Operations Research

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

Percentage and Relative Error Measures in Forecast Evaluation

Victor Richmond R. Jose

To cite this article:
Victor Richmond R. Jose (2017) Percentage and Relative Error Measures in Forecast Evaluation. Operations Research 65(1):200-211. https://doi.org/10.1287/opre.2016.1550

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

OPERATIONS RESEARCH
Vol. 65, No. 1, January–February 2017, pp. 200–211
ISSN 0030-364X (print), ISSN 1526-5463 (online)

METHODS

Percentage and Relative Error Measures in Forecast Evaluation

Victor Richmond R. Jose[a]

[a] McDonough School of Business, Georgetown University, Washington, DC 20057
Contact: vrj2@georgetown.edu (VRRJ)

Received: September 1, 2014
Revised: January 31, 2016, May 18, 2016
Accepted: June 21, 2016
Published Online in Articles in Advance: November 17, 2016

Subject Classifications: decision analysis: forecasting; statistics: data analysis
Area of Review: Decision Analysis

https://doi.org/10.1287/opre.2016.1550

Abstract. Properties of two large families of scale-free forecast accuracy measures that include popular measures such as mean absolute percentage error, relative error, and squared percentage error, are examined in this paper. We describe the optimal reports when forecasts are evaluated using these measures. We also provide analytic expressions for the optimal Bayes' act associated with these measures under a general power transformation for several well-known probability distributions. We then show that using measures from these two families may inadvertently provide incentives for either pessimism or optimism among forecasters, i.e., rewarding underforecasts or overforecasts relative to some reference measure of central tendency. As an illustration of these concepts, we examine the use of these measures for model selection in a forecast aggregation example using stock price forecasts derived from the Thomson Reuters Institutional Brokers' Estimate System. This example illustrates how aggregation methods that always yield lower estimates relative to the mean or median generally exhibit better scores using percentage error-based measures, while those that yield higher estimates compared to the mean or median will effectively rank higher when relative error-based measures are used.

Keywords: forecast evaluation • forecast accuracy • scale-free measures • scale-independence • pessimism • optimism

1. Introduction

The reliance on both point and probability forecasts is pervasive in the theory and practice of decision analysis. Though probabilistic forecasts contain more information about uncertainties, there are many applications in decision analysis where a point estimate is preferred or perhaps the only piece of information available.

Over the years, the gathering of forecasts has led to the development of many scoring rules (or loss functions) that are useful in judging the quality of point forecasts provided by individuals or models. These include popular measures, such as mean squared error (MSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE), and relative error (RE), to name a few. Though many measures exist, many of these measures are closely related. For example, MAPE is defined as the average percentage deviation of forecasts from their corresponding observed outcomes, whereas RE is the (average) percentage deviation of observed outcomes from their corresponding forecasts. Among these measures, MAPE is perhaps the most widely used measure among forecasters. Two recent surveys by McCarthy et al. (2006) and Fildes and Goodwin (2007) show that MAPE was the most frequently used forecasting accuracy measure among practitioners in businesses and organizations.

A likely reason for the MAPE's popularity is its clear interpretation. In addition, this measure has the ability to take into consideration the scale of data sets. This means that if quantities are expressed using a different scale or unit (e.g., millions instead of thousands, meters instead of centimeters, etc.), the score remains the same. In the literature, this type of measure is referred to as scale-free, scaled, or scale-independent. The choice for making the scores scale-free is often dependent on what benchmark or normalizing factor is used for deviations between the realization and forecasts.

Negligence to take scale into account can often mislead decision makers about the quality of a forecasting source or model (e.g., see Zellner 1986 and Chatfield 1988 on the first M-competition). This typically happens when a few large observations dominate the remaining smaller observations when analyzing the average performance of a forecasting source using measures such as MSE and MAD. Fildes and Goodwin (2007, p. 574) recognized and highlighted this issue when they included the following advice in their list of forecasting principles: organizations should "use error measures that adjust for scale in data."

Despite the popularity of MAPE and the emphasis several authors have made about the importance of scale-free measures, little research has been done to understand the properties of these measures. Some exceptions do exist. For example, the properties of percentage error-based measures have been studied in a regression setting (e.g., Chen et al. 2010). Similar studies for relative error and its squared variants also exist

in the literature (Narula and Wellington 1977, Farnum 1990, Park and Stefanski 1998).

In the realm of forecast evaluation, the importance of taking scale into account and of using scale-free measures has been recognized (e.g., Hyndman and Koehler 2006, Patton 2011). There has also been recent work examining problems related to existing scale-free measures used in practice. For example, Davydenko and Fildes (2013) explain that in demand forecasting the "use of percentage errors is often inadvisable, due to the large number of extremely high percentages which arise from relatively low demand values" (p. 511). They illustrate using a large demand data set at the stock keeping unit level that alternative scaled measures may yield more reasonable conclusions in terms of forecast evaluation.

what is the difference between scale free and percentage

The aim of this paper is to help improve our understanding of several well-known and commonly used scale-free measures. We argue that their use could be misleading when evaluators do not fully understand the consequences associated with using some of these measures in practice. Here, we focus on two large generalized families of scale-free measures. These families include several well-known measures such as absolute percentage error, relative error, and squared percentage error.

In the next section, we study properties of these two families by looking at the relevant statistics that yield the best expected score (i.e., lowest expected loss) for these functions. Under a general power transformation and for several well-known probability distributions, we provide analytic expressions for these relevant statistics. We then show that these families of scale-free measures indirectly provide incentives for forecasts that exhibit pessimism or optimism. To further illustrate these concepts, we then provide a numerical example in a forecast aggregation setting using real-world data. Finally, we provide some concluding remarks about the implication of these results in the theory and practice of forecast evaluation.

2. Properties

Suppose an expert or model reports a point forecast r for some uncertain quantity or random variable Y. A *scoring rule* $S(r, y)$ evaluates a forecast r based on the realization y of Y. These rules are also referred to in the literature as *reward* or *loss functions*. In the point forecasting case, these functions are typically designed such that the score is 0 when $r = y$ and $S(r_1, y) \leqslant S(r_2, y)$ whenever r_1 is closer to y than r_2 based on some notion of distance.

A statistic or report function associated with cumulative distribution function (cdf) F, denoted by $r(F)$, is said to be *consistent* with S if $r(F)$ minimizes $\mathbb{E}_F S(r(F), Y)$ for every distribution function F (e.g., see Gneiting 2011a, b). This is different from the notion of consistency often discussed when dealing with estimators. In decision theory, such a report function is referred to as an *optimal Bayes act* for S.

As an example, consider the squared error function $S^{\mathrm{SE}}(r, y) = (r - y)^2$, which is commonly used in forecast evaluation. When averaged over multiple instances or realizations, this average score becomes MSE. This scoring function is consistent with the mean report, since the mean of the distribution F minimizes the expected score $\mathbb{E}_F S^{\mathrm{SE}}$ for every F. The MAD, another popular measure given by averaging $S^{\mathrm{AD}}(r, y) = |r - y|$ over multiple realizations, has the median instead of the mean as its optimal Bayes' act. In the probability setting (i.e., reported forecasts are probabilities or distributions), consistency is equivalent to the notion of properness in the scoring rules literature.

A scoring rule is said to be *scale-free* if $S(ar,ay) = S(r, y)$ for all $a > 0$ (Hyndman and Koehler 2006). Functions with this property are also referred to as *scale-independent* or *scaled functions* in the literature.

