
Minimum Description Length and Cognitive Modeling

Yong Su

*Department of Electrical Engineering
Ohio State University
2015 Neil Avenue Mall
Columbus, Ohio 43210-1272
USA
su.85@osu.edu
<http://peekaboo.hopto.org>*

In Jae Myung

*Department of Psychology
Ohio State University
1885 Neil Avenue Mall
Columbus, Ohio 43210-1222
USA
myung.1@osu.edu
<http://quantrm2.psy.ohio-state.edu/injae/>*

Mark A. Pitt

*Department of Psychology
Ohio State University
1885 Neil Avenue Mall
Columbus, Ohio 43210-1222
USA
pitt.2@osu.edu
<http://lpl.psy.ohio-state.edu>*

The question of how one should decide among competing explanations of data is at the heart of the scientific enterprise. In the field of cognitive science, mathematical models are increasingly being advanced as explanations of cognitive behavior. In the application of Minimum Description Length (MDL) principle to the selection of these models, one of the major obstacles is to calculate Fisher information. In the present study we provide a general formula to calculate Fisher information for models of cognition that assume multinomial or normal distributions. We also illustrate the usage of the formula for models of categorization, information integration, retention, and psychophysics. Further, we compute and compare the complexity penalty terms of two recent versions of MDL [Rissanen 1996; Rissanen 2001] for a multinomial model. Finally, the adequacy of MDL is demonstrated in the selection of retention models.

3.1 Introduction

The study of cognition is concerned with describing the mental processes that underly behavior and developing theories that explain their operation. Often the theories are specified using verbal language, which leads to an unavoidable limitation: The lack of precision. Mathematical modeling represents an alternative approach to overcoming this limitation by inferring structural and functional properties of a cognitive process from experimental data in explicit mathematical expressions.

Formally, a mathematical model or model class¹ is defined as a parametric family of probability density functions, $f_{X|\Theta}(x|\theta)$, as a Riemannian manifold in the space of distributions [Kass and Vos 1997], with $x \in \mathcal{X}$ and $\theta \in \Omega$. \mathcal{X} and Ω are the sample space and parameter space respectively. The sample space or parameter space could be an Euclidean space with arbitrary dimension. Thus the dimension of the parameter here corresponds to what is commonly referred to as the number of parameters of the model.

In modeling cognition, we wish to identify the model, from a set of candidate models, that generated the observed data. This is an ill-posed problem because information in the finite data sample is rarely sufficient to point to a single model. Rather, multiple models may provide equally good descriptions of the data. In statistics, this ill-posedness of model selection is overcome by reformulating the inference problem as one of making a best guess as to which model provides the closest approximation, in some defined sense, to the true but unknown model that generated the data. The particular measure of such an approximation, which is widely recognized among modelers in statistics and computer science, is *generalizability*. Generalizability, or predictive accuracy, refers to a model's ability to accurately

1. Strictly speaking, 'model' and 'model class' are not interchangeable. A model class consists of a collection of models in which each model represents a single probability distribution. In the present chapter, however, we will often use these terms interchangeably when the context makes it clear what we are referring to.

predict future, as yet unseen, data samples from the same process that generated the currently observed sample.

A formal definition of generalizability can be given in terms of a *discrepancy function* that measures the degree of approximation or similarity between two probability distributions. A discrepancy function $D(f, g)$ between two distributions, f and g , is some well-behaved function (e.g., Kullback-Leibler information divergence [Kullback and Leibler 1951]) that satisfies $D(f, g) > D(f, f) = 0$ for $f \neq g$. Generalizability could be defined as:

$$E^{f_T}[D(f_T, f_M)] \triangleq \int_{\mathcal{X}} D(f_T(x), f_M(\hat{\theta}(x)))f_T(x)dx \quad (3.1)$$

where f_T and f_M denote the probability distributions of the true and guessing models and $\hat{\theta}(x)$ is the Maximum Likelihood (ML) estimate of the parameter. According to the above equation, generalizability is a mean discrepancy between the true model and the best-fitting member of the model class of interest, averaged across all possible data that could be observed under the true model. The basic tenet of model selection is that among a set of competing model classes, one should select the one that optimizes generalizability (i.e., minimizes the quantity in equation (3.1)). However, generalizability is not directly observable and instead, one must estimate the measure from a data sample by considering the characteristics of the model class under investigation.

Several generalizability estimates have been proposed. They include Bayesian Information Criterion (BIC) [Schwarz 1978] and Cross Validation (CV) [Stone 1974]. In BIC, generalizability is estimated by trading off a model's *goodness of fit* with *model complexity*. Goodness of fit refers to how well a model fits the particular data set whereas model complexity or flexibility refers to a model's ability to fit arbitrary patterns of data. The BIC criterion, which is derived as an asymptotic approximation of a quantity related to the Bayes factor [Kass and Raftery 1995], is defined as

$$\text{BIC} \triangleq -\log f_{X|\Theta}(x|\hat{\theta}) + \frac{k}{2} \log(n)$$

where $\log(\cdot)$ is the natural logarithm function of base e , k is the dimension of the parameter and n is the sample size. The first term represents a goodness of fit measure and the second term represents a complexity measure. From the BIC viewpoint, the number of parameter (k) and the sample size (n) are the only relevant facets of complexity. BIC, however, ignores another important facet of model complexity, namely, the functional form of the model equation [Myung and Pitt 1997]. Functional form refers to the way in which the model's parameters are combined to define the model equation. For example, two models, $x = at + b$ and $x = at^b$, have the same number of parameters but differ in functional form.

CV is an easy-to-use, sampling-based method of estimating a model's generalizability. In CV, the data are split into two samples, the calibration sample and the validation sample. The model of interest is fitted to the calibration sample and

the best-fit parameter values are obtained. With these values fixed, the model is fitted again, this time to the validation sample. The resulting fit defines the model's generalizability estimate. Note that this estimation is done without an explicit consideration of complexity. Unlike BIC, CV takes into account functional form as well as the number of parameters, but given the implicit nature of CV, it is not clear how this is achieved.

The principle of Minimum Description Length (MDL) [Barron, Rissanen, and Yu 1998; Grünwald 2000; Hansen and Yu 2001; Rissanen 1989; Rissanen 1996; Rissanen 2001], which was developed within the domain of algorithmic coding theory in computer science [Li and Vitanyi 1997], represents a new conceptualization of the model selection problem. In MDL, both models and data are viewed as codes that can be compressed, and the goal of model selection is to choose the model class that permits the greatest compression of data in its description.² The shortest code length obtainable with the help of a given model class is called the *stochastic complexity* of the model class. In the present study we focus on two implementations of the stochastic complexity, a *Fisher Information Approximated normalized maximum likelihood* (FIA) [Rissanen 1996] and *Normalized Maximum Likelihood* (NML) [Rissanen 1996; Rissanen 2001]. Each, as an analytic realization of Occam's Razor, combines measures of goodness of fit and model complexity in a way that remedies the shortcomings of BIC and CV: Complexity is explicitly defined and functional form is included in the definition.

