Reliability Engineering and System Safety 158 (2017) 41–49

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Deriving the probability of a linear opinion pooling method being superior to a set of alternatives

Donnacha Bolger*, Brett Houlding

Department of Statistics, Trinity College Dublin, Dublin 2, Ireland

ARTICLE INFO

Keywords:
Decision analysis
Expert judgment
Reliability
Risk analysis
Bayesian analysis

ABSTRACT

Linear opinion pools are a common method for combining a set of distinct opinions into a single succinct opinion, often to be used in a decision making task. In this paper we consider a method, termed the Plug-in approach, for determining the weights to be assigned in this linear pool, in a manner that can be deemed as rational in some sense, while incorporating multiple forms of learning over time into its process. The environment that we consider is one in which every source in the pool is herself a decision maker (DM), in contrast to the more common setting in which expert judgments are amalgamated for use by a single DM. We discuss a simulation study that was conducted to show the merits of our technique, and demonstrate how theoretical probabilistic arguments can be used to exactly quantify the probability of this technique being superior (in terms of a probability density metric) to a set of alternatives. Illustrations are given of simulated proportions converging to these true probabilities in a range of commonly used distributional cases.

1. Introduction

In realistic decision making scenarios DMs will commonly have access to a wide range of opinions about the true value of the uncertain aspect(s) inherent within the decision task, e.g., the value of a stock price in a month, or the amount of cars that will pass along a motorway in an hour period. In addition to these opinions she (we assume the DM to be female, a common convention) also has an opinion that she herself holds. Numerous publications discuss how an amalgamated opinion can potentially outperform several of the opinions comprising it. Bates and Granger [1] consider using a weighted sum in point estimate setting with weights inversely proportional to absolute errors from previous predictions, with Newbold and Granger [27] comparing this method to several others models using a collection of eighty financial data sets and providing strong evidence in its favour. Bunn [3] considers conjugate updating (either Beta/Binomial or Dirchlet/Multinomial) for assigning weights, with performance of this scheme appearing strong in comparision to linear and exponential methods. Several of these discussions arise within an expert judgment context, where a DM requires the knowledge of domain-specific experts to assist her in her decision making task, perhaps most notably the classical method of Cooke [6,7]. This provides a framework for assessing the respective merits of the opinions of experts, using a set of seed variables (whose true values are known to the DM but unknown to the experts) to gauge their relative reliabilities. Empirical justification has been provided for this technique by Cooke and Goossens [8] and Eggstaff et al. [13], with Clemen [4] and Flandolini et al. [14] discussing the validation method used and potential alternatives. Below we consider a more novel setting, consisting of n non-competing DMs, each of whom possesses an opinion about the inherent decision uncertainty, and is willing to combine this with those of her neighbours in the hope of increasing the accuracy of the opinion that she makes her decision with, and hence her corresponding (personal) decision quality. This environment subtly differs from the expert judgment framework but shares a similar goal, i.e., constructing a combined opinion that is as accurate as possible for use in a decision making task.

Discussion on manners by which a set of opinions can be amalgamated into a single succinct opinion are widespread, with Clemen and Winkler [5] and Genest and Zidek [18] providing details for the interested reader. A fully Bayesian framework is an attractive concept, in which a DM specifies her own prior distribution before viewing the opinions of others as data to be incorporated into a likelihood function, with the product of the two yielding her posterior distribution. Yet major issues arise with the specification of an appropriate likelihood function, and while specific forms have been suggested for specific problems there is no generalised method of supplying such a function. This likelihood function would need to entail the dependence between informaiton sources, a task which will frequently be beyond the computational scope of users. French [16] comments on the attractiveness of the concept but concedes it has vast

* Corresponding author.
E-mail addresses: bolgerdo@tcd.ie (D. Bolger), brett.houlding@tcd.ie (B. Houlding).

http://dx.doi.org/10.1016/j.ress.2016.10.008
Received 1 November 2015; Received in revised form 29 September 2016; Accepted 22 October 2016
Available online 27 October 2016
0951-8320/

implementation problems, with Clemen and Winkler [5] concurring that while the method is compelling it "is also frustratingly difficult to apply". Morris [24–26] is another example of an attempt to implement fully Bayesian updating. However this work uncovers deep problems concerning both partial exchangeability and over-determination, unfortunately making the derived approach troublesome in practice, with Morris commenting that the likelihood function can be "extremely complicated and virtually impossible to assess in all but the simplest cases". We comment that van Noortwijk et al. [31] is one illustration of a specific application of this Bayesian approach and how, given a set of assumptions, it may be successfully implemented. Like our own methodology it is concerned with combining and updating opinions. However it does not extend to a generalised framework, and is reliant upon the user having some knowledge about the accuracy of received opinions prior to witnessing data.

The Bayesian methodology is extremely attractive but is not suitably general for widespread application. Hence various pooling methods are commonly considered, most notably in a linear or logarithmic fashion. The various approaches have differing associated strengths and weaknesses, e.g., linear pooling obeys the marginalisation property, while logarithmic pooling obeys the external Bayesianity property, both of which are discussed in, for instance, Genest and Zidek [18]. In what follows we choose to use linear opinion pooling, for its relative simplicity, ease of interpretation, and in sticking with common practice. The obvious area of interest, upon making this choice, is in determining the appropriate weights to assign, with these weights being constrained to be non-negative and to sum to unity. Genest and McConway [17] discuss various approaches by which this can be done, depending on the setting of interest, the aims of the process, and the underlying philosophy of the individual(s) assigning weights. Our setting of interest is a dynamic one, in which each DM makes a decision, sees a return, and then repeats this process indefinitely. Hence ideally two types of learning will occur over time, with a DM modifying her own opinion about the (latent) unknown decision quantity in light of the noisy realisations of it that have been observed, and also adjusting the degree of consideration that she affords to the opinions of her neighbours, given the contrast between their predictions and the witnessed reality. This dynamicity further increases the novelness and applicability of our method. We comment that DeGroot [11] and DeGroot and Bayarri [12] discuss similar settings with weight updating, but do so in a different context to us - in their context there is one set of weights to be determined/updated at each new observation, while in our case there are n sets, one for each DM.

The problem context which we highlight and provide a solution for in what follows is a substantive one which combines the fields of dynamic reliability and expert opinion. Below we include several examples of practical risk and reliability environments of this nature, illustrating the relevance of the research undertaken and some potential areas for application. This paper can be seen as providing both a methodology (and some considerable justification for its use), as well as some theoretical results which demonstrate when it is most appropriate for use. The work that is presented within this paper is highly relevant to a wide range of realistic problems across a broad spectrum of fields that pertain to technical risk and reliability.

- A collection of stockbrokers may communicate amongst each other to predict the behaviour of a financial stock, i.e., whether its price will rise or fall. Here the risk is over their personal monetary fortunes.
- A similar problem exists in the area of weather forecasting, where it has been demonstrated that a combined (ensemble) forecast can commonly outperforms single forecasts (e.g., Gneiting and Raftery, [19]). Hence there is a need to determine the weights to assign the respective forecasts. A government may require these predictions to determine if urgent anti-flood measures should be carried out to prevent against the risk of dwellings being damaged as well as individuals being injured.
- Policy developers with a non-government organisation (NGO) which provides aid to countries in the developing world may pool their respective predictions concerning the prospective Gross Domestic Product (GDP) of these nations in order to determine the ratio by which funds should be divided. Risk in this situation can be seen as one nation being allocated too much of a particular resource while another is allocated too little (e.g., food and water, vacciniations).
- A group of computer programmers may wish to pool their beliefs about the average number of bugs occurring per one thousand lines of code, in order to aid their individual decisions on whether to release their software or to continue testing it (e.g., Wilson and McDaid, [33]). The risks are clear here: if they release too early then they may need to recall software and hence lose money, as well as suffering damage to their brand reputation.
- Nuclear power stations may wish to confer between each other as to the perceived risk of a fault occurring to assist in their decision of determining whether additional safety devices need be installed or not (e.g., Starr, [30]). Potential risk to local and national safety are evident here.
- Medical practitioners may seek the opinion of peers as to the probability of a diagnosis being correct given some symptoms witnessed, or the efficacy of a novel drug treatment (and similar problems, e.g., Cox, [9]). In cases like this the risk is defined over the future health state of patient(s) using the medicine prescribed.
- Several companies may wish to exchange their opinions on what proportion of a particular demographic (e.g., males under twenty-one) buy a particular product (e.g., computer games, laptops, jeans) to ascertain how large a quantity they should respectively produce. If the proportion is overestimated then they risk too large a quantity of the goods being produced and not sold, leading to unneccesary manufacturing cost and hence a waste of captial.

