International Statistical Review

International Statistical Review (2018), 86, 3, 442–468 doi:10.1111/insr.12250

Ranking Forecasts by Stochastic Error Distance, Information and Reliability Measures

Omid M. Ardakani[1], Nader Ebrahimi[2] and Ehsan S. Soofi[3]

[1]*Department of Economics, Georgia Southern University, Savannah, Georgia, USA*
Email: oardakani@georgiasouthern.edu
[2]*Division of Statistics, Northern Illinois University, DeKalb, Illinois, USA*
Email: nebrahim@niu.edu
[3]*Sheldon B. Lubar School of Business and Center for Research on International Economics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA*
E-mail: esoofi@uwm.edu

Summary

The stochastic error distance (SED) introduced by Diebold and Shin (2017) ranks forecast models by divergence between distributions of the errors of the actual and perfect forecast models. The basic SED is defined by the variation distance and provides a representation of the mean absolute error, but by basing ranking on the entire error distribution and divergence, the SED moves beyond the traditional forecast evaluations. First, we establish connections between ranking forecast models by the SED, error entropy and some partial orderings of distributions. Then, we introduce the notion of excess error for forecast errors of magnitudes larger than a tolerance threshold and give the SED representation of the mean excess error (MEE). As a function of the threshold, the MEE is a local risk measure. With the distribution of the absolute error as a prior for the threshold, its Bayes risk is the entropy functional of the survival function, which is a known measure in the information theory and reliability. Notions and results are illustrated using various distributions for the error. The empirical versions of SED, MEE and its Bayes risk are compared with the mean squared error in ranking regression and autoregressive integrated moving average models for forecasting bond risk premia.

Key words: Bayes risk; convex order; dispersive order; entropy; mean absolute error; mean residual life; mean squared error; stochastic order.

1 Introduction

Traditionally, forecasting models are evaluated according to risk functions defined by expected values of various loss functions such as the mean squared error (MSE) and the mean absolute error (MAE). Recently, Diebold and Shin (2017), hereafter written as D–S, introduced the notion of stochastic error distance (SED) that ranks forecast models by the divergence between the cumulative distribution function (CDF) of the error of the forecast model from the CDF of the ideal error-free forecast. Let ε denote the random error of a forecast model. For a perfect forecast model, the CDF of ε is a step function at zero:

$$F_0(\varepsilon) = \begin{cases} 0, & \varepsilon < 0 \\ 1, & \varepsilon \geq 0; \end{cases} \tag{1}$$

(the use of ε for random error and its values simplifies notations without creating ambiguity).

The basic SED of D–S is defined by the variation distance between the CDF of a forecast error F_ε and F_0:

$$SED(F_\varepsilon, F_0) = \int_{-\infty}^{\infty} |F_\varepsilon(z) - F_0(z)| dz. \tag{2}$$

The SED ranks forecast models as follows. For two forecast models with random errors ε_k distributed as $F_k, k = 1, 2$, the forecast model with ε_1 is preferred to the forecast model with ε_2 when $SED(F_1, F_0) \leq SED(F_2, F_0)$. Thus, the SED defines a distributional order, which will be denoted as $F_1 \preceq_{sed} F_2$ or $\varepsilon_1 \preceq_{sed} \varepsilon_2$.

D–S also defined a more general version of $SED(F_\varepsilon, F_0)$ in the following form:

$$SED_{p,w}(F_\varepsilon, F_0) = \int_{-\infty}^{\infty} |F_\varepsilon(z) - F_0(z)|^p w(z) dz, \tag{3}$$

where $p > 0$ and $w(\cdot)$ is a weight function. When $p = 1$ and $w(\varepsilon) = 1$ for all ε, we obtain $SED(F_\varepsilon, F_0)$, that is, $SED_{1,1}(F_\varepsilon, F_0) = SED(F_\varepsilon : F_0)$. These authors noted that Cramér-von Mises divergence is $SED_{2,f}(F_\varepsilon, F_0)$, where $w(\varepsilon) = f_\varepsilon(\varepsilon)$ is the probability density function (PDF) of F_ε. They explored the connection between $SED_{2,1}(F_\varepsilon, F_0)$ and Cramér distance and noted a connection between SED and the Kolmogorov–Smirnov distance. They illustrated that the MSE does not always fit in the SED framework. They also noted that the Kullback–Leibler information divergence does not fit in the SED framework, because the distribution of the error of perfect forecast is degenerate.

The basic SED is related to a traditional measure of forecast evaluation, namely,

$$SED(F_\varepsilon, F_0) = E(|\varepsilon|) = MAE(\varepsilon), \tag{4}$$

provided that $E(|\varepsilon|) < \infty$. However, the notion of SED moves beyond the traditional approach, as evident by generalisation (3) that does not always correspond to a traditional measure. The important aspect of SED is evaluation of forecast models by 'basing rankings on the entire distribution' of the forecast error and using 'accuracy measures based directly on the distance between' distributions (D–S). This aspect of SED is in the same spirits as the notions of distribution orders widely used in reliability and many other fields, and the information theory where statistical activities are viewed in terms of divergence between probability distributions (Kullback, 1959; Akaike, 1973; Soofi & Retzer, 2002). This paper builds on this important aspect of SED notion through exploring the connections of the SED ranking of forecast models with the notions of information theory and partial ordering of distributions. We further develop the concept of SED through synthesis of existing results and providing original results.

First, we explore connections between ranking forecast models by the SED, entropy of the forecast error and some well-known partial orderings of the error distributions. In addition, connections with the variance will be also included as the MSE of the unbiased forecast models. Within the information framework of Kullback (1959) and Jaynes (1957), a well-known broad family of maximum entropy models encompasses the Laplace, normal and uniform distributions, which are associated, respectively, with the MAE, variance and entropy as measures of risk. In the partial ordering of distributions approach, the stochastic order and two variability orders provide conditions for identical rankings of forecasts by the SED and entropy. Some conditions for the identical rankings of these measures by the variance are also explored for the case of unbiased forecast models. This expedition is a continuation of Ebrahimi *et al.* (1999) for the identical rankings of random variables by entropy and variance.

Then, we introduce the notion of excess error defined by $|\varepsilon| > \tau$, which are costly and are penalised. This notion connects the SED concept to the dynamic generalisation of mean absolute error, namely, the mean residual life function used in reliability and survival analysis, also known as the mean excess loss function in the actuarial science; see Poynor (2010) for an exposition, several distributional examples and numerous references. In reliability and survival analysis, τ represents the current age of an item, and in actuarial science, τ represents the amount of deductible of an insurance policy. The role of τ in the notion of excess error is similar to the actuarial science where the insurer reimburses costs in excess of the deductible amount. The idea is akin to the use of a prediction interval where all values within the margin of error are acceptable, which implies that a range of forecast errors is tolerated. We formulate a version of (3), where $\tau \geq 0$ represents a tolerance threshold such that only the excess forecast errors are costly. This formulation provides a dynamic generalisation of $SED(F_\varepsilon, F_0)$, which will be referred to as the *mean excess error (MEE)*, in view of (4). The mean residual order defined in the reliability literature is a partial order of random variable, but we will show that it is also the risk associated with a loss function that only penalises the excess errors. This formulation also identifies sufficient conditions for forecast errors such that the weighted SED ranks forecasts identically to the Shannon entropy.

For a given tolerance threshold τ, $MEE(\tau)$ is a local risk function. Its global risk is given by the Bayes risk $E_\tau[MEE(\tau)]$, where the expectation is with respect to a distribution for τ. When the distribution of τ is the same as the distribution of the absolute error, $E_\tau[MEE(\tau)]$ gives a measure that is known as the cumulative residual entropy (Rao *et al.*, 2004), survival entropy (Zografos & Nadarajah, 2005) and the entropy functional of the survival function (Asadi *et al.*, 2014). The SED framework reveals this measure as a risk, which we call the *mean excess error risk (MEER)*. This exploration is a continuation of studies of relationships between the trio of reliability, information and economic notions; see Ebrahimi *et al.* (2014) for the latest developments and references.

We establish connections between the SED and some information and reliability notions theoretically and illustrate the results through distributional examples, which serve the main purpose of this paper. In addition, we also illustrate implementations of the new measures introduced here through an empirical example. Consistent estimators of the mean residual life function and the entropy functional of the survival function are available in the reliability and information theory literatures. These estimators provide empirical versions of $MEE(\tau)$ and $MEER(\tau)$. We use these empirical measures as descriptive statistics along with the MAE and MSE to rank several models for forecasting bond risk premia. This problem has been studied extensively in the literature where the use of three or five principal components (PCs) of the return factors for predictors has been controversial. The new measures and SED confirm the results of Bauer and Hamilton (2015) found based on the MSE. Although the purpose of our example is to examine the agreement/disagreement of the new and traditional measures in an empirical analysis, new time series models provide further insights about forecasting the bond risk premia.

This paper is organised as follows. Section 2 gives the background materials for the rest of the paper, including the broad families of distributions that are used to illustrate various concepts, information theory, definitions and implications of some well-known distributional orders. Section 3 presents implications of the distributional orders for the identical rankings of forecast models by the SED, entropy and variance. Section 4 introduces the notion of excess error, shows the representation of MEE as a weighted SED and gives some sufficient conditions for identical ranking of forecasts by MEE and the entropy. This section also gives MEER as a Bayes risk of MEE and discusses MEER ranking under dispersive order. Section 5 presents implications to mixture models for the forecast error. Section 6 presents empirical MEE and

MEER and illustrates applications to factor models for bond risk premia forecasting. Section 7 gives some concluding remarks. Technical details and R codes for the empirical measures are shown in the Appendix. R codes and instructions for implementing the distributional and empirical examples are available in the Supporting Information.

2 Backgrounds

The notions of SED, information theory and partial ordering of the forecast error distributions will be illustrated via examples that include various types of distributions. Figure 1 presents examples of the PDFs of three groups of models that will be used for illustrating the concepts and measures throughout the paper. The left panel shows three PDFs in the generalised error (GE) family of distributions that has the following PDF:

$$f_\beta(\varepsilon) = \frac{\beta}{2\Gamma(1/\beta)} e^{-|\varepsilon|^\beta}, \ \beta > 0. \tag{5}$$

The GE family is also known as the generalised Gaussian and exponential power distributions. This family includes distributions with various shapes and tail thickness less or more than the Laplace and normal distributions that are given by (5) with $\beta = 1, 2$, respectively. The uniform PDF over the interval $(-1, 1)$ is the pointwise limit of the PDF in (5) as $\beta \to \infty$. The information property of this family is described in Section 2.1. The middle panel shows examples of the set of models with bell-shaped PDFs that includes the normal and distributions with thicker tails (logistic and Student-t family). The right panel shows examples of PDFs that have the double-exponential shape with thicker tails than the Laplace distribution. This set is introduced here and will be referred to as the *double generalised Pareto (DGP)* family, because the distribution of the absolute error is the generalised Pareto. The Laplace (double exponential) distribution is the limiting case in this family.

Table A1 gives the PDFs of the absolute error $|\varepsilon|$ for these three sets of models. In Table A1, $g_{|\varepsilon|}$ gives five PDFs for the absolute errors $|\varepsilon|$. For normal, t, and logistic distributed errors, the distributions of $|\varepsilon|$ are folded normal (half-normal), folded-t and half-logistic. For the GE distributed error, the distribution of $|\varepsilon|$ is the generalised gamma with the gamma shape parameter $1/\beta$ and Weibull shape parameter β. For the DGP distributed error, the distribution of $|\varepsilon|$ is generalised Pareto (Pareto type II for $0 < \alpha < \infty$ and exponential for $\alpha \to \infty$).