Scale-free rules have been popular and considered desirable in many other contexts, because they avoid the problem of a few large data points dominating a large number of smaller observations when averaging over repeated trials. For example, consider a situation where stock prices are estimated and a squared error measure is used. When predicting the price of a stock such as Berkshire-Hathaway, which ranged between \$100,000 and \$230,000 in the last five years, squared error for a relatively small forecast error (say somewhere between \$100 and \$500) will overshadow any realistic price prediction for a stock such as Bank of America, which ranged between \$5 and \$20 over the same period. Using scale-free measures potentially provides a more equitable way of comparing performance across varying scales in a large data set.

Some common examples of scale-free measures in other domains are the coefficient of variation and the correlation coefficient, which are used in comparing relative dispersion across different series and association between pairs of data sets, respectively. In both cases, the measures automatically adjust to differences in scales and units.

Among forecast accuracy measures, some commonly used functions are scale-free. Absolute percentage error (APE), the building block of MAPE, as well as RE are examples of scale-free functions, which both belong to a larger family of scale-free functions. Let ϕ be a strictly monotonic and nonnegative (or nonpositive) function on $\mathbb{R}^+$. One large family of scale-free functions is the generalized absolute percentage error (GAPE) family given by

$$S_\phi^{\mathrm{APE}}(r, y) = \left| 1 - \frac{\phi(y)}{\phi(r)} \right|. \tag{1}$$

This generalizes the *power-transformed absolute percentage error function* S_β^{APE} provided by Gneiting (2011b),

where $\phi(z) = z^\beta$. S_β^{APE} yields APE when $\beta = -1$, while $\beta = 1$ generates RE. In the case of $\phi(z) = e^z$, $S_\phi^{\text{APE}}(r, y) = |1 - \exp(y - r)|$ yields a variant of the exponential loss function.

A second large family of scale-free measures is the generalized squared percentage error (GSPE) family given by

$$S_\phi^{\text{SPE}}(r, y) = \left(1 - \frac{\phi(y)}{\phi(r)}\right)^2. \tag{2}$$

Similarly, under a power transformation (i.e., when $\phi(z) = z^\beta$), S_β^{SPE} becomes the power-transformed squared percentage error function that contains the squared percentage error (SPE) measure and squared relative error (SRE) when $\beta = -1$ and $\beta = 1$, respectively (Park and Stefanski 1998, Khoshgoftaar et al. 1992).

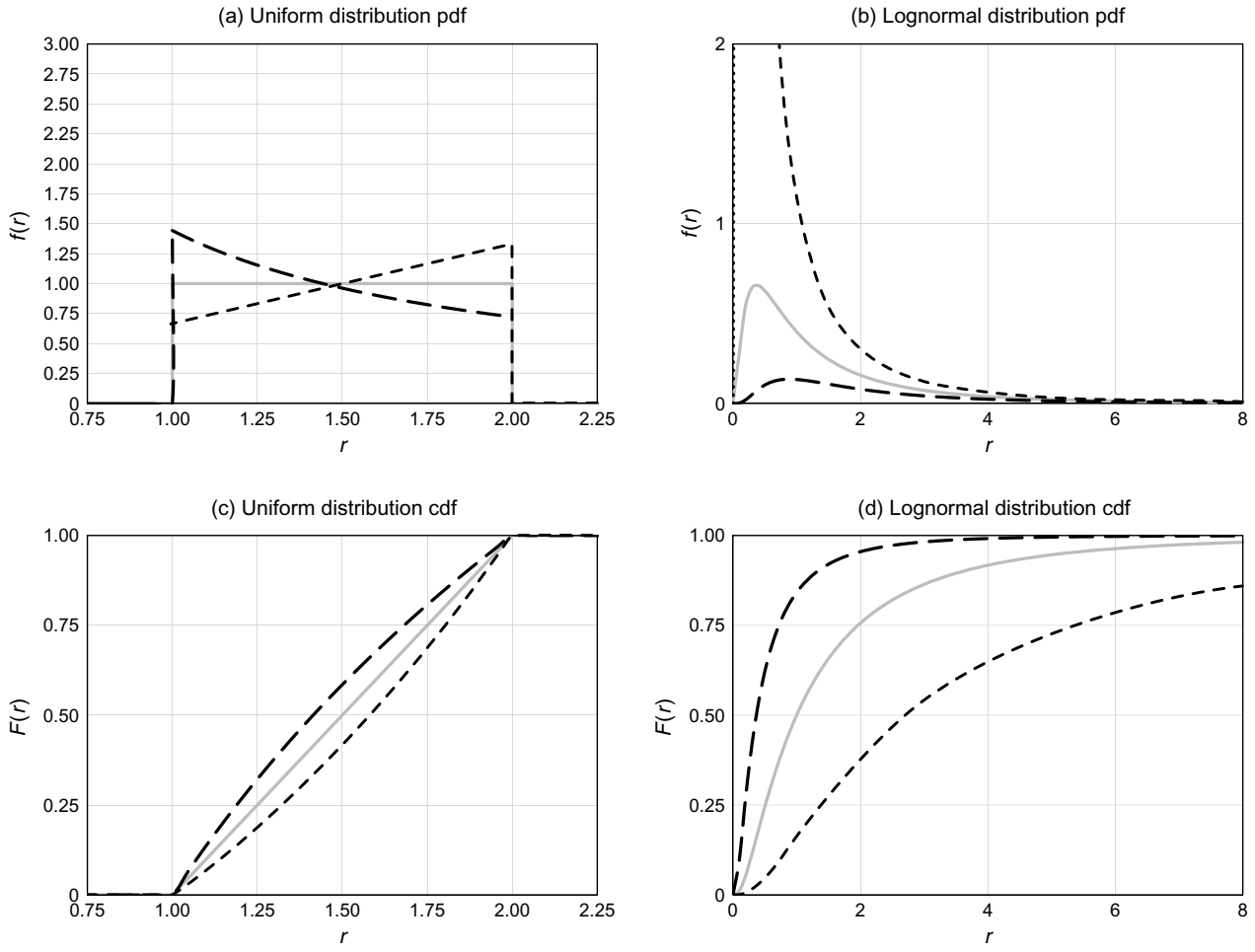
Unlike MSE or MAD, these families are neither consistent with the mean nor with the median. Therefore, reporting the mean or median does not yield the best expected score for the purpose of forecast evaluation. Our first two propositions explain what statistical reports are consistent with these two families.

Proposition 1. *Let ϕ be a strictly monotonic and nonnegative (or nonpositive) function on $\mathbb{R}^+$. Moreover, let f be a continuously differentiable probability density function (pdf) for Y with associated cdf F defined on the same space as ϕ with finite ϕ-moment ($\mathbb{E}_F[\phi(Y)] < \infty$). The consistent statistic that minimizes the expected value of S_ϕ^{APE} is the median of the density proportional to $\phi(y)f(y)$, which we refer to as the median of the ϕ-adjusted density of Y.*

To minimize the expected score under a GAPE function, the optimal report needs to be the median of a renormalized density whose weight is determined by the transformation function ϕ. Due to the GAPE's similarity to MAD, the resulting use of a median estimator is not completely surprising. The need for the renormalization however is not ex ante straightforward.

To explain the need for this weighting function, consider the APE ($\phi(z) = 1/z$) and RE ($\phi(z) = z$) functions when the random variable Y is uniformly distributed on the interval $[1, 2]$, i.e., $Y \sim U[1, 2]$. The ϕ-adjusted density for the APE function places more weight on the

Figure 1. Renormalized densities for the uniform (panels (a) and (c)) and lognormal distributions (panels (b) and (d)) using the APE (long-dash) and RE (short-dash) functions.



lower end of the interval, while less weight is placed on this same region for the RE function. The resulting pdf and cdf are provided in Figure 1 panels (a) and (c), respectively. For APE, the cdf of the uniform distribution given by $F(r) = r - 1$ is transformed by the function $\hat{\phi}(F) = \ln(F+1)/\ln(2)$. This renormalization is equivalent to converting the variables into "percentages," since a log-log transformation implies that a 1% change in r yields a 1% change in $y = F(r)$. It is clear from the cdf of these renormalized distributions that a gap exists between the median of these two distribution functions when a straight line at $F(r) = 1/2$ is drawn because the revised cdf $\hat{\phi}(F)$ is (first-order) stochastically dominated by the original cdf $F(r)$. For the uniform distribution on the interval $[1, 2]$ (with a mean/median of 1.5), the optimal report for the APE is 1.414.

