Bayesians Can Learn from Old Data

William H. Jefferys

Citation: AIP Conference Proceedings **954**, 85 (2007); doi: 10.1063/1.2821303
View online: http://dx.doi.org/10.1063/1.2821303
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/954?ver=pdfcov
Published by the AIP Publishing

Bayesians *Can* Learn from Old Data

William H. Jefferys

University of Texas at Austin, and University of Vermont

Abstract. In a widely-cited paper, Glymour (*Theory and Evidence*, Princeton, N. J.: Princeton University Press, 1980, pp. 63-93) claims to show that Bayesians cannot learn from old data. His argument contains an elementary error. I explain exactly where Glymour went wrong, and how the problem should be handled correctly. When the problem is fixed, it is seen that Bayesians, just like logicians, *can indeed* learn from old data.

Keywords: Logic, Probability Theory, Bayesian Inference, Problem of Old Data
PACS: 02.10.Ab, 02.50.Cw, 02.50.Tt

GENERAL OVERVIEW

Outline of the Paper. I first review some aspects of standard logic that are relevant to this paper. I then discuss the relationship between standard logic and standard probability theory, and in particular point out the fact that standard probability theory contains standard logic in the particular sense that for any argument that reaches a conclusion using standard logic, there exists a parallel argument (calculation) in standard probability theory that reaches the same conclusion, and furthermore, that any *valid* argument by any method (whether logical or Bayesian) must arrive at the same conclusion.

I then introduce a simple "toy example" that is nonetheless sophisticated enough to reveal the problem with Glymour's claim. The toy example is an extension of the example that Glymour used in his paper. I describe Glymour's argument [1], and use the toy example to show that his reasoning leads to a contradiction with ordinary logic, and therefore must be invalid. I then explain, again in terms of the toy example, exactly where Glymour's argument goes wrong, and how to correct it. I conclude with a summary of what we have learned.

Standard Logic

Standard logic tells us how to combine propositions $A, B, C, \ldots$ with logical operations such as $\wedge, \vee, \neg, \rightarrow, \ldots$ to obtain new and valid propositions. The propositional calculus allows us to calculate, using definite rules, the truth value of any proposition that has been constructed from other propositions using these logical operations, given the truth values of the propositions from which they are constructed.

For example, given propositions A, B, we can calculate the truth value of the proposition $C = A \wedge B$ as follows: C is true if both A and B are true, otherwise it is false.

Likewise, the truth value of the proposition $D = A \rightarrow B$ is true if A is false, otherwise it is equal to the truth value of B. That is, if A is true, then B must be true. If A is not true, then it doesn't matter what the truth value of B is, $A \rightarrow B$ is true.

An important feature of standard logic is that it is time-independent (Jaynes [2], p. 89). That is, it describes relationships between propositions that are independent of when we may learn the truth or falsity of the propositions themselves. For example, the truth-values of the expressions $\neg A$, $A \wedge B$, $A \vee B$, and $A \rightarrow B$ depend only on the truth-values of A and B, and not upon when we happen to learn their truth-values.

Probability and Logic

Probability theory extends the basic notions of standard logic to a regime where the *degree of plausibility* of propositions is no longer just "true" or "false", but may be intermediate between the two. That is, to any proposition we can assign a number in the unit interval $[0,1]$ that corresponds to our assessment of how likely it is that the proposition is true, where 1 means that we are certain the proposition is true and 0 means that we are certain that it is false. The larger the degree of plausibility, the more likely it is that we would regard the proposition as true.

A theorem of Cox [3, 4] proves that, up to an isomorphism, standard probability theory is the unique extension of ordinary logic to this regime that satisfies certain obvious requirements necessary for the theory to yield consistent results. Jaynes ([2], p. 19) lists a set of three such requirements, which he calls *desiderata*:

1 *If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.* An important aspect of this desideratum is that if a conclusion can be obtained using ordinary logic, then a *valid* calculation using probability theory must arrive at the same result. If a purported Bayesian calculation arrives at a result different from one that we can derive using standard logic, it must *ipso facto* be invalid. We will see below that Glymour's calculation fails this test.

2 *The calculation takes into account all of the evidence relevant to the question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains. It is, as Jaynes says, completely nonideological.* Glymour's calculation fails this test in a subtle way, muddling the issue by failing use standard probability notation to indicate all the information that was taken into account in a calculation. Indeed, this results in a basic confusion of models that turns out to be at the root of the problem with Glymour's calculation.

