Weighted Scoring Rules for Density Forecast Comparison in Subsets of Interest *

Justinas Pelenis †
peleenis@ihs.ac.at
Institute for Advanced Studies, Vienna
February 15, 2014

PRELIMINARY AND INCOMPLETE

Abstract

We extend the notion of the scoring rules to weighted scoring rules that are designed to assess and rank the accuracy of competing density/probabilistic forecasts on the subsets of interest. We introduce the definitions of locally proper and preference preserving scoring rules. Locally proper scoring rules are designed to reward probabilistic/density forecasts that are correct in the subset of interest and disregard the forecast accuracy outside of the region of interest, that is a weighted scoring rule should not reject forecasts that are equal to the true data generating density on the subset of interest. Preference preserving scoring rules are such that a competing forecast preferred in both subsets A and B must be preferred over the union of A and B. We show that none of the previously proposed scoring rules satisfy these properties as weighted threshold continuous ranked probability score is not a locally proper scoring rule while censored and conditional likelihood scoring rules are not preference preserving. To address such shortcomings of extant scoring rules we introduce two novel scoring rules: likelihood based penalized weighted likelihood score and incremental continuous ranked probability score. Additionally, we present an empirical application that demonstrates that a choice of the preferable density forecast might be driven by a choice of a weighted scoring rule.

Keywords: Density and probabilistic forecasting, forecast comparison, weighted scoring rules, locally strictly proper.

JEL classification: C53, C52, C12.

†I am very thankful to Maximilian Kasy and to seminar participants at CFE 2013 for helpful discussion and suggestions that contributed to improve the quality of this manuscript. All remaining errors are mine.
1 Introduction

There is a growing strand of literature that is interested and addresses the questions of density forecast evaluation and comparison. A majority of these papers are concerned with the forecast density evaluation and comparison with regards to the whole set of observations (Berkowitz (2001), Mitchell and Hall (2005), Bao et al. (2007)). However there is an increasing literature (Galbraith and Norden (2012), Groen et al. (2013), van Dijk and Franses (2003), Diks et al. (2013) and others) that is concerned with the situations where the decision maker might have a targeted interest in a particular subset of the outcome space. Such concerns are applicable in risk management where the objects of interest might be such measures as Value-at-Risk (VaR) or Expected Shortfall, option pricing in computational finance and extreme inflation and unemployment forecasting in macroeconomics. Naturally, there is an increasing interest to provide the tools for evaluation and comparison of competing forecasts when the object of interest is not the full (un)conditional distribution. Furthermore, in the cases when the point estimate of interest is not elicitable as is the case for Expected Shortfall (see Gneiting (2011)) then such point estimates could be addressed by evaluating the relevant density forecasts over the subset of interest.

Amisano and Giacomini (2007), Gneiting and Ranjan (2011) and Diks et al. (2011) are just a few of the papers that address this concern of forecast evaluation and comparison for targeted regions of interests. These papers focus on the density/probabilistic forecast comparison when the region of interest is defined by a particular weight/threshold function that allows the decision maker to base the comparison on the relative performance in the tails or the center. The usual forecast comparison test is based on the difference losses/scores and it is the need to find appropriate score (loss) functions that is addressed in this paper. The scoring rules that are introduced in the two latter papers are promoted based on their property of being proper scoring rules. As it has been forcefully argued
in Gneiting and Raftery (2007), Diebold et al. (1998), Granger and Pesaran (2000) and Garthwaite et al. (2005) among others, proper scoring rules should be preferred in order to elicit honest revelation by the forecast makers. While we agree that propriety is one of the desirable properties of a scoring rule we will argue in this manuscript that it is insufficient when it is applied to the weighted scoring rules.

Generally, we will define weighted scoring rules to be scoring rules that focus on a particular region of interest through a weight function. When a decision maker is interested in the precision of a density forecast on a particular region of interest we will claim that a weighted scoring rule should be such that the true density forecast over the region of interest should be preferred to other density forecasts in that particular subset. Weighted scoring rules which would satisfy such a property would be defined as locally proper weighted scoring rules. We will show that the threshold weighted continuous ranked probability score (CRPS) presented in Gneiting and Ranjan (2011) does not have this property even though it is a strictly proper scoring rule. Given two competing density forecasts and an object of interest (such as the center of the distribution) it is possible that the density forecast that is equal to the true density on the region of interest might be ranked as second best compared to a forecast different from the true DGP over the region of interest. This is feasible when the density forecast is correct in the subset of interest but is incorrect outside of the region of interest and if the CRPS rule is used.

We suggest that another desirable property of the weighted scoring rules is to be preference preserving. Given two competing density forecasts \( f \) and \( g \) and two possible regions of interest \( A \) and \( B \) we will define a scoring rule preference preserving if it is such that if \( f \) is preferred to \( g \) in both \( A \) and \( B \) it must be preferred in the union of subsets \( A \) and \( B \). While one might argue whether this is a necessary or desirable property we believe that this might of interest in some particular situations. For example consider a stylized example of pricing/valuation of two butterfly options that provides variable
payoffs over two disjoint price ranges. For the two regions of interest taken individually, density forecast $f$ might be preferred and each of the butterfly options could be priced according to $f$, while the whole portfolio of two options might need to be priced according to $g$ if $g$ is preferred over the union of the two disjoint price ranges. This seems like an undesirable eventuality and for this we would like to determine which of the weighted scoring rules are preference preserving. We will show that the conditional (CL) and censored (CSL) likelihood scoring rules proposed in Diks et al. (2011) are not preference preserving.

