Combining expert opinions in prior elicitation

Isabelle I. Albert, Sophie S. Donnet, Chantal C. Guihenneuc-Jouyaux, Samantha S. Low-Choy, Kerrie K. Mengersen, Judith J. Rousseau

Bayesian Analysis (2012) **7**, Number 3, pp. 503–532

Combining Expert Opinions in Prior Elicitation

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

Abstract. We consider the problem of combining opinions from different experts in an explicitly model-based way to construct a valid subjective prior in a Bayesian statistical approach. We propose a generic approach by considering a hierarchical model accounting for various sources of variation as well as accounting for potential dependence between experts. We apply this approach to two problems. The first problem deals with a food risk assessment problem involving modelling dose-response for *Listeria monocytogenes* contamination of mice. Two hierarchical levels of variation are considered (between and within experts) with a complex mathematical situation due to the use of an indirect probit regression. The second concerns the time taken by PhD students to submit their thesis in a particular school. It illustrates a complex situation where three hierarchical levels of variation are modelled but with a simpler underlying probability distribution (log-Normal).

Keywords: Bayesian statistics, Hierarchical model, Random effects, Risk assessment

1 Introduction

In this paper we consider the problem of combining opinions from different experts in an explicitly model-based way to construct a valid subjective prior in a Bayesian statistical approach. In many applied problems, it is necessary to construct complex models. In these models some parts are well informed by what we could call *good data*, that is informative data, whereas in other parts, it is very difficult to collect appropriate data to provide the required information. This occurs for instance when considering contamination by ingestion of some bacteria, say campylobacter (Albert et al. 2008). A complex model can be built by specifying sub-models, which are then combined. Data are provided to inform some sub-models, in order to obtain as much information as possible on the global model. However in other sub-models very little data are available so that it is necessary to use expert opinions to supplement the information provided in

[*]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-BA717

the other well-informed sub-models. From a Bayesian perspective, this corresponds to constructing informative priors on some of the parameters for which data can provide little information (Berger 2006). The construction of such informative priors using expert opinions is a delicate problem, because the human mind finds it difficult to quantify qualitative knowledge; and has been reviewed from various perspectives (e.g. Spetzler and Staël von Holstein 1975; Gill and Walker 2005; O'Hagan et al. 2006; Kynn 2008; Low-Choy 2012).

Consider a sampling model with observation X following a probability distribution P_θ, with unknown parameter θ and as in any Bayesian approach a prior π on θ is constructed. The aim of prior elicitation is then to construct such a prior probability distribution for θ using expert knowledge. In most cases, it is more realistic to base the prior probability elaboration on a parametric family, say $\pi \in \{\pi_\gamma, \gamma \in \Gamma\}$ where γ is also estimated from the experts' knowledge. Indeed, it is often the case that we may not be able to feasibly elicit more than a few quantities from experts, which we call the elicited data.

With more than one expert, we may elicit from each expert a different prior (i.e. a different γ) and in many situations it is desirable to combine these different priors into a single "consensus" prior for θ. There are various methods registered in the literature to achieve this, although most are not entirely satisfactory for applications such as the case studies considered here. The prevailing approaches are averaging (e.g. Burgman et al. 2011) and pooling (Genest and Zidek 1986). Averaging emphasizes the consensus on elicited quantities. Diversity among experts can then be expressed in terms of how much experts differ from this consensus (e.g. Lipscomb et al. 1998). The advantage of averaging is its simplicity, making it accessible, especially where rapid feedback is desired (Burgman et al. 2011). A disadvantage is that focussing on deviations from the average can understate variation by ignoring uncertainty, being the range of values considered plausible by each expert. Furthermore, averaging may mis-represent multiple modes, which may arise for example due to distinct rather than gradual differences in schools of thought.

In contrast, the popular linear or logarithmic pooling methods emphasize diversity, since they accumulate, across experts, the plausibility of all values. Pooling has been advocated as a principled approach to combining expert judgments (O'Hagan et al. 2006). Since pooling can be viewed as a construction of an additive or multiplicative mixture (albeit with specified weights w_ℓ) across individual experts, $\sum_{\ell=1}^{L} w_\ell \pi_{\gamma_\ell}$, it can easily be extended to allow expert weights to be unequal (Genest and Zidek 1986). Indeed Cooke's method (as summarized in Cooke and Goossens 2008) provides an integrated classical method for estimating then pooling with these weights that 'has stood the test of time' (French 2011). Weights are based on p-values for evaluating how well expert assessments on *seed* variables align with empirical results. Whilst pooling acknowledges diversity, it does not highlight a consensus nor capture expert differences from the consensus. Here we seek an approach for combining elicited judgments that retains the advantages yet addresses the disadvantages of these two approaches by explicitly modelling *both* the consensus and diversity of expert opinions, whilst acknowledging multiple sources of uncertainty and variation.

In our approach, we deviate from these traditional approaches of averaging and pooling, by treating the elicited information as data, in the spirit of other Bayesian approaches to combining expert judgments (see for instance Winkler 1968; Lindley 1983; French 1985; West 1988). The decision-maker (DM) updates their knowledge according to Bayes' Theorem (French 1985, 2011). The DM begins with some prior opinion ψ_γ on the parameter θ, which may then be updated using information elicited from experts, to provide an updated *posterior* view. In turn, this posterior may then form the prior for a subsequent analysis, and is here denoted π_γ.

Our approach falls under the Parameter-Updating form of the Supra-Bayesian approach (Roback and Givens 2001), since we combine priors on plausible values of parameters, represented by probability distributions, rather than probabilities of events (French 2011). However instead of directly eliciting the hyper-prior parameters γ for θ, we perform indirect elicitation of expert knowledge on more intuitive observables $X|\theta, \gamma$, and then infer γ. This indirect approach to encoding has previously been applied in various situations, including Generalized Linear Models (GLMs), but only for a single expert (e.g. Kadane et al. 1980; Bedrick et al. 1996; Kynn 2005; Denham and Mengersen 2007; James et al. 2010). Here we demonstrate how to combine indirect elicitations across multiple experts.

The method is generic and we consider two applications. One deals with risk assessment using a dose-response model for *Listeria monocytogenes* on mice, the second deals with the time taken to submit a PhD dissertation by students in applied mathematics in an Australian university.

In Section 2 we describe the approach and the hierarchical model. In Section 4 we consider the the dose-response and the PhD example and Section 6 contains some conclusions.

2 Method

Let X be a possible vector of observations from a distribution P_θ, $\theta \in \Theta$, with density $f(X|\theta)$. As described in the introduction, we suppose that X provides only a limited amount of information on θ (for instance X consists of a small number of i.i.d. replicates), so that the prior on θ is likely to have an impact. In this Section we describe a generic approach for combining elicited expert assessments about θ.

Each expert may have their own conceptual model about θ, which we parameterize by γ so that $\pi(\theta|\gamma)$ belongs to a parametric class $\{\pi_\gamma, \gamma \in \Gamma \subset \mathbb{R}^p\}$. The aim is to construct an informative prior probability distribution on θ based on expert information denoted by D_{elicit}, in a way that accounts for our uncertainty in modelling each expert's information. Information on γ can be obtained using the posterior from a Bayesian analysis of the elicited information, that begins with a prior π_0 and treats elicited expert knowledge as data (e.g. Lindley 1983), using the following scheme:

$$\pi(\gamma|D_{\text{elicit}}) \propto f(D_{\text{elicit}}|\gamma)\pi_0(\gamma). \tag{1}$$

Thus the elicited data D_{elicit} is considered conditional on the expert's knowledge, which here is conceptualized as being represented by the distribution π_0 and parameter γ.

The information on θ can be summarized by integrating this posterior over possible parameterizations indexed by γ:

$$\pi(\theta|D_{\text{elicit}}) = \int \pi(\theta|\gamma)\pi(\gamma|D_{\text{elicit}})d\gamma. \tag{2}$$

We thus adopt the so-called supra-Bayesian approach (see for instance Gelfand et al. 1995) where a supra-Bayesian constructs a likelihood for the elicited data $f(D_{\text{elicit}}|\gamma)$. This construction is detailed in Section 2.2.

2.1 About the experts and the elicited data

In the following we interview N experts. To each expert e corresponds an unknown hyperparameter γ_e resulting in their own prior distribution $\pi(\theta|\gamma_e)$. To estimate this hyperparameter, we interview each expert e and encode their knowledge on X. We denote by D_e the set of the elicited quantities provided by expert e. These quantities may vary in nature, reflecting variation within each expert or across experts or both. For the sake of presentation and also because they are often considered more reliable (see O'Hagan et al. 2006, for a review of encoding techniques), we restrict our attention to elicited quantiles (also known as fractile estimation) and probabilities (also known as interval estimation). Therefore here D_e consists of a vector of quantiles Q_e and a vector of probabilities P_e.

In the following we denote by $|x|_o$ the dimension of a vector x. Let $Q_e = (Q_{ek}, k = 1 \dots |Q_e|_o)$ be the vector of elicited quantiles of the distribution of interest $f(X|\theta)$ corresponding to specified cumulative probabilities $(p_{ek}, k = 1 \dots |Q_e|_o)$ for expert e. Then $Q_{\text{elicit}} = (Q_e, e = 1 \dots N)$ is the vector of all the elicited quantiles for all the experts. Similarly, we denote cumulative probabilities by $P_e = (P_{e\ell}, \ell = 1 \dots |P_e|_o)$ the set of the elicited probabilities of $f(X|\theta)$ at the specified quantiles $(q_{e\ell}, \ell = 1 \dots |P_e|_o)$ for expert e. Then $P_{\text{elicit}} = (P_e, e = 1 \dots N)$ is the vector of all the elicited probabilities for all the experts. We denote the complete set of elicited data by $D_{\text{elicit}} = (P_{\text{elicit}}, Q_{\text{elicit}}) = (D_{et}, t = 1 \dots |P_e|_o + |Q_e|_o)$, t being the index of the question. Each block of answers (P_{elicit} or Q_{elicit}) can be used as separate sources to provide the elicited distribution (2) as detailed below.

For each question t, the expert e also provides a measure of uncertainty in their answer in the form of a number $c_{et} \in (0, 1)$ quantifying the expert's confidence in their response. This information allows us to build a measurement error model to quantify each expert's individual accuracy, adopting similar ideas to earlier approaches to elicitation modelling (Lindley et al. 1979; Lindley 1983) as described below.

