

# Minimum Description Length in Cognitive Modeling (Extended Abstract)

Yong Su

February 28, 2003

Model selection can proceed confidently when complexity, a well justified intrinsic property of the model, is available. In this paper we first solve the problem of calculating Minimal Description Length 1996 (MDL1996[1]) complexity measure by deriving a formula to calculate Fisher information for the models with multinomial or independent normal distribution. The resulting formula is attractive as it greatly simplifies the computation of the Fisher information. We also illustrate the usage of the formula for several mathematical models in cognitive science. Then, in the further investigation of the newer MDL2001[2] criterion, we show that the two complexity measures of MDL1996 criterion and MDL2001 criterion are close to each other for a multinomial model. Finally, the application of MDL in cognitive science is illustrated in the selection of retention models. And the advantage of MDL is demonstrated by comparing the model selection performance of MDL and several other commonly used selection criteria. The following describe this three components of recent investigation in detail.

First, for multinomial distribution models, we show that the formula of Fisher information can be written in the form

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

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  $\{p_n(\theta)\}$  is the multinomial distribution parameter specified by a particular model,  $\frac{\partial (p_1(\theta), \dots, p_{NC}(\theta))}{\partial (\theta_1, \dots, \theta_K)} = [P_{i,j}]$  is the  $NC \times K$  Jacobian matrix with  $P_{i,j} = \frac{\partial p_i(\theta)}{\partial \theta_j}$  and  $\Lambda$  is the diagonal matrix with  $p_1(\theta), \dots, p_{NC}(\theta)$  as the diagonal elements. By placing restrictions on the parameter dimension number  $K$ , the random vector dimension number  $N$  and the number of categories  $C$ , we demonstrate its application to the models of categorization, information integration and retention in cognitive science.

Also, for models with independent normal distribution, the form of Fisher information formula turns out to be the same as that of multinomial distribution. Now  $P$  and  $\Lambda$  can be easily determined from the normal distribution parameter which is specified by a particular model. As a simple application, Fechner's logarithmic model of psychophysics is considered.

Second, the crucial difference of MDL1996 and MDL2001 criteria is in the model complexity measure  $C_{1996}$  and  $C_{2001}$  which are defined below. With the formula for Fisher information derived before, calculation of  $C_{2001}$  is normally more challenging

than that of  $C_{1996}$ . Under certain conditions, it turns out that  $C_{1996}$  and  $C_{2001}$  are closely related. To show the relationship, we consider the MDL complexity of a simple model with multinomial distribution. And two complexity measures for this model are

$$\begin{aligned} C_{1996} &\triangleq \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Omega} \sqrt{|I(\theta)|} d\theta \\ &= \frac{C-1}{2} \log \frac{n}{2\pi} + \log \left( \frac{\pi^{\frac{C}{2}}}{\Gamma(\frac{C}{2})} \right) \end{aligned}$$

and

$$\begin{aligned} C_{2001} &\triangleq \log \int_{\hat{\theta}(x) \in \Omega} f_{X|\Theta}(x|\hat{\theta}(x)) dx \\ &= \log \left( \sum_{\substack{0 \leq x_c \leq n \\ x_1 + x_2 + \dots + x_C = n}} \binom{n}{x_1, \dots, x_C} \prod_{c=1}^C \left( \frac{x_c}{n} \right)^{x_c} \right) \end{aligned}$$

where  $C$  is the number of categories and  $n$  is the sample size. We compare the relative difference  $(2(C_{2001} - C_{1996}) / (C_{2001} + C_{1996}))$  with changing sample size in Figure 1 for the models with different categories. The curves in the figure indicate that the two complexity measures are asymptotically converge as the sample size goes to infinity. It also shows the bigger relative difference as the category increases.

