RECONCILING BAYESIAN AND NON-BAYESIAN ANALYSIS

David H. Wolpert
The Santa Fe Institute, 1660 Old Pecos Trail
Santa Fe, NM, 87501, USA (dhw@santafe.edu)

ABSTRACT. This paper is an attempt to reconcile Bayesian and non-Bayesian approaches to statistical inference, by casting both in terms of a broader formalism. In particular, this paper is an attempt to show that when one extends conventional Bayesian analysis to distinguish the truth from one's guess for the truth, one gains a broader perspective which allows the inclusion of non-Bayesian formalisms. This perspective shows how it is possible for non-Bayesian techniques to perform well, despite their handicaps. It also highlights some difficulties with the "degree of belief" interpretation of probability.

1. Introduction

Why should one want to reconcile Bayesian and non-Bayesian analysis? Why be bothered with non-Bayesian techniques? Bayesian analysis forces one to makes one's assumptions explicit; it ensures self-consistency; it provides a single unified approach to all inference problems; if one is very sure of the prior (e.g., as an extreme, you constructed the data-generating mechanism yourself) it is essentially impossible to beat; and in some ways most important of all (sociologically speaking), Bayesian analysis is in some senses more elegant than non-Bayesian analysis.

For these very reasons I have used Bayesian techniques in the past and will do so again in the future. As compelling as these reasons are though, none of them constitute a proof that Bayesian techniques perform better than non-Bayesian techniques in the real world. Indeed, there are many examples — some constructed by self-professed Bayesians — which cast doubt on such a guarantee. For example, as is discussed separately at this conference [1], although the "evidence" procedure sometimes works well in practice [2, 3], and although it is championed by fervent Bayesians, careful scrutiny reveals that it is a non-Bayesian technique. In particular, in [3], in a section entitled "Why Bayes can't systematically reject the truth", MacKay presents a theoretical argument for the evidence procedure's setting hyperparameters by maximum likelihood. However this argument can be extended to "justify" setting *any* parameter by maximum likelihood, not just a hyperparameter. It is hard to imagine a more non-Bayesian line of reasoning. As another example, despite being a self-professed fervent Bayesian, MacKay recently won a prediction competition using an extension of the non-Bayesian technique of cross-validation [4].

Another reason not to dismiss non-Bayesian techniques arises from the problem of setting the probability distribution $P(\text{truth}= t, \text{data}= d)$.[1] If—as is often the case in the real world—we already know the likelihood $P(d \mid t)$, in what ways can we fix the remaining degrees of freedom in the joint distribution while ensuring consistency with the laws of probability theory? One way to do this is to provide the prior distribution $P(t)$—this is the basis for conventional Bayesian analysis. But there are other ways as well. For example, consider the case where there is a data set $d = d'$ such that $P(d' \mid t)$ does not exactly equal zero for any t. If we now provide the values of $P(t \mid d')$ for all possible t, we will have fixed the entire joint distribution, for all possible data sets. (This follows from the equality $P(t_i, d_j) = \frac{P(t_i|d')P(d_j|t_i)/P(d'|t_i)}{\sum_{k,m} P(t_k|d')P(d_m|t_k)/P(d'|t_k)}$.) In particular, by setting $P(t \mid d')$, we will have fixed $P(t \mid d)$ for any $d \neq d'$.

With this alternative scheme we would be assured of self-consistency, our assumptions would be explicit, etc.; this scheme possesses all the formal strengths of conventional Bayesian analysis. However rather than use pseudo-intuitive arguments to set $P(t)$, as in the conventional prior-based Bayesian approach, with this alternative scheme we use such arguments to set $P(t \mid d)$ for one particular d. For example, one could set $P(t \mid d)$ for one particular d using "pseudo-intuitive" cross-validation type arguments. One might even be able to use "desiderata" to set $P(t \mid d)$ for one specific d, rather than to set $P(t)$. There is no reason prior knowledge has to concern a prior probability; one can have prior knowledge that is expressed directly as a posterior. For example, my "prior knowledge" might consist of knowing that cross-validation works well for a certain class of problems.