Similarly, the renormalization for the RE function transforms the cdf by $\check{\phi}(F) = ((F+1)^2 - 1)/3$, which is a reflective image of $\hat{\phi}(F)$ using the uniform distribution's cdf as the line of reflection. This yields an optimal report of 1.581 for RE. This change to percentages is slightly more involved when dealing with other distributions. An asymmetric example is provided in Figure 1 panels (b) and (d) using a lognormal distribution with $\mu = 0$ and $\sigma = 1$.

The next proposition provides the optimal report for the GSPE family of scale-free measures.

Proposition 2. *Let ϕ be a continuous, nonnegative, and strictly monotonic function on $\mathbb{R}^+$ and f be a continuously differentiable pdf for Y with associated cdf F defined on the same space as ϕ with finite first two ϕ-moments ($\mathbb{E}_F[\phi(Y)] < \infty$ and $\mathbb{E}_F[\phi(Y)^2] < \infty$). If the GSPE scoring rule S_ϕ^{SPE} is used to measure forecast accuracy for realizations of Y, then ϕ-adjusted moment ratio functional given by*

$$R_\phi(F) = \phi^{-1}\left(\frac{\mathbb{E}_F[\phi(Y)^2]}{\mathbb{E}_F[\phi(Y)]}\right) \tag{3}$$

minimizes the expected value of S_β^{SPE} with respect to the distribution F.

The (generalized) moment ratio functional is not commonly used in the decision analysis literature, but variants of such have appeared in the reliability testing literature. Consider the simplest case when $\phi(z) = z$, the function $r(F)$ reduces to the ratio of the noncentral second moment to the mean of the distribution, which behaves similarly to the coefficient of dispersion (also known as the variance-to-mean ratio or index of dispersion). The coefficient of dispersion is a popular measure in risk analysis that quantifies how much points are dispersed or clustered relative to the center of the distribution. By the normalization through the mean, this measure allows for better comparison of volatility across different data sets.

The moment ratio functional provides a significantly different report compared to the median of the ϕ-adjusted density corresponding to the GAPE family. There, however, are some similarities between the two families. For example, the expected score under GAPE and GSPE are bounded above by the same constant as the next proposition shows.

Corollary 3. *Let $\mathbb{E}_F\phi(Y) < \infty$ and F have a continuously differentiable density function for Y.*

(i) *If ϕ is a positive increasing or negative decreasing function*

$$\min \mathbb{E}_F[S_\phi^{\mathrm{APE}}(r, y)] = \mathbb{E}_F[S_\phi^{\mathrm{APE}}(r^*, y)] = 2F(\mathrm{med}^{(\phi)}F) - 1 \leqslant 1,$$

while if ϕ is a positive decreasing or negative increasing function

$$\min \mathbb{E}_F[S_\phi^{\mathrm{APE}}(r, y)] = \mathbb{E}_F[S_\phi^{\mathrm{APE}}(r^*, y)] = 1 - 2F(\mathrm{med}^{(\phi)}F) \leqslant 1,$$

where $r^* = \arg\min_r \mathbb{E}_F[S_\phi^{\mathrm{APE}}(r, y)]$.

(ii) *Let F be a distribution for Y in the positive half-line with finite first two ϕ moments.*

$$\min \mathbb{E}_F[S_\phi^{\mathrm{SPE}}(r, y)] = \mathbb{E}_F[S_\phi^{\mathrm{SPE}}(r^*, y)] = 1 - \frac{\mathbb{E}_F[\phi(Y)]^2}{\mathbb{E}_F[\phi(Y)^2]} \leqslant 1,$$

where $r^* = \arg\min_r \mathbb{E}_F[S_\phi^{\mathrm{SPE}}(r, y)]$.

This corollary to Propositions 1 and 2 provides analytic expressions for the expected score under truthful reporting for S_ϕ^{APE} and S_ϕ^{SPE}. In both cases, regardless of the choice of ϕ, these minimum expected scores are bounded above by 100%, which implies that in the long run we expect truthful reporting to lead to an average score less than 100% for both scores. Any scores that exceed this over a long period suggests that individuals and models may not necessarily be taking into account the nature of the scoring function used. Naturally, other factors could also play a role such as biases and outliers.

In terms of the optimal report, it is often difficult to analytically express this quantity for both the GAPE and GSPE families. However, under the general power transformation $\phi(z) = z^\beta$, we are able to derive the optimal Bayes' act corresponding to S_β^{APE} and S_β^{SPE} for several well-known probability distributions. Table 1 provides these expressions using the parametrization used by Johnson et al. (1994) for these families of distributions. These expressions can easily be numerically estimated using standard statistical packages such as Mathematica and R. We note that for some of these distributions the optimal Bayes' act does not exist because $\mathbb{E}_F[Y^\beta]$ is undefined.

To gain some additional insights about these estimators, consider the following example: A forecaster believes that the price of an asset at some future period

t follows a lognormal random variable P_t with parameters μ and σ, i.e., $\ln(P_t) \sim N(\mu, \sigma^2)$. This distribution then has a mean and median of $\exp(\mu + \sigma^2/2)$ and $\exp(\mu)$, respectively. If the forecaster wishes to minimize her expected loss under a GAPE measure with a power generator $\phi(z) = z^\beta$, it is in her best interest to report $r^* = \exp(\mu + \beta\sigma^2)$. When $\beta < 0$, r^* is less than the median with a percentile rank of $F_{P_t}(r^*) = (1/2)[1 + \operatorname{erf}(\beta\sigma/\sqrt{2})] < F_{P_t}(\exp(\mu)) = 0.5$. Conversely, $\beta > 0$ yields an estimate for r^* greater than the median.

To illustrate how far from the median these estimates can be, consider the specific example where $\mu = 0$ and $\sigma = 1$. In this case, the mean and median are 1.649 and 1, respectively. For APE, the optimal report is 0.368, which corresponds to the 15.8th percentile of this distribution, while using RE, the optimal report is 2.718 (percentile = 0.841). Similarly when SPE and SRE are used, the optimal reports yield more extreme optimal reports of 0.223 (percentile = 0.067) and 4.482 (percentile = 0.933), respectively.

This example suggests that GAPE and GSPE incentivize forecasters to either underreport and overreport point forecasts relative to some measure of central tendency such as the mean or median. The same holds true for all distributions presented in Table 1. From a behavioral standpoint, this example illustrates that forecasters who are pessimistic or optimistic in reporting point forecasts relative to the information are likely to score better on average. The next result formalizes this observation.

