Forecast Elicitation with Weighted Scoring Rules [*]

Justinas Pelenis [†]
pelenis@ihs.ac.at
Institute for Advanced Studies, Vienna

January 25, 2015

PRELIMINARY AND INCOMPLETE

Abstract

We investigate the possible advantages of matching the loss function (scoring rule) used for the estimation with the loss function used for the evaluation of density forecasts. We focus on weighted scoring rules for density forecasts which reward forecasting performance on specific regions of support. When forecasting models are correctly specified the choice of the specific scoring rule lead to asymptotically identical results. However, if the models are misspecified density forecasts elicited under different scoring rules might diverge and consequently lead to different decisions. We examine the benefit of this approach in the context of forecasting downside risk in the equity markets.

Keywords: Density and probabilistic forecasting, Value-at-Risk (VaR) forecasting, forecast comparison.
JEL classification: C53, C52, C12.

[*] Current version: December 28, 2015.

[†] I am very thankful to seminar participants at NYUAD for helpful discussions and suggestions that contributed to improve the quality of this manuscript. All remaining errors are mine.

1 Introduction

Probabilistic and density forecasting has been increasing in popularity and relevance. Alessi et al. (2014) describe new framework for prediction and policy guidance at FRBNY as:"The framework has two key ingredients: (1) ... (2) the recognition that one should pay attention to distributional features beyond point forecasts, in line with general notions of macroeconomic risk." There is an interest in the measures and forecasts of the downside risk for investments as well outside of the macro-forecasting world. One could attempt to measure and forecast the distribution on the left tail or alternatively focus on such point measures as Value-at-Risk (VaR) and/or expected shortfall (ES). Even if one measure of the downside is chosen as the target functional it still leaves open the question on how these estimates/forecasts should be evaluated and elicited. In this manuscript we will explore the estimation and forecast elicitation through the use of the proper scoring rules for density forecasts.

More specifically, the purpose of this paper is to explore the advantages of using forecast scoring rules that are targeted to reward the forecasts on the left tail for estimation purposes. Furthermore, we explore whether there is an advantage to obtain VaR forecasts from the model parameters estimated through the use of the newly introduced weighted scoring rules.

The first contribution of this paper is to explore the merits of using weighted scoring rules discussed in the papers by Amisano and Giacomini (2007), Gneiting and Ranjan (2011), Diks et al. (2011) and Pelenis (2015) in order to improve the density forecasts on the left tail. The second contribution is to show that the weighted scoring rules could be used in estimation to improve such point forecasts as Value-at-Risk. Furthermore this manuscript adds to the debate on how to incorporate the information from competing forecasting models. Once the idea that the models are misspecified is accepted one alternative is to focus on forecast combination schemes in combining density forecasts (see Kapetanios et al. (2015) and Opschoor et al. (2015)). Another alternative would be to realize that the estimation based on different scoring rules might lead to different estimates and forecasts (even asymptotically) when the models are misspecified. This is the approach that is taken in this paper and adds to the literature dealing with estimation procedures and forecast evaluation under model misspecification.

In Section 2 we provide an introduction to the definition and use of the scoring rules. In Section 3 we discuss the approach of using scoring rules for estimation purposes. In Section 4 we present the empirical results from the simulation study and an empirical study for stock returns showing that left tail density forecasting and VaR forecasting can be improved through the use of weighted scoring rules in estimation.

2 Scoring Rules

Often the performance of the competing forecasts and/or forecasting procedures are measured based on a particular scoring functions. The most familiar scoring functions for point forecasts probably would be root mean squared error (RMSE) and mean absolute error (MAE). The specification of the scoring rule should be relevant for forecast elicitation. Consider a situation where one is requested to provide a point forecasts $\hat{y}$ for some future realization of $Y \in \mathcal{Y}$ with distribution F. Without specification of the scoring rule or at least the specification of the functional of interest (such as mean, median, quantile) it is not immediately clear on what should be the correct point forecast even if the distribution F is known.

Gneiting (2011) defines a scoring function S consistent for the functional T if

$$\mathbb{E}_F\left[S(t,Y)\right] \leq \mathbb{E}_F\left[S(x,Y)\right]$$

for all $F \in \mathcal{F}$, all $t \in T(F)$ and all $x \in \mathcal{Y}$. A functional is called elicitable if the exists a consistent scoring function for it. Different consistent scoring functions elicit different functionals such as mean, median, quantiles, etc. However, not all point functionals are elicitable. For example, expected shortfall (ES) and marginal expected shortfall (MES) are not elicitable.

