Information Theoretic Model for Subjective Response Data with an Application to Probability Extremization

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

Department of Statistics
The Wharton School of the University of Pennsylvania
Philadelphia, PA 19104- 6340, USA

Abstract

Randomness in scientific estimation is generally assumed to arise from unmeasured or uncontrolled factors. However, when combining subjective probability forecasts, heterogeneity stemming from people's cognitive or information diversity is often more important than measurement noise. This paper presents an information theoretic framework that models the heterogeneity arising from cognitively diverse experts, and applies the model to the task of aggregating probabilities given by a group of experts who forecast whether an event will occur or not. Our model describes the distribution of information across experts 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 experts' information overlap. Our model thus gives a more principled understanding of the historically *ad hoc* practice of extremizing average forecasts.

1 Introduction

Consulting experts are often asked to make estimates under incomplete information. If the experts form their estimates independently of each other, their estimates are likely to be different. To analyze the estimates with statistical methodology, it is mathematically convenient and often necessary to assume that these differences arise from a probability distribution. For instance, consider a group of experts who aim to eyeball the height of a building. Their estimates can be modeled as independent draws from a Gaussian distribution that is centered at the true height. This framework is called the *generated signal framework*. Even though distributional assumptions clearly oversimplify the reality, they are typically useful enough to improve our understanding of the world.

Unfortunately, this is unlikely to be the case with subjective response data. Hong and Page [2009] explain how standard distributional assumptions can be dramatically misaligned with the process that generates subjective responses. As an alternative to the generated signal framwork,

they introduce the *interpreted signal framework*. An estimate is said to be interpreted if the expert first filters reality into set of categories and then makes an estimate by applying active cognitive effort to these categories. For instance, consider an Olympic judge who is evaluating a figure skating performance. The judge first interprets the performance with the aid of categories, and then scrutinizes that experience into a final score. Therefore any differences between estimates is assumed to arise from cognitive diversity instead of an underlying probability distribution.

The interpreted signal framework is clearly a more realistic modeling choice for subjective response data. Given that this kind of data is very common in real-world applications, such as product reviews, online auctions, and voting, this framework shows great potential in improving our understanding of many experimental results. Unfortunately, the interpreted signal framework is more of an abstract concept than a concrete model for subjective response data. Therefore it is not clear how this framework can be applied in complex real world situations.

The first contribution of this paper is to introduce a concrete model that generates heterogeneity from cognitive diversity and can be used in a real-world situations. Our model is considered *information theoretic* as it is based on the distribution of information among the experts. This distribution is completely characterized by the amount of information known by each expert and the amount information shared between any two experts. For instance, experts A and B may know 10% and 5% of the full information, respectively. If they share 2.5% of the full information, then their information sets overlap and their estimates are assumed to be positively correlated. The experts are assumed to give optimal estimates conditional on their private information sets. This means that a larger information set typically leads to a more accurate estimate, and two identical information sets lead to the same estimate. In reality experts, however, do not typically make optimal use of their information. Therefore the information theoretic framework is a simplification of the real world. It is, however, a compromise that strikes a good balance between psychological realism and analytical convenience. Therefore it offers a platform that is particularly convenient for development of future methodology.

The second contribution of this paper is to apply the information theoretic framework on probability aggregation. Combining multiple probability forecasts is an important problem with many applications including medical diagnosis (Pepe [2003], Wilson et al. [1998]), political and socioeconomic foresight (Tetlock [2005]), and meteorology (Baars and Mass [2005], Sanders [1963], Vislocky and Fritsch [1995]). There is strong empirical evidence that bringing together the strengths of different experts by combining their probability forecasts into a single consensus, known as the *crowd belief*, improves predictive performance. For instance, consider the aforementioned experts A and B. The union of their information sets covers 12.5% of the full information. Therefore it seems plausible that some combination of their probability forecasts is more informed than either one of the individual probabilities. The naive approach is to simply average the individual forecasts. To see why this approach can be problematic, recall that A's forecast is based on a larger

information set and hence typically more accurate than B's forecast. Therefore A's forecast is on average closer to the actual outcome of the event (0 if it does not happen and 1 if it does happen) and should be given a higher weight in the final aggregate. The average forecast, however, gives each forecast equal weight and hence ends up being necessarily too close to 0.5. Recent developments suggest that shifting the average probability closer to its nearest extreme (0.0 or 1.0), known as *extremizing*, yields improved forecasting performance. For instance, Satopää et al. [2014] use a linear regression model in the logit-space to derive an extremizing aggregator that performs well on real-world data. Ranjan and Gneiting [2010] propose transforming the average probability with the cumulative distribution function (CDF) of a beta distribution. If both the shape and scale of this beta distribution are equal and constrained to be at least 1.0, the aggregator extremizes and has some attractive theoretical properties (Wallsten and Diederich [2001]). Baron et al. [2013] provide yet another extremizing aggregator in addition to two intuitive justifications for extremizing.