3 *Equivalent states of knowledge are always represented by equivalent plausibility assignments. That is, if in two situations the state of knowledge is the same, then (except for possible relabeling of the propositions), the calculation must assign the same plausibilities to both.* Glymour's calculation fails this test as well.

It turns out that these three desiderata, together with the assumption that degrees of plausibility are represented by real numbers on the unit interval $[0,1]$, are sufficient to

derive standard probability theory as the unique embodiment of these sensible requirements of plausible reasoning.

In particular it turns out, as a consequence of Jaynes' desideratum #1 and Cox's theorem, that standard probability theory contains standard logic as a subset. This means that for every calculation that can be made using standard logic, there is a corresponding calculation in standard probability theory that will arrive at the *same* result, and no *valid* calculation in standard probability theory can yield a different result.

A Toy Example

We consider a situation where there are precisely two theories under consideration, say T and $\overline{T} = \neg T$, and only two observations of evidence are possible, that is E and $\overline{E} = \neg E$. We furthermore presume that $T \rightarrow E$ and $\overline{T} \rightarrow \overline{E}$. This means that if theory T is true, we must observe evidence E, and if theory $\overline{T}$ is true, then we must observe evidence $\overline{E}$.

For example, let T be the theory of general relativity, and $\overline{T}$ be pure Newtonian mechanics. Let E be the (in this case old) evidence that the motion of Mercury's perihelion is anomalous (cannot be explained under Newtonian mechanics). I assume that we can be certain whether we have observed anomalous perihelion motion or not. Then we see immediately that in this toy example $T \rightarrow E$ and $\overline{T} \rightarrow \overline{E}$.

It is important to recognize that these relationships are *defined by the theory*, independently of any data that may have been observed and independently of when those data may have been observed. The relationships are therefore *time-independent*. Newtonian theory *always* predicts that anomalous perihelion motion will *not* be observed, and general relativity *always* predicts that anomalous perihelion motion *will* be observed. This is a consequence of the theories and mathematics.

If we observe evidence E, then standard logic says $\overline{T} \rightarrow \overline{E}$, so $\neg T \rightarrow \neg E$. It follows that $E \rightarrow T$ and $E \rightarrow \neg\overline{T}$. Hence observing E rules out $\overline{T}$ and confirms T.

Note that this result follows from standard logic. Since standard logic is just a calculus on the truth-values of the propositions, and does not depend on when we observe evidence E, it follows that we can certainly learn from old evidence if we use only logic. But, as pointed out above, Jaynes' desideratum #1, together with Cox's theorem, says that the same result *must* be obtainable by a *valid* application of probability theory. If a calculation using probability theory obtains a different result, it is certainly not a valid calculation.

Translated into the language of probability theory, the result $E \rightarrow \neg\overline{T}$ is equivalent to $P(\neg\overline{T} \mid E) = P(T \mid E) = 1$ and $P(\overline{T} \mid E) = 0$. Any purported Bayesian calculation that does not arrive at this result must be invalid. Note also that when we translate the initial assumptions of this toy example into standard probability notation we can calculate the *likelihood* as $P(E \mid T) = 1$ and $P(E \mid \overline{T}) = 0$ for use when we observe E, and $P(\overline{E} \mid T) = 0$ and $P(\overline{E} \mid \overline{T}) = 1$ for use when we observe $\overline{E}$. Since all of these probability assignments are simply translations of statements of ordinary logic into the language of probability theory, they are time-independent, that is, their values are independent of when we happen to observe the evidence.

GLYMOUR'S ARGUMENT

Glymour argues [1] that the Bayesian cannot learn from old evidence E. This article has generated a lively discussion, e.g., [5, 6, 7, 8, 9, 10, 11, 12]. The argument goes as follows[1]: Since we have observed the old evidence E, Glymour claims that

$$P(E) = 1 \quad ??? \tag{1}$$

I put question marks here because I believe this equation to be wrong. Nonetheless, if we grant Eq. (1), Glymour's argument goes through easily. Since $P(E) = 1$, it follows from standard probability theory that $P(E \mid X) = 1$ for all propositions X that are not absurd (tautologically false). In particular, $P(E \mid T) = 1$. Therefore, by Bayes' theorem,

$$P(T \mid E) = \frac{P(E \mid T)}{P(E)} P(T) = P(T)$$

and since the posterior probability is equal to the prior probability, we haven't learned anything.

