Beating the average forecast: Regularization based on forecaster attributes

Edgar C. Merkle
University of Missouri

Geoff Saw
University of Melbourne

Clinton Davis-Stober
University of Missouri

Abstract

In a variety of real-world forecasting contexts, researchers have demonstrated that the unweighted average forecast is reasonably accurate and difficult to improve upon with more complex, model-based aggregation methods. We investigate this phenomenon by systematically examining the relationship between individual forecaster characteristics (e.g., bias, consistency) and aspects of the criterion being forecast (e.g., “signal strength”). To this end, we develop a model inspired by Cultural Consensus Theory (Batchelder & Romney, 1988) that (i) allows us to jointly estimate both forecaster characteristics and environmental characteristics and (ii) contains the unweighted average as a special case. This allows us to use the model as a regularization method for forecast aggregation, where restrictions on forecaster parameters make the model similar to use of an unweighted average. Relatedly, the model allows us to apply existing results on optimal forecaster weighting to real data. We show how the model provides guidance for identifying prediction environments where the average forecast can potentially be beaten. We also conduct two simulation studies and illustrate the model’s practical application using forecasts of Australian Football League point spreads.

Background

A considerable body of work in forecasting has established that taking a simple, unweighted average of forecasters yields a forecast that is not only good, it is extremely difficult to improve upon with more complex aggregation methods (Armstrong, 2001; Cuzán & Bundrick, 2009; Graefe, 2015; Green & Armstrong, 2015). Graefe, Küchenhoff, Stierle, and

Correspondence to Edgar C. Merkle, merkle@missouri.edu. Code to reproduce the models and analyses in this paper can be found at <https://osf.io/by93u/>.

Riedl (2015) compared the accuracy of a simple unweighted average to that of Ensemble Bayesian Model Averaging, a modern model averaging technique used frequently in weather forecasting, over many previously published economic data sets. They found that, on average, Bayesian model averaging yielded 5 percent greater error than the simple unweighted average. Graefe (2015) found similar results in a re-analysis of election forecasting data sets comparing the unweighted average to linear regression estimates. More generally, Green and Armstrong (2015) carried out a literature review examining the relative benefit of simple aggregation forecasts (including the unweighted average) to more complex forecasting methods. They identified a total of 97 such comparisons contained in 32 published papers. Green and Armstrong (2015) concluded that “none of the papers provide a balance of evidence that complexity improves forecast accuracy” (p. 1678), reporting that complexity increased error by an average of 27 percent.

It may seem counter-intuitive at first blush that an unweighted average could strongly outperform more complex aggregation methods that utilize the data to determine weighting relationships. This relationship has been studied within the context of linear regression since the 1970s. Dawes (1979) established that simple nonoptimal weighting, termed “improper linear models,” could routinely outperform ordinary least squares estimates in a variety of social science contexts, e.g., determining how to weight test scores and grade point average when making university admission decisions. Generally speaking, there are two major perspectives on why simple weighting schemes work well compared to more complex methods. Beginning with Wilks (1938), and later Wainer (1976), when variables used for prediction (or forecasters to be aggregated) are positively correlated with one another it simply doesn’t matter what weighting scheme is used. This is the “flat maximum” perspective in which there exists a large class of equivalent weights for predicting the independent variable (Wainer & Thissen, 1976), see also the “fungible weights” of Waller (2008). The second perspective is that of a bias-variance trade-off. We can envision improper linear models, including the unweighted average, as “estimators” in their own right – albeit ones that are not dependent on data and do not vary. When compared to standard estimators, such as ordinary least squares, improper linear models are clearly biased and inconsistent. Yet, by construction, these models have less variance than standard estimators across repeated sampling. In other words, if the empirical application is such that sample size and predictability (i.e., strength of association between predictors and criterion) are modest in size, then standard estimators will be less accurate than a well-chosen improper linear model, such as the unweighted average. This is simply due to the fact that standard estimators are based on noisy, not overly predictive data, and will, on average, produce variable weights that are less accurate than a potentially biased, yet fixed, set of weights. See Davis-Stober (2011) for an analysis with regard to linear regression and Smith and Wallis (2009) for a discussion related to forecasting.

This isn’t to say that improving upon an unweighted average is a hopeless endeavor. Budescu and Chen (2014) developed a weighting scheme that identifies poorly performing forecasters and down-weights (or eliminates) their contributions before aggregation, yielding superior forecasts. Additionally, the ‘Bayesian truth serum’ approach of Prelec (2004) and ‘Surprisingly popular’ approach of Prelec, Seung, and McCoy (2017) differentially weight forecaster contributions by including forecaster beliefs about the overall distribution of forecasts. This approach can outperform majority rule aggregation methods in many contexts

and is applicable even when there does not exist a true criterion value being forecast. It has been shown to be useful in the predicting the outcomes of American football games (Lee, Danileiko, & Vi, 2018). Finally, many researchers (e.g., Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Lee & Danileiko, 2014; Merkle, 2010; Montgomery, Hollenbach, & Ward, 2015; Raftery, Gneiting, Balabdaoui, & Polakowski, 2005; Satopää et al., 2014; Shlomi & Wallsten, 2010; Turner, Steyvers, Merkle, Budescu, & Wallsten, 2014) have discussed “re-calibration” and related methods, whereby forecasts are corrected for systematic forecaster biases and distortions (though also see Kadane & Fischhoff, 2013). Many of these methods were developed and successfully applied in the Aggregative Contingent Estimation (ACE) forecasting tournament, sponsored by IARPA from about 2011–2015.

We take the following approach to better understand when and why differential weighting will or will not be more accurate than an unweighted average. First, we need to decide what sort of differential weighting scheme to compare an unweighted average to. Many of the above models are strong contenders. Ideally, such a comparison model would satisfy two aims. First, the differential weighting model should represent an optimal weighting structure, according to a reasonable objective function. This would place the unweighted average against a strong competitor, and not a weak straw model. If the unweighted average proved superior to an optimal weighting structure, then it would outperform *all* such differential weighting schemes. Second, the differential weighting framework and subsequent analyses should reveal interpretable *cognitive aspects* of the individual forecasters. For example, it may be the case that certain configurations of forecaster biases lead the unweighted average to fail in a given forecasting environment.

In this paper, we develop a model-based approach that satisfies both of these aims by formally linking the unweighted average to an optimal forecaster weighting scheme. The model, inspired by Batchelder’s Cultural Consensus Theory (CCT; e.g., Batchelder & Anders, 2012; Batchelder & Romney, 1988; Karabatsos & Batchelder, 2003; Romney, Batchelder, & Weller, 1987; Romney, Weller, & Batchelder, 1986), allows us to estimate relevant forecaster and environment characteristics within a broad class of forecasting environments. Drawing on analytic work from Davis-Stober, Budescu, Broomell, and Dana (2015), we are then able to translate these characteristics into optimal forecaster weights, using mean squared error as an objective function. This allows us to study the utility of differential forecaster weights as applied to real data, outlining scenarios where it may be beneficial to move beyond the unweighted average. While our model is less general than most existing CCT models, its simple expressions of forecaster characteristics allow us to directly compare different weighted averages in ways that are not possible with other models.

