

John C. Dunn

## Model complexity: The fit to random data reconsidered

Received: 10 October 1998 / Accepted: 22 December 1998

**Abstract** A recent controversy in the field of depth perception has highlighted an important aspect of model testing concerning a model's complexity, defined as the prior propensity of the model to fit arbitrary data sets. The present article introduces an index of complexity, called the *mean minimum distance*, defined as the average squared distance between an arbitrary data point and the prediction range of the model. It may also be expressed as a dimensionless quantity called the *scaled mean minimum distance*. For linear models, theoretical values for the scaled mean minimum distance and the variance of the scaled minimum distance can be readily obtained and compared against empirical estimates obtained from fits to random data. The approach is applied to resolving the question of the relative complexity of the Linear Integration model and the Fuzzy Logic of Perception model, both of which have been the subject of controversy in the field of depth perception. It is concluded that the two models are equally complex.

### Introduction

Progress in experimental psychology depends on the development and evaluation of mathematical models (Ratcliff, 1998). Such models are necessary both to provide a precise description of the essential features of data and to reveal deeper insights into the nature of the underlying mechanisms or psychological processes. The former goal is realised when a model is found that fits the data well. The latter goal is realised when the model says something important about how the data arise.

Whereas model-fitting is relatively objective, evaluation of the theoretical value of a model is more subjective. Terms such as simplicity, parsimony, beauty, and consistency are often used to capture this aspect of model or theory.

Theorists approach model development from differing perspectives. The assumptions they make in formulating a mathematical model reflect their differing ideas concerning how the data have been generated. Ultimately, however, these different approaches must be tested against the data. If a theory is in disagreement with how the world actually is, it must be rejected or modified. For mathematical models, this means they must stand or fall as a consequence of the extent to which they fit the observed data. However, this supposes that the fits of different models may be directly compared. Only when different models have equal *propensities* to fit data can the extent to which they actually do fit the data be used to adjudicate between them. However, what if a model, by its very nature, is more likely to fit any kind of data? In the extreme, we can imagine a model capable of fitting any conceivable set of data. Such models are generally regarded as vacuous, uninformative, or unscientific (Popper, 1959). Therefore, goodness-of-fit cannot be the only criterion by which models are judged. Their relative propensities to fit data must also be taken into account. This issue has been highlighted recently in the field of experimental psychology in relation to two models of depth perception.

The two models in question specify how depth cues are integrated to form an overall judgment of perceived depth. According to the Linear Integration Model (LIM) proposed by Cutting and his co-workers (Bruno & Cutting, 1988; Cutting, Bruno, Brady, & Moore, 1992), parameters representing the value of different depth cues are combined additively to form a judgment of overall depth. According to the Fuzzy Logical Model of Perception (FLMP) proposed by Massaro and his co-workers (eg., Massaro & Cohen, 1993), a corresponding set of parameters are combined non-linearly using

J. C. Dunn  
Department of Psychiatry and Behavioural Science,  
University of Western Australia,  
WA, Australia, 6907;  
Tel.: +618 9346 2251; Fax: +618 9346 3828  
E-mail: jdunn@cyllene.uwa.edu.au;

equations derived from the field of fuzzy logic.<sup>1</sup> Cutting et al. (1992) investigated how well the two models fit data collected by Bruno and Cutting that consist of judgments of overall depth obtained under 16 different combinations of depth cues (defined by an orthogonal combination of the presence and absence of four different cues). Overall, the two models were found to fit these data equally well.

On the face of it, both the LIM and the FLMP appear to provide adequate explanations of the experimental data. However, Cutting et al. (1992) suggested that the relatively good fit of the FLMP was a consequence of its greater propensity to fit any conceivable set of data. Massaro and Cohen (1993) have used the term *superpower* to describe this feature of a model, while Myung and Pitt (1997) refer to it as the *complexity* of a model and define it as “the flexibility inherent in a model that enables it to fit diverse patterns of data” (p. 80). If the FLMP has greater complexity in this sense, then a good fit to the data may not be unexpected, and the fact that it provides as good a fit to the data as a less powerful and less complex model should be counted as evidence *in favour* of the simpler model.

As the example illustrates, model complexity is an important issue in model comparison. Only when models are equally complex, in the sense of Myung and Pitt (1997), is it possible to adjudicate between them solely on the basis of goodness-of-fit. Yet, this issue is often given little attention by researchers interested in comparing or evaluating several different models. If model complexity is taken into account, it is often done so only informally. For example, a researcher may note that one of the models may contain more parameters than another and hence be more complex. Yet, it is known that the number of free parameters of a model is not an unambiguous guide to its complexity or testability (Bamber & van Santen, 1985).

Cutting et al. (1992) attempted to assess the complexity of the LIM and FLMP in a variety of ways (for a summary, see Cutting, in press). In one approach, the two models were fit to random data. That is, the 16 data points were replaced by numbers selected randomly over the range from zero to one, and the two models were fit to these data. It was found that the FLMP enjoyed a slight advantage over the LIM in terms of its average goodness-of-fit to random data. Yet this approach has proven controversial. Li, Lewandowsky and DeBrunner (1996) argued that “the use of arbitrary data appears inelegant and is subject to criticisms about the adequacy of the particular functions chosen for data generation” (p.361). The results of the analysis were also dismissed by Massaro and Cohen (1993), who pointed out that both the LIM and the FLMP provide unacceptably poor fits to almost all the random data. They concluded that “the simulations (with random noise) show that the

FLMP is not superpowerful because it does not give an acceptable description of any possible result” (p.122). Recently, Cutting (in press) has accepted both these criticisms.

