Model selection in cognitive science as an inverse problem

Jay I. Myung[a], Mark A. Pitt[a], and Daniel J. Navarro[b]

[a]Dept. of Psychology, Ohio State Univ., 1885 Neil Avenue, Columbus, OH USA 43210-1222;
[b]Dept. of Psychology, University of Adelaide, Adelaide, Australia, SA 5005

ABSTRACT

How should we decide among competing explanations (models) of a cognitive phenomenon? This problem of model selection is at the heart of the scientific enterprise. Ideally, we would like to identify the model that actually generated the data at hand. However, this is an un-achievable goal as it is fundamentally ill-posed. Information in a finite data sample is seldom sufficient to point to a single model. Multiple models may provide equally good descriptions of the data, a problem that is exacerbated by the presence of random error in the data. In fact, model selection bears a striking similarity to perception, in that both require solving an inverse problem. Just as perceptual ambiguity can be addressed only by introducing external constraints on the interpretation of visual images, the ill-posedness of the model selection problem requires us to introduce external constraints on the choice of the most appropriate model. Model selection methods differ in how these external constraints are conceptualized and formalized. In this review we discuss the development of the various approaches, the differences between them, and why the methods perform as they do. An application example of selection methods in cognitive modeling is also discussed.

Keywords: model selection, cognitive modeling, inverse problems, model complexity, minimum description length, stochastic complexity, normalized maximum likelihood

1. INTRODUCTION

Cognitive science is concerned with the architecture and function of the mind. Researchers from a range of disciplines (e.g., computer science, engineering, linguistics, philosophy, psychology) work to delineate the mental processes and biological mechanisms that enable humans to function in the world, where "functioning" can entail navigating across a room, carrying on a conversation, or studying for an exam. Scientific progress proceeds through the collection of data using traditional behavioral methods as well as more modern brain scanning technologies.

Greater clarity about the design of cognitive processes has often been sought through quantitative modeling. The precision of cognitive models makes it possible to test theoretical assumptions by measuring the ability of the corresponding models to account for (i.e., fit) experimental data. The variety of modeling approaches (e.g., neural networks, simulation-based models) has enabled researchers to build models in a wide range of fields. As a result, modeling has made significant contributions to the field, and its popularity and impact continues to grow. This progress has brought with it a need for methods that can assist in choosing among competing models. This is the problem of *model selection*, in which the goal is to develop a set of well-justified criteria that can be used to discriminate between competing models.

When choosing between models, we would ideally like to identify the one that actually generated the data. Realistically, this is an un-achievable goal because information in a finite data sample is seldom sufficient to point to a single model. Even when the number of data points exceeds the number of dimensions of the data generating function (i.e., true model), there is often so much statistical redundancy in the data that it provides little information about the underlying structure. Moreover, the problem is made harder by the presence of random error in the data. Finally, it is improbable that the true model is included in the particular set of models

Send correspondence to Jay I. Myung (e-mail: myung.1@osu.edu; tel.: +1 614 292 1862). Contact information for other authors: Mark A. Pitt (e-mail: pitt.2@osu.edu, tel.: +1 614 292 4193); Daniel J. Navarro (e-mail: daniel.navarro@adelaide.edu.au, tel.: +61 8 8303 5265).

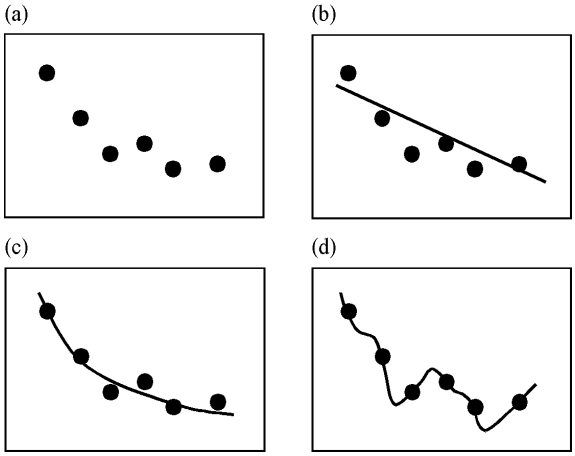

Figure 1. Ill-posedness of model selection.

under comparison. The truth is likely to be much more complicated than one can imagine.[1] Taken together, these difficulties ensure that the model selection problem is invariably ill-posed.

