Ann Oper Res (2015) 229:429–450
DOI 10.1007/s10479-015-1846-0

Calibration, sharpness and the weighting of experts in a linear opinion pool

Stephen C. Hora[1] · **Erim Kardeş**[1]

Published online: 28 March 2015

Abstract Linear opinion pools are the most common form of aggregating the probabilistic judgments of multiple experts. Here, the performance of such an aggregation is examined in terms of the calibration and sharpness of the component judgments. The performance is measured through the average quadratic score of the aggregate. Trade-offs between calibration and sharpness are examined and an expression for the optimal weighting of two dependent experts in a linear combination is given. Circumstances where one expert would be disqualified are investigated. Optimal weights for the multiple, dependent experts are found through a concave quadratic program.

Keywords Probabilistic expert judgment · Subjective probability · Scoring rules · Subject matter expert · Calibration · Quadratic score · Brier score

1 Introduction

The use of subject matter experts is common in risk and decision analyses. These experts are used to help quantify models when data and first principles are insufficient for this purpose. For example, experts have been used in nuclear power safety analyses [(Hora and Iman 1989), (Rasmussen et al. 1975)], studies of health risks (Morgan and Henrion 1990), and

This research was supported by the United States Department of Homeland Security through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) under award number 2010-ST-061-RE0001. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the United States Department of Homeland Security, or the University of Southern California, or CREATE.

✉ Erim Kardeş
erimkardes@hotmail.com

Stephen C. Hora
Hora@USC.edu

[1] University of Southern California (USC), 3710 McClintock Ave., Los Angeles, CA 90089, USA

decisions regarding terrorism countermeasures [(Department of Homeland Security 2006), (National Research Council 2008)]. In each of these studies, the experts were asked to provide probability density functions for quantities in the respective model. These densities, and their corresponding distribution functions, are often referred to as uncertainty distributions and the practice of obtaining these uncertainty distributions is known as probability elicitation.

The value of uncertainty distributions is that they communicate both what the experts know and what they are uncertain about – what they know they do not know. It is this dual role of communicating both what is known and what is uncertain that makes evaluating the quality of the uncertainty distributions somewhat tricky. Two qualities that are sought are sharpness and calibration. Sharpness refers to the ability to delineate a value by providing a relatively narrow range of possibilities. The narrower the range, the sharper the density. We might speak, for example, of a range in which the target quantity falls with 95 % probability. Assigning a probability to a range brings up the concept of calibration, which is a measure of the faithfulness of the elicited probabilities. Intervals for which a well-calibrated expert assigns a probability of p, should, on average, contain the true value with a relative frequency of p.

Attaining uncertainty distributions that are both sharp and have good calibration can be difficult as the two goals may conflict. Densities that are too sharp will produce intervals that are too narrow to encompass the stated values with the requisite relative frequency and therefore have poor calibration. There has been little research, however, into how these properties trade-off against one another.

Understanding how both sharpness and calibration contribute to good judgments can be valuable on a number of levels. First, in the selection of experts. Should one emphasize calibration, disqualifying experts who may be knowledgeable but poor at expressing their knowledge in terms of probabilities or should one emphasize expertise perhaps disqualifying less knowledgeable experts? Perhaps there is an optimal balance between calibration and sharpness in the selection and qualification of experts. Second, where is it best to allocate resources? On the training of the experts to improve their ability to make probabilistic judgments or on the acquisition and analysis of data and models to improve the knowledge on which the judgments are based. Such a decision could entail the allocation of an expert's time in preparing for a probability elicitation. Third, when multiple experts respond with uncertainty distributions for the same quantity, how should sharpness and calibration influence which judgments to use and whether to combine the judgments? If the judgments are to be combined, can information on sharpness and calibration be used to differentially weight the judgments?

1.1 Summary of contributions

Contributions of this paper can be summarized as follows:

1. It is shown, when a linear aggregation is used in the case of multiple possibly dependent experts, the problem of finding optimal weights can be formulated as a mathematical program. An important result is that this mathematical program is concave for square integrable densities and hence the optimal weights can efficiently be found for a given setting.
2. Building on a result given by Hora (2010), this paper investigates how sharpness and calibration trade-off against each other for a single expert, who provides normal densities for a sequence of commensurable uncertain quantities. The average quadratic score, which is also known as Brier score for continuous quantities, (Brier 1950; Matheson and Winkler 1976) is used for this purpose. The sequence of densities is characterized by two

parameters which facilitates the modeling of an expert with various levels of expertise and degrees of calibration. Furthermore, this trade-off is examined via two other scoring rules, the logarithmic and the spherical scoring rules.

3. We extend the single-expert case to two independent experts whose judgments are aggregated by a linear rule. Expressions for the optimal weights for the two experts are given in terms of differences in sharpness and calibration. The question of whether experts should be disqualified on the basis of inferior expertise (lack of sharpness) or because of poor calibration is examined. The conclusion is that there are circumstances where disqualification is not optimal regardless of the differences and there are situations where disqualification is appropriate and optimal. It is also shown that strictly proper scoring rules do not necessarily enforce propriety (answering truthfully) on the components of the linear aggregate.
4. We provide a brief examination of the impact of dependence among experts. A simulation experiment is used to compare the performance of optimal weights to equal weights as a function of the differentiation among experts. Finally, some suggestions for using the theoretical findings in practical situations are given.

1.2 Outline of the paper

The rest of this section presents related articles in the literature. Section 2 introduces existing definitions and methods in the literature that constitutes the preliminaries and background for this paper, facilitating the exposition of new analyses and results in the subsequent section. In particular, the idea of a limiting average performance measure is introduced, followed by the notion of calibration. Subsequent topics of Sect. 2 are the construction of well-calibrated sequences of distribution functions and the definition of the average quadratic scoring rule for an aggregated sequence. Section 3 proves that the optimal weights in a linear opinion pool could be obtained in most meaningful cases by solving a quadratic program that is concave. Sections 4 and 5 discuss the effect of calibration and sharpness properties of the component judgments on the aggregation. The case of multiple dependent experts is briefly presented in Sect. 6. Section 7 presents a simulation study that compares the performance of optimal weights to that of equal weights assigned to each expert. Last two sections present practical considerations and conclusions.

1.3 Literature review

Different methods of aggregating expert judgments and their properties have frequently appeared in the literature. There are two aggregation approaches: mathematical and behavioral. Mathematical aggregation includes axiomatic and Bayesian approaches. Linear opinion pools belong to axiomatic approaches. Comprehensive reviews are given by Genest and Zidek (1986) and Clemen and Winkler (2007). Drawing a clear-cut, generally accepted conclusion as to an approach outperforming another is difficult. Instead, each approach has its own advantages given the circumstances of the probability elicitation and aggregation process. For example, linear opinion pools are simple in nature, easy to use, robust in performance, and are regarded as one of the most widely used aggregation methods. On the other hand, Bayesian approaches allow modeling the quality of individual expert distributions (such as overconfidence) and dependence among experts. Detailed comparisons of the different approaches can be found in Clemen and Winkler (2007).

Several articles appeared in the literature on allocating weights in opinion pools. Cooke's (1991) classical method determines weights by using probability assessments given by experts on variables for which the analyst knows the outcome. The weight given to an expert is determined based on the measure of an expert's calibration, which can be calculated given the known outcomes. Genest and McConway (1990) present a Bayesian mechanism to update experts' weights in a linear opinion pool, in the face of new information. Authors provide a result highlighting the importance of choosing the initial weights carefully. We note that Clemen and Winkler (2007) acknowledge the methods of Cooke (1991) and Genest and McConway (1990) and remark that no foundationally based method for determining the weights in a linear opinion pool exists. Our work attempts to fill this gap in the literature by providing a concave optimization problem, the optimal solution of which provides the optimal weights. The concave program presented in this paper is based on parameters that allow us to model important characteristics of experts such as calibration, sharpness, and dependence.

