Model Selection by Normalized Maximum Likelihood

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

[a] *Department of Psychology*
Ohio State University
238 Townshend Hall
1885 Neil Avenue Mall
Columbus, Ohio 43210-1222
{myung.1, pitt.2}@osu.edu
614-292-1862 (voice)

[b] *Department of Psychology*
University of Adelaide, Australia
daniel.navarro@complex.psych.adelaide.edu.au

[c] *Corresponding Author*

Abstract

The Minimum Description Length (MDL) principle is an information theoretic approach to inductive inference that originated in algorithmic coding theory. In this approach, data are viewed as codes to be compressed by the model. From this perspective, models are compared on their ability to compress a data set by extracting useful information in the data apart from random noise. The goal of model selection is to identify the model, from a set of candidate models, that permits the shortest description length (code) of the data. Since Rissanen originally formalized the problem using the crude 'two-part code' MDL method in the 1970s, many significant strides have been made, especially in the 1990s, with the culmination of the development of the refined 'universal code' MDL method, dubbed Normalized Maximum Likelihood (NML). It represents an elegant solution to the model selection problem. The present paper provides a tutorial review on these latest developments with a special focus on NML. An application example of NML in cognitive modeling is also provided.

Key Words: Minimum Description Length, Model Complexity, Inductive Inference, Cognitive Modeling.

To select among competing models of a psychological process, one must decide which criterion to use to evaluate the models, and then make the best inference as to which model is preferable. For models that can be formulated as probability distributions, there exist a range of statistical model selection methods to assist in this endeavor. The purpose of this paper is to describe the latest advances in an information theoretic model selection method known as Minimum Description Length (MDL). We begin by describing the philosophy and assumptions behind MDL. This is followed by a tutorial on its most recent incarnation Normalized Maximum Likelihood (NML), as well as its relationship with other selection methods. The paper ends with an application of NML to the analysis of category learning data.

1 Statistical Inference as Data Compression: The MDL Approach

The Minimum Description Length principle (Rissanen 1978, 1989) is a statistical inference method that originated in information theory (Cover & Thomas, 1991), in contrast to both classical-frequentist methods and Bayesian methods which stem from probability theory. In classical and Bayesian methods we begin with the assumption that there exists a true probability distribution $f_T(\cdot)$ from which the observed data were sampled, and the goal of statistics is to develop models that approximate this truth. It follows that the goal of model selection is to find the model that is closest to the truth. According to the MDL principle, this foundational assumption is incorrect. Rissanen (2003, p. 4) argues that

> "..such a process can work well if the situation is similar to that in physics, namely, that there is a 'law' which is guessed correctly and which is capable of describing the data sufficiently well for one to be able to account for the unavoidable deviations [due] to small random instrument noise ... In general, however, we do not have enough knowledge of the machinery that generates the data to convert it into a probability distribution, whose samples would be statistically similar to the observed data, and we end up in the impossible task of trying to estimate something that does not exist."

According to this view, the question of whether the true distribution $f_T(\cdot)$ even exists is inherently unanswerable. We are thus ill-advised to base our inferential procedures on an unjustifiable faith in an unknown truth. In response to this concern, the MDL principle adopts a very different approach to modeling. It proposes that the basic goal should be to find regularities in data, and use these regularities to compress the data set so as to unambigiously "describe it using fewer symbols than the number of symbols needed to describe the data literally" (Grünwald 1998, p. 6). The more a model permits data compression, the more the model enables us to discover the regularities underlying the data. Therefore the goal of model selection is to identify the model that allows the greatest compression of the data. Conceptually, the intent is not to discover the truth about the world so much as it is to provide the most concise account of our observations about the

world (i.e., data). Although these two viewpoints sound similar, they lead to somewhat different statistical machinery. [1]

The MDL approach has evolved over the past 30 years from its initial "two-part" code that was limited in applicability, to the more "modern" and sophisticated form that has made its way into the psychological literature (Grünwald, 2000; Navarro & Lee, 2004; Pitt, Myung & Zhang, 2002). Normalized Maximum Likelihood (NML) is currently the endpoint of this journey, providing a theoretically elegant solution. Readers interested in more detailed coverage of the material that follows should consult the tutorial chapter by Grünwald (2005), upon which section 2 of this paper is drawn. A comprehensive summary of recent developments in MDL theory and its application can be found in Grünwald, Myung and Pitt (2005).

2 Normalized Maximum Likelihood

We begin with a discussion of the foundational ideas that underlie MDL and NML. We do so to ensure that this paper is self-contained. An excellent discussion can be found in Grünwald (2000; see also Li & Vitányi, 1997), so there is little need to repeat the material in detail.

2.1 Codes, Codelengths and Probabilities

Suppose we are presented with a data set $\boldsymbol{x}$ that consists of the sequence of n symbols $x_1\, x_2\, \ldots\, x_n$. If we are flipping coins, for instance, this could be the binary sequence

HHHHHHHTHTHHHTHTTHHHHHHHHHHH

The literal description length of this data set, when written in the binary *alphabet* (H,T), is $l(\boldsymbol{x}) = n = 28$ symbols. However, it may be possible to devise a different *encoding* of the data set that is much shorter. Suppose that we adopt a *code* that supposes AA=H, AB=T, BA=HHH, BB=HHHH. This code also uses a binary alphabet, in this case (A,B), but when we write out the data in code, we obtain

BBBAABAAABBAABAAABABBBBBBB

which has a *codelength* of only $l(\boldsymbol{x}) = 26$ binary symbols. Notice that an observer who knows the code can precisely decipher the message. No message written in this code can

[1] It is worth noting that this philosophical position is not inherent to information theory. Other information theoretic approaches to statistics take a "subjectivist" Bayesian view, notably the Minimum Message Length framework (Wallace & Boulton, 1968; Wallace & Dowe 1999a).

correspond to multiple data sets, due to the fact that none of the *codewords* (i.e., AA, AB, BA, BB) is a prefix of any other, making the code a so-called *prefix code* that guarantees unique decodability. Moreover, note that the code makes a statement about the kinds of regularities that we expect to see in the data, namely that there should be more heads than tails.

With this in mind, the key to understanding the MDL principle lies in Shannon's source coding theorem and in the Kraft inequality. The Kraft inequality states that for any computable probability distribution $p(\cdot)$ there exists a corresponding prefix code that encodes the data set $\boldsymbol{x}$ as a sequence of $l(\boldsymbol{x}) = -\lceil \log_2 p(\boldsymbol{x}) \rceil$ binary symbols ("bits"), and vice versa, in such a way that short codewords are assigned to frequent data symbols and long codewords are assigned to infrequent data symbols. The symbol $\lceil z \rceil$ denotes the smallest integer greater than or equal to z. The convention in information theory is to use a slight idealization that allows non-integer codelengths. It is also commonplace to think of these idealized codelengths in "nats" rather than bits, where a nat refers to an "alphabet" that consists of e "symbols", so we use the natural logarithm rather than the binary logarithm. Under these idealizations, the correspondence between codelength functions and probability distributions is much simpler, since $l(\boldsymbol{x}) = -\ln p(\boldsymbol{x})$. In other words, there is an exact correspondence between prefix codes and probability distributions. Moreover, for data sequences generated from the distribution $p(\cdot)$, Shannon's source coding theorem tells us that this coding scheme is optimal in the sense that it minimizes the expected length of an encoded data set (Hansen & Yu, 2001). Using these optimal *Shannon-Fano-Elias* 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 $p(\cdot)$. To be slightly more precise, the Kraft inequality allows us to associate the distribution $p(\cdot)$ with the code, and Shannon's theorem tell us the code would be optimal if the data were actually generated from this distribution. Taken together, these theorems tell us that when the observed data $\boldsymbol{x}$ are compressed using this code, the resulting codelength is $-\ln p(\boldsymbol{x})$ and the minimum expected codelength is achieved for data generated from $p(\cdot)$.

