BAYESIAN INFERENCE WITHOUT POINT ESTIMATES

Paul Snow

Hawthorne College
59 Maple Ave #107
Keene, NH 03431

ABSTRACT

It is conventional to apply Bayes' formula only to point estimates of the prior probabilities. This convention is unnecessarily restrictive. The analyst may prefer to estimate that the priors belong to some set of probability vectors. Set estimates allow the non-paradoxical expression of ignorance and support rigorous inference on such everyday assertions as "one event is more likely than another" or that an event "usually" occurs. Bayes' formula can revise set estimates, often at little computational cost beyond that needed for point priors. Set estimates can also inform statistical decisions, although disagreement exists about what decision methods are best.

I INTRODUCTION

Probabilistic information often comes in forms other than point estimates. "It is more likely to rain to day than not" is an intelligible statement about a probability even though it gives no specific value for the chance of rain. The statement is also useful as it stands; it helps us decide what to wear outdoors.

A point estimate, e.g. "The chance of rain today is seventy percent", might be more useful. If our weather source doesn't know the precise probability, however, then we'd surely rather have the "more than fifty percent" estimate than nothing at all. We might even be grateful that our source did not pretend to have more precise information than was actually warranted.

Such modesty wins no applause from conventional Bayesians, especially those who work in the tradition of Savage. From their vantage, every statement about probabilities ought to assert point estimates for the events of interest.

Researchers in artificial intelligence who use Bayesian inference have largely adopted the point estimate restriction as given. Further, it appears that some researchers reject probability methods in favor of non-additive belief measures partly because they attribute certain shortcomings of point estimates to probability estimates in general.

Freed of the restriction to points, probability estimates can be as expressive as any fuzzy possibility. The liberalization of probability comes at what is often a modest cost in computational effort, and at no cost at all in statistical rigor. Bayes' formula still works, the intuitively meaningful "relative frequency" interpretation of probabilities still holds, and non-point estimates retain considerable power to guide decisions under uncertainty.

II QUALITATIVE ASSERTIONS

One obvious difficulty with point estimates is that the analyst simply may not know the probabilities of the interesting events with much precision. Zadeh (1985) cites the commonness of such imprecise probability knowledge as the key factor motivating a "fuzzy probability". If the analyst is not restricted to point estimates, however, imprecision poses little problem for (crisp) statistical inference.

A dramatic instance of imprecise knowledge occurs when the analyst is totally ignorant of the event probabilities. The conventional, point-bound, representation of utter ignorance is to assign equal probabilities to each possible "state of the world". It is well-known that it's difficult to express ignorance consistently by this method when there are three or more mutually exclusive states.

For definiteness, suppose there are three such states. Each state is assigned a probability of one third. The disjunctive probability of any two states (the sum of the two states' probabilities, or 2/3) is strictly greater than the probability of the third state (i.e. 1/3). If the analyst is truly ignorant, how does one know that any state is less likely than the disjuction of the other two?

Such problems have led some workers to embrace cardinal measures of belief that are point-valued, but not additive (Shackle, 1949; Prade, 1985). Another answer is to allow the analyst to say that the vector of correct state probabilities belongs to some set. For ignorance, that is the "vacuous set", the set of all probability vectors with the right number of states.

In the general case, where the analyst's knowledge is imprecise, but not so completely imprecise as ignorance, the analyst might choose any set that is thought to contain the correct vector. We do not assume that the analyst has an opinion about which member of the set is the right one, only that the correct vector is not to be found outside the chosen set.

If the analyst's imprecise knowledge happens

to involve linear equalities or inequalities among the state probabilities, then the resulting estimate set has a simple and convenient geometry. The linear relations define hyperplanes in the probability vector space, and the estimate set is the intersection of half-spaces bounded by these hyperplanes. The resulting figure is a polytope: a convex set with a finite number of vertices located where the hyperplanes intersect. To construct the estimate set, the analyst simply enumerates the vertices.

