Likelihood-based scoring rules for comparing density forecasts in tails

Cees Diks, Valentyn Panchenko, Dick van Dijk

Accepted Manuscript

Likelihood-based scoring rules for comparing density forecasts in tails

Cees Diks, Valentyn Panchenko, Dick van Dijk

PII: S0304-4076(11)00080-7
DOI: 10.1016/j.jeconom.2011.04.001
Reference: ECONOM 3477

To appear in: *Journal of Econometrics*

Received date: 9 April 2009
Revised date: 10 April 2011
Accepted date: 18 April 2011

Please cite this article as: Diks, C., Panchenko, V., van Dijk, D., Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* (2011), doi:10.1016/j.jeconom.2011.04.001

Likelihood-Based Scoring Rules for Comparing 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 7, 2011

Abstract

We propose new scoring rules based on conditional and censored 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 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 likelihood-based scoring rules avoids this problem.

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

JEL Classification: C12; C22; C52; C53

*Center for Nonlinear Dynamics in Economics and Finance, Department of Quantitive 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

‡Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. E-mail: djvandijk@ese.eur.nl (corresponding author)

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 a full predictive 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.[1]

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).[2] We refer to Clements (2005) and Corradi and Swanson (2006c) for in-depth surveys on specification tests for univariate density forecasts.

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,

[1] 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, 2009) 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.*, 2009; Corradi *et al.*, 2009).

[2] Alternative test statistics based on the PIT are developed in Berkowitz (2001), Bai (2003), Bai and Ng (2005), Hong and Li (2005), Li and Tkacz (2006), and Corradi and Swanson (2006a), mainly to counter the problems caused by parameter estimation uncertainty and the assumption of correct dynamic specification under the null hypothesis.

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 between the density forecast and the true density 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 ability 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 of asset returns. Bao *et al.* (2004) and Amisano and Giacomini (2007) suggest likelihood ratio (LR) tests based on weighting the KLIC-type logarithmic scoring rule for the purpose of evaluating and comparing density forecasts over a particular region. However, as mentioned by Corradi and Swanson (2006c) the accuracy of density forecasts in a specific region cannot be measured in a straightforward manner using the KLIC. The problem that occurs is that by construction the weighted logarithmic scoring rule favors density forecasts with more probability mass in the region of interest, rendering the resulting tests of equal predictive ability biased towards such density forecasts.

In this paper we demonstrate that two density forecasts can still be compared on a specific region of interest by means of likelihood-based scoring rules in a natural way. Our proposed solution is to replace the full likelihood by the conditional likelihood, given that the actual observation lies in the region of interest, or by the censored likelihood, with censoring of the observations outside the region of interest. We show analytically that these new scoring rules can be interpreted in terms of Kullback-Leibler divergences between weighted versions of the density forecast and the actual density. This implies that the conditional likelihood and censored likelihood scoring rules favor density forecasts that approximate the true conditional density as closely as possible in the region of interest. Consequently, tests of equal

predictive accuracy of density forecasts based on these scoring rules do not suffer from spurious rejections against densities with more probability mass in that region. This is confirmed by extensive Monte Carlo simulations, in which we assess the finite sample properties of the predictive ability tests for the different scoring rules. Here we also find that the censored likelihood scoring rule, which uses more of the relevant information present, performs better in most, but not all, cases considered.

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

Gneiting and Ranjan (2008) independently also address the tendency of the weighted LR test of Amisano and Giacomini (2007) to favor density forecasts with more probability mass in the region of interest, but from a quantile forecast evaluation perspective. They point out that this tendency is a consequence of the scoring rule not being proper (Winkler and Murphy, 1968; Gneiting and Raftery, 2007), meaning that an incorrect density forecast may receive a higher average score than the true conditional density. Exactly this gives rise to the problem of spuriously favoring densities with more probability mass in the region of interest. As an alternative Gneiting and Ranjan (2008)propose weighted quantile scoring rules. Our aim in this paper is different in that we specifically want to find alternative scoring rules that generalize the unweighted likelihood scoring rule. The two main reasons for pursuing this are, first, that likelihood-based score differences are invariant under transformations of the outcome space and, second, that they lead to LR statistics, which are known to have optimal power properties, as emphasized by Berkowitz (2001) in the context of density forecast evaluation.

The remainder of this paper is organized as follows. In Section 2, we discuss conventional scoring rules based on the KLIC divergence for evaluating density forecasts and illustrate the problem with the resulting LR tests in case the logarithmic scores are weighted to focus on a particular region of interest. In Section 3 we put forward our alternative scoring rules based on conditional and censored likelihood and show analytically that the new scoring rules are proper. We assess the finite sample properties of tests of

equal predictive accuracy of density forecasts based on the different scoring rules by means of extensive Monte Carlo simulation experiments in Section 4. We provide an empirical application concerning density forecasts for daily S&P 500 returns in Section 5, demonstrating the practical usefulness of the new scores. We summarize and conclude in Section 6.

2 Scoring rules for evaluating density forecasts

We consider a stochastic process $\{Z_t : \Omega \to \mathbb{R}^{k+1}\}_{t=1}^{T}$, defined on a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and identify Z_t with $(Y_t, X_t')'$, where $Y_t : \Omega \to \mathbb{R}$ is the real valued random variable of interest and $X_t : \Omega \to \mathbb{R}^k$ is a vector of observable predictor variables. 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 of Y_{t+1} are denoted by the probability density functions (pdfs) $\hat{f}_t(y)$ and $\hat{g}_t(y)$, 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 a finite number m of most recent observations $Z_{t-m+1}, \ldots, Z_t$. Forecast methods of this type arise naturally, for instance, when density forecasts are obtained from time series regression models, for which 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 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 the one-step-ahead density forecasts $\hat{f}_t(y)$ and $\hat{g}_t(y)$. 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). 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 value y_{t+1}, such that a density forecast that is 'better' receives a higher score. Of course, what is considered to be a better forecast among two competing incorrect forecasts depends on the measure used to quantify divergences between distributions. However, as argued by Diebold *et al.* (1998) and Granger and Pesaran (2000), any rational user would prefer the true conditional density p_t of Y_{t+1} over an incorrect density forecast. This suggests that it is natural to focus, if possible, on scoring rules that are such that incorrect density forecasts $\hat{f}_t$ do not receive a higher average score than the true conditional density, that is,

$$\mathsf{E}_t\left(S(\hat{f}_t; Y_{t+1})\right) \leq \mathsf{E}_t\left(S(p_t; Y_{t+1})\right), \qquad \text{for all } t.$$

Following Gneiting and Raftery (2007), a scoring rule satisfying this condition will be called *proper*.

It is useful to note that the correct density p_t includes true parameters (if any). In practice, density forecasts typically involve estimated parameters. This implies that even if the density forecast $\hat{f}_t$ is based on a correctly specified model, but the model includes estimated parameters, the average score $\mathsf{E}_t\left(S(\hat{f}_t; Y_{t+1})\right)$ may not achieve the upper bound $\mathsf{E}_t\left(S(p_t; Y_{t+1})\right)$ due to nonvanishing estimation uncertainty. This reflects the fact that a density forecast based on a misspecified model with limited estimation uncertainty may be preferred over a density forecast based on the correct model specification having larger estimation uncertainty. Section 4.3 illustrates this issue with a Monte Carlo simulation. The above may seem to suggest that the notion of properness of scoring rules is of limited relevance in practice. Nevertheless, it does appear to be a desirable characteristic. Proper scoring rules are such that density forecasts receive a higher score when they approximate the true conditional density more closely, for example in the Kullback-Leibler sense as with the logarithmic score (2) discussed below. We have to keep in mind though that in the presence of nonvanishing estimation uncertainty, as accounted for in the adopted framework of Giacomini and White (2006), this may be a density forecast based on a misspecified model.

Given a scoring rule of one's choice, there are various ways to construct tests of equal predictive ability. Giacomini and White (2006) distinguish tests for unconditional predictive ability and conditional predictive ability. In the present paper, for clarity of exposition, we focus on tests for unconditional

predictive ability.[3] Assume that two competing density forecasts $\hat{f}_t$ and $\hat{g}_t$ and corresponding realizations of the variable Y_{t+1} are available for $t = m, m+1, \ldots, T-1$. We may then compare $\hat{f}_t$ and $\hat{g}_t$ based on their average scores, by testing formally whether their difference is statistically significant. Defining the score difference

$$d^*_{t+1} = S^*(\hat{f}_t; y_{t+1}) - S^*(\hat{g}_t; y_{t+1}),$$

for a given scoring rule S^*, the null hypothesis of equal scores is given by

$$H_0: \quad \mathsf{E}(d^*_{t+1}) = 0, \qquad \text{for all } t = m, m+1, \ldots, T-1.$$

Let $\overline{d}^*_{m,n}$ denote the sample average of the score differences, that is, $\overline{d}^*_{m,n} = n^{-1} \sum_{t=m}^{T-1} d^*_{t+1}$ with $n = T - m$. In order to test H_0 against the alternative H_a: $\mathsf{E}\left(\overline{d}^*_{m,n}\right) \neq 0$, (or < 0 or > 0) we may use a Diebold and Mariano (1995) type statistic

$$t_{m,n} = \frac{\overline{d}^*_{m,n}}{\sqrt{\hat{\sigma}^2_{m,n}/n}}, \tag{1}$$

where $\hat{\sigma}^2_{m,n}$ is a heteroskedasticity and autocorrelation-consistent (HAC) variance estimator of $\sigma^2_{m,n} = \mathsf{Var}\left(\sqrt{n}\,\overline{d}^*_{m,n}\right)$, which satisfies $\hat{\sigma}^2_{m,n} - \sigma^2_{m,n} \xrightarrow{P} 0$. The following theorem characterizes the asymptotic distribution of the test statistic under the null hypothesis.

Theorem 1 *The statistic $t_{m,n}$ in* (1) *is asymptotically (as $n \to \infty$ with m fixed) standard normally distributed under the null hypothesis if: (i) $\{Z_t\}$ is ϕ-mixing of size $-q/(2q-2)$ with $q \geq 2$, or α-mixing of size $-q/(q-2)$ with $q > 2$; (ii) $\mathsf{E}|d^*_{t+1}|^{2q} < \infty$ for all t; and (iii) $\sigma^2_{m,n} = \mathsf{Var}\left(\sqrt{n}\,\overline{d}^*_{m,n}\right) > 0$ for all n sufficiently large.*

Proof: This is Theorem 4 of Giacomini and White (2006), where a proof can also be found. □

The proof of this theorem as 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 in Theorem 1 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} for all t.

