Partial Information Framework: Aggregating Estimates from Diverse Information Sources

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

Prediction polling is an increasingly popular form of crowdsourcing in which multiple participants estimate the probability or magnitude of some future event. These estimates are then aggregated into a single forecast. Historically, randomness in scientific estimation has been generally assumed to arise from unmeasured factors which are viewed as measurement noise. However, when combining subjective estimates, heterogeneity stemming from differences in the participants' information is often more important than measurement noise. This paper formalizes information diversity as an alternative source of such heterogeneity and introduces a novel modeling framework that is particularly well-suited for prediction polls. A practical specification of this framework is proposed and applied to the task of aggregating probability and point estimates from two real-world prediction polls. In both cases our model outperforms standard measurement-error-based aggregators, hence providing evidence in favor of information diversity being the more important source of heterogeneity.

Keywords: Expert belief; Forecast heterogeneity; Judgmental forecasting; Model averaging; Noise reduction

1 Introduction

Past literature has distinguished two types of polling: prediction and opinion polling. In broad terms, an opinion poll is a survey of public opinion, whereas a prediction poll involves multiple agents collectively predicting the value of some quantity of interest (Goel et al., 2010; Mellers et al., 2014). For instance, consider a presidential election poll. An opinion poll typically asks the voters who they will vote for. A

*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); Shane T. Jensen is a Statistician, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340 (e-mail: stjensen@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 by 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 also like to thank Don Moore for providing us with the weight dataset.

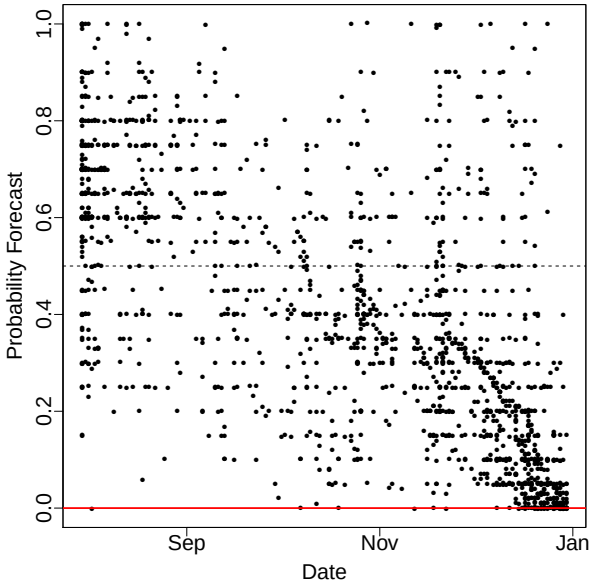


Figure 1: Probability forecasts of the event "Will Moody's issue a new downgrade on the long-term ratings for any of the eight major French banks between 30 July 2012 and 31 December 2012?" The points have been jittered slightly to make overlaps visible.

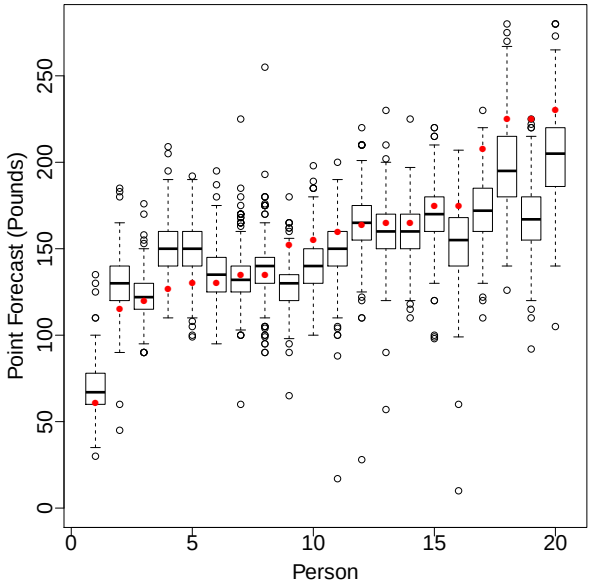


Figure 2: Point forecasts of the weights of 20 different people. The boxplots have been sorted to increase in the true weights (red dots). Some extreme values were omitted for the sake of clarity.

prediction poll, on the other hand, could ask which candidate they think will win in their state. A liberal voter in a dominantly conservative state is likely to answer differently to these two questions. Even though opinion polls have been the dominant focus historically, prediction polls have become increasingly popular in the recent years, due to modern social and computer networks that permit the collection of a large number of responses from both human and machine agents. This has given rise to crowdsourcing platforms, such as MTurk and Witkey, and many companies, such as Myriada, Lumenogic, and Inkling, that have managed to successfully capitalize on the benefits of collective wisdom.

This paper introduces statistical methodology designed specifically for the rapidly growing practice of prediction polling. The methods are illustrated on real-world data involving two common types of responses, namely probability and point forecasts. The probability forecasts were collected by the Good Judgment Project (GJP) (Ungar et al. 2012; Mellers et al. 2014) as a means to estimate the likelihoods of international political future events deemed important by the Intelligence Advanced Research Projects Activity (IARPA). Since its initiation in 2011, the project has recruited thousands of forecasters to make probability estimates and update them whenever they felt the likelihoods had changed. To illustrate, Figure 1 shows the forecasts for one of these events. This example involves 522 forecasters making a total of $1,669$ predictions between 30 July 2012 and 30 December 2012 when the event finally resolved as

"No" (represented by the red line at 0.0). In general, the forecasters reported updates very infrequently. Furthermore, not all forecasters made probability estimates for all the events, making the dataset very sparse. The point forecasts for our second application were collected by Moore and Klein (2008) who recruited 416 undergraduates from Carnegie Mellon University to guess the weights of 20 people based on a series of pictures. This is an experimental setup where each participant was required to respond to all the questions, leading to a fully completed dataset. The responses are illustrated in Figure 2 that shows the boxplots of the forecasters' guesses for each of the 20 people. The red dots represent the corresponding true weights.

Once the predictions have been collected, the challenge is to combine them into a single consensus. This is typically done for the sake of decision-making and improved accuracy. Principled aggregation, however, requires an assumption about the source of heterogeneity among the forecasts. In particular, it is necessary to specify how the forecasts differ from the true value of the target quantity. For the past several decades, potentially due to the early forms of data collection, measurement error has been considered as the main source of heterogeneity. This approach has become the standard and is often applied in practice even when data variation is dominated by causes besides error in measurement. For instance, assuming measurement error may be reasonable in modeling repeated estimates from a single instrument. However, it is unlikely to hold in prediction polling, where the estimates arise from multiple, often widely different sources.

The main contribution of this paper is a new source of forecast heterogeneity, called *information diversity*, that serves as an alternative to measurement error. In particular, any variation in the forecasts is assumed to stem from information available to the forecasters and how they decide to use it. For instance, forecasters studying the same (or different) articles about a company may use separate parts of the information and hence report differing predictions on the company's future revenue. Such diversity forms the basis of a novel modeling framework known as the *partial information framework*. Theory behind this framework was originally introduced for probability forecasts by Satopää et al. (2015); though their specification is somewhat restrictive for empirical applications. The current paper generalizes the framework beyond probability forecast and removes all unnecessary assumptions, leading to a new specification that is more appropriate for practical applications. This allows the decision-maker to build context-specific models and aggregators, instead of relying on the usual mean or median aggregates.

The paper is structured as follows. Section 2 first describes the partial information framework at its most general level and then introduces a practical specification of the framework. The section ends with a brief review of previous work on modeling forecasts. Section 3 derives a numerical procedure that efficiently estimates the information structure among the forecasters. Sections 4 and 5 illustrate the framework on synthetic and real-world forecasts of different types of outcomes. In particular, the framework is used to analyze probability and point forecasts from the two real-world prediction polls discussed above. The resulting partial information aggregators achieve a noticeable performance improvement over the common measurement-error-based aggregators, suggesting that information diversity is the more important source of forecast heterogeneity. Finally, Section 6 concludes with a summary and discussion of future research.

2 Partial Information Framework

2.1 General Framework

Consider N forecasters and suppose forecaster j predicts X_j for some (random) quantity of interest Y. The prediction X_j is simply an estimator of Y. Therefore, as is the case with all estimators, its deviation from the truth can be broken down into two components: bias and noise. On the theoretical level, it is important to separate these two problems because they are addressed by different mechanisms. This paper considers the aggregation problem for conditionally unbiased forecasts, therefore isolating the task to one of noise reduction. In particular, the forecasters are assumed to be conditionally unbiased such that $\mathbb{E}(Y|X_j) = X_j$ for all $j = 1, \ldots, N$. In probability forecasting this assumption is often known as calibration. It is particularly common in the economic, statistical, and meteorological forecasting literature (Ranjan and Gneiting, 2010) and traces back to Murphy and Winkler (1987). However, it is different from the game-theoretic alternative discussed, e.g., in Dawid (1982) and Foster and Vohra (1998). In general, the assumption of conditional biased forecasts is, to some degree, self-fulfilling if the forecasters operate under a loss function that is minimized by the conditional expectation of Y (given the forecaster's information). For this reason such loss functions can be called *revealing*. Even though this group of functions involves many well-known penalties such as the Kullback-Leibler divergence, Mahalanobis distance, and the quadratic loss (Banerjee et al., 2005), the current paper analyzes univariate outcomes and only focuses on minimizing the quadratic loss.

The partial information framework assumes that the observables Y and X_j are measurable random variables under some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The probability measure $\mathbb{P}$ provides a non-informative yet proper prior on Y and reflects the *basic information* known to all forecasters. Such a prior has been discussed extensively in the economics and game theory literature where it is usually known as the *common prior*. Even though this is a substantive assumption in the framework, specifying a prior distribution cannot be avoided as long as the model depends on a probability space. This includes essentially any probability model for forecast aggregation. How the prior is incorporated depends on the problem context: it can be chosen explicitly by the decision-maker, computed based on past observations of Y, or estimated directly from the forecasts.

The principal σ-field $\mathcal{F}$ can be interpreted as all the possible information that can be known about Y. In any Bayesian setup, with a revealing loss function, it is more or less tautological that forecaster j predicts $X_j = \mathbb{E}(Y \mid \mathcal{F}_j)$ based on some partial *information set* $\mathcal{F}_j \subseteq \mathcal{F}$. Therefore $\mathcal{F}_i \neq \mathcal{F}_j$ if $X_i \neq X_j$, and forecast heterogeneity stems purely from *information diversity*. Note, however, that if forecaster j uses a simple rule, $\mathcal{F}_j$ 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. 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. This general point of view was motivated in Satopää et al. (2015) with simple examples that illustrate how the optimal aggregate is not well-defined without assumptions on the information structure among the forecasters.

For the purposes of aggregation any available information discarded by the forecaster may as well

not exist because information comes to the aggregator only through the forecasts. Therefore it is not in any way restrictive to assume that $\mathcal{F}_j = \sigma(X_j)$, where $\sigma(X_j)$ is the σ-field generated by X_j. The form $X_j = \mathbb{E}(Y|\mathcal{F}_j)$ then arises directly from the assumption of conditional unbiasedness and the existence of an underlying probability model. The following proposition uses these assumptions to relate the target quantity to the forecasts. All the proofs are deferred to the Appendix.

Proposition 2.1. *Suppose $\mathcal{F}_j = \sigma(X_j)$ such that $\mathbb{E}(Y|\mathcal{F}_j) = \mathbb{E}(Y|X_j) = X_j$ for all $j = 1, \ldots, N$. Then, the following holds.*

i) The forecasts are marginally consistent: $\mathbb{E}(Y) = \mathbb{E}(X_j)$.

ii) $\mathrm{Cov}\,(X_j, X_i) = \mathrm{Var}\,(X_i)$ *if* $\mathcal{F}_i \subseteq \mathcal{F}_j$.

iii) The variance increases in information: $\mathrm{Var}\,(X_i) \leq \mathrm{Var}\,(X_j)$ *if* $\mathcal{F}_i \subseteq \mathcal{F}_j$.