2.2 A model of error for elicited data

The individual inaccuracies for each expert can be modelled using a measurement error model. Set η as the link function. The link-transformed elicited data may follow a general measurement error distribution h with measurement error σ_{et}, for $e = 1 \dots N$ and $t = 1, \dots |P_e|_o + |Q_e|_o$:

$$\eta(D_{et}) \sim h(\eta(d_t(\gamma_e)); \sigma_{et}). \tag{3}$$

Here $d_t(\gamma_e)$ is the theoretical response to the question relative to t-th quantity of the distribution of X under the model $X|\theta \sim P_\theta$ and $\theta \sim \pi_{\gamma_e}$. For instance the quantile $q_k(\gamma_e)$ and the probability $p_\ell(\gamma_e)$ respectively satisfy:

$$\int P(X \leq q_k(\gamma_e)|\theta) d\pi(\theta|\gamma_e) = p_{ek} \quad \text{and} \quad \int P(X \leq q_{e\ell}|\theta) d\pi(\theta|\gamma_e) = p_\ell(\gamma_e). \tag{4}$$

A common measurement error model is additive:

$$\eta(D_{et}) = \eta(d_t(\gamma_e)) + \epsilon_{et}, \quad e = 1 \dots N \quad t = 1, \dots |P_e|_o + |Q_e|_o. \tag{5}$$

The expressions for $q_k(\gamma_e)$ and $p_\ell(\gamma_e)$ derived from equations (4) depend on the application. Expressions specific to the dose-response parameter in a probit model and the location and scale parameters for a lognormal distribution are detailed for each example in Section 4.

The link function η can be different for quantiles and probabilities. Typical examples of link functions for probabilities are probit or logit link functions.

We assume that the error (on the appropriate scale) with which experts specify elicited quantities are conditionally independent, given the expert's conceptual model γ_e. The ϵ_{et} are therefore independent and have a known distribution h_{et} constructed by the assessor (or supra-Bayesian) from the expert's measures of uncertainty c_{et}, together with measures of individual coherence and precision considered by the assessor (or supra-Bayesian), based for instance on the training of the expert or on previous expertise (this point is detailed in the examples of Section 4). In our examples, we have considered $h_{et} = h(.|v_{et})$ to be centered Gaussian distributions with variance v_{et}. These variances are constructed using the expert's measures of uncertainty c_{et} together with extra information on uncertainty. Thus, the influence of the answers of the expert e is assessed via the error densities h_{et}: an expert whose own measures of uncertainty are large typically would have error densities $h(.|v_{et})$ with large variance inducing a weak influence of D_{et} on the likelihood.

Remark 1. *In practice and in our examples, only a few quantities (say fewer than 10) are generally elicited from each expert. It is however useful to ask for more quantities than $|\gamma|_o$ to check for coherence in each expert's elicitation. Note that the above error model allows for some incoherence in the elicitation of the experts, in the sense that there might not exist a γ_e such that $d_t(\gamma_e) = D_{et}$ for all t. However a sensible choice of the distribution of ϵ_{et} will ensure that the error model provides a reasonably coherent set of quantiles conforming to the order imposed by both $(p_{ek}, q_{e\ell}, k = 1 \dots |Q_e|_o, \ell = 1 \dots |P_e|_o)$.*

2.3 Combining the experts' opinions: a hierarchical model

The key issue is to derive a final unique distribution $\pi(\theta|D_{\text{elicit}})$ taking into account the fact that the elicitations vary among the N experts and that potential dependence between experts may exist. This pooling step relies on the building of the joint likelihood of expert opinions. One option is to model this likelihood using a multivariate distribution, such as a multivariate normal (e.g. Lindley 1983). This highly parameterized approach requires estimation, and therefore specification of hyperparameters for several fixed effects: bias (additive and multiplicative) of individual experts as well as correlations between experts. A random effects model provides a more parsimonious approach.

We assume that the N experts can be grouped into J homogeneity classes (of respective sizes N_j) , corresponding to similar background or similar schools of thought for instance. Thus, from now on, we label the experts according to their class, so that $e = (i, j)$ denotes the i-th expert in class j.

In order to represent variation between and within classes of expert opinions, we consider a hierarchical formulation of a random effects model (e.g. Lipscomb et al. 1998; Lin and Bier 2008). We suggest the following hierarchical model to group the experts:

$$\begin{array}{ccll} \gamma_{ij} & \overset{\text{i.i.d.}}{\sim} & g(\cdot|\gamma_j, b_j), & \forall i = 1, \ldots, N_j, \\ \gamma_j & \overset{\text{i.i.d.}}{\sim} & g(\cdot|\gamma, b), & \forall j = 1, \ldots, J, \\ \gamma & \sim & \pi_0 & \end{array} \tag{6}$$

where π_0 is the assessor's prior. In other words the expert opinions grouped into the same homogeneity class have the same distribution $g(.|\gamma_j, b_j)$. Then the different groups have knowledge that can be linked via a common distribution $g(.|\gamma, b)$. Finally in the last level a prior is used, representing the overall uncertainty on γ prior to the elicitation. Thus γ can be understood as the *true parameter* of model (2), or more realistically as the parameter representing the agreement of experts. In model (6), the γ_j's are location parameters and so is γ. The hyperparameters b_j, b are typically dispersion parameters. In Section 3.1, we consider the following example of model (6) in the case of $\gamma = (\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$:

$$\begin{array}{ccccccc} \forall j \quad \mu_{ij}|\mu_j, \tau_j & \overset{\text{i.i.d.}}{\sim} & \mathcal{N}(\mu_j, \tau_j) & \text{and} & \frac{\sigma_{ij}^2}{\sigma_j^2}|\sigma_j, \xi_j \overset{\text{i.i.d.}}{\sim} \Gamma(\xi_j, \xi_j) & i = 1 \ldots N_j \\ \mu_j|\mu, \tau & \overset{\text{i.i.d.}}{\sim} & \mathcal{N}(\mu, \tau) & \text{and} & \frac{\sigma_j^2}{\sigma^2}|\sigma^2, \xi \overset{\text{i.i.d.}}{\sim} \Gamma(\xi, \xi), & j = 1 \ldots N \\ \mu & \sim & \mathcal{N}(\mu_0, V) & \text{and} & \sigma^2 \sim \sigma_0^2 \Gamma(a, a) & \end{array} \tag{7}$$

corresponding to $\gamma_j = (\mu_j, \sigma_j^2)$, $b_j = (\tau_j, \xi_j)$, $\gamma = (\mu, \sigma^2)$ and $b = (\tau, \xi)$. The hyperparameters are thus $\{b_j, j = 1 \ldots J\}, b, \mu_0, \sigma_0^2, V, a$. This model-based approach seeks to quantify the two variance components, which could be used to inform design of elicitation.

Our approach is also useful when a consensus expert model is desired, since it ex-

plicitly combines potentially disparate expert elicitations in order to specify the prior distribution of interest.

Remark 2. *In model (6) we have implicitly considered that we weight each group equally, which would typically occur in situations where we have no extra knowledge on the quality of the groups of experts that have been interrogated nor on the specific reliability of a specific expert within a group. However it is possible to consider other situations where for instance one of the groups is a priori known to be less reliable or to have an opinion which represents a smaller fraction of the population of experts (outside those interrogated) than the other groups, in which case we can use this extra knowledge (based on previous elicitations made by this group, for instance) to consider a specific distribution for the parameter γ_j corresponding to this group. Such a scenario can occur when the groups correspond to different schools of thought, say you have two groups corresponding to two schools of thought, the first one corresponding to the majority of the population of experts and the second one being more marginal. In such a situation, even though it is important to take into account the second group we may not want to put too much weight on the answers of these experts. One way to take into account this difference is to assume a higher variance parameter for its distribution, in which case γ_1 and γ_2 would follow respectively a distribution in the form $g(\cdot|\gamma, b)$ and $g(\cdot|\gamma, bb')$ with b' is such that the variance of the second distribution is greater. Or, if the second group is known to have a systematic bias of some order of magnitude, we could consider a distribution in the form: $\gamma_2 \sim g(\cdot|\gamma + \Delta, b)$ where Δ is assessed using this extra knowledge. Hence any other knowledge on the behaviour of each expert or group of experts could be and should be included in the model, using variations such as the two just described. For example, a 'supra-Bayesian' could estimate the imprecision b' and systematic bias Δ for each expert (Lindley 1983), or these could be estimated via responses to seed questions where the answer is known (Genest and Zidek 1986).*

Remark 3. *The sizes N_j of the groups play an important role. The larger they are the better in terms of statistical information, nonetheless the approach can be used for small sample sizes as illustrated in the dose-response example. However, the number of experts is often dictated by the area and level of expertise required (Low-Choy et al. 2010), and some experts may provide more relevant or accurate assessments than others (Lin and Bier 2008). Moreover in the situation where a group j contains a single expert, the model for this group is reduced to $\gamma_{ij} \sim g(.|\gamma)$.*

Finally, we have constructed a Bayesian hierarchical framework to model the imprecision and incoherence of individual experts as well as their variability (between experts). We now present two estimation methods deriving from two ways of formulating the model to utilize both sources of elicitation data.

2.4 Estimation of the elicited distribution

The difference between the two following methods (method A and method B) is mainly characterised by how priors on hyperparameters and parameters are determined. In method B, priors for all elicited data are considered while in method A the elicited data

is split into two parts, one part is used to estimate the hyperparameters (b_j, b) and the other part is used to construct the likelihood.

In the following, $P_{ij} = (P_{ij\ell}, \ell = 1 \dots |P_{ij}|_o)$ denotes the elicited probabilities of expert i of group j and $Q_{ij} = (Q_{ijk}, k = 1 \dots |Q_{ij}|_o)$ denotes the elicited quantiles of expert i of group j. $P_{\text{elicit}} = (P_{ij\ell}, i = 1 \dots N_j, j = 1 \dots J, \ell = 1 \dots |P_{ij}|_o)$ is the vector of all the elicited probabilities and $Q_{\text{elicit}} = (Q_{ijk}, i = 1 \dots N_j, j = 1 \dots J, k = 1 \dots |Q_{ij}|_o)$ the vector of all the elicited quantiles.

Method A: Two-Stage Estimation of $\pi(\theta|D_{\text{elicit}})$ in Practice.