In this paper, we will be interested in providing density forecasts of the left-tail distribution and point forecasts of Value-at-Risk (VaR). VaR is a quantile and it is known that there is a number of consistent scoring functions for quantile forecasts. However, if the scoring function is not specified this might lead to a number of problems. The issue is that given two competing forecasts $\hat{y}^A$ and $\hat{y}^B$ and given their ranking based on a scoring function S_i generally this ranking does not extend to a different scoring rule S_j even if it elicits the same functional.

Consider forming an optimal forecast for the random observation Y_t given the forecaster's information set $\mathcal{F}_t$ as:

$$\hat{y}_t^A \equiv \arg\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[S(\hat{y},Y_t)|\mathcal{F}_t^A\right].$$

A number of propositions in Patton (2015) show that if two competing forecasts A and B are optimal under a consistent scoring function and that if the information sets $\mathcal{F}_t^A$ And $\mathcal{F}_t^B$ are nested, then their ranking is the same for all scoring functions for that particular functional (such as a mean, or quantile, or a density). However, Patton (2015) and Elliott et al. (forthcoming) provide counterexamples to show that once the information sets are non-nested or if the forecasts are based on misspecified models, then the ranking of these forecasts is sensitive to the choice of the scoring function:

$$\mathbb{E}\left[S_1(\hat{y}_t^A, Y_t)\right] \geq \mathbb{E}\left[S_1(\hat{y}_t^B, y_t)\right] \nRightarrow \mathbb{E}\left[S_2(\hat{y}_t^A, Y_t)\right] \geq \mathbb{E}\left[S_2(\hat{y}_t^B, y_t)\right]$$

for two consistent loss functions S_1 and S_2 eliciting the same functional.

In this paper we will consider scoring rules for evaluation and elicitation of density forecasts. This is relevant, as given a number of the results in the literature that show that under misspecification (or if the information sets are non-nested) the ranking of density forecasts might vary according to a scoring rule. If different scoring rules are used for elicitation of the forecasts (i.e. model estimation procedures), this might lead to improved forecasting results when evaluated based on a different of scoring rules. We will focus on the density forecasts of the left tail and the implied VaR forecasts given the predictive distribution. Before proceeding to the empirical results we provide a quick introduction into the scoring rules.

A scoring rule is a function $S : \mathcal{P} \times \mathcal{Y} \mapsto \overline{\mathbb{R}}$. That is the scoring rule assigns an extended real number given a probabilistic/density forecast $P \in \mathcal{P}$ and an observation $y \in \mathcal{Y}$. Expected score under the data generating process (DGP) Q when forecast is P is:

$$S(P, Q) = \mathbb{E}_Q\left[S(P, y)\right] = \int S(P, y) dQ(y).$$

Expected score could be used to rank different density/probabilistic forecasts. We will consider scoring rules to be positively oriented, that is the higher the score the better. Scoring rule S is proper relative to $\mathcal{P}$ if

$$S(Q, Q) \geq S(P, Q) \quad \text{for all} \quad P, Q \in \mathcal{P}.$$

Proper scoring rules are such that the expected score of a forecast equal to the true DGP is always at least as large as expected score given any other forecast. Proper scoring rules can be thought as an equivalent to a consistent scoring function for a particular functional. There are a number of examples of proper scoring rules, such as logarithmic score for density forecasts:

$$S(p, y) = \log p(y).$$

and continuous ranked probability score for probabilistic forecasts

$$S(P, y) = -\int_{-\infty}^{\infty} \left(P(y) - \mathbf{1}\{x \geq y\}\right)^2 dx.$$

The reason why one might choose to focus on density/probabilistic forecasts as opposed to point forecasts is that the forecast producer and user might not be the same subject and then it might be the most sensible for the forecast producer to provide probabilistic forecasts. Suppose that the forecast user has a loss function $L(a, y)$ where a is the action choice and y is the realization of the future observation. Given a forecast density $f(\cdot)$ the chosen action is:

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

If forecast user has two density forecasts $f(\cdot)$ and $p(\cdot)$ and the true DGP is $h(\cdot)$, then for $f = h$

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

for any loss function $L(\cdot, \cdot)$. Hence regardless of the loss function the decision maker should prefer density forecasts that are equal to the true data generating process (Diebold et al. (1998)).