DeGroot and Mortera (1991) consider a set-up where a group of experts sharing a common prior distribution for a random variable θ and a common loss function, report their posterior distributions for θ based on different data sets. A single distribution must be chosen, aggregating each expert's individual posterior distributions. When the data observed by the experts are not conditionally independent given θ, the optimal weights to be used in a linear opinion pool are determined for problems involving quadratic loss functions and arbitrary distributions for θ and the data observed. Clearly, the problem set-up and the approach adopted in that paper, which has a Bayesian flavor, differentiates it from our work, which does not involve a Bayesian paradigm. Instead, this paper is based on a construction that allows explicit modeling of expert distributions which, in principle, is a favorable but not-easy-to-implement trait of the Bayesian paradigm (see Clemen and Winkler (2007) for a detailed discussion and comparison of axiomatic and Bayesian approaches). We note that the work by Berger and Mortera (1991) relates to that by DeGroot and Mortera (1991), and takes into account a limited communication in the experts' reports.

Faria and Smith (1997) focus on generalizing the external Bayesian property so that it is suitable for the analysis of multivariate structures, such as partially complete chain graphs. Their approach allows the weights attributed to the joint probability assessments of different experts in the pool to differ across the distinct components of each joint density. Examples on how the weights can be updated based on the experts' relative expertise on causal variables are given by Faria and Smith (1996) and by Faria (1996). Smith and Makov (1978) present a quasi-Bayes approach that adapts the weights based on each experts' relative predictive performances.

Gneiting et al. (2007) proposes three different modes of calibration for a sequence of distributions: a sequence can be probabilistically calibrated, exceedance calibrated, or marginally calibrated. These modes are defined for continuous quantities. They complement each other and may occur in any combination. If an expert satisfies all three modes, the expert is said to be strongly calibrated, or equivalently, ideal. The authors derive conclusions as to whether an expert who is only probabilistically calibrated or only marginally calibrated can be sharper than an ideal expert. The fundamental difference between their work and ours is that they consider a single expert, whereas we consider how calibration and sharpness of multiple experts affect the performance of the aggregation of these experts' judgments.

Ranjan and Gneiting (2010) show that, for binary events, any nontrivial weighted average of two or more distinct, calibrated probability forecasts is necessarily uncalibrated and lacks sharpness. To compensate for this resulting miscalibration, authors propose a transformed linear opinion pool for the aggregation of probability forecasts from distinct, calibrated or uncalibrated sources. Clements and Harvey (2011) consider properties of various combination schemes for binary event probabilities and indicate which types of combinations are likely to be useful.

We would like to acknowledge that there are aggregation approaches in the literature that are not based on probability theory but on other theories such as imprecise probabilities (Walley 1991), evidence theory (Shafer 1976), and possibility theory (Dubois and Prade 1988). Destercke et al. (2009) discusses that these theories allow modeling incomplete or imprecise data, a feature that probability theory arguably cannot account for. They focus on using possibility distributions to model uncertainty and to aggregate data from multiple sources.

2 Background and preliminaries

This section introduces the preliminaries necessary to understand the analyses and results in this paper. Except for Sect. 2.5, contents of this section are not new; they are extracted from published articles in the literature, as cited in appropriate places in the subsections below. Further details on the content of this section could be found in the cited articles. Section 2.5 explicitly depicts the role of the quadratic scoring rule in accounting for miscalibration.

2.1 Calculating the performance of an aggregate distribution function

In this work, we are interested in the performance of a sequence of aggregated distribution functions and how it is affected by certain properties of the component distributions. For this purpose, as a tool, it is assumed here that we have a sequence of densities provided by each expert and one realization for each density. Each density represents a subjective judgment about a random phenomenon. This approach facilitates the asymptotic analysis on how the properties of the component distributions affect the aggregate, and hence provides insights. We note that considering sequences and the limiting average performance measure is a framework similar to that used to find asymptotical relative efficiencies (ARE) for statistical tests or estimators (see Serfling 1980, Chapter 10). Our work is inspired by the ARE methodology and uses sequences as a tool to investigate the behavior of the aggregate sequence.

To this end, let $x_i, i = 1, 2, \ldots$, be a sequence of real numbers and let $F_{ij}(y)$, $i = 1, 2, \ldots, j = 1, \ldots, m$ be m sequences of continuous distribution functions. Note that there is one realization $\mathrm{x_i}$ corresponding to a distribution function. Each $\mathrm{x_i}$ is the realization of an uncertain quantity, where $F_{ij}(y)$ is the probability distribution for this quantity provided by the jth of the m experts. $\mathrm{F_{ij}(y)}$ can be regarded as the distribution elicited from the expert for a random phenomenon, capturing the expert's subjective judgment. We can construct m sequences of distribution functions with certain predefined properties, use an aggregation rule to aggregate the m distribution functions corresponding to each realization x_i, and calculate the performance of the resulting sequence of aggregates using a performance measure. Here, constructing a sequence of distribution functions provides a way to model an expert, as outlined below in Sect. 2.3.

Let us denote a generic performance measure that depends on distribution function H_i and a realization x_i by $G(H_i, x_i)$. Here, each $H_i(y)$ is an aggregation of the functions $F_{ij}(y)$, $j = 1, \ldots, m$. The performance of the aggregates $H_i(y)$, $i = 1, 2, \ldots$ can then be measured using a limiting average, if it exists, by calculating

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} G(H_i, x_i) \tag{1}$$

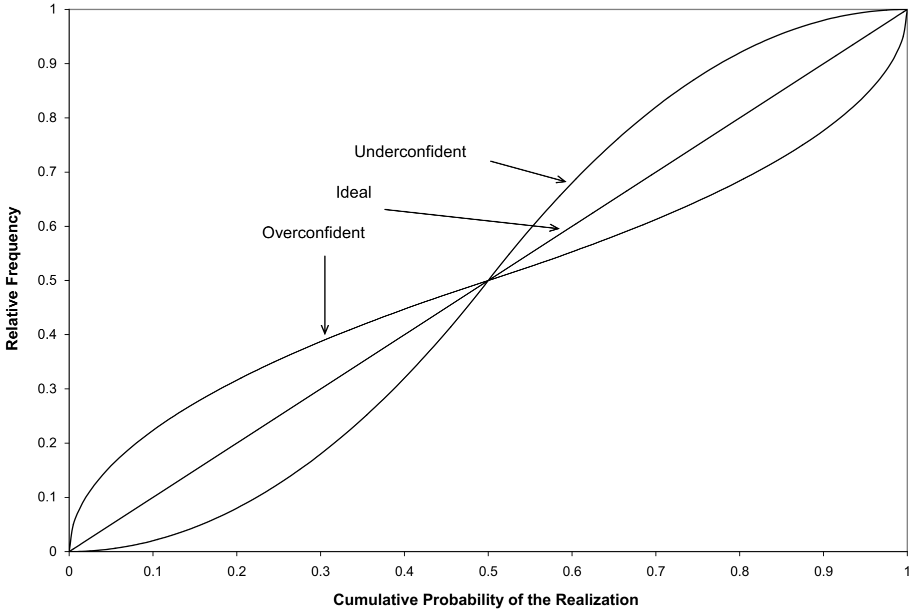

Fig. 1 Representative calibration curves

2.2 Calibration

Definition Let F_i (x), $i = 1, 2, \ldots$ be distribution functions and $x_i, i = 1, 2, \ldots$ be the realizations associated with the functions. We say that a sequence of distribution functions $F_1, F_2, \ldots$ is well-calibrated for the x_i, if, for all $p \in (0, 1)$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I[p - F_i(x_i)] = p$$

where $I(x)$ is the indicator function such that $I(x) = 1$ if $x \geq 0$ and $I(x) = 0$ otherwise.

In other words, for a well-calibrated sequence, the fraction of distribution functions, for which the values x_i fall into the left tail intervals having cumulative probabilities of p, approaches p.

Calibration of a sequence of distributions and realizations can be displayed as a graph, named the *calibration trace* (Hora 2004). Note that in the definition of calibration, each distribution function is associated with a realization. A calibration trace shows, for each fractile p, the proportion of the distribution functions in the sequence that contain its associated realization in its p^{th} fractile. If this proportion is p for every p^{th} fractile, then we have a well-calibrated sequence.

