Bayesian Statistical Inference and Approximate Truth

Olav B. Vassend

October 28, 2016

Abstract

Scientists and Bayesian statisticians often study hypotheses that they know to be false. This creates an interpretive problem because the Bayesian probability of a hypothesis is supposed to represent the probability that the hypothesis is true. I investigate whether Bayesianism can accommodate the idea that false hypotheses are sometimes approximately true or that some hypotheses or models can be closer to the truth than others. I argue that the idea that some hypotheses are approximately true in an absolute sense is hard to square with Bayesianism, but that the notion that some hypotheses are comparatively closer to the truth than others can be made compatible with Bayesianism, and that this provides an adequate and potentially useful solution to the interpretive problem. Finally, I compare my "verisimilitude" solution to the interpretive problem with a "counterfactual" solution recently proposed by Jan Sprenger.

Contents

1 Introduction

According to the standard Bayesian interpretation of probability, the probability of a hypothesis is the probability that the hypothesis is *true*. However, scientists, including scientists who make use of Bayesian statistical methods, often investigate models and hypotheses that they know to be false. In particular, statistical models tend to be constructed on the basis of auxiliary assumptions (e.g. normality and independence of measurement errors) that are often known to be false. Moreover, statistical analysis is often restricted to hypothesis sets, such as the set of linear or exponential functional relationships, that are known to at best be (false) approximations of the actual functional relationships. Presumably, if something is known to be false, then it has a probability of 0 of being true, so all of the preceding practices are hard to reconcile with the standard Bayesian interpretation of probability. Indeed, Bayesian statistical practice apparently is faced with an interpretive problem: on

the one hand, Bayesian probabilities are standardly interpreted as probabilities of truth; on the other hand, Bayesian scientists routinely assign non-zero probabilities to hypotheses they know to be false.

How serious is the interpretive problem and how may it be solved? I argue that there are many cases where the interpretative problem does not arise, even when the statistical model is false. But there are also many cases where the interpretive problem does arise. Many scientific realists have suggested that successful scientific models and hypotheses, though usually false, are nonetheless often approximately true, or – at the very least – that successful hypotheses in general are "closer to the truth" (or have higher "verisimilitude") than hypotheses that are less successful. I argue that, provided we jettison the standard Bayesian interpretation of the probability axioms, Bayesianism can accommodate the insight that some false hypotheses are closer to the truth than others, and that this reinterpretation of the probability axioms is potentially useful. I contrast this solution to the interpretive problem with another recent proposal due to Jan Sprenger (2016), according to which probabilities of false hypotheses are interpreted as "counterfactual degrees of belief," and I argue that the two approaches – when spelled out in detail – are formally inter-translatable and help illuminate each other.

2 The Basics of Standard Bayesian Inference

Bayesianism is a prominent approach in both confirmation theory and in statistical inference. Bayesian confirmation theory and Bayesian statistics clearly have many things in common, but they are also different enough that it pays to discuss them separately. In this paper, I will focus my attention on Bayesian statistical inference, though much of what I will say also has relevance to Bayesian confirmation theory.

In statistical inference, a set of competing hypotheses is usually indexed by a *parameter*, which in general will be a real-valued variable or a vector of real-valued variables. Given a space of candidate hypotheses parameterized by Θ, and given some particular context in which the possible observations or outcomes are x_1, x_2, etc. – or X, for short – a *statistical model* consists of a set of conditional probability

(density) distributions, $p(x|\theta)$, that jointly specify the probability of each possible $x \in X$ given each possible $\theta \in \Theta$.[1]

Almost invariably, the statistical model is premised on various auxiliary assumptions, A, that jointly guarantee that each value of θ *entails* a probability for each x. Sometimes A itself has free parameters – so-called "nuisance parameters," N – that must also be estimated from the data, in which case the conditional probability distributions will be of the form $p(x|\theta\&n)$. Thus, a statistical model may in general be regarded as being composed of two distinct ingredients: the hypotheses of interest, parameterized by Θ; and the auxiliary assumptions, A, consisting of nuisance parameters, N, and background assumptions, B. It follows that a statistical model is "true" if and only if the following conjunction is true: (1) some element of Θ is true and (2) A is true: that is, B is true and some element of N is true.

For example, suppose you are interested in estimating the mass of some object by measuring it a single time using a scale. The hypotheses of interest are the various possible masses of the object, which you may index using a real-valued parameter, m. The possible outcomes, x, are the various possible outcomes of the measurement. In order to probabilistically link m to x, you may, for example, add the auxiliary assumption, A, that the measurement outcome is normally distributed around the true mass with a variance of d. Here d is a nuisance parameter. Then the assumptions of the statistical model generate the following conditional probabilities:

$$p(x|m\&d) = \frac{1}{d\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-m)^2}{2d^2}} \tag{2.1}$$

In this case, the statistical model is true if and only if (1) there is some value m_0 of m that corresponds to the actual mass of the object, and (2) the measurement outcome is actually normally distributed around m_0 with some variance d_0.

What distinguishes Bayesian inference from other sorts of statistical inference is that Bayesians use probability distributions to assess the plausibility of *parameter values*. In addition to requiring a statistical model, a Bayesian analysis therefore

[1]Note: p is a *probability function* over the set $\mathbf{X}$ if and only if the following three axioms are satisfied: (1) $p(\mathbf{X}) = 1$. (2) $p(x_i) \geq 0$ for all $x_i \in \mathbf{X}$. (3) $p(\bigvee x_i) = \sum p(x_i)$, whenever the x_i in the disjunction are mutually exclusive.

requires that the parameters of interest Θ and the nuisance parameters N all be assigned so-called *prior* probabilities; these are probabilities that are assigned before the observation of data. Moreover, if there are multiple candidate statistical models, then all of the models must be assigned prior probabilities as well. In the above example, prior probabilities must therefore be assigned to each possible value of m and to each possible value of d. Once these probabilities have been assigned, the *joint* distribution of the possible observations and the parameters is defined as the product of the likelihood and the prior: $p(x, m, d) = p(x|m, d) * p(m, d)$. The *posterior* probability distribution of m is given by Bayes's theorem, $p(m|x, d) = p(x|m, d) * p(m, d)/p(x, d)$

There is disagreement among Bayesians concerning how prior and posterior probabilities should be interpreted. Some see these probabilities as the subjective or rational degrees of belief of some agent, whereas others interpret them as evidential degrees of support or as representing an objective state of information. However, regardless of whichever more specific interpretation they endorse, Bayesians of all kinds agree that $p(\theta)$ represents the probability that θ is *true*.[2] This interpretation of probability – the standard Bayesian interpretation – leads to problems, however, because the models and hypotheses that scientists investigate are often believed or even *known* not to be true. This problem has not gone completely unnoticed in the philosophical literature,[3] but in general the seriousness of the problem seems not to have been appreciated. The problem seems to be more acknowledged in the statistical literature, but no satisfactory resolution has been offered.

