Applications of MDL to Selected Families of Models

Andrew J. Hanson and Philip Chi-Wing Fu
Computer Science Department
Indiana University, Bloomington

Abstract

Evaluating models that approximate complex data distributions is a core problem in data understanding. In this pedagogical review, we describe how the Minimum Description Length principle (MDL) can be applied to evaluate the *relative appropriateness* of distinct models that defy conventional comparison methods, including models that are obscurely equivalent under functional transformations and inequivalent models with the same number of parameters. The MDL principle provides a concrete approach to identifying models that fit the data, avoid over-fitting the noise, and embody no more functional complexity in the model itself than is necessary. New results on the geometric complexity of several families of useful models are derived, and illustrative examples are worked out in detail.

1 Introduction

The world of experimental science is replete with applications that require suitable approximations in order to model complex data sets that contain excessive apparent detail due to noise. Signal analysis, image analysis, shape detection, modeling data from psychological observations, modeling data from physical observations, and so forth, are only a few of the examples that spring immediately to mind. As each new research paper develops yet another clever technique or proposes yet another functional class of models, one bothersome question remains: how can we distinguish among different approaches? What criterion besides the author's word do we have to conclude that one model is better than another? In other words, how do we distinguish a *suitable* approach from an *optimal* approach? Our purpose in this article is to present a core collection of data models analyzed so that the Minimum Description Length (MDL) principle can be used, *after* a parameter choice has been selected, as a possible means of *comparing the appropriateness of distinct models*.

We outline the practices of MDL for a general scientific audience, derive new results for the geometric complexity of common classes of functional models, and provide a selection of illustrations suitable for a variety of modeling and data reduction problems.

The Idea in a Nutshell. The conundrum that leads us to the ideas presented in this article is simply this: suppose you do a least squares fit to a proposed model for a data sample. You suspect that the 10th-order polynomial you used to fit the data is in fact *nonsense*, even if it has *really* low variance, because you have good reason to believe that a cubic polynomial process actually generated the data. You confirm this by checking your 10th order fit against another attempt at the measurement, and it is ridiculously bad, even though the fit to the first sample was superb. If you check the new data against a cubic fit to the old data, it will *still* be an appropriate fit. *How do you figure this out* a priori *when you cannot* take *a second sample?* The MDL method described here shows how this can be accomplished in well-behaved situations. The essence of the entire argument is illustrated in Figures 2, 3, and 4, which show explicitly that the "best" model using MDL, the lowest point on the graph, is also typically the "right" model, the one used secretly to simulate the noisy data; the lowest-variance models are *almost always* wrong.

General Background. The Minimum Description Length principle appears first in the work of Rissanen [15–17], where it arose in the information-theoretic analysis of stochastic processes. Over the years, a number of refinements have appeared, many also due to Rissanen and his collaborators (see, e.g., [1, 3, 5–14, 18–20]). The definitive formulation, answering many questions regarding comparison to other alternative approaches, is found in Rissanen's paper [21]. The underlying idea is simply to stretch information theory to its limits, and to evaluate all the parts of a data description in the same universal language: the number of bits needed in the description. Thus an excessively simple model would require few bits for its own description, but many bits to describe the deviations of the data from the model, while an excessively complex model could describe the data flawlessly, but would require a huge self-description. Less obvious is the fact that two models with the same number of parameters can differ substantially in the measure of the "descriptive power" of their functional spaces, and the appropriateness of a particular model can be distinguished on that basis as well. If this is done carefully, the theory is insensitive to reparameterizations of the models, a potential source of endless confusion and controversy. Also of interest to some classes of problems is the fact that both the model for the data sample and the model for its error process enter into the evaluation. Thus one intuitively expects the evaluation of all modeling problems to involve a compromise including the model's parameters, the form of its statistical noise, and a description of the intrinsic complexity of the modeling function itself. The best compromise is the most elegant description, the minimal overall amount of required information, the concrete mathematical formulation of Occam's razor.

At this time, there are still some open questions regarding the uniqueness of the "geometric cost" that permits the calculation of the relative complexity of two models, the handling of small, as opposed to nearly-infinite, data samples, and an annoying arbitrariness in the choice of model parameter volumes. However, practical calculations using formulas valid for asymptotically large data samples and a functional metric based on the Fisher information matrix are straightforward in practice and exhibit the most essential desired properties: the results are independent of functional reparameterizations of the models, and favor models that generalize to other samples from the same distribution, as opposed to deceptively accurate models that are in fact overfitting the noise.

2 Computing Model Description Length

The fundamental description-length or "cost" formula that we will use, loosely following [11, 13, 14], takes this form:

$$D = F + G \ , \tag{1}$$

which can be read as

"**D**escription-length **equals** **F**it **plus** **G**eometry."

The first term quantifies the "goodness-of-fit" to the data and takes the general form

$$F = -\ln f(y|\hat{\theta}) \ . \tag{2}$$

To compute this term, we must have some means of making a specific numerical choice for the fitted values $\{\hat{\theta}\}$ of the model parameters. We will restrict our treatment here to models of the form

$$y = g(\theta, \mathbf{x}) + \text{error model} \ , \tag{3}$$

which describes the dependent (measured) variable y in terms of a set of model parameters $\{\theta\}$ and the independent variables $\mathbf{x}$; we assume an additive noise model, although other error models such as multiplicative noise could be specified.

The function $f(y|\theta)$ is a user-chosen statistical likelihood function corresponding to the model of Eq. 3 with its error process, and the $\{\hat{\theta}\}$ are model parameters fixed by some (typically maximum likelihood) fitting procedure. F is thus an information-theoretic measure corresponding to the number of bits of description length attributable to inaccuracy: if $f \approx 0$, the data are not well-described by $\{\hat{\theta}\}$, while if $f \approx 1$, the description is ideal. (Note: we will use natural logarithms denoted by "ln" throughout, although technically perhaps $\log_2$ should be used to express all description lengths directly in bits.)

If we have a *sample*, $\{y_n(\mathbf{x}), n = 1, \ldots, N\}$, then we evaluate $f(y|\hat{\theta})$ as the product of the probabilities for each individual *outcome* y_n at the fixed parameter values $\hat{\theta}$ found from the maximum likelihood fit to the hypothesized model $g(\theta,\, \mathbf{x})$:

$$f(y|\hat{\theta}) \to f(\{y_n\}|\hat{\theta}) \equiv \prod_{n=1}^{N} f(y_n|\hat{\theta}) \; . \tag{4}$$

This makes explicit the intuition that F quantifies the cost of describing the deviation of the set of N measured outcomes in the sample $\{y_n(\mathbf{x})\}$ from the maximum likelihood fit. A critical feature of the approach is that the error distribution *must* be specified to completely define the model; MDL can in fact theoretically distinguish between identical models with *differing* statistical error generation processes.

The second term is the "geometric term" (technically the *parametric complexity* of the model),

$$G = +\frac{K}{2} \ln \frac{N}{2\pi} + \ln \int_{\mathrm{Vol}} d^K\theta \, \sqrt{\det \, I(\theta)} \; , \tag{5}$$

where K is the number of parameters and $\{\theta_k,\, k = 1, \ldots, K\}$ is the parameter set of the model having Vol as the domain of the entire K-dimensional parameter space integration for the model being considered. $I(\theta)$ is the $K \times K$ Fisher Information Matrix averaged over the data samples, but with each y replaced by its expectation; the computation of $I(\theta)$ is complex, and will be discussed in detail in a moment. Note that our choice of upper-case K for the number of model parameters is often written with a lower-case k in the literature.

Intuitively, $I(\theta)$ has many properties of a metric tensor, and in fact $d^K\theta \, \sqrt{\det \, I}$ has precisely the form of a reparameterization-invariant volume element $d^K\mathbf{x} \, \sqrt{g}$ familiar from Riemannian geometry and general relativity. This volume element effectively allows us to count the number of distinct probability distributions the model can generate (see the discussion in [14] and related citations).

The Fisher Information Matrix. We now attend to the definition of the Fisher Information Matrix and the rest of the machinery required to carry out explicit computations of $I(\theta)$, as well as working out a standard example that will serve as our model throughout the rest of the article.