These aggregators, however, are based on ad hoc techniques that learn the amount of extremization by optimizing a scoring rule over a separate training set (Gneiting and Raftery [2007]). It is concerning that extremization does not arise naturally from the underlying model. These aggregators are also too detached from the psychology literature to provide any insight beyond the aggregate probability. Therefore it is still not well-understood when and how much the average probability should be extremized. This paper remedies these shortcomings by developing an aggregator that is based on the information theoretic framework. Under this model the average forecast is always extremized. The amount of extremization is given in a closed-form expression that can be applied to any number of experts with any given information structure. By assuming a simplified information structure, the amount of extremization can be made to depend only on three intuitive parameters. This allows us to visualize extremization and make concrete statements on when and how much extremization should be performed.

This paper is structured as follows. The first section introduces our information theoretic framework and compares it with the generated and interpreted signal frameworks. The second section applies the framework to probability forecasts. The third section derives a closed-form expression for the amount of extremization and analyzes this expression under unstructured and compound symmetric information structures. The paper concludes with a discussion of model limitations and future directions.

2 Information Theoretic Framework

This section discusses the information theoretic framework in comparison to the generated and interpreted signal frameworks. This comparison is by no means comprehensive as signal generation is a large part of the statistical literature. The first subsection builds intuition via a simple example. The second subsection provides a technical comparison.

2.1 Simple Example

Consider two experts 1 and 2 who are observing a hockey tournament. The tournament consists of three games played between teams RED and BLUE. After seeing the outcome of the first game, the experts are asked to report the probability of RED winning the tournament. Assume that RED wins if $G_1 + G_2 + G_3 \geq 0$, where $G_k \in \{-1, 1\}$ indicates whether RED won the kth game. Suppose that all 8 possible combinations are equally likely.

Generated Estimate: Based on the first game, the ith expert believes that RED has an independent chance of q_i winning any of the two remaining games. The probability q_i is assumed to arise from a probability distribution defined on the unit interval. Therefore any individual differences in the way the experts process the first game and turn the acquired information into probabilities q_1 and q_2 are assumed to stem from a probability distribution. The exports report

$$p_i = \begin{cases} q_i(2 - q_i) & \text{if } G_1 = 1 \\ q_i^2 & \text{if } G_1 = -1 \end{cases} \tag{1}$$

Interpreted Estimate: Interpretations are different ways of seeing the first game. Assume that RED's performance can be attributed entirely to its *defense* D and *offense* O that are equally likely to be either good 1 or bad -1. Expert 1 follows only the defense and expert 2 looks only at the offensive play. Based on these attributes the experts construct their final predictive models. Under the generated signal framework the details of these predictive models are abstracted into a probability distribution. Under the interpreted signal framework, however, the details are fixed and known. For instance, the experts may report (1) but with

$$q_1 = \begin{cases} 2/3 & \text{if } D = 1 \\ 1/3 & \text{if } D = -1 \end{cases} \qquad q_2 = \begin{cases} 3/5 & \text{if } O = 1 \\ 2/5 & \text{if } O = -1 \end{cases}$$

Each expert interprets the available information independently and subjectively. Therefore even if the experts observed the same attributes of the game, their probability forecasts do not need to be the same.

Information Theoretic Estimate: The experts know that RED has a 1/2 chance of winning any given game. Suppose that expert 1 knows the number of wins in the first two games, i.e. the value of $G_1 + G_2$, but not necessarily the separate outcomes of the two games. Assume that expert 2 only knows the outcome of the first game G_1. Then,

$$p_1 = \begin{cases} 0 & \text{if } G_1 + G_2 = -2 \\ 1/2 & \text{if } G_1 + G_2 = 0 \\ 1 & \text{if } G_1 + G_2 = 2 \end{cases} \qquad p_2 = \begin{cases} 1/4 & \text{if } G_1 = -1 \\ 3/4 & \text{if } G_1 = 1 \end{cases}$$

Experts 1 and 2 know 2/3 and 1/3 of the full information, respectively. Their predictive models are completely determined by the size of their private information sets. As their information sets overlap by 1/3 of the full information, their probability forecasts are positively correlated. In this example, the correlation coefficient for their forecasts is $\sqrt{2}/2$.

2.2 Technical Details

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where the set Ω contains all possible states of the world, the σ-field $\mathcal{F}$ consists of all subsets of Ω, and $\mathbb{P}$ is a probability measure. Let $A \in \mathcal{F}$ denote an event of interest. The experts aim to forecast the probability $p = \mathbb{P}(A)$. Even though this paper focuses on probability estimates, the following discussion can be easily generalized to different types of estimates.