2.2 Universal Codes and Universal Distributions

If theoretical models consisted only of a single probability distribution $p(\cdot)$, then these basic principles of information theory would have provided a complete (and trivial) solution to the model selection problem. The probability model that best compresses the data is simply the one that assigns highest probability to those data. That is, choose the model that fits the data best (in the maximum likelihood sense). However, this is not how real modeling works. Theoretical models are built with parameters $\boldsymbol{\theta}$ that are allowed to vary from context to context, and what the model specifies is the conditional distribution $f(\cdot|\boldsymbol{\theta})$. As a result, the model is actually a *family* of probability distributions consisting of all the different distributions that can be produced by varying $\boldsymbol{\theta}$. In this situation, the model selection problem is substantially more difficult. From an MDL perspective,

we would like to compare two models on their ability to compress the data. However, in order to do so, we need to know how to optimally encode the data $\boldsymbol{x}$ with the help of an entire family of distributions M. The corresponding codelength is called the *stochastic complexity* (SC) of $\boldsymbol{x}$ with respect to M.

How should we optimally encode the data $\boldsymbol{x}$ with the help of the model family M? Notice that, since the Kraft inequality establishes a correspondence between codes and probability distributions, this is the same as asking what probability distribution should be used as the most suitable substitute for the family M when trying to describe the data $\boldsymbol{x}$. An obvious answer would be to use the code corresponding to $f(\cdot|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate for the data $\boldsymbol{x}$, since this distribution assigns shorter codelength to the data than any of the other distributions in the model family. The problem with doing so is that the code remains unknown until the data are observed. If person A wishes to describe $\boldsymbol{x}$ to person B (who has not yet seen the data) using the code corresponding to $f(\cdot|\hat{\boldsymbol{\theta}})$, then *both* A and B need to know the code, which depends on $\hat{\boldsymbol{\theta}}$, which in turn depends on the data. Since person B has not yet seen the data, this is impossible. In other words, since this "maximum likelihood code" depends on the data themselves, it cannot be used to describe those data in an unambiguous (i.e., unique) manner. Something else is required.

To recap, we wish to identify the single probability distribution that is "universally" representative of an entire family of probability distributions in the sense that the desired distribution mimics the behavior of any member of that family (Barron, Rissanen & Yu, 1998). The MDL principle gives guidelines as to how to construct such a distribution. Formally, a *universal distribution* $p_U(\boldsymbol{x})$ relative to a family of distributions M is defined as a distribution that allows us to compress every data set $\boldsymbol{x}$ almost as well as its maximum likelihood code (i.e., $-\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$) in the following sense (Grünwald, 2005):

$$-\ln p_U(\boldsymbol{x}) \leq -\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + K_n(M) \tag{1}$$

where $K_n(M)$ increases sublinearly in n, that is, $\lim_{n\to\infty} K_n(M)/n = 0$. The *Shannon-Fano-Elias* code corresponding to this universal distribution is called the *universal code*. According the equation (1), the universal code is nearly as good as the maximum likelihood code, with the difference between the two codes being no more than $K_n(M)$, yet importantly, the universal code avoids the pitfalls of the 'maximum likelihood code'. Notice that there may exist multiple universal distributions, each with a different value of $K_n(M)$. We would then be interested in finding the "optimal" universal distribution for the model family M. The NML solution is presented next.

2.3 NML as Optimal Universal Distribution

To understand the NML solution, it is useful to consider how conventional inference methods conceive the problem. In most statistical frameworks, the problem is addressed by trying to find the model that is "closest" to the true distribution $f_T(\cdot)$ in some well-defined sense. One natural choice is to measure the discrepancy between the model and the truth using the Kullback-Liebler divergence (Kullback, 1968),

$$D(p||f_T) = E_{f_T}\left[\ln \frac{f_T(\boldsymbol{x})}{p(\boldsymbol{x})}\right]. \tag{2}$$

Information theoretically, the Kullback-Liebler approach is appealing because it measures the amount of information lost when $p(\cdot)$ is used to approximate $f_T(\cdot)$. This is in the sense that, when data is generated from $f_T(\cdot)$ and encoded using $p(\cdot)$, an average of $D(p||f_T)$ additional nats are required beyond what would have been needed had we used the optimal code specified by $f_T(\cdot)$. Under the information theoretic view, model selection should aim to find the model that best approximates the truth. This approach (under some very strong asymptotic assumptions) was used to derive the Akaike Information Criterion (AIC; Akaike, 1973). As such, the approach relies on the assumption that a true distribution really exists. As mentioned earlier, this assumption is rejected in MDL. It is not simply that the true distribution is unknown but the assumption of the data generating distribution is "quite irrelevant to the task at hand, namely, to learn useful properties from the data" (Rissanen, 1989, p. 17). What this implies is that the goal of model selection is not to estimate an assumed but 'unknown' distribution, but to find good probability models that help separate useful information in the data from noise (Rissanen, 1989, p. 84).

In this theoretically conservative approach, how should the problem of inference be redefined to avoid referring to a true distribution? Firstly, the machinery available to us in statistics are probability distributions, so we must work with what we have. Secondly, in order to make "safe" inferences about an unknown world, a cautious approach would be to assume a worst-case scenario of some kind. With this in mind, we return to the original problem, in which we are given a family of probability distributions to assist us with data coding. Recall that the absolute best performance that the model family is capable of for any data set $\boldsymbol{x}$ is equal to the minus log maximum likelihood value, $-\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$, but that it is not possible to specify the corresponding code before observing the data, making it useless for any practical purpose. We will instead use a universal distribution $p(\boldsymbol{x})$ to encode the data. The excess codelength needed to encode the data set $\boldsymbol{x}$ with this universal distribution, $\{-\ln p(\boldsymbol{x}) + \ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}$, is called the *regret* of $p(\boldsymbol{x})$ relative to M for the data.

From a worst-case perspective, the following question is an obvious one to ask: Given a coding scheme that we can specify in advance which corresponds to the distribution $p(\boldsymbol{x})$, how closely does it imitate the impossible scheme implied by $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$ under the worst conditions (i.e., when the data are generated from a "worst enemy" distribution that makes it hardest for $p(\boldsymbol{x})$ to approximate $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$)? More formally, the worst-case expected regret is given by

$$R(p||M) = \max_q E_q \left[\ln \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{p(\boldsymbol{x})} \right] \tag{3}$$

where the "worst enemy" distribution $q(\cdot)$ is allowed to be (almost) any probability distribution.