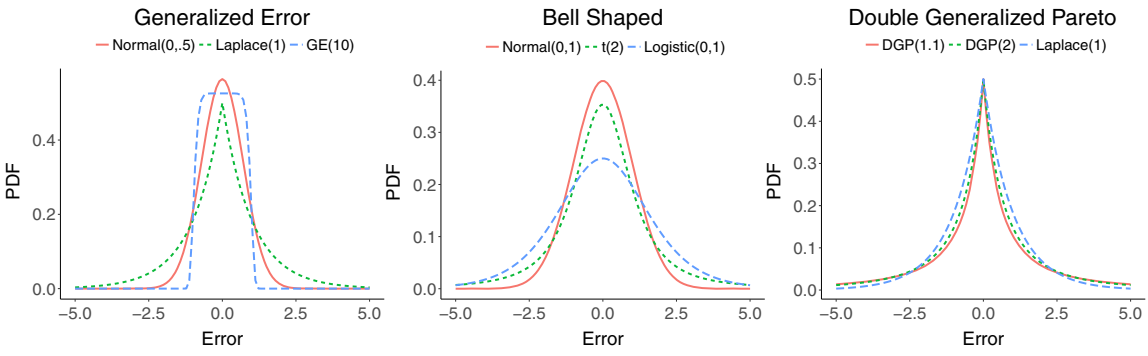

Figure 1. *The PDFs of members of three types of distributions with various shapes and tail thickness models for the forecast error; Normal(0,.5)=GE(2), Laplace(1)=GE(1)=DGP(∞). PDF, probability density function. [Colour figure can be viewed at wileyonlinelibrary.com]*

2.1 Information Divergence

The Kullback–Leibler (KL) information divergent between two distributions with PDFs $f_k, k = 1, 2$ is defined by

$$K(f_1 : f_2) = \int_{\mathcal{S}} f_1(z) \log \frac{f_1(z)}{f_2(z)} dz \geq 0, \tag{6}$$

provided that the integral is finite and requires f_1 to be absolutely continuous with respect to f_2. The inequality in (6) becomes equality if and only if $f_1(z) = f_2(z)$ almost everywhere. The KL divergence has been used as loss and risk functions in various estimation problems; see, for example, Ghosh and Yang (1988) and Clarke and Barron (1994).

Because the distribution of the error of the perfect forecast is degenerate, $K(f_\varepsilon : f_0)$ is not well defined, causing the lack of connection between KL divergence and the SED approach, as noted by D–S. However, ranking distributions by the SED and by the KL divergence may be connected indirectly via the entropy. The Shannon entropy of the random error ε with PDF f_ε is defined by

$$H(\varepsilon) = H(f_\varepsilon) = -\int_{-\infty}^{\infty} f_\varepsilon(z) \log f_\varepsilon(z) dz,$$

provided that the integral is finite. The error variance, denoted as $V(\varepsilon) = E[\varepsilon - \mu_\varepsilon]^2, \mu_\varepsilon = E(\varepsilon)$. It is well known that $V(\varepsilon) < \infty$ implies that $E(|\varepsilon|) < \infty$ and $H(\varepsilon) < \infty$, but the converse does not hold. Also, $H(\varepsilon)$ can be finite when $E(|\varepsilon|)$ is not (e.g. Cauchy distribution, as can be seen in Table A1). For two forecast errors $\varepsilon_k, k = 1, 2$, the entropy order will be denoted as $\varepsilon_1 \preceq_{ent} \varepsilon_2$ and the variance order will be denoted as $\varepsilon_1 \preceq_{var} \varepsilon_2$.

The KL divergence between any distribution with a PDF and the uniform PDF f_ε^* on a common bounded support $\mathcal{S}$ is given by

$$K(f_\varepsilon : f_\varepsilon^*) = H(f_\varepsilon^*) - H(f_\varepsilon) \geq 0, \tag{7}$$

where $H(f_\varepsilon^*) = \log ||\mathcal{S}||$ and $||\mathcal{S}||$ denotes the size of the support. The entropy is a measure of expected loss in terms of (7) where $-H(f)$ is interpreted as a measure of the expected gain of information in terms of the concentration of f over the uniform distribution (Lindley, 1956; Zellner, 1971). By (7), for two forecast models with error PDFs $f_k, k = 1, 2$ on a common bounded $\mathcal{S}$, $\varepsilon_1 \preceq_{ent} \varepsilon_2$ if and only if $K(f_1 : f^*) \geq K(f_2 : f^*)$. The case of unbounded $\mathcal{S}$ can be interpreted similarly in terms of (7) or in terms of its following more general version.

The Laplace and Gaussian models, referred to as the Laplace's first and second laws of errors. These models are often associated with the absolute and squared error loss functions, because minimising $|\varepsilon|^\beta, \beta = 1, 2$ is equivalent to maximising (5) as the likelihood model. The PDF (5) is the maximum entropy model in Ω_θ defined by the single moment constraint $E|\varepsilon|^\beta = \theta, \beta > 0$, where $\beta = \theta^{-1}$. The maximum entropy characterisation of (5) formalises these associations. The maximum entropy error models subject to the risks $E(|\varepsilon|)^\beta \leq \theta, \beta = 1, 2$ are the Laplace and normal models, respectively. The relationship (7) holds for any $f_\varepsilon, f_\varepsilon^* \in \Omega_\theta$, where Ω_θ is defined by a set of moment constraints and f_ε^* is the maximum entropy model; (see Ebrahimi *et al.*, 2010 and references therein). The information divergence between any $f \in \Omega_\theta$ is given by the entropy difference in (7) with $f_\varepsilon^*(z) = f_\beta(z)$. That is, for two forecast models with error distributions $f_k \in \Omega_\theta, k = 1, 2, \varepsilon_1 \preceq_{ent} \varepsilon_2$ if and only if $K(f_1 : f_\beta) \geq K(f_2 : f_\beta)$.

Table 1. *Distributional orders and their implications used in this paper.*

Distributional order	Notation	Implications [condition]
Stochastic order	$Z_1 \leq_{st} Z_2$	$Z_1 \leq_{ent} Z_2$ [if f_2 is decreasing]
Dispersive order	$Z_1 \leq_{disp} Z_2$	$Z_1 \leq_{ent} Z_2$, $Z_1 \leq_{var} Z_2$
		$Z_1 \leq_{st} Z_2$ [if left end of $\mathcal{S}_k = a > -\infty, k = 1, 2^*$]
Convex order	$Z_1 \leq_{cx} Z_2$	$E(Z_1) = E(Z_2)$, $V(Z_1) \leq V(Z_2)$
		$Z_1 \leq_{ent} Z_2$[if f_2 is logconcave]
Mean excess order	$Z_1 \leq_{mee} Z_2$	
Hazard rate order	$Z_1 \leq_{hr} Z_2$	$Z_1 \leq_{mee} Z_2$
Likelihood ratio order	$Z_1 \leq_{\ell r} Z_2$	$Z_1 \leq_{hr} Z_2$

* $\mathcal{S}_k, k = 1, 2$ denotes the support of the distribution of $Z_k, k = 1, 2$.

2.2 Distributional Orders

This section gives the definitions of the orders and statements of implications used in this paper. Table 1 presents the overview of the distributional orders and their implications in terms of two random variables $Z_k, k = 1, 2$ that represent ε_k or/and $|\varepsilon_k|$. More details and results can be found in Shaked and Shanthikumar (2007) (hereafter written as S–S).

Definition 1. *Let $Z_k, k = 1, 2$ be two random variables with CDFs F_k, PDFs f_k and survival functions S_k.*

(a) Z_1 is said to be smaller than Z_2 in stochastic order, denoted as $Z_1 \leq_{st} Z_2$, if $S_1(z) \leq S_2(z)$ for all z in the supports of S_1 and S_2.
(b) Z_1 is said to be smaller than Z_2 in dispersive order, denoted as $Z_1 \leq_{disp} Z_2$, if

$$F_1^{-1}(u) - F_1^{-1}(v) \leq F_2^{-1}(u) - F_2^{-1}(v), \text{ for all } 0 < u < v < 1,$$

where $F_k^{-1}(u) = \sup\{\varepsilon : F_k^{-1}(z) \leq u\}$ is the u-th quantile.
(c) Z_1 is said to be smaller than Z_2 in convex order, denoted as $Z_1 \leq_{cx} Z_2$, if $E[\phi(Z_1)] \leq E[\phi(Z_2)]$ for all convex functions $\phi : \Re \to \Re$.

The distributional order in part (a) is also known as the 'first order stochastic dominance'. The dispersive and convex orders are distributional variability orders (S–S, chapter 3).

It is easy to verify the stochastic and convex orders by the number of sign changes $SC(h)$ of a function h and the dispersive order via the hazard rates $\lambda_k(\tau) = \frac{f_k(z)}{S_k(z)}, k = 1, 2$.

Lemma 1 (S–S, Theorems 1.A.12, 3.A.44, 3.B.18)**.**

(a) $Z_1 \leq_{st} Z_2$ if $SC(f_2 - f_1) = 1$ and the sign sequence is $-, +$.
(b) $Z_1 \leq_{cx} Z_2$ if $SC(f_2 - f_1) = 2$ and the sign sequence is $+, -, +$.
(c) $Z_1 \leq_{cx} Z_2$ if $SC(F_2 - F_1) = 1$ and the sign sequence is $+, -$.
(d) For two absolutely continuous non-negative random variables $Z_1 \leq_{disp} Z_2$ if $\lambda_1(z) \leq \lambda_2(c + z)$ for all $c, z \geq 0$.

The following lemma states some implications of these orderings.

Lemma 2.

(a) If $Z_1 \leq_{st} Z_2$, f_1 is absolutely continuous relative to f_2, and f_2 is decreasing, then $Z_1 \leq_{ent} Z_2$ (Ebrahimi *et al.*, 2013)*.*

(b) If $Z_1 \leq_{disp} Z_2$, then $V(Z_1) \leq V(Z_2), H(Z_1) \leq H(Z_2)$, provided that the measures exist (Oja, 1981, p.160; Shaked, 1982, p.313).
(c) Let $Z_k, k = 1, 2$ be two random variables such that supports of their distributions have a common left endpoint $a > -\infty$. If $Z_1 \leq_{disp} Z_2$, then $Z_1 \leq_{st} Z_2$ (S–S, Theorem 3.B.13).

The following lemma gives ordering results for random variables with a common mean.

Lemma 3.

(a) If $Z_1 \leq_{disp} Z_2$ and $E(Z_1) = E(Z_2) < \infty$, then $E|Z_1 - E(Z_1)| \leq E|Z_1 - E(Z_2)|$, (Oja, 1981, p.160; Shaked, 1982, p.313).
(b) If $Z_1 \leq_{cx} Z_2$, then $E(Z_1) = E(Z_2)$, provided that the expectations exist (S–S, p.110).
(c) If $Z_1 \leq_{cx} Z_2$, then $Z_1 \leq_{var} Z_2$, provided that the variances exist (S–S, p.110).
(d) If $Z_1 \leq_{cx} Z_2$ and f_2 is log-concave, then $Z_1 \leq_{ent} Z_2$ (Yu, 2008).

The aforementioned lemmas will be used for ε or/and $|\varepsilon|$. The following notions and results will be used for $|\varepsilon|$.

Let Z be a non-negative continuous random variable, such as $|\varepsilon|$. The excess (residual) of Z is defined by $Z - \tau | Z > \tau$ and its PDF is given by the truncated (conditional) PDF,

$$f_{z>\tau}(z) = \frac{f_z(z)}{S_z(\tau)}, \quad z > \tau \geq 0. \tag{8}$$

The mean residual life, called here as the mean excess absolute error and denoted as $MEE(\tau)$, is defined by

$$MEE(\tau) = E_{z>\tau}(Z - \tau | Z > \tau) = \frac{1}{S_z(\tau)} \int_\tau^\infty S_z(u) du, \tag{9}$$

where $E_{z>\tau}$ denotes the expectation with respect to (8), provided that $E(Z) < \infty$.

Definition 2. *Let Z_k, k = 1, 2 be two random variables with PDFs on the support $z \geq 0$*

(a) Z_1 is said to be smaller than Z_2 in the MEE order, denoted as $Z_1 \leq_{mee} Z_2$, if $MEE_1(\tau) \leq MEE_2(\tau)$ for all $\tau \geq 0$.
(b) Z_1 is said to be smaller than Z_2 in the hazard rate order, denoted as $Z_1 \leq_{hr} Z_2$, if $\lambda_1(\tau) \geq \lambda_2(\tau)$ for all $\tau \geq 0$.
(c) Z_1 is said to be smaller than Z_2 in the likelihood ratio order, denoted as $Z_1 \leq_{\ell r} Z_2$, if $\frac{f_1(z)}{f_2(z)}$ decreases in z for $z > 0$.

The following lemma gives results for verifying MEE order in terms of the hazard rate and likelihood ratio orders.

