Strathprints Institutional Repository

Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability

Evaluation of elicitation methods to quantify Bayes linear models

T Bedford and L Walls

The online version of this article can be found at:

Published by:

$SAGE

On behalf of:

Institution of MECHANICAL ENGINEERS

Institution of Mechanical Engineers

Additional services and information for *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* **can be found at:**

Email Alerts: http://pio.sagepub.com/cgi/alerts

Subscriptions: http://pio.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://pio.sagepub.com/content/224/4/322.refs.html

>> Version of Record - Dec 1, 2010

What is This?

Evaluation of elicitation methods to quantify Bayes linear models

M Revie*, **T Bedford,** and **L Walls**

Department of Management Science, University of Strathclyde, Strathclyde, UK

Abstract: The Bayes linear methodology allows decision makers to express their subjective beliefs and adjust these beliefs as observations are made. It is similar in spirit to probabilistic Bayesian approaches, but differs as it uses expectation as its primitive. While substantial work has been carried out in Bayes linear analysis, both in terms of theory development and application, there is little published material on the elicitation of structured expert judgement to quantify models. This paper investigates different methods that could be used by analysts when creating an elicitation process. The theoretical underpinnings of the elicitation methods developed are explored and an evaluation of their use is presented. This work was motivated by, and is a precursor to, an industrial application of Bayes linear modelling of the reliability of defence systems. An illustrative example demonstrates how the methods can be used in practice.

Keywords: Bayes linear models, elicitation; expert judgement, reliability analysis, uncertainty analysis

1   INTRODUCTION

The Bayes linear theoretical framework is a quantitative methodology capable of expressing subjective beliefs and reviewing these beliefs once observations have been made [1]. The Bayes linear methodology offers decision makers a method to structure, study, and assess the relationships between coherent partial belief statements. Once observations are made, these beliefs are adjusted by linear fitting (as opposed to the traditional conditioning rules of Bayes theorem) using equations (1) and (2) where X is a vector of quantities of interest and D is the observed data.

$$E_D(X) = E(X) + \text{cov}(X, D)(\text{var}^{-1}(D))(D - E(D)) \tag{1}$$

$$\text{var}_D(X) = \text{var}(X) - \text{cov}(X, D)(\text{var}^{-1}(D))\text{cov}(D, X) \tag{2}$$

It is natural to view the Bayes linear methodology as an inferential tool that is similar in philosophy to 'traditional' Bayesian approaches. However, it has a number of distinct features that give it certain advantages, such as simplicity and tractability, over these 'traditional' approaches when modelling problems [2]. As such, more complex problems could be modelled given the same amount of time and effort. While it is beyond the scope of this paper, it is assumed that those wishing to adopt this modelling are sympathetic to the subjective Bayesian paradigm. Further reading and debate on the objective versus subjective Bayesian paradigm can be found in references [3] and [4], and other papers within that special issue.

The Bayes linear methodology has been applied in projects in many different areas. For example: Craig *et al.* [5] use the Bayes linear methodology to estimate the level of hydrocarbons in a reservoir based on a computer simulation; Coolen *et al.* [6] to reduce the number of software tests required to remove faults from software; O'Hagan *et al.* [7] to estimate the assets of sections of the UK water industry; Farrow *et al.* [8] to support decision making within a brewery; and Revie [9] to support reliability decision making for the UK Ministry of Defence (MOD).

These papers are useful but limited. The majority of the research to date has focused on justifying the adoption of the Bayes linear methodology. Areas such as explaining the modelling process, discussing

**Corresponding author: Department of Management Science, University of Strathclyde, Graham Hills, Building 40, George Street, Strathclyde G1 1QE, UK.*
email: matthew.j.revie@strath.ac.uk

how to build the model, how to elicit the necessary values, or how to carry out sensitivity analysis, all remain relatively underdeveloped. In particular, developing pragmatic elicitation approaches has been largely ignored. For example, reference [5] does not focus on elicitation but refers to it as an area of general interest. In reference [10], the general principles of elicitation for variance and covariance are discussed, but the reader is not given guidance on how this may be carried out. More guidance is given in reference [8]. A comprehensive review of the Bayes linear theory and methods is given by Goldstein and Wooff [11], but they consider the issue of elicitation briefly. A non-exhaustive list of potential methods that could be used to populate a Bayes linear model, ranging from specifying quantiles to using historical data from samples in related populations, is presented in reference [11].

Naturally, Goldstein and Wooff [11] do not wish to give strict guidelines that may not be applicable in all cases; however, there is no clear guidance on how to specify the quantiles and elicit the expectation, variance, and covariances. Only reference [7] devotes substantial effort to discussing how elicitation is carried out. In this case, the expert stated a mean for the expectation while the variance was calculated from an elicited standard deviation. The authors are open in their concerns about this elicitation method as they state 'this expression of standard deviation, as well as the sweeping interpretation of a point estimate as a mean, is not designed to produce the most careful and accurate of prior judgements' [7]. This method may be effective in cases where the output is robust to changes in the prior beliefs, but in other cases a more structured elicitation procedure may be desirable.

To support decision makers in applying Bayes linear models more widely, clear procedures are required. Indeed, the research reported in the current paper is motivated by the authors' involvement in reliability modelling during defence procurement projects [9], which required the development of methods to support implementation of Bayes linear models that could be used with UK MOD decision makers. Prior to practical implementation within the live projects, a study was conducted of methods that could be used to support elicitation and specification of the model. This paper reports the development and evaluation of the methods proposed to overcome one of the main challenges of building a Bayes linear model. Alternative methods for quantifying a Bayes linear model through elicitation of the expectation and covariance structure are investigated and the results used to inform the recommendations. Section 2 gives an introduction to elicitation and the different methods that could be used. Section 3 presents a comparative study to evaluate methods and makes a recommendation on which method should be used. Section 4 presents an illustrative example based on an industrial application. Section 5 discusses future research.

