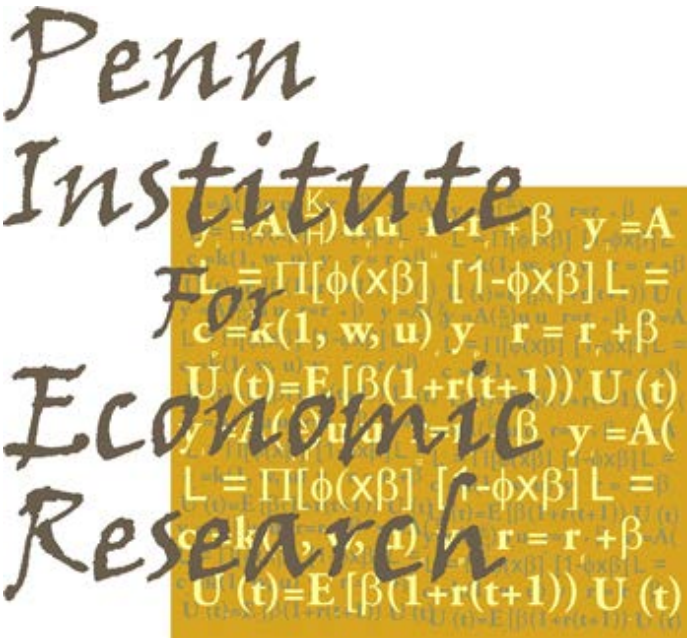


Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper 14-011

"Assessing Point Forecast Accuracy by Stochastic Divergence from Zero"

by

Francis X. Diebold and Minchul Shin

http://ssrn.com/abstract=2418270

Assessing Point Forecast Accuracy by Stochastic Divergence from Zero

Francis X. Diebold
University of Pennsylvania

Minchul Shin
University of Pennsylvania

March 26, 2014

Abstract

We propose and explore several related ways of reducing reliance of point forecast accuracy evaluation on expected loss, $E(L(e))$, where e is forecast error. Our central approach dispenses with the loss function entirely, instead using a "stochastic error divergence" (SED) accuracy measure based directly on the forecast-error c.d.f., $F(e)$. We explore several variations on the basic theme; interestingly, all point to the primacy of absolute-error loss and its generalizations.

Acknowledgments: For helpful comments and/or discussion we are especially grateful to Lorenzo Braccini, Laura Liu, Andrew Patton and Allan Timmermann. We also thank Ross Askanazi, Xu Cheng, Valentina Corradi, Mai Li, Oliver Linton, Essie Maasoumi, Norm Swanson, Mark Watson, Tiemen Woutersen, and the Penn Friday Econometrics Research Group. The usual disclaimer applies.

Key words: Forecast accuracy, forecast evaluation, absolute-error loss, quadratic loss, squared-error loss

JEL codes: C53

Contact: fdiebold@sas.upenn.edu

1 Introduction

One often wants to evaluate (that is, 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 $L(e) \geq 0$, $\forall e$. But (1) the mathematical expectation $E(L(e))$ is only one aspect of the loss distribution, in contrast to the complete summary provided by its c.d.f. $F(L(e))$, and moreover (2) the relevant loss function L is far from obvious in many situations.[1] In this paper we address both (1) and (2).

We make two related contributions; the first addresses (1), and the second addresses (2). First, we develop accuracy measures that incorporate aspects of the entire loss distribution, $F(L(e))$, not just its expectation $E(L(e))$. We do this by assessing the divergence between $F(L(e))$ and the unit step function at 0,

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

because nothing can dominate a benchmark forecast whose errors consistently achieve zero loss; i.e., a forecast whose errors achieve $F(L(e)) = F^*(L(e))$.

