Partial Likelihood-Based Scoring Rules for Evaluating Density Forecasts in Tails

Cees Diks∗
Department of Quantitative Economics
University of Amsterdam

Valentyn Panchenko†
School of Economics
University of New South Wales

Dick van Dijk‡
Econometric Institute
Erasmus University Rotterdam

April 8, 2009

Abstract

We propose new scoring rules based on partial likelihood for assessing the predictive accuracy of competing density forecasts over a specific region of interest, such as the left tail in financial risk management. These scoring rules are proper and can be interpreted in terms of Kullback-Leibler divergence between weighted versions of the density forecast and the true density. Existing scoring rules based on weighted likelihood favor density forecasts with more probability mass in the given region, rendering predictive accuracy tests biased towards such densities. Using our novel partial likelihood-based scoring rules avoids this problem.

Keywords: density forecast evaluation; scoring rules; weighted likelihood ratio scores; partial likelihood; risk management.

JEL Classification: C12; C22; C52; C53

∗Center for Nonlinear Dynamics in Economics and Finance, Department of Quantitative Economics, University of Amsterdam, Roetersstraat 11, NL-1018 WB Amsterdam, The Netherlands. E-mail: C.G.H.Diks@uva.nl
†School of Economics, Faculty of Business, University of New South Wales, Sydney, NSW 2052, Australia. E-mail: v.panchenko@unsw.edu.au
‡Corresponding author: Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. E-mail: djvandijk@ese.eur.nl
1 Introduction

The interest in density forecasts is rapidly expanding in both macroeconomics and finance. Undoubtedly this is due to the increased awareness that point forecasts are not very informative unless some indication of their uncertainty is provided, see Granger and Pesaran (2000) and Garratt et al. (2003) for discussions of this issue. Density forecasts, representing the future probability distribution of the random variable in question, provide the most complete measure of this uncertainty. Prominent macroeconomic applications are density forecasts of output growth and inflation obtained from a variety of sources, including statistical time series models (Clements and Smith, 2000), professional forecasters (Diebold et al., 1999), and central banks and other institutions producing so-called ‘fan charts’ for these variables (Clements, 2004; Mitchell and Hall, 2005). In finance, density forecasts play a pivotal role in risk management as they form the basis for risk measures such as Value-at-Risk (VaR) and Expected Shortfall (ES), see Dowd (2005) and McNeil et al. (2005) for general overviews and Guidolin and Timmermann (2006) for a recent empirical application. In addition, density forecasts are starting to be used in other financial decision problems, such as derivative pricing (Campbell and Diebold, 2005; Taylor and Buizza, 2006; Härdle and Hlávka, in press) and asset allocation (Guidolin and Timmermann, 2007). It is also becoming more common to use density forecasts to assess the adequacy of predictive regression models for asset returns, including stocks (Perez-Quiros and Timmermann, 2001), interest rates (Hong et al., 2004; Egorov et al., 2006) and exchange rates (Sarno and Valente, 2005; Rapach and Wohar, 2006), as well as measures of financial market volatility (Bollerslev et al., in press; Corradi et al., in press).

The increasing popularity of density forecasts has naturally led to the development of statistical tools for evaluating their accuracy. The techniques that have been proposed for this purpose can be classified into two groups. First, several approaches have been put forward for testing the quality of an individual density forecast, relative to the data-generating process. Following the seminal contribution of Diebold et al. (1998), the most prominent tests in this group are based on the probability integral transform (PIT) of Rosenblatt (1952). Under the null hypothesis that the density forecast is correctly specified, the PITs should be uniformly distributed, while for one-step ahead density forecasts they also should be independent and identically distributed. Hence, Diebold et al. (1998) consider a Kolmogorov-Smirnov test for departure from uniformity of the empirical PITs and several tests for temporal dependence. Alternative test statistics based on the PITs are developed in Berkowitz (2001), Bai (2003), Bai and Ng (2005), Hong and Li (2005), Li and Tkač (2006), and Corradi and Swanson (2006a), mainly to counter
the problems caused by parameter uncertainty and the assumption of correct dynamic specification under the null hypothesis. We refer to Clements (2005) and Corradi and Swanson (2006c) for in-depth surveys on specification tests for univariate density forecasts. An extension of the PIT-based approach to the multivariate case is considered by Diebold et al. (1999), see also Clements and Smith (2002) for an application. For more details of multivariate PITs and goodness-of-fit tests based on these, see Breymann et al. (2003) and Berg (in press), among others.

The second group of evaluation tests aims to compare two or more competing density forecasts. This problem of relative predictive accuracy has been considered by Sarno and Valente (2004), Mitchell and Hall (2005), Corradi and Swanson (2005, 2006b), Amisano and Giacomini (2007) and Bao et al. (2004, 2007). All statistics in this group compare the relative distance between the competing density forecasts and the true (but unobserved) density, albeit in different ways. Sarno and Valente (2004) consider the integrated squared difference as distance measure, while Corradi and Swanson (2005, 2006b) employ the mean squared error between the cumulative distribution function (CDF) of the density forecast and the true CDF. The other studies in this group develop tests of equal predictive accuracy based on a comparison of the Kullback-Leibler Information Criterion (KLIC). Amisano and Giacomini (2007) provide an interesting interpretation of the KLIC-based comparison in terms of scoring rules, which are loss functions depending on the density forecast and the actually observed data. In particular, it is shown that the difference between the logarithmic scoring rule for two competing density forecasts corresponds exactly to their relative KLIC values.

In many applications of density forecasts, we are mostly interested in a particular region of the density. Financial risk management is an example in case, where the main concern is obtaining an accurate description of the left tail of the distribution. Bao et al. (2004) and Amisano and Giacomini (2007) suggest weighted likelihood ratio (LR) tests based on KLIC-type scoring rules for evaluating and comparing density forecasts in a particular region. However, as mentioned by Corradi and Swanson (2006c) measuring the accuracy of density forecasts over a specific region cannot be done in a straightforward manner using the KLIC. The problem that occurs with KLIC-based scoring rules is that they favor density forecasts which have more probability mass in the region of interest, rendering the resulting tests biased towards such density forecasts.

In this paper we demonstrate that two density forecasts can be compared on a specific region of interest in a natural way by using unweighted likelihood scores, but with the full likelihood replaced by
partial likelihood (Cox, 1975). We specifically introduce two different scoring rules based on partial likelihood. The first rule considers the value of the conditional likelihood, given that the actual observation lies in the region of interest. The second rule is based on the censored likelihood, where the region of interest determines if an observation is censored or not. We show analytically that these partial likelihood scoring rules are proper, which means to say that a correctly specified density forecast necessarily receives a better average score than an incorrectly specified density forecast (Winkler and Murphy, 1968; Gneiting and Raftery, 2007). In particular, such proper scoring roles do not suffer from spurious rejections against densities with more probability mass in the region of interest. This is confirmed by our Monte Carlo simulations. We find that the scoring rule based on the censored likelihood, which uses more of the relevant information present, performs better in all cases considered.

Gneiting and Ranjan (2008) independently consider the problems with the weighted likelihood scoring rule. They point out that this scoring rule is in fact improper, such that it may happen that an incorrect density forecast receives a better average score than a correctly specified density forecast. They propose alternative scoring rules based on weighted quantile scoring rules, which can be shown to be proper. Our aim in this paper is different in that we specifically want to find alternative proper scoring rules that generalize the (proper) unweighted likelihood scoring rule. There are two main reasons for pursuing this. Firstly, likelihood-based scores are invariant under transformations of the outcome space, so that the scores remain unchanged if the variables are represented in other coordinates. Quantile-based scoring rules are not coordinate-free in this sense. Secondly, the likelihood-based scoring rules, at least in the unweighted case, naturally lead to LR statistics, and it is known that tests based on these have optimal power against specific alternatives, as emphasized by Berkowitz (2001) in the context of density forecast evaluation.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss conventional scoring rules based on the KLIC distance for evaluating density forecasts and illustrate the problem with the weighted versions of the resulting LR tests when these are used to focus on a particular region of the density. In Sections 3 and 4, we develop alternative scoring rules based on partial likelihood, and demonstrate that these do not suffer from this shortcoming. This is further illustrated by means of Monte Carlo simulation experiments in Section 5, where we assess the properties of tests of equal predictive accuracy of density forecasts with the different scoring rules. We provide an empirical application concerning density forecasts for daily S&P 500 returns in Section 6, demonstrating the practical usefulness

3
of our approach. Section 7 summarizes the results.