This proposition raises several important points: First, item i) provides guidance in estimating the prior mean of Y from the observed forecasts. Second, given that $Y = \mathbb{E}(Y|\mathcal{F})$ and $\mathcal{F}_j \subseteq \mathcal{F}$ for all $j = 1, \ldots, N$, item ii) shows that the covariance matrix $\mathbf{\Sigma}_X$ of the X_j's extends to the unknown Y as follows:

$$\mathrm{Cov}\,\left((Y, X_1, \ldots, X_N)'\right) = \begin{pmatrix} \mathrm{Var}\,(Y) & \mathrm{diag}(\mathbf{\Sigma}_X)' \\ \mathrm{diag}(\mathbf{\Sigma}_X) & \mathbf{\Sigma}_X \end{pmatrix}, \tag{1}$$

where $\mathrm{diag}(\mathbf{\Sigma}_X)$ denotes the diagonal of $\mathbf{\Sigma}_X$. This provides leverage for regressing Y on the X_j's without a separate training set of past predictions and known outcomes. The resulting estimator, called the *revealed* aggregator, is denoted with $X'' := \mathbb{E}(Y \mid \mathcal{F}'')$, where $\mathcal{F}'' := \sigma(X_1, \ldots, X_N)$ is the σ-field generated (or revealed) by the X_j's. The revealed aggregator uses the forecasters' information optimally. In particular, if $\mathbb{E}(Y^2) < \infty$, then among all variables measurable with respect to $\mathcal{F}''$, the revealed aggregator minimizes the expected quadratic loss and is therefore considered the relevant aggregator under each specific instance of the framework.

Third, item iii) shows that $\mathrm{Var}\,(X_j)$ increases to $\mathrm{Var}\,(Y)$ as forecaster j learns and becomes more informed. Therefore $\mathrm{Var}\,(X_j)$ quantifies the amount of information used by forecaster j, and $\mathrm{Cov}\,(X_i, X_j)$ can be interpreted as the amount of information overlap between forecasters i and j. Given that being non-negatively correlated is not generally transitive (Langford et al., 2001), these covariances are not necessarily non-negative even though all forecasts are non-negatively correlated with the outcome. Calibrated forecasts with negative correlation can arise in a real-world setting. For instance, consider two forecasters who see voting preferences of two different sub-populations that are politically opposed to each other. Each individually is a weak predictor of the total vote on any given issue, but they are negatively correlated because of the likelihood that these two blocks will largely oppose each other. Observe that under the partial information framework increased variance suggests more information and is generally deemed helpful. This is a clear contrast to the standard statistical models that often regard higher variance as increased noise.

2.2 Gaussian Partial Information Model

The general partial information framework is clearly too abstract to be applied in practice. Therefore a more specific model within the framework is needed. The first step is to choose a joint distribution for the model variables $Y, X_1, \ldots, X_N$. This in general requires an understanding of how these variables change together. One approach is to refer to Proposition 2.1 and parametrize the joint distribution in terms of covariances. This suggests the multivariate Gaussian distribution that is a typical starting point in developing statistical methodology and often provides the cleanest entry into the issues at hand. Furthermore, the Gaussian distribution incorporates conditional distributions that have simple forms also within the Gaussian family. This turns out to be particularly useful in deriving the revealed aggregator.

Intuitively the Gaussian distribution can be motivated by a pool of M information particles. Each particle, which can be interpreted as representing a piece of information, has either a positive or negative effect on Y. The total sum (integral) of these particles determines the final value of Y. Each forecaster, however, observes only the sum of some subset of the particles. Based on this sum, the forecaster makes an estimate of Y. If the particles are independent and have small tails, then as $M \to \infty$, the joint distribution of the forecasters' observations will be asymptotically Gaussian. Therefore, given that the number of such particles in a real-world setup is likely to be large, it makes sense to model the forecasters' observations as jointly Gaussian. This distributional choice has been used to model forecasters before (see, e.g., Broomell and Budescu 2009; Di Bacco et al. 2003).

In many applications, however, Y and X_j may not be supported on the whole real line. For instance, in probability forecasting of binary events typically $Y \in \{0, 1\}$ and $X_j \in [0, 1]$. Fortunately, this limitation can be easily handled by borrowing from the theory of generalized linear models (McCullagh et al., 1989) and utilizing a *link function*. The following *Gaussian model* makes this concrete by introducing $N + 1$ auxiliary variables, called the *information variables*, that follow a multivariate Gaussian distribution with the covariance pattern (1):

$$\begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_N \end{pmatrix} \sim \mathcal{N}_{N+1} \left(\mathbf{0}, \begin{pmatrix} 1 & \operatorname{diag}(\boldsymbol{\Sigma})' \\ \operatorname{diag}(\boldsymbol{\Sigma}) & \boldsymbol{\Sigma} \end{pmatrix} := \left(\begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \ldots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \ldots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \ldots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \ldots & \delta_N \end{array} \right) \right). \tag{2}$$

The target outcome is given by $Y = g(Z_0)$, where $g(\cdot)$ is an application-specific link function that fully specifies the model instance. In general it makes sense to have $g(\cdot)$ map from the real numbers to the support of Y. For instance, if $Y \in \{0, 1\}$, then $g : \mathbb{R} \to \{0, 1\}$. The fact that $\operatorname{Var}(Z_0) = 1$ is irrelevant because this choice can be compensated by the link function. The remaining information variables, namely $Z_1, \ldots, Z_N$ summarize the forecasters' information. In particular, the jth forecaster's prediction and the revealed aggregator are $X_j = \mathbb{E}(Y|Z_j)$ and $X'' = \mathbb{E}(Y|Z_1, \ldots, Z_N)$, respectively. These expectations can be often computed by aggregating the information variables via the following

conditional Gaussian distributions:

$$Z_0|Z_j \sim \mathcal{N}\left(Z_j, 1-\delta_j\right) \text{ and}$$
$$Z_0|\mathbf{Z} \sim \mathcal{N}\left(\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}, 1-\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\text{diag}(\boldsymbol{\Sigma})\right),$$

where $\mathbf{Z} = (Z_1, \ldots, Z_N)'$.

One of the most important instances of the Gaussian model considers point forecasts X_j of a continuous outcome Y. The next proposition analyzes the relationship between X'' and the weighted average of the forecasts under this particular setup.

Proposition 2.2. *Consider the Gaussian model for point forecasts of continuous outcomes with the prior* $Y \sim \mathcal{N}(\mu_0, \sigma_0^2)$. *Denote the set of symmetric positive definite* $N \times N$ *matrices with* $\mathcal{S}_{++}^N$. *If* $\boldsymbol{\Sigma} \in \mathcal{S}_{++}^N$ *and* $\mathbf{X} = (X_1, \ldots, X_N)' \in \mathbb{R}^N$, *then the revealed aggregator is* $X'' = diag(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mu_0\mathbf{1}_N) + \mu_0$. *Consider the weighted average* $\mathbf{w}'\mathbf{X}$, *where* $\mathbf{w} = (w_1, \ldots, w_N)'$ *such that all* $w_1, \ldots, w_N \geq 0$ *and* $\sum_{i=1}^N w_j = 1$. *The revealed aggregator* X'' *reduces to* $\mathbf{w}'\mathbf{X}$ *only if there is a forecaster* j *such that* $w_j = 1$ *and* $\text{Cov}\left(X_i, X_j\right) = \rho_{i,j} = \delta_i$ *for all* $i \neq j$. *In this case* $\mathcal{F}_i \subseteq \mathcal{F}_j$ *for all* $i \neq j$.

Recall from Section 2.1 that X'' minimizes the expected quadratic loss among all functions measurable with respect to $\mathcal{F}''$. Given that the weighted average $\mathbf{w}'\mathbf{X} \in \mathcal{F}''$, under the Gaussian model no weighted average can outperform X'' in terms of expected quadratic loss. In particular, the weighted average is necessarily sub-optimal except in the unlikely case when one forecaster knows everything that all the other forecasters know. This result is illustrated on synthetic and real-world data in Sections 4 and 5.2, respectively.

2.3 Previous Work

2.3.1 Interpreted Signal Framework

The *interpreted signal framework* assumes different predictions to arise from differing interpretation procedures (Hong and Page, 2009). For example, consider two forecasters who visit a company and predict its future revenue. One forecaster may carefully examine the company's technological status while the other pays closer attention to what the managers say. Even though the forecasters receive the same information, they interpret it differently and can end up reporting different forecasts. Therefore forecast heterogeneity is assumed to stem from "cognitive diversity".

This is a very reasonable assumption that has been discussed by many authors. For example, Broomell and Budescu (2009) analyze a model that maps the cues to the individual forecasts via different linear regression functions; Parunak et al. (2013) demonstrate that the optimal aggregate of interpreted forecasts can be outside the convex hull of the forecasts. No previous work, however, has discussed a formal framework that links the interpreted forecasts to their target quantity in an explicit yet flexible manner. Consequently, the interpreted signal framework has remained relatively abstract. Our partial information framework, however, formalizes the intuition behind it, allows quantitative predictions, and provides a flexible construction that can be adopted to a broad range of forecasting setups.

2.3.2 Measurement Error Framework

In the absence of a quantitative interpreted signal model, prior applications have typically explained forecast heterogeneity with standard statistical models. These models are different formalizations of the *measurement error framework* that generates forecast heterogeneity purely from a probability distribution. More specifically, the framework assumes a "true" (possibly transformed) forecast θ, which can be interpreted as the prediction made by an ideal forecaster. This is then somehow measured by the individual forecasters with mean-zero idiosyncratic error. More specifically, the forecasts are unbiased estimates of θ, that is, $\mathbb{E}(X_j|\theta) = \theta$, which is not the same as assuming calibration, namely $\mathbb{E}(Y|X_j) = X_j$. Therefore an unbiased estimation model is very different from a calibrated model. This distinction can be further emphasized by noting that the measurement-error aggregators are often different types of weighted averages of the individual (transformed) predictions. Therefore, according to Proposition 2.2, the set of measurement-error aggregators barely overlaps with the set of partial information aggregators. Consequently, measurement error and information diversity are not only philosophically different but they also require very different aggregators.

The measurement error framework has become standard practice possibly because of its simplicity and familiarity. Unfortunately, there are a number of disadvantages. First, measurement-error aggregators estimate θ instead of the realized value y of the random variable Y. In general these aggregators cannot target y because they do not incorporate a correlation structure between Y and the X_j's. The partial information framework, however, motivates a covariance structure among these variables and hence can target y. Second, the standard assumption of conditional independence of the observations forces a specific and highly unrealistic structure on interpreted forecasts (Hong and Page, 2009). Measurement-error aggregators also cannot leave the convex hull of the individual forecasts, which further contradicts the interpreted signal framework (Parunak et al., 2013) and can result in poor empirical performance. Third, the underlying model is rather implausible. Relying on a true forecast θ invites philosophical debate, and even if one assumes the existence of such a quantity, it is difficult to believe that the forecasters are actually seeing the ideal forecast θ with independent noise. Therefore, whereas the interpreted signal framework proposes a plausible micro-level explanation, the measurement error model does not; at best, it forces us to imagine a group of forecasters who apply the same procedures to the same data but with numerous small mistakes.

3 Model Estimation

This section describes methodology for estimating the *information structure* $\mathbf{\Sigma}$. Unfortunately, estimating $\mathbf{\Sigma}$ in full generality based on a single prediction per forecaster is difficult. Therefore, to facilitate model estimation, the forecasters are assumed to address $K \geq 2$ related problems. For instance, the forecasters may participate in separate yet similar prediction problems or give repeated forecasts on a single recurring event. In such setups $\mathbf{\Sigma}$ is likely to remain stable across problems and hence can be estimated based on multiple predictions per forecaster. See Satopää et al. (2015) for revealed aggregation in setups with only one prediction per forecaster.