In each case outlined above there is clearly technical risk present that the users wish to avoid, with this risk being defined over either financial wealth, national safety or medical health. In the illustrated situations each individual entity supplying its probabilistic opinion (be it a single person or be it a large multinational company) will have their own personal utility function over the possible outcomes which may occur, so even if decisions are made using common beliefs different decisions may well be deemed optimal by different entities.

The outline of the remainder of this paper is as follows. Section 2 introduces the Plug-in (PI) approach, a method achieving these two forms of dynamic learning that allocates weights for a linear opinion pooling. Section 3 discusses a simulation study that has been conducted to provide some validation for this technique by comparing its performance to those of some rational alternatives in different contexts. Section 4 derives the theoretical probabilities that are estimated by the simulated proportions in the previous section, illustrating how these proportions converge asymptotically to the true probabilities of interest. We conclude with discussion regarding the relevance of this material, potential applications, and further research in Section 5.

2. The plug-in approach

Suppose we have n DMs, denoted $P_1, \ldots, P_n$, with $n \geq 2$ in non-trivial cases. Each DM possesses two things: her fully parameterised probability distribution $f_i(\theta)$ over the uncertain quantity θ, and her own subjective utility function, $u_i(r)$ which is indicative of her own personal attitude to risks and gambles as consequence of decision return r. Note we assume all DMs possess some uncertainty over the opinion that they hold, i.e., all distributions have a strictly positive variance. We denote the updated opinion of P_i, having received the opinions of her neighbours, by $\widehat{f}_i(\theta)$:

$$\widehat{f}_i(\theta) = \alpha_{i,1} f_1(\theta) + \cdots + \alpha_{i,i} f_i(\theta) + \cdots + \alpha_{i,n} f_n(\theta) \quad (1)$$

We require $\alpha_{i,j} > 0$ for all i, j and $\sum_{j=1}^{n} \alpha_{i,j} = 1$ for all i, with $\alpha_{i,j}$ being the weight P_i assigns P_j.

Each individual DM must make a decision, the consequences of which will solely impact on herself (e.g., whether they should purchase a stock or not, with the corresponding profit/loss falling solely on their shoulders). The method of maximisation of expected utility is a commonly applied technique in decision problems – the quantitative theory of subjective probability and utility was introduced in Ramsey [29], with von Neumann and Morgernstern [32] later proving the expected utility hypothesis and providing a rigorous axiomatic justification for its use. In order to implement this scheme a DM requires two constructs: a utility function and a probability distribution. As mentioned above, each P_i has her own utility function which reflects her personal attitude towards risk. They also possess their own probability distribution, $f_i(\theta)$, and that which is garnered from listening to their neighbours, $\widehat{f}_i(\theta)$. In light of the discussion in the introduction concerning the merits of amalgamating information sources in decision processes, we propose that each P_i choose her optimal decision, d^*, to be such that

$$d^* = \underset{d \in \mathcal{D}}{\operatorname{argmax}} \mathbb{E}[u_i(d)] = \underset{d \in \mathcal{D}}{\operatorname{argmax}} \left[\int_{\Theta} u_i(d, \theta) \widehat{f}_i(\theta) \mathrm{d}\theta \right] \tag{2}$$

For notational ease we use the same symbol u_i for $u_i(d, \theta) \equiv u_i(r)$ and $u_i(d)$ and assume it is contextually evident which is meant. Note Eq. (2) is a function of $\widehat{f}_i(\theta)$, instead of solely being a function of $f_i(\theta)$, i.e., P_i is explicitly taking the opinions of her neighbours into account. Before making an initial decision a DM may be unsure of the weights to award to her neighbours. Hence she may use the Laplacian Principle of Indifference [21] and afford all information sources (including herself) an equal weight. Optionally she may assign arbitrary weights based upon her prior judgments. As mentioned above, we focus on two types of dynamic learning:

- Learning over θ: We assume that all DMs update their opinion over θ in a Bayesian manner. Hence if P_i sees a new return, r, then she constructs her posterior distribution over θ as being proportional to the product of her prior distribution over θ and an appropriate likelihood function:

$$f_i(\theta|r) \propto f(r|\theta) f_i(\theta) \tag{3}$$

Throughout this paper we shall consider the case where all DMs agree on a common likelihood function - indeed this shall be pivotal in the condition shortly outlined in Eq. (6) and for the conjugate updating in our theoretical calculations of Section 4. Further detailed discussion of how to proceed if this is not the case is defered primarily until our concluding remarks in Section 5. In Sections 3 and 4 we assume conjugate updating, but this is not a necessity, with numerical methods (e.g., Markov Chain Monte Carlo) implementable in intractable cases.

- Learning over reliability: Once a return r has been witnessed a DM will want to compare this to the predictions of her various neighbours. We propose that, for P_j, this reliability measure (termed the PI weight), w_j, is the value she placed on r in her prior predictive distribution, i.e.,

$$w_j = f_j(R = r) = \int_{\Theta} f(R = r|\theta) f_j(\theta) \mathrm{d}\theta \tag{4}$$

This will be high for DMs who placed a high probability on r occurring prior to seeing it (i.e., who appear reliable), and the converse. Note that this metric takes into account not only the central tendency of the distribution (e.g., its mean) but also the associated uncertainty (i.e., its variance). Suppose P_i previously assigned a weight of $\alpha_{i,j}$ to P_j, before seeing some new data, and calculating a latest Plug-in weight, w_j. How should she determine the updated normalised weight, $\alpha^*_{i,j}$ to allocate to P_j? We propose the following scheme:

$$\alpha^*_{i,j} = \frac{w_j \alpha_{i,j}}{\sum_{k=1}^{n} w_k \alpha_{i,k}} \tag{5}$$

An argument for the functional form of the reweighting scheme proposed in Eq. (5) is that it ensures updating of a combined belief is done in a manner adhering to the Bayesian paradigm, as illustrated for instance in Lee [22] in relation to a discussion on mixture distributions. Note that this only holds true in the pre-discussed case in which all participants concur on a common likelihood function, i.e., that $\boldsymbol{f_i}(\boldsymbol{r}|\boldsymbol{\theta}) = \boldsymbol{f}(\boldsymbol{r}|\boldsymbol{\theta})$ for all $\boldsymbol{i} = \boldsymbol{1}, \ldots, \boldsymbol{n}$. Consider a case where a combined belief is written as $\widehat{f}_i(\theta)$, and then a return r is witnessed, with corresponding PI weights of $w_1, \ldots, w_n$ for the n DMs. The posterior distribution, $\widehat{f}_i(\theta|r)$ can be written as

$$\begin{aligned} \widehat{f}_i(\theta|r) &= \frac{f(r|\theta)\widehat{f}_i(\theta)}{f(r)} = \frac{f(r|\theta)\sum_{j=1}^{n} \alpha_{i,j} f_j(\theta)}{\sum_{j=1}^{n} \alpha_{i,j} \int_{\Theta} f(r|\theta) f_j(\theta) d\theta} \\ &= \frac{\sum_{j=1}^{n} \alpha_{i,j} f(r|\theta) f_j(\theta)}{\sum_{j=1}^{n} \alpha_{i,j} \int_{\Theta} f(r|\theta) f_j(\theta) d\theta} = \frac{\sum_{j=1}^{n} \alpha_{i,j} f_j(\theta|r) \int_{\Theta} f(r|\theta) f_j(\theta) d\theta}{\sum_{j=1}^{n} \alpha_{i,j} \int_{\Theta} f(r|\theta) f_j(\theta) d\theta} \\ &= \frac{\sum_{j=1}^{n} \alpha_{i,j} w_j f_j(\theta|r)}{\sum_{j=1}^{n} \alpha_{i,j} w_j} \propto w_1 \alpha_{i,1} f_1(\theta|r) + \cdots + w_n \alpha_{i,n} f_n(\theta|r) \end{aligned} \tag{6}$$

