The Verisimilitude Framework for Philosophy of Science, with an Application to Phylogenetic Inference

Olav B. Vassend

November 29, 2017

Abstract

Bayesianism and likelihoodism are two of the most important frameworks philosophers of science use to analyze scientific methodology. However, both frameworks face a serious objection: much scientific inquiry takes place in highly idealized frameworks where all the hypotheses are known to be false. Yet, both Bayesianism and likelihoodism seem to be based on the assumption that the goal of scientific inquiry is always truth rather than closeness to the truth. Here, I argue in favor of a verisimilitude framework for philosophy of science. In the verisimilitude framework, scientific inquiry is conceived, in part, as a process where inference methods ought to be calibrated to appropriate measures of closeness to the truth. To illustrate the verisimilitude framework, I offer a reconstruction of parsimony evaluations of scientific theories, and I give a reconstruction and extended analysis of the use of parsimony inference in phylogenetics. By recasting phylogenetic inference in the verisimilitude framework, it becomes possible to both raise and address objections to phylogenetic methods that rely on parsimony.
1 Introduction

Suppose you have evidence that bears on a number of competing hypotheses. How should you determine which hypothesis is most plausible in light of the evidence? According to likelihoodists, you should come up with a likelihood function that says how probable each hypothesis makes the evidence. The likelihood of a particular hypothesis is then a measure of the extent to which the evidence supports the hypothesis. Bayesians agree that likelihoods are important, but argue that the likelihood function should be supplemented with a prior probability distribution that reflects how plausible each hypothesis is before the evidence is taken into account, and that the prior and likelihood should be combined using Bayes’s formula in order to produce a posterior probability distribution over the various hypotheses. According to Bayesians, the posterior probability of a given hypothesis reflects how plausible the hypothesis is, all things considered.

---

1Bayes’s formula has the following form, where $p(H)$ is the prior distribution of $H$, $p(E|H)$ is $H$’s likelihood on $E$, and $p(H|E)$ is the posterior probability of $H$ given $E$: $p(H|E) = \frac{p(E|H)p(H)}{\sum_i p(E|H_i)p(H_i)}$.

2This is a simplified description of Bayesian inference. More sophisticated Bayesian statistical inference crucially involves predictive model checking (see, e.g., Gelman et al. (1996), Gelman and Shalizi (2013)) and various forms of robustness analysis. Furthermore, there’s also now a version of Bayesianism that uses a so-called “confirmation measure” as the basis for inferences, rather than
Some sciences use explicitly Bayesian or likelihoodist methodology, but many do not. Many disciplines have organically developed their own ways of evaluating hypotheses. It may be that every scientific discipline would be better off if it adopted Bayesian or likelihoodist methodology—but this is contentious. It is not always easy to come up with well-motivated numerical prior probabilities or likelihoods for hypotheses.

Meanwhile, likelihoodism and Bayesianism have proven to be quite useful in philosophy of science for illuminating the inferential methods that scientists actually employ. By treating these frameworks as normative, it is possible to get a better understanding of when and why, exactly, an inferential rule or principle may be expected to lead scientists towards the truth. In this way, philosophers have in recent years given likelihoodist or Bayesian accounts of, for example, inference to the best explanation (Cabrera, 2017), the comparative method in historical linguistics (Okayasu, 2017), parsimony reasoning (Sober, 2015), “no-alternatives” style arguments (Dawid et al., 2015), and robustness analysis (Schupbach, ming).

Unfortunately, even if they are construed only as normative frameworks rather than inferential procedures that should always be carried out in detail, Bayesianism and likelihoodism face serious objections. The most serious objection is that both frameworks seem to have built into them the assumption that discovering the true hypothesis is always the goal of scientific inquiry. As philosophers of science have realized for decades, scientific inquiry tends to be open-ended—the true hypothesis is rarely one of the hypotheses under consideration. Even worse, the set of hypotheses is often framed on the basis of highly idealized assumptions that essentially guarantee that all the hypotheses under consideration are false. Hence, closeness to the truth—not truth—is the realistic goal of scientific inference in many if not most cases (as also emphasized by Forster (2002)).

But if closeness to the truth replaces truth as the goal of scientific inquiry, then—as I explain in the next section—the usual arguments for using the likelihood as a measure of evidential favoring fail to go through. This raises the possibility that the likelihood may not always be appropriate. In this paper, I show that—indeed
the likelihood is not always an appropriate way of measuring evidential favoring. The evidential principle that is usually associated with the likelihood is the so-called Law of Likelihood; hence my criticism of the likelihood will proceed by way of a criticism of the Law of Likelihood. I will then suggest a different framework for understanding inductive inference which I call the “verisimilitude framework.” In the verisimilitude framework, the goal of scientific inference is conceived of, in part, as identifying a measure of evidential favoring that is calibrated against an appropriate measure of closeness to the truth. I illustrate the verisimilitude framework with two examples: the first example is a speculative reconstruction of parsimony evaluations of scientific theories. The second example is a reconstruction of parsimony inference in phylogenetics. Importantly, we will see that by reconstructing phylogenetic parsimony inference in the verisimilitude framework, we will be in a better position to understand its limitations and possible improvements.

