ECONOMETRIC REVIEWS
2017, VOL. 36, NOS. 6–9, 588–598
http://dx.doi.org/10.1080/07474938.2017.1307247

Taylor & Francis
Taylor & Francis Group

Assessing point forecast accuracy by stochastic error distance

Francis X. Diebold[a] and Minchul Shin[b]

[a]Department of Economics, University of Pennsylvania, Philadelphia, Pennsylvania, USA; [b]Department of Economics, University of Illinois, Champaign, Illinois, USA

ABSTRACT
We propose point forecast accuracy measures based directly on distance of the forecast-error c.d.f. from the unit step function at 0 ("stochastic error distance," or *SED*). We provide a precise characterization of the relationship between *SED* and standard predictive loss functions, and we show that all such loss functions can be written as weighted *SED*s. The leading case is absolute error loss. Among other things, this suggests shifting attention away from conditional-mean forecasts and toward conditional-median forecasts.

KEYWORDS
Absolute error loss; forecast accuracy; forecast evaluation; quadratic loss; squared error loss

JEL CLASSIFICATION
C53

1. Introduction

One often wants to rank competing point forecasts by accuracy. Invariably one proceeds by ranking by expected loss, $E(L(e))$, where e is forecast error and the loss function $L(e)$ satisfies $L(0) = 0$ and $L(e) \geq 0$, $\forall e$.[1] Typically, however, little thought is given to the loss function $L(e)$. Instead, Gauss' centuries-old quadratic loss, $L(e) = e^2$, remains routinely invoked, primarily for mathematical convenience.

Against this background, in this paper, we approach the accuracy ranking problem directly, basing rankings on the entire distribution of e. In particular, recognizing that any reasonable loss function must satisfy $L(0) = 0$, we study accuracy measures based directly on the distance between $F(e)$, the c.d.f. of e, and $F^*(e)$, the unit step function at 0,[2]

$$F^*(e) = \begin{cases} 0, & e < 0 \\ 1, & e \geq 0. \end{cases}$$

$F^*(e)$ is the error cdf that corresponds to perfect forecasts; hence we compare $F(e)$ to $F^*(e)$, and we favor forecasts that minimize the integrated absolute distance between the two, or "stochastic error distance" (*SED*). This approach turns out to yield useful insights with important practical implications. We proceed as follows. In Section 2, we introduce *SED* loss and relate it to traditional loss functions, and in Section 3, we assess the likely empirical relevance of our basic result. In Section 4, we introduce a weighted version *SED* and again relate it to traditional loss functions. In Section 5, we generalize *SED* in a way that facilitates relating it to other divergence and distance measures, such as Cramér-von-Mises divergence and Kolmogorov-Smirnov distance. In Section 6, building on the results of Section 5, for our generalized *SED*, we provide a complete characterization of the relationship between generalized *SED* minimization and traditional expected loss minimization. We conclude in Section 7. All proofs appear in an appendix.

CONTACT Francis X. Diebold fdiebold@sas.upenn.edu University of Pennsylvania, Philadelphia, PA 19104, USA.

[1]More general representations are possible, which recognize that the actual and forecasted values (y and $\hat{y}$, say) need not enter loss only through their difference, which is the forecast error, $e = y - \hat{y}$. See, for example, Patton (2015) and the references therein. We could instead rank by $E(L(y, \hat{y}))$, where the loss function $L(y, \hat{y})$ satisfies $L(y, y) = 0$ and $L(y, \hat{y}) \geq 0$, $\forall y, \hat{y}$. In the vast majority of the literature, however, the simple $L(e)$ form is used, and we shall follow suit here.

[2]In an abuse of notation, throughout this paper, we use "$F(\cdot)$" to denote any cumulative density function. The meaning will be clear from context.