Note without out common likelihood function the above proof would fail, as $\boldsymbol{f}(\boldsymbol{r}|\boldsymbol{\theta})$ could not be taken outside the sum in the numerator as a common factor. The combined posterior distribution is a linear combination of individual posterior distributions, with the associated weights being precisely those derived by Bayes' Theorem. Hence the approach can be deemed rational, assuming one agrees with the fundamental principles of Bayesian statistics. Arising from this result is exchangeability, implying that the PI method assigns identical weights irrespective of the order in which a set of t returns is witnessed, i.e., $\alpha_{i,j}$ is unchanged regardless of if $r_{\sigma_1(1)}, \ldots, r_{\sigma_1(t)}$ or $r_{\sigma_2(1)}, \ldots, r_{\sigma_2(t)}$ are witnessed, where σ_1 and σ_2 are distinct permutations of a common set of t returns. We can write the unnormalised weight assigned by P_i to P_j after t returns, $\alpha^{(t)}_{i,j}$, as a product of the PI weights of P_j up to this point, and her initial weight $\alpha^{(0)}_{i,j}$, i.e.,

$$\alpha^{(t)}_{i,j} \propto \alpha^{(0)}_{i,j} \prod_{k=1}^{t} w_j^{(k)} \tag{7}$$

We briefly mention some asymptotic behaviour which is associated with the PI approach. Firstly, irrespective of the (non-degenerate) prior which a DM supplies, all individual posterior distributions, $f_i(\theta|r_1, \ldots, r_N)$ will converge towards identical distributions which are perfectly accurate models of the true underlying phenomena θ, by the Law of Large Numbers, as N approaches infinity. Secondly, Genest and McConway [17] provide an interesting asymptotic result concerning the convergence of weights in a scheme of this nature: in the limit the weight assigned to P_i will be the product of their initial weight (here $\frac{1}{n}$) and the probability placed in their prior distribution on the true value of the unknown quantity. Finally we note that as all posterior distributions converge towards a common distribution, the combined posterior distribution (used by DMs in their decision making as in Eq. (2)) will also be identical for all DMs in the limiting case. Hence, once a large enough amount of returns have been observed, the choice of weighting scheme is irrelevant.

We list below a set of four desirable properties that the PI approach adheres to. Any method failing to obey these properties would appear problematic. Note that there are issues with the definition and achievement of coherency in pooling techniques of this setting, with French [15] commenting on several theorems which demonstrate that there is no combining rule simultaneously meeting the entirety of a particular set of desirable criteria. Dawid et al. [10] provide a more

formal discussion on this notion of coherency, and which criteria various approaches obey. We comment that while the below properties are not an exhaustive list of rationality characteristics they are nevertheless important ones for a method of our ilk to obey.

- *Property 1:* $w_j \geq 0$ for all $j = 1,\ldots,n$, with $w_j = 0$ if and only if $f_j(R = r) = 0$.
- *Property 2:* If $\alpha_{i,j} < \alpha_{i,k}$ and $w_j < w_k$ then $\alpha^*_{i,j} < \alpha^*_{i,k}$.
- *Property 3:* If $\alpha_{i,j} = \alpha_{i,k}$ and $w_j = w_k$ then $\alpha^*_{i,j} = \alpha^*_{i,k}$.
- *Property 4:* If $\alpha_{i,j} < \alpha_{i,k}$ and $w_j > w_k$ then any of the following may occur, depending on differences between initial weights, and updated reliability measures:
  - $\alpha^*_{i,j} < \alpha^*_{i,k}$
  - $\alpha^*_{i,j} = \alpha^*_{i,k}$
  - $\alpha^*_{i,j} > \alpha^*_{i,k}$

Further justification for the PI approach can be found in Bolger and Houlding [2]. We do not expand upon this in great detail here, but comment that a simulation study of the form discussed in the following section is carried out, which demonstrates that the PI approach is superior to a collection of alternatives in a high proportion of cases. This superiority is particularly evident as the number of individuals involved in the process and/or the number of returns that are witnessed grows. In summary, the PI approach is not formally derived for a collection of basic axioms. However, as discussed in Section 1, such a derivation is often not possible – indeed French [15] comments on an impossibility theorem detailing how there is no combining rule simultaneously meeting all of a set of desirable criteria. Hence while our proposed approach may not be unique in obeying the four properties discussed above, and indeed there may be other properties which it fails to obey, the fact that it meets them while outperforming a set of reasonable alternatives in a broad range of situations makes it an attractive technique for use in suitable problems.

3. Simulation study

The properties above supply some justification for the PI approach, but we aspire to more robust evidence. We conducted a large-scale simulation study, contemplating two different scenarios:

- The individual problem: Here we considered if it was in the best interest of a DM to use the combined PI distribution, or if she would be better served by solely heeding her own opinion, ignoring those of her neighbours.
- The group problem: Here we assumed that a combined distribution of some form must be constructed, e.g., in a group decision making setting (where a single decision is made by the group as a whole) or simply in an instance in which a DM is committed to listening to the opinions of her neighbours, but is unsure of the optimal manner of doing so. An obvious action is to compare the performance of this opinion pooling method to several alternatives, to investigate the cases in which it may be deemed as optimal. Arguably the simplest opinion pooling approach is one which applies equals weights (termed the EQ method) to all individuals at all epochs, independent of their perceived reliability. This "wisdom of crowds" technique has obvious advantages and disadvantages: the knowledge provided by accurate neighbours may potentially be overshadowed by that of inaccurate individuals, but if all neighbours have reasonably accurate opinions then this approach may perform well (especially in a scenario in which DMs learn over time, and hence become more reliable). This PI approach assigns weights based on perceived reliability, while the EQ method keeps weights constant. An extreme weighting scheme would be to assign a weight of one to the individual deemed most reliabile (i.e., returning the highest PI weight in Eq. (4)) and to disregard the opinions of others. Advantages and disadvantages are evident here also: if the outcome witnessed is an unlikely one then all weight will be given to a DM who is actually deeply unreliable, yet if there is only one DM who possesses an accurate opinion (and no extremely unlikely events occur) then it makes sense to listen only to her and to disregard the inaccurate opinions of all others. The PI, EQ and most reliable (MR) approach can be seen as forming a spectrum of opinion pooling, with the MR and EQ methods at the far ends (listening only to one DM and listening to all equally), with the PI method (listening to all DMs but in a weighted fashion) lying between these. Hence the EQ and MR approaches seem rational alternatives to compare the PI performance to. In addition, these are the benchmarks frequently considered (e.g., in Eggstaff et al., [13]) in attempting to provide justification in problems of this ilk.

A natural question regards the choice of metric to determine if the PI method is superior to the alternative(s) in any particular instance. Many metrics/scoring rules exist which take as its arguments the prediction of an individual and the observation witnessed, e.g., the commonly used quadratic scoring rule. Note in our simulation study (whose sole purpose is to provide validation for the PI approach) we may consider the true underlying value of θ (i.e., that which is an argument of the data generating mechanism) rather than noisy realisations of it. Our goal in this instance is to determine how accurately the combined posterior distributions of a set of methods estimate an unknown parameter θ. Hence as our scoring rule we consider comparing the probability density placed *a posteriori* on the true value of θ by the distributions of the various methods. Again we stress that in practice this true value will never be known, and is only assumed to be known here for model validation. An advantage of this metric over alternatives is that it implicitly takes both the mean and variance of the distributions into account (i.e., a distribution with a highly accurate mean would be penalised for having an extremely high variance) and hence can be seen as appropriate in our scenario which considers full probability distributions rather than point estimates.