Second, recognizing that one rarely knows what loss function might be appropriate or realistic, we dispense with the loss function entirely, proposing accuracy measures based directly on the c.d.f. $F(e)$ as opposed to the c.d.f. $F(L(e))$. In particular, we assess the divergence between $F(e)$ and the unit step function at 0,

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

because nothing can dominate a benchmark forecast whose errors are consistently 0, i.e., a forecast whose errors achieve $F(e) = F^*(e)$.

The results differ markedly in the two cases. The first case involving $F(L(e))$ vs. $F^*(L(e))$, which we call "stochastic loss divergence," turns out to be something of a dead end. In sharp contrast, the second case involving $F(e)$ vs. $F^*(e)$, which we call "stochastic error divergence," turns out to yield useful insights with important practical implications.

We proceed as follows. We explore stochastic loss divergence in section 2, and then we move to stochastic error divergence in section 3. In section 4 we explore a weighted version

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

of stochastic error divergence, which allows positive and negative errors of the same absolute magnitude nevertheless to have different costs. In section 5 we propose a generalized stochastic error divergence measure, which allows us to relate our stochastic error divergence to energy distance and Cramer-von-Mises divergence, among others, and we provide a complete characterization of the relationship between generalized stochastic error divergence minimization and expected loss minimization. We conclude in section 6.

2 Ranking Forecasts by Stochastic Loss Divergence

By ranking forecasts by SLD we mean that we prefer the forecast whose loss distribution $F(L(e))$ has smallest divergence from the reference loss function $F^*(\cdot)$, the unit step function at 0. In what follows we make this idea more precise and explore its implications.

2.1 The Basic Idea

The idea is simply that loss should be small. Expected loss, however, is only one aspect of the loss distribution. Hence we use the entire loss distribution, quantifying its divergence from unit probability mass at 0. More precisely, we use loss $L(e)$, but instead of ranking by $E(L(e))$ we rank by stochastic divergence' of $F(L(e))$ from $F^*(\cdot)$, the unit step function at 0. That is, we rank forecasts by the area,

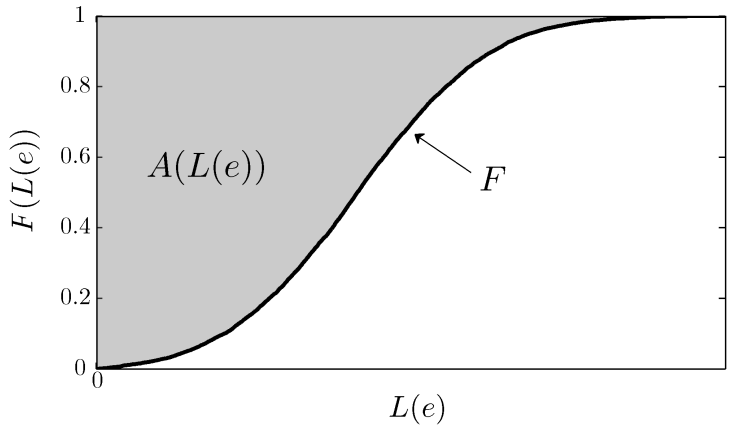
$$A(L(e)) = \int_0^\infty [1 - F(L(e))]\, dL(e), \tag{1}$$

where smaller $A(\cdot)$ is better. In Figure 1a we illustrate the SLD idea, and in Figure 1b we show two loss distributions such that we prefer F_1 to F_2 under the SLD criterion.

