Brit. J. Phil. Sci. **50** (1999), 83–102

Model Selection in Science: The Problem of Language Variance

Malcolm R. Forster

ABSTRACT

Recent solutions to the curve-fitting problem, described in Forster and Sober ([1994]), trade off the simplicity and fit of hypotheses by defining simplicity as the paucity of adjustable parameters. Scott De Vito ([1997]) charges that these solutions are 'conventional' because he thinks that the number of adjustable parameters may change when the hypotheses are described differently. This he believes is exactly what is illustrated in Goodman's new riddle of induction, otherwise known as the grue problem. However, the 'number of adjustable parameters' is actually a loose way of referring to a quantity that is *not* language dependent. The quantity arises out of Akaike's theorem in a way that ensures its language invariance.

1 Introduction

Five years ago Elliott Sober and I published an article on the general problem of selecting from amongst a set of quantitative models (Forster and Sober [1994]). The article described a solution to the problem first worked out in detail by Akaike ([1973]). The idea has not been widely publicized even amongst statisticians, until recently. Since then, there has been a growing interest in the subject, especially on the part of scientists themselves.[1] Unfortunately, there is a certain amount of misinformation surrounding the merits of Akaike's approach. Some common misconceptions held by statisticians are addressed in Forster ([to appear]), whereas this paper is directed at mistakes made by philosophers, especially De Vito ([1997]).

I will start from the beginning by first describing how the problem arises in

[1] This interest is reflected in the proceedings of an econometrics conference held in Tilburg, The Netherlands on 9–11 January 1997, and in a forthcoming special issue on model selection in the *Journal of Mathematical Psychology*.

well-known episodes in the history of science (Section **2**). At the same time, the standard philosopher's understanding of the 'curve-fitting problem' does not fit this example. Philosophy may not provide all the answers, but it can help you ask the right question. The right questions about curve-fitting are asked in Section **3**. There is also a common contention that Goodman's new riddle of induction, also known as the grue problem, is the same as the curve-fitting problem. This is critically examined in Section **5**, where I conclude that it is not the same problem as the problem solved by Akaike. De Vito ([1997]) is right to say that Akaike does not solve the grue problem, but the limitations of Akaike's solution do not undermine its positive achievements in the cases to which it does apply. However, the central question remains. Is Akaike's solution language dependent? The final section lays the foundations of the Akaike framework in way that is manifestly language invariant.

2 A methodological puzzle redescribed

Suppose that we want to accurately predict the future positions of the planets relative to the fixed stars. That is the goal. Copernicus, like his predecessor Ptolemy, was an astronomer who appeared to have an foolproof method for achieving this goal. He modelled the planetary motions as being the result of a compounded series of uniform circular motions, which is mathematically similar to representing a function as a Fourier series. There are mathematical theorems to show that for a sufficient number of circles on circles, or epicycles upon epicycles, the true trajectories of the planets can be matched to an arbitrary degree of accuracy. So, why was Copernicus's theory superseded by the more modern theories of Kepler, Newton, and Einstein? One reason is that Newton's and Einstein's theories have important implications in areas of physics beyond the motion of planets. But is this the *only* reason?

Kepler's laws[2] pose this question in a particularly poignant way because the scope of Kepler's laws is limited to planetary motion. Kepler did not supersede Copernicus because his model had greater scope. Moreover, Kepler's laws do not predict planetary positions extremely well, partly because the sun is not quite at the focus of the Kepler's ellipses, partly because of the disturbances of other planets when they pass close by, and partly because of the slow rotation of the ellipses (the precession of the perihelia), which we know today is a relativistic effect. In contrast, Copernicus's theory is able, in principle, to account for all of these effects. So, if Newton and Einstein never came along, would we have to say that Copernicus's theory is better than Kepler's laws? One popular response is to say that Kepler is preferable to

[2] There are three laws. The first says that all the planets revolve in ellipses with the sun at one focus. The second says that the radius from the sun to a planet sweeps out equal areas in equal times. The third says that the ratio of the mean radius cubed to the period (the time for a complete revolution) squared is the same for all planets.

Copernicus because it is simpler. But what does that mean, and why is simplicity important?

To deepen the puzzle, notice that a naïve empiricist rule of 'choosing the best-fitting curve' has to favour Copernicus over Kepler. For Copernicus's theory of planetary astronomy can always add enough circles to fit any set of observations better than Kepler's laws. In fact, it can even fit better than Einstein's theory of gravitation, which is the best theory currently available.

One reaction to the puzzle is to say that the accuracy predictions is not the goal of science. While I agree that there are other goals of science besides predictive accuracy, I dispute the idea that the puzzle *forces* this conclusion upon us. What it really shows is that the theory that fits the observed data best is not necessarily the one that makes the most accurate predictions. Case in point: Copernicus's theory fitted the data better than Kepler's laws, but Kepler's laws made more accurate predictions. The naïve empiricist tacitly assumes that a precise fit of old data is the best way of guaranteeing an accurate fit of new data. This view is so well entrenched that even those who do not think of themselves as empiricists automatically conflate accuracy-in-fitting-old-data with accuracy-in-fitting-new-data without a second thought.

Forster and Sober ([1994]) explain why the naïve empiricist assumption is false. The basic point is that the in-principle ability of a theory to represent reality is quite different from its in-practice ability to do so. The difference between principle and practice arises because modelling in science introduces adjustable parameters whose values have to be estimated from the data. In the case of Copernican astronomy, each circle introduces a radius, period of motion, and initial positions. It takes many circles to overtake Kepler's laws, and with a large number of circles, there are a large number of adjustable parameters. The problem is that the precision of the parameter estimates goes down as the total number of parameters goes up, and this decrease in the precision of the parameter estimation reduces the accuracy of the predictions.

I will review the details of the explanation in the Sections **4** and **6**. In the meantime, in Section **3**, I want to describe the sense in which this explanation provides a solution to the curve-fitting problem, while in Section **5**, I reply to De Vito's charge that Goodman's famous grue problem shows that this solution is 'conventional' because the number of adjustable parameters depends on our language.