[3] The above inequality in terms of conditional expectations implies the same inequality in terms of unconditional expectations, that is, $\mathsf{E}\left(\mathsf{E}_t\left(S(\hat{f}_t; Y_{t+1})\right)\right) \leq \mathsf{E}\left(\mathsf{E}_t\left(S(p_t; Y_{t+1})\right)\right) \Rightarrow \mathsf{E}\left(S(\hat{f}_t; Y_{t+1})\right) \leq \mathsf{E}\left(S(p_t; Y_{t+1})\right)$

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 conditions required for asymptotic normality of normalized partial sums of dependent heterogeneous random variables have been further explored by De Jong (1997).

2.1 The logarithmic scoring rule and the Kullback-Leibler information criterion

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

$$S^l(\hat{f}_t; y_{t+1}) = \log \hat{f}_t(y_{t+1}), \tag{2}$$

assigning a high score to a density forecast if the observation y_{t+1} falls within a region with high predictive density $\hat{f}_t$, and a low score if it falls within a region with low predictive density. Based on the n observations available for evaluation, $y_{m+1}, \ldots, y_T$, the density forecasts $\hat{f}_t$ and $\hat{g}_t$ can be ranked according to their average scores $n^{-1}\sum_{t=m}^{T-1} \log \hat{f}_t(y_{t+1})$ and $n^{-1}\sum_{t=m}^{T-1} \log \hat{g}_t(y_{t+1})$. The density forecast yielding the highest average score would obviously be the preferred one. The sample average of the log score differences $d^l_{t+1} = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$ may be used to test whether the predictive accuracy is significantly different, using the test statistic defined in (1). Note that this coincides with the log-likelihood ratio of the two competing density forecasts.

Intuitively, the logarithmic scoring rule is closely related to information theoretic goodness-of-fit measures such as the Kullback-Leibler Information Criterion (KLIC), which for the density forecast $\hat{f}_t$ is defined as

$$\text{KLIC}(\hat{f}_t) = \mathsf{E}_t\left(\log p_t(Y_{t+1}) - \log \hat{f}_t(Y_{t+1})\right) = \int_{-\infty}^{\infty} p_t(y_{t+1}) \log\left(\frac{p_t(y_{t+1})}{\hat{f}_t(y_{t+1})}\right) \mathrm{d}y_{t+1}, \tag{3}$$

where p_t denotes the true conditional density. Obviously, a higher expected value (with respect to the true density p_t) of the logarithmic score in (2) is equivalent to a lower value of the KLIC in (3). Under the constraint $\int_{-\infty}^{\infty} \hat{f}_t(y)\,\mathrm{d}y = 1$, the expectation of $\log \hat{f}_t(Y_{t+1})$ with respect to the true density p_t 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,

$$\mathsf{E}_t\left(\log\left(\frac{\hat{f}_t(Y_{t+1})}{p_t(Y_{t+1})}\right)\right) \leq \mathsf{E}_t\left(\frac{\hat{f}_t(Y_{t+1})}{p_t(Y_{t+1})}\right) - 1 = \int_{-\infty}^{\infty} p_t(y)\frac{\hat{f}_t(y)}{p_t(y)}\,\mathrm{d}y - 1 = 0,$$

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

It thus follows that the quality of a normalized density forecast $\hat{f}_t$ can be measured properly by the log-likelihood score $S^l(\hat{f}_t; y_{t+1})$ and, equivalently, by the KLIC in (3). An advantage of the KLIC is that it has an absolute lower bound equal to zero, which is achieved if and only if the density forecast $\hat{f}_t$ is identical to the true distribution p_t. As such, its value provides a measure of the divergence between the candidate density $\hat{f}_t$ and p_t. However, since p_t is unknown, the KLIC cannot be evaluated directly (but we return to this point below). We can nevertheless use the KLIC to measure the *relative* accuracy of two competing densities, as discussed in Mitchell and Hall (2005) and Bao *et al.* (2004, 2007). Taking the difference KLIC($\hat{g}_t$) − KLIC($\hat{f}_t$) the term $\mathsf{E}_t\left(\log p_t(Y_{t+1})\right)$ drops out, solving the problem that the true density p_t is unknown. This in fact renders the logarithmic score difference $d^l_{t+1} = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$.

Summarizing the above, the null hypothesis of equal average logarithmic scores for the density forecasts $\hat{f}_t$ and $\hat{g}_t$ actually corresponds with the null hypothesis of equal KLICs. Given that the KLIC measures the divergence of the density forecasts from the true density, the use of the logarithmic scoring rule boils down to assessing which of the competing densities comes closest to the true distribution.

Bao *et al.* (2004, 2007) discuss an extension to compare multiple density forecasts based on their KLIC values, where the null hypothesis 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 KLIC for testing the null hypothesis of an individual density forecast being correct, that is, H_0 : 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 the density forecast $\hat{f}_t$ relative to p_t is equal to the KLIC of the density of the inverse normal transform of the PIT of $\hat{f}_t$ relative to the standard normal density. Defining $z_{\hat{f},t+1} = \Phi^{-1}(\hat{F}_t(y_{t+1}))$ with $\hat{F}_t(y_{t+1}) = \int_{-\infty}^{y_{t+1}} \hat{f}_t(y)\,\mathrm{d}y$ and Φ the standard normal distribution function, it holds true that

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

where q_t is the true conditional density of $z_{\hat{f},t+1}$ and ϕ is the standard normal density. This result states that the logarithmic scores are invariant to the inverse normal transform of y_{t+1}, which is essentially a consequence of the general invariance of likelihood ratios under smooth coordinate transformations. Of course, in practice the density q_t is not known either, but it may be estimated using a flexible density function. The resulting KLIC estimate then allows testing for departures of q_t from the standard normal.

2.2 Weighted logarithmic 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 ES estimation, an accurate description of the left tail of the distribution of asset returns 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, an obvious way to pursue this is to construct a *weighted* scoring rule, using a weight function $w_t(y)$ to emphasize the region of interest (see Franses and van Dijk (2003) for a similar idea in the context of testing equal predictive accuracy of point forecasts). Along this line Amisano and Giacomini (2007) propose the weighted logarithmic (*wl*) scoring rule

$$S^{wl}(\hat{f}_t; y_{t+1}) = w_t(y_{t+1}) \log \hat{f}_t(y_{t+1}) \tag{4}$$

to assess the quality of density forecast $\hat{f}_t$ on a certain region defined by the properties of $w_t(y_{t+1})$. The weighted average scores $n^{-1} \sum_{t=m}^{T-1} w_t(y_{t+1}) \log \hat{f}_t(y_{t+1})$ and $n^{-1} \sum_{t=m}^{T-1} w_t(y_{t+1}) \log \hat{g}_t(y_{t+1})$ can be used for ranking two competing forecasts, while the weighted score difference

$$d^{wl}_{t+1} = S^{wl}(\hat{f}_t; y_{t+1}) - S^{wl}(\hat{g}_t; y_{t+1}) = w_t(y_{t+1})(\log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})), \tag{5}$$

forms the basis for testing the null hypothesis of equal weighted scores, $H_0 : \mathsf{E}\left(d^{wl}_{t+1}\right) = 0$, for all $t = m, m+1, \ldots, T$, by means of a Diebold-Mariano type statistic of the form (1).

For the sake of argument it is instructive to consider the case of a 'threshold' weight function $w_t(y) = \mathrm{I}(y \leq r)$, with a fixed threshold r, where $\mathrm{I}(A) = 1$ if the event A occurs and zero otherwise. This is a simple example of a weight function we might consider for evaluation of the left tail in risk management applications. In this case, however, the weighted logarithmic score results in predictive

ability tests that are biased towards densities with more probability mass in the left tail. This can be seen by considering the situation where $\hat{g}_t(y) > \hat{f}_t(y)$ for all y smaller than some given value y^*, say. Using $w_t(y) = \mathrm{I}(y \leq r)$ for some $r < y^*$ in (4) implies that the weighted score difference d_{t+1}^{wl} in (5) is never positive, and strictly negative for observations below the threshold value r, such that $\mathsf{E}(d_{t+1}^{wl})$ is negative. Obviously, this can have far-reaching consequences when comparing density forecasts with different tail behavior. In particular, it may happen that a (relatively) fat-tailed distribution $\hat{g}_t$ is favored over a thin-tailed distribution $\hat{f}_t$, even if the latter is the true distribution from which the data are drawn, as the following example illustrates.

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 Student-t distribution with ν degrees of freedom, standardized to unit variance, with pdf

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

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 situation that we have a 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 competing candidates, and we use a threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$, with fixed threshold 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-likelihood score difference $\overline{d}_{m,n}^{wl}$ 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 $\hat{f}_t(y)$. □

2.3 Weighted probability scores

The issue we are signalling has been reported independently by Gneiting and Ranjan (2008). As they point out, the *wl* score does not satisfy the properness property, in the sense that there can be incorrect density forecasts $\hat{f}_t$ that receive a higher average score than the actual conditional density p_t. As a consequence, the associated test of equal predictive accuracy could even suggest that the incorrect density forecast is significantly better than the true density.

As discussed before, it seems reasonable to focus on proper scoring rules to avoid such inconsistencies. However, there are many different proper scoring rules one might use, raising the question which rules are suitable candidates in practice. Our main reason to focus on KLIC-based scoring rules is the close connection with likelihood ratio tests, which are known to perform well in many statistical settings. As mentioned before, the test for equal predictive ability based on the logarithmic scoring rule is nothing but a likelihood ratio test.

Before introducing our proper likelihood-based scoring rules, we briefly summarize the scoring rules proposed by Gneiting and Ranjan (2008), which may also be used for comparing density forecasts in specific regions of interest. Their starting point is the continuous ranked probability score (CRPS), which for the density forecast $\hat{f}_t$ is defined as

$$\text{CRPS}(\hat{f}_t, y_{t+1}) = \int_{-\infty}^{\infty} \text{PS}(\hat{F}_t(r), \text{I}(y_{t+1} \leq r))\, dr, \tag{6}$$

where

$$\text{PS}(\hat{F}_t(r), \text{I}(y_{t+1} \leq r)) = (\text{I}(y_{t+1} \leq r) - \hat{F}_t(r))^2$$

is the Brier probability score for the probability forecast $\hat{F}_t(r) = \int_{-\infty}^{r} \hat{f}_t(y)dy$ of the event $Y_{t+1} \leq r$. Equivalently, the CRPS may be written in terms of α-quantile forecasts $\hat{q}_{t,\alpha} = \hat{F}_t^{-1}(\alpha)$, as

$$\text{CRPS}(\hat{f}_t, y_{t+1}) = \int_0^1 \text{QS}_\alpha(\hat{q}_{t,\alpha}, y_{t+1})\, d\alpha, \tag{7}$$

where

$$\text{QS}_\alpha(\hat{q}_{t,\alpha}, y_{t+1}) = 2(\alpha - \text{I}(y_{t+1} < \hat{q}_{t,\alpha}))(y_{t+1} - \hat{q}_{t,\alpha})$$

is the quantile score (also known as the 'tick' or 'check' score) function, see also Giacomini and Komunjer (2005). As suggested by Gneiting and Ranjan (2008), the CRPS in (7) may be generalized to

emphasize certain regions of interest in the evaluation of density forecasts. Specifically, a weighted quantile scoring rule (*wqs*) may be defined as

$$S^{wqs}(\hat{f}_t; y_{t+1}) = -\int_0^1 v(\alpha)\text{QS}_\alpha(\hat{q}_{t,\alpha}, y_{t+1})\, d\alpha,$$

where $v(\alpha)$ is a nonnegative weight function on the unit interval and the minus sign on the right-hand side is inserted such that density forecasts with higher scores are preferred. Similarly, a weighted probability score (*wps*) is obtained from (6) as

$$S^{wps}(\hat{f}_t; y_{t+1}) = -\int_{-\infty}^{\infty} w_t(r)\text{PS}(\hat{F}_t(r), \text{I}(y_{t+1} \leq r))\, dr, \tag{8}$$

for some weight function w_t. Note that the same *wps* scoring rule was proposed by Corradi and Swanson (2006b) for evaluating density forecasts in case a specific region of the density is of interest rather than its whole support. In the Monte Carlo simulations in Section 4, we include Diebold-Mariano type tests based on $S^{wps}(\hat{f}_t; y_{t+1})$ for comparison purposes.

