THE PREDICTIVE FOCUS ACCOUNT OF THE PRINCIPLE OF SIMPLICITY

BY JUSTIN SHARBER

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Branden Fitelson

and approved by

New Brunswick, New Jersey

October, 2014

ABSTRACT OF THE DISSERTATION

The Predictive Focus Account of the Principle of Simplicity

by Justin Sharber

Dissertation Director: Branden Fitelson

This dissertation presents an account of the Principle of Simplicity, a prominent idea in the philosophy of science. The principle states that when a simple model and a complex model both predict the data, and all else is equal, the data supports the simpler model more. The account, the "Predictive Focus Account," states that simpler models are better confirmed in these contexts because they make narrower, more focused predictions.

The introduction presents Principle of Simplicity and the Predictive Focus Account. It defines key terms and explains the dissertation's methodology. This section also flags background issues that go beyond the scope of the dissertation.

Chapter 1 investigates the philosophical history of the Predictive Focus Account, and the relationship between a model's simplicity and its "global likelihood." On this account, the explanation of the advantage of simplicity is grounded in relations of Bayesian confirmation between competing models. That is, the advantage of simplicity is a subtle, but inherent, feature of standard Bayesian model evaluation. This chapter argues that the Predictive Focus Account is incomplete without a method for fixing prior probabilities. It proposes a new approach to fixing objective priors, the "Data Window Prior," grounded in experimental design. The proposed prior bounds the parameters of statistical models and reigns in their predictions, which are *a priori* unbounded and infinitely extended. So bounded, the models have definite prediction ranges and corresponding degrees of predictive focus. I apply the Data Window Prior to the historical case of Hubble's Law from cosmology, yielding a powerful, intuitive verdict about the confirmation relations between models of varying degree of complexity.

Chapter 2 contrasts the Predictive Focus Account with the more popular Bayesian method of "prior-stacking," whereby Bayesians privilege simpler models and hypotheses with higher prior probabilities. The Predictive Focus Account has distinct advantages over the prior-stacking approach: it shows how simplicity can be favored on *a posteriori*, empirical grounds, and how this favoring relation depends on the nature of the extant data.

Chapter 3 contrasts my account with another, based in Akaike's Information Criterion (AIC), a contemporary, non-Bayesian alternative. The AIC is a statistical model selection criterion that describes estimation error. It is designed to quantify and resist "over-fitting" the data with complex models. The main advantage of the Predictive Focus Account (and corresponding Bayesian method) is its generality. It applies to a wider range of cases and supports a broader range of inferences than the AIC.

Acknowledgements

I am indebted to my exceptional adviser, Branden Fitelson, who helped me wade through ideas and gave me every opportunity to improve. I am indebted to my excellent committee members, Barry Loewer, Tim Maudlin, and Kenny Easwaran, for their advice, help, generosity, and patience on this long endeavor.

The father of the account presented here is Roger Rosenkrantz. In an academic sense, I am grateful to him. Rosenkrantz blended great technical knowledge with philosophical creativity. This dissertation has been a study in humility for me; several times I thought I had an original idea, and then found it tucked away in *Inference, Method, and Decision.*

Special thanks goes to my longtime friend Michael Clark, currently Professor of Chemical Engineering at Lafayette College, who helped me with probability and computation when I was at the drawing board. Thanks for thoughtful discussions to colleagues and friends, including Hanti Lin and my brother, Seth.

Table of Contents

Abstract ii

Acknowledgements iv

List of Figures vii

1. The Data Window Prior for Predictive Focus Calculations 9

1.1. Origin of the Predictive Focus Account 9

1.2. Preliminaries 16

1.3. The Predictive Focus Account 23

1.4. Improper Priors in PFA 26

1.5. Predictive Focus 33

1.6. The Data Window Prior 38

1.7. Calculations with the Data Window Prior 39

1.7.1. Method 1 40

1.7.2. Method 2 41

1.8. A Predictive Focus Calculation on Hubble's Data 41

1.9. Comments on the Data Window Prior 43

1.10. Conclusion 47

2. Can the Bayesian Have It Both Ways? 49

2.1. Introduction 49

2.2. The "Prior-Stacking" Account 52

2.3. Prior-Stacking in the Curve-Fitting Context 56

2.4. Sensitivity to the Dataset 57

2.5. Bayesian Curve-Fitting Methodology 62

2.6. Conclusion 64

3. Simplicity's Strongest Suit 65

3.1. Statistical Accounts 67

3.2. The AIC 70

3.3. The Predictive Focus Account 73

3.4. Extrapolation 77

3.5. "Mere" Hypotheses 82

3.6. Restricted Models 87

3.7. Historical Hypotheses 91

3.8. Conclusion 96

Appendix A. The Hubble–Humason Data Window 97

List of Figures

1.1. Hubble's data. 11

1.2. Best fits on Hubble's data 11

1.3. Rosenkrantz's average likelihoods 14

1.4. Likelihood functions 14

1.5. Likelihoods in curve-fitting 17

1.6. General hypotheses 20

1.7. Example regarding the IUD 28

1.8. GMLs 31

1.9. Kate's snapshot 35

1.10. Approximate counts of images consistent with each model 36

1.11. Possible images 37

1.12. The predictive focus calculation 42

2.1. A stacked prior 54

2.2. Prior-stacking calculation 57

2.3. Two datasets 58

2.4. Prior-stacking calculations 60

2.5. Predictive focus calculations 61

3.1. Resampling 73

3.2. A sampling distribution 77

3.3. Bayesian extrapolation 81

3.4. Dice probabilities 83

3.5. Bayesian coin likelihoods 88

3.6. Frequentist coin likelihoods 89

3.7. Likelihoods on Mendel's ratio 91

Introduction

The core question driving this dissertation is, “What makes simple theories epistemically special?” In philosophy, there is a familiar idea that simplicity is a theoretical virtue, that an agent should never accept or favor a model that is “unnecessarily complex.” Carl Hempel dubs the idea “The Principle of Simplicity” (Hempel, 1998), though the sentiment is most often called “Ockham’s (or Occam’s) Razor,” in reference to an early articulation by William of Ockham.

But why should this principle be true? The obvious philosophical problem with the Principle of Simplicity is that it looks to be chauvinistic, prejudicial. Simple hypotheses are easy to understand, a nice fit for human brains. We might worry: perhaps the principle exists, perhaps scientists and philosophers find it compelling, because they have a fantasy that the world should be easy, straightforward, and convenient. Perhaps we like to think that the world bends to our cognitive limitations, that it will not outstrip our ability to understand it.

On the other hand, we all generally conform to the Principle of Simplicity. We never seriously entertain hypotheses that are both highly complex and have no claim to greater accuracy over simple ones—call this “unnecessary complexity.” Unnecessarily complex hypotheses can be unfamiliar enough to make the Principle of Simplicity, the idea that they are implausible, unfamiliar as well. Ironically, the advantage of simple hypotheses may sometimes go unnoticed because we follow it all too well.

A key case for examining the Principle of Simplicity is Hubble’s Law, which describes the expansion of the universe. Hubble discovered that the distances of other galaxies and other cosmic objects correlated with their velocities. In 1931, he had a critical dataset compatible with a linear relationship (p. 11). While other, more complex relationships were compatible with the data, the linear relationship (a paradigm of simplicity) was clearly the best supported.

Another example is the case of the Nebular Hypothesis of the formation of the solar system, as presented by Laplace. On this hypothesis, the solar system began as a rotating disk of interstellar gas, out of which the sun, planets, and moons coalesced. We might contrast this hypothesis against a fabricated competitor in which the sun gravitationally captures each planet separately, the "Multiple Capture Hypothesis."[1] The Multiple Capture Hypothesis might be a viable hypothesis in some contexts, against some possible datasets. But planets in the solar system all orbit on similar planes, an impressive match for the Nebular Hypothesis. The Multiple Capture Hypothesis is not viable, given that the Nebular Hypothesis also fits the data.

Besides offering an example as to *when* simplicity is a salient consideration, the Nebular Hypothesis also offers a particularly interesting clues as to *why* simplicity bestows a confirmatory advantage. The Multiple Capture Hypothesis is bad in virtue of unnecessary complexity. But it is also bad because we will tend to assign it a low probability, once on we have taken the evidence into account.[2] This hypothesis depends on more independent events, having more adjustable parameters, or having more "moving parts." If we intend to explain why the data fall in such a neat pattern, why they are compatible with the narrow Nebular Hypothesis, we must tell a "just-so" story, an unlikely coincidence where all the "parts" of the complex hypothesis line up just right. On the Multiple Capture Hypothesis, the similarity of planetary orbits is an extraordinary coincidence.

On my account, this dual nature of such complex hypotheses, being both unnecessarily complex and less probable, is at the heart of the Principle of Simplicity. Simple hypotheses are not just-so stories and do not depend on large coincidences to fit the data, so they have much higher probabilities on the data. My account extends the connection between greater complexity and diminished probability to other key cases,

[1] The Multiple Capture Hypothesis is my own invention, a necessary one to show the impact of simplicity in this otherwise historical case.

[2] That the Multiple Capture Hypothesis is improbable on the evidence depends on a Bayesian analysis, and depends on the choice of prior probabilities. I will say a little more about this in a minute.

especially curve-fitting in the historical case of Hubble's Law. Here, some work is required to determine a proper Bayesian probability model over the competing statistical models, but when completed, the simpler model's intuitive advantage is cashed out in cold, hard probability.

Let me make two important notes. First, these probabilities are Bayesian probabilities. The background theory of evidential support is Bayesian Theory, on which the epistemic status of a hypothesis is the subjective probability, the credence, that it is assigned by an agent. This theory requires that hypotheses all start with unconditional probabilities, probabilities that do not take the evidence at hand into account, or *a priori* or "prior" probabilities.

Second, in statistical terms, the kinds of general hypotheses that I have been talking about so far are structured as *models*. In this context, "hypothesis" tends to designate a proposition, describing a feature of or system in the world, that is fully determinate and specific. Models are families of such hypotheses, associated by a similarity or common feature. For Hubble's Law, the inference is a curve-fitting inference, and the hypotheses are specific curves. Models are collections of these curves.[3]

In certain statistical contexts, models and hypotheses are strictly distinct and play different roles in statistical learning and evidential evaluation. But on Bayesian Theory, the parameters of statistical models are treated as random variables, in a subjective or epistemic sense. This allows Bayesian Theory to forge a direct connection between model complexity (often measured in terms of adjustable parameters) and probabilistic confirmation-theoretic verdicts, verdicts which tend to be consonant with our intuitions regarding various cases from the history of science.

This good-making feature of simplicity, couched in probability, is already known to Bayesian statisticians. They call it the "automatic Ockham's Razor" (Smith & Spiegelhalter, 1980). It has also been introduced to philosophy by Roger Rosenkrantz, with an argument in terms of "sample coverage." But both remain less well-known in philosophy, and Bayesians continue to think of simplicity as informing, even biasing our

[3] Each curve graphically represents a function that relates the variables in question. A model replaces certain constants in such a function with adjustable parameters. For more, see Section 1.2.

priors, not as having a more direct (and empirical) relation to evidential support.

The project of this dissertation is to revive, refine, and rebrand this important account of the Principle of Simplicity. I rebrand the account as the "Predictive Focus Account" of the Principle of Simplicity. This account consists of three elementary theses.[4]

Predictive Focus Account

1. A simple model tends to make a prediction (over possible data) with a narrow range.
2. On average, a prediction with a narrow range assigns high confidence over the possible data it contains.
3. Data supports a model through its prediction. All else being equal, if a model predicts the data with high confidence, the data supports the model to a high degree.

Combining these theses yields the probative articulation of the Principle of Simplicity: other things being equal, when a simple and a complex hypothesis both predict the data, the simpler is better confirmed. These theses are not stated as laws, and do not ground an infallible simplicity principle. A *ceteris paribus* clause is in order, with a specific interpretation. I will focus on cases in which competing hypotheses/models are assigned equal prior probabilities.

Some definitions make it possible to unpack this account and translate it into the language of probability. I need six terms for a complete picture. These terms allow me to translate between the qualitative terminology of the Predictive Focus Account and the more quantitative terminology used in the statistical literature on model selection (Section 1.3).

Definitions

Complexity: For most of the dissertation, I am working with curve-fitting models and follow standard statistical practice, taking the complexity of the model to be the

[4] "Predictive Focus" is my term; it is not common in the literature, if it appears anywhere else at all.

number of independently adjustable parameters it contains, k. But some general hypotheses which are not well-described as models with different numbers of adjustable parameters can still exhibit varying degrees of complexity (Section 3.5). Where necessary, I take complexity to be a conceptual basic and follow intuitive judgment.

Prediction Range: The set of events predicted by a hypothesis. I identify this set as a "credible interval" within the relevant model's predictive distribution. These sets are determined using the generalized standard deviation of that distribution.

Prediction Span: A measure of the size of the prediction range. The range is a set; the span is a number.

Predictive Focus: Just the inverse of the prediction span. The smaller the prediction span, the greater the predictive focus. This extra term describes what simple models *have*, bestowing their enigmatic advantage.

Predictive Confidence: The degree to which an event in the prediction range is anticipated by the hypothesis. This is interpreted as likelihood.

Likelihood: The conditional probability of the evidence on the hypothesis. Likelihood plays a key role in Bayesian Theory.

Confirmation: The relation between evidence and a hypothesis where evidence improves the epistemic status of the hypothesis. Here, I adopt the likelihood-ratio measure of degree of confirmation (Fitelson, 2012).[5]

This account explains—in probabilistic terms—the epistemic advantage of simpler hypotheses in paradigm cases from the history of science. For example, it is natural to say that while the Multiple Capture Hypothesis spreads its prediction wide about

[5] Because we are assuming that the priors of competing hypotheses are equal, any measure of confirmation (either as firmness or increase in firmness, in the sense of (Carnap, 1962) will yield the same confirmational verdicts.

the relative orbits of planets, the Nebular Hypothesis gives a narrower, more focused prediction. It associates simplicity with a probative confirmatory property. It gives an answer to why simple hypotheses are better confirmed in the relevant contests: they have more "predictive focus."

Refining the account means placing it not only in clearer terms, but on a stronger foundation. Rosenkrantz and others wanted to focus on the effect of simplicity in a Bayesian context without reference to prior probabilities. But Bayesian Theory essentially involves a prior, and its omission led to problems in the standard accounts. In Chapter 1, I present a new approach for assigning prior probabilities.

The Predictive Focus Account applies to Bayesian evaluation of models and hypotheses, and locates the advantage of simplicity in the model's likelihood. But "likelihood" can be a confusing term in this context: both models and their members have likelihoods, but critically, they can be anti-correlated. On one reading, this is the main point of the Predictive Focus Account; see Section 1.1. To be able to speak of both kinds of likelihood in one breath, I call model likelihoods "global model likelihoods" (GMLs). This term emphasizes that a model's likelihood is a weighted average of its members.

I have to take a few things for granted, philosophically speaking, in order to dive into comparisons of accounts: a language, judgments of complexity, and a theory of evidential support. Let me say something quick about each. First, it is a familiar idea to the philosophy of science that simplicity is dependent on a "language," that is, a lexicon of terms which can feature in hypotheses (Loewer, 1996). Basic terms in the language are simple; complex combinations of these terms are, well, complex. Thus the advantage of simplicity appears to be, in some sense, language-dependent. This might sound like bad news, for the Principle of Simplicity. But in fact, it is completely compatible with the Predictive Focus Account. That compatibility is made possible, in part, in the way that the account does not rely on favoring simple models with higher prior probabilities. The advantage of simplicity happens "downstream" from language choice, and I take our working language of science for granted.[6]

[6] I favor an approach where our choice of language reflects our interests, and take myself to follow Loewer in this regard. Lewis's suggestion, that a language should be constructed such that the simple

Regarding judgments of complexity and simplicity, it is often thought that an account of the Principle should follow directly from a rigorous categorization of simplicity, and there has been concerted philosophical effort dedicated to the descriptive project (Goodman, 1959; Sober, 1975; Swinburne, 1997). But there is a problem with this idea, that follows from the former point: language / conceptual scheme choice determines simplicity. The determination of which terms *are* simple for us would be a descriptive task—it may not even be philosophical. Instead of pursuing this route, I evaluate paradigm cases of simple hypotheses against more complex ones. Along with the Nebular Hypothesis, I take the linear relationship in Hubble's Law and the whole-number ratio in Mendel's Law of Inheritance to be paradigm simple hypotheses.

I adopt the Bayesian Theory of evidence, on which hypotheses are evaluated with subjective probabilities. I assume Joyce's argument (Joyce, 1998, 2005) for the theory. Joyce shows that considerations of epistemic accuracy across possible worlds motivate the theory; one does not need to rest on the potential for irrational bets as in the classic "Dutch Book" arguments. Bayesian Theory also offers a satisfying interpretation of statistical evidence (Royall, 1997; Earman, 1992). I also favor the objective stripe of Bayesian Theory, despite the substantial criticism against it (Seidenfeld, 1979; van Fraassen, 1990). It seems to me that a theory of evidential support must be objective to satisfy the purposes of science, and that an objective theory is possible. I suspect that the keys to such a theory are a proper ground for such distributions, particularly one that has the resources to fix a parametrization on the relevant variables (my prior does this), and a philosophically legitimate interpretation of the role of prior probability. Such interpretations exist: Rosenkrantz (1977, ch. 1) presents a framework in which the goal of the prior probability distribution is to make experiments as informative (in the vein of Shannon information) as possible. Kass & Wasserman (1996) suggests that an objective prior might primarily serve the role of calibrating scientific investigation in a Bayesian context—objective priors are "reference" priors.

To anchor the predictive focus of different models, this dissertation proposes a new

terms refer to natural kinds, is another attractive idea.

approach to fixing prior probabilities, the "Data Window Prior." Chapter 1 describes the method, which uses features of the experiment to fix the priors.[7] This method looks to be a promising prototype for objective Bayesian Theory. It does not imagine that models have primitive prior probabilities, true and pre-existing any evidence. Rather, it looks to the experiment, and consequently the range of hypotheses that it can confirm, to structure and limit how an agent should assess the evidential situation *a priori*, prior to the collection of experimental data.

Some aspects of the Predictive Focus Account have yet to be seen. I have yet to determine the nature of the likelihood function (i.e. a new approximation to this function) in the context of the Data Window Prior. Meanwhile, I rely on a heuristic (but theoretically-driven) approximation to the global model likelihood. Also, although this dissertation describes how experimental design can ground a prior probability distribution, the impact of various experimental parameters on the spread of the predictive distribution has yet to be determined. Nevertheless, the general account presented here, paired with the Data Window Prior, is enough to generate verdicts on scientific cases, and provides a compelling explanation of the enigmatic theoretical advantage of simplicity.

[7] Roger Rosenkrantz is the father of the Predictive Focus Account; he defends it in *Inference, Method, and Decision*. There he suggests relativizing the definition of simplicity to an experiment, but not the prior as such.

Chapter 1

The Data Window Prior for Predictive Focus Calculations

1.1 Origin of the Predictive Focus Account

There is something special about a simple scientific model that fits the data. When the model is simple, the data are plenty, and fit is snug, the model seems to shine with the light of conclusive confirmation. In philosophy of science, curve-fitting is often taken as a paradigm case of this feature of simplicity. Curve-fitting is a visually-accessible kind of inference. Curves represent hypotheses, and datapoints represent observational evidence. There is a persistent tradition in philosophy to understand the role of simplicity in this context (Reichenbach, 1947; Hempel, 1998; Popper, 1992; Quine & Ullian, 1978; Barker, 1966; Glymour, 1980; Sober, 2001) Hermann Weyl (quoted by Popper), makes the first attempt (to my knowledge) to explain why a simple, accurate hypothesis should be so compelling.

> Assume, for example, that twenty coordinated pairs of values of the same function lie (within the expected accuracy) on a straight line, when plotted on square graph paper. We shall then conjecture that we are faced here with a rigorous natural law, and that y depends linearly upon x. And we shall conjecture this because of the *simplicity* of the straight line, or because it would be *so extremely improbable* that just these twenty pairs of arbitrarily chosen observations should lie very nearly on a straight line, and the law in question been a different one.[1] (Popper, 1992, qtd. Sec. 42)

Let us flesh this out using a historical case. In 1931, Edwin Hubble published his second paper describing the "the expansion of the universe:" On the grand scale, cosmic

[1] Weyl goes on to make two important comments that are critical later on. First he says, "If we now use the straight line for interpolation and extrapolation, we get predictions that go beyond what the observations tell us." I count the ability of the account offered here to support extrapolation (in principle, and to a modest degree) as a major point in its favor against a competing account, grounded in a criterion called the AIC, in Section 3.4. Weyl also discusses the problem of language choice. I try to bracket this concern, just touching on it in the Introduction, and saying only a bit more in Section 3.1.

objects are all flying away from each other. Hubble's first dataset (1929) had many observations, but was relatively short-range. The 1931 version was broader and more complete, but featured just ten main observations. Graphically, the data were points, corresponding to the estimated distance and velocity of each cosmic object observed. The data was striking—it was compatible with a linear relationship. Hubble inferred the relationship was linear without further argument (Hubble & Humason, 1931). See Figure 1.1.

To carry Weyl's reasoning over to this case, let us consider three main hypotheses about the relationship between distance and velocity for cosmic objects.

General Hypotheses (Models)[2]

H_1 The relationship is linear.

H_3 The relationship is cubic.

H_5 The relationship is quintic.

Each hypothesis fits the data well, in some sense. Each can accommodate the data well with an instance, a particular curve that comes close to each datapoint (Figure 1.2). But not all are equally favored by Weyl's consideration. To extend the reasoning, we compare each hypothesis against its negation. What is the probability of H accommodating the data, assuming that the truth is anything else? Writing "Fit(H)" to represent the event that hypothesis H fits the data, we can symbolize the pertinent question.[3]

$$P(\,\mathrm{Fit}(H)\,|\,\neg H) = \boxed{?} \tag{1.1}$$

It would be difficult to answer this question without determining the range of possible observations for which Fit(H) holds—the range of observations which each model

[2] We are going to take the first, third, and fifth polynomial families to give us a more dramatic comparison than comparing the first three. So I number the hypotheses 1, 3, and 5. Later, I will explain the models in more detail, Figure 1.6.

[3] Later on, this chapter reorients the discussion from chance probability to predictive focus. It will not be devoting much attention to determining chance-probabilities as such. But I gesture at the form of a chance probability calculation, on page 40.

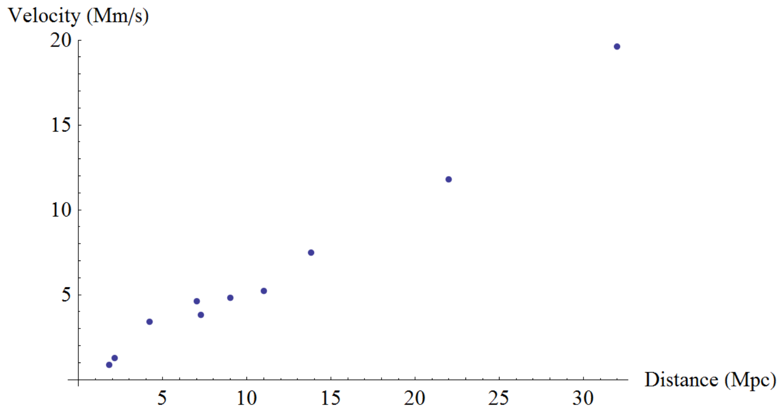


Figure 1.1: Hubble's data.

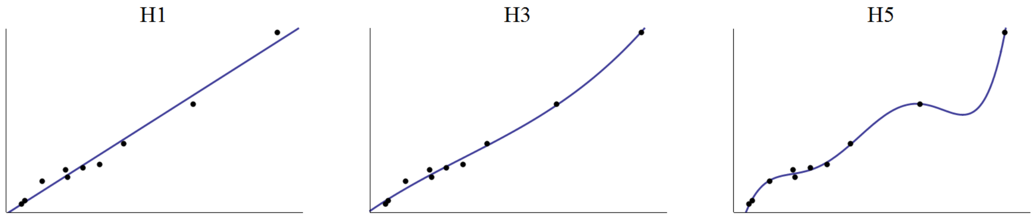


Figure 1.2: The best-fitting curves of each general hypothesis on Hubble's data.

H can accommodate. Figure 1.2 shows a key example of each relationship. All hypotheses are compatible with a broad range of data, but H_1 seems to have the least flexibility. It looks as though H_1 can fit the narrowest range of possible observations. This suggests that H_1 should return the lowest probability in Equation 1.1. In other words, H_1 should have the lowest probability of a "false-positive"—fitting the evidence when it is false.This suggests in turn that H_1's success in fitting the data confirms it over the others.

Fifty years after Weyl's work, Roger Rosenkrantz adopted it directly into a Bayesian framework. Rosenkrantz wrote a seminal book, *Inference, Method, and Decision*, arguing that Bayesian Theory provides the proper framework for evidential support in science. Rosenkrantz argued in terms of a hypothesis's *sample coverage*—the range of observations that the hypothesis can fit by chance, corresponding to a chance probability of fitting the data. A hypothesis' fitting the data by chance has an obvious connection to a hypothesis' admitting a false-positive, and the structure of the two lines of reasoning are the same. Both attempt to explain the confirmatory advantage had by a simple hypothesis. While this inference form is not technically valid, the right background assumptions will support it. Let a chance-probability argument have the following structure.

