# New Semantics for Bayesian Inference: The Interpretive Problem and Its Solutions

October 29, 2018

**Abstract**

Scientists and Bayesian statisticians often study hypotheses that they know to be false. This creates an interpretive problem because the Bayesian probability assigned to a hypothesis is typically interpreted as the probability that the hypothesis is true. I argue that solving the interpretive problem requires coming up with a new semantics for Bayesian inference. I present and contrast two solutions to the interpretive problem, both of which involve giving a new interpretation of probability. I argue that both of these new interpretations of Bayesian inference have the same advantages that the standard interpretation has, but that they have the added benefit of being applicable in a wider set of circumstances. I furthermore show that the two new interpretations are inter-translatable and I explore the conditions under which they are co-extensive with the standard Bayesian interpretation. Finally, I argue that the solutions to the interpretive problem support the claim that there is pervasive pragmatic encroachment on whether a given Bayesian probability assignment is rational.[1]

[1]This paper is forthcoming in *Philosophy of Science*. Thanks to audiences at NTU, University of Wisconsin–Madison, and the 2016 meeting for the Philosophy of Science Association. Thanks, in particular, to Kenny Easwaran, Malcolm Forster, Elliott Sober, Jan Sprenger, Mike Titelbaum, and the referees for *Philosophy of Science*.

# 1 Introduction

Bayesianism is one of the most influential contemporary frameworks for statistical inference, but from a philosophical point of view Bayesian inference faces several difficulties. One particularly serious problem is that statisticians who use Bayesian methods often assign non-zero probabilities over sets of hypotheses that they know are false; yet, as I show in the next section of the paper, this practice is inconsistent with the interpretation of probability that is standardly assumed by Bayesians. Thus there is a tension between the standard Bayesian interpretation of probability and the way the Bayesian framework is often applied, which I will refer to as the "interpretive problem."[2]

Although the problem is primarily interpretive and philosophical, it also has practical consequences. According to most Bayesians, probability distributions ought to incorporate relevant background information – indeed, the fact that Bayesians can do this in a principled way is often touted as a major advantage that Bayesianism has over rival statistical frameworks, such as frequentism. However, in cases where the standard Bayesian interpretation of probability fails, it's unclear how background information should be taken into account in a principled way. Probably in part for this reason, so-called "default priors" that do not even attempt to take into account relevant background information have gained prominence in recent years. But default priors have their own problems (De Heide and Grunwald, 2018). Hence, solving the interpretive problem is not just philosophically interesting; it is also of some practical importance.

I will argue that the only satisfactory solutions to the problem involve reinterpreting what it means to assign a probability to a hypothesis. Accord-

[2]The problem has been noted in the past, e.g. by Box (1980), Bernardo and Smith (1994), Forster and Sober (1994), Forster (1995), Key et al. (1999), Shaffer (2001), Sprenger (2009), Gelman and Shalizi (2013), Walker (2013), and Sprenger (2017) – indeed, Sprenger calls the problem the "scandal of Bayesianism"– but in general the seriousness of the issue seems to be under-appreciated.

ing to one solution (originally proposed by Sprenger (2017)), probabilities are interpreted counterfactually; according to a second solution, probabilities are interpreted as what I will refer to as "verisimilitude probabilities." Much of the paper will be concerned with exploring the features of these two interpretations. In particular, I will argue that the verisimilitude and counterfactual interpretations have the same nice features that the standard interpretation has, but that they have the added benefit of being sensible and useful in situations in which the standard interpretation is not. In particular, the verisimilitude and counterfactual interpretations of probability enable us to incorporate background information in probability distributions in a principled manner, even when all the hypotheses under consideration are known to be false. I will also show that the two interpretations are inter-translatable and that they are therefore – in an intuitive sense – equivalent, and I will explore the relationship between the verisimilitude and counterfactual interpretations, on the one hand, and the standard interpretation on the other.

Although the interpretive problem arises in applied statistics, both the verisimilitude interpretation and the counterfactual interpretation of probability are interesting from an epistemological point of view. In particular, both interpretations have the feature that whether a given Bayesian probability distribution is rational is partly influenced by pragmatic factors. As I argue in Section 10, there are good reasons for suspecting that all solutions of the interpretive problem will have this feature. Thus, I argue, there is an interesting – and unavoidable – form of pragmatic encroachment in Bayesian inference.

# 2 An Abstract Characterization of the Interpretive Problem

The purpose of this section is go give a brief introduction to the fundamentals of Bayesian statistical inference and to provide an abstract characterization

of the interpretive problem; in the next section, I show how the problem arises in practice.

The basic objects of study in Bayesian statistical inference are *statistical models*. Given a set of candidate hypotheses indexed by a *parameter*, $\theta$ in $\Theta$, and given some particular context in which the possible observations or outcomes are $x_1$, $x_2$, etc. in $X$, and given a corpus of background knowledge or background assumptions $K$, a statistical model is a set of conditional probability (density) distributions,[3] $p_K(x|\theta)$, that jointly specify the probability of each possible $x$ in $X$ given each possible $\theta$ in $\Theta$. Given a statistical model or a set of statistical models, Bayesians do inference by following a three-step procedure:

In the first step, a probability is assigned to each $\theta \in \Theta$; these probabilities are supposed to be assigned before looking at the data and are therefore known as "prior" probabilities. If there are multiple candidate statistical models, then all of the models must be assigned prior probabilities as well. The requirement that the numbers assigned to parameters be probabilities rather than just arbitrary real numbers means that the assignment must satisfy the following constraints:

> **Standard probability axioms** Suppose $\Theta$ indexes a set of hypotheses $\{\theta_1, \theta_2, \ldots, \theta_n\}$ considered by some agent, and let $K$ represent a corpus of background knowledge. Then the distribution $p_K$ over $\Theta$ satisfies the probability axioms if and only if:
>
> > 1S. $p_K(\vee\theta_i) = 1$, whenever $K$ entails that at least one hypothesis in the disjunction of hypotheses indexed by $\vee\theta_i$ is true.
> >
> > 2S. $p_K(\theta_i) \geq 0$ for all $\theta_i$ in $\Theta$.

[3]From now on, I will for simplicity simply use "probability" although in practice probability densities are more common.

3S. $p_K(\bigvee \theta_i) = \sum p_K(\theta_i)$, whenever $K$ entails that at most one of the hypotheses in the disjunction of hypotheses indexed by $\vee\theta_i$ is true.

Bayesians divide over how, exactly, $p_K$ should be interpreted. Subjective Bayesians interpret $p_K$ as the degrees of belief of some particular agent and $K$ as that particular agent's background knowledge, whereas objective Bayesians typically interpret $p_K$ as representing a logical degree of support and $K$ as representing a collection of "objective" background information (or intersubjectively shared background knowledge). For our purposes, the differences between subjective and objective Bayesians will not be important. The more important fact, from our point of view, is that both subjective and objective Bayesians agree that $p(\theta)$ represents a probability that the hypothesis indexed by $\theta$ is *true*.

