capture in the data is well approximated by the model (Roberts & Pashler, 2000). Properties of the model itself can enable it to provide a good fit to the data for reasons that have nothing to do with the model's approximation to the cognitive process (Myung, 2000). Two of these properties are the number of parameters in the model and its functional form (i.e., the way in which the model's parameters and data are combined in the model equation). Together they contribute to a model's *complexity,* which refers to the flexibility inherent in a model that enables it to fit diverse patterns of data.[1] The following simulation example demonstrates the independent contribution of these two properties to GOF.

Three models were compared on their ability to fit data. Model M_1 (defined in Table 1) generated the data to fit, and is therefore considered the "true" model. Model M_2 differed from M_1 only in having one additional parameter, two instead of one; note that their

Table 1
Goodness of Fit and Generalizability Measures of Three Models Differing in Complexity

Model	M_1 (true model)	M_2	M_3
Goodness of fit	2.68 (0%)	2.49 (31%)	2.41 (69%)
Generalizability	2.99 (52%)	3.08 (28%)	3.14 (20%)

Note. Each cell contains the average root mean squared error of the fit of each model to the data and the percentage of samples (out of 1,000) in which that particular model fitted the data best (in parentheses). The three models were as follows: M_1: $y = \ln(x + a) +$ error; M_2: $y = b*\ln(x + a) +$ error; and M_3: $y = bx^a +$ error. The error was normally distributed, $M = 0$, $SD = 3$. Samples were generated from model M_1 using $a = 1$ on the same 6 points for x, which ranged from 1 to 6 in increments of 1.

functional forms are the same. Model M_3 had the same number of parameters as M_2, but a different functional form (a is an exponent of x rather than an additive component). Parameters were chosen for each of the three models to give the best fit to 1,000 randomly generated samples of data from the model M_1. Each model's mean fit to the samples is shown in the first row of Table 1 along with the percentage of time that particular model provided a better fit than its two competitors. As can been seen, M_2 and M_3, with one more parameter than M_1, always provided a better fit to the data than M_1. Because the data were generated by M_1, M_2 and M_3 must have overfitted the data beyond what is necessary to capture the underlying regularity. Otherwise, one would have expected M_1 to fit its own data best at least some of the time. After all, M_1 generated the data! The improved fit of M_2 and M_3 occurred because the extra parameter, b, in these two models enabled them to absorb random error (i.e., nonsystematic variation) in the data. Absorption of these random fluctuations is the only means by which M_2 and M_3 could have fitted the data better than the true model, M_1. Note also that M_3 provided a better fit than M_2. This improvement in fit must be due to functional form rather than the number of parameters, because these two models differ only in how the data (x) and the two parameters (a and b) are combined in the model equation.

This example demonstrates clearly that GOF alone is inadequate as a model selection criterion. Because a model's complexity is not evaluated by the method, the model capable of absorbing the most variation in the data, regardless of its source, will be chosen. Frequently this will be the most complex model. The simulation also highlights the point that model selection is particularly difficult in psychology, and in other social sciences, precisely because random error is present in the data. Although this "noise" can be minimized (in the experimental design), it cannot be eliminated, so in any given data set, variation due to the cognitive process and variation due to random error are entangled, posing a significant obstacle to identifying the best model.

To get around this problem, model selection must be based, instead, on a different criterion—that of generalizability. The goal

[1] Cutting, Bruno, Brady, and Moore (1992) used the term *scope,* which is similar to our definition of *complexity.* They proposed to measure the scope by assessing a model's ability "to account for all possible data functions, where those functions are generated by a reasonably large sample of random data sets" (p. 364).

of generalizability is to predict the statistics of new, as yet unseen, samples generated by the mental process being studied. The rationale underlying the criterion is that the model should be chosen that fits all samples best, not the model that provides the best fit to one particular sample. Only when this condition is met can one be sure a model is accurately capturing the underlying process, not also the idiosyncracies (i.e., random error) of a particular data sample. More formally, *generalizability* can be defined in terms of a discrepancy function that measures the expected error in predicting future data given the model of interest (Linhart & Zucchini, 1986; also see their work for a discussion of the theoretical underpinnings of generalizability).

The results of a second simulation illustrate the superiority of generalizability as a model selection criterion. After each of the data samples was fitted in the first simulation, the parameters of the three models were fixed, and generalizability was assessed by fitting the models to another 1,000 samples of data generated from M_1. The average fits are shown in the second row of Table 1. As can be seen, poor generalizability is the cost of overfitting a specific sample of data. Not only are average fits now worse for M_2 and M_3 than for M_1, but these two models provided the best fit to the second sample much less often than M_1. Generalizability should be preferred over GOF because it does a better job of capturing the general trend in the data and ignoring random variation.

This difference between these two selection criteria is shown in Figure 1. Dots in the panel represent observed data points. Lines are the functions generated by two models varying in complexity. The simpler model (thick line) captures the general trend in the data. If new data points (+) are added to the sample, fit will remain similar. The more complex model (thin line) not only captures the general trend in the data, but also captures many of the idiosyncracies of each observation in the data set, which will cause fit to drop when additional observations are added to the sample. Generalizability would favor the simple model, which fits with our intuitions. GOF, on the other hand, would favor the complex model.

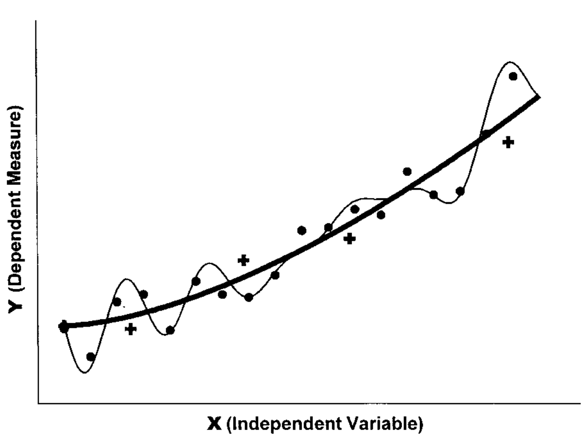

Figure 1. Illustration of the trade-off between goodness of fit and generalizability. An observed data set (dots) was fitted to a simple model (thick line) and a complex model (thin line). New observations are shown by the plus symbol.

The goal of model selection, then, should be to maximize generalizability. This turns out to be quite difficult in practice, because the relationship between complexity and generalizability is not as straightforward as that between complexity and GOF. These differences are illustrated in Figure 2. Model complexity is represented along the horizontal axis and any fit index on the vertical axis, where larger values indicate a better fit (e.g., percent variance accounted for), with the two functions representing the two selection criteria. As was demonstrated in the first simulation, as complexity increases, so does GOF. Generalizability will also increase positively with complexity, but only up to the point where the model is sufficiently complex to capture the regularities in the data caused by the cognitive process being modeled. Any additional complexity will cause a drop in generalizability, because after that point the model will begin to capture random error, not just the underlying process. The difference between the GOF and generalizability curves represents the amount of overfitting that can occur. Only by taking complexity into account can a selection method accurately measure a model's generalizability. The task before the modeling community has been to develop an accurate and complete measure of model complexity, being sensitive not only to the number of parameters in the model but also to its functional form.

Another way to interpret the preceding discussion is that the trademark of a good model is its ability to satisfy the two opposing selection pressures of GOF and complexity, with the end result being good generalizability. These two pressures can be thought of as the two edges of Occam's razor: A model must be complex enough to capture the underlying regularity yet simple enough to avoid overfitting the data sample and thus losing generalizability. In this regard, model selection methods should be evaluated on their success in implementing Occam's razor. The selection method that we introduce in this article, MDL, achieves this goal. Before we describe this method, we review prior approaches to model selection.

Prior Approaches to Model Selection

We begin this section with a formal definition of a model. From a statistical standpoint, data are a sample generated from a true but unknown probability distribution, which is the regularity underlying the data. A statistical model is defined as a collection of probability distributions defined on experimental data and indexed by the model's parameter vector, whose values range over the parameter space of the model. If the model contains as a special case the probability distribution that generated the data (i.e., the "true" model), then the model is said to be correctly specified; otherwise it is misspecified. Formally, define $y = (y_1, \ldots, y_N)$ as a vector of values of the dependent variable, $\theta = (\theta_1, \ldots, \theta_k)$ as the parameter vector of the model, $f(y|\theta)$ as the likelihood function as a function of the parameter θ. N is the number of observations and k is the number of parameters. Often it is possible to write y as a sum of a deterministic component plus random error:

$$y = g(\theta, x) + e. \tag{1}$$

In the equation, $x = (x_1, \ldots, x_N)$ is a vector of an independent variable x, and $e = (e_1, \ldots, e_N)$ is the random error vector from a probability distribution with a mean of zero. Quite often the mean

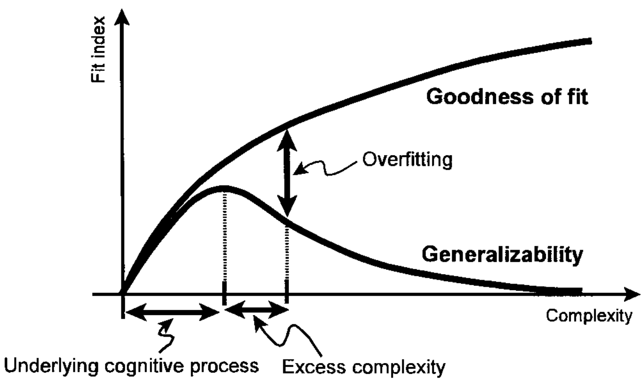

Figure 2. Illustration of the relationship between goodness of fit and generalizability as a function of model complexity (Myung & Pitt, 2001). From *Stevens' Handbook of Experimental Psychology* (p. 449, Figure 11. 4), by J. Wixted (Editor), 2001, New York: Wiley. Copyright 2001 by Wiley. Adapted with permission.

function $g(\theta, x)$ itself is taken to define a mathematical model. However, the specification of the error distribution must be included in the definition of a model. Additional parameters may be introduced in the model to specify the shape of the error distribution (e.g., normal). Often its shape is determined by the experimental task or design. For example, consider a recognition memory experiment in which the participant is required to respond "old" or "new" to a set of pictures presented across a series of n independent trials, with the number of correct responses recorded as the dependent variable. Suppose that a two-parameter model assumes that the probability of a correct response follows a logistic function of the time lag (x_i), for condition i ($i = 1, \ldots, N$), between initial exposure and recognition test, in the form of $g(\theta_1, \theta_2, x_i) = [1 + \theta_1 \exp(-\theta_2 \cdot x_i)]^{-1}$. In this case, the dependent variable y_i will be binomially distributed with probability $g(\theta_1, \theta_2, x_i)$ and the number of binomial trials n, so the shape of error function is completely specified by the experimental task.

Six representative selection methods currently in use are shown in Table 2. They are the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), the root mean squared deviation (RMSD), the information-theoretic measure of complexity (ICOMP; Bozdogan, 1990), cross-validation (CV; Stone, 1974), and Bayesian model selection (BMS; Kass & Raftery, 1995; Myung & Pitt, 1997). Each of these methods assesses a model's generalizability by combining a measure of GOF with a measure of complexity. Each prescribes that the model that minimizes the given criterion be chosen. That is, the smaller the criterion value of a model, the better the model generalizes.[2] A fuller discussion of these methods can be found in Myung, Forster, and Browne (2000; see also Linhart & Zucchini, 1986).