Chance-Probability Argument

1. General hypothesis H fits the data.
2. The probability of H fitting the data by chance is low.
3. So the probability that H does *not* fit by chance is high. (From 1 and 2)
4. So probably, H is true. (From 3)

Rosenkrantz noticed two things about simple hypotheses. First, that simple hypotheses (or at least, those that people are inclined to judge as simple) tend to admit of a smaller range of possible observations than more complex competitors. Second, that when a simple hypothesis fits the data, it often had a higher likelihood on the data than its complex competitors—that is, the simpler hypothesis assigned a higher probability to the data. And these two features seemed to go hand-in-hand: a simple hypothesis assigns significant probability to "fewer" events (those that it fits), so it assigns more probability to those events.

Mathematically, Rosenkrantz's approach offered a new solution to an old problem. Scientists generally believe that simpler hypotheses are better confirmed by compatible data, that simpler hypotheses have some advantage, but naturally how that advantage is expressed depends on one's theory of statistical evidence. Classical statistics cannot evaluate a general hypothesis as a whole (like H_1)—it can only evaluate specific hypotheses (like the linear curve in Figure 1.2). A classical statistician selects the best curve from a general set, and evaluates that curve on its own. But this produces a bit of a paradox, because on the standard approach, a more complex hypothesis *always* fits the data better than a simpler one—the probability of a simpler hypothesis fitting better is 0.[4] In general, the more complex the model, the better the fit of its best-fitting

[4] For the sake of explaining Rosenkrantz's contribution, I am not providing all the details here. Some key components of the "standard approach" in this context are that models (general hypotheses)

curve, regardless of what the data are. We know in advance that the more complex models perform better, in this respect. This fact suggests that there is something wrong with evaluating general hypotheses with their best instances. Because classical statistics cannot directly evaluate general hypotheses, it had to find another way to deal with this paradox.

However, due to the increased structure in Bayesian Theory, Bayesian statistics is capable of evaluating the likelihood of a general hypothesis as a whole.[5] That is, Bayesian Theory can assign a likelihood to a general hypothesis on a dataset. The likelihood is a distribution that must be integrated.[6] Consistent with classical statistics, the likelihood function always has a higher peak with a complex model, so the likelihood of the data on best-fitting curve is always higher. However, it was also narrower and thinner for a more complex model. The likelihood function had more volume and larger (generalized) standard deviation; it was "wider and thicker" for the simpler model. So when taking the likelihood function as a whole, H_1, for example, had a fighting chance. If H_1's likelihood function is almost as tall as that of H_3 and H_5, and it is a good deal broader, then H_1 will naturally have a higher total likelihood on the data (Figure 1.4).

Pragmatically speaking, integrating over this region can easily be prohibitively difficult. Rosenkrantz derived a computation formula for approximating the integral over the likelihood function. Applying his formula to Hubble's Data yields dramatic results. H_1's general likelihood is 2,000 times that of H_3, vindicating the intuition that this simplest hypothesis is best-supported by the data (Figure 1.3).

Another Bayesian statistician, Gideon Schwarz, also took interest in the limiting behavior of simple general hypotheses, or in the language of statistics, simple models. Making certain assumptions about the likelihood function, Schwarz succeeded in producing a formula for ranking global model likelihoods on the data, the popular

have a nested structure (especially the polynomials), more complex models are supersets, and that error probabilities are continuous (especially normally distributed). This all implies that simple models cannot do better than the complex ones.

[5] The likelihood of a hypothesis is the conditional probability of the evidence, given that hypothesis. See Section 1.2.

[6] The likelihood of a hypothesis is the conditional probability of the evidence on the hypothesis; see 1.2. Integration is the continuous analog of summation, studied in calculus.

Hypothesis	Integrated Likelihood
H_1	3.5×10^{-2}
H_3	1.9×10^{-6}
H_5	8.8×10^{-14}

Figure 1.3: Likelihoods for general hypotheses using Rosenkrantz's formula. (Later, these likelihoods will be called "global model likelihoods / GMLs.")

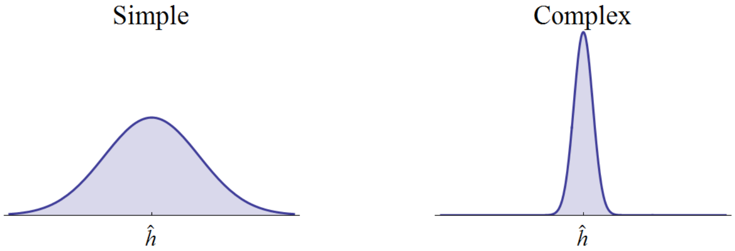


Figure 1.4: A flattened example of the behavior of the likelihood function for a simpler vs. more complex hypothesis. The complex hypothesis always has a higher likelihood for its best instance ($\hat{h}$), but the simpler's likelihood function is wider and tends to have more mass overall. Thus holistic evaluation of the general hypotheses should favor the simpler.

"Bayesian Information Criterion", or BIC (Schwarz, 1978).[7]

The work of Rosenkrantz and Schwarz suggested a general Bayesian account of the advantage of simplicity in confirmatory contexts: the "Predictive Focus Account." The account adopts Weyl's simple reasoning. It connects this reasoning to the behavior of the likelihood function for a model as a whole.

The work of Rosenkrantz and Schwarz was significant, but it never became popular in philosophical circles. For one thing, the work left out some important information. Perhaps part of the problem was what they left unsaid. Rosenkrantz never tells us what the sample coverage of a model is. The reason to prefer a simple hypothesis is explained, but fades into the background. In so doing, he omitted the most interesting piece of the puzzle.

In the same vein, Rosenkrantz's exposition avoids quantifying how much better supported a simple model should be over a complex one on a particular, finite body of evidence.[8] Instead, his argument focuses on the "limiting behavior" of the likelihood function—how the function behaves as the number of datapoints becomes large (approaches infinity), representing what happens as time goes on and more and more experiments have been performed. As long as the contribution of each datapoint favors the simple hypothesis, no matter how little, eventually a large dataset will favor the simple hypothesis significantly. Similarly, Schwarz couched his work in terms of long-run behavior. But Rosenkrantz's omission of sample coverage values meant that the picture was incomplete, and less compelling philosophically. The omission also created a vulnerability in the account. It was all-too-easy to twist the behavior of the Bayesian calculation to reverse the verdict and favor more complex hypotheses (Forster & Sober, 1994).

This chapter will argue that avoiding the question of short-run probabilities on finite evidence results in a subtle incoherence. The only way to resolve the incoherence and secure the Predictive Focus Account is to couple it with a prior probability distribution.

[7] For some philosophical comments on BIC, see Forster (2001).

[8] This chapter will endeavor to do better, but still does not provide an elementary function describing this trade-off.

I propose a prior inspired by the chance-probability argument. Crudely, to build a prior from chance-probability considerations means to build a prior which constrains model predictions consonant with a field of possible observations. Assuming the data will come from an experiment, this essay looks to the experimental apparatus to fix that field of possible observations. The prior is reverse-engineered, as it were, against that field. This prior distribution completes the picture and anchors the Predictive Focus Account of the confirmatory advantage of simple hypotheses.

Section 1.2 provides more detail about Bayesian statistics and curve-fitting. Section 1.3 restates the Bayesian account of simplicity in a preferred version, one that makes its commitments and its effects more transparent. Section 1.4 reevaluates the work of Rosenkrantz and Schwarz with a focus on prior probabilities. Section 1.5 motivates a new prior, the “Data Window Prior.” Section 1.8 applies the new prior to Hubble’s data, evaluating the three general hypotheses introduced here.

1.2 Preliminaries

This section provides background information for Section 1.3. It introduces curve-fitting models, likelihood, and notions of confirmation and prediction. This essay assumes a Bayesian framework, and evaluates models with their global model likelihoods. The standard polynomial models are modified to be disjoint. Probabilistic information is translated into simple qualitative terms using credible intervals, the Bayesian analog to a classical confidence interval.

Curve-fitting is an inference about the relationship between variables. Those variables construct a mathematical space. Curves cutting through that space represent specific hypotheses about the relationship between the variables. Meanwhile, datapoints in that space represent observations. These serve as statistical evidence regarding the curves.

For sets of datapoints like Hubble’s (Figure 1.2), we intuitively assume that the closer a curve comes to the datapoints, the better the data support the curve. This intuitive move corresponds to the standard statistical assumption that experimental

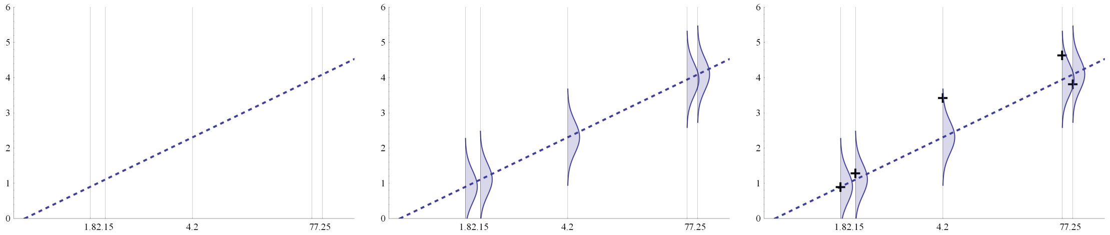


Figure 1.5: Likelihoods for a curve. All experimental apparatuses involve an element of error. Assuming that a curve is the truth, error associates a sequence of probability distributions on that curve, each of which determines how likely an observation is. The farther away an observation is from the supposed true curve, the lower the probability.

error is *normally distributed*, associated with the classic bell-curve-shaped distribution. A curve cutting through the space imparts a probability distribution for observations, assuming that the curve is true. That is to say, error determines a probability distribution for an observation E, assuming that a certain hypothesis h is true. This probability is called the *likelihood*, and it is written $P(E|h)$.[9]

What is the import of the likelihood? That depends on your theory of statistical evidence. This paper assumes the *Bayesian Theory* of evidential support, which holds that the epistemic status of a hypothesis should be measured with a probability. The probability is subjective, at least in the sense that the probability has a distinctly epistemic character; it is not considered to be a feature of the natural world, in the same way that some other probabilities are, such as probabilities associated with dice and roulette wheels. So the verdicts of Bayesian Theory are implicitly relative to an agent who is committed to them.

Bayesian Theory provides a natural meaning for the effect of the likelihood. The probability of the evidence assuming the hypothesis affects the probability of the hypothesis assuming the evidence, $P(H|E)$. The two are connected in "Bayes' Theorem."

[9] On the Bayesian approach, models also have likelihoods—there is no deep difference between models and specific hypotheses. I will tend to write capital "H" for a model, and the likelihood $P(E|H)$.

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \tag{1.2}$$

Bayes' Theorem

The theorem shows that if one wants to know the probability of a hypothesis assuming some evidence, $P(H|E)$, it suffices to have three other bits of information: along with the likelihood, the unconditional probabilities of the hypothesis ($P(H)$) and evidence ($P(E)$). Conceiving of a Bayesian agent as gathering evidence and using it to update the probability of the hypothesis, $P(H)$ is the probability that the hypothesis begins with, so it is called the *prior probability.* Correspondingly, $P(H|E)$ is called the *posterior probability.*

The prior probability term can be a problem for Bayesian Theory as a theory of statistical evidence. Generally, we want a theory of evidence to be objective. But on Bayesian Theory, the posterior depends on the prior, and without additional constraint, agent may assume the prior to take any value. Some Bayesians, strong subjectivists, hold that any prior distribution at all is legitimate. Others, objectivists, hold that there is a unique most rational, most impartial, etc., set of prior probabilities, determining unique posteriors given some evidence. Sociologically, many Bayesians compromise between strong subjectivism and objectivism, for a moderate subjectivism or "tempered personalism" (Earman, 1992).

This paper adopts an objective Bayesian stance. The general position is defended in Williamson (2010). The standard objective rule is MaxEnt, "maximum entropy," on which prior probabilities are supposed to take the least-informative, widest spread possible. But curve-fitting models are unbounded and do not admit of a global least-informative distribution. So the standard approach will not work for this essay.[10]

For a notion of confirmation, I look to the "Likelihood Ratio" measure, defended by Branden Fitelson (Fitelson, 2012).[11] That measure can be defined as a ratio of

[10] My approach will be similar to the standard approach, constructing a prior that nearly maximizes entropy, but on model *predictions*, not hypotheses. Furthermore, my approach restricts and structures the space of predicted events.

[11] This measure ensures that conclusive evidence, evidence that proves a hypothesis true, always

likelihoods: the likelihood of a model being measured, against the likelihood of its complement.

$$\frac{P(E|H)}{P(E|\neg H)} \tag{1.3}$$

Likelihood Ratio

I will assume equal model priors (Section 1.3), which simplifies the behavior of Likelihood Ratio. On this approach, the ordering of likelihoods will be similar to the ordering

However, given that the Principle of Simplicity here assumes equal prior probability for each model, using Likelihood Ratio gives predictable results: the confirmation ordering is the same as the likelihood ordering, and the model with the highest likelihood is the best confirmed.[12]

We can codify this information as its own principle. Call this the "Limited Law of Likelihood" (LLL).[13]

LLL On Likelihood Ratio, with equal priors, the model with the highest likelihood is the best confirmed.

LLL licenses us to evaluate model fitness by likelihood alone.

The use of equal priors also means that likelihoods are exactly proportional to posterior probability. So the likelihoods will quickly give us information about both of Carnap's notions of confirmation: "confirmation as firmness" and "confirmation as increase in firmness" (see Fitelson (2012)). Either way, with equal priors and the

favors a hypothesis or model more than inconclusive evidence does (Fitelson, 2012).

[12] To see this, imagine a set of hypotheses, $\{H_1, H_2, \ldots, H_n\}$, and their set of corresponding likelihoods, $\{L_1, L_2, \ldots, L_n\}$. Suppose further that the likelihoods are in descending order: L_1 is the highest, L_n is the lowest. Now consider the corresponding likelihood ratios, $\{c_1, c_2, \ldots, c_n\}$, "$c$" for "confirmation." Are the likelihood ratios in the same order as the likelihoods themselves? With equal priors, $c_i = \frac{L_i}{\frac{1}{n}\Sigma_j L_j, j \neq i}$. As we go down the sequence, we effectively "swap out" a larger likelihood in the numerator with a smaller likelihood in the denominator. Each successive likelihood ratio c_i is smaller, just like each successive likelihood L from the former series. So the ordering for confirmation here must match that for raw likelihood.

[13] The Law of Likelihood is the general, controversial principle that evidence favors one hypothesis over another if and only if the first has a higher likelihood. This principle does not hold generally across Bayesian measures of incremental confirmation (Fitelson, 2012), but it does when priors are equal.

Model	k	Expression
H_1	2	$y = a_1 x + a_0$
H_3	4	$y = a_3 x^3 + a_2 x^2 + a_1 x + a_0$
H_5	6	$y = a_5 x^5 + a_4 x^4 + \ldots + a_0$

Figure 1.6: Models for this chapter, with number of adjustable parameters and expression. The models are polynomials, sums of powers of x. The complexity of the model is measured by the number k of adjustable parameters a_i. This paper will assume point-exclusions on the models to create a disjoint structure, making a cleaner contest in a Bayesian framework. For example, H_3 is constrained such that $a_3, a_2 \neq 0$.

Likelihood Ratio measure, likelihoods will give us all the information we need about model confirmation.

So far, we have enough information to apply Bayesian Theory to individual curves. But the general hypotheses in the first section (H_1, H_3, H_5) do not correspond to individual curves. H_1 is the hypothesis that the relationship between the variables is linear—that is, some linear curve or other. This hypothesis corresponds to a whole model. In the context of this essay, a *model* is a family of curves. We might think of the function as assigning probabilities only to infinite bundles of curves.

The paradigm models for curve-fitting, especially in this abstract philosophical context, are the polynomials. The main notion of model complexity is a matter of the number of adjustable parameters, k. *Adjustable parameters* are terms that can take different values within a model. Making an assignment to parameters determines a curve. For example, taking LIN and assigning $a_1 \rightarrow 0.5$, $a_0 \rightarrow -1$ yields the curve $y = 0.5x - 1$.

Although I introduced H_1, the hypothesis that the relationship in question is linear, as a general hypothesis, I will start calling it and the others "models" from here on out—this will make it easier to navigate the technical information involved. Like individual curves, models can be evaluated using Bayes' Theorem. Write M for model, h for curve, and $P(M|E)$ for the model posterior. The substitution is almost trivial. The problem is that likelihoods are initially defined only for curves. To upgrade, we need to determine what I will call the *global model likelihood* (GML), the likelihood that evidence imparts on the model as a whole. This is not too difficult. The likelihood for the model is an aggregation of the likelihoods for all the curves in the model on the data. We have a weighted average.[14]

$$P(E|M) = \int P(E|h\&M) \times P(h|M)dh \tag{1.4}$$

Averaging must be done by integration ($\int f(x)dx$), since the hypothesis space is continuous.[15]

Allowing E to vary, the expression for the GML ($P(E|M)$) defines a distribution of likelihoods over possible observations. This distribution is called the model's *predictive distribution.* Equation 1.4 shows directly that the predictive distribution depends on the conditional probability $P(h|M)$. Call this the *conditional prior distribution* (CPD). It will be a multivariate probability distribution over parameters in the model.

$$P(h|M) = P(\langle a_0, a_1, a_2, \ldots \rangle |M) \tag{1.5}$$

The curves defined by the CPD translate a distribution over parameters into one over points in space (the space of the relevant variables). Combining these points with error probabilities results in the predictive distribution.

$$P(E|M) = P(\langle y_1, y_2, y_3, \ldots \rangle |M) \tag{1.6}$$

It will be best to treat the sample size, the number of points, as fixed, and predictions as applying to sets of data.

These probabilities can be unwieldy in philosophical analysis. In order to state the advantage of simplicity in basic, "cash-value" terms, we will want a way to translate from probability distributions to qualitative properties. The error distribution only counts different degrees of error as more or less probable, but it is natural to define

[14] Error is considered independent of the model in curve-fitting cases, and to be i.i.d. Thus the expression simplifies to $P(E|M) = \int P(E|h) \times P(h|M)dh$.

[15] We can restate the global model likelihood (Eq. 1.4) explicitly in terms of the parameters. The result is a multiple integral, one for each parameter. For LIN, it would be this.

$$P(E|M) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(E|\langle a_1, a_0 \rangle) \times P(\langle a_1, a_0 \rangle |M) da_1 da_0$$

an *error range* which draws a sharp cut-off between degrees of error that are considered "live" and those which are not.[16] Similarly, the predictive distribution counts different observations as more or less probable for a model, but it is natural to define the *prediction* of a model, which draws a sharp cut-off between the set of observations compatible with the model and the set that is not. Both cut-offs are achieved with a *credible interval*, which takes some majority of a probability distribution as a "credible level" to define the relevant interval.[17]

The predictive distribution is flexible enough that we can imagine it as making a predictions for single observations which combine, or as making predictions for sets of data all at once. For a basic curve-fitting case, if we suppress the independent variable x, a datum is a single-dimensional value in y; in the same vein, a set of n data will be n dimensional. Although we can think of it either way, for these purposes, we want to think of all the data together as a set.

Again, the global model likelihood depends on the conditional prior distribution across the hypothesis (curve) space within the model. Without such a prior distribution, this important likelihood term is undefined. As a result, Bayesian Theory, the only statistical theory of evidence to invoke priors, is the only theory that can evaluate models in this holistic way. Other theories of evidence cannot directly evaluate H_1 through H_5 as general hypotheses; they must reinterpret them as models for selecting a single curve. Intuitively, it is an advantage of Bayesian Theory that it can evaluate H_1 and others directly. But simultaneously, its dependence on prior probabilities is an intuitive cost. As we have seen, Rosenkrantz and Schwarz try to avoid this problem by setting the prior aside and just examining the impact of complexity on the likelihood, the GML. However, the analysis here suggests that such a move is incoherent. We will investigate the problem in Section 1.4. But now, we are ready to see a better-defined version of the Predictive Focus Account.

[16] The "error range" is similar to the "margin of error," but the margin is only half the credible interval. I want to express the whole range across which the observation is expected, to the degree of confidence specified.

[17] A credible interval is the highest density region with the stated amount of mass in the distribution.

1.3 The Predictive Focus Account

The Predictive Focus Account seeks to explain when and why a simple model has a confirmatory advantage. The Principle of Simplicity states that such an advantage exists. The Principle of Simplicity admits of different formulations; I will formulate it to fit well with this project.

Principle of Simplicity When a simple model and a complex model both predict the data, and all else is equal, the data supports the simple model more.

The previous section provides the resources to unpack this proposition. It states that when all models consider the data to fall within a region of significant probability (defined by a credible interval), and models are equal in other respects, the simplest model will have the highest posterior, the highest probability given the data. Let the clause "all else is equal" suggest that there should be a parity in the performance of the models in some sense. But this phrase does not need precisification to explain the account.

Let me invent a few terms. Let the *prediction range* of a model be the set of observations that fall within the credible interval (of whatever credible level) of a given model's predictive distribution. Remember that nothing important will depend on the choice of credible level, and I am working with 0.99. This means that according to the model, it is 0.99 probable that the observation will fall within that range. Let the *prediction span* of the model be the size of the prediction range. So the range is a set, and the span is a measure. Finally, *predictive focus* is just the inverse of the span—a large span means low focus, and *vice versa.*

Now the *Predictive Focus Account* (PFA) can be stated as a set of three claims.

Predictive Focus Account

1. A simple model tends to make a prediction (over possible data) with a narrow range.
2. On average, a prediction with a narrow range assigns high confidence over the possible data it contains.

3. Data supports a model through its prediction. All else being equal, if a model predicts the data with high confidence, the data supports the model to a high degree.

Combing these claims produces the Principle of Simplicity. When all models predict the data, the simplest is the best confirmed by that data.

Each point of the account can be restated more directly in terms of Bayesian mechanics. The translations are straightforward given the context of Section 1.2.

Statistical Translation of PFA

1. A simpler model tends to have a predictive distribution with a narrow credible interval.
2. Likelihoods across a credible interval are, on average, inversely proportional to the size of the interval. A predictive distribution with a relatively narrow interval places relatively high likelihoods across the events it contains.
3. When models have equal priors, a higher likelihood corresponds to a higher degree of confirmation for the model, as well as a higher posterior probability.

The first point is a substantive element of PFA. It is the only point that is not analytic. This chapter will vindicate that point only in developing an independently attractive conditional prior distribution for which this claim seems to hold. I intend to develop the connection between the prior and this point at a later date.

The second is also obvious. For whatever the credible level C, this is the amount of probability in the given model's prediction range. Then writing S as the size of the prediction range, the average likelihood across that interval is just $0.99/S$. Average confidence is inversely proportional to the size of the prediction range, just by definition. In a sense, this is the core idea of PFA, and it will be important in the next section. To keep track of it, I will call it the "Trivial Theorem of Predictive Focus" (TT).

TT The average likelihood across a model's prediction range is inversely proportional to its span (and directly proportional to its focus).

Equivalently: The average likelihood across a model's prediction range is directly proportional to its focus. Now we can connect Rosenkrantz's reasoning directly into the Bayesian framework. What is the significance of sample coverage, of the chance of fitting random data? If one assumes that the probabilities corresponding to chance are flat, then chance probability is equivalent to measuring the mere spread of an interval—for a flat distribution, the size of a region and the probability falling within that region amount to the same thing. So chance probability can, indirectly, serve as a measure on a prediction range.

The third point corresponds to LLL, presented above (p. 19). When prior probabilities are equal, higher degrees of confirmation correspond to higher posterior probabilities. I will use equal model priors, first because it satisfies my penchant for probabilistic indifference (where this is possible). But it also broadens our options for interpreting confirmation, or perhaps better, a model's *epistemic status.* If one wants to order models according to posterior probability, this will be the same as the confirmational ordering adopted here.[18]

Rosenkrantz and Schwarz noted that the likelihood function was wider for simple models; it gave them more volume.[19] We could think of this fact as being a consequence of the simple model having a narrower predictive distribution. Less of the model is "wasted space," so more of it is close to the data (Trotta, 2008). Relative to the hypothesis space of each model, this makes the likelihood function wider for the simple model, counting more curves as having significant likelihood on the data. PFA gives us a direct way to understand the success of simple models in the Rosenkrantz and Schwarz frameworks. But it also highlights a critical gap in all the reasoning that has come so far: the determination of the prediction ranges for the models in question. The next section shows that this is a critical omission.

[18] If this point sounds too obvious to be worth the time, note that there are multiple Bayesian notions of confirmation in the literature, and that they are not generally equivalent (Fitelson, 2012), nor will they generally represent the ordering of posteriors.

[19] I mean "volume" in a general sense to denote the size of a region, not necessarily a 3-dimensional one.