The purpose of the current chapter is threefold. The first is to address an issue that we have had to deal with in applying MDL to the selection of mathematical models of cognition [Lee 2002; Pitt, Myung, and Zhang 2002]. Calculation of the Fisher information can sometimes be sufficiently difficult to be a deterrent to using the measure. We walk the reader through the application of a straightforward and efficient formula for computing Fisher information for two broad classes of models in the field, those that have multinomial or independent normal distributions. Next we compare the relative performance of FIA and NML for one type of model in cognition. Finally, we present an example application of the MDL criteria in the selection of retention (i.e., memory) models in cognitive science. We begin by briefly reviewing the two MDL criteria, FIA and NML.

3.2 Recent Formulations of MDL

It is well established in statistics that choosing among a set of competing models based solely on goodness of fit can result in the selection of an unnecessarily complex model that over-fits the data [Pitt and Myung 2002]. The problem of over-fitting is mitigated by choosing models using a generalizability measure that strikes the

2. The code here refers to the probability distribution p of a random quantity. The code length is justified as $\log(1/p)$ from Shannon's information theory.

right balance between goodness of fit and model complexity. This is what both the FIA and NML criteria achieve. These two criteria are related to each other in that the former is obtained as an asymptotic approximation of the latter.

3.2.1 FIA Criterion

By considering Fisher information, Rissanen proposed the following model selection criterion [Rissanen 1996]

$$\text{FIA} \triangleq -\log f_{X|\Theta}(x|\hat{\theta}) + C_{\text{FIA}}$$

with

$$C_{\text{FIA}} \triangleq \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Omega} \sqrt{|I(\theta)|} d\theta \quad (3.2)$$

where Ω is the parameter space on which the model class is defined and $I(\theta)$ is the Fisher information of sample size one given by [Schervish 1995]

$$I_{i,j}(\theta) = -E_{\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{X|\Theta}(x|\theta) \right]$$

where E_{θ} denotes the expectation over data space under the model class $f_{X|\Theta}(x|\theta)$ given a θ value. In terms of coding theory, the value of FIA represents the length in “ebits” of the shortest description of the data the model class can provide. According to the principle of minimum description length, the model class that minimizes FIA extracts the most regularity in the data and therefore is to be preferred.

Inspection of C_{FIA} in equation (3.2) reveals four discernible facets of model complexity: the number of parameters k , sample size n , parameter range Ω , and the functional form of the model equation as implied in $I(\theta)$. Their contributions to model complexity can be summarized in terms of three observations. Firstly, C_{FIA} consists of two additive terms. The first term captures the number of parameters and the second term captures the functional form through the Fisher information matrix $I(\theta)$. Notice the sample size, n , appears only in the first term; 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 C_{FIA} to the complexity penalty of BIC. Secondly, 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. Finally, the calculation of the second term depends on the parameter range on which the integration of a nonnegative quantity in the parameter space is required. As such, the greater the ranges of the parameters, the larger the value of the integral, and therefore the more complex the model.

Regarding the calculation of the second term of C_{FIA} , there are at least two nontrivial challenges to overcome: Integration over multidimensional parameter space and calculation of the Fisher information matrix. It is in general not possible to obtain a closed-form solution of the integral. Instead, the solution must be

sought using a numerical integration method such as Markov Chain Monte Carlo [Gilks, Richardson, and Spiegelhalter 1996]. Second, with partial derivatives and expectation in the definition of Fisher information, direct element-by-element hand calculation of the Fisher information matrix can be a daunting task. This is because the number of elements in the Fisher information matrix is the square of the dimension of the parameter. For example, with 100 dimensional parameter, we need to find 10,000 elements of the Fisher information matrix, which would be quite a chore. Efficient computation of Fisher information is a significant hurdle in the application of MDL to cognitive modeling [Pitt, Myung, and Zhang 2002]. This chapter presents a method to overcome it. In section 3.3, we provide a simple algebraic formula for the Fisher information matrix that does not require the cumbersome element-by-element calculations and expectations.

3.2.2 NML Criterion

FIA represents important progress in understanding and formalizing model selection. It was, however, derived as a second-order limiting solution to the problem of finding the ideal code length [Rissanen 1996], with the higher order terms being left off. Rissanen further refined the solution by reformulating the ideal-code-length problem as a minimax problem in information theory [Rissanen 2001].³ The basic idea of this new approach is to identify a single probability distribution that is “universally” representative of an entire model class of probability distributions in the sense that the desired distribution mimics the behavior of any member of that class [Barron, Rissanen, and Yu 1998]. Specifically, the resulting solution to the minimax problem represents the optimal probability distribution that can encode data with the minimum mean code length subject to the restrictions of the model class of interest.

The minimax problem is defined as finding a probability distribution or code p^* such that

$$p^* \triangleq \operatorname{arg\,inf}_p \sup_q \mathbb{E}^q \left[\log \frac{f_{X|\Theta}(x|\hat{\theta})}{p(x)} \right]$$

where p and q range over the set of all distributions satisfying certain regularity conditions [Rissanen 2001], q is the data generating distribution (i.e., true model), $\mathbb{E}^q[\cdot]$ is the expectation with respect to the distribution q , and $\hat{\theta}$ is the ML estimate of the parameter. Given a model class of probability distributions $f_{X|\Theta}(x|\theta)$, the minimax problem is to identify one probability distribution p^* that minimizes the mean difference in code length between the desired distribution and the best-fitting member of the model class where the mean is taken with respect to the worst case scenario. The data generating distribution q does not have to be a member

3. Although formal proofs of the NML criterion were presented in [Rissanen 2001], its preliminary ideas were already discussed in [Rissanen 1996]

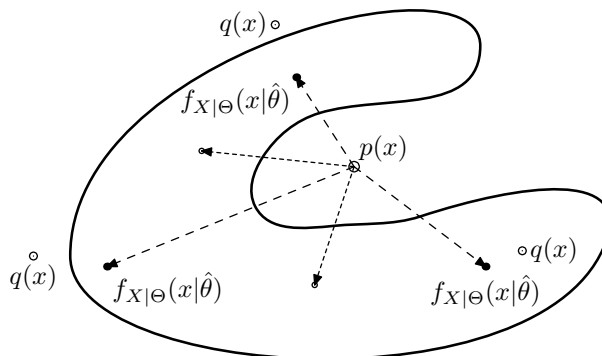


Figure 3.1 The minimax problem in a model manifold.

of the model class. In other words, the model class need not be correctly specified. Similarly, the desired probability distribution, as a solution to the minimax problem, does not have to be a member of the model class. The intuitive account of the minimax problem is schematically illustrated in Figure 3.1.

The solution to the minimax problem [Shtarkov 1987] is given by

$$p^* = \frac{f_{X|\Theta}(x|\hat{\theta})}{C_f}$$

where

$$C_f \triangleq \int_{\hat{\theta}(x) \in \Omega} f_{X|\Theta}(x|\hat{\theta}(x)) dx.$$

Note that p^* is the maximum likelihood of the current data sample divided by the sum of maximum likelihoods over all possible data samples. As such, p^* is called the Normalized Maximum Likelihood (NML) distribution, which generalizes the notion of ML. Recall that ML is employed to identify the parameter values that optimize the likelihood function (i.e., goodness of fit) within a given model class. Likewise, NML is employed to identify the model class, among a set of competing model classes, that optimizes generalizability.