Lemma 4 (S–S, Theorems 1.B.1, 2.A.1, 1.C.1, 2.A.3)**.**

(a) If $Z_1 \leq_{hr} Z_2$, then $Z_1 \leq_{st} Z_2$.
(b) If $Z_1 \leq_{hr} Z_2$, then $Z_1 \leq_{mee} Z_2$.
(c) If $Z_1 \leq_{\ell r} Z_2$, then $Z_1 \leq_{hr} Z_2$.
(d) If $P(Z_k > 0) = 1, k = 1, 2$ and

$$\frac{MEE_1(\tau)}{MEE_2(\tau)} \geq \frac{E(Z_1)}{E(Z_2)} \quad \textit{for all } \tau > 0, \tag{10}$$

then $Z_1 \leq_{mee} Z_2 \Longrightarrow Z_1 \leq_{st} Z_2$.

3 Stochastic Error Distance and Distributional Orders

This section presents implications of the stochastic, dispersive and convex orderings for the equivalent ranking of forecast models by SED, entropy and variance.

3.1 Stochastic Order

The relationship (4) and the well-known representation of the mean of a non-negative continuous random variable by the integral of the survival function provides the following representation of the SED:

$$SED(F_\varepsilon, F_0) = \int_0^\infty S_{|\varepsilon|}(z)dz, \tag{11}$$

where $S_{|\varepsilon|}(z) = P(|\varepsilon| > z)$ is the survival function of the absolute error $|\varepsilon|$. Clearly, in view of (11), for two forecast models, $|\varepsilon_1| \leq_{st} |\varepsilon_2|$ implies that $\varepsilon_1 \leq_{sed} \varepsilon_2$. Therefore, results that are available for the stochastic order are applicable to ranking forecast models by the SED.

By part (a) of Lemma 2 and noting that $H(\varepsilon) = H(|\varepsilon|) + \log 2$, we have the following result.

Proposition 1. *Let* $\varepsilon_k, k = 1, 2$ *denote the errors of two forecast models. If* $|\varepsilon_1| \leq_{st} |\varepsilon_2|$ *and* g_2 *is decreasing, then* $\varepsilon_1 \leq_{sed} \varepsilon_2 \Longleftrightarrow \varepsilon_1 \leq_{ent} \varepsilon_2$.

Some remarks are in order.

1 We should emphasise that the assumption of a decreasing PDF is required only for $|\varepsilon_2|$. This assumption is particularly reasonable when ε_2 is the error of an unbiased forecast (although the unbiasedness is *not* a required assumption).
2 If $\varepsilon_k^* = h(\varepsilon_k), k = 1, 2$, where $h(\cdot)$ is an increasing function, then $|\varepsilon_1| \leq_{st} |\varepsilon_2| \Longleftrightarrow |\varepsilon_1^*| \leq_{st} |\varepsilon_2^*|$ (S–S, 2007, Theorem 1.A.3). The closure under transformation implies that for $\varepsilon^* = \varepsilon/\sigma$, the scale parameter stochastically orders the absolute error and entropy increasingly. Thus, by Proposition 1, the SED and entropy rank the family identically. Clearly, the ranking by the scale parameter is identical to the variance (when defined).

The following example illustrates applications of Proposition 1.

Example 1.

(a) All PDFs of the absolute error $g_{|\varepsilon|}$ *shown in Table A1 are decreasing, hence, by Proposition 1, any error distribution stochastically dominated by these distributions, its SED (when defined) and entropy are less than the measures shown in the table.*
(b) By the sign change rule, the Half-t family is stochastically ordered decreasingly by the degrees of freedom ν*, the generalised gamma family with the PDF shown in Table A1 is stochastically ordered decreasingly by the power parameter* β*, and generalized Pareto family with PDF shown in Table A1 is stochastically ordered decreasingly by the tail index* α*. Thus, by Proposition 1, the SED and entropy order the error models within these families*

identically. This can be verified by the expressions for the $MAE(\varepsilon)$ *and* $H(\varepsilon)$ *shown in Table A1.*

(c) If the distribution of ε_1 *is normal and the distribution of* ε_2 *is logistic with same scale parameter, then, by using the sign change rule, we find that* $|\varepsilon_1| \leq_{st} |\varepsilon_2|$*. Thus, Proposition 1 applies. This can also be verified by the* $MAE(\varepsilon)$ *and* $H(\varepsilon)$ *shown in Table A1.*

D–S illustrated that $SED(F_{|\varepsilon|}, F_0)$ provides an elegant visualisation of the MAE. The following example compares visualisations provided by $SED(F_{|\varepsilon|}, F_0)$ and the survival plots.

Example 2. *In the first row of Figure 2, the left panel shows CDF plots of the GE family for* $\beta = 2, 1.5, 1$ *and the right panel shows CDF plots for Student-t with* $\nu = 2, 5$ *and the normal error model (*$\nu = \infty$*). In the top row, the MAE is depicted by the area between* F_ε *and the CDF of the perfect forecast* F_0*. In the second row of Figure 2, the survival integral representation* (11) *provides the visualisation of the SED in terms of the area under the survival curve of the absolute forecast error. The visualisation for comparing the error models by MAE is more clear in the survival plots.*

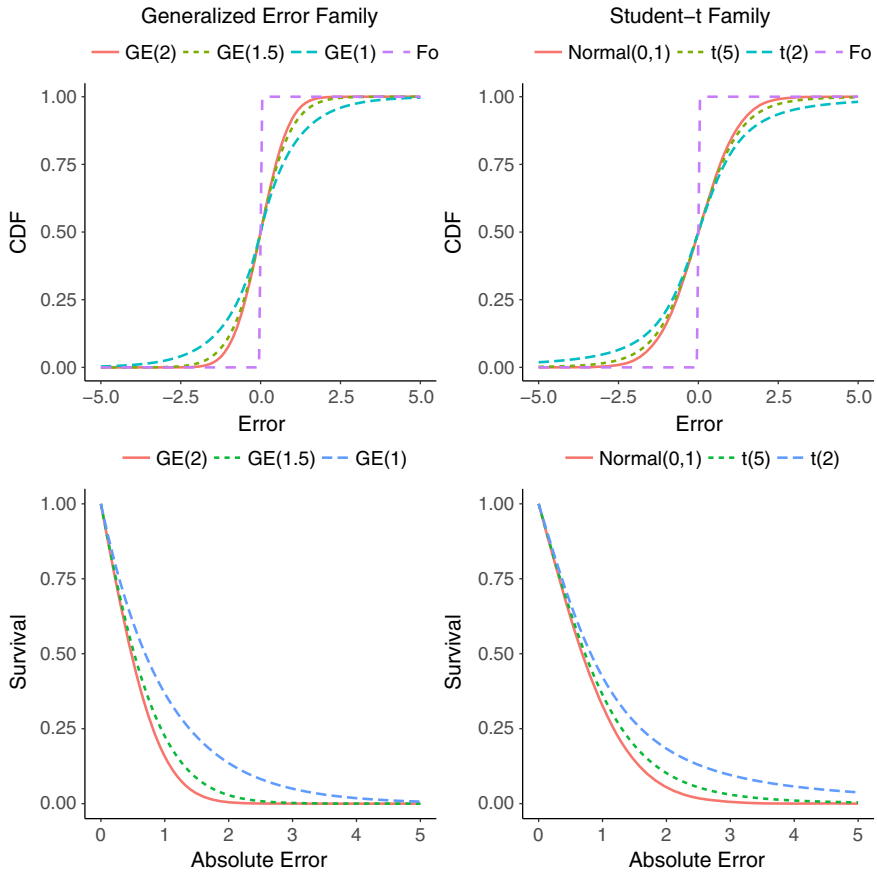

Figure 2. *Visualisations of the MAE provided by the error CDF plots (upper panels) and absolute error survival curves (lower panels) for the generalized error family (left panels) and Student-t family (right panels); GE(2)=Normal(0,.5), GE(1)=Laplace(1), Normal(0,1)=t*(∞)*. MAE, mean absolute error; CDF, cumulative distribution function. [Colour figure can be viewed at wileyonlinelibrary.com]*

3.2 Variability Orders

The orderings of forecast error distributions by SED, entropy and variance can also be connected through the stronger variability orders.

From parts (b) and (c) of Lemma 2, Proposition 1, and noting that the supports of distributions of $|\varepsilon_k|, k = 1, 2$ have a common left endpoint $|\varepsilon| \geq 0$, we have the following implication of the dispersive order for the SED order.

Proposition 2. *For any two forecast models, if* $|\varepsilon_1| \leq_{disp} |\varepsilon_2|$, *then* $\varepsilon_1 \leq_{sed} \varepsilon_2$.

The following proposition gives implications of the convex order for identical ranking of forecasts by SED, entropy and variance.

Proposition 3. *Let* ε_k *be a forecast error with distribution* $F_k, k = 1, 2$.

(a) If $\varepsilon_1 \leq_{cx} \varepsilon_2$, *then* $\varepsilon_1 \leq_{sed} \varepsilon_2$.
(b) If $\varepsilon_1 \leq_{cx} \varepsilon_2$ *and* f_2 *is log-concave, then* $\varepsilon_1 \leq_{sed} \varepsilon_2 \iff \varepsilon_1 \leq_{ent} \varepsilon_2$.

Proof. (a) The convex order $\varepsilon_1 \leq_{cx} \varepsilon_2$ can be represented by the following two integral relationships (S–S, p.110):

$$\int_{-\infty}^{b} F_1(z)dz \leq \int_{-\infty}^{b} F_2(z)dz \quad \text{for all } b, \tag{12}$$

$$\int_{b}^{\infty} [1 - F_1(z)]dz \leq \int_{b}^{\infty} [1 - F_2(z)]dz \quad \text{for all } b, \tag{13}$$

provided that the integrals exist. Summing up (12) and (13) and letting $b = 0$, we obtain

$$\begin{aligned} SED(F_1, F_0) &= \int_{-\infty}^{0} F_1(z)dz + \int_{0}^{\infty} [1 - F_1(z)]dz \\ &\leq \int_{-\infty}^{0} F_2(z)dz + \int_{0}^{\infty} [1 - F_2(z)]dz = SED(F_2, F_0). \end{aligned}$$

(b) This result is obtained from part (a) of this Proposition and part (d) of Lemma 3.

The equality of expectations implies that the results based on convex order are only applicable to equally biased (unbiased) forecast models. The following corollary encapsulates the results for the equally biased and unbiased forecast models.

Corollary 1.

(a) If $\varepsilon_1 \leq_{disp} \varepsilon_2$, *then the MAE, variance and entropy rank equally biased (unbiased) forecast models identically.*
(b) If $\varepsilon_1 \leq_{cx} \varepsilon_2$ *and* f_2 *is log-concave, then the MAE, variance and entropy rank equally biased (unbiased) forecast models identically.*

The inequality (12) is identical to an ordering referred to as the increasing concave order which with strict inequality at some b gives the second order stochastic dominance. The inequality (13) is identical to an ordering referred to as the increasing convex order. Each of

these two orderings is weaker than and implied by the usual (first order) stochastic dominance. However, two random variables with a common mean are convex ordered if and only if one of the two relationships, (12) or (13), holds. Therefore, neither (12) nor (13) alone is sufficient for ranking the forecasts by $SED(F, F_0)$.

The convex order with log-concavity is sufficient but not necessary for identical ranking of forecasts by MAE (SED), variance and entropy. For example, for the case of the Student-t family, the PDF is not log-concave for all ν, so parts (a) and (b) of Proposition 3 are applicable, but part (c) is not applicable to the entire family. However, the three measures order the family identically. The GE family is convex ordered decreasingly by β when, $\beta \leq 2$ and when $\beta_1 < 2, \beta_2 \geq 2$. The PDF is log-concave for $\beta \geq 1$. Yet, as seen earlier, for all $\beta > 0$, the MAE, variance and entropy order forecast errors identically.