3 The curve-fitting problem redefined

Traditionally, in the philosophy of science, the curve-fitting problem is understood as follows: We are interested is representing some variable y as a function of a variable x. For example, y might be the position of a planet and x might be time. Suppose that we have a number of observations of y for

different values of x. These may be plotted as a set of points on an x-y diagram, and for that reason one often refers to the data as data points. Curve-fitting is the procedure by which we fit a curve to the points in the x-y plane, from which we may predict the values of y for new values of x. Many curves can be drawn through any finite set of points, so the points do not determine which curve should be used to represent them. So how do we determine a curve from a finite set of data points?

The traditional answer to the question is to say, in addition to the requirement that the curve passes through the points, that the curve should be simple. The problem with this solution is threefold: (1) The notion of simplicity is left vague. (2) We cannot rely on our intuitions about which curves look simple because that depends on how we represent the curves. For as Priest ([1976]) pointed out, a curve that looks simple in the x–y diagram will look complex if its represented on an x'–y diagram, where x' is some complex transformation of the variable x. Similarly, a complex-looking curve can be made to look simple. (3) Even if we solve these problems, then it is still mysterious why simplicity is valuable. The usual story—that we know *a priori* that the world is simple—seems plainly wrong in light of the fact that we know *a posteriori* that the world is complex, especially at the curve-fitting level. For example, the true trajectories of the planets are chaotic to some degree, and therefore extremely complex.

Forster and Sober ([1994]) do not propose a solution to the traditional curve-fitting problem. They reject the traditional problem because it is ill posed. First, it buys into the naïve empiricist assumption that curves should fit the seen data perfectly. Second, it misdescribes the problem as it arises in science. For instance, in Copernican astronomy, one always fixes the number of circles before attempting to fit the theory to the data. Only a bounded *family* of curves, or trajectories, is ever fitted to the observations. I will refer to such a family as a *model*. Within the context of a fixed model, there is no problem of the underdetermination of a curve from data. In fact, the problem is quite the opposite. There is usually *no* curve that passes through all the data points! Therefore, curve-fitting in science selects the curve that *best* fits the data, where 'best' is defined in terms of a statistical measure of fit, such as in the method of least squares. Under most conditions (known as conditions of *identifiability*), the best-fitting curve in a family is unique. The parameters values associated with this best fitting curve are the parameter *estimates*. That is how statistical estimation works.

There is a problem here, but it takes a different form. The problem is that there are many competing models, each yielding a different best-fitting curve. How do we choose from amongst these? It is a problem of underdetermination, but it occurs at the level of models, not curves. Forster and Sober ([1994]) discuss this problem, and describe a solution based on the work of Akaike ([1973]).

The Akaike solution is similar to the solution of the traditional problem. It says that the competing best-fitting curves should be chosen on the basis of their fit with the data as well as the simplicity *of model* from which it was selected. However, this solution is not subject to any of the three objections listed above. (1) The notion of simplicity used is not vague. Under the previously mentioned condition of identifiability, it is adequately defined as the paucity of adjustable parameters. (2) It is not subject to Priest's problem because the simplicity of a single curve is never defined. (3) The reason why simplicity is valuable is that tends to increase the precision by which parameters may be estimated, which has a positive effect on the accuracy of predictions.

The precision of parameter estimation is not the only factor that affects predictive accuracy. For example, suppose we model planetary orbits as squares centered at the sun. Even if we could estimate the parameters of this model with infinite precision, the model would not make accurate predictions because there is no curve in the family which fits the true orbit very well. It cannot predict the phenomena accurately *in principle*. The discrepancy between the best in-principle curve in a family and true curve is known as the *model bias* or the *approximation discrepancy* of the model. Any model that contains the true curve has zero model bias, whereas modelling planetary orbits as squares introduces a large model bias.

So the accuracy of the predictions made by a model depends on two factors: the error in estimating the parameters and the model bias. Simple Copernican models compare favourably with Kepler's laws with respect to estimation error, because they have about the same number of adjustable parameters. But Kepler provides a better in-principle approximation to the true motion than a simple Copernican model. On balance, Kepler has an advantage over simple Copernican models.

What about complex Copernican models? They are better at approximating the true motion in principle, but this advantage is outweighed by the imprecision of the parameter estimates. So in either case, Kepler does better than Copernicus.

4 Akaike's theorem revisited

Akaike ([1973]) provides a precise quantitative way of correcting the methodology of naïve empiricism. Naïve empiricism, remember, operates under the false assumption that the fit of a model to *seen* data is the best estimate of its predictive accuracy. It fails to take account of the estimation error. If there were no estimation error, then the naïve empiricist assumption would be correct. So, if we could estimate the estimation error, then maybe we could combine this information with the observed fit to provide an improved estimate of predictive accuracy. This is what Akaike succeeded in doing.

First, let me introduce some terminology. Consider any family of curves **F**, and denote the best in-principle curve in **F** by F^*. The discrepancy between F^* and the true curve is the model bias. The curve that best fits the seen data is $\hat{F}$ and the discrepancy between $\hat{F}$ and F^* is the estimation error. The predictive accuracy of $\hat{F}$ is maximized by minimizing the discrepancy between $\hat{F}$ and the true curve. The discrepancy between $\hat{F}$ and the true curve is the sum of the estimation error and the model bias.

This very simple analysis of the problem leads to some deep and insightful consequences. Consider two families of curves, **F** and **G**, such that **F** is *nested* in **G**. That is to say, all the curves in **F** are also in **G**. This subset relation arises frequently in real examples of model selection. For example, if LIN is the family of all straight lines in the x–y plane, and PAR is the family of all parabolas, then LIN is nested in PAR because straight lines are special cases of a parabola (with the coefficient of the x^2 term put to zero). Or consider two Copernican models in which the second is obtained from the first by the addition of one circle. The first is nested in the second because it is a special case of the second with the radius of the added circle equal to zero. In these cases, two important facts about discrepancies.

