Bayesian Analysis (2012) **7**, Number 3, pp. 541–546

Rejoinder

Isabelle Albert *, Sophie Donnet †, Chantal Guihenneuc-Jouyaux ‡, Samantha Low-Choy §, Kerrie Mengersen ¶ and Judith Rousseau ‖

We very much appreciated the interesting comments of S. French and J.P. Gosling on our paper and we briefly address their key points. In agreement with the two discussants, we found existing methods for mathematically pooling opinions to be limiting, and hence proposed a Bayesian framework. We thank the discussants for providing this opportunity to address the philosophical issues that give meaning and context to the use of information elicited from experts to construct informative priors. Our rejoinder to the issues raised by discussants can be classified as follows:

1. Modelling elicitation error using a pooled prior. This includes:
   (a) Practical application of this hierarchical model.
   (b) Confidence: encoding expert *confidence* or *credibility*.
   (c) Model validation.
2. Ownership of the prior: Decision-maker, group or society?
   (a) Does this hierarchical prior apply more when the prior is 'owned' by the decision-maker?
   (b) How does it apply when 'owned' by a group or more broadly by society?
   (c) What is the role of a pooled prior: as an end in itself or as information to inform the experts, the decision-maker and their interactions?

We first discuss the points raised by S. French.

1 Rejoinder on S. French's comments

First we thank Simon French for considering our approach as useful. The main concern as far as we understand is on when and how should such an approach be considered.

*INRA, UR1204, Mét@risk, AgroParisTech, Paris, France, isabelle.albert@paris.inra.fr
†Université Paris Dauphine, Paris, France, donnet@ceremade.dauphine.fr
‡EA 4064, Faculté des sciences pharmaceutiques et biologiques, Université Paris Descartes, chantal.guihenneuc@parisdescartes.fr
§School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia and Cooperative Research Centre in National Plant Biosecurity, Canberra, Australia, s.lowchoy@qut.edu.au
¶School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia, k.mengersen@qut.edu.au
‖CREST-ENSAE, 92245 Malakoff and CEREMADE, Université Paris Dauphine, Paris, France, rousseau@ceremade.dauphine.fr

 DOI:10.1214/12-BA717REJ

More specifically S. French finds such a hierarchical modelling approach appropriate for the expert problem but questions its uses in the group decision and the textbook problems. We essentially agree with this concern since the starting point of our work was really the expert problem, i.e. in an inferential problem various experts are questioned to construct a probability distribution on some parameter.

1.1 Modelling elicitation error using a pooled prior: Practical application

Moreover, we agree with S. French that although our approach is conceptually easy to understand it is not necessarily easy to apply in practice. The main reason behind this is that some choices need to be made at various levels of the modelling. First the family of prior distributions ($\pi(\theta|\gamma)$) has to be chosen, then the higher level distributions on the hyperparameters and finally an error model for the elicited data. In the usual pooling approaches, it is merely the family of prior distributions that is chosen, together with the calibration of the weights of each expert. We would like to point out however that the latter case corresponds also to strong modelling assumptions. To illustrate this, consider the way each expert distribution is considered in the pooling case. Typically a parameter γ_e is constructed to fit as best as possible the elicited quantities of expert e, say D_e. If some kind of least squares approach is used to fit γ_e, then it corresponds implicitly to a Gaussian distribution for the error model. Our approach is slightly more tedious because we have tried to explicitly capture the main sources of variation across both the elicitation and combination process.

Nevertheless, we think that including calibration data in the hierarchical modelling would indeed be extremely beneficial. In our examples, we had no historical data on the experts' elicitation skill. However in the case of the availability of calibration data for each or some experts, a natural way to incorporate it in our model is through the construction of the error model. As it stands, the error model is written as :

$$D_{et}|\gamma_e \sim h(d_t(\gamma_e); \sigma_{et})$$

and in our examples we have considered additive models in the form

$$\eta(D_{et}) = \eta(d_t(\gamma_e)) + \epsilon_{et}, \quad \epsilon_{et} \sim \mathcal{N}(0, \sigma_e^2).$$

Calibration data could be used to learn about the variance σ_e^2 of the noise ϵ_{et}, either formally adding a hierarchical structure to incorporate the calibration data or less formally by plugging-in an estimate $\hat{\sigma}_e$ which would be based on these calibration data. In the case where one expert is found to have a systematic bias in the calibration data, this bias could either be introduced in the error model above by adding a nonzero mean to ϵ_{et} or at the conceptual level γ_e. This is mentioned also in our Remark 2. These two possibilities correspond to different natures of bias. In the first case it is believed that the bias is due to the elicitation process whereas in the second case it is believed that the expert e has a shifted distribution compared to the other experts of his group.

