Gruesome Simplicity: A Guide to Truth

Aidan Lyon*

June 5, 2007

*** DRAFT ***

Abstract

Priest [1976] demonstrated that a more general version of Goodman's grue-paradox exists for the standard solution to the 'traditional' curve-fitting problem (which says that we ought to choose the simplest curve that passes through all observed data points). Forster and Sober [1994] argue that this problem is ill-posed and that their solution (the Akaike Information Criterion (AIC)) to the 'real' curve-fitting problem—the problem of balancing accuracy and simplicity—avoids the kind of paradox described by Priest [1976].

DeVito [1997] argued that there is, in fact, a version of the grue paradox for AIC. However, Forster [1999] and Kieseppä [2001] show that DeVito's arguments do not warrant his conclusion. Despite these results, I will argue that there is indeed a version of the grue paradox for AIC which is quite similar to the problem that Priest originally identified. However, this problem is not unique to AIC. Indeed, I will argue that any solution to the curve-fitting problem will be susceptible to a version of the grue paradox.

1. Introduction

In *Fact, Fiction and Forecast* Goodman introduced his now famous New Riddle of Induction.[1] The riddle raised a serious problem for any hope of a purely syntactic theory of inductive logic, such as the one Carnap was trying to develop

*Philosophy Program, RSSS, Australian National University, Canberra, ACT, 0200, Australia. Email: aidan@coombs.anu.edu.au

[1] One version of Goodman's New Riddle of Induction: Suppose that every emerald so far observed is green. From these observations, it seems we could infer that every emerald is green. But equally compatible with our observations is the hypothesis that all the observed emeralds before 2070 are green and all others are blue. The riddle is commonly taken to be that we want to say that the latter hypothesis is absurd whilst the former is respectable; but why?

(see e.g., Carnap [1945]). Many, including Carnap, tried to resolve this riddle by invoking the use of the notion of simplicity of hypotheses. Goodman responded with what is sometimes called the Grue Paradox: the simplicity of a hypothesis depends on the choice of language it is formulated in [ref]. Many have tried to deflate the paradox by giving one language a privileged status. For example, Lewis [1984] invokes the notion of a *natural* predicate to privilege a language. Goodman used his theory of entrenched predicates to attain a privileged language, in which to compare hypotheses. Such moves make a theory of induction language dependent. While this is perhaps regrettable, it is typically seen as a necessary feature of any theory of induction.

In the contemporary literature, a particular problem of inductive inference—the curve-fitting problem—has been the focus of much attention, and it is claimed that a solution has been provided which does not suffer from the kind of language dependence problems that purely syntactic theories of induction suffer from. In this paper I will argue that this solution to the curve-fitting problem (and any other solution, except subjective Bayesian ones) is language dependent, but that this shouldn't be seen as a reason to reject such solutions. Some of the problems of language dependence for curve-fitting are, in a way, much worse than the same type of problems for purely syntactic theories of inductive logic, but in §3 I outline a way in which they may be overcome through the use of symmetry considerations. First though, it will be necessary to go over a few preliminaries.

1.1. The Traditional Curve-Fitting Problem

Curve-fitting is a very common form of inductive inference. Traditionally, curve-fitting has been the process of inferring from a finite set of particular observations of two quantities—which are represented as points in a coordinate system—to a generalised hypothesis about the relationship between the two quantities—and this hypothesis is represented as a *curve* that passes through each point (e.g., see Popper [1959], p. 124; Glymour [1981], p. 322). What makes this process of inference interesting is that no matter how many observations we have made, (assuming there is only ever a finite number of observations) there are always infinitely many curves which pass through each data point in the coordinate system. We are thus confronted with the problem of deciding which curve, out of the infinite number available, is the curve that represents the true relationship between the two quantities.

1.2. Gruesome Simplicity

The standard solution to this problem—which I will call the *traditional* curve-fitting problem—has been to choose the *simplest* curve which passes through each data point (see e.g., Priest [1976], p. 432; Popper [1959], p. 124). For example, the reason why we ought to choose a straight line which passes through each data point over a 'bumpy' curve, which also passes through each data point, is that the former is simpler than the latter. The idea is that the simplicity/complexity of a curve represents the simplicity/complexity of the hypothesis that the curve itself represents.

However, Priest [1976] demonstrated that the simplicity of the curve that represents a hypothesis depends on the language we choose to formulate the hypothesis in. He did this by using the following example:

> "We observe a moving particle and note its velocity v, and momentum p. It is found that when $v = 2$, $p = 6$ and when $v = 3$, $p = 8$. We now ask what the best prediction is for its momentum when $v = 4$. Obviously the curve that best fits the data is the straight line:
>
> $$p = 2v + 2$$
>
> and hence we predict that when $v = 4$, $p = 10$.
>
> But now suppose we decide to correlate the velocity with the (classical) kinetic energy of the particle $E(= pv/2)$, computed from the same data. We have that when $v = 2$, $E = 6$ and when $v = 3$, $E = 12$. Again the curve that best fits this data is a straight line:
>
> $$E = 6v - 6$$
>
> Hence we predict that when $v = 4$, $E = 18$. But since $E = pv/2$, $p = 2E/v$, so the corresponding value for p is 9. This is clearly incompatible with our previous 'best' prediction." Priest [1976], pp. 432–3

When we translate the ('best') hypothesis about the relationship between momentum and velocity, $p = 2v + 2$, into a hypothesis about the relationship between kinetic energy and velocity, we get a curve more complicated than a straight line, $E = 12(1 - v^2)$. And similarly in the other direction.

As Priest points out (Priest [1976], p. 435), the above example highlights a problem which is a more general version of Goodman's grue paradox. To see this, think of all the green emerald observations we have seen so far as points in the *(time, light frequency)* plane, the green hypothesis as the straight line which passes through all of these data points, and the grue hypothesis as a curve which passes through all of the points but drops to the light frequency of the colour blue after some future time t.

In this representation, the green hypothesis appears to be simpler than the grue hypothesis—it's a straight line whilst the grue hypothesis is a step function.

But, with a suitable change of representation, the grue hypothesis appears to be simpler than the green hypothesis—it is now the grue hypothesis which is the straight line, and the green hypothesis is now a step function.[2]

Actually, in a way, this is worse than Goodman's original Grue Paradox. When comparing 'grue' and 'green', we could at least say that 'green' is more natural or more entrenched etc., than 'grue'. But how could we say that the velocity/momentum system of representation is more natural or entrenched than the velocity/kinetic energy system of representation, or *vice versa*? Both systems seem perfectly natural and equally entrenched. I will return to this issue in §3.

1.3. Bertrand's Paradox

Interestingly, while Priest's example does seem to be very similar to the Grue Paradox, it is also reminiscent of another famous paradox: Bertrand's Paradox. Bertrand's Paradox is a paradox for the Principle of Indifference (which says that in the absence of evidence to the contrary, we should assume each possible outcome equally likely). The paradox is needlessly complicated for the point it makes. The following example adapted from van Fraassen ([1989], p. 303) is simpler, but makes the same point.[3] Suppose there is a box factory which produces cubes of side-length between 0 and 1 metre. We do not know any further details about the box factory. What probability should we assign to the event of the box next produced having a side-length between 1/2 and 1 metre? The Principle of Indifference says that we ought to assign this event a probability of 1/2. This is because we have no reason to suppose it any more probable that the box has a side-length between 0 and 1/2 metre than for it to have a side-length between 1/2 and 1 metre. Since these two options exhaust the space of possibilities, their probabilities must sum to 1. And since their probabilities are equal, they must both be 1/2. An important part of this reasoning—which is often played down—is that the two events have the same number of ways of occurring, even though this requires comparing two sets that both have continuum many points. This is sometimes justified by the fact that the sets $[0, 1/2]$ and $[1/2, 1]$ have the same Lebesgue measure, or length according to the Euclidean metric. But suppose we represent the scenario in a slightly different way (without changing any facts), so that instead of speak of side-lengths, we speak about side-areas. The box factory produces cubes of *side-area* between 0 and 1 square metre. What is the probability that the box next produced as a side-area between 0 and 1/4 square metres? The Principle of Indifference says that we

[2] See Priest [1976], p. 435 for one particular way of formally setting up a similar scenario.

[3] See Gillies [2000], pp. 37–49 for a detailed discussion of Betrand's Paradox

ought to assign this event a probability of 1/4, which obviously contradicts our previous assignment. The reason why is similar to before. There are four possibilities which we have no reason to think are not equi-probable—the side-area is between (i) 0 and 1/4 square metres, (ii) 1/4 and 1/2 square metres, (iii) 1/2 and 3/4 square metres, or (iv) 3/4 and 1 square metre. Since these options exhaust the space of probabilities and are to be equi-probable, they each have to have a probability of 1/4. Which probability assignment we should make—according to the Principle of Indifference—depends on how we represent the scenario, just as which curve we should choose—according to the simplicity solution to the curve-fitting problem—depends on how the curves are represented. Note that the justification of the equi-probability of the events of the side-length being in [0, 1/2] or [1/2, 1] was in terms of the Lebesgue measure, or Euclidean metric. But these provide the very same justification of the equi-probability of the events of the side-area being in [0, 1/4], [1/4, 1/2], [1/2, 3/4], or [3/4, 1]. We see then that the notion of the size of a set of possibilities—measured by a measure function, or metric—is representation dependent.