In method A, we separate the contributions of P_{elicit} and Q_{elicit}. In a first step, the hyperparameters of model (6), namely $(b_j)_{j=1\dots J}, b$ and those entering π_0 (in model (7), these are μ_0, V and a) are estimated from P_{elicit}. In a second step, we "plug" these dispersion hyperparameter estimators into the likelihood of elicited data Q_{elicit} –as described in (9)– and derive the posterior distribution $\pi(\theta, \gamma|D_{\text{elicit}})$ (10) using a Markov Chain Monte Carlo (MCMC) algorithm. More precisely,

- From P_{elicit} we derive preliminary estimators of γ_{ij} by minimizing the least squares objective (Low-Choy et al. 2008):

$$\hat{\gamma}_{ij} = \text{argmin}_\gamma \sum_{\ell=1}^{|P_{ij}|_o} [P_{ij\ell} - p_\ell(\gamma)]^2 \tag{8}$$

  where $p_\ell(\gamma)$ is the theoretical response.

Estimators of $(b_j)_{j=1\dots J}$ and b are then deduced using moment estimators for instance. Various estimates are available, depending on the models and on the elicited quantities. This point is discussed in the two examples (Section 4). We denote by $(\hat{b}_j)_{j=1\dots J}$ and $\hat{b}$ the obtained estimates.

- Using (2), (5) and (6) and plugging in the estimated dispersion hyperparameters, we deduce the likelihood of elicited data Q_{elicit}:

$$\begin{aligned} f\left(Q_{\text{elicit}}|\gamma, (\hat{b}_j)_{j=1\dots J}, \hat{b}\right) &= \int \prod_{ijk} h_{ijk}(\eta(Q_{ijk}) - \eta(q_k(\gamma_{ij}))) \\ &\quad \times \prod_{ij} g(\gamma_{ij}|\gamma_j, \hat{b}_j) \prod_j g(\gamma_j|\gamma, \hat{b}) d\gamma_j \; d\gamma_{ij} \end{aligned} \tag{9}$$

- Finally using:

$$\pi(\theta|D_{\text{elicit}}) \propto \int \pi(\theta|\gamma) f\left(Q_{\text{elicit}}|\gamma, (\hat{b}_j)_{j=1\dots J}, \hat{b}\right) \pi_0(\gamma) d\gamma \tag{10}$$

  we generate Markov realizations of (γ, θ) under the posterior distribution:

$$\pi(\theta, \gamma|D_{\text{elicit}})$$

  through Markov Chain Monte Carlo.

This is similar to an empirical Bayesian procedure, except that there is no double-use of the data since we split the elicited data into two parts used respectively for the estimation of the hyperparameters and for the computation of the likelihood. Note however that elicited data is of a specific nature since each data point conveys (hopefully) a lot of information (knowing a quantile of a distribution is much more informative than knowing only one realisation). Hence, even though we avoid the double use of the data, per se, we do not exactly avoid the double use of the information. This method is useful in situations where the number of experts is small, since it avoids using weakly informative priors on the hyperparmeters. We found this approach well-suited to the dose-response case study on food risk.

In the case of sufficient numbers of experts and so a sufficient amount of elicited data, we could implement a global MCMC approach, avoiding the plug-in step for the dispersion hyperparameters. This is described in Method B.

Method B: All-in-one Estimation of $\pi(\theta|D_{\text{elicit}})$ in Practice.

The second method specifies priors on $(b_j, j = 1 \dots J), b, v = (v_{ijt}, t = 1 \dots |D_{ij}|_o, i \leq N_j, j \leq J)$, where v represents the imprecision parameters involved in the distribution of errors ϵ_{et} (see (5)) and defines a joint Bayesian elicitation model for Q_{elicit} and P_{elicit}:

$$\pi(\theta|D_{\text{elicit}}) \propto \int_{\gamma,c,d} \pi(\theta|\gamma) f\left(Q_{\text{elicit}}, P_{\text{elicit}}|\gamma, v, b\right) \pi_0(\gamma)\pi_0(v) d\gamma \; db \; dv \tag{11}$$

From the conditional independence of the error model (5) we obtain

$$f(D_{ij}|\gamma_{ij}, v_{ij\cdot}) = f(P_{ij}|\gamma_{ij}, v_{ij\cdot}) f(Q_{ij}|\gamma_{ij}, v_{ij\cdot}). \tag{12}$$

Method B gives equal weight to the P-elicitations and Q-elicitations, and is also fully Bayesian, so in this regard is more satisfying; however as it needs to estimate also the scale parameters it is better adapted to more than a few experts. Method A can be interpreted as a two-stage modelling approach, where the second stage is Bayesian, but the first stage utilizes simple Frequentist point-estimates of hyperparameters in the prior. Method A has also the advantage of simplifying the computation.

3 Examples

In this section, we detail a particular case of the hierarchical model (6), namely model (7) and discuss some technical points such as the model for hyperparameters $(b_j)_{j=1\dots J}$, and b and the error model. In a second step, we use this hierarchical model and our methodology on two examples. The first example arose in food risk science and is a model for a dose-response to a pathogen for mice. In the second example we are interested in the time to thesis submission for an applied mathematical PhD student in an Australian university. Although the two problems are of very different natures, the hierarchical structures of the models used for combining the different expert opinions follow a similar pattern, following either Method A or B detailed above.

3.1 Description of a particular hierarchical model (6) used for both examples

In both examples the parameter γ is composed of a mean μ and a variance σ^2: $\gamma = (\mu, \sigma^2)$. We consider the hierarchical model (7) to model the possible interactions between experts.

The hyperparameters $(b_j)_{j=1\dots J} = (\tau_j, \xi_j)_{j=1\dots J}$, $b = (\tau, \xi)$ and $(\mu_0, \sigma_0^2, a, V)$ must be modelled carefully since their influence might be important, especially when the numbers of experts and of elicited quantities are small, which is a common situation. Under method A, we estimate the hyperparameters $(\tau_j, \xi_j), j \leq J$, (τ, ξ), (V, a) and (μ_0, σ^2) using P_{elicit} in the following way. First we derive $\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2$ via least squares (see (8)). The solution of this equation is specific to each example and is detailed in Sections 3.4 and 3.3. Once such estimates have been obtained, we use moment estimators for $(b_j)_{j=1\dots J}$ and b. More precisely, since τ_j represents the conditional variance of μ_{ij} in the group j (given μ_j), a natural estimate is given by $\hat{\tau}_j = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (\hat{\mu}_{ij} - \hat{\mu}_j)^2$, where $\hat{\mu}_j$ is the average of the $\hat{\mu}_{ij}$'s in the group j. The variance τ can also be estimated using $\hat{\tau} = \frac{1}{J-1} \sum_{j=1}^{J} (\hat{\mu}_j - \hat{\mu})^2$, where $\hat{\mu}$ is the average of $\{\hat{\mu}_j, j = 1 \dots J\}$. Similarly ξ_j^{-1} represents the conditional variance of $\frac{\sigma_{ij}^2}{\sigma_j^2}$ given σ_j^2 so that a natural estimate is $\hat{\xi}_j^{-1} = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} \left(\frac{\hat{\sigma}_{ij}^2}{\hat{\sigma}_j^2} - 1 \right)^2$ and ξ can be estimated using $\hat{\xi}^{-1} = \frac{1}{J-1} \sum_{j=1}^{J} \left(\frac{\hat{\sigma}_j^2}{\hat{\sigma}^2} - 1 \right)^2$, where $\hat{\sigma}^2$ is the average of $\{\hat{\sigma}_j^2, j = 1 \dots J\}$. We then use $\hat{\mu}$ and $\hat{\sigma}^2$ as estimates for μ_0 and σ_0^2. Finally since V and a^{-1} are measures of uncertainty (variances) on μ and $\frac{\sigma^2}{\sigma_0^2}$ we replace them by our observed uncertainty, namely $\hat{V} = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_j + \hat{\tau}$ and $\hat{a}^{-1} = \frac{1}{J} \sum_{j=1}^{J} \hat{\xi}_j^{-1} + \hat{\xi}^{-1}$. These hyperparameters are then plugged into the likelihood to obtain an elicited prior distribution using an MCMC algorithm.

The alternative method (method B) uses all elicited data $(P_{\text{elicit}}, Q_{\text{elicit}})$ for the computation of the elicitation likelihood and uses weakly informative priors for the hyperparameters $(\tau_j, \xi_j, \tau, \xi, a, V, \mu_0, \sigma_0^2)$. We apply truncated Normal priors to the standard deviations (hereafter SD) of random effects. As shown by Gelman (2006), in the context of a model comprising an intercept and one variance component, this form of prior distribution is a useful two-parameter simplification of the folded noncentral-t distribution, which is conjugate for these SD parameters:

$$\sqrt{\tau_{ij}} \sim \mathcal{N}_+(0, \zeta_{\tau_{ij}}^2); \quad \sqrt{\tau_j} \sim \mathcal{N}_+(0, \zeta_{\tau_j}^2); \quad \sqrt{\tau} \sim \mathcal{N}_+(0, \zeta_\tau^2)$$
$$\sqrt{\sigma_0} \sim \mathcal{N}_+(0, \zeta_0^2); \quad \sqrt{V} \sim \mathcal{N}_+(0, \zeta_V^2)$$

where $\mathcal{N}_+$ denotes a Normal distribution truncated at zero to include only non-negative values. Similarly, we may replace the gamma prior $(\sigma_0^2 \Gamma(a, a))$ on the lognormal variance σ^2 (7), with a truncated normal prior on the lognormal SD $\sigma \sim \mathcal{N}_+(0, \zeta_\sigma^2)$. In practice setting the ζ parameters to large values provides weakly informative priors. The hyperparameters for the multiplicative effects have exponentially distributed priors $\xi, \xi_j \sim \mathcal{E}xp(1)$. In addition we utilize weakly informative priors for the consensus mean:

$$\mu_0 \sim \mathcal{N}(0, \zeta_\mu^2). \tag{13}$$

We now describe the error model we have considered to construct the elicitation likelihood.