C_f , the normalizing constant of the NML distribution, represents a complexity measure of the model class. It denotes the *sum of all best fits* the model class can provide collectively. Complexity is positively related to this value. The larger the sum of the model class, the more complex the model is. As such, this quantity formalizes the intuitive notion of complexity often referred to as a model's ability to fit diverse patterns of data [Myung and Pitt 1997] or as the “number” of different data patterns the model can fit well (i.e., flexibility) [Myung, Balasubramanian, and Pitt 2000]. It turns out that the logarithm of C_f is equal to the minimax value of the mean code length difference obtained when the NML distribution happens to be the data generating distribution [Rissanen 2001]. In other words, another interpretation of C_f is that it is the minimized worst prediction error the model class makes for the data generated from the NML distribution.

The desired selection criterion, NML, is defined as the code length of the NML distribution p^* (i.e., $-\log(p^*)$)

$$\text{NML} \triangleq -\log f_{X|\Theta}(x|\hat{\theta}) + C_{\text{NML}}$$

where

$$C_{\text{NML}} \triangleq \log \int_{\hat{\theta}(x) \in \Omega} f_{X|\Theta}(x|\hat{\theta}(x)) dx. \quad (3.3)$$

As presented above, the difference between FIA and NML is in their complexity measure, C_{FIA} and C_{NML} . Since C_{FIA} can be obtained from a Taylor series expansion of C_{NML} under the assumption of large sample sizes, C_{NML} captures the full scope of model complexity, thereby being a more complete quantification of model complexity. Like C_{FIA} , C_{NML} is also non-trivial to compute. They both require evaluation of an integral, though different kinds: integration over the parameter space in C_{FIA} and integration over the data (sample) space in C_{NML} . In the next section, we provide an easy-to-use formula to calculate the Fisher information matrix when computing C_{FIA} . Calculation of C_{NML} is more challenging, and requires two steps of heavy-duty computation: (step 1) maximization of the likelihood function, given a data sample, over the parameter space on which the model class is defined; and (step 2) integration of the maximized likelihood over the entire data space. In practice, the first step of parameter estimation is mostly done numerically, which is tricky because of the local maxima problem. The second step is even harder as sample space is usually of much higher dimension than parameter space. Another goal of the present investigation is to compare these two complexity measures for specific models of cognition to examine the similarity of their answers (see Section 3.4).

3.3 Fisher Information

As discussed in section 3.2.1, a major challenge of applying FIA is to compute the Fisher information matrix $I(\theta)$, especially when the dimension of the parameter space is large. Although the standard formula for the Fisher information matrix has been known in the literature, it is often presented implicitly without the detail of its derivation. In this section, we show its derivation in detail and provide a unified, easy-to-use formula to compute it for a model having an arbitrary dimensional parameter defined in terms of multinomial or independent normal distributions—the two most commonly assumed distributions in cognitive modeling. The resulting formula, which is obtained under simplifying assumptions, greatly eases the computation of the Fisher information by eliminating the need for both numerical expectation and second-order differentiation of the likelihood function. We also demonstrate the application of this formula in four areas of cognitive modeling: categorization, information integration, retention and psychophysics.

3.3.1 Models with Multinomial Distribution

We begin by defining the notation. Consider the model $f_{X|\Theta}(x|\theta)$ with a multinomial distribution. The parameter $\theta = [\theta_1, \dots, \theta_K]^T$, and $X|\Theta = [X_1|\Theta, \dots, X_N|\Theta]^T$. It is assumed that $\{X_n|\Theta\}$ is independent and each follows a multinomial distribution with C categories and sample size n' , i.e., $X_n|\Theta \sim \text{Mult}_C(n', p_{n,1}(\theta), \dots, p_{n,C}(\theta))$. K is the parameter dimension number, N is the random vector dimension number and C is the number of categories. Different selection of $\{p_{n,c}(\theta)\}$ yields different model. So

$$f_{X_n|\Theta}(x_n|\theta) = \binom{n'}{x_{n,1}, \dots, x_{n,C}} \prod_{c=1}^C p_{n,c}(\theta)^{x_{n,c}}$$

with respect to a counting measure on $\{(x_{n,1}, \dots, x_{n,C}) : \sum_{c=1}^C x_{n,c} = n', x_{n,c} \in \{0, \dots, n'\}\}$. Since $\{X_n|\Theta\}$ is independent,

$$\begin{aligned} f_{X|\Theta}(x|\theta) &= \prod_{n=1}^N \binom{n'}{x_{n,1}, \dots, x_{n,C}} \prod_{c=1}^C p_{n,c}(\theta)^{x_{n,c}} \\ \log f_{X|\Theta}(x|\theta) &= \sum_{n=1}^N \left(\log \binom{n'}{x_{n,1}, \dots, x_{n,C}} + \sum_{c=1}^C x_{n,c} \log p_{n,c}(\theta) \right). \end{aligned}$$

The first and second derivatives of the log likelihood function are then calculated as

$$\begin{aligned} \frac{\partial \log f_{X|\Theta}(x|\theta)}{\partial \theta_i} &= \sum_{n=1}^N \sum_{c=1}^C \frac{x_{n,c}}{p_{n,c}(\theta)} \frac{\partial p_{n,c}(\theta)}{\partial \theta_i} \\ \frac{\partial^2 \log f_{X|\Theta}(x|\theta)}{\partial \theta_i \partial \theta_j} &= \sum_{n=1}^N \sum_{c=1}^C \frac{-x_{n,c}}{p_{n,c}^2(\theta)} \frac{\partial p_{n,c}(\theta)}{\partial \theta_i} \frac{\partial p_{n,c}(\theta)}{\partial \theta_j} + \frac{x_{n,c}}{p_{n,c}(\theta)} \frac{\partial^2 p_{n,c}(\theta)}{\partial \theta_i \partial \theta_j}. \end{aligned}$$

With the regularity conditions [Schervish 1995, p. 111] held for the model $f_{X|\Theta}(x|\theta)$ in question, we write one element of Fisher information matrix of sample size one as

$$\begin{aligned} I_{i,j}(\theta) &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{X|\Theta}(x|\theta) \right] \\ &= \sum_{n=1}^N \sum_{c=1}^C \frac{1}{p_{n,c}(\theta)} \frac{\partial p_{n,c}(\theta)}{\partial \theta_i} \frac{\partial p_{n,c}(\theta)}{\partial \theta_j} - \frac{\partial^2 p_{n,c}(\theta)}{\partial \theta_i \partial \theta_j} \\ &= \sum_{n=1}^N \sum_{c=1}^C \frac{1}{p_{n,c}(\theta)} \frac{\partial p_{n,c}(\theta)}{\partial \theta_i} \frac{\partial p_{n,c}(\theta)}{\partial \theta_j} \\ &= \sum_{l=1}^{NC} \frac{1}{p_l(\theta)} \frac{\partial p_l(\theta)}{\partial \theta_i} \frac{\partial p_l(\theta)}{\partial \theta_j} \end{aligned}$$

where $p_l(\theta) \triangleq p_{n,c}(\theta)$ and $l = (n-1)C + c$. Rewriting the above result in matrix

form yields

$$I(\theta) = P^T \Lambda^{-1} P \quad (3.4)$$

with

$$P \triangleq \frac{\partial (p_1(\theta), \dots, p_{NC}(\theta))}{\partial (\theta_1, \dots, \theta_K)}$$

$$\Lambda \triangleq \text{diag}(p_1(\theta), \dots, p_{NC}(\theta))$$

where $\frac{\partial (p_1(\theta), \dots, p_{NC}(\theta))}{\partial (\theta_1, \dots, \theta_K)} = [P_{i,j}]$ is the $NC \times K$ Jacobian matrix with $P_{i,j} = \frac{\partial p_i(\theta)}{\partial \theta_j}$ and Λ is the diagonal matrix with $p_1(\theta), \dots, p_{NC}(\theta)$ as the diagonal elements. With the formula (3.4), to calculate $I(\theta)$, we need to evaluate only two sets of functions: $\{p_i(\theta)\}$, which are nothing but the model equations; and the first derivatives $\{\frac{\partial p_i(\theta)}{\partial \theta_k}\}$, which can easily be determined in analytic form.