Proposition 1: If **F** is nested in **G**, then the model bias of **F** is greater than or equal to the model bias of **G**.

Proof: By definition F^* is the curve in **F** with the smallest discrepancy. Since F^* is also in **G**, the curve in **G** with the least discrepancy, G^*, cannot have a greater discrepancy than F^*.

In the Copernicus example, this tells us that we can never increase the bias of a Copernican model when we add an epicycle. Fourier's theorem goes one step further. It implies that we can reduce the bias of a Copernican model to an arbitrary amount by adding a sufficient number of epicycles.

Proposition 2: If **F** is nested in **G**, then the estimation error in **F** is less than or equal to the estimation error in **G**.

Informal Proof: Unfortunately, this property is not as general. But in the Akaike framework, the property follows from the choice of discrepancy measure and the particular assumptions made in the proof of Akaike's theorem. The proof rests on the fact that there is a geometrical representation of the families in which each member is a point in Euclidean space, and the discrepancy between any two curves is the square of the distance between them (see Figure 1). Moreover, the displacement from $\hat{F}$ to F^* is the (orthogonal) projection of the displacement from $\hat{G}$ to G^*. The projection of a displacement can never have a greater length than the original displacement, so the estimation error in **F** can never be greater than the estimation error in **G**.

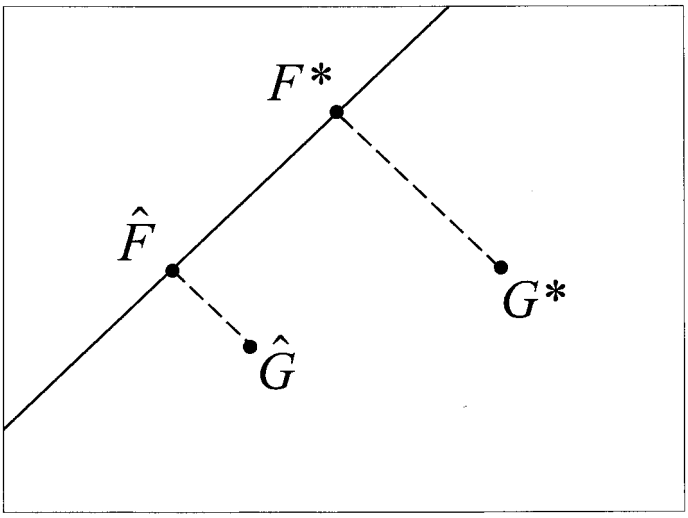

Fig 1. The family **F** is represented by points on the line, while family **G** includes all points in the plane.

All of this proves that there is a kind of dilemma involved in the choice between **F** and **G**. By choosing the more complicated model, **G**, we may reduce the model bias, which is good. But we will also increase the estimation error, which is bad. Whether the move to greater complexity is good or bad depends on whether the reduction in bias outweighs the increased estimation error. This is commonly known as the bias/variance dilemma (e.g. Geman *et al*. [1992]).

Akaike ([1973]) uses the Kullback–Leibler ([1951]) information measure to define the discrepancy, which is why he calls his model selection criterion an information criterion. Note that the concepts of 'discrepancy' and 'fit' are two sides of the same coin. The less the discrepancy between two curves the greater their mutual fit. This definition of discrepancy defines the fit between a hypothesis and a data set as the logarithm of the likelihood, where the likelihood of a hypothesis is a technical term that refers to the probability (density) of the data given the hypothesis. Do not confuse this with the Bayesian notion of the probability of a hypothesis given the data. The log-likelihood of a family of curves **F** is undefined.[3] So when one speaks of the fit of the family **F** to a set of data, one is referring to the *maximum* log-likelihood. A hypothesis is a particular curve, or trajectory in this case, *together with an error distribution*

[3] In logical terms, a family or model **F** is a huge disjunction of hypotheses, each of which does have a well-defined likelihood. The laws of probability show that the likelihood of a model is an average of the likelihood of its members. Therefore, each hypothesis would need to have a probability (density) in order for the likelihood of the model to be well defined. Only Bayesians assume that such probabilities exist.

(for a curve by itself does not assign a probability to data). This will play an important role in Section **6**.

Akaike ([1973]) shows that if fit is defined is that way, then under quite general circumstances[4] we may expect F^* to fit the true curve by an amount $k/2$ *better* than $\hat{F}$, where k is standardly equal to the number adjustable parameters. In other words, the estimated parameter values in $\hat{F}$ are less accurate than those for F^* and this translates into a lower predictive accuracy by an average amount of $k/2$. But how do we use this information to estimate the predictive accuracy? That is the next question.

Assume that the seen data are representative of the data generated by the true curve. However, it does not follow that the degree to which $\hat{F}$ fits the current data is representative of how well $\hat{F}$ will fit other data generated by the true curve. The reason is that F has the special status of being the curve in **F** that *best* fits the current data. That introduces a selection bias, which means that $\hat{F}$ has special features that fit to *accidental* features of this particular data set. That is, we know that $\hat{F}$ *overfits* the current data. Overfitting is dramatically illustrated by the fact that F fits the current data better than F^* even though it will usually do worse than F^* on other data sets. In fact F^* will do better, on average, than *any* other curve in **F** because it is by definition the curve that is closest to the true curve.

If we knew which curve F^* was, we would use it, but we do not. In fact, we have no direct information about where F^* is relative to $\hat{F}$. All we 'see' is $\hat{F}$. So, what can we do? This is where Akaike's theorem introduces a clever trick. If we have a number of competing models of different complexities, each with a different $\hat{F}$, then maybe we can correct for overfitting, and choose the best $\hat{F}$ from the competing models. That is, we may compare the competing curves by their estimated fit with unseen data rather than their fit with seen data. Since predictive accuracy is defined as the expected fit with unseen data, this idea amounts to comparing competing models by their estimated predictive accuracy.