3.2 Description of the likelihood: error model (5)

In our examples we consider Gaussian errors in the elicitation error model for quantiles (5):

$$\eta(Q_{ijk}) \sim \mathcal{N}(\eta(q_k(\gamma_{ij})), v_{ijk}), k = 1 \dots |Q_{ij}|_o \tag{14}$$

where q_k is defined in (4) with $P(X|\gamma)$ specific to each example (detailed below) and η is a link function. In method B, we also consider Gaussian errors in the elicitation error model for probit-transformed cumulative probabilities (5):

$$\Phi(P_{ij\ell}) \sim \mathcal{N}(\Phi(p_\ell(\gamma_{ij})), v_{ij\ell}), l = 1 \dots |P_{ij}|_o. \tag{15}$$

For either method, the variances v_{ijt} may be estimated using all the available information on the precision of the experts. In particular this allows some flexibility so that experts can provide this information in whatever form they find most natural. For instance, when the experts provide confidence measures $c_{ijt} \in (0, 1)$ with the elicited quantities D_{ijt}, as explained in Section 2.1, we interpret c_{ijt} as a coverage probability of a confidence interval and write

$$\begin{aligned} 1 - c_{ijt} &= P\left[|\eta(D_{ijt}) - \eta((d_t(\gamma_{ij}))| > q_{ij}^\star\right] \\ &= P\left[|\eta(D_{ijt}) - \eta(d_t(\gamma_{ij}))| / \sqrt{v_{ijt}} > q_{ij}^\star / \sqrt{v_{ijt}}\right] \\ &= 2(1 - \Phi(q_{ij}^\star / \sqrt{v_{ijt}})), \end{aligned}$$

so that

$$\sqrt{v_{ijt}} = \frac{q_{ij}^\star}{\Phi^{-1}((1 + c_{ijt})/2)}. \tag{16}$$

The reference value $q_{ij}^\star$ reflects the assessor's estimate of the precision. This can be evaluated from the training of the expert, or from other constraints on the precision such as discretization. The choice of the $q_{ij}^\star$'s is illustrated in the two examples. We can understand in the above formulations the c_{ijt}'s, which are the personal evaluations of the experts of their uncertainty, as relative measures of uncertainty within each expert, the q_{ij}^* as global measures of uncertainty for each expert.

In some other cases, the confidence is not assessed by a value but is given in terms of an interval around a given value. Then setting a level for the confidence interval we obtain a value for v_{ijt} using a similar formulation.

Remark 4. *For the elicited quantiles in the PhD example, η is the identity link and in the food risk example, η is the probit link since in that case $Q_{ijt} \in [0, 1]$. In both examples, η is the logit link for the elicited cumulative probabilities.*

Alternatively, the individual elicitation errors can be modelled using a simpler model than specified earlier (5). For each individual, the error in eliciting the cumulative probabilities v_{ijP} can be expressed as a factor C of the error in eliciting quantiles v_{ijQ} (after suitable link functions have been applied):

$$v_{ijP} = Cv_{ijQ}. \tag{17}$$

A typical prior for a coefficient of variation such as C is a Gamma distribution. This approach suits a situation where (i) the relative elicitation error in quantiles and cumulative probabilities C_q is of interest; and (ii) more information has been elicited, both within and between experts, so that these elicitation errors can be estimated from the elicited information.

We now describe the two examples. In the first example we develop a model to describe the mortality rate for mice under a dose δ of *Listeria monocytogenes* strain EGD or EGDe. The second example concerns the time students take to submit their mathematical PhD thesis in an Australian university.

3.3 Dose-response example

The model used for the dose response example is highly nonlinear and the number of experts is expected to be small in practice, hence we have only considered method A. We consider a (typical) bioassay problem where the dose of some treatment affects mortality. We model the dose-response curve for the contamination of three strains of mice - BALB/c, C57 Black/6 or Swiss - from *Listeria monocytogenes* EGD or EGDe by intravenous injection. Let X_n be the number of dead mice out of n mice exposed to a dose δ. Then, the sampling model is, conditionally on the injected dose, a binomial model of parameter $p(\delta, \theta)$

$$X_n \sim Bin(n, p(\delta, \theta)) \quad \text{with} \quad p(\delta, \theta) = 1 - e^{-\theta\delta}, \quad \theta > 0. \tag{18}$$

$p(\delta, \theta)$ is the probability for a mouse to die from a dose δ of Listeria, (see for instance Haas et al. 1999, p. 264). We are interested in the elicitation of the prior distribution of the unobservable parameter θ. We consider a log-normal distribution: $\log \theta \sim \mathcal{N}(\mu, \sigma^2)$ and denote $\gamma = (\mu, \sigma^2)$.

Following advice (Kadane et al. 1980; Low-Choy et al. 2010), we follow an indirect elicitation approach and ask questions on observables rather than directly interrogating experts about parameters. Each expert chooses a dose $\delta = \delta_{ij}$, which he or she finds easier to work with and is then asked questions about the proportions of dead mice out of a sample of n mice exposed to dose δ (X_n/n), which we relate to the probabilities of mortality, to help formulate the distribution of $p(\delta, \theta)$.

We first ask questions regarding the quantiles of $p(\delta, \theta)$, for which we ask questions based on large samples ($n = 100$) of mice exposed to dose δ, so that we can approximate proportions X_n/n as the probabilities $p(\delta, \theta)$. Recall that the answers given by the expert i of group j are denoted Q_{ijk} and the theoretical associated quantile is

$$q_k(\mu_{ij}, \sigma_{ij}) = 1 - \exp\{-\delta_{ij} \times \exp(\sigma_{ij}\Phi^{-1}(p_{ijk}) + \mu_{ij})\}.$$

The second set of questions concerns the probabilities $P[X_{10} \leq q_{ij\ell}|\delta_{ij}]$, $l = 1 \ldots |Q_{ij}|_o$ where X_{10} is the number of dead mice out of 10 mice submitted to dose δ_{ij}. We denote by $P_{ij\ell}$ the answer given by the expert i of group j and the theoretical probabibility is

$$p_\ell(\mu_{ij}, \sigma_{ij}) = \sum_{r=0}^{q_{ij\ell}} C_{10}^r \int_0^\infty (1 - e^{-\theta\delta_{ij}})^r e^{-(10-r)\theta\delta_{ij}} \varphi((\log\theta - \mu_{ij})/\sigma_{ij}) \sigma_{ij}^{-1} \theta^{-1} d\theta \quad (19)$$

where $\varphi(.)$ is the density function of a standard Gaussian random variable. We then apply the methodology described in Section 2 and compare the elicited prior distribution obtained with this hierarchical approach to those obtained with two standard methods. The first standard method is the **plug-in method**, where a global estimate $(\hat{\mu}, \hat{\sigma}^2)$ is obtained from the method of moments described in Section 3.1 on the overall D_{elicit} leading to a *consensus* elicited prior in the form $\log\theta \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ and the second is the **mixture method** where a point estimate $(\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2)$ is derived for each expert from D_{elicit} leading to a *non-consensus* prior $\log\theta \sim N^{-1} \sum_{i,j} \mathcal{N}(\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2)$. The latter is the same as Cooke's method (as described in (Cooke and Goossens 2008)) in the absence of prior calibration of experts, which is the case in our examples.

To study the robustness of our method we first consider various simulated scenarios presented in the following section.

Note that the comparison is not aimed at showing some superiority of our method, compared to other methods, since we are only comparing with two naive methods. Indeed more sophisticated versions of the plug-in or the pooling of experts exist in the literature. Here we merely wish to better understand how the hierarchical modelling stands in terms of consensus of experts.

Simulation study

We now describe four simulated datasets and comment on the results. In each dataset, the doses δ are different for all the experts and fixed arbitrarily between 10^3 and 10^7. These values correspond to realistic situations. We simulate elicitated probabilities associated with the following quantiles for the number of dead mice out of 10: $\{q_{e\ell}, \ell = 1, 2\} = \{3, 8\}$ and we simulated elicited quantiles of the distribution of the probability that a mouse should die subject to a dose δ, associated with the probabilites $\{p_{ek}, k = 1 \ldots 5\} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. We add an error term with variance $v_{ijt} = 0.1$ for all the experts and all the questions. In each simulated case, we simulate the individual parameters (μ_{ij}) and (σ_{ij}) following:

$$\mu_{ij} \sim \mathcal{N}(\mu_j, \tau_j) \quad , \quad \frac{\sigma_{ij}^2}{\sigma_j^2} | \sigma_j \sim \Gamma(\xi_j, \xi_j)$$

and the σ_j's are fixed.

Dataset 1. Balanced case: In this dataset, we consider a balanced case where we interview 10 experts divided into two groups of the same size ($N_1 = N_2 = 5$). We then

set:

$$\begin{array}{llll} N_1 = 5 & N_2 = 5 & & \\ \mu_1 = -2 & \mu_2 = -1.1, & \tau_1 = 0.01 & \tau_2 = 0.01 \\ \sigma_1^2 = 1 & \sigma_2^2 = 1, & \xi_1 = 100 & \xi_2 = 100. \end{array}$$

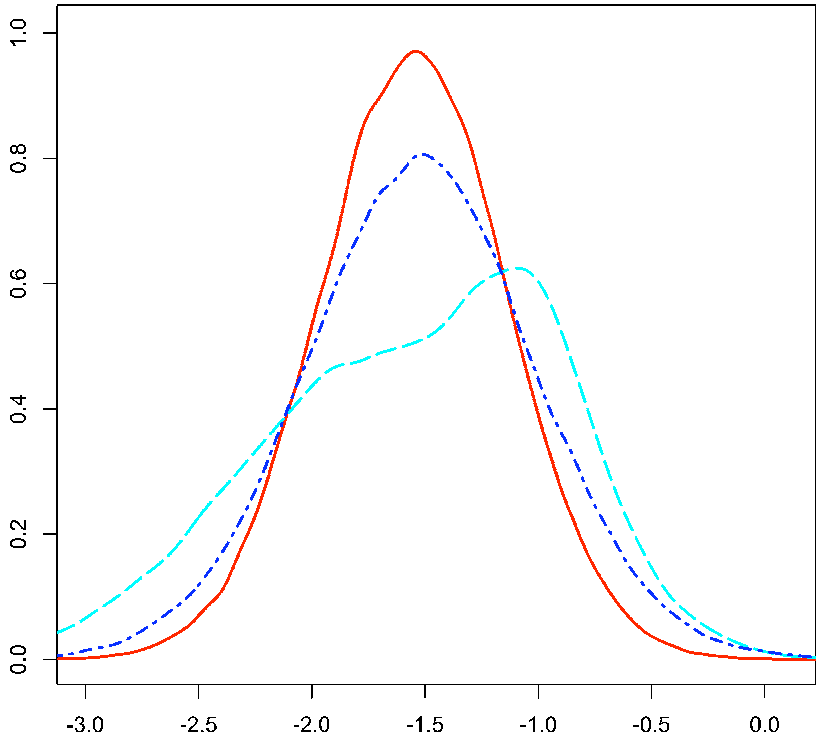

Figure 1: *Dataset* 1. *Balanced case.* Comparison of methods for combination of experts using $p(\log\theta|D_{\text{elicit}})$: mixture $(--)$, plugin (solid line), hierarchical $(-\cdot)$.

