Information Theoretic Alternative to Signal Generation 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

Typically randomness in scientific estimation is assumed to arise from unmeasured or uncontrolled factors. This paper presents an information theoretic framework that introduces an alternative source of randomness. Our framework, which is particularly appropriate for subjective response data, is illustrated on probability aggregation. The probabilities are given by a group of experts who forecast whether an event will occur or not. Our aggregator uses the distribution of information among the experts and depends on easily interpretable parameters. Even though shifting the average probability forecasts closer to its nearest extreme, known as *extremizing*, is known to yield improved forecast performance, it is much less understood when and how much the average forecast should be extremized. By assuming a simplified information structure in our model, the amount of extremizing is studied closely under different values of the parameters. This leads to novel observations and a more principled understanding of extremization.

1 Introduction

Experts are often required to form estimates under incomplete information. These estimates can be made in several ways. Currently the two main approaches classify the estimate either as *generated* or *interpreted* (Hong and Page [2009]). The estimate is considered generated if the expert draws it from a given probability distribution. For instance, estimating the length of an object with a ruler can be assumed to be equivalent to drawing a value from a symmetric distribution, such as the Gaussian distribution, centered at the true length. 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, the score given by an Olympic judge on a figure skating performance can be considered interpreted. The judge has his personal set of criteria (categories) that ultimately decides the score.

The interpretation framework is clearly a more realistic model for subjective response data. Given that such data is very common in real-world applications, such as product reviews, online auctions, and voting, this framework is a great advance in understanding the gap between experimental and real-world results. Unfortunately, it is too general to be useful in developing novel methodology. First, interpretations differ between individuals and arise from a complex cognitive process that can be very difficult to estimate. Second, it is unclear how to construct the model to incorporate all the information that is used in making these interpretations.

The first contribution of this paper is to introduce a novel framework for signal generation. This framework, that is considered *information theoretic*, is based on the distribution of information among the experts. Each expert gives an estimate solely based on his personal information set. The information set includes any internal factors, such as personal preferences and interpretation criteria, and external factors, such as knowledge of the target event, that affect the estimation process. Therefore two overlapping information sets are assumed to produce positively correlated estimates. This framework, which is a good compromise between the psychological and analytical models, is much more amenable for development of future methodology than the interpretation framework.

The second contribution of this paper is to illustrate 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 socio-economic 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. The naive approach is to simply average the individual probability forecasts. Recent developments, however, 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] uses 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 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 two intuitive justifications for extremizing.

Many of the current extremizing aggregators, however, are overly simplistic or too detached from the psychology literature to provide the researcher any insight beyond the aggregate probability. It is mainly for this reason that it is still not well-understood when and how much the average probability should be extremized. Therefore it is necessary to learn the amount of extremizing from a separate training dataset. Furthermore, many of these aggregators assume that the individual probability forecasts arise from the generative framework. Under this assumption the optimal aggregation is accessible by weighted averaging (Parunak et al. [2013]). However, given that extremizing is known to improve the performance of the aggregate, it is unlikely that the generative framework is appropriate for probability aggregation.

These shortcomings are remedied by developing an aggregator based on our information theoretic framework. Under this model the average forecast is always extremized, and the amount of extremization is available in a closed-form. Given that this form depends on three intuitive parameters (the number of experts, the amount of information known by an expert, and the amount of information shared by two experts), it allows us to investigate 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 frameworks. The framework is then illustrated on probability aggregation. After deriving a closed-form expression for the amount of extremization, this form is analyzed in full generality and under a compound symmetric information structure. This simplified structure allows us to discuss and understand extremization in terms of a few intuitive parameters. The paper concludes with a discussion of model limitations and future directions.

2 Information Theoretic Framework

This section illustrates the information theoretic framework. The framework is discussed in comparison with the generative and interpretation frameworks. This comparison is by no means comprehensive as signal generation is a large part of the statistical literature.

Generated estimates can be considered as noisy or distorted versions of the target value. To make this more specific, let Ω be a set of states of the world. The set of possible outcomes is denoted with S. For illustrative purposes, this is assumed to be $S = \{G, B\}$, where G and B denote good and bad outcomes, respectively. The outcome function, $F : \Omega \to S$, maps the state of the world deterministically to the true outcome. If p is the true probability for a good outcome, then a generated probability forecast is realized by $\xi(p)$, where $\xi : [0,1] \to [0,1]$ is a function that randomly distorts the true probability. Therefore any estimate heterogeneity stems from randomness that is typically assumed to be caused by uncontrolled or unmeasured factors. Even though this framework is mathematically convenient, it can be drastically misaligned with the psychology literature and hence lead to results that are not reflective of the actual environment that produces the estimates.

