

Comments

Simon French

Manchester Business School

Manchester, M15 6PB

simon.french@mbs.ac.uk

Expert Judgement Combination using Moment Methods

Bram Wisse, Tim Bedford, John Quigley

Comments

I enjoyed this paper, though disagreeing with its theory. For years I have tried to understand the Bayesian linear methodology; intrinsically some of the ideas seem so sensible. Ultimately, however, I have failed because for me judgements of uncertainty seem more primitive than judgements of prevision. The approach stemming from Savage (1972), DeGroot (1970) and others seems more natural to me than that stemming from De Finetti (1974; 1975), Lad (1996), Goldstein (1981) and others. Thus I am much more comfortable with expert judgement approaches based around probability than about judgements of moments. Moreover, as I have argued at various times (French 1985; French and Rios Insua 2000), I see the expert judgements as *data* rather than *probabilities* and thus tended to be against approaches such as the linear opinion pool which begin by treating their judgements as probabilities and then looking to axioms that draw them into a valid synthesis, where I use valid in a weak sense to capture whatever motivation lies behind a set of axioms.

I guess it is sections 3 and 4 of this paper that gives me the most problems. In French (1985) I listed a range of arguments for and against many of these properties (see also Genest and Zidek 1986). Those arguments translate pretty much to the context here. In particular, I am far from convinced by arguments in favour of *marginalisation* and hence, for instance, do not find linear opinion pools more persuasive than, say, logarithmic ones (McConway 1981). Moreover, I wonder if the *zero preservation* property is even less persuasive in the context of expectations. A statement of zero probability is a statement of impossibility, albeit often a dangerously overconfident one that flies in the face of Cromwell's principle (Lichtenstein *et al* 1982; Lindley 1982; McClelland and Bolger 1994); a statement of zero expectation has no more significance than a statement of an expectation of, say, 2.564. However, it is the *independence preservation* property that concerns me most in a paper such as this. I and others have argued against it in general terms (French 1987; Genest and Wagner 1987). Here, however, it and its sibling, the *zero correlation preservation* property, seem dreadfully out of place. Were they to hold, then any chance of learning from a calibration set would be denied: they prohibit the idea of performance based weighting. One can only learn when there are dependencies and correlations; in performance based weightings one learns about the relative calibration and informativeness of the experts from their assessments in relation to the seed items. This all means that I see section 3 and 4 as a mathematical exploration of the implications

of a series of principles that seem to me to be irrelevant to the context; although, of course, I am glad to see in Example 4 that *zero correlation* is not preserved.

I would have preferred a more pragmatic paper: one in which one jumped from section 2 straight to section 5 with a little intervening hand-waving. Their suggested approach would for me be better justified as an exploration of a particular form of synthesis in terms of its performance against a pragmatic choice of loss function. That it performs similarly to Cooke's classical method renders it worthy of further investigation.

Returning to my initial discomfort with the Bayesian linear methodology, one question does occur to me. I am more comfortable with uncertainty judgements that lead to probability elicitation; others are more comfortable with moment judgements. What happens if one has a mix of experts some of whom are more comfortable with the former, others with the latter? How do we combine both?

TU Delft Expert Judgment Data Base

Roger M. Cooke, and Louis L.H.J. Goossens

Comments

I shall restrict my comments on this paper essentially to: *thank you!* Over the past 20 years or so, the group at Delft and their colleagues have *practised* the combination of expert judgement, while others of us have discussed esoteric theory. They have developed a methodology and applied it consistently to a wide range of risk analyses. Their consistent, careful practice of their approach means that now we have a substantial database on which discussions of the use of expert judgement in risk analysis can be empirically founded. Our profession owes them a huge debt of gratitude.

A Study of Expert Overconfidence

Shi-Woei Lin and Vicki M. Bier

Comments

I found this study fascinating. We all know that experts and non-expert probability assessors are generally poorly calibrated and usually overconfident; yet the starkness of the results shown by Figures 1 to 4 in Lin and Bier's paper still caught me by surprise. Poor calibration and over confidence is so endemic that we really should pause and ask ourselves why we dare use any expert judgement in risk analysis. We should remember that in many of these studies the experts went through prior training to help them overcome biases in the probabilistic encoding of their judgements. This was particularly the case in the USNRC/CEC studies relating to accident consequence modelling (Goossens and Kelly 2000). I guess, therefore, that we should take some comfort from Figure 1: overconfidence in these studies tends to be less than in others, but it is still present and significantly so.

