

Modeling Probability Forecasts via Information Diversity

Ville A. Satopää, Robin Pemantle, and Lyle H. Ungar*

Abstract

Randomness in scientific estimation is generally assumed to arise from unmeasured or uncontrolled factors. However, when combining subjective probability estimates, heterogeneity stemming from people's cognitive or information diversity is often more important than measurement noise. This paper presents a novel framework that uses partially overlapping information sources. A specific model is proposed within that framework and applied to the task of aggregating the probabilities given by a group of forecasters who predict whether an event will occur or not. Our model describes the distribution of information across forecasters in terms of easily interpretable parameters and shows how the optimal amount of *extremizing* of the average probability forecast (shifting it closer to its nearest extreme) varies as a function of the forecasters' information overlap. Our model thus gives a more principled understanding of the historically *ad hoc* practice of extremizing average forecasts. Supplementary material for this article is available online.

Keywords: Expert belief; Gaussian process; Judgmental forecasting; Model averaging; Noise reduction

*Ville A. Satopää is a Doctoral Candidate, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340 (e-mail: satopaa@wharton.upenn.edu); Robin Pemantle is a Mathematician, Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395 (e-mail: pemantle@math.upenn.edu); Lyle H. Ungar is a Computer Scientist, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6309 (e-mail: ungar@cis.upenn.edu). This research was supported in part by NSF grant # DMS-1209117 and a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. The authors would like to thank Edward George and Shane Jensen for helpful discussions.

1. INTRODUCTION AND OVERVIEW

1.1 The Forecast Aggregation Problem

Probability forecasting is the science of giving probability estimates for future events. Typically more than one different forecast is available on the same event. Instead of trying to guess which prediction is the most accurate, the predictions should be combined into a single consensus forecast (Armstrong, 2001). Unfortunately, the forecasts can be combined in many different ways, and the choice of the combination rule can largely determine the predictive quality of the final aggregate. This is the principal motivation for the problem of *forecast aggregation* that aims to combine multiple forecasts into a single forecast with optimal properties.

There are two general approaches to forecast aggregation: empirical and theoretical. Given a training set with multiple forecasts on events with known outcomes, the empirical approach experiments with different aggregation techniques and chooses the one that yields the best performance on the training set. The theoretical approach, on the other hand, first constructs a probability model and then computes the optimal aggregation procedure under the model assumptions. Both approaches are important. Theory-based procedures that do not perform well in practice are ultimately of limited use. On the other hand, an empirical approach without theoretical underpinnings lacks both credibility (why should we believe it?) and guidance (in which direction can we look for improvement?). As will be discussed below, the history of forecast aggregation to date is largely empirical.

The main contribution of this paper is a plausible theoretical framework for forecast aggregation called the *partial information framework*. Under this framework, forecast heterogeneity stems from information available to the forecasters and how they decide to use it. For instance, forecasters studying the same (or different) articles on the presidential election may use distinct parts of the information and hence report different predictions of a candidate winning. Second, the framework allows us to interpret existing aggregators and illuminate aspects that

can be improved. This paper specifically aims to clarify the practice of *probability extremizing*, i.e., shifting an average aggregate closer to its nearest extreme. Extremizing is an empirical technique that has been widely used to improve the predictive performance of many simple aggregators such as the average probability. Lastly, the framework is applied to a specific model under which the optimal aggregator can be computed.

1.2 Bias, Noise, and Forecast Assessment

Consider an event A and an indicator function $\mathbf{1}_A$ that equals one or zero depending whether A happens or not, respectively. There are two common yet philosophically different approaches to linking A with the probability forecasts. The first assumes $\mathbf{1}_A \sim \text{Bernoulli}(\theta)$, where θ is deemed a “true” or “objective” probability for A , and then treats a probability forecast p as an estimator of θ (see, e.g., Lai et al. 2011, and Section 2.2 for further discussion). The second approach, on the other hand, treats p as an estimator of $\mathbf{1}_A$. This links the observables directly and avoids the controversial concept of a “true” probability; for this reason it is the approach adopted in this paper.

As is the case with all estimators, the forecast’s deviation from the truth can be broken into bias and noise. Given that these components are typically handled by different mechanisms, it is important, on the theoretical level, to consider them as two separate problems. This paper focuses on noise reduction. Therefore, each forecaster is considered *calibrated* (conditionally unbiased given the forecast), which means that, in the long run, events associated with a forecast $p \in [0, 1]$ occur with an empirical frequency of p .

A forecast (individual or aggregate) is typically assessed with a loss function $L(p, \mathbf{1}_A)$. A loss function is called *proper* or *revealing* if the Bayesian optimal strategy is to tell the truth. In other words, if the subjective probability estimate is p , then $t = p$ should minimize the expected loss $pL(t, 1) + (1 - p)L(t, 0)$. Therefore, if a group of sophisticated forecasters operates under a proper loss function, the assumption of calibrated forecasts is, to some degree, self-fulfilling.

There are, however, many different proper loss function, and an estimator that outperforms another under one loss function will not necessarily do so under a different one. For example, minimizing the quadratic loss function $(p - \mathbf{1}_A)^2$, also known as the Brier score, gives the estimator with the least variance. This paper concentrates on minimizing the variance of the aggregators, though much of the discussion holds under general proper loss functions. See Hwang and Pemantle (1997) for a discussion of proper loss functions.

1.3 The Partial Information Framework

The construction of the partial information framework begins with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable event $A \in \mathcal{F}$ to be forecasted by N forecasters. In any Bayesian setup, with a proper loss function, it is more or less tautological that Forecaster i reports $p_i = \mathbb{E}(\mathbf{1}_A | \mathcal{F}_i)$, where $\mathcal{F}_i \subseteq \mathcal{F}$ is the *information set* used by the forecaster. This assumption is not in any way restrictive. To see this, recall that for a calibrated forecaster $\mathbb{E}(\mathbf{1}_A | p_i) = p_i$. Under the given probability model, conditioning on p_i is equivalent to conditioning on some \mathcal{F}_i , and hence $\mathbb{E}(\mathbf{1}_A | p_i) = \mathbb{E}(\mathbf{1}_A | \mathcal{F}_i) = p_i$. Consequently, the assumption $p_i = \mathbb{E}(\mathbf{1}_A | \mathcal{F}_i)$ only requires the existence of a probability model, together with the assumption of calibration. Note that the forecasters operate under the same probability model but make predictions based on different information sets. Therefore, given that $\mathcal{F}_i \neq \mathcal{F}_j$ if $p_i \neq p_j$, forecast heterogeneity stems purely from *information diversity*.

The information sets contain only information actually used by the forecasters. Therefore, if Forecaster i uses a simple rule, \mathcal{F}_i may not be the full σ -field of information available to the forecaster but rather a smaller σ -field corresponding to the information used by the rule. For example, when forecasting the re-election of the president, a forecaster obeying the dictum “it’s the economy, stupid!” might utilize a σ -field containing only economic indicators. Furthermore, if two forecasters have access to the same σ -field, they may decide to use different sub- σ -fields, leading to different predictions. Therefore, information diversity does not only

arise from differences in the available information, but also from how the forecasters decide to use it.

The partial information framework distinguishes two benchmarks for aggregation efficiency. The first is the *oracular* aggregator $p' := \mathbb{E}(\mathbf{1}_A | \mathcal{F}')$, where \mathcal{F}' is the σ -field generated by the union of the information sets $\{\mathcal{F}_i : i = 1, \dots, N\}$. Any information that is not in \mathcal{F}' represents randomness inherent in the outcome. Given that aggregation cannot be improved beyond using all the information of the forecasters, the oracular aggregator represents a theoretical optimum and is therefore a reasonable upper bound on estimation efficiency.

In practice, however, information comes to the aggregator only through the forecasts $\{p_i : i = 1, \dots, N\}$. Given that \mathcal{F}' generally cannot be constructed from these forecasts alone, no practically feasible aggregator can be expected to perform as well as p' . Therefore, a more achievable benchmark is the *revealed* aggregator $p'' := \mathbb{E}(\mathbf{1}_A | \mathcal{F}'')$, where \mathcal{F}'' is the σ -field generated (or revealed) by the forecasts $\{p_i : i = 1, \dots, N\}$. This benchmark minimizes the expectation of any proper loss function (Ranjan and Gneiting, 2010) and can be potentially applied in practice.

Even though the partial information framework, as specified above, is too theoretical for direct application, it highlights the crucial components of information aggregation and hence facilitates formulation of more specific models within the framework. This paper develops such a model and calls it the Gaussian partial information model. Under this model, the information among the forecasters is summarized by a covariance structure. Such a model has sufficient flexibility to allow for construction of many application specific aggregators.

1.4 Organization of the Paper

The next section reviews prior work on forecast aggregation and relates it to the partial information framework. Section 3 discusses illuminating examples and motivates the Gaussian partial information model. Section 4 compares the oracular aggregator with the average probit

score, thereby explaining the empirical practice of probability extremizing. Section 5 derives the revealed aggregator and evaluates one of its sub-cases on real-world forecasting data. The final section concludes with a summary and discussion of future research.