3 The Interpretive Problem in Bayesian Statistical Inference

Statistical models, much like other models in science, contain idealizations and approximations that render the models strictly speaking false. Typical examples in-

[2]Or more precisely, the probability that the hypothesis indexed by θ is true.

[3]E.g. Forster and Sober (1994), Shaffer (2001), and more recently Sprenger (2016).

clude, e.g., the assumption that measurement error is bell-shaped, or that measurements are independent and identically distributed. To be sure, these assumptions are often justified because they hold *approximately*, but they rarely hold *exactly*. In other words, the auxiliary assumptions of statistical models are generally false.

For these reasons, the statistician George Box famously said that "all models are false, but some models are useful."[4] More recently, Andrew Gelman and Cosma Shalizi write, "To reiterate, it is hard to claim that the prior distributions used in applied work represent statisticians' states of knowledge and belief before examining their data, if only because most statisticians do not believe their models are true, so their prior degree of belief in all of Θ is not 1 but 0." (Gelman and Shalizi, 2013, p. 19).

A fully Bayesian analysis requires that we assign probabilities to our models and to the parameters inside the models. But according to the standard Bayesian interpretation of probability, the probabilities we assign are supposed to represent the probabilities that the models and parameters are true. If we know that they are all false, it would seem they should therefore be assigned a probability of 0.

Of course, Bayesian statisticians typically do not assign probabilities of 0 to parameters or to models; they assign non-zero probabilities. This practice is what leads to the interpretive problem, which may be phrased in the form of a question: what does it mean to assign a model or hypothesis that is known to be false a non-zero probability? To more precisely diagnose the problem, it helps to state the probability axioms with the standard Bayesian interpretation made explicit:

> Suppose $\mathbf{H}$ is a set of hypotheses $\{H_1, H_2, \ldots, H_n\}$. Then
>
> 1S. $p(\mathbf{H}) = 1$. Interpretation: one of the hypotheses in $\mathbf{H}$ is true.
>
> 2S. $p(H_i) \geq 0$ for all $H_i \in \mathbf{H}$. Interpretation: no hypothesis has a negative probability of being true.
>
> 3S. $p(\bigvee H_i) = \sum p(H_i)$, whenever it is impossible for more than one H_i in the disjunction $\bigvee H_i$ to be true.

[4]This quote is famous enough that it has a Wikipedia page. Box repeated the quote, or variations of it, in several places. e.g. Box (1980)

Here we can see that the interpretive problem is really a problem with the standard interpretation of the first probability axiom. That is, for many of the hypothesis sets that scientists study, it will not be the case that one of the hypotheses is true. Hence, strictly speaking, many hypothesis sets will not satisfy axiom 1S. Axioms 2S and 3S, on the other hand, will generally be satisfied by the kinds of hypothesis sets that Bayesian statisticians study.

One possible remedy to the interpretive problem that might initially seem attractice is to try to change the algebra over which the probability function p ranges.[5] Later, we shall consider a couple of specific proposals along these lines. However, there is a fundamental reason why any such proposal will not work. Briefly, the reason is that if you want to do Bayesian inference on a statistical model that is parameterized by θ, then you need to assign probabilities to θ; you cannot instead assign probabilities to, e.g., propositions of the sort $<\theta = 2$ is the best parameter value$>$ or $<\theta = 2$ is the parameter value that is most predictively accurate$>$, because these propositions are not part of the statistical model. Nor can you amend the statistical model so that it is instead parameterized by these other propositions.

Gelman and Shalizi's (2013) solution to the interpretive problem (to the extent that they see it as a problem) seems to be to refuse to interpret Bayesian probabilities in any standard way. Bayesian probabilities of parameters inside models, they say, are "regularization devices" and models themselves should not really be assigned probabilities at all. This does not seem like a solution so much as an admission of defeat. Morey et al. (2013) pursue a different strategy. They reply to Gelman and Shalizi with the assertion that "...scientific models, including statistical models, are neither true nor false" (p. 71) and that "Box's (1979) famous dictum... ...could be shortened to 'some models are useful' without any loss" (p. 71). They then recommend assigning odds rather than probabilities to models because a "Bayesian who employs odds is silent on whether or not she is in possession of the true model, and, in fact, need not acknowledge the existence of a true model at all" (p. 71). It is, however, unclear how using odds rather than probabilities is supposed to solve

[5]For example, some might be tempted to consider the algebra generated by the associated propositions, $<H_i$ is the best hypothesis$>$, for each H_i, or something similar.

the interpretive problem. And it is not clear how refusing to assign truth values to models solves the problem either. What does it mean to say that your odds are 5 to 1 in a model that is neither true nor false as against another model that is also neither true nor false? The interpretive problem seems to be just as severe here as before.

Moreover, the claim that statistical models do not have truth values seems wrong. As we saw, a statistical model can be regarded as a conjunction of a claim about the hypotheses of interest (namely that one of them is true) and a claim about the auxiliary assumptions (namely that they are all true). It follows that a statistical model is false either if none of the hypotheses of interest is true or if one of the auxiliary assumptions is false. The second situation arguably is less serious than the first.

4 False Auxiliary Assumptions vs False Hypotheses of Interest

If a statistical model is false because one of its auxiliary assumptions is false – which is almost always the case – then the interpretive problem arises on the level of model inference. That is, if there are multiple statistical models that all contain known false auxiliary assumptions, then all of the models will have a probability of 0 of being true, and hence a standard Bayesian who wants to use Bayesian inference to find the best model will run into the problem of how to sensibly assign non-zero probabilities to the models.

How to make sense of model inference and model selection is therefore a serious problem for Bayesians. However, if the statistical model is not itself the hypothesis of interest, then the fact that the statistical model is false does not necessarily mean we are faced with the interpretive problem.

Consider, for example, the previous example involving the estimation of the mass m of some object. A good way of getting an estimate of m is by embedding m in a statistical model. Now, even if the statistical model is false because it is based

on known false (auxiliary) assumptions, probabilistic statements about m will still be completely sensible; thus, in cases like this one, the interpretive problem does not arise for inferences about the parameter m. For example, a statement like "the probability that it's true that m is 2kg is 0.5" is perfectly sensible as long as it is remembered that the probability is premised on the auxiliary assumptions of the model. If those assumptions are seriously wrong, the probability may well be inaccurate or misleading; however, the probability can still sensibly be interpreted as a probability of truth.