For instance, the analyst may know minimum values for the various state probabilities (at least some of the minima being positive in the non-trivial case). If there are n states, then the analyst's knowledge can be expressed as the n inequalities $P1 \geq L1$, $P2 \geq L2$, ... , $Pn \geq Ln$. If T is one minus the sum of the minima ($T > 0$), then it is simple to show that the n vertices are:

(L1+T, L2, ... , Ln), (L1, L2+T, ... , Ln),
..., (L1, ..., Ln+T)

Another common kind of estimate is an ordering of state probabilities, that is, the n-fold linear inequality $P1 \geq P2 \geq ... \geq Pn$. The set representing this assertion also has n vertices, which are

(1/n, 1/n, ... , 1/n),
(1/(n-1), ... , 1/(n-1), 0), ...,
(1, 0, ... , 0)

Not all simple probability statements that assert linear relationships among the probabilities give rise to a small (i.e., comparable to the number of states) number of vertices. The number of vertices needed to represent probability maxima is subject to combinatorial explosion in bad cases. E.g., if there are n states and each probability is no more than 2/n, then each vertex has n/2 elements equal to 2/n and n/2 elements equal to zero. There are C(n, n/2) such linearly independent vectors.

The information possessed by the analyst might vary from state to state; perhaps a point estimate for one, a range for another, an ordering among others and a bound on the disjunction of still others (that is, a bound on the sum of probabilities, also a linear inequality). The basic procedure of defining the estimate set by enumerating the vertices is the same (and one hopes the number of maxima is small, or the maxima are well-behaved).

Although linear relationships are "special cases" of the possible probability knowledge, it is remarkable how easily they mesh with many common qualitative descriptions of the state probabilities. Nilsson (1986) discusses the construction of polytopes from linear relations that arise from certain formal logical statements about probabilities.

Natural language, too, seems rich in linearities. For example, "S1 is the typical outcome" suggests an ordering in which $P1 > Pj$ for all j. Words like "often", "usually" or "almost always" suggest minima; "rarely" and "almost never" connote maxima. The breakpoints for such representations may be arbitrary (does "almost always" mean $P > .8$? $P > .9$?), but not obviously more so than the estimates of membership grades used with fuzzy set methods. Freed of the point restriction, probability estimates are evidently more useful in the face of imprecise qualitative statistical descriptions than some workers have believed.

III BAYESIAN INFERENCE WITH SET PRIORS

Suppose the analyst has chosen a set representation for the probability information available before observing any evidence. It would be helpful if there were some way to revise the estimate later, when some evidence has been observed.

If the analyst knows the conditional probability of seeing the evidence given each of the possible states, then the analyst can apply Bayes' formula point-by-point to the prior set, making a posterior set in the process. If the correct prior belongs to the original estimate set, then clearly the correct posterior vector is in the revised set.

That much is self-evident. Point-by-point Bayesian revision works, but it is apt to be prohibitively cumbersome for large prior sets. We can lower the computational burden quite a bit if the estimate set has a congenial geometry for revision. In the discussion to follow, we assume that the conditionals are available to us as point estimates. We could allow the conditionals to be set estimates, but that would obscure the present argument and add unilluminating complication.

As luck would have it, our old friend the polytope, the hero of the last section, has a congenial geometry for Bayesian revision. It turns out that if the prior set is a polytope, then the posterior set will also be a polytope. The vertices of the posterior polytope are the Bayes' formula posterior values of the prior set's vertices. For proof, see Levi (1980).

To apply Bayes' formula to a polytope, therefore, one need only find the Bayes' posteriors of the prior vertices. As long as the number of vertices is small, polytope revision is simple and cheap. Given that the polytope is also an expressive geometry, this is a heartening result.

Polytopes are so gifted that a word of caution is in order. Polytopes are not the only convenient geometry for Bayesian revision, nor are they the only kind of set estimate that can occur in easily imagined circumstances. Levi goes too far when he offers convex sets as the only defensible geometry. A set of discrete points, for example, is not convex. It isn't hard to imagine cases where the analyst knows that the true probability vector is either V or W, and no value "in between". Bayesian revision of this estimate set is quite efficient.

Polytopes are emphasized here because they are versatile and convenient, but they are not obligatory. Restricting the geometry of estimate sets to polytopes would be as artificially confining as the point restriction has been.