Comparison of (2) and (3) brings out the differences between the conventional and MDL approaches to the inference problem. In (2), we assume a true distribution f_T that plays two distinct roles: it is both the thing to be approximated (in the numerator), and the thing that we must assume produces the data (in the expectation). In MDL, we are not allowed to assume that such a thing exists, and (3) addresses these two roles differently. Firstly, we cannot approximate a true distribution because we are not allowed to assume such a thing in the numerator. Instead, we adopt the more modest goal of seeking an optimal coding scheme based on the various maximum likelihood codes that belong to the model family M. In the second case, we are not allowed to assume that the data are generated by a true distribution, so we adopt the most cautious approach we can think of, and assume that the data are generated from the distribution q under which the approximation is poorest.

Without making any reference to the unknown truth, Rissanen (2001) formulated finding the optimal universal distribution as a minimax problem: Find the coding scheme that minimizes the worst-case expected regret,

$$p^* = \arg_p \min_p \max_q E_q \left[\ln \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{p(\boldsymbol{x})} \right] \tag{4}$$

where q ranges over the set of all probability distributions satisfying $E_q \left[\ln \frac{q(\boldsymbol{x})}{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})} \right] < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Neither p or q is required to be a member of the model family, nor is the solution, p^*. The process by which the NML is derived is illustrated in Figure 1. The distribution that satisfies this minimax problem is the *normalized maximum likelihood*

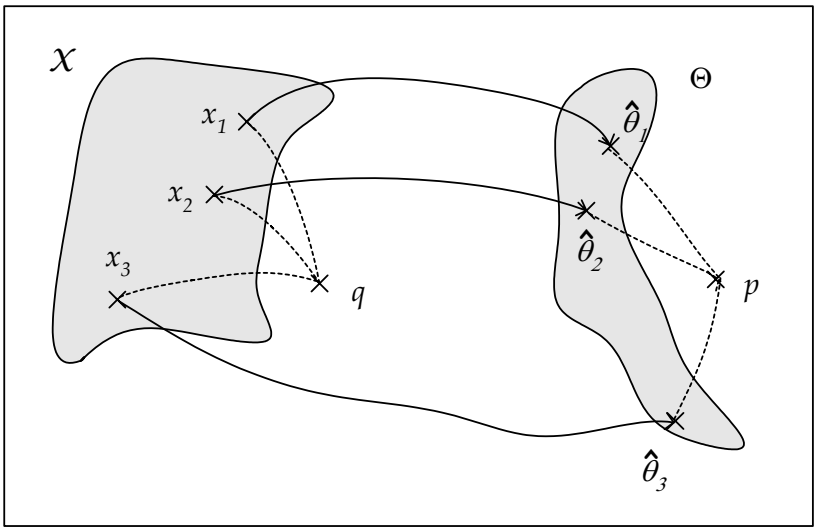


Fig. 1. A schematic illustration of the minimax problem used to derive NML. Data $\boldsymbol{x}$ are treated as if they were generated from q, and the model fit is given by $f(\cdot|\hat{\boldsymbol{\theta}})$. The optimal distribution p^* minimizes the expected discrepancy from p to $\hat{\boldsymbol{\theta}}$, under the assumption that the distribution q that generates $\boldsymbol{x}$ is chosen to maximize this discrepancy.

(NML) distribution (Barron, Rissanen & Yu, 1998; Rissanen, 2001) [2],

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

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{y}}$ denotes the maximum likelihood estimate for the data set $\boldsymbol{y}$. Therefore, the probability that this optimal universal distribution p^* assigns to the data set $\boldsymbol{x}$ is proportional to the maximized likelihood value $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$, and the normalizing constant $\int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\, d\boldsymbol{y}$ 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. When the data $\boldsymbol{x}$ is defined over a discrete sample space (e.g., binomial data), the integration symbol $\int$ in (5) is replaced by the summation symbol $\sum$.

[2] It is interesting to note that the same NML distribution $p^*(\boldsymbol{x})$ can also be derived as the solution to another minimax problem defined as $p^* = \arg_p \min_p \max_{\boldsymbol{x}} \left(\ln \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{p(\boldsymbol{x})} \right)$ (Shtarkov, 1987). Note that in this minimax formulation, the worst-case *individual data* regret is being minimized, rather than the worst-case *expected* regret as in (4).

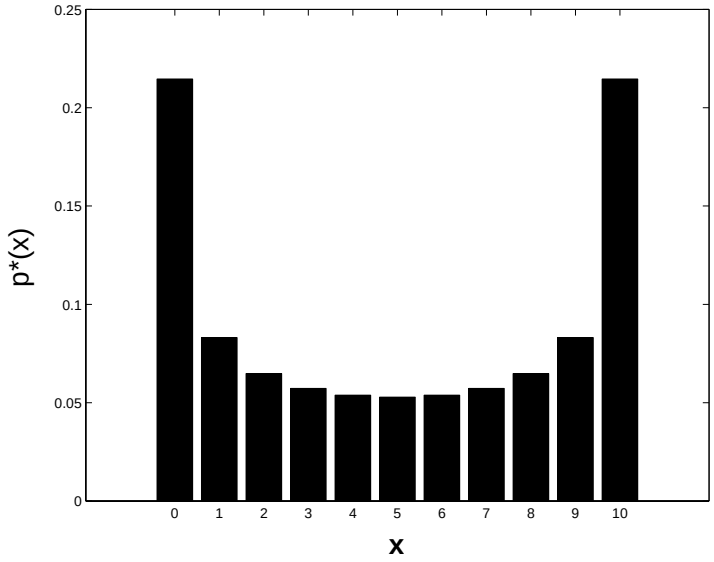


Fig. 2. The NML distribution for the binomial model class with sample size $n = 10$.

The codelength of the normalized maximum likelihood, $-\ln p^*(\boldsymbol{x})$, is referred to as the stochastic complexity of the data set $\boldsymbol{x}$ with respect to the model class M and is given by

$$\mathrm{SC}_1 = -\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + \ln \int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\, d\boldsymbol{y}. \tag{6}$$

In (6) the first term of the right hand side is a lack of fit measure and the second term defines the *complexity* of the model class M. Thus, in SC_1, model complexity is operationalized as the logarithm of the *sum of all best fits* a model class can provide collectively. 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 model's ability to fit random data sets (Myung & Pitt, 1997). Another interpretation of complexity is that it is equal to the minimized worst-case *expected* regret, i.e., the expected regret at $p^*(\boldsymbol{x})$ (Rissanen, 2001),

$$\ln \int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\, d\boldsymbol{y} = E_q \left[\ln \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{p^*(\boldsymbol{x})} \right]. \tag{7}$$

According to the MDL principle, given a set of competing models, we first use the NML distributions to compare their ability to compress the data and then select the one model that minimizes the SC_1 criterion.

