Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

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

Shu Huang, Stephen B. Broomell, Russell Golman
Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, shuh1@andrew.cmu.edu, broomell@cmu.edu, rgolman@andrew.cmu.edu

The wisdom of a crowd can be extracted by simply averaging judgments, but weighting judges based on their past performance may improve accuracy. The reliability of any proposed weighting scheme depends on the estimation precision of the features that determine the weights, which in practice cannot be known perfectly. Therefore, we can never guarantee that any weighted average will be more accurate than the simple average. However, depending on the statistical properties of the judgments (i.e., their estimated biases, variances, and correlations) and the sample size (i.e., the number of judgments from each judge), we may be reasonably confident that a weighted average will outperform the simple average. We develop a general algorithm to test whether there are sufficiently many observed judgments for practitioners to reject using the simple average and instead trust a weighted average as a reliably more accurate judgment aggregation method. Using simulation, we find our test provides better guidance than cross validation. Using real data, we demonstrate how many judgments may be required to be able to trust commonly used weighted averages. Our algorithm can be used for power analysis when planning data collection and as a decision tool given existing data to optimize crowd wisdom.

Key words: estimation error, weighted average, wisdom of crowds, hypothesis test algorithm

1. Introduction

In uncertain contexts, decision makers can improve judgment accuracy by aggregating information from multiple sources. For example, a manager of a company's marketing team might ask her or his team members to predict the percentage change in a product's sales in the next year. In such cases, the manager may rely on the wisdom of crowds, the phenomenon that aggregated judgment

across individuals tends to be more accurate than a random individual's judgment and may even be more accurate than any single individual's judgment (Davis-Stober et al. 2014, Surowiecki 2005). By combining all team members' forecasts, the manager is able to obtain a more accurate sales prediction. In expert elicitation, similarly, a panel of experts drawing on diverse sources of expertise (e.g., independent information cues and varied analytic methodologies) generally produces aggregated judgments that are more accurate than one specific expert's opinion (Bansal et al. 2017, Budescu and Chen 2014, Larrick and Soll 2006, Palley and Soll 2019). The premise of collective wisdom is that individual judgment errors cancel out through judgment aggregation, allowing us to extract the knowledge shared by the members of the crowd (Hong and Page 2008, Makridakis and Winkler 1983, Minson et al. 2017).

There are many methods for aggregating judgments from multiple individuals, and determining the best method in a given context requires analysis of the statistical properties of the judgment context. In other words, the best way to aggregate individuals' judgments depends on how much the decision maker knows about the environmental characteristics (e.g., the uncertainty inherent to quantities of interest and the redundancy of information) and the individual judges (e.g., their judgment abilities and the dependencies between their judgments) (Broomell and Budescu 2009). Particularly, when the decision maker knows little about the judges, a simple average of the judgments is applicable and reasonable (Mannes et al. 2012). Previous work has already demonstrated that the simple average, as an exemplar improper linear model, has a robust performance and can be superior to the single predictor or the standard regression model in many situations (Dawes 1979, Einhorn and Hogarth 1975). Davis-Stober et al. (2010) also derived the upper bound on the mean squared error of the equal-weighting estimator and demonstrated that it has less variance than the ordinary least squares (OLS) estimate.

Although the simple average is better than a random individual's judgment, there may be room for improvement by using a weighted average in order to take full advantage of the wisdom of crowds. For instance, to balance the benefits of information aggregation and the costs of introducing

less accurate judgments, the decision maker could take the simple average of judgments only from the best-performing group of judges (i.e., placing equal weights on these judges and zero weight on other judges), a method that has been shown to work well in practice (Mannes et al. 2014, Yaniv 1997). More generally, if the judges have established track records (i.e., known and stationary judgment accuracy), then the decision maker can use targeted weights to average their judgments, giving better judges higher weights (Budescu and Chen 2014, Olsson and Loveday 2015). For example, a manager in a marketing team could identify each team member's predictive ability according to a collection of team members' forecasts year by year, and then rely more on team members with high accuracy. Furthermore, if the decision maker not only knows about judges' abilities but also the dependencies among judges, he can do better still. Specifically, with known biases, variances, validities and correlations of the judges, he could use a weighted average with weights determined by minimizing the expected squared error of the aggregated judgment, what we will refer to as the theoretically optimal aggregation method. Such a weighting method would place higher weights on judges who tend to be more accurate as well as less correlated (ideally even negatively correlated) with the rest of the crowd (Davis-Stober et al. 2014, 2015, Lamberson and Page 2012).

In practice, however, the judges' abilities and correlations are unknown, and a weighted average can only be computed based on their estimates. For instance, selecting the best-performing group of judges depends on identifying the predictive ability of judges, which may involve estimating individual's validity (i.e., the correlation between individual's judgments and the target value). Determining the theoretically optimal weights requires estimates of all the statistical properties of the judges and the environment. These weights can generate poor performance if the estimates are far from the true values. In such cases, the simple average of the individual judgments (i.e., the equal-weighting method) outperforms the weighted average (Genre et al. 2013, Stock and Watson 2004). The inferior performance of the weighted average based on empirical data is driven by imperfect estimates of the unknown statistical properties of the judgments. Estimates are necessarily

imperfect because of sampling error. Although some sampling error is unavoidable, in many cases the sampling error caused by insufficient sample size is large enough to produce unstable weights from the optimization (Kang 1986). Winkler and Clemen (1992) have investigated the sampling distribution of the best estimated weights in the two-forecaster case and found that even small errors in estimates of variance and correlation result in highly fluctuating weights, sometimes even outside the range from zero to one (i.e., some negative weights). Sampling error is a potential risk for any weighting method, including crowd selection methods that equally weight just a selected subset of the entire crowd, because small samples may be misleading about who should be selected as well as about how much each selected individual should be weighted.

How can we tell if a generalized weighted average based on an observed sample of judgments will be more reliably accurate than the simple average? Due to the estimation error, we can never rule out with complete certainty the possibility that a weighted average based on estimates of the statistical properties of the judges is worse than the simple average. However, in some cases, observed judgments may give us reasonable confidence that the estimation error for the weights is small enough to generate a weighted average that is more accurate than the simple average. Recent work by Blanc and Setzer (2016) has offered a decision threshold to determine when to use the simple average or the naively optimally weighted average, but they considered only two judges, which simplifies the analysis considerably. We develop a more generally applicable method which can compare any weighted average to the simple average and can be used with any number of judges.

Specifically, we propose a hypothesis test algorithm to assist the decision maker with the selection of the aggregation method to achieve the best possible accuracy given the data available. The null hypothesis is that the simple average is better than a given weighted average. But there are many possible true states of the world in which this would hold. Based on observed data, we search for the most likely true state of the world that supports the null hypothesis and adopt it as a presumptive scenario. We then compute a p-value for this presumptive scenario, providing a

quantitative measure of the reliability of the weighted average in light of the uncertainty about the true environment with the given sample size. Thus, our hypothesis test algorithm provides a useful check on estimation error before decision makers decide to use any weighting method instead of the simple average.

Returning to the marketing team example, suppose a manager has collected a sample of prior judgments from all team members, as well as many suggestions on how to assign weights on those judgments. She or he can utilize our algorithm to determine whether there is sufficient information in this sample to trust the given weighted average. When the number of judgments from each team member is small, it is risky for the manager to trust any weighted average due to large sampling errors in estimating the true predictive accuracy and correlation of team members. In this case, the sampling error swamps the potential benefits of weighting, and our algorithm would not reject the simple average. As the number of judgments from each team member increases, the estimates of true predictive accuracy and correlation of team members become more representative of the true state of world such that a weighted average becomes more robust. The manager can obtain a quantitative measure from our algorithm of how confident she or he should be in weighting the judgments of different team members based on these estimates. The algorithm can thus be used as a decision rule, with a pre-specified significance threshold, similar to a traditional hypothesis test.

An alternative approach to decide whether to trust a weighted average is cross validation. In Section 4, we compare our hypothesis test algorithm to cross validation using simulations in which we can tell how well any given weighted average will perform. We find that while cross validation also performs well, it makes more errors than our test when applied to small sample sizes. This phenomenon becomes more pronounced when the proposed weighted average is more significantly different from the simple average. We also demonstrate the application of our algorithm to real data, specifically, an existing dataset from the European Central Bank's Survey of Professional Forecasters (SPF) in which domain experts provide hundreds of forecasts on many macro-economic indicators. This exercise lets us identify the number of observations necessary for different weighted averages to outperform the simple average on this particular dataset.

Previous work either considers the reliability of a weighted average as a function of the true environment, which we can never have perfect knowledge of, or when restrictive assumptions apply, such as contexts with only two unbiased forecasters. Our hypothesis test guides the decision of a judgment aggregation method based on statistical inference from observable data instead of an assumed true environment. The test also accounts for judgment bias, variance and correlation simultaneously, and can be applied with any number of forecasters, and to any method for estimating weights.

This paper is organized as follows. Section 2 briefly reviews the literature on the decomposition of the expected squared error and the performance of some popular weighted averages versus simple averages in judgment aggregation. Section 3 introduces the basic model to compute the expected squared error of aggregated judgments as a function of the true judgment biases, true judgment covariance matrix, and the number of judgments (per forecaster) in the sample. Section 4 presents our hypothesis test algorithm to assist the decision maker in selecting an appropriate aggregation method for any collection of M judges. We validate and demonstrate the effectiveness of our algorithm with simulated judgments, and demonstrate its application to the European Central Bank's SPF dataset. We conclude in Section 5 with a discussion of our framework's advantages, limitations, and future directions. Sample data and code of this paper are also available online (link to Github).

2. Previous Literature on Judgment Aggregation

We apply the definition of crowd wisdom proposed by Davis-Stober et al. (2014) to assess the quality of aggregated judgments. When a collection of M individuals predict a target value of interest, crowd wisdom is defined as a linear aggregate of the members' judgments having less expected Squared Error (SE) than the judgment from one randomly selected individual member. The expected SE, as the measure of judgment accuracy, can be decomposed into four components:

$$E[SE] = (\mu_X^T \mathbf{w} - \mu_y)^2 + \mathbf{w}^T \Sigma_{XX} \mathbf{w} - 2\mathbf{w}^T \sigma_{Xy} + \sigma_y^2 \tag{1}$$

where μ_X is a $M \times 1$ vector indicating the judgment mean for all M individuals and Σ_{XX} is the $M \times M$ covariance matrix of judgments. The target value is also considered as a random variable with mean and variance denoted by μ_y and σ_y^2 respectively. Correlation between the target value and individuals' judgments, represented by a $M \times 1$ vector σ_{Xy}, depicts the validity of individuals' judgments. The weight vector $\mathbf{w}$ represents different aggregation rules and it has a constraint, $\mathbb{1}^T\mathbf{w} = 1$ where $\mathbb{1}$ is a $M \times 1$ vector of ones. To pursue a lower expected squared error, weights are not restricted by non-negativity, meaning negative weights are feasible if higher accuracy of aggregated judgments can be achieved.

The simple average is a special case for linearly combining judgments where equal weights are assigned to a collection of M individuals. There is no information about the judges' precision and dependence required to generate equal weights, as judges are treated as exchangeable in a simple average. Davis-stober et al. (2014) have demonstrated that the simple average can produce a robust wisdom-of-the-crowds effect by comparing the expected SE between the simple average and a random selected individual's judgment.

Previous literature has focused on the expected SE of a weighted average with *theoretically optimal weights* computed from given values of judges' biases, variances, and correlations. The optimal aggregation weights for minimizing the expected SE of aggregated judgments can be obtained by solving the following system of linear equalities:

$$\begin{bmatrix} \Sigma_{XX} + (\mu_X - \mu_y\mathbb{1})(\mu_X - \mu_y\mathbb{1})^T & \mathbb{1} \\ \mathbb{1}^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ \lambda \end{bmatrix} = \begin{bmatrix} \sigma_{Xy} \\ 1 \end{bmatrix} \tag{2}$$

where λ is a real-valued unknown variable, i.e., a Lagrange multiplier (Davis-Stober et al. 2015). If the sub-matrix $\Sigma_{XX} + (\mu_X - \mu_y\mathbb{1})(\mu_X - \mu_y\mathbb{1})^T$ is positive-definite, the equation has a unique solution. The optimal weights solved above apply for a general case: (1) individuals' judgments may be both biased and correlated, and (2) the target value is a random variable as well.

In practice, all the population parameters in Eq. (2), including the true judgment bias, variance, correlation and predictive validity, can never be perfectly known, meaning that the theoretically

optimal weights can only be computed from estimates derived from finite samples. The sampling error inherent in these estimates can result in unstable weights that are no longer truly optimal. The amount of sampling error thus becomes an important factor in the reliability of the theoretically optimal weighting method. To reduce the sampling error and improve the accuracy of aggregated judgment, various alternative weighting methods are proposed. We classify them into three categories: regularization of covariance matrix, constrained regularized regression, and sub-crowds..

2.1. Regularization of Covariance Matrix

The first category is to estimate a smaller number of population parameters by simplifying the formula in Eq. (2). As an analytic solution for the role of sampling error in reducing the reliability of weighted averages is not possible, prior work relies on two simplifying assumptions. First, the judgment errors (i.e., differences between judgments and the target value) and the target value are assumed to be independent, making the covariance matrix of judgment errors equivalent to the covariance matrix of judgments while treating the target value as a fixed number. Second, individuals' judgments are assumed to be unbiased (i.e., to have a zero-mean error). In practical applications, after observing each individual's mean judgment error, we are able to debias the individuals' judgments by shifting each by the equal and opposite amount of its mean error. With these simplifying assumptions, the optimal weights depend only on the variances and correlations of individuals' judgment errors (Clemen and Winkler 1986, Kang 1986). For the two-forecaster case, the weight assigned on the first forecaster would be $w_1 = (1 - \rho\sigma_1/\sigma_2)/(1 + (\sigma_1/\sigma_2)^2 - 2\rho\sigma_1/\sigma_2)$ and the weight on the second forecaster is $1 - w_1$, where σ_1 and σ_2 are the standard deviation of two forecasters, respectively, and ρ denotes the correlation (Winkler and Clemen 1992). Extending to multiple forecasters, the formula of optimal weights becomes:

$$\mathbf{w}^T = \frac{\mathbb{1}^T \Sigma_{XX}^{-1}}{\mathbb{1}^T \Sigma_{XX}^{-1} \mathbb{1}} \tag{3}$$

where Σ_{XX} is the *true* covariance matrix of individuals' judgment errors (Lamberson and Page 2012).