Third, we also illustrate the adequacy of MDL by comparing its performance with other model selection criteria in the selection of three retention models in cognitive science. The model selection result is presented in table 1. It consists of four matrices corresponding to Maximum Likelihood Estimator (MLE), Bayesian Information Criterion (BIC), Cross Validation (CV) and MDL1996 model selection criteria. In each matrix, the element with row  $i$  and column  $j$  represents the percentage that we choose model  $i$  from the data generated by model  $j$  under that criterion. So the bigger the diagonal element, the better the model selection criterion. From the model recovery rates listed in the table, we conclude that MDL1996 has the most model selections correctly and has clear preference for the simple model in contrast to MLE.

In summary, the derived formula for Fisher information greatly simplifies the computation and can be applied to a wide range of models in cognitive science. We also show that MDL1996 complexity measure provides a good approximation to MDL2001 complexity measure at least in the context of a multinomial model. Ideally, we would prefer MDL2001 to MDL1996. But at most time in practice, parameter space has much lower dimensions than the sample space, so MDL1996 might be more computationally feasible than MDL2001. We further illustrate the adequacy of MDL1996 by comparing its performance with other model selection criteria with the selection of retention models. As a conclusion, it is suggested that the derived formula for Fisher information could be applied in scientific enterprises other than MDL model selection as well.

## References

- [1] Jorma Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, Jan 1996.
- [2] Jorma Rissanen. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717, July 2001.

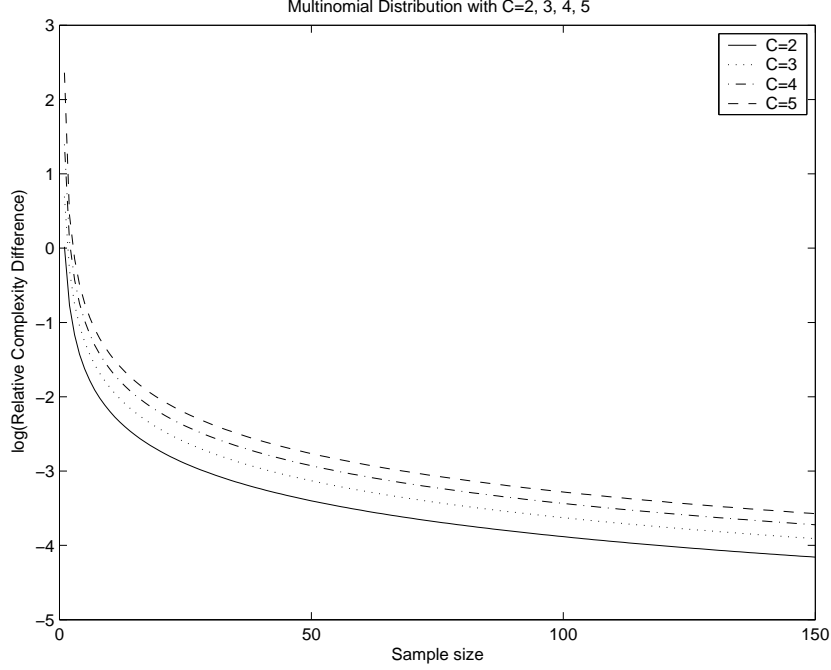


Figure 1: Relative Complexity Difference with Different Categories

Table 1: Model Recovery Rates of Three Retention Models

Selection Method/ Data Fitting Model	Data Generating Model ( $C_{1996}$ )		
	M1(1.2361)	M2(1.5479)	M3(1.7675)
<b>MLE</b>			
M1	22%	11%	0%
M2	41%	88%	4%
M3	37%	1%	96%
<b>BIC</b>			
M1	91%	55%	8%
M2	4%	44%	4%
M3	5%	1%	88%
<b>CV</b>			
M1	52%	40%	7%
M2	28%	53%	19%
M3	20%	7%	74%
<b>MDL1996</b>			
M1	83%	37%	7%
M2	11%	62%	6%
M3	6%	1%	87%

Note. For the retention models, the data sample  $X_k|[a, b]^T \sim Bin(20, f(a, b, t_k))$ , the independent variable  $t_k=1, 2, 4, 8, 16$ . The range of exponential parameter is  $[0, 10]$  and  $[0, 100]$  otherwise and

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