A significant strength of our model is that it contains the unweighted average as a special case. This allows us to directly evaluate the conditions under which estimating individual forecaster attributes leads to superior accuracy. As we demonstrate, the various considerations include forecaster bias, variance of the forecasts, and overall predictability of the criterion being forecast. The goal of our model and approach is to provide a “road map” for investigators for identifying circumstances when the unweighted average will do well, when it is worth estimating individual forecaster aspects and differentially weighting forecasters, and, most importantly, which forecaster aspects lead to superior predictability.

Our article is organized as follows. After presenting a definition of optimal weighting

for a set of forecasters, we go on to define the statistical model that allows us to obtain the weights. We then conduct two simulations to explore the conditions under which it pays to differentially weight forecasters and the conditions under which it does not. Next, we apply our method to real forecasts involving Australian football point spreads, illustrating its performance relative to the unweighted average. These results provide information about aspects of the forecasters and forecasting environment that can be reliably estimated in practice.

Optimal weighting

To obtain a proper definition of optimal weighting, we present the following assumptions on forecaster behavior coupled with an objective function for optimization. Assume that there are J -many forecasters reporting on a single forecasting question. Let y_j denote the realization of forecaster j . We consider y_j as a random variable, such that $\text{var}(y_j)$ denotes the variance of forecaster j 's predictions. To simplify notation, let Σ denote the $J \times J$ covariance matrix of all forecasters and let $\mathbf{y}$ denote a $J \times 1$ vector of forecaster realizations (i.e., reported forecasts for a single question). Let θ be the criterion value being forecast and define $\delta_j = E[y_j - \theta]$, i.e., δ_j is the *bias* of forecaster j . Let $\sigma_{j,\theta}$ be the covariance between forecaster j 's predictions and the criterion value being forecast. Finally, we create group forecasts using a weighted linear aggregate, $G(\mathbf{y}) = \sum_{j=1}^J w_j y_j$, where w_j is the weight assigned to forecaster j and $\sum_{j=1}^J w_j = 1$.

Similar to many previous approaches (Davis-Stober, Budescu, Dana, & Broomell, 2014; Lamberson & Page, 2012), we assess the accuracy of the group-level forecast using *mean squared error* (MSE), defined as

$$\text{Acc}(w, \theta) = E[(G(\mathbf{y}) - \theta)^2].$$

One benefit to using mean squared error as our accuracy definition is that the following optimal weighting results are distribution-free, in that they apply no matter how forecasters' predictions are distributed. We can now state the optimal weighting result from Davis-Stober et al. (2015), where the subscript i corresponds to items ($i = 1, \dots, I$) and j corresponds to forecasters ($j = 1, \dots, J$).

THEOREM (DAVIS-STOBER ET AL., 2015). *The optimal weighted average of a collection of forecasts $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$ with covariance matrix Σ , is the solution to the following system of linear equalities:*

$$\begin{pmatrix} \Sigma + \delta\delta' & \mathbf{u} \\ \mathbf{u}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \lambda \end{pmatrix} = \begin{pmatrix} \sigma_{s,V} \\ 1 \end{pmatrix},$$

where $\mathbf{u}$ is a $J \times 1$ vector of ones, δ is the $J \times 1$ vector of forecaster biases, $\sigma_{\mathbf{y},\theta}$ is the $J \times 1$ vector of covariances between each forecaster and the criterion variable θ , $\mathbf{w}$ is the $J \times 1$ vector of weights constrained to sum to 1, and λ is a real-valued unknown variable. By solving the above system for $\mathbf{w}$, we obtain the optimal set of weights $\mathbf{w}^*$.

Given knowledge of θ , the optimal weights $\mathbf{w}^*$ provide a starting point to compare the performance of the unweighted average with weighted averages based on $\mathbf{w}^*$. But, in order to do this comparison in practice, we need a way to estimate the forecaster traits mentioned in the theorem: forecaster biases, forecaster-criterion covariances, and forecaster

covariances. We propose the use of a model that indirectly estimates these traits, with the estimates then being used to obtain optimal weights via the theorem above. The model and its use in computing optimal weights are described below.

Model for optimal weighting

To estimate the forecaster weights, we use a factor analysis model with the roles of people and questions reversed from their traditional roles. This approach has been taken in multiple CCT models, and it is one that the first author directly learned from Bill Batchelder. The model is also related to some item response models that have been developed for probabilistic forecasts (Bo, Budescu, Lewis, Tetlock, & Mellers, 2017; Merkle, Steyvers, Mellers, & Tetlock, 2016, 2017), to recent extensions of CCT (Aßfalg, 2018; Anders & Batchelder, 2015; Anders, Oravecz, & Batchelder, 2014; Batchelder, Strashny, & Romney, 2010; Oravecz, Vandekerckhove, & Batchelder, 2014), and to the models proposed by Merkle and Steyvers (2011) and Lee, Steyvers, and Miller (2014) for aggregating judgments of magnitude and ranking, respectively. But it is simpler than most of these other models, because the optimal weighting theorem requires us to derive analytic expressions of forecaster biases and covariances under the model.

As above, let y_{ij} be the (continuous) judgment of forecaster j on question i . Let θ_i be the realized value of question i . The model is

$$y_{ij}|\theta_i \sim N(\mu_{ij}, \psi_j^2) \quad (1)$$

$$\mu_{ij} = \delta_j + \lambda_j \theta_i \quad (2)$$

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2). \quad (3)$$

In words, a particular forecaster’s judgment on question i is based on the truth θ_i , but it is perturbed by three forecaster-specific parameters: an “offset” δ_j , a multiplicative parameter λ_j , and a noise term with variance ψ_j^2 . These forecaster-specific parameters are assumed to be stable traits across questions, with the question environment (“reference class”) being described by the $N(\mu_\theta, \sigma_\theta^2)$ distribution.

We consider this to be a cognitive psychometric model because each forecaster has three unique parameters, each of which correspond to a particular aspect of the response process. This differs from traditional psychometric models, which would assign each forecaster a single, generic parameter reflecting his/her overall proficiency at forecasting, while adding extra parameters on the item (question) side. Model extensions could potentially include multiple forecaster and item parameters, though the expressions required for optimal weighting (detailed below) become more difficult to derive in those situations.

Parameters and optimal weights. The model parameters generally have intuitive interpretations, which is helpful for understanding the model and for setting prior distributions. The δ_j parameters reflect forecaster bias, and would assume values close to 0 in environments where forecasters are approximately unbiased. The λ_j parameters reflect the extent to which the forecaster’s response is influenced by the truth (θ_i), and will often be close to 1. Finally, the ψ_j parameters reflect noise in the response process, so their values will be dictated by the questions being forecasted. For reference, an oracle forecaster who knows the outcome of each question would have $\delta_j = 0$, $\lambda_j = 1$, and $\psi_j = 0$.

The question outcomes θ_i have a normal hyperdistribution with parameters μ_θ and σ_θ^2 . The μ_θ parameter is the mean of true values in the reference class, which in many situations could be fixed based on prior knowledge. Interpretation of σ_θ^2 requires further explanation. This parameter does *not* reflect the variance of true values in the reference class because, in the model, it helps to predict the variance of the reported forecasts in addition to the average forecast. We have

$$\text{var}(y_{ij}) = \lambda_j^2 \sigma_\theta^2 + \psi_j^2, \quad (4)$$

which shows that the predicted variability in forecasts is partly due to variability in the signal (σ_θ^2) and partly due to noise (ψ_j^2). We cannot treat σ_θ^2 as the variance of true values because (i) the variance of the y_{ij} will often be similar to the variance of true values; and (ii) the value of ψ_j^2 is almost surely nonzero in practice. When the λ_j equal 1, then σ_θ^2 becomes the variability in observed forecasts due to the signal in forecasting questions. This property will later become useful to us as we set prior distributions.

