SOME SELECTED CONSENSUS APPROACHES

Lee R. Abramson
Applied Statistics Branch
U.S. Nuclear Regulatory Commission
Washington, DC 20555

Summary

A number of approaches to the problem of forming a consensus based on individual assessors' judgments have been proposed. This paper discusses and provides examples of several of these approaches, including the weighted average approach, the Bayesian approach and the performance criterion approach.

1. Introduction

Suppose that a group of n assessors each provides an assessment of a subjective probability distribution, denoted by f_i, $i = 1, \cdots, n$. Winkler[11] discusses three approaches to the problem of forming a consensus assessment f which, in some sense, best represents the group assessment. These are the weighted average approach, the controlled feedback approach and the Bayesian approach.

In the weighted average approach, the consensus f is usually expressed as a linear combination of the f_i's, although non-linear combinations are sometimes used (e.g., median, geometric mean, harmonic mean). The weights are either equal or proportional to some ranking of the assessors which reflects their expertise, e.g., self-ratings. Some studies have shown that equal weights work about as well as other methods of assigning weights. The approach can be a one-step procedure or can be iterative. If iterative, the weights can be constant or change at each step.

In the controlled feedback approach, the consensus f is developed by several rounds of feedback and reassessment. The reassessment can be either group (face-to-face) or individual (Delphi). While this approach is widely used, the resultant consensus may be overly influenced by group dynamics. Experimentation by psychologists has shown that when group interaction involves open discussion, group positions tend toward uniformity and established norms. This can be induced by the influence of discussion leaders as well as the desire to reach agreement. Even with anonymity, feedback can induce pressure towards a consensus.

In the Bayesian approach, the f_i's are viewed as sample information which is then combined with the decision-maker's prior through Bayes' theorem. While this approach has the virtue of directly involving the decision-maker, the results may be difficult to interpret. Furthermore, the f_i's are often chosen for mathematical convenience rather than as an unconstrained expression of the assessors' judgments.

In [3], Dalkey proves that no group decision rule exists which is consistent with all of the postulates of probability theory. For example, the average of a set of probabilities fulfills the requirement that probabilities of exclusive events add; however, it does not fulfill the requirement that the probability of the conjunction of two independent events is the product. The converse is true for the product (or the geometric mean) as an aggregation rule; it does not sum to one for exclusive and exhaustive events but it is multiplicative for conjunctions.

SOME SELECTED CONSENSUS APPROACHES

Lee R. Abramson
Applied Statistics Branch
U.S. Nuclear Regulatory Commission
Washington, DC 20555

As a solution to this dilemma, Dalkey espouses what he terms the Emerson Principle*: Performance is at least as important a criterion for aggregation as consistency ([4], page 10). The Emerson Principle is a rationale for the use of performance criteria, i.e., probabilistic scoring rules which reward the assessors depending on the accuracy of their assessments.

This paper discusses some selected examples of the consensus approaches introduced above. No attempt at completeness is made; the purpose is to illustrate some of the ways in which the consensus problem has been approached.

2. Weighted Averages

An Axiomatic Approach

In [1], Abramson uses an axiomatic approach to the problem of combining subjective probability distributions into a group consensus. A small number of plausible properties which a consensus dist[illegible] should satisfy are specified and it is th[illegible] that there is only one function of the individ[illegible]butions which satisfies these properties.

Assume that each of a group of n assessors provides a subjective probability distribution on a common set of m mutually exclusive and exhaustive events, as indicated in the table below.

	Event						
Assessor	E_1	E_2	$\cdots$	E_j	$\cdots$	E_m	Σ
1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1m}	1
i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{im}	1
⋮							
n	p_{n1}	p_{n2}	$\cdots$	p_{nj}	$\cdots$	p_{nm}	1
Consensus	G_1	G_2	$\cdots$	G_j	$\cdots$	G_m	1

Here p_{ij} is the probability assigned by assessor i to event E_j and G_j is the consensus probability for E_j. For the case of equally weighted assessors, it is assumed that the consensus distribution (if it exists) must satisfy the following three properties:

(1) The consensus probabilities sum to 1.

(2) The consensus probability for any event depends only on the set of probabilities

*"A foolish consistency is the hobgoblin of little minds...." - Ralph Waldo Emerson

for that event, and not on the probabilities for the other events or on which assessor assigns which probability.

(3) If all assessors agree on the probability of an event, then the consensus probability is their common probability.