One solution to the ill-posedness of model selection is to reformulate the problem as one of *inferring* the model that is "closest" to the truth in some well-defined sense. This leads to the notion of generalizability in statistics, which formally conceptualizes the notion of closeness or approximation to truth. Generalizability, or predictive accuracy, refers to how accurately a model predicts the statistics of future (as yet unseen) samples from the same process that generated the observed data sample. Central to the concept of generalizability is the idea is that a model should not be evaluated solely on how well it fits the *observed* data, but on how well it fits *new* data generated from the same process.

Figure 1 illustrates the ill-posedness of model selection. The solid dots in panel (a) represent observed data points, and each of three panels (b) - (d) represents a model's best fit to the same data set. The two-parameter linear model in panel (b) clearly underfits the data and represents an over-simplified description of the data. The three-parameter exponential model in panel (c) does a fairly good job in capturing the general trend of the data and seems likely to generalize well. The model in panel (d), with many more parameters, fits the data better than the previous two models but the improvement seems mostly due to its ability to capture the idiosyncracies that are unique to the current data sample but are unlikely to repeat in new data samples. In other words, this overly sophisticated model would generalize poorly. Accordingly, the generalizability criterion favors the model in panel (c) as the "best" description of the data among the three models considered.

Interestingly enough, the model selection problem bears a striking similarity to the visual perception problem in cognitive psychology. To see this, consider the well-known problem of occlusion in visual object recognition. The presence of depth in the visual world permits one object to cover up a portion of another object in a visual scene. This situation creates ambiguities in object shape because inferences must be made about shape of that portion of the object that is occluded. The four panels in Figure 2 illustrate the relationship between the occlusion problem and the model selection problem. Panel (a) depicts a blob that corresponds to some visual input, and panels (b) - (d) display three possible interpretations or parsings of the blob. Just as the models in Figure 1 fit the data differently depending on their parameters and functional form, the different parsings provide different "fits" to the blob shown in panel (a) of Figure 2. The parsing in panel (d) accounts for the blob "perfectly", simply by copying it exactly. This is precisely what the model in panel (d) of Figure 1 does. In contrast, the parsings depicted in panels (b) and (c) of Figure 2 are analogous to the models shown in panels (b) and (c) in Figure 1: they make explicit assumptions that about the regularities that might underlie the data. As with model selection, the problem is to determine which account is the most accurate representation of the process that generated the data.

This example is meant to illustrate the parallels between model selection and perception. Fundamentally

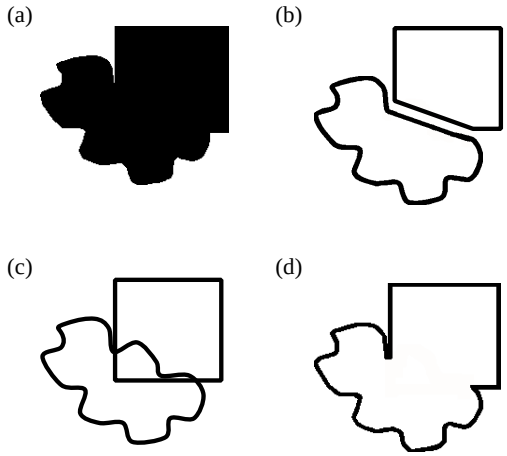

Figure 2. Ill-posedness of visual object recognition.

it is one of ill-posedness: the data are not sufficiently informative to derive an accurate, unique, and robust solution.[2] This paper provides a brief introduction to the field of model selection. We begin with an overview of the problem itself, followed by a discussion of the strengths and weaknesses of various solutions. We finish by providing an illustrative example of the application of model selection methods.

2. MODEL SELECTION AND REGULARIZATION

2.1. Generalizability and Overfitting

For a data set of n observations $\boldsymbol{x} = (x_1, ..., x_n)$, a quantitative model with k parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ is defined as a parametric family of probability distributions $f(\boldsymbol{x}|\boldsymbol{\theta})$. The generalizability of model M is then defined as

$$E[D(M,T)] = \int D(f_M(\boldsymbol{x}|\hat{\boldsymbol{\theta}}), f_T(\boldsymbol{x}))f_T(\boldsymbol{x})d\boldsymbol{x} \tag{1}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood (ML) estimate, f_T denotes the probability distribution that corresponds to the true model, and $D(g,h)$ denotes a discrepancy function between two distributions, g and h, satisfying $D(g,h) \geq D(g,g) = 0$ for $g \neq h$ (e.g., Kullback-Leibler information distance). According to (1), generalizability represents a mean discrepancy between the true model and the best-fitting member of the model class, averaged across all data sets that could be observed under the true model. This is how generalizability formalizes the degree or closeness of a model to the truth. The goal of model selection is then to select the model that minimizes $E[D(M,T)]$ and thereby maximizes the generalizability of the model.