The model proposed here is especially useful to us because it provides simple expressions for the forecaster traits discussed in the optimal weighting theorem. In addition to the expression for $\text{var}(y_{ij})$, we have the following model-based expressions:

$$\text{bias}(y_{ij}) = \delta_j + (\lambda_j - 1)\mu_\theta \quad (5)$$

$$\text{cov}_{j \neq k}(y_{ij}, y_{ik}) = \lambda_j \lambda_k \sigma_\theta^2 \quad (6)$$

$$\text{cov}(y_{ij}, \theta_i) = \lambda_j \sigma_\theta^2, \quad (7)$$

where the latter covariance (between y_{ij} and θ) is sometimes called *forecast validity*. The derivations for these expressions appear in Appendix B. They allow us to obtain optimal weights for a set of forecasters via the following steps:

1. Fit the model to the reported forecasts.
2. Using the fitted model's parameter estimates, obtain estimates of forecaster traits via Equations (5) to (7).
3. Enter the estimated forecaster traits in the system of equations presented in the Davis-Stober et al. (2015) theorem and solve, yielding a set of optimal weights $\mathbf{w}^*$.

The optimal weights can then be used to compute weighted averages of reported forecasts for questions whose outcomes are not yet known. The above shows that the fitted model is an intermediate step towards obtaining optimal weights. The model provides estimates of forecaster traits, which in turn are used to compute optimal weights outside of the model, which in turn are used to compute weighted averages outside of the model.

Parameter identification. Assuming that μ_θ and σ_θ^2 are fixed, the parameters of the proposed model are identified up to a sign change. To resolve the sign change issue, we can constrain one forecaster's λ parameter to be positive (e.g., Peeters, 2012). However, as the outcomes of certain questions become known (i.e., as some θ_i become known), then the sign change issue can resolve itself.

If the question outcomes and the reference class parameters μ_θ and σ_θ^2 are all unknown, we must make further assumptions about the forecasters in order to achieve parameter

identification (as is required of traditional factor analysis models; see, e.g., Little, Slegers, & Card, 2006). The intuition here is that, if we do not know how forecasters performed on any resolved questions, then we cannot distinguish forecaster attributes from question attributes. For the model here, traditional identification constraints amount to assuming that either (i) one forecaster’s responses bear a specific relationship to the truth, or (ii) on average, forecasts are unbiased. In model parameters, the latter assumption can be formulated as

$$\sum_j \delta_j = 0 \quad (8)$$

$$\sum_j \lambda_j = J. \quad (9)$$

This assumption (that, on average, forecasts are unbiased) seems generally preferable to making an assumption about a single forecaster. But it is easy to imagine scenarios where this assumption will lead us astray, leading to poor forecaster weights. For example, forecasters may be systematically biased in one direction, possibly due to forecasters sharing false beliefs about a forecasting domain. Or reported forecasts may bear a weak relationship to the outcomes, making the restriction on the λ_j implausible. For the analyses in this paper, we generally sidestep these issues by assuming that some question outcomes are known. This differs from other CCT models that assume no ground truth, and it undoubtedly aids our model’s performance in the simulations and application described later. However, we are also careful to only examine model performance on questions whose outcomes are not provided to the model.

Regularization and estimation of θ_i . While we have so far focused on the model as an intermediate step towards computing optimal weights, the model’s θ_i ($i = 1, \dots, I$) parameters represent our posterior estimates of each question’s ground truth. This provides a competing approach to optimally-weighted averages: instead of doing optimal weighting outside of the model, we can rely on the θ_i estimates that arise directly from the model. We compare these θ_i estimates to optimal weighting in the simulation and application sections.

Along with providing a competing approach, the θ_i parameters have important theoretical properties that allow us to connect our model to weighted and unweighted averages. It can be shown that the sufficient statistic for each θ_i is $\sum_j \lambda_j y_{ij}$. This implies that the estimated θ_i are proportional to a weighted average, where the weights are based on each forecaster’s multiplicative parameter λ_j . Importantly, if all the λ_j are constrained to be equal, then the θ estimates will be proportional to an unweighted average. In contrast, as the λ_j parameters become unequal, the θ estimates are a weighted average, with some forecasters being weighted more heavily than others. But these weights arise directly from the model, whereas our optimal weights arise indirectly from the model estimates of forecaster traits.

The sufficient statistic for θ_i leads us to propose *regularization* methods for forecast aggregation by directly constraining classes of parameters (the λ , δ , or ψ classes) to be equal across forecasters. Regularization is a common approach in statistics and machine learning to obtain improvements in estimation, prediction and classification. The basic idea is to add information to the model not included in the dataset itself, and this is typically done via constraints on parameters. Regularization methods tend to “bias” how the model

interacts with the data but can lead to substantial improvements by reducing variability of estimates/predictions, etc. Depending on which classes of parameters are constrained, we obtain aggregates that are closer to or further from the unweighted average. Starting with the full model, where each forecaster has three unique parameters, we propose to constrain parameters across forecasters in a specific order: first constraining the λ parameters to be equal across forecasters, then the δ parameters, then the ψ parameters. This ordering of equality constraints is similar to the order in which parameters would often be constrained in psychometric studies of measurement invariance and differential item functioning (e.g., Vandenberg & Lance, 2000). The ordering reflects the fact that the λ s are generally most difficult to estimate because they are involved in an interaction with the θ_i , so those parameters receive constraints before the other types of parameters.

This ordering of equality constraints leads to three special cases of the full model, which can be studied alongside the full model. To fix notation, we let $\mathcal{M}_f$ be the full model, $\mathcal{M}_\lambda$ be the model with λ parameters constrained equal across forecasters, $\mathcal{M}_\delta$ be the model with both λ and δ parameters constrained equal across forecasters, and $\mathcal{M}_0$ be the model with all three classes of parameters constrained equal across forecasters. The latter model represents a model-centric way of taking an unweighted average.

We can potentially obtain additional regularization by considering σ_θ^2 and its interpretation as “variability in reported forecasts due to signal, as opposed to noise.” This involves parameterizing σ_θ^2 as

$$\sigma_\theta^2 = \rho \times s_y^2, \quad (10)$$

where s_y^2 is the observed variance of reported forecasts, and ρ is a parameter in $(0, 1)$ that serves as an R^2 value. In a well-defined reference class, a beta prior on ρ reflects our general expectations about forecaster performance in that reference class. This is practically useful because model performance is often sensitive to the prior on σ_θ^2 , so we want the prior to be as intuitive as possible for applications.

In summary, the above arguments lead us to two routes for averaging forecasts via the proposed model.