Interestingly enough, the above formula (3.4) looks strikingly similar to the well-known formula for the re-parameterization of Fisher information [Schervish 1995, p. 115]. There is, however, an important difference between the two. The formula in (3.4) tells us how to calculate Fisher information for a given model in which the Jacobian matrix P is in general non-square. On the other hand, the re-parameterization formula reveals how to relate Fisher information from one parameterization to another, once the Fisher information has been obtained with a given parameterization. Since the two parameterizations are related to each other through a one-to-one transformation, the corresponding Jacobian matrix P is always square for re-parameterization.

In the following, we demonstrate application of the formula (3.4) for models of categorization, information integration, and retention.

3.3.1.1 Categorization

Categorization is the cognitive operation by which we identify an object or thing as a member of a particular group, called a category. We categorize a robin as a bird, a German Shepherd as a dog, and both as mammals. Since no two dogs are exactly alike, categorization helps us avoid being overwhelmed by the sheer detail of the environment and the accompanying mental operation that would otherwise be required to represent every incident we encounter as a unique event [Glass and Holyoak 1986]. Without categorization, the world would appear to us as an incidental collection of unrelated events. Further, categorization helps us make inferences about an object that has been assigned to a category. For example, having categorized a moving vehicle as a tank, we can infer that it has all-terrain, is armored, is armed with a cannon gun mounted inside a rotating turret, and can damage the road.

In mathematical modeling of categorization, an object, often called a category exemplar, is represented as a point in a multidimensional psychological space in which the value of each coordinate represents the magnitude or presence/absence

of an attribute such as height, weight, color, whether it is an animal or not, etc. In a typical categorization experiment, participants are asked to categorize a series of stimuli, presented on computer screen, into one or more pre-defined categories. The stimuli are generated from a factorial manipulation of two or more stimulus attributes. As a example, we illustrate the application of (3.4) for the Generalized Context Model (GCM) [Nosofsky 1986].

According to the GCM model, category decisions are made based on a similarity comparison between the input stimulus and stored exemplars of a given category. Specifically, this model requires no specific restrictions on K , N and C in computing complexity, and assumes that the probability of choosing category c in response to input stimulus n is given by

$$p_{n,c} = \frac{\sum_{m \in C_c} s_{nm}}{\sum_q \sum_{p \in C_q} s_{np}}$$

where C_c is the set of all indexes of the prototype stimuli in category c and

$$s_{ij} = \exp \left(-s \cdot \left(\sum_{t=1}^{K-1} w_t |x_{it} - x_{jt}|^r \right)^{1/r} \right)$$

In the above equation, s_{ij} is a similarity measure between multidimensional stimuli i and j , s (> 0) is a sensitivity or scaling parameter, w_t is a non-negative attention weight given to attribute t satisfying $\sum_{t=1}^{K-1} w_t = 1$, and x_{it} is the t -th coordinate value of stimulus i . According to the above equation, similarity between two stimuli is assumed to be an exponentially decreasing function of their distance, which is measured by the Minkowski metric with metric parameter r (≥ 1). Note that the parameter θ consists of $\theta = [\theta_1, \dots, \theta_K]^T \triangleq [w_1, \dots, w_{K-2}, s, r]^T$.

The first derivatives of the model equation are computed as

$$\frac{\partial p_{n,c}}{\partial \theta_k} = \frac{\left(\sum_{m \in C_c} \frac{\partial s_{nm}}{\partial \theta_k} \right) \left(\sum_q \sum_{p \in C_q} s_{np} \right) - \left(\sum_{m \in C_c} s_{nm} \right) \left(\sum_q \sum_{p \in C_q} \frac{\partial s_{np}}{\partial \theta_k} \right)}{\left(\sum_q \sum_{p \in C_q} s_{np} \right)^2}$$

with

$$\frac{\partial s_{ij}}{\partial \theta_k} = \begin{cases} s_{ij} \cdot \frac{-s}{r} \cdot T_{ij}^{\frac{1-r}{r}} \cdot (|x_{ik} - x_{jk}|^r - |x_{iK-1} - x_{jK-1}|^r) & k = 1, \dots, K-2 \\ s_{ij} \cdot -T_{ij}^{\frac{1}{r}} & k = K-1 \\ s_{ij} \cdot \log s_{ij} \cdot \left(\frac{-1}{r^2} \cdot \log T_{ij} + \frac{\sum_{t=1}^{K-2} w_t |x_{it} - x_{jt}|^r \log |x_{it} - x_{jt}|}{r T_{ij}} \right) & k = K \end{cases}$$

where $T_{ij} \triangleq \sum_{t=1}^{K-1} w_t |x_{it} - x_{jt}|^r$.

Using $l = (n-1)C + c$, we can easily obtain $\{p_l(\theta)\}$ and $\{\frac{\partial p_l(\theta)}{\partial \theta_k}\}$ from $\{p_{n,c}(\theta)\}$

and $\{\frac{\partial p_{n,c}(\theta)}{\partial \theta_k}\}$ derived above. Plugging these into equation (3.4) yields the desired Fisher information matrix.

3.3.1.2 Information Integration

Models of information integration are concerned with how information from independent sources (e.g., sensory and contextual) are combined during perceptual identification. For example, in phonemic identification we might be interested in how an input stimulus is perceived as /ba/ or /da/ based on the cues presented in one or two modalities (e.g., auditory only or auditory plus visual). In a typical information integration experiment, participants are asked to identify stimuli that are factorially manipulated along two or more stimulus dimensions.

