Albert, Isabelle, Donnet, Sophie, Guihenneuc, Chantal, Low Choy, Samantha, Mengersen, Kerrie, & Rousseau, Judith (2010) *Combining expert opinions in prior elicitation.* Biostatistics. (Submitted (not yet accepted for publication))

Combining expert opinions in prior elicitation

Journal:	*Biostatistics*
Manuscript ID:	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Donnet, Sophie; Université Paris Dauphine, CEREMADE Mengersen, Kerrie; QUT, School of mathematical sciences Albert, Isabelle; INRA, Metarisk Guihenneuc, Chantal; Université Paris 5, MAP5 Low Choy, Samantha; QUT, School of mathematical sciences Rousseau, Judith; Université Paris Dauphine, CEREMADE; ENSAE-CREST, Laboratoire de Statistiques
Keywords:	Bayesian analysis, Hierarchical models, Statistical methods in Toxicology

Biostatistics (2007), **8**, 1, *pp.* 1 33
doi:10.1093/biostatistics/kxl000
Advance Access publication on May 15, 2006

Combining expert opinions in prior elicitation

ISABELLE ALBERT
INRA, UR1204, Métarisk, AgroParisTech
16, rue Claude Bernard, F75231 Paris, France
isabelle.albert@paris.inra.fr

SOPHIE DONNET*
CEREMADE, Université Paris Dauphine, 75 016 Paris, France
donnet@ceremade.dauphine.fr

CHANTAL GUIHENNEUC
MAP5, Université Paris Descartes, 75 006 Paris
chantal.guihenneuc@univ-paris5.fr

SAMANTHA LOW CHOY
School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia
s.lowchoy@qut.edu.au

KERRIE MENGERSEN
School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia
k.mengersen@qut.edu.au

JUDITH ROUSSEAU
CEREMADE, Université Paris Dauphine, 75 016 Paris, France
rousseau@ceremade.dauphine.fr

SUMMARY

A method of eliciting prior distributions for Bayesian models using expert knowledge is proposed. Elicitation is a widely studied problem, from a psychological perspective as well as from a statistical perspective. Here, we are interested in combining opinions from more than one expert using an explicitly model-based approach so that we may account for various sources of variation affecting elicited expert opinions. We

*To whom correspondence should be addressed.

use a hierarchical model to achieve this. We apply this approach to two problems. The first problem involves a food risk assessment problem involving modelling dose-response for *Listeria monocytogenes* contamination of mice. The second concerns the time taken by PhD students to submit their thesis in a particular school.

Keywords: Bayesian statistics; Expert opinions; Hierarchical model; Random effects; Prior elicitation; 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) [1]. A complex model is 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 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. The construction of such informative priors using expert opinions is a delicate problem, because the human mind finds it difficult to quantify qualitative knowledge, see for instance O'Hagan et al. (2006) [10] for a good review on the subject.

To simplify the presentation, consider a sampling model with observation X distributed from a probability distribution P_θ, with unknown parameter θ. In a Bayesian approach θ is considered as random and a probability distribution π, called the prior probability, is considered on θ. 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 γ and in many situations it is desirable to combine these different priors into a single "consensus" prior estimate of θ. There are various methods proposed in the literature to achieve this, although most are not entirely satisfactory for applications such as the case studies considered here. The most common methods, referred to as linear or logarithmic "pooling", define the overall prior as an additive or multiplicative mixture, respectively, of the individual priors see Cooke and Goossens, 2004 [2], for instance. However in these cases, it is hard to account for the various sources of variation affecting the elicited expert opinion. In our approach we propose a Bayesian hierarchical (random effects) model that reflects the elicitation process involving various experts: bias and precision of individual experts as well as consensus and diversity in opinion among experts, both overall and within known groups. We treat the elicited information as data, in the spirit of the other Bayesian approaches to pooling expert opinions (see for instance Winkler, 1968 [17]; Lindley 1983 [7]; West 1983 [16]; Genest and Zidek, 1986 [4]). Unlike other approaches (such as pooling), Bayesian updating provides a model-based framework for capturing elicited information.

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 to thesis submission for PhD students in applied mathematics in an Australian university. The first case study is simpler mathematically since the underlying distributions are log-Normal, but more complex since three levels of variation are considered—between schools-of-thought, between experts within schools of thought, and intra-expert elicitation error. The second case study is more complex mathematically due to inversions required to support indirect encoding of a probit regression, but only two levels of variation are considered—between and within experts.

In Section 2 we describe the approach and the hierarchical model we consider. In Section 3 we consider the the dose-response and the PhD example and Section 5 contains some conclusions.

2. METHOD

In this Section we describe the generic approach. Let X be a possible vector of observations from a distribution P_θ, $\theta \in \Theta$, with density $f(X|\theta)$.

The aim is to construct an informative prior probability distribution on θ based on expert knowledge. Such a prior for θ can be specified as the posterior from a Bayesian analysis that begins with a vague prior π_0 and treats elicited expert knowledge as data (e.g. Lindley, 1983 [7]), from the following scheme:

$$\pi(\theta|\boldsymbol{D}_{\text{elicit}}) = \int \pi(\theta|\gamma)\pi(\gamma|\boldsymbol{D}_{\text{elicit}})d\gamma \qquad (2.1)$$

where $\boldsymbol{D}_{\text{elicit}}$ are the elicited data and $\pi(\gamma|\boldsymbol{D}_{\text{elicit}}) \propto f(\boldsymbol{D}_{\text{elicit}}|\gamma)\pi_0(\gamma)$. In this formula, $\pi(\theta|\gamma)$ belongs to a parametric class $\{\pi_\gamma, \gamma \in \Gamma \subset \mathbb{R}^p\}$.

2.1 *About the experts and the elicited data*

In the following, we assume that we interview N experts who can be grouped into J homogeneity classes (of respective sizes N_j) , corresponding to similar background or similar schools of thought for instance. In each class j, (i, j) denotes the i-th expert. To each expert (i, j) corresponds an unknown hyperparameter γ_{ij} resulting into their own prior distribution $\pi(\theta|\gamma_{ij})$, which reflects their conceptual model for the distribution of X. To estimate this hyperparameter, we interview each expert (i, j) and encode their knowledge. A more complete view of their knowledge is obtained by asking two styles of questions corresponding to different approaches to statistically encoding their information into probability distributions (see O'Hagan et al, 2006 [10] for a review of encoding techniques).

These two styles can, for instance, correspond to (i) eliciting quantiles for specified cumulative probabilities (also known as fractile estimation) and (ii) eliciting cumulative probabilities for specified quantiles (also known as interval estimation). Thus, in the PhD case study (Section 3), 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?". Hereafter, for the sake of presentation, we take these two types of elicited data but other quantities could be elicited without changing the overall method. In the literature, these two approaches –(i) and (ii)– have been used iteratively within a feedback cycle to elicit opinions (e.g. Low Choy et al, in press [14]). 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.