Rejection of the approach of fitting models to random data may be premature. The aim of the present article is to argue that far from being ad hoc or inelegant, this approach is based on sound logic. In particular, it is suggested that the propensity of a model to provide a good fit to any conceivable data is exactly what is meant by the concept of *model complexity*. The idea is that a model can be represented as a high-dimensional response surface embedded in a higher-dimensional space defined by the set of outcome variables. It is proposed that the response surface of a complex model traverses outcome space in such a way that the average discrepancy between it and a random data point is relatively small. From this view, the mean discrepancy between arbitrary data and a model is inversely related to its complexity. The aim of the present article is to explore this idea and use it to determine the relative complexity of the LIM and the FLMP.

---

### Model complexity

Following Myung and Pitt (1997), model complexity is defined as the capacity of a model to fit arbitrary patterns of data. This concept can be most readily appreciated by relating it to the definition of a model proposed by Bamber and van Santen (1985). They distinguish between qualitative and quantitative models. A qualitative model generates ordinal predictions of data. For example, a qualitative model of reaction times may specify that response times in one condition are always greater than response times in another condition. A quantitative model, on the other hand, generates exact predictions of data. According to Bamber and van Santen, a quantitative model can be defined as the ordered triple,  $(\mathbf{P}, F, \mathbf{Q})$ , where, for positive integers  $m$  and  $n$ ,  $\mathbf{P} \subseteq R^m$  is a parameter domain consisting of all conceivable combinations of the  $m$  parameter values of the model,  $F$  is a prediction function defined on  $\mathbf{P}$  such that  $F(\mathbf{P}) \subseteq \mathbf{Q}$ , and  $\mathbf{Q} \subseteq R^n$  is an outcome space consisting of all conceivable combinations of the values of  $n$  different outcome quantities. The set  $\mathbf{R} = F(\mathbf{P})$  is called the model's prediction range.

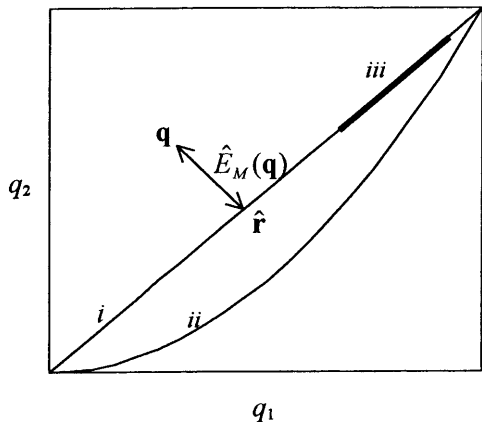
Geometrically, the prediction range of a model corresponds to an  $m$ -dimensional response surface in  $n$ -dimensional outcome space. To illustrate this, consider the following model:

$$\begin{aligned} q_1 &= p \\ q_2 &= p \end{aligned} \tag{1}$$

The model defined by Equation 1 states that two outcome variables,  $q_1$  and  $q_2$ , can be represented by a single parameter  $p$ . The parameter domain of this model is a sub-set of the real number line,  $\mathbf{P} \subseteq R^1$ , the outcome

---

<sup>1</sup> The form of this function is described in Cutting, Bruno, Brady, and Moore (1992), Massaro and Cohen (1993), and Myung and Pitt (1997).



**Fig. 1** The response surfaces of three models: *i* = the linear model defined by Equation 1; *ii* = the non-linear model defined by Equation 5; *iii* = a restricted version of the linear model defined by Equation 1

space is a sub-set of two-dimensional space,  $\mathbf{Q} \subseteq R^2$ , and the prediction range  $\mathbf{R}$  is a one-dimensional subset of the outcome space given by the line  $q_1 - q_2 = 0$ . This is shown in Fig. 1 by the line labelled *i*, which corresponds to the response surface of the model. The model proposes that all outcome quantities,  $q_1$  and  $q_2$ , should lie on this line.

Since the outcome space of a model is defined as the set of all conceivable combinations of outcome quantities, each and every point in this space is a possible data point. Whenever an experiment is performed, one or more points in outcome space are identified, and the model is fitted by finding points in its prediction range that are “closest” to the data in terms of a well-defined discrepancy or error function. An example of a frequently used discrepancy measure is the sum of squared difference. Let  $\mathbf{q} = (q_1, q_2, \dots, q_n)$  be a point in outcome space, and let  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  be a point in the model’s prediction range. The sum of squared difference between  $\mathbf{q}$  and  $\mathbf{r}$  is defined as:

$$E(\mathbf{q}, \mathbf{r}) = \sum_i^n (q_i - r_i)^2 \tag{2}$$

A model’s goodness-of-fit is evaluated by finding the point in the model’s prediction range that minimises the sum of squared difference. Let  $M$  be a model, let  $\mathbf{R}$  be its prediction range, and let  $\mathbf{q} \in \mathbf{Q}$  be a point in its associated outcome space. Then,

$$\hat{E}_M(\mathbf{q}) = \min_{\mathbf{r} \in \mathbf{R}} E(\mathbf{q}, \mathbf{r}) \tag{3}$$