The resulting elicited prior distributions are plotted in Figure 1. This standard dataset clearly illustrates the specific behaviour of our hierarchical method. On the one hand, the plug-in method (solid line) forces an agreement between the experts' answers, smoothing the variabilities due to the origin of knowledge for instance. On the other hand, the mixture model $(--)$ takes into account the variabilities and models the difference between experts. The hierarchical model is an intermediate approach allowing one to consider the interactions between experts: the elicited prior distribution of $p(\log\theta|D_{\text{elicit}})$ $(-\cdot)$ (which is thus a posterior) is smoother than the mixture one but has a wider support than the plug-in elicited prior distribution.

Dataset 2. Unbalanced groups of experts: In this dataset, the numbers of experts

in the groups are strongly unbalanced :

$$\begin{array}{llll} N_1 = 10 & N_2 = 2 & & \\ \mu_1 = -2.5 & \mu_2 = -1, & \tau_1 = 0.01 & \tau_2 = 0.01 \\ \sigma_1^2 = 0.5 & \sigma_2^2 = 0.5 & \xi_1 = 100 & \xi_2 = 100. \end{array}$$

In Figure 2, we see again that the mixture method takes into account the global variability whereas the plug-in method (solid line) encourages a consensus, leading to a narrow distribution; the hierarchical method is a compromise between these two. Note that the hierarchical prior has the additional advantage of taking into account the small group, which has been 'forgotten' by the plug-in method. Indeed, the mode of the hierarchical elicited prior is slightly shifted toward the small group (corresponding to $\mu_2 = -1$). This shows that the hierarchical approach clearly does what it is aimed at: take into account the dependencies between experts to avoid redundancies.

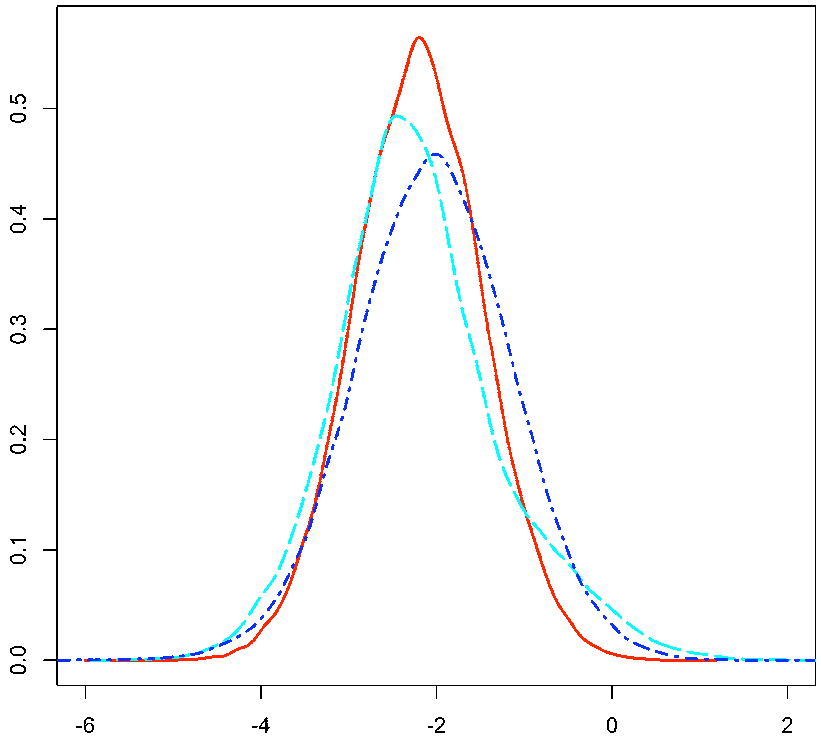

Figure 2: *Dataset* 2. *Unbalanced groups of experts.* Comparison of methods for combination of experts using $p(\log\theta|D_{\text{elicit}})$: mixture $(--)$, plug-in (solid line), hierarchical $(-\cdot)$.

Dataset 3. Mis-specification of the number of groups: In this dataset, we

suppose that the experts are issued from a unique group but the elicitation procedure is performed assuming that there are two groups. This group is simulated under the following set of parameters:

$$\begin{array}{ll} N = 10 & \\ \mu_1 = \mu_2 = -1.5, & \tau_1 = \tau_2 = 0.05, \\ \sigma_1^2 = \sigma_2^2 = 0.5 & \xi_1 = \xi_2 = 100. \end{array}$$

We apply our procedure assuming that the experts are divided into two groups of size $N_1 = N_2 = 5$.

One can see the elicited densities in Figure 3. As expected, we observe similar elicited priors across the three methods: hierarchical, mixture and plug-in. As a consequence, artificially creating a group of experts does not deteriorate the performance of our method.

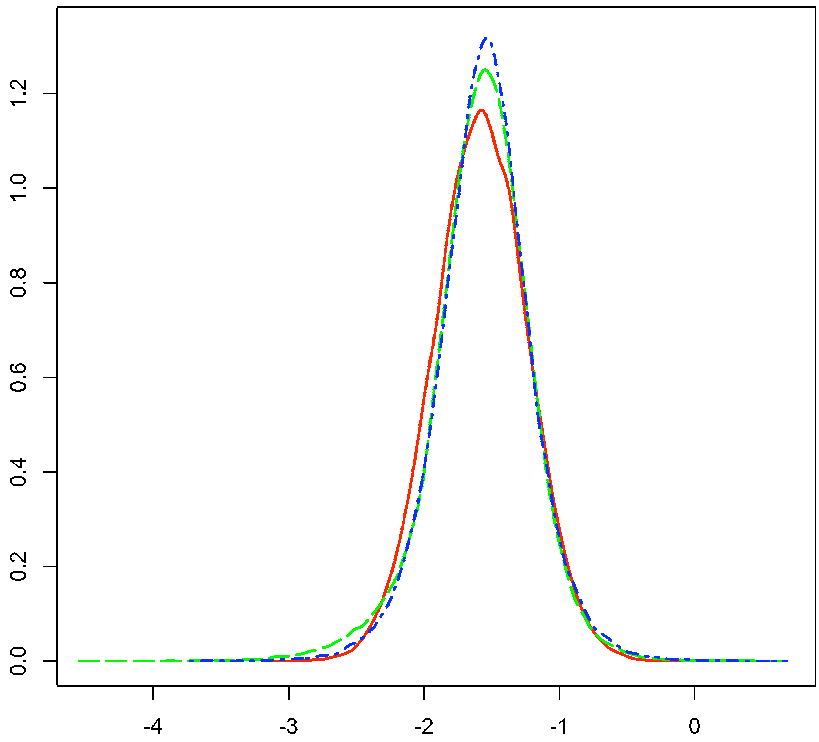

Figure 3: *Dataset* 3. *Mis-specification of the number of groups.* Comparison of methods for combination of experts using $p(\log\theta|D_{\text{elicit}})$: mixture $(--)$, plug-in (solid line), hierarchical $(-\cdot)$.

Real elicited data

A real elicitation has been conducted in this example. Five French experts of Listeria dose-response experiments on mice have been questioned: 3 from Institut Pasteur and 2 from INRA (French National Institute for Agricultural Research). We have asked questions about the quantiles of $p(\delta, \theta)$:

$$P(p(\delta, \theta) \leq Q_k) = k \quad \text{with } k = 1, \ldots, |Q_e|_o \quad \text{and } |Q_e|_o = 3$$

and about the probabilities

$$P_\ell(\mu, \sigma) = P[X_{10} \leq \ell] \quad \text{with } \ell = 1, \ldots, |P_e|_o \quad \text{and } |P_e|_o = 2.$$

The sets T and L have been chosen by the experts and they differ across experts. The doses δ have also been chosen by each expert. For lack of information in the elicited data and since the elicited data lead to very similar estimates of the variances σ_{i2}^2 for the two experts of the INRA group we simplify the model by considering the same variance σ_{i2} in the INRA group : $\sigma_{i2}^2 \equiv \sigma_2^2$. Figure 4 presents the elicited prior densities of $p(\delta, \theta)$ for a fixed usual dose $\delta = 4$ by mixture, plug-in and hierarchical approaches. The density of a Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$ is added as an example of a non-informative prior on $p(\delta, \theta)$. In this case, the higher prior weights on values near 0 or 1 may be interpreted as reflecting an expert's tendency to think concretely of whether mortality occurs or not on a single trial.

As shown on simulation, the mixture approach (i.e. Cooke's method with equal weights) models the differences between experts and the plug-in method encourages an agreement between the experts' answers. The two modes in the mixture model results reflect the large inter-expert variability, indicating that two experts have quite different opinions. The hierarchical model provides results close to the plug-in method results but with a larger support and a slight translation towards the left probably due to a smaller weight on the second group. Practically this could lead to quite different inferences in the lower tail, which could be pivotal for decision-making related to limitations in the efficacy of the dose. After accounting for both intra- and inter-expert variability, the hierarchical model provides a larger estimated probability (compared to the plug-in) that the mortality rate (at fixed dose of 4) is lower than 20%, and consequently weaker evidence that mortality will be greater than 20% at this dosage. The hierarchical formulation is the only model which both reflects this increased chance of low efficacy at low dose, as well as smoothing the estimated probability of survival rate near the mode (approx. 60%). The differences between the non informative prior (Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$) and the elicited prior distributions clearly indicate that experts supply information on the parameter.