In the following $\boldsymbol{Q}_{\text{elicit}} = (Q_{ijt})$ are elicited quantiles of the distribution of interest $f(X|\theta)$, and correspond to specified cumulative probabilities p_{ijt}. Similarly, cumulative probabilities $\boldsymbol{P}_{\text{elicit}} = (P_{ij\ell})$ for this same distribution of X are elicited at specified quantiles $q_{ij\ell}$. $\boldsymbol{D}_{\text{elicit}} = (\boldsymbol{P}_{\text{elicit}}, \boldsymbol{Q}_{\text{elicit}}) = (D_{ijt})$ denote the complete elicited data. Each block of answers ($\boldsymbol{P}_{\text{elicit}}$ or $\boldsymbol{Q}_{\text{elicit}}$) can be used to provide separate sources to estimate the parameter of interest γ_{ij} (see below for more details).

For each question, the experts also provide a measure of uncertainty in their answer. More precisely, to each answer corresponds a number $c_{ijt} \in (0, 1)$ quantifying the expert's confidence in their response. This information allows to build a measurement error model to quantify their individual accuracy, adopting similar ideas to earlier approaches to elicitation modelling (Lindley et al., 1979) [8]; Lindley,1983) [7]) as described below.

2.2 *A model of error for elicited data*

The individual inaccuracies for each expert are modelled via the following error model:

$$\eta(D_{ijt}) = \eta(d_t(\gamma_{ij})) + \epsilon_{ijt}, \tag{2.2}$$

where $d_t(\gamma_{ij})$ 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 $\theta \sim \pi_{\gamma_{ij}}$. For instance the quantile $q_t(\gamma_{ij})$ and the probability $p_\ell(\gamma_{ij})$ respectively satisfy:

$$\int P(X \leqslant q_t(\gamma_{ij})|\theta) d\pi(\theta|\gamma_{ij}) = p_{ijt} \quad \text{and} \quad \int P(X \leqslant q_{ij\ell}|\theta) d\pi(\theta|\gamma_{ij}) = p_\ell(\gamma_{ij}). \tag{2.3}$$

η is a link function – such as the identity or the probit functions– depending on the situation. The ϵ_{ijt} are independent and have a known distribution h_{ijt} constructed using their given measure of uncertainty c_{ijt}, together with measures of individual coherence and precision considered by the assessor, based for instance on the training of the expert or on previous expertises (this point is detailed in the examples of Section 3). Thus, the influence of the answers of the expert (i, j) is assessed via the error density h_{ijt} : an expert whose own measures of uncertainty are large typically would have error densities h_{ijt} with large variance inducing a weak influence of Q_{ijt} on the likelihood.

Remark : In practice and in our examples, only a few quantiles (say $3 \leqslant |T| \leqslant 5$) are generally elicited from each expert, so that a sensible choice of the distribution of ϵ_{ijt} will ensure the error model provides a coherent set of quantiles that conforms to the order imposed by T. Should more quantiles be elicited, or the error in quantiles cause incoherent overlap between quantiles, then a more complex model using a Dirichlet distribution could be applied as suggested by West, 1983 [16].

2.3 *Combining the experts opinions: a hierarchical model*

The key issue is to derive a final unique distribution $\pi(\theta|\boldsymbol{D}_{\text{elicit}})$ taking into account the fact that the elicitations vary among the N experts. 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 et al, 1983) [7]. 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 therefore consider an hierarchical formulation of a random effects model (e.g. Lipscomb, Parmigiani and Hasselblad, 1998) [9], to represent variation between, and within individual, expert opinions. Firstly, variation between experts reflects the level of consistency or diversity amongst expert opinions, and can be as important as the “average” opinion for informing decision-making, particularly where decisions may have high impact. Secondly, variation within experts reflects both elicitation (measurement) error, being the expert’s difficulty in accurately quantifying their knowledge, as well as incoherence, being logical inconsistencies in their underlying knowledge. To help address incoherence we elicit expert opinion using two different constructs of their knowledge: quantiles and cumulative probabilities. Elicitation error of individual experts is modelled at the within-expert level by the error model (2.2). This model-based approach seeks to quantify these two variance components, which could be used to inform design of elicitation. Indeed this is similar in aim to the approach of Osherson and Vardi (2006) [11] who search for the "closest" though coherent set of probabilities that match the expert opinions. To achieve this, these authors apply a simulated annealing algorithm to minimize logical incoherence at these two levels (i.e. within each experts and between experts), whilst accounting for the first variance component between experts. Our approach is also useful when a consensus expert model is desired, since it explicitly combines potentially disparate expert elicitations in order to specify the prior distribution of interest. Concretely, we propose the following hierarchical model:

$$\begin{array}{rcl} \gamma_{ij} & \overset{\text{i.i.d}}{\sim} & g(\cdot|\gamma_j, b_j), \quad \forall i = 1, \dots, N_j, \\ \gamma_j & \overset{\text{i.i.d}}{\sim} & g(\cdot|\gamma, b), \quad \forall j = 1, \dots, J, \\ \gamma & \sim & \pi_0 \end{array} \tag{2.4}$$

where π_0 is typically some weakly informative 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 hierarchical modelling. Thus γ can be understood as the *true parameter* of model (2.1), or more realistically as the parameter representing the agreement of experts. In model (2.4), the γ_j's are location parameters and so is γ. The hyperparameters b_j, b are typically dispersion parameters.

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*

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

In method A, we propose to split the elicited data $\boldsymbol{D}_{\text{elicit}}$ into two natural blocks namely $\boldsymbol{P}_{\text{elicit}}$ and $\boldsymbol{Q}_{\text{elicit}}$. In a first step, from the P-elicitation data $\boldsymbol{P}_{\text{elicit}}$, we estimate the hyperparameters $(b_j)_{j=1\ldots J}$ and b. In a second step, we plug in these dispersion hyperparameter estimators into the likelihood of elicited data $\boldsymbol{Q}_{\text{elicit}}$, as described in (2.7) and derive the posterior distribution $\pi(\theta,\gamma|\boldsymbol{D}_{\text{elicit}})$ using a MCMC algorithm. This is similar to an empirical Bayesian procedure, except that there is no double-use of the data, different data is used in estimating the prior and the likelihood. More precisely,

- from $\boldsymbol{P}_{\text{elicit}}$ we derive preliminary estimators of γ_{ij} by minimizing the least squares objective (Low Choy et al, 2008) [15]:

$$\hat{\gamma}_{ij} = \text{argmin}_{\gamma} \sum_{\ell \in L} \left[P_{ij\ell} - p_\ell(\gamma_{ij})\right]^2 \tag{2.5}$$

  Estimators of $(b_j)_{j=1,\ldots,J}$ 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 3). We denote $(\hat{b}_j)_{j=1\ldots J}$ and $\hat{b}$ the obtained estimates.

