June 29, 2012
Minor formatting December 15, 2017; typos June 17, 2018

DRAFT

Chapter from a book, *The Material Theory of Induction*, now in preparation.

Simplicity in Model Selection

John D. Norton
Department of History and Philosophy of Science
University of Pittsburgh
http://www.pitt.edu/~jdnorton

1. Introduction

In philosophical analyses, simplicity is most commonly introduced as a rather abstruse metaphysical notion whose application in theory appraisal is important but troublesome. For the invocation of simplicity seems to require the highest level of human insight, as opposed to the mechanical application of an unambiguous, even algorithmic rule. Hence it was quite a revelation in the philosophy of science literature when Forster and Sober (1994) pointed out that the model selection literature in statistics had succeeded in incorporating a simplicity condition into rules for model selection that are applied mechanically, that is, without the need for higher level human insight.

This example of model selection is important and interesting. However, my sense is that Forster and Sober were too optimistic in just what they thought we can learn from it. They passed too readily from the case of model selection to broader morals pertaining to other cases in which there are invocations of simplicity, such as the decision between Copernican and Ptolemaic astronomy. That is an overreach. The model selection literature shows how simplicity considerations arise in solving a quite specific problem: the discerning of the true relation obscured by random, statistical noise. The simplicity considerations in Copernican and Ptolemaic astronomy are not dependent essentially on error noise. There is a loose similarity between the two cases, but much more needs to be said before general morals can be recovered from the one case of model selection.

My goal in this chapter is more modest. I seek to recover no universal claims about simplicity from this example. Rather I merely want to show how the literature on model selection provides an important illustration of the central claim of the last chapter: that there is no epistemically potent, universal principle of parsimony and that simplicity considerations in theory appraisal are really surrogates for background facts. I will look at hypothesis selection governed by the Akaike Information Criterion, discussed by Forster and Sober. The criterion directs us to evaluate an hypothesis by determining how likely it makes the data at hand. The danger of overfitting is greater, the larger the hypothesis space of the model from which the hypothesis is drawn. The criterion directs that we correct for this overfitting merely by subtracting the dimension of the hypothesis space from the statistic that expresses the likelihood of the data. This correction is its notable property for it rewards models for their simplicity. However, I will argue, the criterion provides no comfort for metaphysicians of simplicity since:

- The criterion is deduced from straightforward assumptions about the systems investigated. These assumptions include no posit of simplicity and no principle of the parsimony of nature.
- The criterion deduced is simply a formula used to weight the performance of various models in narrowly specified condition. No general principle of parsimony is inferred such as could be applied elsewhere.
- Considerations of simplicity need not enter into the discussion at all. They arise only because we metaphysically-minded readers see a particular formula and find it comfortable to interpret one term in it as a reward for simplicity (or punishment for being complicated).

Finally we shall see that the simplicity correction is merely a surrogate for a correction derived from a background assumption. The most potent of the governing assumptions is that the data are generated by an hypothesis in the model under test.[1] That assumption proves strong enough to allow us to estimate how much overfitting the model permits and, as a result, to correct for it in

[1] For a good account of the Akaike Information Criterion, see Konishi and Kitagawa (2008, ch. 3) and especially their Section 3.3 for an account of additional terms needed if the truth is not assumed to be one of the hypotheses under test.

an especially simple way. We then interpret this correction as what simplicity requires, although that notion played no role in its generation.

The chapter will introduce model selection and the Akaike Information Criterion. It is merely one of many such criteria. For our purposes of identifying how generally simplicity considerations enter model selection, it is as good as any.[2] The early part of the chapter will introduce model selection and try to explain how the criterion is able to generate the simplicity correction. The chapter will then turn to a fully worked out example of the criterion in action and conclude with an account of its relation to the material theory of induction.

2. Model Selection

Model selection deals with data generated by some probabilistic system. A model consists of a set of hypotheses such that each is a candidate description of the probabilistic system. A primary application is the example of curve fitting discussed in the last chapter, in which data is generated by a function confounded by statistical noise. The models are the different families of functions that may be fitted: linear functions, quadratic function, and so on and their associated error distributions. However the methods can deal with more general cases and can be applied whenever data are generated probabilistically. If, for example, one samples the heights, weights, genders and so on of a population, the resulting data are generated by a probability distribution that covers these features of the population. In this case, the models are sets of possible distributions and the parameters sought are means, variance, covariances and other parameters of the distributions.

The model selection literature seeks ways to see past the statistical noise in the data to the true system that generated it. For any particular data set, one can always find a better fitting model by sacrificing simplicity. The more complicated model fits better since it can conform to the confounding statistical noise. The larger the model, that is, the more hypotheses it contains,

[2] There is, for example, an extended version of the Akaike criterion modified to correct for small data sets and large numbers of parameters. (Burnham and Anderson, 2004). Other related criteria include the Bayes Information Criterion ("BIC"), which arises in a Bayesian analysis of model selection (Wasserman, 2000).

the greater its ability to conform to the data and the greater the danger of overfitting. The remedy is to forgo some goodness of fit in favor of a simpler model.

A crude illustration is the problem of identifying the daily arrival times of a bus. We may find the bus to arrive at 11:58, 12:04 and 12:02 on successive days. These data are accommodated well enough by the hypothesis that the bus arrives roughly at 12:00. However if we allow more complicated descriptions, we can find a hypothesis that fits the data perfectly. We might propose that the bus arrival times cycle successively through 11:58, 12:04 and 12:02, thereby eliminating any mismatch between our hypothesis and the data at hand. Informally, we would judge the improvement in goodness of fit to be spurious, a result of overfitting, and revert to the "roughly 12:00 arrival" hypothesis as simpler.