In fact many practicing statisticians do implicitly exploit "prior knowledge" directly concerning the posterior. However they do so in conjunction with a approximation; they have their prior knowledge set all of $P(t \mid d)$ at once, and therefore they (usually) violate strict consistency with the laws of probability. For historical reasons, such approximations are usually called "non-Bayesian". However they are closely analogous to using a conventional (i.e., prior-based) Bayesian analysis which involves calculational approximations, and which therefore also violates strict consistency with the laws of probability. So the question arises of how accurate the approximations in a Bayesian technique must be to "beat" a particular non-Bayesian technique. (From here on the term "Bayesian" will mean conventional, prior-based Bayesian.) To address this and related issues we need to use a new formalism.

2. A Formalism for Reconciling Bayes and Non-Bayes

In most inference problems there are four quantities of interest: the data d, the truth t (which might be a probability distribution), one's "guess for the truth" g, and a real world "cost" or "loss" or "utility" accompanying a particular use of one's inference technique. (For many scenarios cost only depends on t and g, and g is formally called a "decision".) Accordingly, the inference process is governed by $P(t, g, d, c)$.

Now conventional Bayesian analysis doesn't distinguish t from g—it does not analyze joint distributions over those two variables. Therefore one must be careful in relating $P(t, g, d, c)$ to the distributions used in Bayesian analysis. In particular, note that the "posterior" of Bayesian analysis is $P(t \mid d)$, not $P(g \mid d)$. This follows from how a Bayesian uses Bayes' theorem to set the "posterior" in terms of the likelihood. Since the likelihood is $P(d \mid \text{truth} = t)$, not $P(d \mid \text{guess} = g)$, the "posterior" must be $P(t \mid d)$.

$P(g \mid d)$ is a different kind of object which has no analogue in Bayesian analysis. It is the probability of making a guess g given data d. In other words, it is one's algorithm for performing statistical inference. A priori, it need have nothing to do with Bayesian techniques, and need not even be expressible in "Bayesian" terms. (For example, the evidence procedure's $P(g \mid d)$ can not be expressed this way, since there is always necessarily some difference between it and full hierarchical Bayesian analysis—see [1].) As such, $P(g \mid d)$ is the object which allows one to expand the discussion to consider non-Bayesian techniques.

One nice feature of this "extended" Bayesian framework is that in it, the difference between conventional Bayesian analysis and (most forms of) non-Bayesian analysis is no longer some quasi-philosophical preference for different statistical dogmas. Instead that difference reduces to simply what conditional distribution the two formalisms choose to evaluate. Bayesian analysis is concerning with finding the $P(g \mid d)$ that optimizes $P(c \mid d)$, and sampling theory statistics with evaluating $P(c \mid t, m)$ (m being the data set size). It is only with the extended Bayesian framework that one can consider both at once, and thereby investigate the subtle connections between the two [7].

The implicit view in this extended framework is that inference is a 2-person game pitting you, the statistician, against the data-generating mechanism, aka the universe. Your opponent draws truths t at random, according to $P(t)$, and then randomly produces a data set from t, according to $P(d \mid t)$. This d is shown to you. Based on d, you guess a g according to $P(g \mid d)$. We then use some cost function to determine how well g matches t. Note that if you know $P(t)$ and $P(d \mid t)$, then you can use that information to perform optimally. But if you don't know $P(t)$ exactly (!) and therefore have to guess it—as in the real world—you have no such assurance.

In fact, extended Bayesian analysis can be used to prove the following (see [5, 6])):

Theorem 1: $P(c \mid d) = \sum_{g,t} P(g \mid d)\ P(t \mid d)\ M_{c,d}(g,t)$, for some matrix M parameterized by c and d.

(A similar result holds if g and t are not countable.)

In many situations M is symmetric, in which case theorem (1) means that $P(c \mid d)$ is given by an inner product between the posterior and one's inference algorithm. In other words, how well your algorithm performs is determined by how "aligned" it is with the true posterior. In particular, theorem (1) allows that a Bayesian's $P(g \mid d)$ might not be predicated on the actual $P(t \mid d)$, and therefore might perform poorly—perhaps even worse than a non-Bayesian $P(g \mid d)$. Such mismatch between the Bayesian's $P(g \mid d)$ and $P(t \mid d)$ can occur even if the Bayesian somehow knows $P(t)$ and $P(d \mid t)$, if the Bayesian's $P(g \mid d)$ uses those distributions in conjunction with calculational approximations. So in general there are two issues confronting both the Bayesian and the non-Bayesian: i) how accurately $P(g \mid d)$—based as it is on assumptions and approximations—aligns with $P(t \mid d)$, and ii) how probability of cost varies with changes in that accuracy.