Generated estimates can be considered as noisy or distorted versions of p. In other words, if $\xi : [0,1] \to [0,1]$ is a noise function that randomly distorts a probability, then a generated probability forecasts is realized by $p_i = \xi(p)$. Due to its mathematical convenience, this framework is typically applied to subjective response data. Unfortunately, it can be drastically misaligned with the psychology literature and hence lead to results that are not reflective of the actual process that generates the estimates.

Under the interpreted signal framework, the set of states Ω is assume to be finite. The expert partitions Ω into non-overlapping subsets. This partition, know as the *interpretation*, is denoted with $\Pi^i = \{\pi_1^i, \pi_2^i, \ldots, \pi_{n_i}^i\}$, where $\bigcup_{j=1}^{n_i} \pi_j^i = \Omega$ and $\pi_j^i \cap \pi_k^i = \emptyset$ for $j \neq k$. The expert can only associate a state $\omega \in \Omega$ with a set in his partition. Therefore his information is incomplete as long as not all the sets of his partition are singletons. To make probability forecasts, the expert specifies a map $\phi_i : \Omega \to [0,1]$ that is measurable with respect to Π^i. Unfortunately, Hong and Page [2009] do not specify how the expert constructs the map ϕ_i. It is also not clear how to the set of states Ω should be specified in real-world applications.

The information theoretic framework removes these intractable components and provides a model that can be applied in practice. It does not assume detailed knowledge of the expert interpretations nor pose any structure or cardinality restrictions on Ω. Each expert simply forecasts $p_i = \mathbb{P}(A|\mathcal{F}_i)$ based on a private information set $\mathcal{F}_i \subseteq \mathcal{F}$. If two experts i and j share information such that $\mathcal{F}_i \cap \mathcal{F}_j \neq \emptyset$, then the correlation of their forecasts p_i and p_j is positive and proportional to the overlap in their information sets. This means that, similarly to the interpreted signal framework, any heterogeneity in the estimates stems from cognitive diversity. As the details of the information set $\mathcal{F}_i$ cannot be known in practice, the information known by the ith expert is quantified as a fraction of the full information.

3 Model for Probability Forecasts

Consider two experts 1 and 2 who forecast the probability of the event A happening. Assume that the event A is determined by a pool of white noise, and that the experts 1 and 2 see respective δ_1 and δ_2 portions of this noise. These portions form their information sets. The overlap in these information sets is a fixed share ρ of what is seen by either expert. To make this more precise, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. On this space, define a white noise process that is indexed by the unit interval $S = [0, 1]$. A white noise process is a Gaussian process $\{X_B\}$ indexed by Borel measurable subsets B. The unit interval is endowed with the uniform measure μ. This gives the white noise process a covariance structure $\text{cov}(X_B, X_{B'}) = \mu(B \cap B') = |B \cap B'|$, i.e. the length of the intersection. The target event is defined as $A = \{X_S > 0\}$. Let $I_1, I_2 \subseteq S$ be the information sets observed by experts 1 and 2, respectively. Then,

$$\begin{aligned} \mu(I_1) &= |I_1| = \delta_1 \\ \mu(I_2) &= |I_2| = \delta_2 \\ \mu(I_1 \cap I_2) &= |I_1 \cap I_2| = \rho \end{aligned}$$

Call $\tilde{X}_{I_j} = X_{I_j}/\sqrt{1-\delta_j}$ the probit forecast of the jth expert. If Φ denotes the standard normal CDF, then the calibrated forecast given by the jth expert is

$$p_j = \mathbb{P}\left(A|\mathcal{F}_{I_j}\right) = \Phi\left(\tilde{X}_{I_j}\right)$$

Recall that if Z is standard normal random variable, then $\Phi(Z)$ is uniform on $[0, 1]$. Therefore the marginal distribution of p_j is uniform on $[0, 1]$ when $\delta_j = 0.5$, i.e. when the expert knows half of the information. If the expert knows less than half of the information, i.e. $\delta_j < 0.5$, then the marginal distribution of p_j is unimodal at 0.5 with the variance decreasing to 0 as $\delta_j \to 0$. On other hand, if the expert knows more than half of the information, i.e. $\delta_j > 0.5$, then the marginal distribution of p_j is more heavily concentrated around extreme probabilities. In fact, when $\delta_j = 1$, the marginal distribution of p_j is uniform over the set $\{0, 1\}$. Figure 1 illustrates these marginal distributions for δ_j equal to 0.3, 0.5, and 0.7.