In the second step of Bayesian inference, data $x$ are collected and the "likelihood" of each hypothesis is calculated. The likelihood of $\theta$ is the probability that $\theta$ assigns to the data, $p_K(x|\theta)$.

In the third and final step, the *posterior* probability of each parameter and each statistical model is calculated by combining the prior and the likelihood of each hypothesis using Bayes's theorem, $p_K(\theta|x) = p_K(x|\theta) * p_K(\theta)/p_K(x)$.

In what follows, I will refer to the above three-step procedure as "standard Bayesian inference." Although I think each of the three steps of standard Bayesian inference faces difficulties, in this paper I will focus on the first step. What I will refer to as the "interpretive problem" arises whenever scientists assign non-zero probabilities to hypotheses that they know to be false. In such situations, they will, in fact, be violating the probability axioms.

To see why, let's suppose, for simplicity (but without loss of generality), that the parameter $\theta$ can take a finite number of possible values $\theta_1$, $\theta_2$, ..., $\theta_m$. Now suppose we know that each of the hypotheses under consideration is false, i.e. $K$ entails that $\theta_i$ is false, for each $i$. Then $K$ entails that $\neg\theta_i$ is true, for each $i$. 1S then implies that we must – on pain of violating the

probability axioms – assign a probability of 1 to $\neg\theta_i$. Finally, axioms 2S and 3S jointly entail that we must must assign a probability of 0 to $\theta_i$ for every $i$. Hence, if we nonetheless assign non-zero numbers to the various possible values of $\theta$, we will be violating the standard probability axioms.[4]

In the next section, I will argue that scientists often know that all of the hypotheses they consider are false.

# 3 The Interpretive Problem in Practice

Scientists are often interested in studying the functional relationship between multiple quantities. Statisticians call this type of problem "regression analysis." An example of a regression problem that is of obvious practical importance[5] concerns the relationship between minimal pressure and maximal windspeed in tropical storms. Let $X$ represent the minimal pressure of some storm and let $Y$ represent the maximal windspeed of the storm; then we would like to know the true functional dependence of $Y$ on $X$. This relationship is unknown and probably quite complex. However, various idealized assumptions (see Knaff and Zehr (2007)) justify the following model:

$$Y = \alpha(1010 - X)^n + \epsilon \tag{3.1}$$

Here, $\epsilon$, $n$, and $\alpha$ are all parameters that must be estimated from the data.[6] Each triple of values for $\alpha$, $\epsilon$, and $n$ picks out a given hypothesis about the true relationship between $X$ and $Y$. Importantly, the fact that the

[4]A referee points out that one way to undercut this argument is to insist that the probability distribution should only be based on some proper subset of $K$. This is correct, but then the question arises of which proper subpart of $K$ it is legitimate to use. The verisimilitude and counterfactual interpretations that I offer later in the paper provide principled answers to this question.

[5]Discussed, for example, by Choi et al. (2016).

[6]Strictly speaking, $\epsilon$ itself is not a parameter; it is an error term, which in general will have an associated parameter $d$ that will need to be estimated. I will gloss over those nuances here.

model is based on idealized assumptions (i.e. assumptions that are known to be violated in practice—indeed physically impossible) implies that the model in fact is known to be false. That is, the true relationship between $Y$ and $X$ does not belong to the class of hypotheses picked out by the parameters in the model. Hence, every hypothesis picked out by any triple of values for $\alpha$, $n$, and $\epsilon$ is also known to be false, even before any evidence is collected.

It's worth emphasizing that this example is by no means unrepresentative. It is almost invariably the case in regression problems that the hypotheses under consideration will be restricted to very simple functional relationships, such as the set of lines, parabolas, exponentials, etc. Most functional relationships in the world cannot realistically be expected to belong to one of these sets of simple functional relationships, and indeed the choice of functional class is usually justified on the basis of highly idealized scientific assumptions, if it is justified at all. Hence, scientists will generally know that all the functional relationships they consider are false. By the argument at the end of the preceding section, the probability axioms imply that scientists ought to assign a probability of 0 to all of their hypotheses. But that is of course not what they do, and for good reason because in the Bayesian framework assigning a hypothesis a probability of 0 is tantamount to excluding it from further consideration. If scientists were to assign a probability of 0 to all functional relationships they know to be false, they would in effect rule out all of their hypotheses from the get-go.

Bayesian phylogenetics is an example of another major area of statistical inference where scientists generally know that the hypotheses they consider are false. 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 (see, e.g. O'Malley et al. (2010), Heggarty et al. (2010), and Velasco (2012)). For example, a tree phylogeny for a language family is premised on the (false) idea that languages bifurcate instantaneously and

are forever separated thereafter. Again, if Bayesian phylogeneticists took seriously the standard probability axioms, then they would have to assign all of their hypotheses a prior probability of 0. But that is not what they do.

The widespread practice of assigning non-zero prior probabilities to hypotheses that are obviously false 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?

# 4 Unsuccessful Solutions to the Interpretive Problem

One response to the interpretive problem that initially strikes many philosophers as attractive is to try to change the algebra over which the probability function $p$ ranges. For example, some might be tempted to consider the algebra generated by the associated propositions, $<\theta_i$ is the best hypothesis$>$, for each $\theta_i$, or something similar. The idea is that even if $\theta_i$ must be assigned a probability of 0 (because it is known to be false), the standard probability axioms allow us to assign $<\theta_i$ is the best hypothesis$>$ a non-zero probability.

However, this proposal faces several difficulties. The most immediate problem is the fact that scientists do not, in fact, consider hypotheses of the form $<\theta_i$ is the best hypothesis$>$. And for good reason, as we will soon see. The problem is that whereas a parameter $\theta$ in a statistical model will index a set of probability distributions each of which entails probabilities for the various possible observations, an expression such as $<\theta_i$ is the best hypothesis$>$ does not. For example, in the example in Section 3, $\alpha = 1$ picks out a particular class of hypotheses that make probabilistic predictions about the possible observations;[7] but a proposition such as $<\alpha = 1$ is the best hypothesis$>$ is not part of any statistical model and does not make any probabilistic predictions.

[7] In fact, each value of $\alpha$ picks out a class of hypotheses that is itself a statistical model.