1.2 Ownership of the pooled prior: The type of owner

Application of a hierarchical model, such as the one we have proposed in this paper, for the other two contexts, namely the group decision problem and the textbook problem, only makes sense if the existence of an analyst (or a supra-Bayesian) makes sense. Our paper focuses on presenting a new and Bayesian method for pooling expert opinions, in a way that helps combine their mental models. Taking a parametric perspective provides a novel approach with several benefits. However, as noted by discussant French, it is also important to consider what this pooled (or consensus) prior means. As we note in the paper, a subjective view of Bayesian statistics holds that any prior ought to reflect a subjective degree of belief, which implicitly requires that ownership of the prior probabilities must be clear.

From a practical perspective, there is a spectrum from the case where the decision-maker 'owns' the consensus prior to one where the group 'owns' the consensus prior. In an ideal world, expert consultation occurs in several phases, where individual and pooled priors may be provided as feedback to a group of experts, until the group learns from this feedback and reaches the next level of learning, hopefully represented by an improved model. For instance, this may occur in a Delphi-like strategy of group moderation (e.g. Burgman et al. (2011)), or use some other approach for challenging and promoting group discussion, as discussed by French. Such an approach would lead to greater group ownership.

However in practice, resources may be limited so that feedback is perfunctory, so that the decision-maker is assigned more ownership. For example in Ford and Sterman (1998), the goal of sharing and comparing assessments within a group may be to highlight conflict, without necessarily desiring consensus. For instance this could be achieved by inspecting differences with respect to a pooled prior. In any case, our method for producing a pooled prior can be viewed as an alternative to current approaches (such as moment or linear pooling), and so does not necessarily define the endpoint of the Bayesian analysis. Nevertheless, in some fields, where statistical analysis is proscribed before the data arrive (in efforts to preserve objectivity), we can see that the decision-maker approach may be the more easily acceptable approach. Then the hierarchical models can still be useful since they take into account the diversity of the opinions and the posterior distribution of each conceptual parameter γ_e (associated either to an individual or a group) can still be recovered. The same idea applies to the textbook problem. It is true however that more work needs to be done to understand how hierarchical models react to various elicitation problems and what are their limitations.

2 Rejoinder on J.P. Gosling's comments

2.1 Ownership of the pooled prior: The type of owner

We would like to thank J.P. Gosling for his interesting comments and suggestions. These comments tackle some of the details of our modelling approach which in practice turn out to be more than details because they are real difficulties, due to the subjective

nature of prior elicitation. In particular the problem of the interpretation of confidence statements given by each expert is difficult and no real satisfying solution exists to our knowledge.

2.2 Modelling elicitation error using a pooled prior: Assessing confidence

Our paper suggests, in agreement with discussant Gosling, that in Method A, the confidence c_{int} can be encoded from expert estimates of their uncertainty or through other methods such as scoring, if some form of gold standard is available for calibration. In the two case studies illustrated in the paper, no gold standard was available, so expert estimates of uncertainty were used. There is a current ongoing and complex debate about how much expert opinions can be "trusted", which we felt was extremely important, but unfortunately would have substantially changed the main thesis of the paper. However we cited some literature (as did Gosling) should an interested reader wish to pursue this.

The *ideal* situation is when calibration data on each expert is available so that some more objective construction of the error model is possible, as discussed in the previous section. When this is not the case, we believe that every approach, including ours, is only partially satisfying. We are however puzzled by the fact that discussant Gosling does not see $c \in [0, 1]$ as a probability. In our construction when the confidence level c given by the expert lies in $(0, 1)$ (it is often given as a percentage) we interpret it as a probability. The difficulty comes from interpreting the event to which it is related. Our proposition gives one potential answer but we are aware that it is only an interpretation and that the expert may indeed hold a different interpretation. In our experience, we have tried as much as possible to have the expert express what he/she meant by such a confidence level c.