It is then proven that the only consensus distribution which satisfies properties (1), (2) and (3) is the average of the assessors' subjective probability distributions, i.e.,

$$G_j = \frac{1}{n} \sum_{i=1}^{n} p_{ij} \quad \text{for } j=1, \cdots, m .$$

These results are generalized to the case where the assessors have arbitrary known weights. Let w_i = weight of assessor i, where $w_i \geq 0$ and $w_1 + \cdots + w_n = 1$. Properties (1) and (3) remain unchanged and property (2) is generalized to allow the consensus probability for an event to depend on both the probabilities and weights for that event. Two additional properties are assumed:

(4) If two assessors assign the same probability to an event, they can be replaced by a single assessor with weight equal to the sum of their weights.

(5) The consensus probability for any event is a continuous function of the assessors' weights.

It is then proven that the only consensus distribution which satisfies properties (1) - (5) is the weighted average of the assessors' subjective probability distributions, i.e.,

$$G_j = \sum_{i=1}^{n} w_j \, p_{ij} \quad \text{for } j=1, \cdots, m .$$

Dalkey ([4], p.228) proved essentially the same result with a very similar approach. One difference between the approaches is that Abramson assumed that the assessors are assigned weights independently of their subjective assessments and that the significance of these weights is expressed by property (4), while in Dalkey's derivation the weights are implied by the consensus distribution. (The implied weight for each assessor is the consensus probability of an event to which that assessor assigns probability one and to which all other assessors assign probability zero.)

Iterative Weighting

Constant weights. DeGroot[6] considers the following problem.

θ = parameter to be estimated (may be a vector)

F_i = subjective probability distribution assigned by assessor i to parameter θ ($i=1, \cdots, k$)

p_{ij} = weight that assessor i assigns to the distribution of assessor j, where

$$p_{ij} \geq 0 \text{ and } \sum_{j=1}^{k} p_{ij} = 1 \ (i, j=1, \cdots, k) .$$

The F_i are revised by each assessor using the weights p_{ij}. Thus, the first revision of F_i by assessor i is

$F_{i1} = \sum_{j=1}^{k} p_{ij}F_{ij}$. This procedure is then iterated. In matrix notation,

$$F^{(1)} = P\ F$$
$$F^{(n)} = P\ F^{(n-1)} = P^n\ F,\ n=2,\ 3,\ \cdots$$

By definition, a consensus is reached if the revised subjective distributions all approach some limiting distribution F*. A necessary and sufficient condition for a consensus to be reached is that there exists a vector $\pi = (\pi_1, \cdots, \pi_k)$ such that

$p_{ij}^{(n)} \rightarrow \pi_j$ for i, j=1, $\cdots$, k. Then

$$F^* = \sum_{i=1}^{k} \pi_i\ F_i \quad \text{and} \quad \pi P = \pi\ .$$

A sufficient condition for a consensus to be reached is that for some n, every element in at least one column of the matrix P is positive. In other words, for some iteration, there is at least one assessor to whom all of the other assessors give positive weight.

In a validation experiment conducted by Moskowitz and Bajgier,[8] subjects, participating either as a member of a panel discussion or Delphi group, made iterative subjective probability distribution (SPD) assessments using the fractile method on various unknown quantities. Examples included the percentage of Purdue students on academic probation and the number of miles driven per automobile accident fatality. The DeGroot model with constant weights did not appear to predict or describe the panel discussion or Delphi group consensus process. Opinion weights were not stable and appeared to vary inversely with the dispersion of a group member's SPD. These, however tended to stabilize after several iterations. Models in which the weights were inversely proportional to the variance or the .01 to .99 fractile range of each group member's SPD gave considerably better predictions than did the DeGroot model.

Variable weights. Chatterjee and Seneta[2] generalized the DeGroot model[6] to the case of variable weights. Let

$p_{ij}(n)$ = weight assigned by individual i to the distribution of individual j after n iterations.

Then a sufficient condition for a consensus to be reached is that

$$\sum_{n=1}^{\infty} \max_{j} (\min_{i} p_{ij}(n)) = \infty\ . \qquad (1)$$

Three examples where consensus is reached are as follows.

(a) There are an infinite number of occasions when there is at least one assessor to whose opinion everyone attaches a weight of at least $\delta > 0$ (generalizes constant weight criterion).

(b) Open-minded assessors. As the iterations proceed, information is exchanged and the assessors' initial specialized information tends to become group knowledge, i.e., the assessors tend to give equal weight to all opinions. Then $p_{ij}(n) \rightarrow 1/k$, Eq.(1) is satisfied and consensus will be reached.