To see the problem from a different perspective, consider Bayes's formula:

$$p_K(\theta|x) = \frac{p_K(x|\theta) * p_K(\theta)}{p_K(x)} \tag{4.1}$$

Clearly, the likelihood and the prior have to range over the same set of hypotheses in order for Bayes's formula to be applicable. If we change the algebra of hypotheses so that we instead assign probabilities to propositions of the form $<\theta_i$ is the best hypothesis$>$, then we may assign non-zero prior probabilities to our hypotheses without violating the probability axioms. However, now the likelihoods will be of the form $p_K(x|\ <\theta_i$ is the best hypothesis$>)$, but $<\theta_i$ is the best hypothesis$>$ does not entail any probabilistic prediction for $x$, so it's hard to see how we are to come up with a principled estimate for $p_K(x|\ <\theta_i$ is the best hypothesis$>)$.[8]

There is another, related, reason why we cannot just change the algebra over which the probability distribution ranges. The problem is that in replacing $\theta_i$ with $<\theta_i$ is the best hypothesis$>$, important evidential relationships between the hypotheses and evidence will generally be lost. An important special case is parameter estimation with exchangeable evidence,[9] where a theorem due to de Finetti[10] shows that there will be a probability model such that the parameters of the model render the evidence conditionally in-

[8] A similar solution has recently been proposed in the statistics literature. Walker (2013) suggests that in cases where no hypothesis in the model indexed by $\theta$ is true, we ought to construe the goal of Bayesian analysis as finding the hypothesis $\theta^*$ that minimizes statistical divergence from the true data-generating distribution (a similar proposal is adopted by Bissiri et al. (2016)). Hence the prior distribution ranges over the possible values of $\theta^*$. There is a problem, however: the parameter $\theta^*$ and the parameter $\theta$ range over distinct hypotheses; $\theta$ ranges over hypotheses in a statistical model whereas $\theta^*$ ranges over hypotheses of the following form, where $S$ is a statistical divergence and $g$ is the truth: $\theta^* = \min_{\theta \in \Theta} S(\theta, g)$. Hence the likelihood, which Walker derives from the statistical model, is of the form $p(x|\theta)$, whereas the prior is of the form $p(\theta^*)$. But $p(x|\theta)$ and $p(\theta^*)$ cannot be combined using Bayes's formula since they range over different sets of hypotheses. To be fair, Walker (2013) is sensitive to the problem.

[9] Roughly speaking, evidence is exchangeable if the probability of receiving any given sequence of evidence is not dependent on the order in which the evidence is received

[10] Proven in a more general form by Hewitt and Savage (1955).

dependent. Hence, when the evidence is exchangeable, statisticians have an imperative to construct models that render the evidence conditionally independent. But $<\theta_i$ is the best hypothesis$>$ will in general not render the evidence conditionally independent whenever $\theta_i$ does.

As a concrete example, consider coin tossing. Coin tosses are clearly exchangeable (e.g. "Heads, Tails, Heads" is as probable as "Heads, Heads, Tails"), so de Finetti's theorem implies that there exists a model with a parameter that renders the coin tosses conditionally independent. In fact, there is a well known model that does this, namely the model that posits a parameter, Bias, that represents the coin's underlying propensity to land Heads. Each possible bias of the coin renders all future coin tosses conditionally independent.[11] The coin bias model is therefore an adequate statistical model for coin tossing in the sense that it captures the conditional independence relations between evidence and hypotheses that de Finetti's theorem says it's possible to capture. However, note that there is no reason to think that a proposition like $<$Bias $= 0.3$ is the best value for the coin's propensity$>$ will likewise render the coin tosses conditionally independent. Hence, we cannot simply replace the Bias parameter with a different parameter without risking losing important relationships that hold between the evidence and the hypotheses.

The same points holds more generally: statisticians (rationally) prefer hypotheses that (1) entail probabilities for the possible evidence and (2) have suitably informative connections with the evidence. But a proposition such as $<\theta_i$ is the best hypothesis$>$ will generally not satisfy either (1) or (2). And that is probably why such hypotheses do not occur in statistical practice.

Hence, avoiding the interpretive problem by changing the algebra over

[11] For example,

$p(\text{Heads on second toss}|\text{Bias} = 0.3\&\text{Tails on first toss}) = p(\text{Heads on second toss}|\text{Bias} = 0.3).$

which $p$ ranges is not a workable solution to the interpretive problem. Other ways of avoiding the interpretive problem also fail to deliver. For example, Morey et al. (2013) assert that "...scientific models, including statistical models, are neither true nor false" (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 avoid the interpretive problem. And it is not clear how refusing to assign truth values to models avoids 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.

We have to face the interpretive problem head on, and if we are to face interpretive problem head on, then we have to face up to the fact that it really is an *interpretive problem*—the problem is that the standard probability axioms do not fit with how the Bayesian machinery is often applied in practice. To solve the problem, it follows that we will have to come up with a different interpretation of the Bayesian framework. For the remainder of the paper, I will consider two solutions to the interpretive problem. One solution involves interpreting conditional probabilities counterfactually rather than indicatively, while the other interpretation involves interpreting probabilities as what I will refer to as a "verisimilitude probabilities." As we will see, each interpretation necessitates a new version of the probability axioms.

# 5 Verisimilitude Probabilities

In cases where all the hypotheses under consideration are known to be false, the goal of Bayesian inference cannot reasonably be construed as discovering

the hypothesis that most probably is true. A natural proposal is that the goal in such cases changes to discovering which hypothesis is – in some sense – *closest* to the truth. Indeed, scientific realists have long held that the real (achievable) goal of inference is closeness to the truth rather than truth itself.

The idea that the goal of inference is to identify the $\theta$ that is closest to the truth leads to a natural reinterpretation of probability. Instead of interpreting $p_K(\theta)$ as the probability that $\theta$ is true, we interpret $p_K(\theta)$ as the probability that $\theta$ is closest to the truth out of the hypotheses in $\Theta$. I will call this interpretation of probability the "verisimilitude interpretation."

The reader may wonder how the verisimilitude interpretation differs from the earlier rejected suggestion of changing the algebra of hypotheses. Does the verisimilitude interpretation not just say that we ought to assign probabilities to propositions of the form $<\theta$ is closest to the truth$>$ rather than to $\theta$ itself? The answer is no. According to the verisimilitude interpretation, $p_K(\theta)$ is a probability that is assigned to $\theta$ itself, not to $<\theta$ is closest to the truth$>$. Thus, according to the verisimilitude interpretation:

> $p_K(\theta)$ = the probability that $\theta$ is closest to the truth out of the hypotheses in $\Theta$.

In other words, according to the the verisimilitude interpretation, a probability assignment to $\theta$ represents a *complex epistemic attitude* taken towards $\theta$; it does not represent a simple attitude taken towards a complex proposition.[12] This is important, because as we saw in the previous section, avoiding the interpretive problem by changing the algebra of propositions does not work.