3. Maximum Likelihood Criterion

The Akaike Information Criterion "AIC" (Akaike, 1974) is an elaboration of another simpler criterion, the maximum likelihood criterion. Assume we have some probabilistic system that produces data and we wish to infer back to the properties of the system. We identify those properties through the parameters characteristic of the system. These would be the coefficients in the functions we fit to the data in curve fitting; or they might be means and variances if we are trying to find the population parameters from the data of a population sample. To start, we presume some model, that is, some set of hypotheses indexed by the sorts of parameters we believe characteristic of the system. In curve fitting, it would be, say, a linear or quadratic curve confounded by error noise. Different parameters in the model pick out different hypotheses that will make the data actually recovered more or less probable. That conditional probability is called the likelihood L:

$$L = P(\text{data} \mid \text{model parameters})$$

Which parameters should we choose? An obvious choice is those parameters that make the data most probable; that is, we choose to maximize the likelihood L and the resulting parameters are know as "maximum likelihood estimators." It turns out to be convenient not to work with the likelihood L directly but with its logarithm, log L. Since the logarithm function is strictly increasing, maximizing L is equivalent to maximizing log L. And maximizing log L is equivalent to minimizing –log L. This gives us:

Maximum Likelihood Criterion:
Seek the parameters that maximize log L,
that is, that minimize –log L

This criterion works well until we try to use it to compare models with different numbers of parameters. You might expect that we can compare two models by looking at the maximum log-likelihood each supplies. What if best fitting hypothesis H of model M_1 yields a higher log likelihood of the data than does best fitting hypothesis K of model M_2? It would seem straightforward that we should pick the H of model M_1 over the K of model M_2.

That straightforward conclusion is too hasty because the log likelihood delivered by one model can be spuriously inflated by overfitting. For example, in curve fitting, if we use a model with linear functions y=A+Bx, we fit just two parameters, A and B, as well as any parameters characterizing the error noise distribution. If we move to a model with quintic equations $y=A+Bx+Cx^2+ Dx^3+Ex^4+Fx^5$, these two parameters are replaced by six parameters, A, B, C, D, E and F. The larger number of parameters in the second model gives it more flexibility and that gives it an unfair advantage over the first model. The data is generated probabilistically and, as a result, will not perfectly reflect the probabilistic system that generated it. A sample mean will typically differ slightly from a population mean. A maximum likelihood estimator can increase the likelihood of the data by tracking these slight deviations. Selecting the sample mean as the estimator for the population mean will render this particular data set more probable than selecting the true population mean. This unwanted effect is overfitting, once again. As the number of parameters in the model grow, the model becomes more flexible and the extent of overfitting increases.

4. Akaiki Information Criterion

How can we guard against overfitting? Qualitatively, we might seek to protect ourselves by favoring simpler models, that is, models with fewer parameters. That solution is correct at the level of vague generality, but it does not translate into a quantitative procedure with a precise justification that tells us just when to abandon the models with more parameters.

Akaike approached the problem by considering not just performance with the particular data at hand. Instead he asked that we choose estimators that perform well on average over all the data sets that might be produced by the probabilistic system. The motivation is that overfitting produces estimators that work well for one data set to which they are tuned, but will generally fare worse for others that the probabilistic system may produce. A model with a larger set of parameters is more flexible and thus more likely to be overfitted to the data. So, if we seek models that perform well on average, we must penalize the performance of models with larger numbers of parameters to compensate for the inflation in their performance due to overfitting. What Akaike found was that the requirement of best performance on average over all data sets led to a remarkably simple correction to the Maximum Likelihood Criterion. That is, he found that overfitting inflates the log likelihood of the data by the dimension d of the parameter space. We correct the log likelihood function for overfitting merely by subtracting this dimension d from it. What results is:

Akaike Information Criterion (AIC):
Seek the parameters that maximize log L –d;
that is, they minimize[3] –log L + d.

The penalizing factor d automatically favors models with lower numbers of parameters. It expresses in quantitative form the qualitative notion that we should favor the simpler over the more complicated model.

4.1 How it Works: The Essential Assumption

The criterion works by asking not merely how well the estimator performs with the particular data set at hand. Rather it asks how the estimator performs on average with all possible data sets and rewards and penalizes the various models accordingly. For example, if we suspect a population is exactly 50% female, we would not be surprised to find that there are 57 females in a random sample of 100 from the population. We might be tempted by this datum to posit that 57% of population overall is female. The posit would make the datum of 57 females in the sample more probable than the supposition that 50% are female. However, we would likely hesitate. In *this* sample, we might allow, we found 57 females. But what might happen if we draw another random sample of 100; and another; and another? Over the repeated samplings, if

[3] Akaike's original proposal was to minimize - 2log L + 2d, but I have dropped the factor of two since it confounds the simplicity of the formula without any gain.

the 50% hypothesis is correct, we would find a range of sample results scattered around 50 females. The hypothesis of 57% would perform poorly over this range and, on average, the true hypothesis of 50% female would perform best.

The Akaike Information Criterion arises when we correct the performance of an estimator for how it is likely to perform on average over all possible data sets. The great difficulty with this correction is that we do not know the full properties of the true probabilistic system, so, it would seem, we cannot know what all possible data sets are. It is true that we cannot know this without further assumption. We must assume something more. Otherwise the analysis would be performing impossible magic.

The key assumption of the analysis is that the *true probabilistic system lies within the model under consideration*, where a model is simply some collection of hypotheses.[4] So if we are fitting a linear curve y = A + Bx to data, then we assume that some values of A and B are the true values of the system. The remarkable thing about Akaike's analysis is that this assumption is sufficient to allow the analysis to proceed. We do not need to know which values of A and B are the true values. We merely need to assume that there are some values of A and B that coincide with the truth.

What results is a correction to the Maximum Likelihood Condition of impressive simplicity. That simplicity comes at some cost, for it arises only after we have made strong assumptions about the background system and our sampling of it. In addition to the assumption noted above, we also assume that the data set is sufficiently large for the central limit theorem of statistics to be applicable. Nonetheless, it is striking that such a simple correction formula is possible under any conditions. The penalizing factor d merely records the dimension of the space of parameters. The two parameters A and B of linear functions provide two dimensions; the six parameters A, B, C, D, E and F of quintic functions provide six parameters. Nothing else in the details of the space matters.