Because Bayesian parameter inference often makes sense even if the statistical model is false, George Box famously recommended a reconciliation between Bayesian and frequentism. According to Box, frequentist methods should be used to identify a "useful" (albeit false) statistical model; Bayesian inference can then be used to infer plausible parameter values inside the assumed statistical model. This two-step procedure makes sense in cases where the hypotheses of interest are parameters that represent real quantities out in the world, such as for example the mass of an object.

However, it happens not infrequently in science that the hypotheses of interest are themselves known to be false, strictly speaking; but this has not stopped scientists from employing Bayesian methods in their research. For example, phylogeneticists in both biology and linguistics use trees to represent family relationships between species or between languages. In both cases, the trees investigated omit known relationships and introduce false idealizations. For example, a tree phylogeny for a language family is premised on the (false) idea that languages bifurcate instantaneously and are forever separated thereafter. Yet, even though all phylogenetic trees are clearly false, Bayesian phylogeneticists are often interested in discovering which tree has the highest posterior probability. These probabilities cannot comfortably be interpreted as probabilities that the trees are literally true, and thus we are faced with the interpretive problem.

The interpretive problem also arises whenever the hypotheses under consideration posit simple functional relationships that are almost certainly false idealizations. This is usually the case whenever Bayesian linear regression is used, for example, because most functional relationships in the world are not actually linear.

As an example, suppose you are interested in the functional relationship between just two variables, X and Y. For concreteness, suppose X represents some measurement of a complex system, e.g. the barometric pressure of a weather system, and Y represents some quantity of interest, e.g. how much it will rain in the next hour. The true functional dependence of Y on X is in all likelihood very complex.[6] Nonetheless, it is very common in such cases to restrict attention to classes of simple functional relationships, such as the set of linear hypotheses with 0 intercept, which models the relationship between Y and X as follows:

$$Y = \alpha X + \epsilon \tag{4.1}$$

Here, ϵ represents the (hypothesized) random fluctuation around the linear function $Y = \alpha X$; ϵ is generally taken to be a normal distribution with a mean of 0 and standard deviation d. α is the parameter of interest while d is a nuisance parameter (auxiliary assumption); both need to be estimated from data. Note that α does not represent some "real" quantity out there in the world; indeed, if we were to interpret α as representing a real quantity, then presumably that quantity would be a rate. Thus, α would refer to the constant rate at which Y changes (on average) given changes in X. However, if the true functional relationship between X and Y is not actually linear, then there is no constant rate at which Y changes in response to changes in X. Thus, in sharp contrast to the previous example concerning the estimation of the mass of an object, $\alpha = 2$ cannot be true or false in the same way that statements such as $m = 2$ or $m = 3$ are true or false.

But if α does not represent a quantity in the world, then what does it mean for a given value of α to be "true" or "false"? Well, α indexes a set of hypotheses, namely $Y = \alpha X + \epsilon$, so to say that α_0 is "true" in this case is the same as saying that there *exists* some value of ϵ such that the hypothesis $Y = \alpha_o X + \epsilon$ is the true functional relationship between X and Y. To paraphrase Sober (2015), α "lives inside" its

[6]By "the true functional dependence," I mean the functional dependence that would result if we were to keep fixed all other predictively relevant variables and see how Y varies given changes in X. Since X may – indeed probably does – interact with other variables, this definition is too simplistic, but going into the details here is not worth the pay-off.

model; it has a meaning only in the context of the statistical model of which it is a part. Not all parameters are created equal.

Note that in this example, it is pretty much a foregone conclusion that *no* hypothesis of the form $Y = \alpha X + \epsilon$ describes the true functional relationship between Y (i.e. how much it will rain) and X (the barometric pressure). Hence it's not merely the auxiliary assumptions of the model that are false in this case; the very hypotheses that we are interested in are all known in advance to be false. Hence, the interpretive problem hits us again with full force: how are we supposed to understand non-zero probability assignments to values of α?

5 Approximate Truth

This is where the notion of approximate truth may be helpful. More generally, scientific realists would doubt whether any scientific or statistical model could be "useful" (to use Box's term) were it not approximately true in some sense; thus, we should assign a model (or a parameter inside a model) a probability proportional to the extent to which we find it approximately true (in the relevant sense). The question we need to ask is whether and how the idea that hypotheses and models are sometimes approximately true or that some hypotheses are closer to the truth than others can be accommodated within the Bayesian framework. Because model inference and parameter inference are different in some important ways, I will from now on focus only on parameter inference. That is, I will assume that the hypotheses of interest are indexed by a parameter Θ inside some fixed statistical model, and that each $\theta \in \Theta$ picks out some hypothesis that does not itself contain adjustable parameters.

Before we can address properly the question whether some hypotheses can be approximately true or closer to the truth than others, we must make a few assumptions about what approximate truth is and how it can be measured.

The study of approximate truth was initiated by Popper (1963) and has by now

accumulated a large literature.[7] The most influential contemporary approach in the study of approximate truth – known in the literature as the "similarity approach" – takes seriously the idea that approximate truth is a particular kind of approximation. To say that something is a good approximation of something else is to say that the two things are similar in some relevant respect. Thus, to say that a hypothesis or is approximately true is to say that the hypothesis is sufficiently similar to the true hypothesis.

This idea can be formalized if we suppose that there is a (context-appropriate[8]) verisimilitude measure, v, that takes as its input a hypothesis θ and has as its output some real number that represents how similar θ is to the truth. If we presume that such functions are available, we can say that θ is approximately true just in case $v(\theta) < \epsilon$, for some suitably chosen ϵ. There are certain requirements that the verisimilitude measure arguably ought to obey. For example, it arguably ought to be non-negative, and it is also natural to demand that it be continuous whenever the hypothesis space is indexed by a real-valued parameter.

As a concrete example, one non-negative and continuous divergence measure that has been suggested as a verisimilitude measure in a statistical context is the Kullback-Leibler divergence (Forster and Sober, 1994). Supposing that q is the "true" probability distribution that governs the distribution of the data, then the verisimilitude (according to the K-L divergence) of some hypothesis θ (that does not contain adjustable parameters) is $KL(\theta) = -\int q(x) \log \frac{q(x)}{p(x|\theta)} dx$.