The problem associated with generalizability is that it is not directly observable, and must instead be estimated from observed data. However, accurately estimating generalizability of a model is a non-trivial problem. The model that provides the best fit to the observed data set will not necessarily be the one that generalizes best. This is because a model's fit to a particular set of observed data tends to be overly optimistic, as it represents the contribution from capturing the regularity but also the contribution from fitting idiosyncratic patterns or random noise, thereby *overfitting* the data. Unless this effect is compensated for, model selection based solely on overall fit will result in the choice of a model that generalizes poorly.

This ability to fit random noise is closely correlated with the "complexity" of the model. Intuitively, model complexity or flexibility relates to a model's intrinsic ability to describe diverse patterns of data. The wider the range of data patterns that a model can account for, the more complex the model is. There are at least three dimensions of model complexity: (a) the number of a free parameters in the model; (b) the functional form of the model equation; and (c) the parameter range.[3,4] The number of parameters is self-explanatory and well-understood, but the second and third dimensions are less well understood. To illustrate the functional form

dimension, consider two models, a power model $y = ax^b$ and a hyperbolic model $y = a/(1+bx)$. One model may be able to describe more data patterns than the other, even though both have the same number of parameters. In short, they may not be equally complex. Parameter range can also affect model complexity. For example, for the power model, suppose that we extend the range of the exponent parameter from $0 < b < 1$ to say $0 < b < 10$. The model will then be able to describe not only concave curves ($b < 1$) but also convex curves ($b > 1$), and as such, the model will now become more complex.

A complex model with many parameters and highly nonlinear model equation generally absorbs random noise more easily than a simple model with few parameters, thereby improving the fit independently of the model's ability to approximate the underlying regularity. However, we would like to choose a model with good generalizability, not just good fit. For this purpose, the best model is one that is sufficiently complex to capture the regularity in the data, but not so complex that it absorbs the random noise. To achieve this, model fit must be traded off against model complexity: an overly complex model should be penalized to the extent that the extra complexity merely helps the model fit the random error. This trade-off is called complexity regularization.

2.2. Model Selection as Complexity Regularization

Tikhonov's *regularization theory* is a method for solving ill-posed inverse problems that, when applied to model selection, can be used to avoid overfitting.[5,6] The basic idea of regularization is to transform an ill-posed problem into a well-posed one by "augmenting" the data with extraneous, non-data information. This "prior" information is imposed upon the solution of the inverse problem in such a way that a unique solution is guaranteed.

In the context of model selection, for a given model M and a data set $\boldsymbol{x}$, the following equation expresses Tikhonov's regularization theory:

$$R(M, \boldsymbol{x}) = G_M(\boldsymbol{x}) + \lambda H(M) \tag{2}$$

In this equation, $R(M, \boldsymbol{x})$ is called the *Tikhonov functional* to be minimized. The *performance term* $G_M(\boldsymbol{x})$ measures how well or badly the model fits the observed data set, and as such, the value of this term depends upon the data as well as the model. The *complexity-penalty term* $H(M)$ is data-independent but model-dependent, and represents the prior information about the form of the solution (true model). As the name implies, this term measures the inherent complexity of the model. Finally, the *regularization parameter* λ (> 0) represents the relative weight given to the complexity penalty term to the performance term. The regularization parameter is assigned a value that "optimally" trades off between these two opposing factors.