To give an example showing how NML distributions are defined, consider the binomial variable x with the probability mass function,

$$f(x|\theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} \tag{8}$$

where $x = 0, 1, ..., n$ and $0 < \theta < 1$. The maximum likelihood estimate of the parameter θ is given by $\hat{\theta} = x/n$. The NML distribution $p^*(x)$ is then defined as

$$p^*(x) = \frac{\frac{n!}{x!(n-x)!}\left(\frac{x}{n}\right)^x\left(\frac{n-x}{n}\right)^{n-x}}{\sum_{y=0}^{n}\frac{n!}{y!(n-y)!}\left(\frac{y}{n}\right)^y\left(\frac{n-y}{n}\right)^{n-y}}. \tag{9}$$

For sample size n = 10, the normalizing constant in the denominator of (9) is approximately 4.66, the logarithm of which (i.e., $\ln(4.66) = 1.54$) gives the complexity measure of the model class. Figure 2 shows $p^*(x)$ for $n = 10$. As can be seen in the figure, $p^*(x)$ does not resemble any binomial distribution as the NML distribution resides outside of the model class. Despite this "misgiving," it is important to note that the NML distribution is the one that is universally representative of the entire model class in the minimax sense in (4).

There are two important caveats concerning the implementation of the stochastic complexity criterion in (6) in practice. First, if the value of the normalizing constant in (5) is infinite, the NML distribution p^* is then undefined and therefore cannot be used. This is sometimes known as the *infinity problem*, to which there is not yet a fully general solution, though several remedies have been suggested to 'repair' this undesirable situation (Grünwald, 2005). The problem is currently a topic of active research and discussion in the field (Rissanen, 2000; Lanterman, 2005; Foster & Stine, 2005; Wallace & Dowe, 1999b).

Second, although the complexity measure in (7) is invariant to reparameterizations of the model, it is not necessarily invariant to different choices of experimental design. To illustrate, suppose that we have observed 7 heads out of 10 independent Bernouilli trials in a coin tossing experiment and that we wish to compute stochastic complexity. It turns out that this is not possible because additional information is needed. That is, different values of the normalizing constant are obtained depending upon the sampling scheme we assume for the data, whether we had planned to terminate the experiment after 10 trials regardless of the outcomes (i.e., binomial sampling) or to continue the experiment until 7 heads are observed (i.e., negative binomial sampling). Wallace and Dowe (1999b) recently questioned NML-based model selection on the grounds that this violates the Likelihood Principle (LP; e.g., Berger & Wolpert, 1988). The LP states that given a data set $\boldsymbol{x}$, the likelihood function for the data $f(\boldsymbol{x}|\boldsymbol{\theta})$ contains all the information about $\boldsymbol{\theta}$. If information about $\boldsymbol{\theta}$ lies outside the likelihood function such as in how the experiment was carried out or other background information, it is a violation of this principle (Press, 2003, p. 36).

As Edwards, Lindman and Savage (1963, p. 193) argue, "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience". However, it is worth noting that there is some disagreement over the LP in statistics (e.g., Hill, 1987), and with regard to MDL in particular. In particular, given that MDL focuses on sequentially observed data transmitted over some channel, it is not obvious whether the stopping rules should be irrelevant to MDL (Peter Grünwald, personal communication). In short, this remains an issue for theoreticians to address in the future.

2.4 Asymptotic Approximations to NML

Having introduced the NML criterion as an "optimal" implementation of the MDL principle, it is instructive to compare it with a number of other formulae that have, at different points in time, been referred to as "the MDL criterion". The version of MDL that has been used most frequently in psychology is the "Fisher information approximation to the stochastic complexity criterion". This is the criterion used by Pitt et al. (2002) and discussed at length in Grünwald (2000). It has proven to be fairly robust in model selection and is much more tractable than NML. The expression was derived by Rissanen (1996) by taking an asymptotic expansion of the complexity term in (6) for large sample sizes,

$$\mathrm{SC}_2 \;=\; -\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int_{\Theta}\sqrt{\det I(\boldsymbol{\theta})}\,d\boldsymbol{\theta} + o(1) \tag{10}$$

where n denotes the sample size, k is the number of model parameters, and $I(\boldsymbol{\theta})$ is the Fisher information matrix (e.g., Schervish, 1995) of sample size 1 defined as $I(\boldsymbol{\theta})_{i,j} = -E_{f(\cdot|\boldsymbol{\theta})}\left[\frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right]$, $i,j = 1,...,k$. The $o(1)$ term collects all the higher-order terms in the asymptotic expansion, and vanishes as $n \to \infty$. The second and third terms together are often referred to as the complexity of the model, albeit an asymptotic one. Like NML, this complexity measure is reparametrization-invariant and is not necessarily invariant under changes of experimental design (Wallace & Dowe, 1999b).

The second and third terms in (10) reveal 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 second term of the SC_2 criterion captures the number of parameters whereas the third 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 second term but not in the third. This implies that as the sample size becomes large, the relative contribution of the third term to that of the second becomes negligible, further

reducing SC_2 to another asymptotic expression:

$$\mathrm{SC}_3 \;=\; -\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) \tag{11}$$

which is one of the early formulations of the MDL principle (Rissanen, 1978 & 1983). Second, because the second term of SC_2 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 third term in SC_2 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.

Returning the discussion to SC_2 in (10), the comparative tractability and ease of interpretability of this criterion make it a tempting alternative to the NML. However, there are some important caveats that are often neglected in applied contexts, relating to regularity conditions. As Rissanen (1996) remarks, in order to apply this expression, the most important things to verify are that:

- The maximum likelihood estimate must lie "sufficiently" in the interior of the model. That is, for some $\epsilon > 0$ and for all large n, the best fitting parameter value $\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}$ must be further than ϵ from the edge of the parameter space;
- All elements of the Fisher information matrix $I(\boldsymbol{\theta})$ must be continuous in Θ;
- The Fisher information integral term $\int_\Theta \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta}$ must be finite.

If these conditions are violated, then the stochastic complexity SC_2 is not well-defined and may exhibit anomalous behavior (see Navarro, 2004 for an example). Rissanen (1996; see also Lanterman, 2001, 2005) discusses several 'repair' methods for handling such situations. For example, a small neighborhood around singular points of $I(\boldsymbol{\theta}) = \infty$ may be excluded from the computation, or the range of parameters may be restricted in order to make the Fisher information integral finite.