1.4 Improper Priors in PFA

Rosenkrantz and Schwarz were faced with a problem. The Predictive Focus Account (PFA) holds that the advantage of a simple model is in its prediction span. And the predictive distribution, $P(E|M)$, is determined by the conditional prior distribution, $P(h|M)$. So defining a conditional prior distribution is necessary to determine a solution for any problem. But they wanted to show that simplicity presented a *general* advantage in Bayesian Theory, that it would not be dependent on a particular prior distribution. So they faced a dilemma: either select a conditional prior distribution, and make the analysis all too easy to dismiss. Or set the prior aside and try to show that, in some relevant sense, the outcome is independent of the prior chosen.

When it came down to calculating the effect of simplicity for a curve-fitting case, Rosenkrantz and Schwarz both elected for the second option: ignoring the prior distribution and focusing just on the GML. Or, what is equivalent, they used an improper uniform prior.[20] A *improper uniform prior distribution* (I will write "IUD") is a probability-like function with constant density, equal to 1 everywhere. The function is not a true probability distribution since it does not integrate to one. But it can serve as one for many practical purposes—thus "probability-like". An IUD simulates a uniform (flat) distribution that is proper and spread wide. In a basic case, using the one gives nearly an identical result to using the other. Thus the IUD can allow one to use a uniform distribution without the difficult decision of choosing endpoints for the prior.

What is so hard about choosing endpoints? Imagine that a Bayesian alien, Sammy, has arrived at the earth. Sammy wants to determine the distribution of height for

[20] This exposition is somewhat uncharitable to Rosenkrantz and Schwarz. Both are, in a sense, discounting short-term calculations. Each is focused on the *limiting behavior* of the models in a Bayesian context, that is, the behavior each has as the number of observations n approaches infinity. Why should this be valid while the short-range calculations are not? Because as observations pile on, the likelihood distribution shrinks drastically, and all prior distributions approach uniform shape over the increasingly tiny window of significant likelihood. Now the Bayesian convergence result applies. Taking a simple, arbitrary prior (or the IUD for that matter) is in a sense an act of faith to this convergence. This act of faith has been criticized on philosophical grounds. It is not clear what the value of a Bayesian calculation when it is calibrated only to achieve the right result with an infinite number of observations. "Limits are never reached" (Forster, 1995). This is an important philosophical question, but not the central one for our purposes.

human adults. As a Bayesian, Sammy has to start with an initial distribution and update. Suppose further that he wants to start off with a broad prior—intuitively he wants to start out as less opinionated. A fairly attractive choice is a uniform prior: this allows Sammy to be indifferent on a range of heights. But Sammy is one of those really little aliens, just about a foot tall. He thinks that the range of 0 to 8 ft. should be safe, so he makes these the endpoints of his prior. Sammy takes a sample, with an average height is 6 feet, 95% falling within the span of 4.8 to 7.2 ft., and a familiar bell-curve shape.[21] Now Sammy updates. Multiplying the prior distribution by the likelihood distribution given by the sample and normalizing effectively returns the sample distribution back. Since Sammy's prior was broad and even, the posterior distribution is very close to the sample.

It turns out, Sammy was cutting things close. What if we had been taller? Sammy's prior means that his epistemic framework can only accommodate evidence from within a certain range. On the other hand, there is nothing wrong with a uniform distribution that is too wide. Sammy realizes that he can use a technical trick that will ensure he never makes a mistake regarding the endpoints of his priors. He can use an improper uniform distribution to ensure that his prior distribution will never cut into the significant region of the likelihood distribution. See Figure 1.7.

With the IUD, the posterior distribution is identical to the likelihood distribution, because the prior is equal to 1 everywhere. This prior is sometimes called "informationless," because it intuitively does not offer a bias on the data, or lets the data "speak for itself." Thus the IUD is a tempting choice for Rosenkrantz and Schwarz. After all, if using this distribution means that the likelihood distribution shines through the clearest, then using it should show the effect of simplicity in a Bayesian framework free of prejudice. If Bayesian calculations with the IUD favor a simple model, then it may be fair to say that Bayesian Theory favors simplicity in general. Philosophers are rightly suspicious of easy answers for prior probability distributions. Despite its appeal, the IUD is philosophically incoherent for the Predictive Focus Account. This incoherence

[21] Perhaps the distribution really should be bimodal or something else, but it will be easiest to assume a simple normal distribution.

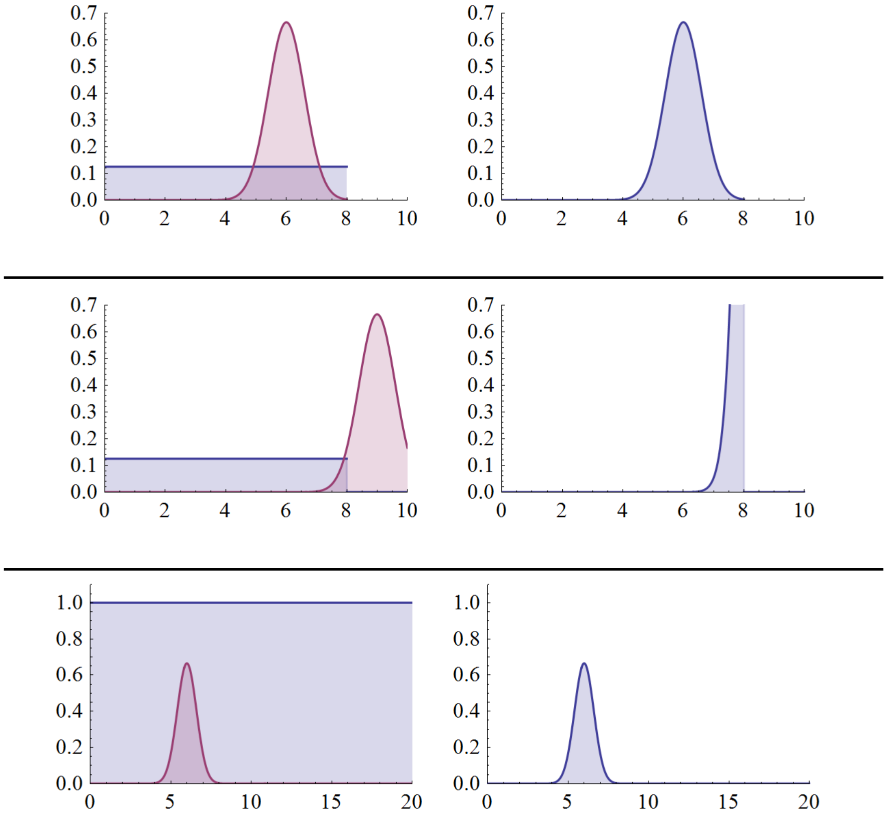


Figure 1.7: Sammy's initial Bayesian calculation, with a proper prior, cut things close (top row). But in a sense, he was lucky. If people had been taller, Sammy's prior would have severely cut off most of the likelihood distribution (middle row). Sammy decides that he could save himself some worry by switching over to the IUD (bottom row), which simulates the effect of the proper prior without making a cut-off.

has opened Rosenkrantz's work to technical criticism.

Recall the Trivial Theorem of PFA: average predictive confidence (likelihood) is inversely proportional to the prediction span. Simple models are supposed to have a narrower range, resulting in higher average GML, and probably resulting in a higher GML on the data. But with the IUD, our models do not admit of bounded prediction ranges—the spans are infinite. There can be no meaningful restriction of curves in the model to a range *via* credible interval. After all, the IUD is not a true probability distribution, so the predictive distribution will not be a true probability distribution either.[22]

The very thing that makes the IUD attractive—not having endpoints—means that the distribution over hypotheses within the model resists any description in terms of credible interval. With an IUD, there is no span over hypotheses. But hypotheses determine predictions. So if there is no span over hypotheses, there will not be a finite span over predicted events either—observations will just belong to a distribution with infinite mass. But the whole point of the Predictive Focus Account is that simpler models tend to have narrower prediction spans. These infinite spans are not meaningfully comparable, and the account no longer applies.

Mathematically, we can identify two important ways in which a comparison of curve-fitting models, like Rosenkrantz's, differs from Sammy's simpler inference about human heights. First, for Sammy, nothing hangs on the wasted mass in his prior distribution—the calculation just shaves it off. But in a sense, wasted prior is what PFA is all about (Trotta, 2008). PFA states that a simple model is focused, and its increased focus (particularly over the region with the data) gives it an advantage. We can make the same point in the negative: the disadvantage of a complex model is its lack of focus, its diffusion, and in particular, its greater expenditure of probability mass on regions irrelevant to the data. The disadvantage of complex model, even while accurate, is its high relative degree of wasted probability. So unlike Sammy, Rosenkrantz needs to take

[22] Note that the predictive distribution will contain infinite (conditional) probability mass. So while it will have a highest density region (used for defining a credible interval), it will not admit of such a (finite) region that constitutes, for example, 99% of the whole.

care with the total size of his priors.

The second difference is that Rosenkrantz pursues a comparison between models, while Sammy does not. And pointedly, Rosenkrantz compares models of different dimension. So Rosenkrantz is committed to the idea that the IUD will be impartial in this respect; it will not favor lower or higher-dimensional models as such. But this commitment is incorrect. The IUD makes a certain choice about the relative rates of probability assignment between parameters. So the IUD is not a recommended choice for this kind of comparison (Kass & Wasserman, 1996).

If the IUD is not philosophically compatible with PFA, then why does Rosenkrantz achieve a higher GML for the simplest model in his calculations? Why does it seem to work? To understand Rosenkrantz's results, we should find the proper prior distribution that most closely matches the IUD and the results—call this a "proper counterpart." This will allow us to "reverse-engineer" prediction ranges that best represent Rosenkrantz's calculations. Let a *proper counterpart* be a prior distribution with the following features.

Proper Counterpart to an IUD

1. It has the same density as the IUD over the region of accuracy.
2. It has the same shape as the improper version.
3. It sums to 1.

We can construct a proper counterpart prior for each model simply by setting independent uniform distributions on each parameter, ranging from 0 to 1.[23] For each model, $y = a_0 + a_1x + \ldots + a_nx^n$, each parameter is uniformly distributed between 0 and 1. This creates square regions of parameter space with unit-length sides. These resulting proper priors are designed to give the same answers as the original calculation.

[23] An IUD's proper counterpart is not unique. To make sure that the proper counterpart is accurate, we only have to make sure that each parameter range includes its best-fit value. With the density that we have to work with, this gives us some wiggle room with the parameter values. We might, for example, shift the region for some parameters slightly below 0, say [-0.2, 0.8]. This change would reduce the variance of predictions a little, making a small difference in the results.

	GML	Proper Counterpart Span
H_1	3.5×10^{-2}	2.5×10^{2}
H_3	1.9×10^{-6}	1.5×10^{7}
H_5	8.8×10^{-14}	2.3×10^{15}

Figure 1.8: The GMLs from Section 1 (Figure 1.3, p. 14), calculated with Rosenkrantz's formula. But these GMLs were derived implicitly with an improper uniform distribution (IUD). We can measure the prediction spans (or proportionately, standard deviations) of the predictive distributions resulting from using the "proper counterpart," and can see a striking inverse relationship. The proper counterpart reveals the IUD's implicit determination of model prediction spans.

However, these priors admit of quantifiable prediction spans (Figure 1.8). So we can see that the IUD was effectively smuggling in prediction ranges.[24] The point here is not that the proper counterpart is a bad prior. It is natural enough—I personally might place it on my second string for prior distributions. But the results from the use of an IUD are not particularly universal or objective. It is no better in this regard than an arbitrary proper prior. In fact, it might be a little worse, since its assumptions about the models are harder to tease out.

So the IUD does not free us from particular commitments about the prior distribution. In fact, the IUD does not free us from such commitments even within the narrow context of uniform distributions. Merely expanding the range over which the parameters vary for each model has a significant impact on the results. For example, expand the range of the parameters from $[0, 1]$ to $[-3, 3]$: the result disproportionately disfavors more complex models. The range of hypotheses is increased by a factor of 6 for each adjustable parameter. As a result, predictive confidence local to data will decrease by that amount, for each parameter. With two parameters, H_1's GML decreases by $6^2 = 36$. But H_3's decreases $6^4 \approx 1300$, H_5's by 6^6, over 46,000. This is a dramatic shift. So the IUD does not even represent all uniform distributions. It effectively gives us the results of one particular choice of prior.

[24] The IUD's proper counterpart will depend on the data. If the data had been substantially different, then the proper counterparts to the IUD would have been different, and the resulting spans would be different. So the IUD does not make a fixed assumption about model prediction spans, and this makes it a little harder to pin the assumption down.

The GML evaluation techniques of Rosenkrantz and Schwarz have been criticized as parametrization-dependent. By rewriting the models of H_1 through H_5, one can arbitrarily change the GMLs (Forster & Sober, 1994). The effect of reparametrization is the same as that of my changing proper ranges above. This dependence on parametrization here might seem both abstruse and purely technical. But the IUD is the problem. This probability-like function has infinite mass upon integration, in a sense, infinite "probability" at its disposal. Changing the description of the model, and then flatly re-applying the IUD, allows local segments of the hypothesis space to soak up more or less of that infinite mass. So the IUD's technical problems are related to its conceptual problems.

The point is even easier to see in the context of the chance-probability arguments. The chance probability of a model fit depends on the size of the field of possibilities. The next section will make this point clearer, but for now consider a quick example. Jason finds Jess interesting—he wonders whether she likes him. At a party, Jason sits to talk to his friends, and Jess takes the seat right next to him. Is she trying to sit near him? Is this good news for Jason? It is not clear. The key bit of missing information is the size of the party, and how many other seats were available. If this is a small get-together in a living room, Jess did not have many options, and her position probably means nothing. Her close positioning is easily explained by chance. On the other hand, if this is a big party spanning multiple rooms, and Jess sat down right next to him, Jason should get ready to make conversation! The probability that Jess is sitting next to Jason by chance, without interest, seems to be a function of the number of free seats.

Chance probability is determined by the range of possible outcomes. If we have a large range of possible outcomes that expands and approaches infinity, and chance-probability is uniformly distributed, then chance probability vanishes for each event in the range. A context that implies infinite ranges of outcomes is incompatible with chance probability reasoning. And the close connection between chance probability and predictive focus reasoning reinforces that infinite ranges do not fit the framework.

In sum, the IUD implies infinite ranges, both for chance outcomes and model predictions. It ultimately does not make sense to use an IUD for the purposes of PFA. It yields a technical vulnerability, parametrization dependence. But that vulnerability is not a stand-alone, purely-technical problem. It is connected to a deep incoherence—the attempt to showcase PFA, on which simple models have narrower prediction spans, with the IUD, on which a model does not have a prediction span (or sample coverage / chance-fit probability). The IUD actually implied prediction spans that were not explicitly accounted. The next section will go back to the drawing board, to derive a new approach to assigning a proper prior for a predictive focus calculation.

1.5 Predictive Focus

So far, I have avoided the key question: *What is the predictive focus for each model?* How much larger is the real, or proper, prediction range for H_5 than for H_1? By this point, the question may look misguided. The preceding section, which works to connect predictive focus with prior distributions, and the definition of a global model likelihood (Equation 1.4), both suggest that there is no such thing as pure predictive focus—only predictive focus given some prior distribution or other. But there may yet be a particularly good way to set up our priors for predictive focus calculations. Previously I focused on priors, and looked to the prediction ranges of the models that result. But perhaps I did things backward.

First, let us think directly about the models in themselves. Remember the three models under consideration (p. 10). The probative difference between these models is the number of adjustable parameters. Remember that the adjustable parameters are written a_i, and for a model, the number of adjustable parameters is k. The inherent flexibility of each model is determined by this number. In the model, the set of parameters defines a *hypothesis space*, and the key difference between each model is the dimension of this space, which is equal to k. H_1 has a 2-dimensional hypothesis space, corresponding to the well-known fact that it takes two points to define a straight line. H_5 has a 6-dimensional hypothesis space.

Intuitively, H_5's hypothesis space is larger. But to really be comparable in terms of proportions, spaces should have the same dimension. Hypothesis spaces of different dimension are awkward to compare. How would you compare a foot of length to a gallon of volume? Is the gallon larger? By how much? The best answer we can give is to translate length into volume terms: this would mean a three-dimensional region of $(1 \times 0 \times 0)$ feet, giving a volume of 0. Thus, all lengths are point-sized in a space of volume, and have no effective size in that context. By extension, all lower-dimensional spaces are point-sized, measuring zero, in a higher-dimensional space, according to the best way we have to compare them. To restate the point, higher-dimensional spaces are counted as infinitely larger than lower-dimensional ones. This looks like bad news for PFA, because we might conclude from this that a complex model's prediction range must also be infinitely larger as a result. PFA cannot run on such extremes.

But is H_5 infinitely broader than H_1 in the relevant sense? That does not sound right. In the Hubble data in Section 1, H_1 clearly had greater than 0 probability of fitting the data— it *did* fit the data after all. This leads to an important point. While the hypothesis spaces of simple models may be point-sized with respect to more complex ones, their prediction ranges are not!

The reason is error—error distributions "deepen" model predictions, expanding their dimension from k (the number of adjustable parameters) to n (the number of data). Mathematically, the dimension of the *sample space*, the space of possible observations, is the same for each model. That dimension is just the number of observations, n. It does not matter how many adjustable parameters a complex model may have. If an experimenter only makes one observation, the dimension, the "depth" of that prediction space is just the one. But the number of observations is not a feature of the models in themselves—it is a feature of experimental design.

Perhaps in order to understand the prediction ranges of different models, we have to consider an experimental situation. At least, we need one particular detail about the experiment, the sample size, in order to determine the sample space, a region of which is the prediction range for each model. In pursuit of a better understanding of these prediction ranges, let us pursue a simplistic curve-fitting example, with a physical curve.

Imagine Kate is a space explorer and photographer. She whizzes away to a remote

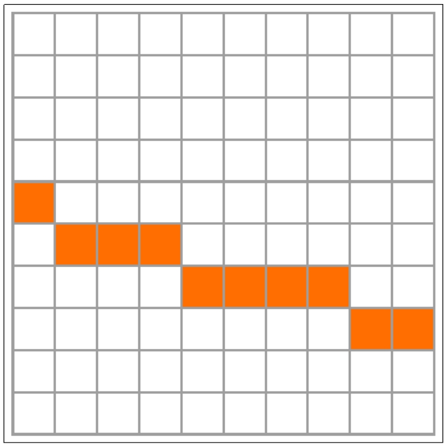


Figure 1.9: Kate's snapshot of the mysterious trail. This picture is consistent with a linear curve. The question is: does the data favor H_1? By how much?

region of space, the Van Fraas Sector, and spots a mysterious object, something like a comet. The object leaves a glowing trail as it flies by. She rushes to snap a picture, but her high-tech camera is malfunctioning. She has to rely on a back-up camera with cartoonishly bad resolution (10x10 pixels, 100 in total), which are bi-valued, either colored or blank. Seconds before it vanishes, she snaps an image of the ephemeral trail, Figure 1.9.

Kate's picture is consistent with a linear shape. She uses this image to test our three models ($H_1 - H_5$), and she assigns probability 1/3 to each. The question is: does the image favor the linear model H_1? By how much?

In this example, images are sets of dark pixels on the screen. The natural way to approach this problem is to assign equal likelihood to each image consistent with a particular model. To fit the curve-fitting paradigm, consistent images will be restricted to those that only have one lit pixel in each column (the camera has ten). The more images consistent with a model, the lower the likelihood of each image, with inverse proportionality.

Start with the total number of possible images. This is a basic problem in the math of combinatorics. The camera has ten columns and ten rows. With no other restrictions on images, this means that each column represents an independent choice of ten pixels

H_1 :	2.6×10^3
H_3 :	1.1×10^6
H_5 :	5.1×10^8

Figure 1.10: Approximate counts of images consistent with each model

to be dark, for a total count of 10^{10}, ten billion possible images. Most of these are disconnected and not compatible with any of the three hypotheses.

Figure 1.10 gives approximate counts of images. How many of these images are consistent with H_1? About 2,600, a small fraction of the former number. What are the odds of one of these images occurring by chance, randomly out of the total set? About one in four million! Here the force of the chance-probability argument is clear. Against H_1, we can dismiss the chance hypothesis without a second thought.

Figure 1.10 provides approximate counts of consistent images for each model. The ratios between these counts are dramatic. Assigning predictive probabilities indifferently means that Kate's image (Figure 1.9) has likelihood 1/2,600 on H_1, and much lower likelihoods on the other two models. And given Kate's indifferent prior probabilities, she can safely infer that the path is linear, even on this small amount of evidence.

In this example, the experiment determined the number of possibilities. With some minor assumptions, the prediction range for each model—the total number of possible images consistent with it—is well-defined.[25] Applying probabilities indifferently gives the strongest expression to TT (p. 24)—the likelihoods are exactly inversely proportional to the prediction spans.

Kate's camera gives H_1, H_3, and H_5 what they lacked on their own—natural prediction ranges. If the predictive focus reasoning in this example is the right kind, then predictive focus is always relative to experimental design. Only with the articulation of an experiment can we say how many observations are possible, or what the range of possible observations is, and what the respective ranges are for each model.

[25] Here, I implicitly use a confidence level of 100%, which is practical given that there are sharp bounds on each set of compatible images.

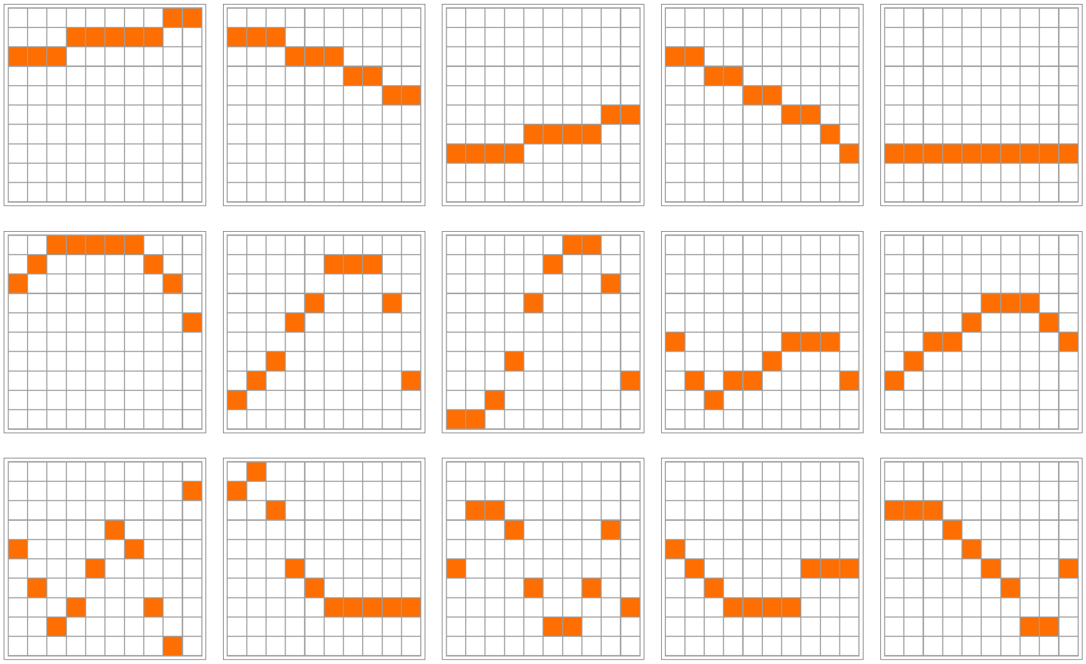


Figure 1.11: Rows of a few possible images for H_1, H_3, and H_5. The more complex models are more flexible and allow for more possible images.

It is tempting to hope that there would be some emergent, natural proportion between $H_1 - H_5$, such that the articulation of the experiment would be an unnecessary detail. We might hope that a camera with a higher resolution might keep the same proportion of images compatible with each. Or perhaps something less—a proportion that changes but that approaches a limit as resolution increases. Unfortunately, this is not the case. For this example, as camera resolution increases, the proportion of images compatible with H_1 to those compatible with H_5 shrinks without limit. It must do this, as the H_1 model is point-sized compared to H_5, as per the considerations given earlier in this section. But an experiment puts "flesh on the bones," and provides a context where the models admit of non-extreme comparisons.

Respecting the impact of experimental design in a Bayesian context means designing a prior probability distribution that is sensitive to the experiment. The next section will outline how to generalize the reasoning from the case of Kate's camera in Bayesian Theory.

1.6 The Data Window Prior

The example of Kate and her digital camera presented an inference problem where counts of possible images were well-defined and relatively accessible, and in these respects the case might seem completely foreign to real science. We do not often have our data and possibilities wrapped up in such a nice, neat package. But real science may not be so distant from this imaginary case as it might first appear.

First of all, most, if not all, experimental apparatuses have limits. Telescopes have limited ranges. Microscopes have maximum magnifications. Thermometers have a limited range of temperatures, barometers a range of pressures. Basic scientific technologies certainly have limits on their ranges of possible observations, and it is plausible that most or all sophisticated technologies have limits as well. Call the range of possible observations the *data window*.

