Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Ensembles of Overfit and Overconfident Forecasts

Yael Grushka-Cockayne, Victor Richmond R. Jose, Kenneth C. LichtendahlJr.

To cite this article:
Yael Grushka-Cockayne, Victor Richmond R. Jose, Kenneth C. LichtendahlJr. (2017) Ensembles of Overfit and Overconfident Forecasts. Management Science 63(4):1110-1130. https://doi.org/10.1287/mnsc.2015.2389

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 4, April 2017, pp. 1110–1130
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Ensembles of Overfit and Overconfident Forecasts

Yael Grushka-Cockayne,[a] Victor Richmond R. Jose,[b] Kenneth C. Lichtendahl Jr.[c]

[a] Darden School of Business, University of Virginia, Charlottesville, Virginia 22903; [b] McDonough School of Business, Georgetown University, Washington, DC 20057; [c] Darden School of Business, University of Virginia, Charlottesville, Virginia 22903

Contact: grushkay@darden.virginia.edu (YG-C); vrj2@georgetown.edu (VRRJ); lichtendahlc@darden.virginia.edu (KCL)

Received: August 1, 2014
Revised: March 21, 2015; August 18, 2015
Accepted: September 28, 2015
Published Online in Articles in Advance: April 20, 2016

https://doi.org/10.1287/mnsc.2015.2389

Abstract. Firms today average forecasts collected from multiple experts and models. Because of cognitive biases, strategic incentives, or the structure of machine-learning algorithms, these forecasts are often overfit to sample data and are overconfident. Little is known about the challenges associated with aggregating such forecasts. We introduce a theoretical model to examine the combined effect of overfitting and overconfidence on the average forecast. Their combined effect is that the mean and median probability forecasts are poorly calibrated with hit rates of their prediction intervals too high and too low, respectively. Consequently, we prescribe the use of a trimmed average, or trimmed opinion pool, to achieve better calibration. We identify the random forest, a leading machine-learning algorithm that pools hundreds of overfit and overconfident regression trees, as an ideal environment for trimming probabilities. Using several known data sets, we demonstrate that trimmed ensembles can significantly improve the random forest's predictive accuracy.

History: Accepted by James Smith, decision analysis.
Supplemental Material: The online supplement is available at https://doi.org/10.1287/mnsc.2015.2389.

Keywords: wisdom of crowds • base-rate neglect • linear opinion pool • trimmed opinion pool • hit rate • calibration • random forest • data science

1. Introduction

Instead of relying on a single expert's opinion, firms today gather forecasts from multiple experts and models. Once forecasts are gathered, these firms face the challenge of forming an aggregate opinion. When aggregating forecasts, the simple average has proven to be a powerful heuristic that captures the wisdom of a crowd (Clemen and Winkler 1986, Armstrong 2001, Larrick and Soll 2006, O'Hagan et al. 2006). Furthermore, the crowd can be wise even when individual forecasts are poor.

Overfitting and overconfidence are two of the primary reasons individual forecasts are poor. Psychologists have shown that humans have a general tendency to overfit to sample data they see and to be overconfident in the forecasts they provide. In other words, they overrely on sample data and underestimate variance. In competitive forecasting environments, economists argue that forecasters have an incentive to respond strategically with overfit and overconfident forecasts. By design, machine-learning algorithms aggregate forecasts from many "weak learners"—models that are often overfit and overconfident.

Because overfitting and overconfidence are such pervasive phenomena, the first question we address in this paper is how these phenomena affect the simple average of point or probability forecasts, also known as the linear opinion pool. Little is known about the combined effects of overfitting and overconfidence on the linear opinion pool. In light of the effects we find, the second question we investigate is how well an alternative heuristic aggregates overfit and overconfident forecasts. The alternative heuristic we consider is the trimmed opinion pool, a close relative of the linear opinion pool.

Psychologists attribute overfitting in humans to cognitive biases such as base-rate or system neglect (Bar-Hillel 1980, Tversky and Kahneman 1982, Massey and Wu 2005, Barbey and Sloman 2007). As Budescu and Yu (2007, p. 173) point out, forecasters "are more sensitive to the external 'signals' generated by probabilistic systems than to their basic underlying parameters because these signals are unambiguous and transparent, while the background parameters are not directly observable and, often, are harder to assess." Overconfidence is one of the most documented findings in the probability judgment literature (Lichtenstein et al. 1982, DeBondt and Thaler 1995). Radzevick and Moore (2011, p. 94) state that "overprecision, the excessive certainty that one has the right answer, is the most robust variety of overconfidence . . . . Yaniv and Foster (1995, 1997) argue that people express overprecision because it increases the informativeness of what they say."

In winner-take-all point forecasting competitions, people have an incentive to stand out from the crowd

by exaggerating their own private sample information (Ottaviani and Sørensen 2006, Lichtendahl et al. 2013a). In exaggerating, forecasters place too much weight on their private information. In binary-event forecasting contests, optimal reports are extreme probabilities that are too near 0 or 1 (Lichtendahl and Winkler 2007). These responses translate into forecasts that are overfit and overconfident, respectively.

Overfitting and overconfidence are present even when we consider models instead of human experts. The random forest, a popular machine-learning algorithm, averages the point predictions of many regression trees (Breiman 2001). The trees in a random forest are notorious for overfitting to the data they are trained on (Hastie et al. 2009). Later we will show that the trees in a random forest can be overconfident as well. Despite the trees' weaknesses, the random forest has been shown to be a very accurate prediction tool. In 2012, Jeremy Howard, chief scientist at Kaggle, the world's leading platform for data prediction competitions, noted that "ensembles of [regression] trees (often known as 'Random Forests') have been the most successful general-purpose algorithm in modern times. For instance, most Kaggle competitions have at least one top entry that heavily uses this approach" (Howard and Bowles 2012).

To study how these phenomena affect pooled forecasts, we introduce a theoretical model that incorporates overfitting and overconfidence into a forecasting environment where forecasters' learning from sample data is characterized by a linear updating equation. This model allows forecasters (i) to learn from data generated from any exponential-family likelihood and its conjugate prior and (ii) to report forecasts that are distorted in an intuitively appealing way. Our model is in the spirit of recent attempts to incorporate cognitive biases and distortions into formal aggregation methods (Turner et al. 2014, Lee and Danileiko 2014). These methods include "psychologically interpretable parameters" that describe the ways in which forecasters are miscalibrated in a binary-event setting. Here, we consider a model with parameters that describe forecasters who are overfit and overconfident in continuous-quantity and discrete-event settings.

Our results support the intuition that although overfitting reduces individual expertise, it increases the crowd's diversity. Overall, overfitting leads to an improved crowd's point forecast. For probability forecasts of continuous quantities, however, we show that an overfitting forecaster will be poorly calibrated, even if she is neither overconfident nor underconfident. In other words, her hit rate—the percentage of realizations that fall within her central prediction intervals—will be too low. Whereas low hit rates have traditionally been diagnosed as overconfidence (Hora 2004, Gneiting et al. 2007, Lichtendahl et al. 2013b), we demonstrate that overfitting may instead be the cause of individual miscalibration.

For aggregate probability forecasts of a continuous quantity, we show that the combination of overfitting and overconfidence leads to mean and median probability forecasts that are poorly calibrated, and in opposite ways, with hit rates of their prediction intervals too high and too low, respectively. Consequently, we prescribe the use of a trimmed average probability forecast, or trimmed opinion pool—a practical alternative to the linear opinion pool introduced by Jose et al. (2014). Whereas Jose et al. (2014) examine these average probability forecasts as a function of experts' reported means and variances, we study average probability forecasts at a more foundational level. We derive forecasters' reported means and variances from the data forecasters use to form beliefs and from the distortions they apply in reporting these beliefs. This theoretical study contributes a richer understanding of the source and nature of individual and aggregate miscalibration. Furthermore, we hope the model in this study will serve as a platform for studying a variety of related phenomena.

Using our theoretical model with a normal prior-normal likelihood information structure, we identify the limiting trimmed opinion pools (LTOPs) as the number of forecasters grows large. This limiting distribution turns out to have the same distribution as the sum of a normal random variable and a truncated normal random variable. The limiting mean and median probability forecasts emerge from trimming at 0% and 50%, respectively. Importantly, the more we trim, the less disperse the limiting trimmed opinion pool becomes. This result allows us to show when a trimming level exists for which the trimmed opinion pool is better calibrated than the linear opinion pool. In addition to these results on forecasting a continuous quantity, we provide results for a binary-classification task when forecasters learn from data distributed according to a beta prior-Bernoulli likelihood information structure.

Finally, we apply trimmed opinion pools to random forests. In an empirical study of a popular synthetic data set, we draw parallels between our theoretical model and properties of the random forest. Then, using several business-relevant data sets from the public Comprehensive `R` Archive Network (`CRAN`) and University of California, Irvine (UCI) data repositories, we demonstrate how the random forest's point and probability forecasts can benefit from trimming. Our routine for trimming the trees in the random forest extends the work of Meinshausen (2006). He applies the linear opinion pool to the random forest and finds that the simple average of the trees' probability forecasts compares favorably to various quantile regression methods. Prior to Meinshausen (2006), the random forest of

regression trees was seen as a point forecasting tool. Meinshausen's work and ours support the broader movement toward working with probability forecasts. Probability forecasts are useful to practicing managers because they contain critical information about a quantity's uncertainty, and point forecasts can always be derived from them.

2. Theoretical Forecasting Environment

We introduce a forecasting environment that includes two major elements: (i) a sampling process characterized by a general information structure and (ii) a reporting process that incorporates the possibility for forecasters to overfit and be overconfident. Such forecasters exhibit more-than-optimal reliance on sample data and report less-than-optimal variance. In this structure, k exchangeable forecasters observe a sample of size n from the data $(y_1, \ldots, y_{kn})$ and forecast the uncertain quantity of interest $y = y_{kn+1}$. Each data point y_l is independently and identically drawn (iid) from a common likelihood $L(\boldsymbol{\theta})$ where the unknown d-dimensional parameter $\boldsymbol{\theta}$ is drawn from a common prior distribution $\pi(\boldsymbol{\tau})$ with $(d+1)$-dimensional hyperparameter $\boldsymbol{\tau}$. This general information structure is denoted by

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi(\boldsymbol{\tau}), \\ (y_l \mid \boldsymbol{\theta}) &\sim_{\text{iid}} L(\boldsymbol{\theta}) \quad \text{for } l = 1, \ldots, kn+1. \end{aligned} \tag{1}$$

Forecaster i's reported beliefs about y given her sample $\mathbf{y}_i = (y_{(i-1)n+1}, \ldots, y_{in})$ are denoted by the cumulative distribution function (cdf) $F_i(y)$. Note that when forecasters have biases, these reported beliefs may not be consistent with this information structure.

Throughout, we assume that the conditional expectations of some relevant statistics $(h_1(y), \ldots, h_d(y))$, used to characterize $F_i(y)$, satisfy the following linearity property given below in Assumption 1. Later, we will choose these relevant statistics to be the sufficient statistics of the likelihood in (1). We also assume in this paper that the first moment will always be relevant.

Assumption 1 (Linear Expectations of Relevant Statistics). *The expectation of the relevant statistic $h_j(y)$ given $\mathbf{y}_i$ is a linear combination of some constant y_{0j} and the sample average of $h_j(y_{(i-1)n+1}), \ldots, h_j(y_{in})$; i.e.,*

$$E[h_j(y) \mid \mathbf{y}_i] = (1-w)y_{0j} + w\frac{1}{n}\sum_{l=1}^{n} h_j(y_{(i-1)n+l}), \tag{2}$$

where $0 < w < 1$ for $j = 1, \ldots, d$. In particular, $h_1(y) = y$.

For example, when $d = 2$ and the first two noncentral moments are expectations of the relevant statistics, we have the following posterior expectations: $E[h_1(y) \mid \mathbf{y}_i] = (1-w)y_{01} + w(1/n)\sum_{l=1}^{n} y_{(i-1)n+l}$ and $E[h_2(y) \mid \mathbf{y}_i] = (1-w)y_{02} + w(1/n)\sum_{l=1}^{n} y^2_{(i-1)n+l}$. Here, when y_{0j} is the prior expectation of the statistic $h_j(y)$, this statistic's posterior expectation is a weighted average of its prior expectation and its sample average, which can be seen as a natural way to learn from sample data. In Section 2.1 below, we provide several examples of specific information structures in which this type of linear learning applies in a Bayesian updating context.

Next, we formally define overfitting and overconfidence in the context of our general information structure and Assumption 1. In these definitions, we will focus on each forecaster's reported mean $\mu_i' = \int y\, dF_i(y)$ and reported precision $\lambda_i' = (\int (y-\mu_i')^2\, dF_i(y))^{-1}$. Let $\bar{y}_i = (1/n)\sum_{l=1}^{n} y_{(i-1)n+l}$.

Definition 1 (Overfitting). A reported forecast F_i with mean $\mu_i' = (1-w')y_{01} + w'\bar{y}_i$ is overfit if $w' > w$, underfit if $w' < w$, and well fit if $w' = w$, where w is the weight in the conditional expectation given in (2).

Definition 2 (Overconfidence). A reported forecast F_i with precision λ_i' is overconfident if $\lambda_i' > \lambda_i$, underconfident if $\lambda_i' < \lambda_i$, and precise if $\lambda_i' = \lambda_i$, where $\lambda_i = (E[(y - E[y \mid \mathbf{y}_i])^2 \mid \mathbf{y}_i])^{-1}$ is the precision (or the reciprocal of the variance) of $(y \mid \mathbf{y}_i)$.

The type of overfitting we consider is overweighting of a forecaster's private sample average. This overweighting can be interpreted as a one-dimensional analog of the traditional notion of overfitting in the statistics literature. Statisticians often illustrate overfitting in two dimensions with regression models that pass through every data point. In these illustrations, the two-dimensional space is the x–y plane, where y is a response variable and x is a predictor (or covariate). The essential similarity between such regressions and our forecasters above is that they overfit by overrelying on the sample data, and thus their out-of-sample predictions are poor.

To make this analogy more concrete, consider the typical supervised learning framework applied to forecaster i's in-sample training set $\mathbf{y}_i$ and her out-of-sample testing set y. If the model she fits to the training set consists of a single point μ_i', and she wants to minimize the sum of squared errors in the training set $\sum_{l=1}^{n}(y_{(i-1)n+l} - \mu_i')^2$, she should select the model $\mu_i' = \bar{y}_i$; i.e., let $w' = 1$. This point, however, will perform poorly (on average) in the testing set, because the expected squared error $E[(y-\mu_i')^2 \mid \mathbf{y}_i]$ in the testing set is minimized by choosing $\mu_i' = (1-w)y_{01} + w\bar{y}_i$. Consequently, the single-point model best fit to the training set is overfit.

To map our model more exactly to the traditional notion of overfitting, we can think of the model in (1) more generally by conditioning the sampling process on some known predictors $\mathbf{x} = (x_1, \ldots, x_q)$. Under such conditioning, our prior and likelihood become $\pi(\boldsymbol{\tau}_{\mathbf{x}})$ and $L(\boldsymbol{\theta}_{\mathbf{x}})$, respectively. Nonetheless, until we get to our

empirical studies in Section 3, we suppress the notation of any covariates and analyze overfitting as defined above in one dimension. Importantly, with Definition 1, we can build a reasonable model that captures the essence of overfitting and allows us to study its effects in closed form.

Finally, we note that our definition of overconfidence utilizes reported precision and is devoid of any notion of calibration. Later, we define calibration separately and show that either overfitting or overconfidence can lead to poor calibration. This result upends the previously assumed equivalence of overconfidence and poor calibration.