For the risk management procedures one might expect that the decision maker/forecast user has interest in forecasts of only a subset of outcome space such as: tails, center, or any other interval. Define the set of possible weight functions as $\mathcal{W} \equiv \{w |\, w : \mathcal{Y} \mapsto [0, 1]\}$ that allows to represent the interest in the tails or center of the distribution. 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:

$$A_w = \{y \in \mathcal{Y} |\, w(y) > 0\}.$$

A weighted scoring rule is a loss/scoring function $S : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \mapsto \overline{\mathbb{R}}$. The scoring function S assigns an extended real value $S(f, y, w)$ for a given density forecast f, realization y of the future observation Y and a chosen weight function w. We define an expected weighted score under Q when the forecast is P for a given weight function w as $S : \mathcal{P} \times \mathcal{P} \times \mathcal{W} \mapsto \overline{\mathbb{R}}$:

$$S(p, q, w) = \mathbb{E}_Q \left[S(p, y, w) \right] = \int S(p, y, w) q(y) dy.$$

There are a number of examples of weighted scoring rules that could be used for the evaluation of the targeted density forecasts such as censored likelihood (CSL) rule by Diks et al. (2011):

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

threshold weighted continuous ranked probability score (CRPS) rule of Gneiting and Ranjan (2011):

$$S^{CRPS}(f, y, w) = - \int_{-\infty}^{\infty} (F(s) - \mathbf{1}\{y \leq s\})^2 w(s) ds;$$

and penalized weighted likelihood (PWL) scoring rule by Pelenis (2015):

$$S^{PWL}(f, y, w) = w(y) \log f(y) - \int w(s) f(s) ds + w(y).$$

We will discuss how the use of different scoring rules might be implemented in the estimation procedure and why the forecasting results might be affected given the use of different scoring rules in estimation and not just evaluation.

3 Estimation using scoring rules

It is difficult to motivate the choice of one particular scoring rule over another based on substantive economic reasoning as the density/probabilistic scoring rules lack generic axiomatic characterizations. From statistical point of view, one can think of the concepts of consistency and efficiency. If the model is correctly specified all the scoring rules eliciting the same functional (or same density) converge to the same value. Then one could choose based on the notion of efficiency. However, if the model is misspecified, the estimates/forecasts might converge to different values and then the choice of the scoring rule is less clear cut. Therefore one might consider exploring the trade-off of bias vs. efficiency (even asymptotically) under misspecification.

Suppose one is given two competing density forecasts $p(\cdot)$ and $g(\cdot)$ and that the true data generating density is $f(\cdot)$ with associated cdfs $P(\cdot)$, $G(\cdot)$ and $F(\cdot)$. Generally, it seems that given a pair of (misspecified) density forecasts $p(\cdot)$ and $g(\cdot)$ it is very easy to find a pair of proper scoring rules that would rank misspecified forecasts in different order. Lambert (2011) consider order-sensitive scoring functions and show that all proper scoring functions for scalar functionals are order sensitive. Order sensitivity implies that if the forecast A is in the interval between the true value and the forecast B, then the scoring rule is order sensitive if it prefers forecast A over forecast B. Similar result does not extend to strictly proper scoring rules for density/probabilistic forecasts. Even if $p(y) > g(y) > f(y)$ and/or $p(y) < g(y) < f(y)$ for all $y \in \mathcal{Y}$ it might be possible that a proper scoring rule would prefer forecast $p(\cdot)$ over forecast $g(\cdot)$ even if at each point y the density function $g(\cdot)$ is closer to the true data generating density.

Let y_t be a random vector of interest and let $\mathcal{F}_t$ denote the forecaster's information set at time t. Suppose that $p(y_t|\mathcal{F}_t, A)$ denotes the predicted probability density function given a particular forecasting method A. To evaluate and compare the forecasts of two methods A and B we will consider using a (strictly) proper scoring rule $S_i(p, y, w)$ and we will compute the average score of method A as:

$$\overline{S}^T_{A,i} = \frac{1}{T}\sum_{t=1}^{T} S_i(p(y_t|\mathcal{F}_t, A), y_t, w).$$