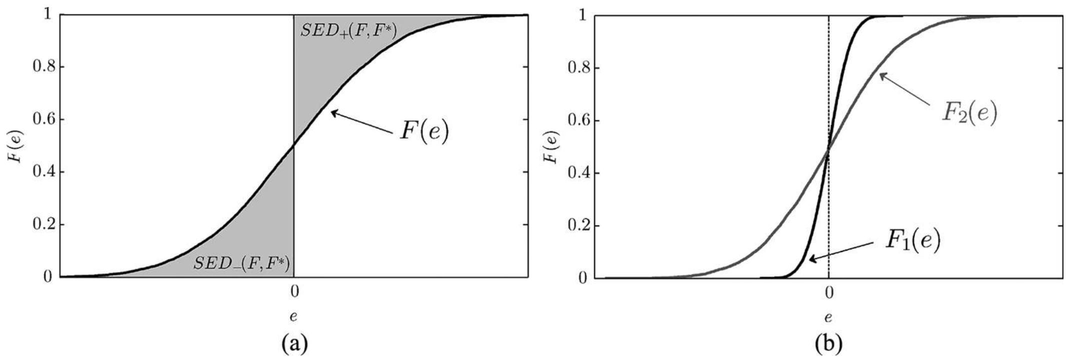

Figure 1. Stochastic error distance ($SED(F, F^*)$). (a) c.d.f. of e. Under the $SED(F, F^*)$ criterion, we prefer smaller $SED(F, F^*) = SED_-(F, F^*) + SED_+(F, F^*)$. (b) Two forecast error distributions. Under the $SED(F, F^*)$ criterion, we prefer $F_1(e)$ to $F_2(e)$.

2. Ranking forecasts by stochastic error distance

We propose simply using the distribution of e directly, ranking forecasts by stochastic distance of $F(e)$ from $F^*(\cdot)$, the unit step function at 0. That is, we rank forecasts by:

$$SED(F, F^*) = \int_{-\infty}^{\infty} |F(e) - F^*(e)|\, de,$$

where smaller is better. We call $SED(F, F^*)$ the *SED*. It will prove useful for what follows to split the $SED(F, F^*)$ integral at the origin, yielding

$$\begin{aligned} SED(F, F^*) &= SED_-(F, F^*) + SED_+(F, F^*) \\ &= \int_{-\infty}^{0} F(e)\, de + \int_{0}^{\infty} (1 - F(e))\, de. \end{aligned} \tag{1}$$

Hence $SED(F, F^*)$ has both "integrated c.d.f." and "integrated survival function" components.[3] In Figure 1a, we show $SED(F, F^*)$ and its components, and in Figure 1b, we provide an example of two error distributions such that one would prefer F_1 to F_2 under $SED(F, F^*)$.

2.1. The relation between $SED(F, F^*)$ and other loss functions

We motivated $SED(F, F^*)$ as directly appealing and intuitive. It turns out, moreover, that $SED(F, F^*)$ is intimately connected to one, and only one, traditionally invoked loss function, and it is not quadratic. We now state a key result.

Proposition 2.1 (Equivalence of *SED* and Expected Absolute Error Loss). *For any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|) < \infty$, we have*

$$SED(F, F^*) = E(|e|). \tag{2}$$

That is, $SED(F, F^*)$ equals expected absolute loss for any error distribution.

Hence if one is comfortable with $SED(F, F^*)$ and wants to use it to evaluate forecast accuracy, then one must also be comfortable with expected absolute error loss and want to use it to evaluate forecast accuracy. The two criteria are *identical*.

[3]Note that in the symmetric case $SED(F, F^*) = 2\int_{-\infty}^{0} F(e)\, de$.

3. On *MAE* vs. *MSE* accuracy rankings

Squared error loss (mean-squared error, *MSE*, or its square root, *RMSE*) and absolute error loss (mean absolute error, *MAE*) are the two great workhorses of point forecast accuracy evaluation. Primary focus, however, is generally on *MSE* rankings, with *MAE* rankings something of a sideshow. Our results argue for a reversal of emphasis, with primary focus on *MAE*. (Recall Proposition 2.1, which says that *SED is MAE*.) But does it matter? That is, do *MAE* and *MSE* rankings agree, in which case it does not matter which is used?

Empirically, *MSE* rankings and *MAE* rankings agree often, but not always. Theoretically, *MSE* rankings and *MAE* rankings certainly do not *have* to agree – they are simply different loss functions – which is why both are typically calculated and examined. However, little is known theoretically about whether and when *MSE* rankings and *MAE* rankings do or do not agree; the question is largely unexplored.[4]

Provision of a full answer turns out to be quite difficult, but we can obtain some results for extremely restrictive cases. If, e.g., forecast errors are Gaussian, $e \sim N\left(\mu, \sigma^2\right)$, then $|e|$ follows the folded normal distribution with mean

$$E(|e|) = \sigma\sqrt{2/\pi}\,\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right].$$

Hence for unbiased forecasts ($\mu = 0$), we have $E(|e|) \propto \sigma$, so that *MAE* rankings and *MSE* rankings must be identical.