To jointly address such shortcomings and to provide feasible alternative scoring rules we introduce two new weighted scoring rules. The first is a penalized weighted likelihood scoring rule presented in Equation (5) is a likelihood based scoring rule that can be seen as a competitor/alternative to the scoring rules presented in Diks et al. (2011). The second is an incremental continuous ranked probability score rule presented in Equation (6) can be seen as an extension of the proposals in Corradi and Swanson (2006) for the interval accuracy tests. We will show that the two newly introduced scoring rules contain all these desirable properties and address the shortcomings of the currently available rules. The summary of our theoretical results can be inferred from Table 2.

The article is structured in the following way. In Section 2 a list of the desirable properties of the weighted scoring rules is introduced and defined. A list of the weighted scoring rules that are present in the current literature and two newly introduced weighted scoring rules are presented in Section 3. Section 4 determines the set of properties that is associated with each of the weighted scoring rules. We present a simple empirical application in Section 5.
2 Properties of Weighted Scoring Rules

We consider probabilistic forecasts on a general sample space $\mathcal{Y}$ equipped with a complete probability space $(\mathcal{Y}, \mathcal{F}, \mathbb{P})$. We will be interested in a case when two (or multiple) competing density (or probabilistic) forecasts are available that generate one-step (or multi-step) ahead density forecasts of the observable $y_{t+1} \in \mathcal{Y}$ and we would be interested in comparing the forecasting performance of each of these forecasts with a possible focus on the tails or the center where the weight function can be dependent on the information at time $t$. To address such concerns of targeting a specific subset we will employ a weight function and define a weighted scoring rule. This will be an extension of the usual definition of a scoring rule that incorporates the weighting function $w(\cdot)$.

Define the set of possible weight functions as $W \equiv \{ w \mid w : \mathcal{Y} \mapsto [0, 1] \}$. Furthermore, we will define a subset of interest for a weight function $w(\cdot)$ as $A_w$ where $A_w$ is the range of $w$, i.e. $A_w = \{ y \mid w(y) > 0 \}$. We will argue that for comparison of density forecasts the only density forecasts of interest are the density forecasts on the subset of interest and thus weighted scoring rules should not be influenced by the behavior of the density or probabilistic forecasts on the complement $A_w^c$ of the subset of interest.

It is widely known that the logarithmic score is strictly proper relative to the class $\mathcal{L}_1$ of the probability measures dominated by $\mu$. So for a given probabilistic forecast $P \in \mathcal{L}_1$ we define its probabilistic density $p$ as the $\mu$-density of $P$. We will be using the scoring rules of probabilistic and density forecasts interchangeably without restricting ourselves to a specific class of probability measures $\mathcal{P}$ at this moment. However, for the probabilistic scoring rules we will restrict ourselves to the case of $\mathcal{Y} = \mathbb{R}$ as multivariate extensions of some scoring rules are not immediate. We define a weighted scoring rule as $S : \mathcal{P} \times \mathcal{Y} \times W \mapsto \mathbb{R}$.

**Definition 1.** A weighted scoring rule is a loss function $S(f, y, w)$ or $S(F, y, w)$ where $f$ is a density forecast and $F$ is a probabilistic forecast, $y$ is the realization of the future
observation $Y$, and $w$ is a chosen weight function.

Following the definitions in Gneiting and Ranjan (2011) we define an expected weighted score under $Q$ when the forecast is $P$ (or $q$ and $p$) for a given weight function $w$ as $S : \mathcal{P} \times \mathcal{P} \times \mathcal{W} \rightarrow \mathbb{R}$:

$$S(P, Q, w) = \int S(P, y, w)dQ(y) \quad \text{and} \quad S(p, q, w) = \int S(p, y, w)q(y)dy.$$ 

Then we can define a proper weighted scoring rule relative to $\mathcal{P}$.

**Definition 2.** A weighted scoring rule $S$ is proper relative to $\mathcal{P}$ if for all $w \in \mathcal{W}$:

$$S(Q, Q, w) \geq S(P, Q, w) \quad \text{for all } P, Q \in \mathcal{P} \quad \text{and/or} \quad S(q, q, w) \geq S(p, q, w) \quad \text{for all } P, Q \in \mathcal{P}.$$ 

The definition above is a natural extension of the definition of a proper scoring rule which states that $S(Q, Q, w_1) \geq S(P, Q, w_1)$ for all probabilistic forecasts $P, Q$ and a uniform weight function $w_1(\cdot) = 1$. The natural extension to the definition of the strictly proper weighted scoring rule follows as well.

**Definition 3.** A weighted scoring rule $S$ is a strictly proper weighted scoring rule for density forecasts if $S$ is a proper weighted scoring rule and if for given a weight function $w$:

$$S(q, q, w) = S(p, q, w)$$ 

if and only if $p(y) = q(y)$ a.e. w.r.t. $q$ for $y \in A_w$.

A weighted scoring rule $S$ is a strictly proper weighted scoring rule for probabilistic
forecasts if $S$ is a proper weighted scoring rule and if for given a weight function $w$:

$$S(Q, Q, w) = S(P, Q, w)$$

if and only if $P_{A_w} = Q_{A_w}$ where $A_{w<y} = A_{w}^c \cap (-\infty, y]$]

$$P_{A_w}(y) = P(Y \leq y) - P(A_{w<y})$$.