3.4 PhD example

Contrary to the first example, in this Section we apply methods A and B, and the relations that are involved are mainly linear. This example illustrates that using a vague prior (on the scale of the parameter) at the lower level of the hierarchy does not

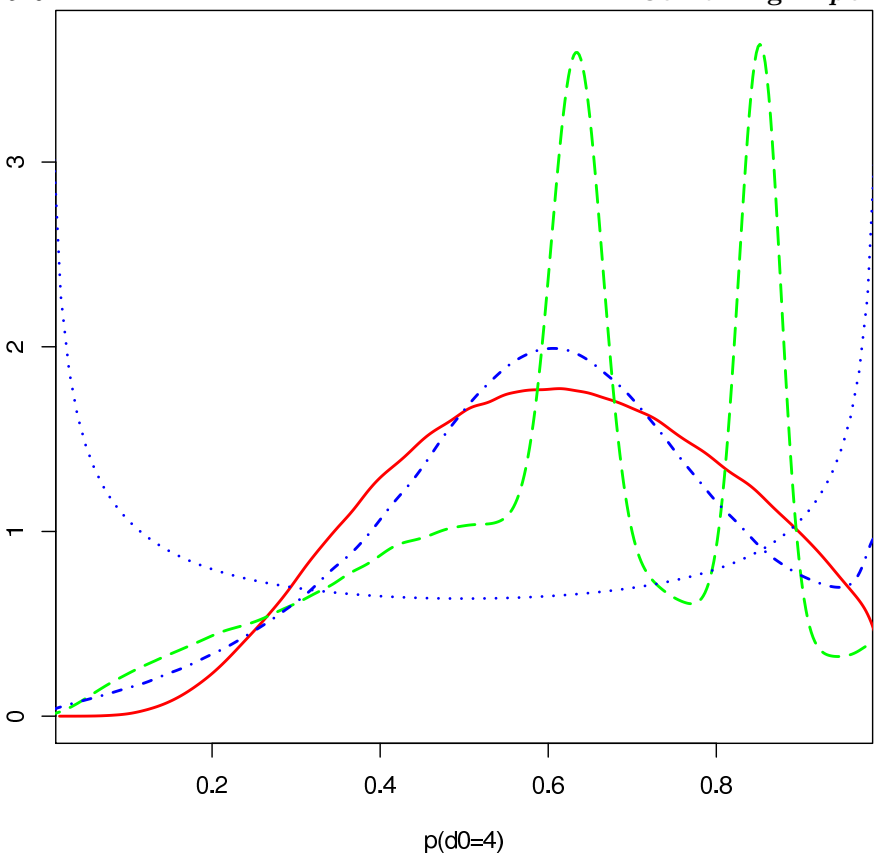

Figure 4: Non-informative prior Beta$\left(\frac{1}{2}, \frac{1}{2}\right)(\cdots)$ and elicited prior densities of $p(\delta = 4)$ with real experts data using: mixture (−−), plug-in (solid line), hierarchical $(-\cdot)$ approaches.

necessarily lead to excessively wide elicited distributions on the time to submission of a PhD thesis.

Let X^* be the time to submission for a PhD student in applied mathematics in the Queensland University of Technology in Australia. The experts were much more comfortable with answering questions based on X^*, which correspond to observable quantities. This agrees with advice on targeting elicitation (Kadane et al. 1980; Low-Choy et al. 2010). Hence we work with the marginal distribution of X^* given μ, σ^2, which differs between experts since they each have their own conceptual model for μ and σ^2.

There is a logical constraint on minimum submission times; experts agreed that except in very rare situations which fall beyond the scope of this model, PhD students would need a minimum of 2 years' candidature before submitting a thesis. This reflects both administrative and practical constraints particular to the university and faculty.

Therefore, the quantity of interest is based on $X^* - 2 > 0$. Also, the time to submission for a PhD is expected to have quite fat tails, as a random variable, we therefore assume that $X = \log(X^* - 2)$ follows a Normal distribution with mean μ and variance σ^2. Such a marginal distribution can be obtained for instance from the following model:

$$X|\theta, v \sim \mathcal{N}(\theta, v), \quad \theta|\mu, v, \rho^2 \sim \mathcal{N}(\mu, v\rho^2), \quad \sigma^2 = v(1+\rho^2). \tag{20}$$

We apply the hierarchical model for describing variation in μ_{ij} and σ_{ij} across experts and groups, as described in (7) in Section 3.1.

Elicitation was conducted in two phases. In each phase different styles of questions were asked. The order of assigning styles to the two phases was randomized for each expert to eliminate anchoring effects. These two styles correspond to (i) Q_{elicit}, eliciting quantiles for specified cumulative probabilities (also known as fractile estimation) and (ii) P_{elicit}, eliciting cumulative probabilities for specified quantiles (also known as interval estimation). To address (i) we asked questions such as "For most students (95 in a hundred), what would you estimate to be the shortest and longest time taken to submit their PhD thesis?" To address (ii) we asked questions such as "In a cohort of one hundred PhD students, how many would you expect to submit their PhD thesis within 4 years?" These two approaches have been used iteratively within a feedback cycle to elicit opinions (Low-Choy et al. 2010). The methodology presented here, however, allows us to retain information from both styles of elicitation, and explicitly model the variability arising from each method separately.

We report results from four experts interviewed in phase I, who were asked for five quantiles associated with probabilities in $\{0.025, 0.25, 0.5, 0.75, 0.975\}$, and two probabilities associated with quantiles in $\{\log(3-2) = 0, \log(4-2) \approx 0.7\}$. We report on results from another five experts interviewed in phase II, who were asked for six quantiles associated with probabilities in $\{0.01, 0.025, 0.25, 0.75, 0.975, 0.99\}$, and four probabilities associated with quantiles in $\{\log 0.5 \approx -0.7, 0, \log 1.5 \approx 0.4, \log 2 \approx 0.7\}$. Only two experts in the latter group could estimate with any level of confidence the cumulative probability associated with the quantile corresponding to the proportion of students that submit in under 2.5 years, we are thus, similarly to the dose-response example, in a case where the experts did not provide the same quantities. Here eliciting three or four cumulative probabilities was satisfactory given that we desired a minimum of two such values. Similarly to before we assume that the error model is Gaussian so that the likelihood associated with the error model for Q_{elicit} is given by $\prod_{i=1}^{N_j} \prod_{j=1}^{J} \phi(Q_{ijk} - q_k(\mu_{ij}, \sigma_{ij}^2)|v_{ijk})$, with $\phi(.|v)$ denoting the density of a centred Gaussian random variable with variance v.

The above model implies that for each k, and corresponding $p_{ijk} \in (0,1)$, the theoretical quantile corresponding to the expert's conceptual model (parameterized by μ_{ij}, σ_{ij}) is $q_k(\mu_{ij}, \sigma_{ij}) = \sigma_{ij}\Phi^{-1}(p_{ijk}) + \mu_{ij}$ and for each $\ell \in \mathbb{R}$, the theoretical probability associated with the quantile $q_{ij\ell}$ is given by $p_\ell(\mu_{ij}, \sigma_{ij}) = \Phi((q_{ij\ell} - \mu_{ij})/\sigma_{ij})$. This provides the basis for both approaches to estimation. For method A, the second set of equations

allow us to determine estimates for μ_{ij} and σ_{ij} by solving for each (i,j)

$$\text{argmin}_{\mu,\sigma} \sum_{\ell=1,\ldots,|P_e|_o} \left(\Phi^{-1}(P_{ij\ell})\sigma + \mu - q_{ij\ell}\right)^2,$$

which leads to:

$$\hat{\mu}_{ij} = \bar{q}_{ij} - \bar{\Phi}^{-1}(P_{ij\ell})\hat{\sigma}_{ij} \quad \text{and} \quad \hat{\sigma}_{ij} = \frac{\sum_\ell (\Phi^{-1}(P_{ij\ell}) - \bar{\Phi}^{-1}(P_{ij\ell}))(q_{ij\ell} - \bar{q}_{ij})}{\sum_\ell (\Phi^{-1}(P_{ij\ell}) - \bar{\Phi}^{-1}(P_{ij\ell}))^2},$$

where $\bar{q}_{ij}$ is the average of the values $q_{ij\ell}$ over ℓ and $\bar{\Phi}^{-1}(P_{ij\ell})$ is the average of the values $\Phi^{-1}(P_{ij\ell})$ over $\ell = 1,\ldots,|P_e|_o$. In other words $(\hat{\mu}_{ij}, \hat{\sigma}_{ij})$ is the least squares estimate associated with the linear model $\Phi^{-1}(P_{ij\ell})\sigma_{ij} + \mu_{ij} + \epsilon_{ij\ell} = q_{ij\ell}$, where $\epsilon_{ij\ell}$ represents the individual error of elicitation. Hence we implicitly consider an error model on the elicitated probabilities similar to the error model on the elicitated quantiles. Then the hyperparameters are estimated as described in Section 3.1.

We consider both the two-stage modelling approach (A), as described in the previous example and the fully Bayes (one-stage) approach (B). For the latter, the likelihood for the Q_{elicit} is supplemented by a likelihood for the P_{elicit}:

$$f(\mathrm{D}_{\text{elicit}}; \gamma_{ij}, v_{ij}, w_{ij}) = \left[\prod_k p(q_{ijk}|\gamma_{ij}, v_{ijt(k)})\right]\left[\prod_\ell p(p_{ij\ell}|\gamma_{ij}, v_{ijt(l)})\right]$$

where

$$\begin{aligned} q_{ijk} &\sim \mathcal{N}\left(q_t(\mu_{ij},\sigma_{ij}), v_{ijt(k)}\right) \\ \Phi^{-1}(p_{ij\ell}) &\sim \mathcal{N}\left(\Phi^{-1}(p_\ell(\mu_{ij},\sigma_{ij}), v_{ijt(\ell)}\right) \\ &\quad t(\ell) = \ell \text{ and } t(k) = k + |P_e|_o \end{aligned}$$

leading to a joint distribution given by

$$\begin{aligned} &p(\mu|\mu_0,\tau_0)p(\sigma^2|\sigma_0^2,\xi_0)\prod_{j=1}^{J} p(\mu_j|\mu,\tau)p(\sigma_j^2|\sigma^2,\xi)\prod_{i=1}^{N_j} p(\mu_{ij}|\mu_j,\tau_j)p(\sigma_{ij}^2|\sigma_j^2,\xi_j) \\ &\times f(\mathrm{D}_{\text{elicit}}; \gamma_{ij}, v)p(v). \end{aligned}$$

Recall that the variances $v = (v_{ijt}, i, j, t = 1,\ldots,|P_e|_o + |Q_e|_o)$ are determined using (16), and in this example we consider the following prior that is vague on the scale of the quantile: $q_{ij}^* \sim \mathcal{N}_+(0,10)$. In this case study v reflects the expert's coherence, among all elicited P and Q quantities, as well as their fidelity to a lognormal distribution for X. In comparison, a deterministic method would assume accuracy of all elicited quantities (P or Q) and the lognormal distribution.

For method B, hyperparameters were set as $\xi_j = 0.5, \xi = 0.5$ for the multiplicative effects on the lognormal variance, within and between groups. Setting $\zeta_0^2 = 10, \zeta_\tau^2 = 10$, also sets a weakly informative prior over the interval $[0,10]$. The coefficient of variation C

relating the error in eliciting cumulative probabilities with respect to quantiles, specified in (17), is thought to be over one, since the latter is an easier task (O'Hagan et al. 2006). However we supposed that the two elicitation tasks are not vastly different in difficulty, so that C is not likely to exceed fifty. This prior assessment led to specification of a Gamma prior on C: $C \sim \mathcal{G}(1, 0.1)$, which has a mean at one, but a non-zero mode (located at 2) with 99.3% of its values falling below fifty.

