Bayesian Analysis (2012) **7**, Number 3, pp. 537–540

Comment on Article by Albert et al.

John Paul Gosling *

Groups of experts are often commissioned by decision makers to help inform policy making in science, commercial and government settings. The typical aim of this is to elicit opinions from across the breadth of a scientific community to help inform decision making. Whether a mathematical, statistical or behavioural technique is used to arrive at some probability distribution that encapsulates an entire group's beliefs, we have to face up to the questions of fairness in the process and of defensibility of any methods employed.

The article by Albert et al. presents another solution to the expert problem (as defined in French 1985) where we have a single decision maker who wants to use multiple expert's opinions to update their own. The proposed method is based on a hierarchical model that allows the ultimate decision maker to group experts and to account for uncertainty in the quality of the experts' judgements. The examples presented by the authors show that the method seems to give viable consensus distributions when compared with other mathematical aggregation techniques. It has been widely accepted in Bayesian circles that this type of modelling approach should be considered the normative approach to pooling expert opinions by an individual decision maker (see Lindley 1985; West 1988; Wiper and Pettit 1996, amongst others). Other mathematical aggregation techniques, as reviewed in Genest and Zidek (1986), seem rather ad hoc in the face of the Bayesian foundations of the present approach.

Confidence in judgements

In the present article, the authors ask the experts to judge their confidence in their probability judgements, which effectively allows them to put uncertainty on their uncertainty judgement. Although probabilities cannot be measured to an arbitrary degree of accuracy, I believe that a probability judgement for an event should encode all of an individual's uncertainty and confidence in making a statement. An expert's specification of c does not just capture their confidence in making such judgement: it is confounded with a reluctance to be pinned to one number. If the aim is to capture the former, then I would argue that there are better ways of judging this (Cooke 1991), or, if the aim is to allow the expert to say they do not know what the outcome should be, they should be encouraged towards specifying a suitably flat probability profile.

On the topic of expert self-weighting, there are many cautionary tales in the elicitation literature about expert over- and under-confidence when self-rating and when experts rate their peers (for instance Cooke 1991; Harvey 1994; O'Hagan et al. 2006). Rather than having the experts do this themselves, I wonder if the decision maker should be making the call about how much credence to give each individual's judgements. Also, in the probability judgement case, it is certainly valid for a decision maker to decide

*School of Mathematics, University of Leeds, Leeds, UK, j.p.gosling@leeds.ac.uk

 DOI:10.1214/12-BA717B

how much an expert's judged probability is going to influence their own judgement.

I think my difficulties stem from my lack of experience in using a variable, c, which lies on the range zero to one, to capture an expert's confidence in each judgement they make. With a probability, I have some faith that experts at least understand what is meant by a value of 0.25 or 0.75. Will they understand the implications of $c = 0.25$ or $c = 0.75$ especially as their judged c value is transformed in some way to be used as error on a link function deep within the hierarchical model? Moreover, will the decision maker understand?

In my experiences of facilitating expert elicitation sessions, the experts have been much more comfortable with ranking their own expertise against their peers (although there are suggestions that they are poor at that too (Burgman et al. 2011)) and are often happy to follow the lead of peers that they feel are more experienced in the topic at hand. Perhaps such ranking results could be used by the decision maker to assign confidence to each individual's judgements.

The decision maker's prior

If we accept the mechanism for specifying an expert's confidence, the suggested prior structure for the decision maker's beliefs about the experts' beliefs seems sensible because it offers a mechanism for judgements to be made about between and within expert group agreement. If the decision maker had a hand in selecting the experts and grouping them (and it should be fair to assume that this has been done in a considered way), then there is a great opportunity to form an informative prior structure here. This is suggested by the authors in **Remark 2**, but, in the examples, noninformative prior structures with many levels of hyperparameters were used. In fact, I believe the "weakly informative" structure of the examples gives far too much prior weight to unrealistic aggregation models. For instance, in the examples, there is a relatively high amount of weight on experts having total agreement both between and within expert groupings.