3 Scoring rules based on conditional and censored likelihood

KLIC-based scoring rules for evaluating and comparing density forecasts in a specific region of interest $A_t \subset \mathbb{R}$ can be obtained in a relatively straightforward manner. Specifically, it is natural to replace the full likelihood in (2) either by the conditional likelihood, given that the observation lies in the region of interest, or by the censored likelihood.

The conditional likelihood (*cl*) score function, given a region of interest A_t, is given by

$$S^{cl}(\hat{f}_t; y_{t+1}) = \text{I}(y_{t+1} \in A_t) \log\left(\frac{\hat{f}_t(y_{t+1})}{\int_{A_t} \hat{f}_t(s)\text{d}s}\right). \tag{9}$$

The main argument for using this scoring rule would be to evaluate density forecasts based only on their behavior in the region of interest A_t. The division by $\int_{A_t} \hat{f}_t(s)\text{d}s$ serves the purpose of normalizing the density on the region of interest, such that competing density forecasts can be compared in terms of their relative KLIC-values, as discussed before.

However, due to this normalization, the *cl* scoring rule does not take into account the accuracy of the density forecast for the total probability of the region of interest. For example, in case A_t is the left tail $y_{t+1} \leq r$, the conditional likelihood ignores whether the tail probability implied by $\hat{f}_t$ matches

with the frequency at which tail observations actually occur. As a result, the scoring rule in (9) attaches comparable scores to density forecasts that have similar tail shapes but may have completely different tail probabilities. This tail probability is obviously relevant for risk management purposes, in particular for VaR evaluation, and therefore it would be useful to include it in the density forecast evaluation. This can be achieved by using the censored likelihood (*csl*) score function, given by

$$S^{csl}(\hat{f}_t; y_{t+1}) = \mathrm{I}(y_{t+1} \in A_t) \log \hat{f}_t(y_{t+1}) + \mathrm{I}(y_{t+1} \in A_t^c) \log \left(\int_{A_t^c} \hat{f}_t(s) \mathrm{d}s \right), \tag{10}$$

where A_t^c is the complement of A_t. This scoring rule uses the likelihood associated with having an observation outside the region of interest, but apart from that ignores the shape of $\hat{f}_t$ outside A_t. In that sense this scoring rule is similar to the log-likelihood used in the Tobit model for random variables that cannot be observed above a certain threshold value (see Tobin, 1958).

The conditional and censored likelihood scoring rules as discussed above focus on a sharply defined region of interest A_t. It is possible to adapt these score functions in order to emphasize certain parts of the outcome space more generally, by going back to the original idea of using a weight function $w_t(y)$ as in (4). For this purpose, note that by setting $w_t(y) = \mathrm{I}(y \in A_t)$ the scoring rules in (9) and (10) can be rewritten as

$$S^{cl}(\hat{f}_t; y_{t+1}) = w_t(y_{t+1}) \log \left(\frac{\hat{f}_t(y_{t+1})}{\int w_t(s) \hat{f}_t(s) \mathrm{d}s} \right), \tag{11}$$

and

$$S^{csl}(\hat{f}_t; y_{t+1}) = w_t(y_{t+1}) \log \hat{f}_t(y_{t+1}) + (1 - w_t(y_{t+1})) \log \left(1 - \int w_t(s) \hat{f}_t(s) \mathrm{d}s \right). \tag{12}$$

At this point, we make the following assumptions.

Assumption 1 *The density forecasts $\hat{f}_t$ and $\hat{g}_t$ satisfy KLIC$(\hat{f}_t) < \infty$ and KLIC$(\hat{g}_t) < \infty$, where KLIC$(h_t) = \int p_t(y) \log \left(p_t(y)/h_t(y) \right) \mathrm{d}y$ is the Kullback-Leibler divergence between the density forecast h_t and the true conditional density p_t.*

Assumption 2 *The weight function $w_t(y)$ is such that (a) it is determined by the information available at time t, and hence a function of $\mathcal{F}_t$, (b) $0 \leq w_t(y) \leq 1$, and (c) $\int w_t(y) p_t(y) \mathrm{d}y > 0$.*

Assumption 1 ensures that the expected score differences for the competing density forecasts are finite.

Assumption 2 (c) is needed to avoid cases where $w_t(y)$ takes strictly positive values only outside the support of the data.

The following lemma shows that the generalized *cl* and *csl* scoring rules in (11) and (12) are proper, and hence cannot lead to spurious rejections against wrong alternatives just because these have more probability mass in the region(s) of interest.

Lemma 1 *Under Assumptions 1 and 2, the generalized conditional likelihood scoring rule given in* (11) *and the generalized censored likelihood scoring rule given in* (12) *are proper.*

The proof of this Lemma is given in Appendix A. The proof clarifies that the scoring rules in (11) and (12) can be interpreted in terms of Kullback-Leibler divergences between weighted versions of the density forecast and the actual density.

We may test the null hypothesis of equal performance of two density forecasts $\hat{f}_t(y_{t+1})$ and $\hat{g}_t(y_{t+1})$ based on the conditional likelihood score (11) or the censored likelihood score (12) in the same manner as before. That is, given a sample of density forecasts and corresponding realizations for n time periods $t = m, m+1, \ldots, T-1$, 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 for computing Diebold-Mariano type test statistics as given in (1).

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 $10,000$ series of $n = 2,000$ independent observations y_{t+1} from a standard normal distribution. For each sequence we compute the weighted logarithmic scores in (4), the conditional likelihood scores in (11), and the censored likelihood scores in (12). We use the threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$, with the threshold fixed at $r = -2.5$. The scores are computed for the (correct) standard normal density $\hat{f}_t$ and for the standardized Student-t density $\hat{g}_t$ with five degrees of freedom. Figure 2 shows the empirical CDF of the mean relative scores $\overline{d}^{*}_{m,n}$, 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 *cl* and *csl* scoring rules both correctly favor the true normal density. The censored likelihood rule appears 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.

To illustrate the behavior of the scoring rules obtained under smooth weight functions we consider the logistic weight function

$$w_t(y) = 1/(1 + \exp(a(y - r))) \qquad \text{with } a > 0. \tag{13}$$

This sigmoidal function changes monotonically from 1 to 0 as Y_{t+1} increases, while $w_t(r) = \frac{1}{2}$ and the slope parameter a determines the speed of the transition. In the limit as $a \to \infty$, the threshold weight function $\mathrm{I}(y \leq r)$ is recovered. We fix the center at $r = -2.5$ and vary the slope parameter a among the values 3, 4, 6, and 10. For $a = 10$, the logistic weight function is already very close to the threshold weight function $\mathrm{I}(y \leq r)$, such that for larger values of a the score distributions essentially do not change anymore. The integrals $\int w_t(y)\hat{f}_t(y)\,\mathrm{d}y$ and $\int w_t(y)\hat{g}_t(y)\,\mathrm{d}y$ are determined numerically by averaging 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 $\overline{d}^*_{m,n}$ obtained with the conditional likelihood and censored likelihood scoring rules for the different values of a. It can be observed that the difference between the scores increases as a becomes larger. In fact, for the smoothest weight function considered ($a = 1$) the two score distributions are very similar. The *cl* and *csl* score distributions become more alike for smaller values of a because, as $a \to 0$, $w_t(y)$ in (13) converges to a constant equal to $\frac{1}{2}$ for all values of y, so that $w_t(y) - (1 - w_t(y)) \to 0$, and moreover $\int w_t(y)\hat{f}_t(y)\,\mathrm{d}y = \int w_t(y)\hat{g}_t(y)\,\mathrm{d}y \to \frac{1}{2}$. Consequently, both scoring rules converge to the unconditional likelihood (up to a constant factor 2) and the relative scores d^{cl}_{t+1} and d^{csl}_{t+1} have the limit

$$\frac{1}{2}(\log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})).$$

□

We close this section with some remarks on the weight function $w_t(y)$ defining the region that is emphasized in the density forecast evaluation. The conditional and censored likelihood scoring rules may be applied with arbitrary weight functions, subject to the conditions stated in Lemma 1. The appropriate choice of $w_t(y)$ obviously depends on the interests of the forecast user. The threshold (or logistic) weight function considered in the example above seems a natural choice in risk management applications, as the left tail behavior of the density forecast is of most concern there. In other applications however the focus may be on different regions. For example, for monetary policymakers aiming to keep inflation within a certain range, the central part of the density may be of most interest, suggesting a weight function such as $w_t(y) = \mathrm{I}(r_l \leq y \leq r_u)$, for certain lower and upper bounds r_l and r_u.

The preceding implies that essentially it is also up to the forecast user to set the parameter(s) in the weight function, such as the threshold r in $w_t(y) = \mathrm{I}(y \leq r)$. For example, when Y_t represents the return on a given portfolio, r may be set equal to a certain quantile of the return distribution such that it corresponds with a target VaR level. In practice, r will then have to be estimated from historical data and might be set equal to the particular quantile of the m observations in the moving window that is used for constructing the density forecast at time t. This makes the weight function dynamic, i.e. $w_t(y) = \mathrm{I}(y \leq r_t)$, while it also involves estimation uncertainty, namely in the threshold r_t. As shown by Lemma 1, as long as the weight function w_t is conditionally (given $\mathcal{F}_t$) independent of Y_{t+1}, the properness property of the conditional and censored likelihood scoring rules is not affected. However, nonvanishing estimation uncertainty in the threshold may affect the power of the test of equal predictive accuracy. In Section 4.3, we verify this numerically with Monte Carlo simulations.