$\hat{E}_M(\mathbf{q})$  is called the minimum distance of  $M$  associated with the point  $\mathbf{q} \in \mathbf{Q}$ . An example of one such distance is illustrated in Fig. 1 with respect to the model defined by Equation 1. If the point in  $\mathbf{R}$  that minimises Equation 2 is denoted by  $\hat{\mathbf{r}} = (\hat{r}_1, \hat{r}_2, \dots, \hat{r}_n)$ , Equation 3 may also be written:

$$\hat{E}_M(\mathbf{q}) = \sum_i^n (q_i - \hat{r}_i)^2 \tag{3}$$

Model complexity can now be thought of in terms of the extent to which the prediction range of a model is “close to” arbitrary points in outcome space. An intuition for this idea can be gained by considering the three aspects of a model that Myung and Pitt (1997) identify as affecting its complexity. These are the number of parameters,  $m$ , the form of the prediction function,  $F$ , and the extension of the parameter domain,  $\mathbf{P}$ .

As the number of parameters of a model increases, its prediction range becomes more extensive and hence, all things being equal, closer to more of the outcome space. The prediction range of the one-parameter model defined by Equation 1 is a one-dimensional line. Consider the following two parameter model:

$$\begin{aligned} q_1 &= p_1 \\ q_2 &= p_2 \end{aligned} \tag{4}$$

Its prediction range corresponds to a two-dimensional surface in outcome space. In fact, since outcome space occupies only two dimensions in this example, the minimum distance between any data point and this model’s prediction range must always be zero. More generally, as the number of parameters of a model increases, the set of minimum distances will tend to decrease.

Complexity is also related to the form of the model’s prediction function,  $F$ . Consider the following one-parameter model:

$$\begin{aligned} q_1 &= p \\ q_2 &= p^2 \end{aligned} \tag{5}$$

The response surface corresponding to the prediction range of this model is described by the line labelled *ii* in Fig. 1. The point to note is that over the chosen range of values<sup>2</sup> for  $q_1$  and  $q_2$ , points on this line are, on average, further from points in the outcome space than are points on line *i*. In this sense, the model defined by Equation 5 is *less complex* than the apparently simpler model defined by Equation 1.

Finally, complexity is also related to the extension of the model domain. This is shown in Fig. 1 by the portion of line *i* that is marked more thickly and is labelled *iii*. This corresponds to the prediction range of the model given by Equation 1 in which the parameter domain  $\mathbf{P}$  is restricted to a smaller interval. It is clear that the effect of this is to increase the minimum distance for many points in the outcome space – of which point  $\mathbf{q}$ , shown in Fig. 1, is an example. As a consequence, the set of minimum distances will tend to increase overall, leading to the restricted version of the model being *less complex* than the unrestricted version.

<sup>2</sup> It is apparent that  $q_1$  and  $q_2$  are limited to the interval  $[0, 1]$  in Fig. 1.

## Mean minimum distance

An obvious indicator of the overall “closeness” of the prediction range of a model  $M$  to arbitrary points in the outcome space is the mean minimum distance,  $\bar{E}_M$ . This is defined as:

$$\bar{E}_M = \frac{\int \hat{E}_M(\mathbf{q}) d\mathbf{q}}{\int d\mathbf{q}} \quad (6)$$

where  $\hat{E}_M(\mathbf{q})$  is defined in Equation 3 and is integrated over each dimension of the outcome space.<sup>3</sup> In order to evaluate  $\bar{E}_M$ , it is necessary to place bounds on the range of possible outcome values. That is, each outcome quantity  $q_i$  is deemed to lie between a lower bound  $a_i$  and an upper bound  $b_i$ . If all of the quantities are commensurable or correspond to measurements of a single variable under  $n$  different conditions, then it is possible to specify a common upper and lower bound. Let  $a$  be the common lower bound, and let  $b$  be the common upper bound. The denominator of Equation 6 simplifies to  $(b - a)^n$ , where  $n = \dim(\mathbf{Q})$  and

$$\bar{E}_M = \frac{1}{(b - a)^n} \int_a^b \hat{E}_M(\mathbf{q}) d\mathbf{q} \quad (7)$$

As defined by Equation 7, the size of  $\bar{E}_M$  depends upon the choice of a measurement scale. Since it is convenient to think of complexity as scale independent, it is useful to express  $\bar{E}_M$  as a dimensionless quantity by re-scaling the outcome space to the unit (hyper) cube, effectively setting  $a = 0$  and  $b = 1$ . This is called the scaled mean minimum distance and is defined as:

$$\bar{S}_M = \frac{\bar{E}_M}{(b - a)^2} \quad (8)$$

Alternatively,  $\bar{S}_M$  may be obtained directly from Equation 3 by integrating over the interval  $[0, 1]$ :

$$\bar{S}_M = \int_0^1 \hat{E}_M(\mathbf{q}) d\mathbf{q} \quad (9)$$

### An example

Consider a model of the inheritance of heights that states that the height of the adult offspring of two parents is equal to the mean of their respective heights. Let  $q_1$  be the height of the father, let  $q_2$  be the height of the mother, and let  $q_3$  be the average height of their (adult) offspring. This model  $M$  is formalised as the triple  $(\mathbf{P}, F, \mathbf{Q})$ , where  $\mathbf{P} \subseteq (R^2)^+$ ,  $\mathbf{Q} \subseteq (R^3)^+$ , and the prediction range  $\mathbf{R} = F(\mathbf{P})$  is given by:

$$\begin{aligned} r_1 &= p_1 \\ r_2 &= p_2 \\ r_3 &= \frac{1}{2}(p_1 + p_2) \end{aligned} \quad (10)$$

where  $\mathbf{r} = (r_1, r_2, r_3)$  is an element of  $\mathbf{R}$ , and  $\mathbf{p} = (p_1, p_2)$  is an element of the domain  $\mathbf{P}$ . Let  $\mathbf{q} = (q_1, q_2, q_3)$  be an element in the outcome space  $\mathbf{Q}$ . The sum of squared difference between  $\mathbf{q}$  and  $\mathbf{r}$  is

$$E(\mathbf{q}, \mathbf{r}) = (q_1 - p_1)^2 + (q_2 - p_2)^2 + \left(q_3 - \frac{1}{2}(p_1 + p_2)\right)^2 \quad (11)$$

Differentiating  $E(\mathbf{q}, \mathbf{r})$  with respect to  $\mathbf{p}$  and setting the result to zero, the minimum distance can be expressed as a function of  $\mathbf{q}$ :

$$\hat{E}_M(\mathbf{q}) = \frac{1}{6}(q_1 + q_2 - 2q_3)^2 \quad (12)$$

To find the mean minimum distance,  $\hat{E}_M(\mathbf{q})$  must be integrated over all possible values of  $\mathbf{q}$ , according to Equation 6. In order to do so, a realistic common upper and lower bound must be specified. For the sake of the exercise, let  $a = 150$  cm be the lower bound, and let  $b = 225$  cm be the upper bound. Substituting these values into Equation 6 gives:

$$\begin{aligned} \bar{E}_M &= \frac{1}{(b - a)^n} \int_a^b \hat{E}_M(\mathbf{q}) d\mathbf{q} \\ &= \frac{1}{(225 - 150)^3} \int_{150}^{225} \int_{150}^{225} \int_{150}^{225} \frac{1}{6}(q_1 + q_2 - 2q_3)^2 dq_1 dq_2 dq_3 \\ &= 468.75 \end{aligned} \quad (13)$$

That is, over all conceivable data triples between 150 cm and 225 cm, the mean fit of the model is 468.75 cm<sup>2</sup>. Since this is range dependent, it is convenient to calculate the scaled mean minimum distance using Equation 8:

$$\begin{aligned} \bar{S}_M &= \frac{\bar{E}_M}{(225 - 150)^2} \\ &\approx 0.083 \end{aligned} \quad (14)$$

In fact, the scaled mean minimum discrepancy is exactly equal to  $1/12$ .

### Scaled mean minimum distance of a linear model

For linear models such as the previous example and the LIM, it is possible to obtain a general expression for the scaled mean minimum distance. A linear model is of the form  $M = (\mathbf{P}, F, \mathbf{Q})$ , where  $\mathbf{P} \subseteq R^m$ ,  $\mathbf{Q} \subseteq R^n$ , and  $F : R^m \rightarrow R^n$  is a linear transformation. Let  $\mathbf{p} \in \mathbf{P}$  and let  $\mathbf{r} \in \mathbf{R}$ . Then,  $\mathbf{r} = F(\mathbf{p}) = \mathbf{A}\mathbf{p}$ , where  $\mathbf{A}$  is an  $n \times m$  matrix of  $F$ . For any  $\mathbf{q} \in \mathbf{Q}$ , the point  $\hat{\mathbf{r}} \in \mathbf{R}$  that minimises the sum of squared difference  $E(\mathbf{q}, \mathbf{r})$ , is given by:

$$\hat{\mathbf{r}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{q} \quad (15)$$

<sup>3</sup>  $d\mathbf{q} = dq_1, dq_2 \dots dq_n$ .

where  $\mathbf{A}'$  is the transpose of  $\mathbf{A}$ . From this, the minimum distance is given by:

$$\hat{E}_M(\mathbf{q}) = \mathbf{q}'\mathbf{B}\mathbf{q} \quad (16)$$

where  $\mathbf{B}$  is the orthogonal projection matrix:

$$\mathbf{B} = \mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \quad (17)$$

Using Equations 9 and 16, the scaled mean minimum distance of a linear model can be calculated:

$$\begin{aligned} \bar{S}_M &= \int_0^1 \hat{E}_M(\mathbf{q}) d\mathbf{q} \\ &= \int_0^1 [\mathbf{q}'\mathbf{B}\mathbf{q}] d\mathbf{q} \\ &= \int_0^1 \cdots \int_0^1 \left[ \sum_i^n \sum_j^n q_i q_j b_{ij} \right] dq_1 \dots dq_n \end{aligned} \quad (18)$$

In order to evaluate this integral, it is helpful to write  $\hat{E}_M(\mathbf{q})$  in the following form:

$$\hat{E}_M(\mathbf{q}) = \sum_i q_i^2 b_{ii} + \sum_i \sum_{j \neq i} q_i q_j b_{ij} \quad (19)$$