Given this simplified representation of the optimal weights, Winkler and Clemen (1992) have explicitly identified the sampling distribution of the estimated weights used to aggregate judgments for the two-forecaster case. They use numerical results to show that the estimated weights can be highly variable, particularly when variances of the two judges are almost the same and the correlation of their judgments is high. This finding reconciles literature on weighting dependent sources by their informativeness where small changes in the correlation of any pair of sources could cause a large variability of informativeness measurement along with the corresponding aggregation weights (Clemen and Winkler 1985, Morrison and Schmittlein 1991, Satopää 2017).

Within this simplified framework, reducing the sampling error in the estimated weights is equivalent to improving the estimation accuracy of the true covariance matrix. A trade-off exists in estimating the covariance matrix: the estimation error brought by an additional estimated parameter might be greater than the reduction in modelling error (i.e., misspecification) with respect to the additional parameter. This tradeoff leads to regularized weighting methods that may, for example, assume identical correlation among all judges or cluster judges into several groups, within which each member is identical (Merkle et al. 2020). Schmittlein et al. (1990) have illustrated the sensitivity of weighted judgment aggregation to various assumed covariance structures. Four operational models of estimating the covariance matrix are applied to compute the weights as well as the aggregated judgment. These models respectively assume exchangeable individuals (i.e., permitting only the equal weighting aggregation method), independent individuals with varied judgment variance, dependent individuals with identical judgment variance and correlation, and dependent forecasters with varied judgment variance and correlation (i.e., the fully general covariance matrix used in the optimal weighting method). Simulation results show that as the number of judgments collected from each judge increases, estimating the full covariance matrix (i.e., using the optimal weighting method) can lead to a more accurate aggregated judgment than assuming exchangeable judges (i.e., the equal weighting method), but the threshold number of judgments required depends on the unknown true covariance matrix. In this paper, we do not apply any simplifying assumptions on the true population parameter to keep the generalization of our algorithm, but our algorithm can still be used to compare the above class of weighting methods to the simple average.

2.2. Constrained Regularized Regression

The second way to reduce the sampling error is regulating judgment weights directly by taking them as the coefficients in a constrained linear regression since the wisdom of crowds' problem to minimize the mean squared error is theoretically a constrained linear regression problem. Thus, we can apply those common regularization techniques in the regularized linear regression to decide the weights. Taking the LASSO regression as an example, due to the collinearity of predictors, a penalty parameter is added to regulate the total number of coefficients to estimate, and as a result only effective predictors are selected and noise is reduced. In the same way, especially for those high-correlated judges, judgment weights from a constrained LASSO regression model are regularized, and more influential judges are naturally selected according to the estimates of weights (Gaines et al. 2018, James et al. 2020). Although these regularized weighting methods have been proven to have a robust performance in empirical data, large sample size is still necessary to decide an appropriate hyper-parameter such as the penalty parameter in the LASSO regression. Our hypothesis test algorithm can compare the regularized weighted averages to the simple average, and the output of our algorithm will tell us whether the current sample size is sufficient to trust a particular implementation of a regularized weighted average.

2.3. Sub-crowds

The last way to deal with sampling error is to constrain the weights according to simple heuristics, such as selecting a subset of the crowd to include in the aggregate judgment (and effectively setting the weights for everybody else to be zero). In practice, this is the most popular, and often most effective, method to improve the accuracy of aggregated judgments. In many cases selected small crowds can outperform larger crowds (Budescu and Chen 2014, Mannes et al. 2014, Olsson and Loveday 2015). Selecting a subset of the crowd can (1) remove poorly performing judges who just add noise or who may be biased; and (2) decrease the risk of introducing sampling errors when determining the weights. We will compare several typical classes of crowd selection methods to the simple average in this paper to validate the effectiveness of our algorithm.

3. Basic Model to Compare a Weighted Average and the Simple Average

To determine a more reliably accurate judgment aggregation method, we compare the expected SE of a weighted average to the simple average. We assume the judgement errors and the target value are independent, as in previous literature. The expected SE is a function of the aggregation rule (i.e., the weights) and the true parameters representing the judgments (i.e., true judgment bias and true covariance matrix between judges). We cannot obtain a perfectly precise estimate of the true parameters with finite samples; the estimates align with the true values only in the limiting case following the Law of Large Numbers. Therefore, our model takes the sampling error introduced by estimated weights into consideration when we compare the expected SE of a weighted average to the simple average.

Let a target value of interest to a decision maker be y. A collection of M individuals provide their judgments or forecasts $\mathbf{X}^T = (X_1, ..., X_M)$ with multivariate-normally distributed errors, that is, $\mathbf{X} = y + \mathbf{e}$ and $\mathbf{e} \sim \text{MVN}(\mu, \Sigma)$ where μ is a $M \times 1$ vector of true bias and Σ is the $M \times M$ true covariance matrix of judgment errors. Holding the assumption that judgment errors and the target value are independent, we can generally represent the estimated weights as a function of the observed judgment bias ($\hat{\mu}$) and the sample covariance matrix ($\hat{\Sigma}$):

$$\hat{\mathbf{w}} = h(\hat{\mu}, \hat{\Sigma}) \tag{4}$$

Then we can denote a weighted average as $f_{wa} = \hat{\mathbf{w}}^T \mathbf{X}$, and the simple average as $f_{sa} = \frac{\mathbb{1}^T \mathbf{X}}{M}$.

The expected SE quantifies the accuracy and reliability of aggregation methods. Given the estimated bias and covariance matrix, the conditional expected SE of a weighted average is:

$$E[(f_{wa} - y)^2 | \hat{\mu}, \hat{\Sigma}] = \hat{\mathbf{w}}^T (\mu^T \mu + \Sigma) \hat{\mathbf{w}} \tag{5}$$

According to the law of total variance, the unconditional expected SE of a weighted average is obtained by taking the joint expectation of the estimated bias and covariance matrix:

$$E[(f_{wa} - y)^2] = E_{\hat{\mu}, \hat{\Sigma}}[\hat{\mathbf{w}}^T (\mu^T \mu + \Sigma) \hat{\mathbf{w}}] = \int_{\hat{\mu}, \hat{\Sigma}} \hat{\mathbf{w}}^T (\mu^T \mu + \Sigma) \hat{\mathbf{w}} g(\hat{\mu}, \hat{\Sigma}) d\hat{\mu} d\hat{\Sigma} \tag{6}$$

where $g(\hat{\mu}, \hat{\Sigma})$ is the joint density probability function of the estimated bias and the sample covariance matrix.

The most common way to estimate the bias and covariance matrix is using the maximum-likelihood estimator (MLE), i.e., the sample mean error and sample covariance matrix. The sample bias vector (i.e., sample error mean, $\hat{\mu}$) and sample covariance matrix ($\hat{\Sigma}$) of judgments from M individuals can be estimated by:

$$\begin{aligned} \hat{\mu} &= \frac{\mathbb{1}_n^T \mathbf{X}}{n} \\ \hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{S}{n} \end{aligned} \tag{7}$$

where n is the number of judgments from each individual, $\mathbb{1}_n$ is n units of one and X_i is a M-vector judgments from all individuals for the ith target value.

When judgments are assumed to be drawn from a multivariate normal distribution, the sampling distribution for the sample mean is still a multivariate normal distribution with shrunken variance and covariance, denoted by $\hat{\mu} \sim N(\mu, \Sigma/n)$. The sampling distribution for the sample covariance matrix is a Wishart distribution, denoted by $S \sim Wishart(\Sigma, n-1)$. Σ is the scale matrix (i.e., the true covariance matrix) and $n-1$ is the degree of freedom (i.e., related to the number of judgments from each individual). Therefore, the expected SE of a weighted average (see Eq. (6)) can further be considered as a function of the true bias μ, true covariance matrix Σ (including the number of individuals M) and the sample size n.

The expected SE of the simple average, on the other hand, only depends on the true bias μ and covariance matrix Σ but not on the sample size:

$$E[(f_{sa} - y)^2] = \frac{\mathbb{1}^T(\mu^T\mu + \Sigma)\mathbb{1}}{M^2} \tag{8}$$

When the sample size (i.e., the number of judgments from each individual) increases to infinity, the sample mean and sample covariance matrix approach to the true bias and covariance matrix, respectively. We have proven that in the limiting case the expected SE of the optimally weighted average would not exceed the expected SE of simple average (see EC.1.), indicating that the

theoretically optimal weighting method is definitely more reliably accurate than the equal weighting method with infinite samples. However, it is unrealistic to collect infinite judgments from each individual, so in practice we need to consider the sampling error in our estimates. Can we tell when the sample size is sufficient to control the sampling error such that a weighted average provides a better performance (i.e., smaller expected square error of the aggregated judgment) than the simple average? We propose a hypothesis test algorithm to answer this question.

4. Hypothesis Test for Deciding When to Weight

A guarantee of better accuracy for a weighted average for any true bias and covariance matrix is impossible since there always exists the possibility that the true bias and covariance matrix make equal weighting optimal. Still, estimates of the biases and the covariance matrix from finite samples may provide useful information about the true values if the sampling errors can be controlled, for example, by sufficiently increasing the sample size. We might be satisfied knowing that a weighted average outperforms the simple average for those true bias and covariance matrix specifications that we consider sufficiently likely. In our framework, we seek an analogue of a hypothesis test, according to which we may reject the simple average in favor of a weighted average if the likelihood of observing the current estimated bias and covariance matrix is sufficiently low for any true bias and covariance matrix suggesting the simple average will be more accurate than the weighted average. Thus, we develop an algorithm to test whether the observed judgments in an empirical data set are sufficient for researchers to reject using the equal weighting aggregation method and instead trust a weighted average.

Our hypothesis test algorithm is slightly different from traditional hypothesis tests. Without any prior knowledge of the true state of the world (i.e., true bias and covariance matrix), we define the null hypothesis as the simple average being more reliably accurate than a proposed weighted average, and then find the most representative bias and covariance matrix to use as the null state in our algorithm. The null bias (μ^*) and covariance matrix (Σ^*) are the pair most likely to generate current estimated bias and covariance matrix among those pairs for which the

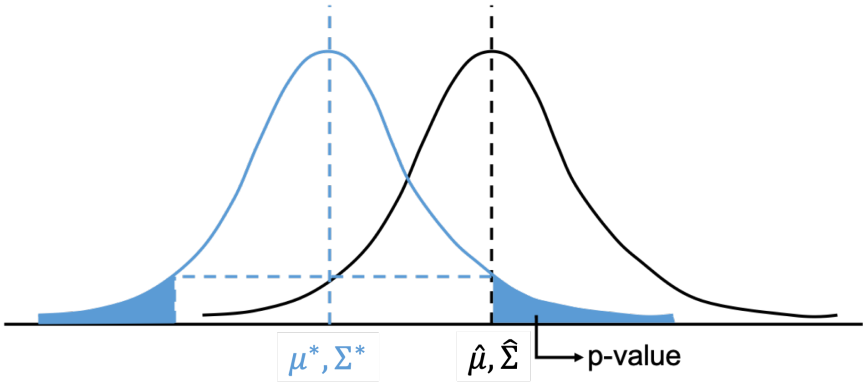


Figure 1 **Conceptual illustration of our analogue P-value. The parameters $\hat{\mu}$ and $\hat{\Sigma}$ are the observed bias and covariance matrix, and the parameters μ^* and Σ^* are the presumed null state of the world, i.e., the parameters most likely to generate $\hat{\mu}$ and $\hat{\Sigma}$, given the constraint that the simple average would have lower mean squared error than the weighted average.**

simple average would outperform the proposed weighted average. The likelihood is computed from the multiplication of a multivariate normal density and a Wishart density given the null bias and covariance matrix. Searching for the null bias and covariance matrix provides flexibility for retaining the simple average. Thus, our algorithm privileges the simple average.

The output of our algorithm is an analogue p-value, which indicates the probability of observing the current estimated bias and covariance matrix (or something more extreme) when the true state of world is the presumptive scenario in which the simple average is more accurate than the proposed weighted average. Figure 1 conceptually shows the p-value given the null state (μ^*, Σ^*) and observed sample size. Small p-values indicate that the probability of observing the realized data would be low if it were the case that the simple average outperforms the weighted average.

4.1. Test Procedure

The first step of the hypothesis test algorithm is determining the bias and covariance matrix consistent with the simple average outperforming the weighted average that would make the observed bias and covariance matrix most likely (i.e., finding the null bias and covariance matrix). Given the sample bias and sample covariance matrix (see Eq. (7)), we need to solve the constrained Maximum

Likelihood Estimation (MLE) problem in Eq. (9):

$$\begin{aligned} \max_{\mu^*,\Sigma^*} \quad & g(\hat{\mu}, S|\mu^*, \Sigma^*, n) \\ s.t., \quad & E[(f_{wa} - y)^2|\hat{\mu}, \hat{\Sigma}] > E[(f_{sa} - y)^2] \end{aligned} \tag{9}$$

The objective function is the likelihood to generate the observed bias ($\hat{\mu}$) and covariance matrix ($\hat{\Sigma} = S/n$) given μ^*, Σ^* and sample size n, so it is the multiplication of a multivariate normal density probability for $\hat{\mu}$ and a Wishart density probability for S due to the independence between $\hat{\mu}$ and S. The constraint represents the condition that in the null state (μ^* and Σ^*) the simple average performs better than the proposed weighted average. The constrained MLE problem is equivalent to the problem in Eq. (10) (more details can be seen in EC.2.):

$$\begin{aligned} \max_{\mu^*,\Sigma^*} \quad & -\frac{1}{2}tr((\Sigma^*)^{-1}S) - \frac{n}{2}\log(|\Sigma^*|) - \frac{n}{2}(\hat{\mu} - \mu^*)^T(\Sigma^*)^{-1}(\hat{\mu} - \mu^*) \\ s.t., \quad & (\frac{\mathbb{1}}{M} - \hat{\mathbf{w}})^T(\mu^*\mu^{*T} + \Sigma^*)(\frac{\mathbb{1}}{M} + \hat{\mathbf{w}}) < 0 \end{aligned} \tag{10}$$

where $\hat{\mathbf{w}}$ represents a weighted average as in Eq. (4), and $\mathbb{1}$ is M units of one.

