A Verisimilitude Framework for Inductive Inference, With an Application to Phylogenetics

Olav B. Vassend

June 20, 2018

Abstract

Bayesianism and likelihoodism are two of the most important frameworks philosophers of science use to analyse 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 inductive inference. In the verisimilitude framework, scientific inquiry is conceived of, 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 favored by 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.[1] 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[2] 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.[3]

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 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 [2018]).

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. In my view, perhaps the most serious objection is that—as I explain in the next section of the paper—both frameworks seem to assume that the true hypothesis is one of the hypotheses under consideration. As philosophers of science realized decades ago, 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. For these reasons, identifying the hypothesis under consideration that is closest to the truth—and not the truth itself—is the realistic goal of scientific inference in many if not most cases (as also emphasized by Forster [2002]).

The study of closeness to the truth—or 'verisimilitude'—was initiated by Popper ([1963]) and has since amassed a large literature.[4] The most influential contemporary approach—and the approach that will be assumed in this paper—understands verisimilitude quantitatively: that is, verisimilitude is conceptualized in terms of a real-valued function v, such that H_1 is closer to the truth than H_2 if and only if $v(H_1) > v(H_2)$. In general, I agree with Northcott ([2013]) that the specific form the verisimilitude function should take will depend on the context.

[1]Importantly, likelihoodists do not consider the likelihood of a hypothesis a measure of the plausibility of the hypotheses—rather, the likelihood is a measure of evidential favoring (see Royall [1997] or chapter 1 of Sober [2008] for more detailed descriptions of likelihoodism).

[2]Bayes'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: $p(H|E) = \frac{p(E|H)p(H)}{\sum_i p(E|H_i)p(H_i)}$.

[3]This is a simplified description of Bayesian inference. More sophisticated Bayesian statistical inference crucially involves predictive model checking (see, for example, Gelman *et al.* [1996], Gelman and Shalizi [2013]) and various forms of robustness analysis.

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

What is the relationship between verisimilitude and inferential frameworks such as likelihoodism and Bayesianism? By and large, this question has been neglected in the literature, although there are important exceptions (Rosenkrantz [1980]; Niiniluoto [1986], Niiniluoto [1987]; Festa [1993]; Cevolani *et al.* [2010]; Oddie [Forthcoming]).[5] For example, Niiniluoto ([1986]) conceives of verisimilitude as a kind of utility, and shows how verisimilitude theory may be integrated into Bayesian decision theory. However, as far I am aware, none of these earlier accounts consider the possibility that the verisimilitude framework, on the one hand, and Bayesianism and likelihoodism, on the other, may be in tension with each other.[6] I will argue that they are.[7]

More precisely, I will argue 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 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), which may be stated in the following way:[8]

> Law of Likelihood, Standard Version: Evidence E favors the proposition that H_1 *is true* over the proposition that H_2 is true if and only if $p(E|H_1) > p(E|H_2)$.[9]

[5]There have been more analyses of the relationship between verisimilitude and AGM belief revision. See, in particular, the *Erkenntnis* september 2011 special issue on the relationship between verisimilitude and belief revision, especially Baltag and Smets ([2011]); Cevolani *et al.* ([2011]); Kuipers ([2011]); Renardel de Lavalette and Zwart ([2011]); and Schurz ([2011]).

[6]Oddie ([Forthcoming]) argues that verisimilitude theory is in conflict with a certain approach within Bayesian epistemology known as the 'accuracy framework' (Joyce ([1998]); Pettigrew ([2016]), but he does not argue that the conflict lies with the Bayesian framework itself.

[7]Interestingly, Niiniluoto's expected verisimilitude account may avoid the tension, but at the cost of often being inapplicable in practice in the kinds of inferential problem scientists typically study, as I explain in footnote 13.

[8]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.

[9]The Law of Likelihood is often phrased in the following way:

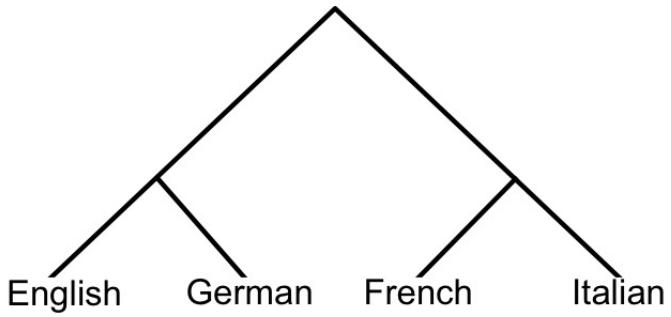

Figure 1. A phylogenetic tree of some European languages.

I have called the above formulation of LL the 'standard version' to contrast it with other versions that will be considered later. Both likelihoodists and Bayesians tend to accept the standard version of LL. Some think it is a primitive postulate about evidential favoring (for example Edwards [1972] or Sober [1988]). Others argue that the standard version of LL is supported by various mathematical results, for example convergence results (Hawthorne [1994]) or the 'Universal Bound' (Royall [1997]), or proofs for the so-called 'Likelihood Principle', which is a related evidential principle (see Birnbaum [1962] and Gandenberger [2015]).