2 Problems with the likelihood and the law of likelihood

The evidential principle that most commonly is taken to justify using the likelihood as a measure of evidential favoring is the Law of Likelihood (LL):

\[
\text{Law of Likelihood, uninterpreted version: Evidence } E \text{ favors } H_1 \text{ over } H_2 \text{ if and only if } p(E|H_1) > p(E|H_2).
\]

The above version of LL is “uninterpreted” in the sense that it’s left imprecise how, exactly, we are to understand the favoring relation. The following is arguably the standard interpretation:

\[\text{Law of Likelihood, interpreted version: }\]

As far as I know, likelihoodists universally accept the Law of Likelihood. In the Bayesian framework, there are multiple versions of the Law of Likelihood (Festa and Cevolani, 2017), and there has been some debate over which version is the best one. Here I will discuss counter-examples to the standard version of the Law of Likelihood, but I hope it’s clear to those concerned that my counter-examples are equally counter-examples to the other variants of the Law of Likelihood.
**Law of Likelihood, standard interpretation**: Evidence \( E \) favors the proposition that \( H_1 \) is true over the proposition that \( H_2 \) is true if and only if \( p(x|H_1) > p(x|H_2) \).

Both likelihoodists and Bayesians tend to accept the standard interpretation of LL. Indeed, Bayes’s theorem all by itself ensures that Bayesianism is committed to at least a “thin” version of LL. This is most easily seen from the odds formulation of Bayes’s theorem, which states that the ratio of the posterior probabilities of two hypotheses is equal to the ratio of their prior probabilities multiplied by the ratio of their likelihoods. That is, given \( E, H_1, \) and \( H_2 \), we have:

\[
\frac{p(H_1|E)}{p(H_2|E)} = \frac{p(E|H_1)}{p(E|H_2)} \cdot \frac{p(H_1)}{p(H_2)} \tag{2.1}
\]

Here, we see that \( E \) “favors” \( H_1 \) over \( H_2 \) if and only if \( p(E|H_1) > p(E|H_2) \) in the rather weak sense that the probability of \( H_1 \) will receive a boost relative to the probability of \( H_2 \) if and only if \( p(E|H_1) > p(E|H_2) \).

However, the odds formula, by itself, does not support the standard interpretation of LL; for that, additional arguments are needed. Typically, various mathematical results are invoked for this purpose. For example, according to one well known type of result, the hypothesis that has the highest likelihood converges (in probability) to the true hypothesis as the amount of evidence increases (under various conditions; see, e.g. Hawthorne (1994)). Another relevant result is the “Universal Bound” (Royall, 1997), which says that the probability that the ratio of the likelihood of the true hypothesis to the likelihood of a false hypothesis will exceed \( k \) is less than or equal to \( 1/k \). In other words, most evidence will not result in a false hypothesis having a much larger likelihood than the true hypothesis. The preceding mathematical theorems – and perhaps a few other ones\(^4\) – justify interpreting the observation that \( H_1 \) has a higher likelihood than \( H_2 \) as favoring the proposition that \( H_1 \) is true over the proposition that \( H_2 \) is true, at least in cases where the assumptions of those theorems

---

\(^4\)For example, purported proofs for the related “Likelihood Principle,” which states that all the evidential information contained in an experiment relevant to a set of hypotheses is contained in the likelihoods of those hypotheses (see Birnbaum (1962); Gandenberger (2015)).
are satisfied. But in cases where the assumptions of the theorems are not satisfied, the standard interpretation of LL is not obviously justified. Hence, Bayesianism is by no means automatically committed to the standard interpretation of LL, even though it is committed to the previously mentioned thin version of LL.

Not only is Bayesianism not automatically committed to the standard interpretation of LL; sometimes, the standard interpretation plainly fails in the sense that the fact that a hypothesis has a high likelihood in no way indicates that the hypothesis is true. As a simple illustration, suppose we have just two hypotheses under consideration, $H_1$ and $H_2$, and that both hypotheses are known to be false. Then if $p(E|H_1) > p(E|H_2)$, $H_1$’s probability will be boosted relative to $H_2$’s probability, but that does not mean that $E$ supports the proposition that $H_1$ is true over the proposition that $H_2$ is true—after all, both hypotheses are already known to be false!

It happens quite often that all the hypotheses under consideration are known to be false, because scientific hypothesizing very often (perhaps usually) takes place in highly idealized frameworks. To take just one example to which I will return later in the paper, phylogenetic hypotheses in both biology and historical linguistics almost universally assume a “tree model” of evolution. Figure 1 gives an example of what a typical linguistic tree looks like.

Even though phylogeneticists use trees to model phylogenetic relationships, it is widely acknowledged among biologists and linguists alike that all trees rest on several highly idealized and false assumptions (See e.g. O’Malley et al. (2010), Heggarty et al. (2010), and Velasco (2012)), including the following three: (1) evolutionary divergence happens through bifurcations, (2) divergences happen instantaneously, (3) descendant lineages do not influence each other after splitting. The last assumption, in particular, is known to be false both in biology and historical linguistics. In biology, species sometimes hybridize, and in linguistics it often happens that languages influence each other through borrowing. Hence, the real question isn’t whether biological or linguistic history can be represented in the form of a tree, but rather how tree-like the actual histories are. Even so, most phylogeneticists restrict themselves to considering only tree hypotheses, and in that case the standard law of likelihood is violated.
When all the hypotheses under consideration are known to be false, Bayesians face the additional problem of how to sensibly interpret the Bayesian prior and posterior probabilities of hypotheses. According to the standard Bayesian interpretation, the prior or posterior probability assigned to a hypothesis is supposed to represent the plausibility that the hypothesis is true. However, this interpretation clearly does not work in cases where all the hypotheses are already known to be false. Recently, Sprenger (2017) and Vassend (manuscript) have suggested two reinterpretations of the Bayesian framework that are each intended to address this interpretive problem. Sprenger’s proposal is that probabilities assigned to known false hypotheses should be understood as counterfactual degrees of belief. That is, \( p(H) \) is construed as a counterfactual degree of belief that \( H \) would be true if it were the case that one of the hypotheses under consideration were true. Vassend’s suggestion is that such probabilities should be understood as verisimilitude degrees of belief; i.e. \( p(H) \) is construed as a degree of belief that \( H \) is closest to the truth out of the hypotheses under consideration. Both of these interpretations have implications for how LL ought to be interpreted. Indeed, on the verisimilitude interpretation, the natural
interpretation of the LL is as follows:

**Law of Likelihood, verisimilitude interpretation:** Evidence \( x \) favors the proposition that \( H_1 \) is closest to the truth over the proposition that \( H_2 \) is closest to the truth (out of the hypotheses under consideration) if and only if \( p(x|H_1) > p(x|H_2) \).