Figure 2 illustrates the model with $N = 2$. The Gaussian process has been partitioned into four parts based on the information sets I_1 and I_2:

$$\begin{aligned} U &= X_{I_1/I_2} & M &= X_{I_1 \cap I_2} \\ V &= X_{I_2/I_1} & W &= X_{(I_1 \cup I_2)^c} \end{aligned}$$

Then,

$$\begin{aligned} X_{I_1} &= U + M \\ X_{I_2} &= M + V \end{aligned}$$

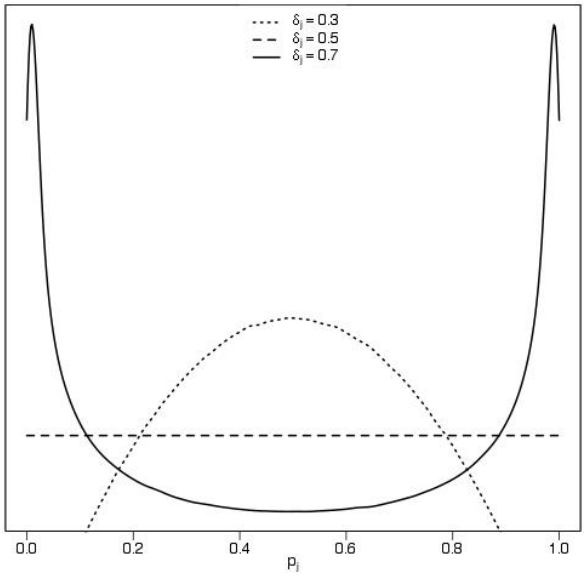


Figure 1: The marginal distribution of p_j under different levels of δ_j.

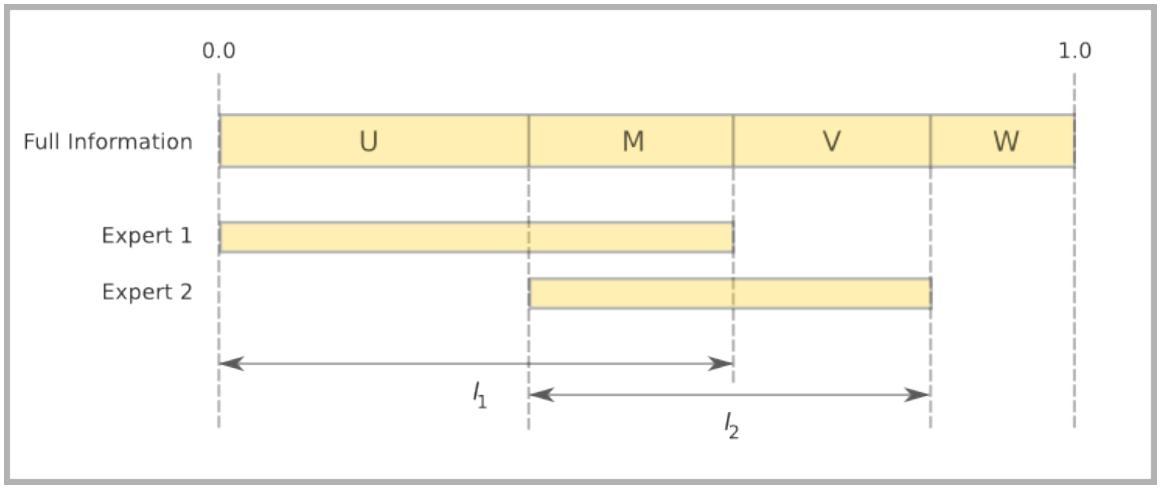


Figure 2: Illustration of the information theoretic model with two experts.

$$X_S = U + M + V + W,$$

where U, V, M, W are independent Gaussians with respective variances $\delta_1 - \rho$, $\delta_2 - \rho$, ρ, $1 + \rho - \delta_1 - \delta_2$. This gives (X_S, X_{I_1}, X_{I_2}) the following multivariate normal distribution.

$$\begin{pmatrix} X_S \\ X_{I_1} \\ X_{I_2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \delta_1 & \delta_2 \\ \delta_1 & \delta_1 & \rho \\ \delta_2 & \rho & \delta_2 \end{pmatrix} \right) \tag{2}$$

Figure 3: Illustration of the information theoretic model with N experts.

Consider now N experts. Let $|I_j| = \delta_j$ be the amount of information known by the jth expert, and $|I_i \cap I_j| = \rho_{ij}$ be the information overlap between the ith and jth experts. Expression (2) generalizes to the vector $(X_S, X_{I_1}, X_{I_2}, \ldots, X_{I_N})$ as follows.

$$\begin{pmatrix} X_S \\ X_{I_1} \\ \vdots \\ X_{I_N} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \ldots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \ldots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \ldots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \ldots & \delta_N \end{array} \right) \right) \tag{3}$$

This extension is illustrated in Figure 3. It is important to notice that the I_j does not have to be a contiguous segment of the unit interval. The sub-matrix Σ_{22} fully describes the information structure among the experts. This matrix has some technical conditions such as symmetry and non-singularity. In addition, Σ_{22} must describe a coherent information structure. The matrix Σ_{22} is coherent if and only if its information can be transformed into a diagram such as the one depicted by Figure 3.