2 METHODS FOR POPULATING A BAYES LINEAR MODEL

Elicitation is a key part of the process in populating a Bayes linear model. The purpose of an elicitation session is to gather quantitative data that represent the belief of the person being elicited [12]. This is important in many problems where empirical data may not exist to quantify the uncertainty of different probabilities. The history of the application of expert judgement is examined by Cooke [13] and a framework for the defensible elicitation and use of expert judgement, methods for the assessment of the quality and usefulness of stated values, and methods to combine judgements are proposed.

Effort has been spent in documenting how elicitation has been carried out [14–18]. A number of books have been written discussing elicitation, notably references [13] and [19–23]. These books present a comprehensive treatment of the processes involved in eliciting expert judgement. They describe the processes of planning and conducting the elicitation, identifying appropriate experts, designing questionnaires, recording judgements and managing biases, and finally analysing and aggregating experts' judgements. These books also provide a comprehensive summary of the biases that may arise during an elicitation session such as availability, representativeness, and anchoring [24], the high-level steps required during an elicitation process, and the different methods available to elicit probability distributions.

In addition, there is extensive research in the field of psychology evaluating the effects of different biases; ordering of questions [25], estimating probabilities [26], and effects of problem decomposition [27], among others. The psychology literature suggests that it is not feasible for experts simply to state their beliefs accurately about subjective probabilities [28] and instead that it is necessary to develop an elicitation process that supports the expert/decision maker in communicating coherent beliefs. A series of essays on the behaviour of people during elicitation processes is given by Kahneman *et al.* [28]. The psychological aspects of the elicitation processes and the cognitive tools people use to help form judgements are discussed, as are the common failings of individuals when making judgements. Different techniques to elicit prior beliefs may, however, produce different output as each 'method of questioning may have some effect on the way the problem is viewed' [29]. The aim when developing an elicitation procedure is

to attempt to minimize these biases and to ensure that the beliefs stated are coherent.

There is consensus in the literature regarding which high-level steps should be carried out during an elicitation process [23]. It is suggested that any elicitation process should include seven stages: background, identification, and recruitment of expert, motivating experts, structuring and decomposition, probability and assessment training, probability elicitation and verification, and aggregation of expert's probability distributions [30]. Much of the work carried out in developing elicitation processes has focused on developing methods for multiple experts [17, 30, 31]. However, on other projects, such as reference [9], only one expert was used during the modelling. As such, high-level steps for projects with only one expert are also required.

Processes which focus on only one expert have been proposed [12, 32]. A four-stage process suitable for scenarios where elicitation is carried out with single experts or decision makers is discussed in reference [12]. This four-stage process is set-up, elicitation of summaries, fit distribution, and assessing adequacy of elicitation. This process was developed for probabilistic models. In order to apply this within the Bayes linear framework, one important change is required. Instead of 'fit distributions', within the developed process, this will be changed to 'calculate mean and covariances'. Alternative processes may elicit expectations and variances directly.

2.1 Eliciting means and variances

Elicitation techniques should be designed to fit the way people think instead of forcing experts to answer questions in a specific format [21]. An expert in a subject domain should not be viewed as an expert in statistics or probability theory; any elicitation method must be as simple as possible for experts to communicate their beliefs accurately. Within a Bayes linear framework, the decision maker's belief about the expectation, variance, and covariance between the parameters in the model must be assessed.

Eliciting means and covariances directly is difficult as experts do not naturally think in these terms – 'experts should not be asked to estimate moments of a distributions (except possibly the first moment); they should be asked to assess quantiles or probabilities' [14]. Others [33] agree that it is necessary to use quantiles to estimate at least the variance of an expert's subjective belief. From these quantiles, the mean and variance can be calculated. Methods have been developed which focus on eliciting different percentiles of a variable and then calculating the mean and variance of the variable directly. A summary

is provided in reference [12] of the different elicitation methods. All of these methods, however, focus on fitting a distribution to the elicited quantiles. As the Bayes linear method does not assume any distributional form on the prior beliefs, the present authors would prefer an elicitation method which does not rely on a distributional form being assumed.

As touched upon earlier, only reference [7] has discussed in detail the different methods to elicit the necessary values. This paper was extended by O'Hagan [15] to describe additional methods to elicit variance. For the expectation, no special elicitation procedure was developed. Instead, only potential pitfalls were identified for senior personnel to pass on to engineers. For the variance, O'Hagan [15] assumed either a normal or log-normal distribution and elicited three percentiles: the median, M_X; the 17th percentile, L_X; and the 83rd percentile, U_X. If $U_X - M_X$ is substantially greater than $M_X - L_X$, then a log-normal distribution is assumed. No guidance is given in reference [15] on the numerical value such that $U_X - M_X$ is substantially greater than $M_X - L_X$.

No justification is given in reference [15] as to why the 17th and 83rd percentiles are used, other than that the author felt that experts were overconservative when specifying the interval between the 25th and 75th percentiles. It is acknowledged by O'Hagan [15] that this is in contrast to other authors, such as Alpert and Raiffa [34] and Murphy and Winkler [35], who found that people were well calibrated at these intervals. Larger intervals, such as a 99 per cent range, were strongly influenced by events of small probabilities [15], and as such, people were not well calibrated at this level. This is supported by evidence [29, 34, 36]. While reference [15] states that this method is 'very crude and simplistic', owing to the number of quantities required to populate the model, a simple approach was adopted. This method appears reasonable if the expert's beliefs are regularly normal or log-normal. In situations where this assumption cannot be made, an alternative method is required.

A method was suggested by Pearson and Tukey [37] specifically for the mean and variance which uses three percentiles for the mean and five for the variance. This method was further developed by Keefer and Bodily [38] for eliciting the variance so that the analyst was only required to specify three percentage points: 'The applicability of this approximation is significantly enhanced by eliminating the need for the 0.025 and 0.975 fractiles, especially in light of the difficulty of assessing these points in the tails' [38]. Given that the decision maker specifies his/her 0.05 ($X_{0.05}$), 0.5 ($X_{0.5}$), and 0.95 ($X_{0.95}$) percentiles for their belief about the variable of interest, the analyst can calculate

a mean, μ_X and variance, σ_X^2 for X. The Pearson and Tukey formulae are as follows

$$\mu_X = 0.63X_{0.5} + 0.185(X_{0.05} + X_{0.95}) \tag{3}$$

$$\sigma_X^2 = \left[\frac{X_{0.95} - X_{0.05}}{3.29 - 0.1\left(\frac{\Delta}{\sigma_0}\right)} \right]^2 \tag{4}$$

where

$$\Delta = X_{0.95} + X_{0.05} - 2X_{0.5}$$

and

$$\sigma_0 = \left(\frac{X_{0.95} + X_{0.05}}{3.25} \right)^2$$

The three-point Pearson and Tukey method performed almost identically to the five-point Pearson and Tukey method when assessing the variance [38].