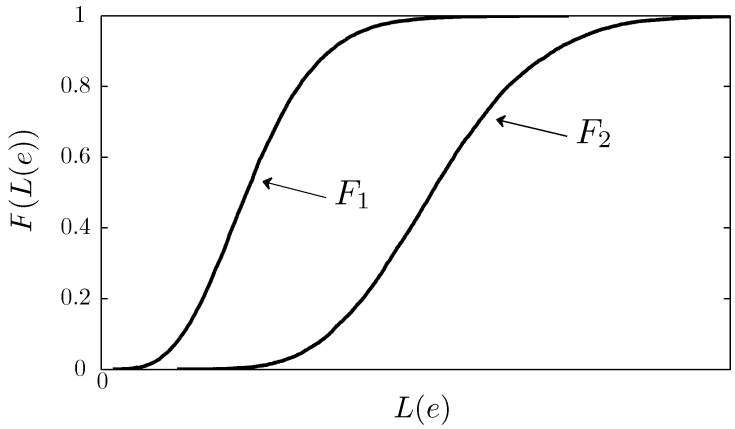
2.2 A Negative Result

Let us begin with a lemma that will feature not only in the negative result of this section, but also in the positive results of subsequent sections.

Lemma 2.1 *For random variable x with c.d.f. $F(x)$, if $E(|x|)) < \infty$,*
$\lim_{c\to\infty} c(1 - F(c)) = 0$.



(a) c.d.f. of $L(e)$. Under the *stochastic loss divergence* (SLD) criterion, we prefer smaller $A(L(e))$.



(b) Two forecast loss distributions. Under the stochastic loss divergence (SLD) criterion, we prefer F_1 to F_2.

Figure 1: Stochastic Loss Divergence (SLD)

Proof We have

$$\begin{aligned} c(1-F(c)) &= cP(X>c) \\ &= c\int_c^\infty dP(x) \\ &= \int_c^\infty c\, dP(x) \\ &\le \int_c^\infty x\, dP(x) \quad \text{(replacing } c \text{ with } x\text{)} \\ &= \int_0^\infty x\, dP(x) - \int_0^c x\, dP(x). \end{aligned}$$

But this converges to zero as $c \to \infty$, because

$$\int_0^\infty x \, dP(x) \leq \int_{-\infty}^\infty |x| \, dP(x) < \infty. \quad \blacksquare$$

Now let us proceed to consider minimization of SLD. Unfortunately it takes us nowhere, insofar as it corresponds to expected loss minimization, as we now show.

Lemma 2.2 *(Equivalence of Stochastic Loss Divergence and Expected Loss) Let $L(e)$ be a forecast-error loss function satisfying $L(0) = 0$ and $L(e) \geq 0$, $\forall$ e, with $E(|L(e)|) < \infty$.*[2] *Then*

$$A(L(e)) = \int_0^\infty [1 - F(L(e))] \, dL(e) = E(L(e)), \tag{2}$$

where $F(L(e))$ is the cumulative distribution function of $L(e)$. That is, SLD equals expected loss for any loss function and error distribution.

Proof To evaluate $E(L(e))$ we integrate by parts:

$$\begin{aligned} \int_0^c L(e) f(L(e)) \, dL(e) &= -L(e)[1 - F(L(e))] \Big|_0^c + \int_0^c [1 - F(L(e))] \, dL(e) \\ &= -c(1 - F(c)) + \int_0^c [1 - F(L(e))] \, dL(e). \end{aligned}$$

Now letting $c \to \infty$ we have

$$\begin{aligned} E(L(e)) = \int_0^\infty L(e) f(L(e)) \, dL(e) &= \lim_{c \to \infty} -c(1 - F(c)) + \int_0^\infty [1 - F(L(e))] \, dL(e) \\ &= 0 + \int_0^\infty [1 - F(L(e))] \, dL(e) \quad \text{(by Lemma (2.1))} \\ &= A(L(e)). \quad \blacksquare \end{aligned}$$

We call this result a "lemma" rather than a "proposition" because we will use it in proving a subsequent proposition. To the best of our knowledge, it has not appeared in the forecasting literature. It does appear, however, in the hazard and survival modeling literature, in whose jargon "expected lifetime equals the integrated survival function."

[2] In another abuse of notation, throughout we use "$L(e)$" to denote either the loss random variable or its realization. The meaning will be clear from context.

