DECISION ANALYSIS
Vol. 8, No. 2, June 2011, pp. 128–144
ISSN 1545-8490 | EISSN 1545-8504 | 11 | 0802 | 0128

informs

DOI 10.1287/deca.1110.0206

Aggregating Large Sets of Probabilistic Forecasts by Weighted Coherent Adjustment

Guanchun Wang, Sanjeev R. Kulkarni, H. Vincent Poor
Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544
{guanchun@princeton.edu, kulkarni@princeton.edu, poor@princeton.edu}

Daniel N. Osherson
Department of Psychology, Princeton University, Princeton, New Jersey 08544, osherson@princeton.edu

Probability forecasts in complex environments can benefit from combining the estimates of large groups of forecasters ("judges"). But aggregating multiple opinions raises several challenges. First, human judges are notoriously incoherent when their forecasts involve logically complex events. Second, individual judges may have specialized knowledge, so different judges may produce forecasts for different events. Third, the credibility of individual judges might vary, and one would like to pay greater attention to more trustworthy forecasts. These considerations limit the value of simple aggregation methods like unweighted linear averaging. In this paper, a new algorithm is proposed for combining probabilistic assessments from a large pool of judges, with the goal of efficiently implementing the coherent approximation principle (CAP) while weighing judges by their credibility. Two measures of a judge's likely credibility are introduced and used in the algorithm to determine the judge's weight in aggregation. As a test of efficiency, the algorithm was applied to a data set of nearly half a million probability estimates of events related to the 2008 U.S. presidential election (∼16,000 judges). Compared with unweighted scalable CAP algorithms, the proposed weighting schemes significantly improved the stochastic accuracy with a comparable run time, demonstrating the efficiency and effectiveness of the weighting methods for aggregating large numbers and varieties of forecasts.

Key words: judgment aggregation; combining forecasts; weighting; incoherence penalty; consensus deviation
History: Received on April 27, 2010. Accepted on April 13, 2011, after 2 revisions.

1. Introduction

Decisions and predictions resulting from aggregating information in large groups are generally considered better than those made by isolated individuals. Probability forecasts are thus often elicited from a number of human judges whose beliefs are combined to form aggregated forecasts. Applications of this approach can be found in many different fields such as data mining, economics, finance, geopolitics, meteorology, and sports (for surveys, see Clemen 1989, Morgan and Henrion 1990, Clemen and Winkler 1999). In many cases, forecasts may be elicited for sets of events that are logically dependent, e.g., the conjunction of two events, and also each event separately. Such forecasts are useful when judges have information about the likely co-occurrence of events or the probability of one event conditional upon another.[1] Including complex events in queries to judges may thus potentially improve the accuracy of aggregate forecasts.

1.1. Aggregation Principles and Practical Algorithms

Combining probabilistic forecasts over both simple and complex events requires sophisticated aggregation because coherence is desired. In particular, mere linear averaging of the probabilities offered for a given event may not yield a coherent aggregate. For one thing, human judges often violate probability axioms (e.g., Tversky and Kahneman 1983, Macchi

[1] Example events might include "President Obama will be reelected in 2012," "the U.S. trade deficit will decrease and the national savings rate will increase in 2011," and "Obama will be reelected if the U.S. unemployment rate drops below 8% by the end of 2011."

et al. 1999, Sides et al. 2002, Tentori et al. 2004), and the linear average of incoherent forecasts is generally also incoherent. Moreover, even if all the judges are individually coherent, when the forecasts of a given judge concern only a subset of events (because of specialization), the averaged results may still be incoherent. Finally, if conditional events are included among the queries, linear averaging will most likely lead to aggregated probabilities that no longer satisfy the definition of conditional probability or Bayes' formula.

To address the foregoing limitations, a generalization of linear averaging was discussed by Batsell et al. (2002) and by Osherson and Vardi (2006). Their idea is known as the coherent approximation principle (CAP). The CAP proposes a coherent forecast that is minimally different, in terms of squared deviation, from the judges' forecasts. Unfortunately, the optimization problem required for the CAP is equivalent to an NP-hard decision problem and has poor scalability as the numbers of judges and events increase. To circumvent this computational challenge, Osherson and Vardi proposed a practical method for finding coherent approximations, termed "simulated annealing over probability arrays" (SAPA). However, the simulated annealing needed for SAPA requires numerous parameters to be tuned and still takes an unacceptably long time for large data sets. In Predd et al. (2008), the problem is addressed by devising an algorithm that strikes a balance between the simplicity of linear averaging and the coherence that results from a full solution of the CAP. The Predd et al. (2008) approach uses the concept of "local coherence," which decomposes the optimization problem into subproblems that involve small sets of related events. This algorithm makes it computationally feasible to apply a relaxed version of the CAP to large data sets, such as the one described next.

1.2. Presidential Election Data

Previously published studies of the CAP and the algorithms that implement it (e.g., Batsell et al. 2002, Osherson and Vardi 2006, Predd et al. 2008) involve no more than 30 variables and 50 judges. To fully test the computational efficiency and practical usefulness of the CAP, it is of interest to compare it to rival aggregation methods using a large data set. The 2008 U.S. presidential election provided an opportunity to elicit a very large pool of informed judgments from knowledgeable and interested respondents. In the months prior to the election, we established a website to collect probability estimates of election related events. To complete the survey, respondents were encouraged to provide basic demographic information (gender, party affiliation, age, highest level of education, state of residence) as well as numerical self-ratings of political expertise (before completing the survey) and prior familiarity with the questions (after completion). The respondents were presented 28 questions concerning election outcomes involving seven randomly selected states and were asked to estimate the probability of each outcome. For example, a user might be asked questions about simple events, such as, "what is the probability that Obama wins Indiana?" and also questions about complex events like, "what is the probability that Obama wins Vermont and McCain wins Texas?" or "what is the probability that McCain wins Florida supposing that McCain wins Maine?" The respondents provided estimates of these probabilities with numbers from 0% to 100%. In the end, nearly 16,000 respondents completed the survey, and approximately half a million estimates were collected.

1.3. Goals of the Study

This is the first study in which the large size of the data set allows us to fully evaluate the computational efficiency of rival aggregation methods. The scope of the study also raises the issue of the varied quality of individual judges and the importance of weighting them accordingly. Hence, our goal is to develop an algorithm that can efficiently implement a relaxed version of the CAP and at the same time allow judges to be weighted by an objective measure of their credibility. There is a long history of debate about whether simple averages or weighted averages work better (for review, see Winkler and Clemen 1992, Clemen 1989, Bunn 1985). We hope to demonstrate from our study the superior forecasting gains that result from combining credibility weighting with coherentization. We will also show theoretically and empirically that smart weighting allows particular subsets of events to be aggregated more accurately. Last, we will compare

the aggregated forecasts (of the simple events) generated from these weighted algorithms with the probabilistic predictions provided from popular prediction markets and poll aggregators.

1.4. Outline

The remainder of this paper is organized as follows. In §2, we first introduce notation. Then we review the CAP and a scalable algorithm for its approximation. Performance guarantees for the algorithm are also discussed. In §3, we propose the *weighted coherentization algorithm* and define two penalty measures to reflect an individual judge's credibility. We also investigate the correlation between these measures and their accuracy measured by Brier score (Brier 1950). Yet other measures of forecasting accuracy are introduced in §4, and they are used for comparing weighted coherentization and other aggregation methods. In §5, we show that coherent adjustment can improve the accuracy of forecasts for logically elementary/simple events provided that judges are good forecasters for complex events. We conclude in §6 with a discussion of implications and extensions.

2. The Scalable Approach to Applying the CAP

2.1. Coherent Approximation Principle

Let $\mathbf{\Omega}$ be a finite *outcome space* so that subsets of $\mathbf{\Omega}$ are *events*. A forecast is defined to be a mapping from a set of events to estimated probability values, i.e., $f\colon \mathscr{E} \to [0,1]^n$, where $\mathscr{E} = \{E_1, \ldots, E_n\}$ is a collection of events. Also we let $\mathbf{1}_E\colon \mathbf{\Omega} \to \{0,1\}$ denote the indicator function of an event E. We distinguish two types of events: simple events and complex events formed from simple events using basic logic and conditioning.[2] Because subjective probability estimates of human judges are often incoherent, it is commonplace to have incoherence within a single judge and among a panel of judges. To compensate for the inadequacy of linear averaging when incoherence is present, the coherent approximation principle was proposed by Osherson and Vardi (2006) with the following definition of coherence.

[2] For ease of exposition in what follows, we often tacitly assume that conditioning events are assigned positive probabilities by forecasters. If not, the estimates of the corresponding conditional events will be disregarded.

Definition 1. A forecast f over a set of events $\mathscr{E}$ is probabilistically coherent if and only if it conforms to some probability distribution on $\mathbf{\Omega}$, i.e., there exists a probability distribution g on Ω such that $f(E) = g(E)$ for all $E \in \mathscr{E}$.