For example, it is feasible that the predictive densities are formed using a misspecified parametric model and the parameter estimates are obtained via maximum likelihood. We will consider parametric models of type $p(y_t|\mathcal{F}_t, \theta, A)$. Then given a (weighted) scoring rule S_j let the estimate $\hat{\theta}^T_{A,j}$ be:

$$\hat{\theta}^T_{A,j} = \arg\max_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} S_i(p(y_t|\mathcal{F}_t, \theta, A), y_t, w)$$

If one considers specific parametric models, then under fairly general conditions, one can show that the parameter estimates converge to some particular pseudo-true value as in $\hat{\theta}^T_{A,j} \xrightarrow{p} \theta^*_{A,j}$. If the models are misspecified, then it is very likely that given two different scoring rules S_j and S_k we would obtain that $\theta^*_{A,j} \neq \theta^*_{A,k}$ even asymptotically. Similarly, given two different forecasting methods A and B it is likely that $\theta^*_{A,j} \neq \theta^*_{B,j}$ and $\hat{\theta}^T_{A,j} \neq \hat{\theta}^T_{B,j}$. A common rule for estimation is the log score (i.e. maximum likelihood), while for evaluation a number of scoring rules have been considered in the literature (such as RMSE, "tick" loss for quantiles, weighted scoring rules, etc.) We discuss whether there might be possible gain from matching the scoring rule for both estimation and evaluation.

Choice of the weighted scoring rule might be situation specific. Weighted scoring rules could be appropriate when the subset of interest is restricted. For example, the left tail for financial risk management, tails of inflation forecasts to assess inflationary and deflationary risk. We will consider using weighted scoring rules for estimation purposes, and in particular we will focus on the PWL scoring rule with different choices of the weight function $w(\cdot)$. For example, given a sample data $\{y_t, x_t\}_{t=1}^{T}$ we might consider splitting the sample into training and evaluation samples (or use a rolling window estimation of size T_0). Given a sample of observations for $t = 1, \dots, T_0$, and a chosen model A and a weighted scoring rule S_w^{PWL} with a chosen weight function w we can find the estimates of the parameters as:

$$\hat{\theta}_{A,PWL,w}^{T_0} = \arg\max_{\theta \in \Theta} \frac{1}{T_0} \sum_{t=1}^{T_0} S^{PWL}(p(y_t|\mathcal{F}_t, \theta, A), y_t, w)$$

Given an evaluation sample for the periods of $T_0 + 1$ to T, we can find an average score $\overline{S}_i^{A,j}$ when the scoring rule for evaluation is S_i and the parameter estimates were obtained via scoring rule S_j for forecast method A:

$$\overline{S}_i^{A,j} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} S_i(p(y_t|\mathcal{F}_t, \hat{\theta}_{A,j}^{t-1}, A), y_t, w).$$

For density forecast comparison one could compare the average scores of various forecasting methods and use the testing procedures of predictive ability as suggested in Diebold and Mariano (1995) and Amisano and Giacomini (2007). When the considered models are misspecified, it has been suggested that it is possible that the average scores might be greater when the scoring rule for estimation and evaluation is matched. However, if the scoring rule used for estimation is inefficient, then it might not be the optimal choice in small samples.

4 Empirical Examples

4.1 Simulation Study

Suppose that the data generating process for the series y_t is given by a two-piece normal distribution

$$p(y_t; \mu, \sigma_1, \sigma_2) = \begin{cases} c \exp\left(-\frac{1}{2\sigma_1^2}(y_t - \mu)^2\right) & \text{if } y_t \leq \mu \\ c \exp\left(-\frac{1}{2\sigma_2^2}(y_t - \mu)^2\right) & \text{if } y_t \leq \mu \end{cases}$$

with a normalizing constant c. Wallis (2004) describes this distribution as one of the choices that has been used by Bank of England to generate inflation forecast fan charts. We will choose the true DGP parameter values as $\mu = 1, \sigma_1 = 1, \sigma_2 = 1.5$. A misspecified model A is fitted to form density forecasts, where density forecast is defined by the parameters $\theta = \{\mu, \sigma\}$ and a normal distribution:

$$p(y_t; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left((-\frac{1}{2\sigma^2}(y_t - \mu)^2\right).$$

We will consider estimating the misspecified model parameters using 4 different weighted scoring rules. We will use S^{PWL} with these alternative weight functions:

- $w_1(y) = 1 \; \forall y$.
- $w_2(y) = \mathbf{1}(y \leq 0)$.
- $w_3(y) = \mathbf{1}(0 \leq y \leq 2)$.
- $w_4(y) = \mathbf{1}(2 \leq y)$.