3 Ranking Forecasts by Stochastic Error Divergence

The conjecture explored in Section 2 was that, because expected loss is only one aspect of the loss distribution, it may be of interest to base point forecast comparisons on suitable functionals of the entire loss distribution. That turned out, however, to lead full circle, with SLD minimization corresponding to expected loss minimization. Here we go farther and arrive at an interesting result.

3.1 The Basic Idea

One rarely has a credible loss function tied to specifics of a situation; rather, quadratic loss is almost always invoked, purely for convenience. The insight of this section is that we can take a more primitive approach, *dispensing* with loss functions, and still rank forecasts (although, as we shall show, we are inescapably pulled back to a *particular* loss function). We simply use e directly, and we rank by stochastic divergence of $F(e)$ from $F^*(\cdot)$, the unit step function at 0. This amounts to ranking forecasts by the shaded area in Figure 2a,

$$A(e) = A_-(e) + A_+(e) = \int_{-\infty}^{0} F(e)\, de + \int_{0}^{\infty} (1 - F(e))\, de, \tag{3}$$

where smaller is better.[3] We call $A(e)$ the *stochastic error divergence* (SED). In Figure 2b we provide an example of two error distributions such that we prefer F_1 to F_2 under SED.

3.2 A Positive Result

We motivated SED as directly appealing and intuitive. It turns out, however, the SED is intimately connected to one, and only one, tradtionally-invoked loss function. And it's not quadratic. We begin with a lemma and then proceed to the main result.

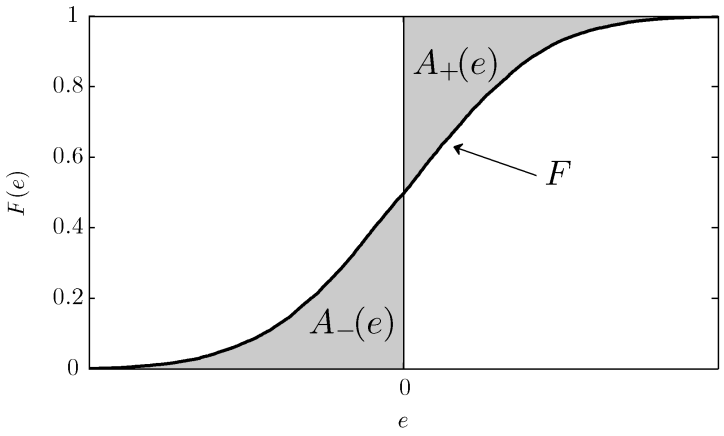
Lemma 3.1 *Let x be a negative random variable such that $E(|x|)) < \infty$.*[4] *Then*

$$E(x) = -\int_{-\infty}^{0} F(x)dx,$$

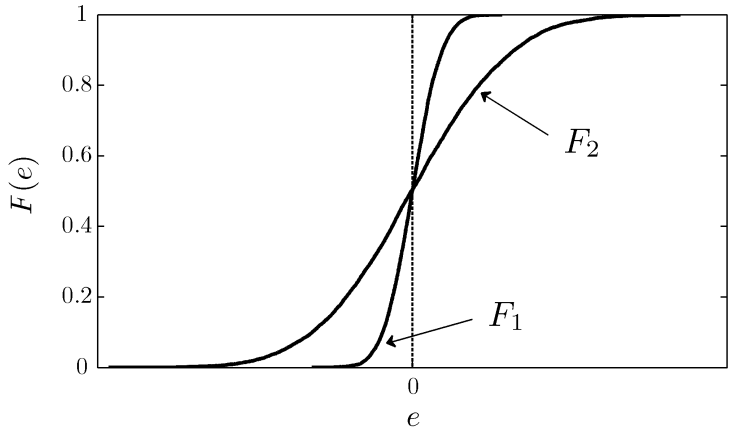
where $F(x)$ is the cumulative distribution function of x.

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

[4] In yet another abuse of notation, throughout we use "x" to denote either a generic random variable or its realization.