Figure 1 displays three calibration traces. The well-calibrated sequence produces a graph that is a forty five degree diagonal line. The trace that is above the forty five degree line at the left, crosses it at 0.5, and is below the forty five degree line thereafter, indicates distribution functions that are too concentrated and have too many small and large cumulative probabilities and too few probabilities near 0.5. This trace corresponds to a condition known as ‘overconfidence’ (Kahneman et al. 1982). The trace that corresponds to underconfidence is produced by distributions that are, on average, too diffuse.

2.3 Constructing well-calibrated sequences densities with location and scale parameters

This subsection presents a method to model a well-calibrated expert. This method is first introduce by Hora (2010) and generates a well-calibrated sequence of distribution functions, associated with an expert. Our goal is to aggregate the distributions corresponding to each realization. We will then use the limiting average performance measure for the aggregate, which helps us analyze how the calibration and sharpness properties affect the aggregate.

Let $F*(x|\mu, \sigma)$ be a member of a family of distribution functions with a location parameter μ and a scale parameter σ. Then we can write $F*(x|\mu, \sigma) = F[(x - \mu)/\sigma]$. Consider a sequence of such distributions, $F[(x - \mu_i)/\sigma]$ and the corresponding realizations x_i for $i = 1, 2, 3, \ldots$. Being well-calibrated implies that the empirical distribution function of the cumulative probabilities of the realizations converges to a uniform [0,1] distribution function. Using this property, we associate each of the distribution functions with a value from a uniform distribution and construct the location parameter by equating the cumulative probability with the uniform value u_i. Thus,

$$F[(x_i - \mu_i)/\sigma] = u_i \text{ or } \mu_{\mathrm{i}} = x_i - \sigma F^{-1}(u_i) \tag{2}$$

where $F^{-1}(u)$ is the inverse of F(x) and assumed to be unique.

The sequence of distribution functions $\{F[(x_1 - \mu_1)/\sigma], F[(x_2 - \mu_2)/\sigma], \ldots\}$ will then produce cumulative probabilities that have the same empirical distribution as the uniform values and thus will be well calibrated. Altering the value of the spread parameter will change the sharpness of the sequence. Generating the location parameters with the spread parameter σ and then replacing σ with another spread parameter, say s, results in a sequence $\{F[(x_1 - \mu_1)/s], F[(x_2 - \mu_2)/s], \ldots\}$ that may lack calibration. If $\sigma > s$, for example, the distributions will be too narrow emulating overconfidence. Hence, the use of two spread measures allows us to model both an expert's calibration and an expert's sharpness or information. An interpretation of these two spread measures is that σ is the spread of the distribution that the well-calibrated expert should provide but s is the one that the expert actually provides.

Multiple sequences of distributions can be generated by this method and Hora (2010) shows how one may introduce dependence between sequences using a copula. Thus the method allows one to construct sequences that differ with respect to both calibration and sharpness and can have varying degrees of dependence.

2.4 Average quadratic scoring rule

Calibration measures the faithfulness of probabilities but not the information, whereas sharpness is a measure of information but not the faithfulness of the probabilities. Scoring rules have been used as summary measures of the quality of probability densities that incorporate both aspects of calibration and information, and can be used to measure the goodness of an aggregated distribution function (Winkler 1969). The best known of these rules is the quadratic scoring rule defined for densities (Matheson and Winkler 1976) as

$$QS(h, x) = 2h(x) - \int_{-\infty}^{\infty} h^2(u)du$$

where $h(u)$ is a density providing information about an uncertain quantity and x is the realization of that uncertain quantity.

This quadratic scoring rule for continuous densities is well known (see Gneiting and Raftery 2007). We would like to note that the literature on the quadratic scoring rule for discrete quantities is large, whereas it is less developed for continuous quantities. In this work, we attempt to make a contribution by considering the above quadratic score in our model for continuous quantities. This score for continuous densities cannot be derived by simple substitution into the score for an event probability because of the differential entailed with probabilities derived from densities. Therefore, analysis considering continuous quantities requires new methods. The quadratic score is a strictly proper scoring rule, which means that it encourages the forecaster to make careful assessments and to be honest. It is positively sensed in that higher scores are preferred and strictly proper in that it rewards experts for expressing their true beliefs (Winkler 1969). The quadratic score is relatively simple in form and admits analytic expressions for some important cases that cannot be obtained with other well-known scoring rules.

In this paper, we use a linear aggregation of distribution functions given by

$$H_i(y) = \sum_{j=1}^{m} \alpha_j F_{ij}(y) \text{ where } \alpha_j \geq 0 \text{ and } \sum_{\text{j}=1}^{\text{m}} \alpha_j = 1.$$

This aggregation always produces a distribution function and when the weights α_j are all equal, the aggregation becomes simple averaging, and may be the most widely used form of aggregation in practice. Note that the linear aggregation of distribution functions and the linear aggregation of densities yield equivalent results. Let us denote the density associated with the aggregated distribution H_i by h_i. Then from (1), a performance measure based on the average quadratic score is given by $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} QS(h_i, x_i)$.

We next describe briefly how the continuous version of the quadratic scoring rule could potentially be used in a security application. The Department of Homeland Security (DHS) establishes certain programs in an effort to protect the U.S. from various security threats. Program activities can be summarized as prevention, detection, interdiction, protection, response, and recovery. Multiple subject matter experts are used to measure the effectiveness of various activities. Effectiveness, which is a continuous quantity indicating how successful an activity is, is measured by decomposing an activity into several steps and conditioning each step on a number of factors that affect it. Experts are then elicited for all combinations of factors. Experts' judgments are then combined to obtain an aggregate distribution function for effectiveness (please see Hora et al. (2013) for details on this study, which focuses on a different aggregation approach). Continuous version of the quadratic scoring rule, for instance, can potentially be used in such a study to measure the performance of the aggregate distribution.

2.5 Quadratic scoring rule and miscalibration

This paper considers the quadratic scoring rule for continuous quantities. Literature on performance measures for aggregated judgments from multiple experts is rather limited for continuous quantities of interest. A natural question in this context is how the continuous version of the quadratic scoring rule represents miscalibration.

To this end, let $g(x)$ be a density having a strictly continuous distribution function for a random variable X. Let x be a realization of X. Two density forecasts are given for the value x. One from a clairvoyant who responds $g(x)$ and one from an expert who is less than well-calibrated and responds $f(x)$.

The expected quadratic scores for each forecaster are

$$E_g[QS(g,X)] = E_g[2g(X) - \int_{-\infty}^{\infty} g^2(u)du] = \int_{-\infty}^{\infty} g^2(u)du$$

$$\begin{aligned}E_g[QS(f,X)] &= E_g[2f(X) - \int_{-\infty}^{\infty} f^2(u)du] = 2\int_{-\infty}^{\infty} f(u)g(u)du - \int_{-\infty}^{\infty} f^2(u)du \\ &= -\int_{-\infty}^{\infty}[f(u)-g(u)]^2du + \int_{-\infty}^{\infty} g(u)^2du = -\int_{-\infty}^{\infty}[f(u)-g(u)]^2du \\ &\quad + \int_{-\infty}^{\infty} g(u)^2du + \int_{-\infty}^{\infty} f(u)^2du - \int_{-\infty}^{\infty} f(u)^2du \\ &= \int_{-\infty}^{\infty} f^2(u)du - \int_{-\infty}^{\infty}(f^2(u)-g^2(u))du - \int_{-\infty}^{\infty}[f(u)-g(u)]^2du]\end{aligned}$$

Note that the first integral in the last expression is the expected density and is a measure of the apparent information in f. The second integral accounts for the difference in information and the third integral is a penalty for miscalibration, delineating how quadratic scores represent miscalibration. The miscalibration penalty is non-negative and bounded from above by

$$\int_{-\infty}^{\infty}[f(u)-g(u)]^2du] \le \int_{-\infty}^{\infty} f^2(u) + g^2(u)du$$

The difference between the expected scores of the clairvoyant and the expert with less calibration is given by

$$E_g[QS(g,X)] - E_g[QS(f,X)] = \int_{-\infty}^{\infty}[f(u)-g(u)]^2du,$$

which is minimized when $f = g$ and thus, as well-known, the quadratic scoring rule is strictly proper.