- Using (2.1), (2.2) and (2.4) and plugging in the estimated dispersion hyperparameters, we deduce the likelihood of elicited data $\boldsymbol{Q}_{\text{elicit}}$:

$$f\left(\boldsymbol{Q}_{\text{elicit}}|\gamma,(\hat{b}_j)_{j=1\ldots J},\hat{b}\right) = \int \prod_{ijt} h_{ijt}(\eta(Q_{ijt}) - \eta(q_t(\gamma_{ij}))) \prod_{ij} g(\gamma_{ij}|\gamma_j,\hat{b}_j) \prod_j g(\gamma_j|\gamma,\hat{b}) d\gamma_j d\gamma_{ij} \tag{2.6}$$

- Finally using:

$$\pi(\theta|\boldsymbol{D}_{\text{elicit}}) \propto \int \pi(\theta|\gamma) f\left(\boldsymbol{Q}_{\text{elicit}}|\gamma,(\hat{b}_j)_{j=1\ldots J},\hat{b}\right)\pi_0(\gamma)d\gamma \tag{2.7}$$

we generate Markov realizations of (γ, θ) under the posterior distribution $\pi(\theta, \gamma|\boldsymbol{D}_{\text{elicit}})$ through Markov Chain Monte Carlo.

By splitting the elicited data into two parts used respectively for the estimation of the hyperparameters and for the computation of the likelihood, we avoid the double use of the elicited data. This method is useful in situations where the elicited cumulative probabilities are obtained in a different phase of elicitation. 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 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 $\pi(\theta|\boldsymbol{D}_{\text{elicit}})$ in Practice.

The second method of utilizing both sources of elicitation data specifies weakly informative priors $\pi_0(b), \pi_0(b_j), \pi_0(\gamma), \pi_0(\gamma_j)$ for the hyperparameters, and defines a Bayesian elicitation model for both $\boldsymbol{Q}_{\text{elicit}}$ and $\boldsymbol{P}_{\text{elicit}}$:

$$\pi(\theta|\boldsymbol{D}_{\text{elicit}}) \propto \int_{\gamma,c,d} \pi(\theta|\gamma) f\left(\boldsymbol{Q}_{\text{elicit}}, \boldsymbol{P}_{\text{elicit}}|\gamma, c, d\right) \pi_0(\gamma)\pi_0(c,d) d\gamma db \tag{2.8}$$

where as before $c = (c_{ijt}; i = 1, \ldots, N_j, j = 1, \ldots, J, t \in T)$ represent imprecision in Q-elicitation and we also introduce $d = (d_{ij\ell}; i = 1, \ldots, N_j, j = 1, \ldots, J, \ell \in L)$ to represent imprecision in P-elicitation. These two likelihood contributions are separable, since the Q-elicitations and P-elicitations provided by each expert can be considered independent when we condition on the their underlying conceptual model γ_{ij}, this leads to the following likelihood for the elicited data:

$$f(\boldsymbol{D}_{\text{elicit},ij}|\gamma_{ij}, c_{ij\cdot}, d_{ij}) = f(\boldsymbol{P}_{\text{elicit},ij}|\gamma_{ij}, d_{ij}) f(\boldsymbol{Q}_{\text{elicit},ij}|\gamma_{ij}, c_{ij\cdot}) \tag{2.9}$$

Thus the likelihood of elicited data $\boldsymbol{Q}_{\text{elicit}}$ and defined in (2.6) is modified to consider elicited data incorporating $\boldsymbol{P}_{\text{elicit}}$ as well as $\boldsymbol{Q}_{\text{elicit}}$, simply by including an additional factor $\prod_{ijt} f_{ij\ell}(P_{ij\ell}|p_\ell(\gamma_{ij}), d_{ij\ell})$. Hence we replace the Q-elicitation likelihood $f\left(\boldsymbol{Q}_{\text{elicit}}|\gamma,\ldots\right)$ by the full Q- and P-elicitation likelihoods $f\left(\boldsymbol{Q}_{\text{elicit}}, \boldsymbol{P}_{\text{elicit}}|\gamma,\ldots\right)$ in the posterior distribution defined in (2.7).

Method B gives equal weight to the P-elicitations and Q-elicitations, and is also fully Bayesian, so in this latter regard is more satisfying. 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 the advantage of simplifying the computation. We also believe that by anchoring, Method A is more adapted to a smaller sets of experts.

2.5 *Remarks and modifications of the method*

This methodology contrasts with a 'hybrid' approach (e.g. Garthwaite and O'Hagan, 2000) [3] where questioning oscillates between both styles of question, providing feedback from an alternative viewpoint, to improve estimates obtained, and iterate towards a single estimated parameter γ. This can be done prior to our analysis or, contrarywise, we can retain estimates obtained using each style. In the latter case, the order in which the styles of questions are delivered is important, due to the opportunity for feedback, as in the hybrid approach. This can be managed as a source of variation in the study, either explicitly estimated or controlled by randomisation. For example, in the PhD case study, the order of the styles of questions was randomly assigned to each expert. Finally, by introducing variability parameters and the measurement error model, we may model cognitive uncertainty. We also reduce the bias arising from the potential for experts to 'anchor' on estimates provided using the first approach encountered. Moreover elicitations obtained using different techniques and uncertainty measures can be interpreted as assessing slightly different perspectives on the expert's knowledge or alternatively as assessing only one version of the expert's underlying knowledge.

In the above formulation we have implicitly considered that 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 scenarii where for instance one of the groups is a priori known to be less reliable 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 happen for instance in a case where the groups correspond to different school of thoughts, say you have two groups corresponding two schools of thoughts, one corresponding to the majority of the population of experts and the other 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 the possible different reliability of this group is to assume a higher level parameter for its distribution, in which case γ_j would follow a distribution in the form $g(\cdot|\gamma, b'_j)$ with b'_j greater than b. Or, if the group is known to have systematically a bias of some order of magnitude we could consider a distribution in the form: $\gamma_j \sim g(\cdot|\gamma + \delta, b)$ 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 one just described. However in the following we focus on the proposal (2.4), assuming that in our examples the groups are not known to behave either much better or worse compared to one another.

An alternative parameterization has proved to be useful: $\gamma_{ij} = \gamma + \Delta_{ij}$ explicitly models bias Δ_{ij} in the expert's description of π. In equation (2.4) all instances of γ_{ij} can then be replaced with Δ_{ij}, γ_j with Δ_j, so that the Δ_j's are typically centered on 0 (unless the assessor decides otherwise).

To illustrate the generic approach we consider 2 examples based on the same formulation of model (2.4).

3. Examples

In this section, we detail a particular case of the hierarchical model (2.4) and discuss some technical points such as the model for hyperparameters $(b_j)_{j=1...J}$, and b and the error model. In a second step, we use this hierarchical model and our methodology on two examples. The first example is issued from 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 nature, the hierarchical structures of the models used for the combination of the different expert opinions follow a similar pattern, following either Method A or B detailed above.