Unfortunately, the various ways one might try to accommodate approximate truth within the Bayesian framework face a severe difficulty having to do with the third probability axiom. Briefly, the problem is that, given a set of hypotheses indexed by a parameter, there will generally be multiple parameter values that meet any verisimilitude threshold we set for "approximate truth." Hence, the different parameter values will not be mutually incompatible in the sense that it will be possible for several of them to be approximately true simultaneously. However, Bayesian infer-

[7] See Niiniluoto (1998) for a survey.

[8] In general I agree with Northcott (2013) that there is little reason to assume a priori that there will be a single distance measure that appropriately measures approximate truth in all contexts.

ence requires that the different parameter values be mutually incompatible. Thus, Bayesian inference will in general be impossible if we change the goal of inference from truth to approximate truth. For a more thorough discussion of these issues, and how exactly a conflict with the third probability axiom is to blame, see the appendix.

The underlying problem is that approximate truth is too coarse-grained a concept since it fails to distinguish between several hypotheses, all of which are approximately true. This problem should motivate us to look for an alternative solution to the interpretive problem.

6 The Verisimilitude Interpretation of Probability

Presumably some hypotheses that are approximately true are closer to the truth than other ones, and – at least in many cases – one of the hypotheses under consideration will be closer to the truth than all the others. This suggests a different interpretation of probability. In particular, it is tempting to interpret $p(\theta)$ as the probability that θ is *closest to the truth* out of the hypotheses in Θ; note that in contrast to both truth and approximate truth, closeness to the truth is fundamentally a comparative notion. I will call this interpretation the "verisimilitude interpretation" of probability, and I will use p_c with a c subscript whenever this is the intended interpretation. It is helpful to write out all of the probability axioms with the new interpretation made explicit:

> 1C. $p_c(\Theta) = 1$. Interpretation: one of the hypotheses in Θ is closest to the truth.
>
> 2C. $p_c(\theta) \geq 0$ for all θ. Interpretation: no hypothesis has a negative probability of being closest to the truth.
>
> 3C. $p_c(\bigvee \theta_i) = \sum p_c(\theta_i)$, whenever it is impossible for more than one θ_i to be closest to the truth.

There are several things to note here. First, and most importantly, just about *any* set of hypotheses will satisfy the verisimilitude interpretation of the probability

axioms. More precisely, given any set of hypotheses that can be compared using some verisimilitude measure, at least one of the hypotheses must be maximally close to the truth according to the verisimilitude measure, so the set of hypotheses will satisfy 1C. Hence, the verisimilitude interpretation avoids the interpretive problem of the standard interpretation, which we saw was really a problem with the first axiom.

The verisimilitude interpretation also avoids the problems with the third probability axiom that we identified with the approximate truth approach. In order for Bayesian inference to be possible on the set of hypotheses, the hypotheses must be mutually incompatible in the sense of 3C; that is, it must be impossible for more than one of the hypotheses to be closest to the truth. This axiom will not always be satisfied. For example, if the hypotheses are models and some of the models are contained in others, it may be possible for several of the models to be equally close to the truth, depending on the verisimilitude measure. However, most of the hypothesis sets that Bayesian statisticians study will satisfy 3C.

Another important thing to note is that, under the verisimilitude interpretation, the probability of a hypothesis is always relative to the set of competing hypotheses under consideration. For example, in the set $\{H_1, H_2\}$, $p_c(H_1)$ is the probability that H_1 is closer to the truth than H_2. On the other hand, in the set $\{H_1, H_3\}$, $p_c(H_1)$ is the probability that H_1 is closer to the truth than H_3. The probability of H_1 is, of course, also relative to the verisimilitude measure. The verisimilitude probability of a hypothesis is therefore not an absolute number; it is context-dependent and contrastive. This is in sharp contrast to the standard Bayesian probability of a hypothesis.

Finally, note that $p_c(\theta)$ describes an epistemic attitude different from a degree of belief in the truth of some proposition. Some might be tempted to interpret $p_c(\theta)$ as a standard probability that attaches to the proposition $<\theta$ is closest to the truth$>$. However, this is a mistake, for the reasons mentioned earlier. The proposition $<\theta$ is closest to the truth$>$ belongs to a different algebra than θ does. θ indexes a set of hypotheses in a statistical model, but $<\theta$ is closest to the truth$>$ does not. If Bayesian inference is to be used on the statistical model that is indexed by θ, the probabilities must be assigned to the parameter θ, not to the associated propositions

$<\theta$ is closest to the truth$>$. Hence $p_c(\theta)$ represents an epistemic attitude towards θ, namely the attitude that θ is closest to the truth out of the hypotheses in Θ.

7 The Verisimilitude Interpretation of Probability is Useful

The verisimilitude interpretation of probability is a logically viable solution to the interpretive problem in the sense that it does not face immediate problems with any of the probability axioms. However, some characteristics of the verisimilitude interpretation may seem objectionable. In particular, the fact that the verisimilitude interpretation makes probability assessments contrastive may be regarded as a serious drawback. Perhaps the appropriate response to the interpretive problem is not to adopt the verisimilitude interpretation, but rather to not use Bayesian methods whenever the hypotheses under consideration are all known to be false. On the other hand, maybe there is an alternative solution to the interpretive problem that is better than the verisimilitude interpretation. In this section and the next, I consider both these alternative responses to the interpretive problem.

In order to determine whether the verisimilitude interpretation is defensible, it is helpful to step back for a moment and ask a more fundamental question: why use Bayesian methods at all? If the benefits of Bayesian methods remain even when he standard interpretation of the probability axioms is replaced with the verisimilitude interpretation, then the verisimilitude interpretation is not just logically viable, but potentially *useful*. The goal of the next subsections is to give a preliminary argument for the claim that the verisimilitude interpretation is useful.

7.1 Why be a Bayesian?

What is the benefit of using Bayesian rather than other statistical methods? Perhaps the greatest selling point of Bayesianism is that the prior distribution gives researchers a principled way of incorporating background information. For example, suppose you are estimating the mass of a small cup of water, and suppose you

model the outcome of your measurement as a likelihood function $p(x|m)$, where x is the outcome of your measurement and m is a possible value of the cup's mass. A standard classical ("frequentist") method of estimating the mass of the cup is to choose as your estimate the value of m that maximizes the probability of the observed measurement. This estimation method is known as "maximum likelihood" estimation.