Returning to the model selection problem, this regularization procedure can be used to avoid overfitting and overcome data sparseness, thereby ensuring good generalizability. Indeed, generalizability and regularization are closely related to each other. Asymptotic expansions of $E[D(M,T)]$ in (1) under a given discrepancy function often yield the form in (2) with corresponding $G_M(\boldsymbol{x})$ and $H(M)$ functions, even the value for the λ parameter.[7] Consequently, one could view the Tikhonov functional as a type of generalizability measure, at least in the asymptotic sense. Accordingly, it is not surprising that many model selection methods take the form of a Tikhonov functional.* For example, the commonly-used *Akaike Information Criterion* (AIC[9]) and *Bayesian Information Criterion* (BIC[10]) are both Tikhonov functionals, since

$$\begin{aligned} \text{AIC} &= -2\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + 2k \\ \text{BIC} &= -2\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + k\ln n \end{aligned} \tag{3}$$

where ln denotes the natural logarithm. For each of these criteria, the first term represents a lack of fit measure that corresponds to the performance term $G_M(\boldsymbol{x})$ in (2). The second term is a complexity measure, and corresponds to both the complexity-penalty term $H(M)$ and the regularization parameter λ in (2). The two criteria have the same performance term, but differ in how model complexity is measured. Unlike AIC, BIC takes into account sample size n as well as the number of parameters k. Each criterion represents a "lack of generalizability" measure so that the lower the criterion value, the better the generalizability. Accordingly, the selection methods prescribe that the model minimizing the given criterion should be preferred.

*O'Sullivan[8] gives an excellent review of a B-spline approach for solving inverse problems in model selection. This approach is also based on regularization theory but will not be discussed in the present paper due to lack of space.

In AIC and BIC, model complexity is defined primarily as the number of parameters, but they both fail to consider other potentially important dimensions of complexity, such as functional form and parameter range, as discussed earlier. The minimum description length approach,[1,11] discussed below, yields a selection criterion that also takes the form of a Tikhonov functional, and importantly, is sensitive to these dimensions of complexity.

3. THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

The minimum description length (MDL) principle originates in algorithmic coding theory[12] in computer science, and represents a new conceptualization of the model selection problem. According to algorithmic coding theory, both data and models are viewed as codes that can be compressed. Any redundancy and regularity in data can be used to compress the data, thereby shortening its description length without loss of information. The more the model permits data compression, the more the model enables us to discover the regularities underlying the data, and therefore, the better the model generalizes. The goal of model selection from this viewpoint is to identify the model, among a set of candidate models, that permits the shortest description length of data.

3.1. Stochastic Complexity and the Normalized Maximum Likelihood

The key to the MDL principle lies in the Kraft inequality, which demonstrates that for any computable probability distribution $p(\boldsymbol{x})$ there exists a corresponding *coding* scheme that encodes the data set $\boldsymbol{x}$ as a sequence of length $-\ln p(\boldsymbol{x})$.[12] Moreover, for data sequences generated from this distribution, this coding scheme is optimal in the sense that it minimizes the expected length of an encoded data set. Using these optimal *Shannon-Fano* codes, we can say that the shortest attainable codelength for the data $\boldsymbol{x}$ is $-\ln p(\boldsymbol{x})$ when encoded with the assistance of the probability distribution $p(\boldsymbol{x})$. If our models consisted only of a single probability distribution, then these basic principles of information theory would have provided a complete solution to the model selection problem. Unfortunately, since models almost always consist of a family of probability distributions, the problem is substantially more difficult.

The MDL approach to model selection introduces the idea of the *stochastic complexity* (SC) code,[13,14] in which we seek to compress the data set $\boldsymbol{x}$ with the assistance of a family of probability distributions $p(\boldsymbol{x}|\boldsymbol{\theta})$. However, it is not immediately apparent how to associate a family of distributions with a single, fixed code. Using the Kraft inequality, an equivalent formulation of the problem is to find the single *universal probability distribution* that "best" approximates the behavior of the entire family of distributions defined by the model. Recently, Rissanen[14] formulated this in terms of the following minimax problem:

$$p^* = \arg\inf_p \sup_q E_q \left[\ln \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}{p(\boldsymbol{x})} \right] \tag{4}$$

where p and q are probability distributions satisfying certain (very weak) regularity conditions and E_q is the expectation under the distribution q. Neither p or q is required to be a member of the model family, nor is the solution, p^*. Under this minimax view, the goal is to find the distribution p^* that minimizes the expected codelength for data generated by the "worst possible" source, when those data are encoded with the help of p^*. The distribution that satisfies this minimax problem[15] is the *normalized maximum likelihood* (NML) distribution,