2.5 Relationships to Bayesian Formulae

Interestingly, one can also interpret the two asymptotic approximations, SC_2 and SC_3, from a Bayesian viewpoint. In Bayesian statistics (Gelman, Carlin, Stern & Rubin, 2004), the goal of model selection is to choose, among a set of candidate models, the one with the largest value of the marginal likelihood defined as

$$p_{_{Bayes}}(\boldsymbol{x}) \;=\; \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{12}$$

where $\pi(\boldsymbol{\theta})$ is a prior density for the parameter $\boldsymbol{\theta}$. The term *Bayes factor* (Kass & Raftery, 1995), which is often mentioned in Bayesian model selection, is referred to as the ratio of the marginal likelihood of one model to the marginal likelihood of a second model. An asymptotic expansion of the minus log marginal likelihood using the 'non-informative' Jeffreys' prior $\pi_J(\boldsymbol{\theta}) = \sqrt{I(\boldsymbol{\theta})}/\int_{\boldsymbol{\Theta}} \sqrt{I(\boldsymbol{\theta})}d\boldsymbol{\theta}$ yields (Balasubramanian, 1997, 2005)

$$-\ln p_{_{Bayes}}(\boldsymbol{x}) \;=\; -\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln \int\limits_{\Theta} \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta} + o(1) \tag{13}$$

which is exactly the same as SC_2 in (10). Hence, for large n, Bayesian model selection with Jeffreys prior and NML become virtually indistinguishable. Obviously, this asymptotic "equivalence" would not hold if a different form of prior is used or if the sample size is not large.

By neglecting the sample-size independent terms in the right-hand side of (13) and then multiplying the result by factor 2, we get another asymptotic expression

$$-2\ln p_{_{Bayes}}(\boldsymbol{x}) \;\approx\; -2\ln f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + k\ln n. \tag{14}$$

This is known as the Bayesian Information Criterion (BIC; Schwarz, 1978) and is essentially the same as SC_3 in (11).

One closing word of caution. Despite the similarities in asymptotic expressions, Bayesian inference and MDL are different in their formulations: the marginal likelihood in Bayesian statistics is not the same as the normalized maximum likelihood in MDL. Therefore they will generally give different results. For further discussions on the relationships between Bayesian inference and MDL, the reader is directed to Grünwald (2005) and Vitányi & Li (2000).

2.6 Predictive Inference and the MDL Principle

Predictive inference and data compression are often interchangeable terms in model selection. Quoting Vitányi and Li (2000, p. 448),"compression of descriptions almost always gives optimal predictions." That said, stochastic complexity and methods of predictive inference have different origins. In the predictive approach to model selection, the ultimate goal is the minimization of prediction errors, and selection criteria differ from one another in how prediction errors are conceptualized and measured (Geisser & Eddy, 1979; Geisser, 1993; Linhart & Zucchini, 1986; Zucchini, 2000). Many of the non-MDL model selection criteria, such as AIC, cross-validation (Stone, 1974) and bootstrap model selection (Efron, 1983) were derived as generalizability criteria in which one is concerned with identifying a model family that yields the best future predictions. Stochastic complexity, on the other

hand, is motivated from a coding-theoretic view of model selection in which the goal is to identify a model family that permits the tightest compression of a data set by effectively filtering out random noise and attending to all of the 'useful' information in the data.

It turns out that there are close ties between these seemingly disparate approaches to model selection, predictive inference in one hand and data compression on the other.The predictive interpretation of the stochastic complexity has its root in the *prequential analysis* pioneered by Dawid (1984, 1992). To motivate this analysis, let us assume that data $\boldsymbol{x}^t = (x_1, ..., x_t)$ are observed sequentially $\{1, ..., t\}$ and that we are interested in predicting the next observation $\boldsymbol{x}^{t+1}$ on the basis of the data observed so far, that is, $\boldsymbol{x}^t$. Let $\hat{\boldsymbol{\theta}}(\boldsymbol{x}^t)$ denote the maximum likelihood estimate of a model class M for the data vector $\boldsymbol{x}^t$. Suppose that we use this estimate to predict the next observation $\boldsymbol{x}^{t+1}$ and further, that we evaluate performance of the maximum likelihood estimate by the logarithmic loss function, $-\ln f(\boldsymbol{x}^{t+1}|\boldsymbol{x}^t, \hat{\boldsymbol{\theta}}(\boldsymbol{x}^t))$. The accumulated prediction error (APE) over a series of observations $t = 1, ..., n$ is then given by

$$\mathrm{APE}(\boldsymbol{x}^n) = -\sum_{t=0}^{n-1} \ln f(\boldsymbol{x}^{t+1}|\boldsymbol{x}^t, \hat{\boldsymbol{\theta}}(\boldsymbol{x}^t)) \tag{15}$$

(See Wagenmakers, Grünwald and Steyvers (2005) for a tutorial on APE, along with application examples in time-series analysis.) It has been shown that the expression in (15) essentially reduces to SC_3 in (11) as $n \to \infty$ under regularity conditions (Rissanen, 1986, 1987; Dawid, 1992; Grünwald & de Rooij, 2005). [3] An implication of this observation is that the model that permits the greatest compression of the data is also the one that minimizes the accumulated prediction error, thereby providing justification for stochastic complexity as a predictive inference method, at least asymptotically.

3 Using NML in Cognitive Modeling

To provide an application example of NML in cognitive modeling, we consider the seminal experiment in human category learning conducted by Shepard, Hovland and Jenkins (1961). In this study, human performance was examined on a category learning task involving eight stimuli divided evenly between two categories. The stimuli were generated

[3] The primary regularity condition required for the equivalence proof is that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{x}^t)$ satisfies the central limit theorem such that the tail probabilities are uniformly summable in the following sense: $P\left(\sqrt{n} \parallel \hat{\boldsymbol{\theta}}(\boldsymbol{x}^t) - \boldsymbol{\theta} \parallel \geq n\right) \leq \delta(n)$ for all $\boldsymbol{\theta}$ and $\sum_n \delta(n) < \infty$ where $\parallel \boldsymbol{\theta} \parallel$ denotes a norm measure (Rissanen, 1986, Theorem 1). Recently, Grünwald and de Rooij (2005) identified another important condition for the asymptotic approximation, i.e., that the model is correctly specified. According to their investigation, under model mis-specification, one can get quite different asymptotic results.

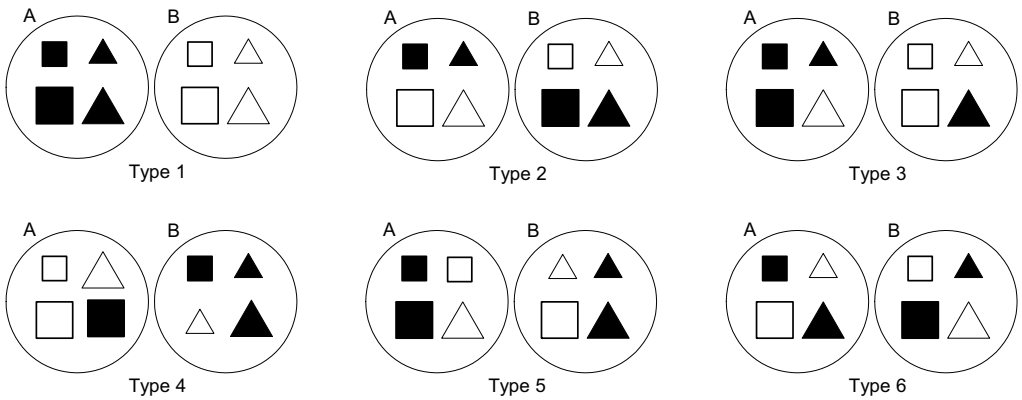


Fig. 3. The six category structures that comprise the Shepard, Hovland and Jenkins task.