With a panel of judges each evaluating a (potentially different) set of events, the CAP achieves coherence with minimal modification of the original judgments, which can be mathematically formulated as the following optimization problem:

$$\min_{f(E)} \sum_{i=1}^{m} \sum_{E \in \mathscr{E}_i} (f(E) - f_i(E))^2 \quad \text{s.t. } f \text{ is coherent.} \tag{1}$$

Here we assume a panel of m judges, where $\mathscr{E}_i$ denotes the set of events evaluated by judge i; the forecasts $\{f_i\}_{i=1}^m$ are the input data, and f is the output of (1), which is a coherent aggregate forecast for the events in $\mathscr{E} = \bigvee_{i=1}^m \mathscr{E}_i$.

2.2. A Scalable Approach

Although the CAP can be framed as a constrained optimization problem with $|\mathscr{E}|$ optimization variables, it can be computationally infeasible to solve using standard techniques when there is a great number of judges forecasting a very large set of events (e.g., our election data set). In addition, the nonconvexity introduced by the ratio equality constraints from conditional probabilities might lead to local minima solutions. In Predd et al. (2008), the concept of local coherence was introduced, motivated by the fact that the logical complexity of events that human judges can assess is usually bounded, typically not going beyond a combination of three simple events or their negations. Hence, the global coherence constraint can be well approximated by sets of local coherence constraints, which in turn allows the optimization problem to be solved using the successive orthogonal projection (SOP) algorithm (for related material, see Censor and Zenios 1997, Bauschke and Borwein 1996). Below, we reproduce the definition of local coherence and the formulation of the optimization program.

Definition 2. Let $f\colon \mathscr{E} \to [0,1]$ be a forecast, and let $\mathscr{F}$ be a subset of $\mathscr{E}$. We say that f is *locally coherent with respect to the subset* $\mathscr{F}$ if and only if f restricted

Table 1 Local Coherence Example

$\mathscr{E}$	E_1	E_2	$E_1 \wedge E_2$	$E_1 \vee E_2$	$E_1 \mid E_2$
f	0.8	0.5	0.2	0.9	0.4

to $\mathscr{F}$ is probabilistically coherent, i.e., there exists a probability distribution g on Ω such that $g(E) = f(E)$ for all $E \in \mathscr{F}$.

We illustrate this via Table 1. We see that f is not locally coherent with respect to $\mathscr{F}_1 = \{E_1, E_2, E_1 \wedge E_2\}$ because $f(E_1) + f(E_2) - f(E_1 \wedge E_2) > 1$, whereas f is locally coherent with respect to $\mathscr{F}_2 = \{E_1, E_2, E_1 \vee E_2\}$ and $\mathscr{F}_3 = \{E_2, E_1 \wedge E_2, E_1 \mid E_2\}$. Note that f is not globally coherent (namely, coherent with respect to $\mathscr{E}$) in this example, because global coherence requires that f be locally coherent with respect to all $\mathscr{F} \subseteq \mathscr{E}$.

With the relaxation of global coherence to local coherence with respect to a collection of sets $\{\mathscr{F}_l\}_{l=1}^L$, the optimization problem (1) can be modified to

$$\min_{f(E)} \sum_{i=1}^{m} \sum_{E \in \mathscr{E}_i} (f(E) - f_i(E))^2 \quad \text{s.t. } f \text{ is locally coherent w.r.t. } \mathscr{F}_l \quad \forall l = 1, \ldots, L. \tag{2}$$

We can consider solving this optimization problem as finding the projection onto the intersection of the spaces formed by the L sets of local coherence constraints so that the SOP algorithm fits naturally into the judgment aggregation framework.

The computational advantage of this iterative algorithm is that the CAP can now be decomposed into subproblems that require the computation and update of only a small number of variables determined by the local coherence set $\mathscr{F}_l$. If the local coherence set consists of only absolute[3] probability estimates, the subproblem becomes essentially a quadratic program, where analytic solutions exist. If the local coherence set involves a conditional probability estimate that imposes a ratio equality constraint, the subproblem can be converted to a simple unconstrained optimization problem after substitution. Both cases can be solved efficiently. So the trade-off between complexity and speed depends on the selection of the local coherence sets $\{\mathscr{F}_l\}_{l=1}^L$, which is a design choice that an analyst needs to make. Note that when each set includes only one event, the problem degenerates to linear averaging, and on the other hand, when all events are grouped into one single set, this case becomes the same as requiring global coherence. Fortunately, we can often approximate global coherence using a collection of local coherence sets in which only a small number of events are involved, because complex events in most surveys are formed with the consideration of the limited logical capacity of human judges and therefore involve a small number of simple events.

It is also shown in Predd et al. (2008) that, regardless of the eventual outcome, the scalable approach guarantees stepwise improvement in stochastic accuracy (excluding forecasts of conditional events) measured by Brier score, or "quadratic penalty," which is defined as follows:

$$\mathrm{BS}(f) = \frac{1}{|\mathscr{E}|} \sum_{E \in \mathscr{E}} (\mathbf{1}_E - f(E))^2. \tag{3}$$

So we choose $\{\mathscr{F}_l\}_{l=1}^L$ based on each complex event and the corresponding simple events to achieve efficient and full coverage on $\mathscr{E}$. However, it should be noted that there is no theoretical guarantee that stepwise improvement and convergence in accuracy hold for the SOP algorithm when it is applied to conditional probability estimates. Nevertheless, empirically we still observe such improvement, as will be shown in §4 below.

3. Weighted Coherent Adjustment

As suggested in Predd et al. (2008), the scalable CAP approach might be extended to "allow judges to be weighted according to their credibility." We now propose an aggregation method that minimally adjusts forecasts to achieve probabilistic coherence while assigning greater weight to potentially more credible judges. This is of particular interest when the number of judges and events is large and the credibility of the original estimates can vary considerably among judges. Even though the exact forecasting accuracy can be measured only after the true outcomes are revealed, it is reasonable to assume that accuracy of the aggregated result can be improved if

[3] *Absolute* is used here to refer to events that are not conditional. Therefore, negation, conjunction, and disjunction events are all absolute.

larger weights are assigned to "wiser" judges selected by a well-chosen credibility measure.

Let w_i be a weight assigned to judge i. We use w_i as a scaling factor to control how much the ith judge's forecast is to be weighted in the optimization problem. It is easy to show that minimizers of the following three objective functions are equivalent:

1. $\sum_{i=1}^{m} (w_i \sum_{E\in\mathscr{E}_i} (f(E) - f_i(E))^2)$;
2. $\sum_{E\in\mathscr{E}} ((\sum_{i:E\in\mathscr{E}_i} w_i) f(E)^2 - 2\sum_{i:E\in\mathscr{E}_i} w_i f_i(E) f(E))$,

where $\mathscr{E} = \bigvee_{i=1}^{m} \mathscr{E}_i$ is the union of all events;

3. $\sum_{E\in\mathscr{E}} Z(E)(f(E) - \hat{f}(E))^2$, where $Z(E) = \sum_{i:E\in\mathscr{E}_i} w_i$ is a normalization factor and $\hat{f}(E) = \sum_{i:E\in\mathscr{E}_i} w_i f_i(E)/Z(E)$ is the weighted average of estimates from judges who have evaluated event E.

Hence, we can revise (2) to the following optimization problem by incorporating the weighting of judges:

$$\min_{f(E)} \sum_{E\in\mathscr{E}} Z(E)(f(E) - \hat{f}(E))^2 \quad \text{s.t. } f \text{ is locally coherent w.r.t. } \mathscr{F}_l \quad \forall l = 1, \ldots, L. \tag{4}$$

3.1. Measures of Credibility

Studies on the credibility of human judges reveal that weights can be determined by investigating judges' expertise and bias (Birnbaum and Stegner 1979). However, such information might be difficult to obtain, especially in a large-scale survey. As an alternative, judges can be encouraged to report their confidence in their own judgments, before and after the survey. Unfortunately, subjective confidence often demonstrates relatively low correlation with performance and accuracy (for discussion, see, e.g., Tversky and Kahneman 1974, Mabe and West 1982, Stankov and Crawford 1997). This phenomenon is also confirmed by our presidential election data set, as will be shown below. Nor do our election data include multiple assessments of the same judge through time; historical performance or "track record" is thus unavailable as a measure of credibility.

The goal of the present study is thus to test weighted aggregation schemes for situations in which

- there is a large number of judges whose expertise and bias information cannot be measured because of required anonymity or resource constraints;
- self-reported credibility measures are unreliable;
- we have data for only one epoch, either because the events involved are one-off in nature or because the track records of the individual judges are not available; and
- each judge evaluates a significant number of events, both simple and complex.

Within this framework, we propose two objective measures of credibility following the heuristics that can be informally described as follows:

1. more coherent judges are more credible in their forecasts;

2. judges whose estimates are closer to consensus are more credible in their forecasts.