A number of alternative methods have been proposed in the literature; the PERT (Program Evaluation Research Task) method [39], Moder–Rodgers [40], Davidson–Cooper [41], and Swanson–Megill [42]. Comparisons with these methods using the beta distribution found that the three-point Pearson and Tukey method performed best when assessing the mean and was second only to the five-point Pearson and Tukey method. The maximum error across a set of beta distributions using the Pearson and Tukey method was found to be −1.7 per cent [43]. This was substantially lower than the maximum error found with the Moder–Rodgers method, −20.7 per cent, Davidson–Cooper method, −17.7 per cent, and the PERT method, 5506 per cent. This analysis was extended by Johnson [44] to measure the accuracy of the Pearson and Tukey method to alternative distributions such as the gamma and log-normal. Johnson [44] found that the Pearson and Tukey method was robust when assessing the mean with a maximum error of 0.05 per cent for the gamma and 0.68 per cent for the log-normal. For the standard deviation, the method was again robust for the gamma distribution with a maximum error of 0.7 per cent; however, for the log-normal distribution, the maximum error for the standard deviation was 11.6 per cent. Since the current analysis is being carried out within a reliability context, the present authors are interested in assessing how the Pearson and Tukey method copes with right-skewed lifetime distributions such as those discussed.

As the method is robust across a range of possible distributions, the authors believe that the three-point Pearson and Tukey method is adequate for eliciting the mean and variance in most situations.

2.2 Eliciting covariances

In almost all complex real-world decision-making problems it is inevitable that there exists a degree of dependency between the variables being modelled. Where there is a lack of data on the relationships between variables, subjective expert judgement is crucial in modelling these problems and in providing information on the strength of the dependencies between variables [21]. While attention has been given to building influence diagrams and modelling the qualitative dependencies between different variables, less attention has been given to how to assess or elicit the quantitative dependency from experts [45]. This is a crucial part of the modelling as the output of any model could be sensitive to the dependency values specified.

Elicitation methods must have strong theoretically rigorous foundations, be capable of generalization to a wide range of different problems, and should have a clear intuitive interpretation [45]. In addition to this, those using the techniques should find them 'easy and credible' [45]. In the Bayes linear framework, covariance is the measure of dependency; however, eliciting the covariance directly from an expert is likely to encounter problems [12]. Instead, a method to calculate the covariance from other elicited values must be used.

Owing to the way in which prior beliefs are specified in a Bayes linear framework, there are a limited number of ways in which the dependency value between two variables, i.e. $\mathrm{cov}(X, Y)$, can be specified. Four different methods of determining the dependency between two variables, X and Y, in a Bayes linear framework have been identified: direct calculation (DC); direct elicitation of correlation (C); adjusted expectation (AE); and adjusted uncertainty (AU). Details of each method are given below. In each example described below, the analyst is attempting to elicit the beliefs from an expert who is referred to as 'she'. For each method, two variables are modelled using the formula $Y = \alpha X + R$, where it is assumed that X and R are uncorrelated. In this case, X is viewed as an explanatory variable of Y.

2.2.1 Direct calculation (DC)

It is assumed that the expert has already specified her $E(X)$ and $\mathrm{var}(X)$ using the Pearson and Tukey equations (3) and (4). The expert then specifies $E(Y_t)$, her new belief of $E(Y)$ given that $E(X)$ has increased by t. From this, α can be calculated using the formula $[E(Y_t) - E(Y)]/t$. She then specifies her 5, 50, and 95 percentiles for the uncertain variable R. This can be achieved by asking her, 'Given that X is known to be $\tilde{x}$ with complete certainty, what are the 5, 50, and 95 percentiles for Y?' $E(R)$ and $\mathrm{var}(R)$ can be calculated using the Pearson and Tukey method and from this

$E(Y) = \alpha E(X) + E(R)$, $\text{var}(Y) = \alpha^2 \text{var}(X) + \text{var}(R)$, and $\text{cov}(X, Y) = \alpha \text{var}(X)$.

During the elicitation, the analyst must specify $\tilde{x}$. A value for $\tilde{x}$ is chosen such that the expert is comfortable specifying quantiles for Y, given that she has observed $\tilde{y}$. If she does not feel comfortable specifying for this $\tilde{x}$, the analyst should explore alternative values.

2.2.2 Direct elicitation of correlation (C)

For C, the expert first has to specify $E(X)$, $E(Y)$, $\text{var}(X)$, and $\text{var}(Y)$. The expert is then asked to state her assessment of the correlation between variables X and Y. To do this, she must have an understanding of correlation. From $\text{corr}(X, Y)$, $\text{cov}(X, Y)$ may be calculated using $\text{cov}(X, Y) = \text{corr}(X, Y)\sigma_X \sigma_Y$.