(a) c.d.f. of e. Under the SED criterion, we prefer smaller $A(e) = A_-(e) + A_+(e)$.



(b) Two forecast error distributions. Under the SED criterion, we prefer F_1 to F_2.

Figure 2: Stochastic Error Divergence (SED)

Proof Integrating by parts, we have

$$\begin{aligned}\int_{-c}^{0} xf(x)dx &= xF(x)\Big|_{-c}^{0} - \int_{-c}^{0} F(x)dx \\ &= cF(-c) - \int_{-c}^{0} F(x)dx.\end{aligned}$$

As in Lemma 2.2, the first term goes to zero as $c \to -\infty$, by Lemma 2.1. ∎

The proof of Lemma 3.1 of course parallels that of Lemma 2.2. The only difference is that Lemma 2.2 treated positive random variables, whereas Lemma 3.1 treats negative random variables.

We now arrive at a positive result.

Proposition 3.2 *(Equivalence of Stochastic Error Divergence (SED) and Expected Absolute Error Loss) For any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|)) < \infty$, we have*

$$A(e) = \int_{-\infty}^{0} F(e)\, de + \int_{0}^{\infty} [1 - F(e)]\, d\, e = E(|e|). \tag{4}$$

That is, SED equals expected absolute loss for any error distribution.

Proof [5]

$$\begin{aligned}
A(e) &= A_{-}(e) + A_{+}(e) \\
&= \int_{-\infty}^{0} F(e)de + \int_{0}^{\infty} (1 - F(e))de \\
&= -\int_{-\infty}^{0} ef(e)de + \int_{0}^{\infty} ef(e)de \quad \text{(by Lemma 3.1 for} A_{-} \text{ and Lemma 2.2 for } A_{+}) \\
&= \int_{0}^{\infty} ef(-e)de + \int_{0}^{\infty} ef(e)de \\
&= \int_{0}^{\infty} e(f(-e) + f(e))de \\
&= \int_{-\infty}^{\infty} |e| f(e)de \\
&= E(|e|). \quad \blacksquare
\end{aligned}$$

Hence, in a certain sense, "I don't know anything about the loss function, but I'm comfortable minimizing SED" is equivalent to "My loss function is absolute loss."

4 Weighted Stochastic Error Divergence

In other circumstances, however, one may feel more along the lines of "I don't know much about the loss function, but I know that I dislike negative errors (say) more than positive." This leads us to the idea of a *weighted* SED (WSED) criterion $\tilde{A}$, given by a *weighted* sum of $A_{-}(e)$ and $A_{+}(e)$.

[5] We provide an alternative proof of Proposition 3.2 in Appendix A.

4.1 A Natural Generalization

In particular, let,

$$\tilde{A} = 2(1-\tau)A_- + 2\tau A_+ = 2(1-\tau)\int_{-\infty}^{0} F(e)de + 2\tau\int_{0}^{\infty}(1-F(e))de$$

where $\tau \in (0,1)$.[6] The following result is immediate.

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

$$\tilde{A}(e) = 2(1-\tau)\int_{-\infty}^{0} F(e)\,de + 2\tau\int_{0}^{\infty}[1-F(e)]\,d\,e = 2E(L_\tau(e)), \tag{5}$$

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

Proof We have

$$\begin{aligned}
\tilde{A} &= 2(1-\tau)\int_{-\infty}^{0} F(e)de + 2\tau\int_{0}^{\infty}(1-F(e))de \\
&= 2(1-\tau)\int_{-\infty}^{0}(-e)f_e(e)de + 2\tau\int_{0}^{\infty} e f_e(e)de \quad \text{(by Lemmas 2.2 and 3.1)} \\
&= 2(1-\tau)\int |e|1\{e\le 0\}f_e(e)de + 2\tau\int |e|1\{e>0\}f_e(e)de \\
&= 2\int \big[(1-\tau)|e|1\{e\le 0\} + \tau|e|1\{e>0\}\big]f_e(e)de \\
&= 2E(L_\tau(e)). \quad \blacksquare
\end{aligned}$$

The loss function $L_\tau(e)$ appears in the forecasting literature as a convenient and simple potentially asymmetric loss function. It is often called "lin-lin" loss (i.e., linear on each side of the origin), and sometimes also called "check function" loss (again in reference to

[6]Note that when $\tau = 0.5$, WSED $\tilde{A}$ is just SED A.

its shape).[7] Importantly, it is the loss function underlying quantile regression; see Koenker (2005).