IV ZERO-FREE VERTICES AND CONVERGENCE

If the prior set contains only vectors that have no zero components (for polytopes, if the vertices are zero-free), then as conditionally independent evidence accumulates, the posterior set will converge toward a single point. The asymptotic limit vector has probability one in the correct state and zeros elsewhere. This follows from a standard result about the ultimate insensitivity of Bayesian inference to different zero-free priors (see, for example, Jeffrey, 1983). The limiting performance of Bayesian updating for set priors, then, is comparable to that for point priors.

Convergence will generally fail to occur if the estimate set does contain vectors with zero elements. The Bayes' posteriors for such vectors will always contain zeros, in the same components as the priors' zeros. If the vector is a polytope vertex, this will distort the posterior set by "tying down" the vertex even if the evidence comes to overwhelmingly support one of its zero-valued states as true.

The worst case occurs when the analyst expresses prior ignorance as the vacuous set, a polytope whose vertices each have zeros in all components except one. Bayes' inference is fruitless in such a case. No amount of evidence (short of certain revelation of the true state) ever vanquishes initial ignorance. The posterior set remains vacuous.

At first glance, this seems to be troublesome. Realistically, however, total ignorance about the states is rare. We can devise artificial instances readily enough, but in the real world, the analyst usually knows something about the states. Just to name the states typically rules out their having a priori zero probabilities, and so eliminates vectors with zero components. As a practical matter, the analyst is probably willing to assert some miniscule positive floor under each state probability (Jeffrey makes a similar remark about point estimates).

As has already been shown, the willingness to assert positive minima gives rise to a convex set whose vertices are zero-free. However modest the departure from strict prior ignorance, conditional evidence revises the prior set, and asymptotic convergence can occur.

Assertion of small minima also suppresses zeros in less drastic circumstances. The vertices of an exhaustive probability ordering also have zero components, as shown earlier. Even though it is no part of the analysts' intention to say that some state may be impossible, the zeros will resist revision as tenaciously as those that arise from prior ignorance. The solution is for the analyst to assert minima L1, ..., Ln in addition to the ordering. If each of the minima is less than $1/n$, then tedious but simple algebra shows that the vertices for the combined assertion of an ordering and the minima are

$$(1-\sum_{2}^{n}Li,\ L2,\ \ldots,\ Ln),$$
$$((1-\sum_{3}^{n}Li)/2,\ (1-\sum_{3}^{n}Li)/2,\ L3,\ \ldots,\ Ln),$$
$$\vdots$$
$$(1/n,\ 1/n,\ \ldots,\ 1/n)$$

In general, it's a good idea to suppress any zeros that occur in the estimate set, in order to avoid the persistent distortion of posterior estimates that zeros cause. Asserting minima is often the simplest way to do this, and since minima are linear relations, they can usually be combined with other information fairly readily.

V IMPLEMENTATION

The essential AI device for dealing with set estimates characterized by a reasonable number of points is already in place. It is the ordinary Bayesian inference network first proposed for PROSPECTOR by Duda, et al. (1976), and developed further by many others, notably Pearl (1982).

Existing networks have two or more exclusive events' (point) probabilities attached to each node. The links are the conditional probabilities relating the events at higher nodes to those at the lower (evidence) nodes.

By convention and practical necessity, the potential evidence is resolved into groups of exclusive events in such a way that observations from different groups are independent of one another, given the states at higher nodes. The geometry of the prior estimates at the higher nodes appears to raise no new issues for this treatment of the evidence.

The alterations to the network needed to accomodate set estimates are straightforward. Where the higher nodes now contain a single probability vector, in the new scheme they would contain several. The amount of calculation needed to update the network to reflect any given evidence configuration increases linearly in the total number of vectors to be updated.

The extra work can be reduced by the efficient handling of intermediate nodes. These nodes contain neither the events of ultimate interest nor the observed evidence. Rather, their role in the network is to aid in its initial construction and to provide explanations of the network's "reasoning" as the evidence is revealed. These nodes do not contribute to the inference itself, and they can be compiled out of the network before run time, to be replaced by conditional probability links directly connecting evidence and conclusions (Snow, 1985). The explanation function of these nodes can be recovered on demand by attaching them distally to the ultimate event nodes, where they wait inertly until asked a question.