2.2.3 Adjusted expectation (AE)

For AE, again the expert must specify $E(X)$, $E(Y)$, $\text{var}(X)$, and $\text{var}(Y)$. She is then told that she is to consider her belief about X given that she can observe the true value for variable Y. Supposing that the true value of variable Y is $\tilde{y}$, she is then asked to specify her new belief about $E(X)$, $E_Y(X) = X_Y$, given that she has observed $\tilde{y}$. By rearranging the formula for adjusted expectation (1), $\text{cov}(X, Y)$ can be calculated using

$$\text{cov}(X, Y) = \left(\frac{E_Y(X) - E(X)}{Y - E(Y)} \right) \text{var}(Y)$$

$$= \left(\frac{X_Y - \overline{X}}{\tilde{y} - \overline{Y}} \right) \sigma_Y^2 \qquad (5)$$

From $\text{cov}(X, Y)$, it is possible to calculate α. As $Y = \alpha X + R$, thus $\text{cov}(X, Y) = \alpha \, \text{var}(X)$ so that

$$\alpha = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

In order to ensure that the expert is coherent, there are limitations on the values that can be specified. As $\text{var}(Y)$ and $\text{var}(X)$ have already been specified, there exists an upper and lower bound on the value for α. Since $\text{var}(R) \geqslant 0$ and $\alpha^2 \text{var}(X) \leqslant \text{var}(Y)$, this implies that

$$\alpha \leqslant \sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}$$

Inserting

$$\alpha \leqslant \sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}$$

and assuming that $\tilde{y} \geqslant \overline{Y}$ into equation (1) gives

$$X_Y \leqslant \overline{X} + \sqrt{\frac{\text{var}(X)}{\text{var}(Y)}} (\tilde{y} - \overline{Y}) \qquad (6)$$

If the expert specifies a belief greater than the value given by equation (6), then

$$\text{cov}(X, Y) > \alpha \sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}$$

However, as this is an upper bound, one of the other values has been incoherently specified. In this case, the analyst must revisit and reassess with the expert the previously elicited values.

As with the DC method, the analyst must specify $\tilde{y}$. A value for $\tilde{y}$ is chosen such that the expert is comfortable specifying X_Y, given that she has observed $\tilde{y}$. As with the DC method, if she does not feel comfortable specifying for this $\tilde{y}$, the analyst should explore alternative values.

2.2.4 Adjusted uncertainty (AU)

For AU, again the expert must specify $E(X)$, $E(Y)$, $\text{var}(X)$, and $\text{var}(Y)$. She is then told that the value that she initially believed variable Y would take, i.e. $\overline{Y}$, has been observed. She is then asked to specify her adjusted variance for X, $\sigma_{X|Y}^2$ given that she now knows the true value of Y. To calculate the variance, equation (4) can be used. Rearranging equation (2), $\text{cov}(X, Y)$ can be calculated using the formula

$$\text{cov}(X, Y) = \sqrt{(\text{var}(X) - \text{var}_Y(X))\text{var}(Y))} \qquad (7)$$

As with AE, there are limitations on the values that the expert can specify in order to maintain coherency. In this case, the expert cannot specify that the adjusted variance, $\text{var}_Y(X)$ is greater than the prior variance, $\text{var}(X)$. In addition, in order that $\text{var}_Y(X) \geqslant 0$, $\text{cov}(X, Y) \leqslant \sqrt{\text{var}(X)\text{var}(Y)}$.

2.2.5 Setting hypothetical values

In DC and AE, the analyst must specify hypothetical values. The expert then states their belief, given that they have hypothetically observed these new values. The analyst should specify a value with which the expert is comfortable. This could clearly be multiple values. In order to assess the consistency of the expert, the analyst could choose multiple hypothetical values and ensure that the covariances across these multiple scenarios are similar.

3 EXPERIMENTAL STUDY TO EVALUATE ALTERNATIVE ELICITATION METHODS

A study has been carried out to assess the effectiveness of the four methods. The purpose of this study is to identify which of the four theoretical methods

given in section 2 is most appropriate to adopt during an elicitation session. In order to achieve this, three dependencies have been presented to participants and for each of the four methods above, a correlation value is calculated. These dependencies are: life expectancy between males and females in the same country; height and weight of male students at a university; and mean miles to failure (MMTF) between a population of vehicles and the miles to failure of a single observed vehicle. These dependencies have been chosen in order that three different scenarios are investigated. Scenario 1 assesses two dependencies with the same unit of measure; scenario 2 assesses two dependencies with different units of measures; and scenario 3 assesses the dependency between a population and a single observation. These scenarios represent the types of dependency that decision makers are likely to have to model during a risk and reliability analysis. For each of these three dependencies, expectations and variances for each of the variables have been gathered and four correlation values have been elicited from each participant using each of the described methods.

A total of 23 participants, all of whom were postgraduate students of the University of Strathclyde, have taken part in the study. While none of the participants received formal training in the Bayes linear methodology, an example was used to explain how to interpret the values. These participants can be partitioned into three distinct groups; ten MSc students in operational research, nine PhD students from the Department of Management Science, and four PhD students from the University of Strathclyde. The MSc students had all recently carried out a class in statistics, and all the PhD students had attended a short course in advanced quantitative methods. All had learned about correlation, means, and variances. Within the MSc students and the PhD students there was a mix of statistical knowledge, as some had undergraduate degrees in either mathematics or statistics.

There has been criticism that much of the empirical evidence gathered to assess elicitation techniques relies heavily on university students as the subject group (O'Hagan *et al.* [23]). It is dangerous to assume that the findings from this group will necessarily translate across to other populations. While it is necessary to be wary about the deductions that can be made from these types of experimental study, it is also necessary to be pragmatic. For example, in the current case it was necessary to pre-test methods so that an appropriate one could be selected for implementation within the industrial projects. The authors judged that the variation in the statistical knowledge of the selected students matched those of typical engineers with whom they would work in the reliability applications.
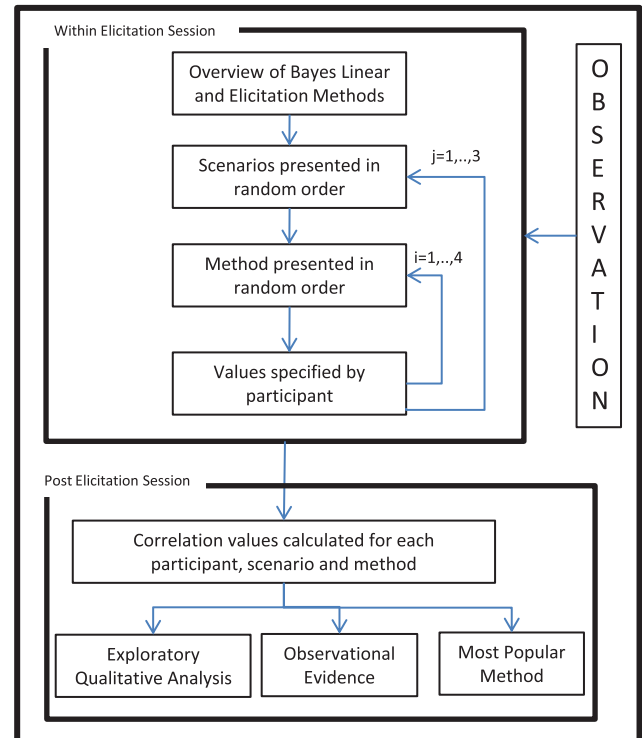