Because WSED is twice expected lin-lin loss, we are led inescapably to the insight that point forecast accuracy evaluation "without taking a stand" on the loss function (SED) or "taking only a small stand" on the loss function (WSED), actually *does* map into point forecast accuracy evaluation by expected absolute or expected lin-lin loss, respectively. The primacy of lin-lin loss, and the leading case of absolute loss, emerges clearly.

4.2 Remarks

Several remarks are in order.

Remark 4.2 (WSED and stochastic dominance (SD)). Our work is related to, yet distinct from, that of Corradi and Swanson (2013), who propose tests for first-order stochastic dominance (SD) of loss distributions, and hence also to earlier work on which Corradi and Swanson build, such as Linton et al. (2005). The WSED 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, and more centrally, both begin as attempts to make point forecast accuracy rankings robust to specific choices of loss functions.

There is, however, a clear difference between SD and WSED, and hence between our approach and that of Corradi and Swanson (2013). If SD holds (whether first- or higher-order), it really *does* imply robustness to certain classes of loss functions. But in our view SD criteria (again whether first- or higher-order) for forecast error loss distributions are so restrictive as to be unlikely ever to hold, which renders SD "tests" – and certainly first-order SD tests – of limited practical relevance for forecast evaluation.

WSED, in contrast, also begins as an attempts at loss-function robustness insofar as it is motivated from first principles without reference to a loss function, but it winds up at the doorstep of lin-lin loss. Indeed we have shown that the WSED criterion *is* the lin-lin loss criterion! Hence, in contrast to SD which strives for robustness to loss function, WSED ultimately *embraces* a loss function and is of immediate practical relevance. But it embraces a *particular* loss function, lin-lin loss and its leading case of absolute error loss, which until now has been something of a sideshow relative to the ubiquitous quadratic loss, thereby strongly suggesting a change of emphasis toward lin-lin.

Remark 4.3 (WSED and optimal prediction under aymmetric loss). By Proposition 4.1,

[7] Christoffersen and Diebold (1996) and Christoffersen and Diebold (1997).

the forecast that optimizes WSED is the forecast that optimizes lin-lin loss, $L_\tau(e)$. That is, the WSED criterion leads directly and exclusively to lin-lin loss.

The important work of Patton and Timmermann (2007) suggests a different and fascinating 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 lin-lin 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 higher-order cases remain elusive), whereas the WSED-optimal forecast is *always* a conditional quantile.

Remark 4.4 (WSED as an estimation criterion). WSED, which of course includes SED as a special case, 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 deviations (LAD) estimator.

Remark 4.5 (WSED as a forecast combination criterion). Because the forecast combination problem is a regression problem (Granger and Ramanathan (1984)), forecast combination under WSED simply amounts to estimation of the combining equation using quantile regression, with the relevant quantile governed by τ.

5 Generalized Stochastic Error Divergence

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 stochastic error divergence (GSED) measure:

$$D(F^*, F; p, w) = \int |F^*(e) - F(e)|^p \, w(e) de, \tag{6}$$

where $p > 0$. Our stochastic error divergence measures are of this form. When $p = 1$ and $w(x) = 1 \; \forall \; x$, we have SED, and when $p = 1$ and

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

we have WSED.

The GSED representation facilitates comparisons of WSED to other possibilities that emerge for alternative choices of p and/or $w(\cdot)$.