(c) Slow hardening of positions. Suppose that the information exchange process causes the assessors to put more weight on their own opinions and less on those of others, with a tendency in the limit to put all the weight on their own opinions. If the hardening of positions is sufficiently slow, even this situation can lead to a consensus. For example, suppose that

$$p_{ii}(n) = 1 - \frac{1}{2n} \ .$$

$$p_{ij}(n) = \frac{1}{2(k-1)n} \ , \ i \neq j \ .$$

Then Eq.(1) is satisfied and consensus will be reached.

3. Qualitative Controlled Feedback

A controlled feedback procedure can be characterized as follows:

(i) Each member of a group of respondents independently answers a battery of related questions. Sometimes, reasons for their answers are also solicited.

(ii) Summary information is presented to each group member, and step (i) is repeated.

(iii) The questioning and feedback process is repeated until it stabilizes (little change from round to round). The stabilization can either be in the form of a group consensus or judgment nuclei, i.e., a hung jury.

The commonly used Delphi procedure is a quantitative controlled feedback procedure, whereby the summary information in step (ii) is in the form of group medians, quantiles and the like. In [9], Press presents a qualitative controlled feedback procedure whereby panelists supply answers and justifying reasons as in step (i) but only a composite of the reasons is fed back in step (ii).

If one question only is asked, Press proposes the following model:

$$z_{i,1} = \sum_{k=1}^{r} x_{ik} \beta_k + u_{i1} \ ,$$

where

$z_{i,1}$ = first-stage response of respondent i

x_{ik} = "cue" variables (demographic and attitudinal characteristics of respondent as well as variables related to the question)

β_k = unknown regression coefficient

u_{i1} = random disturbance with zero mean and constant variance.

For stage $n \geq 2$,

$$z_{i,n} = z_{i,n-1} + \sum_{j=1}^{R_{n-1}} c_{ij}[1 - \delta_{ij}^{(n-1)}]p_{jn} + u_{in}$$

where

$z_{i,n}$ = n-stage response of respondent i

R_n = total number of reasons presented at stage n

c_{ij} = unknown weight coefficients

$$\delta_{ij}^{(n)} = \begin{cases} 1 \text{ if respondent i gives reason j at stage n} \\ 0 \text{ otherwise} \end{cases}$$

p_{jn} = probability of response j at stage n

u_{in} = random disturbance with zero mean, variance σ_n^2 and $E(u_{in}u_{jn}) = \lambda_n$, $i \neq j$.

Remarks

1. The first-stage model is a conventional multiple regression model which assumes independent responses, but the model for the subsequent stages is an autoregressive model which accounts for the dependencies induced by the feedback process.

2. The model for $n \geq 2$ can be interpreted as saying that panelist i's response from stage (n-1) to stage n is proportional to the "importance" of the reasons given by the panel in stage (n-1) which panelist i did not give. (The p_{jn} are not fed back to the panelists.)

3. The assumed error structure expresses the requirement that the same information is fed back to each panelist on each round and assumes the panel is homogeneous.

4. The model can be used to predict response on a given round from responses on earlier rounds. This capability could be used in situations where it is inconvenient, costly or impossible to carry out the next round of questioning.

5. The expected group mean response after n rounds of qualitative controlled feedback is proposed as an estimate of the group judgment.

6. The model can be extended to the case where there is quantitative feedback (e.g., the mean response), either with or without qualitative feedback.

7. In [10], Press generalizes the model to the multivariate case where there are many related questions of simultaneous interest. The generalization consists of the assumption of an arbitrary covariance matrix for any individual's responses. There is no other assumed interaction among the different questions. Panelists are still assumed to respond independently.

8. Many questions remain to be addressed by empirical research.

   (a) Should an analyst edit the panelists' reasons which appear to be duplicates or paraphrases of other reasons or should he not tamper with the semantic issues which might arise?

   (b) Should panelists generate all of the reasons themselves or should a list be provided?

   (c) Should panelists be questioned by mail, telephone, personal interview, or by on-line computer?

   (d) Should models account for round-to-round changes in responses on a relative or absolute basis?

   (e) Does group polarization disappear under qualitative controlled feedback?

   (f) Most important, how well do the models predict?