- Direct model estimation: Estimate the θ_i parameters within the model, which represent each question’s ground truth. These parameter estimates are similar to weighted averages because their sufficient statistics are $\sum_j \lambda_j y_{ij}$.
- Optimal weighting after model estimation: Use the model to estimate the forecaster traits from Equations (5) to (7). Use these estimated traits to solve the system of equalities from the Davis-Stober et al. (2015) theorem, which yield optimal forecaster weights. Use these forecaster weights to compute weighted averages.

Either route can be used with any of the four models $\mathcal{M}_f$, $\mathcal{M}_\lambda$, $\mathcal{M}_\delta$, and $\mathcal{M}_0$, and the regularization of σ_θ^2 can be implemented within any of the four models.

Model estimation. Bayesian estimation of the proposed model is particularly useful because it provides flexibility in model estimation and prior specification. We estimate models using the *runjags* package (Denwood, 2016) in R (R Core Team, 2019), which provides an interface to the MCMC methods implemented in JAGS (Plummer, 2003). The prior distributions vary across the applications below, so we describe them separately in

each section. We generally employ “weakly informative” priors (e.g., Gelman, Jakulin, Pittau, & Su, 2008) that keep model parameters in a plausible range based on the application at hand. MCMC chain convergence is monitored in the usual manners (effective sample sizes, Gelman-Rubin statistics, trace plots), with a focus on the Gelman-Rubin statistics (Gelman & Rubin, 1992). For models fit by MCMC, the latter statistic compares each parameter’s between-chain variability to within-chain variability. If these variabilities are nearly equal, then the statistic is close to 1 and provides evidence of model convergence.

In the following sections, we use the four model variations ($\mathcal{M}_f$, $\mathcal{M}_\lambda$, $\mathcal{M}_\delta$, $\mathcal{M}_0$) to average forecasts within the model (via θ_i) and outside the model (via optimal weights). We initially study these estimates via simulation, comparing the two aggregates to each other as well as to the unweighted average. We examine a best-case scenario in Simulation 1, where forecasters exactly follow the model. We then examine a more realistic scenario in Simulation 2, where forecasters stray from the model in various manners. Finally, we apply the model to real data involving the prediction of point spreads, examining how the simulation results carry over to real data and how the model can be expected to perform in practice.

Simulation 1

In Simulation 1, we assume that forecasters respond according to the proposed model. We then estimate the model, obtain optimally-weighted averages as well as θ estimates, and compare their accuracies. We also consider the unweighted average forecast in this comparison.

Method

We first describe the data-generating model used in the simulation. We assumed that 10% of question outcomes were known, with participant forecasts being observed on both these questions and the remaining 90% of questions whose outcomes were unknown. Across all questions, we fixed the distribution of true values to be $N(0,1)$. As can be seen from Equations (5) to (7), this means that the λ parameters have no influence on bias, which allows us to more easily disentangle factors that contribute to the simulation results.

For forecaster parameters, we assumed that

$$\begin{aligned}\delta_j &\sim N(\delta^*, 1) \\ \lambda_j &\sim N(\lambda^*, .5),\end{aligned}$$

where the values δ^* and λ^* changed depending on the condition (and where the normal distributions are parameterized via standard deviations). Finally, we fixed all ψ_j parameters to 1 for simplicity. Simulation conditions were then defined by the following factors.

- Number of forecasters: 10, 100
- Number of questions: 50, 200
- $\delta^* = 0, 2$
- $\lambda^* = .5, 2$

Figure 1. Simulation 1, distributions of MSE differences between unweighted average vs others. This figure shows simulation conditions where mean bias is 0 (unbiased). Legend notation: Mf is the full model, Ml is the model with λ parameters constrained equal across forecasters, Md is the model with both λ and δ parameters constrained equal, and M0 is the model with λ , δ , and ψ parameters constrained equal.

Figure 2. Simulation 1, distributions of MSE differences between unweighted average vs others. This figure shows simulation conditions where mean bias is 2 (biased). Legend notation: Mf is the full model, Ml is the model with λ parameters constrained equal across forecasters, Md is the model with both λ and δ parameters constrained equal, and M0 is the model with λ , δ , and ψ parameters constrained equal.

The δ^* parameter represents average forecaster bias, and the λ^* parameter influences the correlation between reported forecasts and outcomes as well as correlations between forecasters.

For all 16 combinations of these factors, we generated 100 datasets and used MCMC to fit models to all 100 datasets. Along with the full model proposed earlier ($\mathcal{M}_f$), we estimated the special cases $\mathcal{M}_0$, $\mathcal{M}_\delta$, and $\mathcal{M}_\lambda$. Prior distributions for model parameters were

$$\begin{aligned}\delta_j &\sim \text{N}(0, 3.2) \\ \lambda_j &\sim \text{N}(1, 1) \\ \psi_j &\sim \text{Inv-Gamma}(.1, .1),\end{aligned}$$

with the j subscripts not always being needed for the “special case” models. For example, when the λ_j are constrained to be equal, the prior is on the single free parameter λ instead of on multiple λ_j parameters. These specific prior distributions are particularly suited to sampling in JAGS, where sampling efficiency is often improved under conjugate priors than under non-conjugate priors.

For σ_θ^2 , we used the parameterization from (10). This involved fixing σ_{truth}^2 to the true value implied by the data-generating mechanism, as well as putting a prior on ρ . The prior was either beta(2,8) or beta(8,2), depending on whether λ^* was .5 or 2, respectively. The priors were intended to provide mild information about plausible parameter values, such as may be available in applied forecasting situations with well-defined question domains.

After data generation, we aggregated the forecasts for each question via the methods described previously: the unweighted average, the direct model estimation of the θ_i , and the computation of optimal weights following model estimation. We also computed the “true” optimal weights that could be obtained only if one knew each forecaster’s true δ , λ , and ψ parameters. This leads to four aggregation methods, with the methods’ accuracies being compared on the 90% of questions whose outcomes were not supplied to the model. For each of the 100 replications within a condition, we define accuracy by the MSE between an aggregated forecast and the true value of the associated question.

Results

Across the 6,400 models estimated (16 conditions $\times$ 4 models $\times$ 100 replications), 76 failed to converge. These failures to converge occurred most often for the simplest model $\mathcal{M}_0$, potentially suggesting that our model formulation and priors (which were developed for the full model) did not translate well to the simplest models. The non-convergences happened most often for the conditions with 10 subjects and low validity, with a maximum of 13 non-convergences occurring in a single condition. Non-converged replications are removed from the results reported below.

The primary results are displayed in Figures 1 (no bias conditions) and 2 (bias conditions). For each condition, the figures show the distribution of MSE differences between the unweighted average and each of the other aggregates (the distribution across the 100 generated datasets, as opposed to uncertainty in an average MSE). Positive MSE differences imply that an aggregate was better than the unweighted average. MSE differences close

to zero imply that an aggregate was no better than the unweighted average, and negative MSE differences imply that the unweighted average was better than an aggregate.