There is no closed-form solution for this constrained MLE problem, so we solve it by using the Monte Carlo method. Starting with the state in which judges have identical judgment bias, variances and correlation, we search for parameters that are consistent with the simple average outperforming the weighted average and that would be more likely to generate the observed samples by moving in a direction towards the observed bias and covariance matrix. Specifically, we explore parameters that are a linear combination of the current best parameters and the estimated bias and covariance matrix, or that are drawn from the normal-Wishart distribution determined by the current best parameters. Other optimization algorithms can also be used to solve this problem (e.g., the interior point method) if greater precision is required.

After obtaining the null bias μ^* and covariance matrix Σ^*, we can calculate the analogue p-value:

$$\text{p-value} = \int_{A(\tilde{\mu},\tilde{S})} g(\tilde{\mu}, \tilde{S}|\mu^*, \Sigma^*) d\tilde{\mu} d\tilde{S} \tag{11}$$

where $A(\tilde{\mu}, \tilde{S})$ is the set of $(\tilde{\mu}, \tilde{S})$ such that $g(\tilde{\mu}, \tilde{S}|\mu^*, \Sigma^*) \leq g(\hat{\mu}, S|\mu^*, \Sigma^*)$. (This integral can be evaluated numerically.) The set $A(\tilde{\mu}, \tilde{S})$ is conceptually represented by the shaded area in Figure

1, corresponding to bias and covariance matrix values that are more extreme than those actually observed, relative to the null bias and covariance matrix. The p-value is the cumulative probability of $\tilde{\mu}, \tilde{S}$ in $A(\tilde{\mu}, \tilde{S})$ with a joint multivariate normal $\times$ Wishart distribution (i.e., MVN$(\mu^*, \Sigma^*/n) \times$ Wishart$(\Sigma^*, n-1)$). The p-value will depend on the true population parameters, the sample size, and the proposed weighting method. Naturally, we find that the p-value decreases as the Euclidean distance between the observed bias and covariance matrix and the null bias and covariance matrix increases. If the p-value is sufficiently small, we reject the null hypothesis that the simple average would outperform the weighted average.

4.2. Illustrative Example

For additional clarity, we provide an illustrative example of our algorithm by testing the optimally weighted average against the simple average. For ease of visualization, we assume there are two zero-bias judges (i.e., $M = 2$) in this example. This simplified illustration is meant to convey the intuition underlying the test, but the test can easily be applied in more general contexts (i.e., biased and correlated judgments from multiple judges).

We simulate 10 judgments for each of two judges from a multivariate normal distribution MVN$(\mathbf{0}, \Sigma)$ given the following true covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 4 \end{bmatrix}$$

Then the sample covariance matrix is estimated according to Eq. (7), and the weights can be solved through Eq. 3 by replacing the true covariance matrix with the sample covariance matrix. We utilize our algorithm to solve Σ^* numerically, and then calculate the p-value by drawing Wishart samples based on Σ^*.

All related covariance matrices are visualized in a two-dimensional graph with the x-axis representing the log ratio of standard deviation (i.e., σ_2/σ_1) and the y-axis representing the correlation (i.e., ρ_{12}). Figure 2 provides an example of such illustration. The red star represents the true covariance matrix Σ. Given Σ, we simulate individuals' judgments and estimated the sample covariance

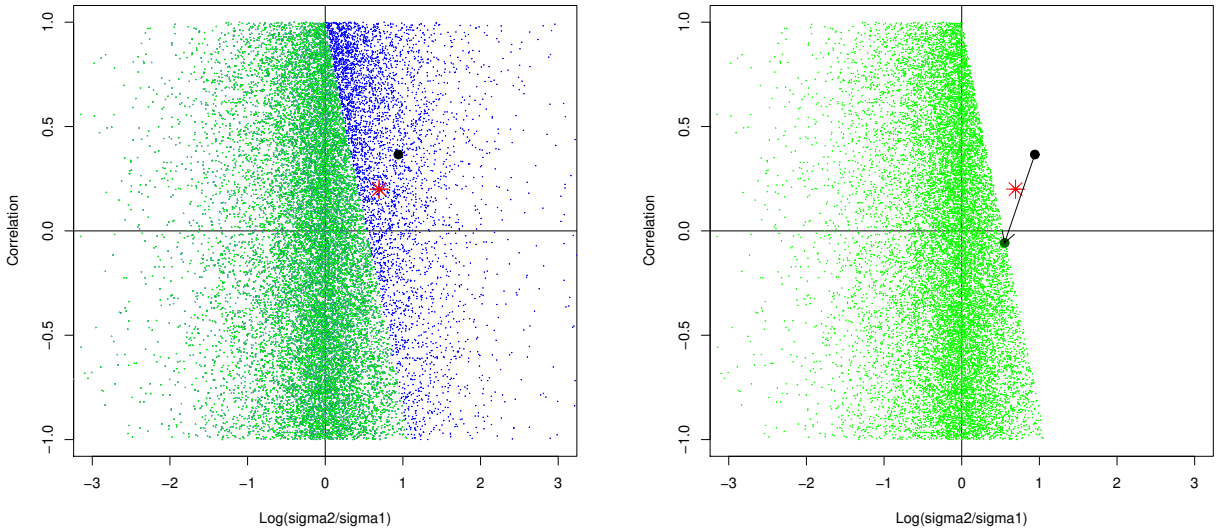
matrix $\hat{\Sigma}$ represented by the black dot. Based on $\hat{\Sigma}$, we find the feasible area (covered by green dots in (a)) for the most representative covariance matrix Σ^* indicating the simple average would outperform the optimally weighted average. In (b), the dark green dot is Σ^*, the most likely covariance matrix to generate the sample covariance matrix such that the simple average outperforms the optimally weighted average. Finally, independent samples (represented by gray dots in (c)) are drawn from a Wishart distribution given Σ^* as the scale matrix with $n-1$ degrees of freedom. The corresponding p-value is calculated as 0.3430, which guides us to not reject the simple average in this situation.

There are two determinants for this p-value calculation. One is the true covariance matrix. The p-value is expected to decline as the true covariance matrix gets further away from the threshold line shown in Figure 2(b). The other is the sample size, impacting the p-value calculation by reducing the variance in the sampling distributions of the sample covariance matrix shown in Figure 2(c). These mechanisms play a similar role in determining the p-value in the more general case with judgment bias.

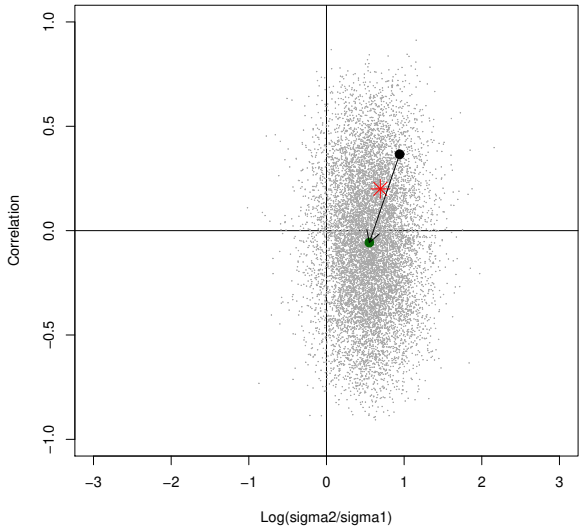
4.3. Comparison with Cross Validation

Next, we compare our hypothesis test algorithm to cross validation using simulated data. Cross validation provides no signal to users about how representative current samples are, which may lead to unreliable conclusions with small sample sizes, whereas our test algorithm may be more reliable with small samples because it relies on statistical inference based on parametric structure.

We use both methods to compare a variety of different weighting methods to the simple average. Besides the theoretically optimal weighting method (OW, as shown in Eq. (2)), we also consider using the regularized estimator of the covariance matrix (tuning the parameters through cross validation, (Fang et al. 2016)) to compute the optimal weights (RegCov), the constrained LASSO method (LAS) (James et al. 2020), and crowd selection methods including taking just the top three forecasters (Top3), the Ranked Performance method with an endogenous number of top forecasters (RP) (Mannes et al. 2014), Contribution Weighted Model (CWM) (Budescu and Chen 2014),



(a) (b)



(c)

Figure 2 **Illustration of hypothesis test algorithm with two zero-bias judges: (a) Red star indicates the true covariance matrix, and Black dot is the sample covariance matrix. Blue dots represent all positive semi-definite matrices in current searching area, and Green dots are matrices that not only satisfy the positive semi-definite condition but also suggest the simple average would outperform the optimally weighted average (i.e., candidates of Σ^*); (b) The dark green dot represents Σ^*; (c) Gray dots are Wishart samples given Σ^* as the true covariance matrix.**

and a Sequential Search method (both the increasing sequential search (SSIN) and the decreasing sequential search (SSDE)) (Olsson and Loveday 2015). Aligning with previous research (Schmittlein et al., 1990; Mannes et al., 2014), we conduct the simulation by taking five judges (i.e., $M = 5$). We generate simulated judgments from a multivariate normal distribution, MVN(μ, Σ), given true bias (μ) and covariance matrix (Σ) parameters. Based on these simulated judgments, we apply our hypothesis test algorithm and cross validation to decide whether to use a weighted average or the simple average. Specifically, we check the out-of-sample MSEs of all weighting methods on held-out testing data, and then compare the Hit Rate (HR) and the False Alarm Rate (FAR) of our algorithm and cross validation. We explore the performance of our algorithm across significance thresholds and present Receiver Operating Characteristic (ROC) curves to fully characterize the detecting capacity of the test algorithm.

We consider a variety of cases of the true state of the world to ensure that there are multiple scenarios in which weighted averages can outperform the simple average and corresponding scenarios in which the simple average is best. Table EC.1 (see EC.3.) presents the precise values of the assumed true biases, variances, and correlation matrices and the corresponding true optimal weights for different cases that we simulate. For half of the trials, the true optimal weights are equal weights and for the other half of the trials, the true optimal weights are approximately $w = [.436, .246, .170, .093, .055]$.

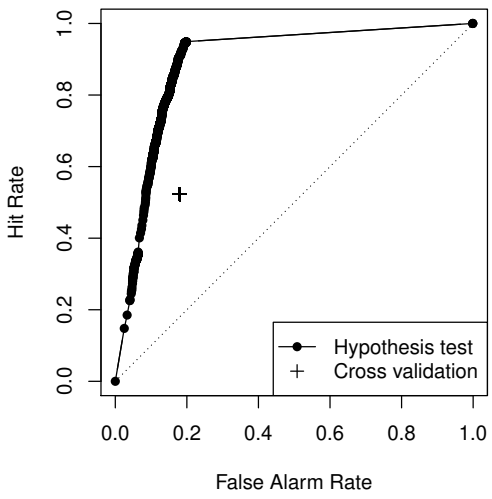
We change the sample size (i.e., the total number of judgments from each individual) from 32 to 256 following rule $2^k, k = 5, 6, 7, 8$. For each set of simulated judgments, we apply our algorithm and cross validation to test each of the weighted averages respectively against the simple average. For our hypothesis test, if the p-value is less than the significance threshold, we consider the test to support the corresponding weighted average, otherwise it sticks with the simple average. For the cross validation, if the average validation error of a weighted average after 100-time and 5-fold validation is lower than that of the simple average, we would say the cross validation supports the weighted average, otherwise the simple average.

For each case that we consider and each sample size, we estimate the FAR and HR for our algorithm and cross validation using 1000 simulation runs. We determine the FAR and HR with 1000 held-out observations, used to evaluate which aggregation method actually performs better. If the simple average (weighted average) performs better on this held-out sample, then deciding to use the weighted average leads to a false alarm (hit). We consider the weighted average to outperform the simple average (or vice versa) if it has significantly lower MSE (in a paired t-test with significance threshold 0.01) on the held-out testing data. In a small number of simulation runs, there may be no significant difference between the weighted and simple average (typically when the weighting rule happens to assign equal weights). We exclude those inconclusive simulation runs when computing the FAR and HR.

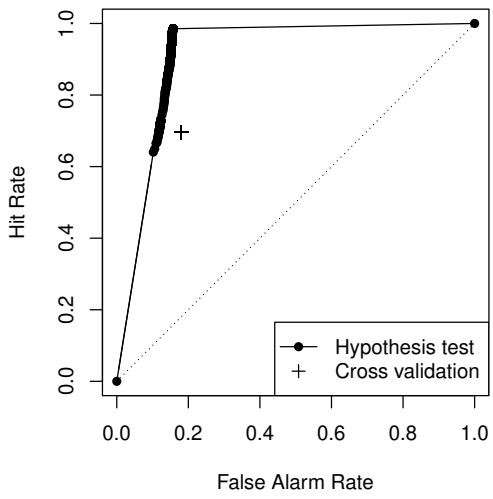
Figure 3 displays the detecting performance of our hypothesis test algorithm (varying the p-value criterion from 0 to 1) and cross validation with different sample sizes in the comparison of the theoretically optimal weights to equal weights. A method with a lower FAR, higher HR and larger Area Under ROC Curve (AUC) would be considered a better detector for the true state of world. We find that our hypothesis test outperforms cross validation at all sample sizes. As the sample size increases, the AUC of both our hypothesis test and cross validation gradually increases, reflecting a more confident decision about which aggregation rule should be applied given the simulated observations.