However, a fundamental problem with the standard version of LL is that it plainly fails to be useful if all the hypotheses under consideration are already known to be false—if we know that all our hypotheses are false, then why would we care about which one is favored to be true? Indeed, in such cases, the standard version of LL is arguably false. To see why, suppose we have just two hypotheses under consideration, H_1 and H_2, that both hypotheses are known to be false, and that $p(E|H_1) > p(E|H_2)$. As Good ([1967]) famously pointed out, all confirmation is relative to background knowledge. If our background knowledge is that both the hypotheses under consideration are false, it would seem to follow that no evidence can support the proposition that H_1 is true over the proposition that H_2 is true. Thus, the left-hand side of the biconditional in the standard version of LL is false. Yet, since $p(E|H_1) > p(E|H_2)$, the right-hand side of the biconditional is true. Hence, the standard version of LL is false in this case.[10]

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 often 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 linguistics alike that tree reconstruction rest on several highly idealized and false assumptions (See, for example, O'Malley *et al.* [2010]; Heggarty *et al.* [2010]; Velasco [2012]), including the following three: (1) evolutionary divergence

---

> Law of Likelihood, standard version: Evidence E favors H_1 over H_2 if and only if $p(E|H_1) > p(E|H_2)$.

By an application of Tarski's T-schema (Tarski [1944]), the two formulations of the Law of Likelihood are equivalent.

[10] A referee points out that there are also counter-examples to the standard version of LL in which the left-hand side is true and the right-hand side is false. I agree, but I will not discuss such cases here.

happens through bifurcations, (2) divergences happen instantaneously, (3) descendent 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 version of the law of likelihood has no sensible application, since all the hypotheses will then already be known to be false.[11]

When all the hypotheses under consideration are known to be false, Bayesians face the additional problem of how to sensibly interpret the 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 how plausible it is that the hypothesis is true. However, this interpretation clearly does not make sense in cases where all the hypotheses are already known to be false. Recently, Sprenger ([2017]) and Vassend ([2018]) 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. rather than interpreting $p(H)$ as a degree of belief that H is true, the verisimilitude interpretation says that $p(H)$ instead should be construed contrastively as a degree of belief that H is closest to the truth (according to a reasonable verisimilitude measure) 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 reading of LL is as follows:

> Law of Likelihood, Verisimilitude Version: Evidence E 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 $p(E|H_1) > p(E|H_2)$.

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

> Law of Likelihood, Counterfactual Version: Evidence E 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 $p(E|H_1) > p(E|H_2)$.

The verisimilitude and counterfactual versions of LL are both better than the standard version in the sense that they are both applicable when all the hypotheses under consideration are false. The question is, however, whether the counterfactual and verisimilitude versions may sometimes be violated in the sense that the likelihood may show a systematic preference for hypotheses that are further from the truth. The answer, as I will argue, is yes. For

[11] Of course, as Gray *et al.* ([2010]) point out, some parts of the actual history may be more tree-like than others: hence, the tree assumption may be more realistic for some parts of the phylogenetic history than for others. Nonetheless, for the phylogenetic history as a whole, the tree-model is clearly unrealistic.

simplicity, I will only discuss the verisimilitude interpretation of probability. As argued by Vassend ([2018]), 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 version of LL happen in statistical model selection. As Forster and Sober ([1994]) point out, predictive accuracy (quantified in terms of Kullback-Leibler divergence (Kullback and Leibler [1951])) 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 with the best likelihood-fit to the data (where the likelihood-fit of a model is quantified as the likelihood score of the best-fitting hypothesis in the model) 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 model that is most predictively accurate (i.e. closest to the truth in the relevant sense) will often not be the model that has the best likelihood-fit. In place of the likelihood, Forster and Sober recommend AIC as a superior measure of evidential favoring for models, where the AIC score of a model is equal to $log(\hat{L}) - k$. Here $\hat{L}$ is the likelihood of the model when the parameters are set at the values that maximize likelihood, and k is the number of parameters in the model.

Although Forster and Sober think the likelihood is a poor measure of evidential favoring in the case of model selection, they maintain that it's a fine 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 version of LL even when the inferential problem is not to select between models that have different numbers of parameters.

For example, suppose we are concerned with identifying the functional relationship between two quantities y and x. To make the example concrete, suppose 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 straight 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.

For example, suppose that we are interested in building a structure that will be capable of withstanding strong winds. In that case, it is very important that we minimize the maximum error we are likely to make in our prediction of windspeed, since we certainly want our structure to be able to withstand worst-case windspeeds.[12] Our average prediction error, on the other hand, is less important. Thus, a sensible measure of closeness to the truth of some given line, $L(x) = \alpha x + \beta$, given this goal, is given by the following formula:
$v_{Max}(L) = -Max_{x\in[a,b]}|t(x) - L(x)|$, where $[a, b]$ is the range of relevant pressures. This verisimilitude measure is also depicted in Figure 2.

Our inferential goal, then, is to identify the hypothesis in our linear model that is closest to the truth in the sense of v_{Max}. Will using the likelihood enable us to achieve this goal? In order to calculate the likelihoods of the hypotheses, we need to make more precise probabilistic assumptions. Suppose that our measurements of windspeed are normally (i.i.d) distributed around $L(x)$, for each x, with a mean centred on $L(x)$ and a (constant) standard deviation of σ.