3.1 *Description of a particular hierarchical model (2.4) used for both examples*

Suppose that the parameter of interest γ is composed of a mean μ and a variance σ^2: $\gamma = (\mu, \sigma^2)$. We build a hierarchical model on γ using the second formulation of the model discussed in Section 2.5. More precisely, we set $\mu_{ij} = \mu + \Delta^{\mu}_{ij}$ and $\sigma^2_{ij} = \sigma^2 \times \Delta^{\sigma}_{ij}$. In each group j, we set:

$$\Delta^{\mu}_{ij} \overset{\text{i.i.d}}{\sim} \Delta^{\mu}_{j} + \mathcal{N}(0, \tau_j) \quad , \quad \Delta^{\sigma}_{ij} \overset{\text{i.i.d}}{\sim} \Delta^{\sigma}_{j} \times \Gamma(\xi_j, \xi_j) \tag{3.10}$$

and the relations between groups are modelled by:

$$\begin{array}{ccc} \Delta^{\mu}_{j} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \tau) & , & \Delta^{\sigma}_{j} \overset{\text{i.i.d}}{\sim} \Gamma(\xi, \xi) \\ \mu \sim \mathcal{N}(\mu_0, V) & , & \sigma^2 \sim \Gamma(a, a/\sigma^2_0). \end{array} \tag{3.11}$$

The hyperparameters $(b_j)_{j=1...J} = (\tau_j, \xi_j)_{j=1...J}$, $b = (\tau, \xi)$ and $(\mu_0, \sigma^2_0, a, V)$ must be modelled carefully since their influence might be important, specially when the number of experts and elicited quantities are small, which is a common situation. As exposed previously under Method A, we may use the group of elicited data $\boldsymbol{P}_{\text{elicit}}$ to derive estimates $\hat{\mu}_{ij}, \hat{\sigma}^2_{ij}$ via least squares (see (2.5)). The solution of this equation is proper to each example and is detailed in Sections 3.4 and 3.3. Once such estimates have been obtained, we propose moment estimators for $(b_j)_{j=1\ldots J}$ and b. More precisely, since τ_j represents the variance of Δ^{μ}_{ij} in the group 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 $\{\hat{\mu}_j, j = 1, ..., J\}$. Similarly ξ_j^{-1} represents the variance of $\frac{\Delta^{\sigma}_{ij}}{\Delta^{\sigma}_{j}}$ so that a natural estimate is $\hat{\xi}_j^{-1} = \frac{1}{N_j-1}\sum_{i=1}^{N_j}\left(\frac{\hat{\sigma}^2_{ij}}{\hat{\sigma}^2_j} - 1\right)^2$ and ξ can be etimated using $\hat{\xi}^{-1} = \frac{1}{J-1}\sum_{j=1}^{J}\left(\frac{\hat{\sigma}^2_j}{\hat{\sigma}^2} - 1\right)^2$, where $\hat{\sigma}^2$ is the average of $\{\hat{\sigma}^2_j, j = 1, ..., J\}$. 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 propose an elicited prior distribution using a MCMC algorithm. The resulting approach can thus be understood as a Bayesian estimation of the prior distribution based on a part of elicited data. Note that we could also have considered all the elicited data together $(\boldsymbol{P}_{\text{elicit}}, \boldsymbol{Q}_{\text{elicit}})$ and used all of them both for the computation of the likelihood and for the computation of the estimates of hyperparameters $(\tau_j, \xi_j, \tau, \xi, a, V, \mu_0, \sigma_0^2)$. However this would have implied a strong double use of the elicited data,

The third alternative (Method B) is to use all elicited data $(\boldsymbol{P}_{\text{elicit}}, \boldsymbol{Q}_{\text{elicit}})$ for the computation of the elicitation likelihood and instead use noninformative priors for hyperparameters $(\tau_j, \xi_j, \tau, \xi, a, V, \mu_0, \sigma_0^2)$.

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

3.2 *Description of the likelihood : error model (2.2)*

In our examples, to make it simple, we consider Gaussian errors in the elicitation error model for quantiles (2.2):

$$\eta(Q_{ijt}) \sim \mathcal{N}(\eta(q_t(\gamma_{ij})), v_{ijt}) \tag{3.12}$$

where q_t satisfies formula (2.3) with $P(X|\gamma)$ specific to each example (detailed below) and η is a link function. In Method B, we may also consider Gaussian errors in the elicitation error model for probit-transformed cumulative probabilities (2.2):

$$\Phi(P_{ij\ell}) \sim \mathcal{N}(\Phi(p_\ell(\gamma_{ij})), \xi_{ij\ell}) \tag{3.13}$$

Similarly to v_{ijt} the variances $\xi_{ij\ell}$ are assigned weakly informative priors.

For either method, the variances v_{ijt} (and potentially $\xi_{ij\ell}$) 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. Some experts are amenable to providing measures of confidence c_{ijt}, for the elicited quantiles. Suppose that $c_{ijt} \in (0, 1)$ is elicited as the expert's degree of confidence in their answer, or the probability of being right, then c_{ijt} can be considered as a coverage probability of a confidence interval and

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

so that

$$\sqrt{v_{ijt}} = \frac{q^\star_{ij}}{\Phi^{-1}((1+c_{ijt})/2)}. \tag{3.14}$$

The reference value $q^\star_{ij}$ 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 $q^\star_{ij}$'s is illustrated in the two examples.

In some other cases the confidence 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 formula similar to before.

Remark: In the PhD example, η is the identity link and in the food risk example, η is the probit link since $Q_{ijt} \in [0,1]$.

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

3.3 *Dose-response example*

Contrary to the PhD 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 dose of some treatment affects a response, often mortality. In this example we model the dose-response curve for the contamination of BALB/c, C57 Black/6 or Swiss mice from *Listeria monocytogenes* EGD or EGDe by intravenous injection. Let X be the number of dead mice out of n mice exposed to a dose do. Then, the sampling model is, conditionally on the injected dose, a binomial model of parameter $p(do,\theta)$ where $p(do,\theta)$ is the probability for a mouse to die from a dose

do of Listeria depending of an unknown parameter θ. We consider an exponential dose-response model (see for instance Haas, Rose and Gerba, 1999, p. 264) [5]. In that case, we have:

$$X \sim \mathcal{B}\text{in}(n, p(do)) \quad \text{with} \quad p(do, \theta) = 1 - e^{-\theta do}, \quad \theta > 0. \tag{3.15}$$

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 $\gamma = (\mu, \sigma^2)$. θ being not observable, we ask questions on the proportions of dead mice in some specific experimental context. Each expert chooses a dose $do = do_{ij}$, which he finds easier to work with and is then asked questions about the probability of mortality, to help formulate the distribution of $p(do, \theta)$.

We suppose that the experts are issued from 2 groups of respective sizes N_1 and N_2. We first ask questions regarding the quantiles of $p(do, \theta)$. The answer given by the expert i of group j is denoted Q_{ijt} and $q_t(\mu_{ij}, \sigma_{ij})$ is the theoretical quantile. $q_t(\mu_{ij}, \sigma_{ij})$ verifies:

$$q_t(\mu_{ij}, \sigma_{ij}) = 1 - \exp\{-do_{ij} \times exp(\sigma_{ij}\Phi^{-1}(t) + \mu_{ij})\}.$$

The second set of questions concerns the probabilities $P[X_{10} \leqslant \ell | do_{ij}]$, where X_{10} is the number of dead mice out of 10 mice submitted to dose do_{ij} and $\ell \in L$. We denote by $P_{ij\ell}$ the answer given by the expert i of group j and by $p_\ell(\mu_{ij}, \sigma_{ij})$ the theoretical probability:

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

where $\varphi(.)$ is the density function of a standard Gaussian random variable.