3.2. Incoherence Penalty

The first heuristic is partially motivated by de Finetti's (1974) theorem, which says that any incoherent forecast is dominated by some coherent forecast in terms of Brier score for all possible outcomes. Thus, we might expect that the more incoherent the judge, the less credible the forecast. Moreover, coherence is a plausible measure of a judge's competence in probability and logic, as well as the care with which she/he responds to the survey. We therefore define a measure of distance of the judge's forecast from coherence. This measure will be termed *incoherence penalty* (IP). It is calculated in two steps. First, we compute the minimally adjusted coherent forecast of the individual judge. Second, we take the squared distance between the coherentized forecast and the original forecast. Note that the first step is a special case of solving (1) with only one judge. Formally, incoherence penalty can be defined as follows.

Definition 3. For a panel of m judges, let f_i be the original forecast on $\mathscr{E}_i$ given by judge i, and let f_i^{IP} be the coherence-adjusted output from solving the CAP on the single judge space $(f_i, \mathscr{E}_i)$. The incoherence penalty of judge i is defined as $\mathrm{IP}_i = \sum_{E\in\mathscr{E}_i} (f_i^{\mathrm{IP}} - f_i)^2$.

We note that a nonsquared deviation measure can also be used. For example, we tested absolute deviation, and it yielded similar forecasting accuracy to that obtained by using squared deviation.

Because the number of events to be evaluated by one judge is usually moderate because of specialization and time constraints (e.g., there are 7 queries on simple events and 28 questions altogether in our presidential election forecast study), f_i^{IP} can be efficiently computed using, for example, SAPA (see Osherson

Figure 1 Correlation Plot (15,940 Judges): Brier Score vs. Incoherence Penalty

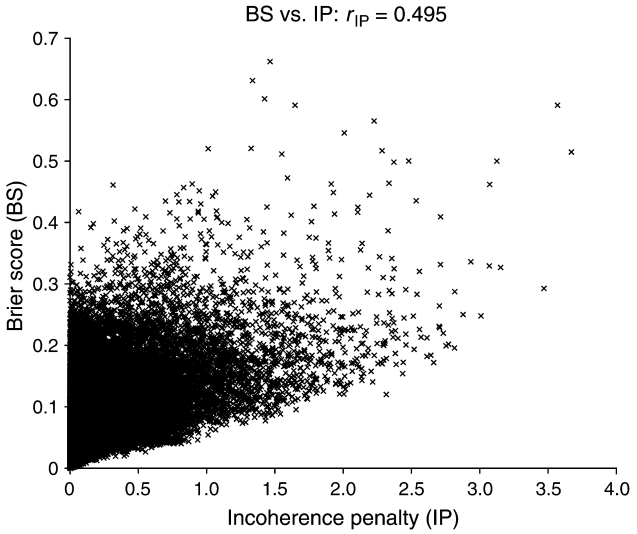


Table 2 Mean Brier Scores of Judges by Quartiles of IP

	First quarter	Second quarter	Third quarter	Fourth quarter
Incoherence penalty	0–0.025	0.025–0.133	0.133–0.432	0.432–3.153
Mean Brier score	0.056	0.088	0.123	0.153

and Vardi 2006) or the scalable algorithm we reviewed in §2.

To test the hypothesis that coherent judges in the election study are more accurate, we computed the correlation between each judge's incoherence penalty and her Brier score ($N = 15{,}940$). We expect a positive coefficient because the Brier score acts like a penalty. In fact, the correlation is 0.495; see the scatter plot in Figure 1. A quartile analysis is given in Table 2, where we see a decrease in accuracy between quartiles.

3.3. Consensus Deviation

Forecasts based on the central tendency of a group are often better than what can be expected from single members, especially in the presence of diversity of opinion, independence, and decentralization (Surowiecki 2004). It may therefore be expected that judges whose opinions are closer to the group's central tendency are more credible. To proceed formally, we use the linear average as a consensus proxy and define *consensus deviation* (CD) as follows.

Definition 4. For a panel of m judges, let f_i be the original forecast given by judge i. For any event $E \in \mathscr{E}_i$,

Figure 2 Correlation Plot (15,940 Judges): Brier Score vs. Consensus Deviation

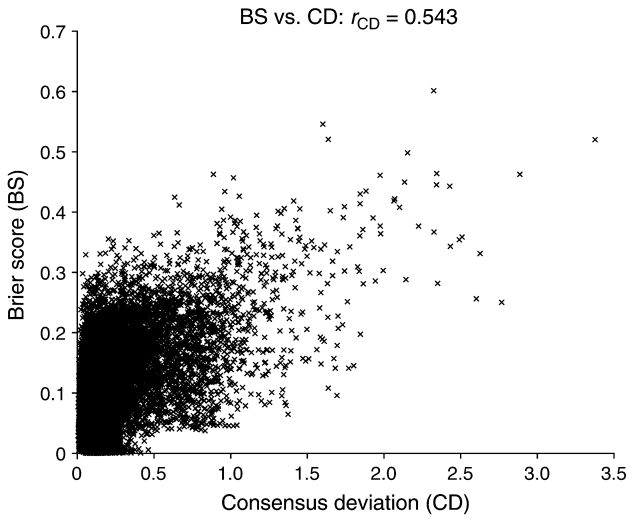


Table 3 Mean Brier Scores of Judges by Quartiles of CD

	First quarter	Second quarter	Third quarter	Fourth quarter
Consensus deviation	0.005–0.078	0.078–0.130	0.130–0.240	0.240–3.376
Mean Brier score	0.074	0.081	0.100	0.165

we let $f^{\mathrm{CD}}(E) = (\sum_{j:E\in\mathscr{E}_j} f_j(E))/N_E$, where N_E is the number of judges that evaluate the event. Then the *consensus deviation* of judge i is $\mathrm{CD}_i = \sum_{E\in\mathscr{E}_i}((f^{\mathrm{CD}}(E) - f_i(E))^2)$.

Again, we use our presidential election data to test the relationship between consensus deviation and forecast accuracy. Because the number of estimates for some complex events is small, we compute consensus deviation relative to simple events only.[4] Across judges, the correlation between CD and Brier score is 0.543; Figure 2 provides a scatter plot. Table 3 shows that accuracy declines between quartiles of CD.[5]

[4] On average, each simple event was evaluated over a thousand times, which is far greater than the average number of estimates for a complex event. The exact average number of estimates for an event of a particular type can be found in Table 6.

[5] However, we note that the empirical success seen from the election data set might not be replicated when the sample size is small and/or the sample is unrepresentative or biased.

Table 4 Comparison in Predicting Power of Credibility Measures for Accuracy

	Pre-Conf.[a]	Post-Conf.[b]	IP	CD
r	−0.154	−0.262	0.495	0.543
R^2	0.024	0.069	0.245	0.295

[a]Self-reported confidence before the survey.
[b]Self-reported confidence after the survey.

3.4. Comparison with Self-Reported Confidence Measures

In our 2008 presidential election forecast study, we asked each respondent to rate his or her level of knowledge of American politics before the survey and his or her prior familiarity with the events presented in the questions after the survey, both on a scale from 0 to 100. These two measures allowed us to roughly learn how confident a judge was in forecasting the election-related events before and after seeing the actual questions. We computed the correlation coefficient and the coefficient of determination between Brier score and the two self-reported confidence measures. The results are summarized in Table 4, which shows that self-reported confidence predicts stochastic accuracy less well than either incoherence penalty or consensus deviation.

3.5. Exponential Weights Using Credibility Penalty Measures

To capture the relationship between accuracy and credibility (as measured inversely by incoherence penalty and consensus deviation), we rely on the following exponential weight function. Given the exponential form of the weighting function, extremely incoherent (or consensus-deviant) judges are given especially low weights.

Definition 5. Let t be either the *incoherence penalty* or the *consensus deviation*. Let the weight function $w\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be defined as

$$w(t) = e^{-\lambda t}, \quad \text{where } \lambda \geq 0 \text{ is a design parameter.} \quad (5)$$

The shape of the exponential function confines weights to the unit interval. To spread them relatively evenly across that interval, we chose λ so that a weight of 0.5 was assigned to the judge with median credibility according to IP (and likewise for CD). This sets $\lambda = 5.2$ for IP and $\lambda = 5.3$ for CD. The incoherence penalty and consensus deviation medians can be seen from Table 2 and Table 3. Later we will see that a sensitivity analysis reveals little impact of modifying λ.

Table 5 The Weighted CAP Algorithm Using IP or CD

Input: Forecasts $\{f_i\}_{i=1}^m$ and collections of events $\{\mathscr{E}_i\}_{i=1}^m$
Step 1: Compute t_i and w_i for all judges
Step 2: Compute the normalizer $Z(E)$ and weighted average $\hat{f}(E)$ for all events
Step 3: Design $\{\mathscr{F}_l\}_{l=1}^L$ for $l = 1, \cdots, L$
Step 4: Let $f_0 = \hat{f}$
  for $t = 1, \cdots, T$
    for $l = 1, \cdots, L$
    $f_t := \arg\min \sum_{E \in \mathscr{E}} Z(E)(f(E) - f_{t-1}(E))^2$
      s.t. f is locally coherent w.r.t. $\mathscr{F}_l$
Output: f_T

3.6. The Weighted CAP Algorithm

Now we have all the pieces for weighted CAP algorithms. The two versions to be considered may be termed the *incoherence-penalty weighted CAP* (IP-CAP) and the *consensus-deviation weighted CAP* (CD-CAP). Their use is summarized in Table 5.