This type of paradox is also used as a standard objection to any theory of logical probability which relies on the Principle of Indifference (see e.g., Gillies [2000], p. 37). This brings us to another point of relevance which can be seen when we look at the details of Popper's simplicity solution to the curve-fitting problem. Instead of attributing simplicity to curves, Popper attributed simplicity to *families* of curves, which I will call *models*. For example, the set of straight lines, LIN, is one model, and the set of parabolas, PAR, is another. Popper wanted to identify the simplicity of a model with its falsifiability. According to Popper, the falsifiability of a model is complementary to its logical probability of containing the true curve (hypothesis) (Popper [1959], p. 102). (From here on, I will simply speak of a model's probability). So for example, LIN has a lower logical probability than PAR because it is a proper subset of PAR, thus it is more falsifiable and simpler than PAR.[4] But this relation between simplicity and falsifiability forced Popper into a dilemma: either he had to adopt a theory of logical probability based on an indifference principle, or restrict the theory so that it only applied to hypotheses that are related by the subset (entailment) relation. Such an indifference principle would ultimately rely on a measure function, or metric over the hypothesis space. On this point Popper writes:

> "I still believe that the attempt to make all statements comparable by introducing a metric must contain an arbitrary, extra logical element. [...] For it can be shown that the metric of content or falsifiability would have to be a function of the metric of the predicate; and the latter must always contain an arbitrary, or

[4] It's a theorem of probability that if $A \subset B$, then $P(A) < P(B)$.

> at any rate an extra-logical element.” Popper [1959] (edition?), pp. 101–2

Popper’s refusal to make use of an extra-logical metric meant that his simplicity solution to the curve-fitting problem could only at best be a partial solution, since curves of the forms $y = \alpha x$ and $y = \alpha x^2 + \beta x^3 + \gamma x^4$ (for instance), could not be compared with respect to their simplicity/falsifiability, as they do not stand in the subset relation. If however, Popper had wanted to make such comparisons, he would have required an indifference principle (or just some general principled way of assigning probabilities to models), which would thus infect his solution to the curve-fitting problem with the probability paradoxes mentioned above.

One might take the upshot of all this to be that the simplicity solution to the traditional curve-fitting problem cannot be correct, since Priest’s example shows that the hypothesis which the simplicity solution dictates depends on which system of representation we choose. I will say more about this in §4.

1.4. The Real Curve-Fitting Problem

The traditional curve-fitting problem is an idealised version of a problem of inductive inference that occurs frequently in scientific practice. It is idealised in the respect that the problem assumes that the curve which ought to be chosen must pass through each data point *exactly.* As Goodman points out, this is not typical:

> “Seldom does the chosen curve pass exactly through each of the points plotted; sometimes it may miss them all. Rather than choosing the simplest among the complex curves that fit the evidence, we choose among simple curves the one that comes nearest to fitting the evidence.” Goodman [1972], p. 346

This is because quite often there is error in the data; the points plotted do not perfectly represent the true values of the quantities in question. The problem then is to somehow ‘see’ through the noise in the data and pick out the trend (assuming there is one) that captures the true relationship between between the two quantities. I will call this problem the *real* curve-fitting problem.

1.5. Gruesome Accuracy

The real curve-fitting problem is interesting, philosophically, because certain conceptual issues arise which were hidden in the traditional version of the problem. The first issue is that we now need to understand what it means for a curve to be the curve that ‘comes *nearest* to fitting the evidence’. Clearly, there are curves that are nearer than others to fitting the data. So we need some way of comparing how close curves are to fitting the data (to be able to find the closest). I will call how close a curve is to fitting the data its *accuracy*.[5]

One common way to measure the distance between a curve and the data is by the sum of squared residuals (SSR):

$$SSR(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2$$

where (x_i, y_i) are the N data points, and f is the function associated with the curve in the (x, y) plane. To find the curve which is closest to the data, we find the curve which minimises this sum.[6]

Miller [1975] famously showed that SSR, as a measure of accuracy, is inappropriate because the measure of accuracy it assigns to any particular curve depends on how that curve is represented. SSR can be written in terms of the Euclidean metric, d, on y:

$$\begin{aligned} SSR(f) &= \sum_{i=1}^{N} d(f(x_i), y_i)^2 \\ \text{where } d &: \mathbb{R} \times \mathbb{R} \to \mathbb{R} \text{ s.t} \\ d(y_1, y_2) &= |y_1 - y_2| \end{aligned}$$

If we represent the data in a different plane, say (x, Y), then SSR can be written in terms of the Euclidean metric, d', on Y. When we transform d' onto the y plane we find that, for some choices of Y, $d' \neq d$. Thus, SSR applied to the data in the (x, Y) plane can give a different result from the result delivered by SSR when it is applied to the data in the (x, y) plane. Hence SSR depends on how we represent the data.

A somewhat disturbing consequence of this representation dependence is that we can make any false theory, f, be as close to the data as we like (i.e., make $SSR(f)$ as small as we want). All we need to do is find a particular (x, Y) so that according to the Euclidean metric on Y, $SSR(f)$ is as small as we want it to be. Even more disturbingly, we can find a particular (x, Y) so that if f is closer to the data in (x, y) than g, then g is closer to the data in (x, Y) than f—our choice of variables can *reverse* the comparative accuracy of two curves. Note the similarity between this problem and the problem for the simplicity solution to the traditional curve-fitting problem, and Bertrand's Paradox.

In light of this result, a popular response has been to measure accuracy by

[5] It is not necessary that a measure of accuracy provide precise numerical values, a comparative measure may suffice. But common measures do in fact provide precise numerical values, as we will see.

[6] Of course, it may not be unique.

the Maximum Likelihood Estimator (MLE), or its cousin: the Maximum Log-Likelihood Estimator (MLLE) (Good [1975], pp. X; Forster [1999], pp. 98–99). Curves are now associated with error distributions. For example, the model LIN—the family of straight lines—is defined as:

$$LIN(x, y) = \{y = f(x)|f(x) = \alpha x + \beta + \mu; \alpha, \beta \in \mathbb{R}\}$$

where μ is an error term, which has an associated probability distribution. MLLE works by selecting the curve, f, which maximises:

$$MLLE(f) = \text{Log-Likelihood}(f; \text{Data}) = \log(P(\text{Data}|f))$$

This will be the curve which makes the observed data most probable. MLLE is more general than SSR and the two measures give the same results when the error term, μ, is Gaussian. It can be shown that MLLE, as measure of accuracy, has some nice transformation invariance properties which SSR lacks (Good [1975], p. X).

However, the drawback is that MLLE requires us to have more information than that which SSR does. In the absence of this information, MLLE is either unusable or becomes dependent on how we choose to represent the data. For example, when we have no information about the error term, it is quite common to just assume that the error term is Gaussian. But it cannot be both Gaussian over y and also Gaussian over some non-affine 1-1 transformations of y. So in a large class of cases we still only have measures of accuracy which are representation dependent.[7] For the most part of the rest of this paper, I will put aside the problems of finding a representation independent measure of accuracy, for I want to focus on the second issue which arises in the real curve-fitting problem and did not in the traditional one.

1.6. Balancing Accuracy and Simplicity

This second issue is that while it is in some sense 'good' for a curve to come close to fitting the data, sometimes a curve can come *too close* to fitting the data. We typically do not want to choose a curve which passes through each data point since this would be tracking the noise and not the truth behind the noise—this is the danger known as 'over-fitting'. Figure 1 illustrates the problem of over-fitting. We don't want the chosen curve to be too close to the data, and we obviously don't want the chosen curve to be nowhere near the data, so a balance must be found. Characterising this balance turns out to be a

[7] I don't want to rule out the possibility of a measure which is not representation dependent, but the prospects appear dim.

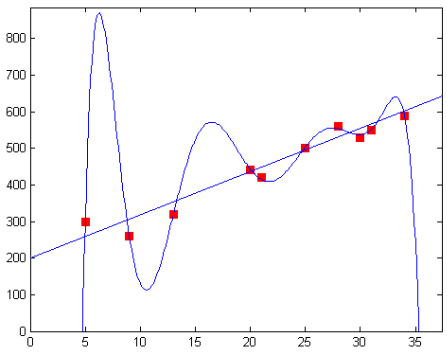

Figure 1: The Danger of Over-Fitting. The 'bumpy' curve clearly does not capture the trend n the data.

tricky business. What we need is a way to recognise patterns, or 'trends' in the data. Curves that fit the data too well, capturing whatever structure happens to be in the noise, tend to be overly complex. But on the other hand, curves that are very simple usually fit the data quite poorly—completely missing the trend (if there is one). So there is a tension between the accuracy of the curve to be chosen and its simplicity, and we need a way to find a balance between these two virtues.

One solution to this problem is to fix upon a particular *family* of curves, Γ, and find the curve in this set which is closest to the data points. By doing this, we decide what the *form* of the curve will be (a straight line, parabola, etc.) and thus we will have chosen not to consider curves that are too complicated. Γ is usually such that there is no curve in it which fits the data perfectly. So our choice of Γ determines how close our curve can get to the observed data.