Table 1 Detecting performance of hypothesis test (taking 0.05 as the p-value criterion) and cross validation across all sample sizes and true states of world for different weighted averages

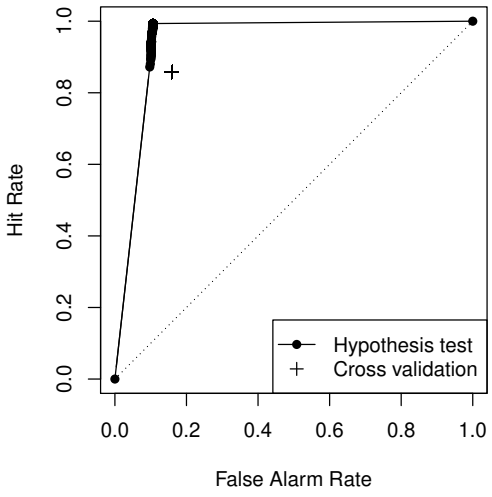
Criterion	Algorithm	Weighting Method							
		TOP3	RegCov	SSDE	OW	SSIN	CWM	RP	LAS
FAR	Hypothesis Test	0.032	0.023	0.093	0.109	0.422	0.411	0.517	0.616
	Cross validation	0.092	0.133	0.173	0.171	0.360	0.371	0.353	0.372
HR	Hypothesis Test	0.858	0.794	0.883	0.914	0.854	0.812	0.837	0.866
	Cross validation	0.827	0.776	0.792	0.802	0.797	0.758	0.814	0.847



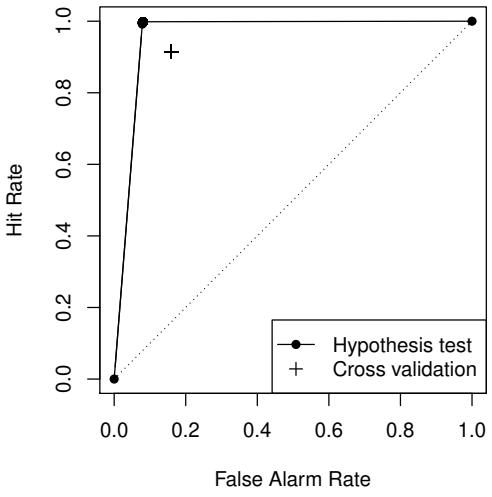
(a) Sample size $n = 32$



(b) Sample size $n = 64$



(c) Sample size $n = 128$



(d) Sample size $n = 256$

Figure 3 ROC curves of hypothesis test and cross validation in the comparison of the optimally weighted average and the simple average

The detecting capacity of our hypothesis test and cross validation varies when we compare the simple average to different weighted averages (as shown in Table 1). The corresponding ROC curves are shown in Figure 4. The hypothesis test and cross validation both performed well in

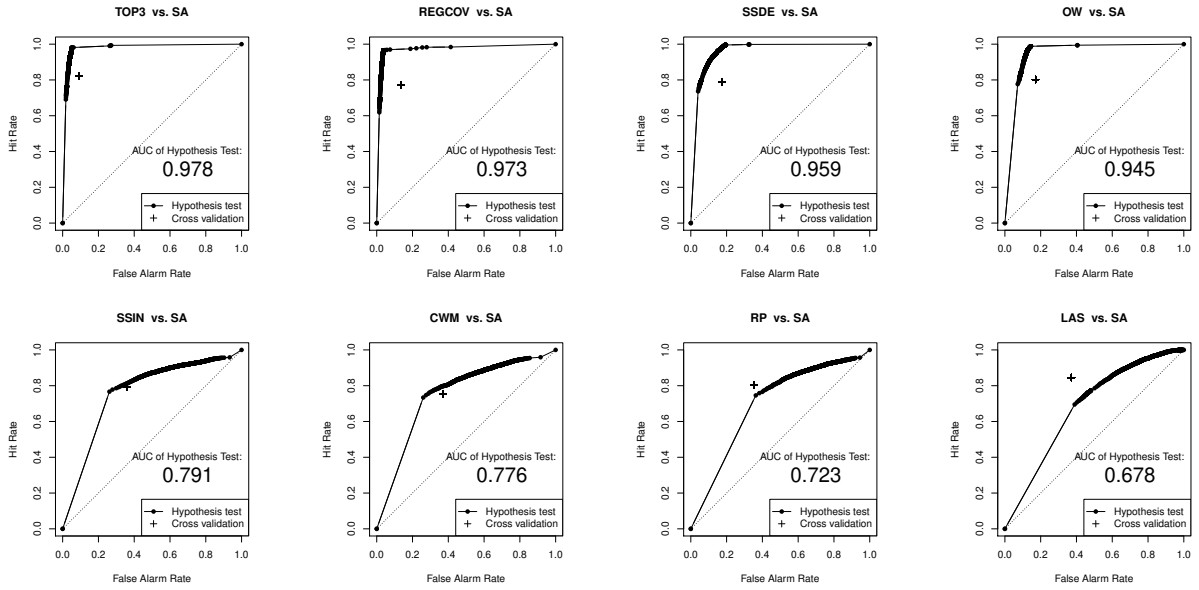


Figure 4 ROC curves of hypothesis test and cross validation for different weighting method across all sample sizes and cases of true state of world

discriminating between using TOP3, SSDE, RegCov, and OW and using the simple average. Our test has even lower FARs and higher HRs, i.e, better performance, than cross validation. Overall performance was much lower for discriminating between using SSIN, CWM, RP and LAS and using the simple average. We believe this is due to the latter group of methods generating weights that are very close to equal. Figure EC.1 (see EC.3.) displays more details on percentages of inconclusive simulation runs, which are due to weighting methods producing equal weights. For RP, SSIN and CWM, there are more than 50% inconclusive comparisons to the simple average (and for LAS more than 20%), including many situations where the weights output by these weighting methods are exactly equal weights. Therefore, we provide additional results related to performance with different sample sizes for these two groups of weighting methods separately.

Table 2 presents the FAR and HR at different sample sizes for TOP3, SSDE, RegCov, and OW. Compared to cross validation, our test has lower FARs for all sample sizes and higher HRs for 3 out of 4 sample sizes. Both methods perform better as the sample size increases. Table 3 presents the FAR and HR ad different sample sizes for SSIN, CWM, RP and LAS. Compared to cross validation, our test has lower FAR and lower HR with a small sample size ($n = 32$), and higher

Table 2 Detecting performance of hypothesis test (taking 0.05 as the p-value criterion) and cross validation across weighted averages (TOP3, RegCov, SSDE, OW) with different sample sizes

Detecting Criterion	Algorithm	Sample Size			
		32	64	128	256
FAR	Hypothesis Test	0.075	0.087	0.046	0.024
	Cross validation	0.202	0.156	0.113	0.075
HR	Hypothesis Test	0.528	0.829	0.987	0.999
	Cross validation	0.625	0.739	0.849	0.910

Table 3 Detecting performance of hypothesis test (taking 0.05 as the p-value criterion) and cross validation across weighted averages (SSIN, CWM, RP, LAS) with different sample sizes

Detecting Criterion	Algorithm	Sample Size			
		32	64	128	256
FAR	Hypothesis Test	0.266	0.731	0.909	0.946
	Cross validation	0.358	0.357	0.395	0.427
HR	Hypothesis Test	0.477	0.823	0.981	1.000
	Cross validation	0.659	0.763	0.845	0.913

FARs and higher HRs with larger sample sizes. Surprisingly, FARs increase with sample size (for both methods). However, as shown in Tables EC.2 - EC.9 in Section EC.3, the absolute number of false alarms remains low because it is quite rare that the simple average significantly outperforms these weighting methods with larger sample sizes. The vast majority of these trials are hits or inconclusive, because the weighting methods typically produce equal weights when equal weights are best and tend not to overshoot when improvement is possible. While this makes it harder to discriminate between these weighted averages and the simple average, it also makes the task of choosing between them somewhat moot.

On the whole, the results of our simulation demonstrate that our hypothesis test performs as well, and sometimes better, than cross validation. As expected, our test less frequently rejects

the simple average than cross validation when the sample size is small for all proposed weighted averages, resulting in a lower FAR, while still achieving a higher HR in some cases. These results also show that the number of necessary judgments depends on the data context as well as the robustness of the weighting method.

4.4. Application to Empirical Judgment Data

We now demonstrate the application of our algorithm to real data to gauge the number of judgments that are sufficient for different weighting methods in one particular, naturalistic context. We analyze a publicly available data set of judgments included in the ECB's Survey of Professional Forecasters (SPF), where professional economic forecasters are organized as experts in their fields to give forecasts on real GDP growth, CPI and unemployment rate. This data set has been used to demonstrate the outperformance of crowd wisdom compared to traditional macroeconomic forecasting methods (Ang et al. 2007, Budescu and Chen 2014, Genre et al. 2013). The data can be extracted from the publicly available databases of the European Central Bank (http://www.ecb.europa.eu).

We seek to show how to apply the hypothesis test algorithm to assess a variety of popular weighting methods given multiple judges and their historical judgments, but we do not attempt to identify the best weighting method or the best aggregate forecasts. Thus, we just use a subset of forecasters with enough shared forecasts (i.e., forecasts of the identical target value in the same prediction and targeted time period) to validate our hypothesis test algorithm, and we sidestep the issue of missing forecasts in the SPF dataset. We filter the data by: (1) only including the prediction for one or two years ahead indicators during the time period 1999-2018; (2) excluding predictions for indicators in crisis years (2008 and 2009); (3) excluding forecasters with more than 90% missing data; (4) filling in remaining missing data by using the AR(1) process proposed in Genre et al. (2013), which assumes that the relative deviation of each forecaster from the simple average of all forecasters in the current period is linked to its relative deviation in the previous period; and (5) excluding unpredictable missing data, e.g., when forecasters did not provide their predictions of the same indicators for more than 2 years. Finally, we are able to find two appropriate sets of

data from EU SPF: One is for the unemployment rate with 8 forecasters and their 222 forecasts (134 1-year-ahead forecasts and 88 2-year-ahead forecasts), and the other is for the inflation rate including 9 forecasters and their 218 forecasts (133 1-year-ahead forecasts and 85 2-year-ahead forecasts).

First, we randomly draw 80% of each individual's judgments as the training data and use the remaining 20% as testing data. Within the training data, a random subset of judgments with different sample size ($n = 16, 32, 48, ..., 80$) are selected to conduct the hypothesis test to decide whether using a weighted average rather than the simple average. We investigate weighted averages based on OW and CWM that were discussed in Section 4.3. The estimated weights from each weighting method are then applied to the testing data and compared to the simple average. By changing the number of judgments from each forecaster in the training dataset for 100 times, we observe how often we can reject using the simple average in favor of the weighted average, as well as how often the choice to use the weighted average would have worked out, i.e., yielded lower out-of-sample Mean Squared Error (MSE) on the testing dataset. We repeat this analysis for five random splits of the dataset into training data and testing data.

Figures 5 and 6 respectively present the proportion of rejections of the simple average at a significance level of 0.05 for different sample sizes in the inflation rate data and unemployment rate data, with shading indicating how many of these rejections were correct based on out-of-sample prediction. Genre et al. (2013) have demonstrated that the simple average is more robust for the unemployment rate data than for the inflation rate data. Our results are consistent with their findings, as for the inflation rate data our hypothesis test begins to (correctly) reject the simple average very often with fewer samples than for the unemployment rate data.

Overall, when the sample size is small, our algorithm provides high p-values such that the total rate of rejection of the simple average is low. As sample size increases, the rate of rejection of the simple average increases rapidly, with most of these rejections ultimately being correct (i.e., in accordance with out-of-sample accuracy). The results verify that the estimated bias and covariance

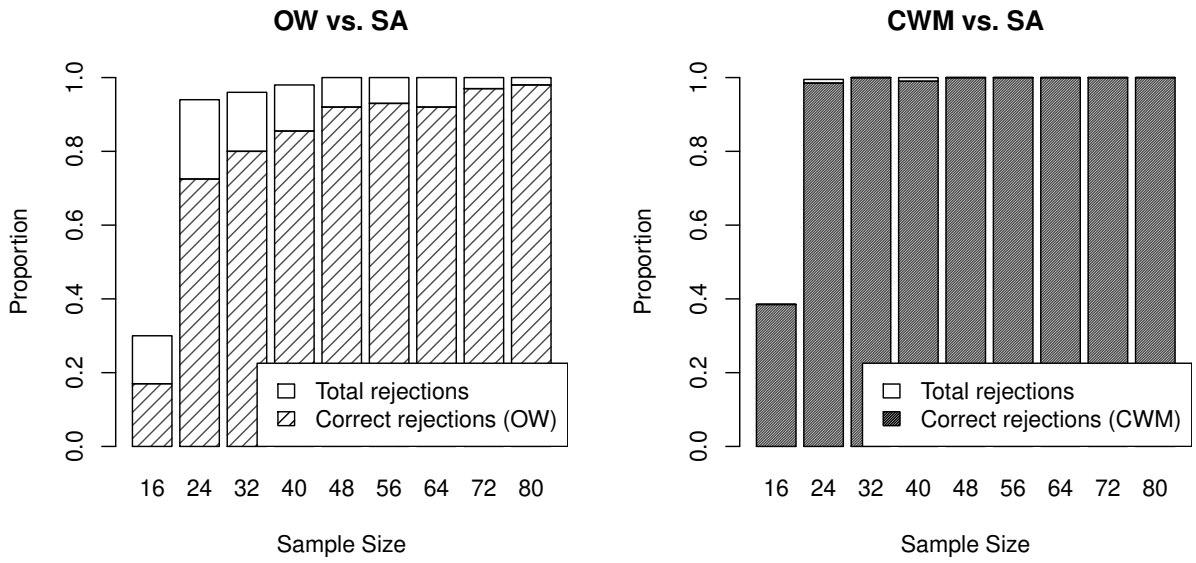


Figure 5 **Proportion of rejection of the simple average given the p-value criterion $\alpha = 0.05$ and proportion of of which are correct based on out-of-sample prediction, for different sample sizes and weighted averages by using the inflation rate data.**

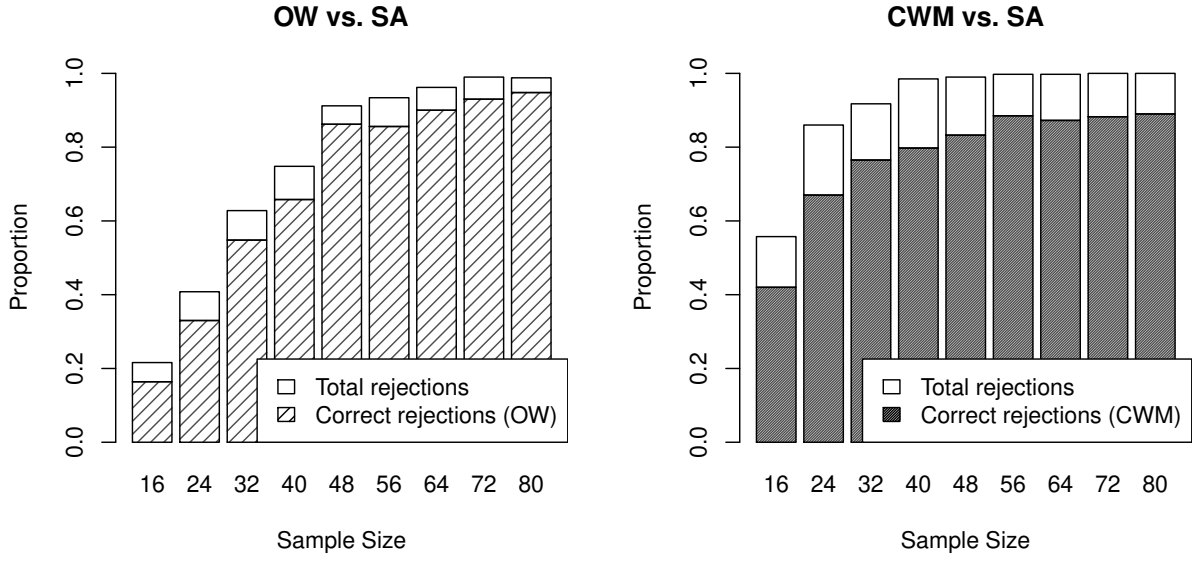


Figure 6 **Proportion of rejection of the simple average given the p-value criterion $\alpha = 0.05$ and proportion of which are correct based on out-of-sample prediction, for different sample sizes and weighted averages by using the unemployment rate data.**

matrix may provide useful information about the true state of the world if we have large enough sample sizes to be confident that sampling errors will not lead us astray.