Information integration models represent a slightly more restricted case compared to categorization models. As with categorization models, the stimulus is represented as a vector in multidimensional space. K is the sum of all stimulus dimensions and N is the product of all stimulus dimensions. To illustrate the application of (3.4), consider models for a two-factor experiment (e.g., FLMP, LIM)⁴. For such models, the response probability p_{ij} of classifying an input stimulus specified by stimulus dimensions i and j as /ba/ versus /da/ can be written as

$$p_{ij} = h(\theta_i, \lambda_j) \quad (3.5)$$

where θ_i and λ_j ($i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$) are parameters representing the strength of the corresponding feature dimensions. So we have $K = I + J$, $N = I \cdot J$ and $C = 2$. With the restriction on C , the above model assumes the binomial probability distribution, which is a special case of multinomial distribution so formula (3.4) can still be used. Now, we further simplify the desired Fisher information matrix by taking into account $C = 2$ and $\sum_{c=1}^C p_{n,c} = 1$ as follows:

$$\begin{aligned} I_{i,j}(\theta) &= \sum_{n=1}^N \sum_{c=1}^C \frac{1}{p_{n,c}(\theta)} \frac{\partial p_{n,c}(\theta)}{\partial \theta_i} \frac{\partial p_{n,c}(\theta)}{\partial \theta_j} \\ &= \sum_{n=1}^N \frac{1}{p_n(\theta)(1-p_n(\theta))} \frac{\partial p_n(\theta)}{\partial \theta_i} \frac{\partial p_n(\theta)}{\partial \theta_j} \end{aligned}$$

with $p_n(\theta) \triangleq p_{n,1}(\theta) = p_{ij}(\theta)$, $n = (i-1)J + j$ and $\theta = [\theta_1, \theta_2, \dots, \theta_{I+J}]^T \triangleq [\theta_1, \theta_2, \dots, \theta_I, \lambda_1, \lambda_2, \dots, \lambda_J]^T$. So

$$I(\theta) = B^T \Delta^{-1} B \quad (3.6)$$

4. Fuzzy Logical Model of Perception [Oden and Massaro 1978], Linear Integration Model [Anderson 1981].

with

$$B \triangleq \frac{\partial (p_1(\theta), \dots, p_{I \cdot J}(\theta))}{\partial (\theta_1, \dots, \theta_{I+J})}$$

$$\Delta \triangleq \text{diag}(p_1(\theta)(1 - p_1(\theta)), \dots, p_{I \cdot J}(\theta)(1 - p_{I \cdot J}(\theta))).$$

It is worth noting that the number of diagonal elements of $\Delta \in \mathbb{R}^{N \times N}$ is one half of that of Λ in (3.4).

Applying the above results to the generic model equation (3.5), we note that since $n = (i-1)J + j$, $i \in \{1, \dots, I\}$ and $j \in \{1, \dots, J\}$, we have $i = \lfloor (n-1)/J \rfloor + 1$, $j = n - J \lfloor (n-1)/J \rfloor$. So

$$p_n(\theta) = h(\theta_{\lfloor (n-1)/J \rfloor + 1}, \theta_{n - J \lfloor (n-1)/J \rfloor})$$

$$\frac{\partial p_n(\theta)}{\partial \theta_k} = \frac{\partial h(\theta_{\lfloor (n-1)/J \rfloor + 1}, \theta_{n - J \lfloor (n-1)/J \rfloor})}{\partial \theta_k}$$

where $\lfloor x \rfloor \triangleq \max\{n \in \mathbb{Z} : n \leq x\}$.

3.3.1.3 Retention

Retention refers to the mental ability to retain information about learned events over time. In a typical experimental setup, participants are presented with a list of items (e.g., words or nonsense syllables) to study, and afterwards are asked to recall or recognize them at varying time delays since study. Of course, the longer the interval between the time of stimulus presentation and the time of later recollection, the less likely the event will be remembered. Therefore, the probability of retaining in memory an item after time t is a monotonically decreasing function of t .

Models of retention are concerned with the specific form of the rate at which information retention drops (i.e., forgetting occurs) [Rubin and Wenzel 1996; Wickens 1998]. For instance, the exponential model assumes that the retention probability follows the form $h(a, b, t) = ae^{-bt}$ with the parameter $\theta = (a, b)$ whereas the power model assumes $h(a, b, t) = at^{-b}$. For such two-parameter models, we have the parameter dimension $K = 2$ and the number of categories $C = 2$. It is then straightforward to show in this case that

$$|I(\theta)| = \sum_{n, l=1; n < l}^N \frac{\left(\frac{\partial p_n(\theta)}{\partial \theta_1} \frac{\partial p_l(\theta)}{\partial \theta_2} - \frac{\partial p_n(\theta)}{\partial \theta_2} \frac{\partial p_l(\theta)}{\partial \theta_1} \right)^2}{p_n(\theta)(1 - p_n(\theta))p_l(\theta)(1 - p_l(\theta))} \quad (3.7)$$

which reduces to the previous result in Appendix A of [Pitt, Myung, and Zhang 2002] for $K = 1$,

$$|I(\theta)| = \sum_{n=1}^N \frac{\left(\frac{dp_n(\theta)}{d\theta} \right)^2}{p_n(\theta)(1 - p_n(\theta))} \quad (3.8)$$

Close inspection and comparison of (3.7) and (3.8) strongly suggests the following form of Fisher information for the general case of $K \geq 1$,

$$|I(\theta)| = \sum_{n_1 < n_2 < \dots < n_K = 1}^N \frac{\left| \frac{\partial(p_{n_1}(\theta), p_{n_2}(\theta), \dots, p_{n_K}(\theta))}{\partial(\theta_1, \theta_2, \dots, \theta_K)} \right|^2}{\prod_{k=1}^K p_{n_k}(\theta)(1 - p_{n_k}(\theta))} \quad (3.9)$$

The above expression, though elegant, is a conjecture whose validity has yet to be proven (e.g., by induction). On the other hand, one might find it computationally more efficient to use the original formula (3.6), rather than (3.9).

3.3.2 Models with Normal Distribution

For models with independent normal distribution, the form of the Fisher information formula turns out to be similar to that of the multinomial distribution, as shown in the following.

Look at a model $f_{X|\Theta}(x|\theta)$ with normal distribution. The parameter $\theta = [\theta_1, \theta_2, \dots, \theta_K]^T$, $X|\Theta = [X_1|\Theta, X_2|\Theta, \dots, X_N|\Theta]^T$, and $X|\Theta \sim \mathcal{N}_N(\mu(\theta), \sigma(\theta))$ with $\mu \in \mathbb{R}^N$ and $\sigma \in \mathbb{R}^{N \times N}$. Different choices of $\mu(\theta)$ and $\sigma(\theta)$ correspond to defining different models. Since $X|\Theta \sim \mathcal{N}_N(\mu(\theta), \sigma(\theta))$, we have

$$f_{X|\Theta}(x|\theta) = (2\pi)^{-\frac{N}{2}} |\sigma(\theta)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu(\theta))^T \sigma(\theta)^{-1} (x - \mu(\theta))\right)$$

with respect to the Lebesgue measure on \mathbb{R}^N . The general expression of Fisher information is then obtained as

$$I_{i,j}(\theta) = \frac{1}{2} \left(\frac{\partial^2 \log |\sigma(\theta)|}{\partial \theta_i \partial \theta_j} + \sum_{m,n=1}^N \sigma_{m,n}(\theta) \frac{\partial^2 (\sigma(\theta)^{-1})_{m,n}}{\partial \theta_i \partial \theta_j} \right) + \frac{\partial \mu(\theta)^T}{\partial \theta_i} \sigma(\theta)^{-1} \frac{\partial \mu(\theta)}{\partial \theta_j}.$$

The above equation is derived without any supposition of independence. With the assumption that $\{X_n|\Theta\}$ is independent (i.e., σ is diagonal with the diagonal element not necessarily equal), the above result can be further simplified to

$$I_{i,j}(\theta) = \frac{\partial \mu(\theta)^T}{\partial \theta_i} \sigma(\theta)^{-1} \frac{\partial \mu(\theta)}{\partial \theta_j} + \sum_{n=1}^N \frac{1}{2\sigma_{n,n}(\theta)^2} \frac{\partial \sigma_{n,n}(\theta)}{\partial \theta_i} \frac{\partial \sigma_{n,n}(\theta)}{\partial \theta_j}.$$