Sometimes we may have a background theory that suggests that the true relationship between the quantities in question is of a particular form. For example, in mathematical psychology, one model proposed for the relationship between a subject's ability to correctly recall some experience and the time elapsed after the experience is the power model:

$$p(t) = w_1 t^{-w_2}$$

where $p(t)$ is the probability of a correct recall at time t, and w_1 and w_2, are adjustable parameters. Given some data, the task is then to estimate values for the parameters w_1 and w_2 so as to choose a curve from the above model which best fits the data.

However, it is not always the case that we have a model for which we are estimating the values of its parameters, as Glymour points out:

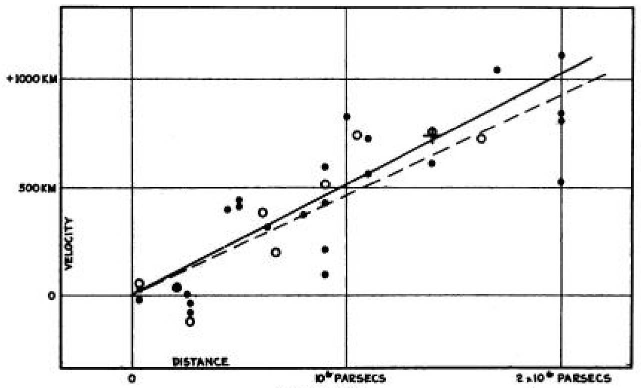

FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Radial velocities, corrected for solar motion, are plotted against distances estimated from involved stars and mean luminosities of nebulae in a cluster. The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.

Figure 2: Graphical version of the data presented in Hubble [1929].

> "The business of parameter estimation is well studied, and its themes are familiar ones to every student of statistics. In contrast, there is very little technical work in the literature on the business of choosing a form of relation, or what is the same, specifying an initial parametric family of relations. Sometimes, no doubt, this family or form is dictated by theoretical considerations, but often it is not, and the exceptions can be important ones: witness Kepler, Boyle, Hubble, and their laws." Glymour [1981], p. 323

Before Hubble discovered that the universe was expanding, it was 'understood' that the universe was static. This meant that there should have been no trend between the distances and recessional velocities of extra-galactic nebulae. However, Hubble found that there was, in fact, a linear relationship between these two quantities. In his famous 1929 paper, Hubble writes:

> "The data in the table indicate a linear correlation between distances and velocities, whether the latter are used directly or corrected for solar motion, according to the older solutions. This suggests a new solution for the solar motion in which the distances are introduced as coefficients of the K term, i.e., the velocities are assumed to vary directly with the distances, and hence K represents the velocity at unit distance due to this effect." Hubble [1929], p. 170

Figure 2 includes the graph from Hubble's paper which shows the linear relationship between extra-galactic nebulae distance and recessional velocity. It is clear that any one of a number of other models could have been fitted to Hubble's data. And each of these would also be in conflict with the static-universe background theory.[8] But it is also clear that there is something right about the straight line fitted to Hubble's data, and something wrong about some other more complicated curve fitted to Hubble's data.

So the problem is to come up with a method for extracting trends out of data which can have noise in it, and formalising what is more 'right' about some curves than others. As Forster and Sober [1994] put it:

> "We know that any curve with perfect fit is probably false, but this does not tell us which curve we should regard as true. What we would like is a method for separating the 'trends' in the data from the random deviations from those trends generated by error. A solution to the curve fitting problem will provide a method of this sort." (Forster and Sober [1994], p. 5)

As we have just seen, such a method must also work in cases where we have just the data, and no background theoretical considerations.

1.7. The Akaike Information Criterion

Forster and Sober [1994] introduced a solution to this problem based on the work of Akaike [1976]. The solution makes use of the Akaike Information Criterion (AIC). AIC is a way of balancing the accuracy of a model, Γ, with its simplicity. It does this by measuring the accuracy of a model by MLLE, the simplicity of a model by the parameter dimension of the model, and characterising the trade-off between accuracy and simplicity as:

$$AIC(\Gamma) = \frac{1}{N}\left(\log(P(\text{Data}|L(\Gamma)) - k\right)$$

where $L(\Gamma)$ is the closest curve in Γ, k is the parameter dimension of Γ, and N is the number of data points.[9] We then choose the model, Γ, which maximises AIC. The parameter dimension, k, of a family of curves, Γ, is (roughly) the number of adjustable parameters of Γ. So for instance, the parameter dimension of LIN is 2 and the parameter dimension of PAR is 3. The idea behind this solution

[8] It may be countered that the linear relationship between extra-galactic nebulae distance and recessional velocity was one of the many models suggested by the background theory at the time—the background theory being Einstein's equations. It is quite common in physics, and other sciences, to ignore solutions to equations which don't 'make sense'. Solutions which permit negative length, negative energy, etc. are ignored since they contradict common sense, or central assumptions that are in place. These solutions are not part of the overall theory. For example, before Dirac seriously entertained negative energy solutions to what is now known as the Dirac Equation, and predicted anti-matter, anti-matter was not part of the theory of physics. Similarly, before Hubble, it was 'common sense' that the universe was static—indeed this is why Einstein introduced the cosmological constant into his equations, making what he later called the "biggest blunder" of his life.

[9] Throughout the rest of this paper I will use the notation $L(\Gamma)$ to pick out the curve in the model Γ which is closest to the data, k to denote the number of parameters of a model, and N to denote the number of data points.

to the curve-fitting problem is that $AIC(\Gamma)$ measures the *expected predictive accuracy* of Γ and it is the goal, or at least, *a* goal of science to make accurate predictions.[10]

An example will help illustrate how AIC works. Suppose we are deciding which of the two models, LIN and PAR, best captures the trend in some data that we have collected. LIN has only two parameters, whilst PAR has three, so, on this definition of simplicity, LIN is simpler than PAR, and it is also contained in PAR. But because LIN is inside PAR, PAR is more 'flexible' than LIN, so it will tend to be able to fit the data better. For us to be justified in choosing a curve in the more complicated family, PAR, the trend in the data will have to be sufficiently more parabolic than linear—and AIC will tell us exactly how much more 'sufficiently more' is. It gives a reward for how close the closest curve in a family is (that's the $\log(P(\text{Data}|L(\Gamma))$ part of the equation) but gives a penalty for complexity (that's the $-k$ part of the equation)—the factor $1/N$ plays no role in the trade-off and can be ignored.

There are many other ways of characterising the trade-off between simplicity and accuracy. For example, the Bayesian Information Criterion (BIC) defines the trade-off between simplicity and accuracy as:

$$BIC(\Gamma) = \frac{1}{N}\left(\log(P(\text{Data}|L(\Gamma)) - \frac{k\log(N)}{2}\right)$$

giving more weight to the simplicity of models. These criteria disagree about how the trade-off between simplicity and accuracy should work, but agree on how these quantities are defined. I do not want to enter the debate about how the trade-off between simplicity and accuracy is meant to work. As Forster [2001] notes, it is not clear that the various information criteria are in conflict with each other, since they arise from different views as to what the goal(s) of science are, or should be. For example, AIC is said to maximise expected predictive accuracy and BIC is said to maximise probability of truth. I won't take a stance on which ought to be the goal(s) of science.

Forster and Sober [1994] argue that their solution to the real curve-fitting problem avoids the kind of problem—demonstrated by Priest [1976]—which the simplicity solution to the traditional curve-fitting problem suffers from (Forster [1999], p. 86). They argue that attributing simplicity to families of curves as opposed to individual curves (like Popper, see §1.2) allows them to avoid Priest's problem:

> "We emphasize that Akaike's Theorem solves the curve-fitting problem without attributing simplicity to specific curves; the quantity k, in the first instance, is a property of families. (Footnote: Thus, the problems of defining simplicity of

[10] See Forster and Sober [1994] for more on the philosophy of science behind the equation.

> curves described by Priest [1976] do not undermine Akaike's proposal.)" Forster and Sober [1994], p. 11

DeVito [1997] argued that Forster and Sober's solution to the real curve-fitting problem is susceptible to a version of the grue paradox. This was very similar to the problem that Priest showed that the simplicity solution to the traditional curve-fitting problem had. However, Forster [1999] and Kieseppä [2001] argued convincingly that DeVito's arguments were flawed. Subsequently, it seems that the final consensus is that there is no grue-like paradox for AIC. For example, Kieseppä writes:

> "[...] the analogy of the riddle [Goodman's new riddle of induction] has little to do with the model selection criteria [AIC] [...]" Kieseppä [2001] p. 787

And Forster [1999] goes into considerable detail to investigate whether or not AIC is language invariant[11], and concludes that it is:

> "In summary, the property of language invariance is an important desideratum for any criterion of model selection. [...] Fortunately, language invariance is built in [to AIC] at the very beginning. [...]" Forster [1999] p. 100

The main point of this paper is to argue that there is indeed a version of the grue paradox for AIC, which is very similar to the one Priest originally identified. I will argue for this point in §2.1 and §2.2. My main target in this paper is AIC, but my objections apply to other approaches to curve-fitting which merely balance accuracy with paucity of parameters (of a family of curves).

