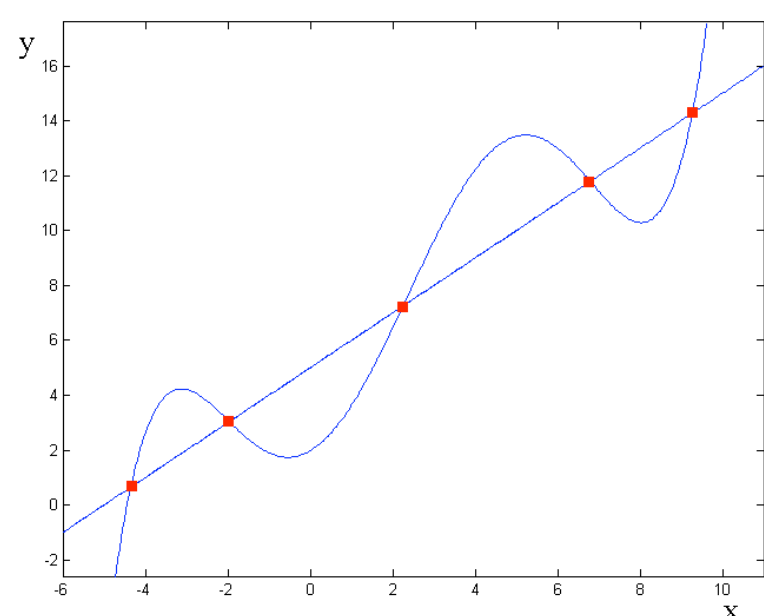


Gruesome Simplicity—A Problem for Akaikean Methodology

The Traditional Curve-Fitting Problem

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. **The problem is that there are infinitely many curves that fit the data. So which should we choose?** I will call this the *traditional* curve-fitting problem. **The standard solution to this problem is to choose the simplest curve that passes through each data point.**



Gruesome Simplicity

Priest [1976] demonstrated that **this solution suffers from an analogue of the grue paradox**. For example, the simplest curve that fits the (v, p) data points, $(2, 6)$ and $(3, 8)$ is:

$$p = 2v + 2 \quad (1)$$

and the simplest curve that fits the (v, E) data points, $(2, 6)$ and $(3, 12)$ is:

$$E = 6v - 6 \quad (2)$$

Supposing that v is the velocity of some particle, p is its momentum, and E is its kinetic energy, then $E = pv/2$. And so (1) becomes:

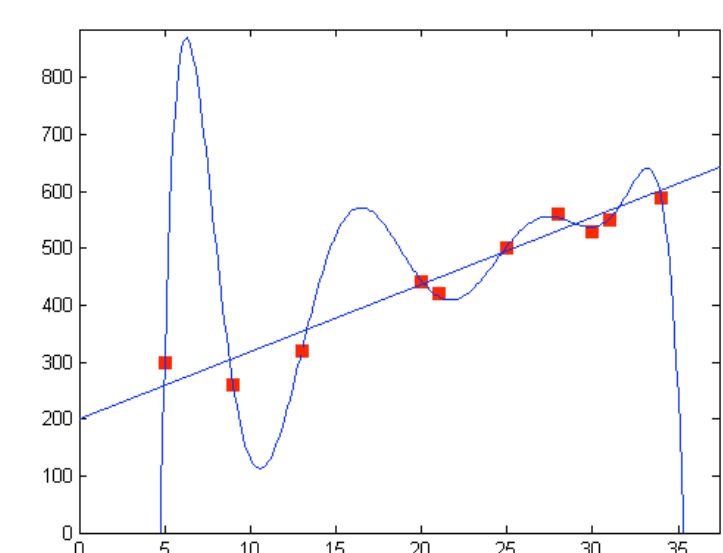
$$E = 12(1 - v^2) \quad (3)$$

which is different to, and more complicated than (2). **So the hypothesis that we should infer from the data depends on how we choose to represent the data.**

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 that ought to be chosen must pass through each data point exactly. But **it is very rare that a curve should fit the data perfectly**. 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 the two quantities.

The real curve-fitting problem is philosophically interesting 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 data’.