Counterexample to Glymour's Argument

We see immediately that Glymour's calculation fails to satisfy Jaynes' desideratum #1, for we have proved that for our toy problem, knowledge of E together with standard logic leads to the conclusion that T is true and $\overline{T}$ is false, regardless of what we may have thought before we did the calculation. But Glymour's calculation allows for no such conclusion: If for example we had adopted $P(T) = 1/2$, Glymour's calculation tells us that $P(T \mid E) = 1/2$, in blatant contradiction to the calculation from ordinary logic. The equation $P(T \mid E) = 1/2$ says that E does *not* entail T, whereas logic says that E *does* entail T. Since Cox's theorem guarantees that any *valid* calculation using probability theory must arrive at the same conclusion that we got using standard logic, this fact by itself demonstrates that Glymour's argument cannot be valid.

It is not hard to pinpoint the source of the problem, again using the toy example as a guide. If $P(E) = 1$, then it follows that $P(E \mid X) = 1$ for any non-absurd proposition X; in particular, $P(E \mid \overline{T}) = 1$, or translated into the language of logic, $\overline{T} \rightarrow E$. That is, according to Glymour's reasoning, if we have observed E, we must conclude that *Newtonian physics predicts that we will observe anomalous motion of the perihelion of Mercury.* But this is absurd. Newtonian physics predicts unambiguously that we will *not* observe anomalous perihelion motion for Mercury, that is, $\overline{T} \rightarrow \overline{E}$. This is a property of the *theory*, which doesn't depend in any way on what observations may or may not have been made.

The absurdity of this situation is compounded when we realize that Glymour's reasoning transforms *every* theory X into Jaynes' dreaded "Sure Thing®" theory [13], which predicts the observed data E *perfectly*.

---

[1] I have altered Glymour's notation to conform to standard probability theory

We have thus arrived at a contradiction. Glymour's reasoning would require us to conclude that $\overline{T} \rightarrow E$, but we know from physics that $\overline{T} \rightarrow \neg E$, independent of time or what we may have observed. Therefore, Glymour's reasoning must be erroneous.

The problem arises from Glymour's assertion that $P(E) = 1$. Without that, the rest of his alleged proof fails.

Glymour's Friend

Physicists are familiar with "Wigner's Friend," a thought experiment named for the late physicist Eugene Wigner, that is designed to help us think about when and under what circumstances the "collapse" of states in quantum mechanics takes place. In this thought experiment, Wigner and his "friend" have different states of knowledge, until Wigner's friend informs Wigner of certain facts, so that they end up with the same state of knowledge, and thus should come to the same conclusions. The details of the physics aren't important here, but the idea that people who start out with different states of knowledge will arrive at the same conclusions, once they have the same state of knowledge, is the key idea that I want to carry over to the present problem.

Let me introduce Glymour's friend Tom. Tom is ignorant of E. Therefore, when Glymour explains the toy problem to Tom, Tom can decide on priors and even calculate in advance what he will think when he learns whether E is true or false, using the usual Bayesian machinery. After he has done this, Tom can tell Glymour what his priors are. Suppose the priors are the same as the ones that Glymour has already adopted, and that $P(T) \neq 1$. Then both are starting with the same priors.

Now Glymour informs Tom that E is true. Tom, upon learning this "new" data, recalls his previous calculations, and concludes that $P(T \mid E) = 1$. Glymour performs the calculation that he advocates (since for him the data are "old") and arrives at $P(T \mid E) = P(T) \neq 1$.

This violates Jaynes' desideratum #3, since at this point both parties have the same state of knowledge, yet they have assigned different plausibilities to $T \mid E$. Since the axioms of probability theory, in virtue of Cox's theorem, cannot violate Jaynes' three desiderata when used validly, we have again arrived at a contradiction. Since it is clear that Tom does not view E as "old" data, and therefore is entitled to carry out the standard Bayesian calculation (which gives the same result as the calculation using logic), his conclusions must be correct and Glymour's wrong.

Where Glymour Went Wrong

Jaynes ([2], pp. 473, 484) points out an important fact: *A fruitful source of error and even apparent paradoxes in probability theory is to fail to condition properly and explicitly on all background information used.* All probability is conditional on every relevant piece of background information, and changing the background information changes the probabilities. To make this crystal clear, let $\mathscr{B}$ represent *all* the relevant background information at our disposal, *except* for any knowledge of E. This includes

our assumptions about mathematics and physics; for example, $\mathscr{B}$ includes the fact that $\overline{T} \rightarrow \overline{E}$.