Proposition 4. *Let $\mathbb{E}_F\phi(Y) < \infty$ and f be a continuously differentiable density function for Y.*

(i) *If ϕ is a positive increasing or negative decreasing function, then $\operatorname{med}^{(\phi)}F \geqslant (\leqslant)\operatorname{med}F$. Similarly, if ϕ is a positive decreasing or negative increasing function, then $\operatorname{med}^{(\phi)}F \leqslant \operatorname{med}F$.*

(ii) *If ϕ is a positive, strictly monotonic, and convex function then $R_\phi(F) \geqslant \mathbb{E}_F Y$. Similarly, if ϕ is a negative, strictly monotonic, and concave function, then $R_\phi(F) \leqslant \mathbb{E}_F Y$.*

Proposition 4 provides a comparison between these new statistics to standard measures of central tendency. For S_ϕ^{APE}, the optimal report is at least as large as than the median of the distribution when ϕ is either a positive increasing or negative decreasing function. It is less than the median in the case where ϕ is either a positive decreasing or negative increasing function. For APE, the generator function ϕ for GAPE is a positive decreasing function that yields estimates that will never exceed the median. For RE, its generator is a positive increasing function, which results in optimal reports that cannot be lower than the median. With respect to S_ϕ^{SPE}, comparisons are made with respect to the mean instead of the median or some other rank order statistic.

To explain this incentive for underforecasting and overforecasting, we can consider the traditional measures such as MSE and MAD. With these traditional measures, the penalty for being under or over by a fixed amount is the same, i.e., a forecast being one unit above the realized value will yield the same score as a forecast one unit below the realization. In the case of APE for example, the symmetric penalty happens with respect to each percentage point. However one percent above a value is not the same as one percent below the same value in real terms. This distortion yields an incentive for pushing the optimal estimate either above or below some reference point. A similar effect can be seen with asymmetric linear loss/check function that yields quantiles (Jose and Winkler 2009). The asymmetric penalty in real terms yields estimates either above or below the median depending on the relative penalty for being over or below the realization.

This discussion of deviations from the mean and median begs the question whether or not scale-free measures that are consistent with the mean and median exist. The next proposition answers this question.

Proposition 5. *Let Y be a nonnegative random variable defined on $\mathbb{R}^+$ with finite first moment.*

(i) *The log-transformed quantile (LogQuantile) or log-transformed piecewise linear scoring rule*

$$S_\alpha^{PL}(r, y) = (\mathbf{1}(r \geqslant y) - \alpha)\log\left(\frac{r}{y}\right) \tag{4}$$

is a scale-free, consistent scoring function for the α-quantile ($\alpha \in (0,1)$) of Y.

(ii) *The log-transformed Bregman (LogBregman) scoring rule given by*

$$S^{Br}(r, y) = -\log\left(\frac{y}{r}\right) - 1 + \frac{y}{r} \tag{5}$$

is a scale-free, consistent scoring function for the mean of Y.

Proposition 5 provides two scoring functions that are scale-free and consistent with the two most popular measures of central tendency. The LogQuantile scoring rule in Equation (4) is well known in the quantile regression literature (Koenker 2005). It is consistent with the median when $\alpha = 1/2$. Jose and Winkler (2009) discuss its scale-free nature and show that among all scoring functions consistent with quantiles, this rule is the only form that has the forecast-to-outcome ratio, r/y, as a sufficient statistic.

On the other hand, the LogBregman score is not as well known, but has been developed in the information theory and engineering literatures (e.g., Itakura and Saito 1968, Fevotte et al. 2009). Its scale-free nature and basic properties have yet to be studied extensively.

Table 1. Table of Bayes estimate for APE and SPE measures under the power transformation $\phi(z) = z^\beta$.

Distribution	Domain	Parameters	Bayes estimate for $S_\beta^{\mathrm{APE}}, \beta \in \mathbb{R}$		Bayes estimate for $S_\beta^{\mathrm{SPE}}, \beta \in \mathbb{R}$	
Uniform	$[a,b]$	$0 \leqslant a < b$	$\left[\frac{1}{2}(b^{1+\beta} + a^{1+\beta})\right]^{1/(\beta+1)}$	[DNE if $\beta \leqslant -1$ when $a = 0$]	$\left[\frac{(b^{2\beta+1} - a^{2\beta+1})(\beta+1)}{(b^{\beta+1} - a^{\beta+1})(2\beta+1)}\right]^{1/\beta}$	[DNE if $\beta \leqslant -1$ when $a = 0$]
Beta	$[0,1]$	$a, b > 0$	$I^{-1}\left(a+\beta, b, \frac{1}{2}\right)$	[DNE if $\beta \leqslant -a$]	$\left[\frac{\Gamma(a+b+\beta)\Gamma(a+2\beta)}{\Gamma(a+\beta)\Gamma(a+b+2\beta)}\right]^{1/\beta}$	[DNE if $a \leqslant -2\beta$]
Special cases: Uniform ($a = b = 1$) Power ($b = 1$) Arcsine ($a = b = \frac{1}{2}$)						
Gamma	$[0,\infty)$	$a, b > 0$	$bQ^{-1}\left(a+\beta, \frac{1}{2}\right)$	[DNE if $\beta \leqslant -a$]	$b\left[\frac{\Gamma(a+2\beta)}{\Gamma(a+\beta)}\right]^{1/\beta}$	[DNE if $a \leqslant -2\beta$]
Special cases: Exponential ($a = 1$) χ^2 ($a = v/2, b = 2$) Erlang ($a \in \mathbb{Z}^+$)						
Rayleigh	$[0,\infty)$	$\sigma > 0$	$\sigma\sqrt{2Q^{-1}\left(1+\frac{\beta}{2}, \frac{1}{2}\right)}$	[DNE if $\beta \leqslant -2$]	$\sqrt{2}\sigma\left[\frac{\Gamma(1+\beta)}{\Gamma(1+\beta/2)}\right]^{1/\beta}$	[DNE if $\beta \leqslant -1$]
Weibull	$[0,\infty)$	$a, b > 0$	$b\left[Q^{-1}\left(1+\frac{\beta}{a}, \frac{1}{2}\right)\right]^{1/a}$	[DNE if $\beta \leqslant -a$]	$b\left[\frac{\Gamma(1+2\beta/a)}{\Gamma(1+\beta/a)}\right]^{1/\beta}$	[DNE if $a \leqslant -2\beta$]
Maxwell	$[0,\infty)$	$\sigma > 0$	$\sigma\sqrt{2Q^{-1}\left(\frac{\beta+3}{2}, \frac{1}{2}\right)}$	[DNE if $\beta \leqslant -3$]	$\sqrt{2}\sigma\left[\frac{\Gamma(\beta+3/2)}{\Gamma(\beta/2+3/2)}\right]^{1/\beta}$	$\left[\text{DNE if } \beta \leqslant -\frac{3}{2}\right]$
Pareto	$[k,\infty)$	$a, k > 0$	$k2^{-1/(\beta-a)}$	[DNE if $\beta \geqslant a$]	$k\left(\frac{a-\beta}{a-2\beta}\right)^{1/\beta}$	[DNE if $a \leqslant 2\beta$]
LogNormal	$[0,\infty)$	$\mu \in \mathbb{R}, \sigma > 0$	$\exp(\mu + \beta\sigma^2)$		$\left[\exp\left(\beta\mu + \frac{3\beta^2\sigma^2}{2}\right)\right]^{1/\beta}$	
Half-normal	$[0,\infty)$	$\sigma > 0$	$\frac{1}{\sigma}\sqrt{\pi Q^{-1}\left(\frac{1+\beta}{2}, \frac{1}{2}\right)}$	[DNE if $\beta \leqslant -1$]	$\frac{\sqrt{\pi}}{\sigma}\left[\frac{\Gamma(\beta+1/2)}{\Gamma(\beta/2+1/2)}\right]^{1/\beta}$	$\left[\text{DNE if } \beta \leqslant -\frac{1}{2}\right]$

Notes. (1) I^{-1} is the inverse of the regularized incomplete beta function, i.e., $I(z,a,b) = B(z,a,b)/B(a,b) = p \Leftrightarrow z = I^{-1}(a,b,p)$. (2) Q^{-1} is the inverse of the regularized incomplete gamma function, i.e., $Q(a,z) = \Gamma(a,z)/\Gamma(a) = p \Leftrightarrow z = Q^{-1}(a,p)$. (3) DNE = Does not exist.