Despite that fact that we cannot 'see' the different F^*, they play an important role in the second part of Akaike's theorem which says that, on average, F^* fits the current data by an amount $k/2$ worse than $\hat{F}$. Therefore, we may estimate the fit of F^* with the current data as equal to the fit of $\hat{F}$ minus $k/2$. This is useful to us

[4] The circumstance goes by the name of the *normality condition*. The normality condition speaks of the way in which repeated parameter estimates would be probabilistically distributed, and is exactly the condition guaranteed by the various versions of the Central Limit Theorem. It does not require that observational errors are normal (Gaussian), although there is a regularity requirement here (see Cramér [1946] for details). It is also worth noting that the theorem does not require that the equation for y is linear in the parameters, although it is true that the assumptions are harder to satisfy when the equation is nonlinear. Note that if the number of data is large, then the multinormal distribution may be concentrated in a narrow region of parameter space in which the function for y is closely approximated by a function that is linear in its parameters.

because F^* is not subject to a selection bias. It is defined independently of the particular data set at hand, and so its fit with the seen data is representative of its fit with any data set generated by the true curve (assuming that it is sampled from the same region of the curve). So we now have a way of estimating the predictive accuracy of F^*, or, equivalently, the discrepancy between F^* and the true curve (modulo a constant that is the same for every model, which can be ignored in making comparisons). Remember that this discrepancy is also called the model bias, or the approximation discrepancy.

Remember that the discrepancy between $\hat{F}$ and the true curve is the sum of the estimation discrepancy and the approximation discrepancy. Therefore the estimated discrepancy between $\hat{F}$ and the true curve is given by its discrepancy with the current data plus k; or equivalently, the estimated predictive accuracy of $\hat{F}$ is equal to its fit with the current data minus k. This corrects the naïve empiricist estimate of predictive accuracy by taking account of the complexity of the model, as measured by k.

In summary: if one is interested in predictive accuracy, one should compare the best fitting curves from *competing* families by their fit with seen data *after it is corrected for overfitting*. This leads to a precise rule for trading off simplicity and goodness-of-fit, known as Akaike's Information Criterion, or AIC (pronounced A-I-C).

It is a quirk of the way statistical fit is defined that it increases as the number of data, n, increases. It has been tacitly assumed that the novel data set has the same number of data as the current data set, so this has not mattered. But it is best to remove this assumption by defining predictive accuracy as the expected *per datum* fit with novel data, which is then estimated by the per datum fit with current data minus k/n. Since n is constant for all models, this modification makes no practical difference in model selection. Nevertheless, it shows that the estimation error, like all *sampling* errors, tends to zero as the number of data increases.[5]

AIC makes one important correction to a very naïve methodology of science—naïve empiricism—by using some background information about the tendency of complex models to overfit the data. In cases where there is other relevant information available, we should expect that AIC will have to be amended.

5 The grue problem as a curve-fitting problem

There is one claim in De Vito's paper that is correct. He's right that this explanation of the relevance of simplicity to curve-fitting does not solve

[5] See Forster ([to appear]) for further discussion of this point as a response to the common allegation that AIC is statistically inconsistent.

Goodman's new riddle of induction (also known as the grue problem). He's also right that the grue problem is a curve-fitting problem. So, it follows that the Akaike solution does not solve all problems commonly included under the title of 'curve-fitting problem'.[6] But he's wrong about the *way* in which the grue problem is represented as a curve-fitting problem, and this mistake leads to a mistaken charge of language variance. I will begin by explaining the sense in which the grue problem is a curve-fitting problem, and show why the example does not justify De Vito's charge of language variance.

First, define 'grue':

DEFINITION: Object x is grue at time t if and only if x is green at time t and $t < 2100$, or x is blue at time t and $t \geq 2100$.

The reason that I relativize the colour predicates to a specific time is that, otherwise, they are not observational predicates. I can observe that a tomato is green now, but I am not thereby observing that it will be green next week. I may be able to infer that, but I cannot observe it. Curve-fitting has to begin with what is observed, so the point is important. In the grue example, the data are that all the emeralds observed to date have been green, which also implies, by the definition, that they have all been observed to be grue (at the time of observation). There are now two competing hypotheses capable of explaining the data: the Green hypotheses says that 'All emeralds are green (at all times)' and the Grue hypothesis says that 'All emeralds are grue (at all times)'. Note that, while they fit all the observations of emeralds to date equally well, they make different predictions. The Grue hypothesis predicts that all emeralds observed after 2100 will be grue, and therefore blue at that time, while the Green hypothesis predicts that they will be green.

The grue *problem* is that the Green hypothesis and Grue hypothesis fit all the observations of emeralds equally well, so fit is not what tells us that we should use the Green hypothesis in favor of the Grue hypothesis. Yet we seem to have no hesitation in preferring the Green hypothesis. It is yet another demonstration that naïve empiricism is wrong. The problem is to say what else marks the difference between the hypotheses. Many people are inclined to say that the Green hypothesis is better because it is simpler. But what is 'simplicity' and why does it matter? There is no doubt that the problem is similar to Akaike's problem in many ways, but that does not imply that it is the same problem.

It is not obvious that the grue problem can be understood as a curve-fitting problem at all. A curve represents a function of some dependent variable y on an independent variable x, where a function is a many-to-one mapping from x-values in a set (called the domain) into y-values (called the range). The Grue or

[6] Forster and Sober ([1994]) do not claim to have solved the grue problem. It is not mentioned there for exactly this reason. See also Sober ([1994]).