So far the discussion of the verisimilitude interpretation has proceeded on an informal and intuitive level. To make the verisimilitude interpretation precise, more needs to be said about verisimilitude. The study of verisimilitude was initiated by Popper (1963) and has by now accumulated a large

[12] *Cf.* the point made by Moss (2018), although the the lesson drawn here is different.

literature.[13] The most influential contemporary approach in the study of verisimilitude – known in the literature as the "similarity approach" – understands verisimilitude as 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 is close to the truth is to say that the hypothesis is similar to the true hypothesis.

This idea can be formalized if we suppose that there is a (context-appropriate[14]) verisimilitude measure, $v$, that ranks hypotheses by how similar they are to the true hypothesis. If we presume that such functions are available, we can say that $\theta_1$ is closer to the truth than $\theta_2$ if and only if $v(\theta_1) > v(\theta_2)$. Here, we can be quite liberal in what we count as a "verisimilitude measure," though as a minimal requirement it is reasonable to suppose that $v$ be maximized by the true hypothesis, if the true hypothesis is one of the hypotheses under consideration. Later in the paper I will suggest a simple verisimilitude measure that makes sense in the earlier example concerning the relationship between windspeed and pressure.

Given a measure of verisimilitude, $v$, I will use $p_K^v$ with a $v$ superscript to indicate that the intended interpretation of $p_K^v$ is the verisimilitude interpretation with measure $v$. That is:

$$p_K^v(\theta) = \text{the probability that } \theta \text{ maximizes } v.$$

Note that the verisimilitude interpretation is consistent with either a subjective or objective Bayesian philosophy. On a subjective Bayesian reading,

[13] See Niiniluoto (1998) for a survey. Some of this literature has dealt with relationships between verisimilitude and Bayesianism (e.g. Rosenkrantz (1980), Niiniluoto (1986), Niiniluoto (1987), Festa (1993), Cevolani et al. (2010) and Oddie (ming)). However, no one in the verisimilitude literature has – to my knowledge – discussed the interpretive problem for Bayesian statistical inference.

[14] 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.

$p^v_K(\theta)$ would be interpreted as some particular agent's epistemic state, $K$ as that agent's background knowledge, and $v$ as the agent's preferred verisimilitude measure. On an objective reading, $p^v_K(\theta)$ would instead be interpreted as expressing a logical probability, $K$ as some objectively shared background knowledge, and $v$ as a verisimilitude measure that is "objectively proper" given the purpose at hand.

Moving from the standard interpretation of probability to the verisimilitude interpretation necessitates a suitable change in the probability axioms. Here is the verisimilitude version of the probability axioms:

> **Verisimilitude Probability Axioms** Suppose $\Theta$ indexes a set of hypotheses $\{\theta_1, \theta_2, \ldots, \theta_n\}$, let $v$ be a verisimilitude measure defined over the hypotheses indexed by $\Theta$, and let $K$ be a corpus of background knowledge. Then a distribution $p$ over $\Theta$ satisfies the verisimilitude probability axioms with respect to $v$ if and only if:
>
> > 1V. $p^v_K(\vee\theta_i) = 1$, whenever $K$ entails that at least one hypothesis in the disjunction of hypotheses indexed by $\vee\theta_i$ maximizes $v$.
> >
> > 2V. $p^v_K(\theta_i) \geq 0$ for all $\theta_i$ in $\Theta$.
> >
> > 3V. $p^v_K(\bigvee \theta_i) = \sum p^v_K(\theta_i)$, whenever $K$ entails that at most one of the hypotheses in the disjunction of hypotheses indexed by $\vee\theta_i$ maximizes $v$.

It is clear that by adopting the verisimilitude probability axioms we avoid the interpretive problem, because the fact that $K$ entails that all the hypotheses under consideration are false does not mean that $K$ will entail that none of the hypotheses under consideration will be closest to the truth. On the contrary, under commonly satisfied conditions, e.g. when the hypothesis space is closed and bounded and $v$ is continuous, then one of the hypotheses will be mathematically guaranteed to maximize $v$.

Note that, on the verisimilitude interpretation, the probability assigned to a hypothesis is relative to a given way of measuring verisimilitude. Consequently, in contrast to what is the case in standard Bayesian analysis, the verisimilitude prior probability of a hypothesis does not simply reflect background information. Instead, on the verisimilitude interpretation, the prior probability distribution is fundamentally goal-relative; its functional role in statistical analysis is to assign less weight to hypotheses that are likely to be further from the truth, given one's background knowledge and given the verisimilitude measure of interest.

# 6 The Verisimilitude Interpretation in Practice

The main purpose of this section is to illustrate, through an example, the abstract remarks made at the end of the previous section. More precisely, the goal is to show how it's possible to combine background information with a verisimilitude measure in a principled manner in order to derive rational constraints on verisimilitude probability distributions in a way that is very analogous to how background information leads to rational constraints on standard probability distributions. Thus, verisimilitude prior probability functions can play a role in inference that is very similar to the role played by standard prior probability functions in standard Bayesian inference. On the other hand, the example will also serve to show how pragmatic factors may influence what the rational constraints on the prior probability function turn out to be, and will thereby prepare the way for the argument in Section 10.

In order to get a sense of how this will work, it is helpful to first look at a simple example of how background knowledge can be incorporated in the prior distribution in a simple case where there is no interpretive problem. Suppose we are estimating the mass of a small cup of water, and suppose

we model the outcome of the measurement as a likelihood function $p_K(x|m)$, where $x$ is the outcome of the measurement and $m$ is a possible value of the cup's mass. The traditional frequentist (non-Bayesian) way of estimating the value of $m$ would be to take as our best estimate the value of $m$ that maximizes the probability of $x$—this is the maximum likelihood estimate. From a Bayesian point of view, maximum likelihood estimation is clearly suboptimal in this case because it fails to take into account background knowledge that we have about the reasonable masses of cups of water.

In particular we know that $m$ cannot be any negative value (the mass of an object cannot be a negative number). Furthermore, we know that a small cup of water will not weigh more than, say, 1kg. Therefore, at a minimum, our background knowledge entails that $m$ lies somewhere in the interval $[0, 1]$. The standard probability axioms, 1S-3S, then entail that we ought to assign every value of $m$ that lies outside of this interval a probability of 0. From a Bayesian point of view, this prior probability function can be expected to improve upon maximum likelihood estimation because it restricts the analysis to an area of the parameter space that is consistent with background knowledge. I will ague that verisimilitude probability distributions can play a similar role in cases where we face the interpretive problem.

Consider again the example concerning the relationship between barometric pressure ($X$) and maximum windspeed ($Y$). Let's use $f$ to denote the true (unknown) functional dependency of $Y$ on $X$. Now, suppose one of the things we know about the relationship between barometric pressure and windspeed is that changes in maximum windspeed are relatively insensitive to changes in barometric pressure, and suppose we also know the amount of maximal windspeed associated with the minimal pressure of interest.