In fact, if the inference problem is to build a classifier, so that both g and t are mappings from features vectors to classification labels, and if one's cost is determined by how well g matches t for feature vectors outside of the data set, one has the following theorem [6]:

Theorem 2: Let $E(\cdot)$ indicate an expectation value, and m the size of the "training set" d. For any two inference algorithms $P_1(g \mid d)$ and $P_2(g \mid d)$, independent of the noise,

i) if there exists a t such that $E(c \mid t, m)$ is lower for $P_1(g \mid d)$, then there exists a different t such that $E(c \mid t, m)$ is lower for $P_2(g \mid d)$;

ii) if there exists a t and a d such that $E(c \mid t, d)$ is lower for $P_1(g \mid d)$, then there exists a different t and d such that $E(c \mid t, d)$ is lower for $P_2(g \mid d)$;

iii) if there exists a $P(t)$ and a d such that $E(c \mid d)$ is lower for $P_1(g \mid d)$, then there exists a different $P(t)$ such that $E(c \mid d)$ is lower for $P_2(g \mid d)$;

iv) if there exists a $P(t)$ such that $E(c \mid m)$ is lower for $P_1(g \mid d)$, then there exists a different $P(t)$ such that $E(c \mid m)$ is lower for $P_2(g \mid d)$.

All of this holds whether or not the inference algorithms in question are constructed in a Bayesian manner. Moreover these (and associated) results don't just say that a non-Bayesian algorithm might beat a Bayesian in one particular trial, by luck. Rather a non-Bayesian algorithm might win on average. In fact, not only does theorem (2) not rely on pathological trials; it also doesn't rely on pathological processes generating the trials. For example 2(iv) can be recast as "averaged over all $P(t)$, $E(c \mid m)$ is the same for all learning algorithms". So for any two inference algorithms, there are "just as many" $P(t)$'s (loosely speaking) for which algorithm one has a lower expected cost as there are for which algorithm two's expected cost is lower. Unless you somehow know $P(t)$ rather than just guess it, your being a Bayesian provides no guarantees.

From this perspective, the Bayesian approach is the approach of choice only if there is no alternative (non-Bayesian) approach which is sufficiently compelling in comparison. (The comparison being between how compelling is a $P(g \mid d)$ based on a guess for $P(t)$ vs. a $P(g \mid d)$ based on other considerations.) Those (not at all uncommon) scenarios in which the Bayesian approach works well compared to non-Bayesian techniques do not reflect some inherent "*a priori* superiority" of Bayesian techniques. Rather they reflect the fact that at least some aspects of the non-Bayesian techniques considered in those scenarios are not sufficiently powerful in comparison to the corresponding Bayesian techniques. Indeed, the utility of using "sufficiently powerful" non-Bayesian approaches when possible is explicitly acknowledged in several variations of Bayesian analysis, like empirical Bayes and ML-II [9].

A particularly important implication of this is that there is nothing inherently bad about using a non-Bayesian algorithm to choose between Bayesian and/or non-Bayesian techniques. For example, if we have little information concerning $P(t)$ (and especially in the limiting case of no knowledge—a case sometimes dealt with via an "uninformative prior"), then it makes sense to be suspicious of any guess for $P(t)$ (even a guess that $P(t)$ is "uninformative"). Therefore it is reasonable to be suspicious of any $P(g \mid d)$ constructed under that guess. In such a scenario, one need not be shy about using something like cross-validation to choose amongst the techniques, or even about using stacked generalization to combine them [8].