First, we define the general notion of an expectation of a function $h(y)$ with respect to a statistical likelihood function as follows:

$$E(h(y)) = \int dy \, h(y) \, f(y|\theta) \; . \tag{6}$$

Thus, any coefficient in a polynomial expansion of $h(y)$ will be multiplied by the expectation corresponding to the appropriate m-th moment,

$$E(y^m) = \int dy \, y^m \, f(y|\theta). \tag{7}$$

To compute the Fisher Information Matrix, one begins by considering the expectation of the second derivative of the chosen log likelihood function for continuous variables and parameters,

$$L_{ij}(\theta, \mathbf{x}) = E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j}\left[\ln f(y|\theta, \mathbf{x})\right]\right) \ , \tag{8}$$

where we explicitly include the possible dependence of f on the dependent variable $\mathbf{x}$ through $g(\theta, \mathbf{x})$. When the expectation is computed, the dependent variable y is integrated out; however, the values of the dependent variables $\mathbf{x}$ remain, and, in particular, will be known for each outcome of a particular sample $\{y_n\}$. This leads to the definition of the Fisher Information Matrix, which is the average of L_{ij} over the actually obtained outcomes in the data sample; using Eq. (4) to expand $\ln f(y|\theta)$ as the sum of the logs of the individual components for the dependent variable y_n measured at the location $\mathbf{x}_n$ in the space of independent variables, we obtain the basic definition

$$\begin{aligned} I_{ij}(\theta) &= \frac{1}{N}\sum_{n=1}^{N} L_{ij}(\theta, \mathbf{x}_n) \\ &= \frac{1}{N}\sum_{n=1}^{N} E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j}\left[-\ln f(y|\theta, \mathbf{x}_n)\right]\right) \end{aligned} \tag{9}$$

for the Fisher Information Matrix of a measured sample.

The Normal Distribution. The easiest way to understand $I(\theta)$ is to choose a specific error model and work out an example. The Gaussian describing the usual normal distribution,

$$f(y|\theta, \mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left(-\frac{1}{2\sigma^2}(y - g(\theta,\ \mathbf{x}))^2\right) \ , \tag{10}$$

is by far the most common error distribution, and is easy to compute with. The error is modeled by the Gaussian width σ, and the relevant expectations may be computed explicitly:

$$\begin{aligned} E(1|\mathbf{x}) &= 1 \\ E(y|\mathbf{x}) &= g(\theta,\ \mathbf{x}) \\ E(y^2|\mathbf{x}) &= \sigma^2 + g(\theta,\ \mathbf{x})^2 \\ \cdots &\quad . \end{aligned} \tag{11}$$

To get the Fisher Information Matrix for the normal distribution, we find from direct differentiation and Eq. (11) that

$$\begin{aligned} L_{ij}(\theta,\ \mathbf{x}) &= -\frac{1}{\sigma^2}E(y - g(\theta,\ \mathbf{x}))\,\partial_i\,\partial_j\,g(\theta,\ \mathbf{x}) + \frac{1}{\sigma^2}\partial_i g(\theta,\ \mathbf{x})\partial_j g(\theta,\ \mathbf{x}) \\ &= 0 + \frac{1}{\sigma^2}\frac{\partial g(\theta,\ \mathbf{x})}{\partial\theta_i}\frac{\partial g(\theta,\ \mathbf{x})}{\partial\theta_j} \ . \end{aligned} \tag{12}$$

We obtain the corresponding Fisher Information Matrix by computing the average over outcomes,

$$I_{ij}(\theta) = \frac{1}{N}\sum_{n=1}^{N} L_{ij}(\theta, \mathbf{x}_n) = \frac{1}{N\sigma^2}\sum_{n=1}^{N} \partial_i g_n(\theta)\partial_j g_n(\theta) \ , \tag{13}$$

where we use the convenient abbreviation $g_n(\theta) = g(\theta_1, \ldots, \theta_K; \mathbf{x}_n)$. Note that the $\{\theta\}$ are free variables, not the maximum likelihood values, since we must integrate over their domains; however, the values $\{\hat{\theta}\}$ may still be of importance, since they can in principle determine the dominant contribution to the integral. It is important to realize that $\det I$ will vanish unless the number N of linearly independent measurements is at least equal to the dimension of the parameter space, $N \geq K$; the geometric term is undefined unless there are enough measurements to determine a fit to the model parameters.

When computing the determinant in the integral of the geometric term for the normal distribution, it is sometimes convenient to rearrange the terms in Eq. (13) using

$$\begin{aligned} \ln \int d^K\theta \sqrt{\det I(\theta)} &= \ln \int d^K\theta \sqrt{\left(\frac{1}{N\sigma^2}\right)^K} \sqrt{\det |\sum_n \partial_i g_n(\theta)\partial_j g_n(\theta)|} \\ &= -\frac{K}{2}\ln N\sigma^2 + \ln \int d^K\theta \sqrt{\det |\sum_n \partial_i g_n(\theta)\partial_j g_n(\theta)|} \ . \end{aligned}$$

This permits us to cancel the factors of N and re-express the geometric term as

$$\begin{aligned} G &= \frac{K}{2}\ln\frac{N}{2\pi} - \frac{K}{2}\ln N\sigma^2 + \ln \int d^K\theta \sqrt{\det |\sum_n \partial_i g_n(\theta)\partial_j g_n(\theta)|} \\ &= -K\ln\sigma\sqrt{2\pi} + \ln \int d^K\theta \sqrt{\det |\sum_n \partial_i g_n(\theta)\partial_j g_n(\theta)|} \ . \end{aligned}$$

3 Piecewise Constant Models

Suppose that a particular data set is sampled at intervals corresponding to power-of-two subdivisions of the domain. Then we can identify the "simplest" model — the global mean, the most complex model, where each data point is itself a model parameter, and a complete set (the binary tree) of power-of-two models between these two extremes. We now treat each in turn.

Global Mean Model. The simplest possible model is just a constant

$$y = \mu$$

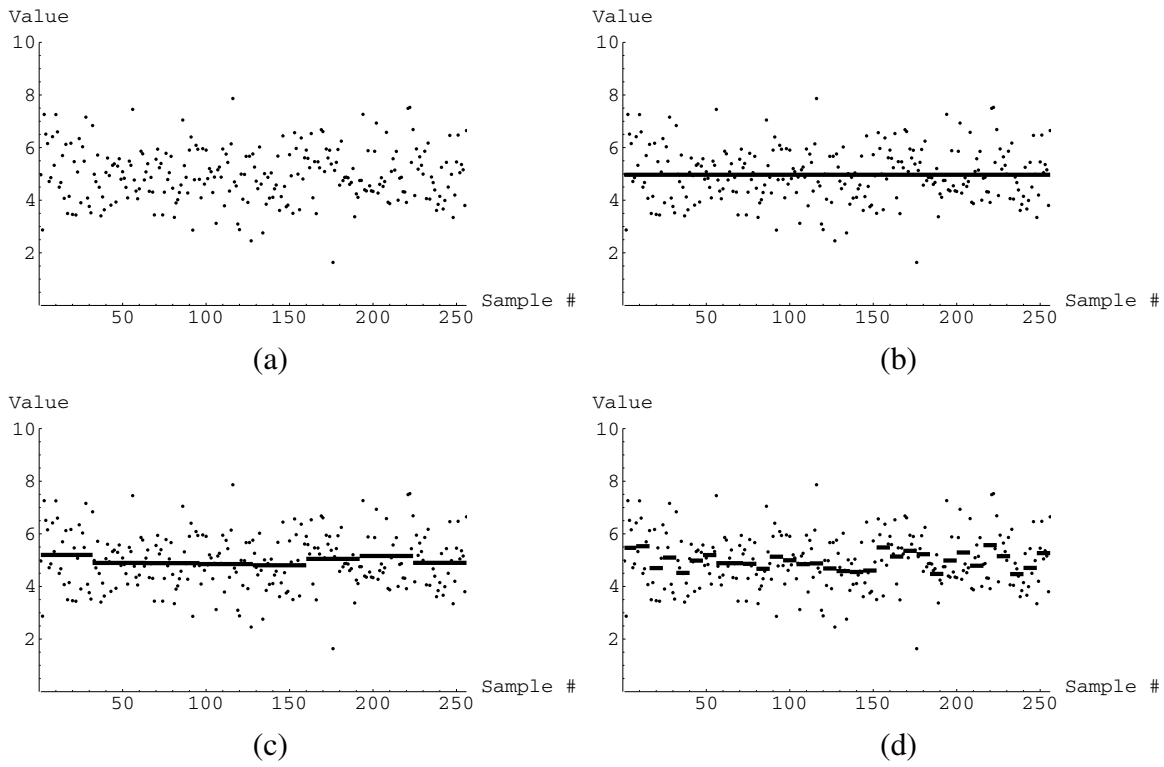


Figure 1: (a) Data set generated from a constant plus noise (identical to the $N = K$-parameter "perfect" piecewise fit). (b) Single-parameter mean value fit. (c) Overfitting with an evenly spaced set of 8 piecewise constants. (d) Overfitting with 32 piecewise constants.