Note that the computational efficiency is achieved by a smart choice of local coherence sets $\{\mathscr{F}_l\}_{l=1}^L$, because the complexity of the optimization is determined by the size of $\mathscr{F}_l$. Within the innermost loop, only the probability estimates of events involved in $\mathscr{F}_l$ are revised and updated. The number of iterations T is a design parameter that needs to be tuned. As $T \to \infty$, our algorithm converges to the solution to (4). In practice, the convergence takes place within a few iterations, and $T = 10$ is often adequate for this purpose. The potential accuracy gain over the simple CAP (sCAP) will come from the weighting effects, as the forecasts from less credible and presumably less accurate judges are discounted.

4. Experimental Results

In the previous section, we presented an algorithm that enforces approximate coherence and weights individual judges according to two credibility penalty measures during aggregation. In this section, we use our presidential election data set to empirically demonstrate the computational efficiency and forecasting accuracy gains of the IP-CAP and CD-CAP compared to rival aggregation methods.

Table 6 Statistics of Different Types of Events

Event type	p	$p \wedge q$	$p \vee q$	$p \mid q$	$p \wedge q \wedge s$	$p \vee q \vee s$	$p \mid q \wedge s$	$p \wedge q \mid s$
No. of questions[a]	7	3	3	3	3	3	3	3
Avg. no. of estimates[b]	1,115.8	9.8	9.8	9.8	1.2	1.6	1.6	1.2

[a]Number of questions of a particular type in a questionnaire.
[b]Average number of estimates for one event of a particular type in the election data set.

4.1. Data

Compared to the previously collected data sets used by Osherson and Vardi (2006) and Predd et al. (2008), the presidential election data set is richer in three ways:[6] (i) the total number of judges (15,940 respondents completed our survey); (ii) the number of simple events (50 variables, one for each state of the union), which induces an outcome space Ω of size 2^{50}; and (iii) the number of different event types (three-term conjunction, disjunction and conditional events are also included). Table 6 lists all the event types as well as the number of questions of each particular type in one questionnaire and the average number of estimates for one event of a particular type in the pooled data set. Note that p, q, and s, represent simple events or their corresponding complements, which are formed by switching candidates.[7]

The data set consists of forecasts from only the judges who completed the questions and provided nondogmatic probability estimates for most events to ensure data quality.[8] Each participant was given an independent, randomly generated survey. A given survey was based on a randomly chosen set of 7 (out of 50) states. Each respondent was presented with 28 questions. All the questions related to the likelihood that a particular candidate would win a state, but involved negations, conjunctions, disjunctions, and conditionals, along with elementary events. Up to three states could figure in a complex event, e.g., "McCain wins Indiana given that Obama wins Illinois and Obama wins Ohio." Negations were formed by switching candidates (e.g., "McCain wins Texas" was considered to be the complement of "Obama wins Texas"). The respondent could enter an estimate by moving a slider and then pressing the button underneath the question to record his or her answer. Higher chances were reflected by numbers closer to 100%, and lower chances by numbers closer to 0%. Some of the events were of the form X AND Y. The respondents were instructed that these occur only when both X and Y occur. Other events were of the form X OR Y, which occur only when one or both of X and Y occur. The respondent would also encounter events of the form X SUPPOSING Y. In such cases, he or she was told to assume that Y occurs and then give an estimate of the chances that X also occurs based on this assumption. As background information for the survey, we included a map of results for the *previous* presidential election in 2004 (with red for Republican and blue for Democratic). The respondent could consult the map or just ignore it as he or she chose. The survey can be found at http://electionforecast.princeton.edu/.

4.2. Choice of λ in the Two Weighting Functions

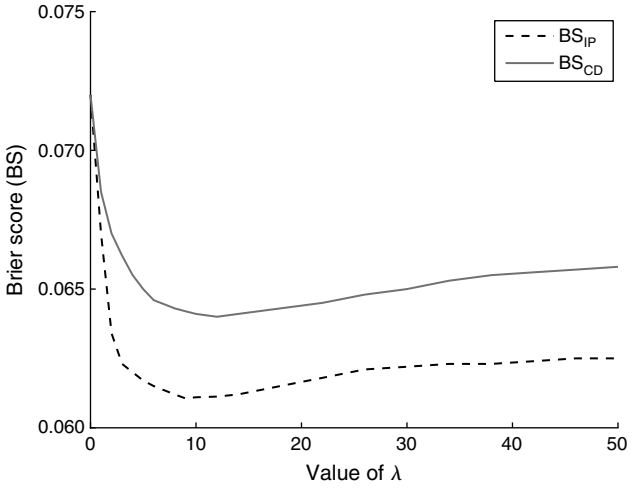
As discussed above, we compared three versions of the CAP, called sCAP, IP-CAP, and CD-CAP. The first employs no weighting function; all judges are treated equally. The IP-CAP weights judges according to their coherence, using the exponential function in Equation (5). The CD-CAP weights judges according to their deviation from the linear average, again using Equation (5). Use of the weighting schemes, however, requires a choice of the free parameter λ; as noted earlier, we have chosen λ so that a weight of 0.5 was assigned to the judge with median credibility according to the IP (and likewise for CD). The choice spreads

[6] To attract judges, we put advertisements with links to our Princeton website on popular political websites such as fivethirtyeight.blogs.nytimes.com and realclearpolitics.com.

[7] In our study, we assume that the probability that any candidate not named "Obama" or "McCain" wins is zero.

[8] Judges who assigned zero or one to more than 14 of 28 questions are regarded as "dogmatic," and their data were excluded from the present analysis. We took this step because we consider those who assigned extreme estimates to more than half of the events to not have understood the instructions of our study. The forecasting accuracy (following coherentization) is actually slightly better if dogmatic judges are left in the pool.

Figure 3 **Sensitivity Analysis in Brier Score w.r.t λ**



the weights relatively evenly across the unit interval. This yields $\lambda = 5.2$ for IP, and $\lambda = 5.3$ for CD. It is worth noting the relative insensitivity of resulting Brier scores to the choice of λ. Indeed, Figure 3 reveals little impact of modifying λ provided that it is chosen to be greater than 5.

4.3. Designing Local Coherence Constraints

As pointed out in Predd et al. (2008), linear averaging and the full CAP are at the opposite extremes of a speed–coherence trade-off, and a smart choice of local coherence sets should strike a good compromise. This is particularly important when there are tens of thousands of judges assessing the probabilities of hundreds of thousands of unique events, as in our presidential election study.

One design heuristic (implemented here) is to leverage the logical relationships between the complex events and their corresponding simple events. The intuition behind such a choice is that most probabilistic constraints arise from how the complex events are described to the judges in relation to the simple events. Following this design, the number of complex events included in a given local coherence set determines the size of the set and thus influences the computation time to achieve local coherence. We illustrate the speed–coherence trade-off spectrum with four kinds of local coherence designs, as follows. The first is linear averaging of the forecasts offered by each judge for a given event. This is the extreme case in which different events do not interact in the aggregation process (see Figure 4). The second aggregation method goes to the opposite extreme, placing all events into one (global) coherence set; we call this the "full CAP" (see Figure 5). The third method is a compromise between the first two, in which each local

Figure 4 **Speed–Coherence Trade-off Spectrum—Linear Averaging**

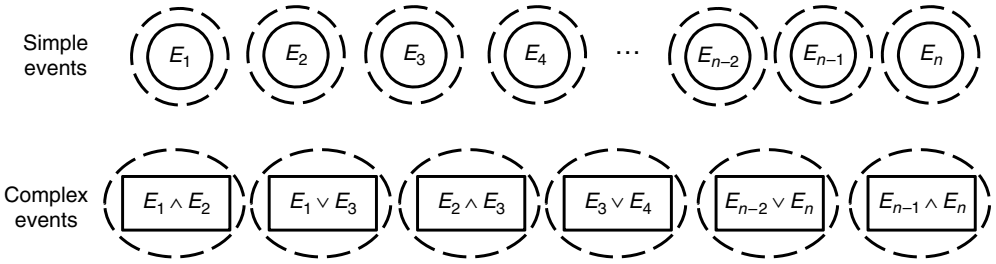


Figure 5 **Speed–Coherence Trade-off Spectrum—Full CAP**

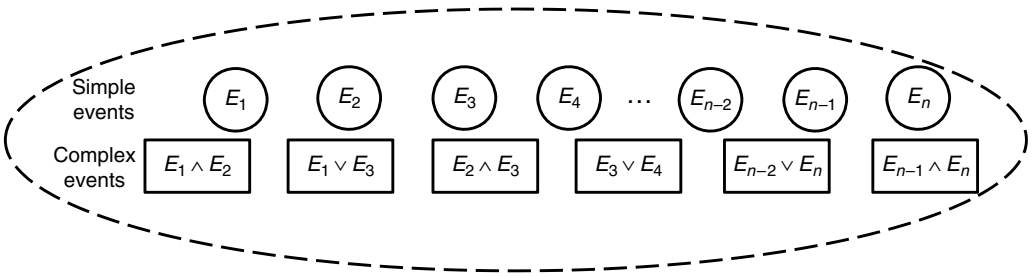


Figure 6 Speed–Coherence Trade-off Spectrum—sCAP(1)