2. A Gruesome Problem For Curve-Fitting

According to Akaikean methodology, we should choose the model that maximises the AIC function (this will be the model with the highest estimated predictive accuracy), and choose the curve in this model which is closest to the data:

> "A literal reading of Akaike's Theorem is that we should use the best fitting curve from the family with the highest estimated predictive value." Forster and Sober [1994], p. 18

To literally do this though, we need to consider *all* families of curves. Practical problems aside, this cannot be right, because we get absurd results.

[11] Language invariance is the property of not having what I have been calling representation dependence.

2.1. Russian Families

The first of the absurd results is known as the sub-family problem (Forster and Sober [1994], p. 18).[12] The problem is that in any sufficiently complicated family of curves (i.e., a family with a large number of adjustable parameters) there is a sub-family which contains only one curve with all of its parameters set so that the only curve in this sub-family passes through each data point. This sub-family has no adjustable parameters—all of its parameters are adjust*ed*—so, by the lights of AIC, it is a very simple family whose closest curve (i.e., the only curve in the family) fits the data perfectly. Thus, it is a family with a very high expected predictive accuracy.

Another way to see the problem is to suppose that AIC has chosen a particular family, Γ, which has (say) four adjustable parameters and has consequently selected the curve γ. Then there is another family inside Γ—let's call it Δ—which has only three adjustable parameters *and* contains the curve γ. For example, let:

$$\Gamma = \{y = f(x) | f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3\}$$

and:

$$\gamma = 1 + 2x + 3x^2 + 4x^3$$

and let:

$$\Delta = \{y = f(x) | f(x) = 1 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3\}$$

It is easy to see that $\gamma \in \Gamma$, $\gamma \in \Delta$ and that Δ has fewer *adjustable* parameters than Γ. Since Δ contains γ, the log-likelihood of Δ is the same as the log-likelihood of Γ. Yet, Δ, by construction, is simpler than Γ, so AIC must recommend Δ over Γ. But there is another family:

$$\Theta = \{y = f(x) | f(x) = 1 + 2x + \alpha_2 x^2 + \alpha_3 x^3\}$$

[12] In the literature, the name 'the sub-family problem' appears to refer to two distinct problems. For example, Dowe *et al* [forthcoming] write:

> "The sub-family problem is a generalised version of the curve-fitting problem. Consider any model selection problem in which one family of models, A, is a subset of another family, B. Then for any consistent assignment of priors and for any possible data $p(B|\text{data}) \geq p(A|\text{data})$. How, ask Forster and Sober [...], can Bayesians explain the fact we sometimes prefer model family A to model family B?" Dowe *et al* [forthcoming] p. 45

While Forster and Sober do ask this question of the Bayesian, this is not what they take the sub-family problem to be. I will describe their sub-family problem below, and when I refer to the sub-family problem, I mean the one defined by Forster and Sober.

which is inside Δ and has only two adjustable parameters *and* contains the curve γ. So AIC has to recommend Θ over Δ. It's easy to see that AIC is forced to play this game of russian dolls—or rather, Russian *families*—until it is left with a family that has no adjustable parameters *and* still contains γ, i.e., the singleton family $\{\gamma\}$. Such a singleton family appears to be *ad hoc*, since the family is constructed around a curve which fits the data really well—too well.

As Forster and Sober note, if AIC has to play this game of Russian familes, then we are pushed back to selecting complicated curves that fit the data exactly (ibid). This is precisely what a solution to the curve-fitting problem shouldn't do. To avoid trouble with Russian families, Forster and Sober introduce what they call the Error Theorem. They use the Error Theorem to show that:

> "[...] The Akaike estimates of the predictive accuracy of [the curve, f] obtained by viewing [f] as the best fitting case in the *ad hoc* hierarchy of subfamilies of F tend to be *too high*. [...] [So, we] have good reason not to trust the Akaike accuracy estimates for *ad hoc* subfamilies constructed by fixing adjustable parameters at their maximum likelihood values. We emphasize that this has nothing to do with *when* subfamilies are constructed, or *who* constructs them."
> Forster and Sober [1994], p. 21

To understand the Error Theorem, we first need a definition:

Definition: The **error** of the estimated predictive accuracy of a family, F, is the AIC estimation of the predictive accuracy of family F minus the true predictive accuracy of F. This error is written as Error[Estimated(A(F))].

The Error Theorem allows us to decompose this error into three parts:

The Error Theorem: Error[Estimated(A(F))] = Residual Fitting Error(F) + Common Error + Sub-Family Error(F)

The Common Error is a constant so we can ignore it when making comparisons across families. The Residual Fitting Error is both statistically and epistemically unbiased.[13] The trouble, according to Forster and Sober, is with the Sub-Family Error.

To understand the Sub-Family Error, consider a family, K, with n parameters that contains every family we would like to apply AIC to, and the true curve. (The following paragraphs are drawn largely from Foster and Sober [1994] p. 20). Every curve in this family can be represented as a point in a n-dimensional vector space—one dimension for each parameter. So for example, if K = LIN, then $n = 2$ and the vector (c, m) represents the curve $y = mx + c$.

[13] An estimator is epistemically biased if it over or under estimates the quantity it estimates (see Forster and Sober [1994], p. 16]).

Suppose that the true curve is represented by the vector $\tau = (\tau_1, \tau_2, ..., \tau_n)$ and let us center a coordinate system on this vector (as Forster and Sober do [ibid]). So now, to represent the curve $y = mx + c$ we use the vector $(c - \tau_1, m - \tau_2)$. Let $\boldsymbol{L(K)}$ denote the vector which represents the best fitting curve in K, and let $\boldsymbol{A(F_{max})}$ denote the vector which represents the curve in F with the highest predictive accuracy. The Sub-Family Error of F is then defined as:

$$\text{Sub-Family Error}(F) = \boldsymbol{L(K)} \cdot \boldsymbol{A(F_{max})}$$

where '$\cdot$' denotes the dot product. Forster and Sober claim that the danger here is that the tips of these two vectors (which represent curves in K) may be *close* together, in which case the Sub-Family Error is large and positive (Forster and Sober [1994], p. 20). (This reference to a measurement of *distance* between curves should ring alarm bells). So for example, (to take an extreme case) if we let Θ be the *ad hoc* singleton family of curves, $\{L(K)\}$, then $\boldsymbol{A(\Theta_{max})} = \boldsymbol{L(K)}$, and so:

$$\text{Sub-Family Error}(\Theta) = \boldsymbol{L(K)} \cdot \boldsymbol{A(\Theta_{max})} = ||\boldsymbol{L(K)}||^2$$

will be large and positive (where $|| \cdot ||$ is the usual norm). Thus *ad hoc* families have large and positive sub-family errors.

The problem with all this is that the dot product of two vectors is relative to a choice of basis vectors for the vector space. A (trivial) mathematical fact is that there can be more than one basis for a vector space.[14] Although the dot product is invariant under certain transformations of the basis vectors, it is not invariant under *all* transformations. For example, consider the vectors $\alpha = \alpha_1 \cdot e_1 + \alpha_2 \cdot e_2$ and $\beta = \beta_1 \cdot e_1 + \beta_2 \cdot e_2$, where e_1 and e_2 are the usual basis vectors for $\mathbb{R}^2$. Their dot product is: $\alpha \cdot \beta = \alpha_1\beta_1 + \alpha_2\beta_2$. However, consider the same vectors but expressed in terms of the basis vectors $e'_1 = e_1$ and $e'_2 = e_1 + e_2$: $\alpha = \alpha_1 \cdot e'_1 + (\alpha_2 - \alpha_1) \cdot e'_2$ and $\beta = \beta_1 \cdot e'_1 + (\beta_2 - \beta_1) \cdot e'_2$. Now the dot product of the two vectors is: $\alpha \cdot \beta = \alpha_1\beta_1 + (\alpha_2 - \alpha_1)(\beta_2 - \beta_1)$. It's easy to see that for some choices of α and β, this transformed dot product is not the same as the original. Since the Sub-Family Error is defined as a dot product, it too is subject to this lack of invariance.

Another way to see this is to look at the definition of the Sub-Family Error when a definition of the dot product is used:

$$\text{Sub-Family Error}(F) = ||\boldsymbol{L(K)}|| \cdot ||\boldsymbol{A(F_{max})}|| \cos(\theta)$$

where θ is the angle between the vectors $\boldsymbol{L(K)}$ and $\boldsymbol{A(F_{max})}$. If our coordinate

[14] For example, both $\{(0,1),(1,0)\}$ and $\{(0,1),(1,1)\}$ are bases for $\mathbb{R}^2$.

system on the parameter space is centered on the true curve, then $||\boldsymbol{L(K)}||$ will be non-zero, but if the coordinate system is centered on the best fitting curve in K, then $||\boldsymbol{L(K)}||$ will be zero. In the former case, the Sub-Family Error(F) can be non-zero, and in the latter the Sub-Family Error(F) *has* to be zero. If AIC is meant to be representation independent, then surely the results it delivers should not depend on *our choice* of coordinate system on the parameter space.