$$p^*(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{\int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\, d\boldsymbol{y}} \tag{5}$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{y}}$ denotes the ML estimate for the data set $\boldsymbol{y}$. Therefore, the probability that this optimizing distribution p^* assigns to the data set $\boldsymbol{x}$ is proportional to the maximized likelihood $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$, and the normalizing constant is the sum of maximum likelihoods of all potential data sets that could be observed in an experiment. It is for this reason that p^* is called the normalized maximum likelihood distribution. The corresponding coding scheme is called the stochastic complexity code, and the codelength for the data set $\boldsymbol{x}$ is referred to as the stochastic complexity of $\boldsymbol{x}$ with respect to the model class $p(\boldsymbol{x}|\boldsymbol{\theta})$. Thus, the stochastic complexity is given by the negative logarithm of the NML,

$$\mathrm{SC} = -\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + \ln \int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\, d\boldsymbol{y}. \tag{6}$$

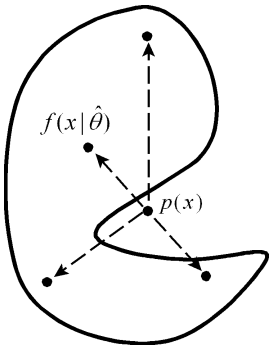

Figure 3. NML as a solution to the minimax problem in the space of probability distributions.

Once again we have a Tikhonov functional. In this equation the first term is the same lack of fit measure as in (3), and the second term is a complexity measure, corresponding to the $\lambda H(M)$ in the regularization equation in (2). Thus in SC, model complexity is operationalized as the logarithm of the *sum of all best fits* the model can provide collectively, over the whole range of data patterns. A model that fits almost every data pattern very well would be much more complex than a model that provides a relatively good fit to a small set of data patterns but does poorly otherwise. This is how the complexity measure captures the intrinsic flexibility of a model that enables it to fit a wide range of data patterns.[3]

A second-order Taylor series expansion of the complexity term in (6) for large sample sizes yields[13, 16]

$$\ln \int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\, d\boldsymbol{y} \;\approx\; \frac{k}{2} \ln \left(\frac{n}{2\pi}\right) + \ln \int_{\Theta} \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \tag{7}$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix[17] of sample size 1 defined as $I(\boldsymbol{\theta})_{i,j} = -E_{f(\cdot|\boldsymbol{\theta})}\left[\frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right], i, j = 1, ..., k$. This asymptotic approximation reveals the three dimensions of model complexity: the number of parameters k, the functional form of the model equation as implied by $I(\boldsymbol{\theta})$, and the parameter range given by the domain of the integral, Θ. The first term in (7) captures the number of parameters whereas the second complexity term captures the functional form and the parameter range. Three observations summarize their contributions to model complexity. First, note that the sample size n appears in the first term but not in the second. This implies that as the sample size becomes large, the relative contribution of the second term to that of the first becomes negligible, essentially reducing the complexity measure to that of the BIC. Second, because the first term is a logarithmic function of sample size but a linear function of the number of parameters, the impact of sample size on model complexity is less dramatic than that of the number of parameters. Third, the second term depends on the parameter ranges implied by Θ. Since $\sqrt{\det I(\boldsymbol{\theta})}$ is a positive scalar, the larger the parameter range, the greater the complexity of the model.

3.2. A Geometric Interpretation of the SC Criterion

Besides the coding-theoretic view, the SC can also be interpreted geometrically. This approach not only sheds new light on the model selection problem but also offers an intuitive understanding of its central concept, complexity. From a geometric perspective, a parametric model family of probability distributions forms a Riemannian manifold embedded in the space of all distributions. Every point on this manifold is a distribution, and the collection of points created by varying the parameters of the model gives rise to a hypersurface in which "similar" distributions are mapped to "nearby" points.[18]

To construct a geometric interpretation of SC, recall that this criterion is defined as the minus logarithm of the NML distribution $p^*(\boldsymbol{x})$ in (5), which represents a solution to the minimax problem in (4). Note that $E_q\left[\ln \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}{p(\boldsymbol{x})}\right]$ in the minimax problem is a discrepancy measure between the two distributions f and p in the sense required by (1). We can interpret this discrepancy geometrically, in terms of the "distance" from f to p