Hence, a strictly proper weighted scoring rule is such that if two density forecasts mismatch over the subset of interest they do not obtain the same expected score if one of these forecasts is the true DGP. For probabilistic forecasts we require that if probability increments mismatch over the subset of interest then if one the probability forecasts is true then they must obtain different scores.

Occasionally, the forecaster might have a specific interest only in the forecasting ability on the left or right tail or just the center. To address such situations when the region of interest is a strict subset of the domain, we introduce the notion of the locally (strictly) proper weighted scoring rules\(^1\). While the previous discussion focused on (strictly) proper weighted scoring rules, we might be interested in weighted scoring rules which do not reject the density/probabilistic forecasts based on their performance outside of the subset of interest. Given a weight function $w$ and the subset of interest $A_w$ of the weight function we would call a weighted scoring rule locally proper if it prefers the density forecasts that are equal to the true density on the subset $A_w$ regardless of their performance on $A^c_w$. For example, as suggested by Corradi and Swanson (2006), “if we want to evaluate the accuracy of different models for approximating the probability that the rate of inflation tomorrow, given the rate of inflation today, will be between 0.5% and 1.5%” we might prefer weighted scoring rules that do not reject the density forecasts which are equal to

\(^1\)This should not be confused with the notion of the local scoring rules
the true density forecasts over this region of interest. Much as proper scoring rules are preferred for the arguments of truthful revelation, thus the scoring rules that are locally proper are preferred for the sake of truthful revelation in the subset of interest. To address such concerns we define *locally (strictly) proper* weighted scoring rules.

**Definition 4.** A weighted scoring rule $S$ is locally proper relative to $\mathcal{P}$ if for all $w \in \mathcal{W}$ for density forecasts:

$$S(q, h, w) \geq S(p, h, w) \text{ for all } P, Q, H \in \mathcal{P} \text{ such that } q(y) = h(y) \text{ a.e. w.r.t. } h \text{ for } y \in A_w,$$

and for probabilistic forecasts:

$$S(Q, H, w) \geq S(P, H, w) \text{ for all } P, Q, H \in \mathcal{P} \text{ such that } Q_{A_w} = H_{A_w}.$$

Note that for a weighted scoring rule to be locally proper it is sufficient if the rule is proper and if $S(h, h, w) = S(q, h, w)$ when $h(y) = q(y)$ a.e. w.r.t. $h$ for $y \in A_w$ or if $S(H, H, w) = S(Q, H, w)$ when $Q_{A_w} = H_{A_w}$. Similarly, the definition to the locally strictly proper weighted scoring rules is extended by replacing weak inequalities with strict inequalities.

The local properness implies that scoring rules should favor generated density forecasts which are equal to the true density forecasts on the subset of interest and should disregard the performance of the density forecasts on the outcome space outside of subset of interest. We will show that it is possible to have strictly proper scoring rules that are not locally proper. However, it seems that if a weighted scoring rule is strictly proper and locally proper, then it must be locally strictly proper weighted scoring rule.

The interest in locally proper scoring rules can be driven by the fact that if the forecast coincides with the true data generating process on this particular subset, then it would be preferred by all decision makers regardless of their loss functions as long as
the loss function is such that the losses outside of the subset of interest are constant, \( L(a, y) = c \) when \( y \in A^c_w \).

In more generality, suppose that the forecast user has a loss function \( L(a, y) \) where \( a \in A \) is an action choice and \( y \) is the realization of the variable of interest. Conditional on the user’s choice of the forecast density \( f(y) \), the chosen action will be:

\[
a^*(f(y)) = \arg \min_{a \in A} \int L(a, y) f(y) dy.
\]

In order to rank the two density forecasts \( f(y) \) and \( p(y) \), the user will evaluate whether

\[
\int L(a_f^*, y) h(y) dy \leq \int L(a_p^*, y) h(y) dy
\]

when the true density function is \( h(y) \). While generally it is impossible to rank \( f(y) \) and \( p(y) \) for all possible loss functions, however it is know that if \( f(y) = h(y) \) then \( f(y) \) would be preferred by all forecast users regardless of their loss functions (as pointed out in Diebold et al. (1998)). Now suppose that the loss function is such that \( L(a, y) = c \) for all \( y \in A^c_w \). Then the forecast user would compare

\[
\int_{A_w} L(a_f^*, y) h(y) dy + H(A^c_w) c \leq \int_{A_w} L(a_p^*, y) h(y) dy + H(A^c_w) c
\]

Then if one of the forecasts, for example \( f \), is the same as the true data generating process \( h(\cdot) \) over \( A_w \), then the decision maker should prefer \( f \) over \( p \) regardless of the loss function. Hence, one might be interested in locally proper scoring rules as the means to rank density forecasts for decision makers that have loss functions that are constant outside of the region of interest.
An additional property that we might desire of the weighted scoring rules is one of preference preserving. Given a weighted scoring rule and two competing density/probabilistic forecasts \( f \) and \( g \) we might desire that if \( f \) is preferred over \( g \) in the region of interest \( A_1 \) and \( f \) is preferred to \( g \) over subset \( A_2 \) one should expect that density forecast \( f \) should be preferred over \( g \) when evaluating its performance over the union \( A_1 \cup A_2 \) of regions of interest. We might want to evaluate whether existing weighted scoring rules satisfy this preference preserving property which is defined below.