[4] This is an awkwardness of the application of AIC. This assumption can fail for at least some of the models we may compare. It must fail, for example for all but one, when we compare models with disjoint sets of hypotheses.

4.2 Kullback-Leibler Discrepancy, Predictive Accuracy and the Truth

This discussion has been kept as simple as possible, so a technical note is required, for those who want it. This characterization of how the Akaike Information Criterion works will at first seem different from the way the criterion is normally motivated. Akaike (1974) and later authors (e.g. Zucchini, 2000; Konishi and Kitagawa, 2008, ch.3) employ what is variously called the Kullback-Leibler discrepancy or the Kullback-Leibler information. In seeking to identify a probabilistic system, we seek to identify the probability the system assigns to each possible outcome datum **x**, where the datum **x** is a vector since it will, in general, consist of several numbers. That true but unknown probability is labeled as the probability density g(**x**). The models we fit are also probability densities over the same space of possible outcomes, f(**x**|**θ**), where the vector valued **θ** is the set of parameters characterizing the model. The Kullback-Leibler discrepancy is

$$\mathrm{I}(\mathrm{g};\mathrm{f}) = \int_{\mathrm{all}\ \mathbf{x}} \mathrm{g}(\mathbf{x})[\log \mathrm{g}(\mathbf{x}) - \log \mathrm{f}(\mathbf{x}|\boldsymbol{\theta})]\ \mathrm{dx}$$

It measures how closely the model f(**x**|**θ**) comes to the target g(**x**). It achieves its minimum value of 0 when g(**x**) = f(**x**|**θ**) almost everywhere. The goal is to find that f(**x**|**θ**) that achieves this minimum value. Since the target g(**x**) is fixed, this goal is equivalent to maximizing the integral

$$\int_{\mathrm{all}\ \mathbf{x}} \mathrm{g}(\mathbf{x}) \log \mathrm{f}(\mathbf{x}|\boldsymbol{\theta})\ \mathrm{dx}$$

This integral computes a measure of average performance. The term log f(**x**|**θ**) is the log-likelihood of some particular datum **x**. The density g(**x**) tells us how frequently that datum will appear in repetitions of whatever procedure or experiment generates the data. So the integral is the average log likehood of a datum over many repetitions. Selecting a parameter **θ** that maximizes the integral identifies that density f(**x**|**θ**) that will have the best performance on average in the sense that it renders the data we expect in multiple repetitions most probable.

The f(**x**|**θ**) that is selected by this performance criterion is commonly described as selecting the probability density that has the best "predictive accuracy." In general, it will not be the distribution that makes the data at hand most probable. That distribution may have been eliminated by a penalty for a larger number of parameters. However the one selected will have the property of making the accumulated data most probable over very many repetitions of the procedure. Since these procedures have yet to happen, this feature is labeled “predictive accuracy.”

While predictive accuracy is desirable, it is less than the goal of finding the truth. False theories can enjoy considerable predictive accuracy. The Demeter-Persephone myth of ancient Greece successfully predicts endless repetitions of fertile and barren seasons. Also some model selection problems may preclude prediction. At an archaeological site, we may collect and map the positions of bone fragments. We want to know if their spatial distribution has one or two peaks, which would correspond to one or two sources. In this problem, we are indifferent to prediction, since there are no further bone fragment locations to be predicted. All we really want is the true distribution.

In the particular case of the Akaike Information Criterion, we can see that the maximization is a condition that will return the true probability distribution to us. For the Akaike Information Criterion proceeds from the assumption that the true distribution g(**x**) coincides with one of the distributions in the model. That is,

$$g(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}_0)$$

for $\boldsymbol{\theta}_0$ the true parameter value. Then we seek to optimize the integral

$$\int_{\text{all } \mathbf{x}} f(\mathbf{x}|\boldsymbol{\theta}_0) \log f(\mathbf{x}|\boldsymbol{\theta})\, dx$$

and this integral achieves its maximum value when we set $f(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}_0)$.[5]

The common justification of the Akaike Information Criterion is that it selected the probability distribution that has greatest predictive accuracy. We can now see that this undersells the criterion. It is designed to seek the true probability distribution. Its justification should be given in terms of truth not predictive accuracy.

5. How It Works: An Oversimplified Analogy

That the Akaike Information Criterion can correct for overfitting may seem mysterious and even magical. It is not so. The correction results from implementing a prosaic standard: seek

[5] This follows since the Kullback-Leibler discrepancy I(g:f) has its minimum value of zero when g(**x**) = f(**x**) almost everywhere.

the best performance over all data on average. The correction does not explicitly set out to reward simplicity. That is does so is merely a consequence of the analysis. A greatly oversimplified analogy shows that this sort of correction is far from mysterious.

In this analogy, we will consider the near trivial problem of fitting linear, quadratic, cubic, … curves to data *without error*. That is, we require that the fitted curve must pass through all the data points without error. We seek a criterion that directs us to the unique curve appropriate to the data. We might initially choose the scoring criterion

Number of hits

That is not a good criterion. If we have three data points for (x,y): {(0,0), (1,1), (2,2)}, then the straight line y = x scores three hits. But so do many cubic curves (as shown in Figure 1) and so do many more quartics.