2 Scoring rules for evaluating density forecasts

We consider a stochastic process \( \{ Z_t : \Omega \rightarrow \mathbb{R}^{k+1} \}_{t=1}^{T} \), defined on a complete probability space \((\Omega, \mathcal{F}, \mathbb{P})\), and identify \( Z_t \) with \((Y_t, X_t')'\), where \( Y_t : \Omega \rightarrow \mathbb{R} \) is the real valued random variable of interest and \( X_t : \Omega \rightarrow \mathbb{R}^k \) is a vector of additional observables. The information set at time \( t \) is defined as \( \mathcal{F}_t = \sigma(Z_1', \ldots, Z_t') \). We consider the case where two competing forecast methods are available, each producing one-step ahead density forecasts, i.e. predictive densities of \( Y_{t+1} \), based on \( \mathcal{F}_t \). The competing density forecasts are denoted by the probability density functions (pdfs) \( \hat{f}_t(Y_{t+1}) \) and \( \hat{g}_t(Y_{t+1}) \), respectively. As in Amisano and Giacomini (2007), by ‘forecast method’ we mean the set of choices that the forecaster makes at the time of the prediction, including the variables \( X_t \), the econometric model (if any), and the estimation method. The only requirement that we impose on the forecast methods is that the density forecasts depend only on the \( m \) most recent observations \( Z_{t-m+1}, \ldots, Z_t \). Forecast methods of this type arise for instance when model-based density forecasts are made and model parameters are estimated with a moving window of \( m \) observations. The reason for focusing on forecast methods rather than on forecast models is that this allows for treating the parameter estimation uncertainty as an integral part of the density forecasts. Requiring the use of a finite (rolling) window of \( m \) past observations for parameter estimation then considerably simplifies the asymptotic theory of tests of equal predictive accuracy, as demonstrated by Giacomini and White (2006). It also turns out to be convenient as it enables comparison of density forecasts based on both nested and non-nested models, in contrast to other approaches such as West (1996).

Our interest lies in comparing the relative performance of \( \hat{f}_t(Y_{t+1}) \) and \( \hat{g}_t(Y_{t+1}) \), based on the sequence of \( n = T - m \) available density forecasts for \( t = m, m+1, \ldots, T-1 \). One of the approaches that has been put forward for this purpose is based on scoring rules, which are commonly used in probability forecast evaluation, see Diebold and Lopez (1996). Lahiri and Wang (2007) provide an interesting application of several such rules to the evaluation of probability forecasts of gross domestic product (GDP) declines, that is, a rare event comparable to VaR violations. In the current context, a scoring rule is a loss function \( S(\hat{f}_t; y_{t+1}) \) depending on the density forecast and the actually observed data, such that density forecast that is ‘better’ receives a higher score. Given a sequence of one-step-ahead density forecasts and corresponding realizations of the time series variable \( Y_{t+1} \), competing density forecasts may then be
compared based on their average scores.

Mitchell and Hall (2005), Amisano and Giacomini (2007), and Bao et al. (2004, 2007) focus on the logarithmic scoring rule

$$S_l(\hat{f}_t; y_{t+1}) = \log \hat{f}_t(y_{t+1}),$$ (1)

where $y_{t+1}$ is the observed value of the variable of interest. The logarithmic scoring rule assigns a high score to a density forecast if an observation falls within a region with high predictive density, and a low score if it falls within a region with low predictive density. Based on the $n$ observations available for evaluation, $y_{m+1}, \ldots, y_T$, where $T = m + n$, the density forecasts $\hat{f}_t$ and $\hat{g}_t$ can be ranked according to their average scores

$$n^{-1} \sum_{t=m}^{m+n-1} \log \hat{f}_t(y_{t+1}) \quad \text{and} \quad n^{-1} \sum_{t=m}^{m+n-1} \log \hat{g}_t(y_{t+1}).$$

The density forecast yielding the highest average score would obviously be the preferred one.

We may also test formally whether differences in average logarithmic scores are statistically significant. Defining the score difference

$$d_{t+1} = S_l(\hat{f}_t; y_{t+1}) - S_l(\hat{g}_t; y_{t+1}) = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1}),$$

the null hypothesis of equal scores is given by $H_0$: $E[d_{t+1}] = 0$, for all $t = m, m+1, \ldots$. Let $\hat{\sigma}^2_{m,n}$ denote the sample average of the score difference, that is, $\hat{\sigma}^2_{m,n} = n^{-1} \sum_{t=m}^{m+n-1} d_{t+1}^2$. Following Giacomini and White (2006) the null hypothesis may be tested against the alternative $H_a$: $E[d_{m,n}] \neq 0$, (or $< 0$ or $> 0$) for $n$ sufficiently large, by means of a Diebold and Mariano (1995) type statistic

$$t_{m,n} = \frac{\hat{\sigma}^2_{m,n}}{\sqrt{n \cdot \hat{\sigma}^2_{m,n}/n}},$$ (2)

where $\hat{\sigma}^2_{m,n}$ is a suitable estimator of $\sigma^2_{m,n} = \text{Var} \left( \sqrt{n} d_{m,n} \right)$. Giacomini and White (2006) show that a conventional heteroskedasticity and autocorrelation-consistent (HAC) variance estimator can be used, as it satisfies $\sigma^2_{m,n} - \sigma^2_{m,n} \xrightarrow{P} 0$.

Theorem 4 of Giacomini and White (2006) states that $t_{m,n}$ in (2) is asymptotically (with $m$ fixed) standard normally distributed under the null hypothesis if: (i) $\{ Z_t \}$ is mixing with $\phi$ of size $-r/(2r-2)$, $r \geq 2$, or $\alpha$ of size $-r/(r-2)$, $r > 2$; (ii) $E[d_{t+1}^2]^{2r} < \infty$ for all $t$; (iii) $\sigma^2_{m,n} = \text{Var} \left( \sqrt{n} d_{m,n} \right) > 0$ for all $n$ sufficiently large. These conditions are rather weak in that they allow for nonstationarity and heterogeneity. However, note that conditions (i) and (ii) jointly imply the existence of at least the fourth moment of $d_{t+1}^2$ for all $t$. Theorem 1.3 of Merlevède and Peligrad (2000) shows that asymptotic normality
can also be achieved under weaker distributional assumptions (existence of the second moment plus a condition relating the behavior of the tail of the distribution of $|d_{t+1}|$ to the mixing rate). However, strict stationarity is assumed by Merlevède and Peligrad (2000). The proof of Theorem 4 given by Giacomini and White (2006) is based on the central limit theorems for dependent heterogeneous processes given in Wooldridge and White (1988). The conditions required for asymptotic normality of normalized partial sums of dependent heterogeneous random variables have been further explored by de Jong (1997).

2.1 Kullback-Leibler information criterion

Intuitively, the logarithmic scoring rule is closely related to information theoretic measures of ‘goodness-of-fit’. In fact, as discussed in Mitchell and Hall (2005) and Bao et al. (2004, 2007), the sample average score difference $D_{m,n}$ in (2) may be interpreted as an estimate of the average difference of the Kullback-Leibler Information Criterion (KLIC), which for the density forecast $\hat{f}_t$ is defined as

$$
\text{KLIC}(\hat{f}_t) = \int p_t(y_{t+1}) \log \left( \frac{p_t(y_{t+1})}{\hat{f}_t(y_{t+1})} \right) dy_{t+1} = E[\log p_t(Y_{t+1}) - \log \hat{f}_t(Y_{t+1})],
$$

(3)

where $p_t$ denotes the true conditional density. Note that by taking the difference between KLIC($\hat{f}_t$) and KLIC($\hat{g}_t$) the term $E[\log p_t(Y_{t+1})]$ drops out, which solves the problem that the true density $p_t$ is unknown. Hence, the null hypothesis of equal logarithmic scores for the density forecasts $\hat{f}_t$ and $\hat{g}_t$ actually corresponds with the null hypothesis of equal KLICs. Bao et al. (2004, 2007) discuss an extension to compare multiple density forecasts, where the null hypothesis to be tested is that none of the available density forecasts is more accurate than a given benchmark, in the spirit of the reality check of White (2000). Mitchell and Hall (2005) and Hall and Mitchell (2007) also use the relative KLIC-values as a basis for combining density forecasts.

It is useful to note that both Mitchell and Hall (2005) and Bao et al. (2004, 2007) employ the same approach for testing the null hypothesis of correct specification of an individual density forecast, that is, $H_0 : \text{KLIC}(\hat{f}_t) = 0$. The problem that the true density $p_t$ in (3) is unknown then is circumvented by using the result established by Berkowitz (2001) that the KLIC of $\hat{f}_t$ relative to $p_t$ is equal to the KLIC of the density of the inverse normal transform of the PIT of the density forecast $\hat{f}_t$ relative to the standard normal density. Defining $z_{f,t+1} = \Phi^{-1}(\hat{F}_t(y_{t+1}))$ with $\hat{F}_t(y_{t+1}) = \int_{-\infty}^{y_{t+1}} \hat{f}_t(y) dy$ and $\Phi$ the standard normal distribution function, it holds true that

$$
\log p_t(y_{t+1}) - \log \hat{f}_t(y_{t+1}) = \log q_t(z_{f,t+1}) - \log \phi(z_{f,t+1}),
$$

where $q_t$ is the true conditional density of $z_{f,t+1}$ and $\phi$ is the standard normal density. This result states that the KLIC takes the same functional form before and after the inverse normal transform of $y_{t+1}$, which is essentially a consequence of the general invariance of the KLIC under invertible measurable coordinate transformations. Of course, in practice the density $q_t$ is not known either, but if $\hat{f}_t$ is correctly specified, the sequence $\{z_{f,t+1}\}$ should consist of independent standard normal random variables. As discussed in Bao et al. (2004, 2007), $q_t$ may be estimated by means of a flexible density function to obtain an estimate of the KLIC, which then allows testing for departures of $q_t$ from the standard normal.