(plus noise) corresponding to the simulated data shown in Figure 1(a). A least squares fit to the data gives the maximum likelihood solution

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} y_n,$$

as shown in Figure 1(b). $\hat{\mu}$ is expected to be very close to the value of μ used to simulate the data, but will virtually never match it exactly. The cost of representing the deviations from this fit is given by

$$\begin{aligned} F &= -\ln f(\{y_n\}|\hat{\mu}) \\ &= -\ln \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y_n - \hat{\mu})^2/2\sigma^2) \\ &= N \ln \sigma\sqrt{2\pi} + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \hat{\mu})^2 \end{aligned}$$

$$= \quad N \ln \sigma\sqrt{2\pi} + \frac{N}{2\sigma^2}(\text{variance}) \; . \tag{14}$$

Since $K = 1$ and the Fisher matrix is 1×1, we have simply

$$I(\hat{\mu}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\sigma^2} = \frac{1}{\sigma^2} \; , \qquad (\text{ Note: } (\partial\hat{\mu})^2 = 1 \;)$$

so the geometric term becomes (from Equation 5)

$$\begin{aligned} G &= \frac{K}{2} \ln \frac{N}{2\pi} + \ln \frac{1}{\sigma} \int_{\text{min}}^{\text{max}} d\mu \\ &= \frac{1}{2} \ln \frac{N}{2\pi} + \ln \frac{\mu_{\text{max}} - \mu_{\text{min}}}{\sigma} \; . \end{aligned} \tag{15}$$

Data-Perfect Model. On the other hand, the most complex model is effectively no model at all, the model with one parameter for each measured value ($K = N$),

$$y = \sum_{n=1}^{N} y_n \delta(x, x_n) \; ,$$

where $\delta(x, x_n)$ is the Kronecker delta (unity for $x = x_n$, zero otherwise). There is nothing to fit: assuming the choice of $\{x_n\}$ is a regularly-spaced sequence, so $\{x_n\}$ is not a choice of parameters, then we have N parameters $\{y_n\}$; if the $\{x_n\}$ are specified independently in the measurement, then we would have $2N$ parameters, $\{(x_n, y_n)\}$. For simplicity, we treat the former case, so the model graph is the same as the data plot in Figure 1(a), and

$$\begin{aligned} F &= -\ln f(\{y_n\}|\{y_n\}) \\ &= N \ln \sigma\sqrt{2\pi} + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - y_n)^2 \\ &= N \ln \sigma\sqrt{2\pi} + 0 \; . \end{aligned} \tag{16}$$

As promised, this has no cost corresponding to deviations of the data from the model. The Fisher information matrix, however, is now $N \times N$, and (from Equation 12)

$$\begin{aligned} L_{ij} &= \frac{1}{\sigma^2} \left(\sum_n \frac{\partial(y_n \delta(x, x_n))}{\partial y_i} \right) \left(\sum_{n'} \frac{\partial(y_{n'} \delta(x, x_{n'}))}{\partial y_j} \right) \\ &= \frac{1}{\sigma^2} \delta(x, x_i) \delta(x, x_j) \end{aligned} \tag{17}$$

$$I_{ij} = \frac{1}{N\sigma^2} \sum_{n=1}^{N} \delta(x_n, x_i) \delta(x_n, x_j) \;=\; \frac{1}{N\sigma^2} \delta(i, j) \; . \tag{18}$$

This is tricky because $\sum_{n=1}^{N} \delta(x_n, x_i)\delta(x_n, x_j)$ equals 1 only if i equals j, so it equals $\delta(i,j)$. Since $K = N$ and $\delta(i,j)$ represents the $N \times N$ identity matrix, the geometric contribution (assuming identical parameter domain sizes V) is (from Equation 5)

$$\begin{aligned} G &= \frac{N}{2}\ln\frac{N}{2\pi} + \ln \underbrace{\int_V \cdots \int_V}_{N} d^N\mathbf{y}\,\sqrt{\det \frac{1}{N\sigma^2}\delta(i,j)} \\ &= \frac{N}{2}\ln\frac{N}{2\pi} - \frac{N}{2}\ln N + \ln\frac{V^N}{\sigma^N} \\ &= +\frac{N}{2}\ln\frac{1}{2\pi} + N\ln\frac{V}{\sigma}\ . \end{aligned} \tag{19}$$

Binary-Tree Model. Data are often approximated by a binary tree generated by applying recursive 2-element box filters. We can represent an entire family of models in this fashion, each with 2^M parameters, where $M = 0$ is the single parameter (global mean) model treated first, and $M = \log_2 N$ ($N = 2^M$) is the zero-error model. The models each take the form

$$y = \sum_{n=1}^{N} g_n(M)\delta(x, x_n)\ .$$

The power-of-two subdivision is represented by requiring the $g_n(M)$'s to be repeated $N/2^M$ times, and defining the 2^M independent parameters to be $\{z_n(M), n = 1, \ldots, 2^M\}$. The best-fit values $\hat{\mathbf{z}}$ then are computed from the means over the repeated occurrences (e.g., the box-filtered means at each level). To be explicit, if $\{y_n\}$ is a sample, the M independent parameter sets giving the best fit at each level are:

$M = \log_2 N$	$\rightarrow$	$\hat{z}_n = y_n$
$M = \log_2(N/2)$	$\rightarrow$	$\hat{z}_1 = (\hat{g}_1 = \hat{g}_2) = (1/2)(y_1 + y_2)$,
		$\hat{z}_2 = (\hat{g}_3 = \hat{g}_4) = (1/2)(y_3 + y_4), \ldots$
$M = \log_2(N/4)$	$\rightarrow$	$\hat{z}_1 = (\hat{g}_1 = \hat{g}_2 = \hat{g}_3 = \hat{g}_4) =$
		$(1/4)(y_1 + y_2 + y_3 + y_4), \ldots$
$\ldots$	$\ldots$	$\ldots$
$M = 0$	$\rightarrow$	$\hat{z}_1 = (\hat{g}_1 = \ldots = \hat{g}_n) = \hat{\mu}$

In Figure 1(a,b,c,d), we show a single data set generated by a distribution with constant mean along with the fits for $M = 0$, $M = 3$, and $M = 5$, respectively, to illustrate overfitting.

The goodness-of-fit term becomes

$$F(M) = -\ln \prod_{n=1}^{N} f(y_n|\hat{\mathbf{z}}(M))$$

$$= \quad N \ln \sigma\sqrt{2\pi} + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \hat{z}_{m(n)}(M))^2 \tag{20}$$

where $m(n) = \lceil n2^M/N \rceil$, so the $\hat{z}_m(M)$ are understood as the 2^M independent constants (not N independent values) giving the best-average fit to the measurements at binary-tree level M.

The geometric term, which gives the measure of the functional space spanned by the $z_m(M)$ considered as *variable* parameters, is based on the $2^M \times 2^M$ matrix

$$\begin{aligned} I_{ij}(M) &= \frac{1}{N\sigma^2} \sum_{n=1}^{N} \Big[\frac{\partial}{\partial z_i} \left(\sum_{l=1}^{N} z_{m(l)}(M) \delta(x_n, x_l) \right) \\ &\qquad \frac{\partial}{\partial z_j} \left(\sum_{l'=1}^{N} z_{m(l')}(M) \delta(x_n, x_{l'}) \right) \Big] \\ &= \frac{1}{\sigma^2 2^M} \delta(i,j) \,, \end{aligned} \tag{21}$$

where we have used the fact that $N/2^M$ occurrences of $z_m(M)$ are replicated at the level M.