4 Monte Carlo simulations

In this section we examine the implications of using the weighted logarithmic scoring rule in (4), the conditional likelihood score in (11), the censored likelihood score in (12), and the weighted probability score in (8) for constructing a test of equal predictive ability of two competing density forecasts in finite samples. Specifically, we consider the size and power properties of the Diebold-Mariano type statistic as given in (1) for testing the null hypothesis that the two competing density forecasts have equal expected

scores, or

$$H_0: \quad \mathsf{E}\left(d^*_{t+1}\right) = 0, \qquad \text{for } t = m, m+1, \ldots, T-1$$

under scoring rule $*$, where $*$ is either *wl*, *wps*, *cl* or *csl*. As before m denotes the length of the rolling window used for constructing the density forecast and $n = T - m$ denotes the number of forecasts. Throughout we use a HAC-estimator for the asymptotic variance of the average relative score $\overline{d}^*_{m,n}$, that is $\hat{\sigma}^2_{m,n} = \hat{\gamma}_0 + 2\sum_{k=1}^{K-1} a_k \hat{\gamma}_k$, where $\hat{\gamma}_k$ denotes the lag-k sample covariance of the sequence $\{d^*_{t+1}\}_{t=m}^{T-1}$ and a_k are the Bartlett weights $a_k = 1 - k/K$ with $K = \lfloor n^{1/4} \rfloor$. We focus on one-sided rejection rates to highlight the fact that some of the scoring rules may favor a wrong density forecast over a correct one.

Concerning the implementation of the *wps* rule in (8), it is useful to note that it is in fact not essential for the properness of this score function to use an integral. As mentioned by Gneiting and Ranjan (2008), a weighted sum over a finite number of y-values also renders a suitable scoring rule. With this in mind, we do not attempt to obtain an accurate numerical approximation to the integral in (8), which is computationally very demanding, but simply use a discretized version with a discretization step of the y-variable of 0.1.

Initially, we examine the size and power properties of the test of equal predictive ability in an environment that does not involve parameter estimation uncertainty in Sections 4.1 and 4.2, to demonstrate the pitfalls when using *wl* scoring rule and the benefits of the *cl* and *csl* alternatives most clearly. The role of estimation uncertainty, both in the density forecasts and in the weighting function, are addressed explicitly in Section 4.3.

4.1 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 $w_t(y)$ that is used in the scoring rules. This complicates the simulation design, 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 $w_t(y) = \mathrm{I}(y \leq r)$ 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 when focusing

on the central part of the distribution by means of the weight function $w_t(y) = \mathrm{I}(-r \leq y \leq r)$. As mentioned before, in some cases this region of the distribution may be of primary interest, for instance to monetary policymakers targeting 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 different means equal to -0.2 and 0.2 and identical variance equal to 1. In this case, independent of the value of r the competing density forecasts have equal predictive accuracy, 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 density forecasts). 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 $w_t(y) = \mathrm{I}(-r \leq y \leq r)$ 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 about here

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, based on $10,000$ replications for sample size $n = 500$. 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 $n = 100$ observations. Hence, the size properties of the predictive ability test appear to be satisfactory.

4.2 Power

We evaluate the power of the test based on the various scoring rules 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 first return to the threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$.

In order to make the rejection frequencies of the null obtained for different values of r more comparable, we make the sample size n dependent on the threshold value in such a way that the expected number of observations in the region of interest, denoted by c, is constant across the various values of r. This is achieved by setting $n = c/\mathsf{P}(Y < r)$. Given that in typical risk management applications there may be only a few tail observations, we consider relatively small values of c.

Figure 5 about here

Figure 6 about here

Figures 5 and 6 show the observed rejection rates for $c = 5$ and $c = 40$, respectively, based on $10,000$ replications, for data drawn from the standard normal distribution (left column) or the standardized Student-$t(5)$ distribution (right column). In both cases, the null hypothesis being tested is equal predictive accuracy of the standard normal and standardized Student-$t(5)$ density forecasts. The top (bottom) panels in these Figures show rejection rates at nominal significance level 5% against superior predictive ability of the standard normal (standardized Student-$t(5)$) 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).

Figure 7 about here

Several interesting conclusions emerge from these graphs. First, the power of the *wl* scoring rule depends strongly on the threshold parameter r. For the normal DGP, for example, the test has excellent power for values of r between -2 and 0, but for more negative threshold values the rejection rates against the correct alternative drop to zero. In fact, for threshold values less than -2, we observe substantial spurious power in the form of rejection 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 for the *wl* rule increases as the sample size becomes larger. This behavior of the test based on the *wl* scoring rule for large negative values of r 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-monotonic nature of these power curves more fully, we use numerical integration to obtain the expected relative score $\mathsf{E}\left(d_{t+1}^{wl}\right)$ 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 Student-$t(5)$ density) is positive the associated test has high power, while it has high spurious power for negative mean scores. The *wl* scoring rule thus 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 bottom-left graph.

Second, the top-right and bottom-left panels of Figure 5 suggest that the *wps*, *cl* and *csl* scores also display some spurious power for certain regions of threshold values. However, in stark contrast to the weighted logarithmic scoring rule, this appears to be due to the extremely small sample size, as it quickly disappears as c increases. Already for $c = 40$ the rejection rates for these scoring rules against the incorrect alternative remain below the nominal significance level of 5%, see Figure 6. This clearly demonstrates the advantage of using a proper scoring rule for comparing the predictive accuracy of density forecasts.

Third, for small values of the threshold r the power for the *csl* scoring rule is higher than that of the *cl* rule, for the standard normal (top left panel) as well as for the standardized Student-$t(5)$ distributions (bottom right panel), especially for $c = 5$ (see Figure 5). Obviously, the additional information concerning the coverage probability of the left tail region helps to distinguish between the competing density forecasts, in particular when the number of observations in the region of interest is extremely small.

Fourth, for $c = 5$, the power of the different tests behaves similarly for large values of r. This should be expected on theoretical grounds for the *wl*, *cl* and *csl* scoring rules, since they become identical in the limit as $r \to \infty$. This is not the case for the *wps* scoring rule though, so its similar power for large r might be coincidental. In fact, for $c = 40$ it is visible that the *wps* rule has slightly deviating power from the other rules for large r; it is somewhat smaller for the normal DGP (top left panel of Figure 6) while

it appears to be somewhat larger for the Student-$t(5)$ DGP (lower right panel of Figure 6).

Figure 8 about here

Next, we perform the same simulation experiments but with the weight function $w_t(y) = \mathrm{I}(-r \leq y \leq r)$ to study the power properties of the tests when they are used to compare density forecasts on the central part of the distribution. 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 Student-$t(5)$ distributions against the alternative that either of these density forecasts has better predictive ability, for $c = 200$ (the number of observations in the region of interest needed to obtain a reasonable power strongly depends on the relative differences between densities). The format is the same as in Figures 5 and 6, with the left (right) column showing results when the DGP is the standard normal (standardized Student-$t(5)$) distribution. The top (bottom) panels in these Figures show rejection rates at nominal significance level 5% against superior predictive ability of the standard normal (standardized Student-$t(5)$) distribution, as a function of the threshold parameter r. It can be clearly observed that the *wl* rule displays spurious power, and that in the full information case (i.e. large values of r) the likelihood-based rules provide more powerful tests than the *wps* rule.

4.3 Estimation uncertainty and time-varying weight functions

In the remaining simulation experiments, we examine the effects of parameter estimation uncertainty. We start with a simulation addressing the effect of non-vanishing estimation uncertainty on the tests of equal predictive accuracy. In particular, we demonstrate that a forecast method using an incorrect model specification but with limited estimation uncertainty may produce a better density forecast than a forecast method based on the correct model specification but having larger estimation uncertainty. For brevity, we focus only on the (unweighted) logarithmic scoring rule (2). The results generalize to other scoring considered in the paper. The data generating process is the following AR(2) specification: $y_t = 0.8y_{t-1} + 0.05y_{t-2} + \varepsilon_t$, $\varepsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$. We compare the predictive accuracy of the AR(2) specification, which is correct up to two estimated parameters, against a more parsimonious, but incorrect AR(1) specification with one parameter to be estimated. The parameters are estimated by MLE. Recall that we work under a rolling forecast scheme, where the size of the estimation window m is fixed, so that

the estimation uncertainty does not vanish asymptotically. Table 1 shows one-sided rejection rates of the test of the equal predictive density for different rolling estimation window sizes m against the alternatives that the average log-score is higher for the AR(2) model relative to the AR(1) model and *vice versa*. For small estimation windows, $m = 100; 250$, the estimation uncertainty is relatively important and the test often indicates that the incorrectly specified, but more parsimonious AR(1) model produces better density forecasts. For intermediate values $m = 500; 1,000$ the test generally does not reject the null of equal predictive accuracy. For very large estimation windows with $m = 2,500; 5,000$, the estimation error is small enough for the test to favor the correctly specified AR(2) model. We can summarize that with small estimation samples, the density forecasts from the AR(1) model approximate the true density forecast more closely, on average, and this is rightfully detected by the log-score and the associated test.

Table 1 about here

Next, in addition to parameter estimation uncertainty in the density forecasts, we investigate the effect of using a weight function that is time-varying and depends on estimated parameters. In particular, we use a threshold weight function $w_t(y) = \mathrm{I}(y \leq \hat{r}_t^{\alpha})$, where the threshold $\hat{r}_t^{\alpha}$ is given by the empirical α-quantile obtained from a finite window of past observations. As shown by Lemma 1, the *cl* and *csl* scoring rules in (11) and in (12) remain proper in this case and the properties of the associated tests of equal predictive accuracy should not be affected.

We focus on a DGP which is more relevant for finance applications. The DGP is taken to be a GARCH(1,1) process, specified as $y_t = \sqrt{h_t}\eta_t$, with $h_t = 0.01 + 0.1y_{t-1}^2 + 0.8h_{t-1}$ and $\{\eta_t\}$ an i.i.d. standard normal sequence. We evaluate the performance of the available scoring rules in identifying the correctly specified GARCH density forecast when compared with an alternative density forecast, which differs only in the specification of the distribution of the standardized innovations η_t. Specifically, the alternative specification assumes a standardized Student-$t(5)$ distribution for η_t. The model for the conditional volatility is correctly specified in both forecast methods, up to the unknown parameters. The GARCH parameters are estimated by MLE using a rolling window of $m = 2,000$ observations and the threshold, $\hat{r}_t^{\alpha}$, is set equal to the empirical α-quantile of $y_{t-m+1}, y_{t-m+2}, \ldots, y_t$. Similarly to the previous experiments the number of observations for which density forecasts are constructed varies

depending on the number of expected observations falling within the region of interest, i.e. $n = c/\alpha$. We report results for $c = 40$.