AIC and BIC are the two most commonly used selection methods. The first term, $-2 \ln(f(y|\hat{\theta}))$, is a maximum likelihood measure of GOF and the second term, involving k, is a measure of complexity that is sensitive to the number of parameters in the model. As the number of parameters increases, so does the criterion. In BIC, the rate of increase is modified by the log of the sample size, n.[3] *RMSD* uses *RMSE* as the measure of GOF and also takes into account the number of parameters through k.[4]

These three measures, AIC, BIC, and *RMSD*, are all sensitive only to one aspect of complexity, number of parameters, but insensitive to functional form. This is clearly inadequate because, as demonstrated in Table 1, the functional form of a model influences generalizability. ICOMP is an improvement on this shortcoming. Its second and third terms together represent a complexity measure that takes into account the effects of parameter sensitivity through trace($\mathbf{\Omega}$) and parameter interdependence through det($\mathbf{\Omega}$), which, according to Li, Lewandowski, and DeBrunner (1996), are two principal components of the functional form that contribute to model complexity. However, ICOMP is also problematic because it is not invariant under reparameterization of the model, in particular under nonlinear forms of reparameterization.[5]

[2] The model selection methods discussed in the present article do not require the assumption that the models being compared are correct or nested. (A model is said to be *correct* if there is a parameter value of the model that yields the probability distribution that has generated the observed data sample. A model is said to be *nested* within another model if the former can be reduced to a special case of the latter by setting one or more of its parameters to fixed values.) On the other hand, the generalized likelihood ratio test based on the G^2 or chi-square statistics (e.g., Bishop, Fienberg, & Holland, 1975, pp. 125–127), which are often used to compare two models, assumes that the models are nested and, further, that the reduced model is correct. When these assumptions are met, both types of selection methods should perform similarly. However, the methods should not be viewed as interchangeable because their goals differ. The selection methods presented in this article were designed to identify the model that generalizes best in some defined sense. The generalized likelihood ratio test, in contrast, is a null hypothesis significance test in which the hypothesis that the reduced model is correct is tested given a prescribed level of the Type 1 error rate (i.e., α). Accordingly, the model chosen under this test may not necessarily be the one that generalizes best.

[3] Sample size refers to the number of independent data samples (more accurately, errors, i.e., e_is) drawn from the same probability distribution. Data size is the number of observed data points that are being fitted to evaluate a model and that may come from different probability distributions, although from the same probability family. Often, the sample size is equal to the data size. A case in point is a linear regression model, $y_i = \theta x_i + e_i$ ($i = 1, \ldots, N$), where $e_i \sim N(0, \sigma^2)$. Note that errors, e_is, are independent and identically distributed according to the normal probability distribution with mean zero and variance σ^2. On the other hand, if it is assumed that each e_i is normally distributed with zero mean but with a different value of the variance, that is, $e_i \sim N(0, \sigma_i^2)$, ($i = 1, \ldots, N$), then the sample size, n, will now be equal to 1 whereas the data size, N, remains unchanged.

[4] The RMSD defined in Table 2 differs from the RMSD that has often been used in the psychological literature (e.g., Friedman, Massaro, Kitzis, & Cohen, 1995) where it is defined as RMSD $= \sqrt{\mathrm{SSE}/N}$, in which ($N - k$) is replaced by N, and therefore does not take into account the number of parameters. This form of RMSD is nothing more than *RMSE*. As such, it is not appropriate to use as a method of model selection, especially when comparing models that differ in the number of parameters.

[5] *Reparameterization* refers to transforming the parameters of a model so that it becomes another, behaviorally equivalent, model. For example, a one-parameter exponential model with normal error, $y = e^{\theta x} + N(0, \sigma^2)$ is a reparameterization of another model: $y = \alpha^x + N(0, \sigma^2)$. The latter is

Table 2
Six Prior Model Selection Methods

Selection method	Criterion equation
Akaike information criterion (AIC)	$\mathrm{AIC} = -2 \ln f(y\|\hat{\theta}) + 2k$
Bayesian information criterion (BIC)	$\mathrm{BIC} = -2 \ln f(y\|\hat{\theta}) + k \ln n$
Root mean square deviation (RMSD)	$\mathrm{RMSD} = \sqrt{\mathrm{SSE}/(N - k)}$
Information-theoretic measure of complexity (ICOMP)	$\mathrm{ICOMP} = -\ln f(y\|\hat{\theta}) + \frac{k}{2} \ln\left(\frac{trace[\Omega(\hat{\theta})]}{k}\right) - \frac{1}{2} \ln \det(\Omega(\hat{\theta}))$
Cross-validation (CV)	$\mathrm{CV} = -\ln f(y_{Val}\|\hat{\theta}_{Cal})$
Bayesian model selection (BMS)	$\mathrm{BMS} = -\ln \int f(y\|\theta)\pi(\theta)d\theta$

Note. y = data sample of size n; $\hat{\theta}$ = parameter value that maximizes the likelihood function $f(y|\theta)$; k = number of parameters; SSE = sum of the squared deviations between observed and predicted data; N = the number of data points fitted; $\mathbf{\Omega}$ = covariance matrix of the parameter estimates; y_{Val} = validation sample of observed data; $\hat{\theta}_{Cal}$ = maximum likelihood parameter estimate for a calibration sample; ln = the natural logarithm of base e; $\pi(\theta)$ = the prior probability density function of the parameter.

In CV, the observed data are divided into two subsamples of equal sizes, calibration and validation. The former is used to estimate the best-fitting parameter values of a model. The parameters are then fixed to these values and used by the model to fit the validation sample, yielding a model's CV index. CV is an easy-to-use, heuristic method of estimating a model's generalizability (for a brief tutorial, see Myung & Pitt, 2001). The emphasis on generalizability makes it reasonable to suppose that CV somehow takes into account the effects of functional form. If, how, and how well it does this is not clear, however.

BMS is a model selection method motivated from Bayesian inference. As such, the method chooses models based on the posterior probability of a model given the data. Calculation of the posterior probability requires the specification of the parameter prior density, $\pi(\theta)$, creating the possibility that model selection will depend on the choice of the prior density. As with CV, complexity in BMS is elusive. The integral form of the measure indicates that BMS takes into account functional form and the number of parameters, but how this is achieved is not entirely clear. It can be shown that BIC performs equivalently to BMS as a large sample approximation.

It is important to note that these selection criteria are themselves sample estimates of a true but unknown population parameter (i.e., generalizability in the population), and thus their values can change from sample to sample. Under the model selection procedure described above, however, one is forced to choose one model no matter how small the difference is among models, even when the models are virtually equivalent in their approximation of the underlying regularity. One solution to this dilemma is to conduct a statistical test, before applying the model selection procedure, to decide if two given models provide equally good descriptions of the underlying process. Golden (2000) proposed such a methodology, in which one can determine whether a subset of models are equally good approximations of the cognitive process.[6] If the number of comparisons is not small, however, it can be difficult to control experiment-wise error.

The preceding selection methods represent important progress in tackling the model selection problem. All have shortcomings that limit their usefulness to various degrees. The complexity measure in AIC, BIC, and RMSD is incomplete, and the other three are either not invariant under reparameterization (ICOMP) or lack a clear complexity measure (CV, BMS). The remainder of this article is devoted to the development and testing of a model selection approach that overcomes these limitations. We begin by showing that differential geometry provides a theoretically well-justified and intuitive framework for understanding complexity and model selection in general.

Model Complexity: A Distributional Approach

We begin the discussion of complexity with a graphical definition of the term, intended to clarify what it means for a model to be complex. Depicted in the top panel in Figure 3 is the set of all data patterns that are possible given a particular experimental design. Every point in this multidimensional data space represents a particular data pattern in terms of a probability distribution, such as the shape of a frequency distribution of response times. All models occupy a section, or multiple sections, of data space, being able to fit a subset of the possible data patterns that could be observed. It is equally appropriate to think of data space as the universe of all models under consideration, because every model will occupy a region of this space, large or small.

obtained from the former by defining a new parameter, α, as $\alpha = e^{\theta}$. Whenever two models are related to each other through reparameterization, they become equivalent in the sense that both will fit any given data set identically, albeit with different parameter values. Statistically speaking, they are indistinguishable from one another.

[6] This is a null hypothesis significance test, which, as an extension of the Wilke's generalized likelihood ratio test, tests the null hypothesis that all models under consideration fit the data equally well. This test, unlike the generalized likelihood ratio test, is applicable to comparing non-nested and misspecified models for a wide range of discrepancy functions, including the ones with penalty terms, such as AIC, BIC, and MDL. In the standard model selection procedure using these criteria, one is forced to decide between two models under comparison. This test allows for a third decision that both models are equally good or there is not enough evidence yet for choosing one model over the other.

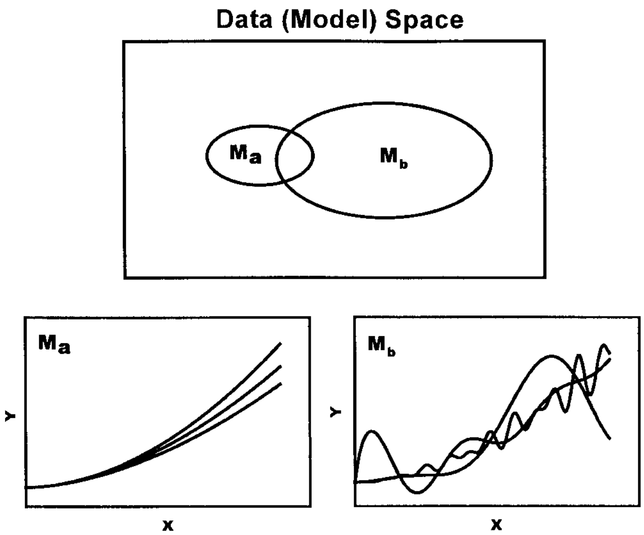

Figure 3. The top panel depicts regions in data space occupied by two models, M_a (simple model) and M_b (complex model), with the range of data patterns that can be generated by each model in the lower panels.

The amount of space occupied by a model is positively related to its complexity. A simple model (M_a) will occupy a small region of data space because it assumes a specific structure in the data, which will manifest itself as a relatively narrow range of similar data patterns. This idea is illustrated in the lower left panel. When one of these few 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 data points outside of its region of data space to disprove the model. In contrast, a complex model (M_b) will occupy a larger portion of data space. Complex models do not assume a single structure in the data. Rather, the structure changes as a function of the parameter values of the model. A slight change in a parameter's value can have a dramatic change in the model's structure. Such chameleon-like behavior enables complex models to be finely tuned to fit a wide range of data patterns. This is illustrated in the lower right panel. Overly complex models are of questionable worth because their ability to fit such a diverse set of data patterns can make them difficult to falsify. In general, a complex model is one with many parameters and a (powerful) nonlinear equation for combining parameters. Complexity is dichotomized in this example for illustrative purposes only. It is more accurate to think of it as a continuum, as depicted in Figure 2.

Response Surface Analysis

Although the examples in Figure 3 are hypothetical, the graphical depiction of mathematical models in this way is not merely illustrative. Response surface analysis (RSA) is a statistical tool that, as in Figure 3, yields graphical representations of models for comparing their relative complexities. In addition, it serves as an informative starting point for the derivation of an elegant quantitative measure of complexity.

RSA is a method for studying geometric relations among responses generated by a mathematical model, often used in nonlinear regression (Bates & Watts, 1988). For a model with k parameters and N observations, the *response surface* is defined as a k-dimensional surface, formed by all possible response vectors that the model can describe. The response surface is embedded in an N-dimensional *data space*, which is the set of all possible response vectors that could be generated independently of a model. The response surface is a hyperplane for a linear model but may be curved when the model is nonlinear. The effects of model complexity on model fit is easily visible when models are compared in the space of response surfaces. This is shown in the following example. See Myung, Kim, and Pitt (2000) for a more detailed discussion.