In practice, these rules are nowhere close to the popularity of the GAPE and GSPE families. Apart from familiarity with these rules, a significant challenge in implementing the LogQuantile and LogBregman rules is finding easy and intuitive ways of explaining such rules to end users and practitioners.

3. Illustration

In this section, we illustrate some of the insights from the previous section on real-world data. In particular, we consider a forecast aggregation setting where scoring rules are often used as a forecast evaluation measure in model selection. The purpose of this empirical investigation is not to show or argue what the best aggregation approach is, rather it is to illustrate how certain methods may be inadvertently preferred when using certain scoring functions. These should highlight the importance of carefully choosing a set of scoring functions and interpreting each of these in practice.

3.1. Combination of Forecasts

When only a single forecast is available, the logical choice is to use this estimate in planning and decision making. However, when more than one forecast is available, a typical approach is to aggregate estimates into a single consensus forecast.

One way to think about the expert forecasts being aggregated is that each one represents a signal from some common data generating process, which might be restrictive in some settings but is a typical assumption in the literature. If we additionally assume that each signal is unbiased and independent, then the optimal way to aggregate these forecasts that minimizes the expected loss is to take the appropriate optimal Bayes' act for the predictive distribution of the unknown quantity using the forecasts collected from experts. Some of these assumptions may not likely hold in many settings but for now, we assume these and later on explain possible ways to handle violations of these assumptions.

To conduct a "horse race" among methods to illustrate model selection, we consider several aggregation mechanisms. There are numerous ways of combining point forecasts and a large body of empirical work has shown that combined forecasts outperform individual forecasters on average and that simpler ways of aggregating forecasts such as using a simple average tend to outperform more sophisticated approaches (Clemen and Winkler 1986, Genest and Zidek 1986, Clemen 1989).

Some improvements to the simple average could be achieved through moderate symmetric trimming (10% on each side) or Winsorizing (15% on each side) by reducing the mean's sensitivity to outliers (Jose and Winkler 2008, Jose et al. 2014). These simple heuristics provide simple, yet effective ways of combining point estimates into a single consensus forecast. To compare rankings between different scoring functions, we shall use these different heuristics as benchmarks, together with the new statistics discussed in the previous section.

In particular, we can compute the finite consistent sample statistics associated with the optimal Bayes' act for the GAPE and GSPE family. For any set of forecasts $\mathbf{r} = \{r_1, \ldots, r_k\}$, we can create a revised empirical density that assigns to the point r_i probability $\phi(r_i)/\sum_{j=1}^{k} \phi(r_j)$ for $i = 1, \ldots, k$. Under this revised density, which we call $\hat{F}$, we can obtain the ϕ-adjusted median estimate by computing the sample median for $\hat{F}$ using linear interpolation when needed to obtain a unique estimate (Scholz 1978, Cormen et al. 1989). For the ϕ-adjusted moment ratio, we can similarly compute the needed statistic to generate the estimate, namely, the first two noncentral moments for $\hat{F}$. For this example, we compute the scores for these new statistics using the power transformation with $\beta = -2, -1, 1,$ and 2. Based on the properties that we have discussed earlier, these new alternative measures will lead to estimates that are more extreme relative to the sample mean or median.

3.2. Data

Thomson Reuters Institutional Brokers' Estimate System (I/B/E/S) is a comprehensive database that provides analysts' forecasts for various economic and financial variables of interest, such as earnings per share, price targets, net income, and operating profit for a large number of firms both inside and outside the United States. For this study, we focus our attention on price targets, which refer to an analyst's projected price level for a stock in a future period. This data set compared to others in the I/B/E/S database is the largest data set with nonnegative values.

In our analysis, we examine forecasts for price targets of stocks that comprise the S&P 500 index aggregated on a monthly basis for the 10-year period between January 2003 and December 2012. Since the composition of this index changes over time, we use the stocks that comprise the index on the last trading day of 2012. To remove ambiguity on whether analysts had the latest information about price adjustments, we only consider forecasts in this data set prior to any stock/reverse stock split within the said time frame. In addition, we only consider nonzero forecasts that have a horizon of at least three months. This yields more than 155,548 estimates for 28,891 unique firm months, with 81.17% of these firm months containing more than one price target/forecast.

Since there is no distinction between all of these aggregation schemes when there is only one forecast, we consider only firm months where there are at least three forecasts. For our data set, this corresponds to

Table 2. Performance accuracy measures for several simple aggregation methods using the I/B/E/S data set.

					ϕ-weighted median, $\phi(z) = z^\beta$				ϕ-weighted moment ratio, $\phi(z) = z^\beta$			
	Mean	Median	Trimmed mean	Winsorized mean	($\beta = -2$)	($\beta = -1$)	($\beta = 1$)	($\beta = 2$)	($\beta = -2$)	($\beta = -1$)	($\beta = 1$)	($\beta = 2$)
MAPE	0.581	0.504	0.559	0.537	0.483	0.490	0.869	0.917	0.473	0.476	0.787	0.876
MSPE	1.437	0.878	1.250	1.109	0.478	0.534	6.446	7.210	0.442	0.464	5.182	6.656
RE	0.553	0.750	0.561	0.608	3.356	2.955	0.451	0.440	3.380	2.905	0.463	0.448
MSRE	0.882	2.781	0.936	1.329	115.2	93.48	0.562	0.521	114.5	85.10	0.608	0.561
Other measures												
MSE	1,904.5	1,217.9	1,700.9	1,563.9	912.2	948.2	8,350.9	8,913.1	898.3	899.5	6,521.5	8,410.1
MAD	22.59	19.73	21.90	21.14	20.39	20.41	32.31	33.61	20.29	20.21	29.49	32.46
RMSE	43.64	34.90	41.24	39.55	30.20	30.79	91.38	94.73	29.97	29.99	80.76	91.71
LogBregman	0.256	0.404	0.258	0.293	2.689	2.340	0.247	0.246	2.699	2.270	0.243	0.247
LogQuantile	0.242	0.253	0.241	0.243	0.386	0.366	0.244	0.247	0.388	0.370	0.239	0.244

Table 3. Rankings for several simple aggregation methods using the I/B/E/S data set.

					ϕ-weighted median, $\phi(z) = z^\beta$				ϕ-weighted moment ratio, $\phi(z) = z^\beta$			
	Mean	Median	Trimmed mean	Winsorized mean	($\beta = -2$)	($\beta = -1$)	($\beta = 1$)	($\beta = 2$)	($\beta = -2$)	($\beta = -1$)	($\beta = 1$)	($\beta = 2$)
MAPE	8	5	7	6	3	4	10	12	1	2	9	11
MSPE	8	5	7	6	3	4	10	12	1	2	9	11
RE	5	8	6	7	12	10	3	1	11	9	4	2
MSRE	5	8	6	7	12	10	3	1	11	9	4	2
Other measures												
MSE	8	5	7	6	3	4	10	12	1	2	9	11
MAD	8	1	7	6	4	5	10	12	3	2	9	11
RMSE	8	5	7	6	3	4	10	12	1	2	9	11
LogBregman	5	8	6	7	11	10	4	2	12	9	1	3
LogQuantile	3	8	2	4	11	9	6	7	12	10	1	5

n = 18,903 firm months with at least three forecasts (average = 7.46, median = 6, std dev = 5.38 forecasts per firm month).