Figure 9 about here

Given the same number of parameters in the model specifications underlying the competing forecasts and the relatively large rolling window size, $m = 2000$, we may expect that the density forecasts based on the correct specification with standard normal innovations are closer to the true conditional density than the forecasts using the standardized Student-$t(5)$ innovations. Figure 9 shows one-sided rejection rates of the null hypothesis of equal predictive abilities against better predictive ability of the forecast based on the standard normal innovations compared to the forecast based on standardized Student-$t(5)$ innovations and *vice versa* for values of α ranging between 0.01 and 0.5. The right panel shows that the *cl* and *csl* scoring rules do not display spurious power, while the *wl* rule has rejection rates substantially above the nominal level of 5%, in particular for small threshold values. This confirms that the *cl* and *csl* scoring rules can be used in combination with time-varying weight functions without introducing spurious rejections of the null hypothesis. The *cl* and *csl* scoring rules display comparable power for quantiles $\alpha > 0.10$, approximately. For lower quantiles, thus focusing more on the left tail, the additional information used by the *csl* rule again leads to improved power compared to the *cl* rule. While the *wps* scoring rule has power comparable to the *csl* rule for the lowest quantiles, its (relative) performance rapidly deteriorates as α becomes larger.

5 Empirical illustration

We examine the empirical relevance of the proposed 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 7,115 observations (source: Datastream).

For illustrative purposes we define two forecast methods based on GARCH models in such a way that *a priori* one of the methods is expected to be superior to the other. Examining a large variety of GARCH models 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 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 for the conditional mean return together with a GARCH(1,1) model for the conditional variance, that is

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

where the conditional mean μ_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 η_t are i.i.d. with mean zero and variance one.

Following Bollerslev (1987), a common finding in empirical applications of GARCH models has been that a normal distribution for η_t is not sufficient 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 η_t is specified as a (standardized) Student-t distribution with ν degrees of freedom, while for the other forecast method we use the (standardized) Laplace distribution. Note that for the Student-t distribution the degrees of freedom ν is a parameter that is to be estimated. The 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 daily returns, using a rolling window scheme for parameter estimation. The length of the estimation window is set to $m = 2,000$ observations, so that the number of out-of-sample observations is equal to $n = 5,115$. For comparing the density forecasts' accuracy we use the Diebold-Mariano type test based on the weighted logarithmic scoring rule in (4), the weighted probability scores in (8), the conditional likelihood in (11), and the censored likelihood in (12). We concentrate on the left tail of the distribution by using the threshold weight function $w_t(y) = \mathrm{I}(y \leq \hat{r}_t^{\alpha})$ for the *wl*, *wps*, *cl* and *csl* scoring rules. The time-varying threshold $\hat{r}_t^{\alpha}$ is set equal to the empirical α-quantile of the return observations in the relevant estimation window, where we consider $\alpha = 0.10, 0.05$ and 0.01. The score difference d_{t+1}^* is 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 2 about here

Table 2 shows the average score differences $\overline{d}_{m,n}^*$ with the accompanying tests of equal predictive accuracy as in (1), where we use a HAC estimator for the asymptotic variance $\hat{\sigma}_{m,n}^2$ 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 thresholds based on $\alpha = 0.05$ and 0.01 the *wl* scoring rule suggests superior predictive ability of the forecast method based on Laplace innovations, while for $\alpha = 0.1$, it fails to reject the null of equal predictive ability. By contrast, the *cl* scoring rule suggests that the performance of the GARCH-t density forecasts is superior for all three values of α. The *csl* scoring rule points towards the same conclusion as the *cl* rule, although the evidence for better predictive ability of the forecast based on the GARCH-t specification is somewhat weaker. The *wps* rule also indicates the superior performance of the GARCH-t especially for $\alpha = 0.01$, but evidence is weak when we consider less extreme quantiles $\alpha = 0.05$ and 0.1. 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.

Figure 10 about here

For most estimation windows, the degrees of freedom parameter in the Student-t distribution is estimated to be (slightly) larger than 6, such that the Laplace distribution implies fatter tails than the Student-t distribution. Hence, it may very well be that the *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 may be relevant here. To see this from a slightly different perspective, we compute one-day 90%, 95% and 99% VaR and ES estimates as implied by the two forecast methods. The $100 \times (1-\alpha)\%$ VaR is determined as the α-th quantile of the density forecast $\hat{f}_t$, that is, through $\mathsf{P}_{\hat{f},t}\left(Y_{t+1} \leq \mathrm{VaR}_{\hat{f},t}(\alpha)\right) = \alpha$. The ES is defined as the conditional mean return given that $Y_{t+1} \leq \mathrm{VaR}_{\hat{f},t}(\alpha)$, that is $\mathrm{ES}_{\hat{f},t}(\alpha) = \mathsf{E}_{\hat{f},t}\left(Y_{t+1}|Y_{t+1} \leq \mathrm{VaR}_{\hat{f},t}(\alpha)\right)$. Figure 10 shows the VaR estimates against the realized returns. We observe that typically the VaR estimates based on the Laplace innovations are more extreme, confirming that it has fatter tails than the Student-t innovations. The same conclusion follows from the sample averages of the VaR and ES estimates, as shown in Table 3.

The VaR and ES estimates also enable us to assess which of the two innovation distributions is the most appropriate in a different way. For that purpose, we first of all compute the frequency of 90%, 95% and 99% VaR violations, which should be close to 0.1, 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 $\mathrm{I}_{\hat{f},t+1}(y_{t+1} \leq \mathrm{VaR}_{\hat{f},t}(\alpha))$ for $\alpha = 0.1, 0.05$ and 0.01, which take the value 1 if the condition in brackets is satisfied and 0 otherwise. Independence of the VaR exceedances is tested against a first-order Markov alternative, that is, the null hypothesis is given by $H_0 : \mathsf{E}(\mathrm{I}_{\hat{f},t+1}|\mathrm{I}_{\hat{f},t}) = \mathsf{E}(\mathrm{I}_{\hat{f},t+1})$. 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 CCC test statistic simply being the sum of the CUC and IND LR statistics. For evaluating the adequacy of the ES estimates we employ the test

suggested by McNeil and Frey (2000). For every return y_{t+1} that falls below the $\widehat{\mathrm{VaR}}_{\hat{f},t}(\alpha)$ estimate, define the standardized 'residual' $e_{t+1} = (y_{t+1} - \widehat{\mathrm{ES}}_{\hat{f},t}(\alpha))/h_{t+1}$, where h_{t+1} is the conditional volatility forecast obtained from the corresponding GARCH model. If the ES predictions are correct, the expected value of e_{t+1} is equal to zero, which can be assessed by means of a two-sided t-test with HAC variance estimator.

Table 3 about here

The results reported in Table 3 show that the empirical VaR exceedance probabilities are very close to the nominal levels for the Student-t innovation distribution. For the Laplace distribution, they are considerably lower for $\alpha = 0.05$ and $\alpha = 0.01$. This is confirmed by the CUC test, which for these quantiles 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 90% and 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 conditional and censored likelihood scoring rules. 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 'over-hedging'.

6 Conclusions

In this paper we have developed new scoring rules based on conditional and censored likelihood for evaluating the predictive ability of competing density forecasts. It was shown that these scoring rules are 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 weighting the (KLIC-based) logarithmic scoring rule is not suitable for this purpose. By construction this tends 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 do not suffer from this problem.

We argued that likelihood-based scoring rules can be extended for comparing density forecasts on a specific region of interest by using the conditional likelihood, given that the actual observation lies in the region of interest, or the censored likelihood, with censoring of the observations outside the region of interest. Furthermore, we showed that the conditional and censored likelihood scoring rules can be extended in order to emphasize certain parts of the outcome space more generally by using smooth weight functions. Both scoring rules can be interpreted in terms of Kullback-Leibler divergences between weighted versions of the density forecast and the true 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. 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 often leads to more powerful tests. This is due to the fact that more information is used by the censored likelihood scores. Additionally, the censored likelihood scoring rule outperforms the weighted probability score function of Gneiting and Ranjan (2008).

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 weighted logarithmic scoring rule and the newly proposed scoring rules can lead to the selection of different density forecasts. The density forecasts preferred by the conditional and censored likelihood scoring rules appear to be more appropriate as they result in more accurate estimates of VaR and ES.

A Appendix

This Appendix provides a proof of Lemma 1.

Generalized conditional likelihood score It is to be shown that $E_t(d^{cl}_{t+1}(p_t, \hat{f}_t)) \geq 0$, where $d^{cl}_{t+1}(p_t, \hat{f}_t) = S^{cl}(p_t; Y_{t+1}) - S^{cl}(\hat{f}_t, Y_{t+1})$. Define $P_t \equiv \int w_t(s)p_t(s)\,\mathrm{d}s$ and $\hat{F}_t \equiv \int w_t(s)\hat{f}_t(s)\,\mathrm{d}s$.

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

$$\begin{aligned}
\mathsf{E}_t\left(d^{cl}_{t+1}(p_t, \hat{f}_t)\right) &= \int p_t(y)\left(w_t(y)\log\left(\frac{p_t(y)}{P_t}\right)\right)dy \\
&\quad - \int p_t(y)\left(w_t(y)\log\left(\frac{\hat{f}_t(y)}{\hat{F}_t}\right)\right)dy \\
&= P_t\int \frac{w_t(y)p_t(y)}{P_t}\log\left(\frac{w_t(y)p_t(y)/P_t}{w_t(y)\hat{f}_t(y)/\hat{F}_t}\right)\mathrm{d}y \\
&= P_t \cdot K\left(\frac{w_t(y)p_t(y)}{P_t}, \frac{w_t(y)\hat{f}_t(y)}{\hat{F}_t}\right) \geq 0,
\end{aligned}$$

where $K(\cdot,\cdot)$ represents the Kullback-Leibler divergence between the pdfs in its arguments, which is finite as a consequence of Assumption 1.

Assumption 1 implies the existence of the Radon-Nikodym derivative of the density forecasts with respect to the true predictive density p_t, i.e. $0\hat{f}_t(y)/p_t(y) < \infty$ and $0\hat{g}_t(y)/p_t(y) < \infty$, which in turn implies support$(\hat{f}_t) =$ support$(\hat{g}_t) =$ support(p_t). This, together with Assumption 2 (c) guarantees that $w_t(y)p_t(y)/P_t$ and $w_t(y)\hat{f}_t(y)/\hat{F}_t$ can be interpreted as pdfs, while Assumption 2 (a) ensures that $w_t(y)$ can be treated as a given function of y in the calculation of the expectation, which is conditional on $\mathcal{F}_t$.