3.1 General Estimation Problem

Denote the outcome of the kth problem with Y_k and the jth forecaster's prediction for this outcome with X_{jk}. For the sake of generality, this section does not assume any particular link function but instead operates directly with the corresponding information variables, denoted with Z_{jk}. In practice, the forecasts X_{jk} can be often transformed into Z_{jk} at least approximately. This is illustrated in Section 5. If the forecasters' problem-specific information are collected into the vectors $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{Nk})'$, then the estimation of $\boldsymbol{\Sigma}$ is performed only based on $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}$. That is, neither the outcomes $\{Y_1, \ldots, Y_K\}$ nor the corresponding information variables $\{Z_{01}, \ldots, Z_{0K}\}$ are assumed known.

The estimation is performed under the covariance pattern (2). More specifically, if $\mathcal{S}_+^N$ denotes the set of $N \times N$ symmetric positive semidefinite matrices and

$$h(\mathbf{M}) := \begin{pmatrix} 1 & \text{diag}(\mathbf{M})' \\ \text{diag}(\mathbf{M}) & \mathbf{M} \end{pmatrix},$$

then the final estimate must satisfy the condition $h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}$. Intuitively, if there exists a random variable Y for which the forecasts X_j are jointly conditionally unbiased, then the corresponding $\boldsymbol{\Sigma}$ satisfies $h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}$.

An accurate yet highly sensitive estimate of $\boldsymbol{\Sigma}$, however, can lead to imprecise aggregation. To see this, recall from Section 2.2 that the conditional mean of Z_{0k} given $\mathbf{Z}_k$ is $\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_k$. This term is important because it is generally found in the revealed aggregator. Re-express the term as $\mathbf{v}'\mathbf{Z}_k$, where $\mathbf{v}$ is the solution to $\text{diag}(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}\mathbf{v}$. The rate at which the solution changes with respect to a change in $\text{diag}(\boldsymbol{\Sigma})$ depends on the condition number $\text{cond}(\boldsymbol{\Sigma}) := \lambda_{max}(\boldsymbol{\Sigma})/\lambda_{min}(\boldsymbol{\Sigma})$, i.e., the ratio between the maximum and the minimum eigenvalues of $\boldsymbol{\Sigma}$. If the condition number is very large, a small error in $\text{diag}(\boldsymbol{\Sigma})$ can cause a large error in $\mathbf{v}$. If the condition number is small, $\boldsymbol{\Sigma}$ is called *well-conditioned* and error in $\mathbf{v}$ will not be much larger than the error in $\text{diag}(\boldsymbol{\Sigma})$. Thus, to prevent error from being amplified during aggregation, the estimation procedure also requires $\text{cond}(\boldsymbol{\Sigma}) \leq \kappa$ for a given threshold $\kappa \geq 1$.

The general estimation problem can then be written as

$$\begin{aligned} &\text{minimize } f_0\left(\boldsymbol{\Sigma}, \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}\right) \\ &\text{subject to } h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}, \\ &\qquad\qquad \text{cond}(\boldsymbol{\Sigma}) \leq \kappa, \end{aligned} \tag{3}$$

where f_0 is some objective function. The feasible region defined by the two constraints is convex. Therefore, if f_0 is convex in $\boldsymbol{\Sigma}$, expression (3) is a convex optimization problem. Typically the global optimum to such a problem can be found very efficiently. Problem (3), however, involves $\binom{N+1}{2}$ variables. Therefore it can be solved efficiently with standard optimization techniques, such as the interior point methods, as long as the number of variables is not too large, say, not more than 1,000. Unfortunately, this limits the maximum problem size to about $N = 45$. The next subsections illustrate how choosing the loss function carefully permits dimension reduction and hence makes estimation possible for much larger N.

3.2 Maximum Likelihood Estimator

Under the Gaussian model the information structure, $\mathbf{\Sigma}$ is a parameter of an explicit likelihood. Therefore estimation naturally begins with the maximum likelihood approach (MLE). Unfortunately, the Gaussian likelihood is not convex in $\mathbf{\Sigma}$. Consequently, only a locally optimal solution is guaranteed with standard optimization techniques. Furthermore, it is not clear whether the dimension of this form can be reduced. Won and Kim (2006) discuss the MLE under a condition number constraint. They are able to transform the original problem with $\binom{N+1}{2}$ variables to an equivalent problem with only N variables, namely the eigenvalues of $\mathbf{\Sigma}$. This transformation, however, requires an orthogonally invariant problem. Given that the constraint $h(\mathbf{\Sigma}) \in \mathcal{S}_+^{N+1}$ is not orthogonally invariant, the same dimension-reduction technique cannot be applied. Instead, the MLE must be computed with the $\binom{N+1}{2}$ variables, making estimation slow for small N and undoable even for moderately large N. For these reasons the MLE is not discussed further in this paper.

3.3 Least Squares Estimator

Past literature has discussed many covariance estimators that can be applied efficiently to large amounts of data. Unfortunately, these estimators are generally not guaranteed to satisfy the conditions in (3). This section introduces a correctional procedure that inputs any covariance estimator $\mathbf{S}$ and then modifies it minimally such that the end result satisfies the conditions in (3). More specifically, $\mathbf{S}$ is projected onto the feasible region. This approach, sometimes known as the least squares approach (LSE), motivates a loss function that guarantees a globally optimal solution and facilitates dimension reduction. Furthermore, it does not rely on a parametric model and hence supports estimation even for future non-Gaussian partial information models.

From the computational perspective, it turns out to be more convenient to project $h(\mathbf{S})$ instead of $\mathbf{S}$. The LSE is then given by $h^{-1}(\mathbf{\Omega})$, i.e., $\mathbf{\Omega}$ without the first row and column, where $\mathbf{\Omega}$ is the solution to

$$\begin{aligned} &\text{minimize } ||\mathbf{\Omega} - h(\boldsymbol{S})||_F^2 \\ &\text{subject to } \mathbf{\Omega} \in \mathcal{S}_+^{N+1}, \\ &\qquad \text{cond}(\mathbf{\Omega}) \leq \kappa, \\ &\qquad \text{tr}(\mathbf{A}_j\mathbf{\Omega}) = b_j, \qquad (j = 1, \ldots, N+1). \end{aligned} \tag{4}$$

In this problem $\text{tr}(\cdot)$ is the trace operator and $||\cdot||_F^2$ is the squared Frobenius norm $||\mathbf{M}||_F^2 = \text{tr}(\mathbf{M}'\mathbf{M})$. Even though many other norms could be used, for the sake of simplicity, this paper only considers the Frobenius norm. The linear constraints in (4) maintain the structure induced by the function $h(\cdot)$. More specifically, if $\mathbf{e}_j$ denotes the jth standard basis vector of length $N+1$, then

$$\begin{aligned} &b_1 = 1, \mathbf{A}_1 = \mathbf{e}_1\mathbf{e}_1', \text{ and} \\ &b_j = 0, \mathbf{A}_j = \mathbf{e}_j\mathbf{e}_j' - 0.5(\mathbf{e}_1\mathbf{e}_j' + \mathbf{e}_j\mathbf{e}_1') \text{ for } j = 2, \ldots, N+1. \end{aligned}$$

Any $\mathbf{\Omega}$ that satisfies the other two conditions gives $\mathbf{\Sigma} = h^{-1}(\mathbf{\Omega})$ that also satisfies them. This follows from the fact that $\mathbf{\Sigma}$ is a principal sub-matrix of $\mathbf{\Omega}$. Therefore $\mathbf{\Omega} \in \mathcal{S}_+^{N+1}$ implies $\mathbf{\Sigma} \in \mathcal{S}_+^N$. Furthermore, Cauchy's interlace theorem (see, e.g., Hwang 2004) states that $\lambda_{min}(\mathbf{\Omega}) \leq \lambda_{min}(\mathbf{\Sigma})$ and

$\lambda_{max}(\mathbf{\Sigma}) \leq \lambda_{max}(\mathbf{\Omega})$ such that $\text{cond}(\mathbf{\Sigma}) \leq \text{cond}(\mathbf{\Omega}) \leq \kappa$. Of course, requiring $\text{cond}(\mathbf{\Omega}) \leq \kappa$ instead of $\text{cond}(\mathbf{\Sigma}) \leq \kappa$ shrinks the region of feasible $\mathbf{\Sigma}$'s. At this point, however, the exact value of κ is arbitrary and merely serves to control $\text{cond}(\mathbf{\Sigma})$. Section 3.4 introduces a procedure for choosing κ from the data. Under such an adaptive procedure, problem (4) can be considered equivalent to directly projecting $\mathbf{S}$ onto the feasible region.

Problem (4) can be solved by expressing the feasible region as an intersection of two sets, namely

$$\mathcal{C}_{sd} = \left\{ \mathbf{\Omega} : \mathbf{\Omega} \in \mathcal{S}_+^{N+1}, \text{cond}(\mathbf{\Omega}) \leq \kappa \right\}, \text{ and}$$
$$\mathcal{C}_{lin} = \left\{ \mathbf{\Omega} : \text{tr}(\mathbf{A}_j \mathbf{\Omega}) = b_j, j = 1, \ldots, N+1 \right\}.$$

Given that both of these sets are convex, projecting onto their intersection can be computed with the Directional Alternating Projection Algorithm (Gubin et al., 1967). This method makes progress by repeatedly projecting onto the sets $\mathcal{C}_{sd}$ and $\mathcal{C}_{lin}$. Consequently, it is efficient only if projecting onto each of the individual sets is fast. Fortunately, as will be shown next, this turns out to be the case.

First, projecting an $(N+1) \times (N+1)$ symmetric matrix $\mathbf{M} = \{m_{ij}\}$ onto $\mathcal{C}_{lin}$ is a linear map. To make this more specific, let $\mathbf{m} = \text{vec}(\mathbf{M})$ be a column-wise vectorization of $\mathbf{M}$. If $\mathbf{A}$ is a matrix with the jth row equal to $\text{vec}(\mathbf{A}_j)$, the linear constraints in (4) can be expressed as $\mathbf{Am} = \boldsymbol{e}_1$. Then, the projection of $\mathbf{M}$ onto $\mathcal{C}_{lin}$ is given by $\text{vec}^{-1}(\mathbf{m} + \mathbf{A}'(\mathbf{AA}')^{-1}(\boldsymbol{e}_1 - \mathbf{Am}))$. This expression simplifies significantly by close inspection. In particular, the projection is equivalent to setting $m_{11} = 1$ and for $j \geq 2$ replacing m_{j1}, m_{1j}, and m_{jj} by their average $(m_{jj} + m_{j1} + m_{1j})/3$. Denote this projection with the operator $\mathcal{P}_{lin}(\cdot)$.

Second, Tanaka and Nakata (2014) describe a univariate optimization problem that is almost equivalent to projecting $\mathbf{M}$ onto $\mathcal{C}_{sd}$. The only difference is that their solution set also includes the zero-matrix $\mathbf{0}$. Assuming that such a limiting case can be safely handled in the implementation, their approach offers a fast projection onto $\mathcal{C}_{sd}$ even for a moderately large N. To describe this approach, consider the spectral decomposition $\mathbf{M} = \mathbf{Q}\text{Diag}(l_1, \ldots, l_{N+1})\mathbf{Q}'$ and the univariate function

$$g(\mu) = \sum_{i=1}^{N+1} \left[(\mu - l_i)_+^2 + (l_i - \kappa\mu)_+^2 \right],$$

where $\text{Diag}(\mathbf{x})$ is a diagonal matrix with diagonal $\mathbf{x}$ and $(\cdot)_+$ is the positive part operator. The function $g(\mu)$ can be minimized very efficiently by solving a series of smaller convex problems, each with a closed form solution. The result is a binary-search-like procedure described by Algorithm 2 in the Appendix. If $\mu^* = \arg\min_{\mu \geq 0} g(\mu)$ and

$$\lambda_j^* := \begin{cases} \mu^* & \text{if } l_j \leq \mu^* \\ \kappa\mu^* & \text{if } \kappa\mu^* \leq l_j \\ l_j & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, N+1$, then $\mathbf{Q}\text{Diag}(\lambda_1^*, \ldots, \lambda_{N+1}^*)\mathbf{Q}$ is the projection of $\mathbf{M}$ onto $\mathcal{C}_{sd}$. Call this projection $\mathcal{P}_{sd}(\cdot : \kappa)$.