3.1 Multinomial Outcomes

If the target event can take upon $K > 2$ outcomes, the the white noise process must be extended to a $K-1$-dimensional process $\{\boldsymbol{X}_B\}$, where $\boldsymbol{X_B} = (X_{1,B}, X_{2,B}, \ldots, X_{K-1,B})' \in \mathbb{R}^{K-1}$. The $K-1$-dimensional space is partitioned into K equal-sized cones with apexes at the origin. The target event results in the outcome k if $\boldsymbol{X}_S$ is in the kth cone. Figure 4 illustrates this for an event that can take upon values A, B, and C. In this case, the process $\{\boldsymbol{X}_B\}$ is 2-dimensional, starts at the origin $(0,0)$, and moves around the 2-dimensional Cartesian plane. Let the top, bottom-right, and bottom-left cones represents A, B, and C, respectively. Given that X_S is in the bottom-right cone, the final outcome of the target event is B.

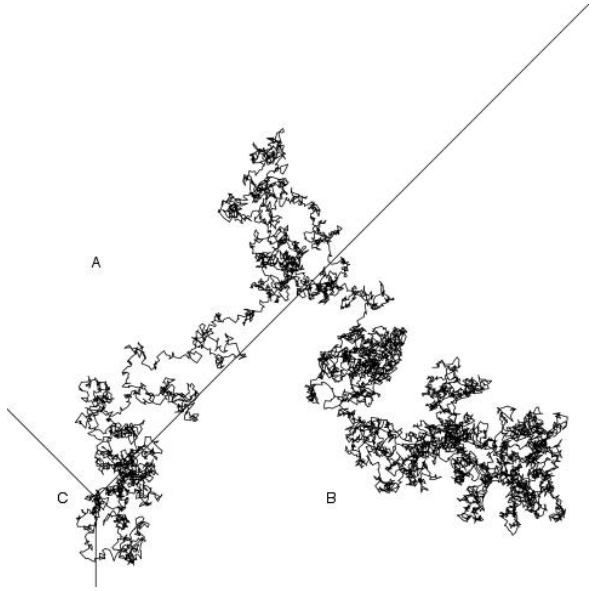


Figure 4: Illustration of the model for a target event with 3 outcomes.

Expert j observes a fixed share δ_j of the process. If the information sets of two experts i and j overlap, i.e. $|I_i \cap I_j| = \rho_{ij} > 0$, then the dependency between X_{I_i} and X_{I_j} is described by the cross-covariance $\text{cov}(X_{I_i}, X_{I_j}) = \rho_{ij} I_{K-1}$. This specifies a multivariate normal distribution for the vector $(X_S, X_{I_1}, X_{I_2}, \ldots, X_{I_N})'$. (EXPLAIN HOW THE MULTINOMIAL PROBIT PARTITIONS THE SPACE; WHAT IS THE CONNECTION BETWEEN X AND P). The remainder of this paper focuses on the binary case as this is the most common case in practice.

4 Extremization

The best in-principle forecast given the knowledge of N experts is $P(X_S > 0|\mathcal{F}')$, where $\mathcal{F}' = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_N$. This aggregate, however, assumes knowledge of the union of the information sets. Understanding the union $\mathcal{F}'$ is very difficult in practice, especially when the number of experts in

the group is large. Therefore the best aggregate probability that can be realistically hoped for is $\mathbb{P}(X_S > 0|p_1, \ldots, p_N)$.

To derive this aggregator under the information theoretic model, let $\boldsymbol{X}$ be a column vector of length N such that $X_j = X_{I_j}$ for $j = 1, \ldots, N$. If Σ_{22} is a coherent overlap structure and Σ_{22}^{-1} exists, then $X_S|\boldsymbol{X} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, where

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_2) = \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X} \tag{4}$$

and

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{5}$$

See Result 5.2.10 on p. 156 in Ravishanker and Dey [2001] for the formulas of the conditional multivariate normal distribution. This gives us the following aggregator.

$$\mathbb{P}\left(A|\boldsymbol{X}\right) = \mathbb{P}\left(X_S > 0|\boldsymbol{X}\right) = \Phi\left(\frac{\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}}{\sqrt{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}}\right) \tag{6}$$

Let α represent the amount of extremization that is performed for the average probit forecast. If $\bar{X} = \left(\sum_{j=1}^{N} \tilde{X}_{I_j}\right)/N$, then

$$\alpha\bar{X} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}}{\sqrt{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}} \qquad \Leftrightarrow \qquad \alpha = \frac{N\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}}{\left(\mathbf{1}_N'\tilde{\boldsymbol{X}}\right)\sqrt{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}} \tag{7}$$

This extremizing constant α is not necessarily greater or equal to 1. Therefore the information theoretic aggregator (6) is not guaranteed to give an extremized probability. The following examples illustrate two typical scenarios when the aggregator shifts the average probability closer to the furthest extreme instead. For the sake of illustration, both examples involve only two experts.