Consider the following one-parameter power model:

$$y = t^{-\theta} \text{ (power model)}, \quad (2)$$

where y is the response probability (e.g., proportion correct), t is a presentation or retention interval greater than 1, and θ (≥ 0) is a parameter. Suppose that y is measured at two different time intervals, t_1 and t_2. Given two fixed values of t_1 and t_2, the response surface is a line or a curve in a two-dimensional data space composed of (y_{t_1}, y_{t_2}) created by plotting the y values at t_1 against the corresponding y values at t_2 for the full range of the parameter θ, similar to phase plots in dynamical systems research (Kelso, 1995). In essence, a model is represented graphically as a plot of y_{t_1} versus y_{t_2} in data space. For example, for the parameter $\theta = 1$, the y value at $t_1 = 2$ is obtained as $y_{t_1} = (t_1)^{-\theta} = (2)^{-1} = 0.500$. Similarly, the y value at $t_2 = 8$ is obtained as $y_{t_2} = (t_2)^{-\theta} = (8)^{-1} = 0.125$. These two values are then represented as a single point (0.500, 0.125) on the (y_{t_1}, y_{t_2}) plane. Additional points are obtained by varying the full range of the parameter (i.e., $0 \leq \theta < \infty$) to form a continuous curve, which is called the response curve of a model, shown in the middle panel of Figure 4. The equation that describes this relationship can be derived analytically as follows:

$$y_{t_2} = y_{t_1}^{\ln t_2/\ln t_1}. \quad (3)$$

Note that the parameter θ has been removed from the equation. The model is now parameter free, having been redefined as the relationship between two y values instead of a parameter and a y value. Each point on the response curve describes the relationship between two y values that are themselves described perfectly by a power function. Similarly, the response curves for the following one-parameter models can be obtained and are graphed in the adjacent panels in Figure 4:

$$y = 1 - \theta t \text{ (linear model)}$$

$$y = [1.102^{-\theta} \sin(5\theta + \pi t/12) + 1]/2 \text{ (blackhole model)}. \quad (4)$$

RSA provides two valuable insights into model complexity. First, RSA makes the meaning of complexity tangible. The response curve of a model represents a complete visual description of the model (i.e., all of the data patterns it can describe). The curve is the model. Any point that falls on the curve can be perfectly fit by the model. Thus, RSA clearly reveals what patterns of data a model can describe and what patterns it cannot. For example, the response curve of the linear model reveals that the

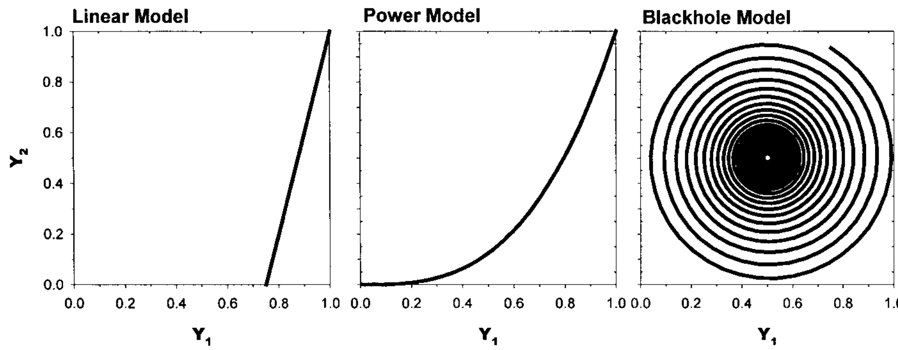

Figure 4. Response curves of three one-parameter models that have the same number of parameters but differ in functional form, each obtained for $t_1 = 2$ and $t_2 = 8$.

model can describe only those (y_2, y_1) points satisfying the equation, $y_2 = 4y_1 - 3$ ($0.75 \leq y_t \leq 1$), no others.

Second, the contributions of functional form to model complexity become evident when models are compared in RSA space. All three models in Figure 4 have one parameter, but their response curves differ greatly, indicating that their functional forms must also differ. This observation leads to an intuitive measure of model complexity: Given that the response surface of a model represents the collection of all possible data patterns that the model can describe, one could define a natural complexity measure as the total length of the model's response curve. For example, for the three response curves in Figure 4, one can conclude that the black-hole model is most complex with its line length of 25.74, followed by the power model (length = 1.50), and then linear model (length = 1.03). Dunn (2000) presents another RSA-based complexity measure.

Despite the possible ways of quantifying complexity within RSA, any such measure would be incomplete because it would not take into account the stochastic nature of the process underlying the data. That is, RSA ignores random variation in the data. The response curves in Figure 4 depict the three models without an error term. Recall that data represent a sample from an unknown probability distribution, the shape of which must be specified by the model. A complete measure of complexity must take into account the distributional characteristics of a model (e), not only that of the mean function, that is, $g(\theta, x)$ in Equation 1. Only the latter is considered in RSA. Thus, any RSA metric would yield only an approximate measure of complexity. To incorporate random error into a complexity measure requires that RSA be extended into a space of probability distributions, to which we now turn.

Differential Geometric Approach to Model Complexity

In this section we show that differential geometry, a branch of mathematics, provides a theoretically well-justified and intuitive measure of model complexity. A more technically rigorous presentation of the topic can be found in Myung, Balasubramanian, and Pitt (2000).

Within differential geometry, a model forms a geometric object known as a *Riemannian manifold* that is embedded in the space of all probability distributions (Amari, 1983, 1985; Rao, 1945). As in the data space depicted in Figure 3, every distribution is a point in this space, and the collection of points created by varying the parameters of the model gives rise to a hypervolume in which similar distributions are mapped to nearby points, as illustrated in Figure 5.

Earlier, we defined complexity as that characteristic of a model that enables it to fit a wide range of data patterns. In a geometric context, this translates into an inherent characteristic of a model that enables it to describe a wide range of probability distributions. Models that are able to describe more distributions should be more complex. Model complexity would therefore seem to be related to the number of probability distributions that a model can generate. This intuition immediately runs into trouble: The number of all such distributions is uncountably infinite, making the value indeterminable. Or is it?

Given that not all distributions are equally similar to one another, one solution is to count only distinguishable distributions. That is, if two or more probability distributions on a model's manifold are sufficiently similar to one another to be statistically indistinguishable, they are counted as one distribution, with a cluster of such distributions occupying a local neighborhood on the manifold. This procedure yields a countably infinite set of distinguishable distributions, the size of which is a natural measure of complexity. More precisely, two probability distributions should be considered indistinguishable if one is mistaken for the other

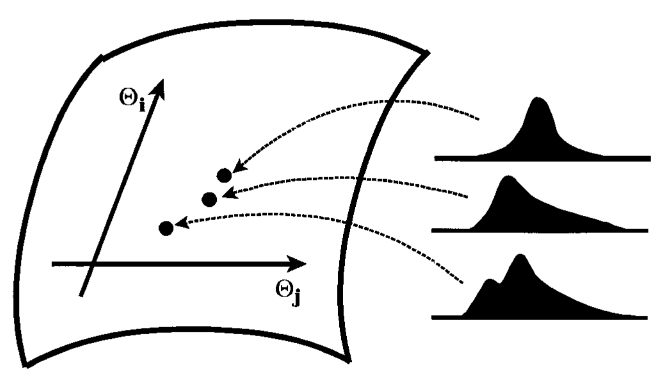

Figure 5. The space of probability distributions forms a manifold on which "similar" distributions are mapped to nearby points.

even in the presence of an infinite amount of data. A measure of volume that counts only distinguishable distributions must be devised to achieve this goal.

The following mental exercise shows how this can be done. Draw data from one distribution, which is indexed by a specific parameter, say θ_p, in the model, and ask how well one can guess whether the data came from θ_p rather than from a nearby θ_q. The ability to distinguish between these distributions increases with the amount of available data. However, it can be shown that for any fixed amount of data there is a little ellipsoid around θ_p where the probability of error in the guessing game is large. In other words, within this ellipsoid, distributions are not very distinguishable in the statistical sense. To count distinguishable distributions, one should then tile the model manifold with such ellipsoids, counting one distribution for each ellipsoid. This procedure turns the manifold into an ellipsoid-covered lattice with a distinguishable distribution at each lattice point. Then the limit of infinite sample size should be taken so that the ellipsoids of indistinguishability shrink and the associated lattice becomes finer, forming a continuum in the limit. Taking this limit recovers a continuum measure that counts only distinguishable distributions. When this computation is carried out, the number of distinguishable distributions turns out to be equal to $d\theta\{\det[I(\theta)]\}^{1/2}$ where $I(\theta)$ is the Fisher information matrix of a sample of size 1, $\det(I)$ is the determinant of the matrix I, and $d\theta$ the infinitesimal parameter volume (see Footnote 6 for a definition of the Fisher Information matrix; see also Schervish, 1995).

The number of all distinguishable probability distributions that a model can generate or describe is obtained by integrating $d\theta\{\det[I(\theta)]\}^{1/2}$ over the entire parameter manifold as follows:

$$V_M = \int d\theta \sqrt{\det[I(\theta)]}, \tag{5}$$

where the subscript M denotes a particular model under consideration. This measure is known as the *Riemannian volume* in differential geometry. A highly desirable property of the volume measure is that it is invariant under reparameterization. This property is an outgrowth of models being represented as manifolds in the space of all probability distributions. In this context, the parameters of a model simply index the collection of distributions a model describes. The choice of the parameters themselves is irrelevant. The manifold is the model, which will never change, regardless of how the model is specified in an equation (see Equation 10 and accompanying text).

The Riemannian volume makes good sense as a complexity measure. Because complexity is related to the volume of a model in the space of probability distributions, the measure of volume should count only different, or distinguishable, distributions, and not the coordinate volume ($\int d\theta$) of the manifold. The Riemannian volume, therefore, is a direct function of the number of distinguishable distributions that a model can generate, with a complex model generating more distributions than a simple model.

Relation to RSA

The differential geometric approach to model complexity is similar to RSA in that a mathematical model is viewed as a geometric shape embedded in a hyperdimensional space, albeit different spaces (probability distributions vs. response vectors). This correspondence is not accidental, because the differential geometric approach is a logical extension of RSA. To understand the connection between the two, think of model selection as an inference game: The goal is to determine, out of a set of probability distributions that index data patterns, which model is most likely to have generated a data sample drawn from an unknown probability distribution. Referring back to Equation 1, the main yardstick used in this selection process is the likelihood function, $f(y|\theta)$. The value of the likelihood function depends upon not only the mean function, $g(\theta, x)$, but also the distributional characteristics of the error term (e). Any justifiable measure of complexity should take into account these two factors. RSA considers only the first term, whereas the differential geometric approach considers both.

