Getting More Wisdom from the Crowd: When Weighting Individual Judgments Reliably Improves Accuracy or Just Adds Noise

SHU HUANG, RUSSELL GOLMAN AND STEPHEN BROOMELL, Carnegie Mellon University

1. INTRODUCTION

The *wisdom of crowds* describes the phenomenon that aggregated judgment across individuals tends to be more accurate than a typical individual's judgment and may even be more accurate than any single individual's judgment (Surowiecki, 2004; Davis-Stober et al., 2014). The premise of collective wisdom is that errors may cancel out through judgment aggregation, allowing us to extract the knowledge shared by the members of the crowd (Makridakis and Winkler, 1983; Hong and Page, 2008; Minson et al., 2018). The theoretically optimal aggregation method to maximize the accuracy of the aggregated judgment is a weighted average of the individual judgments where the optimal weights are determined by the accuracy (biases), reliability (variances) and dependency (correlations) of the individual judgments (Lamberson and Page, 2012; Davis-Stober et al., 2014; 2015). However, a simple average of the individual judgments (i.e., equal-weighting) may outperform this weighted average because the judgment biases, variances, and correlations are unknown, and estimating them from empirical data produces unstable weights (Kang, 1986; Winkler and Clemen, 1992). Here we investigate how many judgments we need for the estimated optimal weights to be more reliable than a simple average.

We explore the conditions under which the optimal-weighting method (OW) with weights computed from an estimated judgment covariance matrix is more reliable than the equal-weighting method (EW). When using the sample covariance matrix to compute the weights, there will be some error in the weighted average judgment due to the mis-specified weights (which depends on how far this particular sample covariance matrix is from the true covariance matrix), along with some error due to the variability of the judgments (which depends on the true covariance matrix). The accuracy of the weighted average judgment can be represented as a function of the sample size (i.e., the number of judgments per person in the dataset) and the true covariance matrix, while the accuracy of the simple average judgment depends only on the true covariance matrix. Thus, whether OW outperforms EW depends on both the sample size and the unknown true covariance matrix. Of course, as the sample size goes to infinity, the sample covariance matrix approaches to the true covariance matrix, resulting in an optimally accurate weighted average. However, there are true covariance matrices for which the optimal weights are actually equal weights (e.g., if correlations are all zero and variances are all identical), so we can never guarantee that OW strictly outperforms EW. Instead, we can compute the sample size necessary to guarantee that OW is almost as accurate as EW, regardless of the true covariance structure. Using numerical methods, we find that a surprisingly large sample size is required to ensure that OW is even 95% as accurate as EW for any true covariance matrix.

A guarantee of high accuracy for OW for any true covariance matrix is a high standard of reliability. The sample covariance matrix provides useful information about the true covariance matrix, and we might be satisfied with a guarantee that OW outperforms EW for true covariance matrices that we consider sufficiently likely. Since we do not have a prior distribution for the true covariance matrix, we seek an analogue of a hypothesis test, according to which we may reject EW in favor of OW if the
likelihood of observing the sample covariance matrix is sufficiently low for any true covariance matrix for which EW is more accurate than OW. We develop an algorithm to test whether the observed judgments in an empirical dataset are sufficient for researchers to reject EW and instead trust OW as more reliable than EW. The output of our test provides a quantitative measure of the reliability of the weighted average method. By using simulated judgments, we find that our hypothesis test algorithm displays clear diagnostic value for deciding how to combine judgments, and moderate sample sizes may sometimes be sufficient to generate reliably accurate optimal weights.