Another way to see how the Sub-Family Error is representation dependent is to re-parameterise the vector space which represents the curves in K. For example, in the above example where K = LIN, use $(d = 2c, m)$ instead of (c, m). $(4, 3)$ in the (d, m) space picks out the same curve $(y = 3m + 4/2)$ that $(2, 3)$ picks out in the (c, m) space $(y = 3m + 2)$. Suppose the curve in F which is most predictively accurate is $y = 1$. The vector in the (d, m) space which picks out this curve is $(2, 0)$ and in the (c, m) space it is $(1, 0)$. So, in the (d, m) space the dot product between $\boldsymbol{L(K)}$ and $\boldsymbol{A(F_{max})}$ is: $4 \times 2 + 3 \times 0 = 8$, and in the (c, m) space it is: $2 \times 1 + 3 \times 0 = 2$.

One may object that there could be some other way of formally showing that there is something wrong with these *ad hoc* singleton families. However, there is a very nice argument due to Kieseppä [2001] which shows that resorting to any such formal result must, ultimately, be unsuccessful. As we will see, the main thrust of his argument will appear very familiar.

Before we see Kieseppä's argument, we first need to look at a problem which is very similar to the sub-family problem. Consider a case where we use AIC to choose among the various polynomial models: $M_{pol\text{-}0}, M_{pol\text{-}1}, M_{pol\text{-}2}, ..., M_{pol\text{-}N}$; where we have a large number of data points, and N is also large. (I have switched notation to match Kieseppä's. $M_{pol\text{-}n}$ = POLY-n = $\{y = f(x)|f(x) = \sum_{i=0}^{n} \alpha_i x^n\}$.) Assume that the curve which passes closest to the data is in $M_{pol\text{-}N}$—call this curve h—and further assume that it doesn't fit the data perfectly. Due to the nature of the data, h will be a very 'bumpy' curve. Denote the model which contains all vertical transformations of h (i.e., h plus a constant) $M_{h+const}$. Both $M_{h+const}$ and $M_{pol\text{-}0}$ have only one adjustable parameter, but there is a curve in $M_{h+const}$ which passes closer than any other curve in the polynomial models, namely h, and there is no such curve in $M_{pol\text{-}0}$, so AIC chooses $M_{h+const}$ over $M_{pol\text{-}0}$. This is like the sub-family problem because we constructed the model $M_{h+const}$ around the curve h so that AIC treats it as though it is as simple as $M_{pol\text{-}0}$. From now on I will refer to this problem—the problem that AIC will choose models like $M_{h+const}$—as the Ad Hoc Family Problem (the sub-family problem is a special case of the Ad Hoc Family Problem). Clearly we, or at least Forster and Sober, would like a mathematical result (such as the Error Theorem) which would rule out $M_{h+const}$ from the scope of AIC whilst leaving $M_{pol\text{-}0}$ in. We are not so lucky, as Kieseppa points out:

> "Unfortunately, no such mathematical results can exist. This is because the supposedly relevant feature of the models $M_{h+const}$ and $M_{pol\text{-}0}$—i.e. the number of 'bumps' of their curves—*depends on the way one chooses to represent these models.*" (my emphasis) Kieseppä [2001], p. 783

As Kieseppä notes, this is a consequence of the fact that a straight line can be transformed into a curve, and at the same time, a curve transformed into a straight line by simply changing the coordinate system which they are in—this was originally pointed out by Priest [1976]. However, later on Kieseppä also writes:

> "This point can also be formulated more positively by stating that there is nothing inherently wrong with the unusual model $M_{h+const}$, and that any argument which shows that it is more rational to include in [the scope of AIC] the model $M_{pol\text{-}0}$ than to include in it the model $M_{h+const}$ must be specific for the particular application that one has in mind. Such an argument must show that *in that particular application* it is more rational to assume that the true curve is approximately a horizontal line than to assume that it is approximately a curve which has the unusual shape that all curves of $M_{h+const}$ have, instead of showing that models like $M_{pol\text{-}0}$ have some mathematical feature which models like $M_{h+const}$ lack." (emphasis in original) Kieseppä [2001], p. 791

The problem with this though is that if we have an argument (in a particular application) that the true curve is approximately a horizontal line, then we have an argument for why all complicated models, such as $M_{pol\text{-}N}$, should not be in the scope of AIC in the first place (for the particular application). In *some* cases we may have such arguments but it clearly isn't the case that we *always* have some argument for why complicated models should not be in the scope of AIC. For what might such an argument look like? The argument may be an argument from background theory to a particular form which the true curve must be (as in the mathematical psychology example mentioned in §1.5). But as we have already seen, sometimes there are no (relevant) background theories (e.g., the discovery of Kepler's, Boyle's, and Hubble's laws (Glymour [1981], p. 323)). Forster in particular, also thinks that AIC can be applied to cases where there are no background theories:

> "In fact, we may suppose that there are no background theories. All that is required is that the models share the common goal of predicting the same data." Forster [2001], p. 101

Alternatively, the argument may be one from a trend in the data to the form of the true curve—but this is exactly what a solution to the (real) curve-fitting problem is to provide. (I can think of no other reasonable way to argue that the true curve should be of a particular form). Moreover, to suppose we always have an argument for why the true curve is of a particular form, would be to suppose that we never have to choose between simple and complicated models! The model to be chosen would already be given to us by the argument.

Since there is no general mathematical argument for why models like $M_{h+const}$ should be banned from the scope of AIC, and since it is absurd to suppose that in every application we will have an argument for why the true curve should be of a particular form (or form$s$) we are left only with the option of ruling $M_{h+const}$ out on the basis of $M_{h+const}$ having certain representation dependent properties. Of course, any scientist would reject $M_{h+const}$ straight away, but as Forster notes, the challenge is to understand the *rationality* of such practices (Forster [1995], p. 35).

Note that this problem of representation dependence is not unique to AIC. *Any* solution to the curve-fitting problem that is subject to the Ad Hoc Family Problem will be representation dependent. The Bayesian Information Criterion (BIC), for example, is also subject to the Ad Hoc Family Problem, so it too must be representation dependent.

In this section I have argued that for AIC to be a solution to the curve-fitting problem, it must avoid the Ad Hoc Family Problem. I have also argued that any general way of doing this will result in the representation dependence of AIC. In the next section, I will demonstrate how Priest's original problem comes back at the level of families.

2.2. Gruesome Families

Forster and Sober's running example in their 1994 paper is the application of AIC to the choice between LIN and PAR:

> "A typical inference problem is that of deciding, given a set of seen data (a set of number *pairs*, where the first number is a measured x-value, and the second number is a measured y-value), whether to use LIN or whether PAR is better for the purpose of predicting new data (a set of unseen (x, y) pairs). Since LIN and PAR are competing models, the problem is a problem of *model selection.* [...] The philosophical problem is to understand exactly how scientists should compare models." Forster [2001], p.85

Let's consider how AIC is meant to work for the choice between LIN and PAR. First we have some data, $D(x, y) = \{(x_i, y_i)|i = 1, ..., m\}$, over continuous variables x and y, where m is the number of data points. Then we identify the two families, LIN and PAR, over x and y:

$$LIN(x, y) = \{y = f(x)|f(x) = \alpha + \beta x + \mu\}$$
$$PAR(x, y) = \{y = f(x)|f(x) = \alpha + \beta x + \gamma x^2 + \mu\}$$

where paramters α and β are assumed to be real, and μ is an error term. Having identified these two families, we then measure the estimated predictive accuracy of each family using AIC:

$$AIC(LIN(x,y)) = \frac{1}{m}\left(\log(P(\text{D(x,y)}|L(LIN(x,y))) - 2\right)$$
$$AIC(PAR(x,y)) = \frac{1}{m}\left(\log(P(\text{D(x,y)}|L(PAR(x,y))) - 3\right)$$

If $AIC(LIN(x,y)) \geq AIC(PAR(x,y))$, then AIC recommends $L(LIN(x,y))$ as our final theory, and if $AIC(LIN(x,y)) < AIC(PAR(x,y))$, then AIC recommends $L(PAR(x,y))$ as our final theory.

But we didn't have to represent the data using the x and y variables. We could represent the data as $D(X,Y)$, where X and Y are non-affine one-to-one transformations of x and y respectively. If we had represented the data this way, then we identify the two families, LIN and PAR, over X and Y, instead of x and y:

$$LIN(X,Y) = \{y = f(X)|f(X) = \alpha + \beta X + \mu\}$$
$$PAR(X,Y) = \{y = f(X)|f(X) = \alpha + \beta X + \gamma X^2 + \mu\}$$

Again, we then need to measure the estimated predictive accuracy of each family using AIC:

$$AIC(LIN(X,Y)) = \frac{1}{m}\left(\log(P(\text{D(X,Y)}|L(LIN(X,Y))) - 2\right)$$
$$AIC(PAR(X,Y)) = \frac{1}{m}\left(\log(P(\text{D(X,Y)}|L(PAR(X,Y))) - 3\right)$$

Now either $AIC(LIN(X,Y)) \geq AIC(PAR(X,Y))$, in which case AIC recommends $L(LIN(X,Y))$ as our final theory, or $AIC(LIN(X,Y)) < AIC(PAR(X,Y))$, in which case AIC recommends $L(PAR(X,Y))$ as our final theory. It is easy to find examples where both $L(LIN(X,Y)) \neq L(PAR(x,y))$ and $L(PAR(X,Y)) \neq L(PAR(x,y))$, so that the curve AIC recommends depends on how we choose to represent the data.