In the group problem we declare the optimal method as that maximising this posterior density, i.e., we examine the density placed on the true value of θ by the PI, EQ and MR posteriors, with the technique considered superior being that which thought this value was most likely to be correct. In the individual problem we consider the density placed by the PI posterior, and the respective posteriors of individuals, on this true θ, declaring the PI approach optimal if it places more density than the distributions of over half the DMs. If this is the case then if an individual is randomly chosen from within the group there is a probability exceeding 0.5 that the PI distribution will lead to better estimation then her own. Recall that, *a priori*, DMs do not know if they are accurate information sources or not.

We use three common conjugate cases, outlined in Table 1. For each of these cases we contemplate three sub-cases, in which DMs prior distributions on average respectively overestimate, underestimate, and are centred on, the true value of θ. In each of these nine instances we consider a variety of settings, varying the number of DMs inherent in the neighbourhood and the number of returns observed. For each particular case we firstly simulate prior parameters from the outlined distributions. Note that the uniform distributions from which these parameters are chosen are simply a device which we use in this study to help us consider various cases, such as that in which DMs on average underestimate θ, and is not reflective of any inherent consensus between DMs or indeed any facet of the true data generating mechanism. Next we simulate a return from the associated data generating mechanism, update weights and distributions, and record which technique is superior under the metrics discussed above. We continue to do this iteratively in order to see how results vary with the number of data points observed. In each case we repeat this process a suitably large number of times (5,000 times for the results fiven below) and record the proportion of times that the various approaches are super-

Table 1
True data mechanisms and parameters, prior structures and distributions over the simulation of prior parameters, as well as the corresponding average prior means.

	Beta-Binomial	Normal-Normal	Gamma-Poisson
Data Mechanism	$R \sim Binomial(5, \theta)$	$R \sim \mathcal{N}(\theta, 1)$	$R \sim Poisson(\theta)$
θ	0.5	0	5
Priors	$f_i(\theta) \sim Beta(\beta_1, \beta_2)$	$f_i(\theta) \sim \mathcal{N}(m_i, s_i)$	$f_i(\theta) \sim Gamma(\beta_1, \beta_2)$
Overestimation	$\beta_1 \sim U(1, 16)$	$m_i \sim U(-2, 8)$	$\beta_1 \sim U(1, 4)$
	$\beta_2 \sim U(1, 4)$	$s_i \sim U(0, 3)$	$\beta_2 \sim U(1, 32)$
	$\mathbb{E}_{fi}(\theta) = 0.8$	$\mathbb{E}_{fi}(\theta) = 3$	$\mathbb{E}_{fi}(\theta) = 8$
Underestimation	$\beta_1 \sim U(1, 6)$	$m_i \sim U(-6, 2)$	$\beta_1 \sim U(1, 4)$
	$\beta_2 \sim U(1, 14)$	$s_i \sim U(0, 3)$	$\beta_2 \sim U(1, 16)$
	$\mathbb{E}_{fi}(\theta) = 0.3$	$\mathbb{E}_{fi}(\theta) = -2$	$\mathbb{E}_{fi}(\theta) = 4$
Mean-Centred	$\beta_1 \sim U(1, 10)$	$m_i \sim U(-8, 8)$	$\beta_1 \sim U(1, 4)$
	$\beta_2 \sim U(1, 10)$	$s_i \sim U(0, 4)$	$\beta_2 \sim U(1, 20)$
	$\mathbb{E}_{fi}(\theta) = 0.5$	$\mathbb{E}_{fi}(\theta) = 0$	$\mathbb{E}_{fi}(\theta) = 5$

ior, declaring the overall superior technique for that particular instance to be that which is superior most frequently. Pseudo-code for this process is included in the Appendix, highlighting the step-by-step process underlying this simulation study.

An illustration of the results is given in Fig. 1 for the group Normal overestimation case. As the number of DMs and/or returns increases the prevalence of the PI approach being superior to alternatives also increases. This is as we would expect, given the learning that is inherent within the technique. The conclusions inferred here are consistent with those found under the various other initialisations. We see that the PI approach is certainly meritorious, outperforming the considered alternatives given a suitable amount of DMs in the group/returns witnessed.

4. Theoretical calculations

4.1. Derivation of probabilities

When conducting our study we found that as the number of simulations increased the proportion of times that the PI approach was superior to the alternatives appears to deviate less and less, i.e., it seems to converge asymptotically. Hence the proportion of times the PI approach was superior having ran our simulations 5000 times was very similar to its success proportion after 1000 iterations. The simulation study from the previous section constructed empirical estimates for the probability that the PI approach was the superior technique. In this section we illustrate how the true underlying probabilities that these proportions are tending towards can be explictly calculated using mathematical operations, and illustrate this convergence in practice with a series of numerical examples. For the sake of brevity we limit our calculations to the Beta-Binomial case, but provide comments on extension to the Normal-Normal and Poisson-Gamma cases that follow straightforwardly from this. We assume that n is odd (i.e., no ties can occur in the individual problem) but this material is easily modified if not. In the Beta-Binomial case θ is a Bernoulli success probability that each P_i has a Beta (α_i, β_i) prior over:

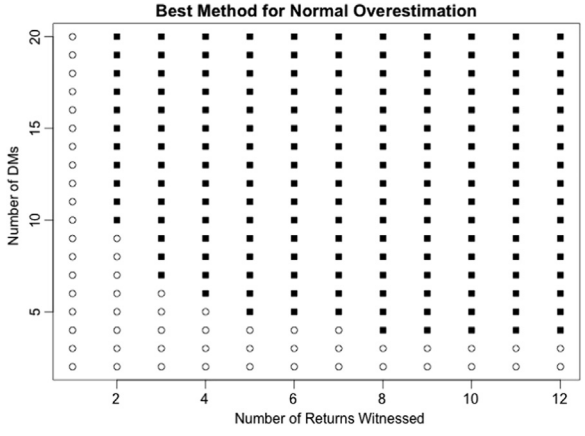

Fig. 1. Plot of the method with the highest success proportion in Normal Overestimation for a varying amount of DMs (x-axis) and returns (y-axis). Filled squares imply that the PI method is superior, with unfilled circles and filled triangles for the EQ and MR methods respectively.

$$f_i(\theta) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\theta^{\alpha_i - 1}(1-\theta)^{\beta_i - 1} \quad \text{with } \theta \in [0, 1] \tag{8}$$

Returns are realisations of a Binomial random variable, R. Each r denotes a number of successes in m independent and identically distributed Bernoulli trials. Each particular value of r occurs with a probability of

$$f(R = r|\theta) = \binom{m}{r}\theta^r(1-\theta)^{m-r} \quad \text{with } r = 0, 1, \ldots, m \tag{9}$$

Over t epochs there are $(m+1)^t$ possible t- tuples of returns that can be witnessed. From Eq. (7), the unnormalised weight assigned to P_i after t returns, $u_{i,t}$, is a product of her PI weights over these t returns, and her initial weight, which we here assume to be $\frac{1}{n}$. Denoting r_k as the return at the k^{th} epoch and $w_{i,k}$ for the PI weight of P_i over r_k we have:

$$u_{i,t} = \frac{1}{n}w_{i,1}\ldots w_{i,t} = \frac{1}{n}f_i(R_1 = r_1)\ldots f_i(R_t = r_t|R_{t-1} = r_{t-1}\ldots, R_1 = r_1) \tag{10}$$