4 Mean Excess Error

In this section, we consider the excess error defined by $|\varepsilon| - \tau$, given that $|\varepsilon| > \tau$, where $\tau \geq 0$ is a threshold for the forecast errors with $|\varepsilon| > \tau$ is costly and the errors with magnitudes smaller than a tolerance threshold being negligible. The loss associated with the excess error can be represented as

$$L(\tau, \varepsilon) = \begin{cases} 0, & |\varepsilon| \leq \tau \\ |\varepsilon| - \tau, & |\varepsilon| > \tau. \end{cases} \tag{14}$$

This is a three-piece linear loss function and clearly shows that forecasts with errors of magnitudes below the threshold τ are not penalised. This approach may be viewed as the counterpart of using a prediction interval where all values within the margin of error are acceptable, hence, a range of forecast errors are tolerated.

The mean excess absolute error $MEE(\tau)$ defined in (9) can also be represented as

$$MEE(\tau) = E_{|\varepsilon|>\tau}[L(\tau, \varepsilon)]. \tag{15}$$

Next, we give the SED representation of $MEE(\tau)$, point out some ordering properties and illustrate by a few examples.

4.1 Stochastic Error Distance Representation

Consider the weighted SED (3) with $p = 1$ and the weight function

$$w_\tau(\varepsilon) = \begin{cases} 0, & |\varepsilon| \leq \tau \\ \frac{1}{P(|\varepsilon|>\tau)}, & |\varepsilon| > \tau, \end{cases} \tag{16}$$

where τ is an error tolerance threshold. This weight function provides the measure given by the following proposition.

Proposition 4. *The weighted SED defined in* (3) *with* $p = 1$ *and weight function* (16) *is*

$$SED_\tau(F_\varepsilon, F_0) = MEE(\tau).$$

Proof. Using (16) in (3) with $p = 1$ gives

$$\begin{aligned} SED_\tau(F_\varepsilon, F_0) &= \int_{-\infty}^{\infty} |F_\varepsilon(z) - F_0(z)| w_\tau(z) dz \\ &= \frac{1}{P(|\varepsilon| > \tau)} \left[\int_{-\infty}^{-\tau} |F_\varepsilon(z) - F_0(z)| dz + \int_{\tau}^{\infty} |F_\varepsilon(z) - F_0(z)| dz \right] \\ &= \frac{1}{P(|\varepsilon| > \tau)} \left[\int_{-\infty}^{-\tau} F_\varepsilon(z) dz + \int_{\tau}^{\infty} [1 - F_\varepsilon(z)] dz \right] \\ &= \frac{1}{S_{|\varepsilon|}(\tau)} \int_{\tau}^{\infty} S_{|\varepsilon|}(z) dz \\ &= MEE(\tau). \end{aligned}$$

$SED_\tau(F_\varepsilon, F_0)$ is a dynamic generalisation of $SED(F_\varepsilon, F_0)$ due to the fact that $MEE(0) = E(|\varepsilon|) = SED_0(F_\varepsilon, F_0)$. The MEE order $\varepsilon_1 \leq_{mee} \varepsilon_2$ provides comparison of forecasts by $SED_\tau(F_\varepsilon, F_0)$. Because the order is defined for all $\tau \geq 0$, we have

$$\varepsilon_1 \leq_{mee} \varepsilon_2 \Longrightarrow \varepsilon_1 \leq_{sed} \varepsilon_2.$$

From this, representation (11), and Lemma 4, we have the following scheme of implications:

$$\begin{array}{ccccccc} Z_1 \leq_{\ell r} Z_2 & \Longrightarrow & Z_1 \leq_{hr} Z_2 & \Longrightarrow & Z_1 \leq_{mee} Z_2 & \Longrightarrow & Z_1 \leq_{sed} Z_2 \\ & & \Downarrow & & & & \\ & & Z_1 \leq_{st} Z_2 & \Longrightarrow & Z_1 \leq_{sed} Z_2. & & \end{array} \tag{17}$$

From part (a) of Lemma 2 and parts (a) and (b) of Lemma 4, we have the following result.

Proposition 5. *Let $\varepsilon_k, k = 1, 2$ be errors of two forecast models and g_k denote the PDFs of $|\varepsilon_k|, k = 1, 2$. If g_2 is decreasing and $|\varepsilon_1| \leq_{hr} |\varepsilon_2|$, then $\varepsilon_1 \leq_{mee} \varepsilon_2 \Longleftrightarrow \varepsilon_1 \leq_{ent} \varepsilon_2$.*

An MEE function can be monotone (decreasing or increasing) or non-monotone. In general, $MEE(\tau)$ is not available in a closed form. But the MEE order can be easily verified by the implications of order relations depicted in (17). The following example illustrates (17) and (10).

Example 3. *Consider the normal and logistic error models shown in Table A1. The expressions for $S(\tau)$ and $MEE(\tau)$ of these models can be written as follows.*

$$\begin{array}{lll} \textit{Normal:} & S_{|\varepsilon|}(\tau) = 2[1 - \Phi(\tau)], & MEE(\tau) = \frac{\phi(\tau)}{1-\Phi(\tau)} - \tau, \quad \tau \geq 0 \\ \textit{Logistic:} & S_{|\varepsilon|}(\tau) = \frac{2e^{-\tau}}{1+e^{-\tau}}, & MEE(\tau) = \frac{2\log(1+e^{-\tau})}{S(\tau)}, \quad \tau \geq 0, \end{array}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF. The expression for the normal model is known (Poynor, 2010)*, and the derivation of expression for the logistic model is shown in the Appendix. Showing the existence or lack of an order relation between the MEE's of these two models is tedious, at best. However, it is easy to show if the distributions of ε_1 and ε_2 are $N(0, 1)$ and $LG(0, 1)$, respectively, then $\varepsilon_1 \leq_{\ell r} \varepsilon_2$. Thus, by* (17)*, $|\varepsilon_1| \leq_{mee} |\varepsilon_2|$. That is, according to $SED_\tau(F_{|\varepsilon|}, F_0)$, the $N(0, 1)$ model is preferred to $LG(0, 1)$.*

Figure 3 illustrates (10) *with $N(0, 1)$ and $LG(0, 1)$ error models. The following points are noteworthy.*

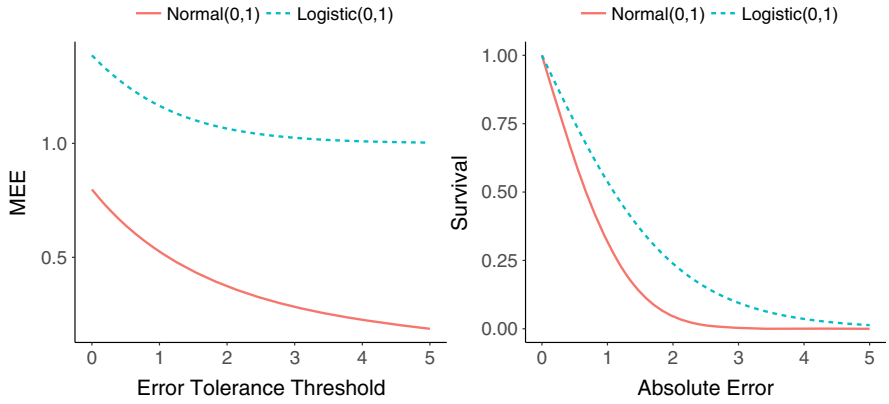

Figure 3. *The mean excess error plots (left) and survival plots (right) of a normal and a logistic error models. MEE, mean excess error. [Colour figure can be viewed at wileyonlinelibrary.com]*

(a) For both models, $MEE(\tau)$ is decreasing, which implies that the more tolerance is allowed, the lower will be the loss. Hence, the MAE is the maximum loss, $SED_\tau(F_{|\varepsilon|}, F_0) \leq SED(F_{|\varepsilon|}, F_0)$ for all $\tau \geq 0$.

(b) The $MEE(\tau)$ plot also shows that the gap between the two curves is increasing, hence, (10) holds and $SED(F_{|\varepsilon|}, F_0)$ gives the same preference, for which the survival plot provides visualisation.

The MEE function uniquely determines the distribution as follows. For a non-negative random variable Z with a finite mean,

$$S(z) = \frac{MEE(0)}{MEE(\tau)} \exp\left\{-\int_0^z \frac{1}{MEE(\tau)} d\tau\right\}.$$

The following example illustrates a well-known case.

Example 4. *Let $MEE(\tau) = a\tau + b$, $a > -1, b > 0$. Oakes and Dasu (1990) showed that the survival function corresponding to the linear MEE is*

$$S(z) = \left(\frac{b}{az + b}\right)^{1/a+1}, \quad z \geq 0, \quad a > -1, b > 0. \tag{18}$$

The case of $-1 < a < 0$, (18) can include the survival function of a Beta distribution with unbounded PDF, which is not a suitable model for $|\varepsilon|$. The case of $a \geq 0$ gives survival function of the generalized Pareto distribution for the absolute error shown in Table A1, where $a = 1/\alpha$ and $b = \alpha/(\alpha - 1), \alpha > 1$. For $a = b > 0$, (18) is the survival function of the Pareto type II distribution $P(\alpha, \sigma)$ where $\alpha = 1/a$, $\sigma = \alpha$, and

$$MEE(\tau) = \frac{\alpha + \tau}{\alpha - 1}, \quad \alpha > 1.$$

The exponential distribution is the limit as $\alpha \to \infty$ that gives $MEE(\tau) = 1$ for the Laplace error distribution. Figure 4 shows plots of three linear $MEE(\tau)$ and the corresponding survival plots of three DGP error models (Laplace when $\alpha \to \infty$ ($a \to 0$) and two double-Pareto with $\alpha = 2, 4$ ($a = .5, .25$)). For the Pareto models, $MEE(\tau)$ is increasing, which implies that the more tolerance is allowed, the higher will be the loss, and $MEE(\tau) \geq MEE(0) = MAE(\varepsilon)$.

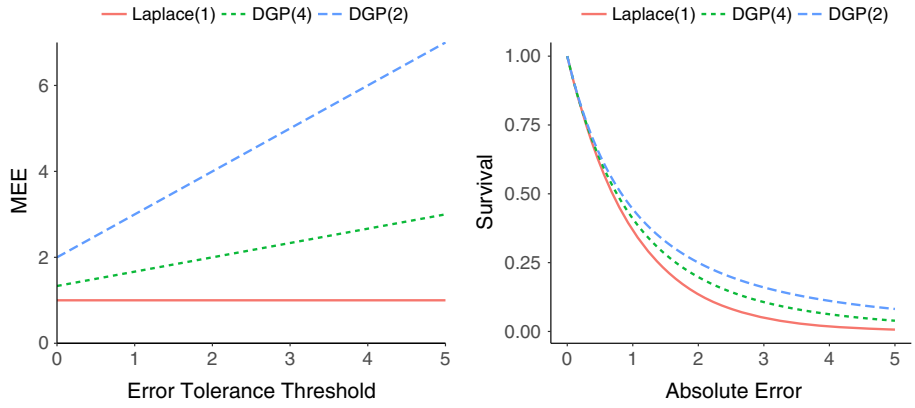

Figure 4. *Plots of three linear mean excess error functions (left) and the corresponding survival plots (right); Laplace(1)=DGP(∞). MEE, mean excess error. [Colour figure can be viewed at wileyonlinelibrary.com]*

Hence, the $MAE(e)$ *is the minimum expected loss,* $SED_\tau(F_\varepsilon, F_0) \geq SED(F_\varepsilon, F_0)$ *for all* $\tau \geq 0$.

4.2 Bayes Risk of MEE

In the decision theoretic framework, as the mean of the PDF (8), $MEE(\tau)$ is the optimal prediction of $|\varepsilon|-\tau \mid |\varepsilon| > \tau$ under the conditional quadratic loss $L^*(\tau, \varepsilon) = (|\varepsilon|-\tau-d)^2 \mid |\varepsilon| > \tau$. That is,

$$d^*(\tau) = \arg\min_d E_{|\varepsilon|>\tau}[L^*(d, |\varepsilon| - \tau)] = MEE(\tau), \quad |\varepsilon| > \tau > 0.$$

As a function of the threshold τ, $MEE(\tau)$ is a local risk measure, conditional on the threshold. By (9), $MEE(\tau)$ is minimum local risk conditional on the threshold, and by (15), $MEE(\tau)$ is a local risk conditional on the threshold.