It is interesting that Cooke's classical method of combining expert judgements seemingly achieves such good results without any attempt to adjust the input judgements to improve their calibration. The method's robustness and ability to overcome poor calibration relies entirely on its draconian approach to giving the more suspect experts zero weight. In French and Wiper (1995) we explored a Bayesian approach which attempted to calibrate

the input judgements before combining them. At a far, far greater computational cost it achieved effectively the same performance as the much simpler classical method. Cooke's original insight that lay behind the mathematics of group scoring rules and significance tests was considerable.

But to return to the import of this paper, I would make a comment and raise one question of significance that our community should address before selling our wares to be used in risk analyses.

The comment: Lin and Bier follow others in concluding that mathematical combination of expert judgement yields better results than behavioural aggregation. In many ways I concur in areas such as risk analysis and then with a proviso. In many decisions, while there are risks, they are not life threatening and, even in terms of monetary or other losses, they do not risk the stability of the organisation itself. In such areas, methodologies such as decision conferencing grew up with the intention of developing a shared understanding of the issues and a commitment to the chosen action (French 1988; Phillips 1984). The commitment is important because poor outcomes from a decision can arise as much from poor implementation as from any poorness of the chosen strategy itself. Thus behavioural approaches to working with groups tend to encourage convergence of views and buy in to the conclusion, building commitment and better implementation. In risk analysis and management this may be very dangerous. Good risk management monitors potential threats and has management strategies ready should they materialise. One needs to maintain an overview of all opinions and risk assessments, indeed to maintain disagreements. Thus in a risk analysis, while one might use Cooke's method to develop the probabilities used in the calculations, one should never discard all the zero weighted experts' opinions in the risk management process: they point to issues that should be monitored. I make these points not to disagree with Lin and Bier: in fact, they are tangential to their study. Rather I am anxious to emphasise that one should always look at the context of a study and then design the analysis and the management processes around it to fit with that context. Rules of thumb that mathematical aggregation is generally better than behavioural may blind us to that need.

The question: if the public see results such as those in Figures 1 to 4, would they trust any analysis that relies on them? And if not, are we professionally responsible in continuing to use such expert judgements without debating the issue with the public? I ask this in the context of all the knowledge we have of public trust and engagement. The impact of many events is magnified if public trust and confidence is lost. Many psychological and social studies have shown that these can be lost if there is some doubts about earlier decisions, leading to questions of blame and poor management (see, e.g., Bennett and Calman 1999; Fischhoff 1995). If we build risk analyses on expert judgements as poorly calibrated as these and if, Heaven forbid, some low probability disaster follows, will we be perceived as having conducted analyses poorly and lose public confidence with all the consequences that may be entailed?

On the Performance of Social Network and Likelihood Based Expert Weighting Schemes

Roger M. Cooke, Susie, El Saadany, Xinzheng Huang

Comments

I read this paper with a great deal of interest. As an unreconstructed Bayesian working on the theory of expert judgement in the early 1980's, I often expressed the view that Bayesians should be prepared to assess quantities that corresponded to the relative expertise of those consulted. This suggestion that social networks might offer a method of assessing relative expertise *a priori* could provide a way forward to developing a Bayesian decision maker's likelihood function for the experts' statements. However, if social networks are to be used to assess relative expertise, be it via a Bayesian approach or a more classical one as here, then I suspect a more subtle way of articulating them is needed. Citation scores between members of the expert panel is an obvious, but somewhat simplistic way forward. The following points would, it seems to me, need to be addressed before one can claim to have investigated citations as a mechanism for estimating relative expertise.