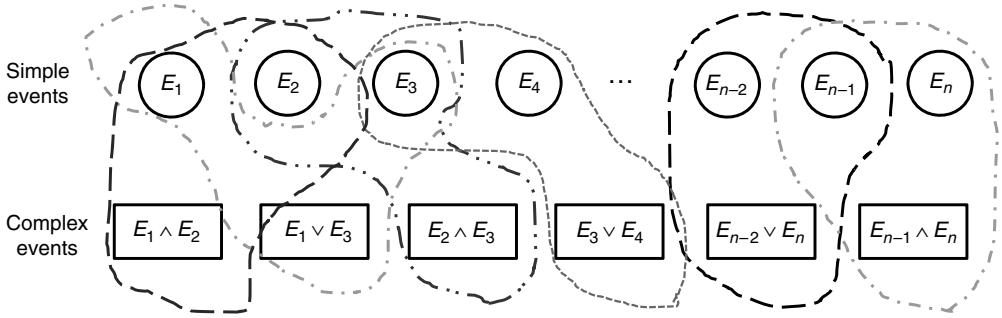
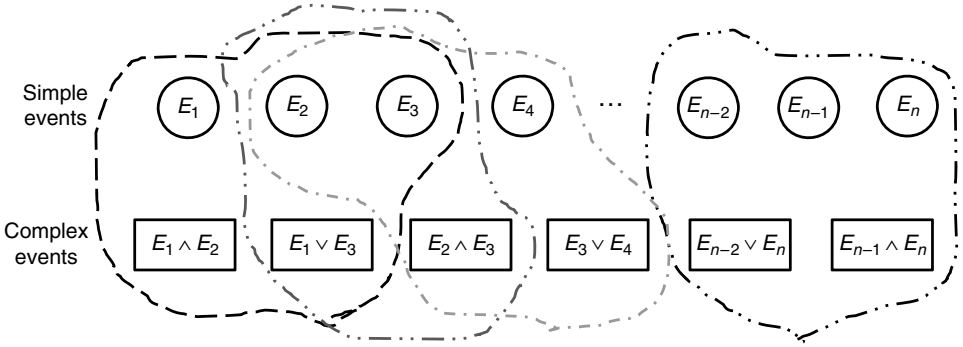


Figure 7 Speed–Coherence Trade-off Spectrum—sCAP(2)



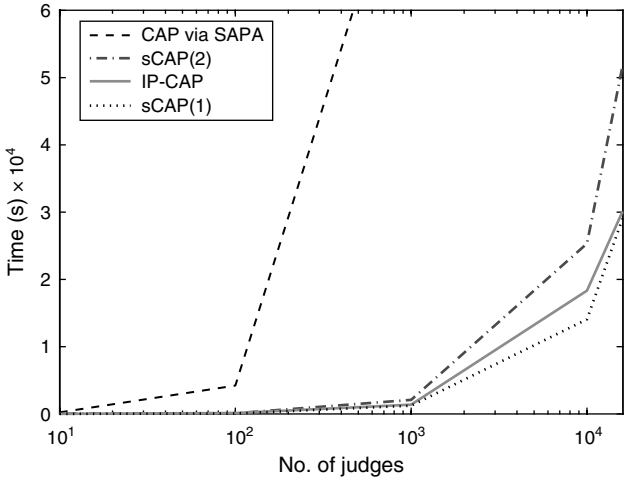
coherence set consists of one complex event and its corresponding simple events; we call this "sCAP(1)" (see Figure 6). The last method is like sCAP(1) except that it leans a little more toward the full CAP. The local coherence set in this case consists of two complex events and their associated and potentially overlapping simple events; this is called "sCAP(2)" (see Figure 7).

4.4. Computational Efficiency and Convergence Patterns

Altogether, we collected 446,880 estimates from 15,940 judges on 179,137 unique events. Such a large data set poses a computationally challenging problem, and therefore provides an opportunity for us to fully evaluate the computational efficiency of various implementations of the scalable CAP (e.g., sCAP(1) and sCAP(2)) versus that of the full CAP (e.g., SAPA; see Osherson and Vardi 2006). Meanwhile, it is also interesting to investigate the trade-off between computation time and forecasting gains (to be discussed in detail in the following subsections) for the unweighted scalable CAP algorithm versus the weighted ones (e.g., IP-CAP). We therefore looked at the overall time spent for each coherentization process to converge with the stopping criterion that the mean absolute deviation of the aggregated forecast from the original forecast changes no more than 0.01%. We varied the size of the data set by selecting estimates from 10, 100, 1,000, 10,000, and all 15,940 judges. The overall time for each aggregation method as a function of the number of judges is the average of five individual runs. All experiments were run on a Dell Vostro with an Intel® Core™ Duo Processor at 2.66 GHz.

Figure 8 shows that CAP implemented via SAPA quickly becomes computationally intractable as the number of judges scales over 1,000. IP-CAP and sCAP(1) are comparable, and both take about eight hours to coherentize all estimates from the 15,940 judges, whereas sCAP(2) requires about 14 hours for the same task. We know sCAP(2) achieves better coherence and potentially higher accuracy than sCAP(1), but we will show in the following subsections that the weighed coherent aggregation via IP-CAP and CD-CAP can achieve greater forecasting gains with less computation time compared

Figure 8 Comparison of Overall Time Spent



to unweighted sCAP(2). In other words, a suitable weighting scheme for the computationally more tractable sCAP(1) can yield greater forecasting accuracy than (unweighted) sCAP(2), despite the greater coherence achieved with the latter.

Figures 9 and 10 detail the convergence patterns and illustrate how the Brier scores of different combinations of events by type evolve versus the number of iterations (T) in our weighted CAP algorithm, using incoherence penalty and consensus deviation, respectively. In both cases and for all combinations of events, our algorithm converges within 10 iterations,

Figure 9 IP-CAP: Brier Score vs. T

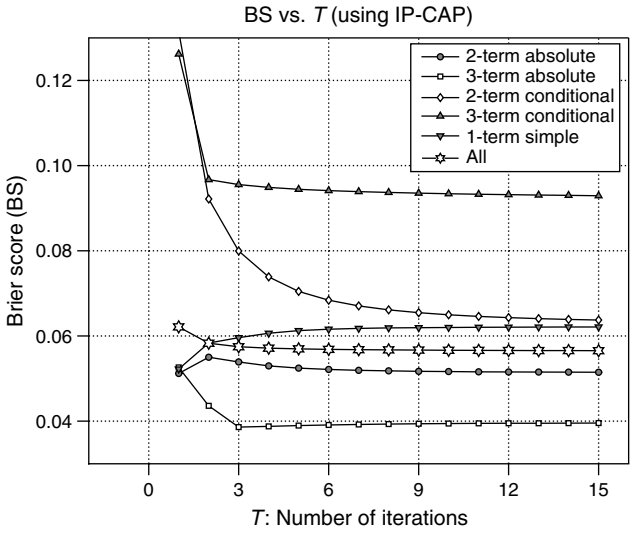
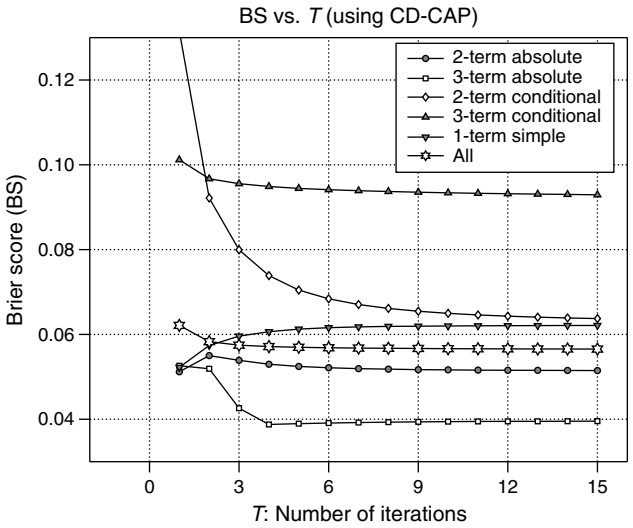


Figure 10 CD-CAP: Brier Score vs. T



with the Brier scores of absolute events stabilizing faster than those of the conditional events.

These two plots also reveal how different types of events interact with each other as our algorithm converges. For one thing, our weighted CAP algorithm reduces the Brier score for the collection of all the events monotonically in both cases, just like the simple CAP algorithm does using the smaller data sets in Predd et al. (2008). Moreover, we can also note that the Brier scores of the less accurate combinations of events (in this case, the complex events, i.e., the two-term and three-term conjunction, disjunction, and conditional events) gradually improve at the expense of the more accurate ones (the simple events). However, the gain achieved for complex events outstrips the loss for simple events, hence the overall improvement in accuracy measured by Brier score.