### 2.2 Weighted scoring rules

In empirical applications of density forecasting it frequently occurs that a particular region of the density is of most interest. For example, in risk management applications such as VaR and Expected Shortfall estimation, an accurate description of the left tail of the distribution obviously is of crucial importance. In that context, it seems natural to focus on the performance of density forecasts in the region of interest and pay less attention to (or even ignore) the remaining part of the distribution. Within the framework of scoring rules, weighting may be achieved by introducing a weight function $w(y_{t+1})$ to obtain a weighted scoring rule, see Franses and van Dijk (2003) for a similar idea in the context of testing equal predictive accuracy of point forecasts.

Amisano and Giacomini (2007) propose the weighted logarithmic (wl) scoring rule

$$S_{wl}(\hat{f}_t; y_{t+1}) = w(y_{t+1}) \log \hat{f}_t(y_{t+1})$$

(4)

to assess the quality of density forecast $\hat{f}_t$, and the weighted average scores

$$n^{-1} \sum_{t=m}^{m+n-1} w(y_{t+1}) \log \hat{f}_t(y_{t+1})$$

and

$$n^{-1} \sum_{t=m}^{m+n-1} w(y_{t+1}) \log \hat{g}_t(y_{t+1})$$

for ranking two competing forecasts. Using the weighted score difference

$$d_{wl,t+1} = S_{wl}(\hat{f}_t; y_{t+1}) - S_{wl}(\hat{g}_t; y_{t+1}) = w(y_{t+1})(\log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})),$$

(5)

the null hypothesis of equal weighted scores, $H_0 : E[d_{wl,t+1}] = 0$, for all $t = m, m+1, \ldots$, may be tested by means of a Diebold-Mariano type test statistic of the form (2), with $d_{d,t+1}$ replaced by $d_{wl,t+1}$, together with an estimate of the corresponding asymptotic variance of $\sqrt{n} d_{wl,m,n}$.

Although Amisano and Giacomini (2007) focus on positive bounded weight functions, it is instructive to consider the limiting case of a ‘threshold’ weight function $w(y) = I(y \leq r)$, where $I(\mathcal{A}) = 1$ if the event $A$ occurs and zero otherwise, for some value $r$. This is a typical weight function one would
consider for evaluation of the left tail in risk management applications. However, we are then confronted
with the problem pointed out by Corradi and Swanson (2006c) that measuring the accuracy of density
forecasts over a specific region cannot be done in a straightforward manner using the KLIC or log scoring
rule. In this particular case the weighted logarithmic score may be biased towards fat-tailed densities.
This can be seen by considering the situation where \( \hat{g}_t(Y_{t+1}) > \hat{f}_t(Y_{t+1}) \) for all \( Y_{t+1} \) smaller than some
given value \( y^* \), say. Using \( w(y) = I(y \leq r) \) with \( r < y^* \) in (4) implies that the weighted score differ-
ence \( d_{t+1}^{rel} \) in (5) is never positive, and strictly negative for observations below the threshold value \( r \), such
that \( E[d_{t+1}^{rel}] \) is negative. Obviously, this can have far-reaching consequences when comparing density
forecasts with different tail behavior. In particular, there will be cases where the fat-tailed distribution \( \hat{g}_t \)
is favored over the thin-tailed distribution \( \hat{f}_t \), even if the latter is the true distribution from which the data
are drawn.

The issue we are signalling has been reported independently by Gneiting and Ranjan (2008). As they
point out, the problem is that there are density forecasts \( \hat{f}_t \) that, under the true conditional distribution,
can receive a higher average score than the actual conditional density \( p_t \). Following these authors, we
define a scoring rule to be proper if, under the true conditional density \( p_t \), no density forecast \( \hat{f}_t \) can
receive a better average score than the actual conditional density \( p_t \).

**Definition 1** A scoring rule \( S(\hat{f}_t;Y_{t+1}) \) is called proper if, for any density \( p_t \) and any density forecast
\( \hat{f}_t \),

\[
E_{p_t}(S(\hat{f}_t;Y_{t+1})) \leq E_{p_t}(S(p_t;Y_{t+1})).
\]

A scoring rule is called improper if it is not proper, and strictly proper if equality is obtained if and
only if \( \hat{f}_t(y) = p_t(y) \). By construction, a proper scoring rule assigns maximum (average) score to the
correctly specified density forecast. The following example illustrates the improperness of the weighted
logarithmic scoring rule.

[Figure 1 about here.]

**Example 1** Suppose we wish to compare the accuracy of two density forecasts for \( Y_{t+1} \), one being the
standard normal distribution with pdf

\[
\hat{f}_t(y) = (2\pi)^{-\frac{1}{2}} \exp(-y^2/2),
\]
and the other being the (fat-tailed) Student-$t$ distribution with $\nu$ degrees of freedom, standardized to unit variance, with pdf

$$
\hat{g}_t(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi\Gamma\left(\frac{\nu}{2}\right)}} 
\left(1 + \frac{y^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \quad \text{with } \nu > 2.
$$

Figure 1 shows these density functions for the case $\nu = 5$, as well as the relative log-likelihood score

$$
\log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1}).
$$

The relative score function is negative in the left tail $(-\infty, y^*)$, with $y^* \approx -2.5$. Now consider the average weighted log score $\bar{d}_{m,n}$ as defined before, based on an observed sample $y_{m+1}, \ldots, y_T$ of $n$ observations from an unknown density on $(-\infty, \infty)$ for which $\hat{f}_t(y)$ and $\hat{g}_t(y)$ are candidates. Using the threshold weight function $w(y) = I(y \leq r)$ to concentrate on the left tail, it follows from the lower panel of Figure 1 that if the threshold $r < y^*$, the average weighted log score can never be positive and will be strictly negative whenever there are observations in the tail. Evidently, the test of equal predictive accuracy will then favor the fat-tailed Student-$t$ density $\hat{g}_t(y)$, even if the true density is the standard normal $\hat{f}_t(y)$.

3 Scoring rules based on partial likelihood

The exposition in the previous section demonstrates that intuitively reasonable scoring rules can in fact favor the wrong density forecast when the evaluation concentrates on a particular region of interest. We argue that this can be avoided by requiring that score functions correspond to the logarithm of a (partial) likelihood function associated with the outcome of some statistical experiment. To see this, note that in the standard, unweighted case the log-likelihood score $\log \hat{f}_t(y_{t+1})$ is useful for measuring the divergence between a candidate density $\hat{f}_t$ and the true density $p_t$, because, under the constraint $\int_{-\infty}^{\infty} \hat{f}_t(y) \, dy = 1$, the expectation of $\log(\hat{f}_t(Y_{t+1}))$ with respect to the true density $p_t$,

$$
E_{p_t}[\log \hat{f}_t(Y_{t+1})] \equiv E[\log \hat{f}_t(Y_{t+1})|Y_{t+1} \sim p_t(y_{t+1})]
$$

is maximized by taking $\hat{f}_t = p_t$. This follows from the fact that for any density $\hat{f}_t$ different from $p_t$,

$$
E_{p_t}\left[\log \left(\frac{\hat{f}_t(Y_{t+1})}{p_t(Y_{t+1})}\right)\right] \leq E_{p_t}\left[\frac{\hat{f}_t(Y_{t+1})}{p_t(Y_{t+1})}\right] - 1 = \int_{-\infty}^{\infty} p_t(y) \frac{\hat{f}_t(y)}{p_t(y)} \, dy - 1 = 0,
$$

where the inequality follows from applying $\log x \leq x - 1$ to $\hat{f}_t/p_t$.

This shows that log-likelihood scores of different forecast methods can be compared in a meaningful way, provided that the densities under consideration are normalized to have unit total probability.
It follows that the quality of a normalized density forecast $\hat{f}_t$ can be measured properly by the KLIC $E_{p_t}[\log \hat{f}_t(Y_{t+1})]$. It is nonnegative and defines a divergence between the true and the predictive distribution. Since $p_t$ is unknown, it cannot be evaluated directly. However, we can use KLIC differences to measure the relative performance of two competing densities, which renders the logarithmic score difference discussed before, $d_{t+1} = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$.

Nothing in the above considerations prevents us from focusing on a specific regions of interest by using the likelihood based on a function of $Y_{t+1}$, rather than $Y_{t+1}$ itself. In fact the (partial) likelihood associated with any (possibly vector-valued) function $\gamma_t(Y_{t+1})$ may be used to construct a scoring rule that is proper in the sense of Gneiting and Raftery (2007). This motivates us to consider scoring rules based on partial likelihood. In the specific case of the threshold weight function $w(y) = I(y \leq r)$ we can break down the observation of $Y_{t+1}$ in two stages. First, it is revealed whether $Y_{t+1}$ is smaller than the threshold value $r$ or not. We introduce the random variable $V_{t+1}$ to denote the outcome of this first stage experiment, defining it as

$$V_{t+1} = I(Y_{t+1} \leq r)$$

In the second stage the actual value $Y_{t+1}$ is observed. The second stage experiment corresponds to a draw from the conditional distribution of $Y_{t+1}$ given the region (below or above the threshold) in which $Y_{t+1}$ lies according to the outcome of the first stage, as indicated by $V_{t+1}$.