From a Bayesian point of view, maximum likelihood estimation is essentially equivalent to Bayesian inference with a flat (improper) prior probability function that assigns a non-zero and equal probability density to every possible value of m from $-\infty$ to $+\infty$, because the maximum likelihood estimate will be equal to the estimate that has the highest posterior probability if and only if the prior is flat. Clearly, the prior implicitly used in maximum likelihood estimation neglects to incorporate common sense background information that we have about m, and is therefore – from a Bayesian and intuitive point of view – deficient. For example, the mass of an object cannot be a negative number, so no prior should assign any probability mass to negative values of m. Furthermore, we can be absolutely certain that a small cup of water is not going to weigh more than, say, 1kg, so we can also assign a probability of 0 to all values of m greater than 1kg. Thus, as a minimal requirement, any prior probability distribution we use should be restricted to the interval $[0, 1]$. Of course, we have additional common sense knowledge that allows us to restrict the class of sensible prior distributions further.

The above example shows how even very obvious background information can be incorporated in a Bayesian prior in order to improve the inference. Indeed, at least to Bayesian statisticians and scientists who make use of Bayesian methods, this is probably the single biggest advantage that Bayesianism has over its competitors. But how are you supposed to take into account your background information when you are trying to come up with a prior probability distribution over a class of false hypotheses? Do the advantages of Bayesianism carry over when the goal of inquiry changes from finding the truth to finding the hypothesis that is closest to the truth? In the next subsection, I will suggest that the answer is "yes." Scientists often have background knowledge that they can use to discriminate between false hypotheses

in a principled way. And a good way of incorporating this background knowledge is through the construction of a Bayesian prior.

7.2 Verisimilitude and Background Knowledge

Consider again the example concerning the relationship between barometric pressure and the expected amount of rainfall. Suppose one of the things you know about the relationship between barometric pressure and precipitation is that the expected amount of precipitation is not *very* sensitive to changes in barometric pressure. Throughout the whole possible range of barometric pressure, a small change in barometric pressure will not lead to a drastic change in the amount of expected precipitation.

So far, this is background knowledge about the actual, unknown function relating barometric pressure and precipitation. What consequences does this background knowledge have for inferences about the hypothesis set actually under consideration? Suppose, as before, that the hypothesis set you are considering is the set of linear functions. That is, you model the relationship between precipitation and barometric pressure by the set of linear functions $l(Y) = \alpha X + \epsilon$, where ϵ is a normally distributed error term. Can you use your background knowledge to discriminate between the various false linear hypotheses in a principled way? Arguably, you can. Intuitively, by any reasonable measure of verisimilitude, linear functions according to which expected precipitation is not very sensitively dependent on barometric pressure are going to be closer to the truth than are linear functions that model expected precipitation as very sensitively dependent on barometric pressure.

How can all of this be captured reasonably in a prior probability distribution? Let us first see how you can formally capture your background information. Suppose f is the true (and unknown) functional relationship between precipitation and barometric pressure. Then the background information that precipitation does not depend sensitively on changes in barometric pressure can be modeled as a claim about the partial derivative of f (with respect to the barometric pressure variable). The simplest and least sophisticated way of translating your background information into a

quantitative restriction on f' is to to suppose that f' is bounded by some interval (a, b). Next, the intuition that insensitive linear hypotheses are closer to the truth than sensitive linear hypotheses can be formalized as follows: there is some suitably large interval (a', b') that contains (a, b) such that every linear hypothesis l for which l' is bounded by (a', b') is closer to the truth than every linear hypothesis that does not satisfy this requirement. Now, since $l' = \alpha$, the requirement that l' be bounded by (a', b') reduces to the simple requirement that every $\alpha \in (a', b')$ is closer to the truth than every $\alpha \notin (a', b')$. This, in turn, translates to a simple rational requirement on the prior distribution over α, namely that every $\alpha \notin (a', b')$ be assigned a prior probability of 0.

There are more refined ways of formalizing the background information that expected precipitation does not depend very sensitively on barometric pressure. In particular, if we assume a specific verisimilitude measure, then we can get tighter constraints on α.[9] Furthermore, if the hypotheses under consideration are more complicated (i.e. contain more parameters), then the background information will not lead to rational requirements on the prior distribution as neatly. My goal in this section is not, however, to demonstrate in full generality how to best translate background information into reasonable requirements on prior distributions over false hypotheses. My goal is rather to show that it is possible to do so, and that it is plausibly useful. I defer a more thorough treatment of these issues to another time.

[9]For example, suppose we use the following reasonable albeit crude distance measure as our measure of verisimilitude: if f is the true function over the range (m, n) and l is a linear function, then the verisimilitude of l is $v(l) = Max_{x \in (m,n)} |f(x) - l(x)|$. In this case, if we assume that we know that f is bounded by (a, b), then it is possible to prove that every linear function l whose derivative is bounded by (a, b) is closer to the truth than every linear function whose derivative is not bounded in this way, where closeness to the truth is measured using v. For the sake of space, I omit the proof.

8 The Counterfactual Interpretation of Probability

The preceding section shows that the verisimilitude interpretation of the probability axioms is a potentially useful solution to the interpretive problem. However, it may be that there is another solution to the interpretative problem that is better. Earlier, we examined two candidate solutions to the interpretive problem and found them wanting. However, in a very recent paper, Jan Sprenger (2016) proposes a new and different solution to the interpretive problem that is more promising. Sprenger's solution also involves reinterpreting the probability axioms, but he offers a reinterpretation that is interestingly different from the verisimilitude interpretation. However, as we will soon see, given certain plausible assumptions, the verisimilitude solution and Sprenger's solution are formally inter-translatable.

Sprenger's suggestion is that the probability of a false hypothesis can sensibly be interpreted as a *counter-factual* degree of belief. More precisely, suppose α is a parameter that indexes a set of hypotheses, all of which are known to be false. Then any probability assigned to some particular α_0 should be construed as a degree of belief in α_0 that is *conditional* on the (false) supposition that one of the hypotheses indexed by α is true. In other words, the probability of α_0 is really the *conditional* probability $p(\alpha_0 | \vee \alpha)$, where the condition $\vee\alpha$ is the false disjunction that says that one of the α's is true.

This idea is less abstract than it may seem at first blush. As an illustration, suppose I have a coin in a locked cabinet. The probability that the coin would land heads given that I *were* to toss the coin is 0.5, even if it is false that I ever toss the coin. Similarly, according to Sprenger, we can evaluate the probability that a hypothesis is true given that the false supposition that the world *were* such that one of the hypotheses under consideration is true.