Second, all experimental apparatuses have some liability toward error. There is always some uncertainty about the true value of an observed quantity, especially due to nuisance variables and round-off error, but also due to stochastic elements in calibration, imperfections in the instruments, etc.

These two features represent features in Kate's camera case. An experiment's data window corresponds to the edges of Kate's screen, beyond which the camera would not record information. Error corresponds to the pixelation of the screen, which infuses uncertainty into the observations that are recorded.

Ultimately, I use the term "data window" in two senses. In the weaker sense, the data window just is the range of possible observations. In the stronger sense, the data window is this range, enriched with the parametrization that corresponds to experimental error. In particular, this is the parametrization (if it exists) on which error is *i.i.d.*—independent and identically distributed. This is the parametrization on which uncertainty is evenly spread across the data window, just as it is intuitively evenly spread across a camera screen through even pixelation. Call this parametrization the experiment's *native parametrization.*

The data window (strong sense) has the resources to ground a fully objective prior probability distribution. This distribution, the *data window prior*, is *the distribution that spreads model predictions as widely as possible across the experimental data window, on the experiment's native parametrization.* That is, it is the prior probability distribution on a model that yields a predictive distribution across possible observations that comes as close as it can to a uniform distribution within the data window, and wastes no predictive confidence outside that window.

1.7 Calculations with the Data Window Prior

In this section, I will describe two methods for computing the data window prior. The first is more complicated, and also more idealized, as it depends on heavy computational resources (those beyond my current capabilities), but is useful conceptually. The second is a quick-and-dirty method for approximating results of that calculation.

1.7.1 Method 1

This method runs mainly by manipulating a large random sample, which represents a probability distribution. This random sample goes through different stages.

With the data window information in hand, including the sample size n, one begins by simulating an *ideal chance distribution* over possible events. This is an n-dimensional distribution, which is just uniform for each datapoint. The simulation is a random sample of datasets. Denote this sample Γ_0. Γ_0 is naturally the same for all models.

The data window prior is stipulated on a particular predictive distribution, so in an unusual turn for Bayesian analysis, the predictive distribution is determined *before* the prior. The simulated predictive distribution is just a subset of Γ_0—whichever subset the model fits.[26] This is easy to determine. We may take a criterion of fit for the model (for example, the expected squared error for such a dataset) and see whether the model's best fit on that set falls within the criterion. If so, the calculating agent preserves that dataset; otherwise she throws it out. The remaining subset of possible datasets represents the model's intended predictive distribution. Let this sample be deliberately denoted Γ_2.

Notice that the ratio of numbers of datasets, Γ_2/Γ_0, *just is* the chance probability for the model on the experiment. This is the mysterious quantity suggested by the early chance-probability arguments of Weyl and Rosenkrantz. This fact highlights the close relationship between chance-probability reasoning and predictive-focus reasoning. Finally, the actual data window prior, represented by a new set Γ_1, is the distribution of parameter values for the model that produces predictive distribution represented by Γ_2.

Although we can articulate a method for determining the data window prior, computation is a critical issue. Numerically integrating model likelihoods in curve-fitting problems is hard. At least here, I will need to use a drastic shortcut.

[26] This definition introduces a dependence on a criterion of fit, a credible level. Since the error distribution will usually not be uniform, it will differ in shape from the ideal chance distribution in the window, and the criterion of fit will make a difference to the results. However, as long as error is narrow compared to the data window itself, the proportions of model fits will be relatively constant. In other words, the dependence on the criterion of fit will not make a difference in relative model evaluation.

1.7.2 Method 2

Rosenkrantz's approximation to Global Model Likelihood suggests that prediction span and maximum likelihood are effectively independent parameters in the determination of GML. In particular, GML is proportional both to ML (the model's "maximum likelihood" value) and the spread of the likelihood function for the model. But in relative terms, the spread model's likelihood function will just be inversely proportional to its prediction span—the more space the model's prediction covers, the less of span will be taken up by the likelihood function on the actual data. These considerations suggest a rough approximation to the GML with the data window prior: model ML divided by the spread of the model's predictive distribution.

$$P(E|M) \approx P(E|\hat{h})/\sigma_M \tag{1.7}$$

It is fairly easy to generate a random sample of curves intended by the DWP—we may, for example, use Method 1 until we get to this point. Combining these with random experimental errors produces a sample predictive distribution. The generalized standard deviation of this distribution tells us the spread of the distribution, and effectively the prediction span (which again is just a credible interval). But this spread is inversely proportional to predictive density, and given the construction of the DWP, this distribution is as even as possible. So the model's confidence is approximately uniform across its predictions, and span is a direct guide to confidence.

1.8 A Predictive Focus Calculation on Hubble's Data

Let us return to the initial question of this paper. Do the data in Hubble & Humason (1931) confirm the linear relationship better than more complex polynomial relationships? Is H_1 better supported than H_3 or H_5? To the best of my ability, I will perform a Bayesian Calculation based on the Data Window Prior, in particular estimating GML with Method 2.

The most challenging consideration for the evaluation of Hubble and Humason's experimental apparatus is its profile for experimental error. Their observatory, Mount

$$\begin{pmatrix} \text{Model} & \text{ML} & \text{St. Dev.} & \text{GML} \\ H_1 & 1.8 \times 10^{-5} & 5.9 \times 10^{1} & 3. \times 10^{-7} \\ H_3 & 2.1 \times 10^{-4} & 1.2 \times 10^{4} & 1.8 \times 10^{-8} \\ H_5 & 6.4 \times 10^{-4} & 2. \times 10^{6} & 3.2 \times 10^{-10} \end{pmatrix}$$

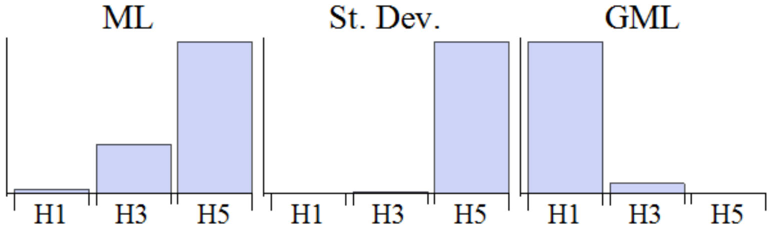


Figure 1.12: Results of my predictive focus calculation. As before, H_1 has the lowest accuracy of any of the models—in terms of maximum likelihood (ML). But H_1 also has the narrowest prediction span, proportional to the standard deviation of its predictive distribution (St. Dev.). These two factors combine to provide a rough approximation of global model likelihood (GML). Accuracy increases with the model complexity, but more gently. By contrast, prediction span increases in a roughly exponential fashion with complexity. As a result, H_1 has the highest GML.

Wilson, had multiple telescopes and spectrographs by the publication of the 1931 paper, and it is not clear how much uncertainty would have existed in their spectrographic readings. But rather than chasing the ideal, sophisticated DWP, we can make a first pass at the calculation by making use of velocity samples from the same nebulae. This information gives the first idea about the error intrinsic in the data.

Similarly, our best information about the data window for the experiment seems to come from the tone of the data presented. It provides a variety of information about nearby nebulae, but seems to strain with the more distant objects. It is plausible that Hubble and Humason were recording objects as far and as fast as they could at the time, supporting a minimalistic approach to the problem.

I determine that the data window for the Hubble–Humason experiment, for this basic treatment, to be [-20, 20] Mm/s (mega-meters per second). This would yield a prediction span of 38 Mm/s for a single point (depending on the credible level), and on the order of 38^{10} for the 10-point set. I determine experimental error to be normally distributed with $\sigma = 0.8$ Mm/s.See the appendix for more information.

The linear model H_1 takes away a satisfying 94% of the total posterior probability between the three (Figure 1.12). While model accuracy increases with complexity, prediction span increases much more dramatically. Prediction span seems to increase exponentially—the more complex model is compatible with dramatically more possible datasets within the data window. H_1's high probability on the data meshes with the sentiment that this important dataset provided "definitive proof of a linear velocity-distance relation" (Sandage, 1994, 34).

In turn, the chance probability of being confirmed dramatically increases with the more complex model. It is indeed much less likely for H_1 to fit random data on these settings, to a degree that we can now quantify. The span of the ideal chance distribution is 4.2×10^{10}. Against this large range of possibilities, even H_5's 2.0×10^6 is quite restrictive. But H_1's is much narrower still. The chance probability of H_1 being confirmed, the probability of its fitting random data, is tiny.

$$\frac{5.9 \times 10^1}{4.1 \times 10^{10}} = 1.5 \times 10^{-9} \tag{1.8}$$

1.9 Comments on the Data Window Prior

Is the Data Window Prior legitimate in the context of Bayesian Theory? The answer will, of course, depend on a complete elaboration of the method, rather than the introductory proposal here, but I want to make just a few comments on this question. Of course, for all experiments that have finite data windows in the relevant sense (and many will), this method will produce a proper probability distribution that can be taken as a prior and updated by likelihoods, satisfying the fundamental requirement of Bayesian Theory. Also notice that this conditional prior distribution does not determine how much probability mass is assigned to each model, but only how that probability is distributed. In this sense, the Data Window Prior is nonprejudicial on the key question of how probable each model is before evaluation. In Chapter 2, I emphasize the general point that the Predictive Focus Account does not depend on assigning more probability mass to simpler models.

But the Data Window Prior is strange in some respects. The most obvious is that this prior is not really *a priori.* It depends on other information. An agent must wait to assign the distribution until after analyzing the experimental equipment in question. A major possible worry is that the Data Window Prior represents a deep misunderstanding of the relation between an experiment and the world. By nature, an experimental outcome is supposed to be determined by an important feature of the world—not *vice versa.* Good experiments do not influence the studied object—or do so as little as possible. Light flowing in from a star channels through a telescope to tell an experimenter where the star is; the telescope does not shoot the light outward to fix the star's position. But in using the experiment to determine prior probabilities on hypotheses, does the Data Window Prior get the basic mechanics of experimentation backward? Does it make the perverse implication that the experiment fixes the world?

True, the Data Window Prior does use general features of the experiment to fix assumptions about the world. But with almost any decent apparatus, these assumptions will be wide-spread. And wide-spread (or "low-information") priors are just the sort that we want. If nothing else, the experimental data window is a merely available, if sometimes complex, basis for accomplishing this objective.

Moreover, we might see the Data Window Prior to represent a natural, philosophical view on which hypotheses are worthy of scientific attention. There is a significant literature in the early philosophy of science on the "demarcation problem," the question over which propositions are worthy of our attention. Most answers were in terms of empirical confirmability.[27] The Data Window Prior represents this general approach to scientific investigation directly in each model's conditional prior distribution. Here, the maxim is: *allocate positive probability to all and only hypotheses that are worthy of scientific investigation, and these are just the hypotheses that are confirmable on the experimental apparatus.*

[27] For just a few, see Russell (2009); Popper (1992); Hempel (1998). Many philosophers were engaged in the more ambitious project of determining which propositions are *meaningful*, and they took an empiricist, scientific approach to that question. This often forced them to speak of confirmability (and verifiability and falsifiability) in principle, divorcing it from the details of current experimental technology and other contingencies of actual experimentation. But for the much less ambitious project of determining just which hypotheses are worthy of scientific attention, which I invoke here, there should be no need to abstract away from the limits of current, particular experimental technology.

Here is an important test question for this approach: Do we want to say that some hypotheses are not scientific now (because they are not testable with current technology), but may become scientific later (because they become testable on future, better technology)? If the answer is "yes," then it is legitimate to demarcate whether a hypothesis is scientific with whether it is confirmable in practice. As the Data Window Prior gives positive probability (really, probability density) to all and only hypotheses within a model that are confirmable on a particular experiment, it represents a clear, attractive approach to scientific demarcation.

To apply the thought to curve-fitting models, we may note that every curve in the linear curve-fitting model is natural, and in that sense, scientifically respectable. But given the context of Hubble's experiment, in another sense, many of these curves do not represent worthy hypotheses, because they are well outside the bounds of the experiment to confirm. If science is committed to only working with empirically confirmable hypotheses, then scientific investigations should exclude some hypotheses: those that may be perfectly natural but experimentally inaccessible. On this rationale, allowing the experimental apparatus to limit the scope of considered hypotheses does not represent a misunderstanding of the connection between world and experiment, but rather a proper philosophical understanding of the limits of scientific investigation. If this is right, then at least on a basic conceptual level, the Data Window Prior should be quite attractive.

Whether the Data Window Prior represents a technically coherent application of Bayesian Theory is another matter. The prior is conceived in the context of a single experiment. But new experiments will yield different data windows. Extending the Data Window Prior to the investigation of the same models across multiple experiments will not be a trivial task. We can anticipate that the best extension of this method will require a fair amount of *retroactive prior revision*: the awkward act where a Bayesian sees that her probability model has some critical problem, replaces her prior probability, and recalculates posteriors based on the new prior. A Bayesian scientific community will tend to collaborate over a question, combining the findings from multiple experiments. Initially there is only one experiment, one data window, and the application of the prior is clear enough. But the addition of the second experiment complicates the method. We need some way to extend the Data Window Prior across experiments.[28] This extension will almost certainly require retroactive prior revision, going back and modifying the prior in some way.

Effectively requiring some retroactive prior revision is a conceptual cost. After all, standard Bayesian Theory, which is conceptually attractive partially because of its fidelity to an initial assignment of probabilities. Famous “diachronic Dutch Book” arguments are supposed to show that one is vulnerable to an irrational act (a combination of bets that will result in a sure loss) if one fails to update using Bayes’ Rule, faithfully, from single prior distribution.

However, objective Bayesians are generally prepared to count a prior-fixing method to be more important than strict adherence to conditionalization across time (Rosenkrantz, 1977; Seidenfeld, 1979; Williamson, 2010). We might construe the subjective Bayesian as asking the question, “how does the evidence modify my initial assessment of the hypotheses?” In a way, the subjective Bayesian always remains true to her initial assessments, although heavily modifying them through evidential likelihoods. But the objective Bayesian is doing something a little different. He agrees that credences should be represented by probabilities, that probabilities are the right way to assess hypotheses, but avoids starting from subjective assessments. The objective Bayesian asks the question, “what is the most rational (especially least-informative) way to assign probabilities to hypotheses, taking the data and all relevant constraints into account?” On such a principle, it should be acceptable to allow a prior distribution to change along with the range of hypotheses that can be confirmed. And on a sunny note, as experimental instruments become more powerful, the Data Window Prior will grow less informative. Our Bayesian mechanics get better along with our instruments!

Thus, whatever retroactive prior revision is necessary for extending the method of the Data Window Prior, it will be at home in the objective Bayesian framework.

[28] I speculate that the right way to extend this method to multiple experiments is constructing a mixture of the Data Window Priors from across all experiments involved, but there are other options as well.

More work may be needed to describe this style of Bayesian Theory in full. And I must admit that a fair degree of retroactive prior revision is a significant theoretical cost for the approach. But the benefits will be clear: a public, invariant, closed-form method for assigning proper prior probabilities to unbounded models, grounding predictive focus calculations, taking a philosophically satisfying rationale for evaluating the science-worthiness of hypotheses, and reflecting the august tradition of chance-probability reasoning.

1.10 Conclusion

The Predictive Focus Account explains why simple general hypothesis has a confirmatory advantage—it has a narrower, highly focused prediction range. If an observation is contained in the narrow prediction range of a simple general hypothesis and the broader range of a complex hypothesis, it lends more support to the simpler. This line of reasoning relates closely to the classic chance-probability argument: a simple general hypothesis is much less likely to fit an observation by chance than a complex general hypothesis, so it is a bigger deal if the simpler one fits.

Bayesian thinkers have previously suggested this account, and determined powerful calculation techniques to go with it. But previous proposals attempt to avoid the question of prior probabilities. In my analysis, this was a mistake. It does not make sense to try to divorce Bayesian Theory from priors, particularly for whole models. Model likelihoods are not defined in that context, and the use of an improper prior obscures the issue by smuggling in effective prediction ranges. Bayesians who hold the view must "roll up their sleeves" and commit to a prior distribution.

The Data Window Prior is designed specifically for this purpose, to provide an objective prior distribution which best expresses the intuitions of the Predictive Focus Account. It begins not with the hypothesis space, but ranges of possible observations on an experiment. The prior is "reverse-engineered" to achieve an even spread of its predictive distribution across the possible observations that the model fits.

The Data Window Prior is a proper prior, a genuine probability distribution. And it is invariant, because the parametrization of the problem is not up to the agent—it is determined by features of the experiment. This method is not a standard implementation of Bayesian Theory, but given its theoretical context, it does not seem to incur any critical problems.

Conceptually, the Data Window Prior is connected to confirmability. It is designed to account all and only curves for each model whose predictions could be observed by the experimental apparatus given. This produces a subtle modification to, or perhaps teases out a previously unstated presupposition of, the chance-probability argument. A simple model's virtue is that it is quite unlikely to fit a random-chance observation *of a particular experiment*; if it does fit, its degree of confirmation is proportional to how unlikely the fit. The philosophical framework proposed in this essay holds that chance observation is only defined in the context of the experiment. If this connection is trivial, if chance observation and experimental context are inseparable, so much the better for the Data Window Prior.

In the case of Hubble's experiment, my basic analysis yields a dramatic advantage for the simple linear model, a likelihood many of times greater than its competitors. Paired with the Data Window Prior, the Predictive Focus Account is on better footing, prepared not only to answer critical philosophical challenges (Forster & Sober, 1994), but to give new articulation to the Principle of Simplicity.

Chapter 2

Can the Bayesian Have It Both Ways?

Comparing Bayesian "Prior-Stacking" to Predictive Focus

2.1 Introduction

In Chapter 1, I presented an account of the Principle of Simplicity, the "Predictive Focus Account." This account explains why simplicity is a theoretical virtue in a Bayesian context of evidential support. On the account, simple models have a basic property that sets them apart from complex competitors in the relevant kind of contest. Simple models have high predictive focus, which gives them high predictive confidence (likelihood) across their prediction ranges. But the advantage of a simple model is not constant; it depends on the conditional prior distribution of hypotheses within the model. Chapter 1 outlines an objective method for specifying conditional prior distributions, to ground global model likelihoods, at least for individual experiments. These likelihoods favor simple, accurate models, because these models have high predictive focus.

In this chapter, we will look at a different take on the advantage of simplicity, but remain "in house" with Bayesian Theory. Contrary to the Predictive Focus Account, it has long been a part of "Bayesian lore" that the advantage of simplicity is to be located in model priors (Earman, 1992). Simple models are thought to deserve higher prior probabilities—perhaps in light of their intuitive plausibility, or because assigning them higher priors accords with standard scientific practice, or perhaps for more sophisticated reasons (to follow).[1] Call the practice of assigning higher priors to simple models *prior stacking.* Early proposals to this effect include Jeffreys (1988) and Salmon (1988),

[1] Model priors are different from the conditional prior distribution. In a sense, model priors are the highest level of prior distribution. The *model prior* is a scalar, the total amount of probability mass that the model is assigned. The model's conditional prior distribution divies up that amount of probability across hypotheses within the model.

which tout the ability of Bayesian Theory to conform the Principle of Simplicity with the prior-stacking mechanic.

Comparing these accounts will gives us an excellent opportunity to probe one of the fundamental philosophical tensions on the subject of the Principle of Simplicity. There seem (to me) to be three popular ideas about the Principle that conflict. One idea is that we should exercise as much indifference as we can in our initial, *a priori* evaluations of models and hypotheses. While the Principle of Indifference, a strong, normative version of this idea, is often considered to be in bad shape (Seidenfeld, 1979; van Fraassen, 1990), philosophers generally have a strong preference for indifference. It seems irrational to be very confident in a particular theory before the evaluation of any evidence at all.

The second idea, the Principle of Simplicity, states that in certain contests between simple and complex models, the simpler model is better supported by the evidence. The relevant kind contest is one where both models can accommodate the evidence with similar degrees of accuracy—the evidence is "equivocal" in a certain sense. It seems crazy to accept a model that is "unnecessarily complex," and often, even irrational to fail to favor the simpler one in such a situation.

The third is that these first two ideas contradict each other. If the data is "equivocal," then it sounds as though it should not favor one model over another. In which case, any probability that one model has over another after updating on the evidence, any *advantage* that it might have, it must have had prior to the evidence.

Again, here are the three ideas.

1. **Weak Indifference**

   The world is not more probable to be simple than complex at the outset, prior to the accumulation of any evidence.

2. **The Principle of Simplicity**

   When a body of evidence E supports both a simpler model H_1 and a more complex model H_2, *other things being equal*, E favors H_1 over H_2.

3. **Inheritance**

   If all models in a set predict the data, then whichever model has the most posterior probability must inherit that advantage from the priors.

We know from the last chapter that the Predictive Focus Account rejects the Inheritance principle above. If models have different prediction ranges, then the fact that they can accommodate the data equally well, the fact that they "equally" contain the data in their respective predictions, still leaves open what the model likelihoods are. But we have seen that it takes some work to ground the Predictive Focus Account, particularly in assigning conditional prior distributions, and it is easy to see how one would be tempted by Inheritance. So many philosophers have a different response to the tension above: balance the force of Weak Indifference and the Principle of Simplicity. This strategy compels one to assign higher prior probabilities to simple models, but conservatively, so the prior bias toward simplicity is modest.

But many others have responded differently. They reject the Principle of Simplicity as such, but look for some other good-making feature of simple models. These philosophers have understood simplicity as a "super-empirical" virtue—a virtue which is not necessarily reflected in how well the empirical evidence confirms or supports simple vs complex hypotheses. For instance, some have argued that simpler theories are "more informative" (Sober, 1975), while others have maintained that simpler theories are more falsifiable (Popper, 1992). Still others have claimed that simpler theories are better predictors (Forster & Sober, 1994); or more efficient (in the long-run) for the purposes of truth-seeking (Kelly, 2007). In effect, this move reinterprets the intuitive appeal that simple models have in the relevant contexts, and enables one to fully accept the other two principles (Weak Indifference, Inheritance).

This chapter examines the Prior-Stacking Account as an alternative to the Predictive Focus Account. The Prior-Stacking Account has several severe limitations. For one, prior-stacking is in tension with Weak Indifference; by its nature, prior-stacking is philosophically awkward. For another, it is actually difficult to make prior-stacking accord with standard Bayesian curve-fitting. Prior-stacking is committed to the idea that all models in a relevant comparison will have similar likelihoods, but we have seen that this is not true. Applying prior-stacking to curve-fitting requires some major methodological fudging. For a third limitation, prior-stacking produces weak results. It fails to reflect the strong confidence we have in simple models in certain contexts. Finally (and related to the second point), prior-stacking is insensitive to important features of the data. Intuitively, the more data that are compatible with a simple model, the better supported the simple model is, even over competitors that can accommodate that data. But prior-stacking gives simplicity a "flat rate" advantage, which is insensitive to the accumulation of additional evidence.

2.2 The "Prior-Stacking" Account

The Principle of Simplicity is tricky. In expressing that simplicity is a theoretical virtue, that simplicity should add something to an agent's assessment of evidential support, it is best expressed in the context where both the simple and complex models are equivalent in their relationship to the data. And the basic statistical notion of a model's performance on the data is *fit*, or likelihood.

The notion of fit, defined as likelihood $P(E|H)$, produces a possible precisification of the Principle of Simplicity. It contains the phrase "other things being equal," and fit is a natural candidate to substitute in.

PS$_1$ When a simpler model H_1 and a more complex model H_2 fit a body of evidence E roughly equally well, (i.e. H_1 and H_2 similar likelihoods on E), then E favors H_1 over H_2.

This principle pigeonholes the Bayesian into an implementation of The Principle of Simplicity. Combining PS$_1$ with Bayes' Theorem implies that simpler hypotheses must have higher prior probabilities. The following is an automatic consequence of Bayes' Theorem (implicitly assuming that H_1 is the simpler hypothesis and H_2 the more complex).

$$P(H_1 \mid E) > P(H_2 \mid E)$$

iff

$$\frac{P(H_1)}{P(E)} \times P(E \mid H_1) > \frac{P(H_2)}{P(E)} \times P(E \mid H_2)$$

PS_1 applies when the likelihoods are equal, and the probability of the evidence is necessarily equal to itself. The previous inequality can only be satisfied if the simpler has a greater prior probability.

By simple algebra, if $P(E|H_1) = P(E|H_2)$, then the above inequality entails that $P(H_1) > P(H_2)$. In other words, PS_1 entails that the only way simplicity can have confirmational value is if simpler hypotheses have greater prior probabilities than complex hypotheses. This naïve analysis suggests that "prior-stacking" is the only way to vindicate the Principle of Simplicity in a Bayesian framework (Good, 1968; Sober, 1994b).

As we saw in Chapter 1, this is not true. The Principle of Simplicity is tricky, and the trick is that there are multiple ways to interpret equivalence on the data, as well as evidential support. The Predictive Focus Account finds a different, weaker notion of model equivalence. Models may be equivalent on the data just in virtue of both predicting it.[2] Here, the Principle of Simplicity is helpful especially when the agent has difficulty assessing fit. After all, these calculations are involved; I had to resort to gross approximation in the preceding chapter. But it is clear why the Prior-Stacking Account is a tempting idea.