3.3. Empirical Results and Discussion

To examine the behavior of these different scoring functions, we examine the performance of several simple heuristics (mean, median, trimmed mean, and Winsorized mean) for aggregating forecasts together with the new consistent statistics (med$^{(\phi)}(F)$ and $R_\phi(F)$). For the trimmed and Winsorized means, we follow the 20% and 30% suggestion for the trimming and Winsorizing found by Jose and Winkler (2008) in the M-3 forecasting competition. If the assumptions of independence and a common likelihood hold, these new estimators should perform well with the scoring/loss functions they are consistent with. However, such assumptions are often violated in practice, because of behavioral anomalies that enter the forecast generation process for each expert.

We begin by examining the average scores for a set of forecasted values $\mathbf{r}$ using the different simple heuristics as well as the sample power-weighted median (med$^\beta\{\mathbf{r}\}$) and power-weighted moment ratio ($RF^\beta\{\mathbf{r}\}$) provided in Table 2. We provide several scale-free measures (MAPE, MSPE, RE, MSRE, LogBregman, and LogQuantile) as well as some conventional non-scale-free measures (MSE, MAD, and root mean squared error RMSE = $\sqrt{\text{MSE}}$). Table 3 provides the rankings of each of these simple aggregation scheme relative to the 12 different methods examined.

Consistent with the literature, the trimmed and Winsorized mean ranked better than the mean across most measures. The mean however did not perform as well as the median. This is likely due to the varying scale for different stocks in this data set. In particular, larger values are likely to dominate the averages for both the MSE and RMSE. To a lesser extent, it still plays a significant role as well in the MAD when the discrepancy between the overall values are large (e.g., stocks less than \$1 versus stocks valued over \$500).

For med$^\beta(F)$ and $R_\beta(F)$, we see that for $\beta < 0$, med$^\beta\{\mathbf{r}\}$ outperforms the simple measures with respect to the MAPE and MSPE. For $\beta > 0$, this approach provides a significantly worse estimate with respect to the MAPE and MSPE, however this performs significantly better than most other approaches when either RE or MSRE is used. A similar pattern is seen with the $RF^\beta\{\mathbf{r}\}$.

These results suggest in part that these measures reward underestimation and overestimation relative to the mean or median. It is interesting to note that for

the MSE and MAD, despite the fact that the mean and median are consistent estimators, the $\text{med}^{\beta}(F)$ and $R_{\beta}(F)$ both with $\beta < 0$ ranked slightly better than the mean and median under MSE and MAD.

One possible explanation for this can be attributed to the general quality of estimates provided by financial analysts. In particular, the issue of unbiasedness and independence of forecasts may not necessarily hold. Research has indicated that analysts tend to have an optimism bias when it comes to reporting firm performance (de Bondt and Thaler 1990, Easterwood and Nutt 1999, Stotz and von Nitzsch 2005). This leads to biased and heavily correlated forecasts. The systematic upward bias for each individual forecast allows the aggregated underforecast to perform better.

This illustration highlights that aggregation methods that always yield lower estimates relative to the mean or median generally exhibit better scores using percentage error-based measures, but not for relative error-based forecast accuracy measures. Though this is not a sufficient condition to prove whether a downward (or pessimism) bias is naturally inherent in a forecasting system or aggregation scheme, it could potentially be used as a tool for detecting such type of a bias. A similar strategy can be used for detecting an upward (or optimism) bias in forecasting methods that always perform well on relative error-based measures, but not necessarily on percentage error-based measures.

Ideally, if this bias can be accurately measured and detected then aggregate forecasts could easily be adjusted. Unfortunately, this often entails significant work and effort. Several ways to combat this bias without directly computing for the expected bias are increasing the diversity of the pool of forecasters, improving the information content of forecasts, or reducing psychological anchors (e.g., debiasing using time unpacking similar to Jain et al. 2013).

With respect to the scale-free measures that are theoretically consistent with the mean and median, the empirical results show that the LogBregman and LogQuantile exhibit patterns that are closer to the RE and MSRE measures rather than MAPE and MSPE. They would rank $\text{med}^{\beta}(F)$ and $R_{\beta}(F)$ with $\beta > 0$ higher than the same statistics with $\beta < 0$. This is surprising since the two other measures consistent with the mean and median (MSE and MAD) would rank these methods in the opposite direction. This result implies that the rescaling of points into "percentages" through the log-log transformation of both variables reversed the overall pattern. For large values, variability in terms of percentages would be significantly smaller than the variability in percentages for smaller values, due to the flattening out of the logarithmic function. Such empirical results suggest that even beyond consistency, the issue of scale is something that must be considered when dealing with different scoring functions.

4. Conclusion

Scale-free forecasting accuracy measures are among the most commonly used tools in forecasting practice. These functions have proven to be useful when comparing aggregate forecasting performance across different quantities, time horizons, and scales.

In this paper, we explore properties of the GAPE and GSPE families, two large classes of scale-free forecasting accuracy measures that include well-known measures such as APE, RE, SPE, and SRE. In particular, we examine statistics that are consistent with these scores. We then derive closed-form analytic expressions for the optimal Bayes' act under several popular distribution functions for the power-transformed members of these two families. One important finding is that these scale-free families provide incentives for underreporting and overreporting relative to certain measures of central tendency such as the mean and median.

These results imply that forecasters that are optimistic or pessimistic may be inadvertently rewarded in performance when using scores from these two families. This suggests the need for additional care when interpreting results using these rules. As theory suggests, aggregation mechanisms that skew forecasts higher (lower) than the median performed better with respect to APE (RE) and other similarly transformed GAPE functions. A similar expected pattern emerges with GSPE measures and the mean.

In terms of overreporting or underreporting with respect to the mean and median, we prove the existence of scoring functions that are consistent with these statistics. These rules however are harder to interpret compared to a percentage or relative error measure. The characterization of all scale-free rules consistent with these statistics still remains a challenging open question.

Some recent work has also recognized other potential issues with APE- and RE-based measures. For example, Davydenko and Fildes (2013) point that these measures when averaged may provide poor insights in forecast evaluation especially when a few observations have outcomes or forecasts that are extremely close to 0. In their paper, they recommended the use of the average relative mean absolute error, which is the geometric mean of the ratio between the mean absolute error of the forecast to the mean absolute error of a baseline statistical forecast. Other papers such as Hyndman and Koehler (2006) have also proposed other scale-free measures such as the mean scaled error, which is the difference between the forecast and the observed outcome normalized by the average in-sample mean absolute error generated from a naïve/random walk forecasting method. These are potentially promising in practice. It is not clear, however, what types of statistics these new measures are

consistent with and what type of incentives they may directly or indirectly provide.

When dealing with probabilistic forecasts, scales often do not play a role, since probabilities are unitless by nature. However, forecasts associated with probabilities such as quantiles may be still affected by scale. In this setting, we can use a scale-free scoring function that is consistent with quantiles such as the LogQuantile scoring rule.

Finally, we comment on the relevance of these results and on the use of these percentage and relative error-based forecasting accuracy measures in practice. First, we reiterate the importance of Fildes and Goodwin's advice of using multiple forecasting accuracy measures and using measures that adjust to scale. Next, when forecast rankings are not consistent between measures then one possible explanation for these discrepancies may arise from the different properties we have highlighted in this paper. Though some of these theoretical results have been demonstrated in empirical papers, these issues may not be well known to practicing forecasters and econometricians.

When a forecasting or aggregation method consistently performs well on percentage error-based accuracy measures but not on relative error-based measures over many different data sets and series, this may suggest to evaluators that the method may systematically generate downward (or pessimism) bias. Alternatively, it may be an optimism or upward bias in the opposite case. This may provide a signal and explanation to forecast evaluators about the quality of the reported or aggregate forecasts provided by a method. Finally, we note that though these scores may inadvertently reward underforecasts or overforecasts relative to the mean or median, the ease of interpretation for these measures is hard to beat, so these measures will likely continue being useful and popular in practice.