Require: Unconstrained estimator $\mathbf{S}$, stopping criterion $\epsilon > 0$, and an upper bound on the condition number $\kappa \geq 1$.

procedure DIRECTIONAL ALTERNATING PROJECTION ALGORITHM
  $\boldsymbol{\Omega}_A \leftarrow h(\boldsymbol{S})$
  repeat
    $\boldsymbol{\Omega}_B \leftarrow \mathcal{P}_{lin}(\boldsymbol{\Omega}_A)$
    $\boldsymbol{\Omega}_C \leftarrow \mathcal{P}_{sd}(\boldsymbol{\Omega}_B : \kappa)$
    $\boldsymbol{\Omega}_D \leftarrow \mathcal{P}_{lin}(\boldsymbol{\Omega}_C)$
    $\Delta \leftarrow ||\boldsymbol{\Omega}_B - \boldsymbol{\Omega}_C||_F^2 / \text{tr}\left[(\boldsymbol{\Omega}_B - \boldsymbol{\Omega}_D)'(\boldsymbol{\Omega}_B - \boldsymbol{\Omega}_C)\right]$
    $\boldsymbol{\Omega}_A \leftarrow \boldsymbol{\Omega}_B + \Delta(\boldsymbol{\Omega}_D - \boldsymbol{\Omega}_B)$
  until $\max\left\{(\boldsymbol{\Omega}_D - \boldsymbol{\Omega}_C)_{ij}^2\right\} < \epsilon$
  return $h^{-1}(\boldsymbol{\Omega}_C)$
end procedure

Algorithm 1: This procedure projects the point $h(\boldsymbol{S})$ onto the intersection $\mathcal{C}_{sd} \cap \mathcal{C}_{lin}$. Denote the projection with $\mathcal{P}_{LSE}(\cdot : \kappa)$. Throughout the paper, the stopping criterion is fixed at $\epsilon = 10^{-5}$.

Algorithm 1 uses these projections to solve (4). Each iteration projects twice on one set and once on the other set. The general form of the algorithm, however, does not specify which projection should be called twice. Therefore, given that $\mathcal{P}_{sd}(\cdot : \kappa)$ takes longer to run than $\mathcal{P}_{lin}(\cdot)$, it is beneficial to choose to call $\mathcal{P}_{lin}(\cdot)$ twice. The complexity of each iteration is determined largely by the spectral decomposition which is fairly fast for moderately large N. Overall time to convergence, of course, depends on the choice of the stopping criterion. Many intuitive criteria are possible. Given that $\boldsymbol{\Omega}_D \in \mathcal{C}_{lin}$ and $\boldsymbol{\Omega}_C \in \mathcal{C}_{sd}$, the stopping criterion $\max\{(\boldsymbol{\Omega}_D - \boldsymbol{\Omega}_C)_{ij}^2\} < \epsilon$ suggests that the return value is in $\mathcal{C}_{sd}$ and close to $\mathcal{C}_{lin}$ in every direction. Based on our experience, the algorithm converges quite quickly. For instance, our implementation in C++ generally solves (4) for $\epsilon = 10^{-5}$ and $N = 100$ in less than a second on a 1.7 GHz Intel Core i5 computer. For the remainder of the paper, projecting $\mathbf{S}$ onto the feasible region is denoted with the operator $\mathcal{P}_{LSE}(\cdot : \kappa)$.

3.4 Selecting the value of κ

Algorithm 1 requires the specification of the threshold κ. This subsection describes two approaches to choosing κ under the Gaussian model. The general principles, however, extend directly to any partial information model.

Won and Kim (2006) propose choosing the value of κ that maximizes the expected predictive log-likelihood. That is, if $\mathbf{Z}_1, \ldots, \mathbf{Z}_K, \tilde{\mathbf{Z}} \overset{i.i.d.}{\sim} \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Sigma})$, they recommend

$$\kappa = \underset{\nu \geq 1}{\arg\max}\, \mathbb{E}\left\{\mathbb{E}_{\tilde{\mathbf{Z}}}\left[\ell(\tilde{\mathbf{Z}}, \mathcal{P}_{LSE}(\mathbf{S} : \nu))\right]\right\},$$

where $\ell(\cdot)$ denotes the log-likelihood and $\mathbf{S}$ is the sample covariance matrix computed only based on $\mathbf{Z}_1, \ldots, \mathbf{Z}_K$. They approximate the expected predictive log-likelihood with cross validation. This par-

titions the data $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_K)'$ into R subsets such that $\mathbf{Z} = \left(\mathbf{Z}_{(1)}, \ldots, \mathbf{Z}_{(R)}\right)'$. During the rth iteration, $\mathbf{Z}_{(r)}$ functions as $\tilde{\mathbf{Z}}$ and the remaining $R-1$ subsets are used for computing the sample covariance matrix, denoted with $\mathbf{S}_{(r)}$. The chosen value of κ maximizes the out-of-sample log-likelihood:

$$\kappa_{out} = \underset{\nu \geq 1}{\arg\max} \sum_{r=1}^{R} \ell\left(\mathbf{Z}_{(r)}, \mathcal{P}_{LSE}(\mathbf{S}_{(r)} : \nu)\right). \tag{5}$$

Denote the final estimate with $\mathbf{\Sigma}_{out} := \mathcal{P}_{LSE}(\mathbf{S} : \kappa_{out})$. Unfortunately, this approach suffers from several drawbacks: a) it introduces a new tuning parameter R; b) it is computationally expensive as $\mathcal{P}_{LSE}(\mathbf{S}_{(r)} : \nu)$ must be called R times for each candidate value ν; c) it can be unstable under small K, which is particularly unfortunate because the number of responses per forecaster in prediction polls is often kept low for the sake of avoiding responder fatigue; and d) if the data is very sparse, finding an appropriate partition of the data may be difficult or require a non-random procedure.

To address these drawbacks, the current paper introduces a new selection procedure called *conditional validation*. This is an in-sample approach that can be used for choosing any tuning parameter under the partial information framework. To motivate, recall that the revealed aggregator X'' uses $\mathbf{\Sigma}$ to regress Z_0 on the rest of the Z_j's. Of course, the accuracy of this prediction cannot be known until the actual outcome is observed. However, apart from being unobserved, the variable Z_0 is theoretically no different to the other Z_j's. Therefore, for $j = 1, \ldots, N$, it is natural to let each of the Z_j's in turn play the role of Z_0, predict its value based on Z_i for $i \neq j$, and choose the value of ν that yields the best overall accuracy. Even though many accuracy measures could be chosen, this paper uses the conditional log-likelihood. Therefore, if $\mathbf{Z}_j^* = (Z_{j1}, \ldots, Z_{jK})'$ collects the jth forecaster's information of the K events, the chosen value of κ is

$$\kappa_{cov} = \underset{\nu \geq 1}{\arg\max} \sum_{j=1}^{N} \ell\left(\mathbf{Z}_j^*, \mathcal{P}_{LSE}(\mathbf{S} : \nu) \middle| \mathbf{Z}_i^* \text{ for } i \neq j\right), \tag{6}$$

where the log-likelihood is now conditional on $\mathbf{Z}_i^*$'s for $i \neq j$ and $\mathbf{S}$ is the sample covariance matrix computed based on all the forecasts $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_N^*$. Denote the final estimate with $\mathbf{\Sigma}_{cov} := \mathcal{P}_{LSE}(\mathbf{S} : \kappa_{cov})$.

The idea in conditional validation is similar to cross-validation but, instead of predicting across rows (observations), the prediction is performed across columns (variables). This is likely to be more appropriate in prediction polling-like setups that typically involve a large number of forecasters (large N) attending relatively few problems (small K). In contrast to cross-validation, however, conditional validation requires only one call to $\mathcal{P}_{LSE}(\cdot : \nu)$ per candidate value ν. Furthermore, it has no tuning parameters and remains more stable when K is small (as will be illustrated in the next section). Unfortunately, both optimization problems (5) and (6) are non-convex in ν. However, as was mentioned before, Algorithm 1 is fast for moderately sized N. Therefore κ can be chosen efficiently (possibly in parallel on multicore machines) over a grid of candidate values.

4 Synthetic Data

Model evaluation naturally begins with synthetic data generated directly from the multivariate Gaussian distribution (2). This provides insight into the behavior of the estimation procedure and also introduces the simplest instance of the Gaussian model.

> **Model Instance.** The link function $g(\cdot)$ is the identity. Consequently, the target quantity is $Y_k = g(Z_{0k}) = Z_{0k}$, and the forecasts are $X_{jk} = \mathbb{E}(Y_k|Z_{jk}) = Z_{jk}$ for $j = 1, \ldots, N$ and $k = 1, \ldots, K$. The revealed aggregator for problem k is $X''_k = \text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{X}_k$, where $\mathbf{X}_k = (X_{1k}, \ldots, X_{Nk})'$.

Simulating forecasts from (2) requires a covariance matrix $\boldsymbol{\Sigma}$ such that $h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}$. One approach is to generate a random $N \times N$ positive definite matrix $\mathbf{M}$ from a Wishart distribution, scale it such that all diagonal entries are within $[0, 1]$, and let $\boldsymbol{\Sigma}$ be equal to the scaled version of $\mathbf{M}$ if this satisfies $h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}$. However, based on a brief simulation study that is not presented in this paper for the sake of brevity, the rate at which $\mathbf{M}$ is accepted decreases in N and is very close to zero already for $N > 5$. Therefore this section adopts a different approach that samples $\boldsymbol{\Sigma}$ with full acceptance rate but only within a subset of all information structures. In particular, the idea is to pick $\delta_j \in [0, 1]$ and set $\rho_{ij} = \delta_i\delta_j$ for all $i \neq j$. If $\boldsymbol{\Sigma}$ is the resulting information structure, then $\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{\Sigma})\text{diag}(\boldsymbol{\Sigma})' = \text{Diag}((\delta_1 - \delta_1^2, \ldots, \delta_N - \delta_N^2)') \in \mathcal{S}_+^N$, which by the Schur complement is equivalent to having $h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}$. This procedure generates a broad range of information structures, both well- and ill-conditioned. In fact, if any $\delta_j = 0$, the resulting $\boldsymbol{\Sigma}$ is singular. To avoid this, the δ_j's are generated uniformly at random from the somewhat more realistic interval $[0.1, 0.9]$. The resulting information structure is used to draw $(Z_{0k}, Z_{1k}, \ldots, Z_{Nk})' \overset{i.i.d.}{\sim} \mathcal{N}_{N+1}(\mathbf{0}, h(\boldsymbol{\Sigma}))$ for each problem $k = 1, \ldots, K$. These forecasts are then aggregated in the following ways:

1. The revealed aggregator X''_k with the sample covariance matrix $\mathbf{S}_Z$. Given that $\mathbf{S}_Z$ is singular when $K < N$, its inverse is computed with the (Moore-Penrose) generalized inverse.

2. The revealed aggregator X''_k with $\boldsymbol{\Sigma}_{cov}$. The condition number constraint κ_{cov} is found over a grid of 100 values between 10 and $1,000$. Denote this aggregator with X''_{cov}.