2. PRIOR WORK ON AGGREGATION

2.1 The Interpreted Signal Framework

Hong and Page (2009) introduce the *interpreted signal framework* in which the forecaster's prediction is based on a personal interpretation of (a subset of) the factors or cues that influence the future event to be predicted. Differences among the predictions are ascribed to differing interpretation procedures. For example, if two forecasters follow the same political campaign speech, one forecaster may focus on the content of the speech while the other may concentrate largely on the audience interaction. Even though the forecasters receive the same information, they interpret it differently and therefore are likely to report different estimates of the probability that the candidate wins the election. Therefore forecast heterogeneity is assumed to stem from "cognitive diversity".

This is a very reasonable assumption that has been analyzed and utilized in many other settings. For example, Parunak et al. (2013) demonstrate that optimal aggregation of interpreted forecasts is not constrained to the convex hull of the forecasts; Broomell and Budescu (2009) analyze inter-forecaster correlation under the assumption that the cues can be mapped to the individual forecasts via different linear regression functions. To the best of our knowledge, no previous work has discussed a formal framework that explicitly links the interpreted forecasts to their target quantity. Consequently, the interpreted signal framework, as proposed, has remained relatively abstract. The partial information framework, however, formalizes the intuition behind it and permits models with quantitative predictions.

2.2 The Measurement Error Framework

In the absence of a quantitative interpreted signal model, prior applications have typically relied on the *measurement error framework* that generates forecast heterogeneity from a probability distribution. More specifically, the framework assumes a “true” probability θ , interpreted as the forecast made by an ideal forecaster, for the event A . The forecasters then “measure” some transformation of this probability $\phi(\theta)$ with mean-zero idiosyncratic error. Therefore each forecast is an independent draw from a common probability distribution centered at $\phi(\theta)$, and a recipe for an aggregate forecast is given by the average

$$\phi^{-1} \left\{ \frac{1}{N} \sum_{i=1}^N \phi(p_i) \right\}. \quad (1)$$

Common choices of $\phi(p)$ are the identity $\phi(p) = p$, the log-odds $\phi(p) = \log \{p/(1-p)\}$, and the probit $\phi(p) = \Phi^{-1}(p)$, giving three aggregators denoted in this paper with \bar{p} , p_{\log} , and p_{probit} , respectively. These *averaging aggregators* represents the main advantage of the measurement error framework: simplicity.

Unfortunately, there are a number of disadvantages. First, given that the averaging aggregators target $\phi(\theta)$ instead of 1_A , important properties such as calibration cannot be expected. In fact, the averaging aggregators are uncalibrated and under-confident, i.e., too close to $1/2$, even if the individual forecasts are calibrated (Ranjan and Gneiting, 2010).

Second, the underlying model is rather implausible. Relying on a true probability θ is vulnerable to many philosophical debates, and even if one eventually manages to convince one’s self of the existence of such a quantity, it is difficult to believe that the forecasters are actually seeing $\phi(\theta)$ with independent noise. Therefore, whereas the interpreted signal framework proposes a micro-level explanation, the measurement error model does not; at best, it forces us to imagine that the forecasters are all in principle trying to apply the same procedures to the same data but are making numerous small mistakes.

Third, the averaging aggregators do not often perform very well in practice. For one thing, Hong and Page (2009) demonstrate that the standard assumption of conditional independence poses an unrealistic structure on interpreted forecasts. Any averaging aggregator is also constrained to the convex hull of the individual forecasts, which further contradicts the interpreted signal framework (Parunak et al., 2013) and can be far from optimal on many datasets.

2.3 Empirical Approaches

If one is not concerned with theoretical justification, an obvious approach is to perturb one of these estimators and observe whether the adjusted estimator performs better on some data set of interest. Given that the measurement error framework produces under-confident aggregators, a popular adjustment is to *extremize*, that is, to shift the average aggregates closer to the nearest extreme (either zero or one). For instance, Ranjan and Gneiting (2010) extremize \bar{p} with the CDF of a beta distribution; Satopää et al. (2014a) use a logistic regression model to derive an aggregator that extremizes p_{\log} ; Baron et al. (2014) give two intuitive justifications for extremizing and discuss an extremizing technique that has previously been used by a number of investigators (Erev et al. 1994; Shlomi and Wallsten 2010; and even Karmarkar 1978); Mellers et al. (2014) show empirically that extremizing can improve aggregate forecasts of international events.

These and many other studies represent the unwieldy position of the current state-of-the-art aggregators: they first compute an average based on a model that is likely to be at odds with the actual process of probability forecasting, and then aim to correct the induced bias via *ad hoc* extremizing techniques. Not only does this leave something to be desired from an explanatory point of view, these approaches are also subject to the problems of machine learning, such as overfitting. Most importantly, these techniques provide little insight beyond the amount of extremizing itself and hence lack a clear direction of continued improvement. The present paper aims to remedy this situation by explaining extremization with the aid of a theoretically

based estimator, namely the oracular aggregator.

3. THE GAUSSIAN PARTIAL INFORMATION MODEL

3.1 Motivating Examples

A central component of the partial information models is the structure of the information overlap that is assumed to hold among the individual forecasters. It therefore behooves us to begin with some simple examples to show that the optimal aggregate is not well defined without assumptions on the information structure among the forecasters.

Example 3.1. Consider a basket containing a fair coin and a two-headed coin. Two forecasters are asked to predict whether a coin chosen at random is in fact two-headed. Before making their predictions, the forecasters observe the result of a single flip of the chosen coin. Suppose the flip comes out HEADS. Based on this observation, the correct Bayesian probability estimate is $2/3$. If both forecasters see the result of the same coin flip, the optimal aggregate is again $2/3$. On the other hand, if they observe different (conditionally independent) flips of the same coin, the optimal aggregate is $4/5$.

In this example, it is not possible to distinguish between the two different information structures simply based on the given predictions, and neither $2/3$ nor $4/5$ can be said to be a better choice for the aggregate forecast. Therefore, we conclude that it is necessary to incorporate an assumption as to the structure of the information overlap, and that the details must be informed by the particular instance of the problem. The next example shows that even if the forecasters observe marginally independent events, further details in the structure of information can still greatly affect the optimal aggregate forecast.

Example 3.2. Let $\Omega = \{A, B, C, D\} \times \{0, 1\}$ be a probability space with eight points. Consider a measure μ that assigns probabilities $\mu(A, 1) = a/4$, $\mu(A, 0) = (1 - a)/4$, $\mu(B, 1) =$

$b/4, \mu(B, 0) = (1 - b)/4$, and so forth. Define two events

$$S_1 = \{(A, 0), (A, 1), (B, 0), (B, 1)\},$$

$$S_2 = \{(A, 0), (A, 1), (C, 0), (C, 1)\}.$$

Therefore, S_1 is the event that the first coordinate is A or B , and S_2 is the event that the first coordinate is A or C . Consider two forecasters and suppose Forecaster i observes S_i . Therefore the i th Forecaster's information set is given by the σ -field \mathcal{F}_i containing S_i and its complement. Their σ -fields are independent. Now, let G be the event that the second coordinate is 1. Forecaster 1 reports $p_1 = \mathbb{P}(G|\mathcal{F}_1) = (a + b)/2$ if S_1 occurs; otherwise, $p_1 = (c + d)/2$. Forecaster 2, on the other hand, reports $p_2 = \mathbb{P}(G|\mathcal{F}_2) = (a + c)/2$ if S_2 occurs; otherwise, $p_2 = (b + d)/2$. If ε is added to a and d but subtracted from b and c , the forecasts p_1 and p_2 do not change, nor does it change the fact that each of the four possible pairs of forecasts has probability $1/4$. Therefore all observables are invariant under this perturbation. If Forecasters 1 and 2 report $(a + b)/2$ and $(a + c)/2$, respectively, then the aggregator knows, by considering the intersection $S_1 \cap S_2$, that the first coordinate is A . Consequently, the optimal aggregate forecast is a , which is most definitely affected by the perturbation.

This example shows that the aggregation problem can be affected by the fine structure of information overlap. It is, however, unlikely that the structure can ever be known with the precision postulated in this simple example. Therefore it is necessary to make reasonable assumptions that yield plausible yet generic information structures.

3.2 Gaussian Partial Information Model

The central component of the Gaussian model is a pool of information particles. Each particle, which can be interpreted as representing the smallest unit of information, is either positive or negative. The positive particles provide evidence in favor of the event A , while the negative

particles provide evidence against A . Therefore, if the overall sum (integral) of the positive particles is larger than that of the negative particles', the event A happens; otherwise, it does not. Each forecaster, however, observes only the sum of some subset of the particles. Based on this sum, the forecaster makes a probability estimate for A . This is made concrete in the following model that represents the pool of information with the unit interval and generates the information particles from a Gaussian process.