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 3.

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) work. Figure 4 shows 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.

In order to claim that a category learning model 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 4. 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

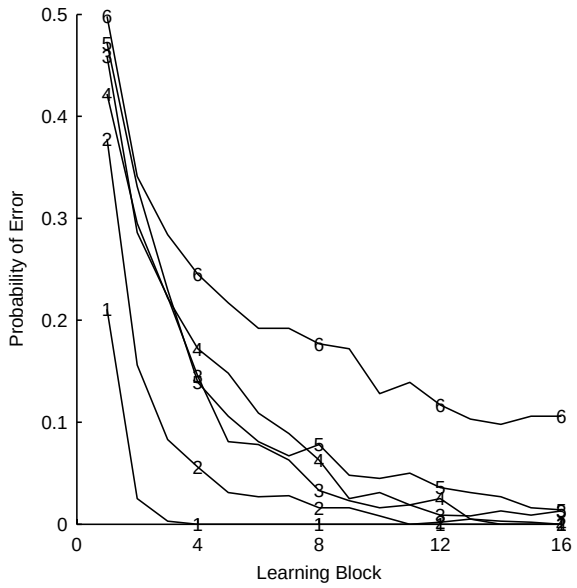


Fig. 4. Empirical learning curves for the Shepard, Hovland and Jenkins task (from Nosofsky et al., 1994).

one that maximizes the NML. 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 NML is known to be an appropriate solution. Under a model-based clustering procedure, the data are treated as the outcome of some random process. A clustering solution is thus treated as a *model* for the data, and the adequacy of that solution can be assessed using statistical model selection tools. In this section we outline a clustering model for discrete data that is appropriate to the applied problem of partitioning learning curves.

3.1 A Partitioning Model

Suppose that we have a discrete data set made up of T samples, each of which is an M-variate discrete probability over H response options. For instance, we might have T participants who solve M different kinds of problems, and each problem has H possible answers. Note that since each class of problem may have a different number of potential responses, H should technically be denoted H_m. However, this subscript will be dropped, since it will be clear from context. A particular partitioning of these T samples might be expressed in the following way. If we assume that there are K clusters, we might let D_k indicate how many of the original samples fall into the kth cluster. So D_k represents the size of the cluster, and thus $\sum_k D_k = T$. As before, we will generally drop the subscript k

when discussing D.

We represent the data $\boldsymbol{x}$ in terms of the statistics $x_{11}^{11} \ldots x_{DH}^{KM}$, where x_{dh}^{km} counts the number of observations that fall into the hth response category on the mth dimension for the dth sample that belongs to the kth cluster. In the example given earlier, x_{dh}^{km} would denote the number of times that participant d of cluster k gave the response h to a problem of type m. It will be convenient to define y_h^{km} and w^{km} as $y_h^{km} = \sum_{d=1}^{D} x_{dh}^{km}$, $w^{km} = \sum_{h=1}^{H} y_h^{km}$. In the example discussed, y_h^{km} is the number of times that someone in the kth cluster gave the answer h to a problem in m, while w_{km} is the total number of times that a problem of type m was presented to group k. A partitioning model for $\boldsymbol{x}$ consists of the set of K clusters $\boldsymbol{C} = (\boldsymbol{c}_1 \ldots \boldsymbol{c}_K)$. In this expression, $\boldsymbol{c}_k$ denotes the set of (indices of) samples that belong to the kth cluster. The model parameters $\boldsymbol{\theta} = (\theta_1^{11}, \ldots \theta_H^{MK})$ correspond to the probabilities with which each of the responses are chosen. Accordingly, θ_h^{mk} gives the probability with which response h is predicted to occur in trials belonging to cluster k and dimension m. Thus the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ is,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D} \prod_{h=1}^{H} (\theta_h^{km})^{x_{dh}^{km}} = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{h=1}^{H} (\theta_h^{km})^{y_h^{km}}$$

Note the y_h^{km} values are sufficient statistics for the data, assuming that the model is $\boldsymbol{C}$.

Besides the stipulation that observations come partially pre-clustered in samples, the main difference between this model class and that used by Kontkanen et al. (2005) is they employ a finite mixture model, in which the assignment of items to clusters is assumed to be the result of a latent probabilistic process. Motivated by the learning curves problem, we assume that a cluster is a *fixed* grouping of samples. Since the category structures that elicit the samples are derived from the fixed representational structure of the stimuli (Shepard et al., 1961), it makes little sense in this context to propose a model class in which object assignments are assumed to result from a probabilistic process. We now discuss how the NML computations are performed, and show that the results obtained by Kontkanen et al. (2005) apply to the current model. Since the current partitioning model is only a very minor variant on the approach adopted by Kontkanen et al. (2005), we provide only the most basic coverage, and refer interested readers to the original paper for a more detailed discussion.

For our clustering model, the MLE is given by $\hat{\theta}_h^{km} = \frac{y_h^{km}}{w^{km}}$. Substituting the MLE values into the likelihood function gives the maximized likelihood,

$$p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \left(\frac{\prod_{h=1}^{H} (y_h^{km})^{y_h^{km}}}{(w^{km})^{w^{km}}} \right).$$

This will enable us to efficiently calculate the NML value for any data set $\boldsymbol{x}$ when described

using a clustering model $\boldsymbol{C}$ as

$$p_{\text{NML}}(\boldsymbol{x}|\boldsymbol{C}) = \frac{p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{\sum_{\boldsymbol{y}} p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})}.$$

In the denominator term, the sum is taken over all possible data sets and represents the normalizing constant, denoted $\mathcal{R}_{\boldsymbol{C}} = \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}})$.

The normalizing constant for a clustering model $\boldsymbol{C}$ is given by,

$$R_{\boldsymbol{C}} = \sum_{y_1^{11}+\ldots+y_H^{11}=w^{11}} \cdots \sum_{y_H^{KM}+\ldots+y_H^{KM}=w^{KM}} \left[\prod_{m=1}^{M}\prod_{k=1}^{K} \frac{w^{km}!}{\prod_{h=1}^{H} y_h^{km}!}\right] \left[\prod_{m=1}^{M}\prod_{k=1}^{K} \frac{\prod_{h=1}^{H}(y_h^{km})^{y_h^{km}}}{(w^{km})^{w^{km}}}\right],$$

where the first square-bracketed term counts the number of data sets that have the sufficient statistics $y_1^{11}\ldots y_H^{KM}$, and the second square-bracketed term gives the maximized likelihood to any such data set. After rearranging:

$$R_{\mathbf{C}} = \sum_{y_1^{11}+\ldots+y_H^{11}=w^{11}} \cdots \sum_{y_1^{KM}+\ldots+y_H^{KM}=w^{KM}} \left[\prod_{m=1}^{M}\prod_{k=1}^{K} \frac{w^{km}!}{(w^{km})^{w^{km}}} \prod_{h=1}^{H} \frac{(y_h^{km})^{y_h^{km}}}{y_h^{km}!}\right].$$