On the other hand, the natural counterfactual interpretation of LL is as follows:

**Law of Likelihood, counterfactual interpretation:** Evidence \( x \) favors the proposition that \( H_1 \) would be true over the proposition that \( H_2 \) would be true (if the world were such that one of the hypotheses under consideration were true) if and only if \( p(x|H_1) > p(x|H_2) \).

The verisimilitude and counterfactual interpretations of LL are both better than the standard interpretation in the sense that they stand a better chance of being satisfied when all the hypotheses under consideration are false. The question is, however, whether the counterfactual and verisimilitude interpretations may still be violated. The answer, as I will argue, is yes. For simplicity, I will only discuss the verisimilitude interpretation. As argued by Vassend (manuscript), the counterfactual and verisimilitude interpretations are inter-translateable—they are two equivalent formulations of what is essentially the same interpretation. Hence, there is no loss in restricting discussion to the verisimilitude interpretation.

Some of the clearest violations of the verisimilitude interpretation of LL happen in model selection. As Forster and Sober (1994) point out, predictive accuracy may be regarded as a kind of verisimilitude measure. Moreover, as Forster and Sober (2004) show, if there are multiple models, and if the competing models have different numbers of parameters, then the model that has the highest likelihood will often not be the model that is most predictively accurate. Hence, the likelihood is not an appropriate measure of evidential favoring if the verisimilitude measure is predictive accuracy, because the verisimilitude interpretation of LL will not be satisfied in those cases. Nonetheless, Forster and Sober maintain that the likelihood is a fine
evidential measure in cases where the competing hypotheses are not models with different numbers of adjustable parameters.

However, contrary to what Forster and Sober claim, there are violations of the verisimilitude interpretation of LL even when the inferential problem is not selecting between models that have different numbers of parameters. Here’s a simple example of an extreme violation of this kind. Suppose we are concerned with identifying the functional relationship between two quantities $y$ and $x$, where – for example – $x$ and $y$ represent the minimal pressure and maximum windspeed in tropical storms, respectively. Suppose we model the relationship between $x$ and $y$ as a line (with measurements distributed stochastically about the line), even though we know that strictly speaking the true relationship, $t$, is not a line. Since $t$ is not among our hypotheses, we need some way of measuring the degree to which our hypotheses fall short of the truth. That is, we need a measure of verisimilitude. There are many candidates and the choice we make should be guided by our interests. Suppose we decide to use the measure in Figure 2.\textsuperscript{5}

\textsuperscript{5}Admittedly, justifications for why we would be interested in this measure are hard to come by, but that is irrelevant to the main point that is being made.
According to this measure, the verisimilitude of line $L$ is equal to the minimal vertical distance between $t$ and $L$ in the range of interest. Obviously, lines that intersect the truth will be maximally close to the truth; moreover, within a given range of pressures, lines that have a very steep slope will be more likely to intersect the truth. Hence lines that have a very steep slope can a priori be expected to be closest to the truth in the relevant sense. Yet, such lines are also sure to have a very low likelihood. Hence, in this case there is no overlap between the set of hypotheses that will have a high likelihood as the evidence comes in (and hence high posterior probability) and the set of hypotheses that are close to the truth in the sense we care about.

The above example is an extreme example of a violation of the verisimilitude interpretation of LL: it’s an example where the likelihood can be expected to pull in the opposite direction of what we want given the verisimilitude measure in which we are interested. However, even in cases where there isn’t an extreme violation, it may still be that the likelihood is not an optimal evidential measure. All of these considerations motivate the search for measures of evidential favoring other than the likelihood. Indeed, in Bayesian statistical inference, modifying or replacing the likelihood has recently been explored by, among others, Zhang (2006), Jiang and Tanner (2008), and Bissiri et al. (2016). Here, the same possibility will be explored in the verisimilitude framework. The guiding idea is that the evidential measure should be calibrated to whatever the appropriate measure of verisimilitude is.

3 Introducing Verisimilitude-Based Inference

As was suggested at the end of the previous section, any decent measure of evidential favoring ought to be suitably calibrated to whatever the appropriate measure of verisimilitude happens to be. But what does it mean for a measure of evidential favoring to be “calibrated to” a verisimilitude measure? To get a better grip on this question, note that given any measure of verisimilitude and measure of evidential favoring

---

6The reason for this is that the likelihood of a hypothesis is a measure of the average fit of the hypothesis over all the evidence.
favoring, there is an associated law of evidential impact:

**Law of Evidential Impact:** Evidence $E$ favors the proposition that $H_1$ is closest to the truth over the proposition that $H_2$ is closest to the truth if and only if $\text{Ev}(E|H_1) > \text{Ev}(E|H_2)$.