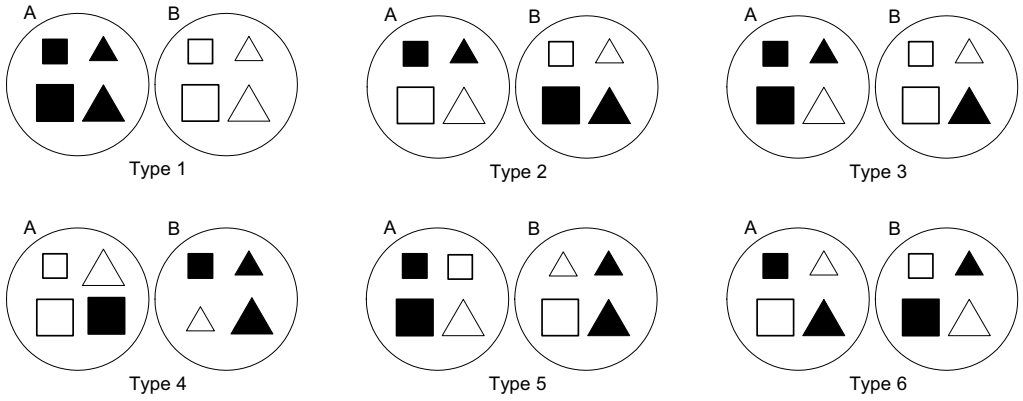

Figure 4. The six category structures that comprise the Shepard, Hovland and Jenkins task.

with respect to q. This geometric perspective allows us to view the minimax problem as one of identifying a distribution that minimizes the maximum distance between that distribution and the best-fitting member of the model family. This geometric interpretation of the minimax problem is schematically illustrated in Figure 3. Under this view, the complexity term of SC is equivalent to the upper-bound of the expected "distance" between the NML distribution $p^*(\boldsymbol{x})$ and the maximum likelihood distribution $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$.[14] Therefore, the complexity term is correlated with the "size" of the model manifold. The larger the manifold, the larger the minimax distance, and thus the more complex the model. In fact, the asymptotic expansion of the complexity term in (7) has an interpretation in terms of the number of "distinguishable" probability distributions within the model manifold that lie "close" to the truth.[19]

In short, what matters in measuring a model's complexity is the "size" of a model manifold in the space of probability distributions, not the apparent constellation of the number of its parameters, the functional form of the model equation, and the parameter range.[19] When examined individually, these features of a model can lead to an incomplete, even misleading understanding of complexity. They are just the mathematical devices that are used to index a collection of distributions defined by the model. 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 manifold.

4. AN ILLUSTRATIVE APPLICATION: PARTITIONING LEARNING CURVES BY MINIMUM DESCRIPTION LENGTH

4.1. Category Learning Curves

To provide an application example of SC in cognitive modeling, we consider the seminal experiment in human category learning conducted by Shepard, Hovland and Jenkins (1961).[20] In this study, human performance was examined on a category learning task involving eight stimuli divided evenly between two categories. The stimuli were generated by varying exhaustively three binary dimensions such as (black, white), (small, large) and (square, triangle). Shepard et al. observed that, if these dimensions are regarded as interchangeable, there are only six possible category structures across the stimulus set. This means, for example, that the category structure that divided all black stimuli into one category, and all white stimuli into the other would be regarded as equivalent to the category structure that divided squares from triangles. These category structures are shown in Figure 4.

Empirically, Shepard et al. found robust differences in the way in which each of the six fundamental category structures was learned. In particular, by measuring the mean number of errors made by subjects in learning each type, they found that Type 1 was learned more easily than Type 2, which in turn was learned more easily than Types 3, 4 and 5 (which all had similar error rates), and that Type 6 was the most difficult to learn. This result was recently replicated in Nosofsky, Gluck, Palmeri, McKinley and Glauthier's (1994).[21] Figure 5 shows

