

Comments on articles in RESS special issue

Tony O'Hagan

December 25, 2006

Lin & Bier, 'A Study of Expert Overconfidence'

The authors are to be congratulated on their attempt to extract results from the unique Delft database. Their conclusions are interesting and thought-provoking. However, I am concerned about the analysis of individual question effects.

As in so many aspects of elicitation, I think it is helpful to disentangle the uncertainties due to lack of knowledge, i.e. epistemic uncertainty, from those due to intrinsic randomness, i.e. aleatory uncertainty. Consider the questions in the Dike Ring study that are listed in Table 2. All of these questions either implicitly or explicitly ask for opinions about individual instances which are subject to aleatory uncertainty. For instance, question Hs asks about wave height for "a randomly chosen occurrence" (of some phenomenon). The randomness here and in question Ts is explicit, but it is implicit in the other questions. Of course, the expert will also have epistemic uncertainty concerning properties of the population of random instances, such as the mean.

In general, I believe it is good practice in elicitation to ask separately about sources of epistemic and aleatory uncertainty, and before continuing to discuss this paper I think it is worthwhile elaborating a little on my view because it is not that of those whose elicitations are found in the Delft database. Theirs is the view that we should only ever ask experts about observable random variables. According to that view, the mean ratio of wave heights (in the population of all occurrences of the phenomenon under study) is not observable and so experts' opinions about it should not be elicited. In contrast, it is my experience that it is perfectly possible to conduct meaningful elicitation about such quantities, and that to do so has the important benefit of separately eliciting epistemic uncertainty. Amongst other advantages, this helps to counter over-confidence (which is major concern of the present paper), and by modelling dependencies appropriately it avoids having to elicit them.

Returning to the Dike Ring study, the authors find better calibration on questions Ts and Hs than on the others, and I suggest that this may be at least in part due to the fact that the aleatory uncertainty is more explicit in the wording of these questions. Respondents to the other questions may have failed to acknowledge the randomness fully, concentrating instead on uncertainty about the mean. This in itself would explain the higher degree of over-confidence found for those questions.

The presence of this aleatory uncertainty also affords a simple way to express my doubts about the authors' modelling of individual question effects. Imagine a question for which aleatory uncertainty dominates. Then the quantity in question will be uncertain because it is a random instance, but there is negligible epistemic uncertainty about its underlying distribution. Consider, for instance, question Mo1 in the Dike Ring study, which asks about "the actual flow rate on occasions when the calculated flow rate is 1 litre per second per meter," and suppose that every expert knows more or less exactly the distribution of flow rate conditional on the calculated rate being 1 l/s/m. There is only the aleatory uncertainty described by that distribution. Then all the experts would give essentially the same percentiles in answer to this question. Now if we have data comprising a sample of observed values of this quantity, what should we expect to see in terms of expert calibration? For 90% of those sample values, the observed value will lie in the expert's 90% intervals and we will observe 100% success. On the other 10% of sample occasions, we will observe 0% success.

It is just about possible to accommodate this phenomenon of only getting 100% or 0% success within the authors' model by making the realization effect variance $\sigma_{realization}^2$ infinite. However, we also have to have an infinite bias term somewhere (in order to get 100% success on 90% of occasions). Although this situation of purely aleatory uncertainty is an extreme, all questions have the potential to exhibit this kind of behaviour. Whenever the experts are in close agreement we can expect their proportion of successes to be close to 0% or 100%. This may be manifested in the model as a high variance for the realization effect or for the question effect (in the case where the uncertainty is purely aleatory and it would not be sensible to think of alternative realizations), but it would be wrong to interpret this (as the authors do) as implying great variability in the difficulty of questions. It is a natural phenomenon.

For this and other reasons, I find the idea of modelling success or failure on individual questions unconvincing. The results in this paper are intriguing, but I think they cast little light on the appropriateness of using experts' performance on seed variables as indicative of their performance on the substantive variables.