The best early statement of the *Prior-Stacking Account* (PSA) is in Harold Jeffreys' *A Theory of Probability.* Jeffreys subscribed to a version of the Principle of Simplicity; which he called "the simplicity postulate."

[2] There are finer ways to construe model equivalence compatible with the Predictive Focus Account. In particular, we could consider models equivalent if they have similar maximum likelihood values (MLs). This criterion is distinct from that of being predicted, but the two are related. But I think it would be of limited value to develop that idea—for the Predictive Focus Account, prediction range is more straightforward.

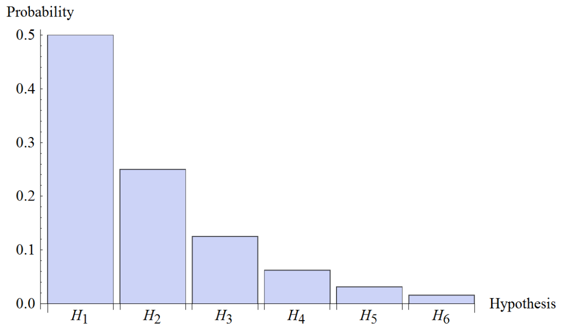


Figure 2.1: A "stacked" prior distribution on hypotheses, ordered by complexity.

The Simplicity Postulate Simpler hypotheses are to be assigned higher prior probabilities than more complex ones.

Jeffreys seems to have two motivations. One is that employing the simplicity postulate accords with intuition and with standard scientific practice. Jeffreys noted that scientists consider the simplest hypotheses first, and only move to higher orders of complexity if the simplest are predictively inadequate. That is, simple hypotheses have a standing advantage until they fail to fit the evidence, and a more complex hypothesis does a better job. Here is how Jeffreys implements his traditional prior-stacking approach. Take an arbitrary set S of hypotheses $H_1, H_2, H_3, \ldots$, which are countably infinite and ordered in complexity—H_1 is the simplest. Then we could assign these hypotheses probabilities in the following way.

$$P(H_1) = 2^{-1} = \frac{1}{2}$$
$$P(H_2) = 2^{-2} = \frac{1}{4}$$
$$P(H_3) = 2^{-3} = \frac{1}{8}$$
$$\ldots$$

See Figure 2.1.

Now, even before taking empirical evidence into account, the simplest hypothesis has the best epistemic standing, the highest probability. Now consider a bit of evidence E on which each of the hypotheses has a similar likelihood: $P(E|H_1) \approx P(E|H_2) \approx P(E|H_3) \approx \ldots$. A quick calculation will show that the posterior probability of each hypothesis will be approximately the same as the prior. Thus given the notion of favoring adopted for this paper, the evidence favors the simpler hypotheses over the more complex.

So prior-stacking accords with intuition and scientific practice—this is one of Jeffreys' motivations. The other is that some degree of prior-stacking is an unavoidable feature for certain sets of subjective probabilities. When a Bayesian agent wants to consider a countably infinite set of hypotheses, $H_1, H_2, \ldots$, she has the problem that she cannot (on pain of violating countable additivity) assign equal probabilities to all hypotheses in the set. When trying to divide the total amount of probability (1) between four possibilities, all can receive an equal share (0.25). When dividing between five, again all can have an equal share (0.20). But dividing between more and more, the share is smaller and smaller. When the number of hypotheses reaches infinity, the purported equal amount drops to zero. The only way to assign probabilities over countably infinitely many hypotheses (while remaining countably additive) is to use a *convergent series*, a series of numbers that tapers off to produce a finite sum. The stacked priors in Figure 2.1 form such a series.[3]

This motivation for stacking priors is limited to cases in which we are comparing infinitely many hypotheses. Not all inference problems are like this. In fact, many inference problems involve only finitely many alternatives. Our Hubble case, for instance, involves only a finite number of alternative hypotheses. As such, this motivation for prior stacking is not probative in the present context.[4]

Various philosophers of science have endorsed PSA as a normative/prescriptive ideal. For instance, Salmon, Earman, and others have taken this stance. In this sense, PSA has become a well entrenched part of Bayesian confirmation-theoretic lore. PSA is easy to use and clearly effective as a form of favoring simple hypotheses.

[3] Philosophers may recognize the factors of $\frac{1}{2}$ from Zeno's classic paradoxes.

[4] In addition, the convergence requirement only justifies prior stacking in a long-run sense. The stacked priors in Figure 2.1 represent only one possible distribution satisfying the convergence requirement. Other, less attractive distributions will also satisfy the requirement. It is possible to construct prior distributions that assign very low probabilities to simple hypotheses, increase with complexity up to point, and then taper off. This sort of distribution would favor simplicity in some rarified, general sense, but it would defy implications of the Principle of Simplicity, such as that H_1 is better confirmed that H_2. So the degree of support that the convergence requirement affords PSA is weak.

2.3 Prior-Stacking in the Curve-Fitting Context

How does prior-stacking work in the paradigmatic case of curve-fitting? Let us do a "prior-stacking calculation" on Hubble's data. In Chapter 1, we performed a "predictive focus calculation" on Hubble's data and saw that the likelihoods for each model were quite different on the data, even though the models were roughly equal in their ability to accommodate the data. In order for prior-stacking to make sense in a curve-fitting context, we will want to focus on a different set of likelihoods—this will be necessary to achieve the parity of fit that the Prior-Stacking Account relies on. And there is a set of likelihoods that give us that parity: the MLs. Thus we can define a prior-stacking procedure this way.

1. Select hypotheses as the best-fitting trends from each model, $H_1 \to \hat{h}_1, H_3 \to \hat{h}_3, H_5 \to \hat{h}_5$.
2. Assign stacked priors to the trends in accordance with the order of complexity of their parent models: $P(\hat{h}_1) = \frac{1}{2}, P(\hat{h}_3) = \frac{1}{4}, P(\hat{h}_5) = \frac{1}{8}$.
3. Use an estimate of experimental error to determine a normal error term (shared by all models).[5]
4. Apply Bayes' Theorem to 1–3 to determine posteriors, and therefore favoring relations, among the best-fitting trends, $\hat{h}_i$.
5. Expropriate this favoring ordering among the trends to obtain a favoring ordering among their parent models, H_i.

Performing the relevant calculations on Hubble's data is relatively straightforward. I use the geometric distribution represented in Figure 2.1 for the priors. The best fit curves are easy to find, and the corresponding trends determine likelihoods. It is an inevitable result that up to a high degree, more complex models have better-fitting, higher-likelihood best-fit trends. Multiplying the prior by this likelihood gives the posterior for this method.

[5] Mine is $N(0, 0.8)$.

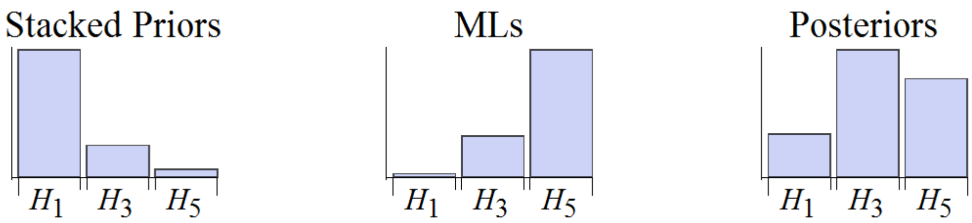


Figure 2.2: A prior-stacking calculation on Hubble's data. Simpler hypotheses are given an advantage in the priors. But more complex models have higher likelihoods. On this approach, the evidence barely favors the linear hypothesis.

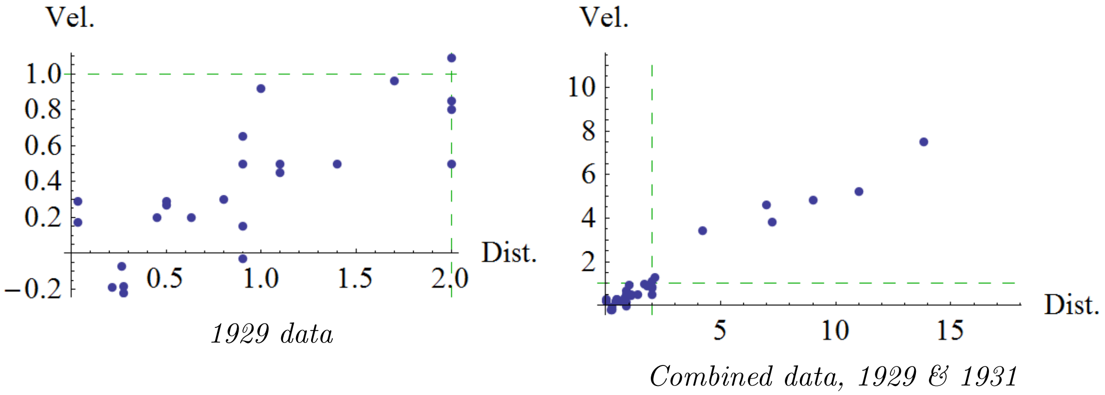
The results are represented in Figure 2.2. The simpler models have higher priors, corresponding to Jeffreys' proposal. But the more complex models can accommodate the data better, as shown by the ML values, and perhaps as something of a fluke, H_3's ability to accommodate the data well outstrips H_1's. As a result, this prior-stacking calculation does *not* favor the linear model.[6]

Prior-stacking successfully favors the simplest model. But the favoring is weak. The simplest model is just barely the best confirmed. If nothing else, the calculation shows a disconnect between the Prior-Stacking Account, or at least its current implementation, and intuition. These results are not the proper grounds for previous astronomers' confidence and excitement over Hubble's Law.

2.4 Sensitivity to the Dataset

Earlier in this chapter, I focused on the most obvious philosophical problem with Prior Stacking Account: it is metaphysically prejudicial. Its ability to respect the Principle of Simplicity comes at the high cost of a suspicious *a priori* presumption. But this is not the only cost of prior-stacking. The advantage of simplicity seems to depend on certain features of the dataset involved. While holistic Bayesian model evaluation, the engine of the Predictive Focus Account, is sensitive to these features, prior-stacking is not. The flat, *a priori* advantage for simple models on the Prior Stacking Account

[6] The proportions of likelihoods here are not robust. They are dependent on an estimate for experimental error. My estimate from Chapter 1 of $\sigma = 0.8Mm/s$ may be too optimistic. But then again, flukes like this can and will happen in statistical contexts.



1929 data

Combined data, 1929 & 1931

Figure 2.3: Hubble's data in the 1929 paper ranged to 2 mega-parsecs (Mpc). But for the 1931 paper, the range was much larger, to 32 Mpc.

prevents it from having much sensitivity to details about the data.

Let me return to the case of Hubble's "Velocity-Distance Relation" in Chapter 1; there I focused on the compelling data from his key paper in 1931. But in fact, Hubble wrote a series of papers on the subject. An earlier paper, in 1929 , offered a dataset with a much more limited scope, relatively speaking. The distances of objects only ranged up to 2 mega-parsecs (Mpc).With a significant amount of "noise" in the data—random motion among the objects—the data supported a linear relationship, but the support was underwhelming. More evidence was needed to make a compelling case.

The 1931 paper provided this evidence. It expanded the range of distances from 2 Mpc to 32—the "data window" was much larger in the second set. And the 1931 data fell nearly along the path suggested in 1929. It was a convincing case; Sandage (1994) comments that the later paper furnished "conclusive proof" of the linear relationship. See Figure 2.3.

Why was the 1931 paper more significant? Intuitively, the scope is the biggest factor. But let us focus just on the number of data. Looking back, Weyl's early argument was that it would be extremely unlikely for a set of 20 data to be compatible with a simple linear relationship, if those data were generated by chance. Of course, the number of data we consider makes a drastic difference, does it not? If the sample size were only 2, it would be trivial that the linear model fit—any 2 points determine a straight line.

If 5 data lined up, it would be interesting, but no cause for excitement. 10 data (as we have seen) can be significant, and 20 even moreso. The more data in the set, the lower the chance probability of fitting a simple model. The more data in a set that is compatible with a simple model, the stronger the advantage of simplicity should be.

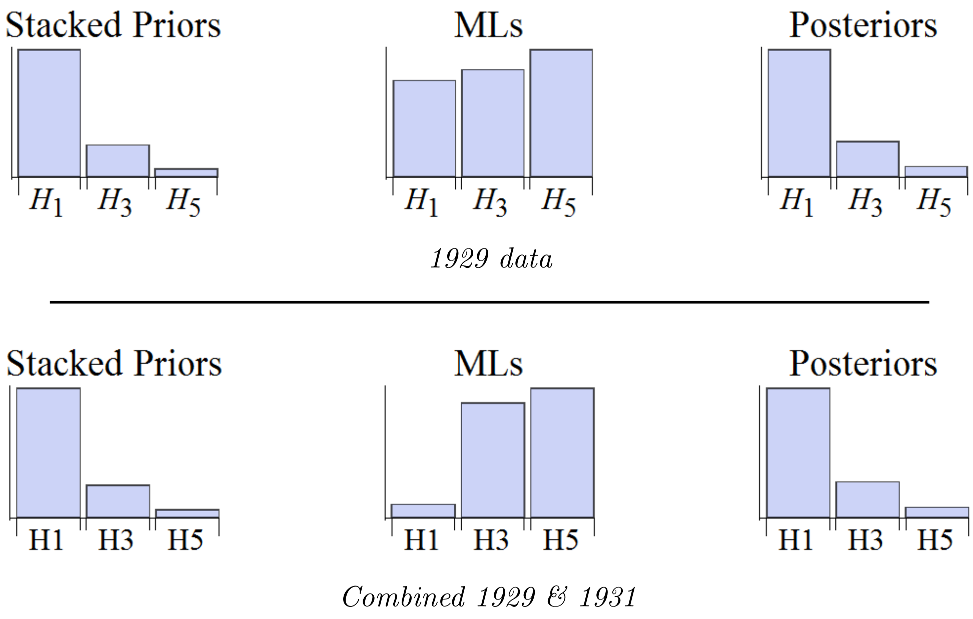
So we should expect our account of the advantage of simplicity to be sensitive to the size of the dataset, and perhaps to other features as well. An account of the advantage of simplicity should judge the combined data in 1931 to favor the simplest model (H_1 from Chapter 1) more than the 1929 data does on its own. Ideally, we want to see the simplest model have a modest advantage on the early dataset, and a dramatic advantage on the combined set.

Carrying over the methods from Chapter 1, as well as the assumptions about the experiment, we can compare calculations. First, let us look at the prior-stacking method. Here, the advantage of simplicity is constant—it is the bias that the agent gives to the simplest model. Let us adopt the bias discussed earlier, where each more complex model has half the prior probability of the former.

On the prior-stacking method described here, models are compared by their best-fitting hypotheses, their "maximum likelihood values" (ML). Adding more data should decrease the ML of all models.[7] As long as each model remains compatible, the ratios of their MLs will be similar. Thus on prior-stacking, when the simplest model is true, we expect little change in the posterior, no matter how much data we have. So prior-stacking will not provide a method of favoring simple models that is sensitive to the data. See Figure 2.4.

However, the Predictive Focus Account has the right behavior on these datasets, at least with the use of the Data Window Prior. The 1929 dataset is not very telling, particularly because error interval is large against the data window. So we should not expect the experiment to be very informative. Combining the datasets both increases the data window and the size of the dataset. Now there is a much wider range of

[7] Since we use continuous likelihood functions and evaluate them by their densities, it is possible for more data to *increase* the likelihood, if the error distributions are narrow. But this is an artifact of convention. Likelihood should decrease, since each additional data point expands the range of possibilities for the set.



1929 data

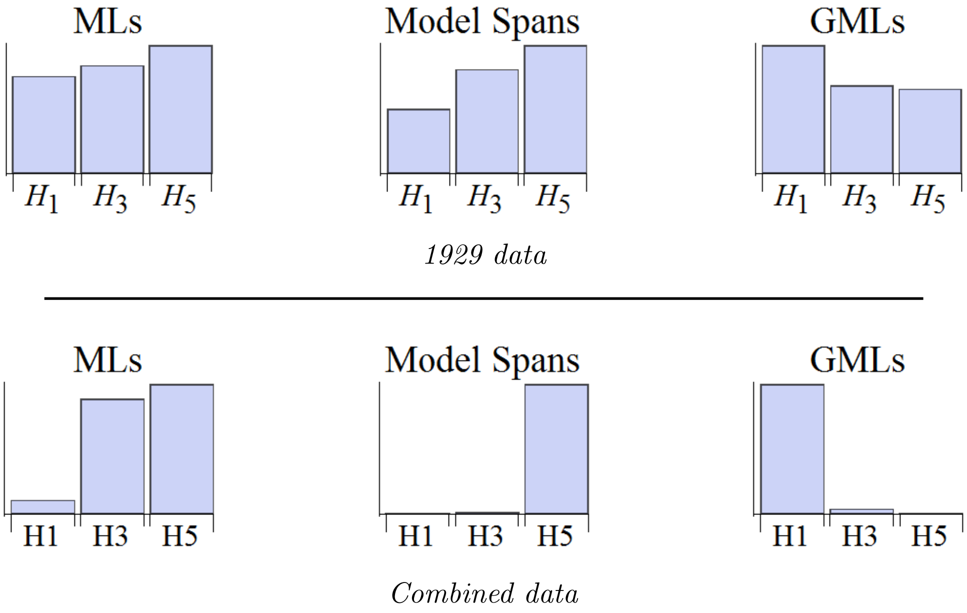
Combined 1929 & 1931

Figure 2.4: Prior-stacking calculations on the two datasets. The results look good on the 1929 dataset—it is a modest advantage for H_1, the linear model. But prior-stacking does badly on the combined dataset. It favors H3, the cubic model, which is intuitively the wrong result. It certainly fails to show an increased advantage for H_1, as we expected.

possibilities—a much broader ideal chance distribution—and the experiment has more opportunity to discriminate between the models. See Figure 2.5.

In a curve-fitting context where all models will fit the data, the advantage of the simple model should grow as the data pile on. The more data that fall within the margin of error for a linear relationship, the more significant the data for the linear relationship. After 20 or 40 or 100 data, and given a powerful experiment, the simple model should reach conclusive confirmation. But this is only possible if the advantage of simplicity is located in the likelihoods. If the advantage of simplicity is in the priors, then the simplest model has, at best, the same advantage across the board, no matter how large or small the dataset is.

Prior-stacking presents a dreadful dilemma for its practitioner. In Figure 2.2, on the combined dataset, prior-stacking favored the wrong model. The failure was due, in part, to the moderate prior bias for the linear model. That bias was overwhelmed by the greater maximum likelihood values for H3 and H5. That difference in maximum likelihood values is normal, par for the course in the vicissitudes of exact likelihood. A



1929 data

Combined data

Figure 2.5: Bayesian global model evaluations / predictive focus calculations the two datasets. The predictive focus calculation gives exactly what we want. For the 1929 dataset, the simplest model has the highest global likelihood, but the advantage is weak, interpeting the data as ambiguous between the models. With a narrow data window and large interval for noise/error, there is less opportunity for model predictions to differ and differentiate themselves. But combining the datasets produces a different situation, on which the simplest model has a much higher global likelihood.

prior-stacking agent may want to prevent such reversals by giving greater prior weight to simple models. She could choose a much steeper rate of decline for complex models—there are probability distributions that accomplish this. But the more she shores up the advantage of the simplest model, the worse the metaphysical prejudice—the more she asserts high confidence that the linear model is correct before gathering any data at all. And *vice versa*: the less biased the stacked priors, the weaker the advantage of the simple model.

Not all equivocal datasets are equal. On the Principle of Simplicity, if a dataset is compatible with (predicted by) all models, then the data supports the simplest the best. But this is not to say that the degree of favoring should always be the same. Intuitively, while Hubble's data in 1929 left it open whether the velocity-distance relation for cosmic objects was linear, his accumulated data by 1931 largely secured that hypothesis. Historically, astronomers shared this intuition; the 1931 paper was the real cause for celebration. So the correct account needs to be sensitive to features of the dataset; for example, its size. But only the Predictive Focus Account does this. The Predictive Focus Account captures the importance of Hubble's new data, while the prior-stacking calculation not only missed it, but allowed an idiosyncracy in the data (an unusually low ML for the linear model) to overturn the simple model's advantage.

2.5 Bayesian Curve-Fitting Methodology

The Prior-Stacking and Predictive Focus Accounts associate with different curve-fitting methods. Each implies a different "angle of attack" at a curve-fitting problem. The first thing to note is that the prior-stacking procedure is not (strictly speaking) a proper Bayesian procedure. In effect, the procedure conflates the probability that a particular linear curve is the correct one with the general hypothesis that the relationship is linear—represented by some linear curve or other. This might be a little like scoring a basketball team, not on the total number of points made by all players throughout the game, but just by the number of points made by the MVP for that game.

But we may interpret the Prior-Stacking procedure's emphasis on particular curves as a philosophical orientation: that of treating specific curves as the direct objects of evidential evaluation. It is the age-old philosophical desire to compare specific curves without reference to a model, “in a vacuum.” Dealing with curves one at a time removes the challenge of managing entire models, which significantly increase the difficulty of performing curve-fitting calculations. Treating curves as the objects of evidential evaluation also represents a degree of theoretical purity, in that the evaluation of the curve will not depend on extraneous associations to other curves. Evaluating the curve in a vacuum can seem more straightforward. Finally, dealing with specific curves in a vacuum also enables one to exercise the “Weak Indifference” principle suggested at the beginning of this paper. All one has to do is gather the set of curves under consideration and assign each equal prior probability.

On the other hand, dealing with models first is standard curve-fitting practice. Grouping curves into infinite models makes all curves epistemically accessible. All curves, or at least large swath of them, are on the same footing within the probability model. Technically, prior-stacking bends the rules so that very high prior probabilities are assigned to just a handful of curves, the best fitting curves of each model. While this seems to preserve the exercise of Weak Indifference in one way, it is a stark violation in another way: very many close-by, similar curves are dismissed and counted as unconfirmable.

Although there is something to be said for both curves and models as the direct bearers of evidential evaluation, models seem to be the better choice. Although global model evaluation is hard, it is straightforward in a Bayesian context. It is the same object that is assigned prior probability and directly bears a likelihood on the data. In addition, some philosophers working on statistical accounts of the Principle of Simplicity have weighed in on this problem. Sober (1994b) points out that scientists always start from models, and it would be improper for the philosophy of science to ignore this descriptive fact. Forster & Sober (1994) points out that it is more appropriate to treat *all* individual curves as having zero probability—only continuous sets of curves have positive probability.

Taken together, individual curves are the wrong objects of evaluation in the first place. The Prior-Stacking Account, despite its popularity, has trouble getting off the ground in the paradigmatic case of curve-fitting because it seems to require an invalid Bayesian calculus.

2.6 Conclusion

The beginning of this chapter introduced a trilemma regarding the Principle of Simplicity. The Principle of Simplicity itself states that equivocal evidence favors simple models. "Weak Indifference" stated that, in the spirit of impartial assignments of prior probability, an agent should not give prior bias to simple models. "Inheritance" stated that a prior advantage is necessary for a posterior advantage when data is equivocal.

Inheritance is a very natural idea, and it seems to drive a wedge between the Principle of Simplicity and Weak Indifference. But on the Predictive Focus Account, *the Bayesian can have it both ways!* She can have both equal priors across models, but also have the simpler models favored by equivocal data. The trick is a conception of "equivocal data" which is weak enough to allow large differences in likelihood, but strong enough to make the Principle of Simplicity relevant.

Holistic Bayesian model evaluation, which features in the Predictive Focus Account, also avoids the other pitfalls of prior-stacking. It is standard Bayesian Theory—we are under no pressure to perform awkward expropriations. In addition, as calculations in the last section show, holistic Bayesian model evaluation can produce strong results in favor of the simplest model.

On the Predictive Focus Account, simple models do not have a head start. They just have higher likelihoods on their predictions than complex models. The resulting interpretation of the Principle of Simplicity is not as a rule to enforce over and above the dictates of the evidence, but more as a rule of thumb, perhaps in contexts where exact likelihoods are difficult to compute, but the general features of the inference problem (especially simplicity) are clear enough.

Chapter 3

Simplicity's Strongest Suit

Comparing the AIC and Predictive Focus

Previous chapters have remained "in house" with Bayesian Theory. But there is a very important account of the Principle of Simplicity that occurs in a completely different context. Based on a formula called the AIC and presented in Forster & Sober (1994), this account holds that more complex models are more prone to overfitting, encoding more observational error in the estimation process. As a result, simpler models will yield more predictively accurate hypotheses. On this account, the advantage of simpler models has nothing to do with probability or even confirmation, but is a matter of reliability for future predictions. While the AIC account is compelling, this essay will emphasize its limitations and argue that the Predictive Focus Account is the better general account of the advantage of simplicity that the Principle of Simplicity (as stated in previous chapters) tries to codify.

Again *model* is a set of fully determinate hypotheses that share a common feature. For example, in curve-fitting, a specific curve represents a determinate hypothesis about the relationship between two variables. A model will contain all curves with a common feature; for example, the model LIN contains all possible linear curves. LIN is an important model in general because it is extremely simple. (See Section 3.1.)[1] Models can be used in different ways. Frequentists use a model as field of hypotheses, from which one might choose the most attractive given the data. Bayesians treat models as a disjunction of hypotheses, constituting a more general hypothesis.