Acknowledgments

The author is grateful to the area editor, associate editor, and three referees for their many constructive comments.

Appendix

Proofs

Proof of Proposition 1. The expected value function $\mathbb{E}_F S_\phi^{\mathrm{APE}}$ can be written as

$$\begin{aligned}
&\mathbb{E}_F S_\phi^{\mathrm{APE}}(r,y)\\
&= \int_0^\infty \left|1-\frac{\phi(y)}{\phi(r)}\right| f(y)\,dy\\
&= \int_0^r K\left[1-\frac{\phi(y)}{\phi(r)}\right] f(y)\,dy + \int_r^\infty K\left[\frac{\phi(y)}{\phi(r)}-1\right] f(y)\,dy\\
&= K\left[\left(2\int_0^r f(y)\,dy-1\right) - \int_0^r \frac{\phi(y)}{\phi(r)} f(y)\,dy + \int_r^\infty \frac{\phi(y)}{\phi(r)} f(y)\,dy\right],
\end{aligned} \tag{6}$$

where

$$K=\begin{cases} 1 & \text{if } \phi \text{ is a positive increasing or negative decreasing function}\\ -1 & \text{if } \phi \text{ is a positive decreasing or negative increasing function}\end{cases}. \tag{7}$$

Taking the first-order condition (FOC) via Liebnitz's rule, we have

$$\begin{aligned}
0 &= K\left[2f(r^*) - \left\{\frac{\phi'(r^*)}{\phi(r^*)^2}\int_0^{r^*}\phi(y)f(y)\,dy + f(r^*)\right\}\right.\\
&\qquad \left. + \left\{\frac{\phi'(r^*)}{\phi(r^*)^2}\int_{r^*}^{\infty}\phi(y)f(y)\,dy - f(r^*)\right\}\right]\\
0 &= \int_0^{r^*}\phi(y)f(y)\,dy - \int_{r^*}^{\infty}\phi(y)f(y)\,dy.
\end{aligned}$$

Normalizing both sides by $(\mathbb{E}_F\phi(Y))^{-1}$, it follows that r^* is the median of the renormalized density $f_\phi(y) = \phi(y)f(y)/\mathbb{E}_F\phi(Y)$. Evaluating the second derivative at r^* leads to

$$\begin{aligned}
&\frac{\partial^2 \mathbb{E}_F S_\phi^{\mathrm{APE}}(r,y)}{\partial r^2}\Big|_{r=r^*}\\
&\quad = K\int_0^{r^*}\left\{\frac{\phi''(r^*)\phi(y)}{\phi(r^*)^2} - 2\frac{\phi'(r^*)^2\phi(y)}{\phi(r^*)^3}\right\} f(y)\,dy\\
&\qquad + K\int_{r^*}^{\infty}\left\{2\frac{\phi'(r^*)^2\phi(y)}{\phi(r^*)^3} - \frac{\phi''(r^*)\phi(y)}{\phi(r^*)^2}\right\} f(y)\,dy\\
&\qquad + \frac{2Kf(r^*)\phi'(r^*)}{\phi(r^*)}\\
&\quad = \frac{2Kf(r^*)\phi'(r^*)}{\phi(r^*)} + K\left[\frac{\phi''(r^*)}{\phi(r^*)^2} - 2\frac{\phi'(r^*)^2}{\phi(r^*)^3}\right]\\
&\qquad \cdot (\mathbb{E}_0^{r^*}[\phi(Y)] - \mathbb{E}_{r^*}^{\infty}[\phi(Y)])\\
&\quad = \frac{2Kf(r^*)\phi'(r^*)}{\phi(r^*)} > 0,
\end{aligned}$$

since the last two partial expectations are both equal to $(\mathbb{E}_F\phi(Y))/2$ from the FOC. This is strictly positive since $K\phi'/\phi > 0$.

Finally, to show that the optimal solution is interior, we note that the first derivative is one-switch, i.e., it decreases (increases) and then increases (decreases) once it reaches r^* since $\phi'(r)(\mathbb{E}_0^{r^*}[\phi(Y)] - \mathbb{E}_{r^*}^{\infty}[\phi(Y)])$ is monotonically decreasing (increasing) for increasing (decreasing) positive-valued and decreasing (increasing) negative-valued functions ϕ and nonnegative random variables Y. □

Proof of Proposition 2. We write the expected score as follows:

$$\begin{aligned}
\mathbb{E}_F S_\phi^{\mathrm{SPE}}(r,y) &= \int_0^\infty \left(1-\frac{\phi(y)}{\phi(r)}\right)^2 f(y)\,dy\\
&= 1 - 2\int_0^\infty \frac{\phi(y)}{\phi(r)} f(y)\,dy + \int_0^\infty \frac{\phi(y)^2}{\phi(r)^2} f(y)\,dy.
\end{aligned} \tag{8}$$

Let $\mu_1^\phi := \mathbb{E}[\phi(Y)]$ and $\mu_2^\phi := \mathbb{E}[\phi(Y)^2]$. Taking the FOC, we have

$$0 = 2\phi'(r)\frac{\mu_1^\phi\phi(r) - \mu_2^\phi}{\phi(r)^3}.$$

Since ϕ is positive and strictly monotonic, the FOC is equal to 0 only when

$$0 = \mu_1^\phi \phi(r) - \mu_2^\phi \quad \Rightarrow \quad r^* = \phi^{-1}\left(\frac{\mu_2^\phi}{\mu_1^\phi}\right) = \phi^{-1}\left(\frac{\mathbb{E}_F[\phi(Y)^2]}{\mathbb{E}_F[\phi(Y)]}\right).$$

We verify that it is a minimum by checking the second derivative as follows:

$$\begin{aligned}
&\left.\frac{\partial^2 \mathbb{E}_F S_\phi^{\mathrm{SPE}}}{\partial r^2}\right|_{r=r^*} \\
&= \frac{6\mu_2^\phi \phi'(r^*)^2}{\phi(r^*)^4} - \frac{4\mu_1^\phi \phi'(r^*)^2}{\phi(r^*)^3} - \frac{2\mu_2^\phi \phi''(r^*)}{\phi(r^*)^3} + \frac{2\mu_1^\phi \phi''(r^*)^2}{\phi(r^*)^2} \\
&= \frac{1}{\phi(r^*)^4}\left[2\mu_2^\phi \phi'(r^*)^2 + 2\phi(r^*)\phi''(r^*)\right. \\
&\qquad \left.\cdot\left(\mu_1^\phi \phi\left(\phi^{-1}\left(\frac{\mu_2^\phi}{\mu_1^\phi}\right)\right) - \mu_2^\phi\right)\right] = \frac{2\mu_2^\phi \phi'(r^*)^2}{\phi(r^*)^4} > 0,
\end{aligned}$$

since $\mu_2^\phi = \mathbb{E}[\phi(Y)^2] > 0$. To verify that this is an interior point, we note that ϕ is continuous and strictly monotonic, which implies that the inverse exists and that the range of the inverse is the same as the domain of ϕ. □

Proof of Corollary 3. (i) Simplification of Equation (6) yields the expression for $\mathbb{E}_F[S_\phi^{\mathrm{APE}}(r^*, y)]$. The bound follows from the fact that $\sup F(r) = 1$. (ii) The first equality follows from using the optimal r^* from Proposition 2 in Equation (8). The inequality comes from the fact that the variance of an r.v. is always nonnegative, i.e., $\mathrm{Var}(\phi(Y)) = \mathbb{E}[\phi(Y)^2] - \mathbb{E}[\phi(Y)]^2 \geqslant 0$. □