All this provides suggestions of what some non-Bayesian formalisms are "getting at". For example, if one knows $P(t)$ exactly, then Bayesian techniques incorporating that knowledge into $P(g \mid d)$ always win, on average (assuming we also know the likelihood). However imagine we have limited information concerning $P(t)$. In this case we will inevitably be off a bit in the guess which we make for $P(t)$ and then incorporate into our $P(g \mid d)$. According to theorem (1), this means we will perform sub-optimally. So there is a correlation between how much we know about $P(t)$ and how assured we are that a Bayesian technique using our guess for $P(t)$ is superior to a particular non-Bayesian technique. This can be viewed as introducing a distinction between an assumption for a probability and one's "confidence" in that assumption.[2] It's conceivable that this is what advocates of Dempster-Schaffer theory, fuzzy logic, and the like are getting at with notions like "plausibility vs. probability".

Another example of what non-Bayesian formalisms might be "getting at" arises if we take $P(t, g, c, d)$ to mean $P(t, g, c, d \mid$ prior information $I)$, so we must define the space of possible I. We *could* say that I fixes the precise statistical problem p that we are considering. As an example, that problem may be predicting the change in the value of the Dow Jones average across some precise date, given the current values of *all* physical variables within the light cone of the space-time coordinate {Wall Street, the date in question}, and all of that information is in I.

However if we ignore quantum mechanical issues for the moment, then physics tells us that for such a "precise problem" the outcome is fixed rather than random, regardless of whether that outcome's already occurred or not. Now as usually defined probability distributions must equal 1 for true events and 0 for false events. (Note that such definitions pay no attention whatsoever to whether we happen to know what's true and false). Accordingly, for a "precise problem" probability distributions are delta functions, and statistics becomes vacuous. (This difficulty is similar to the common complaint of non-Bayesians that Bayesians treat parameters as random variables even though they aren't.)

However, tautologically, we're only interested in that information we have concerning the precise problem p that affects how we would guess for p. Accordingly, one could require that I is only that information we have concerning the problem p such that the g and/or d dependence of $P(g \mid d, I)$ would differ if that information were left out of I. (To agree with common usage, I'm taking d to not be part of the "prior information".) Such a choice of the prior information I fixes $P(g \mid d)$ but not necessarily vice-versa.

Now in practice the extra bits fixing the "precise problem" don't affect how we guess. (E.g., this is true for the bits concerning the vast majority of the physical variables within the light cone of the space-time coordinate {Wall Street, the date in question}). Accordingly, those bits aren't in I. This means that $P(t \mid I)$ is not a delta function, and we don't have the vacuous-statistics problem. (I doesn't even include whether we will actually make a guess, since that information doesn't affect $P(g \mid d)$.)

Under this restriction on I, the posterior "$P(t \mid d, I)$" is a distribution defined for the set of all possible problems with the same guess-affecting information as p. It is not defined solely for the precise problem p. So this restriction suggests a set of multiple problems, just as a frequentist might. In fact, this kind of multiple problem $P(t \mid d, I)$ is exactly the starting point for the conventional frequentist view of statistical physics.

On the other hand, if due to his/her beliefs the guess of statistician A depends on the value of variable Q, whereas that of statistician B does not, they have different I's. (From the frequentist perspective, they are concerned with different sets of problems, in only one of which is the value of Q held constant.) So although it suggests frequentism, the concern of some Bayesians for "beliefs" is also reflected in this definition of "prior information" I.

3. The "degree of belief" interpretation of probability

Some researchers interpret "probability" as synonymous with a subjective "degree of belief" (the precise meaning of this expression — to the degree there is one — isn't relevant for current purposes). Bayesians have often used this interpretation to argue for the superiority of their techniques. The reasoning is that under this interpretation, $P(t)$ is your belief in proposition t, i.e., you automatically know $P(t)$ exactly. Therefore — under this interpretation — if you also know the likelihood you know $P(t \mid d)$, and you can use this to set $P(g \mid d)$ in such a way that you have minimal expected cost (up to calculational approximations). (Sometimes the vague caveat is added to this argument that one's beliefs must be "rational".)

This seems to imply that Bayesian and non-Bayesian analysis are not reconcilable, that Bayesian approaches to statistics are definitionally superior to non-Bayesian ones. However the degree of belief (dob) interpretation justifies non-Bayesian techniques just as readily as Bayesian ones: interpret a non-Bayesian's $P(t \mid d)$ as his/her "degree of belief" in t given d, so (s)he "automatically knows $P(t \mid d)$ exactly", and can use that knowledge to guess with "minimal expected cost", again up to various approximations. In this, the dob interpretation does not play favorites between Bayesian and non-Bayesian approaches.