**Definition 5.** A weighted scoring rule \( S \) is preference preserving if for all weight functions \( w_1, w_2 \in \mathcal{W} \) such that \( w_1 + w_2 \in \mathcal{W} \) and for all \( Q, P, H \in \mathcal{P} \) the following holds true:

\[
S(q, h, w_1) > S(p, h, w_1) \text{ and } S(q, h, w_2) > S(p, h, w_2) \implies S(q, h, w_1 + w_2) > S(p, h, w_1 + w_2) \text{ and }
\]

\[
S(Q, H, w_1) > S(P, H, w_1) \text{ and } S(Q, H, w_2) > S(P, H, w_2) \implies S(Q, H, w_1 + w_2) > S(P, H, w_1 + w_2).
\]

Another property that one might require the weighted scoring rules to possess is the weight scale invariance. We contend that if density forecast \( f \) is preferred over \( g \) for a weight rule \( w \in \mathcal{W} \), then \( f \) must be preferred to \( g \) for a weight rule \( \alpha w \in \mathcal{W} \) for any \( \alpha \in \mathbb{R} \).

**Definition 6.** A weighted scoring rule \( S \) is weight scale invariant if for any \( w \in \mathcal{W} \) and any \( \alpha \in \mathbb{R} \) such that \( \alpha w \in \mathcal{W} \) the following is true for any \( Q, P, H \in \mathcal{P} \):

\[
S(q, h, w) > S(p, h, w) \implies S(q, h, \alpha w) > S(p, h, \alpha w) \text{ and }
\]

\[
S(Q, H, w) > S(P, H, w) \implies S(Q, H, \alpha w) > S(P, H, \alpha w).
\]

Hence, we argue that for scoring rules there are four general properties that one might expect a well functioning weighted scoring rule to possess: properness, local properness, preference preservation and weight scale invariance. In Section 4 we evaluate whether any of the proposed in the literature rules satisfy all or any of these four general types.
of properties. Furthermore, one might desire for scoring rules that are invariant under transformations of the outcome space which is a well known property of likelihood based scoring rules.

3 List of Weighted Scoring Rules

In this subsection we consider a list of the weighted scoring rules that have been proposed in the literature.

First, the weighted likelihood (WL) scoring rule proposed by Amisano and Giacomini (2007)

$$S^{WL}(f, y, w) = w(y) \log f(y).$$  \hspace{1cm} (1)

Second, the threshold weighted continuous ranked probability score (CRPS) rule proposed by Matheson and Winkler (1976) and Gneiting and Ranjan (2011)

$$S^{CRPS}(f, y, w) = -\int_{-\infty}^{\infty} PS(F(z), 1\{y \leq z\})w(z)dz,$$

$$PS(F(z), 1\{y \leq z\}) = (F(z) - 1\{y \leq z\})^2$$ (2)

Third, the conditional likelihood (CL) rule proposed by Diks et al. (2011):

$$S^{CL}(f, y, w) = w(y) \log \left( \frac{f(y)}{\int w(s)f(s)ds} \right)$$  \hspace{1cm} (3)

Fourth, the censored likelihood (CSL) rule proposed by Diks et al. (2011):

$$S^{CSL}(f, y, w) = w(y) \log (f(y)) + (1 - w(y)) \log \left( 1 - \int w(s)f(s)ds \right)$$  \hspace{1cm} (4)

In Section 4 we will show that neither of these four rules satisfy all four idealistic require-
ments for a weighted scoring rule. To address this shortcoming we propose two alternative weighted scoring rules. The first one is a different modification of a likelihood scoring rule and hence can be thought of as a competing weighted likelihood based scoring rule which we will call a penalized weighted likelihood (PWL) scoring rule.

\[ S_{\text{PWL}}(f, y, w) = w(y) \log f(y) - \int w(s)f(s)ds + w(y). \]  

(5)

Note that the last term \(w(y)\) is not really necessary and could be eliminated to present an equivalent scoring rule, however we will contend that in this formulation the weighted scoring rule has an associated expected score that is interpretable as a weighted version of negative differential entropy and that \(S_{\text{PWL}}\) reduces to the regular likelihood scoring rule for \(w(\cdot) = 1\) as desired.

The second proposal is a slight modification of the CRPS rule which simplifies to the CRPS rule in the case when \(A_w = \mathcal{Y}\) and we will denote this weighted scoring rule as incremental CRPS (ICRPS).

\[ S_{\text{ICRPS}}(f, y, w) = -\int_{-\infty}^{\infty} (F(z) - F(z_w) - 1\{z_w \leq y \leq z\})^2 w(z)dz, \]  

(6)

where \(z_w \equiv \sup\{A_w^c \cap (-\infty, z]\}\)

This ICRPS can be seen as an extension of the scores for interval accuracy tests proposed by Corradi and Swanson (2006) and Corradi and Swanson (2005) as a weighted integral of the interval accuracy scores over the subset of interest. This rule requires that the forecast attributes correct probabilistic estimate on each of the intervals that is bounded on the left side by the infimum of \(A_w\) and the upper bounds of the intervals are all possible points belonging to \(A_w\). While such a rule has some desirable properties, nonetheless it prefers the forecasts that correctly evaluate the probabilistic estimates on the left side of \(A_w\) as compared to the right side of \(A_w\) and this leads to the well known fact that neither
CRPS nor, in this case, ICRPS is invariant under the transformation of the outcome space. This is one shortcoming of the ranked probability score rules that is avoided if one uses likelihood based scoring rules.