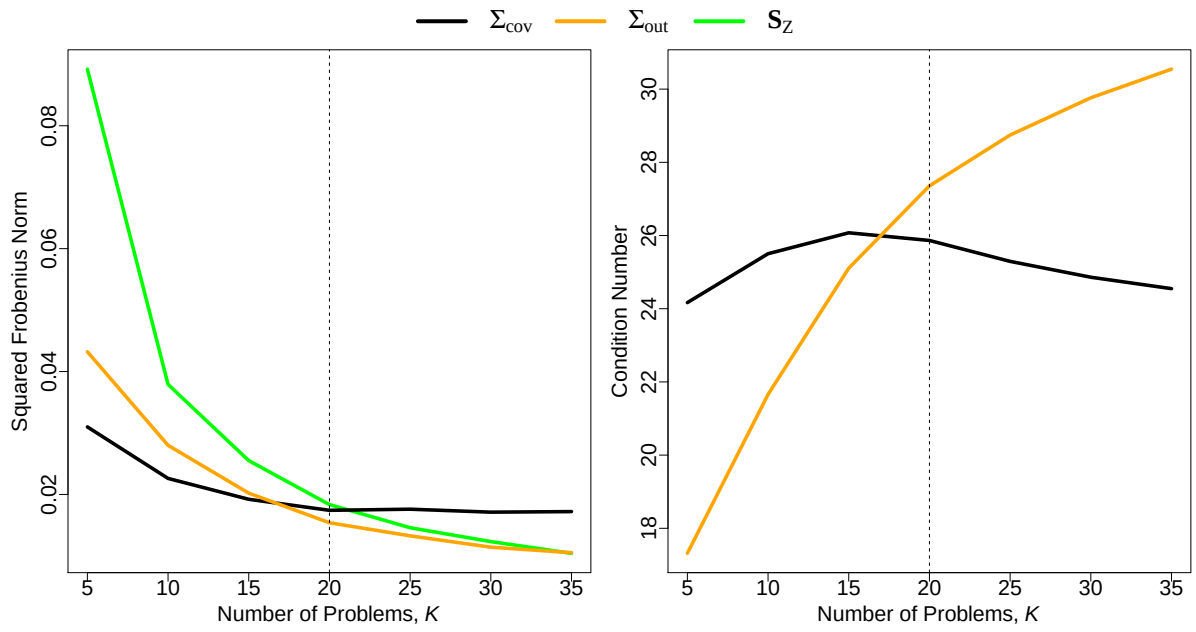
3. The revealed aggregator X''_k with $\boldsymbol{\Sigma}_{out}$. The cross-validation uses five folds and κ is found using the same grid as in X''_{cov}. Denote this aggregator with X''_{out}.

4. The revealed aggregator X''_k with the true information structure $\boldsymbol{\Sigma}$. Denote this aggregator with X''_{true}.

5. The average forecast.

6. The median forecast.

Aggregators 1 to 4 are based on partial information while aggregators 5 and 6 stem from measurement error. The overall process is repeated $5,000$ times under different values of K and N, each ranging



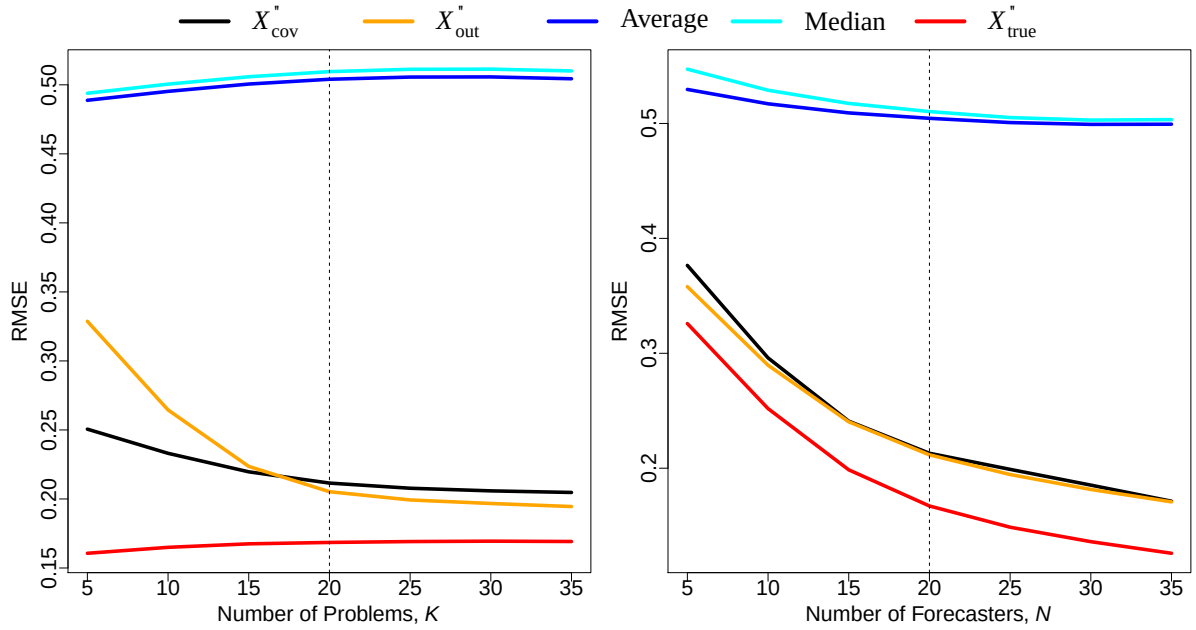
(a) Estimation accuracy of $\mathbf{\Sigma}$

(b) Average condition numbers of the estimates

Figure 3: Estimation of the information structure and the average condition numbers of the estimates. Both are important for accurate prediction of Y_k. The vertical dashed lines represents the number of forecasters fixed at $N = 20$.

from 5 to 35 with constant increments of 5. In the end the results are averaged within each combination of N and K.

Recall that accurate revealed aggregation stems both from a precise estimate of $\mathbf{\Sigma}$ and a low condition number. This allows different strategies for achieving good accuracy. In fact, it turns out that the two selection procedures discussed in Section 3.4 make slightly different tradeoffs. This is illustrated in Figure 3 that varies K between 5 and 35 but keeps N fixed at 20. More specifically, Figure 3a examines how $\mathbf{\Sigma}_{cov}$, $\mathbf{\Sigma}_{out}$, and $\mathbf{S}_Z$ capture the true $\mathbf{\Sigma}$. Even though all estimators become more accurate as K grows, $\mathbf{\Sigma}_{out}$ and $\mathbf{S}_Z$ improve at a higher rate than $\mathbf{\Sigma}_{cov}$. In fact, if $K > N$, $\mathbf{S}_Z$ and $\mathbf{\Sigma}_{out}$ perform better than $\mathbf{\Sigma}_{cov}$. On the other hand, if $K < N$, $\mathbf{\Sigma}_{cov}$ is more accurate than the other two. Figure 3b presents the corresponding (average) condition numbers of these estimates. For the sake of keeping the scale manageable, cond$(\mathbf{S}_Z)$ has been omitted. Notice that cond$(\mathbf{\Sigma}_{out})$ increases while cond$(\mathbf{\Sigma}_{cov})$ generally decreases as K grows larger. In fact, when $K > N$, cond$(\mathbf{\Sigma}_{cov})$ is smaller than cond$(\mathbf{\Sigma}_{out})$. From the prediction perspective, this makes conditional-validation more forgiving towards error in the estimated $\mathbf{\Sigma}$. Therefore, while $\mathbf{\Sigma}_{cov}$ incorporates the actual prediction process and looks for a fine balance between a precise estimate of $\mathbf{\Sigma}$ and a low condition number, $\mathbf{\Sigma}_{out}$ is unaware of the details of revealed aggregation and hence simply focuses on estimating $\mathbf{\Sigma}$ as accurately as possible.



(a) Prediction accuracy under different K but fixed $N = 20$.

(b) Prediction accuracy under different N but fixed $K = 20$.

Figure 4: The accuracy to predict Y_k under different values of N and K. The aggregator X''_{true} assumes knowledge of the true information matrix and hence represents optimal accuracy.

These two strategies lead to slightly different predictive behavior as is illustrated in Figure 4. This plot shows the root-mean-squared-errors (RMSE) of the competing aggregators in predicting Y_k. Figure 4a varies K but fixes $N = 20$. Figure 4b, on the other hand, varies N but fixes $K = 20$. Given that $Y_k = Z_{0k} \sim \mathcal{N}(0, 1)$, the RMSE of the prior mean $\mathbb{E}(\sqrt{(Y_k - 0)^2}) = \mathbb{E}(|Y_k|) = \sqrt{2/\pi} \approx 0.8$ can be considered as the upper bound in prediction error. On the other end, the aggregator X''_{true} is optimal and hence provides the corresponding lower bound. The revealed aggregator X'' with $\mathbf{S}_Z$ typically received a loss much larger than 0.8 and is therefore not included in the figure. Overall, the average and median perform very similarly, with RMSE around 0.5. They both show slight improvement as N increases. In all cases, however, their RMSE is uniformly well above that of the revealed aggregators, suggesting that measurement-error aggregators are a poor choice when forecasts truly arise from a partial information model. The revealed aggregators X''_{cov} and X''_{out} perform very similarly when $K \geq 15$. They collect information and appear to improve at the optimal rate as N increases. This can be seen in the way the performance gap from X_{true} to X''_{out} and X''_{cov} remains approximately constant in Figure 4b. They both, however, approach the optimal as K grows larger. When K is small, say less than 15, conditional validation is more robust and clearly yields better results. Given that conditional validation is also computationally much less demanding, only X''_{cov} is considered in the following section that applies the

Gaussian model to real-world forecasting data.

5 Applications

5.1 Probability Forecasts of Binary Outcomes

5.1.1 Dataset

During the second year of the Good Judgment Project (GJP) the forecasters made probability estimates for 78 events, each with two possible outcomes. One of these events was illustrated in Figure 1. Each prediction problem had a timeframe, defined as the number of days between the first day of forecasting and the anticipated resolution day. These timeframes varied largely among problems, ranging from 12 days to 519 days with a mean of 185.4 days. During each timeframe the forecasters were allowed to update their predictions as frequently as they liked. The forecasters knew that their estimates would be assessed for accuracy using the quadratic loss. This is a revealing loss function that incentivized the forecasters to report their true beliefs instead of attempting to game the system. In addition to receiving \$150 for meeting minimum participation requirements that did not depend on prediction accuracy, the forecasters received status rewards for their performance via leader-boards displaying the losses for the best 20 forecasters. Depending on the details of the reward structure, such a competition for rank may eliminate the truth-revelation property of revealing loss functions (see, e.g., Lichtendahl Jr and Winkler 2007).

This data collection raises several issues: First, given that the current paper does not focus on modeling dynamic data, only forecasts made within some common time interval should be considered. The final outcome is typically the most uncertain in the beginning but becomes fully certain by the resolution date. Second, not all forecasters made predictions for all the events. Furthermore, the forecasters generally updated their forecasts infrequently, resulting into a very sparse dataset. Such high sparsity can cause problems in computing the initial unconstrained estimator $\mathbf{S}$. Evaluating different techniques to handle missing values, however, is well outside the scope of this paper. Therefore, to somewhat alleviate the effect of missing values, only the hundred most active forecasters are considered. This makes sufficient overlap highly likely but, unfortunately, still not guaranteed.

All these considerations lead to a parallel analysis of three scenarios: High Uncertainty (HU), Medium Uncertainty (MU), and Low Uncertainty (LU). Important differences are summarized in Table 1. Each scenario considers the forecasters' most recent prediction within a different time interval. For instance, LU only includes each forecaster's most recent forecast during $30 - 60$ days before the anticipated resolution day. The resulting dataset has 60 events of which 13 occurred. In the corresponding 60×100 table of forecasts, around 42 % of the values are missing. The other two scenarios are defined similarly.

Table 1: Summary of the three time intervals analyzed. Each scenario considers the forecasters' most recent forecasts within the given time interval. The value in the parentheses represent the number of events occurred. The final column shows the percent of missing forecasts.

Scenario	Time Interval	# of Events	Missing (%)
High Uncertainty (HU)	$90-120$	$49\ (10)$	51
Medium Uncertainty (MU)	$60-90$	$56\ (14)$	46
Low Uncertainty (LU)	$30-60$	$60\ (13)$	42

5.1.2 Model Specification and Aggregation

The fact that $Y_k \in \{0,1\}$ is binary and lies on the boundary of the support of the forecasts $X_{jk} \in [0,1]$ makes the model instance slightly more involved. This instance resembles in many ways the latent variable version of a standard probit model.