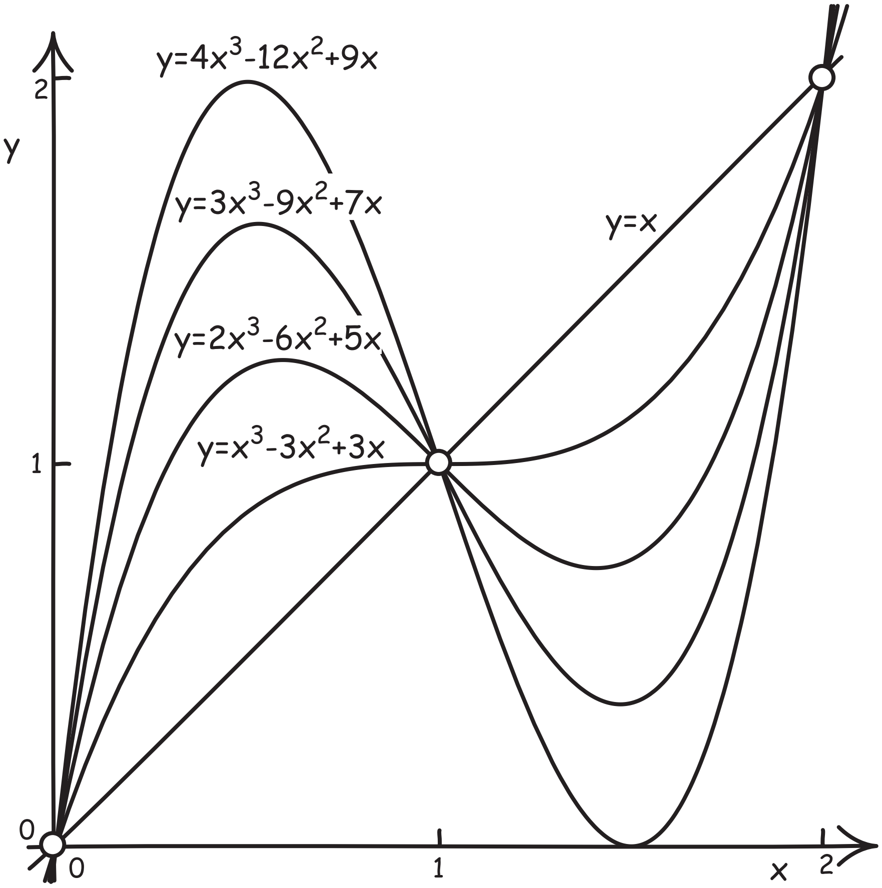

Figure 1. Linear and cubic curve fits

They score equally—3 hits—but they are not equally successful. We discount the cubic and quartic curves since they are not uniquely selected. Cubic curves y = A + Bx + Cx^2 + Dx^3 have 4 free parameters and thus many cubic curves can hit just three data points, but there is only one that can hit four. Quartic curves have five free parameters. Many can hit can three data points, but only one can hit five.

If our interest is uniqueness, instead of counting the number of hits, we should assess whether the number of hits are sufficient to ensure a unique curve. That leads to the new score:

Score = Number of hits – Number of parameters

We have uniqueness if this Score is greater than or equal to zero. For each of the d parameter families of curves mentioned above return a unique curve only when it has a curve that hits d or more points.

This new Score discriminates the linear model from the others in the above case. The linear curve has a Score of 3-2=1, the cubic 3-4=-1 and the quartic 3-5=-1. Only the linear curve has a Score greater than or equal to zero.

The example is elementary, but it manifests two features of model selection methods:

- The score was not derived from a metaphysics of simplicity that demands that more complicated models must be penalized for their lack of simplicity.

Rather all models were held to the same standard: the scoring rewards them only when they produce a unique curve. The result of this requirement was an automatic penalizing of the more complicated models.

- The success of the scoring system depends on background assumptions.

In this case, the curve scoring zero or more is assured to be unique only if the true curve lies in the same model. In the example, if the true curve were actually in the cubic model, then the uniqueness of the straight line y=x for the linear model would be insufficient to assure us that we have found the unique curve. Since we have only three data points, it could be any of many curves in the cubic model.

6. A Coin Tossing Illustration of the Akaike Information Criterion

That the simple correction of the Akaike Information Criterion suffices does seem too good to be true. That it does suffice, under the right conditions, is found merely by working through the statistical analysis that leads to the result. Since this analysis is quite difficult, I have provided a simple application of AIC here and in the Appendix that is designed to display the full analysis and show how it is that a correction merely in the dimension of the parameter space d can be deduced from the requirement of maximizing average performance.

The example pertains to tossing coins. Let us say that we toss N coins and find n heads. What is the chance p of a single toss coming up heads? Our estimation problem is to find that chance. Let us consider models with differing numbers of parameters. Each model assumes independence of the tosses.

6.1 Zero-Parameter Model

The simplest model just posits that our best estimate of p, p̂, is 1/2. It is a rather inflexible model since it allows only one value, but just that is what makes it a zero parameter model. The likelihood L of n heads in N tosses in this model is

$$L_0(1/2) = (1/2)^n\ (1\text{-}1/2)^{n\text{-}N} = 1/2^N.$$

So we have the log likelihood $\log L_0(1/2) = N \log (1/2)$. AIC directs us to maximize

$$L_0(1/2) = N \log (1/2)$$

where no dimensional correction is applied since d = 0.

6.2 One Parameter Model and its Problems

The next simplest model has one parameter, p, which is the chance of a head. The log likelihood of n heads in N tosses is

$$\log L_1(p) = \log (p)^n\ (1\text{-}p)^{n\text{-}N} = n \log p + (N\text{-}n) \log (1\text{-}p)$$

and (as is shown in the Appendix), the value of p that maximizes the log likelihood is

$$\hat{p} = n/N.$$

This model already admits a small amount of overfitting. If, for example, the true value of p is 0.5 = 1/2 and we have N=100 tosses, then n is less likely to be 50 exactly. Rather it will be somewhere in the neighborhood of 50, say n=42 or n=55. Choosing p̂ = 0.42 or 0.55 in these two cases will produce log likelihoods that exceed the log likelihood returned by the zero parameter model, even though, in this case, our supposition is that the zero parameter model happened to have hit upon the true value of p.