[1] Here I am shifting to speaking away from the models (H_1, H_2, etc.), and toward the pure polynomial models (LIN, PAR, etc.). My approach toward Predictive Focus will still employ a disjoint structure (p. 16), but the naturally nested structure of the polynomials will be perspicuous (and familiar) for talking about the AIC. Forster & Sober (1994) have made the question of nested vs. non-nested models a matter of philosophical issue, but I do not consider this to be a deep issue. See Section 3.3 for brief comments.

When a determinate hypothesis's prediction differs from the observed value, the difference is called error. A hypothesis is accurate if it comes close to the data, or error is small. Let *model accuracy* be depend on hypothesis accuracy: a model is accurate if one of its hypothesis-members is accurate. In other words, a model is accurate if it can accommodate the data well. In addition, let model accuracy be restricted to the current data—we assume that we have some data on hand, the current data, even though we are likely to collect more in the future. The model's most accurate hypothesis is often called the "best-fitting hypothesis," or simply *best fit*. In this context, the Principle of Simplicity presents a trade-off between a model's accuracy and its simplicity: the model with the optimal trade-off is best.

The two accounts of the Principle of Simplicity use models in different ways, and it will be helpful to introduce a couple more terms early. Let *model predictive accuracy* be the accuracy of the model's best fit, from the current data, when applied to future data. So a model's predictive accuracy is a matter of a certain kind of process: using the model to determine a single hypothesis, and then using that hypothesis to predict future data. This is the way models are used in the AIC. Call a model *confirmed* if the model's prior probability is less than its conditional probability on the data—if the data increases the probability of the model. Here confirmation is a matter of degree—"confirmed" does not mean "conclusively confirmed" or "verified." Addressing model confirmation treats the model as a general hypothesis in a Bayesian scheme, and is incompatible with the AIC approach.

The truth is predictively accurate, so the goals of confirmation and predictive accuracy are related through it. This is to say, both goals are broadly *alethic*, or directly related to truth. Both of these goals are epistemically valuable, and each has some tendency to support the other.

The comparison of these two accounts implicates major issues in philosophy of science, besides the status of the Principle of Simplicity itself. An obvious implication is the debate between Bayesian Theory and frequentism as theories of evidential support, featured in Loewer *et al.* (1978); Royall (1997); Fitelson (2012). The Predictive

Focus Account is Bayesian. On Bayesian Theory, the epistemic status of a hypothesis is a probability assignment. Typical scientific hypotheses are not considered to be stochastic, but they are nevertheless assigned a probability, reflecting how plausible the hypothesis is, given the current state of evidence. Frequentism, associated with the AIC, is the name given to non-Bayesian theories (classical Neyman-Pearson statistical testing, Likelihoodism). It is so-called because of the common identification of probabilities with frequencies. Frequentism takes advantage of the conditional probabilities that hold for data assuming various hypotheses (likelihoods), but makes no use of unconditional probabilities as such. Here, hypotheses do not gain or drop in probability, but they are selected in response to data.

Another implicated issue is that between scientific realism and instrumentalism, featured in Maxwell (1998); van Fraassen (1980). These are theories about the general goal of hypotheses in science. On realism, the aim of a hypothesis is truth. On instrumentalism, the aim is not truth as such, but predictive accuracy. The AIC has been touted as representing predictive accuracy, taken to be a more pragmatic goal than pure truth (Forster, 2002). Since Bayesian evaluations of hypotheses are probabilities (to truth), Bayesian Theory allies with a realist picture.

3.1 Statistical Accounts In General

The statistical approach to the Principle of Simplicity looks to the statistical properties of models to locate the disadvantage or liability of complexity. As such, it assumes a starting stock of models. The statistical approach takes the prior selection of a stock of models for granted, and does not tell us how to do it. Presumably, these should be natural (Lewis, 1983), projectable (Goodman, 1983), or otherwise plausible or attractive choices for investigation.

A dominant paradigm for PS in philosophy is *curve-fitting*, an inference moving from a set of point-data to a hypothesis about the relationship between variables. The hypothesis cuts through variable space and is visually represented as a curve. The “curve-fitting problem” is the problem of selecting the true, or best, curve on the data, particularly in light of simplicity considerations. Statistical accounts solve the curve-fitting problem by locating some liability in the use of more complex models.

The AIC and Predictive Focus Accounts have different takes on what that liability is. However, the common theme is that complex models are more "flexible." The dominant paradigm for complexity here is the *paucity of parameters criterion*; Popper gives this criterion in *Logic of Scientific Discovery* (Popper, 1992).

PPC The order of complexity of a model is given by its number of adjustable parameters, k.

PPC has been criticized as failing to capture all there is to complexity—indeed, I will stray beyond it later in this essay in discussing interval hypotheses. But it is punchy and fits well with a wide range of cases (particularly when models are constructed in a natural way). Simplicity is the contrary to complexity, so a measure of complexity doubles as a negative measure of simplicity.

Having more adjustable parameters produces a more flexible model. Intuitively, the model has a broader range of hypotheses, and a corresponding broader range of possible datasets that it can accommodate.[2] This breadth is a blessing and a curse. A more flexible model may be more broadly applicable, but it may also be less reliable, or more likely to appear to be a good one when in fact it is not.

It is worth noting here that the statistical approach commits to the evaluation of models (or model–best-fit pairs), rather than specific, determinate hypotheses. This is a substantive move. General statements of the Principle of Simplicity do not commit to models. In addition, the same curve can belong to a number of different models. Thus a hypothesis curve tends not to be evaluated in a pure, extensional sense, but as a part of a broader family.

There are some advantages to approaching complexity through models, rather than individual hypotheses. First, all attempts to describe or measure the complexity of

[2] I say "intuitively" here because ranges are not well-defined at this point. For the AIC, talking of ranges is a little misleading. But in the AIC-frequentist paradigm, we tend to use a nested model structure where more complex models contain simpler ones. Here, we can safely say that the more complex model contains strictly more hypotheses and possible predictions. For PFA, additional input is needed to define the relevant ranges (Chapter 1).

hypotheses, particularly mathematical objects like curves, default to general properties of those objects that effectively determine models.[3] To put this point another way: models are just classes of hypotheses, and the only alternative to using models is to examine each possible curve by rote, one at a time, in isolation. Secondly, using (*prima facie* attractive) models is standard for scientific practice, so philosophical evaluations of specific hypotheses in isolation is not a promising project (Sober, 1994b).

The statistical approach to accounting for the Principle of Simplicity has come under philosophical criticism.[4] McAllister (1991) argues that there is more than one kind of simplicity that scientists favor: a logico-empirical kind (which is the kind that statistical accounts address, and is designated by CPS) and a quasi-aesthetic kind. Ackerman (1966) points out that while model selection criteria have something to say about the right order of complexity for a model, they do not determine the model's form. Kukla (1995) points out that it is possible to construct a model which is syntactically simple—has few or no adjustable parameters—and is accurate, but is "bumpy" and intuitively wrong. And Priest (1976) shows how to construct "gruesome" curves—functional analogues of Goodman's "grue–bleen" predicates (Goodman, 1983).

These challenges can be mostly summarized with two themes.

1. The success of statistical model selection in selecting the intuitively correct model on the data (and critically, taking simplicity into account) depends on the prior selection of intuitive, attractive, or natural models.

2. There is some flexibility in describing a model along the dimensions of complexity and naturalness.

In brief, both of these challenges are correct. But do they really undermine statistical accounts? I take the default, most attractive view of science is in allowing a great many models to be evaluated on the data, but not all possible models, and there is an established philosophical tradition about what makes a model promising (Lewis, 1983;

[3] See the theories of Sober (1975); Goodman (1983); Swinburne (1997).

[4] The AIC has been best-presented and most prominent in recent years in these discussions, so the challenges are often directed at that account. But the challenges apply generally.

Goodman, 1983). In other words, statistical accounts are not the right tool for solving the "grue-bleen" problem, and they do not intend to be.

Even after a choice of language has been made (on whatever proper grounds for that question), there will still be plenty of live models—there will still be plenty of work for the Principle of Simplicity. On these statistical accounts, the statistical properties of models do that work. PAR is an excellent example[5] of a model which is perfectly kosher for science, yet is occasionally unnecessarily complex (particularly against LIN).The challenges above are compatible with this picture, so *prima facie*, statistical accounts are not in trouble.

In any event, these challenges do not affect the AIC and PFA differently. Both accounts are in the same boat. So I will set these aside and move on to analyzing these statistical accounts.

3.2 The AIC

Akaike's Information Criterion, the "AIC," is an estimate about an estimate. It is an estimate of the predictive accuracy of a hypothesis, when the hypothesis is selected from a model with a certain degree of complexity as determined by PPC. We might think of this, not as evaluating not the curve in a vacuum, but as evaluating the result of a process.

Even if a model contains the true hypothesis, we would not really expect to find this hypothesis when fitting the model to the data. The data always contains some degree of error, and it is overwhelmingly likely that this error will favor some other, nearby hypothesis.[6] Unless there is a perfect, errorless dataset, or unless the errors perfectly "cancel each other out," just-so for the model involved, then the data will implicate some false hypothesis (albeit close to the truth). So some error in the data will translate into some error in the chosen hypothesis. In effect, some observational

[5] The French call this an example *par excellence*.

[6] Technically, the assumption that errors are normally distributed entails that the probability distribution is continuous. Thus the probability of having no error, of the observation exactly reflecting the true value, vanishes to zero, even though there is positive probability density at that point. Here we say that the observation "almost surely" differs from the truth.

error is “absorbed” into the model fit. Call the error inherent in fitting a (true) model to data *estimation error*.

How much error to we expect an estimated hypothesis to absorb? The exciting statistical discovery in the AIC is that estimation error is a function of model complexity! More complex, flexible models can accommodate the data better, but in addition, they are more susceptible to estimation error—they absorb more of the noise from the data than simple models do. Accuracy is contrary to error, and the expected predictive error for this hypothesis is cumulative from two sourcess: the current residual error of the hypothesis on the data, and the estimation error that results from choosing it. So the AIC encodes an incredibly natural trade-off between accuracy (on the current data) and simplicity.

I discuss the AIC as the core of a paradigm for PS, but the AIC is actually just a formula which provides the estimate of prediction error. (So the lowest AIC score will correspond to the highest estimated predictive accuracy.) I write “ML” for the model’s maximum likelihood argument—that is, the likelihood of its best fit on the data. Typically, the likelihood on the data is a function of error—error is treated as a random variable, and the more error the data has for a hypothesis, the less probable it is considered on the hypothesis. “ln” is the natural logarithm; the AIC is a function of the logarithm of the likelihood, or just “log-likelihood.”

$$\text{AIC: } 2k - 2\ \ln(\text{ML}) \tag{3.1}$$

Likelihood is a decreasing function with error, so a higher likelihood for the best fit means higher accuracy on the current data.[7]

The importance for the AIC to PS is obvious. But to be philosophically careful, it is important to distinguish the formula from the account that employs it. As admitted in Section 3.1, the procedure only makes sense if the choices for models are *prima facie* attractive, and differ mainly in complexity as such. Accuracy is contrary to error, and in this context, models are used to select a single hypothesis, the best fit. Then the

[7] The decreasing relationship between likelihood and error is a result of the “normality assumption” in the AIC (Forster & Sober, 1994, App. A).

AIC Account of PS can be summarized in just a couple points.

1. The general expected error for a hypothesis, fitted from a model, is the combination of error (inaccuracy) on the current data and inherent estimation error from the model.

2. More complex models have a higher degree of estimation error.

Combining these statements yields the conclusion that complexity degrades, and simplicity enhances, the predictive accuracy of a model.[8] So we have an account which is driven by estimation. Akaike commented that error can be taken to imply run-away complexity, and philosophically remarked, "It can be seen that the subject of statistical identification is essentially concerned with the art of approximation, which is a basic element of human intellectual activity" (Akaike, 1974).

There are a few important details for the AIC. The criterion estimates predictive accuracy, but only directly for the original sample points. In a curve-fitting example, we will have a set of values for the independent variable x. We can write the set as the vector $X_1 = [x_1, x_2, \ldots, x_n]$.[9] Values for y can be written as the vector $Y_1 = [y_1, y_2, \ldots, y_n]$. Combining these vectors gives a more familiar set of data points, $\{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots\}$. Typically, gathering more data means new vectors for both variables: X_2 and Y_2. But the AIC only estimates predictive accuracy for new vectors Y on the original vector X_1.

The relative improbability that the sample points X will be the same makes the restriction to the original points X_1 seem like a strange limitation. But especially from a frequentist standpoint, this limitations is fairly natural. An assumption about the sampling probabilities for X would be a hefty one. Probabilities for Y on X (naturally, dependent on probabilities for single-values of y and x) are a matter of observational error, the general probability distribution for which was the result of a hundred years of mathematical searching (Stigler, 2003). There is no parallel for the

[8] In statistical terms, residual error and estimation error correspond to variance and bias, respectively. See Hastie *et al.* (2008, ch. 7).

[9] I will not worry about vector orientations here, column vs. row.

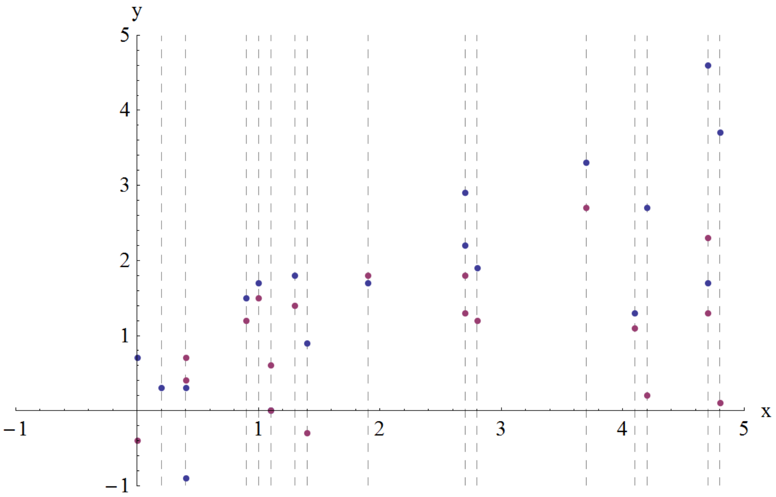


Figure 3.1: The AIC measures predictive accuracy on resampling from the same X vector.

independent variable. In the frequentist scheme, X is supposed to be determinate and non-probabilistic. So the AIC only measures predictive accuracy on *resampling* of the data, taken narrowly to mean gathering a new dependent vector Y_2 at the very same independent vector, X_1.

3.3 The Predictive Focus Account

Previous chapters of this dissertation present the Predictive Focus Account (PFA). The most fundamental difference between PFA and the AIC account is its statistical foundation; the AIC is frequentist, PFA is Bayesian. The statistical foundation determines how each makes use of models. At base, frequentism makes pairwise comparisons of hypotheses.[10] So it is important to select the best hypothesis from a model for comparison. Attempting to evaluate the model as a whole, to evaluate the whole set of hypotheses, would stress the conceptual frequentist framework.

By contrast, Bayesian Theory operates off of a complete, rather than partial, probability model (Fitelson, 2007). In principle, each model has a prior probability, as does

[10] Both Neyman-Pearson hypothesis testing and Likelihoodism are justified in terms of the likelihoods of a pair of hypotheses; see Royall (1997).

each hypothesis within the model. This has the consequence that likelihoods are defined, not just for hypotheses, but for entire *models* on the data. In particular, using the Law of Total Probability, the model likelihood is just an expectation.[11]

$$P(E|M) = \iint_R P(E|h\&M) \times P(h|M)dh \tag{3.2}$$

I call this term the *global model likelihood* (GML), to emphasize that it is the likelihood of the whole model, not the individual hypothesis. Note that a GML just is a likelihood, a likelihood of the whole model. Conceptually, the likelihood of an individual curve and the likelihood of a model are the same kind of mathematical object: the conditional probability of the data on the hypothesis, $P(E \mid H)$. Because model likelihoods are defined in Bayesian Theory, the models can be directly evaluated on the evidence. The Bayesian machinery for accounting likelihoods applies just as well at the general model level as it does on the specific hypothesis level. Thus Bayesian model evaluation can be more holistic and direct than in frequentism.

A model is confirmed just if its probability on the data is higher than its prior probability, $P(M|E) > P(M)$. As such, in order for a simple model to be confirmable over a complex model, they must be non-nested. The complex model should not contain the simpler. Non-nested models represents a change from the standard frequentist framework (especially orthodox Neyman-Pearson hypothesis testing), where successively complex models are supersets of simpler ones. As such, Forster (1995) and Sober (2002) criticize the move to disjoint models as "changing the question," but this is only changing the question from what the frequentist was asking. To my mind, the more natural articulation of a simplicity principle is between exclusive models, and this is the reading that Bayesian Theory prefers.[12]

[11] I write $\iint_R$ to designate the integral over the hypothesis space, whatever the dimension. I am suppressing reference to individual parameters following Rosenkrantz's notation (Rosenkrantz, 1977, ch. 5).

[12] Although there is no need to do so, the Bayesian could bite the bullet and accept that simple models never have higher posteriors than complex models, and so in that sense, are never better confirmed. Yet, simple models could still enjoy higher degrees of *relative confirmation*, especially for the likelihood-ratio measure of confirmation adopted here. Fitelson (2012) shows how the concept of relative confirmation untangles a problem with a similar structure, where a hypothesis seems to be better supported by the evidence, but is necessarily less probable, than a competitor which contains it. Fitelson's solution here naturally also applies to the problem of simple vs. complex models where model structure is nested.

Bayesian Theory can be helpful for a study of simplicity-favoring because of Bayesian calculations can offer rich verdicts. In contrast to the frequentist approach described above, Bayesian calculations do not require definitive acceptance or rejection of hypotheses or models. A Bayesian evaluation of models associates them with a probability model, representing their plausibility or epistemic standing on the data.[13] It is an intuitive consequence of Bayesian Theory that some models can be counted as disconfirmed, but remain within the realm of consideration and even influence an agent's expectations. For example, suppose there are five models, $M_1 - M_5$, ordered in complexity, and assigned probabilities by a rational agent. Suppose that these probabilities decrease with complexity: 0.6, 0.2, 0.1, 0.5, 0.1 (and some probability leftover for other models). While the simplest has the best standing, all models contribute to the agent's general expectations and predictions. There is no need to cut these less probable models out of the general picture of what the data say (Wasserman, 2000). In giving the lion's share of probability to the simplest model, this Bayesian probability model favors simplicity; simultaneously, by accounting a contribution from all models toward prediction, the probability model may be said to "embrace complexity." In this way, the very same evaluation of models can be taken to support multiple, coherent attitudes.

The Bayesian approach to the advantage of simplicity has been heavily criticized, particularly in contrast to the AIC framework. Forster (1995) claims that simplicity is *the* problem for which Bayesian Theory cannot provide an adequate treatment. One criticism regards the proper question for PS, just previously discussed. The central objection is that GMLs depend on priors; in particular the conditional prior distribution $P(h|M)$ (Equation 3.2). If priors are subjective and GMLs depend on them, then GMLs are subjective. But objectivity is a primary desideratum for a theory of confirmation / evidential support. This was the central problem of Chapter 1, and is old news for us.

But since there is neither a technical nor a conceptual problem with the use of disjoint models, I will assume the more basic framework where models are evaluated by posterior probabilities.

[13] Note that models are being used in two senses here: models of physical phenomena, which we want to evaluate, and models of probability, which apply to the first models. Like the two kinds of likelihood described above, both models are two kinds of the same, very general conceptual apparatus. But when I use the word "model" throughout this essay, I will almost always mean those that apply to physical phenomena and that we wish to evaluate.

There we saw a proposal toward objective Bayesian priors.

An alternative response to Forster's challenge is another objective Bayesian model selection criterion, the Strict Minimum Message Length criterion (SMML) (Dowe *et al.*, 2007). The background for SMML is in coding theory. The directive for SMML seems to be less directed at confirmation and more at communication. It is to represent an experimental finding, a dataset, in an efficient code. Simple models are preferable because they provide shorter compressions of the data, and on one version, the trade-off is equivalent with BIC (Vitányi & Li, 2009).

However, coding theory does not seem address the philosophical question in PS, because coding is not a matter of evidential support, not conceptually anyway. In fact, a coding-theoretic approach would seem to give a circular rationale, since efficient, short descriptions are simpler. And efficiency has much in common with elegance, one aspect of simplicity. So taking efficient description of the data as a goal does not seem to justify PS, so much as replace it. It does not enable the synthetic explanation of the Principle of Simplicity contained in PFA.

I take myself to have provided an initial solution to the problem of assigning priors for the relevant kinds of inference problems, especially curve-fitting. The idea is a prior that maximizes entropy over the range of observations possible for the experimental apparatus associated with the data being evaluated. The solution grounds prior choice in a legitimate distinction for science, between those events within the observation range of an apparatus and those outside. I take it this is a natural supplement to PFA.

For the remainder of this paper, I will count PFA as philosophically and technically viable, putting it on equal preliminary footing with the AIC. For the remainder of the essay, I will compare these two accounts. We might expect that the only way to determine which is better is to refer directly to the philosophical foundations: AIC is better just if frequentism is the better account of statistical evidence, etc. In fact, the AIC and Bayesian holistic model evaluation (central to PFA) differ in important ways. While each has something to offer, I will be arguing that PFA is the better choice for a singular account of simplicity as a theoretical virtue.

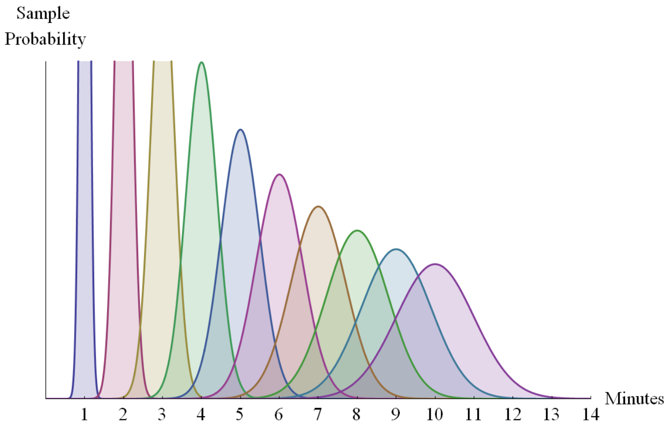


Figure 3.2: A possible probability distribution for X. I imagine a work-at-home scientist checking the pressure of a sealed pot, in between other small kitchen jobs. He intends to check the pot ten times at one-minute intervals, but is more likely to be off from that time for later measurements. I partially choose this example because it makes for an awkward interpolation / extrapolation distinction as presented in Forster (2002).

3.4 Extrapolation

One purpose of model selection is extrapolation. *Extrapolation* means predicting the behavior of a function outside the range of a set of data, for values of x greater than or less than all of those in the initial sample X_1. By contrast, *interpolation* means predicting values of the function within the range of X_1. Forster (2002) claims that AIC estimates interpolative predictive accuracy, not extrapolative. Forster further claims that this is good news; that the limits of the AIC in this regard match with the proper limits of science. This is an attractive proposition. However, there are problems.

Forster suggests that there may be a probability distribution associated with sampling from x. This is plausible enough. As long as there is randomness in the sampling from x, it makes sense to say that there is an associated probability distribution here. Sober (2001) suggests a simple curve-fitting problem on the relationship between the temperature and pressure of a gas, with a sealed pot on a stove-top. We might imagine a (slightly distracted) experimenter checking the pot every so often, and a corresponding probability distribution for x values (Figure 3.2).

The problem is that X probability distribution does not seem to be incorporated into the AIC. Nor is it clear that it could be. AIC must assume that the error distribution (probabilities for Y) must be "nice"—Y must be i.i.d., and each component approximately normally distributed. But neither needs to be true for X. In Figure 3.2, X is not i.i.d. And it is easy to think of a distribution that violates normality; e.g. the geometric distribution.

What the AIC really estimates is predictive accuracy on resampling, i.e. resampling the data at the exact same points, taking a new Y with the same X. We can call this "interpolative prediction," but this is interpolation in quite a narrow sense.

Now common sense has it that an estimate of predictive accuracy on resampling, narrowly construed, should be a good guide to predictive accuracy on nearby points. So it is reasonable to think that the verdicts of the AIC should have some extension beyond these very narrow bounds.

But what is the argument that AIC estimates should extend this way? Let me start off with one that is clearly bad.

1. *Suppose that $\langle x_1, y_1 \rangle$ is in the current dataset, $\langle x_2, y_2 \rangle$ is a new, unobserved datum, and x_2 is close to x_1.*

2. *If x_1 and x_2 are close together, then the corresponding values y_1 and y_2 should be close together.*

3. *The AIC will tend to select simple models, with smooth, flat best fits, relative to possible alternatives.*

4. *If y_1 and y_2 are close together, the AIC will tend to select a hypothesis that assigns close values to each.*

C. *The AIC will tend to select accurate values for y_2, with a similar accuracy (albeit a little less) than for y_1.*

The above argument is intuitive. But as an argument that AIC is an account of PS for interpolation, it is circular. The problem is Premise 2. This premise implies a simplicity principle on its own. That is, it implies that the truth will be smooth and flat, similar to the kind of curve that the AIC selects. This is a circular move.