In Section 4 we will show that the first four weighted scoring rules do not satisfy all the four properties of proprietry, local propriety, preference preservation and weight scale invariance, while PWL and ICRPS do satisfy all the four rules. Between PWL and ICRPS we argue that PWL is preferable on the grounds of it being a likelihood based rule and hence invariant and it is computationally easier to evaluate. On the other hand ICRPS is directly applicable to probabilistic forecasts, while PWL is not.

4 Evaluating Weighted Scoring Rules

In this section we will explore which of the properties are satisfied by the various weighted scoring rules. Based on these properties we can evaluate which scoring rules might be more or less effective in the particular situations of interest. In Table 2 we provide the summary of the weighted scoring rules and of the properties that each of these scoring rules maintain. We will show that none of the WL, CRPS, CSL, CL weighted scoring rules contain these properties while PWL and ICRPS do.

Properness of WSRs.

First, as it has already been pointed out by Gneiting and Ranjan (2011) and Diks et al. (2011) WL is not a proper weighted scoring rule, while CRPS, CL and CSL are. Both newly introduced PWL and ICRPS weighted scoring rules are proper as well as Lemmas 1 and 2 show.

Lemma 1. PWL weighted scoring rule defined in Equation (5) is a (strictly) proper weighted scoring rule.

Proof. Let \( w \in \mathcal{W} \) be given, then define \( F_w = \int w(s)f(s)ds \). Note that if \( F_w = 0 \), then
$S^{PWL}(f, f, w) = S^{PWL}(g, f, w) = 0$ and it is a proper scoring rule, hence assume that $F_w > 0$. Then:

$$d^{PWL}(g, f, w) = S^{PWL}(g, f, w) = 0$$

$$= \int f(y) (w(y) \log f(y) - F_w - w(y) \log g(y) + G_w) \, dy$$

$$= \int w(y) f(y) \log \left( \frac{f(y)w(y)}{w(y)g(y)} \right) \, dy + G_w - F_w$$

$$= \int w(y) f(y) \log \left( \frac{f(y)w(y)}{g(y)F_w/G_w} \right) \, dy + G_w - F_w$$

$$= \int w(y) f(y) \log \left( \frac{f(y)w(y)}{g(y)F_w/G_w} \right) \, dy + F_w \log \frac{F_w}{G_w} + G_w - F_w$$

$$= F_w \int \frac{w(y)f(y)}{F_w} \log \left( \frac{f(y)w(y)}{g(y)F_w/G_w} \right) \, dy + F_w \log F_w - F_w \log G_w + G_w - F_w$$

$$= F_w \cdot K \left( \frac{w(y)f(y)}{F_w}, \frac{w(y)g(y)}{G_w} \right) + F_w \log F_w - F_w \log G_w + G_w - F_w \geq 0$$

where $K(\cdot, \cdot)$ is the Kullback-Leibler divergence between two pdfs. The inequality is true since $K(\cdot, \cdot)$ is non-negative by definition and the remaining term obtains the minimum when $F_w = G_w$, hence it is non-negative as well.² PWL is strictly proper as $d^{PWL}(g, f, w) > 0$ unless $g(y) = f(y)$ a.e. for $y \in A_w$. 

**Lemma 2.** ICRPS weighted scoring rule defined in Equation (6) is a (strictly) proper weighted scoring rule.

**Proof.** Define $F_w(z) = F(z) - F(z_w)$ where $z_w = \sup\{A_w \cap (-\infty, z]\}$. Then for ICRPS

² An interesting fact is that $S^{PWL}(f, f, w) = \int w(y)f(y)\log f(y)dy$.

This could be interpreted as the negative of a “weighted” differential entropy.
simple algebra delivers:

\[
S(G, F, w) = - \int_{-\infty}^{\infty} \left( G_w^2(z) - 2F(z)G_w(z) + F_w(z) \right) w(z) dz.
\]

Hence we can find the associated divergence function

\[
d(G, F, w) = S(F, F, w) - S(G, F, w)
= - \int_{-\infty}^{\infty} \left( -F_w^2(z) + F_w(z) \right) w(z) dz + \int_{-\infty}^{\infty} \left( G_w^2(z) - 2F_w(z)G_w(z) + F_w(z) \right) w(z) dz
= \int_{-\infty}^{\infty} (G_w(z) - F_w(z))^2 w(z) dz
= \int_{-\infty}^{\infty} \left( (G(z) - G(z_w)) - (F(z) - F(z_w)) \right)^2 w(z) dz \geq 0.
\]

Since \(d(G, F, w) \geq 0\) for any \(G, F \in \mathcal{P}\) and for any \(w \in \mathcal{W}\) it is a proper weighted scoring rule. Note that inequality is strict unless \(G(z) - G(z_w) = F(z) - F(z_w)\) a.e. for \(z \in A_w\).

Therefore, CSL, PWL, CRPS and ICRPS are strictly proper weighted scoring rules while CL is not. To show that CL is not a strictly proper weighted scoring rule consider an example where \(w(y) = 1\{y \in A\}\), then given a density \(f\) define \(g\) as \(g(y) = 1\{y \in A\} 0.5f(y) + (1 - 1\{y \in A\})\frac{(1 - 0.5F(A))f(y)}{1 - F(A)}\), then \(S(f, f, w) = S(g, f, w)\) even if \(F_{A_w} \neq G_{A_w}\). For this reason one might prefer the CSL to CL as a strictly proper rule and Diks et al. (2011) do suggest that in simulations CSL exhibit better size and power properties that CL scoring rule.