2.1. Exponential-Family Information Structures

We say an information structure is an exponential-family information structure if the likelihood $L(\boldsymbol{\theta})$ in (1) is a member of the regular exponential family and the prior $\pi(\boldsymbol{\tau})$ is its conjugate. The regular exponential family encompasses a large class of distributions, such as the normal, gamma, log-normal, inverse Gaussian, Weibull, beta, Dirichlet, Bernoulli, categorical, binomial, multinomial, Poisson, geometric, Pareto, and negative binomial. According to Bernardo and Smith (2000, Proposition 5.7), Assumption 1 holds for any exponential-family information structure. Diaconis and Ylvisaker (1979) show that an exponential-family likelihood and Assumption 1 often imply that π must be the exponential-family member's conjugate prior.

The following are four exponential-family information structures that we examine in this paper (for more details on these structures, see Bernardo and Smith 2000, pp. 436–440; Robert 2001, p. 121).

(a) *Normal-normal*: Normal prior $\mathrm{N}(\mu \mid \mu_0, m\lambda)$ and normal likelihood $\mathrm{N}(y \mid \mu, \lambda)$, where $\mathrm{N}(y \mid \mu, \lambda) = (\lambda^{1/2}/(2\pi)^{1/2})\exp(-\frac{1}{2}\lambda(y-\mu)^2)$, $d = 1$, $w = n/(m+n)$, and $y_{01} = \mu_0$.

(b) *Gamma-gamma*: Gamma prior $\mathrm{Ga}(\mu \mid ma+1, m\mu_0)$ and gamma likelihood $\mathrm{Ga}(y \mid a, \mu)$, where $\mathrm{Ga}(y \mid a, \mu) = (\mu^a/\Gamma(a))y^{a-1}\exp(-\mu y)$, $d = 1$, $w = n/(m+n)$, and $y_{01} = \mu_0$.

(c) *Normal-gamma-normal*: Normal-gamma prior $\mathrm{N}(\mu|\mu_0, m\lambda)\mathrm{Ga}(\lambda \mid a, b)$ and normal likelihood $\mathrm{N}(y|\mu, \lambda)$, where $d = 2$, $w = n/(m+n)$, $h_2(y) = y^2$, $y_{01} = \mu_0$, and $y_{02} = 2b/m + \mu_0^2$ when $a = (m+1)/2 + 1$.

(d) *Beta-Bernoulli*: Beta prior $\mathrm{Be}(\mu \mid m\mu_0, m(1-\mu_0))$ and Bernoulli likelihood $\mathrm{Br}(y \mid \mu)$, where $\mathrm{Be}(\mu \mid m\mu_0, m(1-\mu_0)) = (\Gamma(m)/(\Gamma(m\mu_0)\Gamma(m(1-\mu_0))))\mu^{m\mu_0-1} \cdot (1-\mu)^{m(1-\mu_0)-1}$, $\mathrm{Br}(y \mid \mu) = \mu^y(1-\mu)^{1-y}$, $d = 1$, $w = n/(m+n)$, and $y_{01} = \mu_0$.

In a d-dimensional regular exponential family, $(h_1(y), \ldots, h_d(y))$ are referred to as natural sufficient statistics. In the examples above, y_{01} is the prior expectation of $h_1(y)$, or $E[y]$. In the normal-gamma-normal information structure where $d = 2$, y_{02} is the prior expectation of $h_2(y)$, or $E[y^2]$. Consequently, the form of learning with any exponential-family information structure is linear in the prior expectation $E[h_j(y)]$ and the sample sufficient statistic $(1/n)\sum_{l=1}^{n} h_j(y_{(i-1)n+l})$. In all four specific information structures above, the prior parameter m dictates the weight put on the prior expectation in this learning process (because $1 - w = m/(m+n)$ in each).

In the next two subsections, we explore the implications of overfitting and overconfidence on forecasting continuous and discrete quantities. We specifically focus on the normal-normal and the beta-Bernoulli information structures. It is important to highlight that assuming a normal-normal information structure is not as restrictive as it may appear. Consider the typical regression context, where errors are assumed to be normal. Even when errors are not normally distributed, statisticians will often apply a transformation to the response variable before they reject normality (e.g., the popular Box-Cox transformation). With the option of a transformation, the class of distributions associated with the normal-normal information structure is quite large.

2.2. Continuous Quantities Following the Normal-Normal Information Structure

Next we derive several analytical results using the normal-normal information structure. We also provide numerical results associated with the gamma-gamma and normal-gamma-normal information structures.

In the normal-normal information structure, after forecaster i observes her sample, she should report that $(y \mid \mathbf{y}_i) \sim \mathrm{N}(\mu_i, \lambda_i)$, where $\mu_i = (1-w)\mu_0 + w\bar{y}_i$ and $\lambda_i = ((m+n)/(m+n+1))\lambda$. This inference follows from a straightforward application of Bayes' theorem (Bernardo and Smith 2000). Instead, we suppose each forecaster reports that $(y \mid \mathbf{y}_i) \sim \mathrm{N}(\mu_i', \lambda_i')$, where $\mu_i' = (1-w')\mu_0 + w'\bar{y}_i$. We denote forecaster i's reported cdf is $F_i(z) = \Phi(\lambda_i'^{1/2}(z-\mu_i'))$, where Φ is the standard normal cdf. We often use z as a placeholder for a realization of the random variable y so as not to confuse y and its realization. Also let ϕ denote the probability density function (pdf) of the standard normal.

With this information structure, we first examine the impact of overfitting on the accuracy of an individual's point forecast μ_i' as well as the crowd's point forecast, the simple average of the individuals' point forecasts $\bar{\mu} = (1/k)\sum_{i=1}^{k}\mu_i'$. Throughout, mean squared error is used to measure point forecasting accuracy. The smaller an individual's mean squared error $\mathrm{MSE}(\mu_i') = E[(y-\mu_i')^2]$ is, the more expertise an individual has. We define an individual's *expertise* E_i as the reciprocal of her mean squared error, or $E_i = 1/\mathrm{MSE}(\mu_i')$. Because a forecaster's mean squared error is maximized when $\mu_i' = \mu_i$, a forecaster's expertise is highest when she is well fit ($w' = w$).

To gauge the wisdom of the crowd, we consider the percentage improvement in mean squared error the crowd offers over "a randomly chosen member of the crowd." We call this percentage improvement the *wisdom of the crowd*,

$$W = \frac{(1/k)\sum_{i=1}^{k} \mathrm{MSE}(\mu_i') - \mathrm{MSE}(\bar{\mu})}{(1/k)\sum_{i=1}^{k} \mathrm{MSE}(\mu_i')}, \tag{3}$$

where $(1/k)\sum_{i=1}^{k} \mathrm{MSE}(\mu_i')$ can be interpreted as either the expected MSE of a randomly chosen forecaster or the average MSE of members in the crowd. By Jensen's inequality, we know that $(1/k)\sum_{i=1}^{k} \mathrm{MSE}(\mu_i') \geq \mathrm{MSE}(\bar{\mu})$. Thus, $0 \leq W \leq 1$. The greater W is, the wiser the crowd.

Mannes et al. (2014) use a measure similar to W, with mean absolute error in place of mean squared error. Davis-Stober et al. (2014) measure the crowd's wisdom using the numerator in (3), but with unequal weights. Unequal weights allow them to make predictions about the wisdom of a crowd when there are varying levels of expertise. For ease of exposition, we restrict our attention to a crowd of exchangeable experts.

Larrick et al. (2012) stress that a crowd's wisdom is a function of two factors: individual expertise and crowd diversity. We define the crowd's *diversity* D by the expected sample variance in the individual point forecasts, or $E[(1/k)\sum_{i=1}^{k}(\mu_i' - \bar{\mu})^2]$. The concept of diversity here is an ex post one. Although forecasters draw their private samples from the same distribution, their point forecasts will differ from $\bar{\mu}$. The more their point forecasts differ from $\bar{\mu}$, the more diversity we say there is in the crowd. The following result shows that, in the presence of overfitting in our forecasting environment, diversity is the overriding factor. Proofs of all results are given in the appendix.

Proposition 1. *In our forecasting environment with a normal-normal information structure, the wisdom of the crowd is* $W = E_i D$, *and the following statements hold*:

(a) *Individual expertise* $E_i = (1 + w'^2/n + (1-w')^2/m)^{-1}\lambda$ *is decreasing in* $|w' - w|$.

(b) *Diversity* $D = ((k-1)w'^2/(kn))(1/\lambda)$ *is increasing in* w' *for* $w' > 0$.

(c) *The wisdom of the crowd* W *is increasing in* w' *for* $0 < w' \leq 1$.

This result is related to an insight from Larrick et al. (2012, p. 234): "The benefits of diversity are so strong that one can combine the judgments from individuals who differ a great deal in their individual accuracy and still gain from averaging." Here, even when individuals are equally poor, the gain from averaging still exists. Next, we show the role that W plays in the calibration of the aggregate probability forecast.

In evaluating probability forecasts, researchers often examine the distribution of a forecast's probability integral transform (PIT). The PIT of a forecast F is $p = F(y)$, the cdf F evaluated at the realization y. Traditionally, the empirical distribution of PITs from a series of forecasts has been used to diagnose a forecaster's overconfidence/underconfidence. Specifically, a forecaster with a bathtub-shaped empirical PIT distribution has been described as overconfident (Hora 2004, Lichtendahl et al. 2013b, Jose et al. 2014).

When too many realizations fall in the tails of an expert's forecasts, her empirical PIT distribution will be bathtub shaped, or, equivalently, her hit rate for any prediction interval will be too low. In the probability judgment literature, hit rates of reported prediction intervals are typically too low for forecasts of temperature, gross domestic product growth, or stock returns (Lichtenstein et al. 1982, Goldstein and Rothschild 2014). Below we illustrate, however, that a low hit rate could have more to do with overfitting than with overconfidence.

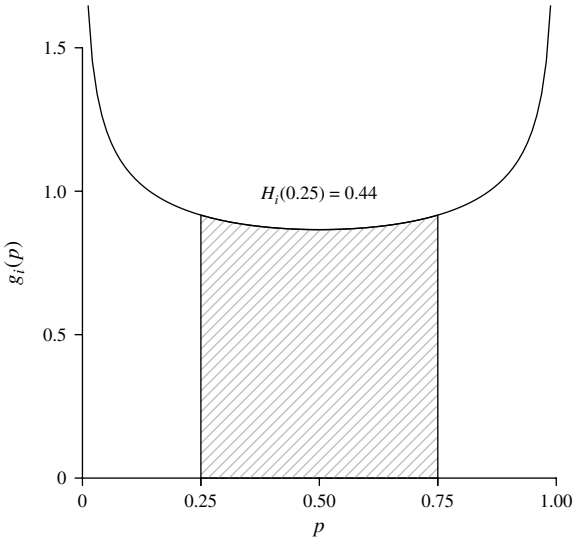
We define poor calibration without any mention of over- or underconfidence. For $0 < u < \frac{1}{2}$, the *hit rate* $H(u)$ of a forecast F's $1-2u$ prediction interval is the probability that the realized PIT $p = F(y)$ falls in the interval $[u, 1-u]$. A forecast is *well calibrated* if $H(u) = 1-2u$, *poorly calibrated with a low hit rate* if $H(u) < 1-2u$, and *poorly calibrated with a high hit rate* if $H(u) > 1-2u$. The result below indicates that a precise forecaster can have too low of a hit rate.

Proposition 2. *The hit rate of forecast* F_i *is* $H_i(u) = 1 - 2\Phi(C_i\Phi^{-1}(u))$, *where* $C_i = (E_i/\lambda_i')^{1/2}$. *This forecast's PIT density is given by* $g_i(p) = C_i\phi(C_i\Phi^{-1}(p))/(\phi(\Phi^{-1}(p)))$.

The constant C_i can be interpreted as the forecaster's *calibration coefficient*, where $C_i = 1$ if and only if the forecast is well calibrated. Clearly, the hit rate is increasing in C_i. By observation and Proposition 1, C_i (and hence the hit rate) is decreasing in λ_i' and $|w' - w|$. The case of a precise forecaster who either overfits or underfits (i.e., $\lambda_i' = \lambda_i$ and $w' \neq w$) is an interesting one. Such a forecaster will be poorly calibrated with a low hit rate. She may appear to be overconfident in the traditional sense, when in fact she is neither overconfident nor underconfident and is either overfitting or underfitting. The intuition behind this result is that overfitting introduces a bias in the location reported by a forecaster. This bias causes a shift in the distribution she reports. As a consequence, too many realizations will fall in the head or tail of her reported cdf. The following example illustrates this possibility.

Example 1. In the normal-normal information structure, let $m = n = \lambda = 1$. We also assume a forecaster reports using $w' = 1$ and $\lambda_i' = \lambda_i$. Figure 1 shows this forecaster's PIT density with a hit rate for the 50% prediction interval. The area of the shaded region under the PIT density is the hit rate $H_i(0.25) = 0.44$ for this interval, which in this case is too low. □

Figure 1. PIT Density with a Hit Rate (Shaded Region) for the Precise and Overfit Forecast from Example 1



The normative way to aggregate probability forecasts is the Bayesian opinion pool (Winkler 1981). It can, however, be difficult to apply. In our forecasting environment, forming the Bayesian opinion pool would require knowledge of the information structure, its four parameters, and the two reporting parameters. In its place, the linear opinion pool, the simple average of individual probability forecasts, has become a popular heuristic. Its form is simple and parameter free: $\bar{F}(z) = (1/k)\sum_{i=1}^{k} F_i(z)$. O'Hagan et al. (2006, p. 190) mentions that it is "hard to beat in practice."

Yet the linear opinion pool suffers from a calibration problem (Hora 2004, Jose et al. 2014). In general, the variance of the linear opinion pool is given by $(1/k)\sum_{i=1}^{k}(\mu_i' - \bar{\mu})^2 + 1/\lambda_i'$ (Lichtendahl et al. 2013b). In our forecasting environment, as crowd diversity increases, the expected variance of the linear opinion pool, which becomes $D + 1/\lambda_i'$, increases. Because overfitting contributes to greater diversity, the linear opinion pool of overfit forecasts runs the risk of being (on average) underconfident. Consequently, the trimmed opinion pool was introduced to counteract the crowd's calibration problem with the linear opinion pool (Jose et al. 2014).

The trimmed opinion pool is the trimmed average of the forecasters' reported cdf values at each support point z: $\hat{F}_\alpha(z) = (1/(k-2j))\sum_{i=j+1}^{k-j} F_{(i)}(z)$, where $j = \lfloor \alpha k \rfloor$ is the greatest integer less than or equal to αk, $0 \le \alpha < \frac{1}{2}$ is the level of trimming, and $F_{(1)}(z) \le \cdots \le F_{(k)}(z)$ are the order statistics of the forecasters' cdfs evaluated at z. We write $\hat{F}_0$ as alternative notation for the linear opinion pool (also known as the mean probability forecast). We let $\hat{F}_{1/2}(z)$ be the median of the forecasters' cdf values at each support point z. This trimmed opinion pool—the median probability forecast—is used by organizations in practice and is the subject of Hora et al. (2013).

To understand the behavior of the linear and trimmed opinion pools in a large crowd, we provide the limiting form of these pools in our forecasting environment. In our empirical studies in Section 3, we consider the random forest, a machine-learning algorithm that pools hundreds of forecasts. The result below makes use of the standard bivariate normal cdf with correlation coefficient ρ:

$$\Phi_B(\gamma, \delta; \rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \cdot \int_{-\infty}^{\delta}\int_{-\infty}^{\gamma} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right) dx_1 dx_2.$$

The standard bivariate normal cdf can be easily evaluated using statistical software, such as R's `mvtnorm` package.