The following is such an example. It is a very simple extension of the problem which Priest [1976] used to demonstrate that a grue problem existed for the simplicity solution to the traditional curve-fitting problem. In fact, the example is exactly the same, except that there is one more data point to consider. Suppose we observe the following velocities: $v = 2$, $v = 3$, $v = 4$, and corresponding momenta: $p = 2$, $p = 4$, $p = 14$ of a moving particle. This data can be seen in figure 3. Since kinetic energy is momentum times velocity divided by two, we also observe the following kinetic energies: $E = 6$, $E = 12$, and $E = 28$. The data in terms of v and E can be seen in figure 3.

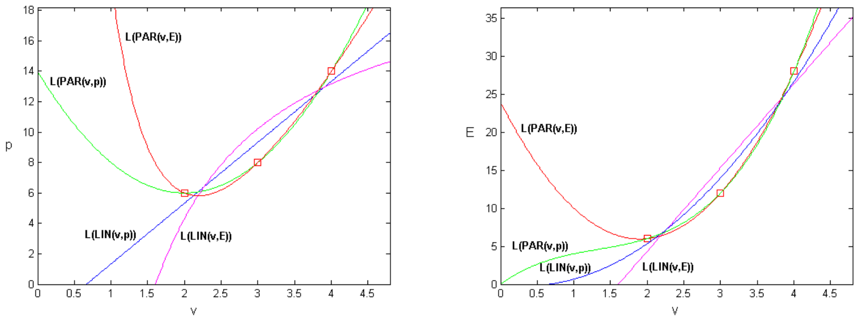

Figure 3: *Left:* The (v, p) plane. *Right:* The (v, E) plane.

When we use AIC to decide between $LIN(v, p)$ and $PAR(v, p)$ and to choose a curve from whichever family has the highest estimated predictive accuracy, AIC selects the curve in figure 3 denoted by $L(PAR(v, p))$. However, when we apply AIC to decide between $LIN(v, E)$ and $PAR(v, E)$ the theorem selects the curve in figure 3 denoted by $L(PAR(v, E))$. It can be seen from both graphs in figure 3 that these curves do not represent the same function. So, AIC gives us inconsistent[15] results, the results depending on which plane we decide to represent the data in.

Note that this problem is different to the problem DeVito [1997] tries to present Forster and Sober with. Forster's reply [1999] to DeVito, I think, adequately shows that DeVito presents no problem for AIC. DeVito makes two claims: (i) that AIC doesn't solve the curve fitting version of Goodman's New Riddle of Induction, and (ii) the notion of simplicity that AIC uses is language dependent. Both claims, I believe, are true. But the truth of (i) is no challenge to AIC, as Forster argues ([1999], pp. 91–95), and the argument DeVito presents for (ii) does not warrant the conclusion, as Forster also argues ([1999], p. 95). I will review Forster's reply to DeVito's second claim here because it is brief and I want to stress that it doesn't apply to my objection, which may easily be confused with DeVito's. This is the problem DeVito presents:

> "Take the data set D, where the curve in family of curves LIN (lines) that best fits the data is $H_1 = \alpha_0 + \alpha_1 x$ and the curve in family of curves PAR (parabolas) that best fits the data is $H_2 = \beta_0 + \beta_1 x + \beta_2 x^2$. Let us assume that the log-likelihood of H_1 equals the log-likeihood of H_2. Then, Akaike's theorem tells us to choose H_1 over H_2 because LIN is simpler than PAR.
>
> [...]
>
> H_1 and H_2 were originally compared in the standard Cartesian coordinate system $[X, Y]$. If we change the coordinate system to $[X', Y]$, where $X' = \beta_1 x + \beta_2 x^2$, we find that H_2 is preferred over H_1. In $[X', Y]$, there is a family of curves PAR', to which H_1' $(\equiv H_1)$ belongs, and a family of curves LIN', to

[15] Not *statistically* inconsistent results.

> which H_2' ($\equiv H_2$) belongs. H_1' is the member of PAR' that best fits the data and H_2' is the member of LIN' that best fits the data. [...]
>
> Since both H_1' and H_2' fit the data equally well and H_1' is a member of a more complex family curves than H_2', Akaike's theorem tells us to choose H_2' over H_1'. But this result is inconsistent with the result achieved in the $[X, Y]$ coordinate system. Once again Akaike's theorem leads to different results when we compare the hypothesis using different conceptualizations of the world." DeVito [1997], p. 394

In reply to DeVito, Forster points out that the transformation which DeVito uses cannot be one to one, since it does not preserve the subset relation (Forster [1999], p. 95). Before the transformation, T, LIN $\subset$ PAR, but after the transformation T(LIN) $\not\subset$ T(PAR). I agree entirely with Forster here. But as I mentioned earlier, my objection is not the same as DeVito's and so cannot be replied to in the same way. My example does not apply the transformation to the families LIN and PAR. My example shows that there at least two coordinate systems (planes), (x, y) and (X, Y) (which are related by a one-to-one non-affine transformation), in which we can construct the families LIN and PAR; hence the extra notation: $LIN(x, y)$, $PAR(x, y)$, $LIN(X, Y)$, and $PAR(X, Y)$.

The problem comes from the fact that, in general, it is not always clear how we are to choose between the coordinate systems, and that the curve which AIC recommends depends upon this choice. Sometimes we may have background information that suggests which coordinate system to construct LIN and PAR in. For example, if X and Y are quantities with no recognised theoretical significance, then we have at least some principled reason for not constructing LIN and PAR in the (X, Y) plane. However, Priest's example, and my extension of Priest's example to the level of families, show that there are cases where both (x, y) and (X, Y) have theoretical significance. As I mentioned earlier, this is actually worse than Goodman's original grue paradox. For we at least have the intuition that somehow the green/blue system of representation is privileged over the grue/bleen system of representation. How could we say that the velocity/momentum system of representation is privileged over the velocity/kinetic energy system of representation, or *vice versa*? Both systems seem perfectly natural.

We saw in the previous section that the space of models to which AIC is applied must be restricted to exclude *ad hoc* models for it to be a viable solution to the curve-fitting problem. In the next section, I will argue that there is yet another way in which the space of models to which AIC is applied needs to be restricted. In doing so, I hope to motivate the claim that *any* solution to the curve-fitting problem will be representation dependent because representation dependence is part of the very job description of a solution to the curve-fitting problem.

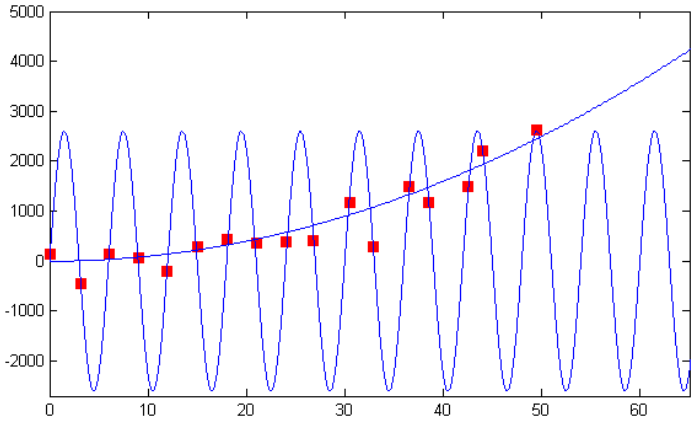

Figure 4: SIN is simpler and can fit the data better than PAR

2.3. SIN is Always Best

Another absurd result arises from the fact that AIC uses a definition of simplicity that is only appropriate for comparing models of a particular type. To see this, consider the family:

$$\text{SIN} = \{y = f(x) | f(x) = \alpha \sin(\beta x);\ \ \alpha, \beta \in \mathbb{R}\}$$

For any consistent[16] set of data points, we can choose a curve in SIN that passes arbitrarily close to the data—by only making use of two adjustable parameters! Hence according to the 'literal' reading of the Akaike Theorem, we should *never* choose a curve in PAR because there will always be a curve in SIN that is simpler—according to simplicity defined as the parameter dimension of the family—and passes closer to the data than $L(PAR)$.[17] Figure 4 compares a SIN curve fitted to the data with a PAR curve fitted to the data. AIC recommends the SIN curve.

In response to this, one might insist that we just restrict our attention to polynomials. However if we do this, then all SIN functions are now infinitely complex. This is because SIN functions are infinite sums of polynomials. For example, the Taylor expansion of $\sin(x)$ is:

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{(2n+1)}$$

[16] Consistent here means that for every x value, there is one unique y value.

[17] Harman and Kulkarni (2003) p.3 make this point.

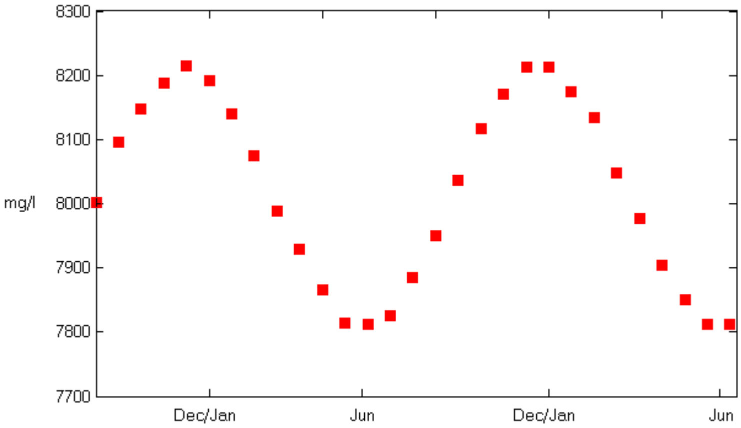

Figure 5: Observed Data

So, instead of *always* choosing a SIN curve over a PAR curve, we now should *never* choose a SIN curve over PAR. It gets much worse than this because SIN is not the only family of curves which are infinite sums of polynomials—all transcendental functions are infinite sums of polynomials. So curves which are, intuitively, quite simple also should never get chosen—for example, $y = e^x$ should never be chosen over any curve in PAR.