The global risk of $MEE(\tau)$ is given by the Bayes risk,

$$MEER_\tau = E_\tau[MEE(\tau)],$$

where E_τ is the expectation with respect to the PDF (prior) for the threshold. (The brackets are used for the expectation operation). When the distribution of the threshold is the same as the distribution of the absolute error, $\pi(\tau) = g_{|\varepsilon|}(\tau), \tau \geq 0$, then we have

$$\begin{aligned} MEER_\tau &= E_\tau[MEE(\tau)] \\ &= \int_0^\infty MEE(\tau) g_{|\varepsilon|}(\tau) d\tau \\ &= h(S_{|\varepsilon|}), \end{aligned} \tag{19}$$

where $h(S_{|\varepsilon|})$ is the entropy functional of survival function defined by

$$h(S_{|\varepsilon|}) = -\int_0^\infty S_{|\varepsilon|}(\tau) \log S_{|\varepsilon|}(\tau) d\tau \geq 0. \tag{20}$$

A proof of (19) is given by Asadi and Zohrevand (2007).

The measure in (20) was introduced by Rao *et al.* (2004) as an alternative to Shannon entropy and was called cumulative residual entropy. They showed the inequality in (20) and the inequality becomes equality if and only if $F(\varepsilon)$ is a degenerate distribution at a point ε_0. As such, (20) is a measure of concentration of the distribution, but unlike Shannon's and other entropies, its global maximum is not the uniform distribution.

Bounds for $h(S_{|\varepsilon|})$ in terms of $H(|\varepsilon|)$, $E(|\varepsilon|)$ and $E(|\varepsilon|^2)$ are available (Rao *et al.*, 2004, Theorems 8 & 10). Using the relationships between these measures for the absolute error and the distributions shown in first column of Table A1, we represent these bounds as follows:

$$e^{H(\varepsilon)-H_\gamma} \leq MEER_\tau \leq \frac{V(\varepsilon)}{2MAE(\varepsilon)}, \tag{21}$$

where $\gamma \approx .5772\ldots$ is the Euler constant and $H_\gamma = \gamma + 1 + \log 2$ is the entropy of the Laplace distribution with scale parameter $e^\gamma \approx .5615$ (proof is given in the Appendix). These bounds are useful when the $MEER_\tau$ is not available in closed form. Note that if $\varepsilon^* = \varepsilon/\sigma$, then the lower and upper bounds of $MEER_\tau$ for ε^* will be σ multiple of the bounds in (21).

The following example illustrates $MEER_\tau$.

Example 5.

(a) Among the error models listed in Table A1, the Laplace, logistic and DGP have closed form $MEER_\tau$. The expressions for $MEER_\tau$ of these models are as follows.

$$\begin{aligned}
&\text{Laplace}: \ MEER_\tau = 1,\\
&\text{Logistic}: \ MEER_\tau = \tfrac{\pi^2}{6} - (\log 2)^2,\\
&\text{DGP}: \quad MEER_\tau = \tfrac{\alpha^2}{(\alpha-1)^2}, \quad \alpha > 2;
\end{aligned}$$

the measure for Laplace error model is easily seen and derivations of other two measures are shown in the Appendix. The bounds for the Laplace give $.56 \leq MEER_\tau \leq 1$, but for the logistic give a much wider interval $.76 \leq MEER_\tau \leq 3.73$. The $MEE(\tau)$ of the logistic model is decreasing, thus $MEER_\tau < MAE(\varepsilon)$ and the $MEE(\tau)$ of the DGP model is increasing, thus $MEER_\tau > MAE(\varepsilon)$.

(b) For $N(0,1)$ error model, (21) *gives the following bounds.*

$$\text{Normal}: \ .341\sqrt{\tfrac{\pi}{2}} \leq MEER_\tau \leq .5\sqrt{\tfrac{\pi}{2}}.$$

These bounds give a range of less than one-fifth of the error standard deviation.

(c) Bounds for the GE and the Student-t families can be easily computed using their measures shown in Table A1. Figure 5 shows plots of the bounds for these models against the error standard deviation. The right panel shows the plots of $MEER_\tau$ and the bounds for the DGP family that can serve as a reference. The DGP plots suggest that $MEER_\tau$ begins at close to the upper bound and moves toward the lower bound as the standard deviation increases.

The MEER order of two forecast errors with $MEER_{\tau,k}, k = 1, 2$, denoted as $\varepsilon_1 \leq_{meer} \varepsilon_2$, is defined by $MEER_{\tau,1} \leq MEER_{\tau,2}$. Clearly, for two forecast errors $\varepsilon_k, k = 1, 2$,

$$\varepsilon_1 \leq_{mee} \varepsilon_2 \Longrightarrow \varepsilon_1 \leq_{meer} \varepsilon_2.$$

Thus, the implications of order relations shown in (17) also apply to $MEER_\tau$.

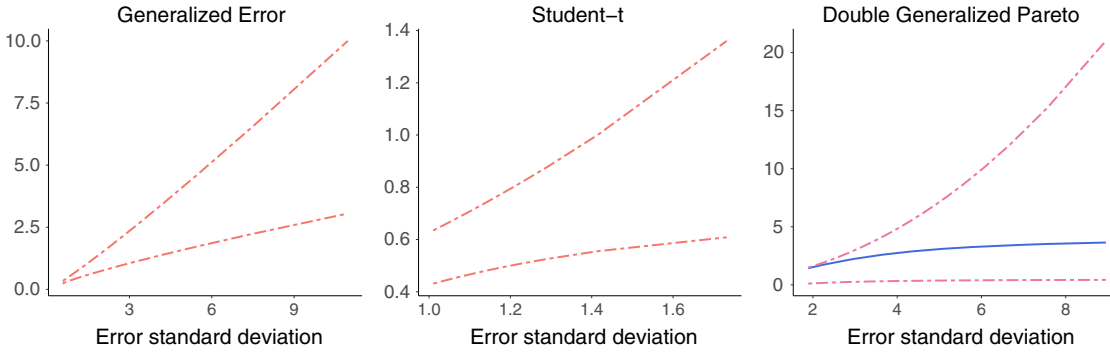

Figure 5. *Upper and lower bounds for MEER for the GE, Student-t ($v > 2$) and DGP ($a > 2$) error distributions as functions of the error standard deviation (right panel includes MEER for the DGP error distribution). MEER, mean excess error risk; GE, generalized error. [Colour figure can be viewed at wileyonlinelibrary.com]*

An implication of stochastic order for $MEER_{\tau,k}, k = 1, 2$ is as follows:

$$|\varepsilon_1| \leq_{st} |\varepsilon_2| \Longrightarrow MEER_{\tau,1} \leq MEER_{\tau,2} - MAE(\varepsilon_1) \log \frac{MAE(\varepsilon_1)}{MAE(\varepsilon_2)};$$

(Navarro *et al*., 2010, Proposition 2.1).

Proposition 6. *Let ε_k be a forecast error with distribution G_k with PDF g_k, $k = 1, 2$. If $\varepsilon_1 \leq_{disp} \varepsilon_2$, then $\varepsilon_1 \leq_{meer} \varepsilon_2$.*

Proof. The dispersive order can be represented in terms of PDFs as follows (S–S, p.149):

$$g_2\left(S_2^{-1}(u)\right) \leq g_1\left(S_1^{-1}(u)\right), \text{ for all } 0 < u < 1. \tag{22}$$

Letting $u = S(\tau)$ in (20) we have

$$\begin{aligned} MEER_{\tau,1} &= -\int_0^\infty S_1(\tau) \log S_1(\tau) d\tau \\ &= \int_0^1 u \log u \left[g_1(S_1^{-1}(u))\right]^{-1} du \\ &\leq \int_0^1 u \log u \left[g_2(S_2^{-1}(u))\right]^{-1} du \\ &= MEER_{\tau,2}. \end{aligned}$$

The inequality is found by noting that $u \log u \leq 0$ and (22) implies that

$$\left[g_1\left(S_1^{-1}(u)\right)\right]^{-1} \leq \left[g_2\left(S_2^{-1}(u)\right)\right]^{-1}, \text{ for all } 0 < u < 1.$$

5 Mixture Models

Mixture models are used for error distribution to capture outliers (Cook, 1999). Comparison of the SED, variance and entropy of mixture models is revealing. Without loss of generality, we illustrate this for mixtures with two components. Let

$$F_\varepsilon(\varepsilon) = \pi F_1(\varepsilon) + (1-\pi)F_2(\varepsilon), \qquad 0 \le \pi \le 1. \tag{23}$$

The following closure property of the stochastic order is useful in the present problem.

Lemma 5 (S–S, Theorem 1.A.3)**.** *If* $Z_k \le_{st} Z_k^*, k = 1, 2$*, then*

$$\pi Z_1 + (1-\pi)Z_2 \le_{st} \pi Z_1^* + (1-\pi)Z_2^*.$$

Using this result, we have

$$\begin{aligned} SED(F_\varepsilon : F_0) &= \int_0^\infty [\pi S_1(z) + (1-\pi)S_2(z)]dz \\ &= \pi SED(F_1 : F_0) + (1-\pi)SED(F_2 : F_0). \end{aligned}$$

For the error variance, we have

$$\begin{aligned} V(\varepsilon) &= \pi\left\{V(\varepsilon_1) + [E(\varepsilon_1) - E(\varepsilon)]^2\right\} + (1-\pi)\left\{V(\varepsilon_2) + [E(\varepsilon_2) - E(\varepsilon)]^2\right\} \\ &\ge \pi V(\varepsilon_1) + (1-\pi)V(\varepsilon_2). \end{aligned}$$

The inequality becomes equality when $E(\varepsilon_k) = E(\varepsilon), k = 1, 2$. For the entropy, we have

$$H(\varepsilon) \ge \pi H(\varepsilon_1) + (1-\pi)H(\varepsilon_2). \tag{24}$$

Thus, $SED(F : F_0)$ is closed under mixtures, the variance is closed under mixtures of unbiased (equally biased) forecasts, but entropy increases under mixtures of errors.

The difference between the two sides of the inequality in (24) gives the Jensen–Shannon (JS) divergence of the mixture error model that has the following representations:

$$\begin{aligned} JS(f_\varepsilon : f_1, f_2, \pi) &= H(\varepsilon) - [\pi H(\varepsilon_1) + (1-\pi)H(\varepsilon_2)] \\ &= \pi K(f_1 : f_\varepsilon) + (1-\pi)K(f_2 : f_\varepsilon) \ge 0. \end{aligned}$$

The inequality becomes equality if and only if $f_k(\varepsilon) = f_\varepsilon(\varepsilon), k = 1, 2$ almost everywhere, implying that there is no outlier. Using (24) and an upper bound for JS given by Asadi *et al.* (2016), we have the following bounds for the error entropy

$$\pi H_1 + (1-\pi)H_2 \le H(\varepsilon) \le \pi H_1 + (1-\pi)H_2 + \pi(1-\pi)J(f_1, f_2),$$

where $H_k = H(\varepsilon_k), k = 1, 2$ and $J(f_1, f_2) = K(f_1 : f_2) + K(f_2 : f_1)$ is the Jeffreys divergence. Estimates of these bounds can be used to assess presence of outliers. For example, in the case of normal distributions, the bounds are functions of $\mu_k, \sigma_k, k = 1, 2$. When the estimates of the bounds do not differ substantially, presence of outliers is ruled out.

The following proposition gives representation of the MEE of a mixture model.

Proposition 7. *Let the distribution of the forecast error be the mixture* (23)*. Then,*

$$MEE_\varepsilon(\tau) = \pi(\tau)MEE_1(\tau) + (1-\pi(\tau))MEE_2(\tau), \tag{25}$$

where $MEE_k(\tau), k = 1, 2$ *is the measure associated with the survival function* S_k *of the component's absolute error* $|\varepsilon_k|, k = 1, 2$ *and* $\pi(\tau) = \frac{\pi S_1(\tau)}{S_{|\varepsilon|}(\tau)}$.