We might try a very general, inductive argument for extending predictive accuracy.

1. *The AIC selects model M^* and hypothesis h^* as predictively accurate for X_1.*

2. *Therefore, h^* is predictively accurate to degree S for each x_i in X_1.*

C. *Therefore, h^* is generally predictively accurate to degree S, and will be accurate on other X.*

This argument is also intuitive. I cannot critique it in any fine way, but the conclusion (in its unqualified formulation) is clearly false. In curve-fitting, the predictive accuracy of a fitted model drops with the distance from the original sample. The farther one gets from the original data set, the worse the prediction, especially if the relationship is a non-linear, non-convergent function.

The reason is obvious: a small error in a parameter now will lead to big errors in prediction later. There is a nice analogue here to rocket science. In the early days, rockets were notoriously difficult to direct because a small error in the direction of the rocket, either due to imperfections in the body, an imperfect initial trajectory, wind turbulence, and other sources of error, will tend to send it way off course.

Notice that we might substitute other properties into the argument above to get a more plausible result. We might think that while a chosen hypothesis h^* will not be *as* predictively accurate in extrapolations as it is on the current data, that it is more likely to be *more* accurate than any of the considered competitors, in accordance with some as-yet-unknown function describing accuracy degeneration. Another intuitive idea, but the argument is incomplete.

So the AIC does not seem to support predictive extension. Its positive, simplicity-favoring verdicts are technically only valid for the original sample points X_1. How does the Predictive Focus Account do better? As a Bayesian scheme, its verdicts are obviously determined and limited by its assumptions, especially the prior probabilities assigned. In a pure-conceptual mood, it is unclear how it can be established *that* a hypothesis is predictively accurate on new data for any new X, any at all, while withholding on the hypothesis's probability to truth or verisimilitude. Because Bayesian Theory attempts to find true hypotheses, interpolative and extrapolative inferences are conceptually supported—even if their verdicts are limited by the priors.

Let us set theory aside. What about actual practice? I suggested just a moment ago that no method can do particularly well in extrapolation. Yet even in the face of failure, Bayesian holistic model evaluation has a pragmatic virtue. Bayesian updating produces nuanced predictions. While the AIC selects a single hypothesis that cuts a rigid path through space, Bayesian updating produces a distribution-graded, infinite set of curves.

On a standard approach, the set will be continuous. But we can think more simply in terms of a sample of curves that represents the whole continuous distribution (Figure 3.3). The predictive distribution for a Bayesian model spreads out for points farther away from the original data. But at points within or close to the data, the predictive distribution is narrow and predictive confidence is high.

This behavior perfectly matches our intuitions. We feel confident for interpolations and short-range extrapolations. We feel less confident for mid-range extrapolations, but will tend to make an "educated guess" about data in this vicinity. For long-range extrapolations, confidence wanes and we may be reluctant to say much. On Bayesian Theory, these intuitions track the predictions that the data support, through a wide range of hypotheses, starting from a wide-spread prior distribution.

But perhaps the best thing about holistic Bayesian model evaluation is its ability to underwrite multiple attitudes about the data. Bayesian Theory can both capture the evidential advantage of simplicity and simultaneously embrace complexity. How? A Bayesian subjective probability model can assign a higher posterior probability to a simple model than a complex one, capturing the evidential advantage of simplicity through PFA. But if the more complex model maintains some positive probability, its predictions will factor into the agent's total assessment of hypotheses and predictions. In so doing, Bayesian Theory can "embrace complexity"—accounting the input of several different models.

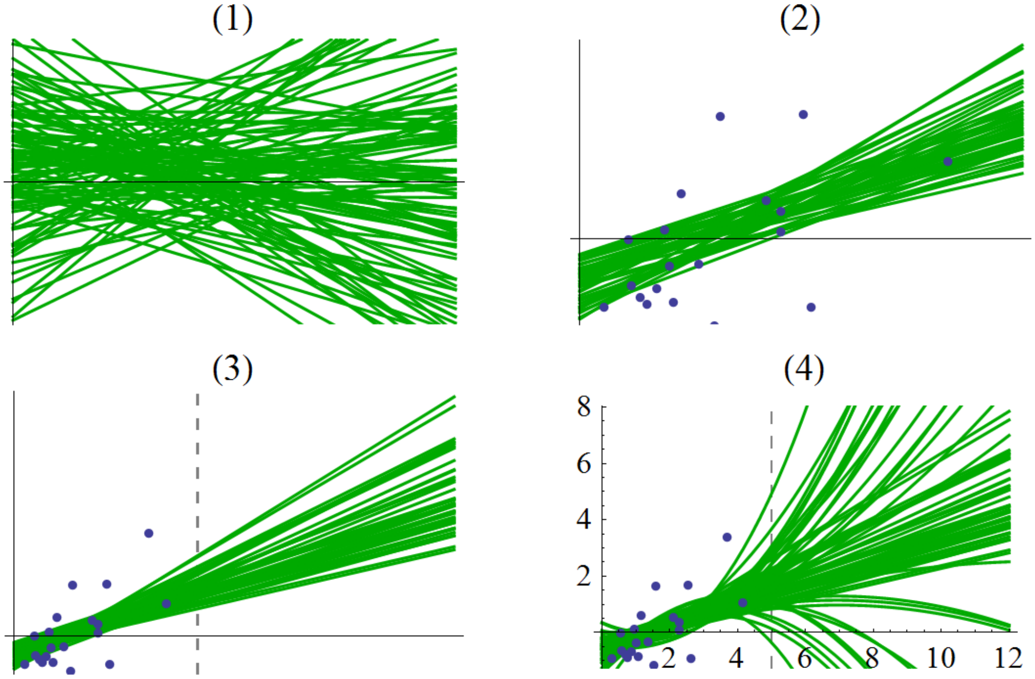


Figure 3.3: (1) We can imagine a Bayesian prior for LIN as a scatter of random linear curves, spread wide. (2) The data drastically narrows the set. The result is the set of all curves which are plausible on the data. (3) Now the set supports predictions beyond the boundary between interpolation and extrapolation. The breadth of the set at any value of x (the range across a vertical cut) represents a range of prediction at that point—inversely proportional to confidence. Generally, the farther out the extrapolation, the wider spread and less confident the prediction. (4) A complete updated probability distribution will include curves from other models as well, adding nuance to the picture that the data supports.

There are good reasons that we would want to combine the predictions of multiple models. McAllister (2007) points out that data can exhibit multiple distinct patterns; mixing multiple models allows an agent to capture these patterns. Raftery & et. al. (2005) points out that combining the predictions of models sometimes gives us *more accurate* predictions than with any single model. These goals might seem contrary to PS, but Bayesian Theory is capable of underwriting these multiple attitudes, both of capturing the advantage of simplicity but also embracing complexity.

No method does particularly well at extrapolation. But holistic Bayesian model evaluation, the methodological basis of the Predictive Focus Account, has several advantages in this arena. First, it licenses extrapolations in principle, while the AIC does not, at least not without a supplementary principle. Second, updated Bayesian models come complete with a measure of confidence, the spread of the predictive distribution, which naturally degrades with distance from the data.

3.5 “Mere” Hypotheses

General hypotheses behave like models, at least in the Bayesian context, by spreading out their predictions over a significant range. Often, such a hypothesis involves a random element. A contest between these hypotheses can implicate considerations of simplicity, even though the objects of evaluation are “mere hypotheses,” as it were, not good examples of models. This is bad news for the AIC Account, on which simplicity enjoys an advantage only at the level of models. By contrast, although the Predictive Focus Account is initially stated in terms of models, has fuzzier boundaries, and can apply to any contest of hypotheses with variable prediction spans.

A hypothesis is just a claim about a natural phenomenon. Like other claims, hypotheses can be more general or more specific. But in frequentism, the role for models is estimation. In this context, a model is something of a field of hypotheses, facilitating the selection of a single one. A hypothesis may be general and contain a fair amount of uncertainty, but if it does not support parameter estimation, then it is not a model in a frequentist context. This is what I mean by a “mere hypothesis”—a hypothesis

	1	2	3	4	5	6	7	8	9	10	11	12
1 Die	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0	0	0
2 Dice	0	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$
3 Dice	0	0	$\frac{1}{216}$	$\frac{1}{72}$	$\frac{1}{36}$	$\frac{5}{108}$	$\frac{5}{72}$	$\frac{7}{72}$	$\frac{25}{216}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{25}{216}$

Figure 3.4: Likelihoods for dice models on scores. More parsimonious dice models—models with fewer dice—usually have higher likelihoods for possible scores.

which is general, but does not involve parameter estimation or further refinement.

Let us start off with a case involving dice. Imagine that Jeff and Jan Anne are waiting for friends to come over to play board games. Their friends are late, and the two invent a quick game to pass the time. Jeff secretly rolls a number of six-sided dice and reports the total score (the sum of the values or pips on the up-side of each die). Jeff is honest and accurate—he never lies about the score, and miscounts less than once in a hundred rolls. Jan Anne simply guesses how many dice Jeff rolled.

Jeff rolls a single die and reports that the score is 4. Jan Anne guesses the simplest hypothesis compatible with the data—Jeff rolled just one die. Her guess was clearly the right one to make: a score of 4 has a likelihood of 1/6 on one die, but just 1/12 on two, and only 1/72 on three.

Of course, this is not a very interesting game, because Jan Anne's best bet is usually the lowest number of dice compatible with the total score. Unless she has a special hunch, Jan Anne should adhere to the Principle of Simplicity in guessing the number of dice used. See Figure 3.4.

Now the Principle of Simplicity states that when a simpler and more complex hypothesis both predict the observed data, the data support the simpler to a greater degree. So the principle is relevant in cases with the following features.

1. There are multiple general hypotheses involved, with a noticeable difference in simplicity.

2. Each hypothesis is compatible with a range of possible data, and the observed data falls within the range for each.

The dice game fits both of these criteria, so *prima facie*, it looks to be a case where the Principle of Simplicity is relevant. By extension, it is a mark in favor of an account of that principle if it can accommodate this case.

Yet key features of the AIC Account are absent. There is no parameter estimation, and no observational error. Again, Jeff is a reliable reporter, so mistakes are rare, but this is just what error would be for this case. And Jan Anne doesn't care about the configuration of pips, just the number of dice, so there is no parameter estimation either. Without these elements, this case does not fit the framework of the advantage of simplicity in the AIC Account.

But the Predictive Focus Account describes advantage of the one-die hypothesis here. Corresponding with the first premise of the account, the (intuitively) simpler hypothesis has the narrowest prediction range. And in accordance with the second, this reduced prediction range corresponds to higher confidence (likelihood) across that range, including on the actual data.

In other words, without having time to exactly calculate the likelihoods involved with different dice-hypotheses, Jan Anne could reason this way:

> *A score of 4 is compatible with one die, or more (up to four). But the more dice in a set, the more possible outcomes for a roll of that dice. The one-die hypothesis allows for the fewest outcomes (6), so a 4 is most likely on that hypothesis.*

But her basic reasoning just is the logic of the Predictive Focus Account. She may be relying on background assumptions and intuitions, but this is the same reasoning that explained the advantage of simplicity in the sophisticated and critical Hubble curve-fitting case in Chapter 1.

In science, there is a project that has some similarity to the dice game case. The project is phylogeny, determining the evolutionary histories of various organisms. A possible particular history constitutes a hypothesis. Models contain multiple hypotheses, and range over them with adjustable parameters, especially regarding the speed of evolutionary change for different traits.

Phylogeny is special for the Principle of Simplicity because it may be the only area of science with an explicit relationship to the Principle of Simplicity. In this context, the principle is simply called "Parsimony," which states that simpler phylogenetic trees tend to have higher likelihoods than more complex ones.

Parsimony fits the paradigm of the Predictive Focus Account well. On that paradigm, the advantage of simplicity is nothing over and above the mere evaluation of hypotheses on evidence. The "Occam's Razor" is "automatic" here (Smith & Spiegelhalter, 1980) because simple models have reliably different statistical properties than complex ones. The connection between simplicity and raw likelihood is so close, they can be difficult to tease apart.

But then if simple trees have higher likelihoods, why need a simplicity principle at all? On this paradigm, the Principle of Simplicity is a rule of thumb to use when calculating likelihoods is too difficult, or requires additional information. Like on the Predictive Focus Account, simplicity is not an extra value imposed on high likelihood; rather, it is a guide to high likelihood.

Parsimony states that the simplest phylogenetic trees tend to have the highest likelihoods. But trees are hypotheses, not models. This best-used example of the Principle of Simplicity is directed at hypotheses, post-estimation, not their corresponding models. Like the case of the dice, the advantage of simplicity for a single phylogenetic hypothesis has its effect on the likelihoods of outcomes, without involving estimation. (Regarding the more sophisticated context of selecting a model and tree, both Bayesian model updating and the AIC have been used to determine the most plausible hypotheses, (Posada & Buckley, 2004).)

Phylogenetic hypotheses also need not involve observational error as such. While the data usually involves noise and so-called "nuisance parameters," some data will be effectively errorless. If, for example, we are tracking broad, qualitative traits, such as the presence of wings or the number of legs on an animal, error may be negligible. Often, a phylogenetic tree's likelihood is primarily a matter of the probabilities of the evolutionary changes it specifies.

Certainly the rhetoric of Parsimony is in terms of Predictive Focus. One textbook explains:

> Parsimony provides one way to identify which branching pattern, among the many that are possible, minimizes the confusing effects of homoplasy [phylogenetically coincidental similarity] and most accurately reflects actual evolutionary history. Parsimony is a general logical criterion. Under parsimony, simpler explanations are preferred over more complex explanations. When parsimony is applied to phylogeny inference, the preferred tree is the one that minimizes the total amount of evolutionary change that has occurred. (Freeman & Herron, 2007, 118)

Why? The easiest explanation is the chance probability argument. For example, forcing a phylogenetic tree to branch too early may imply that two species evolved separately but similarly. In so doing, one often has to double the number of evolutionary changes that occur, resulting in an improbable, "just-so" story about the history of two groups.

> The rationale for invoking parsimony in phylogeny inference is simple and compelling. In many instances, it is valid to assume that convergence and reversal [the two mechanisms of homoplasy] will be rare relative to similarity that is due to modification from a common ancestor . . . Reversals and convergence both require multiple evolutionary changes. It makes sense, then, that the tree that minimizes the total amount of change implied by the data will aslo be the one that minimzes the amount of homplasy. The most parsimonious tree should therefore be the best estimate of the actual phylogeneitc relationships among the species being studied. (Freeman & Herron, 2007, 118)

Making assumptions about the probabilities for various traits to evolve, we can calculate the likelihoods of trees on observed traits. The result is that simple trees tend to have high likelihoods.

Elliot Sober discusses phylogeny in Sober (1994a). His thesis in that piece is that the Principle of Simplicity does not admit of a unified account. Moreover, he stresses that these results depend on assumptions about evolution, on background parameters. The relationship between simplicity and likelihood can be reversed, for example, if one assumes a very high mutation rate. Naturally, if we suppose that organisms are constantly gaining and losing traits, then a tree with a high number of homoplasies can be more likely than a tree with few changes. Nevertheless, Sober's description is an excellent fit with the Predictive Focus Account. He pursues the issue in more depth in Sober (1991), where he discusses the debate over the validity of the Parsimony rule.

I invoke phylogeny here to show that contests of simplicity do not always involve proper models. Sometimes they involve mere (general) hypotheses. And simplicity does seem to have an advantage in these kinds of cases. If so, between the AIC Account and the Predictive Focus Account, only the Predictive Focus Account captures that advantage.

3.6 Restricted Models

Some models have restrictions on their parameters. These restricted models defy the expectations of the AIC–Account, which derives expected model-fit accuracy based on an unrestricted parameter space. But these models are just as natural for the Predictive Focus Account as unrestricted models.

An important kind of restricted model is an "interval hypothesis." An *interval hypothesis* is a hypothesis that a certain, unknown quantity lies within a certain range. For our purposes, such a hypothesis will act as a model.[14] In *Inference, Method, and Decision*, Rosenkrantz shows that interval hypotheses with narrow ranges tend to be better supported than interval hypotheses with wide ranges when both fit the data. Here I will adapt his example from p. 111.

Suppose that Liz is curious about the bias of a coin. She wants to know its probability of landing heads. She has two hypotheses about the coin, giving ranges for the probability of heads p.

- C_1 – The coin is effectively fair.

  $0.49 \leq p \leq 0.51$

- C_2 – The coin is effectively unfair.

  $(0 \leq p < 0.49) \vee (0.51 < p \leq 1)$

Liz flips the coin $n = 4$ times, and the coin comes up heads $x = 2$. Naturally, this result

[14] An interval hypothesis can act as a model because it serves the right role for both Bayesian and frequentist purposes. The Bayesian can use an interval hypothesis as a general hypothesis, which is sufficient. The frequentist can use an interval hypothesis as a field for estimating point hypotheses. Although interval hypotheses as such do not receive much mainstream attention, they represent a conceptually helpful class, and are helpful for understanding the effect of model restriction.

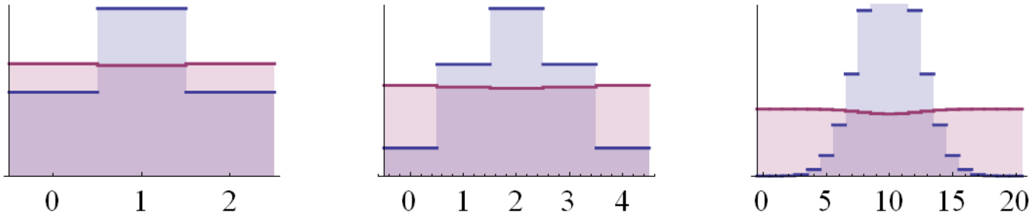


Figure 3.5: Bayesian likelihoods for both narrow and wide interval hypotheses on the bias of a coin. Likelihood distributions are over x heads for $n = \{2, 4, 20\}$ flips. The narrow interval is $0.49 \leq p \leq 0.51$, p the probability of heads. Naturally, when the narrower hypothesis fits the data (that is, when the number of heads is near 1/2 the number of total tosses), it has a higher likelihood.

supports C_1 better, with a likelihood ratio of 2:1. Then she flips the coin an additional 16 times for a total of $n = 20$, with $x = 11$ heads, supporting C_1 all the more. See Figure 3.5 for graphs, and Rosenkrantz (1977) for details.[15]

C_1 is clearly the simpler model. And this case clearly fits the Predictive Focus Account. C_1's simplicity corresponds to a narrower prediction range. The narrow range corresponds to higher confidence, likelihood, over that range. Both interval hypotheses fit the data. So C_1 is better supported.

But the AIC Account is unable to explain why C_1 should have an advantage. On the AIC Account, simpler models tend to absorb less error from the data because they have fewer adjustable parameters. But C_1 and C_2 have just one adjustable parameter, p. The intuitive difference in complexity here is not matched a higher dimension in the hypothesis space. Both models pick out a value for just one parameter—the difference them is just the range of values allowed by each. C_2 is more flexible in some sense, but not in the sense relevant to the AIC.

The AIC approach to the choice between C_1 and C_2 does not involve the use of the AIC (the formula) at all. It is to estimate coin bias within the constraints of each model. In the case where there are 2 successes out of 4, the estimates are nearly identical, 0.50 and 0.51, with nearly identical likelihoods to match. For other cases, the AIC approach will give an answer that seems intuitively wrong. For 11 successes out

[15] The derivation of the likelihoods is straightforward. Rosenkrantz integrates the binomial distribution over the intervals defined with respect to p. The result is an "average likelihood" binomial expression, which still takes one variable, successful trials.

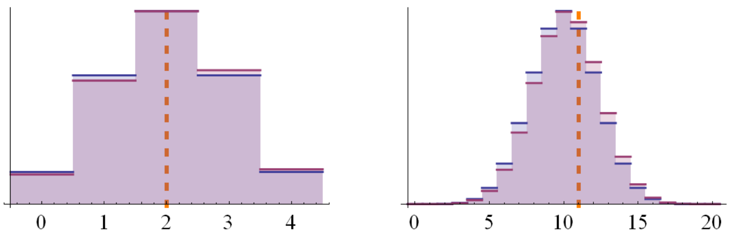


Figure 3.6: The AIC / frequentist approach to C_1 and C_2 is to compare their best fits. With 4 trials and 2 heads, the two are nearly identical, although C_1's best fit maintains a slightly higher likelihood. With 20 trials and 11 successes, C_2 does better.

of 20, the estimates are 0.51 and 0.55, with C_2, the model representing the unfair coin, fitting better. Intuitively, this is a cost. 11 heads out of 20 seems to support C_1 better than C_2.

I should admit that the treatment of statistical evidence by these two paradigms, even the meaning of the verdicts, is different. Later in this section I will suggest the best frequentist response to these critiques. A part of that response is to deny that C_1 and C_2 are tested as wholes, that they are statistically valid objects of evidential evaluation. For the frequentist, these interval hypotheses are models, purely for selecting hypotheses which may be evaluated in turn. Bayesian Theory seems to take a more intuitive approach to these models, but the frequentist may count this approach to be a tempting, misleading one.

Let us turn to a second case of model restriction: Gregor Mendel's experiment and the Law of Independent Assortment. Mendel conducted a famous series of experiments in which he cross-fertilized pea plants with different traits: round peas vs. wrinkled peas, yellow peas vs. green peas, long stems vs. short stems, etc. He found that the offspring of the crosses had one or the other trait with ratios suggestively close to 3:1. Concerning round peas to wrinkled peas, In the peas from the offspring (numbering over 7,000 peas), Mendel observed a ratio of 2.96:1, round to wrinkled.

Naturally, Mendel inferred that the true ratio was 3:1.[16] The 3:1 ratio is clearly the simplest (specific) hypothesis that fits the data. But of course, in a flat sense, the

[16] This hypothesis was corroborated both by other ratios in the first generation, and its predictive success (Mendel, 1951; Olby, 1997). Here I will ignore the other experiments and deal with the round/wrinkled trait alone.

observed ratio of 2.96:1 fits better. To analyze these hypotheses, we have to associate them with models. For an elementary treatment, let each model specify the ratio of round to wrinkled peas as $b : 1$.

- $W - b$ is a whole number.

- $R - b$ is any real number, except for those in W.

Like in the last case, both models have just one adjustable parameter, b. But W is simpler. What makes W simpler is not fewer adjustable parameters, but tighter restrictions on its parameter.Since R does not have more adjustable parameters, it is not subject to more estimation error on the AIC Account. It is not clear whether both models are subject to the same amount of estimation error, but it is hard to imagine that R is subject to more.

By contrast, the Predictive Focus Account gives a natural treatment of the two models. W places high likelihoods on the whole numbers, and through the fluctuations expected do to taking a random sample, on real numbers near those whole numbers. But R places (almost) constant likelihood across all. For a basic case, we can focus just on the segment of the b from 2.5 and 3.5. The probability distribution that describes potential observations for a given probability of success (being a round pea) on a set sample size (7,324 peas) is the binomial. This binomial is well-approximated by the normal distribution.[17] For the complex hypothesis, we assume that all frequencies are equally likely to be observed, described by a uniform distribution.[18] See Figure 3.7. The ratio of their likelihoods is roughly 4.4:1, favoring W.

When simplicity is not a matter of the number of parameters but of parameter restrictions, the AIC Account does not apply. There is an aspect of simplicity that the AIC cannot "see." This is not to say that the AIC-type frequentist could have nothing to say about these cases, but their resolution will not rest on a direct application of the

[17] However, this distribution must be transformed in order to be described in terms of ratios of $b : 1$ rather than relative frequencies, $x\%$. This introduces some skewness, but the effect is small near the mean.

[18] Although a really sensible *a priori* probability model would range over all possible observed ratios, or at least a large range of them, here let us just look at a *limiting analysis*: approximate relative probabilities that would be assumed in the more complete model.

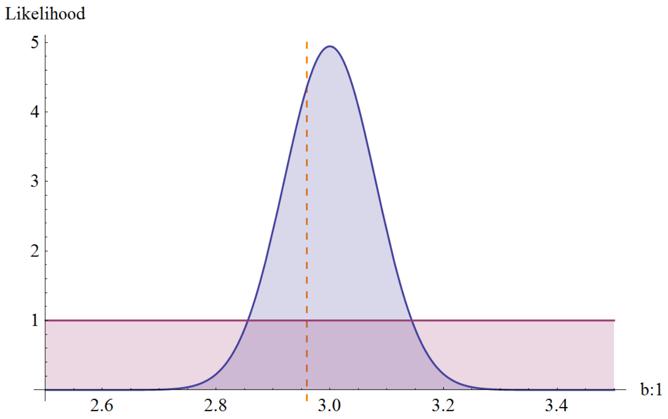


Figure 3.7: Approximate Bayesian predictive distributions for W and R. Both models are assigned uniform prior over the range $2.5 < b < 3.5$. Likelihoods are generated by the binomial distribution, and transformed from success rates as such to proportions. W places a binomial distribution only at $b = 3$. R's predictive distribution spreads out to the uniform. The evidence favors W, with a likelihood ratio of 4.4:1.