To see how the two approaches are related quantitatively, consider the response curve of a one-parameter model, such as the power model in Figure 4 (middle panel). The RSA measure of complexity in this model is the total length of the response curve, which in essence measures the "number" of data points along the curve. In the differential geometric approach, counting is carried out with the additional knowledge of the local distinguishability of data points along the curve. This difference is illustrated in Figure 6 for the one-parameter power model. The response curve is split into segments of different lengths, with the points within each segment being statistically indistinguishable. Note that distinguishability is not uniform along the curve. Points in the middle region are less distinguishable than those at either end. In fact, for any one-parameter model of observed data that follows a binomial probability distribution, one can derive formal expressions for these two measures of complexity as follows (see Appendix A for a complete derivation):

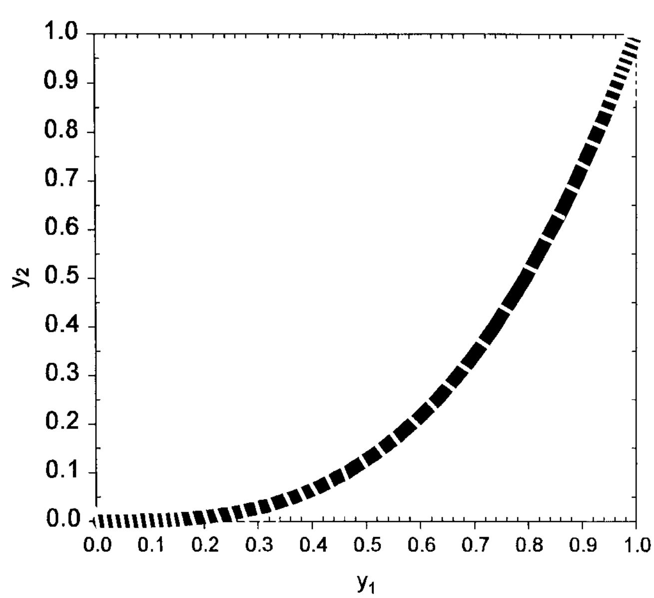

Figure 6. The power model's response curve from Figure 5 divided into local regions of indistinguishability (i.e., the points within each region are statistically indistinguishable).

$$\text{RSA: length } L_M = \int d\theta \sqrt{\sum_{q=1}^{N} \left(\frac{dg(\theta, x_q)}{d\theta} \right)^2} ;$$

Differential geometry: volume V_M

$$= \int d\theta \sqrt{\sum_{q=1}^{N} \frac{1}{g(\theta, x_q)[1 - g(\theta, x_q)]} \left(\frac{dg(\theta, x_q)}{d\theta} \right)^2} . \quad (6)$$

In the equations, it is assumed that observed data, y_q, are distributed binomially, Bin[n, $g(\theta, x_q)$], of sample size n, and probability $g(\theta, x_q)$. Note that the two measures are identical except for the additional term, $1/\{g(\theta, x_q)[1 - g(\theta, x_q)]\}$, in the differential geometric complexity measure. This extra term takes into account local distinguishability and is equal to det[$I(\theta)$] in Equation 5.

MDL Method of Model Selection

Thus far in the article we have introduced a measure of model complexity. Although it is useful for comparing the relative complexities of models, as will be shown below, by itself the measure is insufficient as a model selection method. What is missing is a measure of how well the model fits the data (i.e., a measure of GOF). MDL, a model selection method from algorithmic coding theory in computer science (Grunwald, 2000; Rissanen, 1983, 1996) combines both of these measures.

The MDL approach to model selection was developed within the domain of information theory, where the goal of model selection is to choose the model that permits the greatest compression of data in its description. The assumption underlying the approach is that regularities or patterns in data imply redundancy. The more the data can be compressed by extracting this redundancy, the more we learn about the underlying regularities governing the cognitive process of interest. The full form of the measure is shown below. The first term is the GOF measure and the second and third together form the intrinsic complexity of the model (Rissanen, 1996).

$$\text{MDL} = -\ln f(y|\hat{\theta}) + \frac{k}{2} \ln\left(\frac{n}{2\pi}\right) + \ln \int d\theta \sqrt{\det[I(\theta)]}, \quad (7)$$

where $y = (y_1, \ldots, y_n)$ is a data sample of size n, $\hat{\theta}$ is the maximum likelihood parameter estimate, ln is the natural logarithm of base e, $I(\theta)$ is the Fisher information matrix defined earlier.[7] The integration of the third term is taken over the parameter space defined by the model.

As with prior selection methods, MDL prescribes that the model that minimizes the criterion should be chosen, the assumption being that such a model has extracted the most redundancy (i.e., regularity) in the data and thus should generalize best. In practice, the criterion represents the shortest length of computer code (measured in bits of information) necessary to describe the data given the model. The shorter the code, the greater the amount of regularity in the data that the model uncovered. The soundness of MDL as a model selection criterion has been well documented by Li and Vitanyi (1997), who showed that there is a close relationship between minimizing MDL and achieving good generalizability. From a decision-theoretic perspective, MDL selects the one model, among a set of competing models, that minimizes the expected error in predicting future data in which the prediction error is measured using a logarithmic discrepancy function (Rissanen, 1999; Yamanishi, 1998).

It turns out that minimization of MDL corresponds to maximization of the posterior probability within the Bayesian statistics framework (i.e., BMS). Balasubramanian (1997) showed that the MDL criterion can be derived as a finite series of terms in an asymptotic expansion of the Bayesian posterior probability of a model given the data for a special form of the parameter prior density. This connection between the two suggests that choosing the model that gives the shortest description of the observed data is essentially equivalent to choosing the model that is most likely "true" in the sense of probability theory (see Theorem 1 of Vitanyi & Li, 2000).

The theoretical link between BMS and MDL also suggests that they may perform similarly in practice. Barron and Cover (1991) showed that BMS and MDL are asymptotically equivalent given large sample sizes; that is, both will converge to the true model if the true model is correctly specified. On the other hand, if models are misspecified and sample size is relatively small, they can yield disparate results, especially depending on the form of the parameter prior density used in the calculation of BMS.

Despite these similarities, MDL has at least one advantage over BMS: The complexity measure is well understood. As mentioned above, complexity and GOF are not easily disentangled in the integral form of BMS (Table 2). In contrast, a clear understanding of the complexity term in MDL is provided by its counterpart in differential geometry, the geometric complexity measure. This is described in detail in the following section.

The latter two terms of the MDL criterion (Equation 7) readily lend themselves to a differential geometric interpretation, which is related to the Riemannian volume measure presented earlier. Conceptually, model selection using MDL proceeds by choosing the model that best approximates the true model by counting the number of distinguishable distributions that come close to the true model. Proximity to the true model is assessed by $f(y|\theta)$. Within the differential geometric approach, this corresponds to a volume measure in the space of probability distributions. The following volume, under the assumption of large sample size, is shown to be a valid measure of proximity (Balasubramanian, 1997; Myung, Balasubramanian, & Pitt, 2000): $C_M = (2\pi/n)^{k/2}\, h(\hat{\theta})$, where k is the number of parameters in the model and $h(\hat{\theta})$ is a data-dependent factor that goes to 1 as n grows large (some additional conditions are required; see Balasubramanian, 1997). Essentially, C_M represents the Riemannian volume of a small ellipsoid around $\hat{\theta}$, within which the probability of the data, $f(y|\theta)$, is appreciable. As such, it measures the number of distinguishable distributions

[7] The Fisher information matrix $I(\theta)$ of the MDL criterion is defined as $I_{ij}(\theta) = (-1/n)E(\partial^2 \ln f(y|\theta)/\partial\theta_i\partial\theta_j)$ ($i, j = 1, \ldots, k$) for the data vector $y = (y_1, \ldots, y_n)$ where y_qs are sample values of random variables, Y_qs ($q = 1, \ldots, n$; see, e.g., Rissanen, 1996, Equation 7). Further, if Y_qs are independently and identically distributed, the above $I(\theta)$ reduces to the Fisher information matrix of sample size $n = 1$, that is, $I_{ij}(\theta) = -E(\partial^2 \ln f(y_q|\theta)/\partial\theta_i\partial\theta_j)$ ($i, j = 1, \ldots, k$) for any q.

that come close to the truth, as measured by predicting the data y with relatively high probability.

However, C_M alone is not an adequate measure of proximity because the total number of distinguishable distributions of a model (V_M, the Riemannian volume, Equation 5) must also be considered. Inclusion of this additional measure leads to a volume ratio, V_M/C_M, which penalizes models for having an unnecessarily large number of distinguishable distributions (V_M) or having relatively few distinguishable distributions close to the truth (C_M). Taking the log of this ratio gives

$$\ln\left(\frac{V_M}{C_M}\right) = \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int d\theta\,\sqrt{\det[I(\theta)]} + \ln h(\hat{\theta}). \quad (8)$$

The first and second terms are independent of the true distribution as well as the data, and therefore represent an intrinsic property of the model. Together they will be called the *geometric complexity* of the model, and are invariant under reparameterization of the model. As sample size n increases, the third term, which is data dependent, becomes negligible. When this occurs, the geometric complexity is equal to the complexity penalty in the MDL criterion in Equation 7.

It is also worth noting that the first term of the geometric complexity measure increases logarithmically with sample size n, whereas the second term is independent of n. An implication of this is that as n grows large, the effects of complexity due to functional form, reflected through $I(\theta)$, will gradually diminish compared to those due to the number of parameters (k). Thus, functional form effects will have their greatest impact on model selection when sample size is small. Because small samples are the norm in experiments in much of cognitive psychology, it is imperative that the selection method be sensitive to this property of a model.

Differential geometry provides many valuable insights into model complexity and model selection. One is a new explication of MDL. The MDL selection criterion can be rewritten as follows:

$$\text{MDL} = -\ln\left(\frac{f(y|\hat{\theta})}{V_M/C_M}\right) = -\ln(\text{“normalized } f(y|\hat{\theta})\text{”}). \quad (9)$$

This reinterpretation provides a clearer picture of what MDL does in model selection. It selects the model that gives the highest value of the maximum likelihood per the relative ratio of distinguishable distributions (V_M/C_M). We might call this the *normalized maximum likelihood*. From this perspective, the better model is the one with many distinguishable distributions close to the truth but few distinguishable distributions overall.

Perhaps the most important insight provided by differential geometry is an intuitive understanding of the meaning of complexity in MDL: It measures the minus log of the volume of the distinguishable distributions in a model relative to those close to the truth. In this regard, the size of a model manifold in the space of distributions is what matters when measuring complexity. A model's functional form and its number of parameters can be misleading indicators of complexity because they are simply the apparatus by which a collection of distributions defined by the model is indexed. The geometric approach to complexity presented here makes it clear that neither the parameterization nor the specific functional form used in indexing is relevant so long as the same collection of distributions is catalogued on the model manifold. For example, the following two models, although assuming different functional forms, are equivalent and equally complex in the geometric sense:

$$\text{Model A: } y = \exp(\theta_1 x_1 + \theta_2 x_2) + \text{error},$$

$$\text{Model B: } y = \eta_1^{x_1}\eta_2^{x_2} + \text{error}, \quad (10)$$

where the error has zero mean and follows the same distribution for both models. Here, the parameters of Model A are related to the parameters of Model B through $\eta_i = \exp(\theta_i)$, $i = 1, 2$.

Three Application Examples

Geometric complexity and MDL constitute a powerful pair of model evaluation tools. When used together in model testing, a deeper understanding of the relationship between models can be gained. The first measure enables one to assess the relative complexities of the set of models under consideration. The second builds on the first by suggesting which model is preferable given the data in hand. The following simulations demonstrate the application of these methods in three areas of cognitive modeling: psychophysics, information integration, and categorization. In each example, two competing models with the same number of parameters but different functional forms were fitted to data sets generated by each of these models (human data were not used). Of interest is the ability of each selection method to recover the model that generated the data. A good selection method should be able to discriminate between data generated by a model from those generated by another model. That is, it should be able to “see through” the random variation in the data sample and accurately infer whether the model being tested generated the data it is being fit to. Errors are a sign of overgeneralization and reveal a bias in the selection method, which could be toward either the more complex or simpler model. The ideal pattern of data is one in which each model generalizes best only to data generated by itself, not to data generated by the competing model. In the 2×2 sections of Tables 3–5, this corresponds to a mean selection criterion measure that is lowest in the upper left and lower right quadrants, with perfect recovery rates (100%) in these cells as well.

Four selection methods were compared: AIC, ICOMP, CV, and MDL. Given the close relationship between MDL and BMS, the latter was not included in the comparison. BIC and RMSD were also not included because of their equivalence to AIC in the present testing conditions. AIC can be expressed with BIC as a term in the equation: AIC = BIC + k (2 − ln n). Consequently, both methods will yield the same results when models with the same number of parameters (i.e., equal k) are compared. RMSD will generally yield the same outcome as well.[8] A fuller discussion of the three simulations can be found in Zhang (1999).