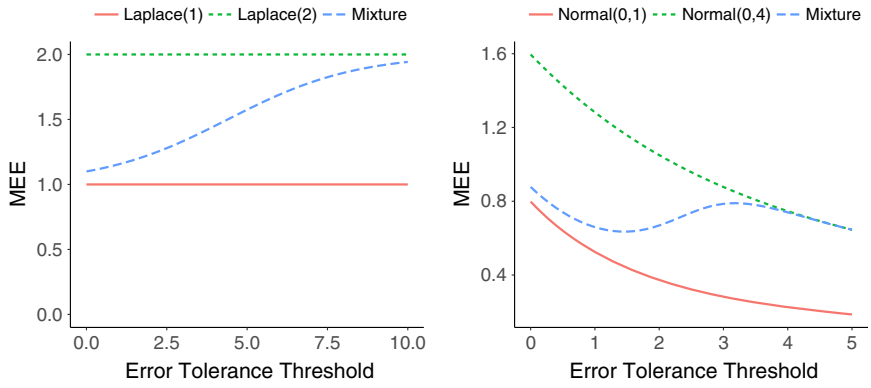

Figure 6. *The mean excess error plots of mixtures of two Laplace models (left) and two normal models (right). [Colour figure can be viewed at wileyonlinelibrary.com]*

Proof. Note that

$$\int_\tau^\infty S_{|\varepsilon|}(z)dz = \pi \int_\tau^\infty S_1(z)dz + (1-\pi)\int_\tau^\infty S_2(z)dz$$
$$= \pi S_1(\tau)MEE_1(\tau) + (1-\pi)S_2(\tau)MEE_2(\tau).$$

The result is obtained upon division by $S_{|\varepsilon|}(\tau)$.

Thus, for each τ, the $MEE_\varepsilon(\tau)$ is also a weighted average of the $MEE_k(\tau), k = 1, 2$. However, over a range of τ, the $MEE_\varepsilon(\tau)$ is a dynamic mixture of $MEE_k(\tau), k = 1, 2$. Proposition 7 implies the following order:

$$\varepsilon_1 <_{mee} \varepsilon_2 \Longrightarrow |\varepsilon_1| \leq_{st} |\varepsilon| \leq_{st} |\varepsilon_2|. \tag{26}$$

This order is known (S–S, Theorem 2.A.18), however, it is a direct consequence of (25). The following example illustrates Proposition 7.

Example 6. *We consider two mixture models.*

(a) Let F_ε be the mixture of two Laplace error models where the survival function of the absolute error $|\varepsilon|$ is

$$S_{|\varepsilon|}(z) = \pi e^{-z/\sigma_1} + (1-\pi)e^{-z/\sigma_2}, \quad 0 \leq \pi \leq 1.$$

It is well known that the exponential distribution is characterised by constant hazard rate and that the mixture of exponential has decreasing hazard rate. Hence, the $MEE(\tau)$ of the mixture of two Laplace error models is increasing. The left panel of Figure 6 illustrates (25) *and* (26) *for $p = .9, \sigma_1 = 1, \sigma_2 = 2$.*

(b) Let $F_\varepsilon = .9N(0,1) + .1N(0,4)$. It is easy to show that for $\sigma_1 \leq \sigma_2$, $|\varepsilon_1| \leq_{\ell r} |\varepsilon_2|$, thus by (17) *$|\varepsilon_1| \leq_{mee} |\varepsilon_2|$. The right panel of Figure 6 shows that for the two normal models, $MEE_\varepsilon(\tau)$ are increasingly ordered by σ. The $MEE_\varepsilon(\tau)$ of the mixture model is bounded between the MEEs of the two normal models, however, it is not monotone. For small τ, $\frac{S_1(\tau)}{S_2(\tau)} \approx 1$, thus $\pi(\tau) \approx .9$ and $S_{|\varepsilon|}(\tau)$ is close to $S_1(\tau)$. But as τ increases, $S_1(\tau)$ decreases much faster than $S_2(\tau)$, hence, $\pi(\tau)$ decreases and $S_{|\varepsilon|}(\tau)$ moves closer to $S_2(\tau)$.*

Finally, $h(S)$ is a concave functional of the survival function. Consequently, if the absolute error $|\varepsilon|$ has a mixture distribution as in (23), then

$$MEER_\tau \geq \pi MEER_{\tau,1} + (1-\pi)MEER_{\tau,2}.$$

That is, the MEER of a mixture is at least as large as the mixture of the MEERs of the constituents.

6 Empirical SED, MEE and MEER

The empirical $SED(F_\varepsilon : F_0)$, $MEE(\tau)$ and $MEER_\tau$ are defined in terms of the empirical distribution of the random error ε based on sample forecast errors, $e_i = \hat{y}_i - y_i, i = 1, \ldots, n$. The empirical CDF of ε and survival function of $|\varepsilon|$ are defined by

$$\hat{F}_\varepsilon(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(e_i \leq \varepsilon) \quad \text{and} \quad \hat{S}_{|\varepsilon|}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(z > |e_i|),$$

where $\mathbf{1}(A)$ is the indicator function of the event A. The empirical $SED(F_\varepsilon : F_0)$ is defined by using $\hat{S}_{|\varepsilon|}$ in (11):

$$\begin{aligned} SED_n(F_\varepsilon : F_0) = SED(\hat{F}_\varepsilon : F_0) &= \int_0^\infty \hat{S}_{|\varepsilon|}(z)dz \\ &= \sum_{i=1}^{n} \int_{|e_{(i-1)}|}^{|e_{(i)}|} \hat{S}_{|\varepsilon|}(z)dz \\ &= \sum_{i=1}^{n} (|e_{(i)}| - |e_{(i-1)}|)\left(1 - \frac{i-1}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} |e_{(i)}| \\ &= MAE_n, \end{aligned}$$

where $0 = |e_{(0)}| < |e_{(1)}| < \cdots < |e_{(n)}|$ are the ordered absolute errors. Under the usual sampling assumption, MAE_n is a consistent estimator of MAE.

The empirical $MEE(\tau)$ corresponding to the MAE_n is defined by using $\hat{S}_{|\varepsilon|}$ in (9):

$$\begin{aligned} MEE_n(\tau) &= \frac{1}{\hat{S}_{|\varepsilon|}(\tau)} \int_\tau^\infty \hat{S}_{|\varepsilon|}(z)dz \cdot \mathbf{1}(\tau < \max_{1\leq i\leq n} |e_i|) \\ &= \frac{1}{n\hat{S}_{|\varepsilon|}(\tau)} \sum_{i=1}^{n} (|e_i| - \tau) \cdot \mathbf{1}(|e_i| > \tau) \\ &= \frac{1}{n_\tau} \sum_{i=1}^{n} (|e_i| - \tau) \cdot \mathbf{1}(|e_i| > \tau), \end{aligned}$$

where $n_\tau = \sum_{i=1}^{n} \mathbf{1}(|e_i| > \tau)$. Under the usual sampling assumption, $MEE_n(\tau)$ is a uniformly strong consistent estimator on the fixed interval $0 \leq |\varepsilon| \leq T < \infty$ (Yang, 1978, Lemma 2). Hall and Wellner (1979) extended Yang's result to the case of $|\varepsilon| > 0$ under some optimal moment conditions.

The empirical $MEER_\tau$ corresponding to $MEE_n(\tau)$ is given by using $\hat{S}_{|\varepsilon|}$ in (20):

$$\begin{aligned} MEER_n = h(\hat{S}_{|\varepsilon|}) &= -\int_0^\infty \hat{S}_{|\varepsilon|}(z)\log \hat{S}_{|\varepsilon|}(z)dz \\ &= -\sum_{i=1}^{n}\int_{|e_{(i-1)}|}^{|e_{(i)}|} \hat{S}_{|\varepsilon|}(z)\log \hat{S}_{|\varepsilon|}(z)dz \\ &= -\sum_{i=1}^{n}\int_{|e_{(i-1)}|}^{|e_{(i)}|} \left(1-\frac{i-1}{n}\right)\log\left(1-\frac{i-1}{n}\right)dz \\ &= -\sum_{i=1}^{n}(|e_{(i)}|-|e_{(i-1)}|)\left(1-\frac{i-1}{n}\right)\log\left(1-\frac{i-1}{n}\right). \end{aligned}$$

Under the usual sampling assumption, the estimator $MEER_n$ is almost surely consistent (Rao *et al.*, 2004, Theorem 9).

The empirical $MEE(\tau)$ and $MEER_\tau$ provide descriptive statistics that can be used for forecast evaluation along with the sample MAE and MSE. (R codes for computing $MEE(\tau)$ and $MEER_\tau$ are given in the Appendix). The statistical properties of these and other descriptive measures hinge on e_i's representing a sample from the distribution of ε. This assumption is reasonable for cross-sectional forecasting where e_i's are computed using a hold-out sample. Nonetheless, this is the underlying assumptions for assortments of all descriptive measures used in forecasting literature for evaluating all sorts of forecast models, including for time series forecasting; see, for example, Hyndman and Koehler (2006) and references therein and Bauer and Hamilton (2015).

6.1 Bond Risk Premia

This empirical example illustrates the use of $MEE_n(\tau)$ and $MEER_n(\tau)$ and compares their results with the sample root mean squared error ($RMSE_n$) and SED (MAE_n) in the context of bond risk premia forecasting. Several models for bond return forecasting have been used in the literature; see, for example, Diebold *et al.* (2006) and Bauer and Hamilton (2015) (hereafter written as B–H). An exhaustive comparison of the models used in the literature is beyond the scope of this paper. For our illustrative purpose, we use the same variables as Cochrane and Piazzesi (2005) and B–H defined in our notations as follows.

Log price of n-year discount bond at time t	$p_t^{(n)}$
Log yield	$X_t^{(n)} = -\frac{1}{n}p_t^{(n)}$
Average four year yields $X_t^{(2)}, X_t^{(3)}, X_t^{(4)}, X_t^{(5)}$	X_t^{AVG}
Log return, buying n-year bond at t and selling it at $t+1$	$r_{t+1}^{(n)} = p_{t+1}^{(n-1)} - p_t^{(n)}$
Excess log returns	$Y_{t+1}^{(n)} = r_{t+1}^{(n)} - X_t^{(1)}$
Average four year returns $Y_{t+1}^{(2)}, Y_{t+1}^{(3)}, Y_{t+1}^{(4)}, Y_{t+1}^{(5)}$	Y_{t+1}^{AVG}
Covariance matrix of $\alpha = (X_t^{(2)}, X_t^{(3)}, X_t^{(4)}, X_t^{(5)}, X_t^{AVG})$	Σ_X
PCs of Σ_X	$W_{1t}, W_{2t}, W_{3t}, W_{4t}, W_{5t}$

Following B–H, we use the CRSP Fama-Bliss yields and bond prices for the period January 1964 to December 2002 for estimating the models and the period January 2003–December 2012 for forecasting. This provides 468 data points for estimation and 120 data points for forecasting. We also follow B–H and compute the PC loadings for the estimation sample period and use them to compute the PCs for the forecasting period. We use a rolling regression scheme to produce one-step-ahead forecasts (B–H have used a fixed scheme to forecast the excess returns for the entire period January 2003–December 2013.) We scale forecast errors by the standard deviation of the forecast error.

The first three PCs, 'commonly labeled level, slope and curvature', have been hypothesized to represent 'all the relevant information for predicting future yields' (B–H). However, Cochrane and Piazzesi (2005) 'reported evidence that information in the fourth and fifth PC of yields has predictive power'. B–H reported that the model with the first three PCs produced better RMSE forecasts than the model with all five PCs and noted that both of these models were outperformed by the unconditional mean model (intercept). But B–H concluded that 'To estimate bond risk premia in a robust way, we recommend using only those predictors that consistently show a strong associations with excess bond returns, namely the level and the slope of the yield curve'.

We rank the forecasting performances of each aspect of the yield curve represented by first three PCs and all five PCs by the sample versions of the RMSE, MAE, MEER and MEE. More formally, we consider the regression models for forecasting the expected log returns using subsets of the principal components defined as follows:

$$Y_{t+1}^{(n)} = \alpha_0 + \sum_{j=1}^{k} \alpha_j W_{jt} + \epsilon_{t+1}, \quad n = 2, 3, 4, 5, \text{“A}VG\text{"}, \; k = 3, 5.$$

We include the conditional mean (intercept) model and the random walk model as benchmarks. In addition, we compare the forecasting performance of the regression models with two $ARIMA$ models for each maturity. The autoregressive integrated moving average (ARIMA) models were prompted by noticing non-stationary and autoregressive behaviours of the maturities.