So far, this is background knowledge about the actual, unknown function relating barometric pressure and windspeed. What consequences does this background knowledge about $f$ have for inferences about the hypothesis set actually under consideration? To simplify the example somewhat, suppose

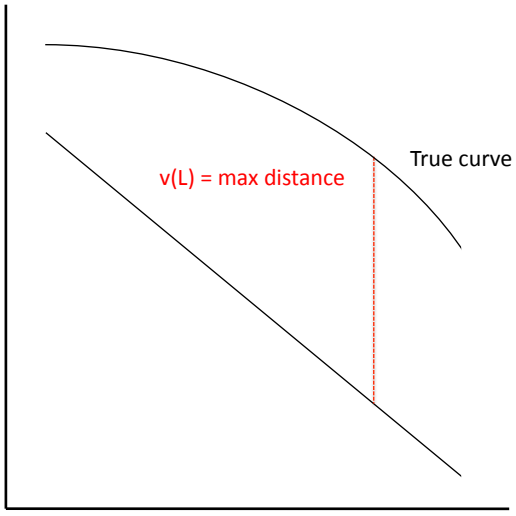


Figure 1: A measure of closeness to the truth

that rather than the hypotheses in (3.1), the set of hypotheses we are considering consists of lines. Suppose, moreover, that we know that $f$ is not a line. Can we use our background knowledge about $f$ to discriminate between the various false lines in a principled way? The answer is yes, but how our background knowledge affects the inferences we are entitled to make will depend on how we measure verisimilitude.

Suppose that our ultimate goal is to build a structure that will be able to withstand strong winds.[15] In that case, it is important that the maximal error we make when we estimate windspeed be as small as possible. In other words, Figure 1 is a natural measure of closeness to the truth given our goal; this is not to say that this is an appropriate way to measure closeness to the truth given other goals.

Mathematically, the verisimilitude of some straight line $L$ is given by the formula $v_{Max}(L) = -Max_{x\in[a,b]}|t(x) - L(x)|$, where $[a, b]$ is the range of relevant pressures. Given that we use $v$ to measure verisimilitude, and given

[15]I thank A for suggesting this example to me.

that we have restricted the analysis to the class of lines, the more immediate goal is to identify lines that are close to the truth according to $v$.

It is in fact easy to show that, under the given conditions, some (identifiable) lines will be further from the truth than others, given the way verisimilitude is measured and given our background knowledge—in particular, our background knowledge entails that certain lines that have a particularly steep slope cannot possibly be closest to the truth.[16]

Hence, the verisimilitude axioms, 1V-3V, entail that such lines ought to be assigned a probability of 0.

However, crucially, if closeness to the truth is measured in a different way, we do not necessarily get the same rational requirements on the prior distribution. Suppose, for example, that we are instead very concerned with the minimal rather than maximal distance of each line from the truth. That is, we use $w_{Min}(L) = -Min_{x\in[a,b]}|t(x) - L(x)|$ to measure the verisimilitude of each line (see Figure 2).

According to $w$, any line that intersects $f$ will be maximally close to the truth, and so our goal now is to identify the lines that intersect $f$. Clearly, lines that have a very steep slope will stand a better chance of intersecting $f$ than lines that do not, and thus if we use $w$ to measure verisimilitude, then it is rational to use a prior distribution that assigns more probability to lines that have a steep slope than to lines that have a more gradual slope; this is opposite of the result we get when we use the verisimilitude measure in

[16]For reasons of space, I have not included a complete demonstration of this fact, but here is a sketch: our background knowledge that changes in maximum windspeed are relatively insensitive to variations in barometric pressure may be formalized as knowledge that the derivative of $f$ is bounded by some known interval, $(a, b)$. Suppose, moreover, that the range of relevant pressures is contained in some known interval $(x_1, x_2)$, and that we know that $f(x_1) = w$. Then it is possible to show that if $L^*(x) = \alpha x + \beta$ is a line such that $L^*(x_1) > w$ and $\alpha > b$, then there is another line $L^1(x) = \alpha^1 x + \beta^1$ such that $L^1(x_1) < w$ and $\alpha^1 \in (a, b)$ such that $L^1$ is closer to the truth than $L^*$, according to the verisimilitude measure $v(L) = -Max_{x\in(x_1,x_2)}|f(x) - L(x)|$. The upshot is that our background knowledge entails that $L^*$ cannot possibly be closest to the truth. $L^*$ should therefore be assigned a probability of 0.

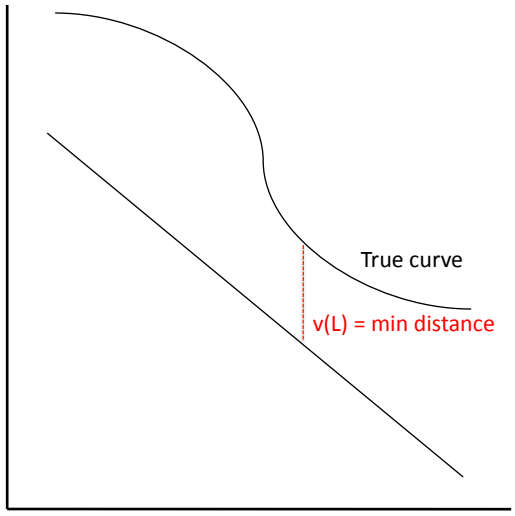


Figure 2: A different measure of closeness to the truth

Figure 1.

In general, how background knowledge interacts with a given measure of verisimilitude in order to induce rational requirements on the prior distribution is a subtle and complex question. 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 sets of known false hypotheses. My goal is rather to show how, in principle, background knowledge can be used to discriminate between multiple false hypotheses, provided we have a verisimilitude measure. As we have seen, the way verisimilitude is measured plays a crucial role in shaping the rational constraints on the prior; moreover, we have also seen that the way verisimilitude ought to be measured is reasonably influenced by the goals that we have.

It is worth emphasizing, once again, that regardless of how verisimilitude is measured, the prior probability distribution ranges over exactly the same set of hypotheses—in this case, the set of lines. The set of hypotheses does not change when we change the verisimilitude measure; rather, on the

verisimilitude interpretation, it is the probability function $p_K^v$ that changes. According to standard Bayesianism, the probability one should assign to any particular hypothesis is independent of one's goals, but this is no longer true for verisimilitude probabilities. Instead, the verisimilitude probability that it is rational to assign to a hypothesis is in part influenced by how verisimilitude is measured.