Proposition 3. *In our forecasting environment with a normal-normal information structure, given μ, as the number of forecasters grows large, the trimmed opinion pool converges in probability to a distribution with the pdf*

$$\vec{f}_\alpha(z) = \frac{\bar{\lambda}^{1/2}\phi(\bar{\lambda}^{1/2}(z-\vec{\mu}))}{1-2\alpha} \cdot \left[\Phi\left(\frac{\Phi^{-1}(1-\alpha) - \rho\bar{\lambda}^{1/2}(z-\vec{\mu})}{(1-\rho^2)^{1/2}}\right) - \Phi\left(\frac{\Phi^{-1}(\alpha) - \rho\bar{\lambda}^{1/2}(z-\vec{\mu})}{(1-\rho^2)^{1/2}}\right)\right] \quad (4)$$

and the cdf

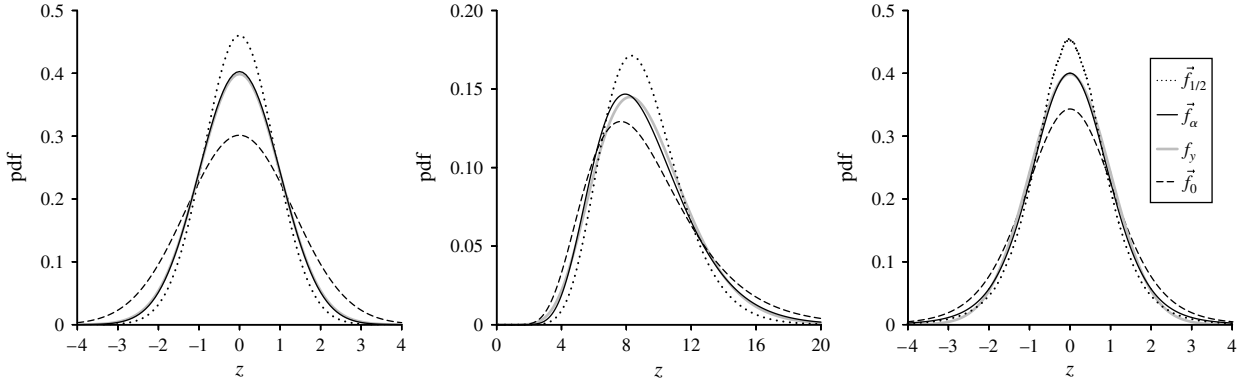
$$\vec{F}_\alpha(z) = \frac{1}{1-2\alpha}[\Phi_B(\Phi^{-1}(1-\alpha), \bar{\lambda}^{1/2}(z-\vec{\mu}); \rho) - \Phi_B(\Phi^{-1}(\alpha), \bar{\lambda}^{1/2}(z-\vec{\mu}); \rho)], \quad (5)$$

where $\vec{\mu} = (1-w')\mu_0 + w'\mu$, $\bar{\lambda} = n\lambda\lambda_i'/(w'^2\lambda_i' + n\lambda)$, *and* $\rho = (\lambda_i' - \bar{\lambda})^{1/2}/\lambda_i'^{1/2}$. *At the two extremes,* $\vec{F}_0(z) = \Phi(\bar{\lambda}^{1/2}(z-\vec{\mu}))$ *and* $\vec{F}_{1/2}(z) = \Phi(\lambda_i'^{1/2}(z-\vec{\mu}))$.

Notice that in this result we condition on μ. This conditioning enables us to express the limiting trimmed opinion pool in closed form by taking advantage of the conditional independence in our information structure. Later, when analyzing hit rates, we marginalize over the distribution of μ, as was done for Proposition 2.

The limiting trimmed opinion pool in Proposition 3 is a type of hidden truncation normal model (Arnold et al. 1993, Arnold 2000). This pool has the same distribution as a normal random variable plus a truncated normal random variable (see the proof of Proposition 5). Although the limiting trimmed opinion pool with $0 < \alpha < \frac{1}{2}$ cannot exactly match the true distribution of y given μ, it has the flexibility, via the tuning parameter α, to closely approximate the truth. The

Figure 2. Probability Density Functions of Three Large-Crowd Trimmed Opinion Pools and the True Distribution From the Three Information Structures in Examples 2–4



(a) Normal-normal ($\alpha = 0.2$) (b) Gamma-gamma ($\alpha = 0.1$) (c) Normal-gamma-normal ($\alpha = 0.15$)

following example illustrates how closely the limiting trimmed opinion pool can match the true distribution (denoted f_y), when forecasts are overfit and overconfident.

Example 2. In the normal-normal information structure, again let $m = n = \lambda = 1$, as in Example 1. Assume that all forecasters report using $w' = 1$ and $\lambda_i' = 2\lambda_i$ and that $\mu = 0$. Figure 2, panel (a) displays the pdfs of three limiting trimmed opinion pools and the true distribution. Notice that the pdf of the limiting linear opinion pool (LLOP), $\vec{f}_0$ (dashed line), is more disperse than that of either limiting trimmed opinion pool. Furthermore, trimming at the 20% level (in black) closely matches the pdf of the true underlying distribution (in gray). □

Before we provide more analytical results for the normal-normal information structure, we investigate the extent to which the results above generalize to other information structures. Because it is difficult to find closed-form expressions for other structures, we numerically simulate a large number ($k = 10{,}000$) of overfit and overconfident forecasts from two other information structures.

Example 3. In the gamma-gamma information structure, let $a = \mu_0 = 10$ and $m = n = 1$. After forecaster i observes her sample, she should report that $(y \mid \mathbf{y}_i) \sim \mathrm{Gg}((m+n)a + 1, m\mu_0 + n\bar{y}_i, a)$, where $\mathrm{Gg}(\alpha, \beta, q)$ is the gamma-gamma distribution, which has the pdf $\mathrm{Gg}(z \mid \alpha, \beta, q) = (\beta^{\alpha}/\Gamma(\alpha))(\Gamma(\alpha+q)/\Gamma(q))(z^{q-1}/(\beta+z)^{\alpha+q})$. Assume that all forecasters report using $n' = 5$ and $a' = 16$ in place of n and a, respectively. The reported mean of $(y \mid \mathbf{y}_i)$ is $\mu_i' = (1 - w')\mu_0 + w'\bar{y}_i$, where $w' = n'/(m+n') > w$, and the reported precision is $\lambda_i' = [\mu_i']^{-2}((m+n')a' - 2)/(m+n'+1)$, which is increasing in a' and on average is 2.7 times greater than λ_i. In addition, let $\mu = (ma+1)/(m\mu_0)$. □

Example 4. In the normal-gamma-normal information structure, let $b = 2$, $m = 1$, and $n = 2$, so that $a = 2$. After forecaster i observes her sample, she should report that $(y \mid \mathbf{y}_i) \sim \mathrm{St}(\mu_i, \nu_i, 2a + n)$, where $\mu_i = (1 - w)\mu_0 + w\bar{y}_i$, $\nu_i = ((m+n)/(m+n+1))(a + n/2)/\beta_i$, $\beta_i = b + \frac{1}{2}ns_i^2 + \frac{1}{2}(mn/(m+n))(\mu_0 - \bar{y}_i)^2$, $ns_i = \sum_{l=1}^{n}(y_{(i-1)n+l} - \bar{y}_i)^2$, and $St(\mu_i, \nu_i, 2a+n)$ is the Student's t distribution, which has pdf $\mathrm{St}(z \mid \mu, \nu, \alpha) = (\Gamma(\frac{1}{2}(\alpha+1))/\Gamma(\frac{1}{2}\alpha))(\nu/(\alpha\pi))^{1/2} \cdot [1 + \alpha^{-1}\nu(z-\mu)^2]^{-(\alpha+1)/2}$. Assume that all forecasters report that $(y \mid \mathbf{y}_i) \sim \mathrm{St}(\mu_i', \frac{3}{2}\nu_i, 2a + n)$, where $\mu_i' = (1 - w')\mu_0 + w'\bar{y}_i$ and $w' = 1$. The reported mean is μ_i', and the reported precision is $\lambda_i' = ((2a + n - 2)/(2a + n))\frac{3}{2}\nu_i$, which is 50% larger than what it should be. □

Similar to panel (a) of Figure 2, panels (b) and (c) depict the effects of trimming at 10% and 15% when samples are drawn from the gamma-gamma and normal-gamma-normal information structures, respectively. These numerical results suggest that the effects of trimming generalize to other information structures.

The following two propositions provide additional properties of the limiting trimmed opinion pool from Proposition 3. These propositions provide further support for the limiting trimmed opinion pool's ability to match the true distribution of y.

Proposition 4. *The mean of the limiting trimmed opinion pool is $\vec{\mu}$, and its precision is $\vec{\lambda} = \bar{\lambda}(1 + 2\rho^2\Phi^{-1}(\alpha) \cdot \phi(\Phi^{-1}(\alpha))/(1-2\alpha))^{-1}$. This precision is increasing in λ_i' and decreasing in w'.*

According to Propositions 3 and 4, the limiting trimmed opinion pool has the same mean as the linear opinion pool. This fact implies that overfitting improves the accuracy of the limiting trimmed opinion pool's point forecast, in a similar fashion to the improvement reported in Proposition 1. Importantly, when $w' = 1$, the means of the limiting linear

and trimmed opinion pools match that of the true distribution.

It is also evident from this result that one possible source of the underconfidence in the linear opinion pool is overfit forecasts, as the precision of any limiting (trimmed or untrimmed) opinion pool decreases in w'. Next, we consider how the trimming level α can be used to address this problem. Proposition 5 below shows that the limiting trimmed opinion pool offers a parameter α that moderates the pool's variance, as well as its higher even central moments.

Let X and Y be random variables with distribution functions F and G, respectively. By Müller and Stoyan (2002) and Shaked and Shanthikumar (2007), F is said to be *smaller than G in dispersive order*, denoted as $X \leq_{\text{disp}} Y$, if $F^{-1}(t) - F^{-1}(s) \leq G^{-1}(t) - G^{-1}(s)$ for all $0 < s < t < 1$.

Proposition 5. *For $0 \leq \alpha_1 \leq \alpha_2 < \frac{1}{2}$, let $Y_1 \sim F_{\alpha_1}$ and $Y_2 \sim F_{\alpha_2}$. Then $Y_2 \leq_{\text{disp}} Y_1$, and Y_2's even central moments are less than or equal to those of Y_1.*

The interpretation of this result is that increasing the trimming level always decreases dispersion or increases precision. We conclude our theoretical investigation with the following result that shows that trimming can be used to form a better-calibrated ensemble.

Proposition 6. *For $0 \leq \alpha_1 \leq \alpha_2 \leq \frac{1}{2}$, the hit rates of forecasts F_i, $\vec{F}_{1/2}$, $\vec{F}_{\alpha_1}$, $\vec{F}_{\alpha_2}$, and $\vec{F}_0$ are ordered $H_i(u) \leq H_{1/2}(u) \leq H_{\alpha_2}(u) \leq H_{\alpha_1}(u) \leq H_0(u)$ for $0 < u < \frac{1}{2}$, respectively. The forms of these hit rates are given in Table 1, where $\Psi_\alpha(\tau) = 1/(1-2\alpha)[\Phi_B(\Phi^{-1}(1-\alpha), \tau; \rho) - \Phi_B(\Phi^{-1}(\alpha), \tau; \rho)]$, and ρ is given in Proposition 3.*

In Table 1, we list the calibration coefficients in short form to suggest that one need not know the parameters of the model to estimate hit rates and PIT densities. With estimates of the quantities C_i and W from data, one can quickly estimate the hit rates and PIT densities of the mean and median probability forecasts. In situations where forecasters completely overfit and are moderately overconfident, the linear opinion pool's hit rate will be too high and the limiting median-trimmed opinion pool's (and the individuals') hit rate will be too low. In such cases, Proposition 6 indicates that increasing the trimming level will reduce a prediction interval's hit rate until it is just right, as the following example illustrates.

Example 5. In the normal-normal information structure, let $m = n = \lambda = 1$, $w' = 1$, and $\lambda_i' = 2\lambda_i$ (which results in a standard deviation that is approximately 70% of what it should be). Hence, forecasters overfit and are overconfident. Figure 3 shows PIT densities with hit rates for the 50% prediction interval. The hit rates $H_0(0.25)$ and $H_{1/2}(0.25)$ bracket 0.5. Trimming at the 0.2 level produces a hit rate for the 50% prediction interval that is almost ideal. This example suggests that in practice, with large, overfitting, and overconfident crowds, a trimmed opinion pool may yield a better-calibrated ensemble. □

Next, we consider the broader range of settings for w' and λ_i' when $m = n = \lambda = 1$ in the normal-normal information structure. The shaded region in Figure 4 depicts the pairs of (w', λ_i') for which the hit rates of the limiting linear and median-trimmed opinion pools bracket the correct hit rate. In particular, the limiting linear and median-trimmed opinion pools' hit rates bracket the correct hit rate (i.e., $H_{1/2}(u) < 1-2u < H_0(u)$) when $C_{1/2} < 1 < C_0$, or, equivalently, when $(1 - C_i^2)/2 < W < 1 - C_i^2$. In this region, because the trimmed opinion pool's hit rate is decreasing (and continuous) in the trimming level, the trimmed opinion pool with some level of trimming can achieve the correct hit rate. The majority of this region is in the upper right quadrant (to the right of and above the dashed lines) where overfitting and overconfidence both occur. Thus, we expect that trimming will be effective more often in settings when both overfitting and overconfidence are present.

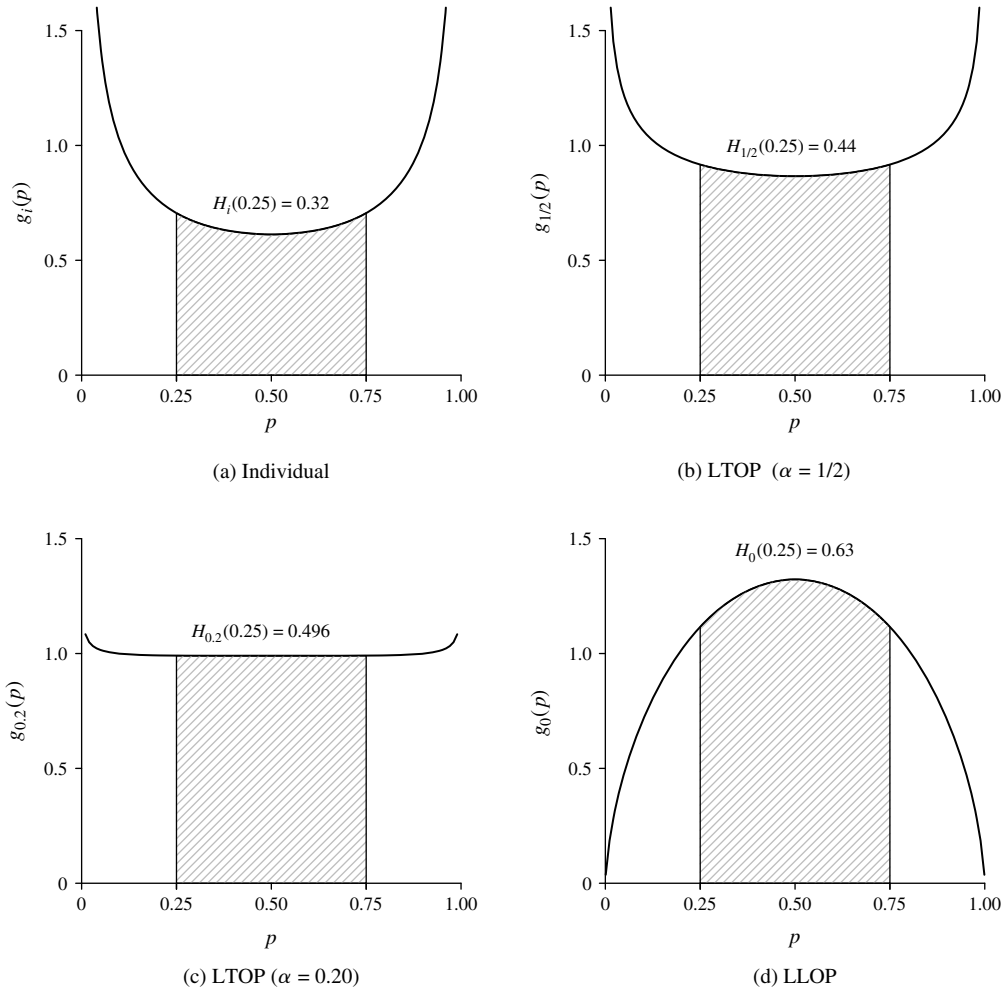
2.3. Discrete Events Following the Beta-Bernoulli Information Structure

In many instances, we are faced with classification problems in which the outcome y is no longer a con-

Table 1. Theoretical Hit Rates and Calibration Coefficients of Proposition 6

Forecast	Hit rate	Calibration coefficient	
		Long form	Short form
$\vec{F}_i$	$1 - 2\Phi(C_i\Phi^{-1}(u))$	$C_i = \left(1 + \frac{w'^2}{n} + \frac{(1-w')^2}{m}\right)^{-1/2}\left(\frac{\lambda}{\lambda_i'}\right)^{1/2}$	$C_i = \left(\frac{E_i}{\lambda_i'}\right)^{1/2}$
$\vec{F}_{1/2}$	$1 - 2\Phi(C_{1/2}\Phi^{-1}(u))$	$C_{1/2} = \left(1 + \frac{(1-w')^2}{m}\right)^{-1/2}\left(\frac{\lambda}{\lambda_i'}\right)^{1/2}$	$C_{1/2} = \left(\frac{C_i^2}{1-W}\right)^{1/2}$
$\vec{F}_\alpha$	$1 - 2\Phi(C_0\Psi_\alpha^{-1}(u))$	$C_0 = \left(1 + \frac{(1-w')^2}{m}\right)^{-1/2}\left(\frac{w'^2}{n} + \frac{\lambda}{\lambda_i'}\right)^{1/2}$	$C_0 = \left(\frac{W + C_i^2}{1-W}\right)^{1/2}$
$\vec{F}_0$	$1 - 2\Phi(C_0\Phi^{-1}(u))$		

Figure 3. PIT Densities with Hit Rates (Shaded Region) for an Individual, the LTOP with $\alpha = 0.2, \frac{1}{2}$, and the LLOP from Example 5



tinuous uncertain quantity, but a discrete event. To analyze the impact of overfitting and overconfidence when forecasting discrete events, we consider the beta-Bernoulli information structure in Section 2.1, where y takes a value on $\{0, 1\}$.