Table 2 reports the forecast error measures of predicting excess returns. Panels (a)–(d) are for the 2–5 years maturities and panel (e) is for the average excess returns of the maturities over 2–5 years. For each maturity, the table gives the results for three and five PC regression, the unconditional mean, the random walk and two ARIMA models. The first ARIMA is identified as the optimal model by an R program that first tests for the unit root and then selects the model using AIC (Hyndman and Khandakar, 2008), but ARIMA(1,1,0) shows better performance for most cases. For each model, the table reports the average and standard deviation of forecast errors, the magnitude and rank (shown in the parentheses) of each forecast error measure and the equivalence of rankings by the three measures. All three measures are computed using the scaled forecast errors by the forecast standard deviation (the scaling makes the measures for various models comparable). The results are as follows.

- All three error measures rank the first three PCs better than all five PCs, but both of these models are outperformed by the unconditional mean model (intercept) that also beats the random walk model. In almost all cases, the two ARIMA models beat the random walk.
- The RMSE and MAE rank ARIMA(1,1,0) as the best for the 2-, 3- and 4-year bonds and ARIMA(0,1,0) as the best for the 5-year bond.

Table 2. *One-step-ahead forecast error measures of predicting excess returns by return-forecasting factors, the unconditional mean and time series models.*

	Forecast error		Forecast error scaled by SD			Equivalent
	AVG	SD	RMSE	MAE	MEER	rankings
a) Two-year bond						
Three PCs	0.758	1.167	1.189 (5)	0.909 (5)	0.769 (5)	All three
All five PCs	0.979	1.145	1.313 (6)	1.031 (6)	0.806 (6)	All three
Unconditional mean	−0.111	0.980	1.002 (3)	0.811 (3)	0.559 (1)	RMSE & MAE
Random walk	0.002	0.004	1.133 (4)	0.882 (4)	0.706 (3)	RMSE & MAE
ARIMA(1,1,1)	−0.011	0.099	1.001 (2)	0.781 (2)	0.638 (2)	All three
ARIMA(1,1,0)	−0.001	0.028	0.996 (1)	0.732 (1)	0.712 (4)	RMSE & MAE
b) Three-year bond						
Three PCs	1.445	2.214	1.191 (5)	0.909 (5)	0.779 (5)	All three
All five PCs	1.976	2.154	1.354 (6)	1.063 (6)	0.832 (6)	All three
Unconditional mean	0.192	1.864	1.001 (3)	0.812 (2)	0.574 (2)	MAE & MEER
Random walk	0.004	0.007	1.138 (4)	0.882 (4)	0.723 (4)	All three
ARIMA(1,1,1)	−0.002	0.220	0.996 (2)	0.821 (3)	0.550 (1)	None
ARIMA(1,1,0)	−0.003	0.094	0.995 (1)	0.756 (1)	0.666 (3)	RMSE & MAE
c) Four-year bond						
Three PCs	2.191	3.111	1.220 (5)	0.943 (5)	0.790 (5)	All three
All five PCs	2.880	3.046	1.373 (6)	1.071 (6)	0.853 (6)	All three
Unconditional mean	0.624	2.699	1.022 (3)	0.836 (3)	0.566 (1)	RMSE & MAE
Random walk	0.006	0.010	1.180 (4)	0.927 (4)	0.729 (4)	All three
ARIMA(1,1,1)	−0.003	0.293	0.996 (2)	0.802 (2)	0.580 (2)	All three
ARIMA(1,1,0)	−0.004	0.122	0.995 (1)	0.776 (1)	0.625 (3)	RMSE & MAE
d) Five-year bond						
Three PCs	2.945	3.793	1.263 (5)	0.989 (5)	0.788 (4)	RMSE & MAE
All five PCs	3.644	3.762	1.389 (6)	1.099 (6)	0.833 (5)	RMSE & MAE
Unconditional mean	1.219	3.459	1.056 (3)	0.873 (3)	0.561 (1)	RMSE & MAE
Random walk	0.008	0.012	1.223 (4)	0.973 (4)	0.742 (3)	RMSE & MAE
ARIMA(0,1,0)	0.006	0.267	0.996 (1)	0.586 (1)	0.850 (6)	RMSE & MAE
ARIMA(1,1,0)	−0.003	0.061	0.997 (2)	0.766 (2)	0.652 (2)	All three
e) Average						
Three PCs	1.835	2.539	1.230 (5)	0.950 (5)	0.793 (5)	All three
All five PCs	2.370	2.497	1.376 (6)	1.075 (6)	0.848 (6)	All three
Unconditional mean	0.481	2.207	1.019 (3)	0.833 (3)	0.560 (1)	RMSE & MAE
Random Walk	0.005	0.008	1.183 (4)	0.929 (4)	0.729 (4)	All three
ARIMA(0,1,2)	−0.025	0.257	1.001 (2)	0.762 (1)	0.650 (3)	None
ARIMA(1,1,0)	−0.003	0.093	0.996 (1)	0.771 (2)	0.637 (2)	MAE & MEER

Measures are calculated using standardised forecast errors; the number in the parentheses is the rank of the forecast model according to the measure. RMSE, root mean squared error; MAE, mean absolute error; MEER, mean excess error risk; PC, principal component.

- The MEER ranks the unconditional mean model as the best for all cases except the 3-year bond where this model is ranked second after the ARIMA(0,1,0).
- In many cases, the rankings by all three error measures agree and by pairs of these measures agree. The table also shows two cases where all the three measures disagree.

Figure 7 shows the plots of the $MEE(\tau)$ of predicting excess returns by the models shown in Table 2. Each panel in Figure 7 shows the plots of $MEE(\tau)$ of all models of a bond maturity for $\tau \leq .5$. In most cases, the plots are decreasing with the same rates. These plots show that the PC models are dominated by the other models and except for the 5-year bond where the plot for ARIMA(0,1,0) is increasing and crosses the plots for the PC models. These switches illustrate that the tolerance level threshold for forecast errors can affect the ranks of forecast models.

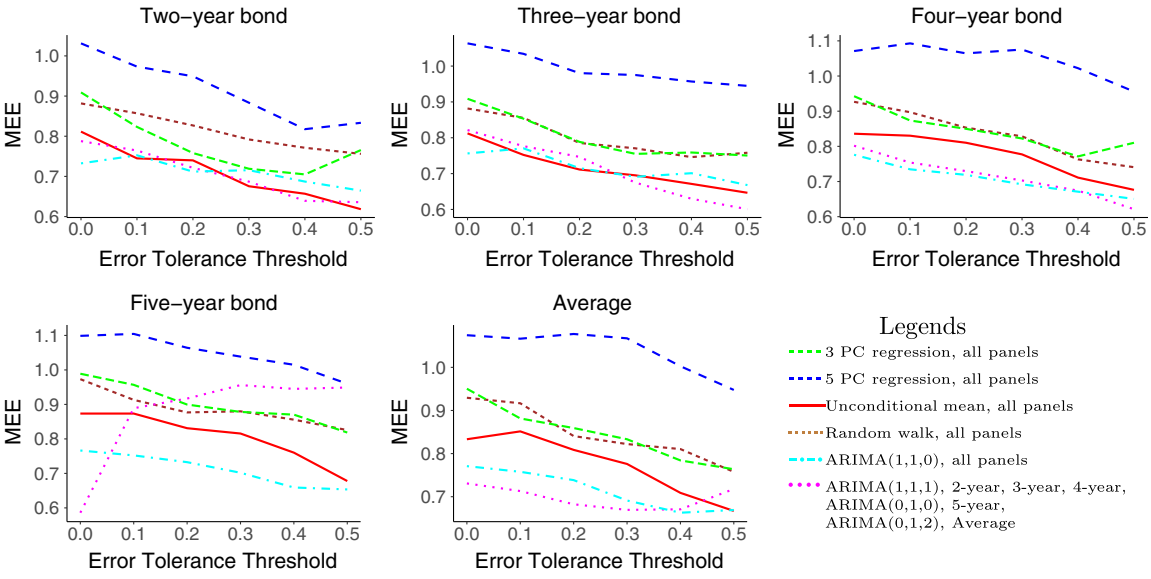

Figure 7. *Mean excess error functions of six forecast models for excess returns. [Colour figure can be viewed at wileyonlinelibrary.com]*

7 Summary and Conclusions

We presented an interface of the econometric, information theory and reliability literatures. By drawing on ideas from the information theory and reliability literatures, we built on Diebold and Shin's (2017) work on ranking point forecasts based on integrated distances between the distribution of the forecast error and the distribution of the error of the perfect forecast. The basic element of D–S is the SED representation of the MAE. We identified a sufficient condition in terms of the stochastic order for the identical rankings of forecasts by the MAE and the error entropy. We also have identified sufficient conditions in terms of the convex and dispersive orders for the identical rankings of forecasts by the MAE, the error variance and the error entropy.

Generalisations of the SED in terms of its weighted versions are presented by D–S. We have introduced the notion of excess error and a weight function that provides the weighted SED representation of the MEE function. The associated loss function allows the user to specify a tolerance threshold τ below which the forecast errors of magnitudes are not penalised. As a function of τ, the MEE absolute error function, $MEE(\tau)$, is a dynamic measure. For each τ, it is a local risk (expected loss) function. The Bayes risk of $MEE(\tau)$, which is defined by the expectation with respect to a distribution for τ, provides a measure of global risk. Using the distribution of the absolute error for τ gives the global risk, as the entropy functional of the survival function of the absolute error. This measure, which has been called in the literature by such names as cumulative residual entropy and survival entropy, appears in the context of this paper as a risk function, hence, we have referred to it as the mean excess error risk ($MEER$). The implication of the dispersive order for ranking forecasts by this measure is noted. Extension to the case of penalising the errors with magnitudes below τ can be obtained through application of the notion of mean past lifetime (down time) of a system. Developing Bayes risks under other distributions for τ will provide interesting and challenging research problems.

We have illustrated the relationships between SED, $MEE(\tau)$ and $MEER$ for some broad families of error distributions such as the generalised error (power exponential), Student-t and the DGP models for the distribution of the forecast error.

We presented empirical $MEE(\tau)$ and $MEER$ and used them as descriptive statistics for ranking forecast models. The illustrative example pertains to the principal components of yields as predictors of the excess bond returns. We used one-step-ahead forecasts for this purpose. In most cases, the rankings by RMSE, MAE and MEER agreed, in several cases, the rankings by two of them agreed, and in two cases, the rankings by none of the three measures agreed. In our one-step-ahead forecast comparisons of the PCs, all three measures ranked the intercept model better than the three and five PC models. These rankings confirmed the RMSE ranking of these models by B–H and their finding regarding the unconditional mean (intercept) model beating both PC models. We showed that the PC models are also outperformed by the random walk models for all the maturities under consideration. As noted by a reviewer of this paper, 'if it is true in the population that the intercept beats all of these indicators, then all of these indicators are irrelevant noise'. However, we also showed that two ARIMA models for the excess returns outperform the unconditional mean model and random walk.

Acknowledgements

The authors thank three reviewers for their constructive comments that led to improving the exposition of this article, the Co-editor, Nalini Ravishanker, for the encouragements, and Francis Diebold for commenting on an earlier draft of this paper. Ebrahimi's research was partially supported by a grant from NSF, DMS 1208273. Soofi's research was supported by a Roger L. Fitzsimonds Distinguished Scholar Award.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory,* Eds. B.N. Petroy & F. Csaki, pp. 267–281. Budapest: Akademiai Kiado.

Asadi, M. & Zohrevand, Y. (2007). On the dynamic cumulative residual entropy. *J. Statist. Plan. Infer.*, **137**, 1931–1941.

Asadi, M., Ebrahimi, N., Soofi, E. S. & Zarezadeh, S. (2014). New maximum entropy methods for modeling lifetime distributions. *Nav. Res. Logistics*, **61**, 427–434.

Asadi, M., Ebrahimi, N., Soofi, E. S. & Zohrevand, S. (2016). Jensen-Shannon information of the coherent system lifetime. *Reliab. Engin. Syst. Safety*, **156**, 244–255.

Bauer, M. D. & Hamilton, J. D. (2015). Robust bond risk premia, Federal Reserve Bank of San Francisco Working Paper.

Clarke, B. S. & Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plan. Infer.*, **41**, 37–60.

Cochrane, J. H. & Piazzesi, M. (2005). Bond risk premia. *Amer. Econ. Rev.*, **95**, 138–160.