Integrating with respect to each element of  $\mathbf{q}$  yields:

$$\begin{aligned} \int \hat{E}(\mathbf{q}) d\mathbf{q} &= \frac{1}{3} \sum_i q_i^3 b_{ii} \prod_{k \neq i} q_k \\ &\quad + \frac{1}{4} \sum_i \sum_{j \neq i} q_i^2 q_j^2 b_{ij} \prod_{k \neq i, j} q_k + C \end{aligned}$$

which, on taking the definite integral over  $[0, 1]$ , simplifies to<sup>4</sup>:

$$\begin{aligned} \bar{S} &= \int_0^1 \hat{E}(\mathbf{q}) d\mathbf{q} \\ &= \frac{1}{3} \sum_i b_{ii} + \frac{1}{4} \sum_i \sum_{j \neq i} b_{ij} \\ &= \frac{1}{12} \text{tr}(\mathbf{B}) + \frac{1}{4} \Sigma(\mathbf{B}) \end{aligned} \quad (20)$$

where  $\Sigma(\mathbf{B})$  is the sum of the elements of  $\mathbf{B}$ , and  $\text{tr}(\mathbf{B})$  is the trace of  $\mathbf{B}$ . To illustrate the application of Equation 20, consider the linear model previously discussed in relation to the heights of parents and their offspring. The design matrix for this model is:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.5 & 0.5 \end{bmatrix} \quad (21)$$

The resultant orthogonal projection matrix is:

$$\mathbf{B} = \frac{1}{6} \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ -2 & -2 & 4 \end{bmatrix} \quad (22)$$

From this it can be readily calculated that  $\bar{S}_M = 1/12$ , in agreement with result found in the previous section.

Variance of the scaled minimum distance of a linear model

Let  $V_M$  be the variance of  $\hat{E}_M(\mathbf{q})$ . This is defined as:

$$V_M = \int_0^1 (\hat{E}_M(\mathbf{q}))^2 d\mathbf{q} - \bar{S}_M^2 \quad (23)$$

For a linear model, the first term on the right-hand side can be obtained using the same logic as that used to obtain  $\bar{S}_M$  in terms of the orthogonal projection matrix  $\mathbf{B}$ . From Equation 16,

$$\hat{E}_M(\mathbf{q}) = \sum_i \sum_j q_i q_j b_{ij} \quad (24)$$

In squaring  $\hat{E}_M(\mathbf{q})$ , each term on the right-hand side of this equation is multiplied by every other term and summed:

$$(\hat{E}_M(\mathbf{q}))^2 = \sum_i \sum_j \sum_k \sum_l q_i q_j q_k q_l b_{ij} b_{kl} \quad (25)$$

It is now necessary to integrate this equation with respect to each element of  $\mathbf{q}$ . Since  $\mathbf{q}$  is bounded by the unit hypercube  $[0, 1]^n$ , the integral depends only on the elements of  $\mathbf{B}$  and the number of equivalent subscripts. The five different ways in which a set of four subscripts may equal one another are listed in Table 1. Associated with each type of combination is a coefficient  $c_{ijkl}$  such that:

$$\int_0^1 (\hat{E}_M(\mathbf{q}))^2 d\mathbf{q} = \sum_i \sum_j \sum_k \sum_l c_{ijkl} b_{ij} b_{kl} \quad (26)$$

Substituting Equations 20 and 26 into Equation 23, the variance of the scaled minimum distance is obtained. Applying this equation to the two-parameter "height" model yields the value of  $V_M \approx 0.009722$ .

Mean minimum distance for the LIM and the FLMP

The linear integration model (LIM) proposed by Cutting and his co-workers (1992) is a linear model. Thus, its theoretical scaled mean minimum distance,  $\bar{S}_{LIM}$ , and the variance of the scaled minimum distance,  $V_{LIM}$ , can be calculated. For the version of the model analysed by Cutting et al., the LIM was defined as a linear mapping from a five-dimensional parameter domain,  $\mathbf{P} = R^5$ , onto a 16-dimensional prediction range,  $\mathbf{R} = (0, 1)^{16}$ . The corresponding design matrix  $\mathbf{A}$  is given

<sup>4</sup> For a common class of linear model, equation 12 simplifies further. If the prediction range of a model,  $M$ , includes the locus of points,  $q_1 = q_2 = \dots = q_n$ , then  $\bar{S}_M = \frac{1}{12}(n - m)$ .

**Table 1** The five coefficients associated with integration of the squared scaled minimum distance (Equation 26)

Paradigm case	Subscript relationship*	Coefficient	No. of instances
All different	$i \neq j \neq k \neq l$	1/16	$n(n-1)(n-2)(n-3)$
Two of a kind	$i = j \neq k \neq l$	1/12	$6n(n-1)(n-2)$
Two pairs	$i = j \neq k = l$	1/9	$3n(n-1)$
Three of a kind	$i = j = k \neq l$	1/8	$4n(n-1)$
Four of a kind	$i = j = k = l$	1/5	$n$

\*The relationships given here are generic and are not to be thought of as literal. For example, “two of kind” also includes  $i \neq j = k \neq l$ ,  $i \neq j \neq k = l$ ,  $i = l \neq j \neq k$ , and so on

in the Appendix<sup>5</sup>. From this, the following values for the scaled mean minimum distance and the variance of the scaled minimum distance can be obtained:

$$\begin{aligned}\bar{S}_{\text{LIM}} &= \frac{11}{12} \\ &\approx 0.91667 \\ V_{\text{LIM}} &\approx 0.08976\end{aligned}\quad (27)$$