The Gaussian Model. Denote the pool of information with the unit interval $S = [0, 1]$. Consider a centered Gaussian process $\{X_B\}$ that is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by the Borel subsets $B \subseteq S$ such that $\text{Cov}(X_B, X_{B'}) = |B \cap B'|$. In other words, the unit interval S is endowed with Gaussian white noise, and X_B is the total of the white noise in the Borel subset B . Let A denote the event that the sum of all the noise is positive: $A := \{X_S > 0\}$. For each $i = 1, \dots, N$, let B_i be some Borel subset of S , and define the corresponding σ -field as $\mathcal{F}_i := \sigma(X_{B_i})$. Forecaster i then predicts $p_i := \mathbb{E}(\mathbf{1}_A | \mathcal{F}_i)$.

The Gaussian model can be motivated by recalling the interpreted signal model of Broomell and Budescu (2009). They assume that Forecaster i forms an opinion based on $L_i(Z_1, \dots, Z_r)$, where each L_i is a linear function of observable quantities or cues Z_1, \dots, Z_r that determine the outcome of A . If the observables (or any linear combination of them) are independent and have small tails, then as $r \rightarrow \infty$, the joint distribution of the linear combinations L_1, \dots, L_N will be asymptotically Gaussian. Therefore, given that the number of cues in a real-world setup is likely to be large, it makes sense to model the forecasters' observations as jointly Gaussian. The remaining component, namely the covariance structure of the joint distribution is then motivated by the partial information framework. Of course, other distributions, such as the t -distribution, could be considered. However, given that both the multivariate and conditional Gaussian distributions have simple forms, the Gaussian model offers potentially the cleanest entry into the issues at hand.

Overall, modeling the forecasters' predictions with a Gaussian distribution is rather common. For instance, Di Bacco et al. (2003) consider a model of two forecasters whose estimated log-odds follow a joint Gaussian distribution. The predictions are assumed to be based on different information sets; hence, the model can be viewed as a partial information model. Unfortunately, as a specialization of the partial information framework, this model is a fairly narrow due to its detailed assumptions and extensive computations. The end result is a rather restricted aggregator of two probability forecasts. On the contrary, the Gaussian model sustains flexibility by specializing the framework only as much as is necessary. The following enumeration provides further interpretation and clarifies which aspects of the model are essential and which have little or no impact.

- (i) **Interpretations.** It is not necessary to assume anything about the source of the information. For instance, the information could stem from survey research, records, books, interviews, or personal recollections. All these details have been abstracted away.
- (ii) **Information Sets.** The set B_i holds the information used by Forecaster i , and the covariance $\text{Cov}(X_{B_i}, X_{B_j}) = |B_i \cap B_j|$ represents the information overlap between Forecasters i and j . Consequently, the complement of B_i holds information not used by Forecaster i . No assumption is necessary as to whether this information was unknown to Forecaster i instead of known but not used in the forecast.
- (iii) **Pool of Information.** First, the pool of information potentially available to the forecasters is the white noise on $S = [0, 1]$. The role of the unit interval is for the convenient specification of the sets B_i . The exact choice is not relevant, and any other set could have been used. The unit interval, however, is a natural starting point that provides an alternative interpretation of $|B_j|$ as marginal probabilities for some N events, $|B_i \cap B_j|$ as their pairwise joint probabilities, $|B_i \cap B_j \cap B_k|$ as their three-way joint probabilities, and so forth. This interpretation is particularly useful in analysis as it links the information structure to many known results in combinatorics and geometry. See, e.g., Proposition

3.3. Second, there is no sense of time or ranking of information within the pool. Instead, the pool is a collection of information, where each piece of information has an *a priori* equal chance to contribute to the final outcome. Quantitatively, information is parametrized by the length measure on S .

- (iv) **Invariant Transformations.** From the empirical point of view, the exact identities of the individual sets B_i are irrelevant. All that matters are the covariances $\text{Cov}(X_{B_i}, X_{B_j}) = |B_i \cap B_j|$. The explicit sets B_i are only useful in the analysis.
- (v) **Scale Invariance.** The model is invariant under rescaling, replacing S by $[0, \lambda]$ and B_i by λB_i . Therefore, the actual scale of the model (e.g., the fact that the covariances of the variables X_B are bounded by one) is not relevant.
- (vi) **Specific vs. General Model.** A specific model requires a choice of an event A and Borel sets B_i . This might be done in several ways: a) by choosing them in advance, according to some criterion; b) estimating the parameters $\mathbb{P}(A)$, $|B_i|$, and $|B_i \cap B_j|$ from data; or c) using a Bayesian model with a prior distribution on the unknown parameters. This paper focuses mostly on a) and b) but discusses c) briefly in Section 6. Section 4 provides one result, namely Proposition 4.1 that holds for any (nonrandom) choices of the sets B_i .
- (vii) **Choice of Target Event.** There is one substantive assumption in this model, namely the choice of the half-space $\{X_S > 0\}$ for the event A . Changing this event results in a non-isomorphic model. The current choice implies a prior probability $\mathbb{P}(A) = 1/2$, which seems as uninformative as possible and therefore provides a natural starting point. Note that specifying a prior distribution for A cannot be avoided as long as the model depends on a probability space. This includes essentially any probability model for forecast aggregation.

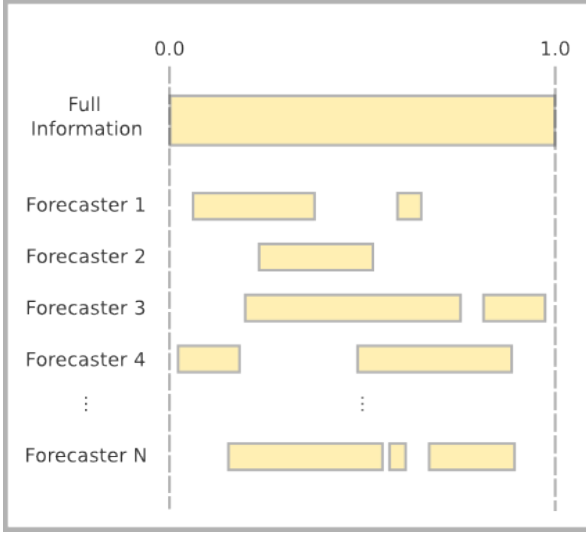


Figure 1: Illustration of Information Distribution among N Forecasters. The bars leveled horizontally with Forecaster i represent the information set B_i .

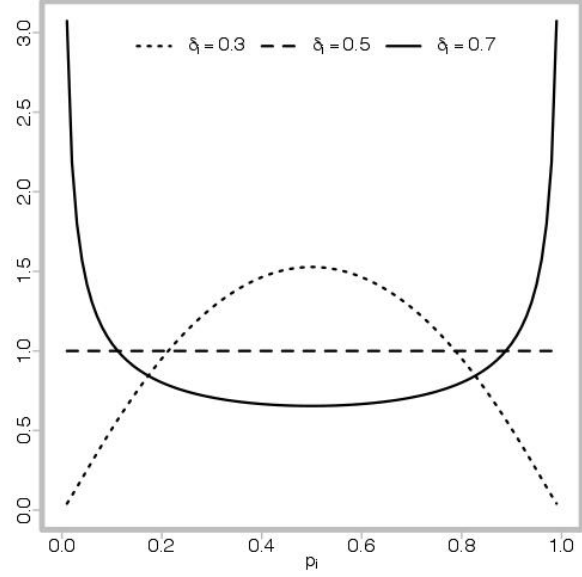


Figure 2: Marginal Distribution of p_i under Different Levels of δ_i . The more the forecaster knows, the more the forecasts are concentrated around the extreme points zero and one.

3.3 Preliminary Observations

The Gaussian process exhibits additive behavior that aligns well with the intuition of an information pool. To see this, consider a partition of the full information $\{C_{\mathbf{v}} := \cap_{i \in \mathbf{v}} B_i \setminus \cup_{i \notin \mathbf{v}} B_i : \mathbf{v} \subseteq \{1, \dots, N\}\}$. Each subset $C_{\mathbf{v}}$ represents information used only by the forecasters in \mathbf{v} such that $B_i = \cup_{\mathbf{v} \ni i} C_{\mathbf{v}}$ and $X_{B_i} = \sum_{\mathbf{v} \ni i} X_{C_{\mathbf{v}}}$. Therefore X_B can be regarded as the sum of the information particles in the subset $B \subseteq S$, and different X_B 's relate to each other in a manner that is consistent with this interpretation. The relations among the relevant variables

are summarized by a multivariate Gaussian distribution:

$$\begin{pmatrix} X_S \\ X_{B_1} \\ \vdots \\ X_{B_N} \end{pmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \dots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \dots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \dots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \dots & \delta_N \end{array} \right) \right), \quad (2)$$

where $|B_i| = \delta_i$ is the amount of information used by Forecaster i , and $|B_i \cap B_j| = \rho_{ij} = \rho_{ji}$ is the amount of information overlap between Forecasters i and j . One possible instance of this setup is illustrated in Figure 1. Note that B_i does not have to be a contiguous subset of S . Instead, each forecaster can use any Borel measurable subset of the full information.