We can also examine how many judgments are necessary to trust these weighted averages. In the inflation rate context, only 32 judgments (per forecaster) seem to be sufficient for us to trust the optimally weighted average rather than the simple average, and even a smaller sample size is sufficient for CWM. As the sample size increases, those weighted averages perform accurately and robustly. In the unemployment rate context, we find that the optimal weights become reliable as the sample size increases, with 48 judgments (per forecaster) usually sufficient for us to trust using it. The contribution weighted model becomes fairly reliable with at least 40 judgments.

We have demonstrated that our hypothesis test can support a decision whether to use a weighted average or the simple average given specific observations, and also can be used to gauge the sufficient sample size for different weighting models in a given data context. In EC.4., we also compare our test algorithm to cross validation using the real SPF data. While most of the trials are inconclusive, we find that our hypothesis test algorithm performs similarly to cross validation with a larger sample size and performs better than cross validation with a smaller sample size.

5. Discussion

Previous literature has provided a variety of weighting models in the field of wisdom of crowds and judgment aggregation, but few of them offer a systematic decision rule to determine whether it is appropriate to use such a weighting method rather than the simple average given current observations. In this paper, we propose a hypothesis test algorithm to assist the decision maker to decide whether to use a weighted average based on the observable data by establishing whether it will be reliably more accurate than the simple average. This test can be applied to different weighting models in which judgment weights are computed from estimates of judges' ability and correlation (e.g., judgment bias and covariance matrix). We believe this is a necessary step before decision makers decide to use any non-equal weighting models.

Our hypothesis test algorithm can only compare a proposed weighted average scheme to the simple average, and cannot determine whether alternative weighting schemes might be the most

accurate. However, a decision maker can test all candidate models for weighted averages, and only consider the application of those that reject equal weights according to our hypothesis test algorithm, and family-wise error rate corrections can be applied for the use of multiple tests. For example, a decision maker might find that optimal weights suffer from sample error and fail to reject the null hypothesis. Then she or he could test regularization methods that balance sampling error by reducing the total number of parameters to estimate, thereby increasing the robustness of the estimated weights. There is no guarantee that this weighting method will be more reliably accurate than equal weights in a given environment. However, our hypothesis test algorithm can provide a quantitative measure of the reliability of this weighted average, reflecting how confident the decision maker could be to trust this weighting strategy.

Our hypothesis test algorithm provides a general approach to prevent overfitting, which is inspired by the imperfect performance of the theoretically optimal weighting method. Optimal weights for crowd wisdom have been established theoretically in prior work (Davis-Stober et al. 2014, 2015, Lamberson and Page 2012) but are not as prevalent as equal weights in practice. Weighting crowds based on observable data may deliver sub-optimal performance because the conditions for optimal performance have not been clearly defined and used to identify situations where weighting should actually work (though for the case of 2 forecasters, see Schmittlein et al. (1990) and Winkler and Clemen (1992)). Although the optimal weighting method based on observable data performs well retrospectively (i.e., when evaluated on the same judgments used to estimate the statistical properties taken as inputs into the optimal weights), insufficient sample size often introduces poor estimates of the statistical properties of the judgments and distorts the estimated weights, resulting in an unreliable out-of-sample performance of the theoretically optimal weighting method. With our method, we now have a way of determining when the optimal weighted average computed with empirical estimates of judgment biases, variances, and correlations will likely outperform the simple average. Thus, our parametric approach guards against overfitting in judgment aggregation.

In our hypothesis test algorithm, we assume there exists a judgment generation model with fixed true parameters, and project the information in the sample estimates onto candidate parameters, which describe our knowledge about the future observations. The null hypothesis specifies candidate parameters, a bias and covariance matrix determined by the sample bias and covariance matrix, reflecting a world in which weighting based on the sample parameters would be less accurate than the simple average. Overfitting is prevented by our algorithm by looking for sample parameters that have a small probability of being encountered under the null hypothesis that equal weighting will have less expected error. When the sample size is too small, the sampling distributions will be wide, making a large collection of sample parameters more likely under the null. On the other hand, when the sample size becomes large, our test algorithm only rejects the simple average when the sampling distribution of the estimated weights indicate that members of the crowd are in-fact different. Rejecting the null hypothesis provides a strong signal that the crowd should be weighted according to the observed statistical properties of the judges.

Cross validation is a common approach to prevent overfitting. In cross validation, there is no assumption of an underlying judgment model and the decision rule is only based on the realized error. Therefore, cross-validation can be used in situations where researchers are unwilling to make assumptions about the parametric structure of their crowd. However, the parametric structure that we introduce has some advantages over cross validation because it facilitates crowd design, experiments, and data collection efforts by making clear what aspects of judgment affect aggregate performance, where cross-validation is limited to a given data set. We describe two advantages of our proposed model in detail.

First, cross-validation may be unreliable with small sample sizes (Piironen and Vehtari 2017). When the variance of the true environment or judgments is large, a small set of unrepresentative samples can potentially mislead a decision maker relying on cross validation. Our model-based hypothesis test is more conservative with small samples because we assume the sampling distribution of the sample covariance matrix is a Wishart distribution where a small sample size makes Wishart samples spread broadly around the observations.

Second, our hypothesis test algorithm can facilitate power analysis, and assist in the planning of data collection. Like other hypothesis tests, our algorithm can help to determine the number of judgements that would be needed to trust a weighted average given an initial guess about the true judgment parameters (i.e., how variable or different the judges are). Such a framework can be used to generate more powerful studies of information aggregation by facilitating the identification of contexts and data sets with enough estimation precision for weighted averages to potentially work well.

In summary, our framework facilitates the weighting of crowds by warning us when we are in danger of overfitting. We present a hypothesis test algorithm that other researchers can download and easily use in the free statistical platform R. This test can be interpreted in the same way as other familiar statistical tests. Our freely available algorithm can be applied generally to decide when to use a weighted average of judgments from a crowd to reliably generate better, more accurate forecasts.

References

Ang A, Bekaert G, Wei M (2007) Do macro variables, asset markets, or surveys forecast inflation better? *Journal of monetary Economics* 54(4):1163–1212.

Bansal S, Gutierrez GJ, Keiser JR (2017) Using experts' noisy quantile judgments to quantify risks: Theory and application to agribusiness. *Operations Research* 65(5):1115–1130.

Blanc SM, Setzer T (2016) When to choose the simple average in forecast combination. *Journal of Business Research* 69(10):3951–3962.

Broomell SB, Budescu DV (2009) Why are experts correlated? decomposing correlations between judges. *Psychometrika* 74(3):531–553.

Budescu DV, Chen E (2014) Identifying expertise to extract the wisdom of crowds. *Management Science* 61(2):267–280.

Clemen RT, Winkler RL (1985) Limits for the precision and value of information from dependent sources. *Operations Research* 33(2):427–442.

Clemen RT, Winkler RL (1986) Combining economic forecasts. *Journal of Business & Economic Statistics* 4(1):39–46.

Davis-Stober CP, Budescu DV, Broomell SB, Dana J (2015) The composition of optimally wise crowds. *Decision Analysis* 12(3):130–143.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1(2):79.

Davis-Stober CP, Dana J, Budescu DV (2010) A constrained linear estimator for multiple regression. *Psychometrika* 75(3):521–541.

Dawes RM (1979) The robust beauty of improper linear models in decision making. *American psychologist* 34(7):571.

Einhorn HJ, Hogarth RM (1975) Unit weighting schemes for decision making. *Organizational behavior and human performance* 13(2):171–192.

Fang Y, Wang B, Feng Y (2016) Tuning-parameter selection in regularized estimations of large covariance matrices. *Journal of Statistical Computation and Simulation* 86(3):494–509.

Gaines BR, Kim J, Zhou H (2018) Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics* 27(4):861–871.

Genre V, Kenny G, Meyler A, Timmermann A (2013) Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* 29(1):108–121.

Hong L, Page SE (2008) Some microfoundations of collective wisdom. *Collective Wisdom* 56–71.

James GM, Paulson C, Rusmevichientong P (2020) Penalized and constrained optimization: An application to high-dimensional website advertising. *Journal of the American Statistical Association* 115(529):107–122.

Kang H (1986) Unstable weights in the combination of forecasts. *Management Science* 32(6):683–695.

Lamberson P, Page SE (2012) Optimal forecasting groups. *Management Science* 58(4):805–810.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science* 52(1):111–127.

Makridakis S, Winkler RL (1983) Averages of forecasts: Some empirical results. *Management Science* 29(9):987–996.

Mannes AE, Larrick RP, Soll JB (2012) The social psychology of the wisdom of crowds. *In J. I. Krueger (Ed.), Frontiers of social psychology. Social judgment and decision making* 227–242, Psychology Press.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *Journal of personality and social psychology* 107(2):276.

Merkle E, Saw G, Davis-Stober CP (2020) Beating the average forecast: Regularization based on forecaster attributes. *Journal of Mathematical Psychology* .

Minson JA, Mueller JS, Larrick RP (2017) The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science* 64(9):4177–4192.

Morrison DG, Schmittlein DC (1991) How many forecasters do you really have? mahalanobis provides the intuition for the surprising clemen and winkler result. *Operations Research* 39(3):519–523.

Olsson H, Loveday J (2015) A comparison of small crowd selection methods. *CogSci*.

Palley AB, Soll JB (2019) Extracting the wisdom of crowds when information is shared. *Management Science* 65(5):2291–2309.

Piironen J, Vehtari A (2017) Comparison of bayesian predictive methods for model selection. *Statistics and Computing* 27(3):711–735.

Satopää VA (2017) Combining information from multiple forecasters: Inefficiency of central tendency. *arXiv preprint arXiv:1706.06006* .

Schmittlein DC, Kim J, Morrison DG (1990) Combining forecasts: Operational adjustments to theoretically optimal rules. *Management Science* 36(9):1044–1056.

Stock JH, Watson MW (2004) Combination forecasts of output growth in a seven-country data set. *Journal of forecasting* 23(6):405–430.

Surowiecki J (2005) *The wisdom of crowds* (Anchor).

Winkler RL, Clemen RT (1992) Sensitivity of weights in combining forecasts. *Operations research* 40(3):609–614.

Yaniv I (1997) Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational behavior and human decision processes* 69(3):237–249.

Electronic Companions

All proof, deviation and supplementary information of simulation and empirical analysis are provided here below.

EC.1. Comparing Simple Average and Weighted Average in the Limiting Case

We ideally would like to find the minimum sufficient number of each individual's judgments that ensures that the optimal weighting method outperforms the equal weighting method. In some special cases, however, the perfect aggregation method is exactly the equal-weighting method. For instance, when all judges have zero bias and the true covariance matrix implies exchangeability among judges, then the weighted average cannot be guaranteed more reliably accurate than the simple average since estimation errors are unavoidably involved in the estimated weights used to compute the weighted average. We define it the "worst" case for the optimal weighting method because even infinite sample size can only ensure that the weighted average matches the simple average, and any finite sample size, no matter how large, cannot guarantee that the weighted average is as good.

When the true judgment parameters are unknown, we prove that in the limiting case the weighted average can has a smaller expected squared error than that of the simple average. Here we start with a simple case where all judges have zero bias, and then the expected SE becomes the error variance (i.e., $E(f-y) = Var(f-y)$).

- Proof : $\lim_{n\to\infty} Var(f_w - y|\hat{\Sigma}(n,\Sigma)) \leq Var(f_s - y)$

  As $n \to \infty$, $\hat{\Sigma} \to \Sigma$, then

$$Var(f_w - y|\hat{\Sigma}(n,\Sigma)) \to \frac{\mathbb{1}^T\Sigma^{-1}\Sigma\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}\mathbb{1}^T\Sigma^{-1}\mathbb{1}} = \frac{1}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}$$

Thus, we want to show $\mathbb{1}^T\Sigma^{-1}\mathbb{1}\mathbb{1}^T\Sigma\mathbb{1} \geq M^2$.

Since Σ is positive definite, we can have $\Sigma = QDQ^T$ where Q is the orthonormal matrix and D is diagonal matrix. Then $\Sigma^{-1} = QD^{-1}Q^T$ and D^{-1} is just the inverse of the diagonal elements of D.