Intuitively, evidential measure $\text{Ev}$ is calibrated to verisimilitude measure $v$ (or, perhaps better, they are calibrated to each other) if and only if their associated law of evidential impact is satisfied, given most pieces of evidence. As a minimal constraint, it needs to be the case that — as the amount of evidence increases to infinity — $\text{Ev}$ favors those hypotheses that are closest to the truth according to $v$. More precisely, the following consistency constraint ought to be satisfied:

**Consistency Constraint** Suppose $H_t$ is a hypothesis that is maximally close to the truth according to $v$ and suppose $H_f$ is a hypothesis that is not maximally close to the truth. Then, as the amount of evidence $E = E_1, E_2, \ldots, E_n$ increases to infinity, $\text{Ev}$ is consistent with $v$ if and only if the following inequality is true:

$$\text{Ev}(E|H_t) > \text{Ev}(E|H_f)$$

(3.1)

For example, to take our earlier case concerning the relationship between min pressure and max windspeed, if minimal vertical distance from the truth is used to measure verisimilitude and the likelihood is used to measure evidential impact, then the Consistency Constraint will be plainly violated because even after arbitrarily large amounts of evidence, the hypotheses that are closest to the truth (i.e. the lines that intersect the truth) will often have significantly lower likelihoods than hypotheses that are further from the truth (e.g. any line that does not intersect the truth, but is close to the truth on average).

Even diehard likelihoodists agree that the likelihood ought to be combined with a prior probability distribution, given that a well-justified prior probability distribution is available. Hence, in addition to being calibrated to the appropriate verisimilitude
measure, any decent evidential measure needs to be capable of replacing the likelihood in Bayesian calculations. Now, Bayesian updating is clearly an instance of the following more general schema:

\[
\text{(Posterior plausibility of } H \text{ given } x_1, \ldots, x_n) \propto \text{(Evidential impact on } H \text{ of } x_1, \ldots, x_n) \times (\text{Initial plausibility of } H)
\]

Here \(\propto\) means “is proportional to.” Let’s call any inference procedure that instantiates the above schema “quasi-Bayesian.” Thus, the additional requirement we make of Ev is that it be capable of entering into quasi-Bayesian inferences. Now, the likelihood has the crucial property that the likelihood of a hypothesis given two pieces of evidence can be decomposed as the evidential impact that the first piece of evidence has on the hypothesis alone multiplied by impact that the second piece of evidence has \textit{given} the first one. In other words, \(p(E_1, E_2|H) = p(E_1|H) \times p(E_2|H & E_1)\). To be capable of replacing the likelihood in Bayesian calculations, any evidential measure therefore needs to have the following formally equivalent property:

**Coherence Requirement** \(Ev(x_1, x_2|H) = Ev(x_1|H, x_2) \times Ev(x_1|H)\)

Note that what the Coherence Requirement in effect demands is that whether we evaluate the evidential impact that \(E_1\) and \(E_2\) have on \(H\) by considering the two pieces of evidence together or one after the other, we end up with the same verdict of how much \(E_1\) and \(E_2\) jointly impact \(H\). Clearly, this is a property that any good measure of evidential impact ought to have, hence the Coherence Requirement is independently motivated.\(^7\)

Whenever \(Ev(x_1|H, x_2) = Ev(x_1|H)\), the Coherence Requirement entails that \(Ev(x_1, x_2|H) = Ev(x_1|H) \times Ev(x_2|H)\), i.e. the evidential impact that \(x_1\) and \(x_2\) jointly have on \(H\) is equal to the product of the evidential impact that they have separately. Now, when \(p(x_1|x_2&H) = p(x_1|H)\), \(x_1\) and \(x_2\) are standardly said to be “probabilistically independent” conditional on \(H\). So let’s define the following more general concept:

\(^7\)Coherence of this type is heavily emphasized by E. T. Jaynes (2003) and also by Bissiri et al. (2016).
Evidential Independence Given evidence $x_1$ and $x_2$ and evidential measure $\text{Ev}$, $x_1$ and $x_2$ are \textit{evidentially independent} conditional on $H$ if and only if $\text{Ev}(x_1|H, x_2) = \text{Ev}(x_1|H)$ and $\text{Ev}(x_2|H, x_1) = \text{Ev}(x_2|H)$.

“Evidential independence” and “probabilistic independence” are often treated as synonymous in the philosophical literature, but as we will see in the next section, there are other ways in which several pieces of evidence can be independent.

The Consistency Constraint and Coherence Requirement are the only general requirements I will make; in the next two sections, we will see how these requirements, and the particulars of the case, lead to particular evidential measures.

4 Examples of verisimilitude-based inference procedures

4.1 Parsimony inference over theories

The verisimilitude framework is general and flexible enough that it allows us to reconstruct inferential methods that cannot be adequately reconstructed in a traditional likelihoodist or Bayesian framework. To illustrate the generality of verisimilitude-based inference, I will start with a rather speculative example. My discussion of this example is mostly intended to illustrate how inferential methods are reconstructed in the verisimilitude framework; I do not claim that the resulting reconstruction will actually be philosophically useful. However, some of the lessons drawn from the example in this section will carry over to the example I discuss in the next section, where I discuss phylogenetic inference, and I do contend that the verisimilitude framework provides a philosophically useful reconstruction of phylogenetic inference.