These comments apply only to the sort of Bayes' network that traffics in traditional probability estimates. They do not apply to the "influence networks" recently proposed by Pearl (1985). In these networks, the structure of the intermediate nodes is crucial to the interpretation of the networks' outputs. The spirit behind Pearl's proposal seems

to be the same as what animates this paper: retention of probability as the basis of uncertain inference while avoiding the limitations inherent in point estimates.

In any case, set estimates can be manipulated by essentially the same techniques that have already been widely proposed for point estimates. Provided that the number of points needed to represent the set is small, the additional cost entailed in using sets instead of points can be modest.

VI DECISIONS

There are several methods for using set estimates to inform decisionmaking. The very diversity is a hint, however, that no one technique has universal acceptance.

The simplest method is to select a single point from the estimate set and to base the decision on that single point. Typically, the point selected will be the vector that displays the most entropy or else the centroid of the estimate set. The chosen point is then used in an expected value or expected utility analysis to determine the best act, or what would be the best act if the chosen vector were the right one. This step would usually be followed by a "sensitivity analysis" to find out whether the choice of an act depends a great deal on which probability point is chosen.

If sensitivity analysis reveals that the final decision is pretty much the same regardless of the point chosen, then all is well. If not, then selecting an arbitrary point and acting according to its counsel defeats the purpose of working with set estimates in the first place. The simplicity of the method, however, makes it suitable for "quick and dirty" analysis of choices other than the final act, e.g. deciding which of several possible experiments ought to be performed first.

Other decision approaches involve looking at the expected utility of each act for every vector in the estimate set. By our earlier assumption, the analyst doesn't know which vector is the correct one, and so is ignorant about which of the expected utilities is the real pay-off for each act. The choice among the acts, therefore, can be made using any of the popular rules for decisions under pure uncertainty.

Once again, the computational task is simpler if the estimate set has a congenial geometry. An especially convenient set occurs when the vertices of the convex hull of the estimate set are themselves members of the estimate set. This family includes not only polytopes, but also discrete points and polytopes with all or part of their interiors removed.

Several standard decision rules consider only the utility values at the hull vertices in such cases. The best known decision rule of this kind is the linear programming and expected utility criterion called "mixed strategy maximin".

If the estimate set has this nice geometry, and we adopt a decision rule that considers only the hull vertices, then we need Bayesian updates only for the vertices (the proof is a specialization of the polytope result discussed earlier). If the estimate itself is not a solid polytope, then we lose information about how much of the interior of the posterior set is included in the estimate. This won't affect the final decision, and considerable information about the precision of the estimate is retained.

Although the maximin rule has a following, its acceptance is far from unanimous, as discussed by Luce and Raiffa (1957). Methods for decisionmaking under pure uncertainty remain an open research topic, and with them, methods for choosing an act informed by a set probability estimate.

VII THE SAVAGE AXIOMS

The exact nature of the "best" decision rule is controversial, but it seems likely that whatever rule does emerge will involve some expected utility calculation. The "choose a point" and maximin rules of the last section both do.

Savage (1972) has proposed axioms that support the conventional point estimate restriction, which also appear to tie that restriction to the commonnest motivation for the adoption of expected utility rules. If rationality (in the sense that expected utility rules are rational) demands point estimates, and we apply "rational" utility rules to "irrational" set estimates, then we court logical contradiction. Even if this were not the case, Savage's axioms are closely reasoned, widely discussed and solidly in the Bayesian mainstream. The case for set estimates must include some explanation of why Savage's prescription is to be ignored.

Savage's first axiom, the complete ordering assumption, is the crucial one for the point restriction (as noted by Smith, 1961). Complete ordering holds that the analyst assigns a specific value to each act, even when the analyst doesn't know the state probabilities that govern which outcome the act will yield. So, for example, if the analyst knows that act A offers either \$5, \$10 or \$20 depending on whether S1, S2 or S3 is true, then the analyst is assumed to assign act A a specific dollar value, perhaps \$8. The first axiom asserts only that an amount like \$8 exists, it does not say how the assignment is made (why \$8 and not \$9). In sum, the first axiom restricts the analyst to point estimates of value.