Thus, $\mathbb{1}^T\Sigma\mathbb{1} = (\mathbb{1}^TQ)D(\mathbb{1}^TQ)^T$ and $\mathbb{1}^T\Sigma^{-1}\mathbb{1} = (\mathbb{1}^TQ)D^{-1}(\mathbb{1}^TQ)^T$. If we write $\mathbb{1}^TQ = \sum a_i$ where $a_i, i = \{1, ..., M\}$ is elements of Q, then we have $\mathbb{1}^T\Sigma\mathbb{1} = \sum a_i^2 d_i$ and $\mathbb{1}^T\Sigma^{-1}\mathbb{1} = \sum \frac{a_i^2}{d_i}$ where d_i is the ith diagonal element of D. By the Cauchy-Schwarz inequality, we obtain

$$\mathbb{1}^T\Sigma\mathbb{1}\mathbb{1}^T\Sigma^{-1}\mathbb{1} = (\sum a_i^2 d_i)(\sum \frac{a_i^2}{d_i}) \geq (\sum (a_i\sqrt{d_i})(\frac{a_i}{\sqrt{d_i}}))^2 = (\sum a_i^2)^2$$

Since we have $\sum a_i = M$ due to the property of orthonormal matrix Q, thus the smallest value of $\sum a_i^2$ would be M when $a_i = 1, \forall i$. Finally, we have proved $\mathbb{1}^T\Sigma^{-1}\mathbb{1}\mathbb{1}^T\Sigma\mathbb{1} \geq M^2$. ■

EC.2. Derivations

- Derivation of the conditional error variance of the optimally weighted average:

$$\begin{aligned} Var(f_w - y|\hat{\Sigma}) &= E[(f_w - y)^2|\hat{\Sigma}] - (E[f_w - y|\hat{\Sigma}])^2 \\ &= \mathbf{w}^T(\hat{\Sigma}) \cdot \Sigma \cdot \mathbf{w}(\hat{\Sigma}) \\ &= (\frac{\mathbb{1}^T\hat{\Sigma}^{-1}}{\mathbb{1}^T\hat{\Sigma}^{-1}\mathbb{1}}) \cdot \Sigma \cdot (\frac{(\hat{\Sigma}^{-1})^T\mathbb{1}}{\mathbb{1}^T\hat{\Sigma}^{-1}\mathbb{1}}) \\ &= \frac{\mathbb{1}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\mathbb{1}}{\mathbb{1}^T\hat{\Sigma}^{-1}\mathbb{1}\mathbb{1}^T\hat{\Sigma}^{-1}\mathbb{1}} \quad \text{(due to symmetric } \hat{\Sigma}^{-1}) \end{aligned}$$

- Derivation of constrained MLE problem in Eq. (10)

The objective function can be represented as follows:

$$\begin{aligned} g(\hat{\mu}, S|\mu^*, \Sigma^*, n) &= h_1(S|\Sigma^*, n-1) \cdot h_2(\hat{\mu}|\mu^*, \frac{\Sigma^*}{n}) \\ &= \frac{|S|^{\frac{n-M-2}{2}} \exp(-\frac{1}{2}tr((\Sigma^*)^{-1}S))}{2^{\frac{(n-1)M}{2}}|\Sigma^*|^{\frac{n-1}{2}}\Gamma_M(\frac{n-1}{2})} \\ &\cdot (2\pi)^{-\frac{M}{2}}|\frac{\Sigma^*}{n}|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\hat{\mu} - \mu^*)^T(\frac{\Sigma^*}{n})^{-1}(\hat{\mu} - \mu^*)) \end{aligned} \tag{EC.1}$$

Taking the log of the likelihood function:

$$\begin{aligned} &\log g(\hat{\mu}, S|\mu^*, \Sigma^*, n) \\ =&\frac{n-M-2}{2}\log(|S|) - \frac{1}{2}tr((\Sigma^*)^{-1}S) - \frac{(n-1)M}{2}\log 2 - \frac{n-1}{2}\log(|\Sigma^*|) - \log(\Gamma_M(\frac{n-1}{2})) \\ &- \frac{M}{2}\log(2\pi) - \frac{1}{2}\log(|\frac{\Sigma^*}{n}|) - \frac{1}{2}(\hat{\mu} - \mu^*)^T(\frac{\Sigma^*}{n})^{-1}(\hat{\mu} - \mu^*) \end{aligned} \tag{EC.2}$$

Given the observed bias and covariance matrix $\hat{\mu}$ and $\hat{\Sigma}$ ($S = n\hat{\Sigma}$), Eq. (EC.2) is also a function of μ^* and Σ^*, so we can obtain:

$$\begin{aligned}
&\log g(\hat{\mu}, S|\mu^*, \Sigma^*, n) \\
=&-\frac{1}{2}tr((\Sigma^*)^{-1}S) - \frac{n-1}{2}\log(|\Sigma^*|) - \frac{1}{2}\log(|\frac{\Sigma^*}{n}|) - \frac{1}{2}(\hat{\mu}-\mu^*)^T(\frac{\Sigma^*}{n})^{-1}(\hat{\mu}-\mu^*) + C \\
=&-\frac{1}{2}tr((\Sigma^*)^{-1}(n\hat{\Sigma})) - \frac{n}{2}\log(|\Sigma^*|) - \frac{n}{2}(\hat{\mu}-\mu^*)^T(\Sigma^*)^{-1}(\hat{\mu}-\mu^*) + C'
\end{aligned} \tag{EC.3}$$

where C and C' are constants.

As for the constraint, from the linear system in Eq. (2) we can solve a more general weights represented by the observed bias and covariance matrix when assuming the target value is invariant, that is, $\hat{w}_{\hat{\mu},\hat{\Sigma}} = \frac{\mathbb{1}^T(\hat{\mu}\hat{\mu}^T+\hat{\Sigma})^{-1}}{\mathbb{1}^T(\hat{\mu}\hat{\mu}^T+\hat{\Sigma})^{-1}\mathbb{1}}$. Meanwhile, the expected SE will count two terms, squared bias and error variance. Thus, the constraint can be represented as follows:

$$\begin{aligned}
&E[(f_s - y)^2] < E[(f_w - y)^2|\hat{\mu}, \hat{\Sigma}] \\
\Rightarrow&\frac{\mathbb{1}^T(\mu^*\mu^{*T} + \Sigma^*)\mathbb{1}}{M^2} < \hat{w}_{\hat{\mu},\hat{\Sigma}}^T(\mu^*\mu^{*T} + \Sigma^*)\hat{w}_{\hat{\mu},\hat{\Sigma}} \\
\Rightarrow&(\frac{\mathbb{1}}{M} - \hat{w}_{\hat{\mu},\hat{\Sigma}})^T(\mu^*\mu^{*T} + \Sigma^*)(\frac{\mathbb{1}}{M} + \hat{w}_{\hat{\mu},\hat{\Sigma}}) < 0
\end{aligned} \tag{EC.4}$$

EC.3. Supplementary of Simulation

True states of world in our simulation

We consider three benchmark cases where the true optimal weights are exactly the equal weights to explore type I errors: (1) Independent judgments with zero bias and identical variance; (2) Independent judgments with non-zero bias (but can be mutually canceled out) and identical variance; and (3) identically correlated judgments with zero bias and identical variance. We also consider three cases where weighting could potentially decrease MSE if the sample size is sufficient. To generate these three cases, we respectively change the true bias, variance and correlation matrix to achieve roughly the same set of non-equal true optimal weights.

Table EC.1 presents details of the assumed true biases, variances, and correlation matrices and the corresponding true optimal weights for different cases that we simulate. Case I, II, and III are benchmark cases where the true optimal weights are equal weights although they differ in bias or correlation matrix. Case IV, V, and VI are cases where the true optimal weights are far away from

the equal weights due to varied variances, bias and correlations. We utilize these six cases to create a balanced true state of world.

Table EC.1 True Bias, Variance and Correlation Matrix in Simulation ($M = 5$)

Case	Bias	Variance	Correlation Matrix	Optimal Weights
I	(0, 0, 0, 0, 0)	(1, 1, 1, 1, 1)	$I_{5\times5}$	(.2, .2, .2, .2, .2)
II	(-.2, -.1, 0, .1, .2)	(1, 1, 1, 1, 1)	$I_{5\times5}$	(.2, .2, .2, .2, .2)
III	(0, 0, 0, 0, 0)	(1, 1, 1, 1, 1)	$\begin{bmatrix} 1 & .1 & .1 & .1 & .1 \\ .1 & 1 & .1 & .1 & .1 \\ .1 & .1 & 1 & .1 & .1 \\ .1 & .1 & .1 & 1 & .1 \\ .1 & .1 & .1 & .1 & 1 \end{bmatrix}$	(.2, .2, .2, .2, .2)
IV	(0, 0, 0, 0, 0)	(.737, .945, 1, 4.934, 4.934)	$I_{5\times5}$	(.355, .277, .262, .053, .053)
V	(0, .5, .7, .9, 1)	(1, 1, 1, 1, 1)	$I_{5\times5}$	(.436, .246, .170, .093, .055)
VI	(0, 0, 0, 0, 0)	(1, 1, 1, 1, 1)	$\begin{bmatrix} 1 & .1 & .1 & .2 & .2 \\ .1 & 1 & .3 & .4 & .5 \\ .1 & .3 & 1 & .5 & .6 \\ .2 & .4 & .5 & 1 & .7 \\ .2 & .5 & .6 & .7 & 1 \end{bmatrix}$	(.355,.277,.262,.143,-.037)

Details of comparing weighted averages to the simple average

In our simulation, for each case of true state of world and each sample size, we compare one common weighted average to the simple average for 1000 times by applying our hypothesis test algorithm and the cross validation. Table EC.2-EC.9 show the type I and II error rates of hypothesis test and cross validation and the number of inconclusive cases respectively for each weighting method. Generally, our hypothesis test algorithm keeps a lower type I error rate than the cross validation, except in the comparison of the LASSO method and the simple average. But with the very small sample size (e.g., $n = 32$), our algorithm still behave conservatively to reject the simple average

Table EC.2 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the optimally weighted average and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	4		23		59		160		**246**	
I	Type I	23/996	117/996	74/977	130/977	70/941	137/941	43/840	118/840	**210/3754**	**502/3754**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	2		13		42		144		**201**	
II	Type I	26/998	115/998	77/987	147/987	76/958	131/958	64/856	139/856	**243/3799**	**532/3799**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	5		24		73		187		**289**	
III	Type I	42/995	143/995	118/976	170/976	108/927	162/927	91/813	139/813	**359/3711**	**614/3711**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	255		87		5		0		**347**	
IV	Type I	104/178	76/178	6/10	5/10	0/0	0/0	0/0	0/0	**110/188**	**81/188**
	Type II	101/567	268/567	4/903	201/903	0/995	43/995	0/1000	6/1000	**105/3465**	**518/3465**
	Inconclusive	359		265		64		1		**689**	
V	Type I	28/367	130/367	33/68	31/68	2/4	1/4	0/0	0/0	**63/439**	**162/439**
	Type II	234/274	121/274	228/667	243/667	18/932	144/932	0/999	51/999	**480/2872**	**559/2872**
	Inconclusive	226		476		549		348		**1599**	
VI	Type I	200/729	185/729	178/397	132/397	55/97	35/97	6/8	5/8	**439/1231**	**357/1231**
	Type II	25/45	34/45	17/127	72/127	16/354	137/354	4/644	168/644	**62/1170**	**411/1170**
	Inconclusive	**851**		**888**		**792**		**840**		**3371**	
Total	**Type I**	**423/4263**	**766/4263**	**486/3415**	**615/3415**	**311/2927**	**466/2927**	**204/2517**	**401/2517**	**1424/13122**	**2248/13122**
	Type II	**360/886**	**423/886**	**249/1697**	**516/1697**	**34/2281**	**324/2281**	**4/2643**	**225/2643**	**647/7507**	**1488/7507**

but trust the regularized weighting method. For the case IV and VI where the true non-equal optimal weights are derived from the varied covariance matrix, our algorithm performs a higher type I error rate than the cross validation but a lower type II error rate simultaneously. Another general conclusion is that as the sample size increases, both errors decrease, making the weighting methods as well as algorithms to test the reliability of the weighting methods more robust.

Figure EC.1 displays the percentages of trials that go to different conclusions for each weighting method across all simple sizes. In general, the increasing sequential search method (SSIN), the ranked performance model (RP) and the Contribution Weighted Model (CWM) and the constrained LASSO regression method (LAS) perform better than the simple average by comparing the

Table EC.3 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the equally weighting top 3 model and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	0		0		0		0		**0**	
I	Type I	22/1000	150/1000	3/1000	90/1000	0/1000	27/1000	0/1000	2/1000	**25/4000**	**269/4000**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	0		0		0		0		**0**	
II	Type I	18/1000	141/1000	4/1000	76/1000	0/1000	18/1000	0/1000	0/1000	**22/4000**	**235/4000**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	0		0		0		0		**0**	
III	Type I	50/1000	213/1000	26/1000	127/1000	3/1000	53/1000	0/1000	8/1000	**79/4000**	**401/4000**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	0		0		0		0		**0**	
IV	Type I	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Type II	167/1000	256/1000	6/1000	107/1000	0/1000	64/1000	0/1000	18/1000	**173/4000**	**445/4000**
	Inconclusive	296		217		110		44		**667**	
V	Type I	1/77	32/77	9/23	10/23	4/6	2/6	0/0	0/0	**14/106**	**44/106**
	Type II	547/627	209/627	283/760	227/760	26/884	159/884	1/956	124/956	**857/3227**	**719/3227**
	Inconclusive	276		286		282		239		**1083**	
VI	Type I	116/630	155/630	99/614	96/614	72/626	81/626	39/670	68/670	**326/2540**	**400/2540**
	Type II	42/94	44/94	7/100	41/100	0/92	35/92	0/91	34/91	**49/377**	**154/377**
	Inconclusive	**572**		**503**		**392**		**283**		**1750**	
Total	**Type I**	**207/3707**	**691/3707**	**141/3637**	**399/3637**	**79/3632**	**181/3632**	**39/3670**	**78/3670**	**466/14646**	**1349/14646**
	Type II	**756/1721**	**509/1721**	**296/1860**	**375/1860**	**26/1976**	**258/1976**	**1/2047**	**176/2047**	**1079/7604**	**1318/7604**

percentage difference at two sides, which represents the percentage of trials where simple average is significantly better than the weighted average (left) and vice versa (right). Using the regularized estimator of covariance matrix to compute weights (RegCov) cannot beat the simple average in most trials, becoming the least trustful weighted average followed by TOP3, OW and SSDE.