[8] When comparing among models with the same number of parameters, model selection under RMSD will be the same as that under AIC and BIC when errors are normally distributed and have equal variances. This is because in such cases the sum of squares error in RMSD is related to the likelihood function in AIC (and BIC) as SSE $\propto -\ln f(y|\theta)$ and hence, minimization of SSE is equivalent to maximization of the likelihood

Table 3
Comparison of Four Selection Methods on Their Ability to Generalize Accurately Using Two Psychophysical Models

Selection method/ model fitted	Data from Stevens's	Data from Fechner's
AIC		
Stevens's	12.42 (100%)	11.92 (47%)
Fechner's	52.22 (0%)	11.86 (53%)
ICOMP		
Stevens's	4.92 (100%)	5.69 (100%)
Fechner's	25.40 (0%)	10.64 (0%)
CV		
Stevens's	9.16 (94%)	10.70 (49%)
Fechner's	25.30 (6%)	11.01 (51%)
MDL		
Stevens's	10.71 (100%)	10.46 (0%)
Fechner's	25.40 (0%)	5.21 (100%)

Note. For each method and model, the mean criterion value and the percentage of samples (in parentheses) in which the particular model was selected under the given method are shown. A thousand samples were generated from each model using the same 6 points for X, which ranged from 1 to 6 in one-step increments. The random error was normally distributed with a mean of zero and a standard deviation of 0.3. The parameter values used to generate the simulated data were $a = 2$ and $b = 2$ for Stevens's model and $a = 2$ and $b = 5$ for Fechner's model. AIC = Akaike information criterion; ICOMP = information-theoretic measure of complexity; CV = cross-validation; MDL = minimum description length.

Psychophysics

Models of psychophysics (Roberts, 1979) were developed to describe the relationship between physical dimensions (e.g., light intensity) and their psychological counterparts (e.g., brightness). Two of the most influential have been Stevens's power model and Fechner's logarithmic model.

$$\text{Stevens's model: } Y = aX^b + \text{error},$$

$$\text{Fechner's model: } Y = a\ln(X + b) + \text{error}. \quad (11)$$

In both models, error is assumed to be normally distributed with a mean of 0 and a standard deviation of 0.3. Data samples were generated from each model using fixed parameter values. Each model was then fitted to both data samples using each of the four selection criteria. The results are displayed in Table 3. Shown are the mean criterion values and the percentage of samples (out of 1,000) by which the specified model bested its competitor. Under AIC, Stevens's model was always selected when fitting its own data (100% vs. 0%) but was selected about equally often as Fechner's model when the data were generated by Fechner's model (47% vs. 53%). This asymmetry demonstrates that AIC overestimated the generalizability of Stevens's model relative to Fechner's model. That is, Stevens's model was shown to generalize better to data generated by Fechner's model than Fechner's model itself. CV and ICOMP performed no better, with model recovery rates for CV comparable to those of AIC but considerably worse for ICOMP.

Because both models have the same number of parameters, the failure of these three selection methods can be attributed to a complexity term that does not adequately incorporate functional form. When MDL was used as the selection method, the model recovery rates were perfect for both models, demonstrating the superiority of the method's complexity measure. (An example of how to calculate MDL and the geometric complexity measure for these two models is provided in Appendix B.)

Calculation of the geometric complexities of the two models reinforces and further elucidates the MDL findings. Stevens's model is more complex than Fechner's, with the complexity difference being equal to 5.52.[9] Given the logarithmic relationship between geometric complexity and the number of distinguishable distributions, this means that for every distribution for which Fechner's model can account, Stevens's model can describe about $e^{5.52} \approx 250$ distributions. These results clearly demonstrate that appropriately accounting for the complexity of a model is essential to model selection. Furthermore, the geometric complexity results validate a long-held suspicion regarding the source of the superior data-fitting abilities of Stevens's model (Townsend, 1975).

Information Integration

In a typical information integration experiment, a range of stimuli is generated from a factorial manipulation of two or more stimulus dimensions (e.g., visual and auditory) and then presented to participants for categorization as one of two or more possible response alternatives. Data are scored as the proportion of responses in one category across the various combinations of stimulus dimensions. For this comparison, we consider two models of information integration, the fuzzy logical model of perception (FLMP; Oden & Massaro, 1978) and the linear integration model (LIM; Anderson, 1981). Each assumes that the response probability (p_{ij}) of one category, say A, on the presentation of a stimulus of the specific i and j feature dimensions in a two-factor information integration experiment takes the following form:

$$\text{FLMP: } p_{ij} = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1 - \theta_i)(1 - \lambda_j)},$$

$$\text{LIM: } p_{ij} = \frac{\theta_i + \lambda_j}{2}, \quad (12)$$

where θ_i and λ_j ($i = 1, \ldots, q_1$; $j = 1, \ldots, q_2$; $0 < \theta_i, \lambda_j < 1$) are parameters representing the corresponding feature dimensions. Again, the two models were fitted to data sets generated by each model. In addition, the sample size, n, was varied in this example to demonstrate its influence on the performance of these selection methods. The results are shown in Table 4.

function. For non-normal or unequal-variance errors, no such relationship exists between SSE and the likelihood function. Therefore, model selection under RMSD will differ from that under AIC and BIC. It has been our experience that RMSD performs worse, never better, than these two methods under such conditions. This appears to be due to RMSD's insensitivity to unequal variances.

[9] In computing geometric complexity measures, the following parameter ranges were assumed: $0 < a < \infty$, $0 < b < 3$ for Stevens's model, and $0 < a, b < \infty$ for Fechner's model.

Table 4
Generalizability Comparisons of Four Selection Methods Over Three Sample Sizes Using Two Information Integration Models

Selection method/ model fitted	Data from FLMP	Data from LIM
Sample size: $n = 20$		
AIC		
FLMP	59.86 (100%)	74.96 (51%)
LIM	89.82 (0%)	75.24 (49%)
ICOMP		
FLMP	25.07 (100%)	30.27 (25%)
LIM	40.32 (0%)	29.63 (75%)
CV		
FLMP	30.43 (100%)	37.78 (29%)
LIM	42.88 (0%)	37.04 (71%)
MDL		
FLMP	34.56 (89%)	42.11 (0%)
LIM	40.80 (11%)	33.51 (100%)
Sample size: $n = 60$		
AIC		
FLMP	77.94 (100%)	93.88 (25%)
LIM	159.72 (0%)	92.32 (75%)
ICOMP		
FLMP	33.55 (100%)	39.45 (14%)
LIM	73.82 (0%)	37.95 (86%)
CV		
FLMP	39.11 (99%)	47.28 (28%)
LIM	76.58 (1%)	45.80 (72%)
MDL		
FLMP	43.70 (100%)	51.71 (0%)
LIM	75.75 (0%)	42.06 (100%)
Sample size: $n = 150$		
AIC		
FLMP	92.98 (100%)	112.22 (17%)
LIM	293.74 (0%)	107.84 (83%)
ICOMP		
FLMP	40.98 (100%)	48.56 (10%)
LIM	140.36 (0%)	45.68 (90%)
CV		
FLMP	47.33 (100%)	55.88 (15%)
LIM	143.56 (0%)	53.32 (85%)
MDL		
FLMP	51.21 (100%)	60.83 (0%)
LIM	142.76 (0%)	49.82 (100%)

Note. For each method and model, the mean criterion value and the percentage of samples (in parentheses) in which the particular model was selected under the given method are shown. Simulated data in a 2×8 factorial design ($q_1 = 2$; $q_2 = 8$) were created from predetermined values of the 10 parameters, $\theta = (0.15, 0.85)$, $\lambda = (0.05, 0.10, 0.26, 0.42, 0.58, 0.74, 0.90, 0.95)$. From these, 16 binomial response probabilities (p_{ij}) were computed using each model equation. For each probability, a series of n (i.e., $n = 20$, 60, or 150) independent binary outcomes (0 or 1) were generated according to the binomial probability distribution. The number of ones in the series was summed and divided by n to obtain an observed proportion. This way, each sample consisted of 16 observed proportions. For each sample size, 1,000 samples were generated from each of the two models. In parameter estimation as well as complexity calculation, θ_1 was fixed to 0.15 so the models became identifiable. FLMP = fuzzy logical model of perception; LIM = linear integration model; AIC = Akaike information criterion; ICOMP = information-theoretic measure of complexity; CV = cross-validation; MDL = minimum description length.

Table 5
Generalizability Comparisons of Three Selection Methods Over Three Sample Sizes Using Two Categorization Models

Selection method/ model fitted	Data from GCM	Data from PRT
Sample size: $n = 20$		
AIC		
GCM	74.34 (98%)	80.36 (15%)
PRT	85.24 (2%)	75.66 (85%)
CV		
GCM	37.70 (86%)	40.33 (23%)
PRT	41.46 (14%)	37.62 (77%)
MDL		
GCM	38.96 (96%)	41.97 (7%)
PRT	43.82 (4%)	39.03 (93%)
Sample size: $n = 60$		
AIC		
GCM	94.26 (98%)	106.90 (4%)
PRT	122.10 (2%)	93.44 (96%)
CV		
GCM	45.93 (99%)	52.37 (9%)
PRT	59.70 (1%)	46.58 (91%)
MDL		
GCM	48.93 (98%)	55.25 (4%)
PRT	62.25 (2%)	47.92 (96%)
Sample size: $n = 150$		
AIC		
GCM	114.06 (99%)	137.36 (1%)
PRT	178.80 (1%)	106.56 (99%)
CV		
GCM	53.30 (99%)	68.22 (4%)
PRT	90.09 (1%)	53.16 (96%)
MDL		
GCM	58.83 (99%)	70.48 (1%)
PRT	90.60 (1%)	54.48 (99%)

Note. The mean criterion value and the percentage of samples (in parentheses) in which the particular model was selected under the given method are shown. Simulated data were created from predetermined values of the seven parameters, $c = 2.5$, $w = (0.2, 0.2, 0.2, 0.2, 0.1, 0.1)$. From these, 27 trinomial response probabilities (p_{iJ}, $i = 1, \ldots, 9$, $J = 1, 2, 3$) were computed using each model equation. For each probability, a series of n (i.e., $n = 20$, 60, or 150) independent ternary outcomes were generated according to the trinomial probability distribution. The number of outcomes of each type in the series was summed and divided by n to obtain an observed proportion. This way, each sample consisted of 27 observed proportions. For each sample size, 1,000 samples were generated from each of the two models. GCM = generalized concept model; PRT = prototype model; AIC = Akaike information criterion; CV = cross-validation; MDL = minimum description length.

When the data were generated by FLMP, regardless of the selection method and sample size, FLMP was recovered 100% of the time, except for MDL when sample size was 20. In this case, MDL performed slightly worse than the other selection methods. When the data were generated by LIM, all prior selection methods faired much more poorly, except MDL, which recovered the correct model (LIM) perfectly across all sample sizes. When AIC was used, FLMP was selected over LIM half of the time with a sample size of 20 (51% vs. 49%). Such errors were reduced considerably by the time the sample size reached 150 (17% vs. 83%). With ICOMP and CV, this erroneous selection bias in favor of FLMP is less severe, but even with the largest sample size, all prior selection methods except MDL failed to recover the correct model at least 10% of the time.