Clearly, this is not the only possible attitude if the analyst hasn't a clue whether S1, S2 or S3 is the true state. The analyst presumably would be willing to make an interval estimate of A's value (between \$5 and \$20 inclusive). Absent further information about the states, however, the analyst might balk at making any stronger, more specific assertion about the value of A.

If the analyst happens to be willing to make point value estimates, then the other Savage axioms allow us to infer point-valued "judgmental proba-

bilities" from the analyst's choices. If the analyst subscribes to all the axioms, then any claim that non-point estimates guide the analyst's choices would result in contradiction.

On the other hand, if the analyst doesn't subscribe to all the axioms (and we have discussed why complete ordering might be denied), then the inference about point estimates is unfounded. No logical difficulty arises, and the axioms are moot.

It is worth noting that Savage resorts to axioms for a reason. A strong restriction (the analyst can make only point probability estimates) is to be justified by its derivation from other, supposedly less restrictive assumptions. In fact, the complete ordering axiom (the analyst can make only point estimates of value) is on its face as strong and as restrictive as the proposition it is called upon to justify.

VIII CONCLUSIONS

Many practical problems, e.g. diagnosis, are fruitfully viewed as probability inference tasks. Here, "probability" means the relative frequency with which some event or condition occurs. Although the probability estimates may reflect the personal opinion of some expert, the goal is typically to match as closely as possible the true relative frequency that prevails in the real world. The loose application of the loaded terms "objective" and "subjective" sometimes obscures this point.

The full exploitation of probability methods has been hindered by the convention that point estimates are the only way to express probability information. Licensing set estimates is not a new idea. Objectivist interval estimation, for instance, has been in the statistician's tool kit for a long time. What may be new is realizing how much well-chosen set representations can overcome the supposed shortcomings of probability estimates. Happily enough, set estimates comport well with common AI techniques, particularly those based on another venerable statistical tool, Bayes' formula.

Using set estimates to inform decisions remains a weak spot. The problem of decision informed by sets is closely related to decisions under ignorance. Progress on set-informed decisions is thus linked to either the invention of new decision rules for ignorance, or the elevation of some existing rule to preeminence. In the meantime, there is no shortage of plausible rules catering to a variety of tastes.

REFERENCES

Duda, R. O., P. E. Hart and N. J. Nilsson, "Subjective Bayesian methods for rule-based inference systems", Proc. Natl. Comp. Conf., 1976, pp. 1075-1082.

Jeffrey, R. C., The Logic of Decision, Chicago: U. of Chicago Press, 1983, chap. 12.

Levi, I., The Enterprise of Knowledge, Cambridge, MA: MIT Press, 1980, chap. 9.

Luce, R. D. and H. Raiffa, Games and Decisions, New York: Wiley, 1957.

Nilsson, N. J., "Probabilistic logic", Artif. Intell. 28 (1986, forthcoming).

Pearl, J., "Reverend Bayes on inference engines: a distributed hierarchical approach", Proc. AAAI Conf. Artif. Intell., 1982, pp. 133-136.

___, "How to do with probabilities what people say you can't", Proc. IEEE Conf. Artif. Intell. Appl., 1985, pp. 6-12.

Prade, H., "A computational approach to approximate and plausible reasoning with applications to expert systems", IEEE Trans. Patt. Anal. & Mach. Intell. 7:3 (1985), pp. 260-283.

Savage, L. J., The Foundations of Statistics, New York: Dover, 1972.

Shackle, G. L. S., Expectation in Economics, Cambridge, UK: Cambridge U. Press, 1949.

Smith, C. A. B., "Consistency in statistical inference and decision", J. Roy. Statist. Soc. B 23:1 (1961), pp. 1-37.

Snow, P., "Tatting inference nets with Bayes' theorem", Proc. IEEE Conf. Artif. Intell. Appl., 1985, pp. 635-640.

Zadeh, L., "Decision analysis and fuzzy mathematics", in M. D. Cohen, et al., "Research needs and the phenomena of decisionmaking and operations", IEEE Trans. Sys. Man & Cyber. 15:6 (1985), pp. 765-767.