□

Generalized censored likelihood score If $d_{t+1}^{csl}(p_t, \hat{f}_t) = S^{csl}(p_t; Y_{t+1}) - S^{csl}(\hat{f}_t, Y_{t+1})$, then

$$
\begin{aligned}
\mathsf{E}_t\left(d_{t+1}^{csl}(p_t, \hat{f}_t)\right) &= \int p_t(y) \log\left((p_t(y))^{w_t(y)} (1-P_t)^{1-w_t(y)}\right) \mathrm{d}y \\
&\quad - \int p_t(y) \log\left(\left(\hat{f}_t(y)\right)^{w_t(y)} (1-\hat{F}_t)^{1-w_t(y)}\right) \mathrm{d}y \\
&= \int p_t(y) \log\left(\left(\frac{p_t(y)}{\hat{f}_t(y)}\right)^{w_t(y)} \left(\frac{1-P_t}{1-\hat{F}_t}\right)^{1-w_t(y)}\right) \mathrm{d}y \\
&= \int p_t(y) \log\left(\left(\frac{p_t(y)\hat{F}_t}{P_t\hat{f}_t(y)}\right)^{w_t(y)} \left(\frac{P_t}{\hat{F}_t}\right)^{w_t(y)} \left(\frac{1-P_t}{1-\hat{F}_t}\right)^{1-w_t(y)}\right) \mathrm{d}y \\
&= \int p_t(y) \left(w_t(y) \log\left(\frac{p_t(y)\hat{F}_t}{P_t\hat{f}_t(y)}\right) + w_t(y) \log\frac{P_t}{\hat{F}_t} + (1-w_t(y)) \log\left(\frac{1-P_t}{1-\hat{F}_t}\right)\right) \mathrm{d}y \\
&= P_t \int \frac{p_t(y)w_t(y)}{P_t} \log\left(\frac{w_t(y)p_t(y)/P_t}{w_t(y)\hat{f}_t(y)/\hat{F}_t}\right) \mathrm{d}y \\
&\quad + P_t \log\frac{P_t}{\hat{F}_t} + (1-P_t) \log\left(\frac{1-P_t}{1-\hat{F}_t}\right) \\
&= P_t \cdot K\left(\frac{w_t(y)p_t(y)}{P_t}, \frac{w_t(y)\hat{f}_t(y)}{\hat{F}_t}\right) + K\left(\mathrm{Bin}(1, P_t), \mathrm{Bin}(1, \hat{F}_t)\right) \geq 0,
\end{aligned}
$$

where $K\left(\mathrm{Bin}(1, P_t), \mathrm{Bin}(1, \hat{F}_t)\right)$ is the Kullback-Leibler divergence between two Bernoulli distributions with succes probabilities P_t and $\hat{F}_t$, respectively. Assumption 2 (b), which requires $w_t(y)$ to be scaled between 0 and 1 for the *csl* rule, is essential for this interpretation because it implies that P_t and $\hat{F}_t$ can be interpreted as probabilities.

Again, Assumptions1 and 2 (c) guarantee that $w_t(y)p_t(y)/P_t$ and $w_t(y)\hat{f}_t(y)/\hat{F}_t$ can be interpreted as pdfs, while Assumption 2 (a) ensures that $w_t(y)$ can be treated as a given function of y in the calculation of the expectation, which is conditional on $\mathcal{F}_t$.

□

Acknowledgments

We would like to thank the associate editor and two anonymous referees, 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 Pennsylvania, 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. Valentyn Panchenko acknowledges the support under Australian Research Council's Discovery Projects funding scheme (project number DP0986718). Usual caveats apply.

References

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

Bai, J. (2003). Testing parametric conditional distributions of dynamic models. Review of Economics and Statistics 85, 531–549.

Bai, J. and S. Ng, (2005). Tests for skewness, kurtosis, and normality of time series data. Journal of Business and Economic Statistics 23, 49–60.

Baillie, R.T. and T. Bollerslev, (1989). The message in daily exchange rates: A conditional-variance tale. Journal of Business and Economic Statistics 7, 297–305.

Bao, Y., T.-H. Lee and B. Saltoğlu, (2004). A test for density forecast comparison with applications to risk management. Working paper 04-08, UC Riverside.

Bao, Y., T.-H. Lee and B. Saltoğlu, (2007). Comparing density forecast models. Journal of Forecasting 26, 203–225.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. Journal of Business and Economic Statistics 19, 465–474.

Bollerslev, T., U. Kretschmer, C. Pigorsch and G. Tauchen, (2009). A discrete-time model for daily S&P500 returns and realized variations: Jumps and leverage effects. Journal of Econometrics 150, 151–166.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. Review of Economics and Statistics 69, 542–547.

Campbell, S.D. and F.X. Diebold, (2005). Weather forecasting for weather derivatives. Journal of the American Statistical Association 100, 6–16.

Christoffersen, P.F. (1998). Evaluating interval forecasts. International Economic Review 39, 841–862.

Clements, M.P. (2004). Evaluating the Bank of England density forecasts of inflation. Economic Journal 114, 844–866.

Clements, M.P. (2005). Evaluating Econometric Forecasts of Economic and Financial Variables. Palgrave-Macmillan, New York.

Clements, M.P. and J. Smith, (2000). Evaluating the forecast densities of linear and nonlinear models: Applications to output growth and inflation. Journal of Forecasting 19, 255–276.

Corradi, V., W. Distaso and N.R. Swanson, (2009). Predictive density estimators for daily volatility based on the use of realized measures. Journal of Econometrics 150, 119–138.

Corradi, V. and N.R. Swanson, (2005). A test for comparing multiple misspecified conditional interval models. Econometric Theory 21, 991–1016.

Corradi, V. and N.R. Swanson, (2006a). Bootstrap conditional distribution tests in the presence of dynamic misspecifation. Journal of Econometrics 133, 779–806.

Corradi, V. and N.R. Swanson, (2006b). Predictive density and conditional confidence interval accuracy tests. Journal of Econometrics 135, 187–228.

Corradi, V. and N.R. Swanson, (2006c). Predictive density evaluation, in: G. Elliott, C.W.J. Granger and A. Timmermann (Eds.), Handbook of Economic Forecasting, Vol. 1. Elsevier, Amsterdam, pp. 197–284.

De Jong, R.M. (1997). Central limit theorems for dependent heterogeneous random variables. Econometric Theory 13, 353–367.

Diebold, F.X., T.A. Gunther and A.S. Tay, (1998). Evaluating density forecasts with applications to financial risk management. International Economic Review 39, 863–883.

Diebold, F.X. and Lopez, J.A. (1996). Forecast evaluation and combination. in: G.S. Maddala and C.R. Rao (Eds.), Handbook of Statistics, Vol. 14. North-Holland, Amsterdam, pp. 241–268.

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

Diebold, F.X., A.S. Tay and K.F. Wallis, (1999). Evaluating density forecasts of inflation: The survey of professional forecasters, in: R.F. Engle and H. White, (Eds.), Cointegration, Causality, and Forecasting: A Festschrift in Honor of C.W.J. Granger. Oxford University Press, Oxford, pp. 76–90.

Dowd, K. (2005). Measuring Market Risk, second edition, John Wiley & Sons, Chichester.

Egorov, A.V., Y. Hong and H. Li, (2006). Validating forecasts of the joint probability density of bond yields: Can affine models beat random walk? Journal of Econometrics 135, 255–284.

Franses, P.H. and D. van Dijk, (2003). Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. Oxford Bulletin of Economics and Statistics 65, 727–744.

Garratt, A., K. Lee, M.H. Pesaran and Y. Shin, (2003). Forecast uncertainties in macroeconomic modelling: An application to the UK economy. Journal of the American Statistical Association 98, 829–838.

Giacomini, R. and I. Komunjer, (2005). Evaluation and combination of conditional quantile forecasts. Journal of Business and Economic Statistics 23, 416–431.

Giacomini, R. and H. White, (2006). Tests of conditional predictive ability. Econometrica 74, 1545–1578.

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

Gneiting, T. and R. Ranjan, (2008). Comparing density forecasts using threshold and quantile weighted scoring rules. Technical Report 533. University of Washington.

Granger, C.W.J. and Z. Ding, (1995). Some properties of absolute return, an alternative measure of risk. Annales d'Economie et de Statistique 40, 67–91.

Granger, C.W.J. and M.H. Pesaran, (2000). Economic and statistical measures of forecast accuracy. Journal of Forecasting 19, 537–560.

Guidolin, M. and A. Timmermann, (2006). Term structure of risk under alternative econometric specifications. Journal of Econometrics 131, 285–308.

Guidolin, M. and A. Timmermann, (2007). Asset allocation under multivariate regime switching. Journal of Economic Dynamics and Control 31, 3503–3544.

Hall, S.G. and J. Mitchell, (2007). Combining density forecasts. International Journal of Forecasting 23, 1–13.

Härdle, W. and Z. Hlávka, (2009). Dynamics of state price densities. Journal of Econometrics 150, 1–15.

Hong, Y. and H. Li, (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. Review of Financial Studies 18, 37–84.

Hong, Y., H. Li and F. Zhao, (2004). Out-of-sample performance of discrete-time spot interest rate models. Journal of Business and Economic Statistics 22, 457–473.

Kendall, M.G. and A. Stuart, (1969). The Advanced Theory of Statistics, Vol. 1, Third edn. Griffin, London

Lahiri, K. and J.G. Wang, (2007). Evaluating probability forecasts for GDP declines. Working paper, University of Albany - SUNY.

Li, F. and G. Tkacz, (2006). A consistent bootstrap test for conditional density functions with time-series data. Journal of Econometrics 133, 863–886.

McNeil, A.J. and R. Frey, (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of Empirical Finance 7, 271–300.

McNeil, A.J., R. Frey and P. Embrechts, (2005). Quantitative Risk Management: Concepts, Techniques, and Tools. Princeton University Press, Princeton

Merlevède, F. and M. Peligrad, (2000). The functional central limit theorem under the strong mixing condition. Annals of Probability 28, 1336–1352.

Mitchell, J. and S.G. Hall, (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation. Oxford Bulletin of Economics and Statistics 67, 995–1033.

Mittnik, S., M.S. Paolella and S.T. Rachev, (1998). Unconditional and conditional distributional models for the Nikkei index. Asia-Pacific Financial Markets 5, 99–128.

Oberhofer, W. and H. Haupt, (2005). The asymptotic distribution of the unconditional quantile estimator under dependence. Statistics & Probability Letters 73, 243 – 250.

Perez-Quiros, G. and A. Timmermann, (2001). Business cycle asymmetries in stock returns: Evidence from higher order moments and conditional densities. Journal of Econometrics 103, 259–306.

Rapach, D.E. and M.E. Wohar, (2006). The out-of-sample forecasting performance of nonlinear models of real exchange rate behavior. International Journal of Forecasting 22, 341–361.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. Annals of Mathematical Statistics 23, 470–472.

Sarno, L. and G. Valente, (2004). Comparing the accuracy of density forecasts from competing models. Journal of Forecasting 23, 541–557.

Sarno, L. and G. Valente, (2005). Empirical exchange rate models and currency risk: Some evidence from density forecasts. Journal of International Money and Finance 24, 363–385.

Taylor, J.W. and R. Buizza, (2006). Density forecasting for weather derivative pricing. International Journal of Forecasting 22, 29–42.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica 26, 24–36.