where

$$S^{PWL}(f, y, w) = w(y) \log f(y) - \int w(s) f(s) ds + w(y).$$

Asymptotic results can be explored first in this simulation study. That is we are interested in the expected scores such as:

$$\mathbb{E}[S^{PWL}(p(y; \theta_j^*), y, w_j)].$$

By definition, it should be the case that the expected score under scoring rule j should be highest if the parameter θ_j^* is used for forming density forecasts. Figure 1 presents the

Figure 1: Probability density functions: true pdf and density forecasts given optimal parameter estimates under various scoring rules

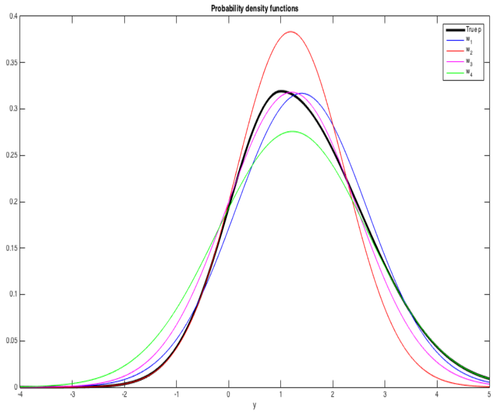

Table 1: Expected weighted scores for different estimation and evaluation scoring rules and parameter pseudo-true values

	Evaluation scoring rule					
Estimation	w_1	w_2	w_3	w_4	μ^*	σ^*
w_1 (uniform)	**-1.6502**	0.56267	-0.79436	0.58147	1.3963	1.2602
w_2 (left tail)	-1.7097	**0.56601**	-0.80064	0.52494	1.1879	1.0457
w_3 (center)	-1.6592	0.55612	**-0.79158**	0.57623	1.2266	1.2644
w_4 (right tail)	-1.677	0.53509	-0.79567	**0.58359**	1.2054	1.4509

plots of the density forecasts given asymptotic estimates of the parameters under various scoring rules. Table 1 presents the results of the forecast evaluation for different scoring rules given different parameter estimates. Unsurprisingly, we do find that the forecasts are optimal for a given weighted scoring rule when the parameters were estimates using the same scoring rule.

Our interest is in exploring the performance of different estimation approaches for different sample sizes. One could worry that the weighted scoring rules do not use the full information/data efficiently and hence might produce forecasts that are inferior even if the forecasts are evaluated using the weighted rules. Suppose that only T_0 observations are used for parameter estimation under various scoring rules. We will consider $T_0 \in \{100, 300, 1000\}$, For forecast evaluation we will consider using $T - T_0 = 100$ observations. We want to investigate the effect of sample size on forecasting performance. We repeat the simulation 10,000 times to obtain average out-of-sample scores.

The results in Table 2 do suggest that for a sample size of 100 observations it is best to use the log-scoring rule (i.e. standard maximum likelihood) even if the evaluation is done

Table 2: Average out-of-sample scores for different estimation and evaluation scoring rules

		Evaluation scoring rule			
T_0	Estimation	w_1	w_2	w_3	w_4
100	w_1	**-1.6618**	**0.56339**	**-0.79685**	**0.57165**
	w_2	-1.9665	0.55738	-0.88958	0.36566
	w_3	-1.6898	0.55355	-0.79897	0.55561
	w_4	-1.8494	0.44998	-0.86544	0.56606
300	w_1	**-1.6532**	0.56411	-0.79373	0.57637
	w_2	-1.7743	**0.56495**	-0.81822	0.479
	w_3	-1.6687	0.55671	**-0.79218**	0.56678
	w_4	-1.7116	0.51881	-0.80715	**0.57677**
1000	w_1	**-1.6513**	0.56509	-0.79456	0.5782
	w_2	-1.7247	**0.56752**	-0.80453	0.51228
	w_3	-1.6628	0.55813	**-0.79195**	0.57099
	w_4	-1.686	0.53328	-0.79913	**0.57989**

using different scoring rules. However, with a sample size of 300 or 1000 observations the optimal performance is obtained by matching the scoring rule for estimation and forecast generation and forecast evaluation.