Any (true or false) probability density function $f_t$ of $Y_{t+1}$ given $\mathcal{F}_t$ can be written as the product of the probability density function of $V_{t+1}$, which is revealed in the first stage binomial experiment, and that of the second stage experiment in which $Y_{t+1}$ is drawn from its conditional distribution given $V_{t+1}$. The likelihood function associated with the observed values $\gamma_t(Y_{t+1}) = (V_{t+1}, Y_{t+1})$ where $V_{t+1} = I(Y_{t+1} \leq r) = v$ and subsequently $Y_{t+1} = y$ can thus be written as the product of the likelihood of $V_{t+1}$, which is a Bernoulli random variable with success probability $F(r)$, and that of the realization of $Y_{t+1}$ given $v$:

$$f_{V_{t+1}|v} f_{Y_{t+1}|V_{t+1}}(y|v) = (F(r))^v (1 - F(r))^{1 - v} \left[ \left( \frac{f(y)}{F(r)} \right)^v \left( \frac{f(y)}{1 - F(r)} \right)^{1 - v} \right].$$

By disregarding either the information revealed by $V_{t+1}$ or $Y_{t+1}$ (possibly depending on the first-stage outcome $V_{t+1}$) we can construct various partial likelihood functions. This enables us to formulate several different proper scoring rules for measuring the relative performance of density forecasts.
**Conditional likelihood scoring rule** If, for a given density forecast \( \hat{f}_t \), we decide to ignore information from the first stage and use the information revealed in the second stage only if it turns out that \( V_{t+1} = 1 \) (that is, if \( Y_t \) is a tail event), we obtain the conditional likelihood (cl) score function

\[
S^{cl}(\hat{f}_t; y_{t+1}) = I(y_{t+1} \leq r) \log(\hat{f}_t(y_{t+1})/F_t(r)).
\] (7)

The main argument for using such a score function would be to evaluate density forecasts based on their behavior in the left tail (values less than or equal to \( r \)). However, due to the normalization with the total tail probability we lose information of the original density forecast on how often tail observations actually occur. This is because the information regarding this frequency is revealed only by the first-stage experiment, which we have explicitly ignored here. As a result, the conditional likelihood scoring rule attaches comparable scores to density forecasts that have similar tail shapes, but completely different tail probabilities. This tail probability is obviously relevant for risk management purposes, in particular for VaR evaluation. Hence, the following scoring rule takes into account the tail behavior as well as the relative frequency with which the tail is visited.

**Censored likelihood scoring rule** Combining the information revealed by the first stage experiment with that of the second stage provided that \( Y_{t+1} \) is a tail event (\( V_{t+1} = 1 \)), we obtain the censored likelihood (csl) score function

\[
S^{csl}(\hat{f}_t; y_{t+1}) = I(y_{t+1} \leq r) \log \hat{f}_t(y_{t+1}) + I(y_{t+1} > r) \log(1 - F_t(r)).
\] (8)

This scoring rule uses the information of the first stage (essentially information regarding the CDF \( \hat{F}_t(y) \) at \( y = r \)) but apart from that ignores the shape of \( f_t(y) \) for values above the threshold value \( r \). In that sense this scoring rule is similar to that used in the Tobit model for normally distributed random variables that cannot be observed above a certain threshold value (see Tobin, 1958).

We may test the null hypothesis of equal performance of two density forecasts \( \hat{f}_t(y) \) and \( \hat{g}_t(y) \) based on the conditional likelihood score (7) or the censored likelihood score (8) in the same manner as before. That is, given a sample of density forecasts and corresponding realizations for \( n \) time periods \( t = m, \ldots, m + n - 1 \), we may form the relative scores \( d^{cl}_{t+1} = S^{cl}(\hat{f}_t; y_{t+1}) - S^{cl}(\hat{g}_t; y_{t+1}) \) and \( d^{csl}_{t+1} = S^{csl}(\hat{f}_t; y_{t+1}) - S^{csl}(\hat{g}_t; y_{t+1}) \) and use these as the basis for computing a Diebold-Mariano type test statistic of the form given in (2).

[Figure 2 about here.]
Example 1  (continued)

We revisit the example from the previous section in order to illustrate the properties of the various scoring rules and the associated tests for comparing the accuracy of competing density forecasts. We generate \(10,000\) independent series of \(n = 2,000\) independent observations \(y_{t+1}\) from a standard normal distribution. For each sequence we compute the mean value of the weighted logarithmic scoring rule in (4), the conditional likelihood in (7), and the censored likelihood in (8). We use the threshold weight function \(w(y) = I(y \leq r)\), with the threshold fixed at \(r = -2.5\). The scores are computed for the (correct) standard normal density \(\hat{f}_t\) and the standardized Student-\(t\) density \(\hat{g}_t\) with five degrees of freedom. Figure 2 shows the empirical CDF of the mean relative scores \(\overline{a}^*_m, n\), where \(*\) is \(wl\), \(cl\) or \(csl\). The average \(wl\) scores take almost exclusively negative values, which means that, on average, they attach a lower score to the correct normal distribution than to the Student-\(t\) distribution cf. Figure 1, indicating a bias in the corresponding test statistic towards the incorrect, fat-tailed distribution. The two scoring rules based on partial likelihood both correctly favor the true normal density. The scores of the censored likelihood rule appear to be better at detecting the inadequacy of the Student-\(t\) distribution, in that its relative scores stochastically dominate those based on the conditional likelihood.

We wish to emphasize that the framework considered here differs from the evaluation of conditional quantile forecasts as considered in Giacomini and Komunjer (2005). That approach focuses on the predictive accuracy for a specific quantile of interest, such as the VaR at a certain level, whereas the partial likelihood scoring rules consider a broader region of the density. We do not claim that our method is a substitute for the conditional quantile forecast evaluation test (or any other predictive accuracy test), but suggest that they may be used in a complementary way.

We close this section with some remarks on the weight function \(w(y)\). First, the threshold weight function considered above seems an obvious choice in risk management applications, as the left tail behavior of the density forecast is of most concern there. In practice the value of \(r\) can be determined on the basis of set policies, e.g. a target VaR level for a given portfolio. If no policy guidelines are available \(r\) can be chosen to correspond to a certain quantile of the data. Second, it is possible to use a dynamic threshold weight function \(w_t(y) = I(y \leq r_t)\), where the time-varying threshold \(r_t\) is allowed to depend on \(\mathcal{F}_t\). The empirical application in section 6 involves such a dynamic threshold weight function. Third, as shown in the next section even more general weight functions may be considered, which may also depend on \(\mathcal{F}_t\) but are neither required to be threshold weight functions nor decreasing weight functions.
Such alternative weight functions may be appropriate in empirical applications of density forecasting where the focus is on a different region of the distribution than the left tail. For example, for monetary policymakers aiming to keep inflation within a certain range, the central part of the distribution may be of most interest, suggesting the use of a weight function such as \( w(y) = I(r_l \leq y \leq r_u) \), with the corresponding binary variable \( V_{t+1} \) as \( V_{t+1} = I(r_l \leq Y_{t+1} \leq r_u) \) for given lower and upper bounds \( r_l \) and \( r_u \).

4 Smooth weight functions and generalized scoring rules

The conditional and censored likelihood scoring rules in (7) and (8) define a precise region for which the density forecasts are evaluated. This is appropriate in case it is perfectly obvious which specific region is of interest. In practice this may not be so clear, and there may be a need for the possibility to use other weight functions than the threshold weight function to gradually vary the emphasis put on different regions. In this section we present possible generalizations of the conditional and censored likelihood scoring rules to such weight functions. We do not claim that these generalizations are optimal in any sense. The main aim of this section is to show that it is possible to generalize the approach based on the threshold weight function, in such a way that the scores satisfy certain intuitively desirable properties. In particular we show that the proposed generalized scores are proper and can, like the scores based on threshold weight functions, be interpreted in terms of Kullback-Leibler divergence.