Model Instance. Identify the kth event with $Y_k \in \{0,1\}$. These outcomes link to the information variables via the following function:

$$Y_k = g(Z_{0k}) = \begin{cases} 1 & \text{if } Z_{0k} > t_k \\ 0 & \text{otherwise,} \end{cases}$$

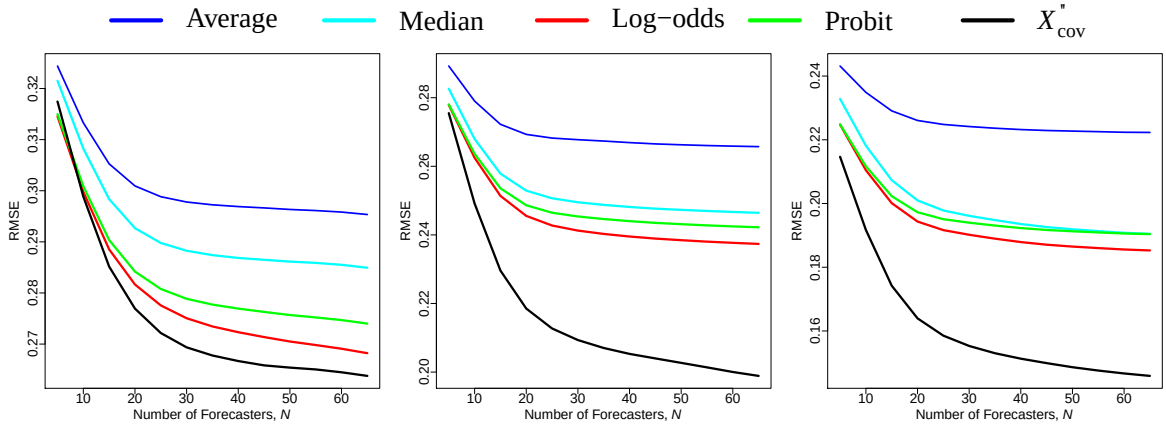
where $t_k \in \mathbb{R}$ is some threshold value. Therefore the link function $g(\cdot)$ is simply the indicator function $\mathbf{1}_{A_k}$ of the event $A_k = \{Z_{0k} > t_k\}$. The prior probability of the kth event is $\mathbb{P}(Y_k = 1) = \Phi(-t_k)$, where $\Phi(\cdot)$ is the CDF of a standard Gaussian distribution. Given that the thresholds are allowed to vary among the events, each event has its own prior. The corresponding probability forecasts $X_{jk} \in [0,1]$ are

$$X_{jk} = \mathbb{E}(Y_k | Z_{jk}) = \Phi\left(\frac{Z_{jk} - t_k}{\sqrt{1-\delta_j}}\right).$$

In a similar manner, the revealed aggregator $X_k'' \in [0,1]$ for event k is

$$X_k'' = \mathbb{E}(Y_k | \mathbf{Z}_k) = \Phi\left(\frac{\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_k - t_k}{\sqrt{1 - \text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\text{diag}(\boldsymbol{\Sigma})}}\right). \tag{7}$$

All the parameters of this model can be estimated from the data. The first step is to specify a version of the unconstrained estimate $\mathbf{S}$. If the t_k's do not change much, a reasonable and simple estimate is obtained by transforming the sample covariance matrix $\mathbf{S}_P$ of the probit scores $P_{jk} := \Phi^{-1}(X_{jk})$. More specifically, if $\mathbf{D} := \text{Diag}(\mathbf{d})\text{Diag}(\mathbf{1}+\mathbf{d})^{-1}$, where $\mathbf{d} = \text{diag}(\mathbf{S}_P)$, then an unconstrained estimator of $\boldsymbol{\Sigma}$ is given by $\mathbf{S} = (\mathbf{I}_N - \mathbf{D})^{1/2}\mathbf{S}_P(\mathbf{I}_N - \mathbf{D})^{1/2}$. Recall that the GJP data holds many missing values. This is handled by estimating each pairwise covariance in $\mathbf{S}_P$ separately based on all the events



(a) High Uncertainty (HU) (b) Medium Uncertainty (MU) (c) Low Uncertainty (LU)

Figure 5: Average prediction accuracy over the 1,000 sub-samplings of the forecasters. See Table 1 for descriptions of the different scenarios.

for which both forecasters made predictions. Next, compute $\boldsymbol{\Sigma}_{cov}$, where κ_{cov} is chosen over a grid of 100 candidate values between 10 and $1,000$. Finally, the threshold t_k can be estimated by letting $\mathbf{P}_k = (P_{1k}, \ldots, P_{Nk})'$, observing that $-\text{Diag}(\mathbf{1} - \text{diag}(\boldsymbol{\Sigma}))^{1/2}\mathbf{P}_k \sim \mathcal{N}_N(t_k\mathbf{1}_N, \boldsymbol{\Sigma})$, and computing the precision-weighted average:

$$\hat{t}_k = -\frac{\mathbf{P}_k'\text{Diag}(\mathbf{1} - \text{diag}(\boldsymbol{\Sigma}_{cov}))^{1/2}\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}}.$$

If $\mathbf{P}_k$ has missing values, the corresponding rows and columns of $\boldsymbol{\Sigma}_{cov}$ are dropped. Intuitively, this estimator gives more weight to the forecasters with very little information. These estimates are then plugged in to (7) to get the revealed aggregator X''_{cov}.

This aggregator is benchmarked against the state-of-the-art measurement-error aggregators, namely the average probability, median probability, average probit-score, and average log-odds. To avoid infinite log-odds and probit scores, extreme forecasts $X_{jk} = 0$ and 1 were censored to $X_{jk} = 0.001$ and 0.999, respectively. The results remain insensitive to the exact choice of censoring as long as this is done in a reasonable manner to keep the extreme probabilities from becoming highly influential in the logit- or probit-space. Similarly to before, the accuracy of the aggregates is measured with the RMSE. Instead of considering all the forecasts at once, the aggregators are evaluated under different N via repeated subsampling of the 100 most active forecasters; that is, choose N forecasters uniformly at random, aggregate their forecasts, and compute the RMSE. This is repeated 1,000 times with $N = 5, 10, \ldots, 65$ forecasters. In the rare occasion where no pairwise overlap is available between one or more pairs of the selected forecasters, the subsampling is repeated until all pairs have at least one problem in common.

Figure 5 shows the average RMSE's under the three scenarios described in Table 1. Notice that the scores improve uniformly from HU to LU. This reflects the decreasing level of uncertainty. In all

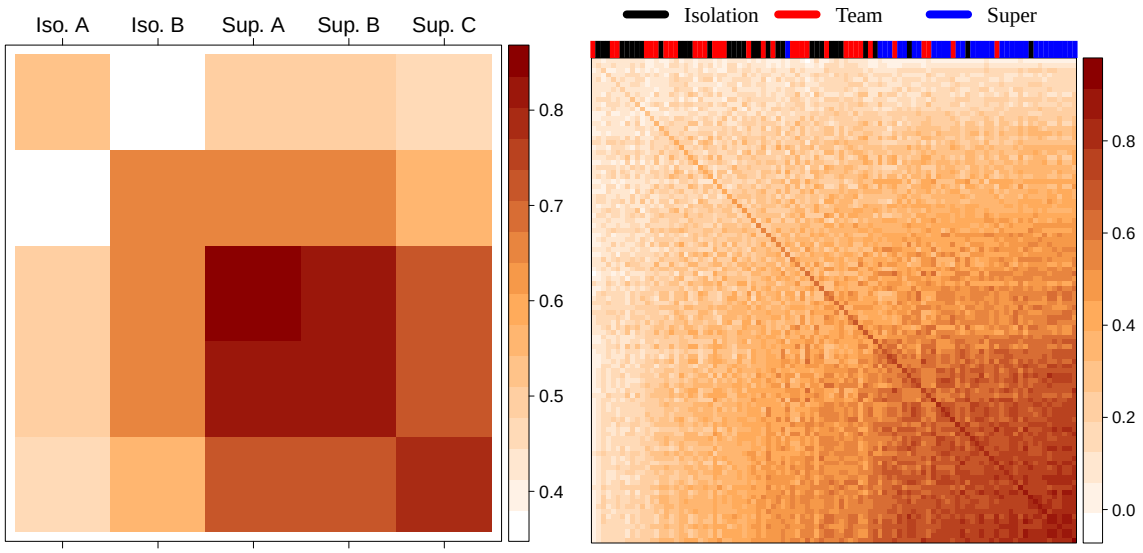
the figures the measurement-error aggregators rank in the typical order (from worst to best): average probability, median probability, average probit, and average log-odds. Despite the level of uncertainty, the revealed aggregator X''_{cov} outperforms the averaging aggregators as long as $K \geq 10$. The relative advantage, however, increases from HU to LU. This trend can be explained by several reasons: First, as can be seen in Table 1, the amount of data increases from HU to LU. This yields a better estimate of $\mathbf{\Sigma}$ and hence more accurate revealed aggregation. Second, under HU the events are still inherently very uncertain. Consequently, the forecasters are unlikely to hold much useful information as a group. Under such low information diversity, measurement-error aggregators generally perform relatively well (Satopää et al. 2015). In the contrary, under MU the events have lost a part of their inherent uncertainty, allowing some forecasters to possess useful private information. These individuals are then prioritized by X''_{cov} while the averaging-aggregators continue treating all forecasts equally. Third, the forecasters are more likely to be conditionally unbiased under MU and LU than under HU.

5.1.3 Information Diversity

This subsection examines how closely the estimated $\mathbf{\Sigma}$ reflects prior knowledge about the forecasters' information structure. In particular, the GJP assigned the forecasters to make predictions either in isolation or in teams. Furthermore, after the first year of the tournament, the top 1% forecasters were elected to the elite group of "super-forecasters". These super-forecasters then worked in exclusive teams to make highly accurate predictions on the same events as the rest of the forecasters. Many of these characteristics directly suggest a level of information overlap. For instance, super-forecasters can be expected to have the highest δ_j's and forecasters in the same team should have a relatively high ρ_{ij}.

For the sake of brevity, only the LU scenario is analyzed as this is where X''_{cov} presented the highest relative improvement. The associated 100 forecasters involve 36 individuals predicting in isolation, 33 forecasting team-members (across 24 teams), and 31 super-forecasters (across 5 teams). Figure 6a displays $\mathbf{\Sigma}_{cov}$ for the five most active forecasters, involving two forecasters working in isolation (Iso. A and B) and three super-forecasters (Sup. A, B, and C). Only the super-forecasters A and B are in the same team and hence have a relatively high information overlap. Overall, the three super-forecasters are more informed than the non-super-forecasters. Such a high level of information unavoidably leads to higher information overlap with the rest of the forecasters.

This pattern generalizes to the entire group of forecasters. To illustrate, Figure 6b displays $\mathbf{\Sigma}_{cov}$ for all the 100 forecasters. The information structure has been ordered with respect to the diagonal such that the more informed forecasters appear on the right. Furthermore, a colored rug has been appended on the top. This rug shows whether each forecaster worked in isolation, in a non-super-forecaster team, or in a super-forecaster team. Observe that the super-forecasters are mostly situated on the right among the most informed forecasters. The average estimated δ_j among the super-forecaster is 0.80. On the other hand, the average estimated δ_j among the individuals working isolation or in non-super-forecaster teams are 0.47 and 0.50, respectively. Therefore working in a team makes the forecasters' predictions, on average, slightly more informed. In general, a plot like the one in Figure 6b is useful for assessing the level of information diversity among the forecasters: the further away the observed plot is from a monochromatic plot, the higher the information diversity. Therefore the colorful Figure 6b suggests that



(a) The five most active forecasters

(b) All 100 forecasters

Figure 6: The estimated information structure $\boldsymbol{\Sigma}$ under the LU scenario. Each forecaster worked either in isolation, in a non-super-forecaster team, or in a super-forecaster team. The super-forecasters generally have more information than the forecasters working in isolation.

the GJP forecasters have high information diversity.

5.2 Point Forecasts of Continuous Outcomes

5.2.1 Dataset

Moore and Klein (2008) hired 415 undergraduates from Carnegie Mellon University to guess the weights of 20 people based on a series of pictures. These forecasts were illustrated in Figure 2. The target people were between 7 and 62 years old and had weights ranging from 61 to 230 pounds, with a mean of 157.6 pounds. All the students were shown the same pictures and hence given the exact same information. Therefore any information diversity arises purely from the participants' decisions to use different subsets of the same information. Consequently, information diversity is likely to be low compared to Section 5.1 where diversity also stemmed from differences in the information available to the forecasters.