3 Optimal weighting

The average score is quadratic in the weights $(w_1, ..., w_m)$ and has the general form

$$\frac{1}{n}\sum_{i=1}^{n} QS\left(\sum_{j=1}^{m} w_j f_{ij}, x_i\right) = \mathbf{w}^t\mathbf{a} - \mathbf{w}^t\mathbf{Q}\mathbf{w} \tag{3}$$

where f_{ij} is the density for the i^{th} quantity provided by the j^{th} expert,

$$\mathbf{w}^t = (w_1, ..., w_m), \mathbf{a}^t = (a_1, ..., a_m), a_j = \frac{1}{n}\sum_{i=1}^{n} f_{ij}(x_i), \underset{(mxm)}{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i, \mathbf{Q}_i = [q_{ijk}],$$

and $q_{ijk} = \int_{-\infty}^{\infty} f_{ij}(u) f_{ik}(u)du$.

Next, it is proved that the average quadratic score is concave. The following lemma is needed for this purpose.

Lemma *Let* $\mathbf{Q}_{kxk} = [q_{ij}]$ *where* $q_{ii} > 0$ *for all* i *and* $q_{ij} = q_{ji} \geq 0$ *for* $i \neq j$. *Then* $\mathbf{Q}_{kxk}$ *is positive definite if* $q_{ij} < \sqrt{q_{ii}q_{jj}}$ *for* $i \neq j$.

Proof Let $\mathbf{D}_{kxk} = [d_{ij}]$ where $d_{ii} = \sqrt{q_{ii}}$ and $d_{ij} = 0$ for $i \neq j$. Then $\mathbf{Q}_{kxk} = \mathbf{D}_{kxk}\mathbf{T}_{kxk}\mathbf{D}_{kxk}$ where $\mathbf{T}_{kxk} = [t_{ij}]$, $t_{ii} = 1$ and $t_{ij} = q_{ij}/\sqrt{q_{ii}q_{jj}}$. Now, $\frac{\alpha^t \mathbf{T}_{kxk}\alpha}{\alpha^t\alpha}$ is minimized (not uniquely) by $\alpha^t = (0, ..., 0, 1/\sqrt{2}, 0, ..., -1/\sqrt{2}, 0, ..., 0)$ where the nonzero elements are at the i^{th} and j^{th} positions and $(i, j) = \underset{i,j}{\arg\max}(t_{ij})$. The minimum is $1 - q_{ij}/\sqrt{q_{ii}q_{jj}}$ which is positive by the condition of the lemma which establishes that $\mathbf{T}_{kxk}$ is positive definite. Since $\mathbf{D}_{kxk}$ is also positive definite symmetric, the product $\mathbf{D}_{kxk}\mathbf{T}_{kxk}\mathbf{D}_{kxk}$ is positive definite symmetric and the sufficiency of the condition is established.

Theorem *Suppose the densities are square integrable and are not equal almost everywhere. The average quadratic score is concave in the weights.*

Proof Now, (3) is quadratic in $\mathbf{w}$ and therefore if $\mathbf{Q}$ is positive definite, **then the** average quadratic score is concave and possesses a unique maximum. From the lemma above, showing that $q_{ijk} < \sqrt{q_{ijj}q_{ikk}}$ for $j \neq k$ is sufficient for each $\mathbf{Q}_{\mathrm{i}}$ to be positive definite provided that the densities are square integrable. This is implied by

$$\int_{-\infty}^{\infty} f_{ij}(u) f_{ik}(u) du \leq \sqrt{\int_{-\infty}^{\infty} f_{ij}^2(u) du \int_{-\infty}^{\infty} f_{ik}^2(u) du}$$

which follows from the nonnegativity of the functions and Holder's inequality which, in turn, holds strictly as an inequality except when the two densities are equal almost everywhere. That $\mathbf{Q}$ is positive definite follows from the sum of positive definite matrices being positive definite.

Note that not all densities are square integrable. For example, the density $f(x) = \frac{1}{2\sqrt{x}}$ for $0 \leq x \leq 1$ is not square integrable.

Next, we examine the important special case where experts have provided sequences of normal densities. Without loss of generality, it is assumed that each target quantity, x_i, is zero. From Hora (2010) equation (22), the limit of the average quadratic score is

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} BS\left(\sum_{j=1}^{m} w_j f_{ij}, x_i\right) = \mathbf{w}^t\mathbf{a} - \mathbf{w}^t\mathbf{Q}\mathbf{w}$$

$$a^t = (a_1, ..., a_m), a_j = \sqrt{\frac{2}{\pi(\sigma_j^2 + s_j^2)}}; \mathbf{Q} = [q_{jk}] \text{ with } q_{jj} = \frac{1}{2s_j\sqrt{\pi}}, \qquad (4)$$

where

$$q_{jk} = \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma_k^2 + s_j^2 + s_k^2 - 2r_{jk}\sigma_j\sigma_k)}} \text{ for } j \neq k.$$

The values σ_j and s_j determine the sharpness and calibration of the j^{th} sequence. Dependence between experts is modeled in terms of the correlation between their means across the sequence of densities and is given by r_{jk}.

The condition required for concavity holds in the limit for the sequences of normal densities except when the following three conditions hold simultaneously: $\sigma_i = \sigma_j$, $s_i = s_j$, $r_{ij} = 1$ for at least one pair (i, j), $i \neq j$. This means that the concavity holds unless there are experts who are completely redundant. Thus, it holds quite generally in the normal case. Note that optimal weights can be efficiently found for a given setting via interior point methods (Nesterov and Nemirovskii 1994), for example, using a solver such as KNITRO (see, Byrd et al. 2006).

4 Calibration and sharpness

To facilitate the analysis of how calibration and sharpness impact densities, this section considers a single expert who has provided normal densities for a sequence of commensurable uncertain quantities. For simplicity, all densities in a sequence have the same sharpness and scale parameter. In later sections, multiple independent and dependent sequences will be used to model the aggregation of multiple experts. Normal densities are assumed here as a probability model so that analytical and numerical results can be obtained to demonstrate how sharpness and calibration are impacted. The Gaussian assumption is not needed for the theorem above, which holds for general densities and states that the average quadratic score is concave in the weights if the densities are square integrable and are not equal almost everywhere.

The analysis begins, then, with a single expert and a sequence of target quantities and densities provided by the expert. Let us recall that in equation (2) of Sect. 2.3, miscalibration can be introduced by using $\sigma_j \neq s_j$. We consider a single expert and hence drop the subscript j when it is appropriate.

So that our analysis is more directly framed in terms of sharpness and calibration, let

$$J = 1/\sigma \text{ and } C = \sigma/s. \tag{5}$$

Then J is positively related to expertise and measures sharpness when the density is corrected for miscalibration. C is an index of calibration. C takes on the value 1 when the densities are well calibrated, $C < 1$ indicates underconfidence, and $C > 1$ overconfidence.

The limiting average quadratic score presented in Equation (4) can then be written, for a single expert, as follows:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} QS(f_i, x_i) &= \frac{1}{2\sqrt{\pi}\sigma} - \left(\frac{1}{2\sqrt{\pi}\sigma} + \frac{1}{2\sqrt{\pi}s} - \sqrt{\frac{2}{\pi(\sigma^2+s^2)}}\right)\\ &= \frac{J}{2\sqrt{\pi}} - \frac{1}{2\sqrt{\pi}}\left[J\left(1 + C - C\sqrt{\frac{8}{C^2+1}}\right)\right]\end{aligned} \tag{6}$$

where the first term in the right side is the reward for sharpness and the second term is the penalty for miscalibration. Of course the second term vanishes when $C = 1$. We note that the quadratic scoring rule is known to be a measure of both calibration and sharpness properties of a density. Consequently, the above expression displays that it accounts for both of these two aspects.