[12]I borrow this example from Vassend ([2018]).

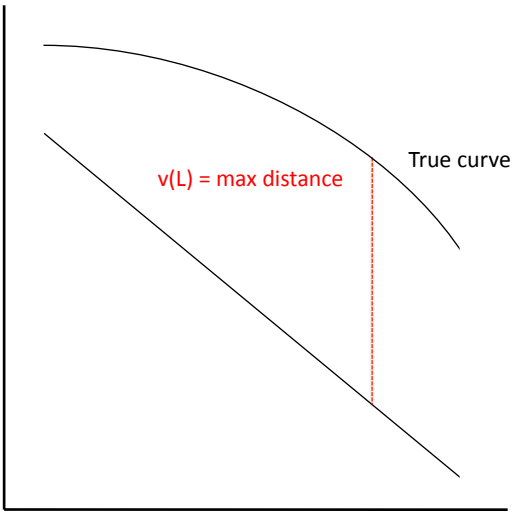

Figure 2. A measure of closeness to the truth.

Suppose, moreover, that we sample values of x uniformly from the interval $[a, b]$. Given these assumptions, we can calculate the likelihood of each line, L, given a set of data points, (x_1, y_1), (x_2, x_2), ..., (x_n, x_n), and we get (Wasserman [2004], p. 213):

$$p((x_1, y_1), (x_2, x_2), \ldots, (x_n, x_n)|L) = \frac{1}{(b-a)\sqrt{2\pi\sigma^2}} e^{\frac{\sum_{i=1}^{n} -(y_i - L(x_i))^2}{2\sigma^2}} \tag{2.1}$$

Now, the line, $\hat{L}$ that has maximal likelihood, given the data, is also the line that has maximal log-likelihood. And thus, ignoring constants, $\hat{L}$ minimizes the sum $\sum_{i=1}^{n}(y_i - L(x_i))^2$, or equivalently, $\frac{1}{n}\sum_{i=1}^{n}(y_i - L(x_i))^2$.

Suppose now that the number of data points increases towards infinity. Then, by the Weak Law of Large Numbers (Wasserman [2004], p. 76), $\frac{1}{n}\sum_{i=1}^{n}(y_i - L(x_i))^2$ converges in probability to the expected value, $\mathrm{E}(y - L(x))^2$. Thus, as the amount of data increases to infinity, the likelihood will increasingly favor the line that minimizes $\mathrm{E}(y - L(x))^2$. But, given our verisimilitude measure, what we are interested in is actually the line that maximizes $v_{Max}(L) = -Max_{x\in[a,b]}|y - L(x)|$. Unless there is a remarkable coincidence, the line that minimizes $\mathrm{E}(y - L(x))^2$ will not be the line that maximizes $-Max_{x\in[a,b]}|y - L(x)|$. Consequently, if we use the likelihood as our measure of evidential favoring, we will never successfully infer the line that is closest to the truth in the sense that we care about.[13]

[13]The previously mentioned (and by now standard) account of expected verisimilitude due to Niiniluoto ([1986]) gets around the preceding problem, but at a cost that often makes the account inapplicable in practice. Glossing over some of the nuances, the account is basically as follows: let v be a verisimilitude measure, then Niiniluoto recommends that we keep track of the evidential standing of hypothesis H given E by calculating its expected verisimilitude, which is given by the following formula: $\mathrm{E}[v(H, E)] = \sum_{w_i} p(w_i|E)Tr(H, w_i)$, where the sum is over every possible world w_i. This might avoid the problem pointed out in this section because using Niiniluoto's formula requires that we calculate $p(w_i|E)$ rather than $p(E|H)$. However, there is a new problem because $p(w_i|E)$ will, in general, not be epistemically accessible. In particular, in the example in this section, calculating $p(w_i|E)$ would require us to have available the partition of every possible functional relationship between pressure and maximum windspeed, but of course scientists do not have this. In general, they restrict

The above example illustrates a violation of the verisimilitude version of LL: it's an example where the likelihood can be expected to pull systematically in a different direction than what we want given the verisimilitude measure in which we are interested. This type of example motivates 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 a verisimilitude framework. The guiding idea will be 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, conceiving of inductive inference in a verisimilitude framework—where the goal is to get close to the truth in some particular sense—motivates the idea that any proposed measure of evidential favoring should 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, 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 (out of the hypotheses under consideration) if and only if $\text{Ev}[E|H_1] > \text{Ev}[E|H_2]$.

Intuitively, evidential measure 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. That is, there should be a systematic fit between the verisimilitude measure and the evidential measure, such that—in general—hypotheses that are close to the truth 'do well' with respect to the evidential measure, and vice versa. As a minimal constraint, it needs to be the case that—as the amount of evidence increases without bound—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, Ev is consistent with v if and only if the following inequality is true:

$$\text{Ev}(E|H_t) > \text{Ev}(E|H_f) \tag{3.1}$$

For example, to take our earlier case concerning the relationship between min pressure and max windspeed, if maximal vertical distance from the truth is used to measure verisimilitude and the likelihood is used to measure evidential impact, then the analysis in the preceding section shows that the Consistency Constraint will be violated because even after arbitrarily large amounts of evidence, the hypothesis that has the best likelihood-score will not be the hypothesis that is closest to the truth. On the other hand, the AIC score of a model satisfies the

themselves to simple classes of functional relationships, none of which will contain the true functional relationship.