Feedback

When designing elicitation protocols, it is good practice to give feedback on the elicited judgements to the experts and to offer them chances to revise their judgements (O'Hagan et al. 2006). In this elicitation scheme, this feedback-revision loop can be handled with the individual experts. However, there is still a question of who should judge if the combined distribution for each group of experts is appropriate: presumably, the experts within a group could help with that. And, more importantly, who should judge the combined effect of the likelihood on the decision maker's prior? Is it possible for the decision maker to judge whether one likelihood structure is more appropriate than another (perhaps, this is more difficult in the case where the decision maker is expected to know next to nothing)? This might not be considered a problem: in more routine Bayesian analyses, how often is it checked that the updating of beliefs through the chosen likelihood is actually appropriate for the individuals and problem that the analysis concerns?

As expert elicitation is often considered a soft science, even greater efforts should be made to defend modelling choices and check that the updating strategy is appropriate. For the present model, this could be done through the use of hypothetical judgements from several experts and assessment of the decision maker's resulting posterior once the machinery of the model has done the updating.

It could also be argued that following a nonparametric approach to distribution fitting that captures the decision maker's or facilitator's beliefs about the form of the resulting group consensus distribution would be beneficial (see Oakley and O'Hagan 2007, for example). This could allow more flexibility in the distributions that the experts want to represent their own beliefs with and provide estimates of uncertainty around the appropriate form of the combined distribution (although it would not be trivial to bring such computations into the present hierarchical model).

Concluding comments

When faced with the expert problem, it is useful to have a defensible mechanism like the one specified in this article. If the group of experts are also charged with making the decision (the group decision problem), I think we need something more than what Bayesian modelling and formal elicitation methods can offer. If a group of experts' judgements are being collected for another purpose (French's text book problem for instance), I would encourage the facilitators of the elicitation exercise to report each individual's judgements and allow future users of this data to make their own judgements about the strength of this evidence.

It is worth stressing that expert elicitation and subsequent pooling is not a precise science and anyone wanting to use mathematics and statistics to aid these processes should pay great attention to (1) justifying their modelling choices, (2) interrogating the judgement sets for each individual, and, (3) above all, being transparent about the elicitation process in their reporting of results.

References

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L., and Twardy, C. (2011). "Expert Status and Performance." *PLoS ONE*, 6(7). 538

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press. 537

French, S. (1985). "Group consensus probability distributions: A critical survey." In *Bayesian Statistics 2 (eds. J. M. Bernardo et al.)*, 183–201. North-Holland, Amsterdam: Valencia University Press. 537

Genest, C. and Zidek, J. V. (1986). "Combining Probability Distributions: A Critique and an Annotated Bibliography." *Statistical Science*, 1: 114–135. 537

Harvey, N. (1994). *Subjective Probability*, chapter Relations between confidence and skilled performance. New York: Wiley. 537

Lindley, D. V. (1985). "Reconciliation of discrete probability distributions." In *Bayesian Statistics 2 (eds. J. M. Bernardo et al.)*, 375–390. North-Holland, Amsterdam: Valencia University Press. 537

Oakley, J. E. and O'Hagan, A. (2007). "Uncertainty in prior elicitations: a nonparametric approach." *Biometrika*, 94(2): 427–441. 539

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. E., Garthwaite, P. H., Jenkinson, D., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: eliciting expert probabilities*. Chichester: Wiley. 537, 538

West, M. (1988). "Modelling expert opinion." In *Bayesian Statistics 3 (eds. J. M. Bernardo et al.)*, 493–508. New York: Oxford University Press. 537

Wiper, M. P. and Pettit, L. I. (1996). "On improving a model for combining experts' forecasts." In *Bayesian Statistics 5 (eds. J. M. Bernardo et al.)*, 809–813. New York: Oxford University Press. 537