The interoperation framework aims to correct these shortcomings by proposing a model that is more cognitive based. See Hong and Page [2009] for the original introduction. Under this framework, the set of states, Ω, is assume to be finite. The expert then partitions this set into non-overlapping subsets. This partition, know as an *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. Therefore any differences in estimates stem from cognitive diversity of the experts. Unfortunately, Hong and Page [2009] does 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 complex real-world applications.

The information theoretic framework abstracts away the intractable components of the interpretation framework. It does not assume expert interpretations nor place restrictions on the cardinality of Ω. To describe this framework in more detail, let $(\Omega', \mathcal{F}, \mathbb{P})$ be a probability space. Each expert is given a set of information $\mathcal{F}_i \subseteq \mathcal{F}$. If the set Ω' has a finite cardinality, $\mathcal{F}_i$ can be considered equivalent to the σ-field generated by the interpretation Π^i. Let $V \in S$ be the true outcome. Then the expert forecasts $p_i = \mathbb{P}(V = G|\mathcal{F}_i)$. This means that the expert makes the optimal probability forecasts given his information. 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 interpretation 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 relative to the amount of full information. For instance, expert i may know 15% of the full information while expert j knows only 8% of the full information. If in addition we specify that their shared information is, say, 3% of the full information, we have established an information structure for these experts. Working with this kind of information structures leads to mathematically convenient computations and psychologically valid inference. The following section illustrates this on probability aggregation.

3 Model for Probability Forecasts

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an event $A \in \mathcal{F}$ to be forecasted. Expert j knows information $\mathcal{F}_j \subseteq \mathcal{F}$ and forecasts $p_j = \mathbb{P}(A|\mathcal{F}_j)$. The best in-principle forecast given the knowledge of the N forecasters is $P(A|\mathcal{F}')$, where $\mathcal{F}' = \mathcal{F}_1 \cup \mathcal{F}_2 \cup .. \cup \mathcal{F}_N$. As in this work we do not allow the experts to exchange information with each other, the best aggregate probability is given by $\mathbb{P}(A|p_1, p_2, \ldots, p_N)$.

For illustrative purposes, we first assume only two experts and then generalize the model to N experts. Under our model the event A is determined by a pool of white noise. Experts 1 and 2 see respective δ_1 and δ_2 portions of the noise. These portions form their information sets. The overlap in their information sets is a fixed share ρ of what is seen by either expert. This can be made more precise by letting $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which we define a white noise process indexed by the unit interval S. A white noise process is a Gaussian process $\{X_B\}$ indexed by Borel measurable subsets B. We endow the unit interval with the uniform measure, μ. This gives the white noise process a covariance structure of $Cov(X_B, X_{B'}) = \mu(B \cap B') = |B \cap B'|$, 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. Thus,

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

Call X_{I_j} the probit forecast of the jth expert. If Φ denotes the standard normal CDF, then

$$p_j = \mathbb{P}(A|\mathcal{F}_{I_j}) = \Phi(X_{I_j})$$

for $j = 1, 2$, and the aggregator is given by $\mathbb{P}(X_S > 0|p_1, p_2)$.

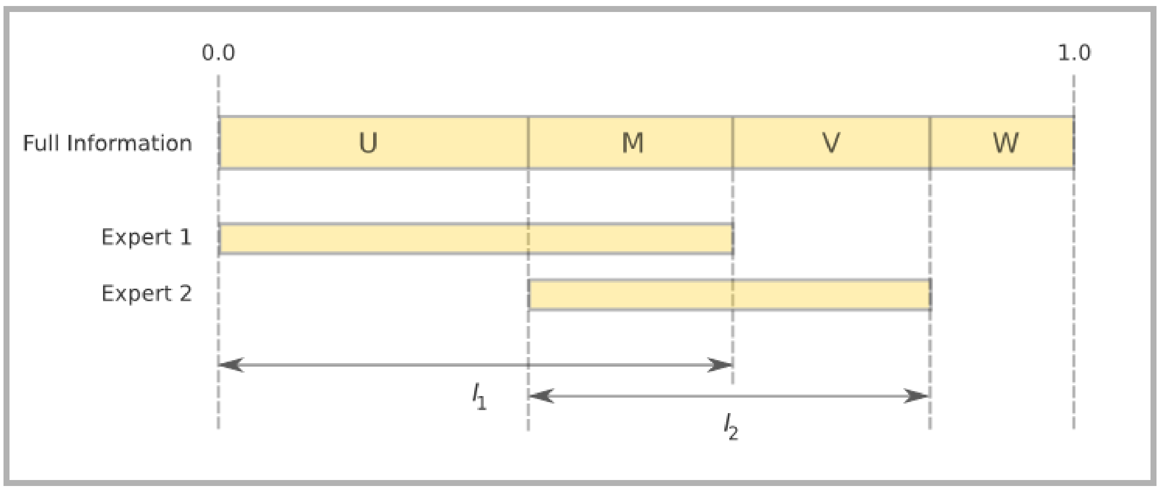

Figure 1: Illustration of the model with two experts.