To understand how information and calibration trade off against one another, consider two experts and denote their respective expertise by J_i and calibration index C_i, $i = 1, 2$. Let the first expert be well calibrated so that $C_1 = 1$. The second expert will be modeled to have the same limiting score as the first expert but to have worse calibration $C_2 \neq 1$ but increased

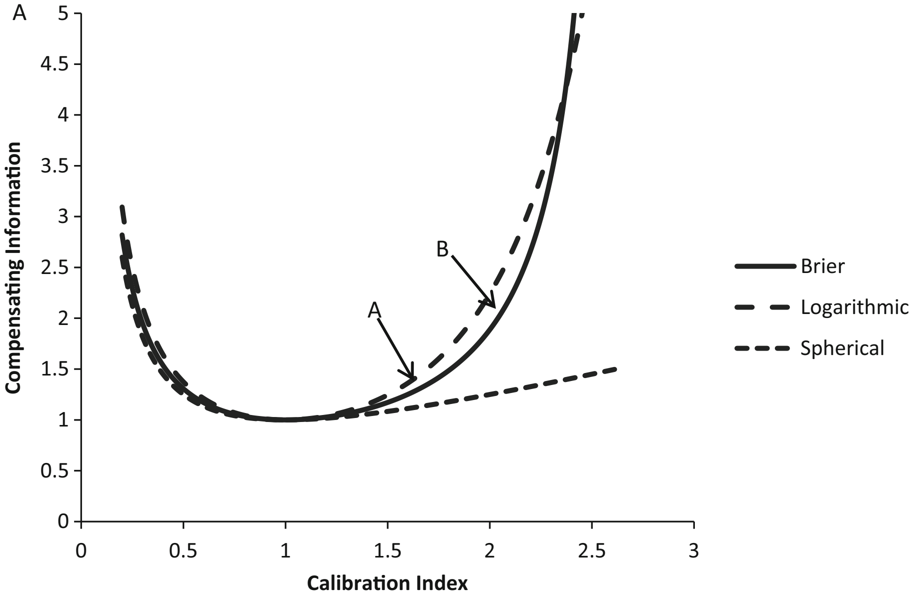

Fig. 2 Relative sharpness vs calibration

sharpness ($J_2 > J_1$) to compensate for the inferior calibration. Using (6), and denoting the relative sharpness of expert 2 to expert 1 by $J_{21} = J_2/J_1$, equal scores are obtained when

$$J_{21} = \frac{1}{\sqrt{\frac{8C_2^2}{C_2^2+1}} - C_2}$$

Figure 2 is a graph of J_{21} as a function of C. The graph displays how much additional sharpness is required to offset a lack of calibration. Similar iso-score lines have been plotted for the logarithmic and spherical scoring rules (Matheson and Winkler 1976). The formulae are

$$Logarithmic\colon J_{21} = \frac{e^{(C_2^2-1)/2}}{C_2} \quad Spherical\colon J_{21} = \frac{C_2^2+1}{2C_2}$$

Combinations of relative sharpness and the calibration index that fall above the curves are preferred to those that fall below the curve while points on a curve are equally preferred. For small departures from perfect calibration, the required compensating increase in information is small regardless of the scoring rule employed. When the departures become large, however, the required compensation increases at an escalating rate except for departure towards overconfidence in the case of the spherical scoring rule. Two points are labeled on the quadratic iso-score curve. Point A ($C_2 = 1.6397$, $J_{21} = 1.2902$) is where the quadratic iso-score curve has a slope of one so that an infinitesimal increase in the calibration index is exactly offset by a numerically equal increase in compensating sharpness. Point B ($C_2 = 2.0581$, $J_{21} = 2.0581$) is where the confidence index and compensating information are equal. The trade-off graph provides an indication that moderate levels of miscalibration are tolerable and do little to degrade the expected score.

What values of the confidence index should be of practical concern? An answer to this question can be found in published studies of overconfidence. Cooke (1991) provides a table of results from many experiments dealing with continuous distributions and measuring overconfidence. Also see Hora et al. (1992). These studies show that about 1/3 of the assessed distributions result in surprises – realizations that fall into the far tails of the densities such as the .01 or .05 probability tails. A rough interpretation of these results is that the standard deviations are about 1/2 what they should be. This would be the case with normal densities. Now, standard deviations that are one-half what they should be correspond to a calibration index of two which is roughly where an equal numerical increase in compensating information would be required to offset the increase in miscalibration. If this is the region where one is operating, close attention should be paid to both the degree of calibration and the level of sharpness. This would not be the case were the miscalibration index closer to one where the trade-off curve is flat. Where the trade-off curve is flat, only small changes in sharpness are needed to compensate for large changes in the calibration index.

5 Two experts

Attention is now turned to how one might aggregate densities from multiple experts. First, consider two experts whose judgments are to be combined with a linear aggregation rule. We start with two independent ($r_{12} = r_{21} = 0$) well calibrated experts ($C_1 = C_2 = 1$) with different levels of expertise ($J_{21} = J_2/J_1 \neq 1$), and consider the weights that should be given to each of the experts. For two experts, setting the derivative with respect to w of the concave expression (4) (or, equivalently, of expression (3)) to 0, using the definitions (5), and simplifying, the average quadratic score is maximized by selecting a weight w for the first expert (and, of course, 1 – w for the second expert) where

$$w = \frac{1 - J_{21}/\left(2\sqrt{J_{21}^2 + 1}\right)}{J_{21}\left(1 - 1/\sqrt{J_{21}^2 + 1}\right) + 1} \tag{7}$$

and J_{21} is the relative sharpness of expert 2 to expert 1, $J_{21} = J_2/J_1 = \sigma_1/\sigma_2$. The solid line in Figure 3 is a graph of w as a function of J_{21}. When $J_{21} = 0$, $w = 1$ and w declines with J_{21} and is asymptotic to zero as J_{21} increases.

We note that as $J_{21} \to 0$, the weight for expert 1 approaches 1 because the relative sharpness of expert 2 to expert 1 approaches 0. On the other hand, if $J_{21} \to \infty$ the weight for expert 1 approaches 0 because the relative sharpness of expert 2 to expert 1 approaches infinity. If the x-axis represents $J_{12} = J_1/J_2$ instead, then the y-axis would represent the weight for the second expert and the above arguments would apply respectively. As long as the relative sharpness is not zero or not infinite, then no expert is disqualified.

Figure 4 shows the average quadratic score when the optimal weight is used to aggregate the densities (solid line) and when equal weighting is used (dashed line). There is no gain from using optimal weighting relative to equal weights when the relative sharpness is one as the optimal weights are 1/2. As the relative sharpness diverges from 1, the ratio of the optimized score to the one obtained with equal weights approaches 4/3.

Next, consider the impact of miscalibration on the weights given to the two independent experts. Assume that they are equally sharp ($J_{21} = 1$) and that expert 1 is well calibrated while expert 2 is less than well calibrated ($C_2 = \sigma_2/s_2 \neq 1$). Again, for two experts,

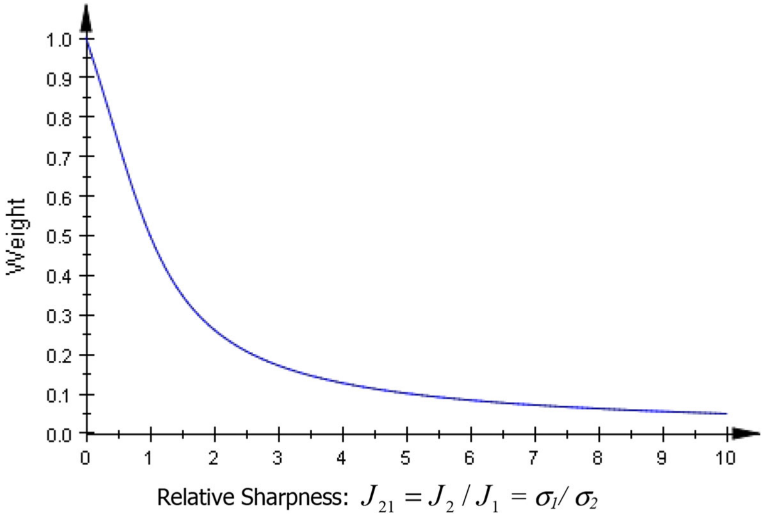

Fig. 3 Relative sharpness vs optimal weights

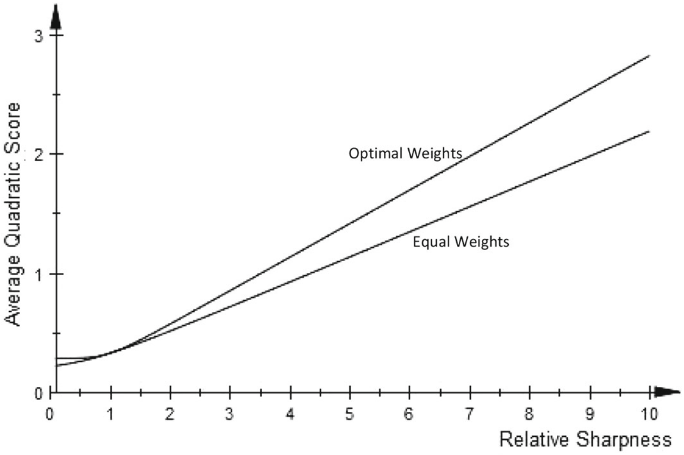

Fig. 4 Relative sharpness vs average score

setting the derivative with respect to w of expression (4) to 0, using the definitions (5), and simplifying, the optimal weight for expert 1 is given by

$$w = \frac{C\left(2 - \sqrt{\frac{2}{3C^2+1}} - 2\sqrt{\frac{2}{C^2+1}}\right) + 2}{2C\left(1 - \sqrt{\frac{2}{3C^2+1}}\right) + 2} \tag{8}$$

The solution was facilitated by the symbolic solver in the MuPad toolbox of Matlab (2011).