The desired Fisher information matrix is then expressed in matrix form as

$$I(\theta) = P^T \Lambda^{-1} P \quad (3.10)$$

with

$$P \triangleq \frac{\partial(\mu_1(\theta), \dots, \mu_N(\theta), \sigma_{1,1}(\theta), \dots, \sigma_{N,N}(\theta))}{\partial(\theta_1, \dots, \theta_K)}$$

$$\Lambda \triangleq \text{diag}(\sigma_{1,1}(\theta), \dots, \sigma_{N,N}(\theta), 2\sigma_{1,1}(\theta)^2, \dots, 2\sigma_{N,N}(\theta)^2)$$

where P is the $2N \times K$ Jacobian matrix and Λ is diagonal matrix. Therefore, we have obtained a computation formula for Fisher information entirely in terms of the mean vector μ and covariance matrix σ for normally-distributed models. Note the similarity between the result in (3.10) and that in (3.4).

To demonstrate the application of (3.10), consider Fechner's logarithmic model of psychophysics [Roberts 1979],

$$X = \theta_1 \log(Y + \theta_2) + E$$

where $X = [x_1, \dots, x_N]^T \in \mathbb{R}^N$ is the data sample, $Y = [y_1, \dots, y_N]^T \in \mathbb{R}^N$ is a vector of independent variables, $\theta = [\theta_1, \theta_2]^T \in \mathbb{R}^2$ is the parameter, and $E \sim \mathcal{N}(0, c)$ is random error with constant variance $c \in \mathbb{R}$. So we have $X|\Theta \sim \mathcal{N}_N(\mu(\theta), \sigma(\theta))$, $\mu(\theta) = \theta_1 \log(Y + \theta_2)$, and $\sigma(\theta) = cI_N$, where I_N denotes the identity matrix and is not to be confused with $I(\theta)$, the Fisher information matrix. Using the formula (3.10), the Fisher information matrix is obtained as

$$\begin{aligned} I(\theta) &= P^T \Lambda^{-1} P \\ &= \begin{bmatrix} \log(Y + \theta_2) & \frac{\theta_1}{Y + \theta_2} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^T \cdot \begin{bmatrix} cI_N & \mathbf{0} \\ \mathbf{0} & 2c^2 I_N \end{bmatrix}^{-1} \cdot \begin{bmatrix} \log(Y + \theta_2) & \frac{\theta_1}{Y + \theta_2} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &= \frac{1}{c} \begin{bmatrix} \sum_{n=1}^N (\log(y_n + \theta_2))^2 & \sum_{n=1}^N \theta_1 \frac{\log(y_n + \theta_2)}{y_n + \theta_2} \\ \sum_{n=1}^N \theta_1 \frac{\log(y_n + \theta_2)}{y_n + \theta_2} & \sum_{n=1}^N \frac{\theta_1^2}{(y_n + \theta_2)^2} \end{bmatrix} \end{aligned}$$

where $\mathbf{0}$ denotes the null matrix of appropriate dimensions. Comparison of the above derivation with one obtained element by element in [Pitt, Myung, and Zhang 2002] nicely illustrates why this method of computing Fisher information is preferable.

3.4 MDL Complexity Comparison

As discussed in section 3.2, the two model selection criteria of FIA and NML differ only in their complexity measure, C_{FIA} and C_{NML} . With the formula of Fisher information derived in section 3.3, the computation of C_{FIA} becomes routine work. On the other hand, C_{NML} is more challenging to calculate and an efficient computation of this quantity has yet to be devised. In certain situations, however, it turns out that one can obtain analytic-form solutions of C_{FIA} and C_{NML} . Taking advantage of these instances, we compare and contrast the two to gain further insight into how they are related to each other.

In demonstrating the relationship between C_{FIA} and C_{NML} , we consider a saturated model with a multinomial distribution (for related work, see [Kontkanen, Bun-

tine, Myllymki, Rissanen, and Tirri 2003]). The data under this model are assumed to be a C -tuple random vector $X|\Theta \sim \text{Mult}_C(n, \theta_1, \dots, \theta_C)$ and $\theta = [\theta_1, \dots, \theta_{C-1}]^T$ is the parameter.

3.4.1 Complexity of C_{FIA}

The complexity penalty term of FIA is again given by

$$C_{\text{FIA}} \triangleq \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Omega} \sqrt{|I(\theta)|} d\theta.$$

For the saturated multinomial model, the dimension of the parameter $k = C-1$ with $\Omega = \{(\theta_1, \theta_2, \dots, \theta_{C-1}) : \theta_c \geq 0 \forall c, \sum_{c=1}^{C-1} \theta_c \leq 1\}$. Using the formula (3.4) of Fisher information for multinomial distribution, we have

$$\begin{aligned} I(\theta) &= \frac{\partial(\theta_1, \dots, \theta_C)}{\partial(\theta_1, \dots, \theta_{C-1})}^T \cdot \text{diag}(\theta_1^{-1}, \dots, \theta_{C-1}^{-1}) \cdot \frac{\partial(\theta_1, \dots, \theta_C)}{\partial(\theta_1, \dots, \theta_{C-1})} \\ &= \left[I_{C-1} \mid -\mathbf{1}_{(C-1) \times 1} \right] \cdot \left[\begin{array}{c|c} A & \mathbf{0} \\ \hline \mathbf{0} & \theta_C^{-1} \end{array} \right] \cdot \left[\begin{array}{c} I_{C-1} \\ \hline -\mathbf{1}_{1 \times (C-1)} \end{array} \right] \\ &= A + \mathbf{1}_{(C-1) \times (C-1)} \cdot \theta_C^{-1} \end{aligned}$$

where $\mathbf{1}_{n \times m}$ is $n \times m$ matrix with all elements equal to one, and $A = \text{diag}(\theta_1^{-1}, \dots, \theta_{C-1}^{-1})$. The determinant of $I(\theta)$ is then calculated as

$$\begin{aligned} |I(\theta)| &= \begin{vmatrix} \frac{1}{\theta_1} + \frac{1}{\theta_C} & \frac{1}{\theta_C} & \cdots & \frac{1}{\theta_C} \\ \frac{1}{\theta_C} & \frac{1}{\theta_2} + \frac{1}{\theta_C} & \cdots & \frac{1}{\theta_C} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\theta_C} & \frac{1}{\theta_C} & \cdots & \frac{1}{\theta_{C-1}} + \frac{1}{\theta_C} \end{vmatrix} \\ &= \begin{vmatrix} \frac{1}{\theta_1} & 0 & \cdots & 0 & \frac{-1}{\theta_{C-1}} \\ 0 & \frac{1}{\theta_C} & \cdots & 0 & \frac{-1}{\theta_{C-1}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\theta_{C-2}} & \frac{-1}{\theta_{C-1}} \\ \hline \frac{1}{\theta_C} & \frac{1}{\theta_C} & \cdots & \frac{1}{\theta_C} & \frac{1}{\theta_{C-1}} + \frac{1}{\theta_C} \end{vmatrix} \\ &= |B| \cdot \left| \frac{1}{\theta_{C-1}} + \frac{1}{\theta_C} - \mathbf{1}_{1 \times (C-2)} \cdot \frac{1}{\theta_C} \cdot B^{-1} \cdot -\mathbf{1}_{(C-2) \times 1} \cdot \frac{1}{\theta_{C-1}} \right| \\ &= \prod_{c=1}^C \frac{1}{\theta_c} \end{aligned}$$