Thus, with $K = 2^M$ and $\delta(i,j)$ representing the $2^M \times 2^M$ identity matrix, we have (assuming identical parameter domain sizes V) (similar to Equation 19),

$$\begin{aligned} G(M) &= +\frac{2^M}{2} \ln \frac{N}{2\pi} + \ln \underbrace{\int_V \cdots \int_V}_{2^M} d^{2^M}\mathbf{z} \sqrt{\det \frac{\delta(i,j)}{2^M \sigma^2}} \\ &= +\frac{2^M}{2} \ln \frac{N}{2\pi} + \ln V^{2^M} + \frac{1}{2} \ln \left(\frac{1}{2^M \sigma^2} \right)^{2^M} \\ &= \frac{2^M}{2} \ln \frac{N}{2\pi\ (2^M)} + 2^M \ln \frac{V}{\sigma} \,, \end{aligned} \tag{22}$$

and, for the M-th level binary-tree model, the total description length is

$$\begin{aligned} D(M) &= F(M) + G(M) \\ &= N \ln \sigma\sqrt{2\pi} + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \hat{z}_{m(n)}(M))^2 \\ &\quad + \frac{2^M}{2} \ln \frac{N}{2\pi\ (2^M)} + 2^M \ln \frac{V}{\sigma} \,. \end{aligned} \tag{23}$$

This form shows explicitly the transition from Eq. (15), with $M = 0$, to Eq. (19), with $2^M = N$.

Quadtrees, Octrees, etc. The binary tree model for piecewise constant 1D data can be easily extended to higher dimensions. For D-dimensional data, the corresponding model is a piecewise constant 2^D-tree ($D = 2$ being a quadtree, $D = 3$ being an octree, etc.). For simplicity, we assume N^D data samples over all D dimensions. We define the model as

$$A(x, y, \ldots) = \sum_{ij\ldots}^{N^D} z_{m(ij\ldots)}(M)\delta(x, x_i)\delta(y, y_j)\ldots .$$

Then the data-fitting and geometric terms become

$$\begin{aligned} F(M, D) &= -\ln \prod_{ij\ldots}^{N^D} f(A_{ij\ldots}|\hat{\mathbf{z}}(M)) \\ &= N^D \ln \sigma\sqrt{2\pi} + \frac{1}{2\sigma^2}\sum_{ij\ldots}^{N^D}(A_{ij\ldots} - \hat{z}_{m(ij\ldots)})^2 . \end{aligned} \tag{24}$$

The number of parameters is $K = (2^M)^D$ and the $K \times K$ Fisher matrix is

$$I_{ij\ldots,i'j'\ldots} = \frac{1}{\sigma^2}\left(\frac{1}{2^M}\right)^D \delta(ij\ldots, i'j'\ldots) ,$$

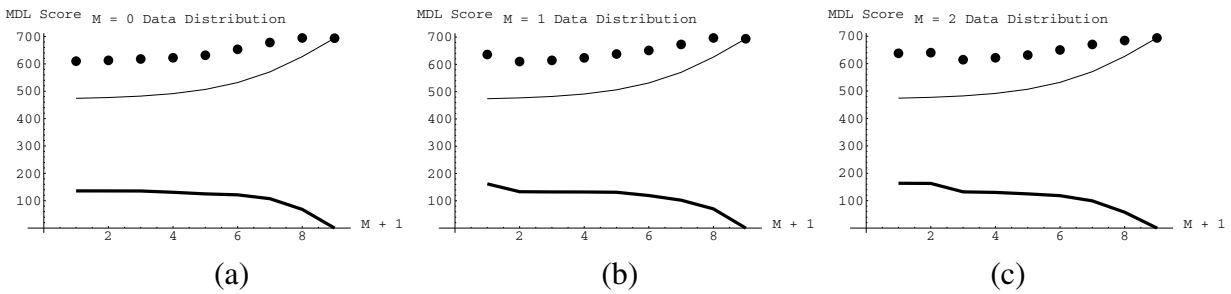
where $i, j, \ldots$ now range to 2^M instead of N. The geometry term becomes

$$G(M, D) = \frac{2^{MD}}{2}\ln\frac{N}{2\pi\,(2^{MD})} + 2^{MD}\ln\frac{V}{\sigma} , \tag{25}$$

which generalizes Equation 23 for $D = 1$. One could even apply this approach to the *independent* modeling of *each* level of a signal's resolution hierarchy to achieve optimal compression.

Numerical Experiments. The principal motivation for the MDL approach is to distinguish between a *good fit* and *overfitting* to achieve what is referred to in the statistics and pattern recognition literature as *generalization*. If the model does not generalize, then additional data sets with the same statistics will *not* be well-described, indicating the presence of an excessively complex model that conforms to the random noise patterns of one isolated sample. We can test this by generating a 256-element sample normally distributed about a mean of zero with $\sigma = 1$, and allowed parameter values $[-3, 3]$ so that $V = 6$ (ambiguities can arise if V/σ is too small). For $M = 0$, the fit to the single parameter is $\hat{z} = \hat{\mu} = (1/N)\sum y_n$; for $M = 1$, we can determine $\hat{z_1}$ and $\hat{z_2}$ to be the means for the left and right halves of the data, etc.

In Figure 2(a), we compare the variance term $(F(M) - \text{const})$ (the heavy curve) vs $(G(M) + \text{const})$ (the light curve) as a function of M, and show the summed description



(a) (b) (c)

Figure 2: Comparison of the variance contribution $F - \text{const}$ (heavy curve), the geometric contribution $G + \text{const}$ (light curve), and their sums $D(M)$ (heavy dots) as a function of level M for three different data distributions: (a) Original data generated by a single mean, $M = 0$. (b) Original data generated by two means, $M = 1$. (c) Original data generated by four means, $M = 2$.

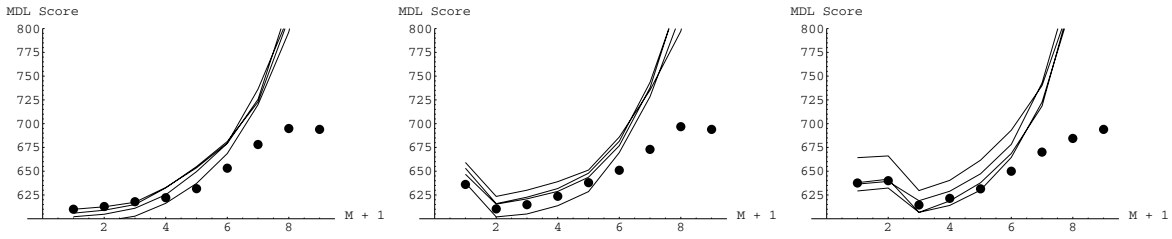
length $D(M)$ as black dots. In Figure 2(b), we repeat the process, except that the data are distributed about a "true" $M = 1$ model, with means 0 and 1 for the left and right halves of the data. Figure 2(c) shows the results for an $M = 2$ model, with means $(0, 1, 0, 1)$. We can now explicitly see the tradeoff between the data fit and the model description: the minimum sum occurs as promised for the true model.

Curiously, we see that the drastic drop in the data error for the "perfect" N-parameter model gives it a slight statistical edge over its neighbors. However, this is an illusory advantage: if we generate several *additional* data sets with the same distributions and evaluate them against the set of fits $\{\hat{z}(M)\}$ determined by the original data sets, we see the results in Figure 3. The overfitted models with excess parameters are extremely poor descriptions of the abstract data distribution. The minimal models generalize perfectly, and the overfitted models are terrible generalizations.

We conclude that choosing the model with the minimum description length avoids both the traps of underfitting and overfitting, and suggests the selection of models close to those generated by the actual data rather than being confused by models with artificially low variance. In principle, models with different statistical distributions and parameter-space geometry can also be distinguished, though non-compact parameter spaces require some externally-imposed assumptions [11, 14].