Figure 5 shows how the optimal weight for the first expert varies with the miscalibration of the second expert. Again, the weight approaches one asymptotically meaning that the second expert will not be totally disqualified because of miscalibration when the experts have similar

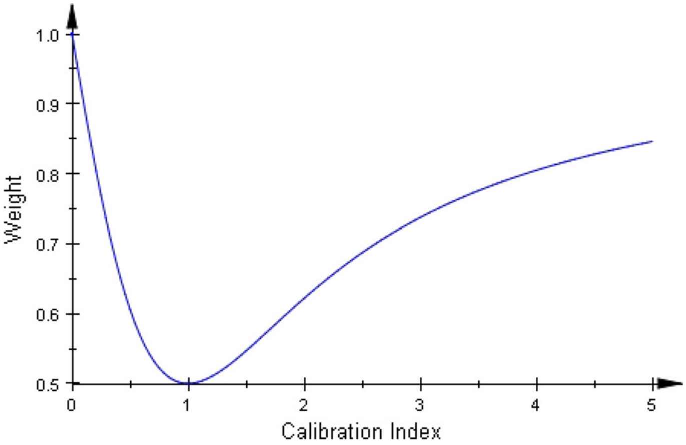

Fig. 5 Calibration vs optimal weights

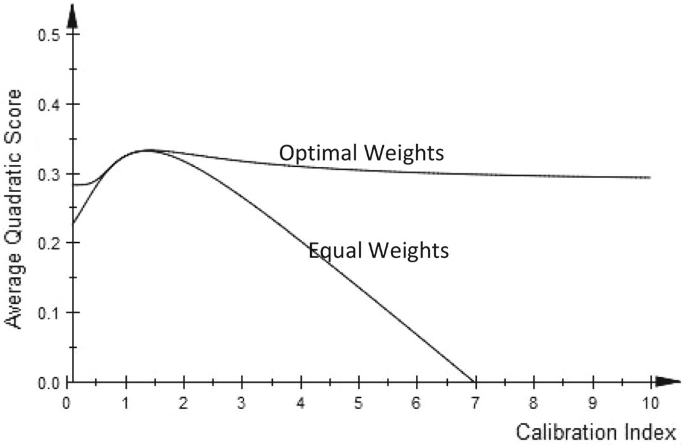

Fig. 6 Calibration vs average score

expertise. Figure 6 shows the average quadratic score when the optimal weight is used to aggregate the densities (solid line) and when equal weighting is used (dashed line).

The asymptotic behavior of the weights in the two cases given above raises the question of whether or not there are conditions where all the weight would be given to one expert and the other expert effectively disqualified. Letting $r_{12} = r_{21} = r$ and setting the derivative of expression (4) to 0, we obtain that if a solution exists such that the weight given to the first expert is an interior point of the unit interval [0,1], then the solution is given by

$$w = \frac{\frac{1}{s_2} + \sqrt{2}\left(\frac{1}{\sqrt{s_1^2+\sigma_1^2}} - \frac{1}{\sqrt{s_2^2+\sigma_2^2}}\right) - \left(\frac{1}{\sqrt{2}}\right)\left(\frac{1}{\sqrt{s_1^2+\sigma_1^2+s_2^2+\sigma_2^2-2r\sigma_1\sigma_2}}\right)}{\frac{1}{s_1} + \frac{1}{s_2} - \left(\sqrt{2}\right)\left(\frac{1}{\sqrt{s_1^2+\sigma_1^2+s_2^2+\sigma_2^2-2r\sigma_1\sigma_2}}\right)} \quad (9)$$

where σ_i and s_i are as before and r is the correlation between the means of the expressed distributions measured across the two series of densities representing the two experts. Thus, r

Table 1 Values corresponding to maximizing score, maximizing weight, and truthful answers

	S_2	C_2	Weight	Ave score
Max score	2.812	0.711	0.226	0.322
Max weight	1.539	1.300	0.272	0.312
Truth	2.000	1.000	0.263	0.318

is a measure of dependence of the experts densities. Letting $\sigma_1^2+s_1^2 \to \infty$, $w \to 1-\sqrt{\frac{2}{1+C_2^2}}$ where $C_2 = \sigma_2/s_2$. Thus, $C_2 < 1$ implies $w < 0$ in (10) so that an interior solution does not exist under these conditions and the first expert would receive zero weight – effectively disqualified. This result holds for all$-1 < r < 1$. The conclusion is that certain circumstances can lead to disqualification of an expert. Here, the calibration of the second expert is instrumental in determining whether the first expert is disqualified and this disqualification does not depend upon the calibration of the first expert. Note that this example entails manipulating both miscalibration and differences in relative sharpness.

Another interesting issue is whether an expert can increase the weight that their density receives by responding with a value s that either overstates or understates their information. This is similar to the "proper" property of strictly proper scoring rules. The answer is yes, it is possible for an expert to improve his weight by, for instance, overconfidence. For example, if the first expert is a calibrated expert with $\sigma_1 = 1$ and the second expert is independent of the first expert and has information $J_{21} = 1/2$, ($\sigma_2 = 2$) relative to the first expert, the second expert will maximize their weight by responding with s_2 = 1.54, equivalent to a calibration index of $C_2 = 1.30$ (moderate overconfidence) thus overstating their information.

Note that maximizing one's weight is very different from maximizing the average quadratic score of the aggregate. In the case described above, the second expert would maximize the average score by responding with $s_2 = 2.81$, resulting in reduced weight and exhibited underconfidence. For comparison, the case where the expert answers truthfully has been analyzed and all three cases presented in the table above (Table 1).

It is apparent, then, that strictly proper scoring rules applied to aggregated densities do not necessarily enforce the truthfulness of the component densities.

6 Dependency

The analysis of dependency in the multiple expert setting entails a large number of potential parameters as the number of distinct correlations is $m(m-1)/2$ where m is the number of experts. So that the situation is manageable, we consider a group of m experts who are symmetric with respect to sharpness, have perfect calibration, but $(m-1)$ of whom have an inter-expert correlation of r > 0 and are mutually independent of the m^{th} expert. We assume that the $m-1$ symmetric experts each have a weight of $(1-w)/(m-1)$ the m^{th} expert has a weight of w. The optimal weight for the m^{th} expert can be found by differentiation and solving for the stationary point and is given by

$$w_m = \frac{(\sqrt{2-r})(\sqrt{2}-m+1)+m-2}{(\sqrt{2-r})(\sqrt{2}m-2m+2)+\sqrt{2}(m-2)} \tag{10}$$

The remaining weight is distributed equally among the remaining $m-1$ dependent experts.