The results show that, for the situations examined in this simulation, the other aggregates often beat the unweighted average. This is most apparent in the conditions/panels where forecasters display high validity ($\lambda^* = 2$) or bias. In Figure 1, optimal weighting under the constrained models (i.e., under all models except $\mathcal{M}_f$) is similar to taking the unweighted average; the boxes for those aggregates all stay close to 0. In contrast, optimal weighting under the full model $\mathcal{M}_f$ displays accuracy that is closer to the model-based θ estimates. The bottom left panel of Figure 1 shows the performance of the aggregates in situations with few forecasters, few questions, and low validity, which may be especially relevant to practice. There, we see that optimal weighting under the full model is as good as the model-based θ estimates, and both aggregates beat the unweighted average. In other panels, we often see that the θ aggregates are as good as or better than the optimal weights, even the optimal weights derived from the truth. This is potentially due to the fact that the form of the estimated model is the same as that of the data-generating model, so that the θ estimates can capitalize off the model’s likelihood in ways that the optimally-weighted average cannot.

As intuition would expect, the unweighted average was beaten most handily in Figure 2, which depicts conditions where forecasters exhibit systematic bias in one direction (i.e., forecasts consistently greater than/smaller than the criterion mean). In these conditions, the unweighted average also exhibited bias whereas most of the other aggregates could partially correct for the bias. The only exceptions to this correction were the optimally-weighted aggregates under $\mathcal{M}_0$ and $\mathcal{M}_\delta$. These are both models where δ , the model parameter related to bias, was constrained equal across forecasters. Thus, individual differences in forecaster bias could not play a role in the computation of optimal weights, because forecasters were forced to have the same magnitude of bias. This means that, for models where δ is constrained to be equal across forecasters, unbiased forecasters could not be weighted more heavily than biased forecasters.

Summary

The results of Simulation 1 show that the unweighted average can be beaten by the proposed aggregates in situations where the data-generating model matches the estimated model. The aggregates generally improve in predictable manners, based on sample sizes (of forecasters and questions), forecaster bias, and validity. Additionally, the model-based aggregates (the θ_i) are as good as or better than the optimally-weighted aggregates in all conditions examined. However, the conditions studied in Simulation 1 are not likely to mimic practice, where the data generating model is unknown and forecasters are less reliable. We further examine these issues in Simulation 2.

Simulation 2

Real forecasters obviously do not behave according to our proposed model. Therefore, in Simulation 2, we study our model’s performance when forecasters behave more haphazardly. This potentially provides relevant information about how the model could be expected to perform in practice.

Figure 3. Simulation 2, distributions of MSE differences between unweighted average vs others. This figure shows simulation conditions where mean bias is 0 (unbiased).

Figure 4. Simulation 2, distributions of MSE differences between unweighted average vs others. This figure shows simulation conditions where mean bias is 2 (biased).

Figure 5. Simulation 2, distributions of MSE differences between full model θ estimates vs other θ estimates. This figure shows simulation conditions where mean bias is 0 (unbiased).

Figure 6. Simulation 2, distributions of MSE differences between full model θ estimates vs other θ estimates. This figure shows simulation conditions where mean bias is 2 (biased).

Method

The data-generating model for Simulation 2 is the same base model used in Simulation 1, with three major modifications:

- For about half of the questions (randomly determined separately for each forecaster), forecasters exhibit a stable level of bias. For the other half, forecasters are unbiased.
- For about half of the questions (randomly determined separately for each forecaster), the forecaster’s judgment is unrelated to the question’s true value θ_i .
- For about half of the questions (randomly determined separately for each forecaster), forecasters exhibit more noise than they do for the other half (a ψ_j of 3 instead of 1). This increase in noise led the total variance in forecasts to at least double, with the specific increase depending on condition.

In using these modifications, we are attempting to mimic an applied situation where forecasters exhibit some regularity that is masked by unpredictable behavior. Aside from these modifications, the remaining simulation method is exactly the same as that used in Simulation 1.

Results

For the 6400 models estimated across all conditions, the Gelman-Rubin statistic indicated that 336 models failed to converge. These failures to converge generally occurred in the conditions with ten forecasters and low validity, and again more often for the constrained models than for the full model. Failures to converge are again removed from the results reported below, though we primarily focus on the full model results for simplicity.

The primary results are displayed in Figures 3 (no bias conditions) and 4 (bias conditions). These figures are similar to the analogous figures from Simulation 1, except that they only show the full model results. Unlike Simulation 1, we now see MSE differences that are often close to 0 or negative, implying that the unweighted average is more competitive with other aggregates. The model-based θ estimates continue to outperform the unweighted average in many conditions, especially when there are only ten subjects. But the magnitude of improvement is lower than that of Simulation 1, and the optimally-weighted average (based on model estimates) is as good as or better than the θ estimates in the conditions with 100 subjects and low validity. However, like Simulation 1, systematic forecast bias continues to make it easier to beat the unweighted average: many of the boxes in Figure 4 are fully above 0, reflecting situations where the unweighted average is always beaten.

In many panels of Figures 3 and 4, we see the green box doing worse than the other aggregates. The green box symbolizes the optimally-weighted average using “true” forecaster weights, computed from the known forecaster parameters without any noise modifications. These boxes help show how the optimally-weighted average can fail in noisy environments. It did not help to know the forecasters’ true characteristics when those characteristics were inconsistently displayed across questions. Instead, it was more helpful to fit an incorrect model to noisy data, where estimates of forecaster characteristics incorporated the noise in the data. In this way, the incorrect model acts as a type of “regularization” by trading bias in estimation for a decrease in variance of its predictions.

Expanding on model regularization, Figures 5 and 6 show MSE differences between the θ estimates arising from $\mathcal{M}_f$ as compared to the θ estimates arising from the reduced models. Positive values imply that a reduced model is better, and negative values imply that the full model is better. We see that, while the reduced models’ performance is generally similar to that of the full model, the reduced models sometimes perform slightly better and sometimes slightly worse than the full model, depending on condition.

Summary

Simulation 2 showed that, in situations closer to reality (where forecasters do not respond according to the proposed model), the unweighted average becomes more difficult to beat. In bias conditions, the model-based θ estimates and optimally-weighted average could sometimes beat the unweighted average by a modest amount. In unbiased conditions, the θ estimates sometimes exhibit a small improvement but often exhibit no improvement. This latter scenario is likely to mimic many applied situations, and it shows how methods that perform well in theoretical situations may not translate to applied situations. Relatedly, the “true” optimally-weighted average using known forecaster characteristics (absent noise) generally failed to beat the unweighted average, illustrating the sensitivity of the optimal weights to noisy forecasting environments.

While the Simulation 2 data generation was intended to mimic applied situations, additional sources of noise are also worthy of consideration. One important, potential source comes from correlated residuals: forecasters agreeing with one another due to common misconceptions, as opposed to forecasters agreeing with one another due to shared knowledge. Such noise may occur, for example, in forecasting scenarios where a major news story causes forecasters to ignore other information relevant to the event of interest (see Broomell & Budescu, 2009, for related discussion). To the extent that these additional sources of noise occur in real data, the Simulation 2 results will still be optimistic. In the following section, we further examine these issues by applying the model to real data.

Application

The simulations showed that, when forecasters do not follow the data-generating model, it becomes difficult to beat the unweighted average by a large amount. They also showed that model-based forecast aggregation (e.g., via θ estimates) often outperformed optimal weighting, and that “regularized” models with some forecaster attributes constrained to be equal could often perform as well as the full model. In this section, we extend these findings to real data. We specifically use point spread forecasts of 2015 Australian Football League (AFL) games, collected by the Monash group (Powell, Seemann, & Dowe, 2016). For each game, the point spread is defined as the number of points that the home team scored minus the number of points that the away team scored.