Under the Gaussian model, the sub-matrix Σ_{22} is sufficient for the information structure. Therefore the exact identities of the Borel sets do not matter, and learning about the information among the forecasters is equivalent to estimating a covariance matrix under several restrictions. In particular, if the information in Σ_{22} can be translated into a diagram such as Figure 1, the matrix Σ_{22} is called *coherent*. This property is made precise in the following proposition. The proof of this and other propositions are deferred to Appendix A of the Supplementary Material.

Proposition 3.3. *The overlap structure Σ_{22} is coherent if and only if $\Sigma_{22} \in \text{COR}(N) := \text{conv} \{ \mathbf{x}\mathbf{x}' : \mathbf{x} \in \{0, 1\}^N \}$, where $\text{conv}\{\cdot\}$ denotes the convex hull and $\text{COR}(N)$ is known as the correlation polytope. It is described by 2^N vertices in dimension $\dim(\text{COR}(N)) = \binom{N+1}{2}$.*

The correlation polytope has a very complex description in terms of half-spaces. In fact, complete descriptions of the facets of $\text{COR}(N)$ are only known for $N \leq 7$ and conjectured for $\text{COR}(8)$ and $\text{COR}(9)$ (Ziegler, 2000). Fortunately, previous literature has introduced both linear and semidefinite relaxations of $\text{COR}(N)$ (Laurent et al., 1997). Such relaxations together with modern optimization techniques and sufficient data can be used to estimate the informa-

tion structure very efficiently. This, however, is not in the scope of this paper and is therefore left for subsequent work.

The multivariate Gaussian distribution (2) relates to the forecasts by

$$p_i = \mathbb{P}(A|\mathcal{F}_i) = \mathbb{P}(X_S > 0|X_{B_i}) = \Phi\left(\frac{X_{B_i}}{\sqrt{1-\delta_i}}\right). \quad (3)$$

The marginal density of p_i ,

$$m(p_i|\delta_i) = \sqrt{\frac{1-\delta_i}{\delta_i}} \exp\left\{\Phi^{-1}(p_i)^2\left(1 - \frac{1}{2\delta_i}\right)\right\},$$

has very intuitive behavior: it is uniform on $[0, 1]$ if $\delta_i = 1/2$, but becomes unimodal with a minimum (maximum) at $p_i = 1/2$ when $\delta_i > 1/2$ ($\delta_i < 1/2$). As $\delta_i \rightarrow 0$, p_i converges to a point mass at $1/2$. On the other hand, as $\delta_i \rightarrow 1$, p_i converges to a correct forecast whose distribution has atoms of weight $1/2$ at zero and one. Therefore a forecaster with no information “withdraws” from the problem by predicting a non-informative probability $1/2$ while a forecaster with full information always predicts the correct outcome with absolute certainty. Figure 2 illustrates the marginal distribution when δ_i is equal to 0.3, 0.5, and 0.7.

4. PROBABILITY EXTREMIZING

4.1 Oracular Aggregator for the Gaussian Model

Recall from Section 1.3 that the oracular aggregator is the conditional expectation of $\mathbf{1}_A$ given all the information used by the forecasters. Under the Gaussian model, this can be emulated with a hypothetical oracle forecaster whose information set is $B' := \bigcup_{i=1}^N B_i$. The oracular

aggregator is then nothing more than the probability forecast made by the oracle. That is,

$$p' = \mathbb{P}(A|\mathcal{F}') = \mathbb{P}(X_S > 0|X_{B'}) = \Phi\left(\frac{X_{B'}}{\sqrt{1-\delta'}}\right),$$

where $\delta' = |B'|$. This construction relies on the fact that A is conditionally independent of the collection $\{X_{B_i}\}_{i=1}^N$ given $X_{B'}$.

This benchmark aggregator provides a reference point that allows us to identify information structures under which other aggregation techniques perform relatively well. In particular, if an aggregator is likely to be near p' under a given Σ_{22} , then that information structure reflects favorable conditions for the aggregator. This idea is used in the following subsections to develop intuition about probability extremizing.

4.2 General Information Structure

A probability p is said to be *extremized* by another probability q if and only if q is closer to zero when $p \leq 1/2$ and closer to one when $p \geq 1/2$. This translates to the probit scores as follows: q extremizes p if and only if $\Phi^{-1}(q)$ is on the same side but further away from zero than $\Phi^{-1}(p)$. The amount of (multiplicative) extremization can then be quantified with the *probit extremization ratio* defined as $\alpha(q, p) := \Phi^{-1}(q)/\Phi^{-1}(p)$.

Given that no aggregator can improve upon the oracular aggregator, it provides an ideal reference point for analyzing extremization. This section specifically uses it to study extremizing of p_{probit} because a) it is arguably more reasonable than the simple average \bar{p} ; and b) it is very similar to p_{log} but results in cleaner analytic expressions. Therefore, of particular interest is the special case $\alpha(p', p_{\text{probit}}) = P' / \left(\frac{1}{N} \sum_{i=1}^N P_i\right)$, where $P' = \Phi^{-1}(p')$. From now on, unless otherwise stated, this expression is referred simply with α . Therefore, the probit opinion pool p_{probit} requires extremization if and only if $\alpha > 1$, and the larger α is, the more p_{probit} should be extremized.

Note that α is a random quantity that spans the entire real line; that is, it is possible to find a set of forecasts and an information structure for any possible value of $\alpha \in \mathbb{R}$. Evidently, extremizing is not guaranteed to always improve p_{probit} . To understand when extremizing is likely to be beneficial, the following proposition provides the probability distribution of α .

Proposition 4.1. *The law of the extremization ratio α is a Cauchy with parameters x_0 and γ , where the location parameter x_0 is at least one, equality occurring only when $\delta_i = \delta_j$ for all $i \neq j$. Consequently, if $\delta_i \neq \delta_j$ for some $i \neq j$, then the probability that p_{probit} requires extremizing $\mathbb{P}(\alpha > 1 | \Sigma_{22}, \delta')$ is strictly greater than $1/2$.*

This proposition shows that, on any non-trivial problem, a small perturbation in the direction of extremizing is more likely to improve p_{probit} than to degrade it. This partially explains why extremizing aggregators perform well on large sets of real-world prediction problems. It may be unsurprising after the fact, but the forecasting literature is still full of articles that perform probability averaging without extremizing. The next two subsections examine special cases in which more detailed computations can be performed.

4.3 Zero and Complete Information Overlap

If the forecasters use the same information, i.e., $B_i = B_j$ for all $i \neq j$, their forecasts are identical, $p' = p'' = p_{\text{probit}}$, and no extremization is needed. Therefore, given that the oracular aggregator varies smoothly over the space of information structures, averaging techniques, such as p_{probit} , can be expected to work well when the forecasts are based on very similar sources of information. This result is supported by the fact that the measurement error framework, which essentially describes the forecasters as making numerous small mistakes while applying the same procedure to the same data (see Section 2.2), results in averaging-based aggregators.

If, on the other hand, the forecasters have zero information overlap, i.e., $|B_i \cap B_j| = 0$ for

all $i \neq j$, the information structure Σ_{22} is diagonal and

$$p' = p'' = \Phi \left(\frac{\sum_{i=1}^N X_{B_i}}{\sqrt{1 - \sum_{i=1}^N \delta_i}} \right),$$

where the identities $\delta' = \sum_{i=1}^N \delta_i$ and $X_{B'} = \sum_{i=1}^N X_{B_i}$ result from the additive nature of the Gaussian process (see Section 3.3). This aggregator can be described in two steps: First, the numerator conducts voting, or range voting to be more specific, where the votes are weighted according to the importance of the forecasters' private information. Second, the denominator extremizes the consensus according to the total amount of information in the group. This clearly leads to very extreme forecasts. Therefore more extreme techniques can be expected to work well when the forecasters use widely different information sets.

The analysis suggests a spectrum of aggregators indexed by the information overlap: the optimal aggregator undergoes a smooth transformation from averaging (low extremization) to voting (high extremization) as the information overlap decreases from complete to zero overlap. This observation gives qualitative guidance in real-world settings where the general level of overlap can be said to be high or low. For instance, predictions from forecasters working in close collaboration can be averaged while predictions from forecasters strategically accessing and studying disjoint sources of information should be aggregated via more extreme techniques such as voting. See Parunak et al. 2013 for a discussion of voting-like techniques. For a concrete illustration, recall Example 3.1 where the optimal aggregate changes from $2/3$ (high information overlap) to $4/5$ (low information overlap).