4.2 Value-at-Risk (VaR) and left-tail density forecasting

We consider forecasting equity market returns with a particular focus on the left tail returns. For risk management purposes such quantities as Value-at-Risk (VaR) and Expected Shortfall (ES) are of interest. We are interested to analyze whether targeted estimation of the left tail via weighted scoring rules could improve the forecasts of the left-tail density and VaR. The empirical example considered is similar to the empirical exercise considered by Opschoor et al. (2015)

We observe daily returns of the stock market index of FTSE 100 for the period of January 03, 2000 until September 29, 2015. (Unreported results for S&P 500, DJIA, Nikkei are available as well.) The daily return data and corresponding realized volatility measures are sourced from Oxford-Man Institute's "realized library". A rolling window scheme is used to estimate the parameters and produce 1-step ahead forecasts with a window size of $T_0 = 750$ observations. For estimation purposes, we will consider a logarithmic scoring rule and a penalized weighted likelihood scoring rule with a weight function $w_t = \mathbf{1}(y_t < \kappa_t)$ where κ_t is the κth quantile of the previous T_0 observations. For evaluation purposes, we will consider both scoring rules and we will evaluate density and VaR forecasts. The setup of this empirical exercise is very similar to the Opschoor et al. (2015), however there the focus is on forecast combinations.

We will consider forecasting methods that specify the predictive density of the observable as:

$$y_t = \mu + \sqrt{h_t} z_t, \text{ where } z_t | \mathcal{F}_t \sim D(0,1).$$

Three specifications for the volatility dynamics will be explored: traditional GARCH(1,1) model -

$$h_t = \omega + \alpha (y_{t-1} - \mu)^2 + \beta h_{t-1};$$

EGARCH(1,1,1) model -

$$\log h_t = \omega + \gamma z_{t-1} + \alpha(|z_{t-1}| - \mathbb{E}[|z_{t-1}|]) + \beta \log h_{t-1};$$

and HEAVY model of Shephard and Sheppard (2010) -

$$h_t = \omega + \alpha R M_{t-1}^2 + \beta h_{t-1}$$

where RM_{t-1} is a realized measure of volatility at day $t-1$. To complete the specification of the return density we will examine three possibilities for the error z_t distribution $D(\cdot)$: standard normal distribution, standardized Student-t distribution with ν degrees of freedom, skewed-t distribution of Hansen (1994).

For evaluation purposes we will focus on both the 1-day ahead out-of-sample density forecasts and 1-day ahead VaR forecasts that are constructed from the predictive distributions. The results are presented in Tables 3 and 4. In this empirical exercise the use of the weighted scoring rules for estimation and forecast elicitation did not improve the density forecasts regardless whether the density forecasts were evaluated via unweighted log-score rule or a weighted scoring rule that focused on the left tail. However, the VaR forecasts were improved via the use of the weighted scoring rules for estimation and forecast elicitation purposes. This suggests that there is some usefulness of using alternative scoring rules for estimation when the target of interest is not the full distribution but a particular functional that depends on the left tail only. For further analysis one should compare the VaR forecasting results with the methods that directly estimate and forecast VaR (i.e. quantiles) rather than the methods that focus on fitting the whole data generating distribution.

Table 3: Evaluation of FTSE 1-day ahead density forecasts

	$\kappa = 0.15$					
Error distribution	Normal		Student-t		Skewed-t	
Estimation scoring rule	LS	PWL	LS	PWL	LS	PWL
	Evaluation - Log score					
GARCH	**-1.3365***	-1.3426	**-1.328***	-1.3352	**-1.3234***	-9.1199
EGARCH	-1.3136	-1.3174	-1.3098	-1.314	**-1.3047***	-1.4458
HEAVY	**-1.3098***	-1.3162	**-1.3058***	-1.3117	**-1.3008***	-188.8787
	Evaluation - PWL with $w = \mathbf{1}(y < \kappa_t)$					
GARCH	0.49488	0.497	0.50254	0.50225	0.50576	0.50348
EGARCH	0.50093	0.50357	0.50543	0.50667	**0.51027***	0.4890
HEAVY	0.50622	0.50868	0.5109	0.51143	0.51409	0.5122