Figure 7 shows w_m in terms of r, for $m = 3, 4, 6$, and 10. Each graph intersects the ordinate at $1/m$ and reaches a value of 1/2 at $r = 1$. This is as expected because at $r = 0$, the experts are mutually independent. When $r = 1$, the $m-1$ dependent experts are perfectly redundant

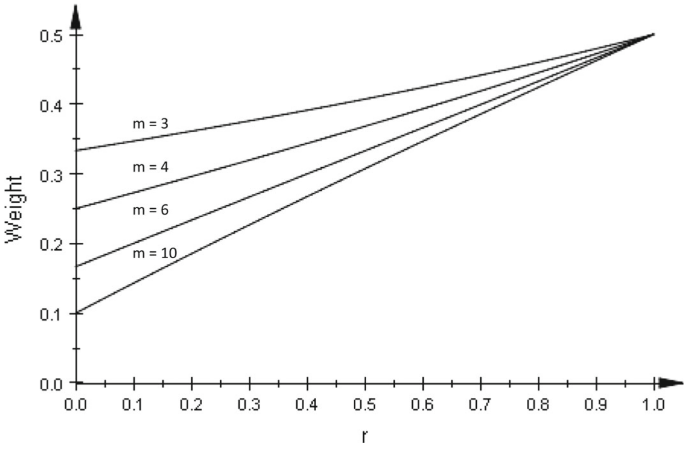

Fig. 7 Optimal weight for the m^{th} SME

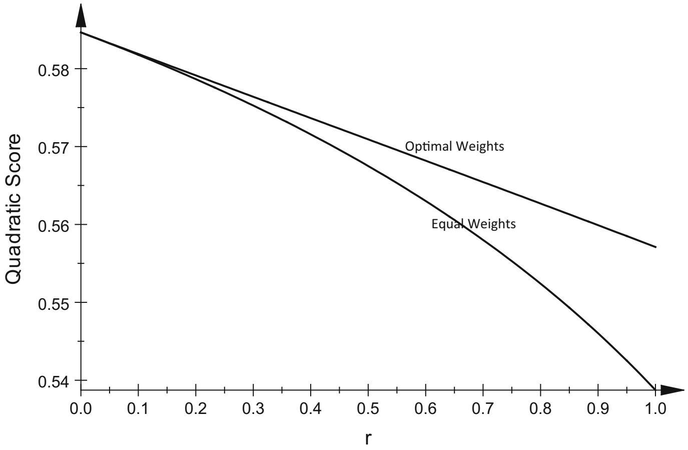

Fig. 8 Limiting quadratic scores with optimal and equal weights $m = 6$

and carry an amount of information equivalent to a single expert. What is surprising is how nearly linear the intermediate weights are, as a function of r. Clearly, (10) is not linear in r. Figure 8 is the companion to Fig. 7 and shows the limiting average quadratic score as a function of r for $m = 6$ for both optimal weights and equal weights. Please note the scale on the vertical axis is truncated and the difference between optimal weighting and equal weighting is slight. Moreover, the impact of dependence on the limiting average score is surprisingly slight.

7 Comparing optimal and equal weights via simulation

The concave quadratic program presented earlier provides optimal weights for judgments for a given set of sharpness and calibration parameters. The purpose of this section is to simulate these parameters and measure the performance of optimal weights found via the concave program, relative to assigning equal weights to experts.

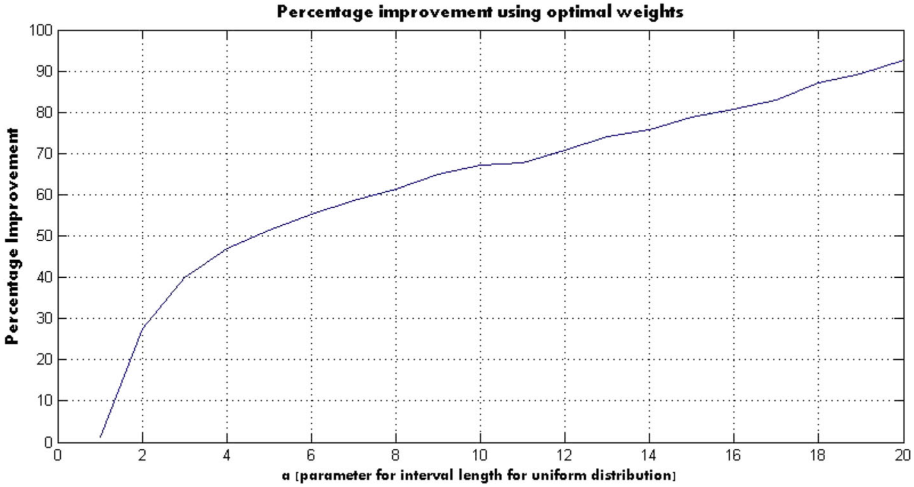

Fig. 9 Percentage improvement via simulation

To this end, consider the aggregation of densities from six experts. For both the sharpness parameter, J, and the calibration index C, a sample is drawn from a log uniform distribution by sampling uniformly on the interval $[\log(1/a), \log(a)]$ and exponentiating the result. The correlation between each pair of experts is assumed to be nonnegative and sampled uniformly from [0,1]. Using the sampled values for six experts, we solve the concave quadratic program to obtain optimal weights and the limiting average score. Using the same sampled values, we then fix each weight at 1/6 to evaluate the score with equal weights. We repeat the same process 10,000 times and obtain a mean value for optimal scores. Similarly, a mean value is obtained for scores with equal weights, averaged over the 10,000 samples, for $a = 1, \ldots, 20$. Thus as a is incremented, the differences among the experts in terms of both sharpness and calibration increases.

Each problem is solved using a commercial nonlinear solver KNITRO (Byrd et al. 2006), on a Dell workstation with a CPU at 3.20 GHz, 2 GB RAM, using the Red Hat Linux 3 operating system. The entire simulation entails solving 200,000 optimization problems and takes about 35 minutes. Let us denote the mean optimal scores by *MO* and mean scores with equal weights by *ME*. A percentage improvement is given by

$$\text{Percentage Improvement} = \left((MO - ME)\Big/ME\right) x100$$

We note in Fig. 9 that percentage improvement increases as sharpness and calibration parameters are given larger sampling intervals, increasing the differentiation among the simulated SMEs, and conclude that considerable improvement could be achieved by using optimal weights as opposed to a rule of thumb of using equal weights.

8 Practical considerations

This section briefly suggests ways to estimate the parameters needed to find optimal weights for experts. We emphasize that the methods suggested in this section are very preliminary and ad-hoc. How well these estimates perform remains as a future research question. Our aim here is to initiate a practical approach to estimate the required parameters, which could possibly be developed and tested in future research.

While trade-offs between calibration and sharpness may be of theoretical interest, there is a significant hurdle to extracting practical value from these insights. Seldom, unfortunately, is information available for qualifying the experts and measuring their dependence. A partial exception is the methodology developed by Cooke (1991). In this methodology, target quantities with known values are interlaced with the real elicitation quantities. These "seed" values and the elicited distributions are then used to conduct a chi-squared goodness-of-fit test to determine if an expert is sufficiently well calibrated. Experts with small p-values are disqualified. Then, a measure of relative information is developed from the span of the elicited distributions and experts are assigned weights proportional to the product of their p-values and the information measure. No information is gathered on dependence nor is any used.

Determining optimal weights under the normal model given here requires that one know the calibration index for each expert, the relative sharpness of their densities, and the dependence among the densities as measured by the correlations of the deviations of the target quantities from the means of the elicited densities. Such information could be obtained from the use of seed values, through the analysis of training quizzes, or through analysis of historical forecasts so that it is not beyond possibility that approximately optimal weights could be calculated.

In what follows, two cases are distinguished. In the first case, one has a sequence of commensurable values such as precipitation amounts and the corresponding forecast densities. The second case entails noncomensurable quantities. Denote the various values by $x_i, i = 1, \ldots, n$, the densities of the j^{th} expert by $f_{ij}(y)$, and the corresponding distribution functions by $F_{ij}(y)$. An estimate of the calibration index, C_j, can be obtained in the first case in the following manner:

1. Reparameterize the distribution functions by inserting a location parameter such as the median of the distribution so that $F^*_{ij}(y - m_{ij}) = F_{ij}(y)$.
2. For some value $0 < t < 1/2$, find C_j such that

$$\sum_{i=1}^{n} \Big(I\{t - F^*_{ij}[(x_i - m_{ij})/C_j]\} + I\{F^*_{ij}[(x_i - m_{ij})/C_j] - (1-t)\} \Big) = 2t$$

where $I(x) = 1$ when $x \geq 0$ and $I(x) = 0$ otherwise.