Example 1: Dominating Expert. Consider the following information structure.

$$\Sigma_{22} = \begin{pmatrix} 0.20 & 0.19 \\ 0.19 & 0.39 \end{pmatrix}$$

If $X_{I_1} = -0.85$ and $X_{I_2} = 0.16$, the information theoretic aggregate probability is 0.54. The average probability and probit forecast are 0.36 and 0.38, respectively. Given that these probabilities are less than 0.5 while the information theoretic aggregate is greater than 0.5, the extremizing constant is negative in both cases. To understand this result, notice that expert 2 knows almost everything that expert 1 knows. Therefore his forecast should be weighted much more heavily in the final aggregate. Given that only the information theoretic aggregator is able to take this into account, its aggregate can differ radically from the average probability and probit forecasts.

Example 2: Voting. Consider the following information structure.

$$\Sigma_{22} = \begin{pmatrix} 0.15 & 0.02 \\ 0.02 & 0.86 \end{pmatrix}$$

If $X_{I_1} = 0.27$ and $X_{I_2} = -0.16$, the information theoretic aggregate probability is 0.68. The average probability and probit forecasts are both equal to 0.47. Therefore the extremizing constant is negative. Notice that the union of the experts' information sets is 0.99. Therefore the experts as a group know almost all of the information.

It is possible to find similar examples where the information theoretic aggregate is on the same side but closer to 0.5 than the average probability and probit forecast. In most cases, however, the aggregator extremizes. The next section studies a class of information structures under which extremization is always guaranteed.

4.1 Compound Symmetric Information Structure

This section assumes that the experts' information sets have the same size and the amount of overlap between any two information sets is constant, i.e. $|I_1| = \cdots = |I_N|$ and $|I_i \cap I_j| = |I_h \cap I_k|$ for all $i \neq j$ and $h \neq k$. This results in the following compound symmetric overlap structure.

$$\begin{pmatrix} X_S \\ X_{I_1} \\ \vdots \\ X_{I_N} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{c|cccc} 1 & \delta & \delta & \dots & \delta \\ \hline \delta & \delta & \rho\delta & \dots & \rho\delta \\ \delta & \rho\delta & \delta & \dots & \rho\delta \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta & \rho\delta & \rho\delta & \dots & \delta \end{array} \right) \right),$$

where $\delta \in [0,1]$ is the fraction known by each expert and $\rho \in \left[\max\left\{\frac{N-\delta^{-1}}{N-1}, 0\right\}, 1\right]$ is the shared proportion of the known knowledge. The positive lower bound for ρ becomes active when $\delta > 1/N$ because then overlap in the information sets is unavoidable. This minimum overlap can be computed by assuming $\delta > 1/N$ and letting the shared information be the same for all N experts. That is, $I_i \cap I_j = I$ and $|I| = \rho\delta$ for all $i \neq j$. The minimum sharing occurs when $\rho\delta + N(\delta - \delta\rho) = 1$, which gives us the lower bound. The quantity $\rho\delta + N(\delta - \delta\rho)$ also describes the maximum coverage of the N experts, i.e. $\max |I_1 \cup I_2 \cup \cdots \cup I_N| = \rho\delta + N(\delta - \delta\rho)$.

Given that Σ_{22} can be written in the form $\Sigma_{22} = I_N(\delta - \rho\delta) + J_{N\times N}\rho\delta$, its inverse is

$$\Sigma_{22}^{-1} = I_N\left(\frac{1}{\delta - \rho\delta}\right) - J_{N\times N}\frac{\rho}{(1-\rho)\delta(1+(N-1)\rho)} \tag{8}$$

See the supplementary material of Dobbin and Simon [2005] for the proof of this fact. The determinant of Σ_{22} is

$$|\Sigma_{22}| = (\delta - \rho\delta)^N \left(1 + \frac{N\delta\rho}{\delta - \delta\rho}\right), \tag{9}$$

which follows from p. 32 in Rao [2009]. As the compound symmetric information structure depends only on two unknown parameters, the values of δ and ρ can be estimated in practice via the maximum likelihood method. That is,

$$\begin{aligned}(\hat{\delta}, \hat{\rho}) = &\arg\max_{\rho,\delta} \log\left[\frac{1}{\sqrt{(2\pi)^N |\Sigma_{22}|}} \exp\left(-\frac{1}{2}\boldsymbol{X}'\Sigma_{22}^{-1}\boldsymbol{X}\right)\right], \\ &\text{s.t. } \delta \in [0,1] \text{ and } \rho \in \left[\max\left\{\frac{N - \delta^{-1}}{N-1}, 0\right\}, 1\right]\end{aligned} \tag{10}$$

where Σ_{22}^{-1} and $|\Sigma_{22}|$ are given by (8) and (9), respectively. Unfortunately, (10) cannot be solved analytically. However, a simple grid-search can be used to find the estimates very efficiently.