Consistency Constraint given that the intended verisimilitude is the Kullback-Leibler divergence (see Forster and Sober [1994]). Note that the statement of the Consistency Constraint is intentionally a bit vague. This is because it is supposed to cover a wide range of situations, including situations in which the hypotheses are probabilistic and situations in which the hypotheses are non-probabilistic.

Even diehard likelihoodists agree that the likelihood ought to be combined with a prior probability distribution if 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:

> (Posterior plausibility of H given $E_1, \ldots, E_n$) $\propto$ (Evidential impact on H of $E_1, \ldots, E_n$) * (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 given the first one. In other words, $p(E_1, E_2|H) = p(E_1|H) * p(E_2|H\&E_1)$. To be capable of replacing the likelihood in Bayesian calculations, any evidential measure therefore needs to have the following formally equivalent property:

> Coherence Requirement: $\mathrm{Ev}(E_1, E_2|H) = \mathrm{Ev}(E_1|H, E_2) * \mathrm{Ev}(E_2|H)$

The Coherence Requirement of the likelihood is not just important in Bayesian calculations; it's also crucial in frequentist inference. For example, the Coherence Requirement is one of the most important properties of the likelihood that allows for the derivation of the previously mentioned AIC measure of evidential favoring.[14]

Furthermore, 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—and regardless of the order in which we consider the two pieces of evidence[15]—we end up with the same verdict of how much E_1 and E_2 jointly impact H. This seems to be a reasonable property to require of a measure of evidential favoring.[16] Thus we see that the Coherence Requirement is motivated from several different perspectives.

[14]In this context, it may be worth noting that even if the likelihood is replaced with a different measure of evidential favoring, the derivation of AIC will still go through in amended form in many cases. Hence, the general form of the AIC score of a model, M, that has k parameters and where $\hat{M}$ is the setting of the parameters of M that maximizes $\mathrm{Ev}(E|M)$ is as follows:

$$AIC(M) = \log \mathrm{Ev}(E|\hat{M}) - k \tag{3.2}$$

[15]I thank a referee for pointing out that the Coherence Requirement also implies an irrelevance of order.

[16]Coherence of this type is heavily emphasized by E. T. Jaynes ([2003]) and also by Bissiri *et al.* ([2016]).

Whenever $\mathrm{Ev}(E_1|H, E_2) = \mathrm{Ev}(E_1|H)$, the Coherence Requirement entails that $\mathrm{Ev}(E_1, E_2|H) = \mathrm{Ev}(E_1|H) * \mathrm{Ev}(E_2|H)$, i.e. the evidential impact that E_1 and E_2 jointly have on H is equal to the product of the evidential impact that they have separately. Now, when $p(E_1|E_2\&H) = p(E_1|H)$, E_1 and E_2 are standardly said to be 'probabilistically independent' conditional on H. So let's define the following more general concept:

> Evidential Independence: Given evidence E_1 and E_2 and evidential measure Ev, E_1 and E_2 are evidentially independent conditional on H if and only if $\mathrm{Ev}(E_1|H, E_2) = \mathrm{Ev}(E_1|H)$ and $\mathrm{Ev}(E_2|H, E_1) = \mathrm{Ev}(E_2|H)$

'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 section, we will see how these requirements, and the details of the cases, 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 may be reconstructed in the verisimilitude framework, and in particular how a reasonable evidential measure may be derived given a completely specified verisimilitude measure. The verisimilitude measure that I will introduce in this section is admittedly not without its difficulties and has been chosen mainly for its theoretical simplicity rather than its plausibility. In the next section, I will discuss an example where the verisimilitude measure is not completely specified.

Suppose we are interested in evaluating scientific theories (for example 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.[17] 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.[18] This understanding of simplicity is consonant with the traditional *vera causa* understanding of parsimony summarized by Isaac Newton in his first first rule for scientific reasoning (Newton

[17] I'm using the rather vague phrase of a theory's 'accounting for the phenomena' because the 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.

[18] I will put aside the vexed and well known question of how to count the number basic entities or the number of auxiliary assumptions.

[1999], p. 794).[19] If you add enough ad hoc auxiliary assumptions, just about any theory can account for just about any empirical phenomenon, but the idea is that 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 (negative[20]) 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 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 $\text{Ev}[D|T] = \text{F}[n(D|T)]$ 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 $\text{Ev}[D_1|T\&D_2] = \text{F}[n(D_1\&D_2|T) - n(D_2|T)]$. 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 set independence.

Now, suppose we have two data sets D_1 and D_2. In order for Ev to satisfy the Coherence Requirement, F needs to satisfy the following equation:

---

[19] Newton arguably endorsed an ontological, not merely methodological, reading of the *vera causa* principle, as evidenced by the following famous quote:

> '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' (Newton [1999], p. 794).