4 Continuous Linear Models

Polynomial Functions Polynomials form the simplest class of differentiable models beyond the piecewise-constant models of the previous section, and can be extended to include piecewise continuous splines in principle. If we choose K-parameter polynomial models



(a) (b) (c)

Figure 3: MDL cost as a function of binary-tree-level M for the model for four data set samplings and three different models. The heavy dots are the same points $D(M)$ as in Figure 2 and denote the costs of the model used to determine the maximum likelihood parameters used in evaluating curves for the remaining models. (a) Original data are distributed about a single mean ($M = 0$); (b) two means ($M = 1$); (c) four means ($M = 2$).

of the form

$$y = \sum_{k=0}^{K-1} a_k x^k \tag{26}$$

and then carry out a least-squares fit to get the maximum-likelihood parameter estimates $\hat{a}_k$, the data-fitting term for N outcomes $\{(x_n, y_n)\}$ is

$$F(K) \;=\; N \ln \sigma\sqrt{2\pi} + \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(y_n - \sum_{k=0}^{K-1} \hat{a}_k (x_n)^k \right)^2 \;, \tag{27}$$

and the $K \times K$ geometric term matrix with $i, j = 0, \ldots, K-1$, is

$$\begin{aligned} I_{ij} \;&=\; \frac{1}{N\sigma^2} \sum_{n=1}^{N} \frac{\partial}{\partial a_i} \left(\sum_{k=0}^{K-1} a_k (x_n)^k \right) \frac{\partial}{\partial a_j} \left(\sum_{k'=0}^{K-1} a_{k'} (x_n)^{k'} \right) \\ &=\; \frac{1}{N\sigma^2} \sum_{n=1}^{N} (x_n)^i (x_n)^j \;=\; \frac{1}{N\sigma^2} \sum_{n=1}^{N} (x_n)^{(i+j)} \;. \end{aligned} \tag{28}$$

The geometric term becomes

$$\begin{aligned} G(K) \;&=\; +\frac{K}{2} \ln \frac{N}{2\pi} + \ln \underbrace{\int_V \cdots \int_V}_{N} d^K a \sqrt{\det \left[\frac{1}{N\sigma^2} \sum_{n=1}^{N} (x_n)^{(i+j)} \right]} \\ &= -\frac{K}{2} \ln 2\pi + K \ln \frac{V}{\sigma} + \frac{1}{2} \ln \det \left[\sum_n (x_n)^{(i+j)} \right] \;, \end{aligned} \tag{29}$$

where we assumed the same domain size V for each parameter. The determinant vanishes even for linearly independent outcomes unless $N \geq K$, excluding underdetermined

models. Note that, unlike the piecewise constant case, $G(K)$ now has an explicit data dependence.

In the specific case where there are sufficiently many outcomes so that the sum in Eq. (29) approximates a Monte Carlo integration over some domain, say $a \leq x \leq b$, we can explicitly compute the last term in Eq. (29), with $\Delta x \approx (b-a)/N$, as

$$\frac{1}{2} \ln \det \frac{N}{b-a} \left[\frac{b^{i+j+1} - a^{i+j+1}}{i+j+1} \right] .$$

We remark on the close relationship of the resulting matrix to the notoriously ill-conditioned Hilbert matrix, which is exact if $a = 0$ and $b = 1$.

An identical exercise to that in Figure 3 can now be carried out. In Figure 4, we show the cost of fits up to $K = 9$ (8th power) for samples generated using constant, linear, quadratic, and cubic models with normally distributed error. We observed that the relative magnitudes of the standard deviation in the error model and the scale of x can affect whether the correct polynomial order is unambiguously selected. Here we used $\sigma = 1$, $a_k = 1$, and $0 \leq x \leq 3$ with 256 outcomes. We see that the optimal K is that used to generate the data.

Orthonormal Functions. If we replace the power series by a set of normalized orthogonal polynomials, we would write

$$y = \sum_{k=0}^{K-1} a_k h_k(x) , \tag{30}$$

where the orthogonality relation between h_k and its conjugate $\overline{h}_k$, using integration domain U, is by definition

$$\int_U dx\, \overline{h}_k(x) h_{k'}(x) = \delta_{k,k'} , \tag{31}$$

so that we may *formally* determine the expansion coefficients from the integrals

$$a_k = \int_U dx\, \overline{h}_k(x) y(x) . \tag{32}$$

Here we in principal have a *choice* of methods to determine the optimal coefficients $\{\hat{a}_k\}$:

- **Maximum Likelihood.** The model Eq. (30) can be fit using least squares methods like any other function. This method is probably preferred for sparse data distributions.

- **Projection.** *Provided* the samples are appropriately distributed or can be selected in such a way that the discretely sampled version of the projection Eq. (32) is a good approximation to the analytic integral, we can take $\Delta x \approx U/N$, and write

$$\hat{a}_k \approx \frac{U}{N} \sum_{n=1}^{N} \overline{h}_k(x_n) y_n . \tag{33}$$

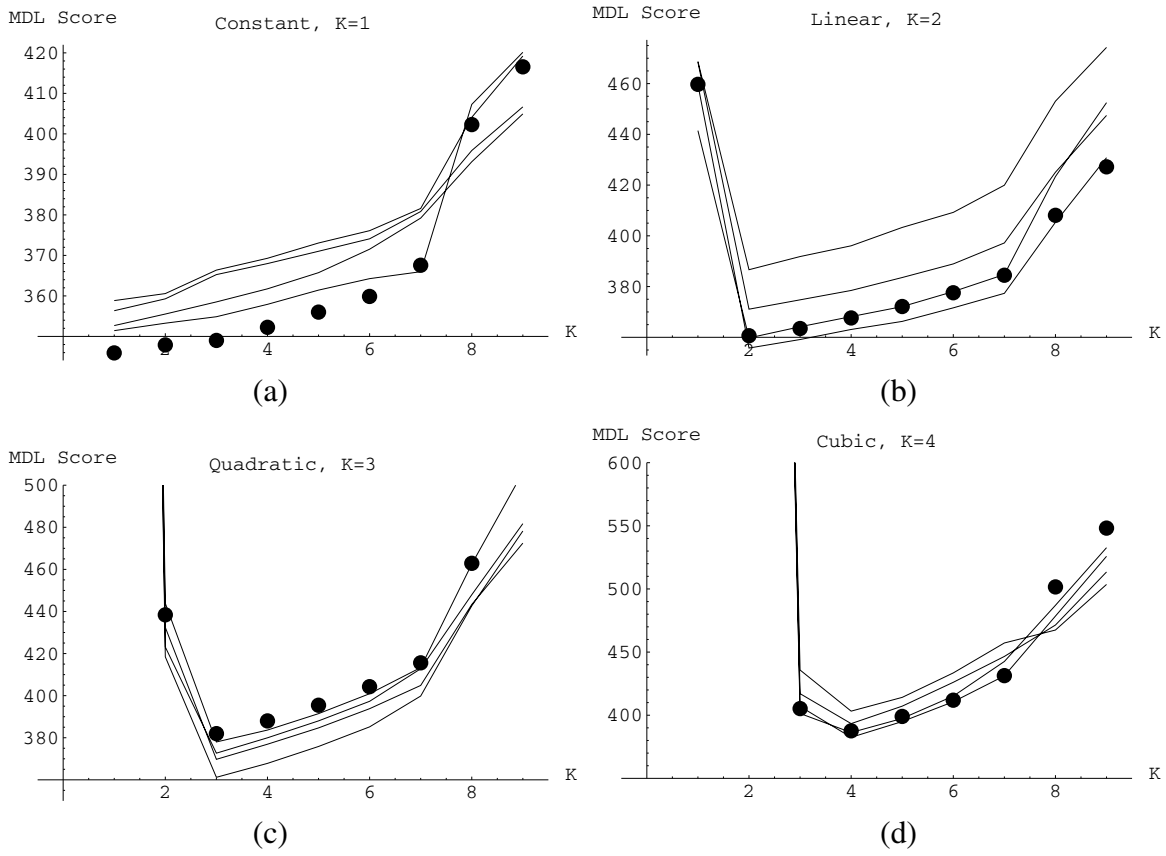


Figure 4: The lowest MDL cost as a function of the number of fitted polynomial parameters K for one data sample (heavy dots) selects the generating model, as well as generalizing to four additional data samples (curves). (a) The simulated data are normally distributed about a constant function; (b) linear function; (c) quadratic function; (d) cubic function.

The polynomials themselves form our first example of this class of functions if we normalize the Legendre polynomials appropriately, e.g.,

$$\begin{aligned} Q_0(x) &= \sqrt{\frac{1}{2}} \cdot 1 \\ Q_1(x) &= \sqrt{\frac{3}{2}} \cdot x \\ Q_2(x) &= \sqrt{\frac{5}{2}} \cdot \left(\frac{3}{2}x^2 - \frac{1}{2}\right) \\ &\vdots \end{aligned}$$

with integration range $-1 \le x \le 1$. Choosing the model sequence up to some maximum K as

$$y = a_0\sqrt{\frac{1}{2}} + a_1\sqrt{\frac{3}{2}}x + a_2\sqrt{\frac{5}{2}}\left(\frac{3}{2}x^2 - \frac{1}{2}\right) + \cdots \ , \tag{34}$$

we can choose either the *least-squares fit* or the *projection* to determine the model coefficients; the projection can be computed using

$$\begin{aligned}
\hat{a}_0 &= \frac{2}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{2}}y_n \\
\hat{a}_1 &= \frac{2}{N}\sum_{n=1}^{N}\sqrt{\frac{3}{2}}x_n y_n \\
\hat{a}_2 &= \frac{2}{N}\sum_{n=1}^{N}\sqrt{\frac{5}{2}}\left(\frac{3}{2}x_n^2 - \frac{1}{2}\right) y_n \\
&\vdots
\end{aligned}$$

Given the model Eq. (34) and the coefficients $\{\hat{\mathbf{a}}\}$, we can compute $f(y|\hat{\mathbf{a}})$ and thus F and G almost exactly as we did for the polynomial example leading to Figure 4, and we expect similar results if the samples are sufficiently well-behaved.