Note that each of $R_1, \ldots, R_t$ is simply a random variable which follows an identical distribution to that in Eq. (9) (as we assume that θ is a static variable which does not change over time), with $r_1, \ldots, r_t$ being realisations of these. Updating can be conducted using Beta-Binomial conjugacy, with the updated hyperparameters of P_i after k returns, denoted by $\alpha_i^{(k)}$ and $\beta_i^{(k)}$, found to be:

$$\alpha_i^{(k)} = \alpha_i + \sum_{j=1}^{k} r_j \quad \text{and} \quad \beta_i^{(k)} = \beta_i + km - \sum_{j=1}^{k} r_j \tag{11}$$

Using the convention that $\alpha_i^{(0)} = \alpha_i$ and $\beta_i^{(0)} = \beta_i$ we may rewrite Eq. (10) as

$$u_{i,t} = \frac{1}{n}\prod_{k=1}^{t}\binom{m}{r_k} \frac{(\alpha_i^{(k-1)}+\beta_i^{(k-1)}-1)!(\alpha_i^{(k-1)}+r_k-1)!(\beta_i^{(k-1)}+m-r_k-1)!}{(\alpha_i^{(k-1)}-1)!(\beta_i^{(k-1)}-1)!(\alpha_i^{(k-1)}+\beta_i^{(k-1)}+m-1)!} \tag{12}$$

By the independence assumption, the probability of any particular return set $\{r_1,\ldots,r_t\}$ is a product of Binomial distributions:

$$f(R_1 = r_1,\ldots,R_t = r_t|\theta) = \prod_{k=1}^{t}\binom{m}{r_k}\theta^{r_k}(1-\theta)^{m-r_k} \tag{13}$$

Any value of Eq. (12) occurs with the associated probability found in Eq. (13). The normalised weight that is afforded to P_i after t returns, here denoted as $\gamma_{i,t}$ to avoid notational confusion, is

$$\gamma_{i,t} = \frac{u_{i,t}}{\sum_{j=1}^{n}u_{j,t}} = \frac{\prod_{k=1}^{t}\binom{m}{r_k}\frac{(\alpha_i^{(k-1)}+\beta_i^{(k-1)}-1)!(\alpha_i^{(k-1)}+r_k-1)!(\beta_i^{(k-1)}+m-r_k-1)!}{(\alpha_i^{(k-1)}-1)!(\beta_i^{(k-1)}-1)!(\alpha_i^{(k-1)}+\beta_i^{(k-1)}+m-1)!}}{\sum_{j=1}^{n}\prod_{k=1}^{t}\binom{m}{r_k}\frac{\left(\alpha_j^{(k-1)}+\beta_j^{(k-1)}-1\right)!\left(\alpha_j^{(k-1)}+r_k-1\right)!\left(\beta_j^{(k-1)}+m-r_k-1\right)!}{\left(\alpha_j^{(k-1)}-1\right)!\left(\beta_j^{(k-1)}-1\right)!\left(\alpha_j^{(k-1)}+\beta_j^{(k-1)}+m-1\right)!}} \tag{14}$$

Weights in Eq. (14) are then merged with distributions in Eq. (8), with the updated hyperparameters from Eq. (11), to give a PI posterior after t returns of

$$\hat{f}_t^{PI}(\theta|r_1,\ldots,r_t) = \sum_{z=1}^{n}\gamma_{z,t}f_z(\theta|r_1,\ldots,r_t) = \sum_{z=1}^{n}\left[\frac{\prod_{k=1}^{t}\binom{m}{r_k}\frac{(\alpha_z^{(k-1)}+\beta_z^{(k-1)}-1)!(\alpha_z^{(k-1)}+r_k-1)!(\beta_z^{(k-1)}+m-r_k-1)!}{(\alpha_z^{(k-1)}-1)!(\beta_z^{(k-1)}-1)!(\alpha_z^{(k-1)}+\beta_z^{(k-1)}+m-1)!}}{\sum_{j=1}^{n}\prod_{k=1}^{t}\binom{m}{r_k}\frac{\left(\alpha_j^{(k-1)}+\beta_j^{(k-1)}-1\right)!\left(\alpha_j^{(k-1)}+r_k-1\right)!\left(\beta_j^{(k-1)}+m-r_k-1\right)!}{\left(\alpha_j^{(k-1)}-1\right)!\left(\beta_j^{(k-1)}-1\right)!\left(\alpha_j^{(k-1)}+\beta_j^{(k-1)}+m-1\right)!}} \times \frac{\Gamma(\alpha_z^{(t)}+\beta_z^{(t)})}{\Gamma(\alpha_z^{(t)})\Gamma(\beta_z^{(t)})}\theta^{\alpha_z^{(t)}-1}(1-\theta)^{\beta_z^{(t)}-1}\right] \tag{15}$$

Over t epochs there are $(m+1)^t$ possible return streams. Denote by $\boldsymbol{r}_x$ the xth of these. Consider an indicator variable, $I_{j,x}$, returning a 1 if the PI posterior places more density on θ than the posterior of P_j, given the return set $\boldsymbol{r}_x$ has been witnessed, and 0 if not, i.e.,

$$I_{j,x} = \begin{cases} 1 & \text{if } \hat{f}_t^{PI}(\theta|\boldsymbol{r}_x) > f_j(\theta|\boldsymbol{r}_x); \\ 0 & \text{if } \hat{f}_t^{PI}(\theta|\boldsymbol{r}_x) < f_j(\theta|\boldsymbol{r}_x). \end{cases}$$

This forms a matrix of zeros and ones, generically outlined in Table 2, with our interest lying in its column sums. If, for a set of returns $\boldsymbol{r}_x$, this sum exceeds $\frac{n}{2}$ then, for $\boldsymbol{r}_x$, the PI posterior distribution gives better estimation than over half of the posteriors of DMs, i.e., it is superior for the individual problem. For the xth column this sum is S_x:

$$S_x = \sum_{k=1}^{n} I_{k,x} \tag{16}$$

We introduce an indicator variable, I_x, which returns a 1 if this sum exceeds $\frac{n}{2}$, and a 0 if not:

$$I_x = \begin{cases} 1 & \text{if} S_x > \frac{n}{2}; \\ 0 & \text{if} S_x < \frac{n}{2}. \end{cases}$$

Finally we consider the quantities $\{I_x\}_{x=1,\ldots,(m+1)^t}$ and their corresponding probabilities for $\boldsymbol{r}_x$, as in Eq. (13). The probability that the PI approach is superior to DM distributions is the cross-product of these two vectors:

$$\mathbb{P}(\text{PI is superior}|\theta) = \sum_{x=1}^{(m+1)^t} I_x f(\boldsymbol{R} = \boldsymbol{r}_x|\theta) \tag{17}$$

We perform similar calculations to find which method is optimal in the group problem. We defined the PI distribution after t returns in Eq. (15), and now turn our attention to the EQ and MR distributions. The former distribution is straightforward:

$$\hat{f}_t^{EQ}(\theta|r_1,\ldots,r_t) = \sum_{z=1}^{n}\frac{1}{n}f_z(\theta|r_1,\ldots,r_t) = \frac{1}{n}\sum_{z=1}^{n}\left[\frac{\Gamma(\alpha_z^{(t)}+\beta_z^{(t)})}{\Gamma(\alpha_z^{(t)})\Gamma(\beta_z^{(t)})}\theta^{\alpha_z^{(t)}-1}(1-\theta)^{\beta_z^{(t)}-1}\right] \tag{18}$$

The MR posterior distribution depends upon which DM was deemed to be the most reliable (i.e., returned the highest PI weight) at the t^{th} epoch. We consider the indicator variable $I_{j,x}^{MR}$, returning a 1 if P_j has the biggest PI weight having seen $\boldsymbol{r}_x$, and 0 if not, where $\boldsymbol{r}_x[t]$ is the tth element of $\boldsymbol{r}_x$:

$$I_{j,x}^{MR} = \begin{cases} 1 & \text{if} f_j(R_t = \boldsymbol{r}_x[t]|\cdot) = \max_{k\in\{1,\ldots,n\}}f_k(R_t = \boldsymbol{r}_x[t]|\cdot); \\ 0 & \text{if} f_j(R_t = \mathrm{r}_x[t]|\cdot) \neq \max_{k\in\{1,\ldots,n\}}f_k(R_t = \boldsymbol{r}_x[t]|\cdot). \end{cases}$$