Here are the values. The zero parameter model yields

$$\log L_0(1/2) = 100 \log (1/2) = -69.31$$

The one parameter estimators do better when employed with the data sets to which they are tuned:

$$\text{For } n=42, \log L_1(.42) = 42 \log (.42) + 58 \log(.58) = -68.03$$

$$\text{For } n=55, \log L_1(.55) = 55 \log (.55) + 45 \log(.45) = -68.81$$

The one parameter estimators yield greater (i.e. less negative) log likelihoods that does the presumed true zero parameter estimator.

The estimators $\hat{p} = 0.43$ or .55 have performed better in these two cases of n=43 or n=55 since they have been tuned specifically to these two cases respectively. They each perform worse than the zero parameter model, however, if we reverse cases and use $\hat{p} = 0.42$ for the case of n=55 and use $\hat{p} = 0.55$ for the case of n=42.

$$\text{For } n=55, \log L_1(.42) = 55 \log (.42) + 45 \log(.58) = -72.23$$

$$\text{For } n=42, \log L_1(.55) = 42 \log (.55) + 58 \log(.45) = -72.73$$

That is, successes of $\hat{p} = 0.43$ or .55 are inflated by overfitting to the specific data at hand. They will perform worse if we employ them with other data sets to which they are not tuned.

6.3 One Parameter Model Repaired

These effects indicate how we can correct our assessments for overfitting. We give up the goal of merely maximizing log likelihood for the data at hand. Instead we seek to optimize the log likelihood over all possible data sets, appropriately weighting each set for its probability. Finding the estimators that perform best by this standard is basis of the Akaike Information Criterion. This fundamental idea is important enough to bear restatement:

> Seek the estimator that gives the best log likelihood when averaged
> over all possible data sets.

To proceed, we need to know which are all possible data sets. For that we assume:

> There is a single true chance of a head, $p^{*,}$ within the hypotheses
> of the one parameter model.

As I noted above, this is the non-trivial assumption of the analysis, for it says that the truth lies somewhere within our present one parameter space of hypotheses.[6] Our calculations are also greatly simplified with the assumption that the number of tosses N in each data set is very large. That means the central limit theorem of statistics can be called up to assure us that the number of heads n is normally distributed around a mean of p^*N with a variance $N\, p^*(1- p^*)$.

Let us fix some particular maximum likelihood estimator $\hat{p} = \pi$ that is derived from one data set. We can ask how the log likelihood of that particular value π will fare over all possible data sets. That is, we compute the expectation

$$E_{\text{all data}}(\log L_1(\pi)) = N[p^* \log(\pi) + (1- p^*) \log(1-\pi)]$$

where the Appendix gives the computation.

We are interested not just in the performance of one particular estimator π, but in all. So we now average over all estimators. Since $\hat{p}= n/N$, we know that $\hat{p}$ will inherit its distribution from n. It is normally distributed about a mean p^* with variance $p^*(1- p^*)/N$. The expectation over all data and over all $\hat{p}$ yields

$$E_{\text{all } \hat{p},\text{ data}}(\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) - 1/2 \tag{1}$$

The first term on the right is the average log likelihood using the true chance p^* over all data:

$$E_{\text{all data}}(\log L_1(p^*)) = N[p^* \log(p^*) + (1- p^*) \log(1- p^*)]$$

The average in (1) is the quantity that measures the success of the maximum likelihood estimators in the one parameter family. It tells us how their log likelihoods fare on average over all possible data sets and thus is corrected for overfitting. We compare this quantity with the corresponding quantity from other families in choosing our final estimate. We read from (1) that the maximum likelihood estimators fare slightly worse overall than the true value p^*, indicating that we have successfully corrected the overfitting of the maximum likelihood estimators.

However, we are not yet in a position to use (1) since we do not know the value of $E_{\text{all data}}(\log L_1(p^*))$. We need to have some estimate of it since it will vary from parameter space to parameter space and thus affect our choices. We will not be able to determine it exactly. The true value p^* is precisely what is unknown and sought. However there is an indirect way that we

[6] It could fail in many ways. The true chance of heads my vary with different tosses; or there may be correlations between successive toss outcomes.

can recover a good estimate of $E_{\text{all data}}(\log L_1(p^*))$. We use the fact, that for each particular data set, the maximum likelihood estimator $\hat{p}$ tuned to that data set will always outperform the true value p^*.

The extent of overperformance will vary from case to case and will be unknown to us in any particular case. However we can compute its average. To do this, we average over a different set from the one used in (1). That is, we average over pairs of data sets and the estimator best tuned to the data set. That is, we look at a data set and the estimator tuned to it and compare that estimator's log likelihood with that of the true value p^*; and we repeat for many cases. The average that results is expressed by the expectation

$$E_{\text{all } \hat{p}\text{, data tuned to } \hat{p}} (\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) + 1/2 \qquad (2)$$

The Akaike Information Criterion is recovered by combining equations (1) and (2). Equation (2) tells us that, on average in the data sets for which it is computed, the log likelihood $\hat{p}$ will yield a log likelihood greater by 1/2 than that of the true chance p^* averaged over all data. Hence we can use $\log L_1(\hat{p})$-1/2 as an estimator of $E_{\text{all data}}(\log L_1(p^*))$. Inserting this into (1), we find that $\log L_1(\hat{p})-1/2-1/2 = \log L_1(\hat{p})-1$ is an estimator of the quantity we seek to optimize, $E_{\text{all } \hat{p}\text{, data}}(\log L_1(\hat{p}))$. That is, $\log L_1(\hat{p})-1$ is an estimator of the average log likelihood of $\hat{p}$, averaged over all possible data sets. Maximizing this quantity $\log L_1(\hat{p})-1$ is what AIC calls for in the case of a one dimensional parameter space.