4.4 Partial Information Overlap

To analyze the intermediate scenarios with partial information overlap among the forecasters, it is helpful to reduce the number of parameters in Σ_{22} . A natural approach is to assume

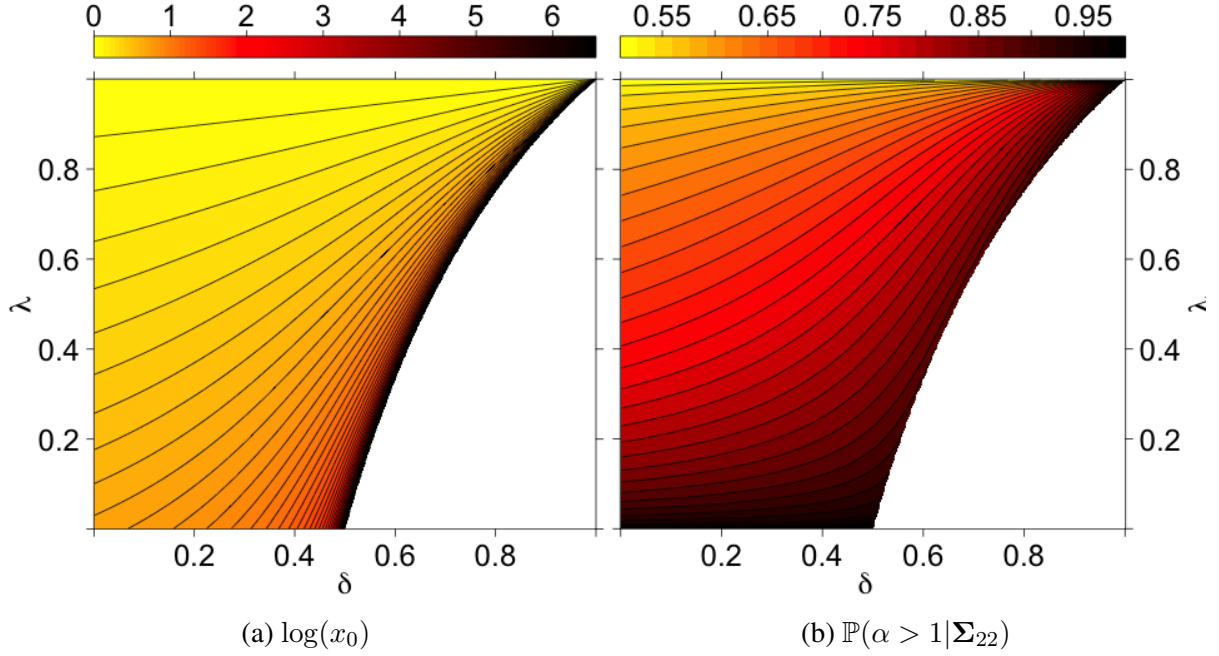


Figure 3: Extremization Ratio under Symmetric Information. The amount of extremizing α follows a Cauchy(x_0, γ), where x_0 is a location parameter and γ is a scale parameter. This figure considers $N = 2$ because in this case δ' is uniquely determined by Σ_{22} .

compound symmetry, where the information sets have the same size and that the amount of pairwise overlap is constant. More specifically, let $|B_i| = \delta$ and $|B_i \cap B_j| = \lambda\delta$, where δ is the amount of information used by each forecaster and λ is the overlapping proportion of this information. The resulting information structure is $\Sigma_{22} = \mathbf{I}_N(\delta - \lambda\delta) + \mathbf{J}_N\lambda\delta$, where \mathbf{I}_N is the identity matrix and \mathbf{J}_N is $N \times N$ matrix of ones. It is coherent if and only if

$$\delta \in [0, 1] \quad \text{and} \quad \lambda|\delta \in \left[\max \left\{ \frac{N - \delta^{-1}}{N - 1}, 0 \right\}, 1 \right]. \quad (4)$$

See Appendix A of the Supplementary Material for the derivation of these constraints.

Under these assumptions, the location parameter of the Cauchy distribution of α simplifies to $x_0 = N / (1 + (N - 1)\lambda) \sqrt{(1 - \delta) / (1 - \delta')}$. Of particular interest is to understand how this changes as a function of the model parameters. The analysis is somewhat hindered by the

unknown details of the dependence between δ' and the other parameters N , δ , and λ . However, given that δ' is defined as $\delta' = |\cup_{i=1}^N B_i|$, its value increases in N and δ but decreases in λ . In particular, as $\delta \rightarrow 1$, the value of δ' converges to one at least as fast as δ because $\delta' \geq \delta$. Therefore the term $\sqrt{(1-\delta)/(1-\delta')}$ and, consequently, x_0 increase in δ . Similarly, x_0 can be shown to increase in N but to decrease in λ . Therefore x_0 and δ' move together, and the amount of extremizing can be expected to increase in δ' . As the Cauchy distribution is symmetric around x_0 , the probability $\mathbb{P}(\alpha > 1 | \Sigma_{22})$ behaves similarly to x_0 and also increases in δ' . Figure 3 illustrates these relations by plotting both $\log(x_0)$ and $\mathbb{P}(\alpha > 1 | \Sigma_{22})$ for $N = 2$ forecasters under all plausible combinations of δ and λ . The white space collects all pairs (δ, λ) that do not satisfy (4) and hence represent incoherent information structures. Note that the results are completely general for the two-forecaster case, apart from the assumption $\delta_1 = \delta_2$. Relaxing this assumption does not change the qualitative nature of the results.

The total amount of information used by the forecasters δ' , however, does not provide a full explanation of extremizing. Information diversity is an important yet separate determinant. To see this, observe that fixing δ' to some constant defines a curve $\lambda = 2 - \delta'/\delta$ on the two plots in Figure 3. For instance, letting $\delta' = 1$ gives the boundary curve on the right side of each plot. This curve then shifts inwards and rotates slightly counterclockwise as δ' decreases. At the top end of each curve all forecasters use the total information, i.e., $\delta = \delta'$ and $\lambda = 1.0$. At the bottom end, on the other hand, the forecasters partition the total information and have zero overlap, i.e., $\delta = \delta'/2$ and $\lambda = 0.0$. Given that moving down along these curves simultaneously increases information diversity and x_0 , both information diversity and the total amount of information used by the forecasters are important yet separate determinants of extremizing. This observation can guide practitioners towards proper extremization because many application specific aspects are linked to these two determinants. For instance, extremization can be expected to increase in the number of forecasters, subject-matter expertise, and human diversity, but to decrease in collaboration, sharing of resources, and problem difficulty.

5. PROBABILITY AGGREGATION

5.1 Revealed Aggregator for the Gaussian Model

Recall the multivariate Gaussian distribution (2) and collect all $X_{B_i} = \Phi^{-1}(p_i)\sqrt{1 - \delta_i}$ into a column vector $\mathbf{X} = (X_{B_1}, X_{B_2}, \dots, X_{B_N})'$. If Σ_{22} is a coherent overlap structure and Σ_{22}^{-1} exists, then the revealed aggregator under the Gaussian model is

$$p'' = \mathbb{P}(A|\mathcal{F}'') = \mathbb{P}(X_S > 0|\mathbf{X}) = \Phi\left(\frac{\Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}}{\sqrt{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}}\right). \quad (5)$$

Applying (5) in practice requires an estimate of Σ_{22} . If the forecasters make predictions about multiple events, it may be possible to model the different prediction tasks with a hierarchical structure and estimate a fully general form of Σ_{22} . This can be formulated as a constrained (semi-definite) optimization problem, which, as was mentioned in Section 3.3, is left for subsequent work. Such estimation, however, requires the results of a large multi-prediction experiment which may not always be possible in practice. Often only a single prediction per forecaster is available. Consequently, accurate estimation of the fully general information structure becomes difficult. This motivates the development of aggregation techniques for a single event. Under the Gaussian model, a standard approach is to assume a covariance structure that involves fewer parameters. The next subsection discusses a natural and non-informative choice.

5.2 Symmetric Information

This subsection assumes a type of exchangeability among the forecasters. While this is somewhat idealized, it is a reasonable choice in a low-information environment where there is no historical or self-report data to distinguish the forecasters. The averaging aggregators described in Section 2, for instance, are symmetric. Therefore, to the extent that they reflect an under-

lying model, the model assumes exchangeability. Under the Gaussian model, exchangeability suggests the compound symmetric information structure discussed in Section 4.4. This structure holds if, for example, the forecasters use information sources sampled from a common distribution. The resulting revealed aggregator takes the form

$$p''_{cs} = \Phi \left(\frac{\frac{1}{(N-1)\lambda+1} \sum_{i=1}^N X_{B_i}}{\sqrt{1 - \frac{N\delta}{(N-1)\lambda+1}}} \right), \quad (6)$$

where $X_{B_i} = \Phi^{-1}(p_i)\sqrt{1-\delta}$ for all $i = 1, \dots, N$. Unfortunately, this version is not as good as the oracular aggregator; the former is in fact a conditional expectation of the latter.

Given these interpretations, it may at first seem surprising that the values of δ and λ can be estimated in practice. Intuitively, the estimation relies on two key aspects of the model: a) a better-informed forecast is likely to be further away from the non-informative prior (see Figure 2); and b) two forecasters with high information overlap are likely to report very similar predictions. This provides enough leverage to estimate the information structure via the maximum likelihood method. Complete details for this are provided in Appendix B of the Supplementary Material. Besides exchangeability, p''_{cs} is based on very different modeling assumptions than the averaging aggregators. The following proposition summarizes some of its key properties.