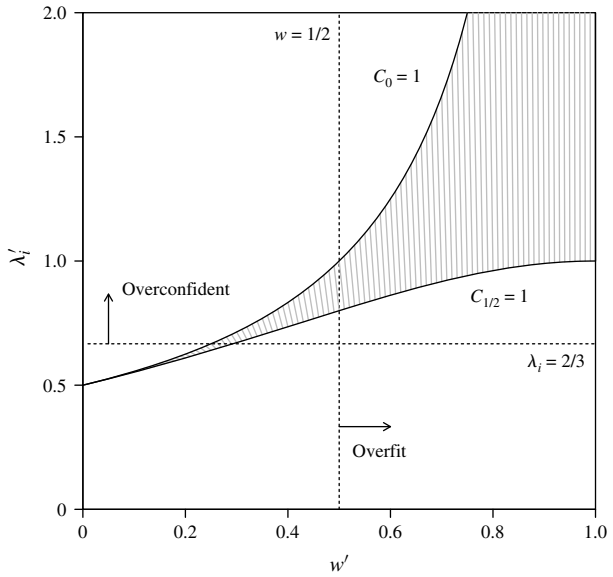
In this information structure, after observing her sample, forecaster i should report that $(y \mid \mathbf{y}_i) \sim \text{Br}(\mu_i)$, where $\mu_i = (1 - w)\mu_0 + w\bar{y}_i$ (Bernardo and Smith 2000, p. 436). Suppose instead the forecaster reports that $(y \mid \mathbf{y}_i) \sim \text{Br}(\mu_i')$, where $\mu_i' = (1 - w')\mu_0 + w'\bar{y}_i$. Her reported variance is then determined to be $\mu_i'(1 - \mu_i')$. Therefore, an overfit forecast may or may not be overconfident. In expectation, however, it is not difficult to show that an overfit forecast is also overconfident; i.e., $E[\mu_i'(1 - \mu_i')] = \mu_0(1 - \mu_0)(1 - (m + n)/((m + 1)n)w'^2) < E[\mu_i(1 - \mu_i)]$ for $w' > w$. The result below provides the probability of $y = 1$ according to the limiting linear opinion pool for this classification problem.

Proposition 7. *In our forecasting environment with a beta-Bernoulli information structure, given μ, as the number of forecasters grows large, the linear opinion pool's $1 - \hat{F}_0(0)$ converges in probability to $(1 - w')\mu_0 + w'\mu$.*

Thus, given μ, the simple average of a crowd of completely overfit forecasts converges to the truth. This result can be generalized to the case of a discrete event with more than two outcomes. The information structure with a categorical likelihood (which is an exponential family member) and its conjugate Dirichlet prior has a similar result: the linear opinion pool converges to the true distribution. Consequently, there is no need for trimming when forecasters' samples are drawn from a Dirichlet-categorical information structure. Nonetheless, we find that trimming can improve the average forecast's predictive accuracy in some empirical settings. In the next section, we provide an explanation for how this effect may arise.

Next, we study the wisdom of the crowd of forecasters predicting a binary outcome by considering W, as defined in Section 2.1. In this case, the MSE of the aggregate forecast is equivalent to the average Brier score of that forecast.

Figure 4. Pairs of (w', λ_i') for Which the Hit Rates of the Limiting Linear and Median-Trimmed Opinion Pools Bracket the Correct Hit Rate (Shaded Region) in the Normal-Normal Information Structure When $m = n = \lambda = 1$



Proposition 8. *In our forecasting environment with a beta-Bernoulli information structure, the wisdom of the crowd is* $W = E_i D$, *and the following statements hold*:

(a) *Individual expertise* $E_i = [(1 + w'^2/n + (1 - w')^2/m) \cdot (m/(m+1))\mu_0(1-\mu_0)]^{-1}$ *is decreasing in* $|w' - w|$.

(b) *Diversity* $D = ((k-1)w'^2/(kn))(m/(m+1)) \cdot \mu_0(1-\mu_0)$ *is increasing in* $w' > 0$.

(c) *The wisdom of the crowd* W *is increasing in* w' *for* $0 < w' \leq 1$.

This result is structurally similar to Proposition 1. Again, although overfitting degrades individual expertise, overall it improves the wisdom of the crowd.

3. Empirical Forecasting Environment

The limiting results in the previous section are fundamental to understanding how the linear and trimmed opinion pools might work with large crowds of forecasters or models. In this section, we apply the trimmed opinion pool to actual large crowds of models from a random forest, using several publicly available and commonly used data sets. As an important parallel to the theoretical hit rates in Figure 3, we present empirical hit rates for the linear and trimmed opinion pools of the trees in a random forest. These parallel results allow us to generalize our conclusions concerning the flexibility of the trimmed opinion pool to generate better-calibrated ensembles.

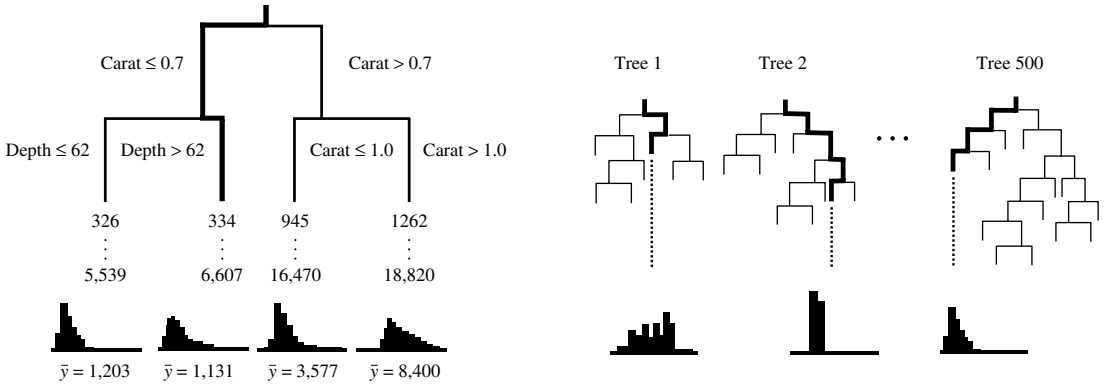
A random forest is an empirical model of multiple forecasts. In particular, it is a collection of regression or classification trees (Breiman 2001) grown up on training data. A random forest is used to predict a response variable y given a set of known predictors $\mathbf{x} = (x_1, \ldots, x_q)$. Regression trees apply to situations where the response variable is a continuous quantity, whereas classification trees apply to discrete response variables. Throughout, we focus on the regression case, which aligns closely with our model above. Each tree in a random forest is fit to, or grown up on, a random subspace of the training data. Next, we briefly describe how regression trees in a random forest are grown up and combined. For more details on regression trees and random forests, see Hastie et al. (2009, pp. 307–308, 588) and Varian (2014).

To grow up each tree on the training data, the random forest algorithm first randomly draws with replacement `sampsize` rows. Using a tree's sampled rows, the algorithm proceeds to recursively partition the predictor space. At each parent node in a tree, `mtry` $< q$ candidate predictors are randomly selected. For each candidate predictor, all possible split points are evaluated. A possible split puts the response values in the parent node into two daughter nodes. For candidate-predictor values less than or equal to the split point, the associated response values go into the left daughter node, and their average is calculated. For candidate-predictor values greater than the split point, the associated response values go into the right daughter node, and their average is calculated. The best predictor/split-point pair minimizes the sum of squares in the left and right daughter nodes. A daughter node's sum of squares is the sum of squared errors in the response values from their average response value. The binary-splitting process attempts to split a node only if the number of responses in that node is above `nodesize`.

To forecast a new y, the associated values of the predictors $(x_1, \ldots, x_q)$ are sent down each tree in the forest until a terminal node is reached. In each tree's relevant terminal node, there will be a set of response values from the training set. A tree's point forecast is the average of these response values. Figure 5, panel (a) presents a simple regression tree grown up on the diamonds data set that we study in more detail below. In this data set, the diamond's price is the response, and its carat weight and depth percentage (depth divided by the average of width and length) are two of the predictors. For a new diamond that weighs 0.4 carat and has 70% depth, the point prediction from this regression tree is \$1,131.

The random forest's point forecast is the ensemble, or average, of `ntree` trees' point forecasts. Because the sampled observations and candidate predictors at each split are drawn independently from tree to tree, the response values in the relevant terminal node (and their averages) will vary. Importantly, the trees' point forecasts will typically be diverse enough for the random forest's point forecast to benefit from the

Figure 5. Regression Tree and Random Forest Diagrams



(a) Regression tree (b) Random forest

wisdom-of-crowds effect. To form the random forest's probability forecast, Meinshausen (2006) proposes taking the empirical distribution of the response values in each tree's relevant terminal node and then averaging these distributions. Thus, Meinshausen's proposal is a linear opinion pool of the trees in the random forest. When pooling forecasts, we also assume that a tree's forecast is the empirical cdf of its relevant terminal node's response values. For a schematic diagram of a random forest, see Figure 5, panel (b). For a new diamond, the path down a tree marked in bold leads to the relevant terminal node and an empirical distribution of prices used to forecast the new diamond's price.

Without tuning its parameters, the popular random forest implementation `randomForest` in `R` has been shown to be an accurate point forecasting tool (Liaw and Wiener 2002). With the default settings, `ntree` = 500 and `nodesize` = 5, a large number of large trees are grown up. Each tree will have many terminal nodes with fewer than five points in them. Such large regression trees are known to overfit the data (Hastie et al. 2009, pp. 307–308). Consequently, the random forest—an ensemble of a large number of overfit forecasts—is a natural forecasting environment in which to study the performance of the trimmed opinion pool. Before we provide our main empirical results, we draw connections between trees in a random forest and the overfit and overconfident individual forecasts from the previous section's forecasting environment.

3.1. Overfitting and Overconfidence of Trees in a Random Forest

To demonstrate how regression trees in a random forest overfit, we simulated the bias-variance decomposition of the mean squared error for a randomly selected tree in a random forest. In the statistics and machine-learning literature, the bias-variance trade-off is a popular way to illustrate overfitting. An overfit model's point forecast has a mean squared error that is low in bias and high in variance. For this simulation, we used a popular synthetic data set called Friedman #1 (Friedman 1991), which is available in `R`'s `mlbench` package.

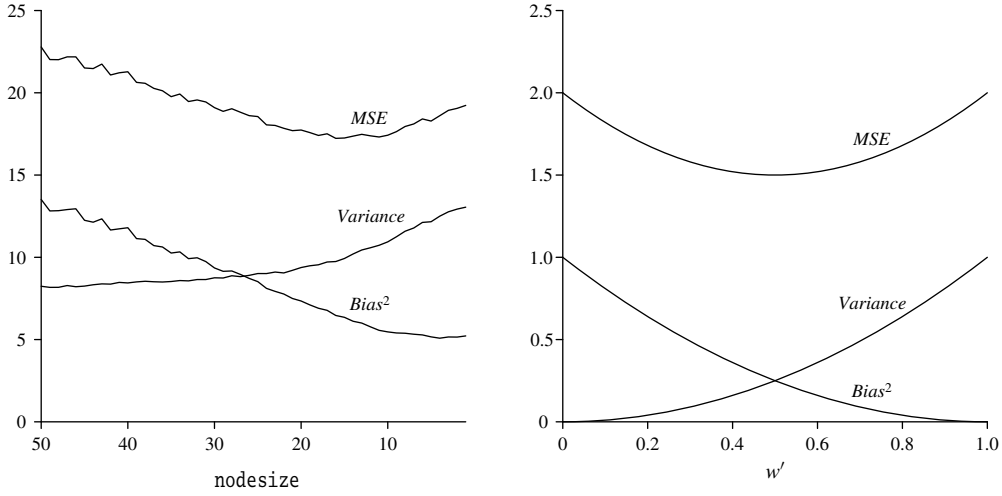
The Friedman #1 data set has 10 predictors, each drawn independently and uniformly on $(0,1)$. We drew 501 rows of predictors, with the goal of using 500 rows to train the random forest's tree and predict the 501st response. The response variable y is a function of only the first five predictor variables: $y = f(x_1,x_2,x_3,x_4,x_5) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon$, where $\epsilon \sim \mathrm{N}(0,1)$. The other five predictors are "noise" variables and have no impact on the response—a common feature in many real-world settings. To estimate the tree's mean squared error and its bias-variance decomposition, we ran a simulation with 100,000 trials. At each trial, a new set of 500 responses was drawn, and one tree from a random forest (at its default settings) was grown up on that trial's training data.

To interpret the results of this simulation, we compare its bias-variance decomposition to our theoretical model's bias-variance decomposition. As we will see, these decompositions turn out to be strikingly similar. By Hastie et al. (2009, Equation 7.9), the conditional mean square error of μ_i', which is an estimate of the true mean μ in our theoretical model, can be decomposed into three terms: the irreducible error $E[(y-\mu)^2 \mid \mu]$, the bias $E[\mu_i' \mid \mu] - \mu$ and the variance $\mathrm{Var}[\mu_i' \mid \mu]$. This decomposition reduces to the following expression:

$$\begin{aligned} E[(y-\mu_i')^2 \mid \mu] &= \textit{Irreducible Error} + \textit{Bias}^2 + \textit{Variance} \\ &= \frac{1}{\lambda} + (1-w')^2(\mu-\mu_0)^2 + \frac{w'^2}{n\lambda}. \end{aligned}$$

We can see that when $w' = 1$, the bias is zero. The trade-off, however, is that the variance is high in this case. Note that the counterpart of μ in our simulation is the function $f(x_1,x_2,x_3,x_4,x_5)$, and the counterpart of μ_i' is the tree's prediction for the 501st response.