Even in the restrictive Gaussian case, however, the rankings can diverge if one (or both) of the forecasts are biased. Consider, for example, two forecast errors, $e_1 \sim N(0, 1)$ and $e_2 \sim N(\mu_2, \sigma_2^2)$, with $\mu_2 \in [-1.3, 1.3]$ and $\sigma_2 \in (0, 1.3]$. By numerical computation, we identify situations *MAE* and *MSE* rankings diverge, which we show in Figure 2. The regions are not large, but they are certainly not negligible. We conjecture, moreover, that divergences may be much more pronounced in *non*-Gaussian situations involving asymmetry and/or fat tails.

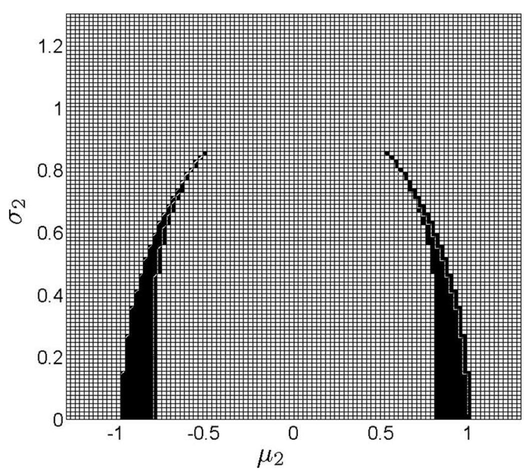

Figure 2. Absolute error loss vs. squared error loss, $e_1 \sim N(0, 1)$, $e_2 \sim N(\mu_2, \sigma_2^2)$. We show the disagreement region in black.

[4]Patton (2015) performs some related, but nevertheless different, explorations.

4. Weighted stochastic error distance

In some circumstances, one may feel that the basic idea behind $SED(F, F^*)$ is appropriate, but that divergence of $F(\cdot)$ from $F^*(\cdot)$ on one side of the origin is more harmful than on the other. This leads to the idea of a *weighted SED* (*WSED*) criterion, given by a *weighted* sum of $SED_-(F, F^*)$ and $SED_+(F, F^*)$.

In particular, let,

$$\begin{aligned} WSED(F, F^*; \tau) &= 2(1-\tau)SED(F, F^*)_- + 2\tau SED(F, F^*)_+ \\ &= 2(1-\tau)\int_{-\infty}^{0} F(e)\, de + 2\tau \int_{0}^{\infty} (1 - F(e))\, de, \end{aligned}$$

where $\tau \in (0, 1)$.[5] The following result is immediate.

Proposition 4.1 (Equivalence of *WSED* and Expected Lin-Lin Loss). *For any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|) < \infty$, we have*

$$WSED(F, F^*; \tau) = 2E(L_\tau(e)), \tag{3}$$

where $L_\tau(e)$ is the loss function

$$L_\tau(e) = \begin{cases} (1-\tau)|e|, & e \le 0 \\ \tau|e|, & e > 0, \end{cases}$$

and $\tau \in (0, 1)$.

The loss function $L_\tau(e)$ appears in the forecasting literature as a convenient and simple potentially asymmetric loss function.[6] It is often called "lin-lin" loss (i.e., linear on each side of the origin) or "check function" loss, again in reference to its shape. Importantly, it is the loss function underlying quantile regression; see Koenker (2005).

Because $WSED(F, F^*; \tau)$ is proportional to expected lin-lin loss as established by Proposition 4.1, we are led inescapably to the insight that point forecast accuracy evaluation by $WSED(F, F^*; \tau)$ is actually point forecast accuracy evaluation by expected lin-lin loss. The primacy of lin-lin loss in the $WSED(F, F^*; \tau)$ case, like the primacy of absolute error loss in the leading special case of $WSED(F, F^*; 1/2)$ ($SED(F, F^*)$), emerges clearly.

Patton and Timmermann (2007) suggest a different route that also leads directly and exclusively to lin-lin loss. Building on the work of Christoffersen and Diebold (1997) on optimal prediction under asymmetric loss, they show that if loss $L(e)$ is homogeneous and the target variable y has no conditional moment dependence beyond the conditional variance, then the L-optimal forecast is always a conditional quantile of y. Hence, under their conditions $WSED(F, F^*; \tau)$, loss is the only asymmetric loss function of relevance.

Our results and those of Patton and Timmermann are highly complementary but very different, not only in the perspective from which they are derived but also in the results themselves. If, for example, y displays conditional moment dynamics beyond second-order, then the L-optimal forecast is generally *not* a conditional quantile (and characterizations in such cases remain elusive), whereas the $WSED(F, F^*; \tau)$-optimal forecast is *always* a conditional quantile.