We can also find that TOP3, SSDE, RegCov and OW have less frequently inconclusive performance as the simple average, and our test algorithm outperforms cross validation in both Type I and Type II error rate. For RP, CWM and SSIN, that produce more frequently inconclusive comparisons (including many situations where weights output from these weighting methods are

Table EC.4 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the ranked performance model and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	789		920		981		1000		**3690**	
I	Type I	48/211	71/211	64/80	20/80	18/19	1/19	0/0	0/0	**130/310**	**92/310**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	788		926		991		1000		**3705**	
II	Type I	42/212	69/212	51/74	20/74	9/9	0/9	0/0	0/0	**102/295**	**89/205**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	617		812		941		999		**3369**	
III	Type I	95/383	140/383	137/188	59/188	53/59	11/59	1/1	0/1	**286/631**	**210/631**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	134		23		2		0		**159**	
IV	Type I	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Type II	248/866	242/866	8/977	122/977	0/998	68/998	0/1000	18/1000	**256/3841**	**450/3841**
	Inconclusive	414		281		142		54		**891**	
V	Type I	3/45	21/45	10/14	7/14	0/0	0/0	0/0	0/0	**13/59**	**28/59**
	Type II	467/541	219/541	321/705	226/705	91/858	186/858	0/946	119/946	**879/3050**	**750/3050**
	Inconclusive	538		625		684		735		**2582**	
VI	Type I	164/290	125/290	149/164	60/164	93/97	43/97	37/37	17/37	**443/588**	**245/588**
	Type II	88/172	72/172	26/211	73/211	8/219	57/219	1/228	37/228	**123/830**	**239/830**
	Inconclusive	**3280**		**3587**		**3741**		**3788**		**14396**	
Total	**Type I**	**352/1141**	**426/1141**	**411/520**	**166/520**	**173/184**	**55/184**	**38/38**	**17/38**	**974/1883**	**664/1883**
	Type II	**803/1579**	**533/1579**	**355/1893**	**421/1893**	**99/2075**	**311/2075**	**1/2174**	**174/2174**	**1258/7721**	**1439/7721**

exactly equal weights), our test has similar performance as cross validation, that is, slightly higher Type I error rate but lower Type II error rate. It's interesting to find that both our test and cross validation have many Type I errors when comparing the LAS to the simple average, and our test is even worse. By looking into Table EC.9, we could find this high Type I error rate only comes from the first three cases. When the sample size is very small (i.e., $n = 32$), our test could keep a lower Type I error rate than cross validation, but it rapidly increases when the sample size increases. We think this is because within the constrained LASSO regression model, overfitting problem has been controlled by choosing an appropriate penalty parameter through generalized cross validation, thus

Table EC.5 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the contribution weighted model and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	612		871		976		1000		**3459**	
I	Type I	89/388	144/388	97/129	40/129	20/24	2/24	0/0	0/0	**206/541**	**186/541**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	607		879		987		1000		**3473**	
II	Type I	84/393	132/393	87/121	39/121	12/13	0/13	0/0	0/0	**183/527**	**171/527**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	413		715		930		998		**3056**	
III	Type I	137/587	209/587	198/285	93/285	61/70	13/70	2/2	1/2	**398/944**	**316/944**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	616		559		481		417		**2073**	
IV	Type I	0/1	1/1	0/0	0/0	0/0	0/0	0/0	0/0	**0/1**	**1/1**
	Type II	98/383	101/383	1/441	68/441	0/519	73/519	0/583	45/583	**99/1926**	**287/1926**
	Inconclusive	349		288		161		119		**917**	
V	Type I	4/42	18/42	4/8	5/8	0/0	0/0	0/0	0/0	**8/50**	**23/50**
	Type II	532/609	219/609	309/704	218/704	31/839	187/839	0/881	113/881	**872/3033**	**737/3033**
	Inconclusive	322		244		174		140		**880**	
VI	Type I	105/211	115/211	73/118	60/118	29/49	32/49	11/24	11/24	**218/402**	**218/402**
	Type II	269/467	213/467	109/638	242/638	20/777	217/777	2/836	165/836	**400/2718**	**837/2718**
	Inconclusive	**2919**		**3556**		**3709**		**3674**		**13858**	
Total	**Type I**	**419/1622**	**619/1622**	**459/661**	**237/661**	**122/156**	**47/156**	**13/26**	**12/26**	**1013/2465**	**915/2465**
	Type II	**899/1459**	**533/1459**	**419/1783**	**528/1783**	**51/2135**	**477/2135**	**2/2300**	**323/2300**	**1371/7677**	**1861/7677**

our test and cross validation would prefer to accept the weighted average. Moreover, as the LAS gets closer to the simple average with more samples, our test would more support for the LAS model.

EC.4. Supplementary of Empirical Analysis

Comparing hypothesis test to cross validation

We apply our hypothesis test and cross validation to two sets of data from EU SPF (used in Section 4.4) and data sets from USA SPF.

Table EC.6 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the sequential search (increasing) method and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	620		872		976		1000		**3468**	
I	Type I	81/380	127/380	91/128	36/128	20/24	2/24	0/0	0/0	**192/532**	**165/532**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	625		880		987		1000		**3492**	
II	Type I	93/375	121/375	81/120	40/120	13/13	0/13	0/0	0/0	**187/508**	**161/508**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	431		717		927		998		**3073**	
III	Type I	138/569	202/569	193/283	97/283	61/73	13/73	2/2	1/2	**394/927**	**313/927**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	144		24		2		0		**170**	
IV	Type I	2/3	0/3	0/0	0/0	0/0	0/0	0/0	0/0	**2/3**	**0/3**
	Type II	228/853	242/853	9/976	126/976	0/998	68/998	0/1000	18/1000	**237/3827**	**454/3827**
	Inconclusive	416		297		148		76		**937**	
V	Type I	5/74	37/74	18/28	14/28	5/5	5/5	2/2	0/2	**30/109**	**56/109**
	Type II	444/510	216/510	311/675	222/675	10/847	185/847	0/922	122/922	**765/2954**	**745/2954**
	Inconclusive	326		322		261		215		**1124**	
VI	Type I	120/237	120/237	70/76	41/76	23/23	13/23	2/2	2/2	**215/338**	**176/338**
	Type II	257/437	202/437	94/602	209/602	9/716	163/716	1/783	123/783	**361/2538**	**697/2538**
	Inconclusive	**2562**		**3112**		**3301**		**3289**		**12264**	
Total	**Type I**	**439/1638**	**607/1638**	**453/635**	**228/635**	**122/138**	**33/138**	**6/6**	**3/6**	**1020/2417**	**871/2417**
	Type II	**929/1800**	**660/1800**	**414/2253**	**557/2253**	**19/2561**	**416/2561**	**1/2705**	**263/2705**	**1363/9319**	**1896/9319**

The data analysis process for the EU SPF data is similar to our simulation. We randomly split each dataset into training data with different sample sizes (40 forecasts or 160 forecasts per forecaster) and testing data (40 forecasts) without replacement. Within the training data, we estimate weights according to different weighting methods and conduct the hypothesis test algorithm and cross validation to decide whether the simple average can be rejected in favor of using each weighted average. The estimated weights are then applied to the testing data set and their performance is compared to that of the simple average. The out-of-sample MSE difference between weighted averages and the simple average from the testing data is taken as evidence about

Table EC.7 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the sequential search (decreasing) method and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	0		0		0		0		**0**	
I	Type I	90/1000	260/1000	110/1000	189/1000	21/1000	126/1000	0/1000	34/1000	**221/4000**	**609/4000**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	0		0		0		0		**0**	
II	Type I	108/1000	263/1000	98/1000	186/1000	13/1000	100/1000	0/1000	42/1000	**219/4000**	**591/4000**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	0		0		0		0		**0**	
III	Type I	162/1000	276/1000	232/1000	239/1000	67/1000	158/1000	2/1000	63/1000	**463/4000**	**736/4000**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	122		21		2		0		**145**	
IV	Type I	3/15	1/15	0/0	0/0	0/0	0/0	0/0	0/0	**3/15**	**1/15**
	Type II	118/863	253/863	3/979	125/979	0/998	68/998	0/1000	18/1000	**121/3840**	**464/3840**
	Inconclusive	415		290		144		74		**923**	
V	Type I	11/70	42/70	16/26	11/26	7/7	5/7	2/2	0/2	**36/105**	**58/105**
	Type II	448/515	216/515	258/684	225/684	10/849	187/849	0/924	120/924	**716/2972**	**748/2972**
	Inconclusive	280		243		152		107		**782**	
VI	Type I	134/225	114/225	64/66	36/66	16/17	11/17	0/0	0/0	**214/308**	**161/308**
	Type II	250/495	224/495	43/691	249/691	9/831	195/831	0/893	138/893	**302/2910**	**806/2910**
	Inconclusive	**817**		**554**		**298**		**181**		**1850**	
Total	**Type I**	**508/3310**	**956/3310**	**520/3092**	**661/3092**	**124/3024**	**400/3024**	**4/3002**	**139/3002**	**1156/12428**	**2156/12428**
	Type II	**816/1873**	**693/1873**	**304/2354**	**599/2354**	**19/2678**	**450/2678**	**0/2817**	**276/2817**	**1139/9722**	**2018/9722**

whether our test algorithm or the cross validation makes type I or type II errors. We again use the paired t-test on the out-of-sample MSEs of the weighted average and the simple average and only count trials with a significant difference at the significance level of 0.05 for error calculation. In accordance with previous literature, we apply the non-negative optimally weighting method to empirical data (OW+). We repeat this random selection process for up to 1000 times to explicitly reveal the error rates.

Tables EC.10 and EC.11 show that overall, our hypothesis test algorithm performs similarly to cross validation. Where they differ, our hypothesis test tends to make fewer type II errors, compared to cross validation. With a small sample size, both our test and cross validation make

Table EC.8 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the regularized weighting method (shrinking the covariance matrix) and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	5		31		78		217		**331**	
I	Type I	18/995	171/995	27/969	156/969	31/922	144/922	18/783	96/783	**94/3669**	**567/3669**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	4		24		95		208		**331**	
II	Type I	11/996	167/996	38/976	173/976	34/905	151/905	36/792	107/792	**119/3669**	**598/3669**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	15		67		154		251		**487**	
III	Type I	21/985	206/985	53/933	206/933	48/846	179/846	19/749	102/749	**141/3513**	**693/3513**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	65		91		87		102		**345**	
IV	Type I	13/824	86/824	10/741	67/741	8/680	52/680	0/634	58/634	**31/2879**	**263/2879**
	Type II	26/111	67/111	9/168	51/168	5/233	73/233	1/264	77/264	**41/776**	**268/776**
	Inconclusive	329		171		34		1		**535**	
V	Type I	0/196	76/196	1/23	5/23	0/0	0/0	0/2	2/2	**1/221**	**83/221**
	Type II	414/475	184/475	319/806	253/806	19/966	121/966	0/997	35/997	**752/3244**	**593/3244**
	Inconclusive	119		13		0		0		**132**	
VI	Type I	11/814	131/814	3/981	22/981	0/1000	0/1000	0/1000	0/1000	**14/3795**	**153/3795**
	Type II	49/67	49/67	2/6	6/6	0/0	0/0	0/0	0/0	**51/73**	**55/73**
	Inconclusive	**537**		**397**		**448**		**779**		**2161**	
Total	**Type I**	**74/4810**	**837/4810**	**132/4623**	**629/4623**	**121/4353**	**526/4353**	**73/3960**	**365/3960**	**400/17746**	**2357/17746**
	653	**489/653**	**300/653**	**330/980**	**310/980**	**24/1199**	**194/1199**	**1/1261**	**112/1261**	**844/4093**	**916/4093**

fewer (more) type I (II) errors in unemployment rate data than that in inflation rate data, which implies our test and cross validation suggest using the simple average more often in unemployment data than in inflation rate data. This finding aligns the conclusion in Genre et al. (2013) but we are different in split the training and testing data.

We also found that our test algorithm has a higher type I error rate than cross validation even with small sample size in the inflation rate data when comparing the optimally weighted average to the simple average. To investigate whether this observation is derived from the noisy data, we reduce the sample size of training data to 20 and check whether our test could produce a lower

Table EC.9 Type I/II error rate of the hypothesis test algorithm (Test) and the cross validation (CV) in the comparison of the regularized weighting method (constrained LASSO regression) and the simple average

Case		Sample Size								**Total**	
		32		64		128		256			
		Test	CV	Test	CV	Test	CV	Test	CV	**Test**	**CV**
	Inconclusive	86		231		488		702		**1507**	
I	Type I	204/914	303/914	560/769	282/769	460/512	228/512	280/298	137/298	**1504/2493**	**950/2493**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	104		248		511		731		**1594**	
II	Type I	195/896	284/896	529/752	271/752	462/489	192/489	264/269	103/269	**1450/2406**	**850/2406**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	101		250		475		701		**1527**	
III	Type I	240/899	292/899	572/750	273/750	483/525	236/525	284/299	128/299	**1579/2473**	**929/2473**
	Type II	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Inconclusive	2		0		0		0		**2**	
IV	Type I	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	**0/0**	**0/0**
	Type II	147/998	205/998	0/1000	63/1000	0/1000	14/1000	0/1000	2/1000	**147/3998**	**284/3998**
	Inconclusive	84		6		0		0		**90**	
V	Type I	0/2	2/2	0/0	0/0	0/0	0/0	0/0	0/0	**0/2**	**2/2**
	Type II	761/914	285/914	287/994	205/994	14/1000	75/1000	0/1000	11/1000	**1062/3908**	**576/3908**
	Inconclusive	474		293		90		20		**877**	
VI	Type I	79/126	55/126	13/13	8/13	0/0	0/0	0/0	0/0	**92/139**	**63/139**
	Type II	198/400	219/400	46/694	271/694	3/910	207/910	0/980	113/980	**247/2984**	**810/2984**
	Inconclusive	**851**		**1028**		**1564**		**2154**		**5597**	
Total	**Type I**	**718/2837**	**936/2837**	**1674/2284**	**834/2284**	**1405/1526**	**656/1526**	**828/866**	**368/866**	**4625/7513**	**2794/7513**
	Type II	**1106/2312**	**709/2312**	**333/2688**	**539/2688**	**17/2910**	**296/2910**	**0/2980**	**126/2980**	**1456/10890**	**1670/10890**

type I error than cross validation. Still, we randomly select data for 1000 times, and the results show that our test algorithm could have a lower type I error rate (25/36) then cross validation (29/36).