Fig. 1 Key stages of study

The students were split over three sessions. Each session lasted approximately 1 h. During this time, none of the students complained about the specifications being cognitively challenging or of fatigue. To try to minimize anchoring and potential learning, the three scenarios described above were presented in a random order. Within each scenario, the four methods for collecting the covariance were presented in a random order. Figure 1 summarizes the study design involving the process for conducting the elicitation with the participants, as well as gathering observational data from them about methods and the follow-up analysis.

3.1 Criteria to evaluate methods

In order to extract as much information as possible about the performance of the alternative elicitation methods from the limited sample of participants, three elements of analysis were conducted: formal checks on coherency; feedback from participants; and observations of the elicitation process.

Traditional methods to validate an elicitation approach are through verification, coherence, and calibration [14]. Verification can be achieved through checking that the expert is satisfied with the overall statements given. Coherency is achieved by ensuring that the values stated conform to the laws of probability. Calibration focuses on ensuring that if an expert specifies a 95 per cent interval for variable X,

then 95 per cent of all observed values of X are within that bound. However, Kadane and Wolfson [14] believe that calibration is not necessary, as elicitation is not used to elicit 'perfect' opinion but, instead, to elicit 'expert' opinion. It is hoped that identifying a 'good' expert will lead to the expert's belief and reality coinciding. If a method captures the subjective belief of an expert, then the method is described as accurate. This poses a significant problem, as traditional techniques cannot be used to determine if the elicited prior is a 'good' fit to the 'true' prior [29].

In addition to verification, coherence, and calibration to measure the accuracy of an elicitation approach, another important consideration is practicality. It is necessary to be flexible and take account of the uniqueness of each project. For complicated models or application-specific models, general elicitation methods are not desirable. Instead, the elicitation method chosen should be determined by 'examining the nature of the problem and whether or not the parameters have intrinsic meaning to the expert' [14]. Therefore, it is unlikely that a single method for eliciting covariance will be recommended. Instead, scenarios where each covariance method might be most appropriate will be highlighted.

Recall the post-elicitation stage of the study demonstrated in Fig. 1. The first stage of analysis is to carry out exploratory quantitative analysis to determine if any of the methods produces values that are inconsistent with the participants' beliefs. Since these elicitation techniques aim to gather the opinion of only the expert, any attempt to validate the techniques can be done only through other statements of belief by the same expert. In this example, it is not possible to determine which of the four techniques is capturing the 'true' belief of the respondent; however, it is possible to discount some of the responses for each of the methods.

Second, the methods are evaluated qualitatively using observational evidence gathered during the elicitation sessions and interviews carried out after the session with participants who volunteered to provide feedback. These interviews focused upon inconsistencies in the participants' responses. Information has been gathered regarding why the participant specified a given value and in which of the four methods they had most confidence.

The third stage of analysis is to determine which method is the most popular among the participants. For each dependency, each respondent has been asked to comment on which method they preferred. This is because it is unlikely that any method with which the participant is uncomfortable is likely to produce accurate results. It is beneficial to gain an understanding of how those unfamiliar with covariance techniques feel about them.

3.2 Analysis and results

For each of the three scenarios, all the participants agree that there is a relationship between the variables, i.e. the correlation between the variables does not equal zero. All the participants also agree that the relationship is not perfect, i.e. the correlation is not equal to one. Finally, all the participants agree that the relationship is positive, i.e. the correlation is positive.

For each of the elicitation methods, 23 participants specified covariance values for each of the three scenarios. From this, 69 correlation values have been calculated for each of the four methods. Each correlation is categorized as one of three: acceptable, inconsistent, or incoherent. Acceptable is defined as being a correlation value of between 0 and 1. Incoherent is defined as producing a correlation value outside -1 and 1. Any correlation value of 0, 1, or negative is defined as inconsistent. Table 1 summarizes the results.

From Table 1, it can be seen that the AU method is unlikely to be a useful method for gathering the covariance value from an expert. On three occasions, it produces correlation values that are beyond the acceptable range and on 25 occasions, it produces values that are inconsistent – all of which were correlation values of 0. From the qualitative evidence gathered, some participants believed that learning about one variable would change their belief about the expectation of the other variable, but not necessarily change their uncertainty about that belief.

A previous study [28] observed that experts do not naturally think within a Bayesian theoretical framework. In the present study, it is apparent that many of the participants do not adjust their beliefs in a similar way to the Bayes linear method updating rule. While they all agree that a relationship exists, they do not all believe that their uncertainty would drop. This is in contrast to the Bayes linear methodology in which the adjusted variance is smaller than the original variance. In addition, out of the 69 responses, only seven participants believed that the AU method was the easiest method to use. Therefore, it is unlikely that this method would be useful in eliciting the

Table 1 Frequency breakdown of correlation characteristics for alternative elicitation methods

Method	Acceptable	Incoherent	Inconsistent	Most popular
Direct calculation (DC)	67	0	2	27
Correlation (C)	57	0	12	18
Adjusted expectation (AE)	58	0	11	17
Adjusted uncertainty (AU)	41	3	25	7

necessary covariance values, other than in specific circumstances. An example where this method may be potentially useful is where making an observation does not change an expert's expectation of another variable, but changes their uncertainty of the other variable. In this case, it may be that this method is most appropriate for eliciting the necessary values.