4. Bayesian Calibration

In [7], Morris proposes a model whereby a decision maker's prior state of information is modified by expert opinion in a Bayesian framework to produce the decision maker's posterior. The key to the model is the decision maker's subjective calibration of the expert(s). For a single expert, the model takes the following form:

$$\{x|f,\ d\} = k \cdot C(x) \cdot f(x) \cdot \{x|d\}$$

$$= k \cdot f_c(x) \cdot \{x|d\}\ ,$$

where

d = decision maker's prior state of information

$\{x|d\}$ = decision maker's prior

$\{x|f,\ d\}$ = decision maker's posterior

$f = f(x)$ = expert's prior

$C(x)$ = Calibration Function (decision maker's subjective calibration of the expert)

$f_c(x) \equiv C(x) \cdot f(x)$

= expert's subjectively calibrated prior

k = normalizing constant.

For several experts, the model becomes:

$$\{x|\underline{f},\ d\} = k \cdot \underline{C}(x) \cdot f_1(x) \cdot \cdots \cdot f_n(x) \cdot \{x|d\}$$

$$= k \cdot f_s(x) \cdot \{x|d\}\ ,$$

where

$f = (f_1, \cdots, f_n)$ = set of expert priors

$C(x)$ = Joint Calibration Function

$f_s(x) = \underline{C}(x) \cdot f_1(x) \cdot \cdots \cdot f_n(x)$

= surrogate prior.

Remarks

1. In principle, in the single-expert case, the Calibration Function can be "measured" by "obtaining a frequency distribution of performance measures on a large set of variables (including the variable of interest), over which the expert's assessment performance is indistinguishable" ([7], p.14).

2. "In the dependent, multi-expert case, measurement [of the Calibration Function] becomes much more difficult. A set of variables must be found, over which, in rough terms, all experts share the same degree of dependence" ([7], p.14). Accordingly, the consensus problem of combining the experts' priors has been replaced by the decision maker's specification of a Joint Calibration Function.

3. If the experts are "independent", then $\underline{C}(x) = \prod_{i=1}^{n} C_i(x)$. "Of course, situations where the experts are independent are rare. The experts need not associate with each other to be dependent in the probabilistic case" ([7], p.11).

4. "The model may also be extended to the situation in which the experts provide event probabilities as opposed to probability densities on continuous variables. ....In the event case, the expert probabilities are combined using a normalized additive rule, as contrasted to the continuous variable case presented here in which the probabilities are combined using a normalized multiplicative formula" ([7], p.15).

5. These results hold under rather general assumptions. In the paper, the results are proved for normal priors, but Morris asserts that they hold for many other priors. Two further assumptions are also made:

   "Invariance to Scale: The variance of the expert's prior alone provides no information about the uncertain quantity. In other words, the expert's stated confidence in his own prediction ability gives no information independent of his actual prediction. For example, if the only information we have from an expert is a statement that he feels quite knowledgeable about the height of the Eiffel Tower, we have no reason to change our own beliefs about the height unless he further provides his actual assessment."

   "Invariance to Shift: The assessment of the location (i.e., the mean) of the expert's prior is directly related to the revealed value of the uncertain quantity. If the reveale value is shifted by some amount, the assessment of the location of the expert's prior must shift by that amount."

Both of these assumptions can be relaxed without affecting the form of the result. "If the uncertain quantity depends upon the variance of the expert's prior alone, then the assessment of this dependence adds another multiplicative term to the likelihood function" ([7], p.15). If the invariance to shift assumption is relaxed, the Calibration Function has the same form but is more difficult to assess.

5. Probabilistic Scoring Rules

As a performance criteria, probabilistic scoring rules are defined as follows (e.g., cf. [4]):

$E = \{E_j\}$ = a set of mutually exclusive and exhaustive events for which probabilities are desired.

$R = \{R_j\}$ = the probabilities which the estimator reports.

$P = \{P_j\}$ = the (unknown) objective probabilities.

$S(R, j)$ = a reward function (scoring rule) which, after the fact, pays the estimator an amount S, depending on the report R and the event j which occurs.

It is crucial that the estimator be motivated to accurately report his assessment. In other words, the scoring rule should not reward the estimator for deliberately distorting his assessment. This requirement can be met by using only scoring rules whose expected value is a maximum when $R = P$. Such rules are called proper scores and satisfy

$$\sum_j P_j \, S(R, j) \leq \sum_j P_j \, S(P, j) \; .$$

Some examples of proper scores are as follows:

1. Logarithmic Score.

$$S(R, j) = \log R_j$$

The logarithmic score has a number of unique properties.

(a) It is the only rule which depends solely on the probability reported for the event that occurs.