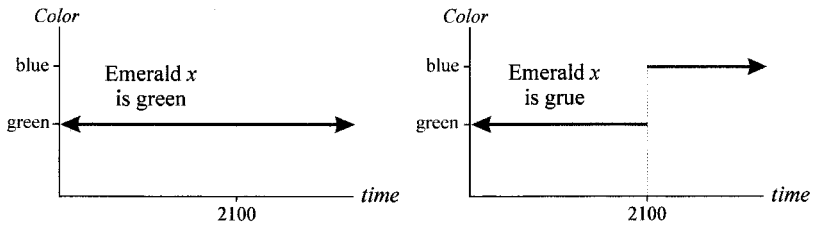

Fig 2. The grue problem as a curve-fitting problem.

the Green hypothesis can be thought of as mapping time values into colour values. Thus, *t* is the independent variable, and 'colour' is the dependent variable. The Grue and the Green hypothesis are each represented as a single curve (Figure 2). Each hypothesis is represented by a *single* 'curve' because it gives a unique prediction for any specification of the independent variable (*time*).

Does the concept of 'adjustable parameter' apply to this example? Here it is important to understand the difference between 'adjustable' and 'adjusted'. The time 2100 AD is a parameter in the Grue hypothesis, but it is an adjusted parameter. In fact, all the parameters associated with an individual curve are adjusted because they have particular numerical values. It is only when parameters are free to take on a range of possible values that they are adjustable, and that occurs only within the context of a *family* of curves. We could place the Green hypothesis and the Grue hypothesis in singleton families in which they are the only member. Then these *families* would have zero adjustable parameters, so the answer is the same. *There are no adjustable parameters in this example*.

Therefore, the Green and the Grue hypotheses do not differ in simplicity in the sense relevant to standard model-selection criteria, including Akaike's solution to the curve-fitting problem. Therefore Akaike's criterion must go by fit alone in this example.

Does the relevant concept of fit apply to this example? The concept of fit in Akaike's theorem is derived from the Kullback–Leibler discrepancy, which requires that the competing hypotheses are probabilistic (so that likelihoods are well defined). The Green and Grue hypotheses are not probabilistic, but that may be fixed by associating them with the same assumption about the probability of observational errors. The bottom line is that the Green and Grue hypotheses both fit the current observations equally well. None of the usual model-selection criteria provide any reason to favour the Green hypothesis over the Grue hypothesis, either by differences of simplicity, or by differences in fit, and Akaike's criterion is no exception.

This makes perfect sense in light of what the Akaike criterion aims to do. Its purpose is to use simplicity to correct for the difference between the in-practice

and in-principle application of a model. When a model has only one member, there is no difference between its in-practice and in-principle application. There is no estimation of parameter values, and therefore no need to use simplicity to correct for estimation error. The only factor involved is model bias, and all standard model-selection criteria assume that fit is the best way of estimating this.

I do not dispute the idea that Akaike's criterion gives the wrong answer. But what follows from that? Some may say that the number of adjustable parameters is not the correct measure of simplicity. But it was never presented as *the* measure of simplicity. It is *one* measure of simplicity, and it is a kind of simplicity that is relevant in the curve-fitting problem because it quantifies estimation error. If other kinds of simplicity are relevant to curve-fitting, then this does not undermine the fact Akaike simplicity is relevant to curve-fitting. Akaike's criterion says that there is no correction for estimation error required, and here it is correct even if it does not provide a complete solution to the problem.

Unfortunately, De Vito ([1997]) reaches a different conclusion. He thinks that there is a serious problem for Akaike in this example that goes well beyond its incompleteness. He thinks that the parameter that appears in the definition of 'grue' is an *adjustable* parameter. This time parameter, which I will denote by θ, should not be confused with t, which is an independent variable. θ is a parameter that gets assigned the value 2100 AD in the definition of grue. True, there is an everyday sense in which it is a parameter, and it is adjustable in some sense, but it is not adjustable in the relevant sense. But if one grants the mistaken premise that it is an adjustable parameter, then one can see how everything goes wrong. Define 'bleen' as: Object x is bleen at time t if and only if x is blue at time t and $t < 2100$, or x is green at time t and $t \geq 2100$. Now we have a complete 'grolour' language in which the terms 'grue' and 'bleen' can replace the terms 'green' and 'blue' entirely. For example, object x is green at time t if and only if x is grue at time t and $t < 2100$, or x is bleen at time t and $t \geq 2100$. Therefore, we may redescribe the Green hypothesis as: All emeralds x are such that, if $t < 2100$ then x is grue at time t and if $t \geq 2100$ then x is bleen at time t. The Grue hypothesis makes no mention of the year 2100 in the grolour language, whereas the Green hypothesis does. If we replotted the 'curves' in Figure 2 using grolour values on the y axis instead of colour values, then the Grue 'curve' would be the simple one, and the Green 'curve' would look complicated. So in the grolour language, the Grue hypothesis is simpler, while in the standard colour language, the Green hypothesis is simpler. Therefore, simplicity is language dependent.

The trouble with this argument is that it uses the wrong notion of simplicity. Akaike's criterion depends on the number of *adjustable* parameters, and not on the number of *adjusted* parameters. None of the parameters in this example is adjustable in the relevant sense because adjusting the value of a parameter should move us from one curve in the family to another. That is impossible

when there is only one curve in the family. Therefore, the Akaike simplicity is not language dependent because no re-description of the hypotheses can turn one curve into many, or many curves into one.

In Section **3**, De Vito presents a different example, which he subjects to the same confusion. In this example, De Vito compares the family of straight-line curves (LIN) with the family of parabolas (PAR) in the x–y plane. Then he considers a transformation of the independent variable x to a new independent variable x'. Under this transformation, LIN transforms to LIN$'$ and PAR transforms to PAR$'$. First, the transformation rule contains adjustable parameters, so there are many possible redescriptions involved. The main problem is that transformations do not map a single member of PAR into a unique member of PAR$'$, so there is no sense in which the transformed families are *equivalent* representations of the old families. De Vito thereby conjures up the appearance that PAR$'$ is simpler than LIN$'$, thereby reversing the simplicity ordering. Again, it is all smoke and mirrors.

In this example, the simplicity ordering is determined by the fact that LIN is *nested* in PAR. For any families **F** and **G**, if **F** is a subfamily of **G**, then **F** is not more complex than **G**. Clearly, any one-to-one transformation of LIN and PAR must preserve the subset relation. Therefore, LIN$'$ is a subfamily of PAR$'$, which implies that LIN$'$ is not more complex than PAR$'$ in the relevant sense. Therefore, the Akaike simplicity cannot be reversed under a one-to-one transformations of the families.