Notice that any particular inner term depends on only a single value of m and k. Thus terms where $m = 1$ and $k = 1$ may be moved forward. Now, notice that all of the nested terms do not depend on the values of $y_1^{11}\ldots y_H^{11}$, so they can be removed as a factor. Repeating this for all m and k allows the normalizing constant to be factorized, yielding

$$R_{\mathbf{C}} = \prod_{m=1}^{M}\prod_{k=1}^{K} \left[\sum_{y_1^{mk}+\ldots+y_H^{mk}=w^{mk}} \frac{w^{km}!}{(w^{km})^{w^{km}}} \prod_{h=1}^{H} \frac{(y_h^{km})^{y_h^{km}}}{y_h^{km}!}\right].$$

Since individual clusters and dimensions are assumed to be independent, it is not surprising to see the normalizing constant factorize. The inner term corresponds to the normalizing constant $R(H, w)$ for a one-dimensional multinomial with H options and a sample size of w. That is, $R_{\mathbf{C}} = \prod_m \prod_k R(H_m, w^{mk})$. The problem of calculating multinomial normalizing constant is addressed by Kontkanen et al. (2005), so it suffices simply to restate their result:

$$R(H, w) = \sum_{r_1+r_2=w} \left(\frac{w!}{r_1!r_2!}\right)\left(\frac{r_1^{r_1}r_2^{r_2}}{w^w}\right) R(J_1, r_1)\, R(J_2, r_2),$$

where J_1 and J_2 are any two integers between 1 and $H - 1$ such that $J_1 + J_2 = H$. They use this result to calculate $R(H, w)$ efficiently using a recursive algorithm. In essence,

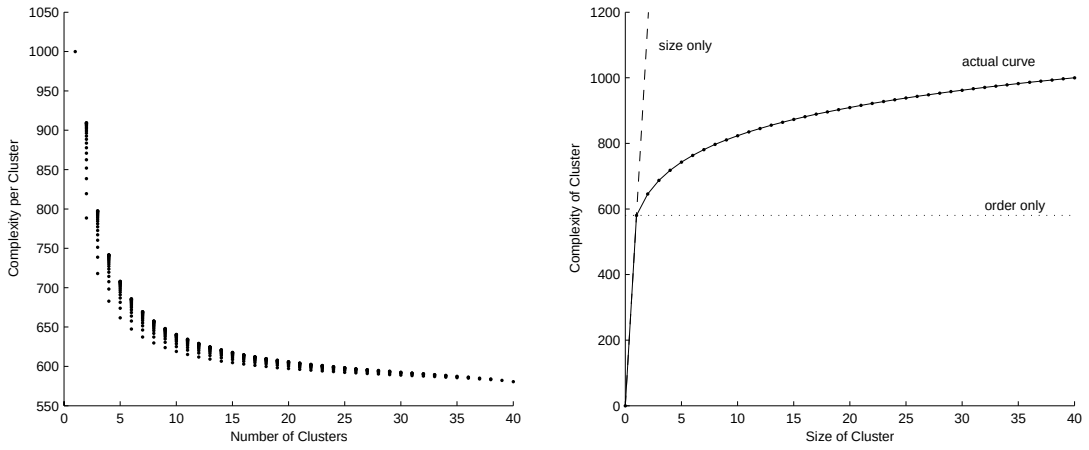
Table 1
Six clustering solutions to the Shepard et al. (1961) problem.

Partition	Lack of Fit $\left(-\ln f(\boldsymbol{x}\|\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right)$	Complexity $\left(\ln \int f(\boldsymbol{y}\|\hat{\boldsymbol{\theta}}_{\boldsymbol{y}}) d\boldsymbol{y}\right)$	SC_1
(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

we start by calculating all the binomial normalizing constants $R(2,1), \ldots R(2,w)$. This is reasonably fast since there are comparatively few ways of dividing a sample across two responses. Once these are known, they can be used to construct the normalizing constants for larger multinomials. For example, if we needed $H = 14$, we would set $J_1 = 2$ and $J_2 = 2$ to arrive at the normalizing constants for $H = 4$. We could then set $J_1 = 4$, and $J_2 = 4$ to get $H = 8$. Then $J_1 = 8$ and $J_2 = 4$ gives $H = 12$, and finally $J_1 = 12$ and $J_2 = 2$ would give the normalizing constant for $H = 14$. Obviously, at each step we need to calculate the sum over r_1 and r_2, but this can be done quickly by constructing tables of normalizing constant values. Once we have the normalizing constants for the various multinomials, we merely need to take the appropriate product to get the normalizing constant for the clustering model.

3.2 Partitioning Learning Curves

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 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



(a) (b)

Fig. 5. Understanding model complexity. In panel a, we see that model complexity per cluster $(1/K) \ln R$ is not constant, either as the number of clusters changes or within a fixed model order. In panel b, we note that complexity associated with a particular cluster increases with size (solid line). The dotted line ("order only") shows the predicted curve if only the number of clusters contributed to complexity. The dashed line ("size only") shows the predicted curve if complexity related only to the size of the clusters.

the closest competitor. Inspection of Figure 4 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.

3.3 Revisiting Model Complexity

Accounting for model complexity is an important topic in statistics (e.g., Hastie et al., 2001; Myung, 2000) with clustering models receiving particular attention in applied work (Lee, 2001; Lee & Navarro, 2005). Unfortunately, many approaches to model selection rely on asymptotic criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978), or else do not provide an explicit measure of model complexity (e.g., Bayes factors; see Kass & Raftery, 1995). As a result, a great deal of the discussion of complexity and model selection has relied on asymptotic measures (e.g., Pitt et al., 2002) that can be misleading in finite samples or when regularity conditions are violated (Lanterman, 2001; Navarro, 2004). In contrast the NML criterion is exact, and optimal (in the minimax coding sense discussed earlier) for data of any sample size. Moreover, it supplies a natural complexity measure (i.e., $\ln R$). Taken together, these two properties allow us to measure complexity properly and discuss it accurately.