Figure 6. Bias-Variance Decompositions from Example 6 for (a) a Randomly Selected Tree in a Random Forest and (b) an Individual Forecaster in the Theoretical Forecasting Environment in Section 2



(a) Randomly selected tree (b) Individual forecaster

Example 6. Suppose an individual forecaster in the theoretical forecasting environment in Section 2 will sample data from an normal-normal information structure with $\mu_0 = 0$ and $\mu = m = n = \lambda = 1$. Figure 6 contains plots of the bias-variance decomposition for a randomly selected tree in a random forest grown up on the Friedman #1 (see Figure 6, panel (a)) and for an individual forecaster in Section 2's theoretical forecasting environment (see Figure 6, panel (b)). Overfitting corresponds to the right-hand sides of these plots: low bias and high variance. In Figure 6, panel (a), `nodesize` varies from 50 down to 1. For Figure 6, panel (b), we change w' from 0 up to 1. In this example, the randomly selected tree's and the individual forecaster's predictions are similarly overfit. □

To demonstrate that the probability forecasts from a tree in a random forest are also overconfident, we use the same simulation, but across 10,000 trees. Because we set $\epsilon \sim \mathrm{N}(0,1)$ for our Friedman #1 data set, we know that the theoretical variance of the response is 1. We can compare this theoretical variance to the variance calculated from a tree's empirical cdf. The average trees' variance turns out to be 0.781, which is too low. This result suggests that the trees in a random forest exhibit some degree of overconfidence, and therefore we might expect the hit rates of the linear opinion pool and the median-trimmed opinion pool to bracket the ideal.

We should note that the variance calculated from a tree's empirical cdf is too low, in part because it is less than the unbiased sample variance, which would result from dividing the sum of squares by $n-1$ instead of n. Each tree's variance is also too low because noise variables are included in the random forest. When a tree splits on a noise variable (a variable that is unrelated to the response), the variance in either child node is less than that in the parent node (according to the usual sum-of-squares decomposition). Such splits, by themselves, cause a tree's estimated variance to be biased downward. These two biases, which tend to underestimate the true variance, are counteracted somewhat by variations in the true model (e.g., $f(x_1, x_2, x_3, x_4, x_5)$ over a terminal node's domain (i.e., the set of $\mathbf{x}$'s that lead to that node). With few response values typically left in each of a tree's terminal nodes, however, this counteracting effect will generally be small in a random forest, resulting in overconfident trees.

3.2. Trimming the Trees in a Random Forest: Continuous Response

In addition to the Friedman #1 data, we examine the performance of the trimmed opinion pool on three other publicly available data sets. We refer to these other data sets as the diamonds, Boston housing, and bike-sharing data sets. They are available through the `R` packages `ggplot2` and `mlbench` and from the UCI data repository, respectively. The diamonds data set has nine predictors with price as the response. For our study, we randomly chose 10,000 observations from this data set. The Boston housing data set has 13 predictors, median house prices as the response, and 506 observations. Finally, the bike-sharing data set has 11 predictors, daily bike-sharing rentals as the response, and 731 observations.

To generate all of our empirical results, we used 25% of each data set as a testing set and applied a standard fivefold cross-validation procedure on the remaining 75% of the data (training set; Hastie et al. 2009, pp. 241–249). The point and probability forecasts reported on below were generated using the `R` package `trimTrees`, an efficient implementation of the trimmed

Table 2. Average MSE and W (in Percentages) of Trimmed Opinion Pools ($\hat{F}_\alpha$) from a Random Forest on Continuous Response Data Sets for the Validation and Testing Sets

Set(s)	Pool	Friedman #1		Diamonds		Boston housing		Bike sharing	
Validation	α	MSE	W	MSE	W	MSE	W	MSE	W
	0.00	6.08	69.23	436,214	62.94	12.87	64.03	487,441	62.25
	0.05	5.92	70.01	428,879	63.58	12.54	64.78	473,226	63.37
	0.10	5.87	70.28	426,682	63.78	12.37	65.21	**470,350**	**63.61**
	0.15	5.84	70.45	**426,358**	**63.81**	**12.31**	**65.34**	470,479	63.61
	0.20	5.81	70.57	427,260	63.73	12.33	65.30	471,974	63.51
	0.25	5.79	70.68	428,944	63.59	12.48	64.95	473,969	63.37
	0.30	**5.78**	**70.71**	430,692	63.44	12.82	64.17	475,675	63.25
	0.35	5.79	70.69	432,201	63.31	13.31	63.04	476,582	63.19
	0.40	5.80	70.61	433,462	63.19	13.78	61.98	477,196	63.14
	0.45	5.82	70.52	434,511	63.10	14.00	61.51	478,281	63.06
	0.50	5.84	70.45	435,416	63.02	14.07	61.44	479,782	62.96
Testing	α	MSE	W	MSE	W	MSE	W	MSE	W
	0.00	5.03	71.65	456,054	59.88	7.25	72.73	444,720	61.36
	α^*	4.92	72.27	454,318	60.03	6.62	75.13	442,370	61.56

Note. Numbers in bold represent best values in the validation set.

opinion pool applied to the trees in a random forest. This package was created by the authors and is available at the `CRAN` repository (with updates available from `GitHub`). Unless stated otherwise, all results below were generated using the `randomForest`'s settings of `nodesize` = 5 and `ntree` = 500 (with all other settings at their default).

Table 2 presents the average MSE and W for the linear opinion pool ($\alpha = 0$) and several trimmed opinion pools. At the top of the table are the average MSE and W, averaged over the five validation sets from the fivefold cross-validation procedure. At the bottom of the table are the average MSE and W on the testing set for the linear opinion pool and the optimally trimmed pool (α^*), as identified in cross-validation. Recall that W is the percentage improvement in MSE of an opinion pool over a randomly selected tree. Note that the point predictions evaluated in Table 2 are the means calculated from the pools' entire distributions.

From Table 2, we see that for Friedman #1 data set, MSE is minimized at 30% trimming. For the diamonds and Boston housing data sets, it is minimized at 15% trimming, and for the bike sharing, at 10% trimming. Not surprisingly, the trimming levels that minimize the MSE also maximize W. Here, W is an attractive measure because it provides a scale-free comparison of performance across domains. On the three real data sets, the crowd of trees is approximately 60%–65% wiser than a randomly selected tree. In all data sets, the MSE on the testing set, when trimmed at an α^* level, is lower than with no trimming at all.

The MSE is useful when the focus is on the accuracy of point predictions. When the decision context requires a distribution, however, we will want to measure the calibration and accuracy of the probability forecast. To measure calibration, we examine the hit rates for the 50% prediction intervals. To measure accuracy, decision analysts and statisticians often prescribe the use of proper scoring rules. These rules reward both calibration and sharpness. One popular proper scoring rule is the linear quantile scoring (LQS) rule (Jose and Winkler 2009). Meinshausen (2006) uses this rule to compare the linear opinion pool from a random forest to other quantile regression approaches.

Table 3 presents average hit rates for the 50% prediction intervals and linear quantile scores for the linear opinion pool ($\alpha = 0$) and several trimmed opinion pools. Hit rates closer to 50% and lower scores are better. The linear quantile score is given by $\mathrm{LQS}(\hat{Q}_\alpha(u_1), \ldots, \hat{Q}_\alpha(u_N), y) = \sum_{j=1}^{N} \mathrm{LQS}_j(\hat{Q}_\alpha(u_j), y)$, where the component score for the u_j-quantile, $\hat{Q}_\alpha(u_j) = \min\{z: u_j \le \hat{F}_\alpha(z)\}$, assigns the score

$$\mathrm{LQS}_j(\hat{Q}_\alpha(u_j), y) = \begin{cases} u_j(y - \hat{Q}_\alpha(u_j)) & \text{for } \hat{Q}_\alpha(u_j) \le y, \\ (1-u_j)(\hat{Q}_\alpha(u_j) - y) & \text{for } \hat{Q}_\alpha(u_j) > y. \end{cases}$$

We use $N = 19$ and $u_j = j/20$ in Table 3. The trimming levels that minimize the linear quantile score are 0.15 for the Friedman #1 data set and 0.1 across the other three data sets.

The hit rates of the linear opinion pool are too high across all data sets and can be improved with some level of trimming. Because scoring rules reward calibration and trimmed opinion pools achieve better calibration through increased sharpness (or less dispersion), it is not surprising that trimming levels that minimize the score are also those that achieve nearly

Table 3. Average Hit Hates (in Percentages) and Linear Quantile Scores for Trimmed Opinion Pools ($\hat{F}_\alpha$) from a Random Forest on Continuous Response Data Sets for the Validation and Testing Sets

Set(s)	Pool	Friedman #1		Diamonds		Boston housing		Bike sharing	
Validation	α	$H_\alpha(0.25)$	LQS	$H_\alpha(0.25)$	LQS	$H_\alpha(0.25)$	LQS	$H_\alpha(0.25)$	LQS
	0.00	66.9	14.63	60.9	2,455	57.2	16.6	59.1	3,579
	0.05	60.3	14.21	55.1	2,411	54.9	16.3	55.3	3,518
	0.10	53.3	13.98	**49.7**	**2,402**	**48.3**	**16.2**	**48.9**	**3,503**
	0.15	**49.1**	**13.91**	44.0	2,417	43.2	16.3	43.4	3,526
	0.20	43.2	14.01	37.7	2,449	36.4	16.6	38.9	3,580
	0.25	36.5	14.24	31.5	2,496	31.4	17.0	33.9	3,638
	0.30	28.3	14.65	26.7	2,553	26.4	17.5	28.5	3,719
	0.35	24.0	15.13	22.6	2,614	20.0	18.2	24.3	3,795
	0.40	18.9	15.69	20.0	2,670	16.9	18.9	23.4	3,863
	0.45	15.5	16.18	20.2	2,711	15.8	19.3	23.2	3,913
	0.50	17.3	16.36	22.5	2,728	18.2	19.5	25.4	3,941
Testing	α	$H_\alpha(0.25)$	LQS	$H_\alpha(0.25)$	LQS	$H_\alpha(0.25)$	LQS	$H_\alpha(0.25)$	LQS
	0.00	70.4	13.50	61.0	2,407	56.7	14.7	62.8	3,366
	α^*	53.6	12.64	48.6	2,364	50.4	14.2	49.2	3,304

Note. Numbers in bold represent best values in the validation set.

ideal hit rates of 50%. Note that optimal trimming levels, when considering either MSE or LQS, are not sensitive to changes in the `randomForest`'s settings. When we change `nodesize` from 5 to 10 and `ntree` from 500 to 100, our trimming recommendations do not change by more than 5%. See the online supplement for detailed results on these robustness checks.

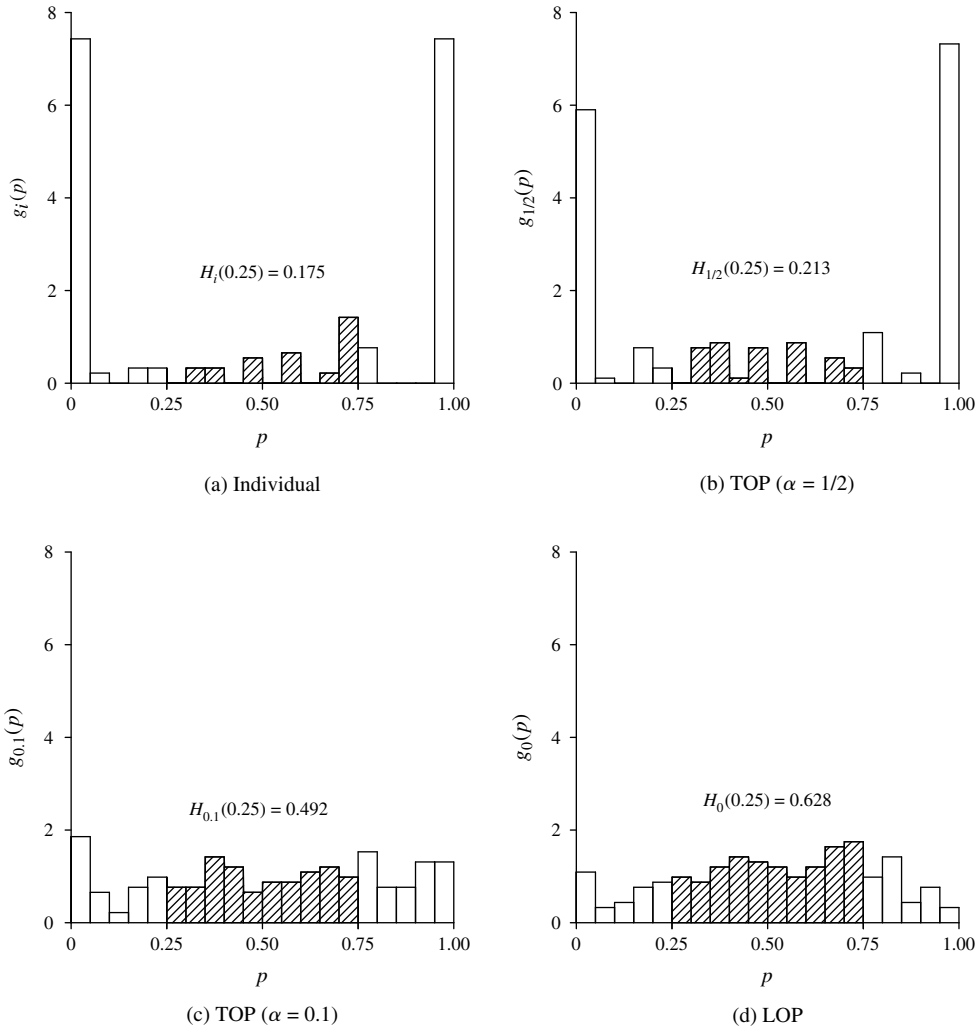
For the bike-sharing data set, Figure 7 provides more detail on empirical hit rates and PIT densities from the testing set. From Figure 7, panel (a), we see that the typical tree is poorly calibrated with an extremely low hit rate for the 50% prediction intervals and other prediction intervals. Trimming with $\alpha = 0.1$ results in the most uniform PIT density and calibrated ensemble. As in Figure 3, the ordering of the hit rates in Figure 7 provides further evidence of the benefits from trimming forecasts and suggests that the trees in a random forest are both overfit and overconfident forecasts.

From Tables 2 and 3, we see that trimming improves the quality of both average point and probability forecasts, albeit in different ways. Specifically, optimal trimming levels are higher for minimizing MSE than for minimizing the linear quantile scores. One explanation for this result may be outliers and the fact that the MSE is particularly sensitive to them. Although outliers in an ensemble of overfit trees would increase diversity, they could negatively impact the accuracy of the average point forecast. The fact that optimal trimming levels for minimizing MSE are high suggests there are outlying trees in a random forest. The optimal trimming levels for the linear quantile score are lower as the absolute difference used in this score is less sensitive to these outliers.

Figure 8 provides insight into the diversity of opinions in a random forest trained on the bike-sharing data set. It also provides a visual check for outliers. In the plot, we see 90% prediction intervals for a single day's bike rentals from 500 trees and the trimmed and untrimmed ensembles. The diversity in the trees' point forecasts is easy to spot. The existence of outliers is also apparent, as too many trees' means fall far to the right of the linear opinion pool's 90% prediction interval, which creates positive skewness in the trees' means. Trimming at the 10% level trims away from the pool, or cleanses it of, some of these outliers and shifts the ensemble's point forecast to the left by 34.4 bike rentals (from 4,433.6 to 4,399.2). This shift brings us closer to the realization of 3,958 bike rentals. For other rows in the testing set, we found negative skewness in the trees' means, and trimming has the opposite effect of shifting the mean to the right. In terms of calibration, the width of the 90% prediction interval for the trimmed opinion pool is 51% narrower than that of the linear opinion pool. Thus, the trimmed opinion pool cleans, aggregates, and recalibrates—all at the same time—using a single parameter.