Cutting et al. (1992, Simulation 2) generated 1000 random data points in the range 0.001 to 0.999 and obtained an empirical mean minimum distance (based on the sum of squared difference) of 0.932 (Table 5, p. 373). Is this statistically different from the theoretical mean? Since the population mean and variance are known, this hypothesis can be examined using a large-sample  $z$ -test:

$$z = \frac{\bar{S}_{\text{obs}} - \bar{S}_{\text{LIM}}}{\sqrt{\frac{V_{\text{LIM}}}{N}}}\quad (28)$$

where  $N = 1000$ . For  $\bar{S}_{\text{obs}} = 0.932$ ,  $z = 1.618$ ,  $p = 0.106$  (2-tailed). However, sampling points over the range 0.001 to 0.999 is not exactly the same as sampling points from 0 to 1 (although it is close). Thus,  $\bar{S}_{\text{obs}} = 0.932$  is an underestimate. Using Equation 8, an adjusted estimate of the observed scaled mean minimum distance is:

$$\begin{aligned}\bar{S}_{\text{obs}}^* &= \frac{0.932}{(0.999 - 0.001)^2} \\ &\approx 0.936\end{aligned}\quad (29)$$

For this value of  $\bar{S}_{\text{obs}}^*$ ,  $z = 2.013$ ,  $p = 0.044$ . Since this meets one of the normal experimental criteria for a significant effect, it raises the question whether the selected points were biased in some way or the calculated values of the minimum distance for each point were slight overestimates. This point is examined below. Cutting et al. (1992) also report the mean minimum distance of the FLMP, obtained in the same way. The value they found was 0.924, which is not significantly different from the theoretical scaled mean minimum

discrepancy  $z = 0.774$ ,  $p = 0.439$ . Since it is also an underestimate due to sampling only the interval 0.001 to 0.999, it was adjusted using Equation 8. The adjusted value is 0.928, which is again not significantly different from the theoretical value,  $z = 1.165$ ,  $p = 0.244$ .

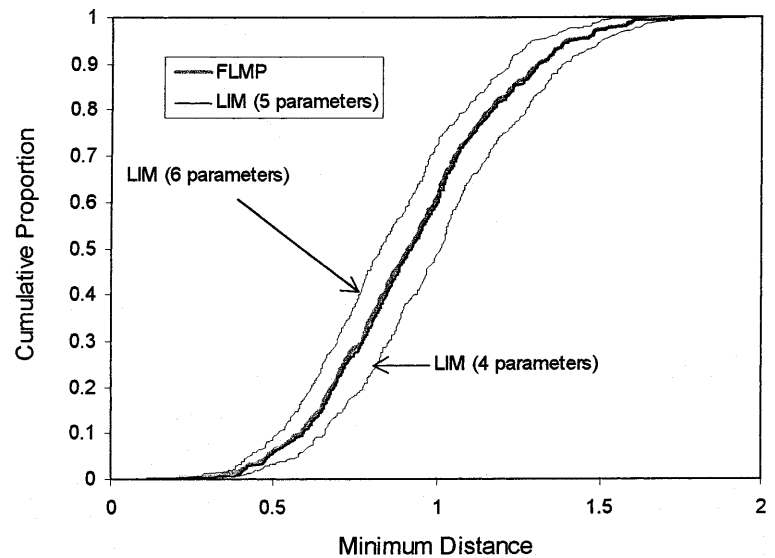
Although the mean minimum distances of both the LIM and the FLMP are close to the theoretical value for a linear model, Cutting et al. (1992) observed that the FLMP provided a better fit to the data on 608 occasions. This is a highly significant departure from the expected number of 500. Similarly, the small difference in the mean minimum distance between the models was also statistically significant.

These observations suggest that the FLMP does enjoy a slight advantage over the LIM in terms of its propensity to fit arbitrary data. However, the mean values obtained by Cutting et al. (1992) are both greater than the theoretical value for a linear model. That is, although  $\bar{S}_{\text{obs}}$  for the FLMP was less than  $\bar{S}_{\text{obs}}$  for the LIM, both of these values are greater than the theoretical value of  $\bar{S}_{\text{LIM}} = 0.917$ , raising the possibility of biased estimation. For this reason, an attempt was made to replicate the analysis performed by Cutting et al. A sample of 500 random data points in the range 0 to 1 were selected and fitted by the LIM and the FLMP. In the case of the LIM, least-squares estimates of the five parameters were calculated directly. In the case of the FLMP, the parameters were estimated using the Microsoft Excel Solver tool. The values of the mean minimum distances were very similar to those obtained by Cutting et al. For the LIM,  $\bar{S}_{\text{obs}} = 0.937$ , and for the FLMP,  $\bar{S}_{\text{obs}} = 0.929$ . In addition, the FLMP fit the data better than the LIM on 294 occasions, a similar proportion to that found by Cutting et al. (0.59 compared to 0.61), and the small difference between the mean minimum distances for the two models was again highly significant,  $F(1,499) = 43.3$ ,  $p < 0.0001$ .