Proposition 5.1. (i) *The probit extremization ratio between p''_{cs} and p_{probit} is given by the non-random quantity $\alpha(p''_{cs}, p_{\text{probit}}) = \gamma\sqrt{1-\delta}/\sqrt{1-\delta\gamma}$, where $\gamma = N/((N-1)\lambda+1)$,*

(ii) *p''_{cs} extremizes p_{probit} as long as $p_i \neq p_j$ for some $i \neq j$, and*

(iii) *p''_{cs} can leave the convex hull of the individual probability forecasts.*

Proposition 5.1 suggests that p''_{cs} is appropriate for combining probability forecasts of a single event. This is illustrated on real-world forecasts in the next subsection. The goal is not to perform a thorough data analysis or model evaluation, but to demonstrate p''_{cs} on a simple example.

5.3 Real-World Forecasting Data

Probability aggregation appears in many facets of real-world applications, including weather forecasting, medical diagnosis, estimation of credit default, and sports betting. This section, however, focuses on predicting global events that are of particular interest to the Intelligence Advanced Research Projects Activity (IARPA). Since 2011, IARPA has posed about 100-150 question per year as a part of its ACE forecasting tournament. Among the participating teams, the Good Judgment Project (GJP) (Ungar et al. 2012; Mellers et al. 2014) has emerged as the clear winner. The GJP has recruited thousands of forecasters to estimate probabilities of the events specified by IARPA. The forecasters are told that their predictions are assessed using the Brier score (see Section 1.2). In addition to receiving \$150 for meeting minimum participation requirements that do not depend on prediction accuracy, the forecasters receive status rewards for good performance via leader-boards displaying Brier scores for the top 20 forecasters. Every year the top 1% of the forecasters are selected to the elite group of “super-forecasters”. Note that, depending on the details of the reward structure, such a competition for rank may eliminate the truth-revelation property of proper scoring rules (see, e.g., Lichtendahl Jr and Winkler 2007).

This subsection focuses on the super-forecasters in the second year of the tournament. Given that these forecasters were elected to the group of super-forecasters based on the first year, their forecasts are likely, but not guaranteed, to be relatively good. The group involves 44 super-forecasters collectively making predictions about 123 events, of which 23 occurred. For instance, some of the questions were: “Will France withdraw at least 500 troops from Mali before 10 April 2013?”, and “Will a banking union be approved in the EU council before 1 March 2013?”. Not every super-forecaster made predictions about every event. In fact, the number of forecasts per event ranged from 17 to 34 forecasts, with a mean of 24.2 forecasts. To avoid infinite log-odds and probit scores, extreme forecasts $p_i = 0$ and 1 were censored to $p_i = 0.001$ and 0.999, respectively.

In this section aggregation is performed one event at a time without assuming any other information besides the probability forecasts themselves. This way any performance improvements reflect better fit of the underlying model and the aggregator’s relative advantage in forecasting a single event. Aggregation accuracy is measured with the mean Brier score (BS): Consider K events and collect all N_k probability forecasts for event A_k into a vector $\mathbf{p}_k \in [0, 1]^{N_k}$. Then, BS for aggregator $g : [0, 1]^{N_k} \rightarrow [0, 1]$ is

$$\text{BS} = \frac{1}{K} \sum_{k=1}^K (g(\mathbf{p}_k) - \mathbf{1}_{A_k})^2.$$

This score is defined on the unit interval with lower values indicating higher accuracy. For a more detailed performance analysis, it decomposes into three additive components: reliability (REL), resolution (RES), and uncertainty (UNC). This assumes that the aggregate forecast $g(\mathbf{p}_k)$ for all k can only take discrete values $f_j \in [0, 1]$ with $j = 1, \dots, J$. Let n_j be the number of times f_j occurs, and denote the empirical frequency of the corresponding events with o_j . Let \bar{o} be the overall empirical frequency of occurrence, i.e., $\bar{o} = \frac{1}{K} \sum_{k=1}^K \mathbf{1}_{A_k}$. Then,

$$\begin{aligned} \text{BS} &= \text{REL} - \text{RES} + \text{UNC} \\ &= \frac{1}{K} \sum_{j=1}^J n_j (f_j - o_j)^2 - \frac{1}{K} \sum_{j=1}^J n_j (o_j - \bar{o})^2 + \bar{o}(1 - \bar{o}). \end{aligned}$$

Confident aggregators exhibit high RES. The corresponding forecasts are likely to be very close to 0 and 1, which is more useful to the decision-maker than the naive forecast \bar{o} as long as the forecasts are also accurate, i.e., calibrated. In the decomposition, good calibration presents itself as low REL. Therefore both improved calibration and higher confidence yield a lower BS. Any such improvements should be interpreted relative to UNC that equals the BS for \bar{o} , i.e., the best aggregate that does not use the forecasters’ predictions.

Table 1 presents results for \bar{p} , p_{\log} , p_{probit} , and p''_{cs} under the super-forecaster data. Empiri-

Table 1: The Mean Brier Scores (BS) with Its Three Components, Reliability (REL), Resolution (RES), and Uncertainty (UNC), for Different Aggregators.

Aggregator	BS	REL	RES	UNC
\bar{p}	0.132	0.026	0.045	0.152
p_{\log}	0.128	0.025	0.048	0.152
p_{probit}	0.128	0.023	0.047	0.152
p''_{cs}	0.123	0.020	0.049	0.152

cal approaches were not considered for two reasons: a) they do not reflect an actual model of forecasts; and b) they require a training set with known outcomes and hence cannot be applied to a single event. Overall, \bar{p} presents the worst performance. Given that p_{probit} and p_{\log} are very similar, it is not surprising that they have almost identical scores. The revealed aggregator p''_{cs} is both the most resolved and calibrated, thus achieving the lowest BS among all the aggregators. This is certainly an encouraging result. It is important to note that p''_{cs} is only the first attempt at partial information aggregation. More elaborate information structures and estimation procedures, such as shrinkage estimators, are very likely to lead to many further improvements.

6. SUMMARY AND DISCUSSION

This paper introduced a probability model for predictions made by a group of forecasters. The model allows for interpretation of some of the existing work on forecast aggregation and also clarifies empirical approaches such as the *ad hoc* practice of extremization. The general model is more plausible on the micro-level than any other model has been to date. Under this model, some general results were provided. For instance, the *oracular* aggregator is more likely to give a forecast that is more extreme than one of the common benchmark aggregates, namely p_{probit} (Proposition 4.1). Even though no real world aggregator has access to all the information of the oracle, this result explains why extremization is almost certainly called for.

More detailed analyses were performed under several specific model specifications such as zero and complete information overlap (Section 4.3), and fully symmetric information (Section 4.4). Even though the zero and complete information overlap models are not realistic, except under a very narrow set of circumstances, they form logical extremes that illustrate the main drivers of good aggregation. The symmetric model is somewhat more realistic. It depends only on two parameters and therefore allows us to visualize the effect of model parameters on the optimal amount of extremization (Figure 3). Finally, the *revealed* aggregator, which is the best in-practice aggregation under the partial information model, was discussed. The discussion provided a general formula for this aggregator (Equation 5) as well as its specific formula under symmetric information (Equation 6). The specific form was applied to real-world forecasts of one-time events and shown to outperform other model-based aggregators.

It is interesting to relate our discussion to the many empirical studies conducted by the Good Judgment Project (GJP) (see Section 5.3). Generally extremizing has been found to improve the average aggregates (Mellers et al., 2014; Satopää et al., 2014a,b). The average forecast of a team of super-forecasters, however, often requires very little or no extremizing. This can be explained as follows. The super-forecasters are highly knowledgeable (high δ) individuals who work in groups (high ρ and λ). Therefore, in Figure 3 they are situated around the upper-right corners where almost no extremizing is required. In other words, there is very little left-over information that is not already used in each forecast. Their forecasts are highly convergent and are likely to be already very near the oracular forecast. The GJP forecast data also includes self-assessments of expertise. Not surprisingly, the greater the self-assessed expertise, the less extremizing appears to have been required. This is consistent with our interpretation that high values of δ and λ suggest lower extremization.

The partial information framework offers many directions for future research. One involves estimation of parameters. In principle, $|B_i|$ can be estimated from the distribution of a reasonably long probability stream. Similarly, $|B_i \cap B_j|$ can be estimated from the correlation of the

two parallel streams. Estimation of higher order intersections, however, seems more dubious. In some cases the higher order intersections have been found to be irrelevant to the aggregation procedure. For instance, DeGroot and Mortera (1991) show that it is enough to consider only the pairwise conditional (on the truth) distributions of the forecasts when computing the optimal weights for a linear opinion pool. Theoretical results on the significance or insignificance of higher order intersections under the partial information framework would be desirable.

Another promising avenue is the Bayesian approach. In many applications with small or moderately sized datasets, Bayesian methods have been found to be superior to the likelihood-based alternatives. Therefore, given that the number of forecasts on a single event is typically quite small, a Bayesian approach is likely to improve the predictions of one-time events. Currently, we have work in progress analyzing a Bayesian model but there are many, many reasonable priors on the information structures. This avenue should certainly be pursued further, and the results tested against other high performing aggregators.

APPENDIX A: PROOFS AND DERIVATIONS

A.1 Proof of Proposition 3.3.