1.1 Determining the minimum sample size to guarantee OW reliability for any true covariance matrix

Let a target value of interest to a decision maker be $\theta$. A set of $M$ calibrated forecasters provide their judgment or forecasts $f^T = (f_1, ..., f_M)$ with multi-normally distributed errors, that is, $f = \theta + e$ and $e \sim \text{MNV}(0, \Sigma)$ where $\Sigma$ is the $M \times M$ variance-covariance matrix of the judgment errors. The optimal weights have been identified to maximize the accuracy of aggregated judgments given the estimate of covariance matrix $\hat{\Sigma}$, that is, $w^T = (1^T \hat{\Sigma}^{-1})/(1^T \hat{\Sigma}^{-1} 1)$, where $1^T = (1, \ldots, 1)$ is the unit $M$-vector. Then the weighted average judgment is $f_c = w^T f$, and the simple average judgment is $f_a = 1^T f / M$.

To determine the more reliable judgment aggregation method, we can compare the mean squared errors of OW and EW, which (because errors are mean zero) are just $\text{Var}(f_c - \theta)$ and $\text{Var}(f_a - \theta)$ respectively. Now assume that the true covariance matrix is unknown. Given an estimate of the covariance matrix $\hat{\Sigma}$, we can obtain the unconditional variance of the OW aggregated judgment error according to the law of total variance as shown in (1). The maximum-likelihood estimator of the covariance matrix is simply the sample covariance matrix. The sampling distribution is the Wishart distribution, $n\hat{\Sigma} \sim \text{W}(\Sigma, n-1)$ where $n$ is the sample size of judgments from each forecaster.

$$
\text{Var}(f_c - \theta) = E_{\hat{\Sigma}} \left[ \frac{1^T \hat{\Sigma}^{-1} \Sigma^{-1} 1}{1^T \hat{\Sigma}^{-1} 1} \right], \quad \text{Var}(f_a - \theta) = \frac{1^T \Sigma 1}{M^2}
$$

We seek to find the minimum sample size $n^*$ of judgments entering into the sample covariance matrix for OW to be reliable relative to EW. Since we cannot ensure OW will be more accurate than EW for any true covariance matrix, we consider OW reliable at level $\alpha$ if the MSE of OW multiplied by $\alpha$ is less than the MSE of EW for any true covariance matrix. The minimum sample size $n^*_\alpha$ that makes OW reliable at any acceptable $\alpha$ level of reliability is given by:

$$
\min \ n \\
\text{s.t.} \ \text{Var}(f_a - \theta) - \alpha \cdot \text{Var}(f_c - \theta) \geq 0, \ n \in \mathbb{Z}^+
$$

In our numerical experiments, we assume the optimal weights are computed using the sample covariance matrices generated from the Wishart distribution, and solve the minimum sample size at three different $\alpha$ levels of reliability, 0.99, 0.95 and 0.90. From simulation results, we find an approximately linear relationship between the minimum sample size $n^*_\alpha$ and the number of forecasters $M$, holding fixed the reliability level $\alpha$ (see Table 1).

### Tab. 1. Minimum sample size at different reliability levels (assuming two types of forecasters)

<table>
<thead>
<tr>
<th>$M$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n^*_0.99$</td>
<td>121</td>
<td>209</td>
<td>326</td>
<td>411</td>
<td>512</td>
<td>613</td>
<td>714</td>
<td>814</td>
</tr>
<tr>
<td>$n^*_0.95$</td>
<td>33</td>
<td>55</td>
<td>76</td>
<td>86</td>
<td>108</td>
<td>126</td>
<td>147</td>
<td>166</td>
</tr>
<tr>
<td>$n^*_0.90$</td>
<td>13</td>
<td>24</td>
<td>36</td>
<td>45</td>
<td>55</td>
<td>69</td>
<td>80</td>
<td>85</td>
</tr>
</tbody>
</table>

1.2 Testing whether the sample covariance matrix is consistent with EW being more accurate than OW

Existing work focusing on the two-forecaster case (Schmittlein et al., 1990; Blanc and Setzer, 2016) or using linear regression to estimate accuracy (Winkler and Clemen, 1992; Chandrasekharan et al., 1994)
provides guidelines for when to use OW from estimated covariance matrices. We provide a theoretical framework analogous to hypothesis testing to extend these guidelines to the general $M$-forecaster case. The precision and reliability of OW depends on the true covariance matrix. We use the sample covariance matrix to assess confidence that the true covariance matrix is inconsistent with EW being more accurate than OW. Given $\hat{\Sigma}$, we identify the most likely estimate of the true covariance matrix restricted such that EW has smaller MSE than OW, denoted by $\Sigma^*$.