One reason for the outliers present in Figure 8 may be that several wild trees were grown up as a result of only `mtry` predictors (out of q total predictors) being selected randomly at each binary split. With a sequence of several unusual sets of variables to split on, the domain of a tree's relevant terminal node may be an unusual (and useless) subset of the entire $\mathbf{x}$ domain. That tree's forecast might then be an outlier. For instance, when predicting bike rentals, a tree could have left out temperature from each set of variables it chose to split on, at each node. We might naturally expect that a tree that did not include temperature at all as a predictor would produce an errant forecast of bike rentals.

Figure 7. Empirical PIT Densities with Hit Rates (Shaded Region) from a Random Forest on the Bike-Sharing Data Set for the Testing Set



Note. TOP, trimmed opinion pool; LOP, linear opinion pool.

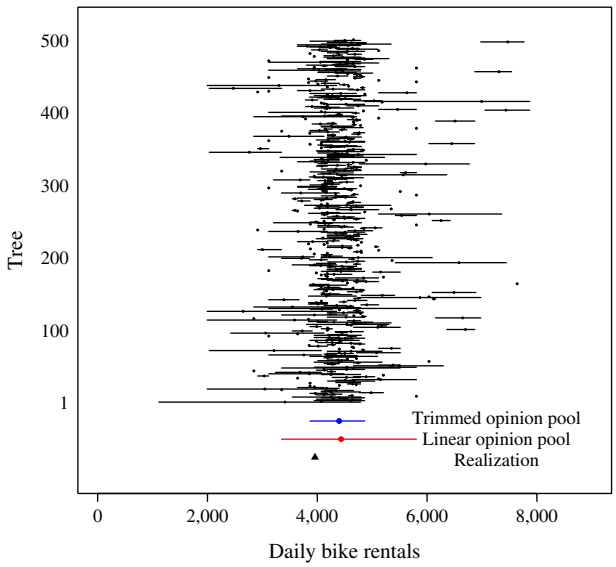
There are other ways in which we might clean, aggregate, and calibrate forecasts of a continuous response. For instance, we could trim the individual trees first and then average their forecasts. This approach would be along the lines of the "calibrate-then-average" method examined in Turner et al. (2014). Trimming the individual trees first would, however, have little effect on the diversity we see in Figure 8. It is this diversity, or decorrelation, among the trees that causes the linear opinion pool to be poorly calibrated with a high hit rate (Jose et al. 2014). Consequently, the "trim-then-average" method is also likely to be poorly calibrated with a high hit rate, similar to the linear opinion pool. Nonetheless, Turner et al. (2014) demonstrate that there are benefits to a calibrate-then-average approach in settings where the response is a binary event and some transformations are applied. In the next subsection, we examine the performance of the trimmed opinion pool in a binary-event setting.

3.3. Trimming the Trees in a Random Forest: Binary-Event Response

Next, we present results on the performance of the trimmed opinion pool on three binary-classification (BC) data sets: the Friedman #1BC, bank, and spambase data sets. The Friedman #1BC data set is a binary classification version of the Friedman #1 data set. We created it by substituting the continuous response with a binary variable ($y = 0$ if below the continuous response's median and $y = 1$ otherwise). The other two data sets are available from the UCI data repository. The bank data set has 20 predictors, 4,119 observations, and a binary response that describes whether or not a contact opens a deposit account as a result of a direct marketing campaign by a Portuguese bank. The spambase data set has 57 predictors of whether or not an email is considered spam and 4,601 observations.

Table 4 presents the average MSE and W for several opinion pools. Recall that the MSE is equivalent to the

Figure 8. (Color online) Ninety Percent Prediction Intervals of Individual Trees, Trimmed Opinion Pool, and Linear Opinion Pool from a Random Forest for One Test Value in the Bike-Sharing Data Set



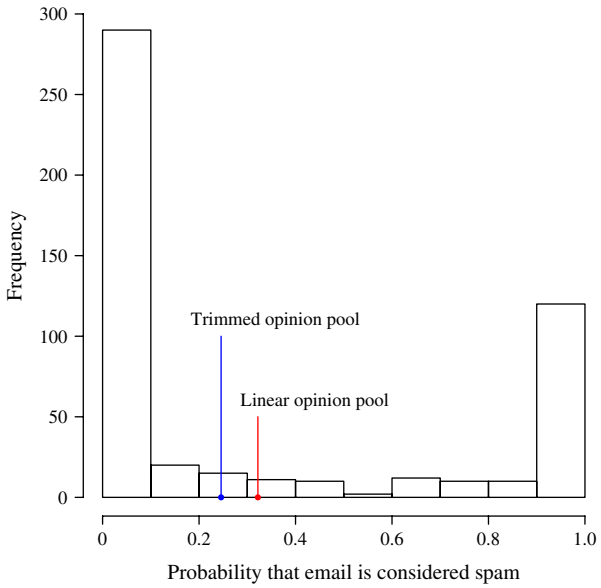
Brier score, which is a proper scoring rule for a binary-event probability forecast. For the Friedman #1BC data set, MSE is minimized at 20% trimming. For the bank data set, there is no need for trimming, and for the spambase data sets, MSE is minimized at 15% trimming. Note that the optimal trimming level, as determined by cross-validation, changes as the settings of the random forest change, although never by more than 5% points. In particular, when `nodesize` changes from 5 to 10, our trimming recommendations for the bank data set changes from 0% to 5%.

Table 4. Average MSE and W (in Percentages) of Trimmed Opinion Pools ($\hat{f}_\alpha$) from a Random Forest on Binary Classification Data Sets for the Validation and Testing Sets

Set(s)	Pool	Friedman #1BC		Bank		Spambase	
Validation	α	MSE	W	MSE	W	MSE	W
	0.00	0.120	56.6	**0.05910**	**44.13**	0.0428	57.6
	0.05	0.115	58.5	0.05915	44.08	0.0403	60.1
	0.10	0.111	60.1	0.05969	43.58	0.0389	61.5
	0.15	0.108	61.2	0.06056	42.76	**0.0383**	**62.1**
	0.20	**0.107**	**61.6**	0.06178	41.62	0.0384	62.0
	0.25	0.108	61.4	0.06330	40.19	0.0389	61.5
	0.30	0.111	60.6	0.06497	38.62	0.0398	60.6
	0.35	0.116	59.0	0.06655	37.13	0.0409	59.5
	0.40	0.122	56.9	0.06781	35.94	0.0420	58.4
	0.45	0.129	54.7	0.06878	35.02	0.0427	57.7
	0.50	0.132	53.5	0.06934	34.48	0.0429	57.5
Testing	α	MSE	W	MSE	W	MSE	W
	0.00	0.117	56.8	0.05752	43.59	0.0450	55.3
	α^*	0.102	62.4	0.05752	43.59	0.0420	58.3

Note. Numbers in bold represent best values in the validation set.

Figure 9. (Color online) Histogram of Probability Forecasts from Individual Trees, Trimmed Opinion Pool, and Linear Opinion Pool from a Random Forest for One Test Value in the Spambase Data Set



Although Proposition 7 predicts no need for trimming when faced with a classification task and forecasts are completely overfit, we see empirical gains from trimming. This result again appears to be caused by outliers. The histogram in Figure 9 depicts the frequencies of the individual trees' probability forecasts for a row in the testing set from the spambase data set. In this figure, we see that outliers on the right (i.e., many probability forecasts of 1) generate positive skewness in the trees' probability forecasts. Trimming at the 15% level trims away some of these outliers and shifts the trimmed ensemble's probability forecast to the left by 0.076 (from 0.322 to 0.246). This shift brings us closer to the realization of 0 (or not spam). When the realization is 0 (1), we might expect more trees to issue probability forecasts less (more) than one-half and thus would consider a probability forecast of 1 (0) to be an outlier.

4. Conclusion

In this paper, we study the impact of overfitting and overconfidence on average point and probability forecasts. We first study their combined effect using a formal model of information gathering and biased reporting. With this model, we provide closed-form expressions for the limiting linear and trimmed opinion pools as the crowd grows large. Using these expressions, we show the systematic ways in which overfitting and overconfidence affect the accuracy of the average point and probability forecasts. Specifically, in the case of a continuous uncertain quantity of interest, overfitting results in accuracy gains for the

average point forecast but at the same time hurts the accuracy of the average probability forecast. We also offer theoretical predictions for how these individual forecaster traits affect individual and ensembles' hit rates. Surprisingly, it turns out that overfitting alone is enough to produce both individual hit rates that are too low and ensemble hit rates that are too high. One of our main theoretical findings is that the trimmed opinion pool is effective when hit rates of the mean and median probability forecasts bracket the correct hit rate. This bracketing typically occurs when the two traits, overfitting and overconfidence, are present.

To test our model's predictions, we consider the random forest, a popular machine-learning algorithm, fit to seven well-known, publicly available data sets, two of which are synthetic. These empirical tests were motivated by the fact that a random forest, by construction, relies on a large number of weak learners. One of our main empirical contributions is in identifying the random forest's trees as candidates for trimming, specifically because its trees may be both overfit and overconfident. Of course, without knowing the true model from which the data were drawn, and because the random forest is a very complicated model, it is difficult to prove that trees in a random forest will be overfit. Nonetheless, researchers have numerically studied many synthetic data sets where the true model is known, and as a result, it is widely believed that the trees in a random forest generally overfit to the data they are trained up on. On the basis of our study of the Friedman data set, we replicated the finding that the random forest's trees can be overfit. Using this same data set, we offer the additional insight that the trees can also be overconfident. Thus, when these two empirical observations are put together with our theoretical results, a testable hypothesis emerges: trimming the trees in a random forest will improve its aggregate probability forecast.

Our hypothesis that trimming offers improvement is confirmed on six of the seven data sets we investigate. For the purposes of this investigation, we developed an augmented random forest algorithm called `trimTrees`, a freely and publicly available `R` package. Our augmented prediction algorithm finds the empirical cdf of each tree in a random forest and calculates a trimmed opinion pool, the trimmed average of the trees' empirical cdfs. This algorithm constitutes a new machine-learning algorithm inspired by aggregation methods developed in the decision analysis literature. At 0% and 50% trimming, the trimmed opinion pool specializes to the linear opinion pool (or mean probability forecast) and median probability forecast, respectively. When we apply our algorithm to the four data sets with continuous responses, the hit rates we find for the mean and median probability forecasts indeed bracket the correct hit rate. Trimming at some level brings us closer to the correct hit rate and improves the accuracy of the aggregate probability forecast. We also see improvement in the accuracy of the trimmed ensemble's point forecast. One explanation for this last improvement is that trimming removes the effects of outliers, or wild trees' point predictions. On the three classification data sets studied, we find that trimming improves the accuracy of binary-event probability forecasts on two of the three data sets. These improvements in binary classification may again be due to the removal of outliers. Consequently, as a result of our theoretical and empirical findings, we expect trimmed opinion pools to be leading contenders for aggregating forecasts in the future.

Appendix

Proof of Proposition 1. We begin with an expression for the mean squared error of the average point forecast $\bar{\mu} = (1/k)\sum_{i=1}^{k}\mu_i'$ in our forecasting environment:

$$\begin{aligned}\mathrm{MSE}(\bar{\mu}) &= E[(y-\bar{\mu})^2] = E[E[(y-\bar{\mu})^2 \mid \mu]]\\ &= E\left[\mathrm{Var}\left[y-\frac{1}{k}\sum_{i=1}^{k}\mu_i' \,\middle|\, \mu\right]\right.\\ &\quad \left. + E\left[y-\frac{1}{k}\sum_{i=1}^{k}\mu_i' \,\middle|\, \mu\right]^2\right]\\ &= E\left[\left(1+\frac{w'^2}{kn}\right)\frac{1}{\lambda}+(1-w')^2(\mu-\mu_0)^2\right]\\ &= \left(1+\frac{w'^2}{kn}+\frac{(1-w')^2}{m}\right)\frac{1}{\lambda}, \end{aligned} \tag{6}$$

where the third and fourth equalities follow because

$$\begin{aligned}(y-\bar{\mu}\mid\mu) &= \left(y-\frac{1}{k}\sum_{i=1}^{k}((1-w')\mu_0+w'\bar{y}_i)\,\middle|\,\mu\right)\\ &\sim \mathrm{N}\left((1-w')(\mu-\mu_0),\left(1+\frac{w'^2}{kn}\right)^{-1}\lambda\right).\end{aligned}$$

When $k=1$, we get an expression for $1/E_i$.

Next we show that the crowd's wisdom W is equal to crowd's diversity divided by the average individual expertise: $W = D/((1/k)\sum_{i=1}^{k}\mathrm{MSE}(\mu_i'))$. This decomposition follows because

$$\begin{aligned}&\frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}(\mu_i')\\ &= E\left[\frac{1}{k}\sum_{i=1}^{k}(y-\mu_i')^2\right] = E\left[\frac{1}{k}\sum_{i=1}^{k}(y-\bar{\mu}+\bar{\mu}-\mu_i')^2\right]\\ &= E\left[\frac{1}{k}\sum_{i=1}^{k}(y-\bar{\mu})^2+\frac{2}{k}\sum_{i=1}^{k}(y-\bar{\mu})(\bar{\mu}-\mu_i')+\frac{1}{k}\sum_{i=1}^{k}(\bar{\mu}-\mu_i')^2\right]\\ &= \mathrm{MSE}(\bar{\mu})+E\left[\frac{2}{k}(y-\bar{\mu})\sum_{i=1}^{k}(\bar{\mu}-\mu_i')\right]+E\left[\frac{1}{k}\sum_{i=1}^{k}(\bar{\mu}-\mu_i')^2\right]\end{aligned}$$

and $\sum_{i=1}^{k}(\bar{\mu}-\mu_i')=0$. Thus, $D=(1/k)\sum_{i=1}^{k}\mathrm{MSE}(\mu_i')-\mathrm{MSE}(\bar{\mu}) = (k-1)w'^2/(kn)$, which is the numerator in the definition of W. Because the individuals in our forecasting environment all have the same mean squared errors, we have $W=E_iD$.