For the vast majority of cases, the AU method will be unsatisfactory and it is necessary to distinguish between the alternative three methods. From Table 1 it is noticeable that the DC method has more acceptable values than the other two methods and no incoherent values have been elicited using this method. This method is also the most popular among participants and has other potential benefits. The method isolates the different uncertainties associated with a dependency relationship and attempts to force the decision maker to think about them individually.

For example, for dependency 3 where the participant has been asked to specify their belief about the relationship between the population MMTF (PMMTF) and the single observed miles to failure, it is clear that this is a causal relationship such that the MMTF of the population influences the miles to failure of the single vehicle. As a result, it is possible to write the observed miles to failure (OMTF), in terms of the population MMTF, PMMTF, such that $OMTF = \alpha PMMTF + R$. If the test measuring the miles to failure is assumed to be unbiased, it would be reasonable to assume that $\alpha = 1$ and that $E(R) = 0$, as did many of the participants. On eliciting var(OMTF), a number of participants stated that they believed that the var(OMTF) = var(PMMTF). When the var(OMTF) is calculated using the DC method, var(OMTF) > var(PMMTF). The DC method forces the participant to consider the different forms of uncertainty with OMTF; both epistemic and aleatory. It is possible that when the participant is directly specifying their belief about var(OMTF), they are underestimating their uncertainty.

Quantitatively, there is little difference between the AE and C method as neither of them produces values that are incoherent, but over 15 per cent of responses are deemed inconsistent. Both of them are relatively popular with participants, with 18 preferring the C method while 17 preferred AE.

In their experiments, Clemen *et al.* [45] found that the C method performed best for assessing the 'true' correlation value. This is surprising as other authors have found that directly assessing moments is a poor method for eliciting an expert's beliefs [14, 20]. One reason for the result in the experiments reported in reference [45] may be attributed to the fact that the study population had just completed a course on correlation. It may be that a problem-domain expert may have similar experience on correlation and be keen in specifying their covariance value via a correlation value or they may have no experience in correlation

and wish not to use it. In this experiment, observational evidence suggests that one reason why the C method is popular is because the participant believes that only one value had to be specified.

As these methods are attempting to model the subjective beliefs of the decision maker, the scope of the analysis that can be carried out is limited. It is impossible to determine which of the four correlation values elicited is the respondents' 'true' subjective belief. However, exploratory quantitative analysis highlighted that on 73 per cent of occasions, the value gathered using DC is greater than AE; on 63 per cent of occasions, the value gathered using DC is greater than C; and on 52 per cent of occasions, the value gathered using AE is greater than C. However, this analysis does not assess which of the three is most accurate. There does appear to be evidence in the literature which suggests that C is a poor method [14, 20].

3.3 Discussion and recommendations

The primary criterion in choosing an elicitation method is practicality. If the expert can answer the questions and feels comfortable, in the end, that to some degree her opinion has been captured, then, provided that the method meets the basic mathematical criteria of coherence, and hopefully involves some reliability testing, it is a good method [14]. Here, reliability refers to 'how well the expert agrees with him or herself in repeated tests' [46]. Out of the three methods that are available, the decision as to which method to use is essentially down to the analyst carrying out the elicitation session and the expert who is specifying their belief. If the expert is particularly comfortable and happy to adopt a specific method, then it could be argued that this method captures their belief the best.

As this study aimed to provide future users of the Bayes linear methodology with guidance for carrying out elicitation, it is beneficial to describe potential scenarios and offer recommendations for these examples. There are two scenarios that may arise when assessing risk and reliability. These are where it is natural to write one variable in terms of another and where it is apparent that only one of the two variables will be observed.

Assume that it is natural to write the relationship between two variables as $Y = \alpha X + R$. This may occur in situations where X is a causal factor of Y or in cases where X and Y have the same scale of measurement. In these cases, it is recommended that the DC method is used to calculate covariance. This method is the most popular among participants and had the most acceptable values. In addition, the elicitation questions should be framed so that the values that are specified by the expert are, in principle, observable. Cooke [13] believes that it is important that the values elicited from an expert can be thought of as observable

values. For the DC method, the expert could be asked, 'Given that you know with certainty that $X = \tilde{x}$, what are your 5, 50, and 95 percentiles for Y?' From this, $E(R)$ and var(R) can be calculated using the Pearson and Tukey method and from this, $E(Y)$ and var(Y) can be calculated.

If it is difficult for the experts to think directly in terms of the explanatory variables, it is recommended that the means and variances for both variables are elicited and the AE method is used to calculate the covariance. By setting

$$\frac{\text{cov}(X, Y)}{\text{var}(Y)} = \alpha$$

the explanatory variable formula above can be written. From this, the covariance between X and Y and other variables in the model can be easily calculated. This method is recommended over direct specification of the correlation because of the amount of literature that recommends that the first-order moments are not directly elicited [14, 20].

4 ILLUSTRATIVE EXAMPLE BASED ON INDUSTRIAL RELIABILITY APPLICATION

As part of the research reported in reference [9], two Bayes linear models were developed to support ongoing decisions by MOD reliability and maintainability (R&M) decision makers. These models supported the MOD in making procurement and entry into service decisions. During the development of the models, decision makers used the above methods to elicit their subjective beliefs. An example is given to demonstrate how the methods may be used in practice.

A decision maker is modelling the following problem. A prototype system is currently undergoing test and this is to be assessed against the required reliability performance of the operational system. The decision maker identifies three variables of interest; the observed reliability of the prototype during the test ($X_{\text{P}i}$); the actual reliability of the prototype (X_P); and the actual reliability of the operational system (X_O). The analyst uses the following two equations as a starting point; $X_\text{O} = \alpha_\text{O}X_\text{P} + R_\text{O}$ and $X_{\text{P}i} = \alpha_\text{P}X_\text{P} + R_\text{P}$ such that R_O and R_P are uncorrelated with everything else in the model.

As methods DC and AE have been recommended to use for elicitation, both are demonstrated in this example.