The MR posterior distribution after $\boldsymbol{r}_x$ is thus defined as

$$\hat{f}_t^{MR}(\theta|\boldsymbol{r}_x) = \sum_{z=1}^{n}I_{z,x}^{MR}f_z(\theta|\boldsymbol{r}_x) = \sum_{z=1}^{n}I_{z,x}^{MR} \times \frac{\Gamma(\alpha_z^{(t)}+\beta_z^{(t)})}{\Gamma(\alpha_z^{(t)})\Gamma(\beta_z^{(t)})}\theta^{\alpha_z^{(t)}-1}(1-\theta)^{\beta_z^{(t)}-1} \tag{19}$$

Consider the following three indicator variables, I_x^{PI}, I_x^{EQ} and I_x^{MR}, which respectively return a 1 if the method is optimal (in terms of maximising posterior density) given $\boldsymbol{r}_x$, and a zero if not:

$$I_x^{PI} = \begin{cases} 1 & \text{if} \hat{f}_t^{PI}(\theta|\boldsymbol{r}_x) > \max\{\hat{f}_t^{EQ}(\theta|\boldsymbol{r}_x), \hat{f}_t^{MR}(\theta|\boldsymbol{r}_x)\}; \\ 0 & \text{if not.} \end{cases} \quad I_x^{EQ} = \begin{cases} 1 & \text{if} \hat{f}_t^{EQ}(\theta|\boldsymbol{r}_x) > \max\{\hat{f}_t^{PI}(\theta|\boldsymbol{r}_x), \hat{f}_t^{MR}(\theta|\boldsymbol{r}_x)\}; \\ 0 & \text{if not.} \end{cases} \quad I_x^{MR} = \begin{cases} 1 & \text{if} \hat{f}_t^{MR}(\theta|\boldsymbol{r}_x) > \max\{\hat{f}_t^{PI}(\theta|\boldsymbol{r}_x), \hat{f}_t^{EQ}(\theta|\boldsymbol{r}_x)\}; \\ 0 & \text{if not.} \end{cases}$$

Hence we can define the probabilities of techniques being optimal in an analogous fashion to as in Eq. (17):

$$\mathbb{P}(\text{PI is superior}|\theta) = \sum_{x=1}^{(m+1)^t} I_x^{PI} f(\boldsymbol{R} = \boldsymbol{r}_x|\theta) \tag{20}$$

$$\mathbb{P}(\text{EQ is superior}|\theta) = \sum_{x=1}^{(m+1)^t} I_x^{EQ} f(\boldsymbol{R} = \boldsymbol{r}_x|\theta) \tag{21}$$

$$\mathbb{P}(\text{MR is superior}|\theta) = \sum_{x=1}^{(m+1)^t} I_x^{MR} f(\boldsymbol{R} = \boldsymbol{r}_x|\theta) \tag{22}$$

The calculation details for the Poisson-Gamma and Normal-Normal

Table 2
Cross-tabulation of DMs $P_1,\ldots,P_n$ and return streams $r_1,\ldots,r_{(m+1)^t}$, with $I_{j,x}$ for each cell $\{j,x\}$.

	r_1	r_2	$\cdots$	$r_{(m+1)^t}$
P_1	$I_{1,1}$	$I_{1,2}$	$\cdots$	$I_{1,(m+1)^t}$
P_2	$I_{2,1}$	$I_{2,2}$	$\cdots$	$I_{2,(m+1)^t}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
P_n	$I_{n,1}$	$I_{n,2}$	$\cdots$	$I_{n,(m+1)^t}$

cases are similar to those outlined above for Beta-Binomial conjugacy, with some minor modifications. In the Binomial case we could produce a (finite) exhaustive list of all possible values of r that can occur at each epoch, i.e., there are $m + 1$. The same cannot be said in the Poisson case were there is theoretically (countably) infinite set of possible values that may occur at each epoch. Eqs. (17) and (20)–(22) must be finite sums to ensure computability. Hence in the Poisson case we must choose a (finite) upper bound value that returns have negligibly small probability of exceeding. In the continuous Normal distribution, which has an (uncountably) infinite set of returns, we perform a similar process, discretising the range and choosing suitable lower and upper bounds, which have negligibly small probabilities of being deceeded or exceeded respectively. In this case we must ensure to perform discretisation in a manner that is not too coarse, to ensure accuracy of results.

4.2. Illustrations of convergence

We show convergence in each of the conjugate cases considered, with parameterisations given below. The true probabilities and simulated proportions are given in Table 3. We use 5000 simulations in the Beta-Binomial/Poisson-Gamma cases, but are required to use twice this to ensure convergence in the (discretised) Normal-Normal case.

- Beta-Binomial Conjugacy: Suppose $R \sim \text{Bin}(2, \theta)$, with $\theta = 0.7$, and that we have five DMs who will witness four returns, with respective Beta(1,3), Beta(3,2), Beta(7,2), Beta(4,3) and Beta(2,1) priors. Convergence is illustrated in Fig. 2.
- Poisson-Gamma Conjugacy: Suppose $R \sim \text{Pois}(\theta)$ with $\theta = 2$, and that we have three DMs who will witness three returns, with respective Gamma(1,3), Gamma(7,2) and Gamma(2,2) priors. We choose $r = 8$ as the upperbound for R, a choice ensuring $\mathbb{P}(R > 8) < 0.001$. Convergence is illustrated in Fig. 3.
- Normal-Normal Conjugacy: Suppose $R \sim \mathcal{N}(\theta, 1)$, with $\theta = 0$, and that we have three DMs who will witness two returns, with respective $\mathcal{N}(-1, 1)$, $\mathcal{N}(4, 4)$ and $\mathcal{N}(5, 3)$ priors. We choose the lower bounds and upper bounds for the return space to be $\theta \pm 3(1)$, i.e., over 99% of returns will lie in this range. We discretise the range into eighty segments of equal length. Convergence is illustrated in Fig. 4.

5. Discussion

The above examples are low-dimensional, as there are issues related to “curse of dimensionality”. If there are N potential returns per epoch then there are N^t potential return streams over t epochs. This term grows quickly, often making computation in a reasonable time impossible. Yet we have seen how accurately simulated proportions mirror probabilities. Hence it is adequate to talk in terms of proportions rather than probabilities, as the former are accurate estimates of the latter, with severely decreased computation time. In addition, in non-conjugate cases it is impossible to give closed form PI weights - simulation is required, i.e., probabilities cannot be calculated, only approximated by proportions. Nevertheless, the material in the previous section represents an attractive mathematical derivation, which confirms the correctness of our simulation method.

Table 3
The true probabilities, and associated simulated proportions, for the three specific distributional parameterisations discussed in the text.