Then, an analogue p-value describes the likelihood of observing such an improbable sample covariance matrix given a true covariance matrix $\Sigma = \Sigma^*$ suggesting EW outperforms OW. We calculate p-value $= \int_{A(\hat{\Sigma})} g(\hat{\Sigma}|\Sigma^*)d\hat{\Sigma}$ where $A(\hat{\Sigma})$ is the set of $\hat{\Sigma}$ such that $\Pr(\hat{\Sigma}|\Sigma^*) \leq \Pr(\hat{\Sigma}|\Sigma^*)$ and $g(\cdot|\Sigma^*)$ is the Wishart density function with scale matrix $\Sigma^*$. The p-value is the cumulative probability of covariance matrices less likely to be sampled than $\hat{\Sigma}$, according to a Wishart distribution with scale matrix $\Sigma^*$. If this p-value is sufficiently small, we reject the hypothesis that $\Sigma^*$ is the true covariance matrix, and more generally, we reject that EW outperforms OW.

We conducted a simulation to validate our hypothesis test algorithm. We simulated judgments of five individuals from a mean-zero multivariate normal distribution using three different true covariance matrices. The first covariance matrix ($\Sigma_1$) is an identity matrix where EW are theoretically the best choice. The second covariance matrix ($\Sigma_2$) assumes independent judges with different judgment variances. The third covariance matrix ($\Sigma_3$) assumes identical judgment variances but varied correlations. For $\Sigma_2$ and $\Sigma_3$, the optimal weights are $w^T = (.271, .220, .183, .163, .163)$. We simulated $n$ judgments for each individual (varying $n$ from 10 to 50), computed a sample covariance matrix, and applied our hypothesis testing algorithm, and then repeated this process 1000 times. Table 2 displays how often our hypothesis test algorithm rejected the null hypothesis to use EW with a 5% threshold for Type I error rate (i.e., p-value less than 0.05) for each of the true covariance matrices. Reassuringly, the test rejects EW less than 5% of the time for $\Sigma_1$. For $\Sigma_2$ and $\Sigma_3$, the test rejects EW more frequently than for $\Sigma_1$ and increasingly more often as the sample size increases. Larger sample sizes would be necessary to reliably reject EW for $\Sigma_2$ and $\Sigma_3$, but for some covariance matrices, moderate sample sizes may suffice. Overall, our hypothesis test algorithm displays clear diagnostic value for deciding how to combine judgments.

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma_1$</td>
<td>0</td>
<td>1.1</td>
<td>0.5</td>
<td>1.2</td>
<td>2.0</td>
</tr>
<tr>
<td>$\Sigma_2$</td>
<td>0</td>
<td>1.4</td>
<td>5.1</td>
<td>7.2</td>
<td>9.1</td>
</tr>
<tr>
<td>$\Sigma_3$</td>
<td>0</td>
<td>1.9</td>
<td>3.6</td>
<td>7.3</td>
<td>10.4</td>
</tr>
</tbody>
</table>

1.3 Summary
We can never be completely sure that the optimal-weighting aggregation method will outperform the simple average method because the true judgment covariance matrix is unknown. Just making sure that OW is almost as accurate as EW requires a surprisingly large sample of previously observed judgments. However, we can be reasonably confident that OW will outperform OW if we have a sufficiently large dataset that suggests that weighting the judgments will help. We develop an algorithm to quantify the confidence that researchers could have in trusting OW to be more reliable than EW given an existing sample of judgments. The measure of reliability derived from our algorithm should play a central role in evaluating aggregation methods.
REFERENCES