For statement (a), E_i is maximized at $w'=w$, where $E_i=\lambda_i$. The necessary first-order condition for maximum E_i involves

the derivative $dE_i/dw' = -[\mathrm{MSE}(\mu_i')]^{-2}(w'/n(1-w')/m) \cdot (2/\lambda)$. This derivative is positive for $w' < w$, equal to 0 for $w' = w$, and negative for $w' > w$, which is a sufficient condition to show that $w' = w$ is a maximizer of individual expertise E_i. It follows that statement (a) holds: E_i is decreasing in $|w' - w|$. For statement (b), the crowd's diversity is given by

$$\begin{aligned}
&E\left[\frac{1}{k}\sum_{i=1}^{k}(\mu_i' - \bar{\mu})^2\right] \\
&\quad = E\left[\frac{1}{k}\sum_{i=1}^{k} E\left[\left(\left(1-\frac{1}{k}\right)w'\bar{y}_i - \frac{1}{k}\sum_{j\neq i} w'\bar{y}_j\right)^2 \middle| \mu\right]\right] \\
&\quad = E\left[\frac{1}{k}\sum_{i=1}^{k}\left(\mathrm{Var}\left[\left(1-\frac{1}{k}\right)w'\bar{y}_i - \frac{1}{k}\sum_{j\neq i} w'\bar{y}_j \middle| \mu\right]\right.\right. \\
&\qquad\qquad \left.\left. + E\left[\left(1-\frac{1}{k}\right)w'\bar{y}_i - \frac{1}{k}\sum_{j\neq i} w'\bar{y}_j \middle| \mu\right]^2\right)\right] \\
&\quad = \left(\frac{(k-1)^2 w'^2}{k^2} + \frac{(k-1)w'^2}{k^2}\right)\frac{1}{n\lambda} = \frac{(k-1)w'^2}{k}\frac{1}{n\lambda}, \qquad (7)
\end{aligned}$$

which is clearly increasing in w' for $w' > 0$. The third equality here follows because

$$\begin{aligned}
&\left(\left(1-\frac{1}{k}\right)w'\bar{y}_i - \frac{1}{k}\sum_{j\neq i} w'\bar{y}_j \middle| \mu\right) \\
&\qquad \sim \mathrm{N}\left(0, \left(\frac{(k-1)^2 w'^2}{k^2} + \frac{(k-1)w'^2}{k^2}\right)^{-1} n\lambda\right).
\end{aligned}$$

For statement (c), we show that the crowd's wisdom is maximized at $w' = m + 1 > 1$ for $w' > 0$. The derivative of $1 - \mathrm{MSE}(\bar{\mu})/\mathrm{MSE}(\mu_i')$ with respect to w' simplifies to the expression

$$\frac{dW}{dw'} = \frac{2nmw'(k-1)(-w'+m+1)}{k(nm+mw'^2+n-2nw'+nw'^2)^2}.$$

This derivative is positive for $0 < w' < m + 1$, equal to 0 for $w' = m + 1$, and negative for $w' > m + 1$, which is a sufficient condition to show that $w' = m + 1$ is a maximizer of W. Statement (c) then holds: W is increasing in w' for $0 < w' \leq 1 < m + 1$. □

Proof of Proposition 2. Because $(y - w'\bar{y}_i \mid \mu) \sim \mathrm{N}((1-w')\mu, n\lambda/(n+w'^2))$, the cdf of $(F_i(y) \mid \mu)$, an individual forecaster's PIT given μ, is

$$\begin{aligned}
&G_i(p \mid \mu) \\
&\quad = \Pr(F_i(y) \leq p \mid \mu) \\
&\quad = \Pr(\Phi(\lambda_i'^{1/2}(y - (1-w')\mu_0 - w'\bar{y}_i)) \leq p \mid \mu) \\
&\quad = \Pr(y - w'\bar{y}_i \leq (1-w')\mu_0 + \lambda_i'^{-1/2}\Phi^{-1}(p) \mid \mu) \\
&\quad = \Phi\left(\left(\frac{n\lambda}{n+w'^2}\right)^{1/2}[(1-w')\mu_0 + \lambda_i'^{-1/2}\Phi^{-1}(p) - (1-w')\mu]\right) \\
&\quad = \Phi(a\Phi^{-1}(p) + b(m\lambda)^{1/2}(\mu - \mu_0)),
\end{aligned}$$

where $a = (n\lambda/(n+w'^2))^{1/2}\lambda_i'^{-1/2}$ and $b = -(n\lambda/(n+w'^2))^{1/2} \cdot (1-w')(m\lambda)^{-1/2}$. Integrating out μ and using the substitution according to $x = (m\lambda)^{1/2}(\mu - \mu_0)$, we have

$$\begin{aligned}
G_i(p) &= \int_{-\infty}^{\infty} \Phi(a\Phi^{-1}(p) + b(m\lambda)^{1/2}(\mu - \mu_0))(m\lambda)^{1/2} \\
&\quad \cdot \phi((m\lambda)^{1/2}(\mu - \mu_0))\, d\mu \\
&= \int_{-\infty}^{\infty} \Phi(a\Phi^{-1}(p) + bx)\phi(x)\, dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\Phi^{-1}(p)\right),
\end{aligned}$$

from which the PIT density immediately follows. The third equality follows from Owen's (1980) Equation 10,010.8. Let $C_i = a/\sqrt{1+b^2}$. The hit rate is given by

$$\begin{aligned}
H_i(u) &= \Pr(u \leq F_i(y) \leq 1-u \mid \mu) = G_i(1-u) - G_i(u) \\
&= 1 - 2G_i(u) = 1 - 2\Phi(C_i\Phi^{-1}(u))
\end{aligned}$$

because $G_i(1-u) = \Phi(C_i\Phi^{-1}(1-u)) = \Phi(-C_i\Phi^{-1}(u)) = 1 - \Phi(C_i\Phi^{-1}(u)) = 1 - G_i(u)$ by the properties of the standard normal cdf Φ. □

Proof of Proposition 3. Given μ, $F_1(z) \leq \cdots \leq F_k(z)$ are iid because $(\bar{y}_1, \ldots, \bar{y}_k)$ are conditionally independent given μ. Thus, we can apply the central limit theorem in Stigler (1973) concerning the asymptotic distribution of the trimmed mean. By Stigler and the fact that a central limit theorem implies a weak law of large numbers (Lehmann 1998), the trimmed opinion pool $(\hat{F}_\alpha(z) \mid \mu)$, being a trimmed mean of the individual forecasters' cdf values, converges in probability to the mean of a truncated distribution of $(F_i(z) \mid \mu)$. Here, the distribution of $(F_i(z) \mid \mu)$ is truncated below at $J^{-1}(\alpha)$ and above at $J^{-1}(1-\alpha)$, where J is the cdf of $(F_i(z) \mid \mu)$. Because $(\bar{y}_i \mid \mu) \sim \mathrm{N}(\mu, n\lambda)$, the cdf of $(F_i(z) \mid \mu)$ is given by

$$\begin{aligned}
J(u) &= \Pr(F_i(z) \leq u \mid \mu) \\
&= \Pr(\Phi(\lambda_i'^{1/2}(z - (1-w')\mu_0 - w'\bar{y}_i)) \leq u \mid \mu) \\
&= \Pr(\bar{y}_i \geq w'^{-1}(z - (1-w')\mu_0 - \lambda_i'^{-1/2}\Phi^{-1}(u)) \mid \mu) \\
&= 1 - \Phi((n\lambda)^{1/2}(w'^{-1}(z - (1-w')\mu_0 - \lambda_i'^{-1/2}\Phi^{-1}(u)) - \mu)) \\
&= 1 - \Phi\left(\frac{(n\lambda)^{1/2}}{w'}(z - \vec{\mu}) - \frac{(n\lambda)^{1/2}}{w'}\lambda_i'^{-1/2}\Phi^{-1}(u)\right) \\
&= 1 - \Phi\left(\frac{\Phi^{-1}(u) - a(z)}{b}\right),
\end{aligned}$$

where $a(z) = \lambda_i'^{1/2}(z - \vec{\mu})$, $b = -(n\lambda)^{-1/2}w'\lambda_i'^{1/2}$, and $J^{-1}(v) = \Phi(a(z) + b\Phi^{-1}(1-v))$. Given μ, the trimmed opinion pool's cdf $\hat{F}_\alpha(z)$ converges in probability to

$$\begin{aligned}
F_\alpha(z) &= E[F_i(z) \mid J^{-1}(\alpha) \leq F_i(z) \leq J^{-1}(1-\alpha), \mu] \\
&= \frac{1}{1-2\alpha}\int_{\Phi(a(z)+b\Phi^{-1}(\alpha))}^{\Phi(a(z)+b\Phi^{-1}(1-\alpha))} u\, dJ(u) \\
&= \frac{1}{1-2\alpha}\int_{\Phi^{-1}(\alpha)}^{\Phi^{-1}(1-\alpha)} \Phi(a(z)+bx)\phi(x)\, dx.
\end{aligned}$$

The third equality follows from a substitution according to $x = (\Phi^{-1}(u) - a(z))/b$, where $dx/du = 1/(b\phi(\Phi^{-1}(u)))$ and $u = \Phi(a(z) + bx)$. By the dominated convergence theorem, differentiation of $F_\alpha(z)$ with respect to z yields $f_\alpha(z) = (a'(z)/(1-2\alpha))\int_{\Phi^{-1}(\alpha)}^{\Phi^{-1}(1-\alpha)} \Phi(a(z)+bx)\phi(x)\, dx$. By the integral $\int \phi(a+bx)\phi(x)\, dx = (1/\sqrt{b^2+1})\phi(a/\sqrt{b^2+1})\Phi(x\sqrt{b^2+1} + ab/\sqrt{b^2+1})$ in Owen (1980, Table I, Equation n10), we obtain the pdf in the result. The cdf $F_\alpha(z)$ can be simplified using the integral $\int_{-\infty}^{\gamma} \Phi((\delta - \rho x)/\sqrt{1-\rho^2})\phi(x)\, dx = \Phi_B(\gamma, \delta; \rho)$ in Owen (1980, Table I, Equation 10,010.2):

$$\begin{aligned}
F_\alpha(z) &= \frac{1}{1-2\alpha}\int_{-\infty}^{\Phi^{-1}(1-\alpha)} \Phi(a(z)+bx)\phi(x)\, dx \\
&\quad - \frac{1}{1-2\alpha}\int_{-\infty}^{\Phi^{-1}(\alpha)} \Phi(a(z)+bx)\phi(x)\, dx
\end{aligned}$$

$$= \frac{1}{1-2\alpha}\left[\Phi_B\left(\Phi^{-1}(1-\alpha), \frac{a(z)}{\sqrt{1+b^2}}; \frac{-b}{\sqrt{1+b^2}}\right) - \Phi_B\left(\Phi^{-1}(\alpha), \frac{a(z)}{\sqrt{1+b^2}}; \frac{-b}{\sqrt{1+b^2}}\right)\right].$$

The cdf $F_\alpha(z)$ with $\alpha = 0$ reduces to $F_\alpha(z) = \Phi(\bar{\lambda}^{1/2}(z-\vec{\mu}))$ using the integral $\int_{-\infty}^{\infty} \Phi(a+bx)\phi(x)\,dx = \Phi(a/\sqrt{b^2+1})$ in Owen (1980, Table I, Equation 10,010.8). The limit of $F_\alpha(z)$ as $\alpha \to \frac{1}{2}$ is given by

$$\begin{aligned}
&\lim_{\alpha\to 1/2} F_\alpha(z) \\
&= \lim_{\alpha\to 1/2} \frac{1}{1-2\alpha} \int_{\Phi^{-1}(\alpha)}^{\Phi^{-1}(1-\alpha)} \Phi(a(z)+bx)\phi(x)\,dx \\
&= \frac{-\Phi(a(z)+b\Phi^{-1}(1-\alpha)) - \Phi(a(z)+b\Phi^{-1}(\alpha))}{-2}\Bigg|_{\alpha=1/2} \\
&= \Phi(a(z)) = \Phi(\lambda_i'^{1/2}(z-\vec{\mu})),
\end{aligned}$$

where the second equality follows from l'Hôpital's and Leibnitz's rules. □

Proof of Proposition 4. By Arnold et al. (1993), the LTOP distribution is the distribution of X given $\gamma_1 < Y < \gamma_2$, where (X, Y) are jointly normally distributed with mean vector $(\mu_1, \mu_2) = (\vec{\mu}, 0)$, variance vector $(\sigma_1, \sigma_2) = (\bar{\lambda}^{-1/2}, 1)$, correlation $\rho = (1-\bar{\lambda}/\lambda_i')^{1/2}$, $\gamma_1 = \Phi^{-1}(\alpha)$, and $\gamma_2 = \Phi^{-1}(1-\alpha)$. Let $U = (X-\mu_1)/\sigma_1$. By Arnold et al. (1993, Equation 13a), the first moment of U given $\gamma_1 < Y < \gamma_2$ is given by $E[U \mid \gamma_1 < Y < \gamma_2] = -(\rho/(1-2\alpha))[\phi(\Phi^{-1}(1-\alpha)) - \phi(\Phi^{-1}(\alpha))] = 0$, because ϕ is symmetric around $\Phi^{-1}(1/2) = 0$, which implies that $\phi(\Phi^{-1}(1-\alpha)) = \phi(\Phi^{-1}(\alpha))$. Consequently, $E[X \mid \gamma_1 < Y < \gamma_2] = E[\mu_1 + \sigma_1 U \mid \gamma_1 < Y < \gamma_2] = \mu_1$. By Arnold et al. (1993, Equation 13b), the second moment of U given $\gamma_1 < Y < \gamma_2$ is given by $E[U^2 \mid \gamma_1 < Y < \gamma_2] = 1 - (\rho^2/(1-2\alpha))[\Phi^{-1}(1-\alpha)\phi(\Phi^{-1}(1-\alpha)) - \Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))] = 1 + 2\rho^2(\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))/(1-2\alpha))$. Using these expressions, we can compute for the variance of X given $\gamma_1 < Y < \gamma_2$ as follows: $\mathrm{Var}[X \mid \gamma_1 < Y < \gamma_2] = \sigma_1^2 \mathrm{Var}[U \mid \gamma_1 < Y < \gamma_2] = \sigma_1^2(E[U^2 \mid \gamma_1 < Y < \gamma_2] - E[U \mid \gamma_1 < Y < \gamma_2]^2)$, which yields the result.

Let $M = 1 + 2(1 - n\lambda/(w'^2\lambda_i' + n\lambda))\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))/(1-2\alpha)$. The derivative of $\vec{\lambda}$ with respect to λ_i' is given by $d\vec{\lambda}/d\lambda_i' = (d\bar{\lambda}/d\lambda_i')(1/M) - (\bar{\lambda}/M^2)(dM/d\lambda_i')$. Clearly, M and $\bar{\lambda}$ are positive. We also have that $d\bar{\lambda}/d\lambda_i' = n\lambda/(w'^2\lambda_i' + n\lambda) - n\lambda\lambda_i' w'^2/(w'^2\lambda_i' + n\lambda)^2 = (n\lambda)^2/(w'^2\lambda_i' + n\lambda)^2 > 0$ and $dM/d\lambda_i' = (2n\lambda w'^2/(w'^2\lambda_i' + n\lambda)^2)\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))/(1-2\alpha)$, which is negative because $\Phi^{-1}(\alpha) < 0$ for any trimming level α. Taken together, we have that $d\vec{\lambda}/d\lambda_i' > 0$. For showing decreasing in w', the derivative of $\vec{\lambda}$ with respect to w' is given by $d\vec{\lambda}/dw' = (d\bar{\lambda}/dw')(1/M) - (\bar{\lambda}/M^2)(dM/dw') < 0$ because $d\bar{\lambda}/dw' = -2n\lambda w'\lambda_i'^2/(w'^2\lambda_i' + n\lambda)^2 < 0$ and $dM/dw' = (2n\lambda\lambda_i'/(w'^2\lambda_i' + n\lambda)^2)\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))/(1-2\alpha) < 0$. □

Proof of Proposition 5. Without loss of generality, let $\vec{\mu} = 0$ and $\bar{\lambda} = 1$. For $0 \le \alpha_1 \le \alpha_2 < \frac{1}{2}$, let X and Y be truncated normal distributions with distribution functions $F(x) = (\Phi(x/\rho) - \alpha_2)/(1-2\alpha_2)$ and $G(y) = (\Phi(y/\rho) - \alpha_1)/(1-2\alpha_1)$, respectively. For $\alpha_2 = \frac{1}{2}$, let $X = 0$ with probability 1, and let Y follow the same truncated normal distribution. Except when $\alpha_2 = \frac{1}{2}$, the random variable X is truncated below at $\rho\Phi^{-1}(\alpha_2)$ and above at $\rho\Phi^{-1}(1-\alpha_2)$ and similarly for Y. In this case, the moment-generating function of X is $M_X(t) = \exp(\rho^2 t^2)((\Phi(\Phi^{-1}(1-\alpha_2) - \rho t) - \Phi(\Phi^{-1}(\alpha_2) - \rho t))/(1-2\alpha_2))$ and similarly for Y (Forbes et al. 2011, p. 147). When $\alpha_2 = \frac{1}{2}$, the moment-generating function of X is 1.