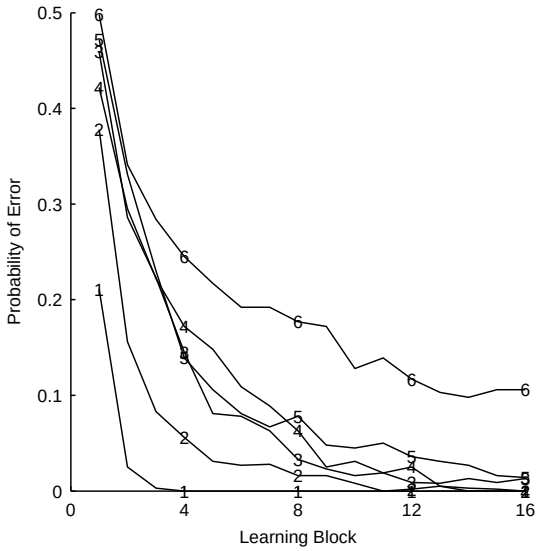

Figure 5. Empirical learning curves for the Shepard, Hovland and Jenkins task (from Nosofsky et al., 1994).

the category learning curves from this experiment.† The consensus in the literature is that the ordinal constraint $1 < 2 < (3, 4, 5) < 6$ represents an important and robust property of human category learning. As a result, the ability to reproduce this ordinal constraint is required in order for a model to be taken seriously by researchers.

4.2. Partitioning Learning Curves by SC

In order to claim that a category learning model such as ALCOVE[22] reproduces this ordinal constraint, we need to be able to find a set of equivalence relations among learning curves (whether these be empirical or predicted curves). This is essentially a partitioning problem. Traditionally, the extraction of the partition from data has been done subjectively, by the visual inspection of the curves in Figure 5. However, this is a somewhat unappealing way to justify the partition, particularly given its importance to category learning. It would be preferable to extract the partition using principled statistical methods. This becomes especially important for data sets that do not lend themselves to simple visual displays.

To address this, we applied a clustering procedure in which the optimal partition is the one that maximizes the SC criterion. In order to do so, we treat a clustering solution as a statistical model for the data, in this case a multivariate binomial model. The problem of choosing the partition now reduces to the problem of choosing among a set of statistical models, a problem for which SC is known to be an appropriate solution. The technical details for calculating the SC values for such models are described elsewhere.[23, 24]

Table 1 shows the results of the SC calculations for a few selected clustering solutions. For each solution, the lack of fit measure and the complexity measure of SC are shown in the second- and the third-columns, respectively, and the overall SC value is shown in the last column. Note that as we increase the number of clusters, the value of the lack of fit goes down (i.e., better fit) while the corresponding value of the complexity term goes up, nicely illustrating the trade-off between these two opposing forces. The SC results in Table 1 agree with the intuition that the correct clustering should be (1)(2)(3,4,5)(6), with the five-cluster solution (1)(2)(3,5)(4)(6) as the closest competitor. Inspection of Figure 5 agrees with this, since the curve for Type 4 is a little different from those for Types 3 and 5, but the discrepancy is not of the same order as those corresponding to Types 1, 2 and 6. In short, the SC-based clustering procedure is "correctly" partitioning this data set.

†Nosofsky et al.'s (1994) data have the following properties: each data point is a pooled set of $n = 40 \times 16 = 640$ binary observations, assumed to be the outcome of independent Bernoulli trials. Each of the six curves consists of 16 data points, corresponding to 16 different measurement intervals.

Table 1. Six clustering solutions to the Shepard et al. (1961) problem.

Partition	Lack of Fit $\left(-\ln f(\boldsymbol{x}\|\hat{\boldsymbol{\theta}})\right)$	Complexity $\left(\ln \int f(\boldsymbol{y}\|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))d\boldsymbol{y}\right)$	SC
(1, 2, 3, 4, 5, 6)	16,337	70	16,408
(1, 2, 3, 4, 5)(6)	15,399	126	15,525
(1, 2)(3, 4, 5)(6)	14,772	185	14,957
(1)(2)(3, 4, 5)(6)	14,597	237	14,834
(1)(2)(3, 5)(4)(6)	14,553	291	14,844
(1)(2)(3)(4)(5)(6)	14,518	343	14,861