Suppose we are interested in evaluating scientific theories (e.g. the theory of evolution, quantum mechanics, etc) and that we accept the following (no doubt controversial) ontological parsimony principle: the true theory is the one that is as simple as possible while accounting for all relevant empirical phenomena.\footnote{I’m using the rather vague phrase of a theory’s “accounting for the phenomena” because the} Of
course, measuring simplicity is a fraught matter, but suppose we choose to measure the simplicity of a theory by the sum of the number of basic entities posited by the theory and the number of ad hoc auxiliary assumptions that must be added to the theory in order for it to account for all the relevant phenomena. This understanding of simplicity is consonant with the traditional *vera causa* understanding of parsimony summarized by Isaac Newton in his first and first rule of reasoning (Clatterbuck 2016). If you add enough ad hoc auxiliary assumptions, just about any theory can account for just about any empirical phenomenon, but a theory that is close to the truth will require comparatively fewer ad hoc assumptions than a theory that is far from the truth. Thus, on this construal of simplicity, our parsimony-motivated measure of verisimilitude is as follows:

**Parsimony verisimilitude measure:** The verisimilitude of a theory $T$ is the sum of the number of basic entities posited by $T$ and the number of auxiliary assumptions that must be added to $T$ in order for $T$ to account for all the relevant empirical phenomena.

Suppose we want a verisimilitude-based inference procedure that will lead us to infer theories that are close to the truth in the above sense. The first thing we need to do is come up with an appropriate measure of evidential impact. For the moment, let’s suppose that all of the theories under consideration posit the same number of basic entities; in that case, the theory that is closest to the truth will simply be the theory that requires the fewest number of auxiliary assumptions. Let $n(D|T)$ be a function that has as its input a data set $D$ and a theory $T$ and that outputs the

\[ n(D|T) \]

criteria for when a theory accounts for a set of phenomena vary between different disciplines. Also the “relevant” qualifier has been added since it would clearly be unreasonable to require, say, quantum mechanics to account for phenomena in economics.

\[ ^9 \text{I will put aside the vexed and well known question of how to count the number basic entities or the number of auxiliary assumptions.} \]

\[ ^{10} \text{Newton arguably endorsed an ontological, not merely methodological, reading of the vera causa principle, as evidenced by the following famous quote:} \]

\[ “\text{Nature does nothing in vain, and more causes are in vain when fewer suffice. For nature is simple and does not indulge in the luxury of superfluous causes}” \text{ (Newton 2003).} \]
minimal number of ad hoc auxiliary assumptions that must be added to $T$ in order to account for $D$. Then, intuitively, any parsimony-tracking evidential measure of the evidential impact that $D$ has on $T$ should be some function of $n$, and in particular it should be a monotonically decreasing function of $n$, since a theory that requires more ad hoc assumptions is worse than one that requires fewer. That is, if $Ev$ is the evidential measure, then we should have $Ev[D|T] = F[n(D|T)]$ where $F$ is some monotonically decreasing function. Any measure of evidential impact that has this form will plausibly satisfy the Consistency Constraint, whereas any measure that does not will not.

How should the conditional impact that $D_1$ has on $T$ given $D_2$ be defined? Intuitively, we should not “double count” any auxiliary assumptions. Hence, the evidential impact that $D_1$ has on $T$ given that we have already considered $D_2$ ought to be a function of the number of auxiliary assumptions required to account for $D_1$ minus the auxiliary assumptions required to account for $D_1$ that are also required to account for $D_2$. In other words, we define $Ev[D_1|T&D_2] = F[n(D_1&D_2|T) - n(D_2|T)]$. Note that this implies that $D_1$ is evidentially independent of $D_2$ conditional on $T$ if and only if $n(D_1&D_2|T) - n(D_2|T) = n(D_1|T)$, i.e. if and only if the number of auxiliary assumptions required to account for $D_1&D_2$ is just the sum of the auxiliary assumptions required to account for $D_1$ and $D_2$ separately. Clearly, this will happen if and only if the auxiliary assumptions required to account for the two data sets are completely non-overlapping. Hence, the relevant notion of evidential independence at work here is not probabilistic independence, but rather independence of auxiliary assumptions.

Now, suppose we have two data sets $D_1$ and $D_2$ that are evidentially independent conditional on $T$. In order for $Ev$ to satisfy the Coherence Requirement, $F$ needs to satisfy the following equation:

$$F[n(D_1&D_2|T)] = F[n(D_1&D_2|T) - n(D_2|T)] * F[n(D_2|T)]$$  \hspace{1cm} (4.1)

Since there is no a priori limitation on which number $n(D|T)$ can be (aside from the fact that it must be a non-negative integer), equation (4.2) implies that $F$ must
satisfy the following equation for all non-negative integers \( x \) and \( y \):

\[
F[x] = F[x - y] \cdot F[y] 
\]  

(4.2)

It is easy to show by induction that if \( F \) is non-decreasing, then equation (4.2) has the following unique solution: \( F[x] = a^{-x} \), where \( a \) is some real number greater than 1.\(^{11}\) Hence we reach the following conclusion:

**Parsimony evidential measure:** Given the parsimony verisimilitude measure, the appropriate way of measuring the evidential impact that \( D \) has on \( T \) is given by the following formula:

\[
\text{Ev}[D|T] = a^{-n(D|T)} \tag{4.3}
\]

Where \( a \) is some real number greater than 1.

Interestingly, but perhaps not surprisingly, (4.3) has a form similar to the evidential measures (implicitly) derived by Bissiri et al. (2016), even though the derivation by Bissiri et al. (2016) is a quite different decision theoretic argument.\(^ {12}\) If the evidential measure is used by itself to evaluate and compare theories, then the choice of \( a \) clearly does not matter. However, if a prior distribution is placed over the alternative theories and a full verisimilitude-based analysis is performed, then the choice of \( a \) might be important.