When $0 \le \alpha_1 \le \alpha_2 < \frac{1}{2}$, the quantile function of X is $F^{-1}(p) = \rho\Phi^{-1}((1-2\alpha_2)p + \alpha_2)$ and similarly for Y. In this case, dispersive order $F^{-1}(t) - F^{-1}(s) \le G^{-1}(t) - G^{-1}(s)$ for all $0 < s < t < 1$ is equivalent to the function $Q(\alpha, p) = \rho\Phi^{-1}((1-2\alpha)p + \alpha)$ being submodular in (α, p). The function Q is submodular if $Q(\alpha_2, t) + Q(\alpha_1, s) \le Q(\alpha_1, t) + Q(\alpha_2, s)$ for all $0 \le \alpha_1 \le \alpha_2 < \frac{1}{2}$ and $0 < s < t < 1$ or $\partial^2 Q(\alpha, p)/\partial\alpha\partial p \le 0$. The cross-partial derivative of $Q(\alpha, p)$ is

$$\frac{\partial^2 Q(\alpha,p)}{\partial\alpha\partial p} = \frac{\rho(1-2\alpha)(1-2p)\Phi^{-1}((1-2\alpha)p+\alpha)}{[\phi(\Phi^{-1}((1-2\alpha)p+\alpha))]^2} - \frac{2\rho}{\phi(\Phi^{-1}((1-2\alpha)p+\alpha))},$$

which is always less than or equal to 0 because $(1-2\alpha)\cdot(1-2p)\Phi^{-1}((1-2\alpha)p+\alpha) \le 0 \le 2\phi(\Phi^{-1}((1-2\alpha)p+\alpha))$. Hence, $Q(\alpha, p)$ submodular implies $X \le_{\text{disp}} Y$. When $\alpha_2 = \frac{1}{2}$, $X \le_{\text{disp}} Y$ clearly holds.

Next we add an independent normal random variable $Z \sim F_{1/2}$ to both X and Y. Then, by Shaked and Shanthikumar (2007, Theorem 3.B.8, p. 152), because Z has a logconcave density, Z is dispersive, and $X + Z \le_{\text{disp}} Y + Z$. Because Y_2 follows the nontruncated marginal distribution of a truncated bivariate normal distribution, its moment-generating function is given by $M_{Y_2}(t) = \exp(t^2)\cdot((\Phi(\Phi^{-1}(1-\alpha_2) - \rho t) - \Phi(\Phi^{-1}(\alpha_2) - \rho t))/(1-2\alpha_2))$ (Arnold et al. 1993, Equation 8) and similarly for Y_1. Because Z has the moment-generating function $\exp(\lambda_i'^{-1}t^2/2)$ and $\rho^2 + \lambda_i'^{-1} = 1$, the moment-generating function of $X + Z$ is equal to $M_{Y_2}(t)$. Similarly, the moment-generating function of $Y + Z$ is equal to $M_{Y_1}(t)$. Thus, $Y_2 \le_{\text{disp}} Y_1$.

To show that the central even moments are ordered, we apply Theorem 1.7.6 in Müller and Stoyan (2002), which states that $Y_2 \le_{\text{disp}} Y_1$ implies $Y_2 - E[Y_2]$ is less than $Y_1 - E[Y_1]$ in convex order, denoted $Y_2 - E[Y_2] \le_{\text{cx}} Y_1 - E[Y_1]$. Because $E[Y_2] = E[Y_1]$ (Proposition 4), $Y_2 - E[Y_2] \le_{\text{cx}} Y_1 - E[Y_1]$ implies $Y_2 \le_{\text{cx}} Y_1$. By Corollary 1.5.4 in Müller and Stoyan (2002), $Y_2 \le_{\text{cx}} Y_1$ implies the central even moments of Y_2 are less than those of Y_1. □

Proof of Proposition 6. First, we show that $H_{\alpha_2}(u) \le H_{\alpha_1}(u)$ for $0 \le \alpha_1 \le \alpha_2 < \frac{1}{2}$. Because $F_\alpha(z) = \Psi_\alpha(\bar{\lambda}^{1/2}(z-\vec{\mu}))$ and $(y \mid \mu) \sim \mathrm{N}(\mu, \lambda)$, the cdf of $(F_\alpha(y) \mid \mu)$, the limiting trimmed opinion pool's PIT given μ, is

$$\begin{aligned}
G_\alpha(u \mid \mu) &= \Pr(F_\alpha(y) \le u \mid \mu) \\
&= \Pr(\Psi_\alpha(\bar{\lambda}^{1/2}(y - (1-w')\mu_0 - w'\mu)) \le u \mid \mu) \\
&= \Pr(y \le (1-w')\mu_0 + w'\mu + \bar{\lambda}^{-1/2}\Psi_\alpha^{-1}(u) \mid \mu) \\
&= \Phi(\lambda^{1/2}[(1-w')\mu_0 + w'\mu + \bar{\lambda}^{-1/2}\Psi_\alpha^{-1}(u) - \mu]) \\
&= \Phi(a\Psi_\alpha^{-1}(u) + b(m\lambda)^{1/2}(\mu - \mu_0)),
\end{aligned}$$

where $a = \lambda^{1/2}\bar{\lambda}^{-1/2}$ and $b = -\lambda^{1/2}(1-w')(m\lambda)^{-1/2}$. Following the same steps in the proof of Proposition 2 for integrating out μ, we get $H_\alpha(u) = 1 - 2\Phi(C_0\Psi_\alpha^{-1}(u))$, where

$$\begin{aligned}
C_0 &= \left(\frac{1}{1+(1-w')^2 m^{-1}}\right)^{1/2}\left(\frac{\lambda}{\bar{\lambda}}\right)^{1/2} \\
&= \left(\frac{1}{1+(1-w')^2 m^{-1}}\right)^{1/2}\lambda^{1/2}\left(\frac{n\lambda\lambda_i'}{w'^2\lambda_i' + n\lambda}\right)^{-1/2}
\end{aligned}$$

$$= \left(\frac{1}{1+(1-w')^2 m^{-1}}\right)^{1/2}\left(\frac{w'^2\lambda_i' + n\lambda}{n\lambda_i'}\right)^{1/2}$$
$$= \left(1+\frac{(1-w')^2}{m}\right)^{-1/2}\left(\frac{w'^2}{n}+\frac{\lambda}{\lambda_i'}\right)^{1/2}.$$

Because $H_{\alpha_2}(u) \le H_{\alpha_1}(u) \Leftrightarrow \Psi^{-1}_{\alpha_2}(u) \ge \Psi^{-1}_{\alpha_1}(u)$, we will show that $\Psi^{-1}_{\alpha_2}(u) \ge \Psi^{-1}_{\alpha_1}(u)$ for $0 < u < \frac{1}{2}$. Let $Y_1 \sim \Psi_{\alpha_1}$ and $Y_2 \sim \Psi_{\alpha_2}$. Because Ψ_α is symmetric about zero, the dispersive order of Y_1 and Y_2 (Proposition 5) implies that $\Psi^{-1}_{\alpha_2}(t) - \Psi^{-1}_{\alpha_2}(s) \le \Psi^{-1}_{\alpha_1}(t) - \Psi^{-1}_{\alpha_1}(s)$ for all $0 < s < t < 1$. Let $t = \frac{1}{2}$ so that this condition becomes $\Psi^{-1}_{\alpha_2}(s) \ge \Psi^{-1}_{\alpha_1}(s)$ for $0 < s < \frac{1}{2}$.

Next we have that $\Psi_0(\tau) = \Phi(\tau)$ because as α goes to 0, $\Phi_B(\Phi^{-1}(1-\alpha), \tau; \rho)$ goes to $\Phi(\tau)$ and $\Phi_B(\Phi^{-1}(\alpha), \tau; \rho)$ goes to 0. Also, $\lim_{\alpha\to 1/2}\Psi_\alpha(\tau) = \Phi((\lambda_i'/\bar{\lambda})^{1/2}\tau)$. Consequently, $H_{1/2}(u) = 1 - 2\Phi(C_0(\bar{\lambda}/\lambda_i')^{1/2}\Phi^{-1}(u))$, or $H_{1/2}(u) = 1 - 2\Phi(C_{1/2}\Phi^{-1}(u))$. Clearly, $C_{1/2} = (1+(1-w')^2/m)^{-1/2}(\lambda/\lambda_i')^{1/2}$ is greater than C_i, and thus $H_i(u) \le H_{1/2}(u)$.

To find expressions for C_0 and $C_{1/2}$ in terms of W and C_i, we use the relationships $W = 1 - E_i\mathrm{MSE}(\bar{\mu})$, $W = E_i D$, $C_i = (E_i/\lambda_i')^{1/2}$, where $\mathrm{MSE}(\bar{\mu}) = (1 + w'^2/(kn) + (1-w')^2/m)(1/\lambda)$ and $D = (k-1)w'^2/(kn)$. In the limit as $k \to \infty$, $\mathrm{MSE}(\bar{\mu}) = (1+(1-w')^2/m)(1/\lambda)$ and $D = w'^2/n$. Thus, we have

$$C_0 = \left(1+\frac{(1-w')^2}{m}\right)^{-1/2}\left(\frac{w'^2}{n}+\frac{\lambda}{\lambda_i'}\right)^{1/2}$$
$$= \left(1+\frac{(1-w')^2}{m}\right)^{-1/2}\lambda^{1/2}\left(\frac{w'^2}{\lambda n}+\frac{1}{\lambda_i'}\right)^{1/2}$$
$$= \mathrm{MSE}(\bar{\mu})^{-1/2}\left(D+\frac{C_i^2}{E_i}\right)^{1/2} = \left(\frac{1-W}{E_i}\right)^{-1/2}\left(D+\frac{C_i^2}{E_i}\right)^{1/2}$$
$$= \left(\frac{E_i D}{1-W}+\frac{C_i^2}{1-W}\right)^{1/2} = \left(\frac{W+C_i^2}{1-W}\right)^{1/2}.$$

The expression for $C_{1/2}$ follows similarly. □

Proof of Proposition 7. Because (i) $1 - \hat{F}_0 = (1/k)\cdot\sum_{i=1}^k \mu_i' = (1-w')\mu_0 + (w'/k)\sum_{i=1}^k \bar{y}_i$; (ii) $(n\bar{y}_i \mid \mu) \sim_{\mathrm{iid}} \mathrm{Bi}(n,\mu)$, where Bi is the binomial distribution with n trials and μ probability of success; and (iii) a central limit theorem implies a weak law of large numbers (Lehmann 1998), we have the result. □

Proof of Proposition 8. The proof follows from taking the steps in the proof of Proposition 1. □

References

Armstrong SJ (2001) Combining forecasts. Armstrong SJ, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Springer, New York), 417–439.

Arnold BC, Beaver RJ (2000) Hidden truncation models. *Sankhyā: Indian J. Statist. Ser. A* 62:23–35.

Arnold BC, Beaver RJ, Groeneveld RA, Meeker WQ (1993) The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* 58:471–488.

Barbey AK, Sloman SA (2007) Base-rate respect: From ecological rationality to dual processes. *Behavioral Brain Sci.* 30:241–297.

Bar-Hillel M (1980) The base-rate fallacy in probability judgments. *Acta Psych.* 44:211–233.

Bernardo JM, Smith AFM (2000) *Bayesian Theory* (John Wiley & Sons, Chichester, UK).

Breiman L (2001) Random forests. *Machine Learn.* 45:5–32.

Budescu DV, Yu HT (2007) Aggregation of opinions based on correlated cues and advisors. *J. Behavioral Decision Making* 20:153–177.

Clemen RT, Winkler RL (1986) Combining economic forecasts. *J. Bus. Econom. Statist.* 4:39–46.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1:79–101.

DeBondt WFM, Thaler RH (1995) Financial decision-making in markets and firms: A behavioral perspective. Maksimovic V, Jarrow RA, Ziemba WT, eds. *Finance: Handbooks in Operations Research and Management Science* (Elsevier, Amsterdam), 385–410.

Diaconis P, Ylvisaker D (1979) Conjugate priors for exponential families. *Ann. Statist.* 7:269–281.

Forbes C, Evans M, Hastings N, Peacock B (2011) *Statistical Distributions*, 4th ed. (John Wiley & Sons, Hoboken, NJ).

Friedman JH (1991) Multivariate adaptive regression splines. *Ann. Statist.* 19:1–67.

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *J. Roy. Statist. Soc.: Ser. B* 69:243–268.

Goldstein DG, Rothschild D (2014) Lay understanding of probability distributions. *Judgment Decision Making* 9:1–14.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag, New York).

Hora SC (2004) Probability judgments for continuous quantities: Linear combinations and calibration. *Management Sci.* 50:597–604.

Hora SC, Fransen BR, Hawkins N, Susel I (2013) Median aggregation of distribution functions. *Decision Anal.* 10:279–291.

Howard J, Bowles M (2012) The two most important algorithms in predictive modeling today. *Strata Conference Tutorials* (O'Reilly Media Inc., Sebastopol, CA), http://strataconf.com/strata2012/public/sv/q/385.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57:1287–1297.

Jose VRR, Grushka-Cockayne Y, Lichtendahl KC Jr (2014) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60:463–475.

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

Larrick RP, Mannes AE, Soll JB (2012) The social psychology of the wisdom of crowds. Krueger JI, ed. *Frontiers of Social Psychology: Social Psychology and Decision Making* (Psychology Press, Philadelphia), 227–242.

Lee MD, Danileiko I (2014) Using cognitive models to combine probability estimates. *Judgment Decision Making* 9:259–273.

Lehmann EL (1998) *Elements of Large-Sample Theory*, Springer Texts in Statistics (Springer, New York).

Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2:18–22.

Lichtendahl KC Jr, Winkler RL (2007) Probability elicitation, scoring rules, and competition among forecasters. *Management Sci.* 53:1745–1755.

Lichtendahl KC Jr, Grushka-Cockayne Y, Pfeifer PE (2013a) The wisdom of competitive crowds. *Oper. Res.* 61:1383–1398.

Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013b) Is it better to average probabilities or quantiles? *Management Sci.* 59:1594–1611.

Lichtenstein S, Fischhoff B, Phillips LD (1982) Evidential impact of base rates. Tversky A, Kahneman D, Slovic P, ed. *Judgment under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 306–334.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107:276–299.

Massey C, Wu G (2005) Detecting regime shifts: The causes of under- and overreaction. *Management Sci.* 51:932–947.

Meinshausen N (2006) Quantile regression forests. *J. Machine Learn. Res.* 7:983–999.

Müller A, Stoyan D (2002) *Comparison Methods for Stochastic Models and Risks* (John Wiley & Sons, Chichester, UK).

O'Hagan AO, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) *Uncertain Judgements: Eliciting Experts' Probabilities* (John Wiley & Sons, Chichester, UK).

Ottaviani M, Sørensen PN (2006) The strategy of professional forecasting. *J. Financial Econom.* 81:441–466.

Owen D (1980) A table of normal integrals. *Comm. Statist.: Simulation Comput.* 9:389–419.

Radzevick JR, Moore DA (2011) Competing to be certain (but wrong): Market dynamics and excessive confidence in judgment. *Management Sci.* 57:93–106.

Robert CP (2001) *The Bayesian Choice* (Springer-Verlag, New York).

Shaked M, Shanthikumar JG (2007) *Stochastic Orders*, Springer Series in Statistics (Springer, New York).

Stigler SM (1973) The asymptotic distribution of the trimmed mean. *Ann. Statist.* 1:472–477.

Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS (2014) Forecast aggregation via recalibration. *Machine Learn.* 95: 261–289.

Tversky A, Kahneman D (1982) Evidential impact of base rates. Tversky A, Kahneman D, Slovic P, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 153–160.

Varian HR (2014) Big data: New tricks for econometrics. *J. Econom. Perspect.* 28:3–28.

Winkler RL (1981) Combining probability distributions from dependent information sources. *Management Sci.* 27:479–488.

Yaniv I, Foster DP (1995) Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *J. Experiment. Psych.: General* 124:424–432.

Yaniv I, Foster DP (1997) Precision and accuracy of judgmental estimation. *J. Behavioral Decision Making* 10:21–32.