That FLMP was selected over LIM when methods such as AIC and CV were used, even when the data were generated by LIM, suggests that FLMP is more complex than LIM. This observation was confirmed when the geometric complexity of each model was calculated. The difference in geometric complexity between FLMP and LIM was 8.74, meaning that for every distribution for which LIM can account, FLMP can describe about $e^{8.74} \approx 6{,}248$ distributions.[10] As the simulation results show, the complexity term in MDL does an excellent job of correcting for this difference and minimizing overgeneralization of the more complex model.

Categorization

Two models of categorization were considered in the present demonstration. They were the generalized context model (GCM; Nosofsky, 1986) and the prototype model (PRT; Reed, 1972). Each model assumes that categorization responses follow a multinomial probability distribution with p_{iJ} (probability of category C_J response given stimulus X_i), which is given by:

$$\text{GCM: } p_{iJ} = \frac{\sum_{j \in C_J} s_{ij}}{\sum_K \sum_{k \in C_K} s_{ik}}$$

$$\text{where } s_{ij} = \exp\left(-c \cdot \left[\sum_{m=1}^{M} w_m |x_{im} - x_{jm}|^r\right]^{1/r}\right),$$

$$\text{PRT: } p_{iJ} = \frac{s_{iJ}}{\sum_K s_{iK}}$$

$$\text{where } s_{iJ} = \exp\left(-c \cdot \left[\sum_{m=1}^{M} w_m |x_{im} - x_{Jm}|^r\right]^{1/r}\right). \tag{13}$$

In the equation, s_{ij} is a similarity measure between multidimensional stimuli X_i and X_j, s_{iJ} is a similarity measure between stimulus X_i and the prototypic stimulus X_J of category C_J, M is the number of stimulus dimensions, c is a sensitivity parameter, w_m is an attention-weight parameter, and r is the Minkowski metric parameter. The two models were fitted to data sets generated by each model using the six-dimensional scaling solution from Experiment 1 of Shin and Nosofsky (1992) under the Euclidean distance metric of $r = 2$. As in the last example, three sample sizes were again used.

The results are shown in Table 5. With AIC, virtually no bias in model recovery rate was observed for $n = 60$ and $n = 150$. Only a small bias toward choosing GCM was found using data generated from PRT when $n = 20$. CV shows a similar pattern of model recovery, with errors being highest with the smallest sample size and there being a slightly larger bias in favor of GCM.[11] When MDL was used to choose between the two models, there was no improvement over the other selection methods when the sample sizes were 60 and 150. Even when sample size was 20, the improvement in model recovery was quite modest relative to AIC. Note that this outcome contrasts with that of the preceding examples, in which MDL was generally superior to the other selection methods when sample size was smallest.

On the face of it, these findings might suggest that MDL is not much better than the other selection methods in measuring generalizability. After all, what else could cause this result? The only circumstance in which such an outcome is predicted using MDL is when the functional forms of the two models are similar, thus minimizing the differential contribution of functional form in the complexity term (recall that the two models have the same number of parameters). Calculation of the geometric complexity of each model confirmed this prediction. GCM is indeed only slightly more complex than PRT, the difference being equal to 0.60, so GCM can describe about two distributions ($e^{0.60} \approx 1.8$) for every distribution PRT can describe.[12]

The results of these three simulations demonstrate the accuracy of MDL in choosing computational models of cognition. MDL's sensitivity to functional form was clearly demonstrated in its superior generalizability (i.e., good model recovery rate) across all three examples, especially when it counts most: when sample size was small and the complexities of the models differed by a nontrivial amount. But no selection method will always perform as well as one would like. This was found to be true for MDL in the information integration simulation when sample size equaled 20: MDL performed slightly worse than the other three selection methods. One point to be made about this outcome is that regions of data space (Figure 3) can always be found in which any selection method will perform less than optimally. Because these regions are not known in advance, it makes the most sense to use the most robust method available. MDL is the clear choice.

The simulations also demonstrate the value of an independent measure of complexity when comparing models. The measure not only identifies which model is more complex, but knowing this information can provide additional insight into the model selection process. For example, the results of an MDL analysis might lead to Model A being chosen over Model B, even though the results of a complexity analysis showed Model A to be the more complex of the two. This outcome would suggest that the additional complexity of A is necessary to capture the underlying regularity in the data. In other words, the mental process may not be as simple as one might have originally thought.

When comparing models, it might be most efficient to compare their relative complexities first before applying a selection method. If the complexities differ significantly, then MDL is the better choice of a selection method. If they are about equally complex,

[10] In computing geometric complexity measures, the following parameter ranges were assumed: $0.001 < \theta_i, \lambda_j < 0.999$ for both FLMP and LIM.

[11] Because ICOMP performed similar to AIC and CV in the first two simulations, it was not included in the third test of the selection methods.

[12] In computing geometric complexity measures, the following parameter ranges were assumed: $0 < c < 5$, $0 < w_i < 1$ ($i = 1, \ldots, 6$), $w_i = 1$ for both GCM and PRT.

then the choice of a selection method will matter less. Short of doing this, MDL is the safest method to use.

In sum, these examples with simulated data in which the true model was known a priori demonstrate the usefulness of MDL and geometric complexity for selecting among computational models of cognition. An obvious next step is to extend their application to human data as well as to other types of models in the discipline. Such endeavors are currently underway.

General Discussion

The purpose of this article is to introduce the psychological community to MDL as a method of selecting among mathematical models of cognition and to demonstrate its advantages over existing methods. We began by discussing shortcomings of the most widely used selection criterion, GOF. The simulation data in Table 1 demonstrate that GOF alone is an insufficient criterion with which to compare models. Instead, generalizability should be the goal of model selection, for it overcomes the problem inherent in GOF of teasing apart random variation in the data sample from variation due to the cognitive process.

Fit and complexity were identified as two properties of a model that any selection method must be sensitive to. As schematized in Figure 2, optimization of these two opposing properties defines the model selection problem. What is needed, then, is a selection method that acts as a fulcrum on which fit and complexity can be balanced. The development of a theoretically justifiable measure of complexity, which includes the effects of functional form as well as the number of parameters, has been a nontrivial problem. Although selection methods such as AIC were important advances in model selection because they are sensitive to one aspect of complexity (the number of parameters in a model), proper use of such techniques is limited to situations in which the effect of functional form on model fit is minimal or can be ignored. Such selection methods are of limited usefulness for comparing models of cognition, most of which differ in functional form.

At the heart of the difficulty of developing a suitable selection method was discovering a meaningful metric that could be used to compare models with radically different functional forms. For example, how does one compare the functional forms of the logarithmic and exponential models in Table 1? The literature has been relatively silent on this issue (but see Cutting et al., 1992; Townsend, 1975). As we show above, differential geometry not only provides a solution, but the solution is intuitive. Complexity is conceptualized as counting explanations (i.e., distinguishable probability distributions that the model can generate) that lie close to the true model that generated the data. The relative complexities of the models, which are defined in Equation 8, can be estimated by comparing these distributions with the total number of distinguishable distributions the models can generate.

There are many attractive features of this complexity measure. To begin with, it is independent of any selection method and can therefore serve as an additional tool with which to evaluate and compare models. Complexity is no longer hidden in the equation of a particular selection method, such as BMS. In addition, geometric complexity equals the complexity term of the MDL criterion (Equation 7). This link between the two measures reveals how model selection works in MDL: It selects the model that maximizes the "normalized" maximum likelihood (Equation 9). The elucidation of MDL by means of differential geometry demonstrates that MDL is a suitable model selection method, and the outcomes of the three application examples support this contention. In the following sections, we discuss some of the limitations of these two tools and the factors that affect their use.

Factors That Affect Geometric Complexity

The number of parameters and functional form are not the only factors that influence geometric complexity. There are at least four others on which the complexity measure depends. Each is described briefly along with a discussion of their implications for model selection.

Extension of the parameter space. As shown in Equation 8, the definition of geometric complexity involves the integration of a non-negative quantity—the square root of the determinant of the Fisher information matrix—over a region of the parameter space on which the model is defined. Model complexity is directly related to the extent of the parameter space, with a larger space resulting in a more complex model. An implication of this is that two models that are identical in all respects, except in the range of their parameters, will have different geometric complexity measures. This makes sense from the standpoint of differential geometry. A model with a wider parameter space will contain more distinguishable probability distributions than a model with a smaller parameter space. The two are different models as far as model selection is concerned. Accordingly, the parameter range must be clearly specified when defining a model.

Sample size. Just as with the extension of the parameter space, sample size is directly related to complexity, as can be seen in Equation 8. Complexity increases with larger sample sizes. The measure's sensitivity to this variable is related to the fact that variance in a probability distribution decreases as sample size increases. With a small sample, two probability distributions on a model's manifold might not be distinguishable because variance is large. They will thus be counted as a single distribution when calculating the geometric complexity of the model. As sample size increases, the variance will decrease and the two distributions will become more discriminable. At some point, the two distributions will be distinguishable and counted separately in the complexity measure. In other words, the number of distinguishable distributions increases as sample size grows, increasing complexity. The first term of the geometric complexity measure in Equation 8 reflects this effect.

Shape of the probability distribution. As described earlier, a model is a parametric family of probability distributions, which are specified by its likelihood function $f(y|\theta)$. Accordingly, a model's geometric complexity will depend upon the particular shape of the probability distribution assumed. To see this, note that the Fisher information matrix, $I(\theta)$, which defines geometric complexity, is determined by the likelihood function $f(y|\theta)$, and that the likelihood function takes different forms depending on the shape of the probability distribution. For example, for two psychophysical models that assume the same power function for the deterministic component $g(\theta, x)$ but different distributions of the error term (e.g., normal vs. uniform), they will in general have different complexity values. It would be interesting to investigate the relative contributions of the two components (structural and distributional) to the overall value of the complexity measure, although the contribution of the latter is predicted to be small.

Experimental design. To the extent that the Fisher information matrix is sensitive to the experimental design represented by the independent variable x in $g(\theta, x)$, the geometric complexity measure can depend upon the specific choices of x values used in an experiment. For example, for a linear model of the form $y_i = \theta x_i + N(0, \sigma^2)$ ($i = 1, \ldots, n$; $0 \leq \theta, \sigma < \infty$), different complexities can be obtained for two designs such as $x = \{1, 2, 3, 4\}$ and $x = \{3, 4, 5, 6\}$ even for the same sample size ($n = 4$) and the same error distribution (see Appendix B for an example).

To summarize, the geometric complexity of a model is determined by a number of factors. Some represent inherent characteristics of the model such as the number of parameters, functional form, and extension of parameter space. Others represent non-model, experimental factors such as the error distribution and the experimental design. As a general rule of thumb, any factor that will affect the number of distinguishable distributions could alter the geometric complexity of the model.

Computing Geometric Complexity

The definition of geometric complexity includes an integral of the determinant of the Fisher information matrix. Therefore, computing geometric complexity involves two stages: derivation of the Fisher information matrix and evaluation of the integral. The Fisher information matrix is obtained by calculating the second-derivatives of the log likelihood function of the model of interest, either by hand derivation or using technical computing software such as Mathematica (Wolfram Research, Inc., Champaign, IL) or Maple (Waterloo Maple, Inc., Ontario, Canada).

Once the Fisher information matrix is obtained, the next step is to integrate its determinant over the extension of the parameter space defined by the model. It is not in general possible to obtain a closed-form solution of the integral, so the solution must be sought numerically instead. This step can be challenging, although recent advancements in Monte Carlo techniques, as well as the availability of statistical packages that implement these techniques, has made the high-dimensional integration problem technically feasible. In Appendix C, we provide a tutorial of Monte Carlo methods. For a more technically rigorous treatment of the topic, consult Gelman, Carlin, Stern, and Rubin (1995), and Gilks, Richardson, and Spiegelhalter (1996).