The results in Tables EC.10 and EC.11 also present how robust different weighting methods perform under given data contexts. For the unemployment rate data, the optimally weighted average with the non-negative constraint is more robust than the contribution weighted model and the regularized weighting method, although they all could outperform the simple average as the sample size increases. For the inflation rate data, the contribution weighted model is more robust than other two models. Particularly for the optimal weighted average, it approaches to the simple

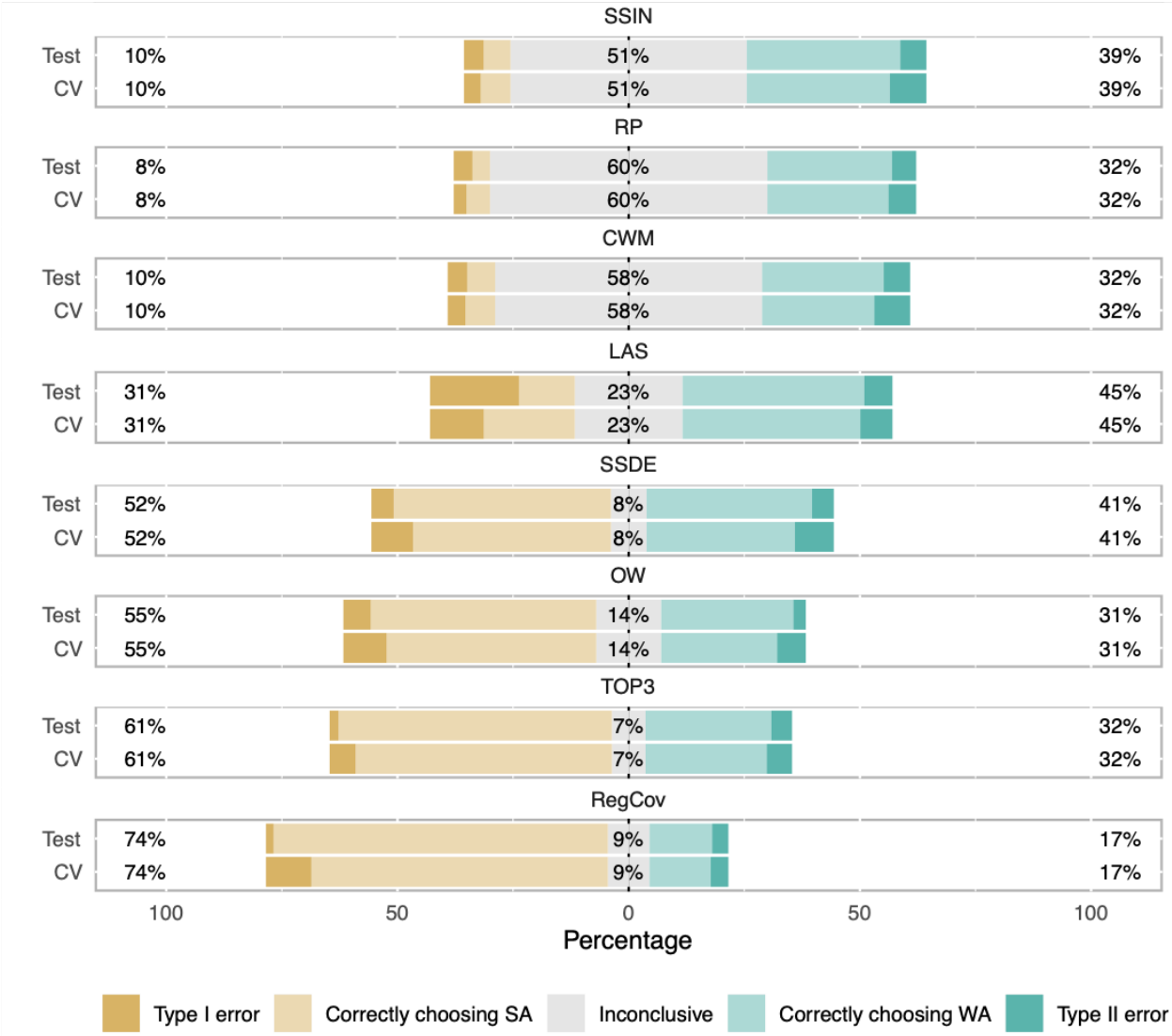


Figure EC.1 **Type I/II error rate results of our hypothesis test algorithm (Test) and cross validation (CV) for different weighting models compared to the simple average across all cases and sample sizes (Percentages from left to right indicate the percent of trials where the simple average outperforms the weighted average in held-out samples, inconclusive trials, and trials where the weighted average outperforms the the simple average in held-out samples)**

average as the sample size increases, resulting in more inconclusive trials. However, for the contribution weighted model and the regularized weighting method which allow subset crowd selection, they keep outperforming the simple average with sufficient samples.

Table EC.10 Confusion matrices of hypothesis test and cross validation (shaded background) in comparison of different weighted averages to the simple average with small and large sample sizes for unemployment rate data

Sample Size = 40

OW+ vs. SA		True State: SA (Null)	WA (Alt.)	Inconclusive	Total
Decision	Retain Null (SA)	2	**14**	204	220
		2	**50**	271	323
	Reject Null (WA)	**4**	76	700	780
		4	40	633	677
	Total	6	90	904	1000

CWM & SA		True State: SA (Null)	WA (Alt.)	Inconclusive	Total
Decision	Retain Null (SA)	1	**3**	30	34
		4	**28**	280	312
	Reject Null (WA)	**9**	59	898	966
		6	34	648	688
	Total	10	62	928	1000

LAS & SA		True State: SA (Null)	WA (Alt.)	Inconclusive	Total
Decision	Retain Null (SA)	3	**18**	247	268
		2	**43**	356	401
	Reject Null (WA)	**11**	62	659	732
		12	37	550	599
	Total	14	80	906	1000

Sample Size = 160

OW+ vs. SA		True State: SA (Null)	WA (Alt.)	Inconclusive	Total
Decision	Retain Null (SA)	0	**0**	0	0
		0	**2**	0	2
	Reject Null (WA)	**0**	168	832	1000
		0	166	832	998
	Total	0	168	832	1000

CWM & SA		True State: SA (Null)	WA (Alt.)	Inconclusive	Total
Decision	Retain Null (SA)	0	**0**	0	0
		0	**1**	0	1
	Reject Null (WA)	**0**	114	886	1000
		0	113	886	999
	Total	0	114	886	1000

LAS & SA		True State: SA (Null)	WA (Alt.)	Inconclusive	Total
Decision	Retain Null (SA)	0	**0**	0	0
		0	**3**	10	13
	Reject Null (WA)	**3**	154	843	1000
		3	151	833	987
	Total	3	154	843	1000

We also apply our hypothesis test algorithm to the US SPF data where domain experts are registered as professional forecasters to make prediction on more than 20 target values (e.g., nominal GDP, average unemployment rate, average level of the index of industrial production and so on). Most forecasts are collected from 1968 quarterly, including both quarterly and yearly forecasts. We download the data from 1968 to 2019, and obtain 161 sub dataset (each dataset is for one specific target value and for one target prediction time period, e.g., current quarter, future 1 year and etc.). If there is no missing data, we are able to get forecasts from 443 forecasters and each of them gives 205 forecasts. However, missing data problem is even more severe in the US SPF data due to the long-term forecasting process. Therefore, we filter data by removing all missing data

Table EC.11 Confusion matrices of hypothesis test and cross validation (shaded background) in comparison of different weighted averages to the simple average with small and large sample sizes for inflation rate data

Sample Size										
40						160				
OW+ vs. SA		True State			Total	OW+ vs. SA		True State			Total
		SA (Null)	WA (Alt.)	Inconclusive				SA (Null)	WA (Alt.)	Inconclusive	
Decision	Retain Null (SA)	0	**0**	11	11	Decision	Retain Null (SA)	0	**0**	0	0
		6	**19**	165	190			0	**0**	0	0
	Reject Null (WA)	**38**	43	908	989		Reject Null (WA)	**19**	18	963	1000
		32	24	754	810			**19**	18	963	1000
	Total	38	43	919	1000		Total	19	18	963	1000
CWM & SA		True State			Total	CWM & SA		True State			Total
		SA (Null)	WA (Alt.)	Inconclusive				SA (Null)	WA (Alt.)	Inconclusive	
Decision	Retain Null (SA)	0	**0**	0	0	Decision	Retain Null (SA)	0	**0**	0	0
		0	**12**	21	33			0	**0**	0	0
	Reject Null (WA)	**4**	125	871	1000		Reject Null (WA)	**1**	157	842	1000
		4	113	850	967			**1**	157	842	1000
	Total	4	125	871	1000		Total	1	157	842	1000
LAS & SA		True State			Total	LAS & SA		True State			Total
		SA (Null)	WA (Alt.)	Inconclusive				SA (Null)	WA (Alt.)	Inconclusive	
Decision	Retain Null (SA)	0	**0**	19	19	Decision	Retain Null (SA)	0	**0**	0	0
		0	**19**	216	235			0	**0**	0	0
	Reject Null (WA)	**10**	55	916	981		Reject Null (WA)	**2**	121	877	1000
		10	36	719	765			**2**	121	877	1000
	Total	10	55	935	1000		Total	2	121	877	1000

and integrate forecasts for different target prediction time periods to guarantee enough samples. Eventually, we obtain the following selected dataset as shown in Table EC.12. The explanation of each item of the target values can be seen in https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/data-files.

Without random selection as empirical analysis for the ECB SPF data, for each subset data we take the first 50% (or 80%) historical data as the training data and the later 50% (or 20%) historical data as the testing data to compare our algorithm and the cross validation. The proposed weighting methods include (1) the theoretically optimal weighting method (OW); (2) the ranked performance method (RP), (3) the contribution weighted model (CWM), (4) the sequential search

Table EC.12 Subset data of US SPF for algorithm validation

Target value	Number of Forecasters	Number of Sample Judgments
NGDP	8	122
PGDP	7	187
CPROF	3	317
RGDP	8	132
RCONSUM	5	387
RNRESIN	5	381
RRESINV	5	363
RFEDGOV	4	429
RSLGOV	5	351
RCBI	5	361
REXPORT	5	358

(decreasing) method (SSD), and (5) the LASSO regularized method (LAS). Paired t-test is still utilized to judge whether the out-of-sample MSE of the weighted average is significant different from that of the simple average.

Table EC.13 presents the analysis results. We could find that our hypothesis test algorithm has a similar performance as the cross validation to decide which weighting model is appropriate for current observed judgments. After combining the performance in the simulation and empirical analysis, we believe our hypothesis test algorithm could be a competitive alternative way of the cross validation if decision makers desire to explore more data characteristics during the selection of a proper aggregation rule.

Table EC.13 Analysis results of the US SPF data

Target	#Training	#Testing	OW-SA			RP-SA			CWM-SA			SSD-SA			LAS-SA		
			Actual	TS	CV	Actual	TS	CV	Actual	TS	CV	Actual	TS	CV	Actual	TS	CV
CPROF	158	159	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD	LAS	LAS	LAS
	253	64	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD	LAS	LAS	LAS
NGDP	61	61	SA	SA	SA	RP	RP	RP		CWM	CWM		SSD	SSD	**SA**	**LAS**	**LAS**
	97	25	SA	SA	SA		RP	RP		CWM	SA		SSD	SSD		LAS	LAS
PGDP	93	94	**SA**	**OW**	**SA**		RP	RP	CWM	CWM	CWM		SSD	SSD		LAS	SA
	149	38		OW	SA		RP	RP		CWM	CWM		SSD	SSD		LAS	LAS
RCBI	180	181	SA	SA	SA		RP	RP		CWM	CWM		SSD	SSD		LAS	LAS
	288	73		SA	SA		RP	RP		CWM	CWM		SSD	SSD		LAS	LAS
RCONSUM	193	194	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD	LAS	LAS	LAS
	309	78	SA	SA	SA		RP	RP		CWM	CWM	SSD	SSD	SSD		LAS	LAS
REXPORT	179	179	SA	SA	SA		RP	RP	CWM	CWM	CWM		SSD	SSD	LAS	LAS	LAS
	286	72	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD		LAS	LAS
RFEDGOV	214	215	SA	SA	SA	**SA**	**RP**	**RP**		CWM	CWM	**SA**	**SSD**	**SSD**	**SA**	**LAS**	**LAS**
	343	86	SA	SA	SA		RP	RP		CWM	CWM	**SA**	**SSD**	**SSD**		LAS	LAS
RGDP	66	66	SA	SA	SA		RP	RP		CWM	CWM		SSD	SSD	**SA**	**SA**	**LAS**
	105	27	SA	SA	SA		RP	RP		CWM	CWM		SSD	SSD	**SA**	**LAS**	**LAS**
RNRESIN	190	191	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD		LAS	LAS
	304	77	SA	SA	SA		RP	RP		CWM	CWM		SSD	SSD		LAS	LAS
RRESINV	181	182	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD	LAS	LAS	LAS
	290	73	SA	SA	SA	RP	RP	RP	CWM	CWM	CWM	SSD	SSD	SSD	LAS	LAS	LAS
RSLGOV	175	176	SA	SA	SA		RP	RP	CWM	CWM	CWM		SSD	SA		LAS	LAS
	280	71	SA	SA	SA		RP	RP		CWM	CWM		SSD	SSD		LAS	LAS