In closing this section, we also note that $WSED(F, F^*; \tau)$ can be used as a forecast model estimation criterion. By Proposition 4.1, this amounts to estimation using quantile regression, with the relevant quantile governed by τ. When $\tau = 1/2$, the quantile regression estimator collapses to the least absolute deviation (LAD) estimator. Similarly, because the forecast combination problem is a regression problem (Granger and Ramanathan, 1984), forecast combination under $WSED(F, F^*; \tau)$ simply amounts to estimation of the combining equation using quantile regression, with the relevant quantile governed by τ.

[5] Note that when $\tau = 0.5$, $WSED(F, F^*; \tau)$ is just $SED(F, F^*)$.
[6] See Christoffersen and Diebold (1997).

5. Generalized weighted stochastic error distance

Here we generalize and represent *SED* in a way that facilitates relating it to other divergence and distance measures.

5.1. A natural extension

As always let $F(e)$ be the forecast-error c.d.f., and let $F^*(e)$ be the unit step function at zero. Now consider the following generalized weighted stochastic error distance (*GWSED*) measure:

$$GWSED(F, F^*; p, w) = \int \left|F(e) - F^*(e)\right|^p w(e)\, de, \tag{4}$$

where $p > 0$. All of our stochastic error distance measures are of this form. When $p = 1$ and $w(e) = 1\ \forall\ e$, we have $SED(F, F^*)$, and when $p = 1$ and

$$w(e) = \begin{cases} 2(1-\tau), & e < 0 \\ 2\tau, & e \geq 0, \end{cases}$$

we have $WSED(F, F^*; \tau)$. The $GWSED(F, F^*; p, w)$ representation facilitates comparisons of $WSED(F, F^*; \tau)$ to other possibilities that emerge for alternative choices of p and/or $w(\cdot)$.

Interesting connections emerge between $GWSED(F, F^*; p, w)$ and various other important distance and divergence measures. We now discuss more on this.

5.2. GWSED and Cramér-von Mises divergence

When $p = 2$ and $w(e) = f(e)$, the density corresponding to $F(e)$, $GWSED(F, F^*; p, w)$ is Cramér-von Mises divergence,

$$CVM(F^*, F) = \int \left|F^*(e) - F(e)\right|^2 f(e)\, de. \tag{5}$$

Note that the weighting function $w(e)$ in Cramér-von Mises divergence $CVM(F^*, F)$ is distribution-specific, $w(e) = f(e)$. We can decompose Cramér-von-Mises divergence as:

$$\begin{aligned}
CVM(F^*, F) &= \int \left|F^*(e) - F(e)\right|^2 f(e)\, de \\
&= \int \big[F(e)(1 - F^*(e)) + (1 - F(e))F^*(e) \\
&\quad - F(e)(1 - F(e)) - F^*(e)(1 - F^*(e))\big] f(e)\, de \\
&= \int_{R_-} F(e) f(e) de + \int_{R_+} (1 - F(e)) f(e) de - \int_R F(e)(1 - F(e)) f(e)\, de \\
&= \int_0^{F(0)} p\, dp + \int_{F(0)}^1 (1 - p)\, dp - \int_0^1 p(1 - p)\, dp \quad \text{(by change of variable, } e = F^{-1}(p)) \\
&= F(0)^2 - F(0) + \frac{1}{3} \\
&\geq \frac{1}{12}.
\end{aligned}$$

$CVM(F^*, F)$ achieves its lower bound of $1/12$ if and only if $F(0) = 1/2$, which implies that $CVM(F^*, F)$ ranks by "closeness to median unbiasedness," just as does $SED(F, F^*)$.

Remark 5.1 (Directional properties of CVM). Although $CVM(F^*, F)$ is well defined, $CVM(F, F^*)$ is not, because

$$CVM(F, F^*) = \int \left|F(e) - F^*(e)\right|^2 f^*(e)\, de,$$

where $f^*(e)$ is Dirac's delta.

Remark 5.2 (Comparative directional properties of CVM and Kullback-Leibler divergence).[7] The Kullback-Leibler divergence $KL(F^*, F)$ between $F^*(e)$ and $F(e)$ is

$$KL(F^*, F) = \int \log\left(\frac{f^*(e)}{f(e)}\right) f^*(e)\, de,$$

where $f^*(x)$ and $f(x)$ are densities associated with distributions F^* and F. Unlike $CVM(F^*, F)$, $KL(F^*, F)$ does not fit in our $GWSED(F, F^*; p, w)$ framework as it is ill-defined in *both* directions.