Viewed from this point of view, the source of Glymour's error becomes embarrassingly obvious. Recall that Eq. (1) was derived in the light of knowledge of the old evidence E and *actually used* that information as background information, even though this dependence was not explicitly noted in the equations. Following Jaynes' advice above, standard notational convention demands that we call out this fact explicitly. If we do this, we obtain the correct Eq. (2):

$$P(E \mid E, \mathscr{B}) = 1 \quad !!! \tag{2}$$

The rest of the proof translates as follows:

$$P(E \mid E, T, \mathscr{B}) = 1 \tag{3}$$

$$P(T \mid E, E, \mathscr{B}) = \frac{P(E \mid E, T, \mathscr{B})}{P(E \mid E, \mathscr{B})} P(T \mid E, \mathscr{B}) \tag{4}$$

But of course, $P(T \mid E, E, \mathscr{B}) = P(T \mid E \wedge E, \mathscr{B}) = P(T \mid E, \mathscr{B})$ by standard logic. Thus we see that when the conditioning that is implicit but unstated in Eq. (1) is explicitly recognized in Eq. (2), what Glymour has actually proved is the (well-known) fact that the Bayesian machinery, quite sensibly, prevents us from using the same evidence twice. He has *not* proved that a Bayesian cannot learn from old evidence, only that he cannot validly manipulate the Bayesian machinery to get additional information out of information that has *already been used*.

We now see that $P(E \mid \mathscr{B})$ and $P(E \mid E, \mathscr{B})$ are entirely different. $P(E \mid E, \mathscr{B})$ has already used evidence E, whereas according to the standard notational convention, $P(E \mid \mathscr{B})$—Glymour's $P(E)$—has *never* used evidence E, not even once. This is because $P(E \mid \mathscr{B})$ is just the *sampling distribution* of E in the mixture model defined by the priors and the likelihood, given that we know $\mathscr{B}$ and *nothing else.* It is a function of the *theory*, which is included in the background knowledge $\mathscr{B}$. It is in fact *entirely ignorant* of our knowledge of E. Thus, there is no reason to suppose that $P(E \mid \mathscr{B}) = 1$, regardless of our state of knowledge of E, and indeed, it usually is not.

Note that the right-hand side of Eq. (4) has its as prior $P(T \mid E, \mathscr{B})$, *not* $P(T \mid \mathscr{B})$. In other words, the prior in Eq. (4) must be constructed from full knowledge of E; it is *not* the same as $P(T \mid \mathscr{B})$, which is (of course) ignorant of E. One cannot substitute $P(T \mid \mathscr{B})$ for $P(T \mid E, \mathscr{B})$ in Eq. (4); the resulting equation is not a valid equation in probability theory.

In order to calculate the value of $P(T \mid E, \mathscr{B})$ for substitution into Eq. (4), we have to start from $P(T \mid \mathscr{B})$ and then apply Bayes' theorem in the usual way, where in this case the right hand side is calculated *unconditioned* on E (which is to say, the right-hand side is ignorant of any knowledge we may have about E). In this case, $P(E \mid \mathscr{B})$ does not know that E has been observed, and is correctly calculated from the priors and the *time-independent* likelihood from the identity:

$$P(E \mid \mathscr{B}) = P(E \mid T, \mathscr{B}) P(T \mid \mathscr{B}) + P(E \mid \overline{T}, \mathscr{B}) P(\overline{T} \mid \mathscr{B}) \tag{5}$$

Thus, in the toy example, where $P(E \mid T, \mathscr{B}) = 1$ and $P(E \mid \overline{T}, \mathscr{B}) = 0$,

$$P(E \mid \mathscr{B}) = P(T \mid \mathscr{B}), \tag{6}$$

which is in general *not* equal to 1.

This tells us the correct way to do the Bayesian calculation, in the case where E has been observed as old data. We still have to assign priors $P(T \mid \mathscr{B})$ and $P(\overline{T} \mid \mathscr{B})$, and this must be done without taking E into account. Although this step might pose some problems of its own (assignation of priors in general requires careful thought), any such problems are unrelated to Glymour's argument, so I will pass over this issue. Suppose, for example, we have assigned $P(T \mid \mathscr{B}) = \alpha$, $P(\overline{T} \mid \mathscr{B}) = 1 - \alpha$, where $\alpha \in (0,1)$. Then the Bayesian calculation goes through in the usual way as follows:

$$P(T \mid E, \mathscr{B}) = \frac{P(E \mid T, \mathscr{B})}{P(E \mid \mathscr{B})} P(T \mid \mathscr{B}) = 1 \tag{7}$$

since in the toy example example $P(E \mid \mathscr{B}) = P(T \mid \mathscr{B})$ and—from the *theory*, not from Glymour's reasoning—$P(E \mid T, \mathscr{B}) = 1$. Thus, independent of α, we obtain the same result as we did using ordinary logic. Thus, Jaynes' desideratum #1 is satisfied: No matter how we do the calculation, whether by ordinary logic or by a *valid* application of probability theory, Cox's theorem guarantees that we *must* arrive at the same result.