**Local Properness of WSRs.**

More surprising results are found when one considers the local properness of the scoring rules. It turns out that WL and surprisingly CRPS is not a locally proper scoring rule, while the others are. Therefore, contrary to the suggestion of Corradi and Swanson (2006), the CRPS is not necessarily an appropriate rule if one wants to compare two density
forecasts with regards to forecasting ability in a particular region/subset of interest.

First we show that CSL, CL, PWL and ICRPS are locally proper scoring rules. Let weight function \( w \in \mathcal{W} \) be given. Consider \( P, Q, H \in \mathcal{P} \) such that \( q(y) = h(y) \) a.e. w.r.t. \( H \) for \( y \in A_w \), then note that since these four are proper scoring rules, hence \( S(H, H, w) \geq S(P, H, w) \). Note that by the definition of CL, CSL, PWL and ICRPS scoring rules in Equations (3), (4), (5) and (6) it is immediate that \( S(Q, H, w) = S(H, H, w) \) whenever \( q(y) = h(y) \) a.e. w.r.t. \( H \) for \( y \in A_w \). Hence for these rules \( S(Q, H, w) = S(H, H, w) \geq S(P, H, w) \). The argument for CSL, PWL and ICRPS being strictly locally proper scoring rules is the same with weak inequalities replaced by strong inequalities.

To show that CRPS is not a locally proper weighted scoring rule we propose to consider this counterexample. Let the true density be \( h \) and let \( h \) be the uniform density on unit interval. Then we will consider two alternative densities \( f, g \) for density forecasting comparison.

\[
\begin{align*}
f(x) &= \begin{cases} 
0 & x \notin [0, 1] \\
0.1 & x \in [0, \frac{1}{1}] \\
1 & x \in (\frac{1}{3}, \frac{3}{4}] \\
1.9 & x \in [\frac{3}{4}, 1]
\end{cases} \\
g(x) &= \begin{cases} 
0 & x \notin [0, 1] \\
0.9 & x \in [0, \frac{1}{4}] \\
1.1 & x \in (\frac{1}{4}, \frac{3}{4}] \\
0.9 & x \in [\frac{3}{4}, 1]
\end{cases}
\end{align*}
\]

(7)

The region of interest will be \( A = [\frac{1}{4}, \frac{3}{4}] \). One can note that \( h(x) = f(x) \) for all \( x \in A \), so if a scoring rule is locally proper then forecast density \( f \) should be preferred to all other density forecasts \( g \) for which \( g(x) \neq h(x) \) for all \( x \in B_w \subseteq A_w \) with \( H(B_w) > 0 \). Hence consider a weight function \( w(y) = 1\{y \in A\} \) and compute \( S^{CRPS}(f, h, w) \) and \( S^{CRPS}(g, h, w) \). Since \( S^{CRPS}(f, h, w) = -0.139 < -0.115 = S^{CRPS}(g, h, w) < -0.114 = S^{CRPS}(h, h, w) \), hence \( g \) is preferred to \( f \) over the subset \( A \) and this implies that CRPS is not a locally (strictly) proper rule. This failure of CRPS to be locally proper stems from the fact that for competing density forecasts when evaluated over the region of interest.
it is important for CRPS rule to have correct estimates of the cumulative distribution function at both\textsuperscript{3} the endpoints of the region of interest. If the region of interest is one of the tails, then CRPS functions well since the cdf at endpoints is always correctly estimated as 0 and 1 regardless of the proposed density forecast, however if the center of the outcome space is the object of interest then CRPS might be considered inappropriate for such a situation. To avoid such unwanted property of the CRPS rule we proposed ICRPS rule that precisely cancels out this requirement to correctly forecast the cdf values of boundary points of the subset of interest.

**Preference Preserving WSRs.**

Next, we consider the preference preservation of the weighted scoring rules. First, since the weighted scoring rules WL, CRPS, PWL and WRPS are linear in the weight function \( w \), hence it is preference preserving since \( S(Q, P, w_1) + S(Q, P, w_2) = S(Q, P, w_1 + w_2) \) for any \( P, Q \in \mathcal{P} \). Second, one can show that CL and CSL are not preference preserving.

We provide a counterexample that shows that CL and CSL are not preference preserving for \( w_1(y) = 1(y \in A) \) and \( w_2(y) = 1(y \in A^c) \). Suppose that the true density \( u(\cdot) \) is uniform on the unit interval. We will consider three competing density forecasts \( f, g \) and \( h \):

\[
\begin{align*}
f(x) &= \begin{cases} 
0 & x \notin [0, 1] \\
1.3 & x \in [0, \frac{1}{2}) \\
0.7 & x \in (\frac{1}{2}, 1]
\end{cases} \\
g(x) &= \begin{cases} 
0 & x \notin [0, 1] \\
1.35 & x \in [0, \frac{1}{2}) \\
0.65 & x \in (\frac{1}{2}, 1]
\end{cases} \\
h(x) &= \begin{cases} 
0 & x \notin [0, 1] \\
\frac{4}{3} & x \in [0, \frac{1}{2}) \\
\frac{2}{3} & x \in (\frac{1}{4}, \frac{3}{4}) \\
\frac{4}{3} & x \in [\frac{3}{4}, 1]
\end{cases}
\end{align*}
\]

The interval of interest \( A = [0, \frac{1}{2}] \) and the weight functions are \( w_1(y) = 1(y \in A) \) and \( w_2(y) = 1(y \in A^c) \) and the weight function \( w_1 + w_2 \) is the uniform weight function which leads to the unweighted log score for both CL and CSL. The results are summarized in