The next example shows that even in very simple, realistic cases, we need a notion of simplicity which is more fine grained than simplicity at level of families of curves. Consider the following scenario. A mine site has an unknown amount of salt as a byproduct from its mining practices. It is illegal for the mine to dump its salt in the nearby freshwater river. However, we suspect that they are discharging salt into the river and are interested in whether or not this is having a detrimental affect to the environment down river from the mine site. To assess the environmental impact of the salt, we need to know how the salt concentration varies over time. We collect some data for the salt content in the water at some point not too far down river from the mine. This data is plotted in figure 5. Intuitively, the curve shown in figure 6 would be an ideal curve for this data. But we will see that AIC picks something ridiculous instead. Consider the following two models[18]:

$$SIN\text{-}1 = \{s = f(t) | f(t) = \alpha \sin(\beta t) + c\}$$

$$SIN\text{-}2 = \{s = f(t) | f(t) = \alpha \sin(\beta t) + \gamma \sin(\delta t) + c\}$$

[18] Let us ignore the practical problem of not being able to consider the entire set of SIN-n models. SIN-1 and SIN-2 will be enough for creating serious trouble.

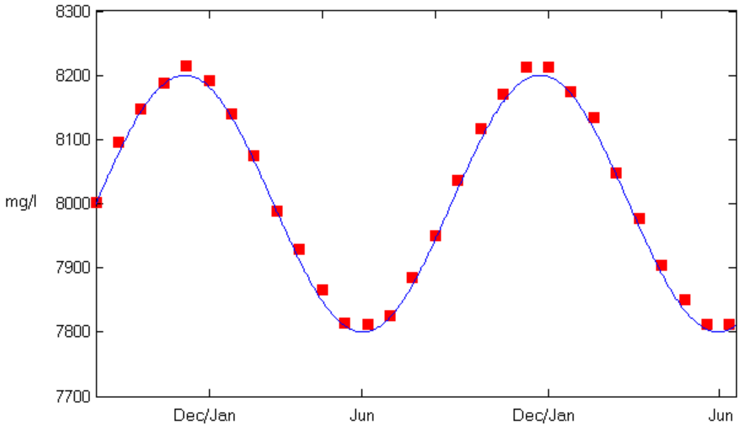

Figure 6: $s = \sin(0.2t)$

Given our data, it is easy to see that AIC will recommend SIN-1 over SIN-2 if L(SIN-2) doesn't fit the data much better than L(SIN-1). This seems right. The problem however, is that AIC delivers absurd results. The curve AIC tells us to choose is *nothing* like the curve in figure 6. This is because a curve from SIN-1 with a very high value for β can fit the data better than a curve from SIN-1 with a more intuitively appropriate value for β. The high frequency curve can be seen in Figure 7. The important point to recognise is that both the high frequency curve and the low frequency curve are in the *same* family, SIN-1. Intuitively, $s = 1.06\sin(3.34t)$ is a much more complicated curve than $s = \sin(0.2t)$.[19] So clearly we want to make a distinction between $s = 1.06\sin(3.34t)$ and $s = \sin(0.2t)$. AIC sees no such distinction. The problem is that within the family SIN-1, there are both the curves which capture the trend in the data, and the curves which fit the data perfectly, giving too much weight to the noise in the data. The above example shows that Forster and Sober's proposed solution to the curve fitting problem is not "a method for separating the 'trends' in the data from the random deviations from those trends generated by error" (Forster and Sober [1994], p. 5). AIC does not recommend the curve (or any nearby curve) which captures the 'trend' in the data. It picks out a curve that is *obsessed* with the random deviations. This is because the definition of simplicity that it uses does not capture the difference between

[19] If you are not willing to grant that it is simplicity at work here, surely you grant that $s = \sin(0.2t)$ is at least *more plausible* to $s = 1.06\sin(3.34t)$, and that any approach to curve-fitting—which is worth its salt!—would capture this difference in plausibility. At any rate, I sympathise with Sober when he writes: "The idea is that choosing the simpler theory means [according to philosophers] regarding it as *more plausible* than its more complex rival." (Sober [2001], p. 13).

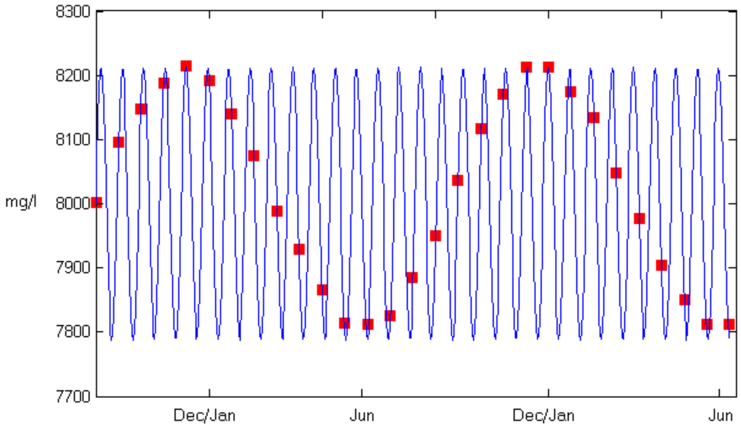

Figure 7: $s = 1.06\sin(3.34t)$

curves such as $s = 1.06\sin(3.34t)$ and $s = \sin(0.2t)$. So much for the claim that AIC solves the curve-fitting problem. Clearly, AIC has only limited successes, if it has any at all.[20]

It seems that in these cases we need to put a simplicity ordering over the SIN curves which correlates with their frequencies—high frequency curves being more complicated than low frequency curves.[21] Something like this does seem right, but such an ordering will depend on the way the curves are represented. Imagine that for the past month we have looked up at the sky at midnight and seen a satellite in roughly the same position every night. Two competing hypotheses which are compatible with these observations are (i) the orbit of the satellite is geostationary and (ii) it has some particular non-geostationary orbit. If we think of these two hypothesis from the point of view of a coordinate system that rotates with the Earth, then we can represent them mathematically as:

$$\begin{aligned} GEOS: \qquad x &= \cos(0t) \\ y &= \sin(0t) \end{aligned}$$

[20] Kieseppä [1999] points out that periodic models do not satisfy certain assumptions built into the Akaike framework, so these results should not be that surprising. One may complain then that I am being unfair to AIC. If AIC were not paraded as a solution to the curve-fitting problem, then I think this would be a reasonable complaint. However, it is, and so the limitations it has need to be pointed out. Also, I will soon use periodic models to argue that *any* solution to the curve-fitting problem must be representation dependent.

[21] Daniel Nolan suggested this to me in personal correspondence. At this point, the meaning of 'simplicity' may be being stretched quite a bit. It does not matter though, all we need is *some* preference ordering. After all, the problem is to get trends out of the data, not to make simplicity solve every aspect of the curve-fitting problem.

$$\neg GEOS: \qquad \begin{aligned} x &= \cos(\omega t) \\ y &= \sin(\omega t) \end{aligned}$$

where ω is a fixed angular velocity. The geostationary hypothesis has a lower frequency than the non-geostationary hypothesis. But of course if we think of the two competing hypotheses in terms of a coordinate system that doesn't rotate with the Earth we get:

$$GEOS: \qquad \begin{aligned} X &= \cos(\omega t) \\ Y &= \sin(\omega t) \end{aligned}$$

$$\neg GEOS: \qquad \begin{aligned} X &= \cos(0t) \\ Y &= \sin(0t) \end{aligned}$$

and now it is the non-geostationary orbit with the lower frequency. Which hypothesis is simpler, depends on how we represent them. We may introduce a richer notion of simplicity to distinguish the two hypothesis. For example, the geostationary hypothesis may entail the existence of more causal mechanisms (orbit stabilizing jets, etc.) than the non-geostationary hypothesis. But that goes beyond the original curve-fitting problem: to make an inference from the data to the true curve—even without the aid of background theoretical considerations.

Another possibility is to compare the number of turning points of the curves.[22] High frequency periodic models have curves with more turning points than the curves in low frequency periodic models (of course this has to be restricted to a bounded domain, otherwise both have infinitely many turning points). However, this too relies on representation dependent properties of the curves that represent the hypotheses, as we saw in Priest's example.