		Ind. problem	Group problem		
Case	Quantity	PI	PI	EQ	MR
Binomial	True Probabilities	0.687	0.460	0.362	0.178
Binomial	Simulated Proportions	0.686	0.462	0.366	0.172
Poisson	True Probabilities	0.963	0.448	0.146	0.406
Poisson	Simulated Proportions	0.959	0.444	0.146	0.410
Normal	True Probabilities	0.508	0.251	0.342	0.401
Normal	Simulated Proportions	0.505	0.250	0.345	0.404

We commented in Section 2 that we often assume for convenience that all DMs agree on a common likelihood function $f(r|\theta)$. What would happen if this, not unreasonably, was not the case i.e., if each participant P_i was to supply her own subjectively perceived likelihood function $f_i(r|\theta)$? This would not impact the practical application of the PI approach. All that would be required is for individuals to agree on the support for the quantity of interest, for instance the unit interval or the whole real line. Complications may arise if this was not the case as it could lead to probabilities of zero being placed on the witnessed outcome in Eq. (4) and hence numerical issues, e.g., if a negative return was witnessed and a DM had represented her opinion via a truncated Normal distribution over the positive real line. Likewise individuals should agree over whether the quantity of interest is discrete or continuous, i.e., whether statements should be made in terms of probability mass or probability density. Each DM has a subjective prior opinion over θ, observes some data (which they may have subjective beliefs over concerning the functional form of the data generating mechanism which it has arisen from) and updates their prior in light of this to gain their posterior distribution. In the case of distinct likelihoods then Eq. (4) can still be applied, with numerical methods used to solve the non-conjugate cases. The theoretical calculations of Section 4 could also be modified to accommodate the increased flexibility of differing likelihoods, although this lack of conjugacy would lead to an approximation of the probability of the PI approach being superior to alternatives rather than an exact value. In summary, under minimal assumptions concerning support our method allows differing likelihoods, although this increases computational complexity and the need for numerical methods.

The assumption has been made that DMs are capable of quantifying their uncertainty over the element of interest by use of a probability distribution. A justification for this assumption is that several elicitation methods do exist for DMs who are not statistically literate (e.g., O'Hagan, [28]) , with techniques allowing DMs to construct full probability distributions from some basic specifications and questions (e.g., “What range do you believe 95% of observations will lie within?”). Nevertheless difficulties may arise for individuals trying to express their opinion. The Bayes Linear method [20] can be seen as a useful in this case, whereby an individual simply supplies a mean and variance for their belief and no higher order moments or distributional assumptions. Wisse et al. [34] provide discussion on the application of this technique to a linear opinion pooling setting. Briefly we mention two areas for further research pertaining to that discussed above. The difficulties discussed above can be viewed as a motivation to generalise the technique in Section 2 to a setting in which DMs can supply their opinions in a more simplistic way, e.g., nonparametrically. This is an area for future consideration. In addition, investigation can be carried out to compare the results obtained from combining probability density functions (as considered here) to those from combining corresponding cumulative distribution functions, in the vein of the research conducted in Lichtendahl et al. [23].

Finally we comment on the context in which we discuss assigning weights for a linear pool. We consider a setting where each individual providing an opinion to be pooled is herself a DM, who will make a decision using her own utility function. Hence there is a novelty inherent within our approach, with the returns witnessed by a DM (and, by the inherent propagation of information, the group as a whole) being dependent upon her utility function. The approach can also be seen as highly objective (assuming it is initialised with equal weights), with weight updating based purely on the discrepancies between witnessed returns and DM's opinions, and fully objective in the case of initial weights being equal. We believe that this method will be

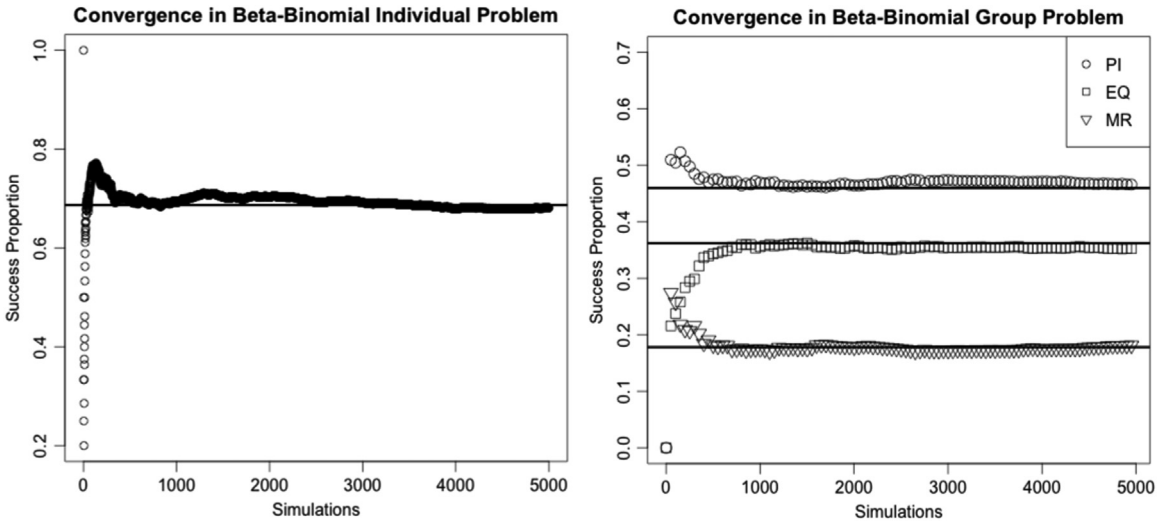

Fig. 2. Convergence in the Beta-Binomial case for the Individual (LHS) and Group (RHS) problems.

extremely useful in realistic implementation, providing increasingly accurate parameter estimation for individuals making decisions, and hence potentially substantially boosting their corresponding respective decision qualities.

Acknowledgements

The authors would like to thank the anonymous reviewers whose suggestions and additional references have helped to greatly improve this research and its presentation.

Appendix

Below we provide pseudo-code for the simulation study discussed in Section 3. We begin by illustrating for the individual problem.

- *Step 1:* Simulate hyperparameters for the n DMs from the appropriate distributions (e.g., those for the Beta-Binomial Overestimation case) in Table 1.
- *Step 2:* Construct an equally weighted combination of the prior distributions resulting from these hyperparameters.
- *Step 3:*. Simulate a return from the corresponding data generating mechanism (e.g., the Binomial distribution) given in Table 1.
- *Step 4:* Update individual distributions and normalised weights in light of this return, and construct the Plug-in posterior distribution.
- *Step 5:* Repeat Steps 3–4 for as many returns as are required in this instance (e.g. , twice if the case we are interested in involves three returns).
- *Step 6:* Compare the density placed on the true value of θ (e.g., 0.5 in the Binomial case) by the combined distribution to the densities placed by the individual distributions, declared the PI approach superior if its posterior places more density than over half of the individual posteriors.
- *Step 7:* Repeat Steps 1–6 a suitably large (e.g., 5000) amount of times.
- *Step 8:* Aggregate results from Step 6 across the number of simulations run, determining what proportion of time the PI approach was declared superior to individual distributions.
- *Step 9:* Declare the Plug-in approach as the best approach if it was superior to the alternative for more than half of the number of simulations run.

The algorithm for the group code is similar to that for the individual problem, with differences only occurring after Step 3:

- *Step 4:* Update individual distributions and normalised weights in

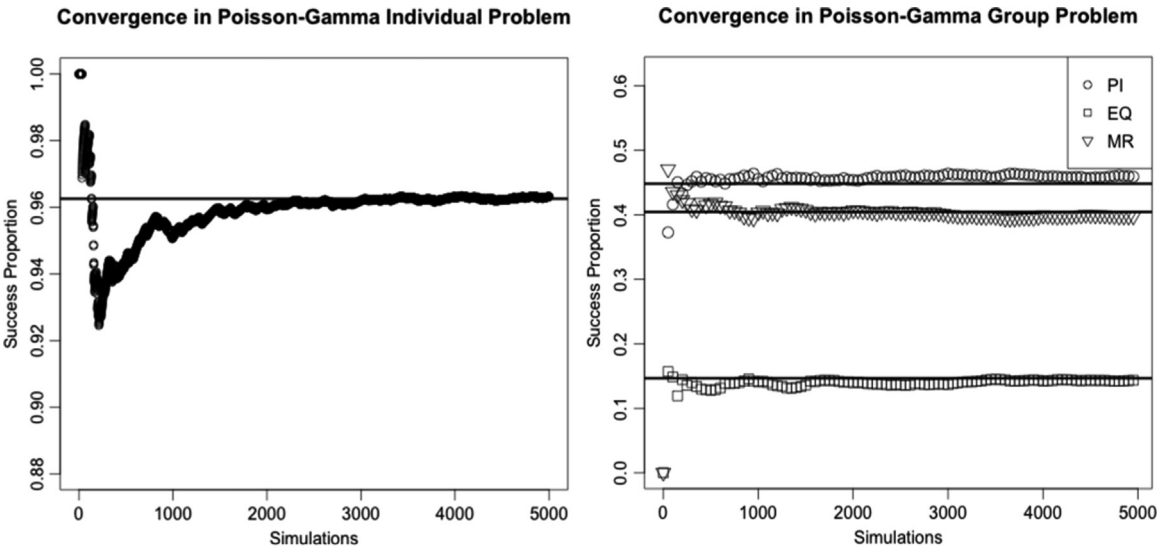

Fig. 3. Convergence in the Poisson-Gamma case for the Individual (LHS) and Group (RHS) problems.