Unlike in Section 5.1, the Gaussian model can be applied almost directly to the data. Only the effect of extreme values was reduced via a 90% Winsorization (Hastings et al., 1947). This handled some obvious outliers. For instance, the dataset contains a few estimates above 1000 pounds and some as low

as 10 pounds. Winsorization generally improved the performance of all the competing aggregators.

5.2.2 Model Specification and Aggregation

Model Instance. Suppose Y_k and X_{jk} are real-valued. If the proper non-informative prior distribution of Y_k is $\mathcal{N}(\mu_{0k}, \sigma_0^2)$, then $Y_k = g(Z_{0k}) = Z_{0k}\sigma_0 + \mu_{0k}$. Consequently, $X_{jk} = \mathbb{E}(Y|Z_{jk}) = Z_{jk}\sigma_0 + \mu_{0k}$ for all $j = 1, \ldots, N$. Therefore $X_j \sim \mathcal{N}(\mu_{0k}, \sigma_j^2)$ for some $\sigma_j^2 \leq \sigma_0^2$. If $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{Nk})'$, then the revealed aggregator for the kth problem is

$$X_k'' = \mathbb{E}\left(Y_k|\mathbf{Z}_k\right) = \text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_k\sigma_0 + \mu_{0k}. \tag{8}$$

Under this model the prior distribution of Y_k is specified by μ_{0k} and σ_0^2. Given that $\mathbb{E}(X_{jk}) = \mu_{0k}$ for all $j = 1, \ldots, N$, the sample average $\hat{\mu}_{0k} = \sum_{j=1}^{N} X_{jk}/N$ provides an initial estimate of μ_{0k}. Theoretically the revealed aggregator (8) does not depend on σ_0^2; that is, the conditional mean is completely defined by how much the forecasters know relative to each other – not by how much they know in absolute terms. Therefore any choice of σ_0^2 should lead to the same aggregator. Unfortunately, the projection procedure discussed in Section 3 is sensitive to this choice. The value of σ_0^2, however, can be estimated by assuming a distribution for the σ_j^2's. More specifically, let σ_j^2 be i.i.d. on the interval $[0, \sigma_0^2]$ and use the resulting likelihood to estimate σ_0^2. For instance, a non-informative choice is to assume $\sigma_j^2 \overset{i.i.d.}{\sim} \mathcal{U}(0, \sigma_0^2)$, which leads to the maximum likelihood estimator $\max\{\sigma_j^2\}$. This has a downward bias that can be corrected by a multiplicative factor of $(N+1)/N$. Therefore, replacing σ_j^2 with the sample variance $s_j = \sum_{k=1}^{K}(X_{jk} - \hat{\mu}_{0k})^2/(K-1)$ gives the final estimate $\hat{\sigma}_0^2 = (N+1)/N \max\{s_j\}$. Using these estimates, the X_{jk}'s can be transformed into the Z_{jk}'s whose sample covariance matrix $\mathbf{S}$ provides the unconstrained estimator for the projection algorithm. The value of κ_{cov} is chosen over a grid of 10 values between 10 and $10,000$. Once $\boldsymbol{\Sigma}_{cov}$ has been found, the prior means are updated with the precision-weighted averages $\hat{\mu}_{0k} = (\mathbf{X}_k'\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}_N)/(\mathbf{1}_N'\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}_N)$. In the end, all these estimates are plugged in (8) to get the revealed aggregator X_{cov}''.

This aggregator is compared against the average, the median, and the average of the median and average (AMA), which was proposed by Lobo and Yao (2010) as a good heuristic for aggregating human judgments. Overall accuracy is measured with the RMSE averaged over $10,000$ sub-samplings of the 416 participants. That is, each iteration chooses N participants uniformly at random, aggregates their forecasts, and computes the RMSE. The size of the sub-samples is varied between 10 and 100 with increments of 10. These scores are presented in Figure 7. Similarly to the synthetic point forecasts in Section 4, the average outperforms the median across all N. The performance of AMA falls between that of average and median, reflecting its nature as a compromise of the two. The revealed aggregator X_{cov}'' is the most accurate once $N > 10$. The relatively worse performance at $N = 10$ suggests that 10 observations is not enough to estimate $\hat{\mu}_{0k}$ accurately. As N approaches 100, however, X_{cov}'' collects information efficiently and increases the performance advantage against the other aggregators. In particular, compared to the lowest RMSE at $N = 10$, namely 21.18 achieved by the average, X_{cov}'' gains an absolute improvement of 0.74 as N reaches 100. This is more than double the corresponding improvement of the average forecast.

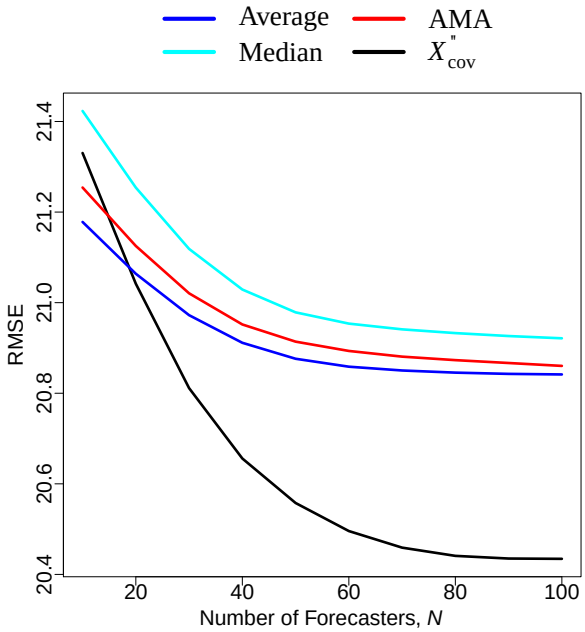


Figure 7: Accuracy of the competing aggregators

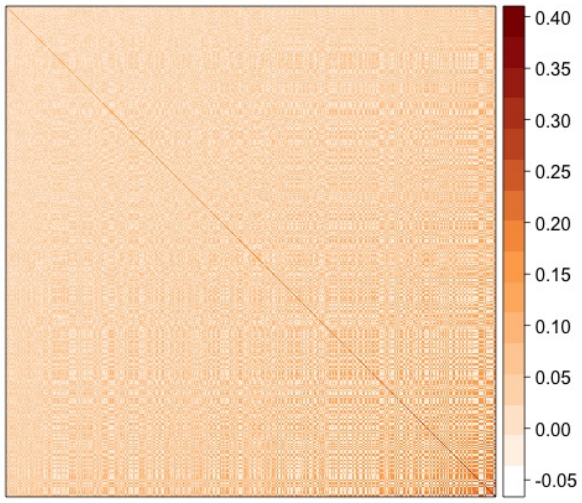


Figure 8: Information diversity among the 416 participants

Figure 8 shows $\boldsymbol{\Sigma}_{cov}$ for all the 416 forecasters. Similarly to before, the structure has been ordered such that the most knowledgeable forecasters are on the right. Overall, this plot is much more monochromatic than the one presented earlier in Figure 6b, suggesting that information diversity among the 416 students is indeed relatively lower. If there were no information diversity, i.e., all the forecasters used the same information, then averaging aggregators, such as the simple average, would perform relatively well (Satopää et al., 2015). Such a limiting case, however, is rarely encountered in practice. Often at least some information diversity is present. The results in the current section show that the revealed aggregator does not require extremely high information diversity in order to outperform the measurement-error aggregators.

6 Discussion

This paper introduced the partial information framework for modeling forecasts from different types of prediction polls. Forecast heterogeneity is assumed to stem purely from differences in the forecasters' information. This is considered more plausible on the micro-level than the historical measurement error. The partial information framework motivates and describes the forecasters' information overlap with a patterned covariance matrix (Equation 1). A correctional procedure was proposed (Algorithm 1) as a general tool for estimating these information structures. This procedure inputs any covariance

estimator and modifies it minimally such that all the associated forecasts can be regarded as (conditionally unbiased) estimates of some common target outcome. Even though the general partial information framework motivates an optimal aggregator, this is generally too abstract to be applied in practice. This paper discusses a specification within the framework, known as the Gaussian model (Section 2.2). The Gaussian model permits a closed-form solution for the optimal aggregator and utilizes a link function to model different types of forecasts and outcomes. In the common case of point forecast of a continuous outcome the optimal aggregator never (except in an unrealistic boundary case) reduces to a weighted average of the individual forecasts (Proposition 2.2). This result was illustrated on synthetic data (Section 4) and verified on real-world forecasts (Section 5.2). The model was also applied to real-world probability forecasts of binary events (Section 5.1). In each application the Gaussian model outperforms the typical measurement-error-based aggregators. This provides evidence that information diversity is the more important source of forecast heterogeneity.

Generally speaking, partial information aggregation works well because it downweights pairs or sets of forecasters that share more information and upweights ones that have unique information (or choose to attend to unique information as is the case, e.g., in Section 5.2, where forecasters made judgments based on pictures). This is very different from measurement-error aggregators that assume each forecaster to have the same information and hence consider the forecasters equally important. In real-world prediction polls, however, participants are unlikely to have equal skill and information sets. Therefore prioritizing is almost certainly called for.

The partial information framework offers both theoretical and empirical directions for future research. One theoretical avenue involves estimation of information overlap. In some cases the higher order overlaps have been found to be irrelevant to aggregation. For instance, DeGroot and Mortera (1991) show that the pairwise conditional (on the truth) distributions of the forecasts are sufficient for computing the optimal weights of a weighted average. Theoretical results on the significance or insignificance of higher order overlaps under the partial information framework would be desirable. Given that the Gaussian model can only accommodate pairwise information overlap, such a result would reveal the need of a specification that is more complex than the Gaussian model. Another theoretical direction involves a continued exploration of the connection between the revealed aggregator and the class of weighted averages. This in particular entails proving the sub-optimality of the weighted average under a more general context.

A promising empirical direction is the Bayesian approach. These techniques are very natural for fitting hierarchical models such as the ones discussed in this paper. Furthermore, 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 in a prediction poll is typically quite small, a Bayesian approach is likely to improve the quality of the final aggregate. This would involve developing a prior distribution for the information structure – a problem that seems interesting in itself. Overall, this avenue should certainly be pursued, and the results tested against other high performing aggregators.

7 Appendix A

7.1 Proof of Proposition 2.1

Proof. Denote the mean of the target outcome Y with $\mu_0 := \mathbb{E}(Y)$. Each item is then proved as follows.

i) Given that $\mathbb{E}(Y|X_j) = X_j$, the law of iterated expectation gives $\mathbb{E}(X_j) = \mathbb{E}(\mathbb{E}(Y|X_j)) = \mathbb{E}(Y)$ for all $j = 1, \ldots, N$.

ii)

$$\begin{aligned}
\text{Cov}\,(X_j, X_i) &= \mathbb{E}((X_j - \mu_0)(X_i - \mu_0)) \\
&= \mathbb{E}(X_j X_i) - \mu_0^2 \\
&= \mathbb{E}(\mathbb{E}(X_j|X_i)X_i) - \mu_0^2 \\
&= \mathbb{E}(\mathbb{E}(\mathbb{E}(Y|X_j)|X_i)X_i) - \mu_0^2 \\
&= \mathbb{E}(\mathbb{E}(Y|X_i)X_i) - \mu_0^2 && \text{(the smallest } \sigma\text{-field wins)} \\
&= \mathbb{E}(X_i^2) - \mu_0^2 \\
&= \text{Var}\,(X_i)
\end{aligned}$$

iii)

$$\begin{aligned}
\text{Var}\,(X_i) &= \text{Cov}\,(X_i, X_j) && \text{(by item ii)} \\
&= \mathbb{E}((X_i - \mu_0)(X_j - \mu_0)) \\
&\leq \mathbb{E}((X_i - \mu_0)^2)^{1/2}\mathbb{E}((X_j - \mu_0)^2)^{1/2} && \text{(by Cauchy-Schwarz' inequality)} \\
&= \sqrt{\text{Var}\,(X_i)\text{Var}\,(X_j)},
\end{aligned}$$

which then provides $\text{Var}\,(X_i) \leq \text{Var}\,(X_j)$. This inequality is tight because $X_i = X_j$ for $\mathcal{F}_i = \mathcal{F}_j$.