See the first chapter of Sober ([2015]) for more discussion.

[20] We need the negative sum here since a theory that has more basic entities or more auxiliary assumptions is further from the truth.

$$\mathrm{F}[n(D_1\&D_2|T)] = \mathrm{F}[n(D_1\&D_2|T) - n(D_2|T)] * \mathrm{F}[n(D_2|T)] \tag{4.1}$$

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 (4.1) implies that F must satisfy the following equation for all non-negative integers x and y:

$$\mathrm{F}[x] = \mathrm{F}[x - y] * \mathrm{F}[y] \tag{4.2}$$

It is easy to show by induction that if F is monotonically decreasing, then equation (4.2) has the following unique solution: $\mathrm{F}[x] = a^{-x}$, where a is some real number greater than 1.[21] 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:
>
> $$\mathrm{Ev}[D|T] = a^{-n(D|T)} \tag{4.3}$$
>
> 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.[22] 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 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

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

[22] 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.

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:

$$p(T|D) = \frac{a^{-n(D|T)} * a^{-n(T)})}{\sum_i a^{-n(D|T_i)} * a^{-n(T_i)}} \tag{4.4}$$

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 version defended by Vassend ([2018]).

The reconstruction of parsimony inference undertaken in this section is 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, since few people these days accept Newton's vera causa principle. In the next section of the paper, I give an application of the verisimilitude framework that is more relevant to contemporary scientific practice.

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 presumably the verisimilitude measure should be some sort of topological or geometric similarity measure—and there are many reasonable candidate measures (see Jurman *et al.* [2015] for some of the possibilities). On the other hand, if the goal is just to estimate some parameter within the true phylogeny, for example 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.

Currently, the inference methods employed by phylogeneticists tend to be based on the likelihood or on parsimony considerations, and there appears to be a consensus that likelihood-based methods are better. Much of the philosophical discussion of phylogenetics has focused on what substantive (for example biological) assumptions the various methods do or do not make (for example Sober [1988]). However, in order to evaluate a given evidential measure, it's also necessary to investigate whether it is calibrated in the appropriate way to the relevant verisimilitude measure. I will not try to determine what verisimilitude measure or evidential measure phylogeneticists ought to use, or whether the evidential measures phylogeneticists already use are adequate to the task. Here I will just note that, as things stand, it is not obvious that likelihood-based measures necessarily are better than parsimony-based measures for all purposes (i.e. regardless of how verisimilitude is measured).

In the remainder of this section, I will discuss a particular inferential framework employed by phylogeneticists, namely cladistic parsimony. As will hopefully become clear, recasting cladistic parsimony in the verisimilitude framework enables us to get a better understanding of its limitations and possible improvements.

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 of 'homoplasies' in order to account for the evidence. So, if T_1 implies that at least m homoplasies have occurred and T_2 implies that at least n homoplasies have occurred, then the principle of parsimony says that we should prefer T_1 to T_2 if and only if $m > n$. Roughly speaking, a homoplasy occurs whenever an evolutionary innovation occurs independently in multiple lineages. For example, in biological phylogenetics, a genetic mutation that occurs in two species, but not in their most recent common ancestor, is one kind of homoplasy. But phenotypic traits can also be homoplasies: if two species both have eyes, but their most recent 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.

Note that, as Farris ([1983]) points out, the principle of parsimony does not assume that the true phylogeny is the most parsimonious tree (as noted earlier, we already know that the true phylogeny is not a tree at all) or even that the true phylogeny is particularly parsimonious. If the true phylogeny were one of the hypotheses under consideration, then the Consistency Constraint would indeed imply that the principle of parsimony is a sound principle only if we assume that the true phylogeny is maximally parsimonious. However, we know that the true phylogeny is not one of the hypotheses under consideration. Hence, as things stand, the Consistency Constraint only implies that among all possible phylogenetic trees (in the limit as the amount of evidence increases), the principle of parsimony can be sensibly applied only as long as we assume that more parsimonious trees are closer to the true phylogeny than less parsimonious trees. But this leaves it open that the principle of parsimony may be compatible with a wide range of different verisimilitude measures.[23] Again, in the rest of the section, I will leave the precise form of the verisimilitude measure unspecified, although I will assume that more parsimonious trees are closer to the truth than less parsimonious trees. The goal is to show that even if the underlying verisimilitude measure is unspecified, we can use the verisimilitude framework (and in particular the verisimilitude version of Bayes's formula) to get a better understanding of cladistic parsimony.

According to the principle of parsimony, then, one should count the number of homoplasies required 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.[24] 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 $n(D|T)$ 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:

[23] I thank a referee for pressing me on this point.

[24] The are also refinements of cladistic parsimony, where some homoplasies are weighted more than others. I will not discuss those refinements here.

$$\text{Ev}[D|T] = a^{-n(D|T)} \tag{4.5}$$

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 with the set of homoplasies that T must posit to account for the second data set.