Although the FLMP provides a marginally better fit than does the LIM, both models were found to fit the randomly generated data more poorly than the theoretical value for a linear model (although this difference was not statistically significant). Thus, neither model is superpowerful with respect to the expected theoretical value. This suggests that some other effect may be operating to increase the mean minimum distance for both models and, potentially, to induce a difference between them. One possibility is that the observed mean distances may be partly due to perturbations in computer random-number generation or to slight errors in

<sup>5</sup> In discussing the LIM, Massaro (1998) points out that it is often necessary to specify an additional parameter or set of parameters that correspond to the relative weighting of each depth cue. This arises in experimental designs in which a given cue assumes several different values or levels of discriminability. In the expanded factorial design used by Cutting et al. (1992), each cue assumed only one level of discriminability when present. Consequently, the LIM and the FLMP require the same number of parameters. This greatly simplifies the issue of comparing their complexity.

**Fig. 2** Cumulative proportion ogives as a function of minimum distance for the FLMP, the LIM, and alternative six-parameter and four-parameter versions of the LIM



parameter estimation. Computers are not perfect, and either of these effects may be sufficient to increase the minimum distance estimates for both models and potentially to induce a small bias in favour of one of them.

There are two issues here. The first is whether there is a slight bias in the generation of random data. To investigate this possibility, and since the LIM can be easily fit to data using least-squares solutions, it was fit to a total of 10,000 random data points. The obtained mean scaled minimum distance  $\bar{S}_{\text{obs}} = 0.918$  is considerably closer to the theoretical value. This suggests that the overestimation of the theoretical mean in the previous simulations are likely due to sampling error.

The second issue is that there may be a bias in the distribution of generated data points such that a slight advantage is gained by the FLMP. That is, although the means of the two models are the same, the distributions of the minimum distances for the two models are different. This would occur if for most data points the minimum distance of the FLMP is slightly less, but for a few data points it is considerably longer than the minimum distance of the LIM.

In order to investigate this possibility, the cumulative proportion ogives for the two models were calculated. These are shown in Fig. 2. If the FLMP provides a generally better fit to random data, its cumulative proportion ogive should be shifted leftwards in Fig. 2 relative to the LIM. It is apparent that while there is some evidence of such a shift, by and large the two ogives are almost coincident.

In order to appreciate the magnitude of this small difference, Fig. 2 also shows the cumulative proportion ogives for two other versions of the LIM. In the four-parameter version, two of the four parameters associated with variable depth cues are constrained to be equal, effectively eliminating one parameter from the model (thereby reducing it from five to four free parameters). The effect of this change is to shift the ogive

for the four-parameter model substantially to the right relative to the five-parameter LIM. The theoretical mean and variance of this distribution are  $\bar{S}_{\text{LIM4}} = 1$  and  $V_{\text{LIM4}} \approx 0.09115$ , respectively.

In the six-parameter version, an additional parameter, corresponding to the interaction of the five free parameters, has been added. As Fig. 2 shows, the cumulative proportion ogive for this model has been shifted substantially to the left. The theoretical mean and variance of this distribution are  $\bar{S}_{\text{LIM6}} \approx 0.833$  and  $V_{\text{LIM6}} \approx 0.08328$ , respectively. The relative fits of both these models show that any difference in complexity between the standard five-parameter LIM and the FLMP is considerably smaller than the change in complexity due to adding or removing a single parameter.

It would appear on the basis of the present analysis that if the FLMP enjoys any advantage relative to the LIM in being able to fit arbitrary data points, the size of this advantage is very small. To place this in its context, the observed minimum distances found for the actual experiment performed by Bruno and Cutting (1988) were 0.017 and 0.034 for the LIM and the FLMP, respectively.<sup>6</sup> Both of these values are far closer to zero than any of the minimum distances generated by random data and cannot readily be attributed to differences in the complexity of either model.

The conclusion that the LIM and the FLMP are equally complex appears to disagree with the results found by Myung and Pitt (1997). In this study, versions of the LIM, the FLMP, and a model based on the theory of signal detection (TSD) were fit to data sets generated using the prediction functions of each of the three models. The aim of the study was not to investigate model complexity directly, but to evaluate a technique of model selection based on the Bayes factor (Kass & Raftery, 1995) which, it was argued, would help to overcome the

<sup>6</sup> For data averaged over all subjects.

apparent advantage of complex models. The FLMP was included as an example of a complex model. That is, Myung and Pitt *assumed* that it was more complex than the LIM. As it turned out, the FLMP tended to provide better fits to most of the generated data sets, including data generated by the LIM itself. Yet this result is not inconsistent with the present analysis, since the points in outcome space to which the models were fit were not selected at random by Myung and Pitt. In fact, they corresponded to points in the prediction range of each of the models under consideration.

---

## Discussion

The aim of the present article was to explore the implications of defining model complexity in terms of the prior propensity of a model to fit arbitrary data sets. Although the term complexity enjoys current usage in this context, it may be something of a misnomer, since it implies that models that look complex in terms of their functional form will necessarily tend to fit arbitrary data better. However, complexity in this sense is very much in the eye of the beholder. A functionally complex model from one standpoint may be a functionally simple model from another. For example, the FLMP becomes a linear model if the outcome quantities are converted into logits, while in this transformed space it is the LIM which is functionally complex.<sup>7</sup>

The meaning of the term complexity in the present analysis is closer to “fittability”, or the extent to which a model is already “close” to any conceivable data, prior to any data being actually collected. In this sense, complexity is analogous the idea of logical probability (Popper, 1959), which refers to the largely unsuccessful attempt to assign a kind of prior probability to different theories. From this view, if Theory A is inconsistent with  $X\%$  of all possible outcomes and Theory B is inconsistent with  $Y\%$  of all possible outcomes, and if  $X > Y$ , then Theory A has a lower logical probability than Theory B. According to Popper’s falsificationist approach, theories with low logical probability are more falsifiable and hence constitute better scientific theories.