\textsuperscript{3}Or multiple endpoints if the subset of interest is a disjoint union of intervals.
Table 1: Weighted Scores for CL and CSL

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_1 + w_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^{CL}(g,u,w)$</td>
<td>0.3466</td>
<td>0.3466</td>
<td>-0.0653</td>
</tr>
<tr>
<td>$S^{CL}(h,u,w)$</td>
<td>0.3171</td>
<td>0.3171</td>
<td>-0.0589</td>
</tr>
<tr>
<td>$S^{CSL}(f,u,w)$</td>
<td>-0.3937</td>
<td>-0.3937</td>
<td>-0.0472</td>
</tr>
<tr>
<td>$S^{CSL}(h,u,w)$</td>
<td>-0.3760</td>
<td>-0.3760</td>
<td>-0.0589</td>
</tr>
</tbody>
</table>

Table 1. One can notice that for the CL rule $g$ is preferred to $h$ for both $w_1$ and $w_2$, but $h$ is preferred to $g$ for $w_1 + w_2$. Similarly, for the weighted CSL rule $h$ is preferred to $f$ for both $w_1$ and $w_2$, but the preference is reversed for the uniform weighting rule.

**Weight Scale Invariance of WSRs.**

We will show that WL, CL, CRPS, PWL and ICRPS are weight scale invariant, while CSL is not. First, note that since WL, CRPS, PWL and ICRPS are linear in $w$, hence for any $P, Q \in \mathcal{P}$ and any $w, \alpha w \in \mathcal{W}$ we know that $S(P, Q, \alpha w) = \alpha S(P, Q, w)$ hence these rules are weight scale invariant. Second, note that for CL we can show that $S(P, Q, \alpha w) = \alpha S(P, Q, w) + r(\alpha, w, Q)$ where $r$ is a function independent of the forecast $P$, therefore CL is weight scale invariant as well. Third, CSL is not weight-scale invariant as the following counterexample illustrates. Suppose that the true density $u(\cdot)$ is the uniform density on the unit interval and consider the two competing density forecasts $f$ and $g$.

$$f(x) = \begin{cases} 0 & x \notin [0, 1] \\ 1.7 & x \in [0, \frac{1}{2}] \\ 0.3 & x \in (\frac{1}{2}, 1) \end{cases} \quad g(x) = \begin{cases} 0 & x \notin [0, 1] \\ 0.4 & x \in [0, \frac{1}{2}] \\ 1.6 & x \in (\frac{1}{2}, 1) \end{cases}$$

And consider two weighting rules $w(\cdot)$ and $0.5w(\cdot)$ for $w(y) = 1\{y \in [0, 0.5]\}$. Simple computation will deliver that $S^{CSL}(f, u, w) - S^{CSL}(g, u, w) = -0.1135$ and that $S^{CSL}(f, u, 0.5w) - S^{CSL}(g, u, 0.5w) = 0.0257$. Therefore $f$ is preferred to $g$ for the weight function $0.5w(\cdot)$ and $g$ is preferred to $f$ for the weight function $w(\cdot)$. Hence, CSL is not a weight invariant scoring rule.
Table 2: Properties of Weighted Scoring Rules

<table>
<thead>
<tr>
<th>WSR</th>
<th>proper (strictly)</th>
<th>locally proper (strictly)</th>
<th>preference preserving</th>
<th>weight scale invariant</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL</td>
<td>No (No)</td>
<td>No (No)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CRPS</td>
<td>Yes (Yes)</td>
<td>No (No)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CL</td>
<td>Yes (No)</td>
<td>Yes (No)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>CSL</td>
<td>Yes (Yes)</td>
<td>Yes (Yes)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PWL</td>
<td>Yes (Yes)</td>
<td>Yes (Yes)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ICRPS</td>
<td>Yes (Yes)</td>
<td>Yes (Yes)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

WSRs are strictly proper only with respect to a particular class of probability measures for different scoring rules.

In conclusion, to the knowledge of the author none of the previously proposed scoring rules satisfied the four main properties of properness, local properness, preference preservation and weight scale invariance. To address this issue we have introduced two new weighted scoring rules PWL and ICRPS which satisfy all these four criteria. The summary of these results are provided in the Table 2. From these results we suggest to use either PWL as a likelihood based density scoring rule or ICRPS as a probabilistic weighted scoring rule. However as it has already been pointed out the ranked probability score based rules are not invariant under transformations of the outcome space, while the likelihood based scoring rules are. Under this consideration one might choose to prefer the PWL weighted scoring rule over the ICRPS. On the other hand, ICRPS can be used for probabilistic forecasts with point mass distributions while PWL would not be applicable to such situations.

5 Empirical Application

To apply the proposed weighted scoring rules for the density forecast comparison in the time series setup we will consider density forecasts that are generated based on a rolling window of past $m$ observations so that the testing procedure suggested in Amisano and Giacomini (2007) can be applied. Suppose that a stochastic process $z_t = (y_t, x_t)$ is observed, where $y_t$ is the variable of interest and $x_t$ are the covariates. Then we will
consider two density forecasts \( \hat{f}_t = f(y_t|\{z_{i}\}_{i=m}^{t-1}) \) and \( \hat{g}_t = g(y_t|\{z_{i}\}_{i=m}^{t-1}) \). For a given weight function \( w(\cdot) \in \mathcal{W} \) density forecasts can be ranked according to the average weighted scores given density forecasts \( \hat{f}_t \) and \( \hat{g}_t \) and realizations of \( y_{t+1} \) for \( t = m, m + 1, \ldots, T - 1 \).