Other examples of this class include the discrete sine-cosine series, $(1/\sqrt{\pi})\cos j\theta$ and $(1/\sqrt{\pi})\sin j\theta$ for $j \neq 0$, and $(1/\sqrt{2\pi})$ for $j = 0$, where, e.g.,

$$\int_{U=2\pi} d\theta \left(\frac{1}{\sqrt{\pi}}\cos j\theta\right)\left(\frac{1}{\sqrt{\pi}}\cos j'\theta\right) = \delta_{j,j'} \ , \tag{35}$$

and the spherical harmonic series with basis functions $Y_{lm}(\theta,\phi)$, detailed below.

Remark. The calculation of the geometric term of the description length for orthonormal functions has one notable peculiarity. If we assume a real basis, so $\overline{h} = h$ (e.g., the cosine), the Fisher matrix can be reduced to

$$I_{ij} = \frac{1}{N\sigma^2}\sum_{n=1}^{N} h_i(x_n)h_j(x_n) \ . \tag{36}$$

Remarkably, just as we saw for the projection, Eq. (32), this is essentially a Monte Carlo approximation to the orthogonality integral, Eq. (31), if the samples are appropriately distributed. Therefore, as $N \to$ (large) (which is, indeed, the condition for the validity of many of the MDL formulas we are using), with $\Delta x \approx U/N$, then

$$I_{ij} \approx \frac{1}{N\sigma^2}\frac{N}{U}\delta_{i,j} \ . \tag{37}$$

If $V_K = \prod_k \int da_k$, then the geometric term is just

$$G = \frac{K}{2} \ln \frac{N}{2\pi U} + \ln \frac{V_K}{\sigma^K} \ . \tag{38}$$

If we assume identical parameter domains, then we can also make the simplification $V_K = (V)^K$.

Similar classes of functions such as wavelets would give results exactly analogous to our findings for orthogonal expansions:

$$\begin{aligned} I_{ij} &= \frac{1}{N\sigma^2} \sum_{n=1}^{N} W_i(x_n) W_j(x_n) \\ &\approx \frac{1}{N\sigma^2} \frac{N}{U} \int_U dx \, W_i(x) W_j(x) \\ &\approx \frac{1}{N\sigma^2} \frac{N}{U} H_{ij} \ , \end{aligned} \tag{39}$$

so

$$G = \frac{K}{2} \ln \frac{N}{2\pi U} + \ln \frac{V_K}{\sigma^K} + \frac{1}{2} \ln \det H_{ij} \ , \tag{40}$$

for some appropriately defined integration domains and overlap functions H_{ij}.

Explicit example: real spherical harmonics. Suppose that we have a model for a radially varying spherical data set that we wish to expand around a fixed origin using an unknown optimal number L of spherical harmonics. Then we can express this radial function for sampled values of the angular coordinates (θ, ϕ) on an ordinary sphere as

$$y = r(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{+l} \left(c_{lm} Y_{lm}^c(\theta, \phi) + s_{lm} Y_{lm}^s(\theta, \phi) \right) \ , \tag{41}$$

where Y_{lm}^c and Y_{lm}^s are the cosine-like and sine-like real spherical harmonics (see, e.g., the *mathworld* web page, `http://mathworld.wolfram.com/SphericalHarmonic.html`, Arfken [2], or Ritchie and Kemp [22] for full details). Note that $\{c_{lm}, s_{lm}\}$ are the model parameters that we previously denoted by $\{\theta\}$, while (θ, ϕ) now corresponds to "physics convention" polar coordinates with ($x = r \cos\phi \sin\theta$, $y = r \sin\phi \sin\theta$, $z = r \cos\theta$) in order to have the correct correspondence to the conventions for $Y_{lm}(\theta, \phi)$: we take $0 \leq \phi < 2\pi$, $0 \leq \theta \leq \pi$, so the integration volume element $d\Omega = d\cos\theta \, d\phi$ has total volume 4π. Our task is to determine the optimal value L of the last useful term in the harmonic series for a body of data using MDL.

For each value of L in a set of attempted data descriptions with $L = 0, 1, 2, 3, \ldots$, we determine by some suitable means (e.g., least squares fit or projection) a corresponding

set of optimal model parameters $\{\hat{c}_{lm}, \hat{s}_{lm}\}$ from the data. The goodness-of-fit term in the MDL expression with normal statistics becomes

$$\begin{aligned} F &= -\sum_{n=1}^{N} \ln f\left(r(\theta_n,\phi_n)|\{\hat{c}_{lm},\hat{s}_{lm}\}\right) \\ &= N\ln\sigma\sqrt{2\pi} + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(r(\theta_n,\phi_n) - \sum_{lm}\left(\hat{c}_{lm}Y^c_{lm}(\theta_n,\phi_n) + \hat{s}_{lm}Y^s_{lm}(\theta_n,\phi_n)\right)\right)^2 \end{aligned} \quad (42)$$

and the geometric complexity term is

$$G = \frac{K}{2}\ln\frac{N}{2\pi} + \ln\int d\{c_{lm}\}\int d\{s_{lm}\}\sqrt{\det I(\{c_{lm},s_{lm}\})}\,. \quad (43)$$

We remark that for even functions, only the c_{lm} survive, and so $K = (L+1)^2$; for odd functions, the $l = 0$ term is absent, and so technically $K = (L+1)^2 - 1$; for mixed functions, we would therefore expect $K = 2(L+1)^2 - 1$ parameters. We will leave K unspecified to allow appropriate adjustments for particular data sets.

Expanding the Fisher Information matrix, we can explicitly write the terms as

$$\begin{aligned} &\det I_{lm,l'm'}(\{c_{lm},s_{lm}\}) = \\ &\qquad \det\frac{1}{N\sigma^2}\begin{vmatrix} \sum_{n=1}^N Y^c_{lm}(\theta_n,\phi_n)Y^c_{l'm'}(\theta_n,\phi_n) & \sum_{n=1}^N Y^c_{lm}(\theta_n,\phi_n)Y^s_{l'm'}(\theta_n,\phi_n) \\ \sum_{n=1}^N Y^s_{lm}(\theta_n,\phi_n)Y^c_{l'm'}(\theta_n,\phi_n) & \sum_{n=1}^N Y^s_{lm}(\theta_n,\phi_n)Y^s_{l'm'}(\theta_n,\phi_n) \end{vmatrix} \\ &= \left(\frac{1}{N\sigma^2}\right)^K \det\sum_{n=1}^{N}\begin{vmatrix} Y^c_nY^c_n & Y^c_nY^s_n \\ Y^s_nY^c_n & Y^s_nY^s_n \end{vmatrix}\,. \end{aligned} \quad (44)$$

Thus we can write the geometric contribution as

$$G = \frac{K}{2}\ln\frac{N}{2\pi} - \frac{K}{2}\ln N\sigma^2 + \ln\int d\{c_{lm}\}\int d\{s_{lm}\} + \frac{1}{2}\ln\det\sum_{n=1}^{N}\begin{vmatrix} Y^c_nY^c_n & Y^c_nY^s_n \\ Y^s_nY^c_n & Y^s_nY^s_n \end{vmatrix}\,.$$

If we denote the parameter integrals as, e.g., $\int dc_{lm} = C_{lm}$, we can finally write

$$G = -K\ln\sigma\sqrt{2\pi} + \sum_{lm}\ln C_{lm} + \sum_{lm}\ln S_{lm} + \frac{1}{2}\ln\det\sum_{n=1}^{N}\begin{vmatrix} Y^c_nY^c_n & Y^c_nY^s_n \\ Y^s_nY^c_n & Y^s_nY^s_n \end{vmatrix}\,. \quad (45)$$

If, as noted above, the sampled values should provide an approximation to the orthogonality relation integral

$$\int d\Omega\, Y^c_{lm}Y^c_{l'm'} = |\text{identity matrix}|_{lm,l'm'}\ ,$$

then with $\Delta\Omega \approx 4\pi/N$, we can obtain the approximate result

$$\begin{aligned} G &= -K\ln\sigma\sqrt{2\pi} + \sum_{lm}\ln C_{lm} + \sum_{lm}\ln S_{lm} + \frac{1}{2}\ln\left(\frac{N}{4\pi}\right)^K \\ &= -K\ln\sigma\sqrt{2\pi} + \sum_{lm}\ln C_{lm} + \sum_{lm}\ln S_{lm} + \frac{K}{2}\ln\frac{N}{4\pi}\,. \end{aligned} \quad (46)$$

Observations. We can see that it is almost trivial to *test* to see whether or not least-squares fitting should be performed rather than numerical projection for orthogonal polynomials: for projection to be a valid approximation, the numerical sum over the independent variables must give a (user-definable) sufficient approximation to the orthogonality relation for the bare basis functions, independent of any measured data or model selection.