4.5. Forecasting Accuracy Measures

Rival aggregation methods were compared in terms of their respective stochastic accuracies. For this purpose, we relied on the Brier score (defined earlier) along with the following accuracy measures.

- *Log score*. Like the Brier score, the Log score (also called the "Good score") is a proper scoring rule (for a discussion of proper scoring rules, see Good 1952, Predd et al. 2009), which means the subjective expected penalty is minimized if the judge honestly reports his or her belief of the event. Log score

Table 7 Forecasting Accuracy Comparison Results

	Raw	Linear	CD-Weighted	IP-Weighted	sCAP(1)	sCAP(2)	CD-CAP	IP-CAP
Brier score	0.105	0.085	0.081	0.079	0.072	0.070	0.065	0.062
Log score	0.347	0.306	0.292	0.288	0.278	0.273	0.255	0.243
Correlation	0.763	0.833	0.836	0.84	0.879	0.881	0.887	0.891
Slope	0.560	0.560	0.581	0.592	0.556	0.562	0.589	0.618

is defined as $-(1/|\mathscr{E}|)\sum_{E\in\mathscr{E}} \ln|1-\mathbf{1}_E - f(E)|$, where $\mathscr{E}$ denotes the set of all events excluding the conditional events whose conditioning events turn out to be false. The Log score can take unbounded values when judges are categorically wrong. Here, for the sake of our numerical analysis, we limit the upper bound to 5, because the Log score of an event is 4.6 if a judge is 99% from the truth. (We note that truncating the Log score in this way renders it "improper" technically speaking.)[9]

- *Correlation*. We consider the probability estimate as a predictor for the outcome and compute the correlation between it and the true outcome. Note that this is a reward measure, and hence a higher value means greater accuracy, in contrast to the Brier score.
- *Slope*. The slope of a forecast is the average probability of events that come true minus the average of those that do not. Mathematically, it is defined as $(1/m_T)\sum_{E\in\mathscr{E}:\mathbf{1}_E=1} f(E) - (1/|\mathscr{E}|-m_T)\sum_{E\in\mathscr{E}:\mathbf{1}_E=0} f(E)$, where m_T denotes the number of true events in $\mathscr{E}$. Slope is also a reward measure.

As usual, conditional events enter the computation of these forecasting accuracy measures only if their conditioning events are true.

4.6. Aggregation Methods

We now compare the respective stochastic accuracies of the aggregation methods discussed above along with *Raw*, i.e., the unprocessed forecasts. Brief explanations of the methods are as follows.

- *Linear*. Replace every estimate for a given event with the unweighted linear average of all the estimates of that event.
- *CD-Weighted*. Replace every estimate for a given event with the weighted average of all the estimates of that event, where the weights are determined by the consensus deviation of each individual judge.
- *IP-Weighted*. Replace every estimate for a given event with the weighted average of all the estimates of that event, where the weights are determined by the incoherence penalty of each individual judge.
- *sCAP*(1). Apply the scalable CAP algorithm with one complex event in each local coherence set to eliminate incoherence, and replace the original forecasts with the coherentized ones.
- *sCAP*(2). Apply the scalable CAP algorithm with two complex events in each local coherence set to eliminate incoherence, and replace the original forecasts with the coherentized ones.
- *CD-CAP*. Apply the weighted CAP algorithm with each judge weighted by consensus deviation.
- *IP-CAP*. Apply the weighted CAP algorithm with each judge weighted by incoherence penalty.

4.7. Comparison Results

Table 7 summarizes the comparison results, which show nearly[10] uniform improvement in all four accuracy measures (i.e., Brier score, Log score, correlation, and slope) from raw to simple linear and weighted average, to simple (scalable) CAP, and, finally, to weighted CAP. Note that we confirm the findings of Osherson and Vardi (2006) and Predd et al. (2008) about CAP outperforming the Raw and Linear methods in terms of Brier score and slope. We also observe the following:

- Weighted averaging and weighted CAP, using weights determined by either CD or IP, perform better than simple linear averaging and CAP with respect to all accuracy measures.
- IP is superior to CD for weighting judges inasmuch as both the IP-CAP and IP-Weighted methods yield greater forecast accuracy than either the CD-CAP or CD-Weighted methods.

[9] In the election data set, less than 0.4% of the estimates from judges are categorically wrong and require bounding when computing their Log scores.

[10] The only exception is that *simple linear averaging* reports the same slope as Raw.

Compared to earlier experiments, judges in the election forecast study have the most accurate (raw) forecasts. Nonetheless, our weighted coherentization algorithm improves judges' accuracy quite significantly. Indeed, IP-CAP is 41% better than Raw, 27% better than Linear, and 14% better than the simple CAP as measured by Brier score.

5. Improving the Accuracy of Simple Events by Coherent Adjustment

5.1. Theoretic Guidelines

In our election forecast study, simple events of the form "Candidate A wins State X" might be considered to have the greatest political interest. So let us consider the circumstances in which eliciting estimates of complex events[11] can improve the forecast accuracy of simple events. The following three observations are relevant; their proofs are given in the appendix. We will use them as guidelines to improve the estimates of simple events using complex events.

Observation 1. For a forecast f of one simple event and its complement, i.e., $\mathscr{E} = \{E, E^c\}$, coherent approximation improves (or maintains) the expected Brier score of the simple event E if the estimate of the complement is closer to its genuine probability than the estimate of the simple event is to its genuine probability, i.e., $|f(E^c) - P_g(E^c)| \le |f(E) - P_g(E)|$, where $P_g\colon \mathscr{E} \to [0, 1]$ is the genuine probability distribution.

Observation 2. For a forecast f of one simple event and one conjunction event involving the simple event, i.e., $\mathscr{E} = \{E_1, E_1 \wedge E_2\}$, coherent approximation improves (or maintains) the expected Brier score of the simple event E_1 if the estimate of the conjunction event is closer to its genuine probability than the estimate of the simple event is to its genuine probability, i.e., $|f(E_1 \wedge E_2) - P_g(E_1 \wedge E_2)| \le |f(E_1) - P_g(E_1)|$, where $P_g\colon \mathscr{E} \to [0, 1]$ is the genuine probability distribution.

Observation 3. For a forecast f of one simple event and one disjunction event involving the simple event, i.e., $\mathscr{E} = \{E_1, E_1 \vee E_2\}$, coherent approximation improves (or maintains) the expected Brier score of the simple event E_1 if the estimate of the disjunction event is closer to its genuine probability than the estimate of the simple event is to its genuine probability, i.e., $|f(E_1 \vee E_2) - P_g(E_1 \vee E_2)| \le |f(E_1) - P_g(E_1)|$, where $P_g\colon \mathscr{E} \to [0, 1]$ is the genuine probability distribution.

Table 8 Average Brier Scores by Event Type

Event type	p	$p \wedge q$	$p \vee q$	$p \wedge q \wedge s$	$p \vee q \vee s$
Brier score (all judges)	0.090	0.102	0.116	0.096	0.126
Brier score (top coherent 1/4)	0.061	0.054	0.055	0.041	0.040

These observations suggest that we attempt to improve the accuracy of forecasts of simple events by making them coherent with the potentially more accurate forecasts of the corresponding negations, conjunctions, and disjunctions. For this purpose, we limit attention to judges who are the most coherent individually because, according to the first heuristic discussed in §3.1, they are likely to exhibit the greatest accuracy in forecasting complex events. Table 8 verifies this assumption with respect to the negation, conjunction, and disjunction events.[12]

So instead of taking into account all complex events in the coherentization process, we can use only those from the more coherent judges. This falls under the rubric of the weighted CAP Algorithm we proposed earlier, because it is the special case in which weights for judges are binary. Essentially, we assign weights of 1 to the top quarter of judges and 0 to the bottom three quarters of judges in terms of their coherence and solve the optimization problem (4). This method is termed TQ-CAP (CAP over the top quarter of the judges by coherence).

Figure 11 shows how the Brier scores of forecasts of different types of events converge during coherentization (including how the Brier score of forecasts of simple events becomes lower), and Table 9 confirms our hypothesis that the accuracy of simple events will improve after coherentization, in terms of all of the four measures discussed earlier. This result has two important implications. From an elicitation and survey design perspective, judges should be encouraged to evaluate the chances of complex events even if the

[11] We limit our attention to complex absolute events, i.e., the negation, conjunction, and disjunction events.

[12] However, the accuracy of the conditional event estimates from the top judges is still worse than that of their simple event estimates.