Remark 5.3 (Kolmogorov-Smirnov distance, CVM, and SED). Kolmogorov-Smirnov distance is

$$KS(F, F^*) = \sup_e \left|F(e) - F^*(e)\right| = max\left(F(0), 1 - F(0)\right).$$

Like $CVM(F^*, F)$, $KS(F, F^*)$ achieves its lower bound at $F(0) = 1/2$, and $KS(F, F^*)$ therefore ranks by "closeness to median unbiasedness," just as does $SED(F, F^*)$.

Remark 5.4 (Preferences for low-variance errors among unbiased forecasts). Note that although CVM divergence and KS distance value median unbiasedness, as emphasized earlier in this section, they do not invoke a notion of variance to rank unbiased forecasts. In contrast, as emphasized in Section 3, SED ranks unbiased forecasts by variance, preferring forecast errors with smaller variance.

Remark 5.5 (Bias-variance tradeoffs). Although CVM and KS do not consider variance among unbiased forecasts, they do consider it among biased forecasts. But they do it in a counter-intuitive way, due to the choice of weighting function.

5.3. GWSED and Cramér distance

When $p = 2$ and $w(e) = 1$, $GWSED(F, F^*; p, w)$ is Cramér distance, also known as Mallows distance, or Monge-Kantorovich distance, or earth-movers distance; see Levina and Bickel (2001). Closely related, moreover, are the "energy distance" used in higher dimensions (Székely and Rizzo, 2013) and the "continuous ranked probability score" of Gneiting and Raftery (2007).[8]

We can decompose Cramér distance as:

$$\begin{aligned}
\int_{-\infty}^{\infty} \left[F(e) - F^*(e)\right]^2 de &= \int \big[F(e)(1 - F^*(e)) + (1 - F(e))F^*(e) \\
&\quad - F(e)(1 - F(e)) - F^*(e)(1 - F^*(e))\big]\, de \\
&= \int_{-\infty}^{0} F(e)de + \int_{0}^{\infty} \left[1 - F(e)\right] de - \int_{-\infty}^{\infty} F(e)(1 - F(e))\, de \\
&= SED(F, F^*) - \int_{-\infty}^{\infty} F(e)(1 - F(e))\, de. \qquad (6)
\end{aligned}$$

[7] There are of course many other distance/divergence measures, exploration of which is beyond the scope of this paper. On Hellinger distance, e.g., see Maasoumi (1993).

[8] The continuous ranked probability score, however, is not used to compare point forecasts, but rather to compare density forecasts.

Equation (6) is particularly interesting insofar as it shows that Cramér distance is closely related to $SED(F, F^*)$, yet not exactly equal to it, due to the adjustment term, $\int F(e)(1 - F(e))\, de$. One can show that[9]

$$\int F(e)(1 - F(e))\, de = \frac{1}{2} E(|e - e'|),$$

where e' is a stochastic copy of e, revealing that the adjustment term, like the leading term, is a measure of forecast error variability.

One can also rewrite Cramér distance solely in terms of two SED's. We state this result as a proposition.

Proposition 5.6 (SED and Cramér distance). *For any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|) < \infty$, we have*

$$\int_{-\infty}^{\infty} \left[F(e) - F^*(e)\right]^2 de = SED(F, F^*) - \frac{1}{2} SED(G, F^*),$$

where G is the distribution function of $e - e'$ and e' is a stochastic copy of e.

6. $GWSED(F(e), F^*(e); p, w(e))$ vs. $E(L(e))$: A complete characterization

Equivalence of $GWSED(F(e), F^*(e); p, w(e))$ minimization and $E(L(e))$ minimization can actually be obtained for a wide class of loss functions $L(e)$. In particular, we have the following proposition.

Proposition 6.1 (Equivalence of $GWSED$ minimization and $E(L)$ minimization). *Suppose that $L(e)$ is piecewise differentiable with $dL(e)/de > 0$ for $e > 0$ and $dL(e)/de < 0$ for $e < 0$, and suppose also that $F(e)$ and $L(e)$ satisfy $F(e)L(e) \to 0$ as $e \to -\infty$ and $(1 - F(e))L(e) \to 0$ as $e \to \infty$. Then*

$$\int_{-\infty}^{\infty} \left|F(e) - F^*(e)\right| \left|\frac{dL(e)}{de}\right| de = E(L(e)). \tag{7}$$

That is, minimization of $GWSED(F(e), F^*(e); p, w(e))$ when $p = 1$ and $w(e) = |dL(e)/de|$ corresponds to minimization of expected loss $E(L(e))$.

Several remarks are in order.