where $B = \text{diag}(\theta_1^{-1}, \dots, \theta_{C-2}^{-1})$. Using Dirichlet integration, we find that

$$\begin{aligned} \int_{\Omega} \sqrt{|I(\theta)|} d\theta &= \int_{\Omega} \prod_{c=1}^C \theta_c^{-1/2} d\theta \\ &= \frac{\Gamma(1/2)^C}{\Gamma(C/2)} \end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function defined as $\Gamma(x) \triangleq \int_0^{\infty} t^{x-1} e^{-t} dt$ for $t > 0$. Finally, the desired complexity of the model is obtained as follows⁵

$$C_{\text{FIA}} = \frac{C-1}{2} \log \frac{n}{2\pi} + \log \left(\frac{\pi^{\frac{C}{2}}}{\Gamma(\frac{C}{2})} \right). \quad (3.11)$$

In the equation (3.11), C_{FIA} is no longer a linear function of the dimension of the parameter (i.e., $k = C - 1$) due to the Gamma function in the second term. This contrasts with BIC, where complexity is measured as a linear function of the dimension of the model parameter.

3.4.2 Complexity of C_{NML}

The following is the complexity penalty of NML

$$C_{\text{NML}} \triangleq \log \int_{\hat{\theta}(x) \in \Omega} f_{X|\Theta}(x|\hat{\theta}(x)) dx.$$

To obtain the exact expression for C_{NML} , we would need the analytical solution of $\hat{\theta}(x)$, which requires solving an optimization problem. The log likelihood function is given by

$$\log f_{X|\Theta}(x|\theta) = \log \binom{n}{x_1, \dots, x_C} + \sum_{c=1}^C x_c \log \theta_c.$$

The ML estimate that maximizes the above log likelihood function is found to be $\hat{\theta}_c = \frac{x_c}{n} \quad \forall c \in \{1, 2, \dots, C\}$. Plugging this result into the earlier equation, we obtain

$$C_{\text{NML}} = \log \left(\sum_{\substack{0 \leq x_c \leq n \\ x_1 + x_2 + \dots + x_C = n}} \binom{n}{x_1, \dots, x_C} \prod_{c=1}^C \left(\frac{x_c}{n} \right)^{x_c} \right). \quad (3.12)$$

C_{NML} can be calculated by considering all possible $\binom{n+C-1}{C-1}$ data patterns in the sample space for a fixed C and a sample size n . To do so, for each data pattern we would need to compute the multinomial coefficient and the multiplication of

5. The same result is also described in Rissanen [Rissanen 1996], and a more general one in Kontkanen et al [Kontkanen, Buntine, Myllymki, Rissanen, and Tirri 2003].

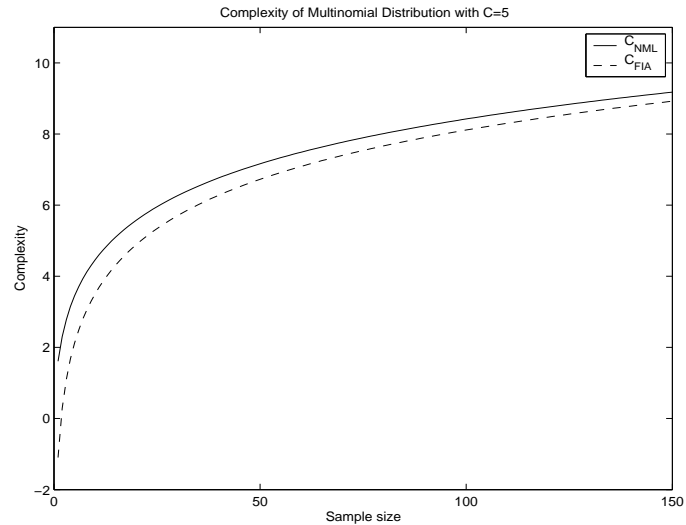


Figure 3.2 MDL complexity as a function of sample size.

C terms. There exists an elegant recursive algorithm based on combinatorics for doing this [Kontkanen, Buntine, Myllymki, Rissanen, and Tirri 2003]. Even so, (3.12) would still be computationally heavier than (3.11).

3.4.3 The Comparison

Shown in Figure 3.2 are plots of C_{FIA} and C_{NML} as a function of sample size n for the number of categories $C = 5$ (i.e., parameter dimension $K = 4$). As can be seen, the two curves follow each other closely, with C_{NML} being slightly larger than C_{FIA} . Both curves resemble the shape of a logarithmic function of n .

In Figure 3.3, we plot the two complexity measures now as a function of C for a fixed sample size $n = 50$. Again, C_{FIA} and C_{NML} are quite close and slightly convex in shape. This nonlinearity is obviously due to the functional form effects of model complexity. In contrast, the complexity measure of BIC, which ignore these effects, is a straight line. Interestingly however, the BIC complexity function provides a decent approximation of C_{FIA} and C_{NML} curves for $C \leq 4$.

To summarize, the *approximate* complexity measure C_{FIA} turns out to do a surprisingly good job of capturing the *full-solution* complexity measure C_{NML} , at least for the saturated multinomial model we examined.

3.5 Example Application

This section presents a model recovery simulation to demonstrate the relative performance of FIA and the other two selection criteria, BIC and CV. We chose three retention models with binomial distribution [Rubin and Wenzel 1996; Wickens

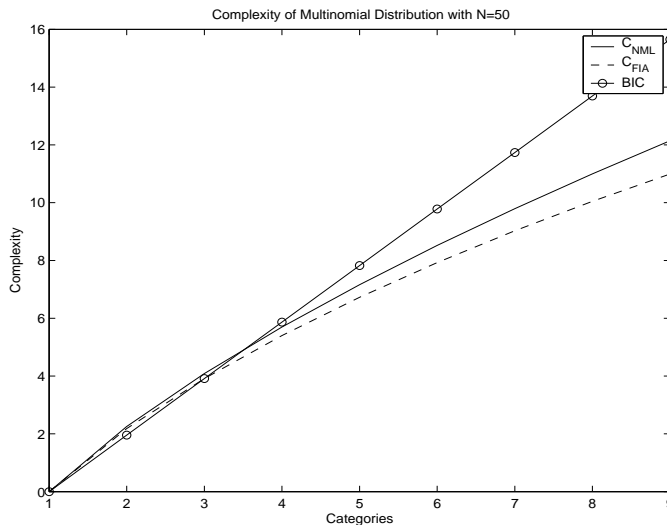


Figure 3.3 MDL complexity as a function of the number of categories.

1998] so $X_k|a, b]^T \sim \text{Bin}(n, h(a, b, t_k))$. The sample size $n = 20$, the independent variable t_k was selected to be $t_k = 1, 2, 4, 8, 16$ and the success probability $h(a, b, t)$ under each model was given by

$$h(a, b, t) = \begin{cases} 1/(1 + t^a) & \text{(M1)} \\ 1/(1 + a + bt) & \text{(M2)} \\ t^{-b}e^{-at} & \text{(M3)} \end{cases}$$

with the range of exponential parameter to be $[0, 10]$ and $[0, 100]$ otherwise.