(b) It is the only rule which is additive over successive estimates,

(c) It is the only rule which is invariant over logically equivalent estimates (e.g., estimates expressed in terms of conditional probabilities, disjunctive combinations, and the like).

2. Quadratic Score.

$$S(R, j) = 2R_j - \sum_k R_k^2$$

The quadratic score is the only one where the difference between the expected score of a perfect forecaster (i.e., one that announces P) and one that announces R is a function solely of R - P.

3. "Scientific" Score.

$$S(R, j) = \begin{cases} 1 \text{ if } R_j = \max \{R_k\} \\ 0 \text{ otherwise} \end{cases}$$

This score can be interpreted as the usual score in an objective test (1 for each correct answer and 0 for each incorrect answer) in which the test-taker checks the answer that he thinks is most likely to be correct.

Some properties of a proper score are as follows:

1. A proper score is operational, i.e., it can be assigned on the basis of a single instance.

2. A proper score rewards the forecaster for accuracy, i.e., the expected score increases as the report R gets closer to the actual probability.

3. A proper score rewards a forecaster for honesty. If the forecaster believes Q and reports R, then his subjective expectation is a maximum when R = Q.

4. A proper score rewards the estimator for increasing his information concerning the events before formulating his report.

If there are several estimators, then an "n-heads" rule should be used. An n-heads rule is a scoring rule such that a group of estimators performs better than the individual members of the group.

If the group estimate is the average, then an n-heads rule requires that the expected score of the group be greater than or equal to the average expected score of the individual estimators:

$$\sum_j P_j \, S(\bar{R}, j) \geq \frac{1}{n} \sum_k \sum_j P_j \, S(R_k, j) \ ,$$

where k=1, ···, n is the index for individual estimators,

$$R_k = \{R_{kj}\} = \text{set of probabilities reported by } k$$

$$\bar{R} = \frac{1}{n} \sum_k R_k \ .$$

A necessary and sufficient condition for S(R, j) to be an n-heads rule is that S(R, j) be concave in R. Examples are the logarithmic score and the quadratic score.

An improved n-heads rule can be derived if the method of aggregation is tailored to the form of scoring rule. For example, the geometric mean "fits" the logarithmic score better than the mean. The expected group score using the geometric mean is equal to the average expected individual score plus a term which is an increasing function of the dispersion of the individual estimates but is independent of the objective probabilities P. If the quadratic score is used and the mean is the aggregation function, the group advantage is the sum of the variances of the individual reports.

In general, individual assessments tend to be correlated because they are often based on the same background experience. This is, perhaps, especially true for groups of experts. In [5], Dalkey considers the problem of aggregating expert assessments without knowing anything about the dependency structure among the experts. Dalkey models the experts as inquiry systems and shows that, if a proper scoring rule is used, it is possible to aggregate the individual assessments so that the group assessment is better than any individual assessment.

References

1. Abramson, Lee R. (1978). Forming a consensus from subjective probability distributions. ORSA/TIMS Joint National Meeting, Los Angeles, California, November 15, 1978.

2. Chatterjee, S. and Seneta, E. (1977). Some convergence theorems on repeated averaging. J. Appl. Prob, 14, 89-97.

3. Dalkey, N. (1972). An impossibility theorem for group probability functions. The Rand Corporation, P-4862.

4. — (1977). Group Decision Theory. UCLA School of Engineering and Applied Science. UCLA-ENG-7749.

5. — (1980). Aggregation of probability estimates. TIMS/ORSA Joint National Meeting, Washington, D.C., May 7, 1980.

6. Degroot, Morris H. (1974). Reaching a consensus. J. Amer. Stat. Assoc., 69, 118-121.

7. Morris, Peter A. (1975). Modeling experts. Xerox Palo Alto Research Center, ARG Tech. Report No. 75-2.

8. Moskowitz, Herbert and Bajgier, Steve M. (1978). Validity of the DeGroot model for achieving consensus in panel and Delphi groups. Krannert Graduate School of Management, Purdue Univ., Paper No. 672.

9. Press, James S. (1978). Qualitative controlled feedback for forming group judgments and making decisions. J. Amer. Stat. Assoc., 73, 526-535.

10. — (1979). Multivariate group judgments by qualitative controlled feedback. Dept of Statistics, Univ. of California, Tech. Report No. 39.

11. Winkler, Robert L. (1968). The consensus of subjective probability distributions. Management Sci., 15, B-16-75.