My argument assumes that the subfamily relation between families constrains their Akaike simplicity ordering. Unfortunately, the substantiation of this premise requires a more detailed understanding of Akaike's theorem. Suffice it to say that it follows from Proposition 2: if **F** is nested in **G**, then the estimation error in **F** is less than or equal to the estimation error in **G**. Akaike uses simplicity to estimate the estimation error, so it should respect the same relationship. The next section of this paper provides more of the details needed to prove this rigorously.

There is another way of setting up the Grue hypothesis so that it does use an adjustable parameter in the Akaike sense. Instead of supposing that the date at which emeralds change color is the year 2100, allow that time to vary. Again, denote that time by θ so that it is not confused with the independent variable t. θ is adjustable because there is a different curve corresponding to each value of θ. The Grue hypothesis is now a non-singleton family of curves, while the Green hypothesis is still a singleton family. This is a different version of the grue problem from the one De Vito describes, but it is worth considering in its own right.

It may appear that Akaike's criterion will now choose the Green hypothesis in favor of the Grue hypothesis; after all, there is a sense in which the Green hypothesis is nested in the Grue family as the special case in t goes to infinity.

The criterion does favour 'All emeralds are green' if it is taken at face value. However, this application of the criterion is not appropriate. Remember that Akaike's notion of simplicity aims to quantify the sampling errors in the parameter estimates. But in this example, there is still no *sampling* error in the estimation of the grue parameter θ. It cannot be estimated *at all* in any principled way because there is a large range of values of θ that makes no difference to the degree of fit. In the language of statistics, the grue model is unidentifiable in the sense that there is no unique value of θ that maximizes the fit with the seen data. There is no overfitting or underfitting in the relevant sense. So Akaike's criterion does not solve the second version of the grue problem either.

There is another point here. You cannot estimate θ because you cannot sample from the future. Akaike's theorem has to do with predicting *new* data from information on *old* data, where both come from the same distribution. Past observations may tell us where θ is not, but predictions require us to say where θ is. The grue problem (set up as problem in which one estimates θ) is like sampling from one urn and using the information to make predictions about a different urn. Without assumptions about how the two urns are related, this is an impossible task.[7]

There are many curve-fitting problems that Akaike's *theorem* does not solve. The subfamily problem in Forster and Sober ([1994]) was one example, and the reply of Forster ([1995]) to Kukla's problem Kukla ([1995]) is another. Kruse's ([1997]) explanation of the value of evidential variety does not rest on Akaike's theorem, even though it is firmly situated within the Akaike framework. None of the limitations in the scope of Akaike's theorem undermines the positive insights it provides in a large class of cases to which it does apply. It is a precise theorem built upon precise assumptions, whose applicability or inapplicability can be examined in particular circumstances.[8] Once the theorem is properly understood, one can see why it is language invariant, and why it does not solve the grue problem.

6 Language invariance restored

The previous section has established that De Vito's arguments are flawed. However, this does not establish that his conclusion is false. So I need to delve deeper into how the quantity k emerges from Akaike's theorem, for k is responsible for introducing the notion of simplicity into the equation. That is the purpose of this section.

[7] I owe this point to Elliott Sober (personal communication).

[8] Kieseppä ([1997]) denies that it is a rigorous theorem, possibly because it approximates functions using truncated Taylor expansions. Almost all of *applied* mathematics and statistics fails this standard of rigour. The rule is: the greater the rigour, the less the applicability. Applied mathematicians make the trade-off differently from pure mathematicians.

The issue at stake is far greater than the health and well-being of Akaike's criterion. For if the charge were true, it would undermine all recent approaches to model selection, including Schwarz's ([1978]) Bayesian Information Criterion (BIC), the Minimum Description Length (MDL) approach, the Minimum Message Length (MML), and a number of others. They all measure simplicity in exactly the same way.

If k were equal to the number of adjustable parameters, then on the surface, there is a problem because the number of adjustable parameters can change when a family of curves is differently *described*. An example in Forster and Sober ([1994], fn. 13) shows this: suppose that we have a family of curves represented by the equation $y = \theta x$, where θ is an adjustable parameter. The same family of curves is equivalently represented by the equation $y = (\alpha + \beta)\, x$, where α and β are two adjustable parameters. This simple re-description of the family changes the number of adjustable parameters from 1 to 2.

However, this does not prove that Akaike's *criterion* is language variant. What it shows is that it is wrong to describe the quantity k as the number of adjustable parameters. It is only the common *description* of the criterion that is wrong, and not the criterion itself. So, what is the true meaning of the quantity k? Its meaning is determined by the role it plays in Akaike's *theorem*. It is important to understand the theorem in greater detail.

In the family of curves $\{y = \theta x | -\infty < \theta < \infty\}$, the parameter θ takes on an infinite range of possible values. For each numerical value of θ there corresponds a curve (a straight line with slope θ passing through the origin in this example) in the family. Conversely, for each member of the family, there corresponds a unique value of θ. So, if we represent the family in the parameter space consisting of all values of the single parameter θ, then there is a one-to-one correspondence between curves in the family and points in the parameter space.

Now consider the parameter space in which points determine an array value assigned to the parameters $\{\theta, \phi, \eta\}$, for example. Our family can still be represented in this parameter space, except that there is no longer a one-to-one correspondence between the family and the points in parameter space. A point in the parameter space determines a unique member of the family, but a member of the family no longer determines a unique point in the parameter space. The mapping from the parameter space on to the family of curves is many-to-one. Or consider the parameter space of possible value for $\{\alpha, \beta, \phi, \eta\}$, where θ is equal to $\alpha + \beta$. Again, there is a many-to-one mapping from points in the parameter space onto the family of curves.