Earlier we made the simplifying assumption that the theories under consideration all posit the same number of basic entities. But suppose they do not. Then, given the parsimony measure of verisimilitude, a theory that posits fewer basic entities will

---

\(^{11}\)Proof: plug in \( x = 2 \) and \( y = 1 \). Then \( F(2) = F(1)^2 \). Now suppose \( F(N - 1) = F(1)^{N-1} \). Then \( F(N) = F(N - 1 + 1) = F(1)^{N-1}F(1) = F(1)^N \). Since \( F \) is monotonically decreasing, \( F(1) \) must be less than 1. Hence, if we put \( a = \frac{1}{F(1)} \), we get that \( F(N) = a^{-N} \), with \( a \) greater than 1.

\(^{12}\)Strictly speaking, the evidential measure in (4.3) cannot be derived in the decision theoretic framework assumed by Bissiri et al. (2016). Among other reasons, \( n(D|T) \) is not an additive loss function: it is not necessarily the case that \( n(D_1&D_2|T) = n(D_1|T) + n(D_2|T) \) for all \( D_1 \) and \( D_2 \). But in Bissiri et. al’s decision theoretic derivation, an additive loss function is assumed. For reasons of space, I cannot pursue a more detailed and complete comparison with Bissiri et al.’s derivation here.
a priori plausibly be closer to the truth than a theory that posits more basic entities. Hence, if we are evaluating multiple theories with a different number of basic entities, the prior probability assigned to any given theory should arguably be a reflection of the number of basic entities, \( n(T_i) \) posited by the theory. Note that, in the parsimony verisimilitude measure, the total verisimilitude score of a theory is determined by the sum of the number of basic entities and auxiliary assumptions posited by the theory. In other words, auxiliary assumptions and basic entities count equally in determining the total verisimilitude score. Hence, the prior distribution should arguably be of the form \( a^{-n(T_i)} \), because then – and only then – will the evidential measure and the prior play an equally important role in determining the posterior probability of the theory. Note that the fact that the sum of the prior probabilities of the theories under consideration should sum to 1 will in fact determine the numerical value of \( a \).

So suppose such a prior \( p \) is assigned over each of the theories under consideration. Then the posterior probability of \( T \) will be given by the following quasi-Bayesian formula:

\[
p(T|D) = \frac{a^{-n(D|T)} \ast a^{-n(T)}}{\sum_i a^{-n(D|T_i)} \ast a^{-n(T_i)}}
\]

(4.4)

Here, the expression \( p(T|D) \) should be interpreted as how plausible it is that \( T \) is closest to the truth out of the theories under consideration, given data \( D \). This interpretation is consonant with the verisimilitude interpretation defended by Vassend (manuscript).

As I noted earlier, the reconstruction of parsimony inference undertaken in this section was mostly intended to illustrate the verisimilitude framework; it is doubtful whether the above verisimilitude-motivated quasi-Bayesian inference procedure will be useful for understanding how actual scientific inference works. However, in the next section of the paper, I give an application of the verisimilitude framework that I believe shows the verisimilitude framework to be philosophically useful.
4.2 Parsimony inference in phylogenetics

As was noted earlier in the paper, phylogeneticists in both biology and linguistics model historical relationships using tree topologies that, in almost every case, will be known to be false even before any evidence is collected, simply because every phylogenetic tree is based on highly idealized assumptions. For that reason, the goal of phylogenetic research cannot reasonably be taken to be inferring the tree topology that is true (or most probably true). Instead, it is closeness to the truth that must be the goal, given some reasonable measure of verisimilitude.

What exactly the intended measure of verisimilitude is supposed to be is rarely made explicit by phylogeneticists. If we assume that the true phylogeny is some type of topological network (not necessarily a tree) and that the goal is to discover the tree that is maximally similar to the true network, then obviously the verisimilitude measure should be some sort of topological or geometric similarity measure. On the other hand, if the goal is just to estimate some parameter within the true phylogeny, e.g. the approximate time when proto-Indo-European was spoken (Bouckaert et al., 2012), then the verisimilitude measure ought to be targeted towards that parameter. In either case, the evidential measure ought to be calibrated to whatever the appropriate verisimilitude measure happens to be. So far, phylogeneticists have not tried to come up with evidential measures that are specifically calibrated to appropriate verisimilitude measures, since they have not – as far as I’m aware – tried to quantify verisimilitude. Nonetheless, every evidential measure will implicitly be calibrated to some measure of verisimilitude; hence, by investigating the inferential methods that phylogeneticists use, we can discover what measures of verisimilitude they implicitly commit themselves to.

One of the inferential frameworks employed by phylogeneticists is cladistic parsimony. The principle of parsimony says that when you are deciding between several possible tree topologies, the evidence favors the tree topology that requires the fewest number “homoplasies” in order to account for the evidence. Roughly speaking, a homoplasy occurs whenever an evolutionary innovation occurs independently

---

13Some parts of the network may well be more tree-like than other parts, as pointed out by Gray et al. (2010).
in multiple lineages. For example, in biological phylogenetics, a genetic mutation that occurs in two species, but not in their common ancestor, is one kind of homoplasy. But phenotypic traits can also be homoplasies: if two species both have eyes, but their common ancestor does not, then having eyes is a homoplasy—it’s a shared evolutionary innovation. In historical linguistics, on the other hand, homoplasies often take the form of phonological traits. For example, if a certain sound occurs in two languages, but not in their most recent common ancestor, then that sound is a homoplasy.

According to the principle of parsimony, then, one should count the number of homoplasies posited by each tree; the most favored tree will be the one that requires the smallest number of homoplasies in order to account for all the data. More formally, let \( D \) be a data set that consists of the distribution of a set of traits among several species/languages/etc., let \( T \) be a tree topology, and let \( n(D|T) \) be the number of homoplasies that must be posited in order for \( T \) to account for \( D \). Then the principle of parsimony says that \( D \) favors \( T_1 \) over \( T_2 \) if and only if \( n(D|T_1) < n(D|T_2) \).