# 7 The Counterfactual Interpretation of Probability

The verisimilitude interpretation has the feature that the prior probability distribution incorporates not just background information, but also what one hopes to accomplish, formalized by way of a verisimilitude measure. Consequently, the verisimilitude probability that it is rational to assign to a hypothesis will be influenced by how verisimilitude is measured, which in turn will generally be influenced by pragmatic factors. In a very recent paper, Jan Sprenger (2017) proposes an alternative solution the interpretive problem. Sprenger's solution also involves reinterpreting the probability axioms, but he offers a reinterpretation that appears to be quite different from the verisimilitude interpretation. However, as we will soon see, given certain plausible assumptions, the verisimilitude solution and Sprenger's solution share many features in common and are even formally inter-translatable.

Sprenger's suggestion is that the probability of a false hypothesis can sensibly be interpreted as a *counterfactual* probability (or, more specifically, a counterfactual degree of belief. However, the counterfactual interpretation, like the verisimilitude interpretation, is consistent with either an objective or subjective reading). More precisely, suppose $\Theta$ is a set of hypotheses, all of which are known to be false. Then any probability assigned to some particular $\theta_i$ should be construed as the probability that $\theta_i$ is true *conditional* on the (false) supposition that one of the hypotheses in $\Theta$ is true. In other

words, the probability of $\theta_i$ is really the counterfactual *conditional* probability $p_K(\theta_i|\Theta)$, where the condition $\Theta$ is construed as the (false) claim that one of the hypotheses in $\Theta$ is true.

Note that $p_K(\theta_i|\Theta)$ cannot simply be replaced with $p_K(\Theta \rightarrow \theta_i)$, i.e. with a probability distribution defined over counterfactual propositions—the discussion on p. 9 applies equally here. $\theta_i$ picks out a hypothesis in a scientific and statistical model that makes probabilistic predictions, but $\Theta \rightarrow \theta_i$ does not.[17]

In order for the counterfactual interpretation to be a rigorous alternative semantics for Bayesian inference, something more substantive needs to be said about how we are supposed to understand and evaluate counterfactual probabilities. Unfortunately, Sprenger does not offer us any guidance. However, a natural thought is that counterfactual probabilities should be evaluated in a way that is analogous to the way counterfactual conditionals are evaluated. According to (a simplified version of) the standard analysis of counterfactuals due to Lewis (1973), evaluating a counterfactual such as "If $A$ were the case, then $B$ would be the case," involves considering the closest possible world in which $A$ is true, and then checking whether $B$ is true in that world. Crucially, Lewis's analysis depends on a ranking of possible 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 pressure and windspeed are linearly related, but presumably most of those possible worlds are not interesting or relevant. As is the case in the 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, we need to be able to rank worlds in terms of their closeness or similarity to the actual world. Suppose we have such a similarity

[17] In addition, replacing $p_K(\theta_i|\Theta)$ with $p_K(\Theta \rightarrow \theta_i)$ might run us into triviality result problems.

measure, $s$. Then we can define the counterfactual probability of $\theta_i$ given $s$, $p^s_K(\theta_i|\Theta)$, where $p^s_K$ must obey the following constraints:

> **Counterfactual probability axioms** Suppose $\Theta$ indexes a set of hypotheses $\{\theta_1, \theta_2, \ldots, \theta_n\}$, let $s$ be a similarity measure defined over the set of possible worlds, and let $K$ represent a corpus of background knowledge. Then a distribution $p$ over $\Theta$ satisfies the probability axioms with respect to $s$ if and only if:
>
> > 1V. $p^s_K(\vee\theta_i|\Theta) = 1$, whenever $K$ entails that one of the hypotheses in the disjunction $\vee\theta_i$ is true in the closest world (according to $s$) in which $\Theta$ is true.
> >
> > 2V. $p^s_K(\theta_i|\Theta) \geq 0$ for all $\theta_i$ in $\Theta$.
> >
> > 3V. $p^s_K(\bigvee \theta_i|\Theta) = \sum p^v_K(\theta_i|\Theta)$, whenever $K$ entails that at most one of the hypotheses in the disjunction of hypotheses indexed by $\vee\theta_i$ is true in the closest world (according to $s$) in which $\Theta$ is true.

The counterfactual interpretation, like the verisimilitude interpretation, solves the interpretive problem, because the fact that $K$ entails that $\theta_i$ is false does not mean that $K$ entails that $\theta_i$ is false in the closest possible world in which $\Theta$ is true. Hence, the counterfactual interpretation allows us to assign non-zero probabilities to hypotheses that we know are false (in the actual world).

It's clear that the counterfactual interpretation has the same broad features as the verisimilitude interpretation. In particular, on the counterfactual interpretation understood in the above Lewisian way, every probability assignment becomes relative to the way similarity between worlds is measured. Moreover, there are many ways of measuring similarity between worlds, but the way in which similarity between worlds should be measured is presumably relative to the features of the world that are relevant, and what features

are relevant is in part determined by the goals of the analysis. Indeed, in the next section we will see that the counterfactual and verisimilitude frameworks are plausibly inter-translatable, so that if verisimilitude probabilities are goal-relative, then so are counterfactual probabilities.

# 8 Relationship Between the Verisimilitude and Counterfactual Interpretations

At this point, we apparently have two viable reinterpretations of the Bayesian framework, both of which solve the interpretive problem. Many philosophers will be tempted to ask which of the two solutions is the better one. My contention is that neither solution is better than the other, and that in fact there is a sense in which the two solutions are equivalent.

Indeed, note that, in general, any similarity ranking of possible worlds straightforwardly induces a natural verisimilitude ranking of hypotheses, and vice versa. More precisely, suppose we are given a similarity ranking function, $s$, on worlds such that $s(w_\alpha) \geq s(w_1) \geq s(w_2) \geq \ldots$, where $w_\alpha$ 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 $s(w) \geq s(w')$.[18]

Conversely, any verisimilitude ranking induces an ordering over possible worlds. Suppose $v(H_0) \geq v(H_1) \geq v(H_2) \geq \ldots$ is a verisimilitude ranking of hypotheses, and for any hypothesis $H$, let $S_H$ denote the set of worlds in which $H$ 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 we define $s$ such that $s(w) \geq s(w')$ if and only if $v(H) \geq v(H)'$.

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

According to the verisimilitude interpretation, agents have to evaluate which hypothesis is plausibly closest to the truth out of the hypotheses under consideration. According to the counterfactual interpretation, agents must instead evaluate which hypothesis is plausibly true in the closest possible world in which one of the hypotheses under consideration is true—in other words, they must evaluate what the closest possible world is plausibly like. Since any verisimilitude ranking may be translated into a ranking of worlds, and vice versa, it's now clear that these two tasks are really one and the same. That is, if $s$ is the similarity ranking that is induced by the verisimilitude ranking $v$, then a hypothesis, $H$, will be closest to the truth according to $v$ if and only if $H$ is also true in the world that is closest to the actual world, according to $s$. Figuring out how probable it is that $H$ is closest to the truth according to $v$ is therefore equivalent to figuring out how probable it is that $H$ is true in the closest possible world according to $s$.