Akaike's theorem begins by defining a discrepancy function between an arbitrary curve $y = \theta x$ and the true curve, which we may denote by $\Delta(\theta)$.[9] The

[9] I am following the notation and terminology of Zucchini ([1999]) and Linhart and Zucchini ([1986]).

function will induce a discrepancy function between points in the parameter spaces $\{\theta, \phi, \eta\}$ and $\{\alpha, \beta, \phi, \eta\}$ as well, except that the induced function will have a built-in redundancy. For example, if $\alpha + \beta = \alpha' + \beta'$, then

$$\Delta((\alpha, \beta, \phi, \eta)) = \Delta((\alpha', \beta', \phi', \eta'))$$

The discrepancy used in Akaike's theorem is the Kullback-Leibler distance (Kullback and Leibler [1951]), but I will keep the account at the most general level possible.

The purpose of Akaike's theorem is to estimate $\Delta(\hat{\theta})$, where $\hat{\theta}$ denotes the curve that best fits the observed data. $\hat{\theta}$ also denotes the estimated value of the parameter, and it is important to understand that a discrepancy function assigns *values* to the *curves* and not to parameter values. Remember that a function is a many-to-one mapping from one set, called the domain, into another set, called the range. Each member of the domain must be mapped to a *unique* member of the range. A discrepancy function maps a set of *curves* into a set of numbers. So if the parameter is transformed so that a particular curve is associated with a different parameter value, the curve will still be assigned the same discrepancy.

What if we transform the x or the y coordinates? The answer is the same. A curve will be assigned the same discrepancy before the transformation as after the transformation. The situation may seem a little more complicated than this. After all, isn't it easy to transform the x and y coordinates to change which curve is closer to the truth than another by the sum of squared deviations (Miller ([1975]) makes this point)? Well, yes, but that just shows that the sum of squared deviations does not define a discrepancy function.

So how is a discrepancy function actually defined? In the Akaike framework, one must first associate an error distribution with each curve. This will then determine a conditional probability density of the form $p_\theta(y/x)$. This conditional density will be different for each curve, as I have indicated by the subscript θ. θ assigns numerical values to all adjustable parameters, including those characterizing the error distribution if there are any. Then we need to add a probability distribution for the independent variable, $p(x)$, which defines the region of the curves from which the data is sampled. This depends on the *kind* of sampled data that we aim to predict. There are many predictive accuracies one could consider, depending on the choice of $p(x)$. But none of them is language dependent in any way. Once $p(x)$ is decided upon, it usually disappears from the equations because it is the same for all curves and all models. From these ingredients, we obtain a joint probability density $p_\theta(x, y)$ for each hypothesis. Let $p^*(y/x)$ be the true conditional probability density, from which we may obtain $p^*(x, y)$ using the same $p(x)$ as before. The discrepancy of hypothesis C is now defined as the Kullback–Leibler distance $p_\theta(x, y)$ between and $p^*(x, y)$. The Kullback–Leibler distance is the difference

between two information or negentropy measures; namely:

$$\Delta(\theta) \equiv \int p^*(x,y) \log p_\theta(x,y) - \int p^*(x,y) \log p^*(x,y).$$

Each term is transformationally invariant modulo a constant, but the constant drops out when one substracts one from the other (Shannon and Weaver [1946]). Therefore, once we represent the hypotheses as probability densities, we may easily define a language invariant discrepancy function (Good ([1975]) made a similar point in reply to Miller ([1975])). There are many other choices besides the Kullback–Leibler distance (see Linhart and Zucchini [1986] or Zucchini [1999]).

Akaike's theorem estimates the discrepancy $\Delta(\hat{\theta})$ in terms of $\Delta(\theta^*)$ by expanding $\Delta(\hat{\theta})$ as a Taylor series around θ^*:

$$\Delta(\hat{\theta}) = \Delta(\theta^*) + \frac{\partial \Delta(\theta^*)}{\partial \theta}(\hat{\theta} - \theta^*) + \frac{1}{2}\frac{\partial^2 \Delta(\theta^*)}{\partial^2 \theta}(\hat{\theta} - \theta^*)^2 + \cdots,$$

where $\partial\Delta(\theta^*)/\partial\theta$ is a shorthand notation for $\partial\Delta(\theta)/\partial\theta$ evaluated at θ^*. The discrepancy function is language invariant, because it assigns discrepancy values to specific *curves*, and not parameter values. Thus, $\Delta(\hat{\theta})$ and $\Delta(\theta^*)$ have the same values even if these curves are picked out by a different value of a rescaled parameter. Moreover, each term in the Taylor expansion is invariant. Consider the second term, for example. It is true that if θ is transformed to θ', then the value of $(\hat{\theta} - \theta^{*\prime})$ will change. But the value of $\partial\Delta(\theta^{*\prime})\partial\theta'$ will change in a way to compensate exactly.

What if we expand the Taylor series in an extended parameter space? In that case, we will have a multivariate Taylor expansion, but it will reduce to something equivalent to the expression above because most of the terms will be zero. For example, the term involving $\partial\Delta(\theta^*)/\partial\eta$ would drop out because the discrepancy function does not vary as a function of η.

What if we define new parameter defined by $\lambda \equiv \theta + \eta$? Then $\partial\Delta(\theta^*)/\partial\lambda$ is not zero. But then by the chain rule for derivatives, $\partial\Delta(\theta^*)/\partial\lambda \equiv \partial\Delta(\theta^*)/\partial\theta$. Adding the term for λ would be to count the dependence on θ twice. So, again, the end result is the same as before, provided that the calculation is done correctly. Likewise, if α and β were used in place of θ, then there would be more terms in the Taylor expansion, but the end result would be the same once we take account of the fact that $\theta \equiv \alpha + \beta$. The point is that language invariance is built into the mathematical theory of functions from the ground up.