Gruesome Accuracy

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

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

Miller [1975] showed that **this measure of accuracy depends on how we choose to represent the data**. In light of this result, a popular response has been to use maximum log-likelihood estimator (MLLE) instead:

$$MLLE(f) = \log(P(\text{Data}|f))$$

Where f now has an error term. MLLE has transformation invariance properties that SSR lacks. The problem however, 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, it is quite common to just assume that the error term is Gaussian. But it can’t be both Gaussian over y and also Gaussian over some non-affine 1-1 transformations of y .

The Akaike Information Criterion

The second issue is that we need to find a balance between the simplicity and accuracy of a curve. Curves that fit the data too well, capturing whatever structure happens to be in the noise, tend to be overly complex curves, but on the other hand curves that are very simple usu-

ally fit the data quite poorly—completely missing the trend. 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.

The Akaike Information Criterion (AIC) is one way of finding this balance. AIC measures the accuracy of a model with MLLE and its simplicity with the parameter dimension of the model, k , and trades them off as follows:

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

where N is the # of data points, and $L(F)$ returns the most accurate curve in model F . The idea is that we choose the model that maximises the AIC function—this will be the model with the highest *expected predictive accuracy*—and then choose the curve in that model that is most accurate. **It is argued by some that this solution to the real curve-fitting problem is not susceptible to the kind of grue-like paradox that the simplicity solution to the traditional curve-fitting problem suffers from. I claim that it is.**

The Sub-Family Problem

The Sub-Family 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 adjusted—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. **The existence of such ad hoc models is a threat to AIC. To resolve this problem, a theorem about how the error behaves is used.**

The Error Theorem & Sub-Family Error

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 $\text{Error}[\text{Estimated}(A(F))]$.

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

The Error Theorem:

$$\text{Error}[\text{Estimated}(A(F))] = \text{Residual Fitting Error}(F) + \text{Common Error} + \text{Sub-Family Error}(F)$$

If we can show that $\text{Error}[\text{Estimated}(A(F))]$ is large and positive when F is *ad hoc*, we will have a principled reason to reject F . It turns out that *ad hoc* families have large Sub-Family Errors. The Sub-Family Error of a family/model F is defined as:

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

where $L(K)$ is the vector that represents the curve that best fits the data in the n -dimensional model K , and $A(F_{\max})$ is the vector that represents the curve in F that is most predictively accurate. **If F is an ad hoc model, then its Sub-Family Error will be large and positive. This is meant to give us reason not to choose ad hoc models.**

Gruesome Sub-Family Errors

The Sub-Family Error of a model, F , is relative to the parameter space used to represent the hypothesis (curve) space. To take a simple example: if $K = \text{LIN}$, and the true curve is (c, m) , then $n = 2$ and the vector $A(F_{\max}) = (c - \tau_1, m - \tau_2)$ represents the curve $y = mx + c$, and the vector $(k_1 - \tau_1, k_2 - \tau_2)$ is $L(K)$. The Sub-Family Error for LIN is then:

$$\text{Sub-Family Error}(\text{LIN}) = (k_1 - \tau_1, k_2 - \tau_2) \cdot (c - \tau_1, m - \tau_2)$$

However, if we use the parameter space (d, m) where $d = 2c$, then $A(F_{\max}) = (d - \tau_1, m - \tau_2)$, so the Sub-Family Error of LIN is then:

$$\text{Sub-Family Error}(\text{LIN}) = (k_1 - \tau_1, k_2 - \tau_2) \cdot (d - \tau_1, m - \tau_2)$$

It is easy to see then that **the Sub-Family Error is language dependent**. Since it is this Error that is used to distinguish *ad hoc* models from other models, **AIC inherits this language dependence**.

Conclusion

This problem of representation dependence is not unique to AIC. **Any solution to the curve-fitting problem which is subject to the Sub-Family Problem will be representation dependent**. The Bayesian Information Criterion (BIC), for example, is also subject to the Sub-Family Problem, so it too must be representation dependent. **Thus, there is indeed a grue-like problem for AIC and similar information criteria.**