5. CONCLUSION

Perhaps it is unsurprising that both perception and model selection can be viewed as inverse problems. In both cases, we are presented with limited information that is consistent with an infinite number of explanations, but are required to infer a single "best" account. Moreover, in both cases the goal of the inference is the same: to appropriately represent the world in spite of our limitations, and make "safe" inferences about future events. There may indeed be "more things in heaven and Earth ... than are dreamt of in [our] philosophy", as Hamlet would have it, but this does not alleviate the fundamental need to understand the environment and behave appropriately within it. From a model selection standpoint, the MDL perspective has the appeal that it avoids making the assumption that the truth ever lies within the set of models that we might consider. In fact, it does not rely on the notion that there even exists any "true" distribution that generates the data. Instead, it relies solely on the efficient coding of observations. By capturing the regular structure in the data that are observed, we seek to generalize better to future data without ever invoking the notion of "the truth". When this notion is formalized through the SC criterion, we arrive at a "complexity regularization" view of model selection, in which extraneous information about the models is used to constrain our inferences in a manner that is strikingly reminiscent of regularization principles in visual perception. Ultimately, this parallel may be no more (or less) than an expression of the "rational"[25,26] view that, conditional on the constraints imposed by time and computational resources, human perception and cognition may be an optimal response to the underlying statistical structure of the environment.

ACKNOWLEDGMENTS

JIM and MAP were supported by NIH grant R01 MH57472. DJN was supported by Australian Research Council grant DP-0451793. We are indebted to Zygmunt Pizlo for introducing us to the literature on inverse problems and for pointing out the similarity between model selection and perception.

REFERENCES

1. P. Grünwald, "Model selection based on minimum description length," *Journal of Mathematical Psychology* **44**, pp. 133–170, 2000.
2. Z. Pizlo, "Perception viewed as an inverse problem," *Vision Research.* **41**, pp. 3145–3161, 2001.
3. I. J. Myung and M. A. Pitt, "Applying Occam's razor in modeling cognition: A Bayesian approach," *Psychonomic Review & Bulletin* **4**, pp. 79–95, 1997.
4. M. A. Pitt and I. J. Myung, "When a good fit can be bad," *Trends in Cognitive Sciences* **6**(10), pp. 421–425, 2002.
5. A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*, Winston, Washington, DC, 1977.
6. S. Haykin, *Neural Networks: A Comprehensive Foundation (Ch. 5)*, Wiley, New York, NY, 1986.
7. H. Linhart and W. Zucchini, *Model Selection*, Prentice Hall, London, UK, 1999.
8. F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statistical Science* **1**(4), pp. 502–527, 1986.

9. H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki, eds., pp. 267–281, Akademiai Kiado, 1973.
10. G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics* **6**(2), pp. 461–464, 1978.
11. A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory* **44**, pp. 2743–2760, October 1998.
12. M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, New York: Springer, second ed., 1997.
13. J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory* **42**, pp. 40–47, Jan 1996.
14. J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory* **47**, pp. 1712–1717, July 2001.
15. Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems in Information Transmission* **23**, pp. 3–17, 1987.
16. A. D. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection," *International Statistical Review* **69**, pp. 185–212, 2001.
17. M. J. Schervish, *Theory of Statistics*, Springer-Verlag, New York, NY, 1995.
18. S. I. Amari, *Differential Geometrical Methods in Statistics*, Springer-Verlag, New York, NY, 1985.
19. I. J. Myung, V. Balasubramanian, and M. A. Pitt, "Counting probability distributions: Differential geometry and model selection," *Proceedings of the National Academy of Sciences* **97**, pp. 11170–11175, 2000.
20. R. N. Shepard, C. I. Hovland, and H. M. Jenkins, "Learning and memorization of classification," *Psychological Monographs* **75**(13), p. 517, 1961.
21. R. M. Nosofsky, M. A. Gluck, T. J. Palmeri, S. C. McKinley, and P. Glauthier, "Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland and Jenkins (1961)," *Memory & Cognition* **22**, pp. 352–369, 1994.
22. J. K. Kruschke, "ALCOVE: An exemplar-based connectionist model of category learning," *Psychological Review* **99**, pp. 22–44, 1992.
23. D. J. Navarro and M. D. Lee, "Applications of minimum description length clustering: Partitioning learning curves," *Manuscript submitted for publication.*
24. P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri, "An MDL framework for data clustering," in *Advances in Minimum Description Length: Theory and Applications*, P. Grünwald, I. J. Myung, and M. A. Pitt, eds., in press.
25. J. R. Anderson, *The Adaptive Character of Thought*, Lawrence Erlbaum, Hillsdale, NJ, 1990.
26. R. N. Shepard, "Perceptual-cognitive universals as reflections of the world," *Psychonomic Bulletin and Review* **1**, pp. 2–28, 1994.