Table 4: Evaluation of 1-day ahead VaR forecasts

	$\kappa = 0.15$					
Model and estimation	95% VaR			99% VaR		
	% violations	p_{CC}	p_{DQ}	% violations	p_{CC}	p_{DQ}
GARCH-N-LS	0.059	0.012	0.020	0.020	0.000	0.000
EGARCH-N-LS	0.065	0.001	0.004	0.023	0.000	0.000
HEAVY-N-LS	0.062	0.002	0.010	0.021	0.000	0.000
GARCH-N-PWL	0.048	**0.384**	**0.437**	0.016	0.001	0.000
EGARCH-N-PWL	0.056	**0.275**	**0.377**	0.016	0.002	0.000
HEAVY-N-PWL	0.048	**0.284**	**0.584**	0.015	0.007	0.000
GARCH-T-LS	0.060	0.008	0.005	0.017	0.000	0.000
EGARCH-T-LS	0.065	0.000	0.002	0.021	0.000	0.000
HEAVY-T-LS	0.064	0.000	0.001	0.019	0.000	0.000
GARCH-T-PWL	0.052	**0.622**	**0.142**	0.013	0.014	0.000
EGARCH-T-PWL	0.059	**0.060**	**0.094**	0.013	**0.154**	0.025
HEAVY-T-PWL	0.053	**0.110**	**0.271**	0.013	**0.115**	0.000
GARCH-ST-LS	0.055	**0.228**	**0.138**	0.014	0.034	0.000
EGARCH-ST-LS	0.059	0.041	**0.091**	0.015	0.020	0.000
HEAVY-ST-LS	0.055	**0.104**	**0.315**	0.013	**0.115**	0.000
GARCH-ST-PWL	0.051	**0.803**	**0.667**	0.012	**0.117**	0.000
EGARCH-ST-PWL	0.061	0.009	0.049	0.019	0.000	0.000
HEAVY-ST-PWL	0.051	**0.178**	**0.474**	0.013	**0.154**	0.000

The table reports the percentage of VaR violations, and the p-values of Conditional Coverage (CC) and the Dynamic Quantile (DQ) tests.

In summary, we do not find that the estimation with the PWL weighted scoring rule improves the out-of-sample density forecasts of the left tail, however it does decrease the forecasting performance over the whole domain. However, forming VaR forecasts via the esimates obtained using PWL weighted scoring rule improves VaR forecasting. Similar results are obtained if we consider using the $\kappa = 0.25$ quantile as the cut-off for our weight function. The results for the other equity indices S&P 500, DJIA, and Nikkei are comparable. The estimation and forecast construction via the use of the PWL scoring rule does not improve nor decrease the predictive ability of the density forecasts for the left tail, however the VaR forecasts are improved.

References

Alessi, L., Ghysels, E., Onorante, L., Peach, R., and Potter, S. Central Bank Macroeconomic Forecasting During the Global Financial Crisis: The European Central Bank and Federal Reserve Bank of New York Experiences. *Journal of Business & Economic Statistics*, 32(4):483–500, 2014.

Amisano, G. and Giacomini, R. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2):177–190, 2007.

Diebold, F. X., Gunther, T. A., and Tay, A. S. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–83, 1998.

Diebold, F. X. and Mariano, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Diks, C., Panchenko, V., and van Dijk, D. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230, 2011.

Elliott, G., Ghanem, D., and Krüger, F. Forecasting conditional probabilities of binary outcomes under misspecification. *Review of Economics and Statistics*, forthcoming.

Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Gneiting, T. and Ranjan, R. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

Hansen, B. E. Autoregressive Conditional Density Estimation. *International Economic Review*, 35(3):705–730, 1994.

Kapetanios, G., Mitchell, J., Price, S., and Fawcett, N. Generalised density forecast combinations. *Journal of Econometrics*, 188(1):150–165, 2015.

Lambert, N. S. Elicitation and evaluation of statistical forecasts. *working paper*, 2011.

Machete, R. L. Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, 143(10):1781–1790, 2013.

Merkle, E. C. and Steyvers, M. Choosing a Strictly Proper Scoring Rule. *Decision Analysis*, 10(4):292–304, 2013.

Opschoor, A., van Dijk, D., and van der Wel, M. Combining density forecasts using censored likelihood scoring rules. *working paper*, 2015.

Patton, A. J. Evaluating and Comparing Possibly Misspecifed Forecasts. *working paper*, 2015.

Pelenis, J. Weighted scoring rules for comparison of density forecasts on subsets of interest. *working paper*, 2015.

Shephard, N. and Sheppard, K. Realising the future: forecasting with high-frequency-based volatility (heavy) models. *Journal of Applied Econometrics*, 25(2):197–231, 2010.

Wallis, K. F. An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review*, (189):64–71, 2004.