The complexity measure C_{FIA} of each model was computed using (3.7) and was evaluated by simple Monte Carlo integration. Its value was 1.2361, 1.5479, and 1.7675 for models M1, M2, and M3, respectively. Model M1 is the simplest with one parameter whereas models M2 and M3 have two parameters, with their complexity difference of 0.2196 being due to the difference in functional form. For each model, we first generated 1,000 random parameter values sampled across the entire parameter space according to Jeffreys' prior [Robert 2001]. For each parameter, we then generated 100 simulated data samples with binomial sampling noise added. Finally, we fit all three models to each of 100,000 data samples and obtained their best-fitting parameter values. The three selection methods were compared on their ability to recover the data generating model. Maximum Likelihood (ML), a goodness of fit measure, was included as a baseline.

The results are presented in Table 3.1, which consists of four 3×3 sub-matrices, each corresponding to the selection method specified on the left. Within each sub-matrix, the value of each element indicates the percentage of samples in which the particular model's fit was preferred according to the selection method of interest. Ideally, a selection criterion should be able to recover the true model 100 percent of

Table 3.1 Model Recovery Rates of Three Retention Models.

Selection Method/ Fitted Model	Data Generating Model (C_{FIA})		
	M1 (1.2361)	M2 (1.5479)	M3 (1.7675)
ML			
M1	22%	11%	0%
M2	41%	88%	4%
M3	37%	1%	96%
BIC			
M1	91%	55%	8%
M2	4%	44%	4%
M3	5%	1%	88%
CV			
M1	52%	40%	7%
M2	28%	53%	19%
M3	20%	7%	74%
FIA			
M1	83%	37%	7%
M2	11%	62%	6%
M3	6%	1%	87%

the time, which would result in a diagonal matrix containing values of 100 percent. Deviations from this outcome indicate a bias in the selection method.

Let us first examine recovery performance of ML. The result in first column of the 3×3 sub-matrix indicates that model M1 was correctly recovered only 22% of the time; the rest of the time (78%) models M2 and M3 were selected incorrectly. This is not surprising because the latter two models are more complex than M1 (with one extra parameter). Hence such over-fitting is expected. This bias against M1 was mostly corrected when BIC was employed as a selection criterion, as shown in the first column of the corresponding 3×3 sub-matrix. On the other hand, the result in the second column for the data generated by M2 indicates that BIC had trouble distinguishing between M1 and M2. CV performed similarly as BIC, though its recovery rates (52%, 53%, 74%) are rather unimpressive. In contrast, FIA, with its results shown in the bottom sub-matrix, performed the best in recovering the data-generating model.⁶

6. A caveat here is that the above simulations are meant to be a demonstration, and as such the results are not to be taken as representative behavior of the three selection methods.

3.6 Summary and Conclusion

Model selection can proceed most confidently when a well-justified and well-performing measure of model complexity is available. C_{FIA} and C_{NML} of minimum description length are two such measures. In this chapter we addressed issues concerning the implementation of these measures in the context of models of cognition. As a main contribution of the present study, we provided a general formula in matrix form to calculate Fisher information. The formula is applicable for virtually all models that assume the multinomial distribution or the independent normal distribution—the two most common distributions in cognitive modeling.⁷ We also showed that C_{FIA} represents a good approximation to C_{NML} , at least for the saturated multinomial probability model. This finding suggests that within many research areas in cognitive science, modelers might use FIA instead of NML with minimal worry about whether the outcome would change if NML were used instead. Finally, we illustrated how MDL performs relative to its competitors in one content area of cognitive modeling.

Acknowledgments

This research was supported by the National Institute of Health Grant R01 MH57472. We wish to thank Peter Grünwald, Woojae Kim, Daniel Navarro, and an anonymous reviewer for many helpful comments.

7. The reader is cautioned that the formula should not be used blindly. In some cases, it might be more efficient to use a simpler formula. For example, instead of (3.4), sometimes it may be easier to use (3.7) for binomial models with a two dimensional parameter.

Bibliography

- Anderson, N. H. (1981). *Foundations of Information Integration Theory*. New York: Academic Press.
- Barron, A., J. Rissanen, and B. Yu (1998, October). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory* 44, 2743–2760.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Glass, A. L. and K. J. Holyoak (1986). *Cognition* (second ed.), Chapter 5. Random House.
- Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology* 44, 133–170.
- Hansen, M. H. and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96, 746–774.
- Kass, R. E. and A. E. Raftery (1995). Bayes factor. *Journal of the American Statistical Association* 90, 773–795.
- Kass, R. E. and P. W. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. John Wiley and Sons.
- Kontkanen, P., W. Buntine, P. Myllymki, J. Rissanen, and H. Tirri (2003). Efficient computation of stochastic complexity. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pp. 181–188. Edited by Christopher M. Bishop and Brendan J. Frey. Society for Artificial Intelligence and Statistics.
- Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics* 91, 79–86.
- Lee, M. D. (2002). Generating additive clustering models with minimal stochastic complexity. *Journal of Classification* 19, 69–85.
- Li, M. and P. Vitanyi (1997). *An Introduction to Kolmogorov Complexity and its Applications* (second ed.). New York: Springer.
- Myung, I. J., V. Balasubramanian, and M. A. Pitt (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences* 97, 11170–11175.

- Myung, I. J. and M. A. Pitt (1997). Applying occam's razor in modeling cognition: A bayesian approach. *Psychonomic Review & Bulletin* 4, 79–95.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115, 39–57.
- Oden, G. C. and D. W. Massaro (1978). Integration of featural information in speech perception. *Psychological Review* 85, 172–191.
- Pitt, M. A. and I. J. Myung (2002). When a good fit can be bad. *Trends in Cognitive Sciences* 6(10), 421–425.
- Pitt, M. A., I. J. Myung, and S. Zhang (2002). Toward a method of selecting among computational models of cognition. *Psychological Review* 109(3), 472–491.
- Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific.
- Rissanen, J. (1996, Jan). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* 42(1), 40–47.
- Rissanen, J. (2001, July). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory* 47(5), 1712–1717.
- Robert, C. P. (2001). *The Bayesian Choice* (second ed.). New York: Springer.
- Roberts, F. S. (1979). *Measurement Theory with Applications to Decision Making, Utility, and the Social Sciences*. Reading, MA: Addison-Wesley.
- Rubin, D. C. and A. E. Wenzel (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review* 103, 734–760.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.
- 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 [with discussion]. *Journal of Royal Statistical Society, Series B* 36, 111–147.
- Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel(1996): A qualitative description of retention. *Psychological Review* 105, 379–386.

Index

Bayesian Information Criterion, 11

cognitive modeling
 categorization, 18
 information integration, 20
 retention, 21

complexity
 model, 11

cross validation, 11

discrepancy function, 11

Fisher information, 16
 multinomial distribution, 17
 normal distribution, 22

generalizability, 10

goodness of fit, 11

minimax problem, 14

model selection, 10

NML, *see* Normalized Maximum Likelihood

Normalized Maximum Likelihood, 14
 approximate, 13

stochastic complexity, 12