The aggregator can be derived by applying (8) and (9) to the general formulas (4) and (5). The resulting conditional mean and variance are

$$\bar{\mu} = \frac{1}{(N-1)\rho + 1}\sum_{j=1}^{N} X_j \qquad\qquad \bar{\Sigma} = 1 - \frac{\delta N}{(N-1)\rho + 1}$$

The resulting aggregator is

$$\mathbb{P}\left(X_S > 0 | \boldsymbol{X}\right) = \Phi\left(\frac{\frac{1}{(N-1)\rho+1}\sum_{j=1}^{N} X_{I_j}}{\sqrt{1 - \frac{N\delta}{(N-1)\rho+1}}}\right)$$

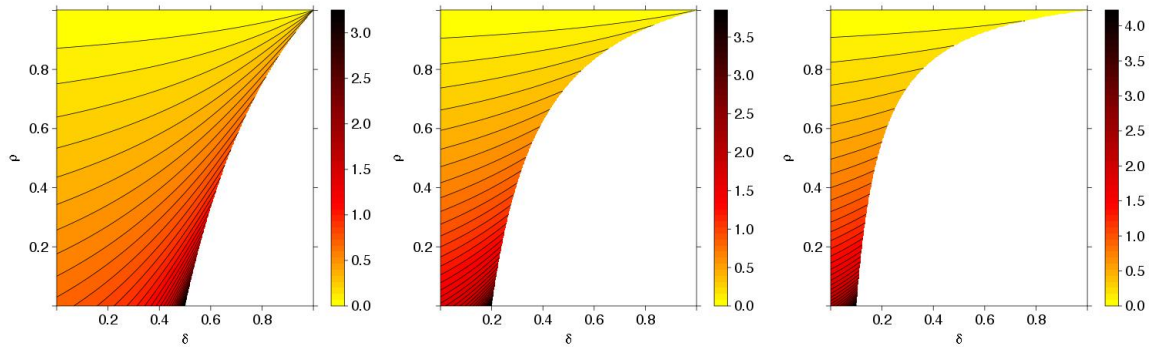
It is crucial to notice that this aggregator can learn the amount of extremization without a separate training set. Therefore it can be applied to a wide range of applied problems. Equating the aggregate probit forecast with $\bar{X}$ results in the following extremization factor.

$$\alpha = \frac{\frac{N\sqrt{1-\delta}}{(N-1)\rho+1}}{\sqrt{1 - \frac{N\delta}{(N-1)\rho+1}}} \tag{11}$$

Notice that, unlike (7), the extremizing constant in (11) does not depend on the forecasts $\boldsymbol{X}$. As the term inside the square-root must be non-negative, another technical restriction must be placed on ρ. That is, in addition to $\rho \in \left[\max\left\{\frac{N-\delta^{-1}}{N-1}, 0\right\}, 1\right]$, it is required that

$$1 - \frac{N\delta}{(N-1)\rho + 1} \geq 0 \qquad\qquad \Leftrightarrow \qquad\qquad \rho \geq \frac{N\delta - 1}{N-1}$$

Notice, however, that $N\delta - 1 > N - \delta^{-1}$ only when $\delta < 1/N$. But when $\delta < 1/N$, both $N\delta - 1$ and $N - \delta^{-1}$ are negative. Therefore this technical condition is redundant and can be ignored.



(a) N = 2 (b) N = 5 (c) N = 10

Figure 5: The amount of log-extremization $\log(\alpha)$ under different combinations of N (the number of experts), δ (the amount of information known by one expert), and ρ (the amount of information shared by any two experts).

Observation 4.1. *Under the compound symmetric information structure, the extremizing factor, α, is always greater or equal to 1. This means that the average probit forecast is always extremized.*

Proof. For a given δ, the extremizing constant α is minimized when $(N-1)\rho+1$ is maximized. This happens at $\rho = 1$. Plugging this into (11) gives

$$\alpha = \frac{\frac{N\sqrt{1-\delta}}{(N-1)\rho+1}}{\sqrt{1-\frac{N\delta}{(N-1)\rho+1}}} \geq \frac{\sqrt{1-\delta}}{\sqrt{1-\delta}} = 1$$

□

Expression (11) is particularly convenient because it only depends on three intuitive parameters. Therefore it can be analyzed graphically. Figures 5a to 5c describe the amount of log-extremization $\log(\alpha)$ under different values of ρ, δ, and N. By Observation 4.1 the amount of extremizing is always greater or equal to 1.0. Notice that most extremization occurs when $\delta = 1.0$ and $\rho = 1.0$, or when $\delta = 1/N$ and $\rho = 0$. In the first case, all experts know whether the target event materializes or not. In the second case, the group's information sets form a partition of the full information. Therefore as a group the experts know all the information. Such a group can re-construct X_S by simply adding up their individual probit forecasts. This means that aggregation becomes voting: if the sum of the probit forecasts is above 0, the event A materializes; else it does not. A similar observation has been made under the interpreted signal framework (see the example on information aggregation in Hong and Page [2009]). Therefore in the real-world voting can be expected to work well when the voters form a very knowledgable and diverse group of people.