Method

We focus on 27 forecasters who each predicted the point spreads of all 206 AFL games that took place during 2015. Outcomes from the first three rounds of the season (27 games) were sent to the proposed model, reflecting a situation where we wish to weight forecasters based on their early season performance. We specifically chose the first three rounds because

it is close to 10% of the total number of games, which is similar to the 10% condition from the simulations. For each game, we examined the three aggregates from the simulations: the unweighted average, the optimally-weighted average, and the model-based θ estimates. For all aggregates, we compute their mean squared error across all AFL games after the first three rounds.

We estimated the full model from Equations (1) to (3), along with the three constrained versions of the model that were described earlier. For each model, we drew 2,000 posterior samples after 1,000 adaptation and 2,000 burn-in iterations. We obtained aggregates via both optimal weighting and θ estimates.

Prior distributions were again mildly informative, being set to

$$\begin{aligned}\delta_j &\sim N(0, 10) \\ \lambda_j &\sim N(.5, 1) \\ \psi_j &\sim \text{Inv-Gamma}(.1, .1) \\ \rho &\sim \text{Beta}(2, 8),\end{aligned}$$

with the beta prior reflecting the fact that the reported forecasts are expected to exhibit a high proportion of noise. Related to the beta prior, s_y^2 from Equation (10) was set to 1,000, which was approximately equal to the observed variance of reported forecasts.

The simulation results most relevant to this application involve the upper left panels of Figures 1 and 3, depicting situations with many questions, few forecasters, and lower validity. Those panels generally showed the optimal weighting and θ estimates outperforming the unweighted average, with θ potentially being the best.

Results

In the paragraphs below, we first provide a preliminary overview of the data and of the models. We then conduct more detailed study of model performance and extensions.

Preliminary overview. The unweighted average had an MSE of 1501.14 for the games in Round 4 and later, which translates to an average “miss” of about 6 or 7 goals. This is a relatively large miss, and Figure 7 shows a histogram of these differences to provide a clearer picture of the unweighted average’s performance. It is seen that, while the unweighted average was often close to the truth, there are also some high-scoring games that have a large impact on the MSE metric (when the unweighted average misses by 100 points or more).

Before comparing the models’ forecasting performance relative to the unweighted average, we did an initial comparison of models $\mathcal{M}_f$, $\mathcal{M}_\lambda$, and $\mathcal{M}_\delta$ via DIC. For each model, there are two unique DICs that can be computed, which differ by whether or not the θ parameters are integrated out of the model likelihood (see Appendix A for details). We call these “marginal” and “conditional” DIC. Table 1 shows both of these DICs for all four models; the models can be compared by either metric (but marginal DICs should not be compared to conditional DICs). It is seen that both types of DIC prefer $\mathcal{M}_f$ over the other two models, suggesting that we should prefer the full model. But it is also the case that the DICs focus on how well the models account for individual forecasts (the models

Figure 7. Histogram of differences between the unweighted average forecast and AFL game outcomes.

Table 1

DIC and effective number of parameters (p_d) for three models fitted to the AFL data.

Model	Marginal		Conditional	
	DIC	p_d	DIC	p_d
$\mathcal{M}_f$	47589.43	80.01	46703.28	249.73
$\mathcal{M}_\lambda$	47919.98	53.95	47054.55	223.37
$\mathcal{M}_\delta$	48049.58	28.97	47192.97	197.6

are being fit to individual forecasts), whereas we are most interested in aggregated forecasts here.

Performance of average forecasts. Next, we study the performance of the aggregated forecasts resulting from the full model. The optimally-weighted averages based on this model had an MSE of 1654.29, as compared to the unweighted average MSE of 1501.14. Further, the model-based θ estimates from this model had the worst MSE of 2427.75, a result that was not seen in the simulation results. The model-based aggregates are clearly not doing as well as the unweighted average.

To further study the full model, we examined the optimal weights resulting from the parameter estimates. The histogram of these weights is displayed in Figure 8, with the red line showing the weights that would be employed in the unweighted average (i.e., there are 27 forecasters, so equal weighting would be $1/27$). We see that the optimal weights are widely dispersed around the red line, reflecting the fact that the forecaster weights are now very unequal. These unequal weights seem to stray too far from equal weighting to be useful.

In contrast to what DIC suggests, the constrained versions of the model yield better

Figure 8. Optimal weights resulting from the full model’s application to the AFL data. The red line reflects equal weights that are associated with the “unweighted” average.

aggregates here, with optimal weighting under $\mathcal{M}_\lambda$ obtaining the best MSE of 1449.31. Optimal weights under $\mathcal{M}_\delta$ also beat the unweighted average in MSE, obtaining a value of 1449.8. To expand on these results, Figure 9 shows the distributions of optimal weights under each of the three models ($\mathcal{M}_f$, $\mathcal{M}_\lambda$, $\mathcal{M}_\delta$), with the corresponding MSEs in the panel titles. We see that, when λ is held equal across forecasters (as in $\mathcal{M}_\lambda$ and $\mathcal{M}_\delta$), the optimal weights are generally close to equal weighting. Conversely, when λ is unique across forecasters, the weights exhibit more dispersion and a worse MSE. This suggests that the data are too noisy to estimate a unique λ parameter per forecaster, a result that we did not directly encounter in our simulations.

Unlike our simulations, we also found that the θ estimates were worse than the optimally-weighted average across all models. The MSEs of the θ estimates were 2427.75, 2404.22, 2392.12 for $\mathcal{M}_f$, $\mathcal{M}_\lambda$, and $\mathcal{M}_\delta$, respectively. These poor MSEs resulted from the fact that the θ estimates exhibited more extremity than were warranted by the data, which can be seen by comparing the optimally-weighted averages to the θ estimates (Figure 10). The figure shows that the optimally-weighted averages and θ estimates are highly correlated with one another, but the variability in the θ estimates is much larger than the variability of the optimally-weighted averages. This shows that the θ estimates are not constrained by the reported forecasts in the same way as the optimally-weighted average, and the increased variability proves detrimental for MSE.

Improving the Model. Given the mismatch between our simulation and application results, we explored specific aspects of the model that contributed to its failure. It is not clear that these modifications will lead to general improvements in future applications; we are only considering modifications because the original model performed poorly. But we still find it useful to highlight particular aspects of the model’s failure, especially in light of the large variability observed in the θ estimates.

Figure 9. Optimal weights resulting from the three constrained models. Panels reflect pairs of models, with the diagonal titles showing the model that led to the weights. The red point reflects equal weights that are associated with the “unweighted” average. Numbers on the upper triangle show correlations between weights for analogous panels in the lower triangle.

Figure 10. Application, optimally-weighted averages vs θ estimates for the full model. Each point represents the aggregates for a single game.

We found that the model can be greatly improved by allowing for randomness in the games whose outcomes are known (i.e., whose outcomes are supplied to the model). That is, in the original form of our models, the outcomes of resolved games were directly inserted into Equation (2) in place of θ_i . In that case, the θ_i for resolved games are no longer parameters; they are instead fixed quantities. In the version of our model examined here, we model these θ_i as arising from a normal whose mean is the realized value:

$$\theta_i \sim N(\text{answer}_i, \sigma_\theta^2) \text{ for resolved games,} \quad (11)$$

where answer_i is the outcome of game i . The θ_i associated with unresolved questions then have the same hyperdistribution as before:

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2) \text{ for unresolved games.}$$

This allows for randomness in all game outcomes, while also allowing the model to capitalize off the outcomes of resolved games.