It would be however more satisfactory to have some more concrete implications of his/her choice of c given to the experts during the elicitation process, as suggested by J. P. Gosling. This would indeed lead to greater transparency. As a side remark we do not understand the difference between the lack of confidence in making a statement and the *reluctance to be pinned to one number*. We note that in the two examples, experts were effectively asked to provide intervals for the targeted random variable, as defined by elicited quantiles q. To then elicit lower and upper bounds on these quantiles, akin to tolerance rather than credible intervals, could lead to confusion. Indeed in the way that the PhD case study was undertaken, experts (themselves mathematicians) were encouraged to express their uncertainty in whichever way they felt most comfortable. Only one expert, a probabilist, chose to express their uncertainty in the form of such tolerance intervals. This suggests that it is natural to few mathematicians. More research is required to determine whether such an approach is effective with non-mathematicians and even mathematicians.

2.3 Model validation

The comment on the validation of the various aspects of the elicitation model, be it the likelihood or the different levels of hierarchy, is highly relevant. For the PhD case study, with the more extensive hierarchy, and larger number of experts, the parametric model was assessed using posterior predictive checks, but not reported in this paper. Indeed this was how the need for the offset was confirmed. Some details for how this style of model checking was applied, and illustrated for one expert, are presented elsewhere (Low-Choy (2012)).

The idea of hypothetical data, similar in spirit to predictive posterior checks, seems interesting but we have not tested it. We are a little more skeptical about the use of nonparametric approaches. In Oakley and O'Hagan (2007) flexibility is proposed by considering a *nonparametric* prior on the distribution of θ, but they do not consider any error model on the elicited data. Part of the feasibility of their approach is due to the exact expression of the conditional ditribution of the density of θ given elicited data, which is only possible if the latter can be written as a linear operator of the density. We do believe however that extending our approach to more complex multivariate problems would be interesting.

Finally, we note that assessing the quality of the measurement instrument, in this case the elicitation questions, has been the subject of substantial portions of the expert elicitation literature. Pioneering work by Tversky and Kahneman (1974) focused on the heuristics that people used, which led to undesired interpretations of the probabilities sought. This work has substantially been updated since then, with confirmatory or non-confirmatory evidence now available (see review by Kynn (2008)). In this paper, we rely on aspects of this research about effective preparation, content, wording and ordering of questions. It is cited by referring to a core approach for eliciting quantiles and cumulative probabilities (Oakley and O'Hagan (2007); Low-Choy et al. (2010)). In addition all elicitation ought to invest heavily in preparation (Spetzler and Staël von Holstein (1975); O'Hagan et al. (2006) and Low-Choy et al. (2010), e.g.) which in this paper we assume is undertaken, since it focuses on the modelling of the elicited information.

We hope this provides more detail about how model validity was and can be assessed for this type of model. In fact one benefit of pooling based on an explicit model, such as proposed in our paper, is that it is more conducive to formal model testing compared to the previous methodologies.

References

Burgman, M. A., MacBride, 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): e22998: 1–7. 543

Ford, D. M. and Sterman, J. D. (1998). "Expert Knowledge elicitation to improve formal and mental models." *System Dynamics Review*, 14: 309–340. 543

Low-Choy, S. (2012). "Priors: Silent or Active Partners in Bayesian inference?" In C., A., Mengersen, K., and Pettitt, A. N. (eds.), *Case Studies in Bayesian Statistical Modelling and Analysis*. John Wiley & Sons, Inc, London. 545

Low-Choy, S., Murray, J., James, A., and Mengersen, K. (2010). "Indirect elicitation from ecological experts: from methods and software to habitat modelling and rock-wallabies." In O'Hagan, A. and West, M. (eds.), *Oxford Handbook of Applied Bayesian Analysis*. Oxford University Press, UK. 545

Oakley, J. E. and O'Hagan, A. (2007). "Uncertainty in prior elicitations: a nonparametric approach." *Biometrika*, 94(2): 427–441. 545

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley. 545

Spetzler, C. S. and Staël von Holstein, C.-A. S. (1975). "Probability encoding in decision analysis." *Management Science*, 22(3): 340–358. 545

Tversky, A. and Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases." *Science*, 185(4157): 1124–1131. 545