The use of MDL requires that the determinant of the Fisher information matrix be nonsingular, meaning that its value must remain finite for the full range of the parameters. If the determinant becomes infinite for certain values of the parameters, the range of parameters must be restricted to ensure the integral of the determinant is finite.[13] This is, in effect, equivalent to redefining the model itself, as discussed earlier. For example, for FLMP, the determinant of the Fisher information matrix becomes singular at the two endpoints of each parameter, that is, $\lambda_i, \theta_j = 0$, and 1. To avoid this problem, we restricted the range of the parameters to $0.001 \leq \lambda_i, \theta_j \leq 0.999$ for all is and js. As noted above, different parameter ranges will yield different complexity values.

A challenge in computing geometric complexity arises for algorithmic models, such as random-walk models (e.g., Ratcliff, 1978). The likelihood function that predicts a model's performance for any given stimulus condition is not defined a priori. Rather, a prediction is obtained only by simulating the model for each given stimulus condition. Further, to get a reliable estimate of the probability of a particular prediction, the procedure must be repeated over a large number of trials. In short, obtaining the likelihood function (or the Fisher information matrix), which is a prerequisite for computing complexity, would be computationally unfeasible, if not impossible. An alternative is to try some sort of an algorithm-based estimate of geometric complexity that in essence implements MDL in principle but does not require the derivation of the Fisher information matrix or its integration (e.g., Hochreiter & Schmidhuber, 1997).

Future Work and Other Issues

Testing qualitative models of cognition. Application of MDL and geometric complexity require that each of the models being compared be quantitative models that can be expressed as a parametric family of probability distributions. Because of the large number of qualitative (i.e., verbal) models in cognitive psychology, it would be ideal to extend these two selection tools to this domain as well. In a qualitative model, predictions are made verbally or graphically at the level of ordinal scales without necessarily making use of mathematical equations or the specification of the error structure. For example, models of word recognition state that lexical decision response times will be faster to high-frequency than low-frequency words, but these models make few statements about the magnitude of the time difference between frequency conditions, how frequency is related to response latency (e.g., linearly or logarithmically), or the shape of the response time distribution. The axiomatic theory of decision making (e.g., Fishburn, 1982) is another example of qualitative modeling. The theory is formulated in rigorous mathematical language and makes precise predictions about choice behavior given a set of hypothetical gambles, but it lacks an error theory. Without an appropriate error theory, it is not possible to express the axiomatic theory as a parametric family of probability distributions.

Preliminary work by Grunwald (1999) suggests that extension of the two tools to such cases is possible. Most importantly, there may be no need to develop a statistical form of the model to apply MDL. This line of research may hold great promise.

The scope of the selection methods. The focus of this article has been on how to evaluate a quantitative model's account of behavioral data. The tools presented here for doing so (MDL and the complexity measure) can give the impression that model selection can be highly automated and purely objective. Nothing could be further from the truth. The outcomes of these tests contribute only a few pieces of evidence to the model selection process and should by no means be the sole selection criteria when comparing models. Not only are they silent on crucial issues such as plausibility, explanatory adequacy, and falsifiability (see Bamber & van Santen, 1985), but on other issues pertinent to the particular testing situation, such as the quality and significance of the data being modeled. MDL and geometric complexity should be

[13] For other regularity conditions, see Rissanen (1996, pp. 41–42).

used in conjunction with other criteria in making informed decisions about which model to choose.

Conclusion

The goal of model selection as outlined early in the article is to satisfy two opposing goals. Choose the model that provides a sufficiently good fit to the data in the least complex manner, thus ensuring good generalizability. MDL does just this. Geometric complexity provides deeper insight into the models under consideration by helping us understand why one model is chosen over another. One model might provide a better fit but only at the cost of additional complexity. This excess complexity might not yet be justified given what is known about the regularities underlying the cognitive process. On the other hand, a model's additional complexity might be justified by its vastly superior fit to the data relative to its competitor. Viewed in this light, the purpose of this article was to make the case that it is no longer enough to select a model on the basis of its superior fit to a sample of data. An additional property of the model must be justified as well, namely its complexity. Doing so will help ensure success in selecting among mathematical models of cognition.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox & F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Amari, S. I. (1983). A foundation of information geometry. *Electronics and Communications in Japan, 66A,* 1–10.

Amari, S. I. (1985). *Differential geometrical methods in statistics.* New York: Springer-Verlag.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation, 9,* 349–368.

Bamber, D., & van Santen, J. P. H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29,* 443–473.

Barron, A. R., & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transaction on Information Theory, 37,* 1034–1054.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications.* New York: Wiley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

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 and Methods, 19,* 221–278.

Cutting, J. E., Bruno, N., Brady N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121,* 364–381.

Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research, 63,* 174–182.

Fishburn, P. C. (1982). *The foundations of expected utility.* Dordrecht, The Netherlands: Reidel.

Friedman, D., Massaro, D., Kitzis, S. N., & Cohen, M. M. (1995). A comparison of learning models. *Journal of Mathematical Psychology, 39,* 164–178.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* New York: Chapman & Hall.

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology, 44,* 153–170.

Grunwald, P. (1999, July 29–August 1). *Determining the complexity of arbitrary model classes.* Paper presented at the 32nd annual meeting of the Society for Mathematical Psychology, Santa Cruz, CA.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44,* 133–170.

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation, 9,* 1–42.

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.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Kelso, S. J. A. (1995). *Dynamic patterns: The self-organization of brain and behavior.* Cambridge, MA: MIT Press.

Li, M., & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications.* New York: Springer-Verlag.

Li, S.-C., Lewandowski, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology: General, 125,* 360–369.

Linhart, H., & Zucchini, W. (1986). *Model selection.* New York: Wiley.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA, 97,* 11170–11175.

Myung, I. J., Forster, M., & Browne, M. W. (Eds.). (2000). Model selection [Special issue]. *Journal of Mathematical Psychology, 44.*

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power-law artifact: Insights from response surface analysis. *Memory & Cognition, 28,* 832–840.

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 & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 327–355). Mahwah, NJ: Erlbaum.

Myung, I. J., & Pitt, M. A. (2001). Mathematical modeling. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology. Vol. 4: Methodology* (pp. 429–459). New York: Wiley.

Nosofsky, R. M. (1986). Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172–191.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society, 37,* 81–91.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

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 Transactions on Information Theory, 42,* 40–47.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal, 42,* 260–269.
Roberts, F. S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences.* Reading, MA: Addison-Wesley.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367.
Schervish, M. J. (1995). *The theory of statistics.* New York: Springer-Verlag.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.
Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-patten classification and recognition. *Journal of Experimental Psychology: General, 121,* 278–304.
Stevens, S. S. (1960). The psychophysics of sensory function. *American Scientist, 48,* 226–253.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B, 36,* 111–147.
Townsend, J. T. (1975). The mind–body problem revisited. In C. Cheng (Ed.), *Philosophical aspects of the mind–body problem* (pp. 200–218). Honolulu, HI: Honolulu University Press.
Vitanyi, P., & Li, M. (2000). Minimum description length, Bayesianism and Kolmogorov complexity. *IEEE Transactions on Information Theory, 46,* 446–464.
Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory, 44,* 1424–1439.
Zhang, S. (1999). *Applications of geometric complexity and the minimum description length principle in mathematical modeling of cognition.* Unpublished doctoral dissertation, Department of Psychology, Ohio State University.

Appendix A

Relationship Between Response Surface Analysis (RSA) and Differential Geometry

This appendix derives the equations that show the relation between the RSA length measure L_M in response-surface analysis and the volume measure V_M in the differential geometric approach for a one-parameter model.

Suppose that a random variable v_q ($q = 1, \ldots, N$) is independently and binomially distributed, Bin[n, $g(\theta, x_q)$], with probability $g(\theta, x_q)$ and sample size n where $0 \le g(\theta, x_q) \le 1$. Define a data variable $y_q = v_q/n$, which is the proportion of ones over n binary trials (see Footnote 2 for the distinction between N and n).

The response surface of the one-parameter model for data size N is in fact a curve embedded in the N-dimensional data space and formed by the expected response vector $E(y) = [E(y_1), \ldots, E(y_N)]$, which is equal to $[g(\theta, x_1), \ldots, g(\theta, x_N)]$ for any n. The total length of the model's response curve can be calculated using the standard calculus technique. From Pythagorean theorem, the infinitesimal length ds of the response curve is given by

$$ds = \sqrt{[dE(y_1)]^2 + \cdots + [dE(y_N)]^2}$$
$$= d\theta \sqrt{\sum_{q=1}^{N} \left(\frac{dg(\theta, x_q)}{d\theta}\right)^2} \qquad \left(\because dE(y_q) = \frac{dg(\theta, x_q)}{d\theta} d\theta\right).$$

Then, the desired total length L_M is obtained by integrating the above ds over the parameter space defined by the model.

To derive the differential geometric volume measure, we first need to find the Fisher information matrix $I(\theta)$ of the model. The Fisher information matrix for a multi-parameter model is defined as

$$I_{ij}(\theta) = -E_\theta\left[\frac{\partial^2 \ln f(y|\theta)}{\partial \theta_i \partial \theta_j}\right],$$

where y is the data vector of sample size $n = 1$, and the expectation is taken over the data. Because our model has one parameter, $I(\theta)$ becomes a scalar quantity instead of a matrix. Note that for $n = 1$, the data variable y_q is equal to the binomial random variable v_q, ($q = 1, \ldots, N$) where $v_q \in \{0, 1\}$ is binomially distributed with probability $g(\theta, x_q)$. The log likelihood of the data vector $y = (y_1, \ldots, y_N)$ is then given by

$$\ln f(y|\theta)$$
$$= \sum_{q=1}^{N}\left(\ln \frac{1!}{y_q!(1-y_q!)} + y_q \ln g(\theta, x_q) + (1-y_q)\ln[1 - g(\theta, x_q)]\right).$$

From the log likelihood, the first and second derivatives are obtained as follows:

$$\frac{d \ln f(y|\theta)}{d\theta} = \sum_{q=1}^{N}\left(y_q \frac{\left(\frac{dg(\theta, x_q)}{d\theta}\right)}{g(\theta, x_q)} + (1-y_q)\frac{\left(-\frac{dg(\theta, x_q)}{d\theta}\right)}{[1 - g(\theta, x_q)]}\right) \text{ and}$$

$$\frac{d^2 \ln f(y|\theta)}{d\theta^2}$$
$$= \sum_{q=1}^{N}\left(\frac{\left(\frac{dg(\theta, x_q)}{d\theta}\right)^2}{g(\theta, x_q)^2(1 - g[\theta, x_q)]^2}(y_q^2 - 2y_q g(\theta, x_q) + g(\theta, x_q)^2)\right).$$

By taking the minus expectation of the second derivative and substituting the relations $E[y_q^2] = g(\theta, x_q)[1 - g(\theta, x_q)]$ and $E[y_q] = g(\theta, x_q)$ for the binomial distribution of sample size 1, we get the following form of the Fisher information matrix:

$$I(\theta) = -E\left[\frac{d^2 \ln f(y|\theta)}{d\theta^2}\right] = \sum_{q=1}^{N} \frac{1}{g(\theta, x_q)[1 - g(\theta, x_q)]}\left(\frac{dg(\theta, x_q)}{d\theta}\right)^2.$$

Finally, the geometric volume measure V_M is obtained by integrating the square root of the determinant of the above quantity, $\sqrt{\det[I(\theta)]}$, over the parameter space defined by the model.

Appendix B

Calculation of Minimum Description Length (MDL) and Geometric Complexity for the Two Psychophysics Models

This appendix shows the steps involved in calculating geometric complexity for Fechner's and Stevens's models of psychophysics.