West, K.D. (1996). Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.

White, H. (2000). A reality check for data snooping. Econometrica 68, 1097–1126.

Winkler, R.L. and A.H. Murphy, (1968). Good probability assessors. Journal of Applied Meteorology 7, 751–758.

Wooldridge, J.M. and H. White, (1988). Some invariance principles and central limit theorems for dependent heterogeneous processes. Econometric Theory 4, 210–230.

Table 1: Tests of equal predictive accuracy under parameter estimation uncertainty

	m	100	250	500	1000	2500	5000
$H_a:$	$E(d^l_{t+1}) > 0$	0.000	0.000	0.024	0.134	0.339	0.463
$H_a:$	$E(d^l_{t+1}) < 0$	0.982	0.239	0.026	0.004	0.001	0.000

Note: The table presents one-sided rejection rates (at nominal significance level 5%) of the null hypothesis of equal predictive accuracy against the indicated alternative by the Diebold-Mariano type test statistic defined in (1) when using the logarithmic scoring rule (2), based on the sample average of the score difference $d^l_{t+1} = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$. The DGP is an AR(2) process: $y_t = 0.8y_{t-1} + 0.05y_{t-2} + \varepsilon_t$,with $\varepsilon_t \sim$ i.i.d. N$(0, 1)$. The competing density forecasts $\hat{f}_t$ and $\hat{g}_t$ are based on AR(2) and AR(1) specifications, respectively. The residuals in both specifications are assumed to be normally distributed and the coefficients are estimated using a moving window of m observations. The estimation window size is varied from $m = 100$ to $5,000$. The number of out-of-sample evaluations is $n = 5,000$ and the number of replications is $10,000$.

Table 2: Average score differences and tests of equal predictive accuracy

Scoring rule	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	$\overline{d}^*$	Test stat.	$\overline{d}^*$	Test stat.	$\overline{d}^*$	Test stat.
Threshold weight function						
wl	-1.69×10^{-4}	-0.14	-5.12×10^{-3}	-4.74	-3.21×10^{-3}	-3.75
wps	4.29×10^{-7}	0.69	7.75×10^{-7}	1.56	8.68×10^{-7}	4.28
cl	1.47×10^{-3}	1.48	1.58×10^{-3}	2.32	7.78×10^{-4}	1.81
csl	2.21×10^{-3}	1.89	1.63×10^{-3}	1.53	1.16×10^{-3}	1.35

Note: The table presents the average score difference $\overline{d}^*$ and the corresponding test statistics for the weighted logarithmic (*wl*) scoring rule in (4), the weighted probability score (*wps*) in (8), the conditional likelihood (*cl*) in (11), and the censored likelihood (*csl*) in (12). All scoring rules are based on the indicator weight function $w_t(y) = \mathrm{I}(y \leq \hat{r}_t^{\alpha})$, where $\hat{r}_t^{\alpha}$ is the α-th quantile of the empirical (in-sample) CDF, where $\alpha = 0.1, 0.05$ or 0.01. 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 3: VaR and ES characteristics

	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	$t(\nu)$	Laplace	$t(\nu)$	Laplace	$t(\nu)$	Laplace
Average VaR	-0.0110	-0.0112	-0.0149	-0.0162	-0.0243	-0.0279
Coverage ($y_t \leq \mathrm{VaR}_t$)	0.1056	0.1001	0.0530	0.0405	0.0104	0.0055
CUC (p-value)	0.1876	0.9814	0.3324	0.0012	0.7961	0.0004
IND (p-value)	0.1082	0.2315	0.0465	0.3658	0.5809	0.5788
CCC (p-value)	0.1156	0.4887	0.0861	0.0036	0.8304	0.0015
Average ES	-0.0168	-0.0185	-0.0209	-0.0235	-0.0312	-0.0351
McNeil-Frey (test stat.)	-0.7538	3.1164	-0.8504	0.3639	-1.1899	-2.3174
McNeil-Frey (p-value)	0.4510	0.0018	0.3951	0.7159	0.2341	0.0205

Note: The average VaRs reported are the observed average 1%, 5% and 10% 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 α. 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 ESs (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 $\mathrm{ES}_{\hat{f},t}(\alpha)$.

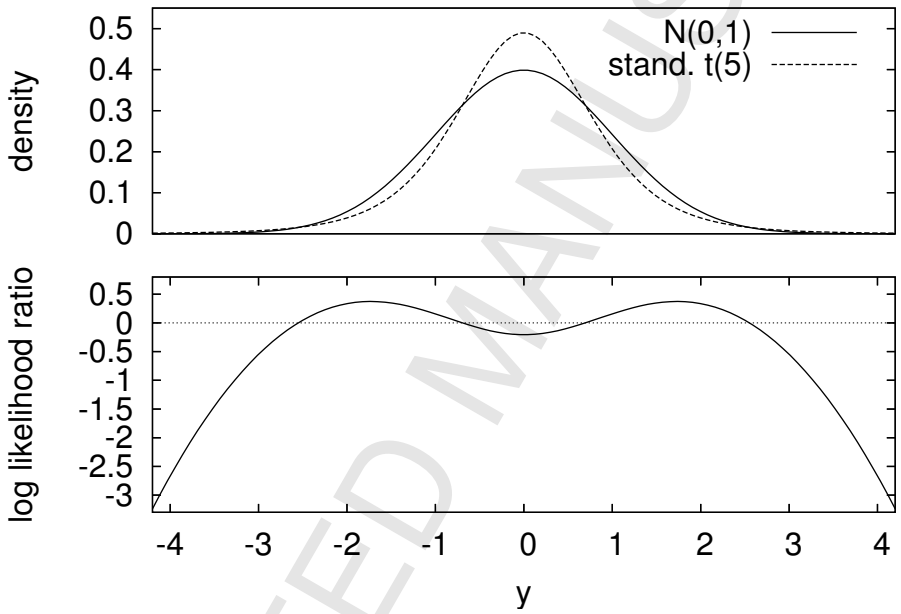

Figure 1: Probability density functions of the standard normal distribution $\hat{f}_t(y_{t+1})$ and standardized Student-$t(5)$ distribution $\hat{g}_t(y_{t+1})$ (upper panel) and corresponding relative log-likelihood scores $\log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$ (lower panel).

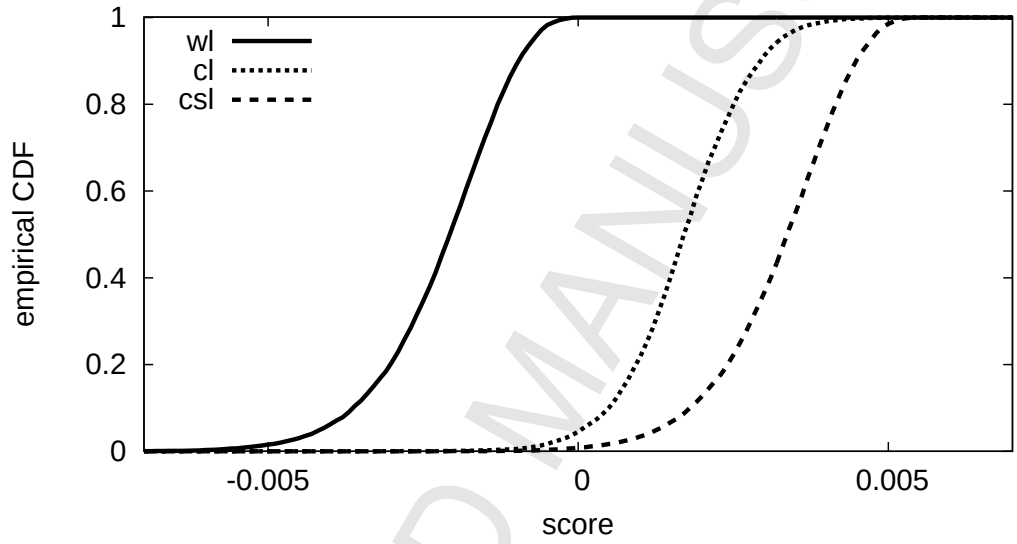

Figure 2: Empirical CDFs of mean relative scores $\overline{d}^*_{m,n}$ for the weighted logarithmic (*wl*) scoring rule in (4), the conditional likelihood (*cl*) in (11), and the censored likelihood (*csl*) in (12) for series of $n = 2,000$ independent observations from a standard normal distribution. The scoring rules are based on the threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$ 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.

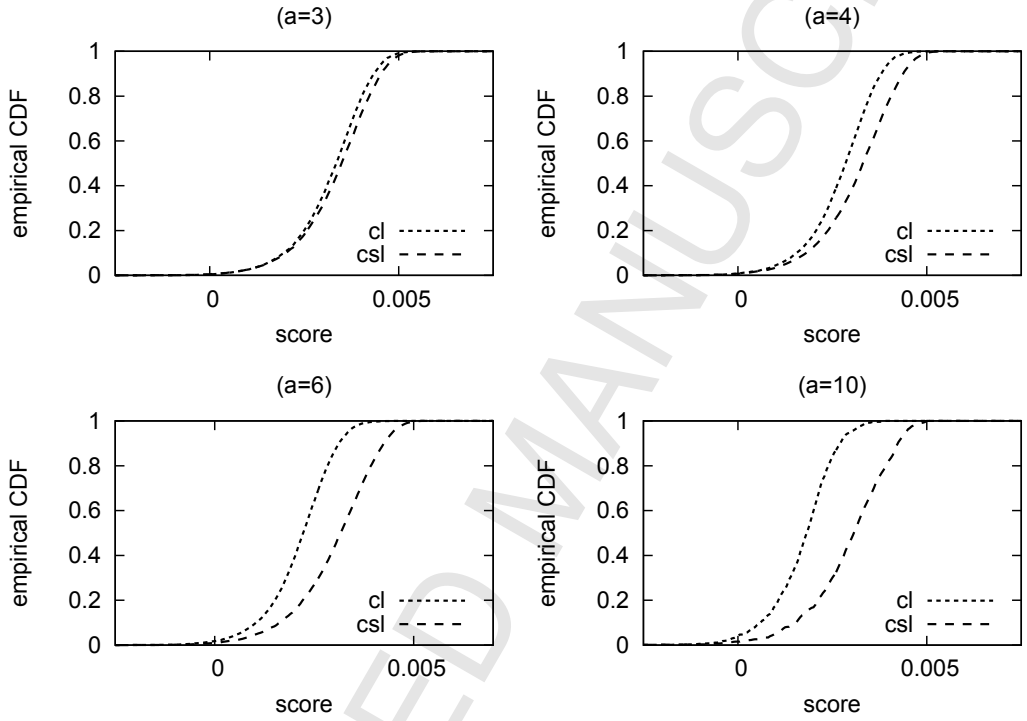

Figure 3: Empirical CDFs of mean relative scores $\overline{d}^*_{m,n}$ for the generalized conditional likelihood (*cl*) and censored likelihood (*csl*) scoring rules for series of $n = 2,000$ independent observations from a standard normal distribution. The scoring rules are based on the logistic weight function $w_t(y)$ defined in (13) 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.