According to Sprenger, the counterfactual interpretation of probability offers a simple solution to the interpretive problem that avoids the "muddy waters of verisimilitude." However, in order to actually evaluate counterfactual probabilities in a principled manner, it seems we have to enter waters that are at least as muddy

as the verisimilitude waters. Consider again the example concerning the set of linear hypotheses relating X (barometric pressure) to Y (precipitation in the next hour). We have already agreed that your actual degrees of belief in all of these linear hypotheses is 0. Your degree of belief (or probability density, rather) in some particular linear hypothesis conditional on the disjunction of all the linear hypotheses may still be different from 0, but how are you supposed to figure out what it is? You somehow have to figure out what your probabilities *would* be on the assumption that the world were such that barometric pressure and precipitation were perfectly linearly related. In order for the counterfactual interpretation of probability to be a viable alternative, guidance on how to evaluate counterfactual probabilities is necessary, in the same way that some assumptions about verisimilitude are necessary in order for the verisimilitude interpretation to be viable.

The standard way of evaluating ordinary counterfactuals is by appealing to possible worlds. According to (a simplified version of) Lewis's analysis of counterfactuals (Lewis, 1973), in order to evaluate a counterfactual such as "If A were the case, then B would be the case," you have to go to the closest possible world in which A is true, and then see whether B is true in that world. Crucially, Lewis's analysis depends on a ranking of worlds, where worlds are ranked by how similar they are to the actual world.

Presumably counterfactual probabilities should be assessed in a similar manner. It is not hard to imagine very strange and fanciful possible worlds in which barometric pressure and precipitation are linearly related, but presumably most of those possible worlds are not interesting or relevant. As is the case in counterfactual analysis of conditionals, it is presumably the closest possible worlds that are the interesting ones. But which possible worlds are those? To answer this question, you need to be able to rank worlds in terms of their closeness or similarity to the actual world. But a ranking of possible worlds is hardly easier to come up with than a verisimilitude ranking of hypotheses.

8.1 Relationship Between the Verisimilitude and Counterfactual Solutions

Indeed, in general, any similarity ranking on possible worlds straightforwardly induces a natural verisimilitude ranking on hypotheses, and vice versa.[10] More precisely, suppose we are given a similarity ranking on worlds $w_\alpha \geq w_1 \geq w_2 \geq \ldots$, where w_α is the actual world. Then we can define a verisimilitude ranking on hypotheses as follows: suppose w is the closest world in which H is true and w' is the closest world in which H' is true, then $v(H) \geq v(H')$ if and only if $w \geq w'$.[11]

Conversely, any verisimilitude ranking induces an ordering of possible worlds. Suppose $v(H_0) > v(H_1) > v(H_2) > \ldots$ is a verisimilitude ranking of hypotheses, and for any hypothesis p, let S_p denote the set of worlds in which p is true. Then we can define an ordering of possible worlds in the following way: suppose H is the hypothesis with the highest verisimilitude such that that $w \in S_H$ and suppose H' is the hypothesis with the highest verisimilitude such that $w' \in S'_H$, then $w \geq w'$ if and only if $v(H) \geq v(H)'$.

Thus, although they appear very different, the verisimilitude interpretation and the counterfactual interpretation of probability are formally inter-translatable.

Although the two approaches are formally inter-translatable, they provide different perspectives and help illuminate each other. In particular, it is arguably easier to come up with a verisimilitude measure than a ranking over possible worlds; for example, the Kullback-Leibler measure is a well known verisimilitude measure over statistical models, and this verisimilitude measure will induce a partial ranking over possible worlds. Thus, the verisimilitude approach helps explain where rankings over possible worlds are supposed to come from.

On the other hand, the counterfactual approach helps explain several features of the verisimilitude interpretation as well. For example, earlier we saw that the

[10]For simplicity, the following informal demonstration presupposes the so-called "Uniqueness Assumption" according to which, for every A, there is a unique closest possible world in which A is true. This is a strong and implausible assumption. However, the demonstration does not depend on this assumption.

[11]Hilpinen (1976) uses a similar approach to define a specific verisimilitude measure.

verisimilitude probability of a hypothesis H is relative to the set of hypotheses under consideration. If H is considered as part of the set $\{H, H'\}$, the verisimilitude probability of H is the probability that H is closer to the truth than is H'. But if H is considered as part of the set $\{H, H''\}$, the verisimilitude probability of H is the probability that H is closer to the truth than is H'. The counterfactual interpretation clarifies what is going on here. In the first case, the counterfactual probability that corresponds to $p_c(H)$ is $p(H|H \vee H')$; in the second case, the counterfactual probability that corresponds to $p_c(H)$ is instead $p(H|H \vee H'')$. As can be seen, the two counterfactual probabilities are conditional on different disjunctions, and it is therefore not mysterious that the corresponding verisimilitude probabilities are also different.

9 Summary and Future Research

I have argued that the interpretive problem is a serious problem, but that the problem does not necessarily arise just because the statistical model under consideration is wrong; rather, the interpretive problem arises whenever the *hypotheses of interest* are false. Next, focusing on parameter inference, I have argued that the verisimilitude reinterpretation of the probability axioms provides a logically viable and potentially useful solution to the interpretive problem. Finally, I have contrasted the verisimilitude reinterpretation with another reinterpretation due to Jan Sprenger, and I have argued that the two reinterpretations are formally inter-translatable, but that they nevertheless shed interestingly different lights on the interpretive problem and on each other.

Several important questions remain unanswered, however. In particular, I have not discussed the problem of Bayesian model inference or model selection when all the models are all false. Nor have I discussed in any detail how researchers can come up with principled prior probabilities that discriminate between false hypotheses. Finally, I have not said anything about what consequences reinterpreting the probability axioms has for evidential principles like the Likelihood Principle or the Law of Likelihood. All of this is work for the future.

References

Box, George E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness." *Journal of the Royal Statistical Society. Series A (General)*, 143, 383–430.

Carnap, Rudolph (1950), *Logical Foundations of Probability*. The University of Chicago Press.

Easwaran, Kenny (2014), "Regularity and Hyperreal Credences." *Philosophical Review*, 123, 1–41.

Festa, Roberto (1999), "Bayesian Confirmation." In *Experience, Reality, and Scientific Explanation* (Maria Carla Galavotti and Alessandro Pagnini, eds.), volume 61 of *The Western Ontario Series in Philosophy of Science*, 55–87, Springer Netherlands.

Forster, Malcolm and Elliott Sober (1994), "How To Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science*, 45, 1–35.

Gelman, Andrew and Cosma Rohilla Shalizi (2013), "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology*, 66, 8–38.