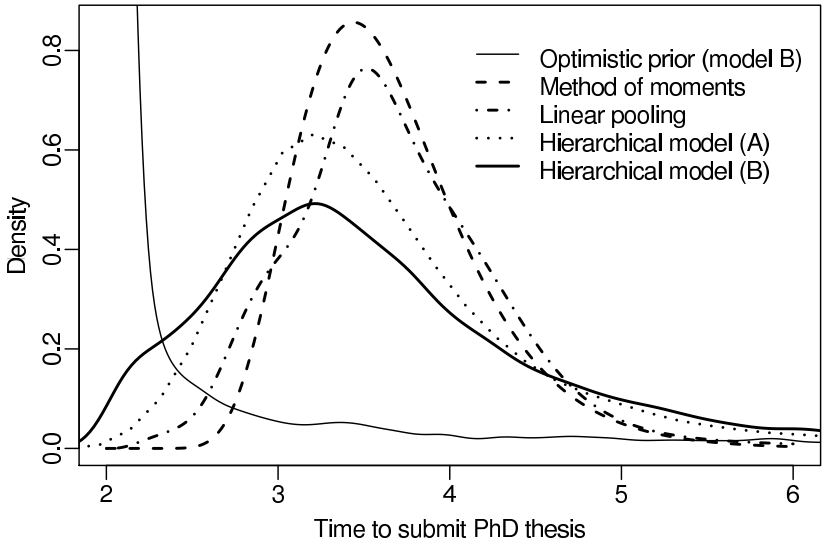

Figure 5: Marginal elicited prior predictive densities of X based on: pooled (mixture model) $(--)$, plug-in $(-\cdot)$, hierarchical approaches, method A $(\cdots)$ and method B (thick solid line) and prior predictive density (thin solid line).

We group the experts depending on their domain of interest and of their formation, an important consideration for their estimation of PhD thesis submission times. One group is formed of applied statisticians (3 individuals), another group is formed of more theoretical mathematicians (4 individuals), a third group is formed of computational mathematicians (2 individuals). Results for the marginal elicited prior predictive distributions of the time to submission are presented in Figure 5. Here it is evident that although the methods-of-moment approach (plug-in) provides a consensus opinion, it overstates the confidence in that opinion, by not addressing variability across and within experts. The pooled estimate (mixture model) focuses on diversity of opinions at the expense of consensus, and also does not adjust for within-expert variation, nor for dependence between experts. In contrast, the hierarchical approaches (methods A and B) distribute the weight of expert opinion more widely across potential submission times than the pooling or method-of-moments approaches. Consensus is concentrated on a mode of 3 years (Method B) or 3.12 years (Method A), much lower than the modal estimate of approximately 3.5 years provided by the other methods. However the weight of expert opinion on the mode is much lower, indicating that there is a wider possibility of submission times away from that most commonly achieved. Interestingly, the expected

submission time is fairly similar across all methods (all means lie between 3.55 and 3.72 years), regardless of the shift in the weight of expert opinion for shorter and longer submission times.

Following the hierarchical model (method B) results suggest that the administration should be ready for the majority of students to submit around the 3 year (rather than 3.5 year) mark, however a fairly large (rather than small) minority takes longer than 4.5 years to submit (about 17%). In addition, the administration should be ready to accept a non-negligible (rather than negligible) proportion of theses to be submitted within 2.3-2.7 years (9%). This suggests that it may be important to account for covariates responsible for shorter or longer submission times. From a more theoretical viewpoint, we comment that the hierarchical models provide a skewed consensus distribution, whilst accounting for within expert as well as between expert variation. This contrasts with the more symmetric consensus distributions encoded using the other methods, which have ignored within-expert variation.

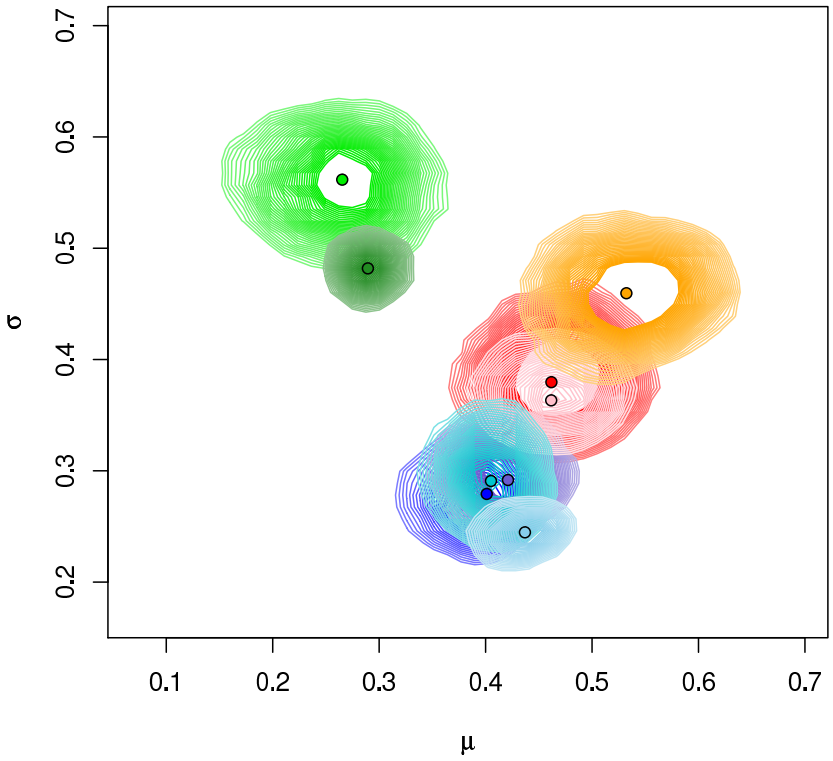

Figure 6: Contour plots of the individual prior distributions on (μ_{ij}, σ_{ij}).

It is interesting to note that the hierarchical approaches lead to a wider elicited prior distribution on X and that it is shifted to the left compared to the other two methods, taking into account the smaller group of more mathematical experts. Note that this method still allows for the individual experts' prior distributions, since we can recover

them from the MCMC algorithm. Figure 6 displays such distributions, corresponding to the hierarchical model using method B. The groups can be easily recognized, forming three different clusters.

4 Discussion

4.1 Applications

In the examples we have considered in this paper, some practically important differences appear in inferences, when compared to more standard methods for combining expert judgments. In the PhD case study, compared to other methods, the hierarchical approach to combining opinions led to a much lower typical thesis submission time, but a greater minority with shorter or longer thesis submission times. In the dose-response case study, the hierarchical approach permitted the possibility that the dose could be of lower efficacy compared to the plug-in approach, but provided smoother estimates of efficacy for mid-range probabilities of mortality. These differences between approaches were evident even though elicitation was based on a small number of parameters; we estimate these differences to be further magnified under higher dimensional models. We believe that the approach described in this paper has potential even for larger dimensional setups.

For a very limited number of experts (only five in the dose-response case study), method A provided an interesting combined prior distribution, acting as a compromise between the plug-in and mixture approaches and being interpreted as a posterior distribution given the elicited data. Interestingly the fully Bayesian approach (method B) also leads to very reasonable priors, at least in the particular case of the PhD example considered here. Hence, even with a few elicited quantities per expert the information is good enough to compensate for the complexity of the hierarchical model.

More generally this method can underpin aggregation of expert assessments in three broad contexts— the decision maker (DM) problem, the group decision problem, and the textbook problem (French 1985, 2011). These two case studies can be viewed as exemplars of the DM problem. The approach could also contribute a quantitative component to group deliberations, to formulate a decision or estimate model parameters (the group decision problem), or else provide a synthesis of available knowledge (the textbook problem). This is particularly important in fields, such as ecology, which rely heavily on expert panels (Low Choy et al. 2009; Perera et al. 2011). We note that any quantitative exploration and aggregation of expert opinions also benefits greatly from a qualitative component, not least to overcome the inherent "impossibility" that group decision-making can achieve a rational, democratic decision (French 2007).

4.2 Managing variation

The most popular methods of aggregation focus on one or two main avenues for managing variation during the elicitation process. Pooling is primarily focused on retaining

diversity among experts, on the entire encoded distribution. Pooling with weights, for instance via Cooke's method, places a strong emphasis on calibration. A hierarchical Bayesian model provides many avenues for managing variation during the elicitation process, which we summarize here.

Encoding variability as well as means. We have proposed a hierarchical model which allows us to combine information elicited from experts, not only on means (of PhD submission times, or of important doses) but also on variability. This was key to practical concerns about the tails of the distributions of interest in each example. The risk of mortality at low doses was of considerable concern in the food risk assessment example. Similarly the percentage of PhD students submitting earlier or later "than usual" had considerable practical implications. The model proposed here provides information on the consensus about both the means and the variability. This contrasts with other applications of random effects models to elicited information, where the hierarchical model applies solely to the mean (Lipscomb et al. 1998; Lin and Bier 2008).

Eliciting different summary statistics across experts. A challenge posed in French (2008) was "I am more comfortable with uncertainty judgments that lead to probability elicitation; others are more comfortable with moment judgments. 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?" In this paper we provide a means for combining different types of information elicited from the same or potentially different experts. The modelling framework combines information on different aspects of the distribution (here Ps and Qs). This is achieved by exploiting the conditional independence of elicited data given the expert's underlying conceptual model, here encapsulated by γ_{ij}, through an error model in the general form of (3). In fact, the likelihood for the elicited data models the difference between each elicited datum and the theoretical value under γ_{ij}, rather than modelling the elicited data themselves D_{elicit}. Thus we may consider these errors conditionally independent given the conceptual model. In this paper we have only combined quantiles and probabilities, each expert considering both types but potentially associated with different nominal values. However the same methodology could be applied to other quantities, as soon as a likelihood is constructed using an error model for the elicited data in the form (3).