AIC. By contrast, the Predictive Focus Account is more flexible. It applies to the case of restricted models, and yields an intuitive result.

3.7 Historical Hypotheses

The AIC Account locates the advantage of simplicity in prediction. That is, hypotheses fitted from simple models tend to be more predictively accurate. But there are cases where simplicity fails to improve predictive accuracy, and yet retains its theoretical value. These are cases that the AIC Account cannot take account of, but the Predictive Focus Account can.

In general, questions about the past tend to be more a matter of truth and falsity than a matter of prediction. The history of human society, for example, may be enlightening, but it is not easy to say exactly how it informs or predictions and judgments about the future. As curious people, we care about history—where the human race originated, how English settlers colonized and expanded in the "New World"—not because we think these questions will give us much in the way of direct predictive power, but just to know the truth. And so it seems to be the case with reconstructions of the past of the natural world: the history of the cosmos, the planet and its geography, and evolutionary history. Some aspects of natural history have no impact on our predictions, but they still represent valuable theoretical pursuits where simplicity matters.

For example, think back to the case of phylogeny. I previously discussed the principle "Parsimony" in phylogeny as one that only fit with the Predictive Focus Account. Phylogeny is also a case where the best hypothesis is has limited predictive power. When we do phylogeny, we are not trying to hone in on a natural law with which we have constant contact and for which accurate estimation has a high payoff (e.g. the law of gravitation). The main principle of evolution is well understood. Instead, the project of phylogeny is "reconstructing the past." We seek to learn something that happened before, a particular evolutionary pathway for a number of organisms, and will not happen again. As a project primarily about history, not a regular natural occurrence, phylogeny does not have the same predictive power that physics has. So phylogeny challenges the AIC Account in another way, besides exhibiting an advantage for simplicity in "mere hypotheses."

For a new case, let us look at Laplace's *Nebular Hypothesis* of the formation of the solar system. Although this is a hypothesis that provides some predictive power, these predictions are outside of the domain of the evidence that supports it. The hypothesis is absolutely unnecessary for the purposes of prediction, and unhelpful for the case of resampling the data.

Pierre-Simon Laplace was a genius, and is well known in philosophy for his important philosophical and mathematical work. He was critical in the development of probability theory and the Bayesian approach to uncertainty (Stigler, 2003). His *System of the World* (originally in French, *Exposition du Système du Monde*), is a five-part treatise concerning the solar system. In the last chapter, Laplace poses a daring hypothesis about the solar system's formation. On this, the *Nebular Hypothesis*, the solar system began as a large rotating disk of interstellar gas. Over time, the sun and planets coalesced in the disk, eventually gaining enough gravity to draw in and sweep out the remaining gas.

Laplace's hypothesis was driven by several telling pieces of evidence (Laplace, 1809, 356)

1. All observed planets orbit, revolve around the sun, in the same fixed direction. (Call this direction "counterclockwise.")
2. All observed planets orbit on nearly the same plane.
3. All observed satellites (moons) revolve counterclockwise around their planets in the same direction.
4. All observed satellites orbit on nearly the same plane.
5. The sun, planets, and all satellites rotate, counterclockwise.

For Laplace, the evidence was striking because of its homogeneity, and because of the simplicity of the hypothesis that it suggested. And in particular, he contrasts evidence for the simple, universalizing hypothesis against chance. In one translation, "A phenomenon so extraordinary is not the effect of chance, it indicates an universal cause" (Laplace, 1809, 356). To flesh out the case, let me stipulate a competing hypothesis that involves a high degree of chance, the *Multiple Capture Hypothesis*. On the Multiple Capture Hypothesis, the sun, planets, and satellites all pre-existed, but were disconnected and had random velocities through space, which at the time was relatively dense with such objects. Over time, planets passed near the sun with the right trajectories and fell into orbit, and their satellites around them.

Naturally, we do not expect the nice, uniform behavior of the solar system that constitute Laplace's five key piece of evidence. But such nice, uniform behavior *is* compatible with the Multiple Capture Hypothesis. Having all planets orbit within the same 7 degrees is no less likely on Multiple capture than any other orbital orientation (with the same degree of precision). This kind of comparison is at the heart of the Predictive Focus Account: the complex model fits the data just fine, and might be plausible on a different context, but it is soundly beaten by a simpler model. The simpler model does in fact fit much better (in a Bayesian context), but not because the complex model fits *badly*—only because the complex model spreads itself too thin.

The Multiple Capture Hypothesis depends on a number of independent occurrences (each capture). Translating this general hypothesis into a model would require designating independent adjustable parameters for each bit of behavior we wanted to capture; for example, we would probably want to assign parameters for each orbital inclination. By contrast, the Nebular Hypothesis, while involving complex physical interactions, is unified. The most natural way to describe it as a model is to assign just one adjustable parameter for the orbital plane of the original nebular disk.

Laplace gives an explicit argument for the Nebular Hypothesis over chance (and by extension, the Multiple Capture Hypothesis) from three of the five pieces of evidence: points 1, 3, and 5. Call this set E. E says: all observed objects in the solar system rotating and revolving in the same direction. Laplace determines the directions of revolution for 25 objects, and rotations for 9, totaling 38 (Laplace, 1798).On the Multiple Capture Hypothesis, for rotation or revolution, the probability of a counterclockwise direction is pretty clearly $\frac{1}{2}$. So the probability of E_1 on the Multiple Capture Hypothesis is $\left(\frac{1}{2}\right)^{37}$, a tiny number. By contrast, the probability of E_1 on the Nebular Hypothesis is nearly one. The global likelihoods for each general hypothesis here are dramatically different, in accordance with the Predictive Focus Account.

But the Nebular Hypothesis adds nothing in terms of predictive accuracy over the Multiple Capture Hypothesis. Once we have established directions of rotation and revolution for the various objects in the solar system, there is no more predictive work to do. Having determined, for example Mars's path around the sun, or Jupiter's spin, the Nebular Hypothesis has nothing to add. Thus the AIC Account fails to capture the advantage of simplicity for the Nebular Hypothesis.

The reason that the Nebular Hypothesis has no predictive impact is a virtual absence of observational error. We have seen that the AIC presents a trade-off between current model accuracy and model predictive accuracy. The liability of complex models is in over-fitting, "absorbing" and encoding too much observational error into its best-fit hypothesis. But especially on E, Laplace's evidence is virtually errorless. One can observe Jupiter's direction of rotation, for example, by visually keeping track of its large red spot. This sort of observation is quite different from point data on continuous variables in noisy contexts. For this sort of observation, error is possible—human observers can always make a dreaded, inexplicable mistake—but the possibility is low enough to be safely ignored. With no observational error to speak of, the AIC Account does not apply.

Nor does the AIC favor the Nebular Hypothesis. Treating both hypotheses as models, the Nebular Hypothesis has just one adjustable parameter in the relevant context for E: the direction of rotation of the disk itself. By contrast, the Multiple Capture Hypothesis has 38 adjustable parameters, one for each direction. With virtually no error, fitting a model produces a specific hypothesis, giving a binary value to each direction. Perfect fit is quite achievable in this case, and both models achieve it. Then, with error known (and known to be negligible), the statistic reduces just to the SOS for both models. So the AIC score on both models is the same.

There are a number of nuances to the relation between the AIC and this case, and a number of ways that we might try to repair the frequentist/AIC approach in this context. For the most part, considering them will not forward the discussion, but let us consider one possible fix. We might take both hypotheses not to be models, but to be "mere hypotheses." In this case, the frequentist treatment looks like the Bayesian treatment: making probabilistic assumptions regarding each hypothesis and calculating global likelihoods. But then the AIC Account cannot consider the advantage of the Nebular Hypothesis to be a matter of simplicity as such, since it only locates this advantage between models with adjustable parameters.

Taking a step back, it is no surprise that the AIC Account does not favor the Nebular Hypothesis. The Nebular Hypothesis does nothing to inform our future expectations about the locations of planets and moons in the sky. Recent telescopic data and the laws of physics are all we need for this purpose; the history is irrelevant. The data tell so strongly for the Nebular Hypothesis because it is, to toe a Bayesian line, very probable given the data. It is would be difficult to deny that the Nebular Hypothesis is true, given the conformity of the data. And given that the Nebular Hypothesis is also much simpler than the Multiple Capture Hypothesis, simpler in the sense of being more unified and depending on fewer independent events, this appears to be a case where the Principle of Simplicity applies.

The fact that the Principle of Simplicity applies to historical hypotheses implies that simplicity is a theoretical virtue with a confirmatory mood—not predictive, as the AIC Account holds. As pertaining to matters of the past, historical hypotheses have limited import for our predictions about the future. Although an accurate picture of the past can inform our future expectations, it often contributes only in roundabout ways, and in particular, outside of the domain of the past data. Knowing the history of our evolutionary decent tells us almost nothing about what the phylogenetic tree will do next, although it enriches our understanding of evolution. Knowing the primordial state of the solar system tells us nothing about where to expect Venus, although it does tell us what to expect when we look to other star systems in formation. Historical hypotheses are unhelpful on exactly the task that the AIC is designed to evaluate: resampling the same data. It appears that our interest in the past is straightforwardly in raw truth, consonant with the Predictive Focus Account.

3.8 Conclusion

The Predictive Focus Account is the best single, all-around account of the Principle of Simplicity. It has a variety of advantages. It is closely connected to chance-probability arguments, which are rooted in the history of science, including with Laplace. In providing a rich posterior probability distribution over hypotheses, it is capable of favoring simplicity and embracing complexity at the same time. It licenses short-range extrapolations in principle, while it comes complete with a measure of predictive confidence over those extrapolations. It has a broader range of applications than the AIC. It does not require significant observational error to favor simplicity. It applies to comparisons of interval hypotheses. It provides a unified conception of the advantage of simplicity for a broader range of cases—the AIC-frequentist has to write some of these off as cases where the simpler hypothesis has an enhanced likelihood, unrelated to the advantage of simplicity. Only the Predictive Focus Account captures the critical case of the nebular disk hypothesis. These considerations favor predictive focus as the better general account of simplicity as a theoretical virtue.

Appendix A

The Hubble–Humason Data Window

The two important assumptions I make about the Hubble–Humason data window are the range of possible observations (the window as such) and observational error. I assume the window to be $[-20, 20]$ Mm/s, and observational error to be normal with standard deviation $\sigma = 0.8$ Mm/s.

The Hubble–Humason data was recorded near 1930 from the Mount Wilson Observatory—a then state-of-the-art astronomical observatory built in the clear skies of Southern California. The observatory featured two large reflectors[1], a 60-inch and a 100-inch, and had two spectrographs with multiple cameras for recording the redshifts of nebulae. Hubble used brightness to estimate the distances of the galaxies, and redshifts to calculate recessional velocities through the Doppler-like behavior of light.

Distance should be a random element by itself, since the astronomers could not previously determine at what distances they would find nebulae. However, since the hypothesis in question is a matter of the dependence of velocity on distance, nebula distances do not play a (significant) confirmatory role by themselves. And the prediction spans of the various models for velocity are comparable whether or not the distance vector X is treated as random. The main effect that X has on ranges seems to be in the clustering of x-points. If x-points cluster together, complex models have less ability to accommodate them independently, and complex model prediction spans suffer more.

Astronomical spectrography is complicated, and determining the proper error profile for the Mount Wilson equipment in 1931 would be a major project on its own. My

[1] Reflectors are mirrors that are shaped to act the same way that glass lenses do. It would not be possible to make and support a smooth glass lens 100-inches in diameter. But by bouncing light *back* instead of channeling it through, a reflector can be curved to have the same focusing effect, and be made out of lighter and more robust material. In addition, a reflector avoids the problem of chromatic aberration that with glass lenses. We use the same design principle today in satellite dishes.

treatment is designed to be representative based on information, and defaults to velocity as the native parametrization of the experiment.

Mount Wilson featured a blazed-grating spectrograph, on which dispersion is a linear function of the wavelength of the light involved. Alone, this fact implies that true native parametrization would be a linearly decreasing function of recessional velocity. That is to say, the data window would be more squished and less precise at higher velocities. So in this way, my supposition about error would not be representative.

The error profile for velocity is further complicated by the fact that one can sacrifice spectral dispersion to gain more light intensity across a shorter spectrum by using cameras with different focal lengths, and the blazed-grating spectrograph at Mount Wilson had three (Hearnshaw, 2009, 33). This factor suggests something of a patchwork data window, with each camera responsible for a different region of the whole.

The distance of the object also contributes to spectrographic error. If the light is faint, it will not support a high-dispersion spectrogram. Thus there is a complicated interplay between the independent and dependent variables in this experiment—error in velocity is also a function of distance.

So there are many factors that should be taken into account for a precise predictive focus calculation. But for a first pass at a probability model, I ignore special issues with spectroscopy and treat error mainly as a function of noise, random motion. Table IV of Hubble & Humason (1931) shows that noise velocity is substantial because there is relatively large variation among nearby objects: Virgo, and the isolated "nebula" groups.[2] I choose the second of the isolated nebula groups, stars and nebulae with photographic magnitudes available (p. 58), to estimate noise velocity: the estimate is $\sigma_*^2 = 1.9$ Mm/s, $\sigma_* \approx 1.4$, for individual stars. However, this estimate is incomplete for Hubble's objects, because it ignores the sampling distribution.

Hubble averages star velocities within each cosmic object. This means that error will tend to cancel out, in a way that is well understood in statistical theory. Sample

[2] "Nebula" here simply means cosmic object, especially galaxies and galaxy clusters. It is a blank early astronomical designation, a little in the spirit of "UFO" for aircraft. The term has shifted in meaning—now it refers to coherent gas clouds in space.

sizes for each object range from 9 (the isolated nebula group I use as a base) to 1 (the farthest objects, Ursa Majoris and Leo). My computational resources are limited, so I need to find a single representative value for σ. I choose a flat average: assuming $\sigma_* \approx 1.4$ for individual stars, average error for collective objects is $\sigma \approx 0.8$. This fits with a standard sample size of 3.

Perhaps my most liberal assumption is to make the data window symmetric about 0 Mm/s. The data window is at least as high as 20 Mm/s recessional, so I assume that it is as low as -20 Mm/s recessional, or 20 Mm/s approaching. In other words, I take a relatively minimal data window, but impose symmetry about 0 velocity (relative to earth).

There is some reason to think that this assumption is too generous for possible blueshifts. Hubble and Humason depended critically on the Calcium$^+$ "H" and "K" spectral lines (for example, see Hubble & Humason (1931, 69)). We are lucky that these lines exist. They are at the very edge of the blue side of the spectrum, and easy to identify. Even as the rest of a spectrogram may blur together and become completely indistinct, these lines can remain uncannily recognizable (Sandage, 1994, 510).

A more conservative estimate for the velocity data window might be the compromise [-10, 20]. In this case, H1's share of the total likelihood drops from 94% to 90%. Cutting the data window even narrower will result in even lower degrees of simplicity favoring. As a gloss, this is to be expected, since the advantage of the simple model is its narrow predictive focus. But the data window shrinks, there is a smaller range of possible observations for the simplest model to accommodate the smallest proportion of.

In this section, I have pointed to a number of ways in which my treatment oversimplifies the real Data Window Prior for Hubble's experiment. These subtleties either introduce too much computational complexity to be worth taking account of, or require information that is not readily accessible, or both. But the calculation here should serve as a good pass on the advantage of the simplest model on the DWP.

Bibliography

Ackerman, Robert. 1966. Inductive Simplicity. *Pages 322–31 of:* Foster, Marguerite H., & Martin, Michael L. (eds), *Probability, Confirmation, and Simplicity.* New York: The Odyssey Press.

Akaike, Hirotugu. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **AC-19**(6).

Barker, Stephen F. 1966. On Simplicity in Empirical Hypotheses. *In:* Foster, Marguerite H., & Martin, Michael L. (eds), *Probability, Confirmation, and Simplicity.* New York: Odyssey Press.

Carnap, Rudolf. 1962. *Logical Foundations of Probability.* Chicago: University of Chicago.

Dowe, David L., Gardner, Steve, & Oppy, Graham. 2007. Bayes Not Bust! Why Simplicity Is No Problem for Bayesians. *British Journal of the Philosophy of Science*, **58**, 709–754.

Earman, John. 1992. *Bayes Or Bust: A Critical Examination of Bayesian Confirmation Theory.* Cambridge: MIT.

Fitelson, Branden. 2007. Likelihoodism, Bayesianism, and Relational Confirmation. *Synthese*, **156**(3), 473–489.

Fitelson, Branden. 2012. Contrastive Bayesianism. *In:* Blaauw, Martijn (ed), *Contrastivism in Philosophy.* New York: Routledge.

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

FORSTER, MALCOLM. 2001. The New Science of Simplicity. *Pages 83–119 of:* ZELLNER, ARNOLD, & KEUZENKAMP, HUGO A. (eds), *Simplicity, Inference, and Modelling.* New York: Cambridge University.

FORSTER, MALCOLM. 2002. Predictive Accuracy as an Achievable Goal of Science. *Philosophy of Science*, **69**, S125–34.

FORSTER, MALCOLM, & 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.

FREEMAN, SCOTT, & HERRON, JON C. 2007. *Evolutionary Analysis.* 4 edn. Boston: Pearson Benjamin Cummings.

GLYMOUR, CLARK. 1980. *Theory and Evidence.* Princeton: Princeton University.

GOOD, I.J. 1968. Corroboration, Explanation, Evolving Probability, Simplicity, and a Sharpened Razor. *British Journal for the Philosophy of Science*, **19**(2), 123–43.

GOODMAN, NELSON. 1959. Recent Developments in a Theory of Simplicity. *Philosophy and Phenomenological Research*, **19**, 429–46.

GOODMAN, NELSON. 1983. *Fact, Fiction, and Forecast.* Cambridge: Harvard University Press.

HASTIE, TREVOR, TIBSHIRANI, ROBERT, & FRIEDMAN, JEROME. 2008. *Elements of Statistical Learning.* 2 edn. Springer-Verlag.

HEARNSHAW, JOHN. 2009. *Astronomical Spectrographs and Their History.* Cambridge: Cambridge University.

HEMPEL, CARL. 1998. Criteria of Confirmation and Acceptability. *In:* CURD, MARTIN, & COVER, J.A. (eds), *Philosophy of Science: The Central Issues.* New York: W.W. Norton and Company.

HUBBLE, EDWIN. 1929. A Relation Between Distance and Radial Velocity Among Extra-Galactic Nebulae. *Proceedings of the National Academy of Sciences*, **15**(3), 168–173.

HUBBLE, EDWIN, & HUMASON, MILTON. 1931. The Velocity-Distance Relation Among Extra-Galactic Nebuale. *Astrophysical Journal*, **74**, 43.

JEFFREYS, HAROLD. 1988. *A Theory of Probability.* New York: Oxford University.

JOYCE, JAMES. 1998. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, **65**(4), 575–603.

JOYCE, JAMES. 2005. Probabilities Reflect Evidence. *Philosophical Perspectives*, **19: Epistemology**.

KASS, ROBERT E., & WASSERMAN, LARRY. 1996. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**(435), 1343–1370.

KELLY, KEVIN T. 2007. *How Simplicity Helps You Find the Truth Without Pointing at It.* Carnegie Mellon University Showcase, Department of Philosophy, Paper 381.

KUKLA, ANDRÈ. 1995. Forster and Sober on the Curve-Fitting Problem. *The British Journal for the Philosophy of Science*, **46**, 248–52.

LAKATOS, IMRE. 1998. Science and Pseudoscience. *In:* CURD, MARTINMARTIN, & COVER, J. A. (eds), *Philosophy of Science The Central Issues.* New York: W.W. Norton and Company.

LAPLACE, PIERRE-SIMON. 1798. *Exposition du Système du Monde.* Accessed: 2014-07-01.

LAPLACE, PIERRE-SIMON. 1809. *The System of the World / Translated from the French.* Old Bailey: w. Flint. Accessed: 2014-07-01.

LEWIS, DAVID. 1983. New Work for a Theory of Universals. *Australiasian Journal of Philosophy*, **61**, 343–377.

Loewer, Barry. 1996. Humean Supervenience. *Philosophical Topics*, **24**, 101–27.

Loewer, Barry, Laddaga, Robert, & Rosenkrantz, Roger. 1978. On the Likelihood Principle and a Supposed Antinomy. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **1**, 279–286.

Maxwell, Grover. 1998. The Ontological Status of Theoretical Entities. *In:* Curd, Martin, & Cover, J.A. (eds), *Philosophy of Science: The Central Issues.* New York: W.W. Norton & Company.

McAllister, James W. 1991. The Simplicity of Theories: Its Degree and Form. *Journal for General Philosophy of Science*, **22**, 1–14.

McAllister, James W. 2007. Model Selection and the Multiplicity of Patterns in Empirical Data. *Philosophy of Science*, **74**(December), 884–94.

Mendel, Gregor. 1951. Experiments in Plant Hybridization. *Journal of Heredity*, **42**(1).

Olby, Robert C. 1997. *Mendel, Mendelism and Genetics.*

Popper, Karl R. 1992. *The Logic of Scientific Discovery.* London: Routledge.

Posada, David, & Buckley, Thomas R. 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike InfoInformation Criterion and Bayesian ApApproach over Likelihood Ratio Tests. *Systematic Biology*, **53**(5), 793–808.

Priest, Graham. 1976. Discussion: Gruesome Simplicity. *Philosophy of Science*, **43**(3), 432–37.

Quine, W. V., & Ullian, J. S. 1978. *The Web of Belief.* New York: Random House.

Raftery, Adrian E., & et. al. 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**(5), 1155–74.

Reichenbach, Hans. 1947. *Experience and Prediction.* Chicago: University of Chicago.

ROSENKRANTZ, ROGER D. 1977. *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*. Dordrecht: D. Reidel.

ROYALL, RICHARD M. 1997. *Statistical Evidence: A Likelihood Paradigm*. 71 edn. Monographs on Statistics & Applied Probability. New York: Chapman & Hall/CRC.

RUSSELL, BERTRAND. 2009. *Human Knowledge: Its Scope and Limits*. Taylor & Francis Routledge.

SALMON, WESLEY C. 1988. Rationality and Objectivity in Science *or* Tom Kuhn Meets Tom Bayes. *In:* CURD, MARTIN, & COVER, J.A. (eds), *Philosophy of Science: The Central Issues*. New York: W.W. Norton & Company.

SANDAGE, ALLAN. 1994. *The Mount Wilson Observatory: Breaking the Code of Cosmic Evolution*. Centennial History of the Carnegie Institution of Washington, vol. 1. New York: Cambridge University.

SCHWARZ, GIDEON. 1978. Estimating the Dimension of a Model. *Annals of Statistics*, **6**(2), 461–4.

SEIDENFELD, TEDDY. 1979. Why I Am Not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz. *Theory and Decision*, **11**(4), 413–40.

SMITH, A. F. M., & SPIEGELHALTER, D. J. 1980. Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society*, **42**, 213–20.

SOBER, ELLIOTT. 1975. *Simplicity*. Oxford: Clarendon Press.

SOBER, ELLIOTT. 1991. *Reconstructing the Past*. Bradford Books.

SOBER, ELLIOTT. 1994a. *From a Biological Point of View*. Cambridge: Cambridge University. Chap. Let's Razor Ockham's Razor, pages 136–57.

SOBER, ELLIOTT. 1994b. No Model, No Inference. *In:* STALNAKER, DOUGLASS (ed), *Grue! The New Riddle of Induction*. Chicago: Open Court.

Sober, Elliott. 2001. What is the Problem of Simplicity? *In:* Zellner, Arnold, Keuzenkamp, Hugo A., & McAleer, Michael (eds), *Simplicity, Inference and Modelling.* Cambridge: Cambridge University Press.

Sober, Elliott. 2002. Instrumentalism, Parsimony, and the Akaike Framework. *Philosophy of Science*, **69**, S112 – 23.

Stigler, Stephen M. 2003. *The History of Statistics: the Measurement of Uncertainty before 1900.* Cambridge: Belknap Press of Harvard University.

Swinburne, Richard. 1997. *Simplicity as Evidence of Truth.* Milwaukee: Marquette University Press.

Trotta, Roberto. 2008. Bayes in the Sky: Bayesian Inference and Model Selection in Cosmology. *Contemporary Physics*, **49**, 71–104.

van Fraassen, Bas. 1980. *The Scientific Image.* Oxford: Clarendon.

van Fraassen, Bas. 1990. *Laws and Symmetry.* Oxford: Clarendon.

Vitányi, Paul, & Li, Ming. 2009. Simplicity, Information, Kolmogorov Complexity, and Prediction. *In:* Zellner, Arnold, Keuzenkamp, Hugo A., & McAleer, Michael (eds), *Simplicity, Inference, and Modelling.* New York: Cambridge University.

Wasserman, Larry. 2000. Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

Williamson, Jon. 2010. *In Defense of Objective Bayesianism.* Oxford: Oxford University.