4.1 Application of DC method to elicit covariance

The first step is to elicit the mean and variance for X_P. To do this, the expert must state their 5th, 50th, and 95th percentile for X_P. If it is assumed that the reliability measure of interest is mean time between failures (MTBF), the expert may state 1000, 2000, and 3000 h. Using the Pearson and Tukey formulae (3) and (4), $E(X_\text{P}) = 2000$ and var$(X_\text{P}) = 607.9027$. The decision maker believes that the $E(X_\text{P})$ increases or decreases at the same rate as $E(X_\text{O})$. Thus, α_O can be set to equal 1. The next step for the decision maker is to assess $E(R_\text{O})$ and var(R_O). The Pearson and Tukey formulae are used to elicit $E(R_\text{O})$ and var(R_O). Questions may be asked such as 'given that we know that X_P was equal to 2000 with complete certainty, what are your 5, 50, and 95 percentiles for $X_{\text{P}i}$?' Assuming that the decision maker specifies 1750, 2000, and 2250, $E(R_\text{O}) = 0$ and var$(R_\text{O}) = 151.9757$. This suggests that the decision maker does not expect any difference between the prototype version and the operational version. From this, $E(X_\text{O}) = 2000$, var$(X_\text{O}) = 607.9027 + 151.9757 = 759.8784$, and cov$(X_\text{P}, X_\text{O}) = 607.9027$.

4.2 Application of AE method to elicit covariance

The AE method is used to elicit the covariance between X_P and $X_{\text{P}i}$. The first step is to elicit $E(X_{\text{P}i})$ and var$(X_{\text{P}i})$ using the Pearson and Tukey method. Assume that the decision maker specifies values such that $E(X_{\text{P}i}) = 3000$ and var$(X_{\text{P}i}) = 1000$. The decision maker believes that the test is not capturing all failure modes and, as such, the test will output a high value. To calculate cov$(X_\text{P}, X_{\text{P}i})$, the decision maker specifies his/her $E(X_\text{P})$ given that he/she has observed $\tilde{x}$. In this case, $\tilde{x} = 2500$, the decision maker specified that $E(X_\text{P}) = 1750$. Using formula (6), cov$(X_\text{P}, X_{\text{P}i}) = 500$. Hence, $\alpha_\text{P} = 0.8225$.

These are all the values that the expert is required to specify. Currently, however, cov$(X_\text{O}, X_{\text{P}i})$ has not been elicited. Owing to the way that the problem was constructed, it is not necessary to elicit any more values to assess cov$(X_\text{O}, X_{\text{P}i})$. The following formula is used

$$\text{cov}(X_\text{O}, X_{\text{P}i}) = \text{cov}(X_\text{O}, \alpha_\text{P}X_\text{P} + R_\text{P})$$
$$= \text{cov}(\alpha_\text{O}X_\text{P} + R_\text{O}, \alpha_\text{P}X_\text{P} + R_\text{P})$$
$$= \alpha_\text{O}\alpha_\text{P}\text{var}(X_\text{P})$$

Thus, cov$(X_\text{O}, X_{\text{P}i}) = 500$.

5 SUMMARY AND FUTURE WORK

This example presented is a small part of a larger Bayes linear model that has been developed and presented in reference [9] to support decision making within the MOD. The modelling has been applied on two projects with different decision makers. Feedback gathered from both regarding the elicitation methods has been largely positive. For example, quantifying

Proc. IMechE Vol. 224 Part O: J. Risk and Reliability

JRR304

beliefs has been useful in establishing the current state of knowledge about the system and the process of quantifying beliefs was considered very useful. In particular, decision makers found the process of specifying their dependency straightforward, with neither one indicating that they found any of the process difficult.

There are many other areas in which Bayes linear modelling could support decision making. As discussed in section 1, Bayes linear methods have been applied in many different domains; in particular, where decision makers may be keen on constructing a full Bayesian analysis but do not have sufficient resources. This paper provides guidance in specifying elicitation techniques for those who are new to Bayes linear methods.

Future research could focus on extending this analysis by gathering more participants. These participants should closely represent those decision makers or experts who would be specifying their beliefs during real projects. Alternatively, if additional participants were available but not meeting this criteria, a comparison could be carried out assessing the difference between those with detailed statistical knowledge and those with only basic knowledge. If there was a difference between the two groups, methods for each could be developed and applied.

ACKNOWLEDGEMENTS

REFERENCES

1 **Goldstein, M.** Exchangeable belief structures. *J. Am. Statist. Assoc.*, 1986, **81**(396), 971–976.

2 **Goldstein, M.** and **Bedford, T.** The Bayes linear approach to inference and decision-making for a reliability programme. *Reliability Engng System Saf.*, 2006, **92**(10), 1344–1352.

3 **Goldstein, M.** Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 2006, **1**(3), 403–420.

4 **Berger, J.** The case for objective Bayesian analysis. *Bayesian Analysis*, 2006, **1**(3), 385–402.

5 **Craig, P., Goldstein, M., Seheult, A.,** and **Smith, J.** Bayes linear strategies for matching hydrocarbon resevoir history. In *Bayesian statistics 5* (Eds J. Bernardo, J. Berger, A. Dawid, and A. Smith), 1996, pp. 69–95 (Oxford University Press, Oxford).

6 **Coolen, F. P. A., Goldstein, M.,** and **Munro, M.** Generalized partition testing via Bayes linear methods. *Inf. Software Technol.*, 2001, **43**, 783–793.

7 **O'Hagan, A., Glennie, E.,** and **Beardsall, R.** Subjective modelling and Bayes linear estimation in the UK water industry. *Appl. Statistician*, 1992, **41**(3), 563–577.

8 **Farrow, M., Goldstein, M.,** and **Spiropoulos, T.** Developing a Bayes linear decision support system for a brewery. In *The practice of Bayesian analysis* (Eds S. French and J. Smith), 1997 (Wiley, New York).

9 **Revie, M.** *Evaluation of Bayes linear modelling to support reliability assessment during procurement.* PhD Thesis, Department of Management Science, University of Stratclyde, Glasgow, UK, 2008.

10 **Farrow, M.** Practical building of subjective covariance structure for large complicated systems. *The Statistician*, 2003, **52**(4), 553–573.

11 **Goldstein M.** and **Wooff, D.** *Bayes linear statistics: Theory and methods.* 2007 (John Wiley, Chichester).