Denote the set of all coherent information structures with \mathcal{Q}_N . Consider $\Sigma_{22} \in \mathcal{Q}_N$ and its associated Borel sets $\{B_i : i = 1, \dots, N\}$. Given that Σ_{22} is coherent, its information can be represented in a diagram such as the one given by Figure 1 in the main manuscript. Keeping the diagram representation in mind, partition the unit interval S into 2^N disjoint parts $C_{\mathbf{v}} := \cap_{i \in \mathbf{v}} B_i \setminus \cup_{i \notin \mathbf{v}} B_i$, where $\mathbf{v} \subseteq \{1, \dots, N\}$ denotes a subset of forecasters and each $C_{\mathbf{v}}$ represents information used only by the forecasters in \mathbf{v} . Given that $\sum_{\mathbf{v}} |C_{\mathbf{v}}| = 1$, it is possible

to establish a linear function L from the probability simplex

$$\begin{aligned}\Delta_N &:= \text{conv}\{\mathbf{e}_{\mathbf{v}} : \mathbf{v} \subseteq \{1, \dots, N\}\} \\ &= \left\{ \mathbf{z} \in \mathbb{R}^{2^N} : \mathbf{z} \geq \mathbf{0}, \mathbf{1}'\mathbf{z} = 1 \right\}\end{aligned}$$

to the space of coherent information structures \mathcal{Q}_N . In particular, the linear function $L : \mathbf{z} \in \Delta_N \rightarrow \Sigma_{22} \in \mathcal{Q}_N$ is defined such that $\rho_{ij} = \sum_{\{i,j\} \subseteq \mathbf{v}} z_{\mathbf{v}}$ and $\delta_i = \sum_{i \in \mathbf{v}} z_{\mathbf{v}}$. Therefore $L(\Delta_N) = \mathcal{Q}_N$. Furthermore, given that Δ_N is a convex polytope,

$$\begin{aligned}L(\Delta_N) &= \text{conv}\{L(\mathbf{e}_{\mathbf{v}}) : \mathbf{v} \subseteq \{1, \dots, N\}\} \\ &= \text{conv}\{\mathbf{x}\mathbf{x}' : \mathbf{x} \in \{0, 1\}^N\} \\ &= \text{COR}(N),\end{aligned}\tag{6}$$

which establishes $\text{COR}(N) = \mathcal{Q}_N$. Equality (6) follows from the basic properties of convex polytopes (see, e.g., McMullen and Shephard 1971, pp. 16). Each $\Sigma_{22} \in \text{COR}(N)$ has $\frac{N(N+1)}{2} = \binom{n+1}{2}$ parameters and therefore exists in $\binom{n+1}{2}$ dimensions. \square

A.2 Proof of Proposition 4.1.

Given that

$$\begin{aligned}P' &\sim \mathcal{N}\left(0, \sigma_1^2 := \frac{\delta'}{1 - \delta'}\right) \\ \frac{1}{N} \sum_{i=1}^N P_i &\sim \mathcal{N}\left(0, \sigma_2^2 := \frac{1}{N^2} \left\{ \sum_{i=1}^N \frac{\delta_i}{1 - \delta_i} + 2 \sum_{i,j:i < j} \frac{\rho_{ij}}{\sqrt{(1 - \delta_j)(1 - \delta_i)}} \right\}\right),\end{aligned}$$

the amount of extremizing α is a ratio of two correlated Gaussian random variables. The Pearson product-moment correlation coefficient for them is

$$\kappa = \frac{\sum_{i=1}^N \frac{\delta_i}{\sqrt{1-\delta_i}}}{\sqrt{\delta' \left\{ \sum_{i=1}^N \frac{\delta_i}{1-\delta_i} + 2 \sum_{i,j:i < j} \frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}} \right\}}}$$

It follows that α has a Cauchy distribution as long as $\sigma_1 \neq 1$, $\sigma_2 \neq 1$, or $\kappa \pm 1$ (see, e.g., Cedilnik et al. 2004). These conditions are very mild under the Gaussian model. For instance, if no forecaster knows as much as the oracle, the conditions are satisfied. Consequently, the probability density function of α is

$$f(\alpha|x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(\alpha - x_0)^2 + \gamma^2},$$

where $x_0 = \kappa\sigma_1/\sigma_2$ and $\gamma = \sqrt{1 - \kappa^2}\sigma_1/\sigma_2$. The parameter x_0 represents the location (the median and mode) and γ specifies the scale (half the interquartile range) of the Cauchy distribution. The location parameter simplifies to

$$x_0 = \kappa \frac{\sigma_1}{\sigma_2} = \frac{N \sum_{i=1}^N \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}}}{\sum_{i=1}^N \frac{\delta_i}{1-\delta_i} + 2 \sum_{i,j:i < j} \frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}}$$

Given that all the remaining terms are positive, the location parameter x_0 is also positive. Compare the N terms with a given subindex i in the numerator with the corresponding terms in the denominator. From $\delta' \geq \delta_i \geq \rho_{ij}$, it follows that

$$\frac{\delta_i}{1-\delta_i} = \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta_i)}} \leq \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \quad (7)$$

$$\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}} \leq \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \quad (8)$$

Therefore

$$N \sum_{i=1}^N \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \geq \sum_{i=1}^N \frac{\delta_i}{1-\delta_i} + 2 \sum_{i,j:i < j} \frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}},$$

which gives that $x_0 \geq 1$. Given that the Cauchy distribution is symmetric around x_0 , it must be the case that $\mathbb{P}(\alpha > 1 | \Sigma_{22}, \delta') \geq 1/2$. Based on (7) and (8), the location $x_0 = 1$ only when all the forecasters know the same information, i.e., when $\delta_i = \delta_j$ for all $i \neq j$. Under this particular setting, the amount of extremizing α is non-random and always equal to one. Any deviation from this particular information structure makes α random, $x_0 > 1$, and hence $\mathbb{P}(\alpha > 1 | \Sigma_{22}, \delta') > 1/2$. \square

A.3 Proof of Proposition 5.1.

(i) This follows from direct computation:

$$\begin{aligned} \alpha &= \left(\frac{\frac{1}{(N-1)\lambda+1} \sum_{i=1}^N X_{B_i}}{\sqrt{1 - \frac{N\delta}{(N-1)\lambda+1}}} \right) / \left(\frac{1}{N} \sum_{i=1}^N \frac{X_{B_i}}{\sqrt{1-\delta}} \right) \\ &= \frac{\frac{N\sqrt{1-\delta}}{(N-1)\lambda+1}}{\sqrt{1 - \frac{N\delta}{(N-1)\lambda+1}}}, \end{aligned} \quad (9)$$

which simplifies to the given expression after substituting in γ . Given that this quantity does not depend on any X_{B_i} , it is non-random.

(ii) For a given δ , the amount of extremizing α is minimized when $(N-1)\lambda+1$ is maximized. This happens as $\lambda \uparrow 1$. Plugging this into (9) gives

$$\alpha = \frac{\frac{N\sqrt{1-\delta}}{(N-1)\lambda+1}}{\sqrt{1 - \frac{N\delta}{(N-1)\lambda+1}}} \downarrow \frac{\sqrt{1-\delta}}{\sqrt{1-\delta}} = 1$$

(iii) Assume without loss of generality that $\bar{P} > 0$. If $\max\{p_1, p_2, \dots, p_N\} < 1$, then setting $\delta = 1/N$ and $\lambda = 0$ gives an aggregate probability $p'' = 1$ that is outside the convex hull of the individual probabilities. \square

A.4 Derivation of Equation 3

Clearly, any $\delta \in [0, 1]$ is plausible. Conditional on such δ , however, the overlap parameter λ must be within a subinterval of $[0, 1]$. The upper bound of this subinterval is always one because the forecasters may use the same information under any δ and N . To derive the lower bound, note that information overlap is unavoidable when $\delta > 1/N$, and that minimum overlap occurs when all information is used either by everyone or by a single forecaster. In other words, if $\delta > 1/N$ and $B_i \cap B_j = B$ with $|B| = \lambda\delta$ for all $i \neq j$, the value of λ is minimized when $\lambda\delta + N(\delta - \delta\lambda) = 1$. Therefore the lower bound for λ is $\max\{(N - \delta^{-1})/(N - 1), 0\}$, and Σ_{22} is coherent if and only if $\delta \in [0, 1]$ and $\lambda|\delta \in [\max\{(N - \delta^{-1})/(N - 1), 0\}, 1]$.