To prove the robustness of our method and to illustrate its behaviour, we propose various simulated scenarii. On each simulated data set, we apply the methodology described in Section 2 and compare the elicited prior distribution obtained with this hierarchical approach to those obtained with standard methods namely the plugin and mixture methods. For these two last methods, the posterior distributions are respectively:

$$\log(\theta)|\boldsymbol{D}_{\text{elicit}} \sim \mathcal{N}\left(\frac{1}{N}\sum_{i,j}\tilde{\mu}_{ij}, \frac{1}{N}\sum_{i,j}\tilde{\sigma}^2_{ij}\right) \quad \text{and} \quad \log(\theta)|\boldsymbol{D}_{\text{elicit}} \sim \frac{1}{N}\sum_{i,j}\mathcal{N}\left(\tilde{\mu}_{ij}, \tilde{\sigma}^2_{ij}\right).$$

where $\tilde{\mu}_{ij}$ and $\tilde{\sigma}^2_{ij}$ are estimates of (μ_{ij}, σ_{ij}) minimizing $\sum_{t\in T}\left[Q_{ijt} - q_t(\mu_{ij}, \sigma_{ij})\right]^2 + \sum_{\ell\in L}\left[P_{ij\ell} - p_\ell(\mu_{ij}, \sigma_{ij})\right]^2$. 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 and more sophisticated versions of the plug-in or the mixture combination of experts exist in the literature, but merely to understand better how the hierarchical modelling stands in terms of consensus of experts.

3.3.*1* *Simulation study*

We now describe four simulated datasets and comment the results. In each dataset, the doses do 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 with $L = \{3, 8\}$ and quantiles with $T = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. We add an error term of variance $v_{ijt} = 0.1$ for all the experts and all the questions.

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 set:

$$\begin{array}{llll} N_1 = 5 & N_2 = 5 & & \\ \mu_1 = -2 & \mu_2 = -1.1, & \xi_1 = 100 & \xi_2 = 100, \\ \rho_1 = 1 & \rho_2 = 1 & \tau_1 = 0.01 & \tau_2 = 0.01 \\ \sigma = 1 & & & \end{array}$$

and we simulate the individual parameters (μ_{ij}) and (σ_{ij}) following:

$$\mu_{ij} \sim \mathcal{N}(\mu_j, \tau_j) \quad , \quad \rho_{ij} \sim \Gamma(\xi_j, \frac{\xi_j}{\rho_j}) \quad , \quad \sigma_{ij} = \sigma\rho_{ij}$$

The resulting elicited prior distribution are plotted in Figure 1. This standard dataset clearly illustrates the specific behaviour of our hierarchical method. On the one hand, the plugin method (solid line) proposes a

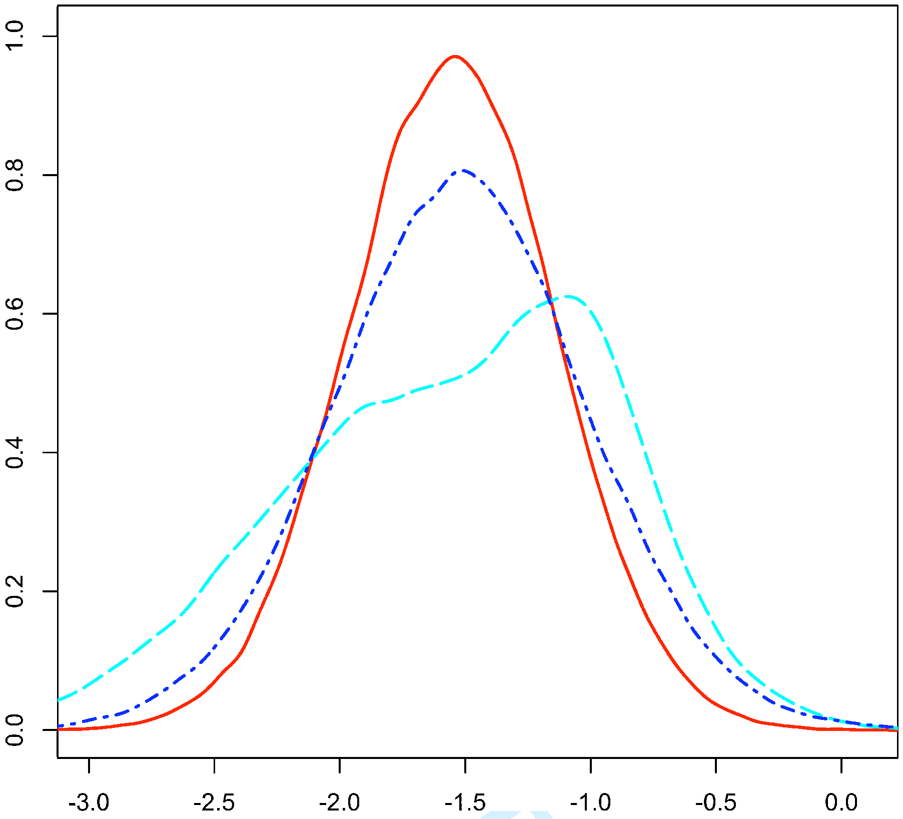

Fig. 1. *Dataset* 1. *Balanced case.* Comparison of methods for combination of experts using $p(\log\theta|\boldsymbol{D}_{\text{elicit}})$: mixture $(--)$, plugin (solid line), hierarchical $(-\cdot)$

forced 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 to consider the interactions between experts: the elicited prior distribution of $p(\log\theta|\boldsymbol{D}_{\text{elicit}})$ $(-\cdot)$ (which is thus a posterior) is smoother than the mixture one but has a wider support than the plugin posterior distribution.

Dataset 2**. Unbalanced groups of experts**: On this dataset, we suppose that the numbers of experts in the groups are strongly unbalanced. More precisely, we set:

$$\begin{array}{llll} N_1 = 10 & N_2 = 2 & & \\ \mu_1 = -2.5 & \mu_2 = -1, & \xi_1 = 100 & \xi_2 = 100, \\ \rho_1 = 1 & \rho_2 = 1 & \tau_1 = 0.01 & \tau_2 = 0.01 \\ \sigma = 0.5 & & & \end{array}$$

On Figure 2, we obviously see again that the mixture method takes into account the global variability whereas the plugin method proposes a forced consensus, leading to a narrow posterior distribution; the hierarchical method is a compromise between the two previous methods. In that particular case, it has the additional advantage to take into account the small group, which has been 'forgotten' by the plugin method. Indeed, the mode of the hierarchical distribution is slightly shifted toward the small group (corresponding to $\mu_2 = -1$) considering the members of this group in the global posterior distribution. This shows that the hierarchical approach clearly does what it is aimed at: take into account the dependencies between experts to avoid redundancies.