Hilpinen, Risto (1976), "Approximate Truth and Truthlikeness." In *Formal Methods in the Methodology of Empirical Sciences*, number 103 in Synthese Library, 19–42, Springer Netherlands.

Lewis, David K. (1973), *Counterfactuals*. Blackwell Publishers.

Morey, Richard D., Jan-Willem Romeijn, and Jeffrey N. Rouder (2013), "The Humble Bayesian: Model Checking From a Fully Bayesian Perspective." *British Journal of Mathematical and Statistical Psychology*, 66, 68–75.

Niiniluoto, Iikka (1986), "Truthlikeness and Bayesian Estimation." *Synthese*, 67, 321–346.

Niiniluoto, Iikka (1998), "Verisimilitude: The Third Period." *British Journal for the Philosophy of Science*, 49, 1–29.

Northcott, Robert (2013), "Verisimilitude: A Causal Approach." *Synthese*, 190, 1471–1488.

Popper, Karl (1963), *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, Hutchinson.

Pruss, Alexander R. (2014), "Infinitesimals Are Too Small for Countably Infinite Fair Lotteries." *Synthese*, 191, 1051–1057.

Shaffer, Michael J. (2001), "Bayesian Confirmation of Theories That Incorporate Idealizations." *Philosophy of Science*, 68, 36–52.

Sober, Elliott (2015), *Ockham's Razors: A User's Manual*. Cambridge University Press.

Sprenger, Jan (2016), "Conditional Degree of Belief." Manuscript.

Wenmackers, Sylvia and Leon Horsten (2013), "Fair Infinite Lotteries." *Synthese*, 190, 37–61.

A Approximate Truth and Bayesianism

There are two natural ways of trying to accommodate approximate truth within Bayesianism. The first way is to expand the algebra of propositions that p ranges over, so that it also ranges over propositions such as $<\theta$ is approximately true$>$ – or P_θ for short. Thus, even though strictly speaking we assign each θ a probability of 0 of being true, we can consistently assign its associated proposition P_θ a non-zero probability, and moreover this probability represents the probability that P_θ is true

and not just approximately true, since the approximation claim is in the proposition itself. This way, the standard Bayesian interpretation of the probability axioms is preserved.

The other natural way of attempting an accommodation is to abandon the standard Bayesian interpretation of the probability axioms, so that $p(\theta)$ is interpreted as the probability that θ is approximately true rather than true. This line of reasoning is pursued by Niiniluoto (1986) and Festa (1999). Let p_a be a potential probability function where the a subscript indicates that the intended interpretation of $p_a(\theta)$ is the probability that θ is approximately true rather than true. For concreteness, we may imagine that p_a represents the degrees of belief that some agent has in all hypotheses (and models, theories, etc) that the agent takes to potentially be approximately true. By contrast, p can be taken to represent the same agent's degrees of belief in propositions that the agent takes to potentially be true.[12] p_a is therefore defined over a much more expansive set of hypotheses, models, theories, etc. than is p. However, if we allow propositions such as $<\theta$ is approximately true$>$, then presumably there will be a simple correspondence between p_a and p in that we should have $p_a(\theta) = p(P_\theta)$.

There is some reason to prefer working with p_a rather than with propositions such as P_θ. Bayes's formula requires that we assign unconditional probabilities to data x. If we stay inside the original distribution p, this means we have to calculate $p(x) = \sum p(x|P_{\theta_i})p(P_{\theta_i})$, but then we are faced with having to make sense of $p(x|P_{\theta_i})$, or in other words the probability of x conditional on the assumption that θ_i is approximately true. But this is hard to make sense of. In statistical practice, each θ_i will, as was mentioned earlier, in general be part of a fully specified statistical model, which means it will entail a probability for each of the possible outcomes. The associated proposition, P_{θ_i}, however, does not entail any probabilities for data, and it is hard to see how to come up with reasonable conditional probabilities of the form $p(x|P_{\theta_i})$. One might try to argue that it is reasonable to hold that $p(x|P_{\theta_i}) \approx p(x|\theta)$, and this will provide a rough value for $p(x|P_{\theta_i})$, but not a precise one.

[12] Although I hasten to add that a subjective Bayesian perspective will not really play any significant role here.

If, on the other hand, we move to the distribution p_a, then we can expand the probability of x as $p_a(x) = \sum p_a(x|\theta_i)p_a(\theta_i)$. Now, if we suppose that the statistical model stays the same, then it is reasonable to suppose that $p_a(x|\theta_i) = p(x|\theta_i)$; i.e. θ_i still entails the same probability for x in the p_a distribution as it does in the p distribution. The final thing we need to do is to define the joint probability of θ_i and x, which we can naturally define as follows: $p_a(\theta_i \& x) = p(x|\theta_i)p_a(\theta_i)$. Thus, we can write $p_a(x) = \sum p(x|\theta_i)p_a(\theta_i)$.

Introducing the p_a distribution has problems of its own, however, since it's not immediately clear whether such a function can actually satisfy the probability axioms. For example, physicists use both the liquid drop model (L) and the shell model (S) of the nucleus in order to generate predictions, even though these models are logically inconsistent. Presumably, both L and S should be taken to be "approximately true" since they are both auxiliary assumptions used by scientists to generate predictions; hence we should expect it to be the case (at least) that $p_a(L) > 0.5$ and $p_a(S) > 0.5$. However, since L and S are logically inconsistent, the third axiom tells us that $p_a(S \vee L) = p_a(S) + p_a(L) > 0.5 + 0.5 = 1$, which is impossible because (by the first axiom) no probability can be greater than 1. Thus, there is apparently a very foundational problem with trying to change our interpretation of probability so that probabilities are interpreted as probabilities of approximate truth rather than probabilities of truth.

However, on closer inspection, this objection fails. The third probability axiom applies to sets of "logically incompatible" hypotheses; but what does it mean for a set of hypotheses to be logically incompatible? On the standard interpretation, it means that it is not possible for more than one of the hypotheses to be true; i.e. the third axiom is interpreted as follows:

> 3S. $P(\theta_i) = \sum P(\theta_i)$ whenever it is impossible for more than one θ_i to be true.

However, in contexts where approximate truth rather than strict truth is the target, this is arguably not how the axiom should be interpreted. Instead, the axiom should be interpreted in the following way:

3A. $P_a(\theta_i) = \sum P_a(\theta_i)$ whenever it is impossible for more than one θ_i to be approximately true.

On the new reading, the earlier objection loses its grip, for – as was pointed out earlier – it is possible for both the shell model and the drop model to be approximately true, so the condition for applying the formula in the third axiom is not met—the two models are not logically incompatible in the sense of $3A$.