**Generalized conditional likelihood scoring rule** We propose the following generalization of the conditional likelihood score in (7):

\[
S_{cl}(\hat{f}_t; Y_{t+1}) = w_t(Y_{t+1}) \log \left( \frac{\hat{f}_t(Y_{t+1})}{\hat{F}_t} \right) .
\]  

(9)

where \( \hat{F}_t = \int_{-\infty}^{\infty} w_t(s) \hat{f}_t(s) \, ds \). It can be readily seen that this is a generalization of the cl scoring rule given in (7), which is obtained by choosing \( w_t(y) = I(y \leq r) \). The weight function \( w_t(y) \), which is assumed to take values in \([0, 1]\), is allowed to depend on \( F_t \). Moreover, we assume that \( \hat{f}_t, \hat{g}_t \) and \( w_t(y) \) are such that undefined scores (e.g. due to the argument of the logarithm being zero) have zero probability under \( p_t \). It is shown in the Appendix that this scoring rule is proper, and that it can be interpreted in terms of the KL-divergence between \( w_t \)-weighted versions of \( p_t \) and the density forecasts.
Generalized censored likelihood scoring rule We propose the following generalization for the censored scoring rule:

$$S_{csl}(\hat{f}; Y_{t+1}) = w_t(Y_{t+1}) \log \hat{f}_t(Y_{t+1}) + (1 - w(Y_{t+1})) \log \left( 1 - \int_{-\infty}^{\infty} \hat{f}_t(x) w(x) \, dx \right).$$  \hspace{1cm} (10)

For the case $w_t(y) = I(y \leq r)$ the $csl$ scoring rule (8) is recovered. We impose the same conditions on the weight function as for the conditional scoring rule before. Properness of this scoring rule is also shown in the Appendix. There it is also shown that the $csl$ score can be interpreted in terms of the KL-divergence between $w_t$-weighted versions of $p_t$ and the density forecasts. \hfill \Box

Example 1 (continued)

Returning to the numerical example concerning the comparison of the normal and Student-\(t\) density forecasts, we consider logistic weight functions

$$w_t(y) = 1/(1 + \exp(a(y - r))) \quad \text{with} \quad a > 0. \hspace{1cm} (11)$$

This sigmoidal function changes monotonically from 1 to 0 as $Y_{t+1}$ increases, with $w(r) = \frac{1}{2}$ and the slope parameter $a$ determining the speed of the transition. In the limit as $a \to \infty$, the threshold weight function $I(y \leq r)$ is recovered. We fix the center at $r = -2.5$ and vary the slope parameter $a$ among the values 1, 2, 5, and 10. The integrals $\int \hat{f}_t(y) w(y) \, dy$ and $\int \hat{g}_t(y) w(y) \, dy$ are determined numerically by averaging $w(y_{t+1})$ over a large number ($10^6$) of simulated random variables $Y_{t+1}$ with density $\hat{f}_t$ and $\hat{g}_t$, respectively.

[Figure 3 about here.]

Figure 3 shows the empirical CDFs of the mean relative scores $\overline{d}_t^k$ obtained with the conditional likelihood and censored likelihood scoring rules for the different values of $a$. It can be observed that for the smoothest weight function considered ($a = 1$) the two score distributions are very similar. The difference between the scores increases as $a$ becomes larger. For $a = 10$, the logistic weight function is already very close to the threshold weight function $I(y \leq r)$, such that for larger values of $a$ essentially the same score distributions are obtained. The score distributions become more similar for smaller values of $a$ because, as $a \to 0$, $w(y)$ in (11) converges to a constant equal to $\frac{1}{2}$ for all values of $y$, so that $w(y) - (1 - w(y)) \to 0$, and moreover $\int w(y) \hat{f}_t(y) \, dy = \int w(y) \hat{g}_t(y) \, dy \to \frac{1}{2}$. Consequently, both
scoring rules converge to the unconditional likelihood (up to a constant factor 2) and the relative scores $d_{t+1}^{cl}$ and $d_{t+1}^{csl}$ have the limit
\[ \frac{1}{2}(\log \hat{g}_t(y_{t+1}) - \log \hat{f}_t(y_{t+1})). \]

5 Monte Carlo simulations

In this section we examine the implications of using the weighted logarithmic scoring rule in (4), the conditional likelihood in (7), and the censored likelihood in (8) for constructing a test of equal predictive ability of two competing density forecasts. Specifically, we consider the size and power properties of the Diebold-Mariano type test as given in (2). The null hypothesis states that the two competing density forecasts have equal expected scores, or
\[ H_0 : E[d^*_t+1] = 0, \quad \text{for } t = m, m + 1, \ldots, m + n - 1 \]
under scoring rule $\ast$, where $\ast$ is either $wl$, $cl$ or $csl$, where as before $m$ denotes the length of rolling window used for constructing the density forecast and $n$ denotes the number of forecasts. We focus on one-sided rejection rates to highlight the fact that some of the scoring rules may favor a wrongly specified density forecast over a correctly specified one.

Throughout we use a HAC-estimator for the asymptotic variance of the average relative score $\overline{d}_{m,n}^*,$ that is $\hat{\sigma}_{m,n}^2 = \hat{\gamma}_0 + 2 \sum_{k=1}^{K-1} a_k \hat{\gamma}_k,$ where $\hat{\gamma}_k$ denotes the lag-$k$ sample covariance of the sequence $\{d^*_{t+1}\}_{t=m+n-1}$ and $a_k$ are the Bartlett weights $a_k = 1 - k/K$ with $K = \lfloor n^{1/4} \rfloor$.

5.1 Size

In order to assess the size properties of the tests a case is required with two competing predictive densities that are both ‘equally (in)correct’. However, whether or not the null hypothesis of equal predictive ability holds depends on the weight function $w(y)$ that is used in the scoring rules. This complicates the simulation design, also given the fact that we would like to examine how the behavior of the tests depends on the specific settings of the weight function. For the threshold weight function $w(y) = I(y \leq r)$ it appears to be impossible to construct an example with two different density forecasts having identical predictive ability regardless of the value of $r$. We therefore evaluate the size of the tests by focusing on the central part of the distribution using the weight function $w(y) = I(-r \leq y \leq r)$. As mentioned before, in some cases this region of the distribution may be of primary interest, for instance to monetary policymakers.
aiming to keep inflation between certain lower and upper bounds. The data generating process (DGP) is taken to be i.i.d. standard normal, while the two competing density forecasts are normal distributions with means equal to $-0.2$ and $0.2$, and variance equal to 1. In this case, independent of the value of $r$ the competing density forecasts have equal predictive ability, as the scoring rules considered here are invariant under a simultaneous reflection about zero of all densities of interest (the true conditional density as well as the two competing density forecasts under consideration). In addition, it turns out that for this combination of DGP and predictive densities, the relative scores $d_{t+1}^*$ for the $wl$, $cl$ and $csl$ rules based on $w(y) = I(-r \leq y \leq r)$ are identical; observations outside the interval $[-r, r]$ do not support evidence in favor of either density forecast, which is reflected in equal scores for the two forecasts, under any of the scoring rules considered.

Figure 4 displays one-sided rejection rates (at nominal significance levels of 1, 5 and 10%) of the null hypothesis against the alternative that the $N(0.2, 1)$ distribution has better predictive ability as a function of the threshold value $r$, for sample size $n = 500$ (based on 10,000 replications). The rejection rates of the tests are quite close to the nominal significance levels for all values of $r$. Unreported results for different values of $n$ show that this holds even for sample sizes as small as 100 observations. Hence, the size properties of the tests appear to be satisfactory.

5.2 Power

We evaluate the power of the test statistics by performing simulation experiments where one of the competing density forecasts is correct, i.e. corresponds exactly with the underlying DGP. In that case the true density always is the best possible one, regardless of the region for which the densities are evaluated, that is, regardless of the weight function used in the scoring rules. Given that our main focus in this paper has been on comparing density forecasts in the left tail, in these experiments we return to the threshold weight function $w(y) = I(y \leq r)$ for the $wl$, $cl$ and $csl$ rules.

Figures 5 and 6 shows the observed rejection rates for sample sizes $n = 500$ and 2,000, respectively (again based on 10,000 replications), for data drawn from the standard normal distribution (top row) or
the standardized Student-$t(5)$ distribution (bottom row). In both cases, the null hypothesis being tested is equal predictive accuracy of the standard normal and standardized $t(5)$ density forecasts. The left (right) panels in these Figures show rejection rates (at nominal significance level 5%) against superior predictive ability of the standard normal (standardized $t(5)$) distribution, as a function of the threshold parameter $r$. Hence, the top left and bottom right panels report true power (rejections in favor of the correct density), while the top right and bottom left panels report spurious power (rejections in favor of the incorrect density).

Several interesting conclusions emerge from these graphs. First, for large values of the threshold $r$, the tests based on the $wl$, $cl$ and $csl$ scoring rules behave similarly (recall that they become identical in the limit as $r \to \infty$) and achieve rejection rates against the correct alternative of around 80% for $n = 500$ and nearly 100% for $n = 2,000$.

Second, the power of the weighted logarithmic scoring rule depends strongly on the threshold parameter $r$. For the normal DGP, for example, the test has excellent power for values of $r$ larger than 2 and between $-2$ and 0, but for other threshold values the rejection rates against the correct alternative drop to zero. In fact, for these regions of threshold values, we observe substantial rejection rates against the incorrect alternative of superior predictive ability of the Student-$t$ density. Comparing Figures 5 and 6 shows that this is not a small sample problem. In fact, the spurious power increases as $n$ becomes larger. For large negative values of $r$, this behavior can be understood from the bottom graph of Figure 1, showing that the logarithmic score is higher for the Student-$t$ density than for the normal density for all values of $y$ below $-2.5$, approximately. To understand the non-monotonous nature of these power curves more fully, we use numerical integration to obtain the expected relative score $E[w_{t-1}]$ for various values of the threshold $r$ for i.i.d. standard normal data. The results are shown in Figure 7. It can be observed that the mean changes sign several times, in exact accordance with the patterns in the top panels of Figures 5 and 6. Whenever the mean score difference (computed as the score of the standard normal minus the score of the standardized $t(5)$ density) is positive the associated test has high power, while it has high spurious power for negative mean scores. This implies that the $wl$ scoring rule cannot be relied upon for discriminating between competing density forecasts. For example, a rejection of the null hypothesis in favor of superior predictive accuracy of the Student $t$ density for $r \approx -2.5$ could be due to the considerable ‘true’ power of the test, as shown in the bottom-right graph in Figure 6. However,
it may equally likely be the result of the spurious power problem shown in the top-right graph.