We note also that if *complex* functions such as the classical spherical harmonics are used instead of the real cosine-like and sine-like harmonic combinations, one finds experimentally that it is necessary to use complex conjugate pairs of Y_{lm}'s in the matrix $I_{lm,l'm'}$ in order to get positive definite numerical results.

If *continuous* Fourier expansions are used as models, the determination of quantities such as the functional integrals over the coefficients required in the MDL procedure appears to be an open question for future research.

5 Gaussian Models

Our examples so far have all been linear in the coefficients, so that the derivatives in the Fisher matrix computation eliminate the parameter dependence, and nothing particularly interesting happens in the integration. In this section, we treat a new class, the Gaussian models, which are very important data models in their own right, and exhibit new and non-trivial behavior in their parameter derivatives. Unfortunately, it is also much more difficult to determine reliable least-squares fits; a single Gaussian's parameters can be determined by a polynomial least squares fit to the logarithm, but sums of Gaussians require more general methods such as Levenberg-Marquardt optimization (see, e.g., [4]).

We choose as our general model a sum of $K/3$ Gaussians, with K parameters in total, of the following form:

$$y = g(x, \theta) = \sum_{k=1}^{K/3} a_k \exp\left(-\frac{(x - b_k)^2}{2c_k^2}\right) . \tag{47}$$

This can easily be generalized to use a D-dimensional independent variable $\mathbf{x}$ by extending b_k to a D-dimensional vector $\mathbf{b}_k$. This increases the number of parameters per Gaussian to $2 + D$ instead of 3.

The calculation of the description length follows the usual procedure: assume that the Gaussian distribution itself has random errors described by a normal distribution with standard deviation σ and carry out a least-squares fit procedure to get the maximum-likelihood parameter estimates $\{\hat{a}_k, \hat{b}_k, \hat{c}_k\}$. (We assume the total error model is given by a single σ, though we could choose different ones for different values of K if we wished.) The data-fitting term for N outcomes $\{(x_n, y_n)\}$ is

$$F(K) \;=\; N \ln \sigma\sqrt{2\pi}$$

$$+\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n-\sum_{k=1}^{K/3}\hat{a}_k\exp(-\frac{(x_n-\hat{b}_k)^2}{2\hat{c}_k^2}\right)^2 . \tag{48}$$

The $K\times K$ geometric term matrix is

$$I_{ij} \;=\; \frac{1}{N\sigma^2}\sum_{n=1}^{N}\left|\begin{array}{c} A_1(x_n)\\ B_1(x_n)\\ C_1(x_n)\\ \vdots\end{array}\;\otimes\;A_1(x_n)\,B_1(x_n)\,C_1(x_n)\ldots\right| \tag{49}$$

where

$$\begin{aligned} A_k(x) \;&=\; \frac{\partial g}{\partial a_k} \;&=\;& e^{\frac{-(x-b_k)^2}{2\,c_k^2}}\\ B_k(x) \;&=\; \frac{\partial g}{\partial b_k} \;&=\;& \frac{a_k\,(x-b_k)}{c_k^2}\,e^{\frac{-(x-b_k)^2}{2\,c_k^2}}\\ C_k(x) \;&=\; \frac{\partial g}{\partial c_k} \;&=\;& \frac{a_k\,(x-b_k)^2}{c_k^3}\,e^{\frac{-(x-b_k)^2}{2\,c_k^2}} . \end{aligned}$$

We denote the allowed integration domains by $a_{\min}\le a\le a_{\max}$, $b_{\min}\le b\le b_{\max}$, $c_{\min}\le c\le c_{\max}$, and note that, for each triple of parameters, there is an overall factor of a^4/c^{10} in the determinant of I_{ij}; thus the argument of the logarithm in $G(K)$ is an integral of the form

$$\begin{aligned} V(K) \;&=\; \int d^{K/3}\mathbf{a}\int d^{K/3}\mathbf{b}\int d^{K/3}\mathbf{c}\sqrt{\prod_{k=1}^{K/3}a_k^4/c_k^{10}\,\det|sum\ of\ exponentials|}\\ &=\; \prod_{k=1}^{K/3}\int da_k\,a_k^2\int db_k\int dc_k\,c_k^{-5}\sqrt{\det|sum\ of\ exponentials|}\,. \end{aligned}$$

6 Models with the Same Number of Parameters

For completeness, we summarize here the comparison of the Fechner and Stevens models presented by Pitt, Myung, and Zhang [14]; these models each have only two parameters, and the problem of whether one or the other is a better description of a given body of psychophysics data had long been an unanswerable question. We shall see that, while standard analysis overwhelmingly favors one model over the other, no matter what the source of the data, MDL can clearly distinguish them.

Goodness of Fit for the Fechner and Stevens Models. The Fechner model,

$$y = a\ln(x+b)\ ,$$

and the Stevens model,

$$y = ax^b$$

both have two parameters and can in principle describe the same data. Assuming corresponding probability distributions

$$\begin{aligned} f^{\rm F}(y|a,b) &= \frac{1}{\sigma\sqrt{2\pi}}\exp -\frac{1}{2\sigma^2}\left(y - a\ln(x+b)\right)^2 \\ f^{\rm S}(y|a,b) &= \frac{1}{\sigma\sqrt{2\pi}}\exp -\frac{1}{2\sigma^2}\left(y - ax^b\right)^2\ , \end{aligned}$$

the goodness of fit term for a body of data with maximum-likelihood parameters $(\hat{a},\hat{b})$ is

$$\begin{aligned} F_{\text{Fechner GOF}} &= -\ln f(\{y_n\}|\hat{a},\hat{b}) \\ &= N\ln\sigma\sqrt{2\pi} + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \hat{a}\ln(x_n+\hat{b}))^2 \\ &= N\ln\sigma\sqrt{2\pi} + \frac{N}{2\sigma^2}(\text{variance}) \end{aligned} \tag{50}$$

for the Fechner model, with an obvious analogous expression for the Stevens model.

Geometric Terms. The geometric term is easily seen to take the general form

$$L_{ij}(a,b,\,\mathbf{x}) = \frac{1}{\sigma^2}\begin{vmatrix} (\frac{\partial}{\partial a}g(a,b,\,\mathbf{x}))^2 & \frac{\partial}{\partial a}g(a,b,\,\mathbf{x})\cdot\frac{\partial}{\partial b}g(a,b,\,\mathbf{x}) \\ \frac{\partial}{\partial a}g(a,b,\,\mathbf{x})\cdot\frac{\partial}{\partial b}g(a,b,\,\mathbf{x}) & (\frac{\partial}{\partial b}g(a,b,\,\mathbf{x}))^2 \end{vmatrix}\ .$$

For the Fechner model, with $E(y_n) = a\ln(b+x_n)$, the relevant matrix term becomes

$$L_{ij}^{\text{Fechner}}(a,b,x_n) = \frac{1}{\sigma^2}\begin{vmatrix} (\ln(b+x_n))^2 & a\frac{\ln(b+x_n)}{b+x_n} \\ a\frac{\ln(b+x_n)}{b+x_n} & \frac{a^2}{(b+x_n)^2} \end{vmatrix}\ , \tag{51}$$

while for the Stevens model, with $E(y_n) = ax_n^b$, the matrix is

$$L_{ij}^{\text{Stevens}}(a,b,x_n) = \frac{1}{\sigma^2}\begin{vmatrix} x_n^{2b} & ax_n^{2b}\ln x_n \\ ax_n^{2b}\ln x_n & a^2x_n^{2b}(\ln x_n)^2 \end{vmatrix}\ . \tag{52}$$