Figure 11 Coherentizing Forecasts from the Top 25% of Judges Ranked by Coherence: Brier Score vs. T

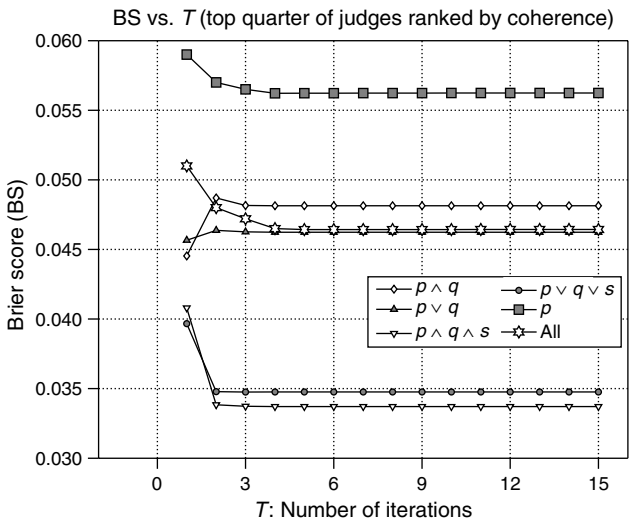


Table 9 Accuracy Improvement by Coherentization in Forecasts of Simple Events from Top Judges

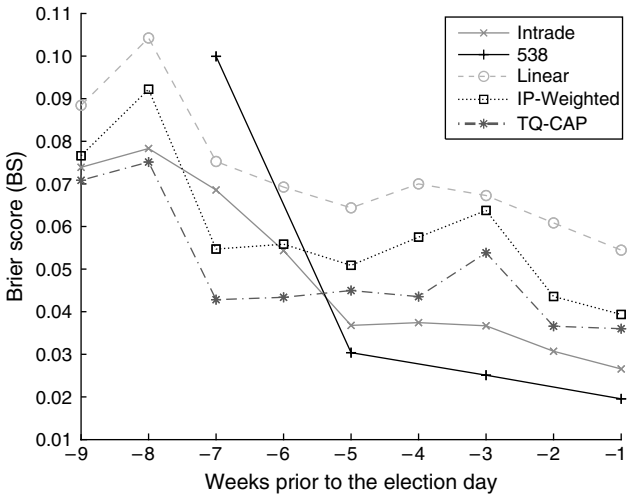
Accuracy measure	Brier score	Log score	Correlation	Slope
Before coherentization	0.061	0.243	0.876	0.590
After coherentization	0.056	0.218	0.928	0.671

primary concern is accurate forecast of simple events, because judges might provide more accurate estimates of complex events that can then be used to improve the accuracy of the forecast of simple events as shown previously. From a judgment aggregation perspective, our results suggest the value of intelligently weighting judges when applying the CAP, notably, via IP.

5.2. Comparison with Poll Aggregators and Prediction Markets

In recent years, there has been growing interest in forecasting elections using public opinion polls and prediction markets as probability aggregators. In this subsection we compare the accuracy of group estimates derived from the election data set with probability estimates provided by http://fivethirtyeight.blogs.nytimes.com (a poll aggregator run by Nate Silver; hereafter, 538) and Intrade (a prediction market). Both sites forecasted state outcomes at several points in time and were highly successful at predicting the 2008 election. Overall, our weighted coherently aggregated

Figure 12 CD-CAP: Brier Score vs. T



forecasts, 538, and Intrade all just predicted one state incorrectly on the eve of the election.[13]

To compare more fully with Intrade and 538, we break the 60-day span prior to election day into nine weeks and compare the Brier scores of all states (i.e., simple events in our election study). For Intrade, we compute the weekly mean of the daily average of the bid and ask prices for each state-level contract and interpret this mean as the event probability expected by the market. For 538, we have four data points at two-week intervals. We compare these forecasts to weekly aggregations of our respondents' predictions, using three methods: Linear, IP-Weighted, and TQ-CAP. The Linear and IP-Weighted methods are defined in §4.6, and TQ-CAP is defined in the previous subsection.

Figure 12 compares the accuracy (measured by Brier score) of the five aggregation methods across time. The first fact to note is that the TQ-CAP always outperforms (uncoherentized) the IP-Weighted method, which in turn outperforms the (unweighted) Linear method. Second, TQ-CAP records higher accuracy than Intrade through five weeks prior to the election day, and IP-Weighted is also comparable with Intrade in that period. Third, 538 performs very well close to the election, but is the worst of the five methods seven

[13] We consider the candidate with a winning probability over 50% as the winner of the state and compare with the true outcome.

weeks prior to election day.[14] To summarize, weighted algorithms (notably TQ-CAP) yield forecasts of simple events that can outperform the most sophisticated forecasting methods, which might require many more participants than our study, especially early in the election period.

6. Conclusion

Making good decisions and predictions on complex issues often requires eliciting probability estimates of both logically simple and complex events from a large pool of human judges. The CAP and the scalable algorithm that implements it overcome the limitation of linear averaging when dealing with incoherence caused by incoherent and specialized judges, and offer the computational efficiency needed for processing large sets of estimates. On the other hand, the credibility of individual judges from a large group could vary significantly because of many factors such as expertise and bias. Hence, incorporating weighting into the CAP framework to aggregate probabilistic forecasts can be beneficial. In this paper, we have introduced two objective penalty measures, namely, the incoherence penalty and consensus deviation, to reflect a judge's credibility and hence determine the weight assigned to his or her judgments for aggregation. Empirical evidence indicates that these measures are more highly correlated with accuracy than self-evaluated confidence measures.

In our 2008 U.S. presidential election forecast study, we collected roughly half a million probability estimates from nearly 16,000 judges to form a very rich data set for empirical evaluation of rival aggregation methods in terms of both efficiency and accuracy. Using the election data set, we show that both broadening the local coherence sets and weighting individual judges during coherentization increase forecasting gains over linear averaging and the simple scalable CAP, i.e., sCAP(1). However, a suitable weighting scheme like IP-CAP or CD-CAP can yield greater forecasting accuracy and remain more computationally tractable compared to unweighted CAP methods with broader local coherence sets like sCAP(2). Overall, four standard forecasting accuracy measures were used to determine the performance of the weighted CAP algorithms in comparison with simple linear averaging, simple/unweighted CAP methods, etc. The results show that coherent adjustments with more weight given to judges who better approximate individual coherence or group consensus consistently produce significantly greater stochastic accuracy measured by Brier score, log score, slope, and correlation. These two objective weighting schemes are also shown to be more effective than using the self-reported confidence measures.

Weighting also allows us to improve the expected forecasting accuracy of the simple events if complex events involving them are more accurate. For the election data set in particular, simple events representing which candidate wins a given state have significant political implications. It may therefore be useful to exploit estimates for complex events to improve the accuracy of predictions of simple events. Three observations have been proved to support this approach with the assumption that estimates of absolute complex events are more accurate. In practice, we have seen that this is possible by limiting attention to the more coherent judges, and can yield more accurate forecasts than popular prediction markets and poll aggregators. Because the weighted coherentized forecast of the election outcomes comes from a moderate number of judges (particularly compared to the probabilistic forecast based on large-scale polls), our algorithm might allow our presidential candidates to make more economic and rational decisions over time (instead of devising campaign strategies blindly after poll results or contract prices on prediction markets). Possible future work can include deriving more general conditions under which coherentization improves the accuracy of a subset of forecasts and studying how to figure in forecasts of conditional events.

Acknowledgments

This research was supported by the U.S. Office of Naval Research under Grant N00014-09-1-0342, the U.S. National Science Foundation (NSF) under Grant CNS-09-05398 and NSF Science & Technology Center Grant CCF-09-39370, and the U.S. Army Research Office under Grant W911NF-07-1-0185. Daniel Osherson acknowledges the Henry Luce Foundation. The authors thank two generous referees who pro-

[14] This is in accord with the observation that polls (the basis of 538's predictions) are highly variable several weeks or more prior to the election but rapidly approach actual outcomes close to election day (see Wlezien and Erikson 2002).

vided many helpful suggestions in response to an earlier draft of this paper.

Appendix. Proofs of Observations Concerning Simple Events

Proof of Observation 1. Let $f_o = (x_o, y_o) = (f(E), f(E^c))$ be the original probability estimates; let $f_g = (x_g, y_g) = (P_g(E), P_g(E^c))$ be the genuine probabilities; and let $f_c = (x_c, y_c)$ be the coherentized probabilities. The coherence space $\mathscr{C}$ for the forecast f is $\{(x, y)\colon x \geq y\}$, and then $x_g + y_g = 1$, because genuine probabilities are always coherent. Because f_c is the projection of f_o onto the coherent space $\mathscr{C}$, we can get $x_c = (x_o - y_o + 1)/2$ and $y_c = (1 - x_o + y_o)/2$. Also by our assumption that the estimate of the complementary event is more accurate than that of the simple event, we have $|x_o - x_g| \geq |y_o - y_g|$. We discuss the following four cases:

1. If $x_o \geq x_g$ and $y_o \geq y_g$, then $x_o - x_g \geq y_o - y_g = y_o - 1 + x_g$. So $x_g \leq (x_o - y_o + 1)/2 = x_c$, i.e., $|x_c - x_g| = x_c - x_g$. Also, because $1 - y_o \leq 1 - y_g = x_g \leq x_o$, $x_c = (x_o - y_o + 1)/2 \leq x_o$. Hence, $|x_c - x_g| = x_c - x_g \leq x_o - x_g = |x_o - x_g|$;

2. If $x_o \geq x_g$ and $y_o < y_g$, then $x_o - x_g \geq y_g - y_o$. So $x_o \geq x_g + y_g - y_o = 1 - y_o$ and $x_c = (x_o - y_o + 1)/2 \leq x_o$. Also, because $1 - y_o > 1 - y_g = x_g$ and $x_o \geq x_g$, $x_c = (x_o - y_o + 1)/2 \geq x_g$. Hence, $|x_c - x_g| = x_c - x_g \leq x_o - x_g = |x_o - x_g|$.