In making this single change, all versions of the model perform better than the unweighted average MSE of 1501.14. The MSEs of the θ estimates are best of all, exhibiting values of 1431.51, 1433.91, and 1432.26 for $\mathcal{M}_f$, $\mathcal{M}_\lambda$, and $\mathcal{M}_\delta$, respectively. The optimally weighted averages are not far behind, exhibiting MSEs of 1436.64, 1437.92 and 1437.91, respectively. These all beat the unweighted average and better match the simulation results.

Our (posthoc) explanation for this difference involves the fact that the AFL dataset contains a large amount of variability in outcomes, relative to reported forecasts: the variance in game outcomes is over twice the variance in reported forecasts across all games, whereas nearly the opposite was true of the artificial data from our simulations. Due to this large variability in outcomes relative to forecasts, fixing the θ_i to realized values for resolved AFL games placed a heavy burden on the model’s forecaster parameters; for resolved games, the model could only reproduce variability in reported forecasts via the forecaster parameters. This burden proved suboptimal for aggregating forecasts, which was not a direct goal of the model estimation. In contrast, use of (11) for resolved games allowed σ_θ^2 to help reproduce variability in those games’ reported forecasts, giving the forecaster parameters more freedom to reflect forecaster traits. This clearly led to improved aggregations here, highlighting the fact that subtle model developments can end up having large influences on model estimates.

General Discussion

In this paper, we developed a model of forecasters that was inspired by Cultural Consensus Theory and related models. We used the model to apply theoretical results on optimal weighting of forecasters to real data, obtaining a new aggregation method based on an “optimally-weighted” average. We then compared the optimally-weighted average to other forms of aggregation, including parameter estimation (of θ) and unweighted averaging. Optimal weighting could beat the unweighted average and compete with the model-based parameter estimates when the proposed model matched the data-generating process, and the three aggregates performed more similarly under noisier data-generating processes. In applying the results to real forecasting data, we found that optimal weighting generally outperformed the model-based parameter estimates, but a constrained version (or modification) of our model was necessary to beat the unweighted average.

Recommendations. Many unique aggregation methods are available under our proposed model, and the results in this paper fail to provide unequivocal evidence about a single best method. The simulations generally pointed towards θ estimates of our full model, while the optimally-weighted average beat the θ estimates in the application (when using the “unmodified” model). We offer the following recommendations based on these findings and on our experience with the model.

- The optimally-weighted average is directly connected to the reported forecasts in a way that the θ estimates are not. Despite the fact that the optimally-weighted average was not the best in all simulation conditions, we prefer it over the θ estimates for applications.
- In the simulations, aggregation under constrained models often performed similarly to aggregation under the full model. In the application, aggregation under constrained models was sometimes better than aggregation under the full model, and never considerably worse than the full model. We therefore recommend use of $\mathcal{M}_\lambda$, which constrains the λ parameters to be equal across subjects. This allows for bias correction while keeping the forecaster weights closer to equal, as compared to the full model.
- Aggregation results can be sensitive to the prior distribution on σ_θ^2 (or equivalently on ρ ; see Equation (10)). Because many forecasting environments of interest are inherently difficult (such as predicting point spreads), we recommend strong priors reflecting a large proportion of noise in the reported forecasts.

Of course, it will also be useful to incorporate domain knowledge into the aggregations. For example, based on the scoring of Australian football games, some point spreads are more common than other point spreads. It would therefore be possible to modify our model so that aggregates are smoothed towards the common point spreads. While this strategy does not easily scale to situations with diverse forecasting topics, it can clearly lead to improvements in aggregated forecasts.

Incorporating such information into the model raises some interesting conceptual issues. First and foremost, our model is designed to be *cognitive* in nature, in that the model’s features are intended to be informative regarding individual forecasters’ behavioral attributes (e.g., bias in forecasts). One interpretation of including such point spread information could be that it reflects a group of individuals’ prior exposure to Australian football games, which would still be cognitive in nature. On the other hand, such information could also be included within a more applied forecasting framework where such an interpretation was not warranted and the goal was simply to provide the most accurate forecast possible, regardless of psychological interpretation. We note that, depending upon the research question, it may be beneficial to ‘blur these lines’ between model interpretations, but the overall goals and development of the current model are directed towards understanding forecaster behavior through cognitive modeling.

Model Extensions. There exist other extensions that can potentially improve our model’s performance and utility. One extension involves a move beyond point aggregates, towards distributions that reflect uncertainty in the aggregates. That is, the Bayesian models proposed in this paper make it possible to push uncertainty about individual model

parameters through to the optimal weights derived after model estimation. Specifically, we could obtain a posterior distribution of each forecaster’s optimal weight by computing optimal weights for each posterior draw of model parameters. This would then lead to a posterior distribution of the optimally-weighted averages for each forecasting question. These posteriors could be summarized in multiple manners, and large posterior uncertainty might make us cautious about use of the optimally weighted average. Related to this point, Carpenter (2019) provides some recent evidence that averaging over the posterior distribution of aggregates could be better than aggregating predictions based on point estimates of model parameters. Computational complexity is potentially problematic here, however, especially in the simulation-based scenarios that we examined.

Another obvious extension entails extending optimal weighting beyond continuous forecasts, towards probabilistic forecasts and others. For example, instead of predicting point spreads, forecasters may report the probability that the point spread is exceeded. While it is relatively easy to specify a logit (or probit) version of the model that can be applied to probabilistic forecasts, the θ parameters could never be observed there. This is because the θ parameters in that model would reflect the “true” logit-probability of event occurrence, and we only observe event occurrence or non-occurrence (as opposed to event probabilities). Further, the Davis-Stober et al. (2015) optimal weighting theorem only covered continuous forecasts, so we would require extensions of that theorem in order to apply optimal weighting to other types of forecasts.

Concluding Remarks. In conclusion, we follow previous researchers who have remarked that, in many forecasting environments, no method will consistently beat the unweighted average by a large margin. We do not expect that the model and recommendations made here will consistently beat the unweighted average by a large amount (unless, e.g., forecasters are excessively biased in one direction); we instead are interested in averaging methods that do not stray too far from equal weighting but that can capitalize on stable individual differences across forecasters. We think the methods proposed here are suited to the task, balancing model complexity, forecaster traits, and (un)equal weighting in an interpretable manner.

References

- Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*, 151–181.
- Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13.
- Armstrong, J. S. (2001). *Principles of forecasting*. Norwell, MA: Kluwer Academic.
- Aßfalg, A. (2018). Consensus theory for mixed response formats. *Journal of Mathematical Psychology*, *86*, 51–63.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145. doi: 10.1287/deca.2014.0293
- Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332.
- Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92.
- Batchelder, W. H., Strashny, A., & Romney, A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. In S.-K. Chai, J. J. Salerno, & P. L. Mabry (Eds.), *Social computing, behavioral modeling, and prediction (SBP) 2010* (pp. 98–107). Lecture Notes in Computer Science 6007. doi: 10.1007/978-3-642-12079-4_15
- Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making*, *12*(2), 90–103.
- Broomell, S., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, *74*, 531–553.
- Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.
- Carpenter, B. (2019). For probabilistic prediction, full Bayes is better than point estimators. *Working paper*. Retrieved from <https://github.com/bob-carpenter/case-studies/blob/master/bayes-versus/bayes-versus.pdf>
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Thomson Learning.
- Cuzán, A. G., & Bundrick, C. M. (2009). Predicting presidential elections with equally weighted regressors in Fair’s equation and the fiscal model. *Political Analysis*, *17*, 333–340.
- Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika*, *76*, 650–669.
- Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., & Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*, 130–143.
- Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.
- Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*, 1–25. doi: 10.18637/jss.v071.i09
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–511.

- Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, 68, 1792–1799.
- Graefe, A., Küchenhoff, H., Stierle, V., & Riedl, B. (2015). Limitations of ensemble Bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3), 943–951. doi: <https://doi.org/10.1016/j.ijforecast.2014.12.001>
- Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68, 1678–1685.
- Kadane, J. B., & Fischhoff, B. (2013). A cautionary note on global recalibration. *Judgment and Decision Making*, 8(1), 25–27.
- Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, 68, 373–389.
- Lamberson, P., & Page, S. E. (2012). Optimal forecasting groups. *Management Science*, 58(4), 805–810.
- Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9(3), 259–273.
- Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, 13(4), 322–333.
- Lee, M. D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PLOS ONE*, 9(5), 1–9. doi: 10.1371/journal.pone.0096431
- Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 59–72. doi: 10.1207/s15328007sem1301_3
- Merkle, E. C. (2010). Calibrating subjective probabilities using hierarchical Bayesian models. In S.-K. Chai, J. J. Salerno, & P. L. Mabry (Eds.), *Social computing, behavioral modeling, and prediction (SBP) 2010* (pp. 13–22). Lecture Notes in Computer Science 6007. doi: 10.1007/978-3-642-12079-4_4
- Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84, 802–829.
- Merkle, E. C., & Steyvers, M. (2011). A psychological model for aggregating judgments of magnitude. In J. Salerno, S. J. Yang, D. Nau, & S.-K. Chai (Eds.), *Social computing and behavioral-cultural modeling 2011* (pp. 236–243). Lecture Notes in Computer Science 6589. doi: 10.1007/978-3-642-19656-0_34
- Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision*, 3, 1–19.
- Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, 33, 817–832.
- Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2015). Calibrating ensemble forecasting models with sparse data in the social sciences. *International Journal of Forecasting*, 31, 930–942.
- Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian cultural consensus theory. *Field Methods*, 26, 207–222.
- Peeters, C. F. W. (2012). Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, 77, 288–292.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*.
- Powell, D., Seemann, T., & Dowe, D. (2016). *Monash University probabilistic footy tipping competition*. (Retrieved from <http://probabilistic-footy.monash.edu>)
- Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306, 462–466.

- Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*, 532–535. doi: 10.1038/nature21054
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155–1174.
- Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist*, *31*, 163–177.
- Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313–338.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. doi: <https://doi.org/10.1016/j.ijforecast.2013.09.009>
- Shlomi, Y., & Wallsten, T. S. (2010). Subjective recalibration of advisors’ probability estimates. *Psychonomic Bulletin & Review*, *17*, 492–498.
- Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, *64*, 583–639.
- Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*, 261–289. doi: 10.1007/s10994-013-5401-4
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.
- Wainer, H. (1976). Estimating coefficients in linear models: It don’t make no nevermind. *Psychological Bulletin*, *83*, 213–217.
- Wainer, H., & Thissen, D. (1976). Three steps towards robust regression. *Psychometrika*, *41*(1), 9–34.
- Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika*, *73*(4), 691.
- Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*(1), 23–40.

Appendix A

DIC Computations

For models with hierarchical (“random”) parameters, there exist multiple DICs for a single model, differing based on the specific likelihood used for the computation. This issue was mentioned in the original DIC paper (Spiegelhalter, Best, Carlin, & van der Linde, 2002), where it was called the model “focus”. More recently, the issue was elaborated upon by Merkle, Furr, and Rabe-Hesketh (2019) in the context of psychometric and multilevel models.

As applied to the models in this paper, we can compute a “conditional” DIC by conditioning on the θ parameters associated with unresolved questions, and we can compute a second, “marginal” DIC by integrating those θ parameters out of the model likelihood. The conditional DIC involves a univariate normal likelihood for each reported forecast, with the likelihood being described by Equations (1) to (3). The marginal DIC, on the other hand, involves a multivariate normal likelihood for each unresolved question:

$$\mathbf{y}_i \sim N(\boldsymbol{\delta} + \boldsymbol{\Lambda}\mu_{\theta}, \sigma_{\theta}^2\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \mathbf{D}_{\psi}), \quad (12)$$

where $\mathbf{y}_i$ is a vector of reported forecasts for question i , $\boldsymbol{\delta}$ and $\mathbf{\Lambda}$ are $J \times 1$ vectors of forecaster parameters, and $\mathbf{D}_\psi$ is a diagonal matrix of forecaster noise variances.

As described by Merkle et al. (2019), the two DIC metrics measure the model's ability to generalize to two different types of new data. The conditional DIC gets at the model's ability to generalize to new forecasters who are reporting on the same questions that were present in the original dataset, whereas the marginal DIC gets at the model's ability to generalize to new forecasters who are reporting on new questions that were not present in the original dataset. Neither metric directly measures the model's ability to produce good aggregated forecasts, which likely explains why neither DIC selected the model that performed best in the application section.

Appendix B Model Expressions

In this appendix, we show how the properties of expectations and covariances can be used to obtain the model expressions for forecaster biases, covariances, and validities that are shown in Equations (4) to (7). For further detail about use of expectation and covariance operators, see, e.g., Casella and Berger (2002).

$$\begin{aligned} \text{bias}(y_{ij}) &= E(y_{ij} - \theta_i) \\ &= E(\delta_j + \lambda_j \theta_i - \theta_i) \\ &= \delta_j + (\lambda_j - 1)E(\theta_i) \\ &= \delta_j + (\lambda_j - 1)\mu_\theta \end{aligned}$$

$$\begin{aligned} \text{var}(y_{ij}) &= \text{var}(\delta_j + \lambda_j \theta_i + e_{ij}) \\ &= \lambda_j^2 \text{var}(\theta_i) + \text{var}(e_{ij}) \\ &= \lambda_j^2 \sigma_\theta^2 + \psi_j^2 \end{aligned}$$

$$\begin{aligned} \text{cov}_{j \neq k}(y_{ij}, y_{ik}) &= \text{cov}(\delta_j + \lambda_j \theta_i + e_{ij}, \delta_k + \lambda_k \theta_i + e_{ik}) \\ &= \text{cov}(\lambda_j \theta_i, \lambda_k \theta_i) \\ &= \lambda_j \lambda_k \sigma_\theta^2 \end{aligned}$$

$$\begin{aligned} \text{cov}_{j \neq k}(y_{ij}, \theta_i) &= \text{cov}(\delta_j + \lambda_j \theta_i + e_{ij}, \theta_i) \\ &= \text{cov}(\lambda_j \theta_i, \theta_i) \\ &= \lambda_j \sigma_\theta^2. \end{aligned}$$