Unfortunately, this feature also leads to a serious problem, because the hypothesis spaces that scientists generally use will not be logically incompatible in the sense of axiom $3A$, precisely because it will in general be possible for multiple hypotheses in the hypothesis space to be approximately true. But this is bad news, because in order for Bayes's formula to be applicable, the hypothesis space we use *must* consist of logically incompatible hypotheses, since the denominator of Bayes's formula requires that $p_a(x)$ (or $p(x)$) be expanded in terms of hypotheses that are logically incompatible. Consider, for concreteness, the class of one-variable linear hypotheses, $y = ax$, indexed by the parameter $a \in \mathbb{R}$, and suppose we have available a continuous verisimilitude measure v. Now suppose the true relationship between y and x is not actually linear. Suppose moreover that we set the approximation threshold at $\epsilon > 0$, so that $y = ax$ counts as approximately true if and only if $0 < v(a) < \epsilon$, i.e. if and only if $v(a)$ is in the open interval $S = (0, \epsilon)$. Then the set of hypotheses that are approximately true is indexed by $A = \{a \in \mathbb{R} \mid v(a) \in S\}$. Moreover, $v^{-1}(S) = \{a \in \mathbb{R} \mid v(a) \in S\} = A$, which means A is also an open interval because v is continuous. Since A is an open interval, it has either no members or infinitely many. But this means either none or infinitely many of the hypotheses will be approximately true. In neither case will Bayesian inference be possible. If *none* of the hypotheses are approximately true, then clearly the goal of the inference cannot be to find a hypothesis that is approximately true. If, on the other hand, infinitely many of the hypotheses under consideration count as approximately true, then the hypotheses cannot be used to calculate an unconditional probability for x. But from this it follows that Bayes's formula cannot be applied, and so Bayesian inference will not be possible.

The above problem arises whenever the verisimilitude measure v is continuous and the hypotheses we are considering are parameterized by a real-valued parameter. But many of the hypotheses spaces that applied statisticians make use of *are* parameterized by continuous parameters; hence the problem arises very widely.

There are, as far as I can see, two ways we can try to get out of this problem. As was mentioned earlier, there are two ways the unconditional probability of x can be calculated, depending on whether we use p_a or p with an expanded algebra of propositions. In the p_a distribution we have $p(x) = \sum p_a(x|\theta_i)p_a(\theta_i)$. In the p distribution, we instead have $p(x) = \sum p(x|P_{\theta_i})p(P_{\theta_i})$, where P_{θ_i} is the proposition $<\theta_i$ is approximately true$>$.

If we expand the unconditional probability of x in the first way, we can try to coarse-grain the hypothesis space; if we expand the unconditional probability of x in the second way, we can try to create a partition out of the P_{θ_i} propositions. Neither alternative is very promising.

Let us consider the second way out first. Carnap (1950) taught us how to create a partition out of any set of propositions. The method is as follows: given any set of propositions – A and B, let's say – we form the *state descriptions* $A\&B$, $A\&\neg B$, $\neg A\&B$, $\neg A\&\neg B$. The resulting state descriptions then form a partition. Now, given a set of hypotheses $\{\theta_i\}$, Carnap's method can be used to make a partition out of the set of associated propositions, $<\{\theta_i$ is approximately true$\}>$; the resulting state descriptions will then be logically incompatible (in the sense of $3S$), and we can therefore use Bayes's formula on the resulting partition of state descriptions. There are, however, two major problems with this proposed solution. First, note that if there are n hypotheses in the hypotheses set, then the partition of state descriptions will have 2^n propositions. But that means that if the hypothesis space is parameterized by a continuous parameter – so that its cardinality is $\aleph_1$ – the partition of state descriptions will have cardinality $2^{\aleph_1}$. But it is not possible to assign a regular probability (density) distribution over a set with cardinality $2^{\aleph_1}$. The resulting probability distribution will have to make use of "hyperreal" numbers (Wenmackers and Horsten, 2013), but there are significant difficulties associated with hyperreal probabilities—see, e.g., Easwaran (2014) and Pruss (2014).

The other problem is perhaps even worse. In order to do use Bayes's formula, Bayesians who make use of the above proposed solution will have to somehow assign likelihoods to each of the state descriptions, each of which is a heinous conjunction of propositions of the form $<\{\theta_1$ is approximately true$\}>\&<\{\theta_2$ is approximately true$\}>\&\neg<\{\theta_3$ is approximately true$\}>\&\ldots$etc. It is very hard to see how reasonable probabilities can be assigned conditional on such complicated expressions.

The other possible way out of the problem is to coarse-grain the hypothesis space. If the hypothesis space is parameterized by a continuous parameter, then – as we have seen – infinitely many hypotheses will in general count as approximately true if any hypothesis counts as approximately true. However, if we make the hypothesis space *discrete* by throwing out most of the hypotheses, then the remaining hypotheses may well all be logically incompatible (in the sense of $3A$). For example, if the parameter that indexes the hypotheses ranges over the interval $(0, 1)$, then we could coarse-grain the parameter to $(0.2, 0.4, 0.6, 0.8, 1.0)$, which may well range over hypotheses that are logically incompatible. However, coarse-graining the hypothesis space in this way is not very attractive because (1) how to coarse-grain the space would depend on which ϵ threshold we use, (2) there are multiple ways to coarse-grain a hypothesis space, and each way arbitrarily throws out most of the viable hypotheses. Needless to say, no Bayesian statisticians actually coarse-grain the hypothesis spaces they use in this way; nor, for that matter, do they create state descriptions in the way suggested in the previous solution. Hence, accommodating approximate truth within the Bayesian framework does not seem to be feasible when the hypothesis space is indexed by a continuous parameter.

The above considerations do not show that all is lost for the approximate truth interpretation of probability, however. In particular, if the hypothesis space is discrete, then the above problems may not arise. On the other hand, the problems will arise even with discrete hypotheses spaces, provided there are multiple hypotheses that all meet the verisimilitude threshold that is set for approximate truth. So to prevent these problems from arising, it is necessary to make sure that the hypotheses (or models) under consideration are sufficiently distinct from each other so that only (and precisely) one of them will count as approximately true. Otherwise,

Bayesian methods will not be applicable because the hypotheses (or models) will not be mutually exclusive in the requisite sense (i.e. in the sense of 3A).

But this is an awkward problem to have to deal with. And it points to a defect with the concept of approximate truth: approximate truth is intrinsically too coarse-grained a concept since it fails to distinguish between several hypotheses, all of which are approximately true.