Third, we find that the tests based on our partial likelihood scoring rules have reasonable power when a considerable part of the distribution is taken into account. For positive threshold values, rejection rates against the correct alternative are between 50 and 80% for \( n = 500 \) and close to 100% for the larger sample size \( n = 2,000 \). For the \( cl \)-based test, the power declines as \( r \) becomes smaller, due to the reduced number of observations falling in the relevant region that is taken into consideration. The power of the \( csl \)-based statistic remains higher, in particular for the normal DGP, suggesting that the additional information concerning the actual coverage probability of the left tail region helps to distinguish between the competing density forecasts.

Fourth, we find that the partial likelihood scores do not suffer from spurious power, in contrast to the weighted logarithmic scoring rule. For both the \( cl \) and \( csl \) rules, the rejection rates against the incorrect alternative remain below the nominal significance level of 5%. The only exception occurs for the i.i.d. standardized \( t(5) \) DGP, where the \( csl \)-based exhibits spurious power for small values of the threshold parameter \( r \) when \( n = 500 \). Comparing the bottom-left panels of Figures 5 and 6 suggests that this is a small sample problem though, as the rejection rates decline considerably when increasing the number of forecasts to \( n = 2,000 \).

Finally, to study the power properties of the tests when they are used to compare density forecasts on the central part of the distribution, we perform the same simulation experiments but using the weight function \( w(y) = I(-r \leq y \leq r) \) in the various scoring rules. Only a small part of these results are included here; full details are available upon request. Figure 8 shows rejection rates obtained for an i.i.d. standard normal DGP, when we test the null of equal predictive ability of the \( N(0, 1) \) and standardized \( t(5) \) distributions against the alternative that either of these density forecasts has better predictive ability, for sample size \( n = 500 \). The left panel shows rejection rates against better predictive performance of the (correct) \( N(0, 1) \) density, while the right panel shows rejection rates against better predictive performance of the (incorrect) standardized \( t(5) \) distribution. Clearly, all tests have high power, provided that the observations from a sufficiently wide interval \((-r, r)\) are taken into account. It can also be observed that the tests based on the \( wl \) score suffer from a large spurious power even for quite reasonable values of \( r \), while the spurious power for the tests based on the partial likelihood scores remains smaller than the nominal level (5%).

[Figure 8 about here.]
6 Empirical illustration

We examine the empirical relevance of our partial likelihood-based scoring rules in the context of the evaluation of density forecasts for daily stock index returns. We consider S&P 500 log-returns \( y_t = \ln(P_t/P_{t-1}) \), where \( P_t \) is the closing price on day \( t \), adjusted for dividends and stock splits. The sample period runs from January 1, 1980 until March 14, 2008, giving a total of 7115 observations (source: Datastream).

For illustrative purposes we define two forecast methods based on GARCH models in such a way that \textit{a priori} one of the methods is expected to be superior to the other. Examining a large variety of GARCH specifications for forecasting daily US stock index returns, Bao et al. (2007) conclude that the accuracy of density forecasts depends more on the choice of the distribution of the standardized innovations than on the choice of the volatility specification. Therefore, we differentiate our forecast methods in terms of the innovation distribution, while keeping identical specifications for the conditional mean and the conditional variance. We consider an AR(5) model to capture small low-order autocorrelation in the returns together with a GARCH(1,1) model to accommodate conditional heteroskedasticity, that is

\[
y_t = \mu_t + \varepsilon_t = \mu_t + \sqrt{h_t}\eta_t,
\]

where the conditional mean \( \mu_t \) and the conditional variance \( h_t \) are given by

\[
\mu_t = \rho_0 + \sum_{j=1}^{5} \rho_j y_{t-j},
\]

\[
h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1},
\]

and the standardized innovations \( \eta_t \) are i.i.d. with mean zero and variance one.

Following Bollerslev (1987), a common finding in empirical applications has been that GARCH models with a normal distribution for \( \eta_t \) are not able to fully account for the kurtosis observed in stock returns. We therefore concentrate on leptokurtic distributions for the standardized innovations. Specifically, for one forecast method the distribution of \( \eta_t \) is specified as a (standardized) Student-\( t \) distribution with \( \nu \) degrees of freedom, while for the other forecast method we use the (standardized) Laplace distribution. For the Student-\( t \) distribution we treat the number of degrees of freedom \( \nu \) as a parameter that is to be estimated. The number of degrees of freedom directly determines the value of the excess kurtosis of the standardized innovations, which is equal to \( 6/(\nu - 4) \) (assuming \( \nu > 4 \)). Due to its flexibility, the Student-\( t \) distribution has been widely used in GARCH modeling (see e.g. Bollerslev (1987), Baillie...
and Bollerslev (1989)). The standardized Laplace distribution provides a more parsimonious alternative with no additional parameters to be estimated and has been applied in the context of conditional volatility modeling by Granger and Ding (1995) and Mittnik et al. (1998). The Laplace distribution has excess kurtosis of 3, which exceeds the excess kurtosis of the Student-\(t(\nu)\) distribution for \(\nu > 6\). Because of the greater flexibility in modeling kurtosis, we may expect that the forecast method with Student-\(t\) innovations gives superior density forecasts relative to the Laplace innovations. This is indeed indicated by results in Bao et al. (2007), who evaluate these density forecasts ‘unconditionally’, that is, not focusing on a particular region of the distribution.

Our evaluation of the two forecast methods is based on their one-step ahead density forecasts for returns, using a rolling window scheme for parameter estimation. The width of the estimation window is set to \(m = 2000\) observations, so that the number of out-of-sample observations is equal to \(n = 5115\). For comparing the density forecasts’ accuracy we use the Diebold-Mariano type tests with test statistics as in (2), based on scores from the weighted logarithmic scoring rule in (4), the conditional likelihood in (7), and the censored likelihood in (8). We concentrate on the left tail of the distribution by using the threshold weight function \(w(y) = 1(y \leq r_t)\) for the \(wl\), \(cl\) and \(csl\) scoring rules. We consider two time-varying thresholds \(r_t\), that are determined as the one-day VaR estimates at the 95% and 99% level based on the corresponding quantiles of the empirical CDF of the return observations in the relevant estimation window. The score differences \(d^*_{t+1}\) are computed by subtracting the score of the GARCH-Laplace density forecast from the score of the GARCH-\(t\) density forecast, such that positive values of \(d^*_{t+1}\) indicate better predictive ability of the forecast method based on Student-\(t\) innovations.

Table 1 shows the average score differences \(\bar{d}^*_{m,n}\) with the accompanying tests of equal predictive accuracy as in (2), where we use a HAC estimator for the asymptotic variance \(\hat{\sigma}^2_{m,n}\) to account for serial dependence in the \(d^*_{t+1}\) series. The results clearly demonstrate that different conclusions follow from the different scoring rules. For both choices of the threshold \(r_t\) the \(wl\) scoring rule suggests superior predictive ability of the forecast method based on Laplace innovations. By contrast, the \(cl\) scoring rule suggests that the performance of the GARCH-\(t\) density forecasts is superior. The \(csl\) scoring rule points towards the same conclusion as the \(cl\) rule, although the evidence for better predictive ability of the GARCH-\(t\) specification is somewhat weaker. In the remainder of this section we seek to understand the reasons for these conflicting results, and explore the consequences of selecting either forecast method for risk management purposes. In addition, this allows us to obtain circumstantial evidence that shows
which of the two competing forecast methods is most appropriate.

[Table 1 about here.]

For most estimation windows, the degrees of freedom parameter in the Student-\textit{t} distribution is estimated to be (slightly) larger than 6, such that the Laplace distribution implies fatter tails than the Student-\textit{t} distribution. Hence, it may very well be that the \textit{wl} scoring rule indicates superior predictive ability of the Laplace distribution simply because this density has more probability mass in the region of interest, that is, the problem that motivated our analysis in the first place. To see this from a slightly different perspective, we compute one-day 95\% and 99\% Value-at-Risk (VaR) and Expected Shortfall (ES) estimates as implied by the two forecast methods. The $100 \times (1 - \alpha)\%$ VaR is determined as the $\alpha$-th quantile of the density forecast $\hat{f}_t$, that is, through $P_{\hat{f}_t}(Y_{t+1} \leq \text{VaR}_{\hat{f}_t}(\alpha)) = \alpha$. The Expected Shortfall is defined as the conditional mean return given that $Y_{t+1} \leq \text{VaR}_{\hat{f}_t}(\alpha)$, that is $\text{ES}_{\hat{f}_t}(\alpha) = \mathbb{E}_{\hat{f}_t}(Y_{t+1}|Y_{t+1} \leq \text{VaR}_{\hat{f}_t}(\alpha))$. Figure 9 shows the VaR estimates against the realized returns. We observe that typically the VaR estimates based on the Laplace innovations distribution are more extreme, confirming that it has fatter tails than the Student-\textit{t} innovations distribution. The same conclusion follows from the sample averages of the VaR and ES estimates, as shown in Table 2.