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. In the simulated (elicitation) data all the individuals belong to the same group which is caracterized by the following parameters:

$$\begin{array}{ll} N = 10 & \\ \mu_1 = \mu_2 = -1.5, & \xi_1 = \xi_2 = 100, \\ \rho_1 = \rho_2 = 1 & \tau_1 = \tau_2 = 0.05 \\ \sigma = 0.5 & \end{array}$$

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

Results: We obtain the following graph (see Figure 3). The mixture and plugin methods are expected to lead to similar posterior distributions. And, as expected, we observe the same behaviour for the hierarchical modelling: the three candidate elicited priors are similar. As a consequence, artificially creating a group of experts does not deteriorate the performance of our method.

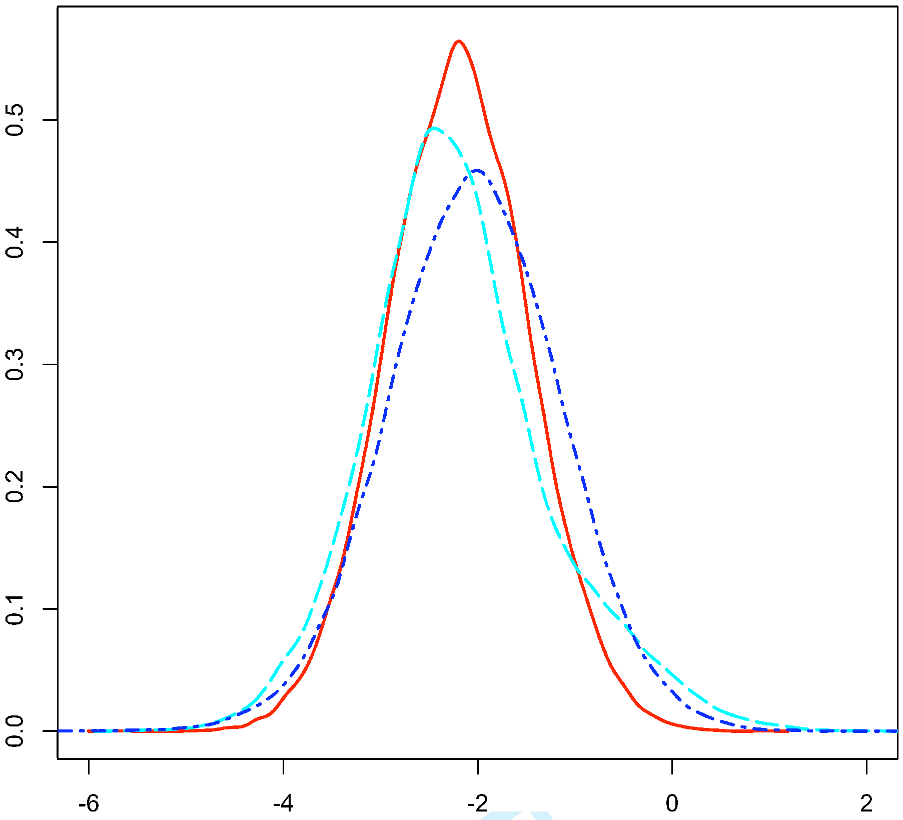

Fig. 2. *Dataset* 2*. Unbalanced groups of experts.* Comparison of methods for combination of experts using $p(\log\theta|\boldsymbol{D}_{\text{elicit}})$: mixture ($--$), plugin (solid line), hierarchical ($-\cdot$)

3.3.*2* ***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 asked questions about the quantiles of $p(do)$:

$$P(p(do) \leqslant Q_t) = t \quad \text{with } t \in T$$

($|T| = 3$) and about the probabilities

$$P_\ell(\mu, \sigma) = P[X_{10} \leqslant \ell] \quad \text{with } \ell \in L \quad (|L| = 2).$$

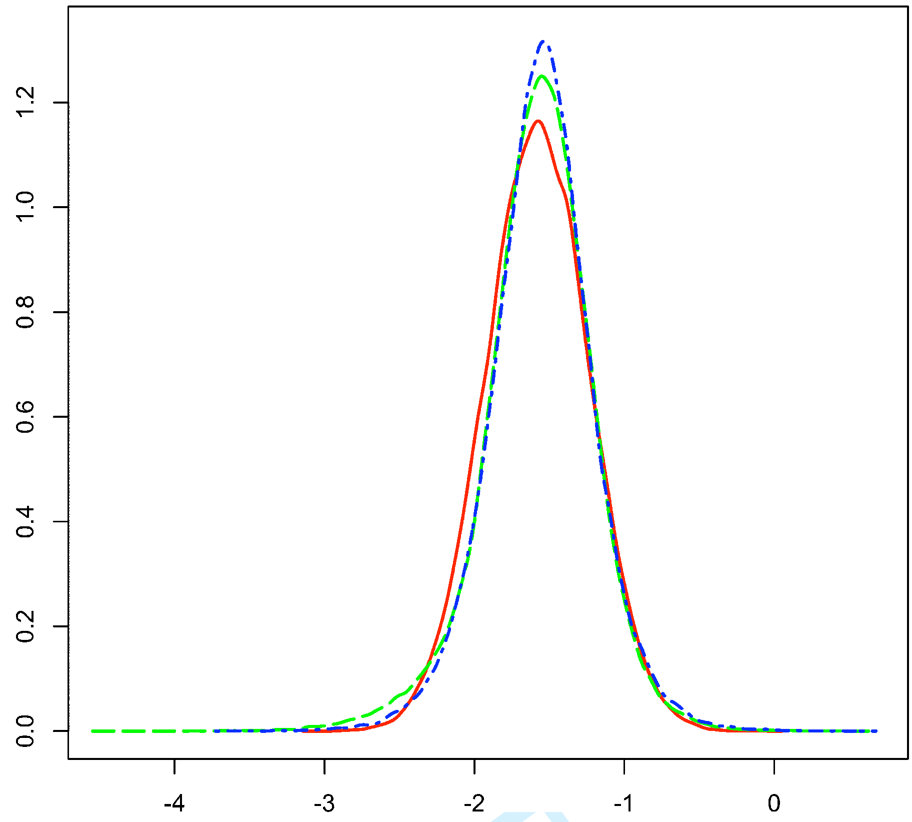

Fig. 3. *Dataset* 3. *Mis-specification of the number of groups.* Comparison of methods for combination of experts using $p(\log\theta|\boldsymbol{D}_{\text{elicit}})$: mixture ($--$), plugin (solid line), hierarchical ($-\cdot$)

T and L have been chosen by the experts, and then are different for an expert to another, depending on their knowledge. The doses do have also been chosen by each expert. For lack of information in the elicited data and so for convergence reasons, we simplify the model by considering the same variance σ_{i2} in the second group (smaller one): $\sigma_{i2}^2 \equiv \sigma_2^2$. To illustrate results, Figure 4 presents posterior densities of $p(do)$ for a fixed usual dose $do = 4$ by mixture, plugin and hierarchical approaches. The density of Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$ $(\cdots)$ is added as an example of non-informative prior on $p(do)$ classically used when no expert opinion is available. In this case, the higher a priori weights on values of 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 models the differences between experts and the plugin method proposes 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. In general, the hierarchical model provides results close to the plugin 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 leads to quite different inferences in the lower tail, which could be pivotal for decision-making related to limitations in the efficacy of the dose, reflected by estimated probabilities of mortality in both the lower and the upper tails. 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 posterior distributions clearly indicate that experts supply information on the parameter.