5.1 Connections Between WSED and Other Distance and Divergence Measures

Several connections emerge. First, when $p = 2$ and $w(x) = 1$, D is the so-called "energy distance,"[8]

$$E(F^*, F) = \int |F^*(e) - F(e)|^2 \, de.$$

We can decompose the energy distance as

$$\begin{aligned} \int_{-\infty}^{\infty} \big[F^*(e) - F(e)\big]^2 de &= \int \big[F(e)(1 - F^*(e)) + (1 - F(e))F^*(e) \\ &\qquad - F(e)(1 - F(e)) - F^*(e)(1 - F^*(e))\big] de \\ &= \int_{-\infty}^{0} F(e) de + \int_{0}^{\infty} \big[1 - F(e)\big] de - \int_{-\infty}^{\infty} F(e)(1 - F(e)) de \\ &= E\left(|e|\right) - \int_{-\infty}^{\infty} F(e)(1 - F(e)) de, \end{aligned} \tag{7}$$

where e and e' are random variables independently and identically distributed with distribution function $F(\cdot)$. Equation (7) is particularly interesting insofar as it shows that energy distance is prominently related to expected absolute error loss, yet not exactly equal to it, due to the adjustment term, $\int F(e)(1 - F(e))de$. However, one can show that

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

[8] On energy distance see Székely and Rizzo (2005) and their recent survey Székely and Rizzo (2013), as well as Gneiting and Raftery (2007).

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

Second, when $p = 2$ and $w(e) = f(e)$, the density corresponding to $F(e)$, D is Cramer-von-Mises divergence,

$$CVM(F^*, F) = \int |F^*(e) - F(e)|^2 f(e)de. \tag{8}$$

Note that the weighting function $w(e)$ in Cramer-von-Mises divergence $CVM(F^*, F)$ is distribution specific, $w(e) = f(e)$.

We can decompose Cramer-von-Mises divergence as

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

Note that $CVM(F^*, F)$ achieves its lower bound of 1/12 if and only if $F(0) = 1/2$, which implies that, like SED, $CVM(F^*, F)$ ranks forecasts according to expected absolute error.

5.2 A Complete Characterization

Equivalence of $D(F^*, F)$ 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 5.1 *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*

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

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

Proof

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

where we obtain the second line by integrating by parts and exploiting the 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 $e \to \infty$. ∎

We hasten to emphasize the key point, however, namely that the $E(L(e))$ minimizers that match various $D(F^*, F)$ minimizers generally correspond to non-standard and intractable loss functions $L(e)$ in all cases but the ones we have emphasized, namely WSED and its leading case SED.

5.3 Remarks

Several additional remarks are in order.

Remark 5.1 (Kolmogorov-Smirnov distance and expected absolute error). Kolmogorov-Smirnov distance is

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

Like $CVM(F^*, F)$, $KS(F^*, F)$ achieves its lower bound at $F(0) = 1/2$. Hence $KS(F^*, F)$ also ranks forecasts according to expected absolute error.

Remark 5.2 (Directional properties of CVM). Although $CVM(F^*, F)$ is well-defined, $CVM(F, F^*)$ is not, because

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

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

Remark 5.3 (Comparative directional properties of Kullback-Leibler divergence)[9]. The Kullback-Leibler divergence $KL(F^*, F)$ between $F^*(e)$ 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 $D(F^*, F)$ framework as it is ill-defined in *both* directions.

Remark 5.4 (Relationship between GSED and Elliott et al. (2005) loss). The GSED measure (6) resembles the Elliott et al. (2005) (ETK) loss function,

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

However, it differs fundamentally in that GSED is based on (integrated) *distributional* divergence, $\int(F^* - F)$, whereas ETK loss is based on the usual *forecast error* divergence, $(y - \hat{y})$. Ultimately, ETK loss is a special case of GSED; corresponding to a particular choice of GDED exponent p and GSED weighting function $w(e)$, as per Proposition 5.1, as are *all* $L(e)$ loss functions that satisfy the regularity conditions of the proposition.