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 the verisimilitude version of Bayes's formula:

$$p(T|D) = \frac{a^{-n(D|T)} * p(T)}{\sum_i a^{-n(D|T_i)} * p(T_i)} \tag{4.6}$$

This is significant, for it means that we are now in a position to critique cladistic parsimony in the same way that Bayesians sometimes critique methods that rely on using only the likelihood. In particular, a standard 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 a flat prior is used in the calculation.[25] Depending on the details of the problem, a flat prior may not be appropriate. The point of this criticism isn't that maximum likelihood estimation is 'wrong', but rather that—from a Bayesian point of view—it can be improved upon in cases where a flat prior is clearly suboptimal or unreasonable.

We can now level the same criticism against cladistic parsimony, because given the verisimilitude version of Bayes's formula, 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 $p(T_i) = p(T_j)$ for all i and j, i.e. if the prior probability assigned to each tree topology is the same. This can most easily be seen from the versimilitude version of the odds formulation of Bayes's formula:

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

From 4.7, we see that if $p(T_i) = p(T_j)$ for all i and j, then $p(T_i|D) > p(T_j|D)$ if and only if $a^{-n(D|T_i)} > a^{-n(D|T_j)}$, which in turn is true if and only if $-n(D|T_i) > -n(D|T_j)$. Hence, if we rank the posterior verisimilitude probabilities of all the hypotheses from highest to lowest, their order will be equivalent to the order we get if we rank the hypotheses by their parsimony scores. Thus, from the point of view of the verisimilitude framework, using the principle of parsimony amounts to doing quasi-Bayesian inference with a flat prior. The question to ask, then, is this: is it reasonable to assign the same prior probability to every tree topology?

[25]To see why, consider the odds formulation of Bayes's formula: $\frac{p(H_i|E)}{p(H_j|E)} = \frac{p(E|H_i)}{p(E|H_j)} * \frac{p(H_i)}{p(H_j)}$. If the prior is flat, $p(H_i) = p(H_j)$, and so $p(H_i|E) > p(H_j|E)$ if and only if $p(E|H_i) > p(E|H_j)$, and so the hypothesis with the highest likelihood will also be the hypothesis that has the highest posterior probability.

In the case of biological phylonegetics, 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. A labeled history is a tree topology in which the internal vertices of the tree are time-ordered. 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 primarily concerned with critiquing Bayesian phylogeneticists, but he notes that his criticism will apply to any inferential method that does not explicitly take account of the prior probabilities of trees, including cladistic parsimony. Our reconstruction of cladistic parsimony in the verisimilitude framework has allowed us to raise Velasco's criticism of cladistic parsimony in a more explicit manner.

However, the verisimilitude framework does not just enable us to raise Velasco's criticism in a more explicit manner, it also enables us to respond to the criticism. This is because our verisimilitude reconstruction of cladistic parsimony in (4.6) allows for the explicit inclusion of any prior, including the prior recommended by Velasco. 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:

$$p(T_i|D) = \frac{a^{-n(D|T_i)} * n(T_i)}{\sum_j a^{-n(D|T_j)} * n(T_j)} \tag{4.8}$$

Here, a is a constant (again, greater than 1) that implicitly balances how bad positing an extra evolutionary innovation is compared to positing a tree that is consistent with fewer labeled histories. Or, on a more abstract level, a 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.[26] 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. Indeed, if we had a completely specified verisimilitude measure, then the verisimilitude measure would uniquely determine the value of a, as was the case in the preceding section. However, if we do not have a completely specified verisimilitude measure, then we do not know what the value of a 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 whatever verisimilitude measure is appropriate. Importantly, $p(a = 1.5) = 0.3$ should not be interpreted as a degree of belief that

[26] If a is greater than $Max_{i,j}(\frac{n(T_i)}{n(T_j)})$, then the prior will have no influence on the posterior ranking in the following sense: if $\mathrm{Ev}[E|T_1] > \mathrm{Ev}[E|T_2]$, then $p(T_1|E) > p(T_2|E)$ and if $\mathrm{Ev}[E|T_1] < \mathrm{Ev}[E|T_2]$, then $p(T_1|E) < p(T_2|E)$. These facts can most easily be seen by considering the odds formulation of the verisimilitude version of Bayes's formula (4.7).

1.5 is the true value of a—there is no clear sense in which any value of a can be said to be 'true'.

Without a specific verisimilitude measure, it's unclear how we should go about determining a prior distribution over the possible values of a, since the value of a that is closest to the truth is inextricably linked to how verisimilitude is measured. However, a conservative option that makes sense regardless of how verisimilitude is measured is to assign a flat prior over all the possible values of a, keeping in mind the restriction that a should be between 1 and $Max_{i,j}(\frac{n(T_i)}{n(T_j)})$. That way, the prior will at least play some role in the inference, which is a good thing if we think the prior suggested by Velasco is reasonable.

To give a concrete example, in the case with four taxa, we may choose to assign a flat prior over all possible values of a between 1 and 2. If we do that, then $\mathrm{Ev}[D|T_i]$ takes the following form:[27]

$$\mathrm{Ev}(D|T_i) = \frac{1 - 2^{1-n(D|T_i)}}{n(D|T_i) - 1} \tag{4.9}$$

The posterior probability of T_i can then be calculated by combining the evidential measure in (4.9) with the prior suggested by Velasco.