3. A Guide to Truth

Representation dependence seems to be taken to be quite a bad feature to have:

> "Unfortunately, this kind of simplicity depends on the mode of representation—it is language dependent, whereas truth and other epistemic virtues such as predictive accuracy are language independent. That is why pragmatic simplicity is no good as an indicator of truth." Forster (MS), p.36

The idea, I think, is that if a solution to the curve-fitting problem depends on how the data and curves are represented, then a degree of subjectivity is intro-

[22] This was suggested to me by Alan Hájek.

duced, and subjectivity is bad. Similarly for ways of measuring accuracy. SSR is taken to be a bad measure of accuracy because we can change the representation of the data to make any theory as close to the observations we like. If we order theories according to how close they are to the observations, then we can even change this ordering by changing the way the data is represented. But note that scientists were successfully using SSR as a measure of accuracy long before any apparently representation independent measures of accuracy (like MLLE) entered the picture (and it is not clear that MLLE is representation independent anyway, §1.4). So representation dependence can't be *that* bad. Also note that the Principle of Indifference (which is also representation dependent) has been used with great success in science. For instance, Jaynes uses an interesting example to make this point:

> "[G]iven the average particle density and total energy of a gas, predict its viscosity. The answer, evidently, depends on the exact spatial and velocity distributions of the molecules (in fact, it depends critically on position-velocity correlations), and nothing in the given data seems to tells us which distribution to assume. Yet physicists *have* made definite choices, guided by the Principle of Indifference, and they *have* led us to correct and nontrivial predictions of viscosity and many other physical phenomena." Jaynes [1973], pp. 478–9

Of course, if we cannot avoid representation dependence, then some system of representation must be privileged. Actually, this point has already been made (even in the context of curve-fitting) long ago by Post [1960] (an important paper that appears to be overlooked by the contemporary literature):

> "The need for some restriction on the choice of basis-language is apparent from the following: We could, of course, always reduce a curve to a straight line and at the same time change a straight line into a curve by an appropriate change of co-ordinates, thus inverting the order of simplicity." Post [1960], p. 38

However, this cannot be the whole story, as we have seen that on the standard way of determining a choice of basis-language (i.e., naturalness and entrenchment), there can still be competing systems of representation. For instance, in Priest's example, how are we to decide between the momentum/velocity and kinetic energy/velocity systems of representation?

One way in which we may solve this problem for curve-fitting comes from a solution to Betrand's Paradox in probability theory. Remember that van Fraassen's box factory version of the problem was that by using the Principle of Indifference over possible side-lengths we get a different probability assignment to the one we get when we use the principle over possible side-areas. Poincaré [ref] and Jaynes [ref] introduced a method of symmetry considerations to solve Bertrand's Paradox. Van Fraassen [ref] uses this method to solve the box factory problem. The method is essentially that we suppose that there is a solution, and consider what symmetries should be true of the solution. For example, whatever the distribution over possible boxes is, it should dilation invariant. It turns out that the only distribution which has this property is the log uniform distribution. So we get a unique solution to the problem.

Similar reasoning can help us solve Priest's problem. We suppose that there is a unique hypothesis and consider what symmetries should be true of it. Whatever the hypothesis we should choose in Priest's problem is, it should be dilation invariant (i.e., it should not matter if we use kinetic energy instead of velocity, and *vice versa*). It turns out that if we choose the simplest curve in a logarithmic scale, then we get a unique hypothesis.

In the literature on the Principle of Indifference, this success was limited. For it was quickly pointed out that there were other paradoxes which this method of symmetry considerations could not solve. One such paradox is von Mises's Water & Wine Paradox. Suppose we have a glass with 10 cc watered down wine in it. There is at least 1 cc of water in the glass, and at least 1 cc of wine. What is the probability that there is at least 5 cc of water? If we are to use the Principle of Indifference here, then there at least four different quantities we could apply it to: the proportion of (i) wine to total, (ii) water to total, (iii) wine to water, (iv) water to wine. And on each application the Principle of Indifference gives us a different answer (see van Fraassen [1989], p. 314 for details). As van Fraassen points out, symmetries won't help us here (ibid). This because the four quantities in question are related by both dilations and translations, and there is no distribution that is invariant under both transformations.

Mikkelson [2004] has proposed a solution to the von Mises Water & Wine paradox by suggesting that water/wine ratios *supervene* on water and wine quantities, and we should privilege the application of the Principle of Indifference over the more fundamental quantities, water and wine. Ratios supervene on quantities because we cannot change a ratio without changing the quantities of which it is a ratio of, but we can change the quantities (double each, for instance) without changing the ratios. But what are the quantities? Mikkelson seems to think that it is a settled matter that they are water and wine. But we can easily carve the world up using different quantities. For example, consider the quantities *wane* and *witer*, defined as:

$$wane = water/wine$$
$$witer = wine/water$$

Wane and witer ratios supervene on the quantities wane and witer, and of course the two wane and witer ratios correspond to what we normally call water and wine. This problem can be avoided if we are prepared to say that water and wine are more natural (or entrenched, etc.) than wane and witer.

These two different classes of paradoxes—grue-like paradoxes and Bertrand-style paradoxs—are thus quite related, and progress made on one class can be turned into progress on another. I hope to have shown in this section that for both types of problems, if we privilege a natural (or entrenched, etc.) language, make use of symmetry considerations and use simplicity, then we make a great deal of headway. Of course I have not shown that if we do this, then there are no more remaining problems to overcome.

4. Conclusion

In §2.1, I argued that AIC suffers from the Ad Hoc Family Problem, and as a result, is representation dependent. In §2.2, I argued that even when we restrict AIC to a choice between LIN and PAR, the results it delivers turns out to be representation dependent. In §2.3, I argued that AIC fails to handle periodic models and that any approach to the curve-fitting problem which *could* handle these models, would need to rely on properties of the models that are representation dependent. The overall purpose of these arguments was to try to convince the reader that all existing approaches to the curve-fitting problem are representation dependent, and that all future ones will also be representation dependent. What I have hope to have shown is that the idea that we can create a representation independent statistical 'black box' into which we can pump raw data and get the best of theories out (the 'trends' in the data) is doomed to failure. Just as is the idea of a purely syntactic theory of inductive logic is doomed to failure. At some point, some kind of semantic content needs to be added. But if this is done via the use of the naturalness or entrenchment of predicates, then problems still remain. In §3 I showed that the use of symmetry can help overcome these problems. In the end, it seems we need a combination of a privileged basis language, symmetry, and simplicity.[23]

REFERENCES

DeVito, S. [1997] 'A Gruesome Problem for the Curve-Fitting Solution', *Brit. J. Phil. Sci.*, Vol. 48, No. 3., pp. 391–396

Forster, M. R. and Sober, E. [1994] 'How to Tell when Simpler, More Unified, or Less *ad hoc* Theories will Provide More Accurate Predictions', *Brit. J. Phil. Sci.* 45, pp. 1–35

[23] I'd like to thank Alan Hájek, Daniel Nolan, and John Matthewson for helpful comments and discussion.

Forster, M. R. [1995] ‘The Golfer’s Dilemma: A Reply to Kukla on Curve-Fitting’, *Brit. J. Phil. Sci.* 46, pp. 368–360

– – – – – [1999] ‘Model Selection in Science: The Problem of Language Variance’, *Brit. J. Phil. Sci.* 50, pp. 83–102

– – – – – [2001] ‘The New Science of Simplicity’, in *Simplicity, Inference and Modelling* eds. Zellner, Keuzenkamp, and McAleer ...

– – – – – [MS] ‘Theories, Models, and Curves’, http://philosophy.wisc.edu/forster/520/Chapter%202.pdf

Friedman, K. [1973] ‘Son of Grue: Simplicity vs. Entrenchment’, *Nous*, Vol. 7, No. 4., pp. 366–378.

Goodman, N. [1972] *Problems and Projects.* New York: Bobbs-Merrill Co.

Harman, G. and Kulkarni, ?. [2003] ‘Inductive Simplicity and the Martix[???]’

Kieseppä, I. A. [1997] ‘Akaike Information Criteria, Curve-fitting, and the Philosophical Problem of Simplicity’, *Brit. J. Phil. Sci.* 48, pp. 21–48

Kieseppä, I. A. [2001] ‘Statistical Model Selection Criteria and the Philosophical Problem of Underdetermination’, *Brit. J. Phil. Sci.* 52, pp. 761–794

Mikkelson, J. M. [2004] ‘Dissolving the Wine/Water Paradox’, *Brit. J. Phil. Sci.* 55, pp. 137–145

Lewis, D. [1984] ‘Putnam’s Paradox’, *Australian Journal of Philosophy* 62:3 pp.221–236.

Miller, D. [1975] ‘The Accuracy of Predictions’, *Synthese* 30, 1/2, pp. 159-191

Post, H. R. [1960] ‘Simplicity in Scientific Theories’, *Brit. J. Phil. Sci.* 41, pp. 32–41

Popper, K. [2006] *The Logic of Scientific Discovery* bib details here.

Priest, G. [1976] ‘Gruesome Simplicity’, *Philosophy of Science*, Vol. 43, No. 3., pp. 432–437.

Sober, E. [2001] ‘What is the Problem of Simplicity?’ in *Simplicity, Inference and Modelling* eds. Zellner, Keuzenkamp, and McAleer ...

Turney, P. [1990] ‘The Curve Fitting Problem: A Solution’, *Brit. J. Phil. Sci.* 41, pp. 509–530

Van Fraassen, B. [1989] *Laws and Symmetry*, Oxford: Oxford University Press.