However there is another more major flaw in this supposed irreconcilability implication: there are foundational problems with the dob interpretation of probability itself. This flaw is the subject of the rest of this section. Fortunately, we don't have to adopt any particular alternative (invariably contentious) interpretation of probability to address it.

The first such foundational problem is that the analysis of the previous section is exactly correct if you're playing a real two-person game. So if in the honored tradition of probability theory one is investigating gambling, then the analysis of the previous section and all of its implications are tautologically correct. In particular, in gambling your "degree of belief" involves $P(g)$ and *a priori* need have nothing to do with $P(t)$, which is instead determined by the other player (the house). Even if you arrive at your beliefs through sophisticated, almost "indisputable" group/information - theoretic arguments, if it turns out that your priors disagree with those of the house, well, then you lose. Your beliefs might be a good *approximation* to $P(t)$, if arrived at rationally and based on extensive prior knowledge (presumably this is the case in those scenarios where Bayesian analysis works well). But that doesn't mean the two quantities are definitionally equal. It doesn't somehow mean that you rather than your opponent fix the probability that your opponent is bluffing.[3]

So the question arises of whether there is a fundamental distinction, with concrete ramifications, between gambling and all real world statistical problems. If there isn't—and it's hard to imagine how there could be—then Thm.'s 1 and 2 imply that there are no guarantees of optimality for the dob Bayesian.

More generally, a "truth" t and a guess g are different objects. Therefore their distributions need not be related *a priori*. (This is reflected in the theorems of the previous section.) Accordingly, a formal statement connecting $P(t \mid d)$ and $P(g \mid d)$ corresponds to an extra assumption concerning $P(t, g, c, d)$, an assumption not demanded by the mathematics. In particular, the dob interpretation is such an extra unjustified assumption.

Another difficulty with the dob interpretation is that if we had sufficient knowledge of the laws of physics (in particular, of the boundary conditions of the universe) and of the (resultant) laws of human psychology, and if we were sufficiently competent to perform the appropriate quantum mechanical calculations, then we might say that we could calculate $P(t)$ exactly. In other words, one possible interpretation is that $P(t)$ is the "real" $P(t)$, determined by the laws of quantum mechanics applied to the universe as a whole. *A priori*, such a $P(t)$ need have nothing to do with one's (pre-calculation) degrees of belief.

Indeed, anyone can *imagine* that quantum mechanics is correct, even if they don't *believe* that to be the case. So we can self-consistently imagine that the universe evolves in accord with equations governing "absolute, objective" probabilities, since those are the building blocks of quantum mechanics. This simple fact that we can self-consistently (!) imagine quantum mechanics shows that there is no formal problem with quantum mechanics' implicit notion of absolute, objective probabilities, which exist independently of any particular person's degree of belief. So there is nothing mathematically necessary about the dob interpretation of probability.

In this regard, note that nothing in Cox's axioms forces a particular interpretation of probability. Those axioms only say that any (reasonable) calculus of uncertainty must obey the laws of probability theory. They do not tell us how to assign the probability values in the first place. One *could* interpret probability as degree of belief. In such a case, Bayesian analysis becomes a set of rules for telling you what structure your beliefs must have to be self-consistent. But the math does not force us to that interpretation.

All of this agrees with Bayesianism as practiced; the actions of a practicing dob Bayesian are indistinguishable from those of someone who thinks $P(t)$ is independent of $P(g \mid d)$, and therefore is not "automatically known" but rather has to be discovered. It's just that for a dob Bayesian, the to-be-discovered $P(t)$ is rather disingenuously considered to be the distribution which "best reflects prior knowledge". To the dob Bayesian, as our understanding of statistics improves, as we get a better understanding of what "uninformative" means, etc., we get a more accurate idea of that $P(t)$. To an outsider, the dob Bayesian is simply changing his/her guess for $P(t)$.