Calibrating experts. Another important source of variation is mis-calibration (Cooke and Goossens 2008). In our context this means that the error model (3) might be miscalibrated by a bad choice of the variances v_{ijt}. In our examples we have combined prior information on uncertainty such as discretization to define the values q^*_{ij} appearing in (16) with the experts' evaluations of their own uncertainty c_{ijt}. We are aware that the latter is amenable to criticism, since the experts can be poor judges of their own ability to make judgments, see for instance Burgman et al. (2011). Where gold standards exist, scoring methods can be used to calibrate experts and define more reliable values for c_{ijt} or q^*_{ij}. In our case studies however, no empirical data yet exists which could be used to calibrate the expert judgments and this is common in various fields of applications such

as ecology where gold standards are difficult to obtain (e.g. Kuhnert et al. 2010; Martin et al. 2012), in contrast to the long-term databases that are emerging to monitor experts on safety and reliability (Cooke and Goossens 2008). Nevertheless, this hierarchical model can accept expert weights through the quantities q_{ij}^* for instance or through their individual distribution $g(.|\gamma_j, b_{ij})$ on γ_{ij}, depending on the form of calibration available. Currently the most popular method for developing these expert weights is Cooke's method (as summarized in Cooke and Goossens 2008). Expert contributions to the pooled assessment are weighted by their performance on estimating quantiles of *seed* variables. Focusing on elicited quantiles aligns with natural propensities to estimate categories better than quantities (e.g. Kynn 2008), and also elegantly leads to a χ^2 test to compare elicited with empirical quantiles. This takes advantage of the frequentist property that in the long run we expect that, for an accurate expert, the elicited values ought to accurately reflect the true values across many elicitation topics (Bayarri and Berger 2004). Nevertheless, this suffers from the limitations imposed by then using the p-values of this χ^2 test (initially designed to help experts understand their own short-comings) to define weights for each expert in the pooled distribution (Clemen 2008).

4.3 Mathematical (encoding) issues

Eliciting natural counts. The dose-response case study shares the same structure as the PhD case study in that we are asking experts about summary statistics (quantiles and cumulative probabilities) of the possible response (the number of mortalities among n mice at specific dose), rather than focussing on the parameters governing the response, which are not interpretable. This approach was chosen to be consistent with a recent review of cognitive biases in elicitation (see Kynn 2008, and references therein), which has confirmed that elicitation based on counts is less prone to cognitive errors than elicitation of probabilities. However the method we use here, of deliberately structuring the elicitation model to relate observable counts to the underlying probability, is quite new; typically a probability is imputed from a count, without accounting for the sampling issues inherent in counts (Low-Choy et al. 2010).

Different prior formulations. In the examples we have considered we have assumed an independent prior on the mean and variance (μ, σ): $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma)$. For other applications, it may be fruitful to instead assume a conditionally conjugate prior which explicitly models dependence between the mean and variance via $p(\mu, \sigma) = p(\mu|\sigma)p(\sigma)$; this is still possible in our framework.

Parametric vs Non-parametric encoding. This approach also combines different types of elicited summary statistics (here quantiles and cumulative probabilities), which is facilitated by the imposition of a parametric model. In these examples, the parametric assumptions seemed feasible. In other cases, a non-parametric approach may be more appropriate. Multiple types of elicited data could be incorporated by extending existing Bayesian methods for non-parametric encoding of information, such as the roulette

approach (Oakley and O'Hagan 2007), the use of the Dirichlet distribution (West 1988), or a Dirichlet process (Merrick 2008).

Sequential versus Simultaneous Encoding. Importantly the γ_{ij} deduced from both Ps and Qs can be different, which is why we propose both a sequential Bayesian approach that treats each type of judgment in sequence (Method A), and a fully Bayesian approach that models both types of judgment.

4.4 Summary

In conclusion, the approach we describe in the paper is quite generic in the sense that it does not depend on the particular distributions involved in the elicitation process, nor does it depend on the questions that are asked of the experts. In particular the experts could be asked questions of a very different nature, without changing the overall hierarchical approach to combining expert elicitations. It does however require some extra information on the nature and the sources of their knowledge to form the different groups. However this information is usually asked of the experts, since it helps them remember all (or at least most) of their knowledge on the subject (Fisher et al. 2012). To our minds, one of the substantial advantages of such a method is that it does not suffer from the various paradoxes that the other (ad-hoc) approaches might suffer, since it is a fully probabilistic and coherent approach.

A critical aspect of our method compared to mixture model or plug-in approaches is that it is computationally more demanding, in order to facilitate the hierarchical combination of opinions whilst accounting for within-expert error. Nevertheless the model for the PhD example can be implemented in WinBUGS (Spiegelhalter et al. 2003), so is widely accessible beyond the statistical research community.

References

Albert, I., Grenier, E., Denis, J.-B., and Rousseau, J. (2008). "Quantitative risk assessment from farm to fork and beyond: a global Bayesian approach concerning foodborne diseases." *Risk Analysis*, 28: 557–571. 503

Bayarri, M. J. and Berger, J. O. (2004). "The interplay of Bayesian and Frequentist Analysis." *Statistical Science*, 19: 58–80. 527

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). "A new perspective on priors for generalized linear models." *Journal of the American Statistical Association*, 91(436): 1450–1460. 505

Berger, J. (2006). "The Case for Objective Bayesian Analysis." *Bayesian Analysis*, 1: 385–402. 504

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): e22998: 1–7. 504, 526

Clemen, R. T. (2008). "Comment on Cooke's classical method." *Reliability Engineering and System Safety*, 93: 760–765. 527

Cooke, R. M. and Goossens, L. L. H. J. (2008). "TU Delft expert judgment data base." *Reliability Engineering and System Safety*, 93: 657–674. 504, 515, 526, 527

Denham, R. and Mengersen, K. (2007). "Geographically Assisted Elicitation of Expert Opinion for Regression Models." *Bayesian Analysis*, 2(1): 99–136. 505

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K., and Caley, M. J. (2012). "A software tool for elicitation of expert knowledge about species richness or similar counts." *Environmental Modelling and Software*, 30. 528

French, S. (1985). "Group consensus probability distributions: a critical survey." In Berger, J., Bernardo, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 2*, 183–201. North-Holland, Amsterdam: Oxford University Press. 505, 525

— (2007). "Web-enabled strategic GDSS, e-democracy and Arrow's theorem: A Bayesian perspective." *Decision Support Systems*, 43: 1476–1484. 525

— (2008). "Comments by Prof. French." *Reliability Engineering and System Safety*, 93: 766–768. 526

— (2011). "Aggregating expert judgement." *Revista de la real academia de ciencias exactas, fisicas y naturales. Serie A Matematicas*, 105: 181–206. 504, 505, 525

Gelfand, A., Mallick, B., and Dey, D. (1995). "Modelling expert opinion arising as a partial probabilistic specification." *Journal of the American Statististical Association*, 90: 598–604. 506

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–533. 512

Genest, C. and Zidek, J. V. (1986). "Combining probability distributions. A critique and annoted bibliography." *Statistical Science*, 1: 114–148. 504, 509

Gill, J. and Walker, L. D. (2005). "Elicited Priors for Bayesian Model Specifications in Political Science Research." *The Journal of Politics*, 67(3): 841–872. 504

Haas, C. N., Rose, J. B., and Gerba, C. P. (1999). *Quantitative microbial risk assessment*. Wiley. 514

James, A., Low Choy, S., and Mengersen, K. (2010). "Elicitator: An Expert Elicitation Tool for Regression in Ecology." *Environmental Modelling & Software*, 25(1): 129–145. 505

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). "Interactive elicitation of opinion for a normal linear model." *Journal of the American Statististical Association*, 75: 845–854. 505, 514, 520

Kuhnert, P., Martin, T. G., and Griffiths, S. P. (2010). "A guide to eliciting and using expert knowledge in Bayesian ecological models." *Ecology Letters*, 13: 900–914. 527

Kynn, M. (2005). "Eliciting expert opinion for a Bayesian logistic regression model in natural resources." PhD thesis, School of Mathematical Sciences, Faculty of Science, Queensland University of Technology.
URL http://adt.library.qut.edu.au/adt-qut/public/adt-QUT20050830.084943 505

— (2008). "The ' heuristics and biases' bias in expert elicitation." *Journal of the Royal Statistical Society, Series A*, 171: 239–264. 504, 527

Lin, S.-W. and Bier, V. M. (2008). "A study of expert overconfidence (with discussion)." *Reliability Engineering and System Safety*, 93: 711–721. 508, 509, 526

Lindley, D. V. (1983). "Reconciliation of probability distributions." *Operations Research*, 31: 866–880. 505, 506, 508, 509

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). "On the reconciliation of probability assessments (with discussion)." *Journal of the Royal Statistical Society A*, 142: 146–180. 506

Lipscomb, J., Parmigiani, G., and Hasselblad, V. (1998). "Combining expert judgment by hierarchical modeling: an application to physician staffing." *Management Science*, 44: 149–161. 504, 508, 526

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. 504

Low-Choy, S., Mengersen, K., and Rousseau, J. (2008). "Encoding Expert Opinion on Skewed Non-Negative Distributions." *Journal of Applied Probability and Statistics*, 3: 1– 21. 510

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. 509, 514, 520, 521, 527

Low Choy, S., O'Leary, R., and Mengersen, K. (2009). "Elicitation by Design for Ecology: using expert opinion to inform priors for Bayesian statistical models." *Ecology*, 90: 265–277. 525

Martin, T., Burgman, M., Fidler, F., Kuhnert, P., Low-Choy, S., MacBride, M., and Mengersen, K. (2012). "Elicitation of Expert Knowledge in Conservation Biology." *Conservation Biology*, 26(1): 29–38. 527

Merrick, J. R. W. (2008). "Getting the Right Mix of Experts." *Decision Analysis*, 5(1): 43–52. 528

Oakley, J. E. and O'Hagan, A. (2007). "Uncertainty in prior elicitations: a nonparametric approach." *Biometrika*, 94(2): 427–441. 528

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. 504, 506, 523

Perera, A. H., Drew, C. A., and Johnson, C. J. (eds.) (2011). *Expert Knowledge and Its Applications in Landscape Ecology*. Springer, NY. 525

Roback, P. J. and Givens, G. H. (2001). "Supra-Bayesian pooling of priors linked by a deterministic simulation model." *Communications in Statistics - Simulation and Computation*, 30(3): 447–476. 505

Spetzler, C. S. and Staël von Holstein, C.-A. S. (1975). "Probability encoding in decision analysis." *Management Science*, 22(3): 340–358. 504

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). "WinBUGS version 1.4 user manual." Technical report, MRC Biostatistics Unit, Cambridge. 528

West, M. (1988). "Modelling Expert Opinion." In Bernardo, J., Degroot, M., Lindley, D., and Smith, A. (eds.), *Bayesian Statistics 3*, 493–508. Clarendon Press. 505, 528

Winkler, R. L. (1968). "The consensus of subjective probability distributions." *Management Science*, 15: 361–375. 505