Mazzuchi, Linzey & Brunin, 'A Paired Comparison Experiment for Gathering Expert Judgment for An Aircraft Wiring Risk Assessment'

Paired comparison methods have a long history in the field of statistics. However, it is not a topic on which I feel myself to be knowledgeable, so the following remarks may be viewed as my attempt to understand the authors' application of these methods to expert judgement.

The underlying idea of paired comparisons is that we observe which is the winner in each of a number of contests between pairs of players from some collection, and wish to rank the players. The statistical models represent the probability that player i beats player j as a monotone function of the difference

$\lambda_i - \lambda_j$ between the average abilities of the players. This can be seen as induced by the variability in the performance of the players in any given contest around their average abilities. If there were no such variation, then player i would always beat j if $\lambda_i > \lambda_j$, whereas if the variation is large then the chance of the player with the higher average ability winning a contest may be little more than 0.5.

A nice feature of the models is that we can ignore the magnitude of the performance variability by scaling the λ_i s. That is, we can assume a fixed variability. If the real variability is low, then a clear ranking will emerge from the data and this will show as widely separated estimates of the λ_i s. Higher variability will induce λ_i estimates that are closer together. So from such an analysis we can read nothing into the magnitudes of the λ_i s; their scale is arbitrary. Also, since the model only depends on the differences in λ_i s their location is also arbitrary. If we only wish to rank the players, of course, these kinds of arbitrariness do not matter.

When the authors apply paired comparison models to judgements of relative failure risks in aircraft wiring, a couple of complications arise. First, note that there are no real players with variable performance in contests. The contest is in the expert's mind, and the analogue of performance variability is the imperfection of the expert's judgement/knowledge. In this analysis, however, we have several experts. In pooling the data from all the experts, there is an implicit assumption that the arbitrary scaling of the λ_i s is the same for all experts, or equivalently that they are all equally good at making the comparisons. That seems a highly implausible situation. Indeed, the authors exclude some experts because the quality of their judgements is not good. Effectively, as I read it, they are being rejected because their 'variabilities' are too high (leading to too many 'circular triads').

The second complication arises because the authors go on to use the λ_i estimates in a regression analysis to estimate relative failure rates. They say that the final answer needs to be scaled by a set of actual failure data on one condition, which in view of the log-linear modelling is equivalent to fixing the arbitrary location of the λ_i s. However, I see no way that the arbitrary scaling is also resolved. I wonder if the authors have implicitly assumed that the experts compare different conditions by notionally observing a pair of failure times. This would fix the magnitude of judgement error and so resolve the scaling issue. My reading of their paper suggests (although it is not really clear to me) that this may be what they have done. If so, this assumption seems particularly hard to defend.

Let me end by reiterating that I am no expert in this area of statistics, so I may have misunderstood some key points. In any case, I commend the authors on an ambitious piece of work tackling an important problem.

Morales, Kurowicka & Roelen, ‘Eliciting Conditional and Unconditional ...’

It is nice to see the continuing development of methods for eliciting complex multivariate distributions based on Bayesian networks and vines. This paper makes an important contribution to that literature.

Eliciting any multi-dimensional distribution is a very challenging task, and so this paper achieves as much as it does by making some very sweeping simplifying assumptions. The basic approach has two key features. One is to decompose the joint distribution into a set of bivariate elicitation. These are typically distributions for a pair of variables conditional on one or more other variables. Although the authors happily elicit distributions conditional on up to five other variables without comment, these are cognitively extremely difficult tasks for the expert. I have serious doubts about whether the elicited values represent genuine beliefs in any meaningful sense. Even an extremely experienced applied statistician would struggle to think about how variable A helps to predict Y after allowing for the explanatory power of variables B, C, D, E and F.

The second feature is that each bivariate (conditional) distribution is elicited by eliciting marginal distributions and a single extra value. This extra value is a conditional quadrant probability – the probability that Y is above its median given that A, B, C, D, E and F are all above their medians. The distribution is then completed by assuming a convenient copula. (The process is presented as first using the conditional quadrant probability to determine a rank correlation coefficient and the copula being fitted to this value, but the intermediate step is not really necessary.) The difficulty of this elicitation task is mentioned above, but its parsimony is also deserving of mention. I am always wary of elicitation methods that elicit only just enough quantities from the expert to determine the parameters of some arbitrarily chosen distributional family. It is surely not much to ask that the suitability of this assumed family might be checked by eliciting at least one more quantity, or by feeding back to the expert some of the conclusions derived from these assumptions.