□

7.2 Proof of Proposition 2.2

Proof. Given that $Y \sim \mathcal{N}(\mu_0, \sigma_0^2)$, the link function is $Y = g(Z_0) = Z_0\sigma_0 + \mu_0$. The forecasts are $X_j = \mathbb{E}(Y|Z_j) = Z_j\sigma_0 + \mu_0$, which gives $Z_j = (X_j - \mu_0)/\sigma_0$. The revealed aggregator is $X'' = \mathbb{E}(Z_0|\mathbf{Z})\sigma_0 + \mu_0 = \text{diag}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{Z}\sigma_0 + \mu_0 = \text{diag}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mu_0\mathbf{1}_N) + \mu_0$. Now, suppose $X'' = \mathbf{w}'\mathbf{X}$ for some vector of weights $\mathbf{w}$. Then,

$$\begin{aligned}
&\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mu_0\mathbf{1}_N) + \mu_0 = \mathbf{w}'\mathbf{X} \\
\Leftrightarrow\ &(\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1} - \mathbf{w}')(\mathbf{X} - \mu_0\mathbf{1}_N) = 0 \qquad (9)
\end{aligned}$$

Given that $\mathbf{\Sigma} \in \mathbf{S}_{++}^N$, $\mathbf{X}$ cannot be degenerate at $\mu_0 \mathbf{1}_N$. Therefore (9) shows that $\text{diag}(\mathbf{\Sigma})' \mathbf{\Sigma}^{-1} = \mathbf{w}'$ or equivalently that $\text{diag}(\mathbf{\Sigma})' = \mathbf{w}' \mathbf{\Sigma}$. This means that a weighted average of the column (or row) elements equals the corresponding diagonal element.

Let $m = \arg\max_j \{\delta_j : j = 1, \ldots, N\}$ index the maximum variance among the forecasts and denote the mth column of $\mathbf{\Sigma}$ with $\boldsymbol{\sigma}_m$. No element in $\boldsymbol{\sigma}_m$ can be larger than δ_m. This is verified as follows:

$$\begin{aligned} |\rho_{im}| &= |\mathbb{E}((X_i - \mu_0)(X_m - \mu_0))| \\ &\leq \mathbb{E}((X_i - \mu_0)^2)^{1/2} \mathbb{E}((X_m - \mu_0)^2)^{1/2} \qquad \text{(by Cauchy-Schwarz' inequality)} \\ &= \sqrt{\delta_i \delta_m} \\ &\leq \delta_m \end{aligned}$$

for all $i \neq m$. Given that $\mathbf{w}' \boldsymbol{\sigma}_m = \delta_m$, either a) $w_m = 1$ or b) $\rho_{i,m} = \delta_m$ for all $w_i > 0$. Next it is shown that only case a) is possible:

a) If $w_m = 1$, then $\mathbf{w} = \mathbf{e}_m$, where $\mathbf{e}_m$ is the mth standard basis vector. Therefore $\text{diag}(\mathbf{\Sigma})' = \mathbf{e}_m' \mathbf{\Sigma}$ such that $\rho_{i,m} = \delta_i$ for all $i \neq m$. An information structure $\mathbf{\Sigma} \in \mathbf{S}_{++}^N$ with this pattern can be constructed by picking $\delta_i \sim [0, \delta_m)$ and setting $\rho_{i,j} = \delta_i \delta_j / \delta_m$ for all $i \neq j$. To see that the resulting $\mathbf{\Sigma} \in \mathbf{S}_{++}^N$, consider the principal sub-matrix $\mathbf{\Sigma}_{(-m)}$ formed by removing the mth row and column from $\mathbf{\Sigma}$. By the Schur complement $\mathbf{\Sigma} \in \mathbf{S}_{++}^N$ if and only if $\delta_m > 0$ and $\mathbf{\Sigma}_{(-m)} - \text{diag}(\mathbf{\Sigma}_{(-m)}) \text{diag}(\mathbf{\Sigma}_{(-m)})' / \delta_m \in \mathbf{S}_{++}^{N-1}$. This final expression becomes $\text{Diag}((\delta_i - \delta_i^2 / \delta_m : i \neq m)')$, which is clearly in $\mathbf{S}_{++}^{N-1}$. Generally, if $\rho_{i,m} = \delta_i$ for all $i \neq m$, the revealed aggregator simplifies to

$$\text{diag}(\mathbf{\Sigma})' \mathbf{\Sigma}^{-1} (\mathbf{X} - \mu_0 \mathbf{1}_N) + \mu_0 = \mathbf{e}_m' (\mathbf{X} - \mu_0 \mathbf{1}_N) + \mu_0 = X_m.$$

Thus $\mathbb{E}(Y|\mathbf{Z}) = \mathbb{E}(Y|Z_m)$. In other words, the forecasts X_i for $i \neq m$ do not provide any further information beyond X_m, and $\mathcal{F}_i \subseteq \mathcal{F}_m$ for all $i \neq m$.

b) Recall that $\delta_m \geq \delta_i$ for all $i \neq m$. If some $\rho_{j,m} = \delta_m$, the correlation coefficient between X_j and X_m is

$$\text{Corr}\,(X_j, X_m) = \frac{\rho_{j,m}}{\sqrt{\delta_j \delta_m}} = \sqrt{\frac{\delta_m}{\delta_j}} \geq 1.$$

Given that the correlation coefficient is always within $[-1, 1]$, it must be the case that $\delta_j = \delta_m$. This means that the jth and mth columns and rows form a 2×2 principal sub-matrix whose elements are all equal to δ_m. This sub-matrix is singular, contradicting the initial assumption $\mathbf{\Sigma} \in \mathbf{S}_{++}^N$.

□

Require: Condition number threshold $\kappa \geq 1$ and sample eigenvalues in ascending order $l_1 \leq l_2 \leq \cdots \leq l_{N+1}$.

procedure BINARY-SEARCH OPTIMIZATION
  Initialize $D \leftarrow \max\{l_1, 0\}$ and $U \leftarrow l_{N+1}/\kappa$.
  $\mu_0 \leftarrow (D+U)/2$
  for $n = 0, 1, \ldots$ **do**
    Compute μ_n^*, $\mathfrak{d}_n$, and $\mathfrak{u}_n$.
    if $\mu_n^* < 0$ and $\mathfrak{d}_n < 0$ **then**
      return 0
    else if $\mu_n^* < \mathfrak{d}_n$ **then**
      $U \leftarrow \mathfrak{d}_n$
    else if $\mu_n^* > \mathfrak{u}_n$ **then**
      $D \leftarrow \mathfrak{u}_n$
    else
      return μ_n^*
    end if
    $\mu_{n+1} \leftarrow (D+U)/2$
  end for
  return μ_n^*
end procedure

Algorithm 2: This procedure solves (10) efficiently using the structure of the problem and binary-search.

7.3 Finding μ^* for $\mathcal{P}_{sd}(\cdot : \kappa)$

This section describers a binary-search-like algorithm to solve

$$\mu^* = \operatorname*{arg\,min}_{\mu \geq 0} g(\mu) = \operatorname*{arg\,min}_{\mu \geq 0} \sum_{i=1}^{N} \left((\mu - l_i)_+^2 + (l_i - \kappa\mu)_+^2 \right) \tag{10}$$

First, it can be assumed that $\text{cond}(h(\mathbf{S}_Z)) \notin [1, \kappa]$; otherwise, the projection can simply return $h(\mathbf{S}_Z)$. Second, $\max\{0, l_1\} \leq \mu \leq l_{N+1}/\kappa$ because otherwise moving μ closer to the nearest sample eigenvalue decreases $g(\mu)$. Now, consider some value $\mu_n \geq 0$ and two index sets $\mathfrak{D}_n = \{i : l_i \leq \mu_n\}$ and $\mathfrak{U}_n = \{i : \mu_n \kappa \leq l_i\}$. Then,

$$g(\mu_n) = \sum_{i \in \mathfrak{D}_n} (\mu_n - l_i)^2 + \sum_{i \in \mathfrak{U}_n} (l_i - \kappa\mu_n)^2,$$

which has a global minimum at

$$\mu_n^* = \frac{\sum_{i \in \mathfrak{D}_n} l_i + \kappa \sum_{i \in \mathfrak{U}_n} l_i}{|\mathfrak{D}_n| + \kappa^2 |\mathfrak{U}_n|}$$

The operator $|\mathfrak{A}|$ denotes the number of elements in the set $\mathfrak{A}$. Let $\mathfrak{d}_n$ and $\mathfrak{u}_n$ denote the minimum and maximum, respectively, of the interval where any value of μ gives the index sets $\mathfrak{D}_n$ and $\mathfrak{U}_n$. To make this specific, define two operators:

$$d(\mu) = \max\{l_i : l_i \leq \mu\} \qquad \text{and} \qquad u(\mu) = \min\{l_i : l_i \geq \mu\}.$$

If no value is found, then $d(\mu) = 0$ and $u(\mu) = +\infty$. Then,

$$\begin{aligned}\mathfrak{d}_n &= \max\{d(\mu_n), d(\mu_n\kappa)/\kappa\} \\ \mathfrak{u}_n &= \min\{u(\mu_n), u(\mu_n\kappa)/\kappa\}\end{aligned}$$

Of course, μ_n^* is the solution to (10) as long as $\mu_n^* \in (\mathfrak{d}_n, \mathfrak{u}_n]$. If, on the other hand, μ_n^* is less than $\mathfrak{d}_n$ (or greater than $\mathfrak{u}_n$), the global minimum μ^* must be smaller than $\mathfrak{d}_n$ (or greater than $\mathfrak{u}_n$). If μ_n^* is, say, less than $\mathfrak{d}_n$, then a natural approach is to update μ_n to μ_{n+1} that is somewhere between $\mathfrak{d}_n$ and some known lower bound of μ. This gives rise to a binary-search-like algorithm described in Algorithm 2.

References

Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *Information Theory, IEEE Transactions on*, 51(7):2664–2669.

Broomell, S. B. and Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74(3):531–553.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

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.

Foster, D. P. and Vohra, R. V. (1998). Asymptotic calibration. *Biometrika*, 85(2):379–390.

Goel, S., Reeves, D. M., Watts, D. J., and Pennock, D. M. (2010). Prediction without markets. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 357–366. ACM.

Gubin, L., Polyak, B., and Raik, E. (1967). The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24.

Hastings, C., Mosteller, F., Tukey, J. W., and Winsor, C. P. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3):413–426.

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

Hwang, S.-G. (2004). Cauchy's interlace theorem for eigenvalues of hermitian matrices. *American Mathematical Monthly*, pages 157–159.

Langford, E., Schwertman, N., and Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55(4):322–325.

Lichtendahl Jr, K. C. and Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755.

Lobo, M. S. and Yao, D. (2010). Human judgement is heavy tailed: Empirical evidence and implications for the aggregation of estimates and forecasts. *Fontainebleau: INSEAD*.

McCullagh, P., Nelder, J. A., and McCullagh, P. (1989). *Generalized linear models*, volume 2. Chapman and Hall London.

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.

Moore, D. A. and Klein, W. M. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes*, 107(1):60–74.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

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.

Satopää, V. A., Pemantle, R., and Ungar, L. H. (2015). Modeling probability forecasts via information diversity. Under Review.

Tanaka, M. and Nakata, K. (2014). Positive definite matrix approximation with condition number constraint. *Optimization Letters*, 8(3):939–947.

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.

Won, J. H. and Kim, S.-J. (2006). Maximum likelihood covariance estimation with a condition number constraint. In *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*, pages 1445–1449. IEEE.