- What about how often they are cited by the community outside the panel. Does that not say something about relative expertise too?
- Some papers are cited a lot because the authors referring to them *disagree* with them. To take one example, the authors of the original paper which postulated a link between autism and the MMR triple vaccine were heavily cited by a medical profession who greatly doubted their views.
- The number of citations one gets may well be correlated with age rather than expertise. One is noticed the more that one has been around. In many cases, one might claim that wisdom and experience come with age; but if one is dealing with emergent technologies and risks, is this so? Kuhn (1970) warns us that new paradigms tend to emerge among the younger scientists and must wait ascendancy not until those who subscribe to the old orthodoxy change their thinking, but until they literally die out.

A final thought. Can we combine social network ideas and calibration scores to provide a better overall assessment? A Bayesian approach could, in principle, do this, but articulating appropriate prior and likelihood functions would be a major stumbling block. I wonder if a more *ad hoc* approach is possible. Would it be possible to 'bias' the calibration/relative information quotients in the classical model in some way so that experts with high social network scores would be less likely to given zero weight on comparison with the α level?

References

- P.G. Bennett and K.C. Calman, Eds. (1999). *Risk Communication and Public Health: Policy Science and Participation*. Oxford, Oxford University Press.
- B. De Finetti (1974). *Theory of Probability*. Chichester, John Wiley and Sons.
- B. De Finetti (1975). *Theory of Probability*. Chichester, John Wiley and Sons.
- M.H. DeGroot (1970). *Optimal Statistical Decisions*. New York, McGraw-Hill.

- B. Fischhoff (1995). 'Risk perception and communication unplugged: twenty years of process.' *Risk Analysis* **15** 137-145.
- S. French (1985). Group consensus probability distributions: a critical survey. *Bayesian Statistics 2*. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Eds, North-Holland 183-201.
- S. French (1987). Conflict of belief: when advisers disagree. *Analysing Conflict and its Resolution: Some Mathematical Contributions*. P.G. Bennett, Ed. Oxford, Oxford University Press 93-111.
- S. French, Ed. (1988). *Readings in Decision Analysis*. London, Chapman and Hall.
- S. French and D. Rios Insua (2000). *Statistical Decision Theory*. London, Arnold.
- C. Genest and C.G. Wagner (1987). 'Further evidence against independence preservation in expert judgement synthesis.' *Aequationes Mathematicae* **32** 74-86.
- C. Genest and J.V. Zidek (1986). 'Combining probability distributions: a critique and annotated bibliography.' *Statistical Science* **1** 114-148.
- M. Goldstein (1981). 'Revising previsions: a geometrical interpretation (with discussion) ' *Journal of the Royal Statistical Society* **B43** 105-130.
- L.H.J. Goossens and G.N. Kelly (2000). 'Special: Issue: Expert Judgement and Accident Consequence Uncertainty Analysis.' *Radiation Protection Dosimetry* **90**(3) 293-381.
- T.S. Kuhn (1970). *The Structure of Scientific Revolutions*. Chicago, The University of Chicago Press.
- F. Lad (1996). *Operational Subjective Statistical Methods*. New York, John Wiley and Sons.
- S. Lichtenstein, B. Fischhoff and L.D. Phillips (1982). Calibration of probabilities: the state of the art to 1980. *Judgement under Uncertainty*. D. Kahneman, P. Slovic and A. Tversky, Eds. Cambridge, Cambridge University Press 306-334.
- D.V. Lindley (1982). The Bayesian approach to statistics. *Some Recent Advances in Statistics*. J.T. Oliveira and B. Epstein, Eds. New York, Academic Press 65-87.
- A.G.R. McClelland and F. Bolger (1994). The calibration of subjective probabilities: theories and models 1940 - 94. *Subjective Probability*. G. Wright and P. Ayton, Eds. Chichester, John Wiley and Sons.
- K. McConway (1981). 'Marginalisation and linear opinion pools.' *Journal of the American Statistical Association* **76** 410-414.
- L.D. Phillips (1984). 'A theory of requisite decision models.' *Acta Psychologica* **56** 29-48.
- L.J. Savage (1972). *The Foundations of Statistics*. New York, Dover.
- M.W. Wiper and S. French (1995). 'Combining experts' opinions using a normal-Wishart model.' *Journal of Forecasting* **14** 25-34.