Moving away from these two extreme points towards the upper left corner, where $\delta = 0.0$ and $\rho = 1.0$, decreases the amount of extremizing monotonically to 1.0. This trend follows directly

from Observation **??**. The decrease in the amount of information in $\boldsymbol{X}$ is caused by a combination of a) reduction in the amount of information that each individual expert holds and b) increase in the amount of shared information. Therefore the more knowledgable and diverse the group of experts is, the more their average probit forecast should be extremized. Contrast this with the generated signal framework where higher variance is typically considered negative. Under the information theoretic and interpreted signal frameworks, however, higher variance implies broader diversity among the experts and hence is considered helpful.

From Figures 5a to 5c it is clear that the feasible set of (δ, ρ)-values becomes smaller as N increases. This limitation arises from assuming a compound symmetric overlap structure. Having many experts, each with a considerable amount of information, simply leads to unavoidable overlap in the information sets. From the domain restriction on ρ, it is clear that $\rho \to 1$ as $N \to \infty$. Therefore in the limit the group of experts is equivalent to a single expert. This observation clearly does not reflect the real-world. When $N = 2$, on other hand, the compound symmetric overlap is completely general IT IS NOT BECAUSE DELTA IS STILL THE SAME. Therefore assuming a compound symmetric information structure can be appropriate for small numbers of experts but becomes overly restrictive as more experts enter the group.

5 Conclusion

This paper introduced a novel framework for analyzing subjective response data. Under this framework any response heterogeneity is assumed to arise from cognitive diversity. The mathematical tractability and real-world applicability of this model were illustrated by deriving an information theoretic aggregation rule for multiple probability forecasts. The aggregator resulted in closed-form expression for the amount of extremization that should be performed for the average probit forecast. By assuming a simplified information structure, the amount of extremization was studied graphically. This led to many insights on extremization. Given that these insights tend to align with common sense, the framework appears to be appropriate for probability aggregation.

Part of our future work is to continue developing the aggregator under the information theoretic framework. It is possible to place flexible priors on the unknown parameters and marginalize them with respect to their posterior distributions. This would lead to a principled aggregator that does not require any training. Instead, it could be applied directly to the data and would replace suboptimal methods such as the mean or median. As was discussed in Section 4, assuming a compound symmetric information structure is hardly a realistic choice. Therefore it will be necessary to develop a class of information structures that reflect the reality more closely. As it is unlikely that such a structure will lead to an aggregator with a closed-form solution, the aggregator will be provided in the form of an efficient algorithm.

Another future direction is to derive an information theoretic aggregator for subjective distribu-

tions. This is an important problem in Bayesian statistics where the analysis heavily depends on the choice of the prior distribution. Often the prior distribution is picked subjectively by the scientist who has previous experience on the problem at hand. If, however, the experiment is conducted by a group of scientists, their prior distributions must aggregated before the statistical analyses can be carried out.

The information theoretic framework is clearly a simplification of the reality. For instance, assuming that each expert produces an optimal probability forecast given his information set may not be a realistic assumption. The experts may believe in false information, hide their true beliefs, or be biased for many other reasons. This could be incorporated in the model by introducing an error term, possibly with a mean of zero, that is applied to the experts' probit forecasts. The resulting model, which is a hybrid of the generated signal and information theoretic frameworks, could lead to more realistic results. This improvement, however, may require a sacrifice in mathematical convenience.

References

Jeffrey A Baars and Clifford F Mass. Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. *Weather and Forecasting*, 20(6): 1034–1047, 2005.

J. Baron, L. H. Ungar, B. A. Mellers, and P. E. Tetlock. Two reasons to make aggregated probability forecasts more extreme. Manuscript submitted for publication (A copy can be requested by emailing Lyle Ungar at ungar@cis.upenn.edu), 2013.

Kevin Dobbin and Richard Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38, 2005.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

Lu Hong and Scott Page. Interpreted and generated signals. *Journal of Economic Theory*, 144(5): 2174–2196, 2009.

Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press Oxford, 2003.

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

C Radhakrishna Rao. *Linear statistical inference and its applications*, volume 22. John Wiley & Sons, 2009.

Nalini Ravishanker and Dipak K Dey. *A first course in linear model theory*. CRC Press, 2001.

Frederick Sanders. On subjective probability forecasting. *Journal of Applied Meteorology*, 2(2): 191–201, 1963.

V. A. Satopää, J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356, 2014.

Philip E Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

Robert L Vislocky and J Michael Fritsch. Improved model output statistics forecasts through model consensus. *Bulletin of the American Meteorological Society*, 76(7):1157–1164, 1995.

T. S. Wallsten and A. Diederich. Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, 18:1–18, 2001.

Peter WF Wilson, Ralph B DAgostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.