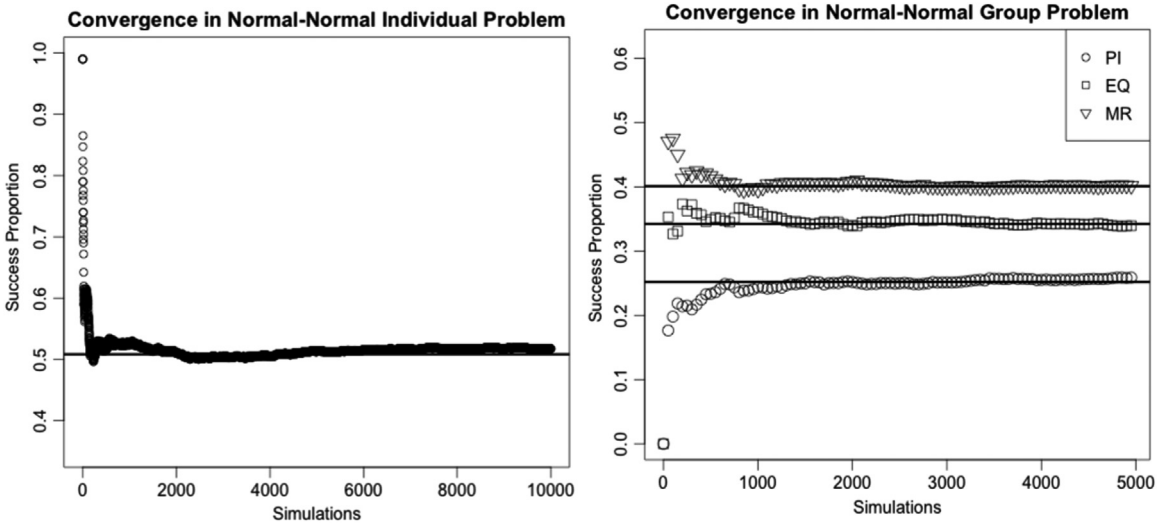

Fig. 4. Convergence in the Normal-Normal case for the Individual (LHS) and Group (RHS) problems.

light of this return. Determine which individual is "most reliable", i.e., the individual with the highest Plug-in weight, who will receive a weight of one under the Most Reliable method. Construct the Plug-in, Equal Weights and Most Reliable posterior distributions.

- *Step 5:* Repeat Steps 3–4 for as many returns as are required in this instance.
- *Step 6:* Compare the density placed on the true value of θ by the Plug-in, Equal Weights and Most Reliable posterior distributions, declared the superior technique as that return the highest value.
- *Step 7:* Repeat Steps 1–6 a suitably large (e.g., 5000) amount of times.
- *Step 8:* Aggregate results from Step 6 across the number of simulations run, determining what proportion of time each of the three methods was superior to the other.
- *Step 9:* Declare the best technique in this instance as that which was superior to the others the greatest proportion of times.

References

[1] Bates JM, Granger CWJ. The combination of forecasts. Oper Res 1969;20:451–68.
[2] Bolger D, Houlding B. Reliability updating in linear opinion pooling for multiple decision makers. Proc Inst Mech Eng, Part O: J Risk Reliab 2016;230:309–22.
[3] Bunn DW. A Bayesian approach to the linear combination of forecasts (1975). Oper Res Q 1975;26:325–9.
[4] Clemen RT. Comments on cooke's classical method. Reliab Eng Syst Saf 2008;93:760–5.
[5] Clemen RT, Winkler RL. Combining probability distributons from experts in risk analysis. Risk Anal 1999;19:187–203.
[6] Cooke R. Experts in uncertainty, opinion and subjective probability in science, environmental ethics and science policy series. Oxford University Press; 1991.
[7] Cooke R. Expert judgement studies. Reliab Eng Syst Saf 2007;93:655–77.
[8] Cooke R, Goossens LLHJ. TU delft expert judgement data base. Reliab Eng Syst Saf 2008;93:657–74.
[9] Cox LA, Jr T. Confronting deep uncertainties in risk analysis. Risk Anal 2012;32:1607–29.
[10] Dawid AP, DeGroot MH, Mortera J, Cooke R, French S, Genest C, et al. Coherent combinations of experts' opinions. Test 1995;4:263–313.
[11] DeGroot MH. Reaching a Consensus. J Am Stat Assoc 1974;69:121–81.
[12] DeGroot M, Bayarri M. Gain Weight: A Bayesian approach. department of statistics, Carnegie Mellon University. Technical report; 1987.
[13] Eggstaff JW, Mazzuchi TA, Sarkana S. The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model. Reliab Eng Syst Saf 2014;121:72–82.
[14] Flandoli F, Giorgi E, Aspinall WP, Neri A. Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. Reliab Eng Syst Saf 2011;96:1292–310.
[15] French S. Group consensus probability distributions: a critical study. Bayesian Stat 1985;2:183–97.
[16] French S. Revista de la real academia de ciencias exactas, fisicas y naturales, serie a. matematicas. Aggregating Expert Judgm 2011;105:181–206.
[17] Genest C, McConway KJ. Allocating the weights in the linear opinion pool. J Forecast 1990;9:53–73.
[18] Genest C, Zidek JV. Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1986;1:114–35.
[19] Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359–78.
[20] Goldstein, Wooff. Bayes linear statistics, theory and methods. wiley series in probability and statistics; 2007.
[21] Laplace S. Theorie analytique des probabilites, Paris, Ve. Courcier; 1812.
[22] Lee PM. Bayesian statistics: an introduction, 4th ed.. John Wiley and Sons; 2012.
[23] Lichtendahl KC, Jnr, Grushka-Cockayne Y, Winkler RL. Is it better to average probabilities or quantiles?. Manag Sci 2013;59:1594–611.
[24] Morris PA. Decision analysis expert use. Manag Sci 1977;20:1233–41.
[25] Morris PA. Combining expert judgments: a Bayesian approach. Manag Sci 1977;23:679–93.
[26] Morris PA. An axiomatic approach to expert resolution. Manag Sci 1983;29:24–32.
[27] Newbold P, Granger CWJ. Experience with forecasting univariate time series and the combination of forecasts. J R Stat Soc, Ser A 1974;137(2):131–65.
[28] O'Hagan A. Eliciting expert beliefs in substantial practical applications. J R Stat Soc, Ser D 1998;47:21–35.
[29] Ramsey FP. Foundations - essays in philosophy, logic, mathematics and economics. Humanities Press; 1931.
[30] Starr C. Risk criteria for nuclear power plants: a pragmatic proposal. Risk Anal 1981;1:113–20.
[31] van Noortwijk JM, Dekker R, Cooke R, Mazzuchi T. Expert judgment in maintenance optimization. IEEE Trans Reliab 1992;41:427–32.
[32] von Neumann J, Morgenstern O. Theory of games and economic behaviour, 2nd ed.. Princeton University Press; 1944.
[33] Wilson S, McDaid K. Deciding how long to test software. Statistician 2001;50:117–34.
[34] Wisse B, Bedford T, Quigley J. Expert judgement combination using moment methods. Reliab Eng Syst Saf 2008;93:675–86.