6.4 d Parameter Model

It might seem that a major step must be taken from this last case of a one parameter model to the case of a d parameter model. However all the hard work has already been done in computing the one parameter case. It is a small step to a d parameter case. To get there, we divide the N tosses into d subsets of tosses. We posit different true chances, p^*_1 for the first M_1 tosses, p^*_2 for the next M_2 tosses, …, p^*_d for the final M_d tosses. We have now introduced a d parameter model, with parameters $p_1, p_2, \ldots , p_d$. Each subset of tosses can be treated as a separate one dimensional parameter space problem. So in each subset of tosses M_i, we estimate the average of the maximum likelihoods of $\hat{p}_i$ by computing $\log L_1(\hat{p}_i)-1$. The estimate for the average maximum likelihood associated with all d parameters is just the sum of these individual estimators, that is

$$\sum_{i=1}^{d} \log L_1(\hat{p}_i)-1 = \log L_d(\hat{p}_1, \ldots , \hat{p}_1)-d$$

But this last quantity is just the quantity to be maximized in applying Akaike's Information Criterion in the d dimensional parameter space of a d parameter model.

The result still depends upon restrictive assumptions: all of the M_i must be large enough for the central limit theorem to take effect; and we have assumed that some set of values for the p_i expresses the truth exactly. What the calculation also shows is that the character of the parameter space is of lesser importance. The particular magnitudes of the subsets M_i played no role in the final result. They can each be different in size, as a long as they are each large enough to support an application of the central limit theorem. All that matters is that they open new dimensions in the parameter space. It is this fact that enables the criterion to be expressed so simply in terms of parameter space dimension only.

6.5 Akaike Information Criterion Computed

The analysis is specific enough for us to be able to use AIC to compare the zero and one parameter models in a context in which we have an independent intuitive grasp of the competing factors. In 100 coin tosses, if the coin is fair so that the chance of a head is 1/2, we expect the number of heads to lie in the range 40 to 60.[7] When do we choose the hypothesis from the zero or the one parameter models?

For the zero parameter model, the quantity maximized in the Akaike Information Criterion is

$$\log L_0(1/2) = 100 \log (1/2) = - 69.31$$

For the one parameter model, it is

$$\log L_1(\hat{p})-1$$
$$= 100([\hat{p} \log (\hat{p}) +(1- \hat{p}) \log (1- \hat{p})] -1$$

[7] The mean number is 50 and the standard deviation is $\sqrt{100.1/2.1/2} = 5$, so the two standard deviation interval 40-60 and will contain the outcome with probability 0.954.

$$= 100([(n/100) \log (n/100) + (1 - n/100) \log (1 - n/100)]) - 1$$

where n is the number of heads and $\hat{p} = n/100$. If we plot these two quantities as a function of n, we find Figure 2.

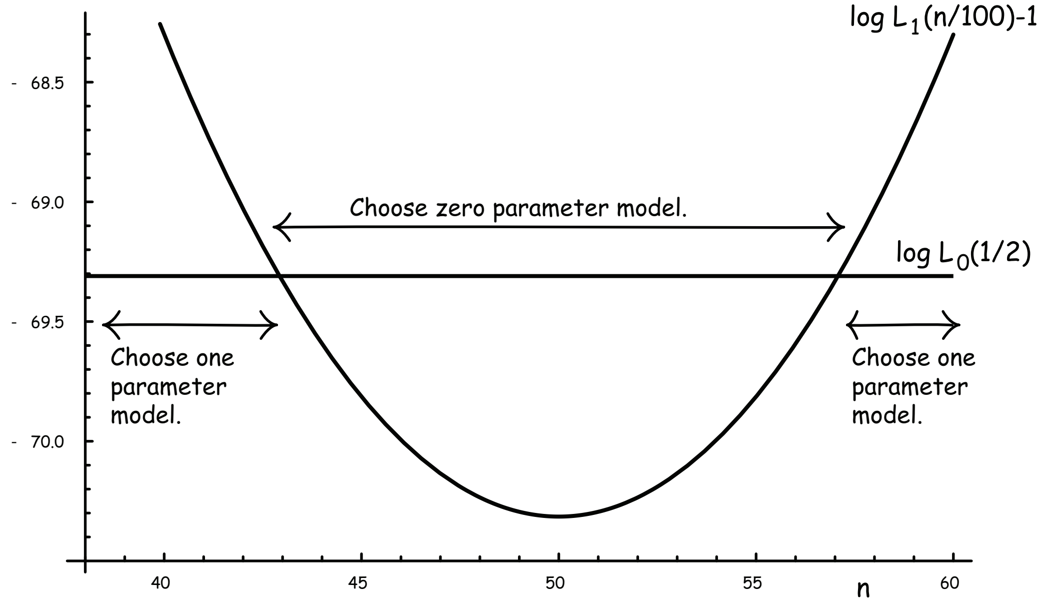

Figure 2. Comparing the zero and one parameter models.

From it, we see that the zero parameter model returns a higher value when n lies between 43 and 57, so we choose the zero parameter estimator p=1/2 for those values. Otherwise, when n falls outside this range, we choose the one parameter estimator $\hat{p}$= n/100.

Here is how we can interpret these results. When we have a datum n=49, the outcome is close enough to the expected value n=50 of the zero parameter model that we prefer the zero parameter model. The one parameter model would give us $\hat{p}$= 0.49 and, as a result, a log likelihood of the data slightly greater than that of p=1/2. However the gain is due to overfitting and not sufficiently great to lead us to switch from the zero parameter value of p=1/2.
If, however, the outcome were to be n=40, then the situation is reversed. The one parameter model gives us $\hat{p}$= 0.40 and a log likelihood for the data that so exceeds the one from p=1/2 that we switch to the one parameter model. These decisions conform with what our vaguer notions would dictate in this case.

7. Relation to the Material Theory of Induction

The main ideas of the connection have already been seen above. I collect them and develop them here. The material theory of induction denies that there is any universal schema for inductive logic. A candidate for such a schema is the idea that we should choose the simpler hypothesis over the more complicated. We have already seen the difficulty with positing this as an independent rule. We still lack any universal characterization of what is simple. At best we can identify the simpler cases on an ad hoc basis according to domains we encounter. The schema also raises the deeper issue of whether it requires us to presume some sort of metaphysics of simplicity. It would assert that the world is, in its essential construction, parsimonious. Are we willing to accept this metaphysics of simplicity? If not, how do we justify the universal schema just described?