As an example, some of the more prominent attendees at this conference have spent much of their careers looking for arguments to establish what priors to use for certain scenarios. Moreover, they've changed their views on this several times. Each time they act as though they were assuming an "incorrect" $P(t)$ before, despite the fact that that old assumption for $P(t)$ properly reflected their old degrees-of-belief. And each time they tend to look askance at any laggards still using the old guess for $P(t)$, despite the fact that said laggards are directly following along with their beliefs. This behavior is consistent with the idea that degree-of-belief Bayesians do not, deep down, view probabilities as just degrees of personal belief, but rather view them as possessing some degree of objective reality.

Indeed, for a century Bayesianism was in disrepute, and the current consensus is that is was in disrepute because it was used with "bad choices of priors". Just translate "bad choice of" to "incorrect assumption for", and you have the theorems of the previous section, with their implication that Bayesianism can be sub-optimal.

Alternatively, note that transferring from an "incorrect" $P(t)$ to a better one is really nothing more than the process accompanying the (in)famous "opportunity to learn" which one encounters when one's Bayesian analysis leads to poor results. Or to put it another way, having $P(t \mid d)$ poorly reflected in $P(g \mid d)$ is an opportunity to learn. If you assume these distributions are always "automatically" connected, you're assuming you never have an opportunity to learn. (As an aside, note that a mismatch in the distributions is an "opportunity to learn" whether or not $P(g \mid d)$ is based on Bayesian analysis—Bayesians have no monopoly on the use of the concept "opportunity to learn" as a cover for poor performance of their statistical algorithms.)

Finally, note that there might well be a way to embed the reasonableness/desiderata arguments often used by dob Bayesians to set priors inside a complete mathematical framework (e.g., there might be a framework which maps any (!) I to a unique prior distribution). If we had such a framework, *then* one might claim that such reasonableness arguments are a well-principled way to assign probabilities. Without that framework in hand though, we have no assurance that any particular reasonableness argument assigns the same values to probabilities as that framework would. In particular we have no assurance that there isn't some lurking reasonableness argument which contradicts our current arguments. In short, at present "degrees of belief" set by desiderata arguments do not constitute mathematics. They constitute philosophy.

Endnotes

[1] For the purposes of this paper there is no reason to specify whether the notation "$P(.)$" refers to a probability, a probability density function, or some other similar object.

[2] I'm speaking loosely here, and have not defined "confidence" formally. In particular, I have not defined confidence in a probability with probability of a probability.

[3] Note that assigning a "degree of belief" to a proposition and making an assumption for the probability of that proposition (as one might do in gambling against the house) are very similar things. Both are subjective declarations concerning how reasonable the researcher thinks the proposition is. This might be why people have confused them so easily. There is an important distinction between the two concepts however: in declaring one's degree of belief in a proposition one is tautologically correct, whereas there is no such notion of tautological correctness to making an assumption. It is the claim of this paper that $P(t)$ is something one can assume as opposed to something one can simply declare.

ACKNOWLEDGMENTS: This work was supported in part by NLM grant F37 LM00011, by the Santa Fe Institute, and by TXN Inc.

References

[1] D.H. Wolpert, C.E.M. Strauss, C.E., and D.R. Wolf, "What Bayes has to say about the evidence procedure," These proceedings, 1994.

[2] S. Gull, "Developments in maximum entropy data analysis," in "Maximum-entropy and Bayesian methods," J. Skilling (Ed.). Kluwer Academics publishers, 1989.

[3] D.J.C. MacKay, "Bayesian Interpolation," "A Practical Framework for Backpropagation Networks," *Neural Computation*, Vol. 4, pp: 415 and 448, 1992.

[4] D.J.C. MacKay, "Bayesian non-linear modeling for the energy prediction competition," These proceedings.

[5] D.H. Wolpert, "On the connection between in-sample testing and generalization error," *Complex Systems*, Vol. 6, pp: 47-94, 1992.

[6] D.H. Wolpert, "On overfitting avoidance as bias," SFI TR 93-03-016.

[7] D.H. Wolpert, "The Relationship Between PAC, the Statistical Physics framework, the Bayesian framework, and the VC framework", in *The Mathematics of Generalization*, D.H. Wolpert (Ed.), Addison Wesley, 1994.

[8] L. Breiman, "Stacked regression," *University of California, Berkeley, Dept. of Statistics*, TR-92-367, 1992.

[9] J. Berger, "Statistical Decision Theory and Bayesian Analysis," Springer-Verlag, 1985.