Proof of Proposition 4. (i) First, we note that since $|1 - \phi(y)/\phi(r)|f(y) \geqslant 0$ for all r, y then $\mathbb{E}_F[S_\phi(r, y)] \geqslant 0$. Hence, $0 \leqslant K(2F(\mathrm{med}^{(\beta)}F) - 1) \Leftrightarrow KF(\mathrm{med}^{(\beta)}F) \geqslant K\frac{1}{2}$. This means that $\mathrm{med}^{(\beta)}F \geqslant F^{-1}(\frac{1}{2}) = \mathrm{med}\,F$, when $K = 1$ (i.e., ϕ is a positive increasing or negative decreasing function). Similarly, we get $\mathrm{med}^{(\beta)}F \leqslant F^{-1}(\frac{1}{2}) = \mathrm{med}\,F$, when ϕ is a positive decreasing or negative increasing function. (ii) Consider the case when ϕ is a positive monotonic convex function. First, we note that

$$\begin{aligned}
&\mathrm{Var}_F(\phi(Y)) = \mathbb{E}_F[\phi(Y)^2] - \mathbb{E}_F[\phi(Y)]^2 \geqslant 0 \\
&\Rightarrow \quad \frac{\mathbb{E}_F[\phi(Y)^2]}{\mathbb{E}_F[\phi(Y)]} \geqslant \mathbb{E}_F[\phi(Y)]
\end{aligned}$$

since $\mathbb{E}_F[\phi(Y)] > 0$. Next, we note that since ϕ is convex, then by Jensen's inequality $\mathbb{E}_F[\phi(Y)] \geqslant \phi(\mathbb{E}_F[Y])$, which implies that

$$\frac{\mathbb{E}_F[\phi(Y)^2]}{\mathbb{E}_F[\phi(Y)]} \geqslant \phi(\mathbb{E}_F[Y]).$$

Similarly, the case for negative, strictly monotonic, and concave functions will yield $\mathbb{E}_F[\phi(Y)^2]/\mathbb{E}_F[\phi(Y)] \leqslant \phi(\mathbb{E}_F[Y])$. Finally, we get the desired result by taking the inverse of both sides and using the fact that the inverse of a continuous, strictly monotonic function always exists and must also be monotonic. □

Proof of Proposition 5. (i) By Thomson's characterization of quantile scoring rules (Thomson 1979; Gneiting 2011a, b), a proper scoring/loss function for the α quantile must be in the form $(\mathbf{1}(r \geqslant y) - \alpha)(g(r) - g(y))$ for some increasing function g. Set $g(z) = \log \phi(z)$, which is strictly increasing since $g'(z) = \phi'(z)/\phi(z) > 0$. This means that $\phi(r^*) = \mathrm{med}\,\{\phi(Y)\} \Rightarrow r^* = \phi^{-1}(\mathrm{med}\,\{\phi(Y)\})$. Since ϕ^{-1} is a monotonic function, it follows that $\phi^{-1}(\mathrm{med}\,\{\phi(Y)\}) = \mathrm{med}\,\{\phi^{-1}(\phi(Y))\} = \mathrm{med}\,\{Y\}$. From Jose and Winkler (2009), we say that the ratio r/y is a sufficient statistic for this scoring rule, which means that the score is scale-free. (ii) From Savage (1971), any scoring rule consistent with the mean must be of a Bregman divergence form, i.e., $S(r, y) = g(y) - g(r) - \langle \nabla g(r), y - r \rangle$ for some convex function g. Setting $g(z) = \log z$ yields the desired result. Again, this function is scale-free, since the unitless quantity y/r is a sufficient statistic. □

References

Chatfield C (1988) Apples, oranges and mean squared error. *Internat. J. Forecasting* 4(4):515–518.

Chen K, Guo S, Lin Y, Ying Z (2010) Least absolute relative error estimation. *J. Amer. Statist. Assoc.* 105(491):1104–1112.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–583.

Clemen RT, Winkler RL (1986) Combining economic forecasts. *J. Bus. Econom. Statist.* 4(1):39–46.

Cormen TH, Leiserson CE, Rivest RL (1989) *Introduction to Algorithms* (MIT Press, Cambridge, MA).

Davydenko A, Fildes R (2013) Measuring forecast accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *Internat. J. Forecasting* 29(3):510–523.

de Bondt WFM, Thaler RH (1990) Do security analysts overreact? *Amer. Econom. Rev.* 80(2):52–57.

Easterwood JC, Nutt SR (1999) Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism? *J. Finance* 54(5):1777–1797.

Farnum NR (1990) Improving the relative error of estimation. *Amer. Statist.* 44(4):288–289.

Fevotte C, Bertin N, Durrieu JL (2009) Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. *Neural Comp.* 21(3):793–830.

Fildes R, Goodwin P (2007) Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces* 37(6):570–576.

Genest C, Zidek JV (1986) Combining probability distributions: A critique and annotated bibliography. *Statist. Sci.* 1(1):114–135.

Gneiting T (2011a) Quantiles as optimal point forecasts. *Internat. J. Forecasting* 27(2):197–207.

Gneiting T (2011b) Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* 106(494):746–762.

Hyndman RJ, Koehler AB (2006) Another look at forecast accuracy measures. *Internat. J. Forecasting* 22(4):670–688.

Itakura F, Saito S (1968) Analysis synthesis telephony based on the maximum likelihood method. Kohasi Y, ed. *IEEE Proc. 6th Internat. Congress Acoustics* (IEEE, Los Alamitos, CA), 17–20.

Jain K, Mukherjee K, Bearden JN, Gaba A (2013) Unpacking the future: A nudge toward wider subjective confidence intervals. *Management Sci.* 59(9):1970–1987.

Johnson NL, Kotz S, Balakrishnan N (1994) *Continuous Univariate Distributions*, 2nd ed., Vols. I and II (John Wiley & Sons, New York).

Jose VRR, Winkler RL (2008) Simple robust average of forecasts: Some empirical results. *Internat. J. Forecasting* 24(1):163–169.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57(5):1287–1297.

Jose VRR, Grushka-Cockayne Y, Lichtendahl K Jr (2014) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60(2):463–475.

Khoshgoftaar TM, Bhattacharyya BB, Richardson GD (1992) Predicting software errors, during development, using nonlinear regression models: A comparative study. *IEEE Trans. Reliability* 41(3):390–395.

Koenker R (2005) *Quantile Regression* (Cambridge University Press, Cambridge, MA).

McCarthy TM, Davis DF, Golicic SL, Mentzer JT (2006) The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practice *J. Forecasting* 25(5):303–324.
Narula SC, Wellington JF (1977) Prediction, linear regression, and the minimum sum of relative errors. *Technometrics* 19(2):185–190.
Park H, Stefanski LA (1998) Relative-error prediction. *Statist. Probab. Lett.* 40(3):227–236.
Patton AJ (2011) Volatility forecast comparison using imperfect volatility proxies. *J. Econometrics* 160(1):246–256.
Savage LJ (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66(336):783–810.
Scholz FW (1978) Weighted median regression estimates. *Ann. Statist.* 6(3):603–609.
Stotz O, von Nitzsch R (2005) The perception of control and the level of overconfidence: Evidence from analyst earnings estimates and price targets. *J. Behav. Finance* 6(3):121–128.
Thomson W (1979) Eliciting production possibilities from a well-informed manager. *J. Econom. Theory* 20(3):360–380.
Zellner A (1986) A tale of forecasting 1001 series: The Bayesian knight strikes again. *Internat. J. Forecasting* 2(4):491–494.

Victor Richmond R. Jose is an associate professor and the William Charles Sonneborn Term Chair in the operations and information management area of the Robert Emmett McDonough School of Business at Georgetown University. His main research interests lie in decision analysis and the use of statistical methods in management science, operations research, and risk analysis. His recent work has been in the areas of data science and machine learning.