For example, with $t = 1/4$, C_j will adjust the distribution functions so that the percentage of values falling within their corresponding interquartile range is exactly 50 %. This does not guarantee however, that each quartile will have 25 % of the values.

Once the estimate of the calibration index is calculated, the sharpness index, J_j, can be estimated by $1/(C_j \bar{s}_j)$ where $\bar{s}_j$ is the average standard deviation of the densities given by the j th expert. The dependence parameter, r_{ik}, may be estimated by calculating the product moment correlation between the deviations from the mean for the i th and k th experts.

The second case entails densities for values that are not commensurable and might occur when experts are asked to respond to question about quantities on various scales such as pressure, temperature or time. Typically, questions in an almanac quiz will have varying scales. Once again, the given distribution functions are reparameterized. The estimate of C_j is found from the following two steps:

1. Standardize the distribution functions by inserting both location and scale parameters such as the median and interquartile range or mean and standard deviation of the distribution so that $F^*_{ij}\left[\frac{(y-m_{ij})}{s_{ij}}\right] = F_{ij}(y)$.

2. For some value 0 < t < 1/2, find C_j such that

$$\sum_{i=1}^{n}\Big(I\{t - F_{ij}^{*}[(x_i - m_{ij})/(C_j s_{ij})]\} + I\{F_{ij}^{*}[(x_i - m_{ij})/(C_j s_{ij})] - (1 - t)\}\Big) = 2t$$

where $I(x) = 1$ when $x \geq 0$ and $I(x) = 0$ otherwise.

The index of sharpness can be developed by rescaling sharpness for each i and averaging the values:

$J_j = \frac{1}{n}\sum\limits_{i=1}^{n}\frac{\frac{1}{m}\sum\limits_{k=1}^{m} C_k s_{ik}}{C_j s_{ij}}$. Dependence parameters are estimated as in the first case.

Armed with estimates of J_j and C_j, letting $\sigma_j = 1/J_j$ and $s_j = J_j/C_j$ one can solve for optimal weights using (11).

9 Conclusions

Sharpness and good calibration are both desirable properties and are modeled here through sequences of normal densities. The findings indicate that moderate levels of miscalibration are well tolerated in that they are offset by small increases in sharpness. The implication is that if experts are moderately well calibrated, emphasis is better placed on improving their informativeness than attempting to improve calibration. This finding suggests that more attention be paid to assisting experts in obtaining information to support their judgments than to improving their coding of judgments into probability distributions.

A second finding is that sometimes it is better to completely ignore the judgment of one expert rather than linearly aggregate that judgment with that of an expert who gives "better" densities. In other situations, neither very bad calibration nor lack of sharpness alone may be enough to disqualify an expert. It is also shown that applying a strictly proper scoring rule to an aggregated density does not necessarily encourage truthful answers. Answering falsely may increase the weight an expert receives in an aggregation. Moreover, an expert may increase the expected score of the aggregate by deliberately responding untruthfully.

An important result presented here is that optimal weights for square integrable densities can be found through a quadratic program. A simulation experiment was conducted to examine the level of improvement one might obtain if optimal weights were used in a linear aggregation rather than equal weights. The judgments of six experts were simulated and a single parameter was used to increase the differentiation of the experts in terms of both sharpness and calibration. These properties were sampled independently. As expected, the relative performance of the optimal weights increased with this differentiation and the improvement in the expected score rose to 92 % above the expected score with equal weights in the most extreme case.

Implementation of optimal weighting will require evidence about performance that is usually not collected during a probability elicitation exercise. This evidence is needed to provide estimates of calibration, relative expertise, and dependence. While requiring additional effort, developing optimal weights will not only improve the quality of the aggregated judgments but may also enhance the credibility of probability elicitation as a scientifically based discipline.

References

Bayarri, M. J., & DeGroot, M. H. (1988). Gaining weight: A Bayesian approach. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 3* (pp. 25–44). Oxford: Oxford Univ. Press.

Berger, J. O., & Mortera, J. (1991). Bayesian analysis with limited communication. *Journal of Statistical Planning and Inference*, *28*(1), 1–24.

Brier, G. (1950). Verification of weather forecasts expressed in terms of probabilities. *Monthly Weather Review*, *76*, 1–3.

Byrd, R. H., Nocedal, J., & Waltz, R. A. (2006). KNITRO: An integrated package for nonlinear optimization. In G. di Pillo & M. Roma (Eds.), *Large-scale nonlinear optimization* (pp. 35–59). Berlin: Springer.

Clemen, R. T., & Winkler, R. L. (2007). Aggregating probability distributions. In W. Edwards, R. Miles, & D. von Winterfeldt (Eds.), *Advances in decision analysis*. Cambridge, UK: Cambridge University Press.

Clements, M. P., & Harvey, D. I. (2011). Combining probability forecasts. *International Journal of Forecasting*, *27*, 208–223.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford: Oxford University Press.

Degroot, M., & Fienberg, S. (1983). The comparison and evaluation of forecasters. *The Statistician*, *32*, 12–22.

DeGroot, M. H., & Mortera, J. (1991). Optimal linear opinion pools. *Management Science*, *37*, 546–558.

Department of Homeland Security (2006). *Bioterrorism Risk Assessment*. Biological threat characterization center of the national biodefense analysis and countermeasures center, Fort Detrick, MD.

Destercke, S., Dubois, D., & Chojnacki, E. (2009). Possibilistic information fusion using maximal coherent subsets. *IEEE Transactions on Fuzzy Systems*, *17*, 79–92.

Dubois, D., & Prade, H. (1988). *Possibility theory: An approach to computerized processing of uncertainty*. New York: Plenum Press.

Faria, A. E. (1996). Graphical Bayesian models in multivariate expert judgments and conditional external Bayesianity. Ph.D. dissertation, Dept. Statistics, University of Warwick.

Faria, A. E., & Smith, J. Q. (1996). Conditional external Bayesianity in decomposable influence diagrams. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 5* (pp. 551–560). Oxford: Clarendon Press.

Faria, A. E., & Smith, J. Q. (1997). Conditionally externally Bayesian pooling operators on chain graphs. *Annals of Statistics*, *25*(4), 1740–1761.

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

Genest, C., & McConway, K. J. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting*, *9*, 53–73.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, *69*, 243–268.

Hora, S. C. (2004). Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, *50*(5), 597–604.

Hora, S. C. (2011). An analytic method for evaluating the performance of aggregation rules for probability densities. *Operations Research*, *58*(5), 1440–1449.

Hora, S. C., & Iman, R. L. (1989). Expert opinion in risk analysis: The NUREG-1150 experience. *Nuclear Science and Engineering*, *102*(4), 323–331.

Hora, S. C., Hora, J. A., & Dodd, N. (1992). Assessment of probability distributions for continuous random variables: A comparison of the bisection and fixed value methods. *Organizational Behavior and Human Decision Processes*, *51*, 133–155.

Hora, S. C., Fransen, B. R., Hawkins, N., & Susel, I. (2013). Median aggregation of distribution functions. *Decision Analysis*, *10*(4), 279–291.

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

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096.

MATLAB, 7.12, (2011). The MathWorks Inc., Natick, MA.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. New York: Cambridge University Press.

National Research Council (2008). *The Department of Homeland Security Bioterrorism Risk Assessment: A Call for Change*. Washington D.C.: The National Academies Press.

Nesterov, Y., & Nemirovskii, A. (1994). *Interior-point polynomial algorithms in convex programming*. Philadelphia: SIAM.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B*, *72*(1), 71–91.

Rasmussen, N., et al. (1975). *Reactor safety study: WASH-1400, NUREG-751014*. Washington: U.S. Nuclear Regulatory Commission.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.

Smith, A. F. M., & Makov, U. E. (1978). A quasi-Bayes sequential procedure for mixtures. *Journal of the Royal Statistical Society, Series B*, *40*, 106–112.

Stone, M. (1961). The linear opinion pool. *Annals of Mathematical Statistics*, *32*, 1339–1342.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. UK: Chapman and Hall.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, *64*(3), 1073–1078.