By taking the expected value of both sides, and dropping higher than quadratic terms, we arrive at:

$$E[\Delta(\hat{\theta}] \approx \Delta(\theta^*) + 1/2\Omega E[\hat{\theta} - \theta^*)^2].$$

where

$$\Omega \equiv \frac{\partial^2 \Delta(\theta^*)}{\partial^2 \theta}.$$

This introduces expected values, but these are also language invariant features of a probability distribution.[10] Akaike then argues that $E(\hat{\theta} - \theta^*)^2$ is equal to $1/(n\Omega)$ under the conditions assumed by the theorem. This step depends the Keibler–Leibler discrepancy measure, and appeals to the Central Limit Theorem. So he ends up with something equivalent to:[11]

$$E[\Delta(\hat{\theta})] \approx \Delta(\theta^*) + \frac{1}{2n}.$$

If we were to change the example so that the discrepancy has a real dependence on k parameters, instead of just one, then the k parameters could be transformed so that Taylor expansion of the discrepancy would involve k quadratic terms, each of which look like the one above. By the same reasoning, each term would contribute and expectation of $1/(2n)$, adding to a total of $k/(2n)$. This is the first part of Akaike's theorem.

The point is that the value of k is not simply the number of adjustable parameters, but the number of parameters *that contribute to the expected discrepancy in a certain way*. Given the fact that the Taylor expansion is language invariant, and expected values are language invariant, there is no way in which this number can change by any redescription of the families of curves. It is convenient to describe k as equal to the number of adjustable parameters only because this equality holds *in most cases*.

One could, in principle, encode the information contained in k variables in a single variable.[12] However, if the discrepancy function were not a sufficiently smooth function of that variable for the derivatives in the Taylor expansion to be well defined, then the derivation would not go through in its present form. If it can be made to work then the answer will be the same as before. There are no tricks that are going to change the bottom line.

In summary, the property of language invariance is an important desideratum for any criterion of model selection. It is not acceptable for someone to say that, in English, Newton's theory of motion is true, but in French it is false (or the other way around). It is equally unacceptable to say that a hypothesis is closer to the truth in one language than it is in another. Fortunately, language invariance is built in at the very beginning. The only problem is in explaining it to a non-mathematical reader.

[10] Not all features of a probability distribution are invariant. Forster ([1995]) criticizes some Bayesians for failing to notice that probability densities are not language invariant. But expected values of an invariant function are always invariant.

[11] For those following the proof, it is important to note that the discrepancy assumed here is defined in a way that does not depend on n. The expected value of $\Delta(\hat{\theta})$ depends on n because the curve $\hat{\theta}$ moves closer to the true curve as the number of data increases. $\Delta(\theta^*)$ does not depend on n in any way.

[12] I owe this point to an anonymous referee.

Acknowledgements

I am grateful to Scott De Vito and the participants of the Central Division meeting of the APA discussion on this topic at Pittsburgh in 1997 for helping me to improve a previous draft of this reply. Special thanks go to Elliott Sober and to an anonymous referee of this journal for helpful advice.

Department of Philosophy
5185 Helen C. White Hall
University of Wisconsin
Madison, WI 53706
USA

References

Akaike, H. [1973]: 'Information Theory and an Extension of the Maximum Likelihood Principle', in B. N. Petrov and F. Csaki (*eds*), *2nd International Symposium on Information Theory*; Budapest: Akademiai Kiado, pp. 267–81.

Cramér H. [1946]: *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

De Vito, Scott [1997]: 'A Gruesome Problem for the Curve Fitting Solution', *British Journal for the Philosophy of Science*, **48**, pp. 391–6.

Forster, Malcolm R. [1995]: 'Bayes and Bust: The Problem of Simplicity for a Probabilist's Approach to Confirmation', *British Journal for the Philosophy of Science*, **46**, pp. 399–424.

Forster, Malcolm R. [1995]: 'The Golfer's Dilemma: A Reply to Kukla on Curve-Fitting', *British Journal for the Philosophy of Science*, **46**, pp. 348–60.

Forster, Malcolm R. [submitted]: 'The New Science of Simplicity', in H. Keuzenkamp, M. McAleer, and A. Zellner (eds) *Simplicity, Inference, and Econometric Modelling*, Cambridge: Cambridge University Press.

Forster, Malcolm R., and Sober Elliott [1994]: 'How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions', *British Journal for the Philosophy of Science***, 45**, pp. 1–35.

Geman, Stuart, Bienenstock, E., and Doursat René [1992]: 'Neural Networks and the Bias/Variance Dilemma', *Neural Computation*, **4**, pp. 1–58.

Good, I. J. [1975]: 'Comments on David Miller', *Synthése*, **30**, pp. 205–6.

Kieseppä, I. A. [1997]: 'Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of Simplicity', *British Journal for the Philosophy of Science*, **48**, pp. 21–48.

Kukla, Andre [1995]: 'Forster and Sober on the Curve-Fitting Problem', *British Journal for the Philosophy of Science*, **46**, pp. 248–52.

Kruse, Michael [1997]: 'Variation and the Accuracy of Predictions', *British Journal for the Philosophy of Science*, **48**, pp. 181–93.

Kullback, S. and Leibler, R. A. [1951]: 'On Information and Sufficiency', *Annals of Mathematical Statistics*, **22**, 79–86.

Linhart, H. and Zucchini, W. [1986]: *Model Selection*, New York: John Wiley & Sons.

Miller, David [1975]: 'The Accuracy of Predictions', *Synthése*, **30**, pp. 159–91.

Priest, Graham [1976]: 'Gruesome Simplicity', *Philosophy of Science*, **43**, 432–7.

Schwarz, Gideon [1978]: 'Estimating the Dimension of a Model', *Annals of Statistics*, **6**, pp. 461–5.

Shannon, Claude and Weaver, W. [1949]: *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.

Sober, Elliott [1994]: 'No Model, No Inference: A Bayesian Primer on the Grue Problem', in Stalker [1994], pp. 225–40.

Stalker, Douglas [1994]: *GRUE!*, Chicago, IL: Open Court.

Zucchini, Walter [1999]: 'An Introduction to Model Selection', To appear in the *Journal of Mathematical Psychology*.