The material theory of induction asserts that we should not accept this simplicity schema as universal. Rather it asserts that any schema for inductive inference is warranted by facts and the schema is applicable only in the domains in which those facts obtain. In the case of the Akaike Information Criterion, the essential posit is that the true hypothesis lies somewhere among the hypotheses of the model that we seek to fit. That assumption in turn gives us sufficient access to all possible data sets that the true probabilistic system may generate for us to correct for overfitting by the models.

The derivation of the criterion makes no prior supposition of parsimony or simplicity of the world. It merely asks that we choose estimators that perform well over all possible data sets, not just the ones to which they were initially tuned. The Akaike Information Criterion then follows. That there is any connection to simplicity understood as a general and abstract notion is an interpretation we supply after the analysis is complete. We look at the correction factor d applied to the log likelihood. It reminds us of a vaguer idea that we find it apt to penalize more complicated models with larger numbers of parameters. So it may seem to us that the criterion is somehow vindicating some broader metaphysics of simplicity. That is an illusion and a mistake. The success of the criterion supplies nothing of the sort. We make a mistake in connecting a statistical data analysis procedure, grounded in quite specific assumptions about the case at hand, to some ill-formulated and dubious metaphysics of simplicity.

The following consideration shows how dependent the approach is on the selection of models and how little it can be said to understand deeper notions of simplicity and complexity.

Consider two models. The first is a two parameter model with parameters p_1 and p_2. Call the model $M_2(p_1, p_2)$ and assume that the AIC directs us to select the particular hypothesis with parameters $\hat{p}_1$ and $\hat{p}_2$, chosen since they maximize the penalized log likelihood $\log L_2(p_1, p_2) - 2$. Now consider a second, one parameter model M_1 defined by

$$M_1(p_1) = M_2(p_1, \hat{p}_2)$$

where the log likelihoods of the two models will be related by

$$\log L_1(p_1) = \log L_2(p_1, \hat{p}_2)$$

It is immediately clear that AIC will direct us to favor the one parameter model M_1 over the two parameter model M_2. We can readily find values for which the one parameter model's penalized log likelihood outperforms that of the two parameter model. For example, if in both we set p_1 to the same value $\hat{p}_1$ returned for the two parameter model, we find

$$\log L_1(\hat{p}_1) - 1 > \log L_2(\hat{p}_1, \hat{p}_2) - 2$$

since $\log L_1(\hat{p}_1) = \log L_2(\hat{p}_1, \hat{p}_2)$.

From our elevated perspective, we know that the case is an unfair contrivance. The model M_1 is really just the same as M_2 with one of its parameters artificially hidden by the contrivance of setting it to the estimator value in advance. We would want to say that it is unfair to ask any method to do well against examples precisely contrived to confound them. But that is the point. Calling up some higher perspective, *we* know that the example is contrived. The AIC analysis itself has no way of knowing that. All it can know is that there are two models, a one parameter M_1, and two parameter M_2, which it treats by its rules. The method has no access to which model is really simple and which is maliciously contrived to look simple and has no provisions for treating them differently.

Finally, Forster and Sober's introduction of the Akaike Information Criterion into philosophy of science attracted some spirited responses. For example, De Vito (1997) urged that it could not overcome the language dependence brought by "grue-like" problems. Myrvold and Harper (2002) have pointed out cases in which AIC fails to pick hypotheses that successfully extrapolate.

These are all worthy complaints in so far as they are leveled against the idea that the AIC has somehow vindicated a broader metaphysics of simplicity. However, once one realizes that the real power and proper ambitions of the AIC analysis are much more modest, these concerns pass. Forster (1999) has responded that variant, grueified descriptions cannot change the dimension of the parameter space that is central to the AIC analysis. Also, I will note here, we can only expect the hypothesis selected by an AIC analysis to fare well in extrapolations if the true hypothesis lies within the models considered. Counterexamples in which the AIC selection fails in extrapolation are easily found merely by contriving examples in which the true hypothesis lies outside the models. Then failure of extrapolation is untroubling since the AIC approach, properly understood, has no power to estimate a truth that lies outside its compass. Understood materially, an AIC analysis can only achieve ends authorized by the assumptions made in the analysis. These assumptions fall far short of the positing of a metaphysics of simplicity that can provide universal guidance whenever philosophical issues of simplicity of are raised.

Appendix: Computations for the Akaike Information Criterion in a Simple Coin Tossing Problem

A coin is tossed N times, where N is very large, and the outcome of n heads is reported as the data. In the one parameter model, we assume that the probability of a head in each toss is equal to some undermined probability p, so that the probability of a tail is (1-p). With independence of the tosses, it now follows that the probability of n heads in N tosses is $(p)^n(1-p)^{n-N}$. Hence the one parameter log likelihood is

$$\log L_1(p) = \log (p)^n (1-p)^{n-N} = n \log p + (N-n) \log (1-p)$$

The maximum likelihood estimator is that value of p that maximizes this likelihood. That is, $\hat{p}$ solves the equation

$$0 = (d/dp) \log L_1(p) = n.(1/p) – (n-N).(1/1-p)$$

which leads to

$$\hat{p}= n/N.$$

Thus, the log likelihood of any data set with n heads according to this estimator is

$$\log L_1(\hat{p}) = N [(n/N) \log \hat{p} +(1-n/N) \log (1-\hat{p})]$$

We now seek to assess how well some particular estimator, say $\hat{p}=\pi$, fares when we consider all possible data sets. We assume that the true value of p is p^* and that n/N will differ from its mean value p^* by an amount δ. Writing $n/N = p^*+\delta$, we have

$$\log L_1(\pi) = N\,[(n/N) \log \pi + (1-n/N) \log (1-\pi)]$$

$$= N\,[(p^*+\delta) \log \pi + (1-p^*-\delta) \log (1-\pi)]$$

$$= N\,[p^* \log \pi + (1-p^*) \log (1-\pi)] + \delta \log (\pi/(1-\pi))$$

We now average this quantity over all possible data sets. The number of heads n/N is distributed about the mean p^*. Hence $\delta = n/N - p^*$ has a mean of 0 and vanishes under the expectation operator $E_{\text{all data}}$. Thus we find:[8]

$$E_{\text{all data}}(\log L_1(\pi)) = N[p^* \log (\pi) + (1-p^*) \log (1-\pi)]$$