[Figure 9 about here.]

The VaR and ES estimates also enable us to assess in a different way which of the two innovation distributions is most appropriate. For that purpose, we first of all compute the frequency of 95\% and 99\% VaR violations, which should be close to 0.05 and 0.01, respectively, if the innovation distribution is correctly specified. We compute the likelihood ratio (LR) test of correct unconditional coverage (CUC) suggested by Christoffersen (1998) to determine whether the empirical violation frequencies differ significantly from these nominal levels. Additionally, we use Christoffersen’s (1998) LR tests of independence of VaR violations (IND) and for correct conditional coverage (CCC). Define the indicator variables $I_{f,t+1}(y_{t+1} \leq \text{VaR}_{f,t}(\alpha))$ for $\alpha = 0.05$ and 0.01, which take the value 1 if the condition in brackets is satisfied and 0 otherwise. Independence of the VaR violations is tested against a first-order Markov alternative, that is, the null hypothesis is given by $H_0 : E(I_{f,t+1}|I_{f,t}) = E(I_{f,t+1})$. In words, we test whether the probability of observing a VaR violation on day $t + 1$ is affected by observing a VaR violation on day $t$ or not. The CCC test simultaneously examines the null hypotheses of correct unconditional coverage and of independence, with the LR test statistic simply being the sum of the CUC and
IND statistics. For evaluating the adequacy of the Expected Shortfall estimates $\text{ES}_{f,t}(\alpha)$ we employ the test suggested by McNeil and Frey (2000). For every return $y_t$ that falls below the $\text{VaR}_{f,t}(\alpha)$ estimate, define the standardized ‘residual’ $e_{t+1} = (y_{t+1} - \text{ES}_{f,t}(\alpha))/\sqrt{h_{t+1}}$, where $h_{t+1}$ is the conditional variance forecast obtained from the corresponding GARCH model. Under the null of correct specification, the expected value of $e_{t+1}$ is equal to zero, which can easily be assessed by means of a two-sided $t$-test with HAC variance estimator.

[Table 2 about here.]

The results reported in Table 2 show that the empirical VaR violation frequencies are very close to the nominal levels for the Student-$t$ innovation distribution. For the Laplace distribution, they are considerably lower. This is confirmed by the CUC test, which convincingly rejects the null of correct unconditional coverage for the Laplace distribution but not for the Student-$t$ distribution. The null hypothesis of independence is not rejected in any of the cases at the 5% significance level. Finally, the McNeil and Frey (2000) test does not reject the adequacy of the 95% ES estimates for either of the two distributions, but it does for the 99% ES estimates based on the Laplace innovation distribution. In sum, the VaR and ES estimates suggest that the Student-$t$ distribution is more appropriate than the Laplace distribution, confirming the density forecast evaluation results obtained with the scoring rules based on partial likelihood. In terms of risk management, using the GARCH-Laplace forecast method would lead to larger estimates of risk than the GARCH-$t$ forecast method. This, in turn, could result in suboptimal asset allocation and hedging decisions.

7 Conclusions

In this paper we have developed new scoring rules based on partial likelihood for evaluating the predictive ability of competing density forecasts. It was shown that these scoring rules are particularly useful when the main interest lies in comparing the density forecasts’ accuracy for a specific region, such as the left tail in financial risk management applications. Directly weighted scoring rules based on KLIC are not suitable for this purpose. By construction they tend to favor density forecasts with more probability mass in the region of interest, rendering the tests of equal predictive accuracy biased towards such densities. Our novel scoring rules based on partial likelihood functions do not suffer from this problem.

We argue that likelihood-based scoring rules can be extended directly to partial likelihood. By using deliberately censored observations and the associated partial likelihood emphasis can be put on specific
regions of interest. After developing two scoring rules based on partial likelihood for the threshold weight function with uniform weight on the left tail, the scores were generalized to more general weight functions. All scoring rules can be interpreted in terms of Kullback-Leibler divergences between weighted versions of the density forecast and the actual conditional density.

Monte Carlo simulations demonstrated that the conventional scoring rules may indeed give rise to spurious rejections due to the possible bias in favor of an incorrect density forecast. The simulation results also showed that this phenomenon is virtually non-existent for the new scoring rules, and where present, diminishes quickly upon increasing the sample size. When comparing the scoring rules based on conditional likelihood and censored likelihood it was found that the latter leads to more powerful tests. This is related to the fact that more information is used by the scores based on censored likelihood.

In an empirical application to S&P 500 daily returns we investigated the use of the various scoring rules for density forecast comparison in the context of financial risk management. It was shown that the scoring rules based on KLIC and the newly proposed partial likelihood scoring rules can lead to the selection of different density forecasts. The density forecasts preferred by the partial likelihood scoring rules appear to be more appropriate as they were found to result in more accurate estimates of Value-at-Risk and Expected Shortfall.

Acknowledgments

We would like to thank Michael Clements, Frank Diebold, and Wolfgang Härdle, seminar participants at Humboldt University Berlin, Monash University, Queensland University of Technology, University of Amsterdam, University of New South Wales, University of Pennsylvanica, and the Reserve Bank of Australia, as well as participants at the 16th Society for Nonlinear Dynamics and Econometrics Conference (San Francisco, April 3-4, 2008) and the New Zealand Econometric Study Group Meeting in honor of Peter C.B. Phillips (Auckland, March 7-9, 2008) for providing useful comments and suggestions.
References


A Appendix

In this appendix it is explicitly shown that the proposed generalized cl and csl scores are proper, as argued in Section 4, and can be interpreted in terms of the Kullback-Leibler divergence between weighted density forecasts and the true conditional density.

Let \( w_t(y) \) denote the (nonnegative) weight function, which may depend on \( F_t \), and \( p_t(y_{t+1}) \) the true conditional density of \( Y_{t+1} \) given \( F_t \). Moreover, let \( P_t = \int w_t(s)p_t(s)\,ds \), \( \hat{F}_t = \int w_t(s)\hat{f}_t(s)\,ds \) and \( \hat{G}_t = \int w_t(s)\hat{g}_t(s)\,ds \).

### Generalized conditional likelihood score

The time-\( t \) conditional expected score difference for the density forecasts \( \hat{g}_t \) and \( \hat{f}_t \) is

\[
E_t \left( d_{t+1}^{cl} \right) = E_t \left( w(Y_{t+1}) \log \left( \frac{\hat{f}_t(Y_{t+1})}{\hat{F}_t} \right) - w(Y_{t+1}) \log \left( \frac{\hat{g}_t(Y_{t+1})}{\hat{G}_t} \right) \right) = M^{cl}(p_t, \hat{g}_t) - M^{cl}(p_t, \hat{f}_t),
\]

where, for a generic density \( h_t \) with \( H_t \equiv \int w_t(s)h_t(s)\,ds \),

\[
M^{cl}(p_t, h_t) = \int p_t(y) \log \left( \frac{p_t(y)}{P_t} \frac{H_t}{h_t(y)} \right) w_t(y) \, dy
= \int w_t(y)p_t(y) \log \left( \frac{p_t(y)}{P_t} \frac{H_t}{h_t(y)} \right) dy
= P_t \int \frac{w_t(y)p_t(y)}{P_t} \log \left( \frac{w_t(y)p_t(y)}{P_t} \frac{H_t}{w_t(y)h_t(y)} \right) dy
\]

Up to a positive constant \( P_t \), this equals the Kullback-Leibler divergence between the two probability densities \( w_t(y)p_t(y)/P_t \) and \( w_t(y)h_t(y)/H_t \). If \( \hat{f}_t \) is correctly specified (\( \hat{f}_t(y) = p_t(y) \)) this leaves

\[
E_t \left( d_{t+1}^{cl} \right) = M^{cl}(p_t, \hat{g}_t) - M^{cl}(p_t, p_t) = M^{cl}(p_t, \hat{g}_t) \geq 0.
\]

\[\square\]

### Generalized censored likelihood score

In this case we have

\[
E_t \left( d_{t+1}^{csl} \right) = E_t \left( \log \left( \frac{\hat{f}_t(Y_{t+1}) w(Y_{t+1}) (1 - \hat{F}_t)^{-w(Y_{t+1})}}{H_t (1 - H_t)^{-w(Y_{t+1})}} \right) - \left( \log \left( \frac{\hat{g}_t(Y_{t+1}) w(Y_{t+1}) (1 - \hat{G}_t)^{-w(Y_{t+1})}}{H_t (1 - H_t)^{-w(Y_{t+1})}} \right) \right) \right)
= M^{csl}(p_t, \hat{g}_t) - M^{csl}(p_t, \hat{f}_t),
\]

where

\[
M^{csl}(p_t, h_t) = \int p_t(y) \log \left( \frac{p_t(y) H_t}{h_t(y) w(Y_t) (1 - H_t)^{-w(Y_t)}} \right) \, dy
= \int p_t(y) \log \left( \frac{p_t(y) H_t}{H_t H_t} \frac{P_t}{P_t} \frac{1 - w(Y_t)}{1 - H_t} \right) \, dy
= M^{cl}(p_t, h_t) + \int p_t(y) \left( \log \frac{P_t}{H_t} - (1 - w_t(y)) \log \frac{1 - P_t}{1 - H_t} \right) \, dy
= M^{cl}(p_t, h_t) + K(P_t, H_t),
\]