For sample sizes two or greater, we average over the values of x_n to find the corresponding 2×2 matrix

$$I_{ij}(a,b) = \frac{1}{N}\sum_{n=1}^{N} L_{ij}(a,b,x_n)\ . \tag{53}$$

The geometric term for each model is determined from the integral over the $K = 2$ dimensional parameter space in the expression

$$G = \ln \frac{N}{2\pi} + \ln \int da \int db \sqrt{\det I(a,b)} \,. \tag{54}$$

Thus we find for the Fechner model, $y = a \ln(x + b)$,

$$\begin{aligned} G_{\text{Fechner}} &= -\ln 2\pi\sigma^2 + \ln \int a\, da \int F(b)\, db \\ F^2(b) &= \left(\sum_{n=1}^{N} (\ln(x_n + b))^2 \right) \left(\sum_{n=1}^{N} (x_n + b)^{-2} \right) - \left(\sum_{n=1}^{N} \frac{\ln(x_n + b)}{(x_n + b)} \right)^2 , \end{aligned} \tag{55}$$

and for the Stevens model, $y = ax^b$,

$$\begin{aligned} G_{\text{Stevens}} &= -\ln 2\pi\sigma^2 + \ln \int a\, da \int S(b)\, db \\ S^2(b) &= \left(\sum_{n=1}^{N} (x_n)^{2b} \right) \left(\sum_{n=1}^{N} (x_n)^{2b} (\ln x_n)^2 \right) - \left(\sum_{n=1}^{N} (x_n)^{2b} \ln x_n \right)^2 . \end{aligned} \tag{56}$$

Comparison and Analysis. The comparison of these two two-parameter models can now be seen to reduce to the comparison of the two integrals over b: we may assume that, if the model choice is ambiguous, the two variance terms are comparable, and that the overall contribution of the scaling coefficient a is also the same. Hence the difference is

$$\Delta_{(\textbf{Stevens} - \textbf{Fechner})} = \ln \int S(b)\, db - \ln \int F(b)\, db \,. \tag{57}$$

Pitt et al. [14] observe that the integral over b from $(0 \to \infty)$ diverges for $S(b)$, requiring an ad hoc choice of finite integration domain, while the integral converges for $F(b)$, so no such choice is necessary. With a reasonable choice of integration domain ($0 \leq b \leq 3$, to be precise), and random samples drawn from the Stevens and Fechner distributions, respectively, the full MDL cost equation clearly prefers the model that created the distribution, while the Stevens model is overwhelmingly chosen over Fechner in all cases if only the goodness-of-fit is taken into account.

7 Remarks and Future Work

The Minimum Description Length criterion for model selection has the remarkable property that it can be formulated in a way, e.g., using the Fisher information matrix as a metric, that does not depend in any essential way on reparameterizations of the models; unlike

many standard methods, the MDL procedures presented here are not deceived by disguises, and so confusions that can arise from subtle transformations are avoided. Furthermore, it is often possible to distinguish *a priori* among competing models to select the model that was most likely to have produced the original distribution, even when a much lower maximum-likelihood fitting error results from overfitting with a more complex model.

However, there are a number of overall problems to be addressed in practical applications of the method. Among these, we note particularly the following:

- **Parameter Ranges.** As we have seen in many examples such as the Stevens model, the parameter values must often be restricted to obtain finite integrals for the geometric term. This technically invalidates the reparameterization invariance. This problem is well-known and various attempts have been made to address it: Rissanen [18], for example, discusses the issue and suggests possible correction terms; other solutions (Myung, private communication) might be to approximate the integral over the determinant using the value of the determinant at the maximum-likelihood point (though this again invalidates reparameterization invariance), or to seek alternative metrics to replace the Fisher information matrix, optimally selected according to some meta-criteria that are consistent with the rest of the MDL procedure. An elegant solution for regression problems has been found by Liang and Barron (this volume); see also Lanterman's contribution (this volume).

- **Sample Sizes.** The MDL formalism that is the basis for the equations we have used is the result of a very sophisticated mathematical analysis, and is valid only for asymptotically large sample sizes. The accuracy of the basic formulas is therefore suspect for the frequently-occurring case of small samples. Correction terms are known, but just how to handle small data samples has not been completely understood.

- **Model Complexity Computation.** The mathematical foundation of the geometric complexity terms we have used is deeply rooted in the mathematics of functional forms, functional integrals, and functional measures (see, e.g., Balasubramanian [3]); while these methods are used extensively in relativistic quantum field theory for simple subclasses of integrands, the general analysis is very poorly understood and lies at the limits of current mathematical methods. There are very likely many details, such as the treatment of unusual probability distributions and error distributions, that remain to be properly analyzed.

Acknowledgments

We are indebted to In Jae Myung, Yvan Leclerc, and Pascal Fua for helpful comments. A.J.H. is particularly grateful to Pascal Fua for his kind hospitality at EPFL while this

work was being completed, and C.W.F. thanks P.A. Heng and T.T. Wong for their corresponding hospitality at CUHK. As this manuscript was being prepared, we learned with great sadness that Yvan Leclerc, an old friend as well as one of the pioneers of practical MDL applications, had passed away; he will be missed.

References

[1] AGGRAWAL, R., GEHRKE, J., GUNOPULOS, D., AND RAGHAVAN, P. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. ACM SIGMOD Int. Conf. on Managment of Data* (1998), pp. 94–105.

[2] ARFKEN, G. *Mathematical Methods for Physicists*, 3rd ed. Academic Press, Orlando, FL, 1985. Sections 12.6 and 12.9: Spherical Harmonics and Integrals of the Products of Three Spherical Harmonics.

[3] BALASUBRAMANIAN, V. Statistical inference, occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation 9* (1997), 349–368.

[4] BATES, D. M., AND WATTS, D. G. *Nonlinear Regression and Its Applications*. Wiley, New York, 1988.

[5] FUA, P., AND HANSON, A. An optimization framework for feature extraction. *Machine Vision and Applications 4* (1991), 59–87.

[6] HANSEN, M. H., AND YU, B. Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*, 454 (2001), 746–774.

[7] LECLERC, Y. Constructing simple stable description for image partitioning. *International Journal of Computer Vision 3* (1989), 73–102.

[8] LECLERC, Y., LUONG, Q., AND FUA, P. Self-consistency and MDL: a paradigm for evaluating point correspondences and detecting change, 2003.

[9] MEHTA, M., RISSANEN, J., AND AGRAWAL, R. MDL-based decision tree pruning. In *Int'l Conf. on Knowledge Discovery in Databases and Data Mining (KDD-95)* (August 1995), pp. 216–221. Montreal, Canada.

[10] MYUNG, I. The importance of complexity in model selection. *Journal of Mathematical Psychology 44*, 1 (2000), 190–204.

[11] MYUNG, I., BALASUBRAMANIAN, V., AND PITT, M. Counting probability distributions: Differential geometry and model selection. *Proc. Natl. Acad. of Sci. 97*, 21 (October 2000), 11170–11175.

[12] MYUNG, I., FORSTER, M., AND BROWNE, M. A special issue on model selection. *Journal of Mathematical Psychology 44* (2000).

[13] MYUNG, I. J., PITT, M. A., ZHANG, S., AND BALASUBRAMANIAN, V. The use of MDL to select among computational models of cognition. In *Advances in Neural Information Processing Systems 13* (2001), MIT Press. (Accepted for publication).

[14] PITT, M. A., MYUNG, I. J., AND ZHANG, S. Toward a method of selecting among computational models of cognition. *Psychological Review 109*, 3 (2002), 472–491.

[15] RISSANEN, J. A universal prior for integers and estimation by minimal description length. *Annals of Statistics 11* (1983), 416–431.

[16] RISSANEN, J. Stochastic complexity and modeling. *Annals of Statistics 14*, 3 (1986), 1080–1100.

[17] RISSANEN, J. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Teaneck, NJ, 1989.

[18] RISSANEN, J. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42* (1996), 40–47.

[19] RISSANEN, J. Hypothesis selection and testing by the MDL principle. *The Computer Journal 42* (1999), 260–269.

[20] RISSANEN, J. Lectures on statistical modeling theory, 2001. http://www.cs.tut.fi/%7Erissanen/papers/lectures.ps.

[21] RISSANEN, J. Strong optimality of the normalized ML models as universal codes and information in data. *IEEETIT: IEEE Transactions on Information Theory 47 (5)* (July 2001), 1712–1717.

[22] RITCHIE, D., AND KEMP, G. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comp. Chem. 20* (1999), 383–395. http://www.math.chalmers.se/ kemp/publications/.