This expectation depends explicitly on the value of $\hat{p}=\pi$. To suppress it, we now average over the possible values of $\hat{p}$. Writing $\hat{p} = p^*+\Delta$, where we now assume that Δ is small, we have

$$E_{\text{all data}}(\log L_1(\hat{p})) = N[p^* \log (p^*+\Delta) + (1-p^*) \log (1-p^*-\Delta)]$$

We expand the two log terms in a power series:

$$\log (p^*+\Delta) = \log p^* + \log (1 + \Delta/p^*) \approx \log p^* + \Delta/p^* - (1/2)(\Delta/p^*)^2$$

$$\log (1 - p^*- \Delta) = \log (1-p^*) + \log (1 - \Delta/(1-p^*)) \approx \log (1 - p^*) - \Delta/(1-p^*) - (1/2)(\Delta/(1-p^*))^2$$

After substituting, multiplying terms and saving terms up to Δ^2, we have

$$E_{\text{all data}}(\log L_1(\hat{p})) \approx N[p^* \log (p^*) + (1- p^*) \log (1- p^*)] - (1/2)N\Delta^2/(p^*(1-p^*))$$

The quantity Δ is a random variable that inherits its probability distribution from n. When N is large, n is normally distributed[9] with a mean p^*N and a variance $Np^*(1-p^*)$. Since $\hat{p} = n/N$ and $\Delta = \hat{p}-p^* = n/N-p^*$it now follows that $Z = \Delta/\sqrt{p^*(1-p^*)/N}$ is a standard normal variable with mean 0 and variance 1. Hence $Z^2 = \dfrac{N\Delta^2}{p^*(1-p^*)}$ is chi–squared distributed with one degree of

[8] This computation does *not* require the assumption that N is large and that n is normally distributed.

[9] This follows since the exact distribution of n is a binomial distribution with these same parameters. The central limit theorem tells us that this distribution approaches a normal distribution of the same mean and variance for large N.

freedom. This distribution has the property that its mean is unity. Hence taking the expectation of $E_{\text{all data}}(\log L_1(\hat{p}))$ over all values of $\hat{p}$, we recover:

$$E_{\text{all } \hat{p}, \text{ data}} = N[p^* \log p^* + (1-p^*) \log (1-p^*)] - 1/2$$

To identify the first term on the right hand side, note that that the likelihood of n heads according to the correct chance p^* is

$$\log L_1(p^*) = N [(n/N) \log p^* + (1-n/N) \log (1- p^*)]$$

We also have the expectation

$$E_{\text{all data}}(n/N) = p^*$$

so that

$$E_{\text{all data}}(\log L_1(p^*)) = N[p^* \log p^* + (1-p^*) \log (1-p^*)]$$

Combining, we have

$$E_{\text{all } \hat{p}, \text{ data}}(\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) - 1/2 \tag{1}$$

of the main text.

To arrive at (2) we compute the behavior of $\log L_1(\hat{p})$ over the data sets to which each $\hat{p}$ is tuned. To limit ourselves to these data sets, we set $n/N = \hat{p}$ in

$$\log L_1(\hat{p}) = N [(n/N) \log \hat{p} + (1-n/N) \log (1-\hat{p})]$$

and write $\hat{p} = p^*+\Delta$ as before, so that

$$\log L_1(\hat{p}) = N [(p^*+\Delta) \log (p^*+\Delta) + (1-p^*-\Delta) \log (1- p^*-\Delta)]$$

Expanding the log terms as a power series in Δ as before, multiplying out terms and saving terms up to Δ^2, we have

$$\log L_1(\hat{p}) \approx N[p^* \log p^* + (1-p^*) \log (1-p^*)] + N\Delta \log (p^*/(1-p^*) + (1/2)N\Delta^2/(p^*(1-p^*))$$

From above, we have that Δ is a standard normal variable with mean zero and $N\Delta^2/(p^*(1-p^*))$ is chi-squared distributed with one degree of freedom and thus has a mean of 1. Hence we recover the expectation:

$$E_{\text{all } \hat{p}, \text{ data tuned to } \hat{p}} (\log L_1(\hat{p})) = E_{\text{all data}}(\log L_1(p^*)) + 1/2 \tag{2}$$

The quantity to be maximized in AIC is recovered from (1) and (2) as described in the main text.

References

Akaike, Hirotugu (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control*, **19** (No. 6): pp. 716–723.

Burnham, Kenneth P. and Anderson, David R. (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods and Research*, **33**, pp. 261-304.

De Vito, Scott (1997) "A Gruesome Problem for the Curve Fitting Solution," *British Journal for the Philosophy of Science,* **48**, pp. 391-396.

Konishi, Sadanori and Kitagawa, Genshiro (2008) *Information Criteria and Statistical Modeling*, New York, NY: Springer.

Forster, Malcolm (1999), "Model Selection in Science: The Problem of Language Variance," *British Journal for the Philosophy of Science*, **50**, pp. 83-102

Forster, Malcolm 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**, 1–35.

Myrvold, Wayne C. and Harper, William L. (2002) "Model Selection, Simplicity, and Scientific Inference," *Philosophy of Science*, **69**, pp. S135-49.

Wasserman, Larry (2000) "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology,* pp. 92-107.

Zucchini, Walter (2000) "An Introduction to Model Selection," *Journal of Mathematical Psychology,* **44**, pp. 41-61.