I will not discuss further the practical issue of how the value of a should be determined.[28] The more important theoretical upshot for our purposes is that verisimilitude-based cladistic parsimony apparently occupies an intermediate spot between traditional cladistic parsimony and a traditional 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.

Furthermore, the example in this section shows how we may improve our inference procedures even if we do not have a fully specified verisimilitude measure. As was noted earlier, the only assumption that we have made about the verisimilitude measure is that more parsimonious trees are closer to the truth than less parsimonious trees. Given this assumption, the principle of parsimony is a reasonable evidential principle, and the principle of parsimony and Coherence Requirement jointly entail that the evidential measure takes the form shown in (4.5). If we do not have a precise idea of the form that the verisimilitude measure ought to take, then a reasonable fully specified evidential measure is given in (4.9).[29] However, to the extent that we do have an idea of what sort of verisimilitude is relevant, the verisimilitude framework enables us to optimize the inference procedure by assigning a more opinionated prior probability distribution over the parameter a in (4.5). In the case where the verisimilitude measure is completely specified, the precise numerical value of a may even be determined, as was the case in (4.4) of the preceding section. Hence, in the same way that standard Bayesian inference allows us to incorporate relevant background information through a prior and thereby improve our inferences—provided that we have background knowledge—the verisimilitude framework allows us to improve our evidential measures, provided that we have some idea of what the verisimilitude measure we care about looks like.

[27]This form of Ev is arrived at by integrating $\mathrm{Ev}[D|T_i, a]p(a|T_i)$ over a, on the assumption that a and T_i are independent.

[28]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]).

[29]Given that there are four taxa; otherwise, the evidential measure will obviously look a bit different.

5 Conclusion

I have argued that the standard version of the Law of Likelihood often fails to make sense in practice. The verisimilitude version is better than the standard version, but even the verisimilitude version 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 the hypothesis that has the highest likelihood will not necessarily be the hypothesis that is closest to the truth in the relevant sense. Hence, the likelihood sometimes needs to be replaced by a different measure of evidential impact, and I have argued that the alternative evidential measure ought to be chosen so that it is calibrated against a verisimilitude measure 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.

My main contention is that the verisimilitude framework is a better—and more general—framework for philosophy of science than either Bayesianism or likelihoodism. The verisimilitude framework is better because it allows us to reconstruct in a broadly probabilistic framework inferential methods where the goal is merely closeness to the truth rather than truth itself, and were Bayesian or likelihoodist reconstructions are consequently not possible. Reconstructing inferential methods in the verisimilitude framework allows us to better understand the presuppositions and limitations of those inferential methods, as well as possible ways in which the methods may be improved.

This paper has focused on Bayesianism and likelihoodism in philosophy of science. However, as was noted in the introduction, Bayesianism and likelihoodism are not just frameworks that are useful in philosophy of science; they are also frameworks for statistical inference. Furthermore, in statistical inference—as in scientific inference in general—it's often the case that all the hypotheses under consideration are known to be false. The following question therefore arises: is the verisimilitude framework a viable—and better—framework statistical inference as well? That is a question for another paper.

Acknowledgements

I am grateful to Elliott Sober for reading a draft of this paper, and to the philosophy department at the National University of Singapore for helpful feedback. I am also grateful to three anonymous referees for very extensive and incisive comments. Research for this paper was supported by Nanyang Technological University Start-Up Grant M4082134.

Philosophy Programme, School of Humanities
Nanyang Technological University
Singapore, Singapore
vassend@ntu.edu.sg

References

Baltag, A. and Smets, S. [2011]: 'Keep Changing Your Beliefs, Aiming for the Truth', *Erkenntnis*, **75**(2), pp. 255–270.

Birnbaum, A. [1962]: 'On the Foundations of Statistical Inference', *Journal of the American Statistical Association*, **57**(298), pp. 269–326.

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

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. and Atkinson, Q. [2012]: 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, **337**(6097), pp. 957–960.

Cabrera, F. [2017]: 'Can There Be a Bayesian Explanationism? On the Prospects of a Productive Partnership', *Synthese*, **194**(4), pp. 1245–1272.

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

Cevolani, G., Festa, R. and Kuipers, T. A. F. [2011]: 'Verisimilitude and Belief Change for Conjunctive Theories', *Erkenntnis*, **75**(2), pp. 183–202.

Dawid, R., Hartmann, S. and Sprenger, J. [2015]: 'The No Alternatives Argument', *British Journal for the Philosophy of Science*, **66**(1), pp. 213–234.

Edwards, A. W. F. [1972]: *Likelihood: An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*, New York: Cambridge University Press.

Farris, J. S. [1983]: 'The Logical Basis of Phylogenetic Analysis', in N. I. Platnick and V. A. Funk (*eds*), *Advances in Cladistics*, vol. 2 of *Proceedings of the Second Meeting of the Willi Hennig Society*, Columbia University Press, New York, pp. 7–36.

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

Festa, R. and Cevolani, G. [2017]: 'Unfolding the Grammar of Bayesian Confirmation: Likelihood and Antilikelihood Principles', *Philosophy of Science*, **84**(1), pp. 56–81.