6 Conclusions and Directions for Future Research

We have proposed and explored several "stochastic error divergence" measures of point forecast accuracy, based directly on the divergence between the forecast-error c.d.f., $F(e)$, and the unit step function at 0. Our results make clear that one can't escape focus on expected loss minimization, even when moving to stochastic divergence accuracy measures.

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

Simultaneously, however, they 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. Put bluntly, our message is that "expected absolute loss (and its lin-lin generalization) is more important than you think." (Or at least more important than you used to think.)

Several interesting directions for future research are apparent. One concerns multivariate extensions, in which case it's 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's not clear what to do when the signs diverge.

Another interesting direction for future research concerns the coherence of absolute- and squared-error loss. We have implicitly argued for absolute-error loss (or its lin-lin generalization). How important is the distinction between absolute-error loss and other loss functions? In particular, under what conditions will absolute-error loss and the ubiquitous squared-error loss agree? If, for example, the forecast error is Gaussian, $e \sim N(\mu, \sigma^2)$, 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 absolute and quadratic loss rankings are identical. In other cases, however, absolute and quadratic rankings diverge, and it would be useful to have a complete characterization.

Appendices

A Alternative Proof of Proposition 3.2

Here we supply a different and shorter, if less instructive, proof.

Proposition

$$E(|e|) = \int_0^\infty [1 - F(e)]\, d\, e = A(e).$$

Proof

$$\begin{aligned} A(e) &= -\int_0^c F^{-1}(p)dp + \int_c^1 F^{-1}(p)dp \quad \text{(where } c = F(0)\text{)} \\ &= \int_{-\infty}^0 -ef(e)de + \int_0^\infty ef(e)de \quad \text{(change of variables with } p = F(e)\text{)} \\ &= \int_{-\infty}^\infty |e|f(e)de \\ &= E(|e|). \quad \blacksquare \end{aligned}$$

References

Christoffersen, P.F. and F.X. Diebold (1996), "Further Results on Forecasting and Model Selection Under Asymmetric Loss," *Journal of Applied Econometrics*, 11, 561–572.

Christoffersen, P.F. and F.X. Diebold (1997), "Optimal Prediction Under Asymmetric Loss," *Econometric Theory*, 13, 808–817.

Corradi, V. and N.R. Swanson (2013), "A Survey of Recent Advances in Forecast Accuracy Comparison Testing, with an Extension to Stochastic Dominance," In X. Chen and N. Swanson (eds.), *Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions, Essays in honor of Halbert L. White, Jr.*, Springer, 121-143.

Elliott, G., A. Timmermann, and I. Komunjer (2005), "Estimation and Testing of Forecast Rationality under Flexible Loss," *Review of Economic Studies*, 72, 1107–1125.

Gneiting, T. and A. E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Granger, C.W.J. and R. Ramanathan (1984), "Improved Methods of Forecasting," *Journal of Forecasting*, 3, 197–204.

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

Linton, O., E. Maasoumi, and Y.J. Whang (2005), "Consistent Testing for Stochastic Dominance Under General Sampling Schemes," *Review of Economic Studies*, 72, 735–765.

Maasoumi, E. (1993), "A Compendium to Information Theory in Economics and Econometrics," *Econometric Reviews*, 12, 137–181.

Patton, A.J. and A. Timmermann (2007), "Testing Forecast Optimality Under Unknown Loss," *Journal of the American Statistical Association*, 102, 1172–1184.

Székely, G.J. and M.L. Rizzo (2005), "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method," *Journal of Classification*, 22, 151–183.

Székely, G.J. and M.L. Rizzo (2013), "Energy Statistics: A Class of Statistics Based on Distances," *Journal of Statistical Planning and Inference*, 143, 1249–1272.