3.4 *PhD example*

Contrarywise to the first example, in this Section we apply Method B and the relations that are involved are mainly linear. This example illustrated that using a vague prior (on the scale of the parameter) at the lower level of the hierarchy does not necessarily lead to excessively wide elicited distributions on the time to submissions 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

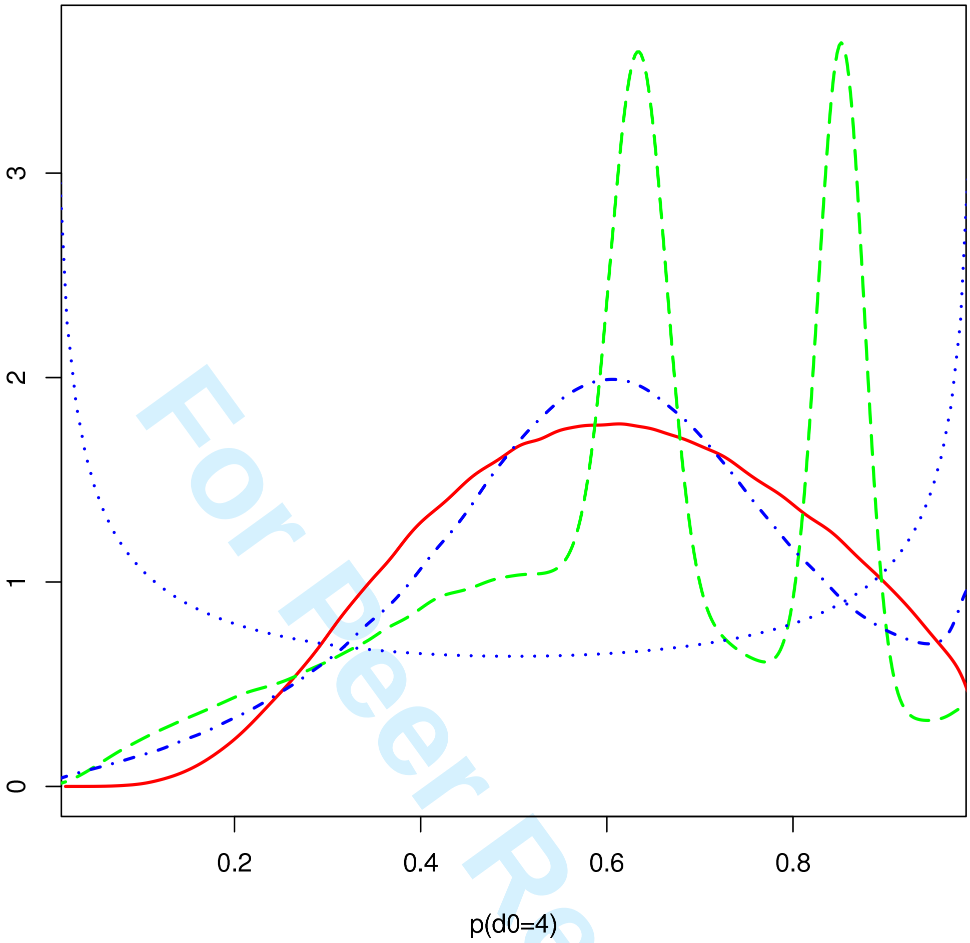

Fig. 4. Non-informative prior Beta$\left(\frac{1}{2},\frac{1}{2}\right)(\cdot\cdot\cdot)$ and posterior densities of $p(do = 4)$ with real experts data using: mixture (−−), plugin (solid line), hierarchical (−·) approaches

based on X^*, which correspond to observable quantities. This agrees with advice (Kadane et al. (1980) [13]; Low Choy et al., in press [14]). 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 $\delta = 2$ 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:

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

Recall that $\mu_{ij} = \mu + \Delta^{\mu}_{ij}, \sigma^2_{ij} = \sigma^2 \times \Delta^{\sigma}_{ij}$. We apply the hierarchical model for describing variation in μ_{ij} and σ_{ij} across experts and groups, as described 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) eliciting quantiles for specified cumulative probabilities (also known as fractile estimation) and (ii) 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 (e.g. Low Choy et al, in press [14]). 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, and since elicitation best targets information that experts can conceptualize [13]. Similarly to before we assume that the error model is Gaussian so that the likelihood associated with the error model for Q-elicitations is given by $\prod_{i=1}^{N_j} \prod_{j=1}^{J} \phi(Q_{ijt} - q_t(\mu_{ij}, \sigma_{ij}^2)|w_{ijt})$. with $\phi(.|w)$ denotes the density of a centered Gaussian random variable with variance w.

The above model implies that for each t, and corresponding $p_{ijt} \in (0,1)$, the theoretical quantile corresponding to the expert's conceptual model (parameterized by μ_{ij}, σ_{ij}) is $q_t(\mu_{ij}, \sigma_{ij}) = \sigma_{ij}\Phi^{-1}(p_{ijt}) + \mu_{ij}$ 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)

$$\text{argmin}_{\mu,\sigma} \sum_{\ell \in L} \left(\Phi^{-1}(P_{ij\ell})\sigma + \mu - q_{ij\ell}\right)^2,$$

which leads to:

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

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 $\ell \in L$. In other words $(\hat{\mu}_{ij}, \hat{\sigma}_{ij})$ is the least square 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 implicitely 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-elicitations is supplemented by a likelihood for the P-elicitations:

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

where

$$q_{ijt} \sim \mathcal{N}\left(q_t(\mu_{ij}, \sigma_{ij}), v_{ijt}\right) \tag{3.16}$$

$$\Phi^{-1}(p_{ijt}) \sim \mathcal{N}\left(\Phi^{-1}(p_\ell(\mu_{ij}, \sigma_{ij}), w_{ij\ell}\right) \tag{3.17}$$

leading to a joint distribution given by

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

Recall that the variances $v = (v_{ijt}, i, j, t \in T)$ and $w = (w_{ij\ell}, \ell \in L)$ are determined using (3.14), and in this example we consider the following prior that is vague on the scale of the quantile: $q_{ij}^* \sim \mathcal{N}_+(0, 10)$.

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. A 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). Here it is evident that although the methods-of-moment approach provides a consensus opinion, it overstates the confidence in that opinion, by not addressing variability across and within experts. The pooled estimate focuses on diversity of opinions at the expense of diversity, and also does not adjust for within-expert variation. In contrast, the hierarchical approaches 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.

Practically this translates to quite different inferences from expert opinion. Following the hierarchical model (Method B) results suggests that 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 take longer than 4.5 years to submit (about 17%). In addition, 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 5 displays the marginal posterior predictive and prior predictive distributions of the time to submission of each expert, resulting from the hierarchical model based on methods A and B. It is interesting to note that the hierarchical approaches lead to a wider posterior 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.