It has often been argued (e.g., Lee, 2001; Lee & Navarro, 2005; Pitt et al. 2002) that model complexity is not the same as model order. However, these assertion have usually relied on asymptotic criteria: In a clustering context, Lee (2001) used a Laplace approximation to the Bayes factor (see Kass & Raftery, 1995), while Lee and Navarro (2005) used the Fisher information approximation to MDL. Using the recursive algorithm to calculate exact NML complexities for clustering models, it is worth briefly revisiting the question. Figure 5a plots NML complexity per cluster $(1/K) \ln R_{\boldsymbol{C}}$ against the number of clusters K for every possible partition of $T = 40$ samples, with $H = 20$ response options, $N = 100$ observations per cell, and $M = 16$ dimensions. If complexity is well-captured by the number of parameters, $(1/K) \ln R_{\boldsymbol{C}}$ should be constant. Figure 5 shows that complexity per cluster is not constant as K increases, nor is it constant across models with the same number of clusters. As suggested by Lee (2001), some partitions are indeed more complex than others even when the total number of clusters remains constant. The reason for this pattern becomes clearer when we consider the relationship between the size of a cluster (i.e., the number of samples assigned to it) and its complexity. Figure 5b plots this relationship for clusters of the same data sets referred to in Figure 5a (i.e., $T = 40$, $H = 20$, $N = 100$ and $M = 15$). The dotted line is the predicted curve if complexity were a constant function of model order, and the dashed line shows the prediction if complexity were a constant function of cluster size (in fact, if the dashed line were accurate, then each observation would contribute equally to complexity irrespective of how they were partitioned, and all clustering solutions would be of equal complexity). However, the figure shows that complexity is a concave increasing function of cluster size. If model complexity were equivalent to model order, this function would be constant, ensuring that all clusters contribute the same amount of complexity irrespective of size. Since the function is increasing, two clusters of size 1 are simpler than two clusters of size 2. Moreover, since the function is concave, complexity is subadditive. As a result, complexity is always decreased by transferring an observation from a small cluster to a large one, implying that the least complex solution is one in which all clusters except one are of size 1, while the remaining cluster is of size $T - K + 1$. This agrees with results based on Laplacian approximations (Lee, 2001).

4 Conclusion

In any scientific context we are presented with limited information that is consistent with an infinite number of explanations, but are required to infer the "best" account 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".

Author Notes

Correspondence address: Jay Myung, Department of Psychology, 238 Townshend Hall, 1885 Neil Avenue Mall, Columbus, Ohio 43210-1222, USA. E-mail: myung.1@osu.edu; Voice: 614-292-1862; Fax: 614-688-3984. JIM and MAP were supported by NIH grant R01 MH57472. DJN was supported by Australian Research Council grant DP-0451793. Portions of this research were presented at the 2005 Electronic Imaging Conference held in San Jose, CA and were subsequently published in the Proceedings. Other parts of the work have been submitted to the 2005 IEEE Information Theory Symposium held in Adelaide, Australia. The authors wish to thank three anonymous reviewers for many helpful suggestions on an earlier version of this manuscript.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (eds), *Second International Symposium on Information Theory, pp. 267–281.* Budapest: Akademiai Kiado.

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

Balasubramanian, V. (2005). MDL, Bayesian inference and the geometry of the space of probability distributions. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 81–98). Cambridge, MA: MIT Press.

Barron, A., Rissanen, J. & Yu, B (1998). The minimum description length principle in coding and modeling, *IEEE Transactions on Information Theory, 44*, 2743–2760.

Berger, J. O. & Wolpert, R. L. (1988). *The Likelihood Principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory* New York, NY: Wiley Interscience.

Dawid, P. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A, 147*, 278–292.

Dawid, P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics, Vol. 4*, pp.

109-125. Oxford, UK: Oxford University Press.

Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Efron, B. (1983). Estimating the error rate of a prediction rule: Some improvements on cross-validation. *Journal of the American Statistical Society, 78*, 316–331.

Foster, D. P. & Stine, R. A. (2005). The contribution of parameters to stochastic complexity. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 195–214). Cambridge, MA: MIT Press.

Geisser, S. (1993). *Predictive Inference: An Introduction*. New York, NY: Chapman & Hall.

Geisser, S. & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Society, 74*, 153–160.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. O. (2004). *Bayesian Data Analysis (2nd edition)*. New York, NY: Chapman & Hall.

Grünwald, P. (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph.D. Thesis, ILLC Dissertation Series DS 1998-03, CWI, The Netherlands.

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

Grünwald, P. (2005). Minimum description length tutorial. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 23–80). Cambridge, MA: MIT Press.

Grünwald, P. & de Rooij, S. (2005). Asymptotic log-loss prequential maximum likelihood codes. *Proceedings of the Eighteenth Annual Conference on Computational Learning Theory (COLT 2005)*.

Grünwald, P, Myung, I. J. & Pitt, M. A., eds.(2005) *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Hansen, M. H. & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96*, 746-774.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning*. New York: Springer.

Hill, B. M. (1987). The validity of the likelihood principle. *American Statistician, 41*, 95-100.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Society, 90*, 773-795. *Journal of the American Statistical Association, 91*, 1343-1370.

Kullback, S. (1968). *Information Theory and Statistics* (2nd Ed.). New York: Dover.

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J. & Tirri, H. (2005). An MDL framework for data clustering. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 323–354). Cambridge, MA: MIT Press.

Lanterman, A. D. (2001). Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection, *International Statistical Review, 69*, 185–212.

Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathe-*

matical Psychology, 45, 131–148.

Lanterman, A. D. (2005). Hypothesis testing for Poisson vs. geometric distributions using stochastic complexity. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 99–125). Cambridge, MA: MIT Press.

Lee, M. D. & Navarro, D. J. (2005). Minimum description length and psychological clustering models. In P. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications* (pp. 355-384). Cambridge, MA: MIT Press.

Li, M. & Vitányi, P. M. B. (1997). *An Introduction to Kolmogorov Complexity and its Applications* (2nd edition). New York, NY: Springer-Verlag.

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

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

Myung, I. J. & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

Navarro, D. J. (2004). A note on the applied use of MDL approximations. *Neural Computation, 16*, 1763-1768.

Navarro, D. J. & Lee, M. D. (2004). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin and Review, 11*, 961-974.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C. & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland and Jenkins (1961), *Memory & Cognition, 22*, 352–369.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472-491.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics* (2nd ed.). New York: Wiley.

Rissanen, J. (1978). Modeling by the shortest data description, *Automatica, 14*, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics, 11*, 416–431.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics, 14*, 1080-1100.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B, 49*, 223-239.

Rissanen, J. (1989) *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*, 40-47.

Rissanen, J. (2000). MDL denoising. *IEEE Transactions on Information Theory, 46*, 2537–2543.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47*, 1712–1717.

Rissanen, J. (2003). *Lectures on statistical modeling theory*. October 2003. Available online at www.mdl-research.org.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Shepard, R. N., Hovland C. I., & Jenkins H. M. (1961). Learning and memorization of classification, *Psychological Monographs, 75*, whole no. 13.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages, *Problems in Information Transmission, 23*, 3–17.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B, 39*, 44–47.

Vitányi, P. M. B., & Li, M. (2000). Minimum description length induction, Bayesianism and Kolmogorov complexity. *IEEE Transactions on Information Theory, 46*, 446–464.

Wagenmakers, E.-J., Grünwald, P. and Steyvers, M. (2005) Accumulative prediction error and the selection of time series models. Submitted to *Journal of Mathematical Psychology*.

Wallace, C. S. & Boulton, D. M. (1968). An information measure for classification. *Computer Journal*, 11, 185-194.

Wallace, C. S. & Dowe, D. L. (1999a). Minimum Message Length and Kolmogorov complexity. *Computer Journal, 42*, 270–287.

Wallace, C. S. & Dowe, D. L. (1999b). Refinements of MDL and MML coding. *Computer Journal, 42*, 330–337.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44*, 41–61.