APPENDIX B: PARAMETER ESTIMATION UNDER SYMMETRIC INFORMATION

This section describes how the maximum likelihood estimates of δ and λ can be found accurately and efficiently. Denote a $N \times N$ matrix of ones with \mathbf{J}_N . A matrix Σ is called compound symmetric if it can be expressed in the form $\Sigma = \mathbf{I}_N A + \mathbf{J}_N B$ for some constants A and B . The inverse matrix (if it exists) and any scalar multiple of a compound symmetric matrix Σ are also compound symmetric (Dobbin and Simon, 2005). More specifically, for some constant c ,

$$\begin{aligned} c\Sigma &= \mathbf{I}_N(cA) + \mathbf{J}_N(cB) \\ \Sigma^{-1} &= \mathbf{I}_N \frac{1}{A} - \mathbf{J}_N \frac{B}{A(A + NB)} \end{aligned} \quad (10)$$

Define

$$\begin{aligned}
\boldsymbol{\Sigma}_{22} &:= \text{Cov}(\mathbf{X}) = \mathbf{I}_N A_X + \mathbf{J}_N B_X \\
\boldsymbol{\Sigma}_P &:= \text{Cov}(\mathbf{P}) = \boldsymbol{\Sigma}_{22}/(1 - \delta) = \mathbf{I}_N A_P + \mathbf{J}_N B_P \\
\boldsymbol{\Omega} &:= \boldsymbol{\Sigma}_P^{-1} = \mathbf{I}_N A_\Omega + \mathbf{J}_N B_\Omega
\end{aligned} \tag{11}$$

To set up the optimization problem, observe that the Jacobian for the map $\mathbf{P} \rightarrow \Phi(\mathbf{P}) = (\Phi(P_1), \Phi(P_2), \dots, \Phi(P_N))'$ is $J(\mathbf{P}) = (2\pi)^{-N/2} \exp(-\mathbf{P}'\mathbf{P}/2)$. If $h(\mathbf{P})$ denotes the multivariate Gaussian density of $\mathbf{P} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Sigma}_P)$, the density for $\mathbf{p} = (p_1, p_2, \dots, p_N)'$ is

$$f(\mathbf{p}|\delta, \lambda) = h(\mathbf{P})J(\mathbf{P})^{-1} \propto |\boldsymbol{\Sigma}_P|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{P}'\boldsymbol{\Sigma}_P^{-1}\mathbf{P}\right],$$

where $\mathbf{P} = \Phi^{-1}(\mathbf{p})$. Let $\mathbf{S}_P = \mathbf{P}\mathbf{P}'$ be the (rank one) sample covariance matrix of \mathbf{P} . The log-likelihood then reduces to

$$\log f(\mathbf{p}|\delta, \lambda) \propto -\log \det \boldsymbol{\Sigma}_P - \text{tr}(\mathbf{S}_P^{-1}\boldsymbol{\Sigma}_P)$$

This log-likelihood is not concave in $\boldsymbol{\Sigma}_P$. It is, however, a concave function of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_P^{-1}$. Making this change of variables gives us the following optimization problem:

$$\begin{aligned}
&\text{minimize} && -\log \det \boldsymbol{\Omega} + \text{tr}(\mathbf{S}_P \boldsymbol{\Omega}) \\
&\text{subject to} && \delta \in [0, 1] \\
&&& \lambda \in \left[\max\left\{\frac{N - \delta^{-1}}{N - 1}, 0\right\}, 1\right),
\end{aligned} \tag{12}$$

where the open upper bound on λ ensures a non-singular information structure $\boldsymbol{\Sigma}_{22}$. Unfortunately, the feasible region is not convex (see, e.g., Figure 3 in the main manuscript) but can be made convex by re-expressing the problem as follows: First, let $\rho = \delta\lambda$ denote the amount

of information known by a forecaster; that is, let $A_X = (\delta - \rho)$ and $B_X = \rho$. Solving the problem in terms of δ and ρ is equivalent to minimizing the original objective (12) but subject to $0 \leq \rho \leq \delta$ and $0 \leq \rho(N - 1) - N\delta + 1$. Given that this region is an intersection of four half-spaces, it is convex. Furthermore, it can be translated into the corresponding feasible and convex set of (A_Ω, B_Ω) via the following steps:

$$\begin{aligned}
& \Sigma_{22} \in \{ \Sigma_{22} : 0 \leq \rho \leq \delta, 0 \leq \rho(N - 1) - N\delta + 1 \} \\
\Leftrightarrow & \Sigma_{22} \in \{ \Sigma_{22} : 0 \leq B_X, 0 \leq A_X, 0 \leq 1 - B_X + NA_X, \} \\
\Leftrightarrow & \Sigma_P \in \{ \Sigma_P : 0 \leq A_P \leq 1/(N - 1), 0 \leq B_P \} \\
\Leftrightarrow & \Omega \in \{ \Omega : 0 \leq A_\Omega - N + 1, 0 \leq A_\Omega + B_\Omega N, 0 \leq -B_\Omega \}
\end{aligned}$$

According to Rao (2009), $\log \det(\Omega) = N \log A_\Omega + \log(1 + NB_\Omega/A_\Omega)$. Plugging this and the feasible region of (A_Ω, B_Ω) into the original problem (12) gives an equivalent but convex optimization problem:

$$\begin{aligned}
& \text{minimize} \quad -N \log A_\Omega - \log \left(1 + \frac{NB_\Omega}{A_\Omega} \right) + A_\Omega \text{tr}(\mathbf{S}_P) + B_\Omega \text{tr}(\mathbf{S}_P \mathbf{J}_N) \\
& \text{subject to} \quad 0 \leq A_\Omega - N + 1 \\
& \quad \quad \quad 0 \leq A_\Omega + B_\Omega N \\
& \quad \quad \quad 0 \leq -B_\Omega
\end{aligned}$$

The first term of this objective is both convex and non-decreasing. The second term is a composition of the same convex, non-decreasing function with a function that is concave over the feasible region. Such a composition is always convex. The last two terms are affine and hence also convex. Therefore, given that the objective is a sum of four convex functions, it is convex, and globally optimal values of (A_Ω, B_Ω) can be found very efficiently with interior point algorithms such as the barrier method. There are many open software packages that implement

generic versions of these methods. For instance, our implementation uses the standard R function `constrOptim` to solve the optimization problem. Denote optimal values with $(A_{\Omega}^*, B_{\Omega}^*)$. They can be traced back to (δ, λ) via (10) and (11). The final map simplifies to

$$\delta^* = \frac{B_{\Omega}^*(N-1) + A_{\Omega}^*}{A_{\Omega}^*(1 + A_{\Omega}^*) + B_{\Omega}^*(N-1 + NA_{\Omega}^*)} \quad \text{and} \quad \lambda^* = -\frac{B_{\Omega}^*}{B_{\Omega}^*(N-1) + A_{\Omega}^*}$$

REFERENCES

- Armstrong, J. S. (2001). Combining forecasts. In Armstrong, J. S., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 417–439. Kluwer Academic Publishers, Norwell, MA.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145.
- Broomell, S. B. and Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74(3):531–553.
- Cedilnik, A., Kosmelj, K., and Blejec, A. (2004). The distribution of the ratio of jointly normal variables. *Metodoloski Zvezki*, 1(1):99–108.
- DeGroot, M. H. and Mortera, J. (1991). Optimal linear opinion pools. *Management Science*, 37(5):546–558.
- Di Bacco, M., Frederic, P., and Lad, F. (2003). Learning from the probability assertions of experts. Research Report. Available at: <http://www.math.canterbury.ac.nz/research/ucdms2003n6.pdf>.
- Dobbin, K. and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38.

- Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3):519–527.
- Hong, L. and Page, S. (2009). Interpreted and generated signals. *Journal of Economic Theory*, 144(5):2174–2196.
- Hwang, J. and Pemantle, R. (1997). Estimating the truth of an indicator function of a statistical hypothesis under a class of proper loss functions. *Statistics & Decisions*, 15:103–128.
- Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, 21(1):61–72.
- Lai, T. L., Gross, S. T., Shen, D. B., et al. (2011). Evaluating probability forecasts. *The Annals of Statistics*, 39(5):2356–2382.
- Laurent, M., Poljak, S., and Rendl, F. (1997). Connections between semidefinite relaxations of the max-cut and stable set problems. *Mathematical Programming*, 77(1):225–246.
- Lichtendahl Jr, K. C. and Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755.
- McMullen, P. and Shephard, G. C. (1971). *Convex Polytopes and the Upper Bound Conjecture*, volume 3. Cambridge University Press, Cambridge, U.K.
- Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115.
- Parunak, H. V. D., Brueckner, S. A., Hong, L., Page, S. E., and Rohwer, R. (2013). Characterizing and aggregating agent estimates. In *Proceedings of the 2013 International Conference*

- on Autonomous Agents and Multi-agent Systems*, pages 1021–1028, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.
- Rao, C. R. (2009). *Linear Statistical Inference and Its Applications*, volume 22 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, New York.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014a). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.
- Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., Ungar, L. H., et al. (2014b). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics*, 8(2):1256–1280.
- Shlomi, Y. and Wallsten, T. S. (2010). Subjective recalibration of advisors’ probability estimates. *Psychonomic Bulletin & Review*, 17(4):492–498.
- Ungar, L., Mellers, B., Satopää, V., Tetlock, P., and Baron, J. (2012). The good judgment project: A large scale test of different methods of combining expert predictions. The Association for the Advancement of Artificial Intelligence Technical Report FS-12-06.
- Ziegler, G. M. (2000). Lectures on 0/1-polytopes. In Kalai, G. and Ziegler, G. M., editors, *Polytopes - Combinatorics and Computation*, volume 29, pages 1–41, Basel. Springer, Birkhäuser.