3. If $x_o < x_g$ and $y_o \geq y_g$, then $x_o - x_g \leq y_g - y_o$. So $x_o \leq x_g + y_g - y_o = 1 - y_o$ and $x_c = (x_o - y_o + 1)/2 \geq x_o$. Also, because $1 - y_o < 1 - y_g = x_g$ and $x_o \leq x_g$, $x_c = (x_o - y_o + 1)/2 \leq x_g$. Hence, $|x_c - x_g| = x_g - x_c \leq x_g - x_o = |x_o - x_g|$.

4. If $x_o < x_g$ and $y_o < y_g$, then $x_g - x_o \geq y_g - y_o = 1 - x_g - y_o$. So $x_g \geq (x_o - y_o + 1)/2 = x_c$, i.e., $|x_c - x_g| = x_g - x_c$. Also, because $1 - y_o \geq 1 - y_g = x_g > x_o$, $x_c = (x_o - y_o + 1)/2 > x_o$. Hence, $|x_c - x_g| = x_g - x_c < x_g - x_o = |x_o - x_g|$.

In all cases, the coherentized probability for simple event x_c is either closer to the genuine probability than the original estimate x_o is, or is not changed, i.e., $|x_c - x_g| \leq |x_o - x_g|$.

Also we know the expected Brier score for an event with a genuine probability x_g and an estimate x_o is $\mathbb{E}[\mathrm{BS}(x)] = x_g(1 - x)^2 + (1 - x_g)x^2 = (x - x_g)^2 + x_g - x_g^2$. So by getting closer to the genuine probability through coherentization, the expected Brier score will decrease, i.e., $\mathbb{E}[\mathrm{BS}(x_c)] \leq \mathbb{E}[\mathrm{BS}(x_o)]$. □

Proof of Observation 2. Let $f_o = (x_o, y_o) = (f(E_1), f(E_1 \wedge E_2))$ be the original probability estimates; let $f_g = (x_g, y_g) = (P_g(E_1), P_g(E_1 \wedge E_2))$ be the genuine probabilities; and let $f_c = (x_c, y_c)$ be the coherentized probabilities. The coherence space $\mathscr{C}$ for the forecast f is $\{(x, y)\colon x \geq y\}$, and then $x_g \geq y_g$, because genuine probabilities are always coherent. Also by our assumption that the estimate of the conjunction event is more accurate than that of the simple event, we have $|x_o - x_g| \geq |y_o - y_g|$. There are four cases to consider:

1. If $x_o \geq x_g$ and $y_o \geq y_g$, then $x_o - x_g \geq y_o - y_g$. So $x_o \geq y_o - y_g + x_g \geq y_o$. Hence, $(x_o, y_o) \in \mathscr{C}$ and $f_c = f_o$.

2. If $x_o \geq x_g$ and $y_o < y_g$, then $x_o > x_g \geq y_g > y_o$. Hence, $(x_o, y_o) \in \mathscr{C}$ and $f_c = f_o$.

3. If $x_o < x_g$ and $y_o \geq y_g$, then $x_g - x_o \geq y_o - y_g$, i.e., $x_o + y_o \leq x_g + y_g$. Also we need to consider only the case when $x_o < y_o$, because otherwise $f_c = f_o$. If $x < y$, f_c will be the projection of f_o onto the coherent space $\mathscr{C}$. Then $x_c = y_c = (x_o + y_o)/2 \leq (x_g + y_g)/2 \leq x_g$. Hence, $|x_c - x_g| = x_g - (x_o + y_o)/2 \leq x_g - x_o = |x_o - x_g|$.

4. If $x_o < x_g$ and $y_o < y_g$, also we need to consider only the case when $x_o < y_o$. Then $x_c = (x_o + y_o)/2 < y_o < y_g \leq x_g$. Hence, $|x_c - x_g| = x_g - (x_o + y_o)/2 \leq x_g - x_o = |x_o - x_g|$.

In all cases, the coherentized probability for simple event x_c is either closer to the genuine probability than the original estimate x_o is, or is not changed, i.e., $|x_c - x_g| \leq |x_o - x_g|$. Hence, $\mathbb{E}[\mathrm{BS}(x_c)] \leq \mathbb{E}[\mathrm{BS}(x_o)]$. □

Proof of Observation 3. Coherentizing the forecasts on $\{E_1, E_1 \vee E_2\}$ is equivalent to coherentizing on $\{E_1^c, E_1^C \vee E_2^c\}$ following De Morgan's laws. Also it is easy to show the closeness to genuine probabilities is invariant with respect to negation, and hence, by Observation 2, coherentization brings the $f(E_1^c)$ closer to its genuine probability, which, in turn, improves $f(E_1)$. □

As a matter of fact, the converse of all the three observations can be proved as well, i.e., coherentizing with more accurate simple events can improve the accuracy of its complement, conjunction, and disjunction. The complementary case is straightforward. And we can prove the later two by realizing $\{E_1, E_1 \vee E_2\} = \{(E1 \vee E2) \wedge E1, E_1 \vee E_2\}$ and $\{E_1, E_1 \wedge E_2\} = \{(E1 \wedge E2) \vee E1, E_1 \wedge E_2\}$ and treating $(E1 \vee E2)$ and $(E1 \wedge E2)$ as the "simple events" of each case. Therefore, the proofs of Observations 2 and 3 can be applied.

References

Batsell, R., L. Brenner, D. Osherson, M. Y. Vardi, S. Tsavachidis. 2002. Eliminating incoherence from subjective estimates of chance. *Proc. 8th Internat. Conf. Principles of Knowledge Representation and Reasoning* (*KR 2002*), Morgan Kaufmann, San Mateo, CA, 353–364.

Bauschke, H. H., J. M. Borwein. 1996. On projection algorithms for solving convex feasibility problems. *SIAM Rev.* **38**(3) 367–426.

Birnbaum, M. H., S. E. Stegner. 1979. Source credibility in social judgment: Bias, expertise, and the judge's point of view. *J. Personality Soc. Psych.* **37**(1) 48–74.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* **78**(1) 1–3.

Bunn, D. W. 1985. Statistical efficiency in the linear combination of forecasts. *Internat. J. Forecasting* **1**(2) 151–163.

Censor, Y., S. A. Zenios. 1997. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* **5**(4) 559–583.

Clemen, R. T., R. L. Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk Anal.* **19**(2) 187–203.

de Finetti, B. 1974. *Theory of Probability*, Vol. 1. John Wiley & Sons, New York.

Good, I. J. 1952. Rational decisions. *J. Royal Statist. Soc.* **14**(1) 107–114.

Mabe, P. A., S. G. West. 1982. Validity of self-evaluation of ability: A review and meta-analysis. *J. Appl. Psych.* **67**(3) 280–296.

Macchi, L., D. Osherson, D. H. Krantz. 1999. A note on superadditive probability judgment. *Psych. Rev.* **106**(1) 210–214.

Morgan, M. G., M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, UK.

Osherson, D. N., M. Y. Vardi. 2006. Aggregating disparate estimates of chance. *Games Econom. Behav.* **56**(1) 148–173.

Predd, J. B., D. N. Osherson, S. R. Kulkarni, H. V. Poor. 2008. Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Anal.* **5**(4) 177–189.

Predd, J. B., R. Seiringer, E. H. Lieb, D. N. Osherson, H. V. Poor, S. R. Kulkarni. 2009. Probabilistic coherence and proper scoring rules. *IEEE Trans. Inform. Theory* **55**(10) 4786–4792.

Sides, A., D. Osherson, N. Bonini, R. Viale. 2002. On the reality of the conjunction fallacy. *Memory Cognition* **30**(2) 191–198.

Stankov, L., J. D. Crawford. 1997. Self-confidence and performance on tests of cognitive abilities. *Intelligence* **25**(2) 93–109.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday Books, New York.

Tentori, K., N. Bonini, D. Osherson. 2004. Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psych. Rev.* **28**(3) 467–477.

Tversky, A., D. Kahneman. 1974. Judgment and uncertainty: Heuristics and biases. *Science* **185**(4157) 1124–1131.

Tversky, A., D. Kahneman. 1983. The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Sci.* **90**(4) 293–315.

Winkler, R. L., R. T. Clemen. 1992. Sensitivity of weights in combining forecasts. *Oper. Res.* **40**(3) 609–614.

Wlezien, C., R. S Erikson. 2002. The timeline of presidential election campaigns. *J. Politics* **64**(4) 969–993.