Note that \( n(D|T) \) does not satisfy the Coherence Requirement. Indeed, \( n(D|T) \) has exactly the same form and properties as \( n(D|T) \) from the previous subsection (hence the same notation). For the same reasons as before, in the verisimilitude framework, the evidential measure therefore takes the following form:

\[
\text{Ev}[D|T] = a^{-n(D|T)}
\]  

Note moreover that, once again, the relevant sort of “evidential independence” here is set-independence, not probabilistic independence. Hence, two data sets are evidentially independent, conditional on \( T \), if and only if the set of homoplasies \( T \) must posit in order to account for the first data set does not overlap the set of homoplasies that \( T \) must posit to account for the second data set.

Which measure of verisimilitude is \( \text{Ev}[D|T] \) implicitly calibrated against? The

\[14\] The are also refinements of cladistic parsimony, where some homoplasies are weighted more than others. I will not discuss those refinements here.
answer seems obvious. Since trees are penalized in proportion to how many independent evolutionary innovations they need to posit in order to account for the evidence, the implicit verisimilitude measure is clearly committed to the thesis that a hypothesis is closer to the truth insofar as it posits that there have been fewer evolutionary innovations. Hence, the relevant verisimilitude measure looks something like this:

**Phylogenetic parsimony verisimilitude measure:** The verisimilitude of a phylogenetic hypothesis $H$ is proportional to the amount of evolutionary innovation it has to posit in order to account for the distribution of all traits.

Note that the phylogenetic verisimilitude measure applies to any sort of phylogenetic hypothesis, including – but not limited to – trees. Note also that the verisimilitude of a hypothesis is proportional to the amount of evolutionary innovation it has to posit in order to explain the distribution of all traits, not simply the traits of which scientists are currently aware.

On the assumption that the phylogenetic parsimony verisimilitude measure is the appropriate verisimilitude measure for cladistic parsimony, $Ev[D|T]$ plausibly satisfies the Consistency Constraint. Indeed, if we make the reasonable assumptions that the objects being compared (whether species or languages) have only finitely many traits that may be compared, then – as the number of independent pieces of evidence increases – eventually every trait will be included among the data sets. Hence, the tree that performs the best on $Ev[D|T]$ given all the data will, by definition, be the tree that is closest to the truth according to the phylogenetic parsimony verisimilitude measure.

In contrast to traditional cladistic parsimony analysis, verisimilitude-based cladistic parsimony allows for the inclusion of a prior probability distribution. Thus, given a prior probability distribution, $p$, over the various possible trees $T_i$ under consideration, we may calculate the posterior probability of each tree using Bayes’s formula:

$$p(T|D) = \frac{a^{-n(D|T)} * p(T)}{\sum_i a^{-n(D|T_i)} * p(T_i)}$$  (4.6)
This is significant, for it means that we are now in a position to critique cladistic parsimony in the same way that Bayesians often critique methods that rely on using only the likelihood. In particular, a common Bayesian criticism is that maximum likelihood estimation – i.e. inferring the hypothesis that has the highest likelihood given the evidence – in general is equivalent to doing Bayesian inference with a “flat” prior that assigns the same probability to each hypothesis, because on a Bayesian calculation, the hypothesis that receives the highest posterior probability will – in general – be equivalent to the hypothesis that has the highest likelihood if and only if a flat prior is used in the calculation. Depending on the details of the problem, a flat prior may not be appropriate.

We can now level the same criticism against cladistic parsimony, for clearly the tree that has the best parsimony score, \( n(D|T) \), will – in general – be equivalent to the tree that has the highest posterior probability if and only if \( p(T_i) = p(T_j) \) for all \( i \) and \( j \), i.e. if and only if the prior probability assigned to each tree topology is the same. The question to ask, then, is this: is it reasonable to assign the same prior probability to every tree topology?

In the case of biological phylonenetics, Joel Velasco (2008) convincingly argues that the answer is “no.” We need not go into the details of Velasco’s argument here, but the upshot of the argument is that if a flat prior is to be used at all, then the prior should be flat over all possible labeled histories. Importantly, some tree topologies are consistent with more labeled histories than others. Hence, a flat prior over all the labeled histories will necessarily induce a non-flat prior over the possible tree topologies.

Velasco is concerned with critiquing Bayesian phylogeneticists, but now that we have reconstructed cladistic parsimony in the verisimilitude framework, we can see that his criticism applies equally to phylogeneticists who use cladistic parsimony, since – as already noted – cladistic parsimony implicitly assumes a flat prior over all the tree topologies.

However, in the verisimilitude framework, it would seem that we are not just in a position to raise Velasco’s criticism—we can also respond to it. Because whereas traditional cladistic parsimony does not mention any priors (although implicitly it
is committed to one), verisimilitude-based cladistic parsimony allows for the explicit inclusion of any prior. In particular, let $n(T_i)$ be the number of labeled histories consistent with $T_i$ and let $L$ be the total number of labeled histories. Then, in light of Velasco’s argument, a reasonable prior to assign to $T_i$ is $p(T_i) = \frac{n(T_i)}{L}$.