None of the above should really be that surprising since a similar fact is true of standard Bayesianism. There is a well known duality between propositions and possible worlds: a proposition may be construed as a set of possible worlds, and a possible world may be construed as a conjunction of propositions. Hence, an agent who has a degree of belief in a certain proposition may be regarded as implicitly having a degree of belief that the actual world is in a certain set of possible worlds, and vice versa. The correspondence between verisimilitude rankings and possible worlds rankings shown in this section demonstrates that the same is true of counterfactual and verisimilitude probabilities: any counterfactual probability may be regarded as an implicit verisimilitude probability, and vice versa.

Thus, although they may appear different, the verisimilitude interpretation and the counterfactual interpretation of probability are, in a sense, two sides of the same coin. This means that if there is pragmatic encroachment in the verisimilitude framework, there will also be pragmatic encroachment in the counterfactual framework. In particular, if the reader agrees that the ex-

ample in Section 6 plausibly shows that verisimilitude rankings are sometimes goal-relative, then the same example will also show that rankings of worlds are sometimes goal-relative, since the verisimilitude ranking may simply be translated into a ranking of possible worlds using the recipe provided in this section. It follows that the rational status of counterfactual probabilities will in general be goal-relative.

# 9 Relationship Between the Verisimilitude, Counterfactual, and Standard Interpretations

The preceding section investigated how the counterfactual and verisimilitude interpretations of probability relate to each other. But how do either of these interpretations relate to the standard interpretation? Recall that cccording to the standard interpretation, $p_K(H)$ is the probability that $H$ is true, relative to background knowledge $K$. Ideally, the verisimilitude and counterfactual interpretations should both be generalizations of the standard interpretation, so that both are extensionally equivalent to the standard interpretation in cases where the standard interpretation is applicable; i.e. in cases where $K$ entails that one of the hypotheses under consideration is true. Is that the case?[19]

The answer is that it depends on characteristics of the verisimilitude and counterfactual similarity measures. Let's first consider the verisimilitude interpretation. Let's call the true – but unknown – hypothesis $t$. Suppose $v$ is such that it has a *unique* maximum over the set of hypotheses under consideration, and that the unique maximum is $t$. According to the verisimilitude interpretation, $p^v_K(H)$ is the probability that $H$ is a maximum of $v$, relative to $K$, which, under the conditions specified, means that $p^v_K(H)$ is the prob-

[19]I thank C for pressing me on this issue.

ability that $H = H_t$ (since $H_t$ is the only maximum of $v$); in other words, $p^v_K(H)$ is simply the probability that $H$ is true, relative to $K$. Thus we have $p^v_K(H) = p_K(H)$. Hence, the verisimilitude interpretation is extensionally equivalent to the standard interpretation under the specified conditions in the sense that the the verisimilitude and standard probability distributions assign the same probabilities to all hypotheses. However, if $v$ has several maxima or if the truth is not among the maxima of $v$, then clearly $p^v(H)$ will not necessarily equal $p_K(H)$. Hence, the verisimilitude interpretation is extensionally equivalent to the standard interpretation just in case the following conditions are met: (1) $v$ has a unique maximum over the set of hypotheses, (2) that unique maximum is the truth.

Now let's consider the counterfactual interpretation of probability. Suppose the similarity ranking over possible worlds satisfies the following conditions: (1) there is a unique world that is closest to the actual world, (2) the actual world is closest to itself. Then, by essentially the same reasoning as above, it follows that we will have $p^s_K(H) = p_K(H)$. Hence, the counterfactual interpretation is extensionally equivalent to the standard interpretation just in case one of the hypotheses under consideration is true and the similarity ranking over possible worlds satisfies the constraint known in the counterfactuals literature as *strong centering*.

# 10 Pragmatic Encroachment in Bayesian inference

I have argued that the only adequate solutions to the interpretive problem in Bayesian statistical inference involve reinterpreting probability, and I have proposed two candidate reinterpretations. Both the counterfactual and verisimilitude interpretation have the following two important features: (1) they both depend on a ranking over some sort of object (either hypotheses or possible worlds), (2) the ranking that it is rational for an agent to have

is influenced by pragmatic factors, such as what the agent's goals are. The upshot is that whether a given probability assignment (i.e. verisimilitude or counterfactual probability) is rational is influenced by pragmatic factors.

Of course, the standard Bayesian interpretation also allows for pragmatic factors to play a role. According to standard Bayesian decision theory, we ought to have both a probability function and a utility function; any pragmatic factor – such as what we are interested in – should be relegated to the utility function. This neat separation between the purely epistemic and the pragmatic fails in cases where we face the interpretive problem. In those cases, I have argued that pragmatic factors should directly influence the probability function, not just the utility function.

The reader may wonder whether there are other potential solutions to the interpretive problem that would avoid having features (1) and (2). In Section 4, I argued that any solution to the interpretive problem needs to offer a reinterpretation of the probability axioms. A moment's reflection should make it clear that any re-interpretation that allows us to assign a non-zero probability to a known false hypothesis needs to involve a ranking of some sort: if $H_1$ and $H_2$ are both known to be false, and yet we assign a higher probability to $H_1$ than to $H_2$, there must be some sense in which $H_1$ is "better" than $H_2$. The remaining question, then, is whether there is a ranking of hypotheses (or other objects—of course, any ranking must implicitly be a ranking of the hypotheses, since we are ultimately assigning probabilities to the hypotheses) that can plausibly count as "objectively correct." Here, thinking about concrete examples – such as the example in Section 6 – should convince us that the answer is "no." Anyone who disagrees will have to explain why, say, the way you rank various lines in the example in Section 6 should be independent of your interests. Hence, my conjecture is that all adequate solutions to the interpretive problem will have features (1) and (2).

By combining the above considerations with a reasonable bridge premise, the following argument may now be formulated:

P1: All satisfactory solutions to the interpretive problem involve reinterpreting what it means to assign a probability to a hypothesis.

P2: Any satisfactory reinterpretation that solves the interpretive problem will have the following two features: (1) it will depend on a ranking over some sort of object, (2) whether a given ranking is rational will in part be determined by pragmatic factors.

P3: If P1 and P2, then whether a given Bayesian probability distribution is rational will, in general, partly be determined by pragmatic factors.

C: Whether a given Bayesian probability distribution is rational will, in general, partly be determined by pragmatic factors.

The upshot of this argument is that there is an important – and hitherto unnoticed – kind of pragmatic encroachment on Bayesian inference.