Cook, D. (1999). Graphical detection of regression outliers and mixtures. In *Proceed. ISI 99*, Vol. 2. Helsinki: Edita Ltd., pp. 103-106.

Diebold, F. X. & Shin, M. (2017). Assessing point forecast accuracy by stochastic error distance. *Econometric Rev.*, **36**, 588–598.

Diebold, F. X., Rudebusch, G. D. & Aruoba, S. B. (2006). The macroeconomy and the yield curve: A dynamic latent factor approach. *J. Econometrics*, **131**, 309–338.

Ebrahimi, N., Maasoumi, E. & Soofi, E. S. (1999). Ordering univariate distributions by entropy and variance. *J. Econometrics*, **90**, 317–336.

Ebrahimi, N., Soofi, E. S. & Soyer, R. (2013). When are observed failures more informative than observed survivals. *Nav. Res. Logistics*, **60**, 102–110.

Ebrahimi, N., Soofi, E. S. & Soyer, R. (2010). Information measures in perspective. *Internat. Statist. Rev.*, **78**, 383–412.

Ebrahimi, N., Jalali, N. Y., Soofi, E. S. & Soyer, R. (2014). Importance of components for a system. *Econometric Rev.*, **33**, 395–420.

Ghosh, M. & Yang, M. C. (1988). Simultaneous estimation of Poisson means under entropy loss. *Ann. Statist.*, **16**, 278–291.

Hall, W. J. & Wellner, J. A. (1979). *Estimation of mean residual life*. Unpublished manuscript, University of Washington.
Hyndman, R. J. & Khandakar, Y. (2008). Automatic time series forecasting: the forecast Package for R. *J. Statist. Software*, **27**, 1–22.
Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *Internat. J. Forecasting*, **22**, 679–688.
Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.
Kullback, S. (1959). *Information Theory and Statistics.* New York: Wiley. (reprinted in 1968 by Dover).
Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.*, **27**, 986–1005.
Navarro, J., del Aguila, Y. & Asadi, M. (2010). Some new results on the cumulative residual entropy. *J. Statist. Plan. Infer.*, **140**, 310–322.
Oakes, D. & Dasu, T. (1990). A note on residual life. *Biometrika*, **77**, 409–410.
Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scand. J. Statist.*, **8**, 154–168.
Psarakis, S. & Panaretoes, J. (1990). The folded t distribution. *Comm. Statist. A, Theory Meth.*, **19**, 2717–2734.
Poynor, V. A. (2010). *Bayesian Inference for Mean Residual Life Functions in Survival Analysis*, M.S Thesis, University of California Santa Cruz.
Rao, M., Chen, Y., Vemuri, B. C. & Wang, F. (2004). Cumulative residual entropy: a new measure of information. *IEEE Trans. Inform. Theory*, **50**, 1220–1228.
Shaked, M. (1982). Dispersive ordering of distributions. *J. Appl. Probab.*, **19**, 310–320.
Shaked, M. & Shanthikumar, J. G. (2007). *Stochastic Orders.* Netherlands: Springer.
Soofi, E. S. & Retzer, J. J. (2002). Information indices: unification and applications. *J. Econometrics*, **107**, 17–40.
Yang, G. L. (1978). Estimation of a biometric function. *Ann. Statist.*, **6**, 112–116.
Yu, Y. (2008). On an inequality of Karlin and Rinott concerning weighted sums of i.i.d. random variables. *Adv. Appl. Probab.*, **40**, 1223–1226.
Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley. (reprinted in 1996 by Wiley).
Zografos, K. & Nadarajah, S. (2005). Survival exponential entropies. *IEEE Trans. Inform. Theory*, **51**, 1239–1246.

Appendix

A.1 Error Models

Table A1 lists five error models and gives the PDF of the absolute errors, $g_{|\varepsilon|}$, and the expressions for the $MAE(\varepsilon)$, $H(\varepsilon)$ and $V(\varepsilon)$ of the error distributions. The first row of the table gives the relationships between f_ε and $g_{|\varepsilon|}$ and $MAE(\varepsilon)$, $H(\varepsilon)$ and $V(\varepsilon)$ and the moments and entropy of $g_{|\varepsilon|}$. The expressions for the mean of half-normal, half-logistic, generalised gamma and generalized Pareto are well known and the mean of folded-t is given by Psarakis and Panaretoes (1990). The expressions for entropies and variance of first four error models are well known, and these measures for the DGP are found from the entropy and second moment of the generalized Pareto distribution.

A.2 Derivation of the Lower Bound in (21)

Direct application of Theorem 8 of Rao *et al.* (2004) gives the lower bound (21) in the form of $LB = cH(|\varepsilon|)$, where

$$c = \exp\left\{\int_0^1 \log(u|\log u|)du\right\}.$$

The representation in (21) is obtained by noting that $\log u < 0$ for $0 < u < 1$ and

$$\int_0^1 \log(u|\log u|)du = \int_0^1 \log u du + \int_0^1 \log(-\log u)du = -1-\gamma,$$

where the integrals are given in Wikipedia's List of definite integrals. Using the entropy transformation formula, $H(\varepsilon/\sigma) = H(\varepsilon) + \log\sigma$, with $H(\varepsilon)$ of the GE shown in Table A1 and $\sigma = e^\gamma$ gives the result.

Table A1. *Error models with decreasing absolute error PDFs.*

Error model	PDF of $\lvert\varepsilon\rvert$	$SED(F_{\lvert\varepsilon\rvert} : F_0)$	Error entropy	Error variance
$f_\varepsilon(z) = \frac{1}{2} g_{\lvert\varepsilon\rvert}(z),\ z \in \Re$	$g_{\lvert\varepsilon\rvert}(z), z \geq 0$	$MAE(\varepsilon) = E(\lvert\varepsilon\rvert)$	$H(\varepsilon) = H(\lvert\varepsilon\rvert) + \log 2$	$V(\varepsilon) = E(\lvert\varepsilon\rvert^2)$
Generalized error, $GE(\beta)$ Laplace, $\beta = 1$, Normal, $N(0, .5)$, $\beta = 2$	$\frac{\beta}{\Gamma(1/\beta)} e^{-z^\beta},\ \beta > 0$	$\frac{\Gamma(2/\beta)}{\Gamma(1/\beta)}$	$\log \frac{2\Gamma(1/\beta)}{\beta} + \frac{1}{\beta}$	$\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}$
Bell-shaped				
Normal, $N(0, 1)$	$\sqrt{\frac{2}{\pi}} e^{-z^2/2}$	$\sqrt{\frac{2}{\pi}}$	$\frac{1}{2}\log(2\pi) + \frac{1}{2}$	1
Logistic, $LG(1)$	$\frac{2e^{-z}}{(1+e^{-z})^2}$	$\log 4$	2	$\frac{\pi^2}{3}$
Student-$t^*, t(\nu)$ Normal, $\nu \to \infty$	$\frac{2C_\nu}{(1+z^2/\nu)^{(\nu+1)/2}}$	$\frac{2\nu C_\nu}{\nu-1},\ \nu > 1$	$-\log C_\nu + \frac{(\nu+1)D_\nu}{2}$	$\frac{\nu}{\nu-2}, \nu > 2$
Double Generalized Pareto $DGP(\alpha)$ Laplace, $\alpha \to \infty$	$\frac{1}{(1+z/\alpha)^{\alpha+1}},\ \alpha > 0$	$\frac{\alpha}{\alpha-1},\ \alpha > 1$	$\log \frac{2}{\alpha} + \frac{1}{\alpha} + 1$	$\frac{2\alpha^2}{(\alpha-1)(\alpha-2)},\ \alpha > 2$

*Notes: $C_\nu = \frac{\Gamma(\nu/2+1/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}$, $D_\nu = \psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)$, $\psi(u) = \frac{d \log \Gamma(u)}{du}$

*A.3 Derivations of **MEE** and **MEER** of Logistic Model*

Let $u = 1 + e^{-z/\sigma}$. Then,

$$\begin{aligned}\int_\tau^\infty \frac{2e^{-z/\sigma}}{1+e^{-z/\sigma}} dz &= -2\sigma \int_{1+e^{-\tau/\sigma}}^0 \frac{du}{u} \\ &= -2\sigma \left[\log u\right]_{1+e^{-\tau/\sigma}}^0 \\ &= -2\sigma \left[\log(1+e^{-z/\sigma})\right]_\tau^\infty \\ &= 2\sigma \log(1+e^{-\tau/\sigma}).\end{aligned}$$

Upon division by the survival function, we obtain,

$$\begin{aligned}MEE(\tau) &= \frac{2\sigma(1+e^{-\tau/\sigma})\log(1+e^{-\tau/\sigma})}{2e^{-\tau/\sigma}} \\ &= \sigma(1+e^{\tau/\sigma})\log(1+e^{-\tau/\sigma}).\end{aligned}$$

Taking expectation with respect to $f_{\lvert\varepsilon\rvert}(\tau)$ gives

$$\begin{aligned}MEER_\tau &= -\int_0^\infty \frac{\sigma(1+e^{-\tau/\sigma})\log(1+e^{-\tau/\sigma})}{e^{-\tau/\sigma}} \frac{2e^{-\tau/\sigma}}{\sigma(1+e^{-\tau/\sigma})^2} d\tau \\ &= -\int_0^\infty \frac{2\log(1+e^{-\tau/\sigma})}{1+e^{-\tau/\sigma}} d\tau = 2\sigma \int_0^1 \frac{\log(1+v)}{v(1+v)} dv \\ &= 2\sigma \int_0^1 \left[\frac{\log(1+v)}{v} - \frac{\log(1+v)}{1+v}\right] dv = 2\sigma \left[\frac{\pi^2}{12} - \frac{\log^2 2}{2}\right] \\ &= \frac{\sigma}{6}[\pi^2 - 6(\log 2)^2] \approx 1.164\sigma,\end{aligned}$$

where $v = e^{-\tau/\sigma}$ and $dv = -\frac{1}{\sigma}e^{-\tau/\sigma}d\tau$; the value of the first term in the last integral is from Wikipedia's List of definite integrals.

A.4 Derivation of ***MEER*** *of the DGP Model*

Let $y = \tau/\alpha$ and $\beta = \alpha - 1$. Then,

$$\begin{aligned} MEER_\tau &= -\int_0^\infty \frac{1}{(1+\tau/\alpha)^\alpha} \log \frac{1}{(1+\tau/\alpha)^\alpha} d\tau \\ &= -\alpha \int_0^\infty \frac{1}{(1+y)^{\beta+1}} \log \frac{1}{(1+y)^{\beta+1}} dy \\ &= \frac{\alpha}{\beta}\left[-\int_0^\infty \frac{\beta}{(1+y)^{\beta+1}} \log \frac{\beta}{(1+y)^{\beta+1}} dy\right] + \frac{\alpha \log \beta}{\beta} \\ &= \frac{\alpha}{\beta}\left[\frac{1}{\beta} - \log \beta + 1\right] + \frac{\alpha \log \beta}{\beta}; \end{aligned}$$

the bracketed integral is the Shannon entropy of Pareto type II distribution given by the last bracketed expression. Letting $\beta = \alpha - 1$ gives the result.

A.5 R Codes for Empirical ***MEE*** *and* ***MEER***

Empirical MEE and MEER measures are computed by the **mee** and **meer** functions defined in R. The R codes are shown in the succeeding text. The **mee** has arguments (requires two inputs) as follows:

err: A vector of out of sample forecast errors with mean zero and variance one. For example, in regression forecasting, **err** consists of a set of standardised deleted residuals, also known as the Studentised residuals.

tau: A sequence of threshold, for instance, 0,.1,.2, . . .,5.

The **mee** function uses **err**.

mee computes MEE	**meer** computes MEER
mee <- function(err, tau) {	meer <- function(err) {
pabserr <- abs(err) - tau	value <- diff(sort(abs(err)))
value <- pabserr[pabserr> 0]	no <- c(1:length(err))
mee <- mean(value, na.rm = TRUE)	surv <- 11 - ((no-1)/length(err))
print(mee) }	surv <- surv[1:(length(err) - 1)]
	lsurv <- log(surv)
	meer <- -sum(value∗surv∗lsurv)
	return(meer) }

Supporting Information

Additional supporting information may be found online in the supporting information tab for this article.

[*Received September 2016, accepted November 2017*]