12 **Garthwaite, P. H., Kadane, J. B.,** and **O'Hagan, A.** Statistical methods for eliciting probability distributions. *J. Am. Statist. Assoc.*, 2005, **100**(470), 680–700.

13 **Cooke, R. M.** *Experts in uncertainty. Opinion and subjective probability in science,* Environmental Ethics and Science Policy Series, 1991 (Oxford University Press, New York).

14 **Kadane, J. B.** and **Wolfson, L. J.** Experiences in elicitation. *The Statistician*, 1998, **47**(1), 3–19.

15 **O'Hagan, A.** Eliciting expert beliefs in substantial practical applications. *The Statistician*, 1998, **47**(1), 21–35.

16 **Craig, P., Goldstein, M., Rougier, J. C.,** and **Seheult, A.** Bayes linear strategies for matching hydrocarbons resevoir history. *J. Am. Statist. Assoc.*, 2001, **96**(454), 717–729.

17 **Walls, L.** and **Quigley, J.** Building prior distributions to support bayesian reliability growth modelling using expert judgement. *Reliability Engng System Saf.*, 2001, **74**, 117–128.

18 **Bedford, T., Quigley, J.,** and **Walls, L.** Expert elicitation for reliable system design. *Statist. Sci.*, 2006, **21**(4), 428–450.

19 **Yates, J. F.** *Judgement and decision making,* 1990 (Prentice-Hall, New Jersey).

20 **Morgan, M.** and **Henrion, M.** *Uncertainty,* 1990 (Cambridge University Press, Cambridge).

21 **Meyer, M. A.** and **Booker, J. M.** *Eliciting and analyzing expert judgement,* 2001 (Academic Press Limited, London).

22 **Koehler, D.** and **Harvey, N.** *Blackwell handbook of judgement and decision making,* 2004 Blackwell Publishers, Oxford).

23 **O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H.,** and **Jenkinson, D. J.** *Uncertain judgements: Eliciting experts' probabilities,* 2006 (John Wiley, Chichester).

24 **Tversky, A.** and **Kahneman, D.** Judgement under uncertainty: Heuristics and biases. *Science*, 1974, **185**, 1124–1131.

25 **Payne, S.** *The art of asking questions,* 1951 (Princeton University Press, Princeton, New Jersey).

26 **Hogarth, R. M.** *Judgement and choice: The psychology of decisions,* 1980 (Wiley-Interscience, Chicago).

27 **Hayes-Roth, B.** Estimation of time requirements during planning: interactions between motivation and

cognition. Technical Report, Rand Corporation Project, 1980.

28 **Kahneman, D., Slovic, P.,** and **Tversky, A.** *Judgement under uncertainty: Heuristics and biases*, 1982 (Cambridge University Press, Cambridge).

29 **Winkler, R. L.** The assessment of prior distributions in Bayesian analysis. *J. Am. Statist. Assoc.*, 1967, **62**(319), 776–800.

30 **Clemen, R.** and **Reilly, T.** *Making hard decisions with decision tools*, 2001 (Duxbury Press, Grove, California).

31 **Phillips, L.** Group elicitation of probability distributions: Are many heads better than one? In *Decision science and technology: Reflections on the contributions of Ward Edwards* (Eds J. Shanteau, B. A. Mellers, and D. A. Schum), 1999, pp. 313–330 (Kluwer Academic, Boston/London).

32 **Merkhoffer, M. W.** Quantifying judgemental uncertainty: Methodology, experiences and insights. *IEEE Trans., Systems, Man and Cybernetics*, 1987, **17**(5), 741–782.

33 **Garthwaite, P. H.** and **O'Hagan, A.** Quantifying expert opinion in the UK water industry: An experimental study. *The Statistician*, 2000, **49**(4), 455–477).

34 **Alepert, M.** and **Raiffa, H.** A progress report on the training of probability assessors. In *Judgement under uncertainty: Heuristics and biases* (Eds D. Kahneman, P. Slovic, and A. Tversky), 1982, pp. 294–305 (Cambridge University Press, Cambridge).

35 **Murphy, A.** and **Winkler, R. L.** Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Statistician*, 1977, **26**, 41–47.

36 **Hora, S. Hora, J.,** and **Dodd, N.** Assessment of probability distributions for continuous random variables: a comparison of the bisection and fixed-value methods. *Organizational Behavior Hum. Decision Processes*, 1992, **51**, 133–155.

37 **Pearson, E.** and **Tukey, J.** Approximate means and standard deviations based on distances between percentage points of frequency curves. *Biometrika*, 1965, **52**(3, 4), 533–546.

38 **Keefer, D. L.** and **Bodily, S. E.** Three-point approximations for continous random variables. *Mgmt Sci.*, 1983, **29**(5), 595–610.

39 **Malcolm, D., Roseboom, J., Clark, C.,** and **Fazar, W.** Application of a technique for research and development program evaluation. *Ops Res.*, 1959, **2**, 646–669.

40 **Moder, J.** and **Rodgers, E.** Judgement estimates of the moments of pert type distributions. *Mgmt Sci.*, 1968, **15**, B76–B83.

41 **Perry, C.** and **Greig, I.** Estimating the mean and variance fo subjective distributions in pert and decision analysis. *Mgmt Sci.*, 1975, **21**, 1447–1480.

42 **Megill, R.** *An introduction to risk analysis*, 1977 (Petroleum Publishing Company, Tulsa).

43 **Keefer, D.** and **Verdini, W.** Better estimation of PERT activity time parameters. *Mgmt Sci.*, 1993, **39**, 1086–1091.

44 **Johnson, D.** Triangular approximations for continuous random variables in risk analysis. *J. Opl Res. Soc.*, 2002, **53**, 457–467.

45 **Clemen, R., Fischer, G. W.,** and **Winkler, R. L.** Assessing dependence: Some experimental results. *Mgmt Sci.*, 2000, **46**(8), 1100–1115.

46 **Kynn, M.** The 'heuristics and biases' bias in expert elicitation. *J. R. Statist. Soc. Part A*, 2008, **171**(1), 239–264.