28
The term denoted by $K(P_t, H_t)$ is the Kullback-Leibler divergence between two Bernoulli distributions with success probabilities $P_t$ and $H_t$, respectively. If $\hat{f}_t$ is correctly specified, we find

$$
\mathbb{E}_t \left( d_{i+1}^{sd} \right) = M^d(p_t, \hat{g}_t) - M^d(p_t, p_t) + K(P_t, \hat{G}_t) - K(P_t, \hat{G}_t)
$$

$$
= M^d(p_t, \hat{g}_t) + K(P_t, \hat{G}_t) \geq 0.
$$
Figure 1: Probability density functions of the standard normal distribution $\hat{f}(y_{t+1})$ and standardized Student-t(5) distribution $\hat{g}(y_{t+1})$ (upper panel) and corresponding relative log-likelihood scores $\log \hat{f}(y_{t+1}) - \log \hat{g}(y_{t+1})$ (lower panel).
Figure 2: Empirical CDFs of mean relative scores $\tilde{a}_{m,n}$ for the weighted logarithmic (wl) scoring rule in (4), the conditional likelihood (cl) in (7), and the censored likelihood (csl) in (8) for series of $n = 2,000$ independent observations from a standard normal distribution. The scoring rules are based on the threshold weight function $w(y) = 1(y \leq r)$ with $r = -2.5$. The relative score is defined as the score for the (correct) standard normal density minus the score for the standardized Student-$t(5)$ density. The graph is based on 10,000 replications.
Figure 3: Empirical CDFs of mean relative scores $\bar{\tau}_{m,n}^*$ for the generalized conditional likelihood (cl) and censored likelihood (csl) scoring rules for series of $n = 2,000$ independent observations from a standard normal distribution. The scoring rules are based on the logistic weight function $w(y)$ defined in (11) for various values of the slope parameter $a$. The relative score is defined as the score for (correct) standard normal density minus the score for the standardized Student-$t(5)$ density. The graph is based on 10,000 replications.
Figure 4: One-sided rejection rates of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (2) when using the weighted logarithmic (wl), the conditional likelihood (cl), and the censored likelihood (csl) scoring rules, under the weight function \( w(y) = I(-r \leq y \leq r) \) for sample size \( n = 500 \), based on 10,000 replications. The DGP is i.i.d. standard normal. The test compares the predictive accuracy of \( N(-0.2, 1) \) and \( N(0.2, 1) \) distributions. The graph shows rejection rates against the alternative that the \( N(0.2, 1) \) distribution has better predictive ability.
Figure 5: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (2) when using the weighted logarithmic (WL), the conditional likelihood (CL), and the censored likelihood (C(S)L) scoring rules, under the threshold weight function \( w(y) = I(y \leq r) \) for sample size \( n = 500 \), based on 10,000 replications. For the graphs in the top and bottom rows, the DGP is i.i.d. standard normal and i.i.d. standardized \( t(5) \), respectively. The test compares the predictive accuracy of the standard normal and the standardized \( t(5) \) distributions. The graphs in the left (right) panels show rejection rates against superior predictive ability of the standard normal (standardized \( t(5) \)) distribution, as a function of the threshold parameter \( r \).
Figure 6: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (2) when using the weighted logarithmic (wl), the conditional likelihood (cl), and the censored likelihood (csl) scoring rules, under the threshold weight function $w(y) = I(y \leq r)$ for sample size $n = 2,000$, based on 10,000 replications. For the graphs in the top and bottom rows, the DGP is i.i.d. standard normal and i.i.d. standardized $t(5)$, respectively. Specifications of the simulation experiments are identical to Figure 5.
Figure 7: Mean relative \( \text{wl} \) score \( E[\text{wl}_{r+1}] \) with threshold weight function \( w(y) = I(y \leq r) \) for the standard normal versus the standardized \( t(5) \) density as a function of the threshold value \( r \), for the standard normal DGP.
Figure 8: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (2) when using the weighted logarithmic (wl), the conditional likelihood (cl), and the censored likelihood (csl) scoring rules, under the weight function $w(y) = I(-r \leq y \leq r)$ for sample size $n = 500$, based on 10,000 replications. The DGP is i.i.d. standard normal. The graphs on the left and right show rejection rates against better predictive ability of the standard normal distribution compared to the standardized $t(5)$ distribution and vice versa.
Figure 9: Daily S&P 500 log-returns (black) for the period December 2, 1987 – March 14, 2008 and out-of-sample 95% and 99% VaR forecasts derived from the AR(5)-GARCH(1,1) specification using Student-t innovations (light gray) and Laplace innovations (dark gray).
Table 1: Average score differences and tests of equal predictive accuracy

<table>
<thead>
<tr>
<th>Scoring rule</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overline{d}$</td>
<td>Test</td>
<td>$\overline{d}$</td>
</tr>
<tr>
<td>$wl$</td>
<td>$-0.0053$</td>
<td>$-4.820$</td>
</tr>
<tr>
<td>$cl$</td>
<td>$0.0016$</td>
<td>$2.328$</td>
</tr>
<tr>
<td>$csl$</td>
<td>$0.0016$</td>
<td>$1.537$</td>
</tr>
</tbody>
</table>

Note: The table presents the average score difference $\overline{d}$ for the weighted logarithmic ($wl$) scoring rule in (4), the conditional likelihood ($cl$) in (7), and the censored likelihood ($csl$) in (8). The $wl$, $cl$ and $csl$ scoring rules are based on the threshold weight function $w(y_{t+1}) = I(y_{t+1} \leq r_t)$, where $r_t$ is the $\alpha$-th quantile of the empirical (in-sample) CDF, where $\alpha = 0.01$ or 0.05. The score difference $d_{t+1}$ is computed for density forecasts obtained from an AR(5)-GARCH(1,1) model with (standardized) Student-$t(\nu)$ innovations relative to the same model but with Laplace innovations, for daily S&P500 returns over the evaluation period December 2, 1987 – March 14, 2008.
### Table 2: Value-at-Risk and Expected Shortfall characteristics

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.05$</th>
<th></th>
<th>$\alpha = 0.01$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t(\nu)$</td>
<td>Laplace</td>
<td>$t(\nu)$</td>
<td>Laplace</td>
</tr>
<tr>
<td>Average VaR</td>
<td>$-0.0149$</td>
<td>$-0.0162$</td>
<td>$-0.0247$</td>
<td>$-0.0279$</td>
</tr>
<tr>
<td>Coverage ($y_t \leq \text{VaR}_t$)</td>
<td>0.0532</td>
<td>0.0407</td>
<td>0.0104</td>
<td>0.0055</td>
</tr>
<tr>
<td>CUC ($p$-value)</td>
<td>0.3019</td>
<td>0.0016</td>
<td>0.7961</td>
<td>0.0004</td>
</tr>
<tr>
<td>IND ($p$-value)</td>
<td>0.0501</td>
<td>0.3823</td>
<td>0.5809</td>
<td>0.5788</td>
</tr>
<tr>
<td>CCC ($p$-value)</td>
<td>0.0861</td>
<td>0.0046</td>
<td>0.8304</td>
<td>0.0015</td>
</tr>
<tr>
<td>Average ES</td>
<td>$-0.0209$</td>
<td>$-0.0235$</td>
<td>$-0.0312$</td>
<td>$-0.0351$</td>
</tr>
<tr>
<td>McNeil-Frey (test stat.)</td>
<td>1.0678</td>
<td>0.0851</td>
<td>1.0603</td>
<td>1.9730</td>
</tr>
<tr>
<td>McNeil-Frey ($p$-value)</td>
<td>0.2856</td>
<td>0.9322</td>
<td>0.2890</td>
<td>0.0485</td>
</tr>
</tbody>
</table>

**Note:** The average VaRs reported are the observed average 5% and 1% quantiles of the density forecasts based on the GARCH model with $t(\nu)$ and Laplace innovations, respectively. The coverages correspond with the observed fraction of returns below the respective VaRs, which ideally would coincide with the nominal rate $\alpha$. The rows labeled CUC, IND and CCC provide $p$-values for Christoffersen’s (1998) tests for correct unconditional coverage, independence of VaR violations, and correct conditional coverage, respectively. The average ES values are the expected shortfalls (equal to the conditional mean return, given a realization below the predicted VaR) based on the different density forecasts. The bottom two rows report McNeil-Frey test statistics and corresponding $p$-values for evaluating the expected shortfall estimates $\hat{ES}_{f,t}(\alpha)$. 

40