Forster, M. and Sober, E. [2004]: 'Why Likelihood?', in M. Taper and S. Lee (*eds*), *The Nature of Scientific Evidence*, Chicago: University of Chicago Press, pp. 153–165.

Forster, M. R. [2002]: 'Predictive Accuracy as an Achievable Goal of Science', *Philosophy of Science*, **69**, pp. S124–S134.

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), pp. 1–35.

Gandenberger, G. [2015]: 'A New Proof of the Likelihood Principle', *British Journal for the Philosophy of Science*, **66**(3), pp. 475–503.

Gelman, A., Meng, X.-L. and Stern, H. [1996]: 'Posterior Predictive Assessment of Model Fitness via Realized Discrepancies', *Statistica Sinica*, **6**, pp. 733–807.

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

Good, I. J. [1967]: 'The White Shoe is a Red Herring', *The British Journal for the Philosophy of Science*, **17**(4), pp. 322.

Gray, R. D., Bryant, D. and Greenhill, S. J. [2010]: 'On the Shape and Fabric of Human History', *Philosophical Transactions of the Royal Society B*, **365**(1559), pp. 3923–3933.

Hawthorne, J. [1994]: 'On the Nature of Bayesian Convergence', *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **1**, pp. 241–249.

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), pp. 3829–3843.

Ibrahim, J. G., Chen, M.-H., Gwon, Y. and Chen, F. [2015]: 'The Power Prior: Theory and Applications', *Statistics in Medicine*, **34**(28), pp. 3724–3749.

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

Jiang, W. and Tanner, M. A. [2008]: 'Gibbs Posterior for Variable Selection in High-Dimensional Classification and Data Mining', *The Annals of Statistics*, **36**(5), pp. 2207–2231.

Joyce, J. [1998]: 'A Non-Pragmatic Vindication of Probabilism', *Philosophy of Science*, **65**(4), pp. 575–603.

Jurman, G., Visintainer, R., Filosi, M., Riccadonnam, S. and Furlanello, C. [2015]: 'The HIM Glocal Metric and Kernel for Network Comparison and Classification', *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, **36678**, pp. 1–10.

Kuipers, T. A. F. [2011]: 'Basic and Refined Nomic Truth Approximation by Evidence-Guided Belief Revision in AGM-Terms', *Erkenntnis*, **75**(2), pp. 223–236.

Kullback, S. and Leibler, R. [1951]: 'On Information and Sufficiency', *Annals of Mathematical Statistics*, **22**(1), pp. 79–86.

Newton, I. [1999]: *The Principia: Mathematical Principles of Natural Philosophy*, University of California Press: Berkeley and Los Angeles, California.

Niiniluoto, I. [1986]: 'Truthlikeness and Bayesian Estimation', *Synthese*, **67**(2), pp. 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), pp. 1–29.

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

Oddie, G. [Forthcoming]: 'What Accuracy Could Not Be', *British Journal for the Philosophy of Science*.
<https://doi.org/10.1093/bjps/axx032>

Okayasu, E. [2017]: *Justifying the Comparative Method in Historical Linguistics*, Ph.D. thesis, University of Wisconsin – Madison.

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

Pettigrew, R. [2016]: *Accuracy and the Laws of Credence*, Oxford University Press.

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

Renardel de Lavalette, G. R. and Zwart, S. D. [2011]: 'Belief Revision and Verisimilitude Based on Preference and Truth Orderings', *Erkenntnis*, **75**(2), pp. 237–254.

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

Royall, R. [1997]: *Statistical Evidence: A Likelihood Paradigm*, CRC Press.

Schupbach, J. N. [2018]: 'Robustness Analysis as Explanatory Reasoning', *British Journal for the Philosophy of Science*, **69**(1), pp. 275–300.

Schurz, G. [2011]: 'Verisimilitude and Belief Revision. With a Focus on the Relevant Element Account', *Erkenntnis*, **75**(2), pp. 203–221.

Sober, E. [1988]: *Reconstructing the Past: Parsimony, Evolution, and Inference*, Cambridge, M.A.: MIT Press.

Sober, E. [2008]: *Evidence and Evolution: The Logic Behind the Science*, Cambridge University Press.

Sober, E. [2015]: *Ockham's Razors: A User's Manual*, Cambridge University Press.

Sprenger, J. [Unpublished]: 'Conditional Degree of Belief', <http://philsci-archive.pitt.edu/13515/>

Tarski, A. [1944]: 'The Semantic Conception of Truth: and the Foundations of Semantics', *Philosophy and Phenomenological Research*, **4**(3), pp. 341–376.

Vassend, O. B. [Unpublished]: 'New Semantics for Bayesian Inference: The Interpretive Problem and Its Solutions', <https://sites.google.com/site/olavbvassend/research>

Velasco, J. D. [2008]: 'The Prior Probabilities of Phylogenetic Trees', *Biology & Philosophy*, **23**, pp. 455–473.

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

Wasserman, L. [2004]: *All of Statistics: A Concise Course in Statistical Inference*, Springer.

Zhang, T. [2006]: 'From e-Entropy to KL-Entropy: Analysis of Minimum Information Complexity Density Estimation', *The Annals of Statistics*, **34**(5), pp. 2180–2210.