Note in particular that Glymour's argument does not use, and in fact denies, the *one key fact* that allows us to calculate the correct result using logic: that $P(E \mid \overline{T}, \mathscr{B}) = 0$. From this fact, we first derive $P(\overline{E} \mid \overline{T}, \mathscr{B}) = 1$, which in turn implies $\overline{T} \wedge \mathscr{B} \rightarrow \overline{E}$ and then (since $\mathscr{B}$ is true) $\overline{T} \rightarrow \overline{E}$ and $E \rightarrow T$. But the correct Bayesian calculation makes full use of that information by using the time-independent likelihood in the calculation of $P(E \mid \mathscr{B})$ to arrive at the *same* result that we got using logic. Glymour's calculation thus violates Jaynes' desideratum #2.

SUMMARY AND CONCLUSIONS

As Jaynes ([2], p. 89) points out, probability theory, like logic, is time-independent. All of the relationships in probability theory are *logical* relationships and have nothing to do with the order in which we happen to learn about the evidence or recognize it in the Bayesian equations. When we calculate $P(T \mid E, \mathscr{B})$ from $P(T \mid \mathscr{B})$, it does not matter when we have actually observed E; the relationship between the two is purely a logical relationship, and the quantities that go into the calculation (likelihoods, priors) are time-independent and will be the same, regardless of when E happens to have been observed. As Tom Loredo observed when I showed him Glymour's argument, "Time plays the same role in probability theory as it does in logic: That is to say, no role whatsoever." [14]

A *valid* Bayesian calculation takes ones knowledge of a particular piece of data into account in just one uniform way, by conditioning on the data. It is essential that this conditioning be called out *explicitly* in the notation, as Jaynes advises. Using data without explicitly calling it out in the notation, as Glymour did, is a reliable route to disaster.

Glymour's error resulted from a failure to follow these basic principles. Using the principles of his argument I was able to derive a contradiction with logic that seems not to have been noticed up to this point, but which is sufficient to demonstrate that Glymour's argument is invalid. The bottom line is that Bayesians can and do learn from old data, when they do the calculation carefully and correctly.

ACKNOWLEDGMENTS

I thank Jim Berger, David van Dyk and especially Rob Pennock and Tom Loredo for their valuable comments and suggestions. I dedicate this paper to the memory of Edwin T. Jaynes. I was not fortunate to know him personally, but I have learned much from his writings.

REFERENCES

1. C. N. Glymour, "Why I Am Not a Bayesian," in *Theory and Evidence*, Princeton, N. J.: Princeton University Press, 1980, pp. 63–93.
2. E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press, 2003.
3. R. T. Cox, *American Journal of Physics* **14**, 1–13 (1946).
4. K. S. Van Horn, *International Journal of Approximate Reasoning* **34**, 3–24 (2003).
5. M. Curd, and J. A. Cover, *Philosophy of Science: The Central Issues*, New York: W. W. Norton & Co., 1998, pp. 656–659.
6. J. Earman, *Bayes or Bust?*, Cambridge, MA: MIT Press, 1992, chap. 5.
7. D. Garber, "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory," in *Minnesota Studies in the Philosophy of Science, Volume X: Testing Scientific Theories*, edited by J. Earman, Minneapolis: University of Minnesota Press, 1983, pp. 99–131.
8. C. Howson, *British Journal of the Philosophy of Science* **42**, 547–555 (1991).
9. C. Howson, and P. Urbach, *Scientific Reasoning: The Bayesian Approach*, La Salle, IL: Open Court, 1989, pp. 270–275.
10. R. Jeffrey, "Bayesianism with a Human Face," in *Minnesota Studies in the Philosophy of Science, Volume X: Testing Scientific Theories*, edited by J. Earman, Minneapolis: University of Minnesota Press, 1983, pp. 133–156.
11. R. T. Pennock, *Annals of the Japan Association for the Philosophy of Science* **13**, 1–26 (2004).
12. R. Rosenkrantz, "Why Glymour Is a Bayesian," in *Minnesota Studies in the Philosophy of Science, Volume X: Testing Scientific Theories*, edited by J. Earman, Minneapolis: University of Minnesota Press, 1983, pp. 69–97.
13. E. T. Jaynes, "Bayesian Methods: General Background," in *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice, Cambridge: Cambridge University Press, 1985, pp. 1–25.
14. T. Loredo, *Private communication* (2006).