The importance of the proposition is its highlighting the fact that for any L there is a corresponding and easily calculated $GWSED$, and similarly, for any $GWSED$ there is a corresponding and easily calculated L. Hence the result clarifies what it means to chose a loss function: choosing a loss function amounts to choosing a $GWSED$ weighting function. This may be helpful not only in helping with introspection as to what loss function might be "reasonable" in a given situation but also in obtaining new results by switching from "L-representations" to "$GWSED$-representations."

Remark 6.2 ($GWSED$ weightings other than those corresponding to $WSED$ and SED). Note that the $E(L)$ minimizers that match various $GWSED(F, F^*; p, w)$ minimizers generally correspond to nonstandard and intractable loss functions $L(e)$ in all cases but the ones we have emphasized, namely, $WSED(F, F^*; \tau)$ and its leading case $SED(F, F^*)$.

Remark 6.3 (The $GWSED$ weighting that produces quadratic loss). The weighting function in (7) that produces expected squared-error loss ($E(L(e)) = E(e^2)$) is immediately $|dL(e)/de| = |2e|$. It is not obvious why one would generally want to adopt such a weighting, other than for mathematical convenience.

[9] See Gneiting and Raftery (2007).

Remark 6.4 (Relationship between *GWSED* and Elliott et al. (2005) loss). $GWSED(F, F^*; p, w)$ somewhat resembles the Elliott et al. (2005) (ETK) loss function,

$$L_{ETK}(e; p, \alpha) = |e|^p \ (\alpha + (1 - 2\alpha)I(e < 0)) .$$

It differs fundamentally, however, in that $GWSED(F, F^*; p, w)$ is based on *distributional* distance, $|F - F^*|$, whereas ETK loss is based on the usual *forecast error* distance, $(y - \hat{y})$. Ultimately, ETK loss is a special case of $GWSED(F, F^*; p, w)$, corresponding to a particular choice of exponent p and weighting function $w(e)$, as per Proposition 6.1, as are *all* $L(e)$ loss functions that satisfy the regularity conditions of the proposition.

7. Concluding remarks

Starting from first principles, we have proposed and explored several "stochastic error distance" (*SED*) measures of point forecast accuracy, based directly on the distance between the forecast-error c.d.f. and the unit step function at 0. *SED*-type criteria sharply focus attention on a *particular* loss function, absolute loss (and its lin-lin generalization), as opposed to the ubiquitous quadratic loss, or anything else. Our results elevate the status of absolute and lin-lin loss for both point forecast evaluation and for estimation.

SED is related to important recent work on tests of stochastic dominance (SD) of loss distributions, including Corradi and Swanson (2013), Lee et al. (2014), and Jin et al. (2015). The SED and SD approaches are related in at least two ways. First, and obviously, both are based on comparative properties of certain c.d.f.'s. Second, the SD literature focuses on achieving robustness of accuracy rankings to loss function choices, and SED initially appears that way too, insofar as it is motivated from first principles without reference to an $L(e)$-type loss function.

Yet there is also a clear difference between SD and SED. If SD holds (whether first or higher order), it really *does* imply robustness to certain classes of loss functions. SED, in contrast, leads one inextricably to absolute error loss. Indeed we have shown in Proposition 2.1 that the SED criterion *is* the absolute error loss criterion. Hence, in contrast to SD, which focuses attention on whether one forecast dominates another regardless of loss function, SED ultimately *embraces* a particular loss function, absolute error loss, which until now has been something of a sideshow relative to the ubiquitous quadratic loss.

Several interesting directions arise for future research. One direction concerns multivariate extensions, in which case it is not clear how to define the unit step function at zero, $F^*(e)$. Consider, for example, the bivariate case, in which the forecast error is $e = (e_1, e_2)'$. It seems clear that we want $F^*(e) = 0$ when both errors are negative and $F^*(e) = 1$ when both are positive, but it is not clear what to do when the signs diverge.

This difficulty is not surprising insofar as it parallels difficulties with multivariate extensions of the median. Various notions of the multivariate median are available, but it seems that no single measure dominates others (Small, 1990; Gneiting, 2011). It would be interesting to investigate how various multivariate versions of SED (for various versions of the unit step function at zero) relate to various versions of multivariate median.

Another direction is further the study of *MAE* vs. *MSE* rankings as introduced in Section 3. In work building on the research reported here, Ardakani et al. (2015) take interesting steps in that direction, obtaining analytic results under a "convex ordering" assumption weaker than a normality. Necessary and sufficient conditions remain elusive, however, for the general case of non-Gaussian, nonzero-mean forecast errors.