Figure 1 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 $X_{I_1} = U + M$, $X_{I_2} = M + V$, and $X_S = U + M + V + W$, where U, V, M, W are independent Gaussians with respective variances $\delta_1 - \rho$, $\delta_2 - \rho$, ρ, $1 + \rho - \delta_2 - \delta_3$. This gives

(X_S, X_{I_1}, X_{I_2}) a multivariate normal distribution. More specifically, we have

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

Figure 2: Illustration of the 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. Then expression (1) generalizes to the vector $(X_S, X_{I_1}, X_{I_2}, \ldots, X_{I_N})$. This gives us

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

This is illustrated in Figure 2. It is important to notice that the I_j does not have to be 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 2.

4 Extremization

Let $\boldsymbol{X}$ be a column vector of length N such that $X_j = X_{I_j}$. If Σ_{22} is a coherent overlap structure such that Σ_{22}^{-1} exists, then $X_S|\boldsymbol{X} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, where

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_2) = \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X} \tag{2}$$

and

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{3}$$

See Result 5.2.10 on p. 156 in Ravishanker and Dey [2001] for the formulas of a conditional multivariate normal distribution. The aggregator then becomes

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

Let α represents the amount of extremization that is performed for the average probit forecast. If $\bar{X}$ denotes the sample average, then

$$\alpha\bar{X} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}}{\sqrt{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}} \qquad \Leftrightarrow \qquad \alpha = \frac{N\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{X}}{(\boldsymbol{1}_N'\boldsymbol{X})\sqrt{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}} \tag{4}$$

Even though this expression assumes no structure on Σ_{22} and depends on $N + \frac{N(N-1)}{2}$ unknown parameters, it can be used to gain insight on the behavior of the extremizing constant, α. One useful development is to determine the amount of information in $\boldsymbol{X}$.

Lemma 4.1. *If δ_0 denotes the information in $\boldsymbol{X}$, then*

$$\alpha = \frac{1}{\sqrt{1-\delta_0}} \qquad \Leftrightarrow \qquad \delta_0 = 1 - \alpha^{-2}$$

Proof. Let α_N denote the extremizing constant for $\boldsymbol{X}$. Consider a single expert whose probit forecast is $\bar{X}$. Denote the size of his information set by δ_0. The extremizing constant for his forecast, as is given by (4), simplifies to

$$\alpha_1 = \frac{1}{\sqrt{1-\delta_0}}$$

Setting $\alpha_1 = \alpha_N$ gives us the final result.

□

Based on Lemma 4.1 there is a monotonic and positive relationship between α and δ_0. This means that the more the sample average is extremized the more information its corresponding $\boldsymbol{X}$

contains, and *vice versa*. Lemma 4.1 is interesting for two reasons: (a) it allows the researcher to use black-box models from existing literature to determine the extremizing constant and then use it to analyze the amount of information in $\boldsymbol{X}$, and (b) it allows us to easily show that $\alpha \geq 1$ under any information structure.

Theorem 4.2. *Under the model described in Section 3, the extremizing factor, α, is always greater or equal to 1. This means that the average probit forecast is always extremized.*

Proof. By Lemma 4.1 we have that

$$\alpha = \frac{1}{\sqrt{1-\delta_0}}$$

Given that $\delta_0 \in [0,1]$, it follows that $\alpha \in [1, \infty)$.

$\square$

To continue our analysis of extremization, it is necessary to reduce the number of degrees of freedom by assuming a simpler form for the overlap structure Σ_{22}.

4.1 Compound Symmetric Information Structure

In this section we assume that any two experts know and share the same amount of information. This gives us the compound symmetric overlap structure.

$$\begin{pmatrix} S \\ X_1 \\ \vdots \\ X_N \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{c|cccc} 1 & \delta & \delta & \dots & \delta \\ \hline \delta & \delta & \rho\delta & \dots & \rho\delta \\ \delta & \rho\delta & \delta & \dots & \rho\delta \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta & \rho\delta & \rho\delta & \dots & \delta \end{array} \right) \right),$$

where $\delta \in [0,1]$ and $\rho \in \left[\max\left\{\frac{N-\delta^{-1}}{N-1}, 0\right\}, 1\right]$. The lower bound for ρ becomes necessary when $\delta > 1/N$ because then overlap is unavoidable. This minimum can be computed by assuming $\delta > 1/N$ and letting the shared information be the same for all N experts. That is, $|I_i \cap I_j| = |I| = \rho\delta$ 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. That is, $\rho\delta + N(\delta - \delta\rho) = \max|I_1 \cup I_2 \cup \dots \cup I_N|$.