In recent years, there has been much debate over whether there is sometimes "pragmatic encroachment" on the epistemic, i.e. whether pragmatic factors can sometimes influence whether an agent, for instance, knows whether a proposition is true.[20] As Mark Schroeder (2017) point outs, it seems to be almost universally agreed among participants of this debate that although there may be pragmatic encroachment on knowledge or rational (full) belief, there is no pragmatic encroachment on Bayesian probability functions. Prominent experts on Bayesian statistical theory agree, including adherents of the subjective (Lindley, 1972, p. 71) and objective (Jaynes, 2003, p. 19) schools of Bayesianism. However, despite this theoretical consensus, in practice Bayesian statisticians tend to use different prior probability distributions

[20]See e.g. Stanley (2005), Fantl and McGrath (2002), Ross and Schroeder (2014), Rubin (2015), or Roeber (2016)

depending on what they are interested in.[21] The arguments in this paper partially undermine the theoretical consensus and lend a justification of statistical practice. Whereas it may be true that there is no pragmatic encroachment on standard Bayesian probability functions, there is – and ought to be – significant pragmatic encroachment on both counterfactual and verisimilitude probabilities, and those are the types of probability distributions that are frequently (implicitly) used in statistical practice.

# 11 Conclusion

This paper has mainly been concerned with the implications of the interpretive problem for our interpretation of the prior probability distributions that are used in Bayesian statistical practice. I have not said anything about the likelihood, but in fact the interpretive problem arguably has even greater implications for how we are to interpret, and use, the likelihood function and associated principles such as the Law of Likelihood and Conditionalization. In particular, although I will not argue this here, the counterfactual and verisimilitude interpretations open the door to the possibility that it may sometimes be rational to use an evidential measure other than the likelihood and an updating procedure other than Conditionalization. This is because the standard arguments for Conditionalization turn out to depend crucially on the standard interpretation of probability. Thus, although this paper has been concerned with showing that we sometimes need to change the standard Bayesian *semantics*, once we have a new semantics, it becomes apparent that we may sometimes be justified in also changing the standard Bayesian *syntax*.

[21] I thank a referee for pointing this out.

# References


Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, New York, NY.

Bissiri, P. G. and Holmes, C. and Walker, S. (2016). A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 78(5):1103-1130.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430.

Cevolani, G., Crupi, V., and Festa, R. (2010). The Whole Truth About Linda: Probability, Verisimilitude, and a Paradox of Conjunction. In D'Agostino, M., Laudisa, F., Giorello, G., Pievani, T., and Sinigaglia, C., editors, *New Essays in Logic and Philosophy of Science*, pages 603–615. College Publications.

Choi, J. W., Cha, Y., Kim, H.-D., and Lu, R. (2016). Relationship between the Maximum Wind Speed and the Minimum Sea Level Pressure for Tropical Cyclones in the Western North Pacific. DOI: 10.4172/2332-2594.1000180.

De Heide, R. and Grunwald, P. D. (2018). Why Optional Stopping Is a Problem for Bayesians.

Fantl, J. and McGrath, M. (2002). Evidence, Pragmatics, and Justification. *Philosophical Review*, 111(1):67–94.

Festa, R. (1993). *Optimum Inductive Methods: A Study in Inductive Probability, Bayesian Statistics, and Verisimilitude*. Synthese Library. Springer Netherlands.

Forster, M. R. (1995). Bayes and bust: Simplicity as a problem for a probabilist's approach to confirmation. *British Journal for the Philosophy of Science*, 46(3):399–424.

Forster, M. R. and Sober, E. (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):1–35.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38.

Heggarty, P., Maguire, W., and McMahon, A. (2010). Splits or Waves? Trees or Webs? How Divergence Measures and Network Analysis can Unravel Language Histories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3829–3843.

Hewitt, E. and Savage, L. J. (1955). Symmetric Measures on Cartesian Products. *Transactions of the American Mathematical Society*, 80:470-501.

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

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Key, J. T., Pericchi, L. R., and Smith, A. F. M. (1999). Bayesian Model Choice: What and Why? In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 343–370. Oxford: Oxford University Press.

Knaff, J. A. and Zehr, R. M. (2007). Reexamination of Tropical Cyclone Wind–Pressure Relationship. *Weather and Forecasting*, 22(1):71–88.

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

Lindley, D. V. (1972). *Bayesian Statistics, A Review*. Capital City Press, Montpellier, Vermont.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2013). The Humble Bayesian: Model Checking From a Fully Bayesian Perspective. *British Journal of Mathematical and Statistical Psychology*, 66(1):68–75.

Moss, S. (2018). *Probabilistic Knowledge*. Oxford University Press.

Niiniluoto, I. (1986). Truthlikeness and Bayesian Estimation. *Synthese*, 67(2):321–346.

Niiniluoto, I. (1987). *Truthlikeness*. Synthese Library. Springer Netherlands.

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

Northcott, R. (2013). Verisimilitude: A Causal Approach. *Synthese*, 190(9):1471–1488.

Oddie, G. (Forthcoming). What Accuracy Could Not Be. *British Journal for the Philosophy of Science*.

O'Malley, M. A., Martin, W., and Dupre, J. (2010). The Tree of Life: Introduction to an Evolutionary Debate. *Biology & Philosophy*, 25:441–453.

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

Roeber, B. (2016). The Pragmatic Encroachment Debate. DOI: 10.1111/nous.12156.

Rosenkrantz, R. (1980). Measuring Truthlikeness. *Synthese*, 45(3):463–487.

Ross, J. and Schroeder, M. (2014). Belief, Credence, and Pragmatic Encroachment. *Philosophy and Phenomenological Research*, 88(2):259–288.

Rubin, K. (2015). Total Pragmatic Encroachment and Epistemic Permissiveness. *Pacific Philosophical Quarterly*, 96(1):12–30.

Schroeder, M. (2017). Rational Stability under Pragmatic Encroachment.

Shaffer, M. J. (2001). Bayesian Confirmation of Theories That Incorporate Idealizations. *Philosophy of Science*, 68(1):36–52.

Sober, E. (2009). Absence of Evidence and Evidence of Absence – Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads. *Philosophical Studies*, 143:63–90.

Sprenger, J. (2009). Statistics Between Inductive Logic and Empirical Science. *Journal of Applied Logic*, 7(2):239–250.

Sprenger, J. (2017). Conditional Degree of Belief. Unpublished manuscript.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford: Oxford University Press.

Velasco, J. D. (2012). The Future of Systematics: Tree Thinking Without the Tree. *Philosophy of Science*, 79(5):624–636.

Walker, S. G. (2013). Bayesian Inference with Misspecified Models. *Journal of Statistical Planning and Inference*, 143:1621?1633.