Appendices

A. A useful lemma

We begin with a lemma.

Lemma A.1.

(i) *Let y be a positive random variable such that $E(|y|)) < \infty$. Then*

$$E(y) = \int_0^\infty [1 - F(y)]dy,$$

where $F(y)$ is the cumulative distribution function of y.

(ii) *Let y be a negative random variable such that $E(|y|)) < \infty$.*[10] *Then*

$$E(y) = -\int_{-\infty}^0 F(y)dy,$$

where $F(y)$ is the cumulative distribution function of y.

Lemma A.1 *(i)* is well known in the mathematical statistics literature (Block and Fang, 1988; Rao, 2009), and it also features prominently in the hazard/survival literature (Neumann, 1999).

B. Proofs of Propositions

We now prove the propositions stated in the paper.

Proposition 2.1 (Equivalence of *SED* and Expected Absolute Error Loss). *For any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|) < \infty$, we have*

$$SED(F, F^*) = \int_{-\infty}^0 F(e)\, de + \int_0^\infty [1 - F(e)]\, de = E(|e|).$$

That is, $SED(F, F^*)$ equals expected absolute loss for any error distribution.

Proof of Proposition 2.1.

$$\begin{aligned}
SED(F, F^*) &= SED_-(F, F^*) + SED_+(F, F^*)\\
&= \int_{-\infty}^0 F(e)\, de + \int_0^\infty (1 - F(e))\, de\\
&= -\int_{-\infty}^0 ef(e)\, de + \int_0^\infty ef(e)\, de \quad \text{(by Lemma A.1 (i) for } SED_- \text{ and (ii) for } SED_+)\\
&= \int_0^\infty ef(-e)\, de + \int_0^\infty ef(e)\, de\\
&= \int_0^\infty e(f(-e) + f(e))\, de\\
&= \int_{-\infty}^\infty |e| f(e)\, de\\
&= E(|e|).
\end{aligned}$$

Proposition 4.1 (Equivalence of *WSED* and Expected Lin-Lin Loss). *For any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|) < \infty$, we have*

$$WSED(F, F^*; \tau) = 2(1 - \tau) \int_{-\infty}^0 F(e)\, de + 2\tau \int_0^\infty [1 - F(e)]\, de = 2E(L_\tau(e)), \tag{A.1}$$

[10] In another abuse of notation, we use "y" to denote either a generic random variable or its realization.

where $L_\tau(e)$ is the loss function

$$L_\tau(e) = \begin{cases} (1-\tau)|e|, & e \leq 0 \\ \tau|e|, & e > 0, \end{cases}$$

and $\tau \in (0,1)$.

Proof of Proposition 4.1.

$$\begin{aligned}
WSED(F, F^*; \tau) &= 2(1-\tau)\int_{-\infty}^{0} F(e)\,de + 2\tau \int_{0}^{\infty} (1-F(e))\,de \\
&= 2(1-\tau)\int_{-\infty}^{0} (-e)f_e(e)\,de + 2\tau \int_{0}^{\infty} ef_e(e)\,de \quad \text{(by Lemma A.1)} \\
&= 2(1-\tau)\int |e|1\{e \leq 0\}f_e(e)\,de + 2\tau \int |e|1\{e > 0\}f_e(e)\,de \\
&= 2\int \big[(1-\tau)|e|1\{e \leq 0\} + \tau|e|1\{e > 0\}\big]f_e(e)\,de \\
&= 2E(L_\tau(e)).
\end{aligned}$$

Proposition 6.1 (Equivalence of *GWSED* minimization and $E(L(e))$ minimization). *Suppose that $L(e)$ is piecewise differentiable with $dL(e)/de > 0$ for $e > 0$ and $dL(e)/de < 0$ for $e < 0$, and suppose also that $F(e)$ and $L(e)$ satisfy $F(e)L(e) \to 0$ as $e \to -\infty$ and $(1-F(e))L(e) \to 0$ as $e \to \infty$. Then*

$$\int_{-\infty}^{\infty} \left|F(e) - F^*(e)\right| \left|\frac{dL(e)}{de}\right| de = E(L(e)).$$

That is, minimization of $GWSED(F, F^*; p, w)$ when $p = 1$ and $w(e) = |dL(e)/de|$ corresponds to minimization of expected loss $E(L(e))$.