Note that this model constrains the marginal distribution of $p_j = \Phi(X_{I_j})$ to be uniform on $[0,1]$ when $\delta_j = 1$. To see this, recall that if $X_{I_j} \sim \mathcal{N}(0,1)$, then $\Phi(X_{I_j})$ is uniform on $[0,1]$. If this does not hold empirically, it is a sign that the model cannot be correct on the micro-level. If X_{I_j} appears more (respectively less) concentrated about 0.5, then the model can be adjusted by changing δ_j to a smaller fraction.

Notice that Σ_{22} can be written as $\Sigma_{22} = I_N(\delta - \rho\delta) + J_{N\times N}\rho\delta$. Therefore its inverse is $\Sigma_{22}^{-1} = I_N\left(\frac{1}{\delta-\rho\delta}\right) - J_{N\times N}\frac{\rho\delta}{(\delta-\rho\delta)((\delta-\rho\delta)+N\rho\delta)}$ (see the supplementary material for Dobbin and Simon [2005]). Applying this to equations (2) and (3) gives us the conditional mean

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

and variance

$$\bar{\Sigma} = 1 - \frac{\delta N}{(N-1)\rho+1}$$

The aggregation rule then becomes

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

From this aggregator we obtain an expression for the amount of extremization under the compound symmetric information structure.

$$\alpha = \frac{\frac{N}{(N-1)\rho+1}}{\sqrt{1-\frac{N\delta}{(N-1)\rho+1}}} \tag{5}$$

Unlike (4), the extremizing constant under the compound symmetric information structure does not depend on the forecasts, $\boldsymbol{X}$. As the term inside the square-root must be non-negative, we have another technical restriction on ρ. That is, in addition to $\rho \in \left[\max\left\{\frac{N-\delta^{-1}}{N-1}, 0\right\}, 1\right]$, we require

$$\rho \geq \frac{N\delta - 1}{N-1}$$

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.

Expression (5) is particularly convenient because it only depends on three intuitive parameters. Therefore it can be examined graphically. Figures 3 to 5 describe the amount of log-extremization, $\log(\alpha)$, under different values of ρ, δ, and N. By Theorem 4.2 the amount of extremizing is always greater or equal to 1.0. Notice that most extremization occurs when $\delta = 1.0$ and $\rho = 1$, or when $\delta = 1/N$ and $\rho = 0$. In the first case, all N experts know whether the event A materializes or not. In the second case, each expert holds an independent set of information such that the group knows all the information. Such a group of experts 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 framework (see the example on information aggregation in Hong and Page

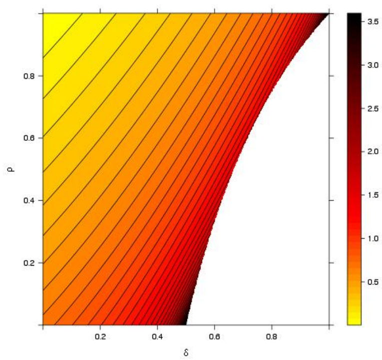

Figure 3: N = 2

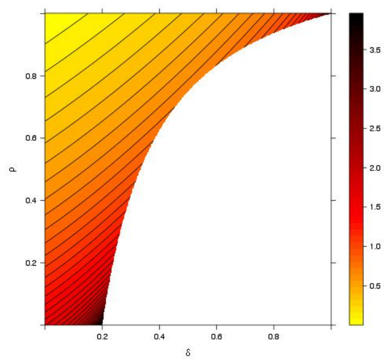

Figure 4: N = 5

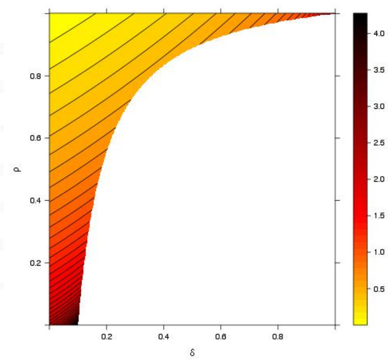

Figure 5: N = 10

[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.

As we move from these two extreme points towards the upper left corner, where $\delta = 0.0$ and $\rho = 1.0$, the amount of extremizing decreases monotonically to 1.0. This trend can be deduced directly from Lemma 4.1. The decrease in the amount of information in $\boldsymbol{X}$ is caused by a combination of (i) decrease in the amount of information that each individual expert holds and (ii) 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 framework where higher variance is typically considered negative. Under the information theoretic and interpreted frameworks, however, higher variance implies broader diversity among the experts and hence is considered helpful.

From Figures 3 to 5 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.

5 Conclusion

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.

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

H Parunak, Sven A Brueckner, Lu Hong, Scott E Page, and Richard Rohwer. Characterizing and aggregating agent estimates. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1021–1028. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

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.

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.