For Fechner's model, the probability distribution of data y_i given x_i and parameters (a, b) is given by

$$f(y_i|a, b) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - a\ln(x_i+b))^2},$$

with its log likelihood

$$L = \ln f(y_i|\theta) = -\frac{1}{2\sigma^2}[y_i - a\ln(x_i + b)]^2 - \ln\sigma - \frac{1}{2}\ln 2\pi.$$

First, we find the first and second derivatives of the log likelihood as follows:

$$\frac{\partial L}{\partial a} = \frac{1}{\sigma^2}[y_i - a\ln(x_i + b)]\ln(x_i + b),$$

$$\frac{\partial^2 L}{\partial a^2} = -\frac{1}{\sigma^2}[\ln(x_i + b)]^2,$$

$$\frac{\partial L}{\partial b} = \frac{2a}{\sigma^2}[y_i - a\ln(x_i + b)]/(x_i + b),$$

$$\frac{\partial^2 L}{\partial b^2} = -\frac{a}{\sigma^2}[y_i - a\ln(x_i + b) + a]/(x_i + b)^2, \text{ and}$$

$$\frac{\partial^2 L}{\partial a\partial b} = \frac{1}{\sigma^2}[y_i - 2a\ln(x_i + b)]/(x_i + b).$$

Next, expectations of minus log likelihoods over y_i are sought using $E(y_i) = a\ln(x_i + b)$ as

$$-E\left(\frac{\partial^2 L}{\partial a^2}\right) = \frac{1}{\sigma^2}[\ln(x_i + b)]^2,$$

$$-E\left(\frac{\partial^2 L}{\partial b^2}\right) = \frac{a^2}{\sigma^2}\frac{1}{(x_i + b)^2}, \text{ and}$$

$$-E\left(\frac{\partial^2 L}{\partial a\partial b}\right) = \frac{a}{\sigma^2}\frac{\ln(x_i + b)}{(x_i + b)}.$$

Therefore, we obtain the Fisher information matrix given x_i as

$$I_1(a, b; x_i) = \frac{1}{\sigma^2}\begin{bmatrix} (\ln(x_i + \mathrm{b}))^2 & a\dfrac{\ln(x_i + b)}{(x_i + b)} \\ a\dfrac{\ln(x_i + b)}{(x_i + b)} & \dfrac{a^2}{(x_i + b)^2} \end{bmatrix},$$

where the subscript 1 stands for sample size 1. The desired Fisher information matrix I in Equation 7 for the entire data $y = (y_1, \ldots, y_n)$, is then obtained as

$$I(\theta = (a, b)) = \left(I_{ij} = -\frac{1}{n}E\left[\frac{\partial^2 \ln f(y = (y_1, \ldots, y_n)|\theta)}{\partial\theta_i\theta_j}\right], i, j = 1, 2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} I_1(a, b; x_i)$$

$$= \frac{1}{n\sigma^2}\begin{bmatrix} \sum_{i=1}^{n}(\ln(x_i + b))^2 & \sum_{i=1}^{n} a\dfrac{\ln(x_i + b)}{(x_i + b)} \\ \sum_{i=1}^{n} a\dfrac{\ln(x_i + b)}{(x_i + b)} & \sum_{i=1}^{n}\dfrac{a^2}{(x_i + b)^2} \end{bmatrix},$$

assuming independent observations, that is, $\ln f(y|\theta) = \sum_{i=1}^{n} \ln f(y_i|\theta)$.

From the above result, the geometric complexity (GC) of Fechner's model is obtained as

$$GC_F = \frac{k}{2}\ln\frac{n}{2\pi} + \ln\int d\theta\sqrt{\det(I(\theta))}$$

$$= \ln\left(\int\int da\, db\frac{a}{\sigma^2}H(b)\right) - \ln n + \frac{k}{2}\ln\frac{n}{2\pi}$$

$$= \ln\int a\, da + \ln\int H(b)\, db - 2\ln\sigma - \ln 2\pi$$

for $k = 2$, with

$$H(b) = \sqrt{\left(\sum_{i=1}^{n}(\ln(x_i + b))^2\right)\left(\sum_{i=1}^{n}\frac{1}{(x_i + b)^2}\right) - \left(\sum_{i=1}^{n}\frac{\ln(x_i + b)}{(x_i + b)}\right)^2},$$

where $x = \{1, 2, \ldots, 6\}$ is assumed in the present simulation.

Similarly, for Stevens's model, the Fisher information matrix for the data $y = (y_1, \ldots, y_n)$ can be obtained as follows:

$$I(a, b) = \frac{1}{n}\sum_{i=1}^{n} I_1(a, b; x_i) = \frac{1}{n\sigma^2}\begin{bmatrix} \sum_{i=1}^{n} x_i^{2b} & a\sum_{i=1}^{n} x_i^{2b}\ln x_i \\ a\sum_{i=1}^{n} x_i^{2b}\ln x_i & a^2\sum_{i=1}^{n} x_i^{2b}(\ln x_i)^2 \end{bmatrix}$$

From this, the geometric complexity of Stevens's model is given by

$$GC_S = \ln\left(\int\int da\, db\frac{a}{\sigma^2}W(b)\right) - \ln n + \frac{k}{2}\ln\frac{n}{2\pi}$$

$$= \ln\int a\, da + \ln\int W(b)db - 2\ln\sigma - \ln 2\pi$$

for $k = 2$, with

(*Appendix continues*)

$$W(b) = \sqrt{\left(\sum_{i=1}^{n} x_i^{2b}\right)\left(\sum_{i=1}^{n} x_i^{2b}(\ln x_i)^2\right) - \left(\sum_{i=1}^{n} x_i^{2b} \ln x_i\right)^2},$$

where $x = \{1, 2, \ldots 6\}$.

We note that in choosing between two models using MDL, only the difference in geometric complexity between the two models matters, not their absolute values. The difference for the two psychophysics models is as follows:

$$\Delta GC = GC_S - GC_F$$

$$= \ln \int W(b)db - \ln \int H(b)db.$$

Therefore, we only need to evaluate two one-dimensional integrals.

For Stevens's model, the integral $\ln \int W(b)db$ over the entire range of the exponent parameter b (i.e., $0 < b < \infty$) turns out to be infinity. Consequently, the range of the parameter must be restricted to make the integral finite. The range of $0 < b < 3$ was chosen based on typical values of the parameter observed in experiments (e.g., Stevens, 1960). The integral was then evaluated numerically by the simple Monte Carlo method (see Appendix C) in which 100,000 random samples were drawn from the uniform distribution on [0, 3]. The result was 8.00. For Fechner's model, the integral $\ln \int H(b)db$ is finite over the entire range of the parameter ($0 < b < \infty$). The integral was evaluated numerically by the simple Monte Carlo method, and its value runs were 2.48. Combining these two results, the difference in geometric complexity between Stevens's and Fechner's models was calculated as $\Delta GC = GC_S - GC_F = 8.00 - 2.48 = 5.52$.

From this point, calculation of MDL is straightforward. Plug the complexity value into Equation 7 and combine it with the calculation of the log likelihood measure; the common constant term $C_0 = \ln \int ada - 2\ln\sigma - \ln 2\pi$ can be ignored:

$$\mathrm{MDL}_S = -\ln f(y|\hat{\theta}_S) + 8.00$$

$$\mathrm{MDL}_F = -\ln f(y|\hat{\theta}_F) + 2.48.$$

Appendix C

Numerical Integration Methods

The definition of the geometric complexity measure includes an integral of the determinant of the Fisher information matrix. As it is, in general, not possible to obtain an analytic solution of the integral, integration by sampling (i.e., Monte Carlo methods) is the only option. This appendix provides a tutorial of Monte Carlo integration methods.

Simple Monte Carlo Method

Let $f(x)$ denote the function to be integrated (e.g., the square root of the determinant of the Fisher information matrix in calculating geometric complexity). Suppose we wish to evaluate the following one-dimensional integral:

$$I = \int_a^b f(x)dx,$$

where $a < b$. The value of this integral is equal to $(b - a) \cdot \bar{f}$, where $\bar{f}$ is the mean value of $f(x)$ over the interval $[a, b]$. The mean can be estimated based on a set of n samples $\{x_1, x_2, \ldots, x_n\}$ randomly selected from the uniform density over $[a, b]$. From the estimated mean, the desired integral can then be approximated as

$$I_n = (b - a) \cdot \bar{f}_n,$$

where $\bar{f}_n = 1/n \sum_{i=1}^{n} f(x_i)$. This method is called the *simple Monte Carlo method*. Generalization of the method to multidimensional integration problems is straightforward.

The rationale for the simple Monte Carlo method is the strong law of large number in probability theory, and therefore, the accuracy of the simple Monte Carlo method is improved as sample size increases. Specifically, the standard error of I_n is given by

$$S_n = \sigma/\sqrt{n},$$

where σ is the population standard deviation. Although in theory s_n approaches zero as n goes to infinity, in practice the quantity may not be quite close to zero even for an extremely large but finite n (e.g., million) because σ may be quite large, depending on the form of $f(x)$. Other techniques must then be developed to increase the accuracy of the numerical approximation.

Importance Sampling Monte Carlo Method

Importance sampling improves on the simple Monte Carlo method by directly controlling σ. Let us rewrite the original integration problem as

$$I = \int_a^b f(x)dx = \int_a^b \frac{f(x)}{g(x)} g(x)dx$$

in terms of some probability density $g(x)$, from which we know how to sample. Based on a set of n independent samples $\{x_1, x_2, \ldots, x_n\}$ drawn from $g(x)$, the desired integral I can be estimated as

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)}. \quad \text{(C1)}$$

For this estimator, σ can be shown to be effectively close to zero if $g(x)$ is chosen such that the importance ratio, $f(x_i)/g(x_i)$, is roughly close to a constant for all x_is. Under this condition, a fairly precise estimate of the integral I can be obtained. On the other hand, importance sampling may be poor if the importance ratio varies greatly across x values. When a uniform density is chosen for $g(x)$, the importance Monte Carlo reduces to the simple Monte Carlo.

Markov Chain Monte Carlo

The simple and importance sampling Monte Carlo methods prescribe that samples $\{x_1, x_2, \ldots, x_n\}$ be drawn independently. This may not always be feasible, for $g(x)$ can be quite nonstandard. In such cases, a dependent sample could be used to estimate the integral by applying Markov chain Monte Carlo methods.

The basic idea of this method is to generate samples $\{x_1, x_2, \ldots, x_n\}$, possibly dependent, from a Markov chain process whose stationary distribution is $g(x)$. Because the samples are not necessarily independent, the strong law of large numbers is not applicable. Instead, the asymptotic consistency of the series in Equation C1 is guaranteed by the ergodic property of Markov chains, which ensures $I_n \rightarrow I$ as n goes to infinity.

Among many algorithms that are currently in use to generate samples from a Markov chain, Gibbs sampling is the most popular method because of its simplicity. This algorithm, as a special case of the Metropolis-Hastings algorithm, proceeds as follows:

Step 1: Let $t = 0$. Draw a starting point $x(t) = [x_1(t), \ldots, x_m(t)]$ from any starting distribution, say $g(x)$, whose full conditional distributions are known.

Step 2: For $i = 1$ to n, sample $x_i(t + 1)$ from the one-dimensional conditional distribution $g[x_i(t)|x_1(t + 1), \ldots, x_{i-1}(t + 1), x_{i+1}(t), \ldots, x_m(t)]$.

Step 3: Let $t = t + 1$. Go to Step 2.

For further detail on this and other Monte Carlo methods, see Gilks et al. (1996) and Gelman et al. (1995), on which the present tutorial is based.

Received January 5, 2000
Revision received June 13, 2001
Accepted June 13, 2001 ■