Define a weighted score difference as

\[
\Delta_{fg}(y_{t+1}, w) = S(\hat{f}_t, y_{t+1}, w) - S(\hat{g}_t, y_{t+1}, w). 
\]

Then we can consider tests of equal forecast performance based on a Diebold and Mariano (1995) test statistic

\[
t_{m,T} = \frac{\sum_{t=m+1}^{T} \Delta_{fg}(y_{t+1}, w)}{\sqrt{(T-m)\sigma_{m,T}^2}} 
\]

where \( \sigma_{m,T}^2 \) is variance estimator of \( \Delta_{fg}(y_{t+1}, w) \). Then under standard regularity conditions that are stated in Giacomini and White (2006), Amisano and Giacomini (2007) and Diks et al. (2011) the statistic \( t_{m,T} \) is asymptotically standard normal as \( T \to \infty \) under the null hypothesis that \( \mathbb{E}[\Delta_{fg}(y_{t+1}, w)] = 0 \). Since the weighted scoring rules are presented in positive orientation, then a forecaster/decision maker should prefer \( \hat{f} \) to \( \hat{g} \) when \( t_{m,T} \) is positive and vice versa. This simple testing procedure can be used to determine which density forecasts are preferably and/or significantly superior for different weight functions and different subsets of interest.

For the empirical application we consider the evaluation of the density forecast for the daily stock index returns and for foreign exchange rates. We consider DAX stock index log returns \( y_t = \ln(P_t/P_{t-1}) \) where \( P_t \) is the closing price on day \( t \). The sample period runs from November 26, 1990 until December 9, 2013 with a total of 5836 observations. For this simple exercise we will consider a comparison of two benchmark GARCH models. We will compare density forecasts generated by a GARCH(1,1) model with Gaussian errors.
and a \( t \)-GARCH(1,1) model with student-t innovations as proposed by Bollerslev (1987). It is a well established empirical fact that the GARCH models with normally distributed errors is not sufficiently leptokurtic, while on other hand the fat-tailed \( t \)-GARCH(1,1) is considered a good benchmark model that is not often outperformed. The common reasoning for the superior performance of the \( t \)-GARCH(1,1) model is that it’s fat-tailed distribution provides a better fit to the empirical observations of log returns. By applying our weighted scoring rules we would like to test whether this increased fit is driven by the superior performance in the left or right tail of the log-return distribution and/or the center of the distribution. To evaluate our density forecasts we will use Diebold-Mariano test statistic based on the estimation window of \( m = 1250 \) which will provide 4586 out of sample observations.

We will consider a set of various weight functions to evaluate the performance of the two competing GARCH models over alternative regions of interest. Furthermore, we will consider five different weighted scoring rules (PWL, CSL, ICRPS, CRPS, and WL) to determine whether the forecast comparison results match up across the different scoring rules and which of the competing GARCH models is preferable over alternative subsets of interest. The results are summarized in the Table 3.

First, note that WL rule indicates that \( t \)-GARCH(1,1) forecasts are superior for the central part of the return distribution and GARCH(1,1) forecasts are preferable for the weight functions \( w(y) = 1 \{-2 < y < -1\} \) and \( w(y) = 1 \{1 < y < 2\} \). However, given

<table>
<thead>
<tr>
<th>( w(y) )</th>
<th>( 1 {y&lt;-2} )</th>
<th>( 1 {-2&lt;y&lt;-1} )</th>
<th>( 1 {-1&lt;y&lt;0} )</th>
<th>( 1 {0&lt;y&lt;1} )</th>
<th>( 1 {1&lt;y&lt;2} )</th>
<th>( 1 {2&lt;y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWL</td>
<td>-1.8161</td>
<td>-2.3459</td>
<td>-1.4768</td>
<td>-3.1834</td>
<td>-0.0178</td>
<td>-0.2352</td>
</tr>
<tr>
<td>CSL</td>
<td>-1.8162</td>
<td>-2.4721</td>
<td>-0.9972</td>
<td>-3.3007</td>
<td>0.0120</td>
<td>-0.2098</td>
</tr>
<tr>
<td>ICRPS</td>
<td>0.0535</td>
<td>-1.5340</td>
<td>0.7988</td>
<td>-3.5561</td>
<td>0.0986</td>
<td>0.5320</td>
</tr>
<tr>
<td>CRPS</td>
<td>0.0535</td>
<td>-0.0038</td>
<td>-3.2280</td>
<td>-0.5844</td>
<td>-0.9888</td>
<td>-1.1735</td>
</tr>
</tbody>
</table>

Positive values of \( t \)-statistics suggest superiority of forecasts generated from normal GARCH(1,1) model, while negative \( t \)-values support superiority of \( t \)-GARCH(1,1) forecasts.
that WL is not a proper scoring rule, we will choose to disregard the results of the WL scoring rule. Second, note that PWL and CSL provide very similar results, while ICRPS results are probably the next most similar of the alternative weighted scoring rules. Third, forecast comparison results based on the CRPS rule suggest different density forecast choice for the regions of \((-2, -1), (-1, 0)\) and \((0, 1)\) as compared to the other weighted scoring rules. Given that CRPS is not a locally proper scoring rule, we would advise to choose the density forecasts based on the PWL, CSL and/or ICRPS scoring rules as opposed to the CPRS weighted scoring rule.

References


Corradi, V. and Swanson, N. R. A test for comparing multiple misspecified condi-