If no such checking is done, then I would expect to see analysis of the robustness of conclusions to the assumed copula form. In particular, I am prepared to believe that the elicited conditional quadrant probability might give similar rank correlation values in different copulas, and yet the complex conclusions in Section 5 could conceivably be quite sensitive to that choice.

Wisse, Bedford & Quigley, ‘Expert Judgement Combination using Moment Methods’

This is technically an impressive paper, but I wonder about the wisdom of basing an elicitation approach on moments. Psychological research indicates generally that people do not evaluate moments well, particularly variances. Covariances will surely be even less well elicited in practice. Although I have a lot of

respect for the Bayes linear approach, and have used it myself on occasions, this has always been a concern. When I used it in a substantial practical application (O’Hagan, 1998), I did not actually elicit moments, but derived them from elicited credible intervals based on assumptions of underlying normal or lognormal distributions. Covariances were deduced from elicited intervals for averages of variables. The authors do something similar, deriving moments from elicited quantiles by assuming an underlying distribution from the extended Pearson-Tukey family. They say, however, that they do this because the dataset did not include elicited moments. The implication is that they would rather elicit moments directly, but this seems dangerous to me.

However, if moments are to be derived from quantiles, then this raises another invariance property that one would wish to have, that the moments derived from combining experts’ quantiles should equal those obtained by combining the derived moments. Does this hold? If so, then is there any point in working with moments instead of probabilities? If not, shouldn’t we be concerned?

I should note that although the authors cite de Finetti, Goldstein and Lad in presenting the Bayes linear approach, all of these people hold (or held, in the case of de Finetti) quite fundamentalist subjective Bayesian views. I doubt if any of them would approve of the mechanistic combination of different people’s beliefs!

The authors’ reference to higher moments at the end of the second paragraph of their introduction prompts me to add some further comments. First, I am not sure what they mean by approximating full probabilistic modelling more closely by the use of higher moments. There is no extension to the Bayes linear framework that encompasses higher moments than the second order. Perhaps what the authors are referring to is the idea that in addition to the variable X itself, one can include X^2 , X^3 and so on as members of the collection of quantities to which the Bayes linear modelling is applied. This means that we include the expectations, variances and covariances of these variables. However, this leads to more complications regarding coherent updating and combination.

Formally, in the notation of section 2, let the vector \mathbf{X} comprise a variable Z and its square Z^2 . Then

$$E(\mathbf{X}) = \begin{pmatrix} E(Z) \\ E(Z^2) \end{pmatrix}, \quad \text{cov}(\mathbf{X}) = \begin{pmatrix} \text{var}(Z) & \text{cov}(Z, Z^2) \\ \text{cov}(Z, Z^2) & \text{var}(Z^2) \end{pmatrix}.$$

The five distinct elements of these two arrays depend on the first four moments of Z , so there is a redundancy. Specifically, of course, $\text{var}(Z) = E(Z^2) - E(Z)^2$. If we include higher moments of Z in \mathbf{X} , we get increasing numbers of redundancies. These logical dependencies between elements of the mean vector and the variance matrix are nonlinear and it is well known that they are not preserved by Bayes linear updating. They will also not be preserved by combination of experts using any linear opinion pool.

Another way of getting closer to full probabilistic Bayesian methods might be by including indicator variables like $I_A(Z)$ in \mathbf{X} . However, these bring their own problems because the variance of a binary random variable is always determined

by its mean in another nonlinear dependency. Bayes linear methodology should not be seen as applicable in all situations. It should primarily be used for continuous, unbounded variables whose distributions are not too far from joint normality.

Reference

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician* 47, 21–35 (with discussion, pp 55–68).