Hence, if we use the prior suggested by Velasco and update the probability of $T_i$ with the parsimony-version of Bayes’s formula, the result will look like this:

$$p(T_i|D) = \frac{a^{-n(D|T_i)} \cdot n(T_i)}{\sum_j a^{-n(D|T_j)} \cdot n(T_j)}$$ (4.7)

Here, $a$ is a constant (greater than 1) that implicitly balances how weighty the evidential measure is when compared to the prior. If $a$ is bigger, then the data will influence the posterior more than if $a$ is smaller. If $a$ is set so that is greater than $Max_{i,j}(\frac{n(T_i)}{n(T_j)})$, then the prior will have no influence on the posterior ranking of hypotheses. Hence, $a$ should presumably be set somewhere lower. For example, if there are four taxa, then $Max_{i,j}(\frac{n(T_i)}{n(T_j)}) = 2$ (Velasco, 2008, p. 469), so $a$ should be set somewhere between 1 and 2.

From a verisimilitude perspective, the goal is to infer the hypothesis that is closest to the truth. Presumably, there is some value of $a$ that will accomplish this more efficiently than any other value, but we do not know what that value is. Hence, it’s natural to regard $a$ as a parameter. From a Bayesian perspective, it’s reasonable to quantify uncertainty concerning the value of $a$ with a prior probability function. Note that $p(a = 1.5) = 0.3$, say, is a verisimilitude probability—it reflects a degree of belief of 0.3 that 1.5 is the value of $a$ that is closest to the truth in the sense that $a = 1.5$ is the value of $a$ that will most efficiently lead one to infer hypotheses that are closest to the truth according to the phylogenetic parsimony verisimilitude measure. Importantly, $p(a = 1.5) = 0.3$ should not be interpreted as a degree of belief that 1.5 is the true value of $a$—there is no true value of $a$.

To give a concrete example, in the case with four taxa, we may assign a flat prior over all possible values of $a$ between 1 and 2. If we do that, then $Ev[D|T_i]$ takes the following form.\(^\text{15}\)

\(^{15}\)This form of $Ev$ is arrived at by integrating $Ev[D|T_i, a]p(a|T_i)$ over $a$, on the assumption that
The posterior probability of $T_i$ can then be calculated by combining the evidential measure in (4.8) with the prior suggested by Velasco. The resulting posterior will of course be sensitive to our choice of prior distribution over the different possible values of $a$. The problem of how to set “tuning” constants such as $a$ has been explored recently by, for example, Ibrahim et al. (2015) and Bissiri et al. (2016); I will not discuss the issue further here. The important upshot for our purposes is that verisimilitude-based cladistic parsimony apparently occupies an intermediate spot between traditional cladistic parsimony and a fully Bayesian analysis: it allows for the incorporation of background information through a prior, while also using a parsimony-motivated evidential measure rather than the likelihood. Thus, by recasting cladistic parsimony in the verisimilitude framework, we have apparently strengthened it.

But things are not that simple. Because the question now arises as to which verisimilitude measure the inference procedure in (4.7) is implicitly calibrated against. Just like the evidential measure, the prior probability distribution ought to make sense in light of the verisimilitude measure, but it’s hard to see how the phylogenetic parsimony verisimilitude measure could possibly justify assigning more prior probability to some tree topologies than others; it’s hard to see how that would lead us to more efficiently infer the tree that posits the smallest amount of evolutionary innovation.

So Velasco’s objection resurfaces in a new version: if we grant that trees that are consistent with more labeled histories are more biologically plausible than trees that are consistent with fewer, then the phylogenetic parsimony verisimilitude measure is deficient. It’s deficient because it doesn’t give us any justification for using the sorts of prior probability distributions that Velasco is suggesting are more biologically realistic, and clearly the prior distribution we use ought to incorporate biological background knowledge. The problem is not mathematical; mathematically, $\text{Ev}(D|T)$

\[
\text{Ev}(D|T_i) = \frac{1 - 2^{1-n(D|T_i)}}{n(D|T_i) - 1}
\]  

\[ 4.8 \]

$a$ and $T_i$ are independent.
can be combined with any prior. The problem is that the prior should be justified in light of the verisimilitude measure, but such a justification is not forthcoming if we use the phylogenetic parsimony verisimilitude measure. There may be some reasonable way of amending the phylogenetic parsimony verisimilitude measure that would address this problem, but I will not pursue that here. The upshot here will just be the following more general lesson: the verisimilitude measure, evidential measure, and prior distribution should all ideally be calibrated to *each other*. The evidential measure needs to be calibrated to the verisimilitude measure, but the verisimilitude measure also ought to make sense in the light of the evidential measure and scientifically reasonable priors.

5 Conclusion

I have argued that the Law of Likelihood often fails to be satisfied in practice, and that the standard interpretation therefore needs to be replaced by a different one. The verisimilitude interpretation is better than the standard interpretation, but even the verisimilitude interpretation often faces difficulties. The problem is that, when the hypotheses under consideration are all known to be false and the goal is simply to find the hypothesis that is closest to the truth, then likelihood-based inference will not necessarily pick the hypothesis that is closest to the truth in the relevant sense. Hence, the likelihood sometimes needs to be replaced by another measure of evidential impact, and I have argued that such measures ought to be chosen so that they are calibrated against a measure of verisimilitude that is appropriate in the given context. Finally, I have illustrated the verisimilitude framework by reconstructing parsimony evaluations of theories and parsimony inference in phylogenetic inference.

This paper has focused on Bayesianism and likelihoodism in philosophy of science. I have argued that the verisimilitude framework is a better – and more general – framework for philosophy of science than either Bayesianism or likelihoodism. However, as was noted in the introduction, Bayesianism and likelihoodism are not just frameworks for philosophy of science; they are also frameworks for statistical inference. So a natural question arises: is the verisimilitude framework a viable – and
better – framework statistical inference as well? That is a question for another paper.

References