Proof of Proposition 6.1.

$$\begin{aligned}
\int_{-\infty}^{\infty} \left|F(e) - F^*(e)\right| \left|\frac{dL(e)}{de}\right| de &= -\int_{-\infty}^{0} F(e)\frac{dL(e)}{de}\,de + \int_{0}^{\infty} (1-F(e))\frac{dL(e)}{de}\,de \\
&= \int_{-\infty}^{0} f(e)L(e)\,de + \int_{0}^{\infty} f(e)L(e)\,de \\
&= \int_{-\infty}^{\infty} f(e)L(e)\,de \\
&= E[L(e)],
\end{aligned}$$

where we obtain the second line by integrating by parts and exploiting the assumptions on $L(e)$ and $F(e)$. In particular,

$$\int_{-\infty}^{0} F(e)\frac{dL(e)}{de}\,de = F(e)L(e)\Big|_{-\infty}^{0} - \int_{-\infty}^{0} f(e)L(e)\,de,$$

by integration by parts, but the first term is zero because $F(e)L(e) \to 0$ as $e \to -\infty$, and similarly,

$$\int_{0}^{\infty} (1-F(e))\frac{dL(e)}{de}\,de = (1-F(e))L(e)\Big|_{0}^{\infty} + \int_{0}^{\infty} f(e)L(e)\,de,$$

again by integration by parts, and again the first term is zero because $(1-F(e))L(e) \to 0$ as $e \to \infty$.

Acknowledgments

We are especially grateful to the Editors (Peter C.B. Phillips and Aman Ullah) and two anonymous referees for helpful guidance and comments. We are also grateful to Ross Askanazi, Alex Belloni, Lorenzo Braccini, Xu Cheng, Peter Christoffersen, Valentina Corradi, Ed George, Roger Koenker, Mai Li, Oliver Linton, Laura Liu, Essie Maasoumi, Andrew Patton, Ehsan Soofi, Norm Swanson, Mike Steele, Allan Timmermann, Aman Ullah, Mark Watson, and Tiemen Woutersen. We also thank seminar participants at Federal Reserve Bank of San Francisco, the Emory University Conference in Honor of Essie Maasoumi, the European University Institute, and the University of Pennsylvania. The usual disclaimer applies.

References

Ardakani, O. M., Ebrahimi, N., Soofi, E. S. (2015). Ranking forecast models by stochastic error distance and survival information risk. Manuscript, Armstrong State University, Northern Illinois University, and University of Wisconsin-Milwaukee.

Block, H. W., Fang, Z. (1988). A multivariate extension of hoeffding's lemma. *Annals of Probability* 16:1803–1820.

Christoffersen, P. F., Diebold, F. X. (1997). Optimal prediction under asymmetric loss. *Econometric Theory* 13:808–817.

Corradi, V., Swanson, N. R. (2013). A survey of recent advances in forecast accuracy comparison testing, with an extension to stochastic dominance. In: Chen, X., Swanson, N., eds. *Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions, Essays in honor of Halbert L. White, Jr.*, Berlin: Springer, pp. 121–143.

Elliott, G., Komunjer, I., Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies* 72:1107–1125.

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting* 27:197–207.

Gneiting, T., Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102:359–378.

Granger, C. W. J., Ramanathan, R. (1984). Improved methods of forecasting. *Journal of Forecasting* 3:197–204.

Jin, S., Corradi, V., Swanson, N. R. (2015). Robust forecast comparison. Manuscript, Singapore Management University, University of Surrey, and Rutgers University. Available at SSRN: http://ssrn.com/abstract=2605927.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monograph Series. Cambridge: Cambridge University Press.

Lee, T. H., Tu, Y., Ullah, A. (2014). Nonparametric and semiparametric regressions subject to monotonicity constraints: Estimation and forecasting. *Journal of Econometrics* 182:196–210.

Levina, E., Bickel, P. (2001). The earth mover's distance is the mallows distance: Some insights from statistics. *Proceedings Eighth IEEE International Conference on Computer Vision*, pp. 251–256. Vancouver, BC, Canada, 7–14 July 2001.

Maasoumi, E. (1993). A compendium to information theory in economics and econometrics. *Econometric Reviews* 12:137–181.

Neumann, G. R. (1999). Search models and duration data. In: Pesaran, M. H., Schmidt, P., eds. *Handbook of Applied Econometrics*, Vol. 2. Blackwell, Oxford and Cambridge, MA, pp. 300–351.

Patton, A. J. (2015). Evaluating and comparing possibly misspecified forecasts. Manuscript, Duke University.

Patton, A. J., Timmermann, A. (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* 102:1172–1184.

Rao, C. R. (2009). *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York, NY.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review* 58:263–277.

Székely, G. J., Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* 143:1249–1272.