The problem with this idea is that all quantitative theories are inconsistent with an infinite number of outcomes. As Bamber and van Santen (1985) have shown, unless a quantitative model is saturated (i.e., is inconsistent with no conceivable outcome), it must have a Lebesgue measure of zero. In other words, no matter what the form of the model, the ratio of the set of consistent outcomes to the set of inconsistent outcomes is always zero.

The concept of model complexity, or fittability, overcomes this difficulty. While the prior probability that a randomly selected data point lies in the prediction range of a quantitative model is always zero, the minimum discrepancy or distance between the data and the

prediction range is well defined. If the average discrepancy between a random data point and a model’s prediction range is greater for Model A than for Model B, then Model A is less complex and, by implication, more falsifiable.

The present analysis of model complexity makes several assumptions about the form of outcome space. One assumption is that this space is bounded. For most practical situations, this can be assumed to be true, since outcomes are either logically bounded (e.g., if they are proportions) or are arbitrarily defined as such. For example, while an infinite response time is conceivable (just), in practice researchers tend to exclude response times that are greater than a predefined maximum.

A second assumption, of perhaps greater concern, is that the set of conceivable outcomes is distributed uniformly between the upper and lower bounds. Against this view, experience leads us to the certain knowledge that some outcomes are more likely than others. For example, in a simple reaction-time task, response times of the order of, say, 200–1000 ms are much more likely than response times that lie outside this interval. This consideration suggests that model complexity ought to be evaluated against a prior probability distribution of the data that is not uniform but takes into account knowledge of the likelihood of different data points.

While this approach is feasible, it raises a logical problem due to the fact that expectations concerning the distribution of possible data are not themselves atheoretical. By assuming a uniform distribution of outcomes, model complexity can be evaluated independently of any actual outcome. That is, a complex model that is “close” to any conceivable outcome can be said to enjoy, a priori, an advantage over a less complex model that is, on average, further from these data. If model complexity is evaluated in the context of prior constraints on the distribution of possible outcomes, then model complexity and model fit become inextricably linked. In the limit, as the set of possible outcomes is constrained to approximate more and more the way the world actually is, model complexity will cease to have any meaning. A model will be highly “fittable” simply because it is the correct model.

In conclusion, the average fit of a model to random data, at least over a specified range, rather than being “inelegant” or providing “unacceptable” fits, may convey considerable information concerning the propensity of the model to fit *any* conceivable data. By comparing different models in terms of their scaled mean minimum distances, researchers may alert themselves to, or put their minds at rest concerning, the possibility that their differential fits to data are due to differences in model complexity. If the models in question differ in complexity, then more sophisticated model selection procedures, such as the one proposed by Myung and Pitt (1997), will need to be explored.

**Acknowledgements** I wish to thank my friend and colleague Ralph James for many thoughtful discussions and without whose assis-

---

<sup>7</sup> The FLMP is a linear model in logistic space defined as,  $q_i = \ln(q_i/(1 - q_i))$ .

tance this research would not have been possible. An additional necessary condition was an Australian Research Council Small Grant granted to the author.

## Appendix

The design matrix associated with the Linear Integration Model (LIM) investigated by Cutting et al. (1992). The weights B, S, H, O, and P indicate the presence (1) or absence (0) of depth information derived from background factors, relative size, height in the visual field, occlusion, and motion perspective, respectively

Condition	Weights				
	B	S	H	O	P
1	1	0	0	0	0
2	1	1	0	0	0
3	1	0	1	0	0
4	1	1	1	0	0
5	1	0	0	1	0
6	1	1	0	1	0
7	1	0	1	1	0
8	1	1	1	1	0
9	1	0	0	0	1
10	1	1	0	0	1
11	1	0	1	0	1
12	1	1	1	0	1
13	1	0	0	1	1
14	1	1	0	1	1
15	1	0	1	1	1
16	1	1	1	1	1

## References

- Bamber, D., & van Santen, J.P.H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, *29*, 443–473.
- Bruno, N., & Cutting, J.E. (1998). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, *117*, 161–170.
- Cutting, J.E. (in press). Model selection: Revisiting a case study in the perception of layout. *Journal of Mathematical Psychology: Special issue on model selection*.
- Cutting, J.E., Bruno, N., Brady, N.P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgement of perceived depth. *Journal of Experimental Psychology: General*, *121*, 364–381.
- Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
- Li, S-C., Lewandowsky, S., & DeBrunner, V.E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology: General*, *125*, 360–369.
- Massaro, D.W. (1998). *Perceiving talking faces*. Cambridge, MA: MIT Press.
- Massaro, D.W., & Cohen, M.M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, *122*, 115–124.
- Myung, I.J., & Pitt, M.A. (1977). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.
- Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Ratcliff, R. (1998). The role of mathematical psychology in experimental psychology. *Australian Journal of Psychology*, *50*, 129–130.