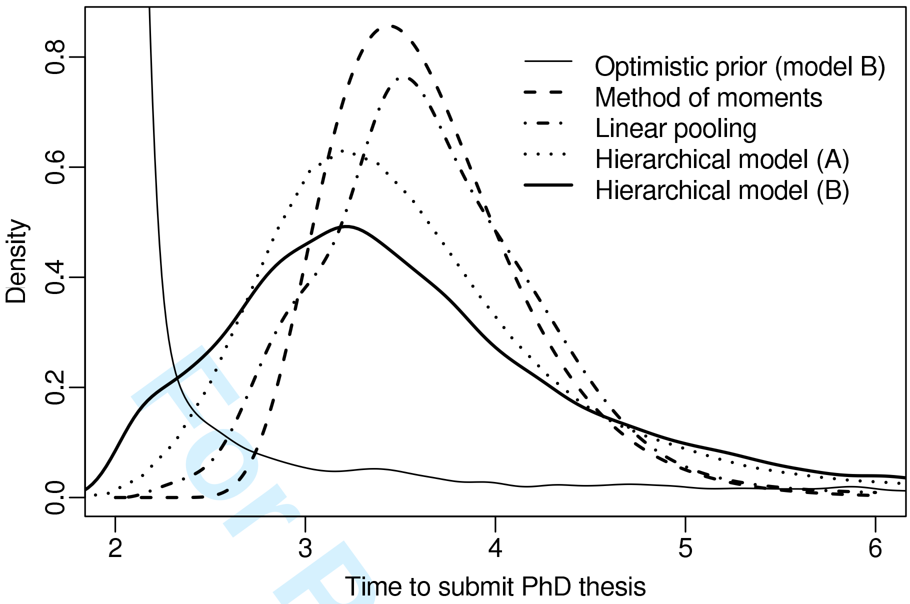

Fig. 5. Marginal posterior predictive densities of X based on: pooled (mixture model) (——), plug-in (—·), hierarchical approaches, method A (· · ·) and method B both posterior predictive density (thick solid) and prior predictive density (thin solid).

4. Discussion

As a 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, neither does it depend on the questions that are asked to the experts. In particular the experts could be aksed 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.

To our mind, one of the great 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

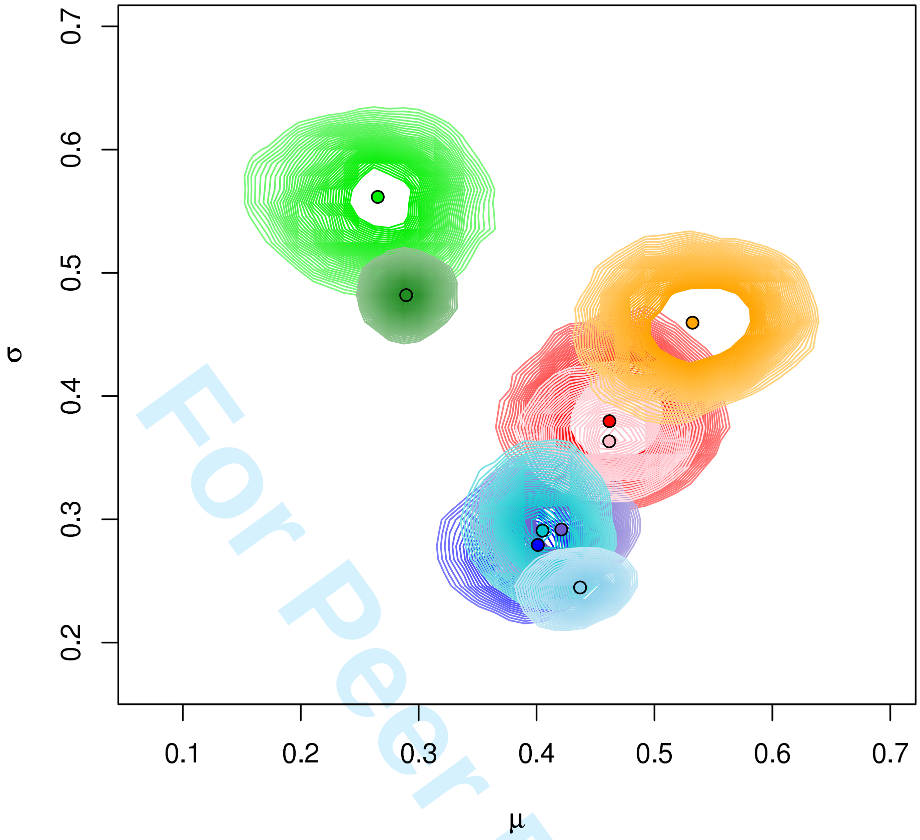

Fig. 6. Contour plots of the individual prior distributions on (μ_{ij}, σ_{ij})

approach.

A critical aspect of our method compared to pooled (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. In the examples we have considered in this paper, this led to some practically important differences in inferences. 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 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 two in the dose-response case study), Method A provided stable estimates of parameters. 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 experts the information is good enough to compensate for the complexity of the hierarchical model.

This hierarchical approach takes an "independence prior" approach , whereby the priors for γ components are independent, e.g. the μ and σ components in the PhD case study. For other applications, it may be fruitful to instead assume a conditionally conjugate prior which explicitly models dependence between the mean and variance via $p(\mu, \sigma) = p(\mu|\sigma)p(\sigma)$, and similarly for the components.

The dose-response case study shares the same structure as the PhD case study in that we are asking experts about (quantiles and cumulative probabilities) of the possible response (the number of mortalities among n mice at specific dose), rather than focussing on the parameter governing the response, here iterpretable as the proportion of dead mice. This approach was chosen to be consistent with recent elicitation research (Kynn, 2008) [6], 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 (e.g. Low Choy et al, 2010) [14].

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 food-borne diseases. *Risk Analysis* **28,** 557-571.

Cooke, R. M. and Goossens, L. H. J. (2004). Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research* **7,** 643-656.

Garthwaite, P.H. and O'Hagan, A. (2000). Quantifying expert opinion in the UK water industry: an experimental study. *The Statistician* **49,** 455-477.

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

Haas, C.N., Rose, J. B., Gerba, C.,P. (1999). *Quantitative microbial risk assessment*. New York: Wiley.

Kynn, M. (2008). The heuristics and biases bias in expert elicitation. *Journal of the Royal Statistical Society, Series A*, **171,** 1: 239 264.

Lindley, D. V. (1983). Reconciliation of probability distributions. *Operations Research* **31,** 866-880.

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.

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

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*. New York: Wiley.

Osherson, D. and Vardi, M. (2006). Aggregating disparate estimates of chance. *Games and Economic Behavior* **56,** 148-173.

Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*, Oxford University Press, UK.

Kadane, J. B., J. M. Dickey, R. L. Winkler, W. S. Smith, and S. C. Peters (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* **75,** 766:845-854.

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 "Oxford Handbook* of Applied Bayesian Analysis, eds. A. O'Hagan and M. West, Oxford University Press, UK.

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.

West, M. (1983) *Modelling Expert Opinion* in Bayesian Statistics 3, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Oxford University Press, pages 493 508.

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