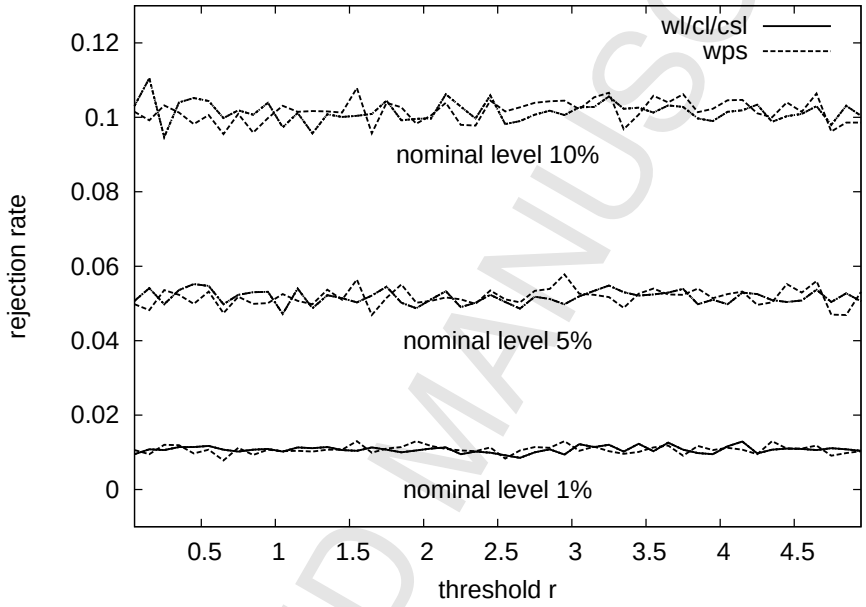

Figure 4: One-sided rejection rates of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (1) when using the weighted logarithmic (*wl*), the conditional likelihood (*cl*), and the censored likelihood (*csl*) scoring rules, under the weight function $w_t(y) = \mathrm{I}(-r \leq y \leq r)$ 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.

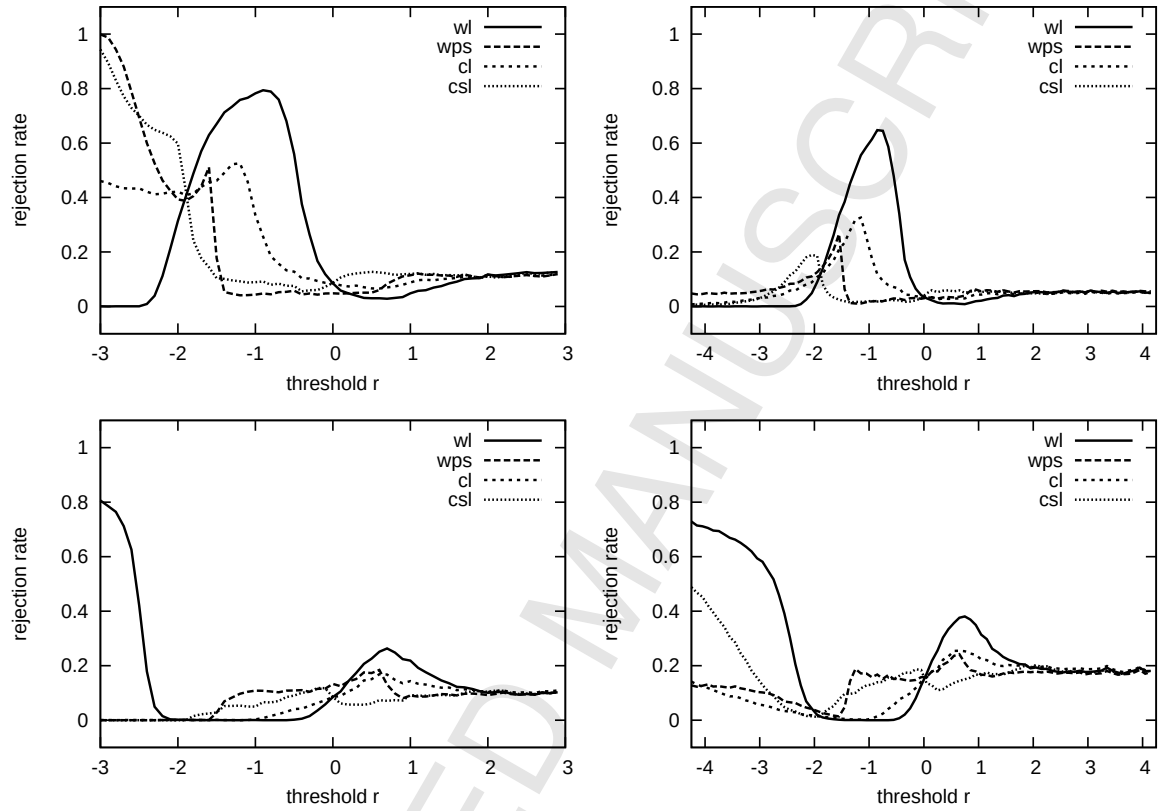

Figure 5: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (1) when using the weighted logarithmic (*wl*), the conditional likelihood (*cl*), and the censored likelihood (*csl*) scoring rules, under the threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$ for $c = 5$ expected observations in the region of interest, based on 10,000 replications. For the graphs in the left and right columns, the DGP is i.i.d. standard normal and i.i.d. standardized Student-$t(5)$, respectively. The test compares the predictive accuracy of the standard normal and the standardized Student-$t(5)$ distributions. The graphs in the top (bottom) panels show rejection rates against superior predictive ability of the standard normal (standardized Student-$t(5)$) distribution, as a function of the threshold parameter r.

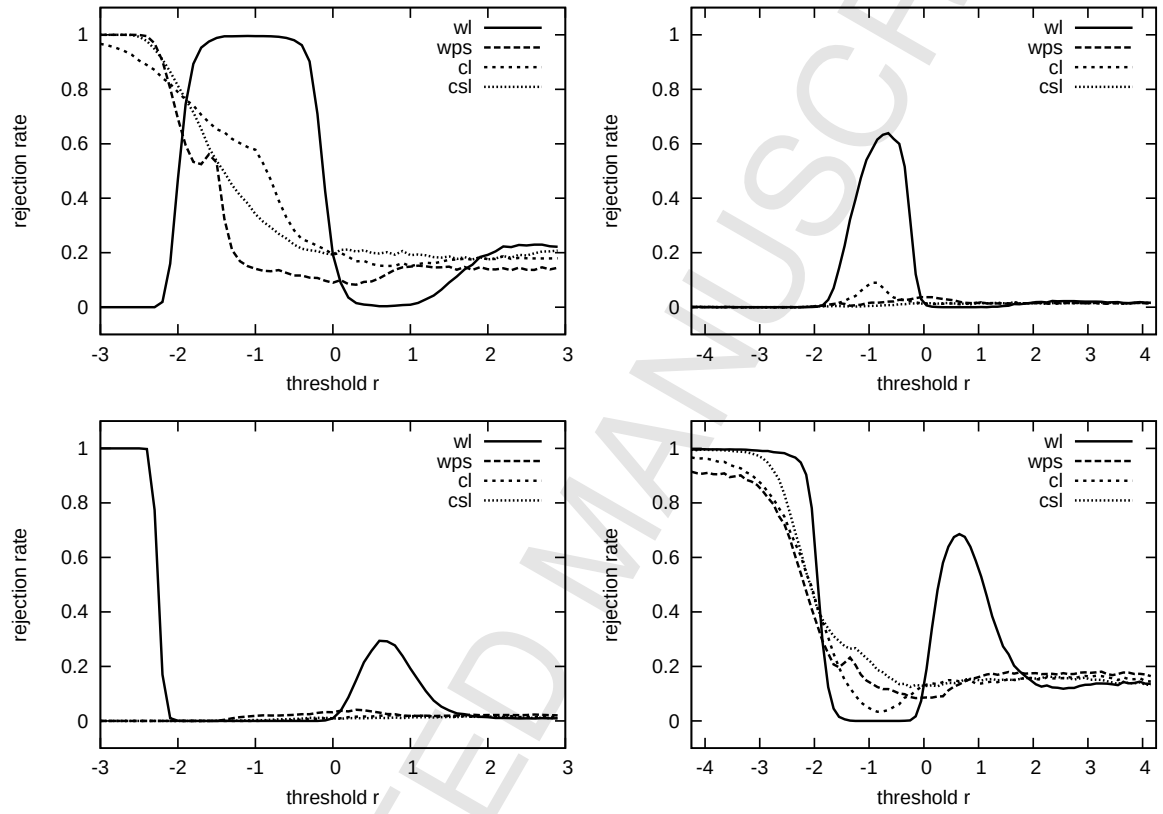

Figure 6: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (1) when using the weighted logarithmic (*wl*), the conditional likelihood (*cl*), and the censored likelihood (*csl*) scoring rules, under the threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$ for $c = 40$ expected observations in the region of interest, based on 10,000 replications. For the graphs in the left and right columns, the DGP is i.i.d. standard normal and i.i.d. standardized Student-$t(5)$, respectively. Further details are identical to those given in Figure 5.

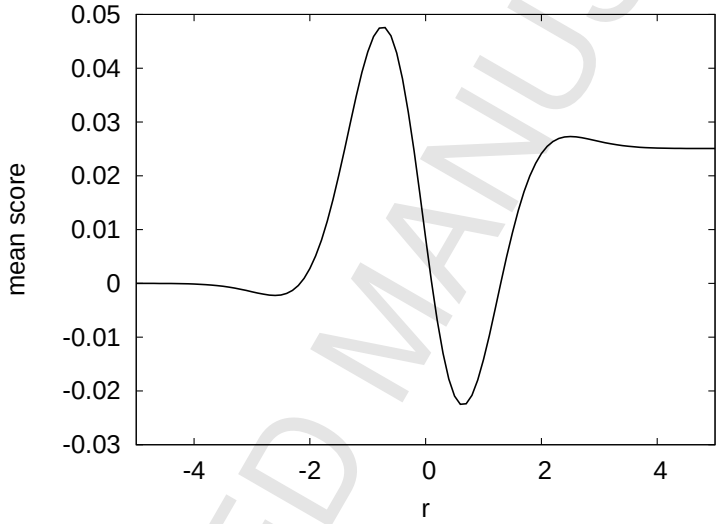

Figure 7: Mean relative *wl* score $E[d_{t+1}^{wl}]$ with threshold weight function $w_t(y) = \mathrm{I}(y \leq r)$ for the standard normal versus the standardized Student-$t(5)$ density as a function of the threshold value r, for the standard normal DGP.

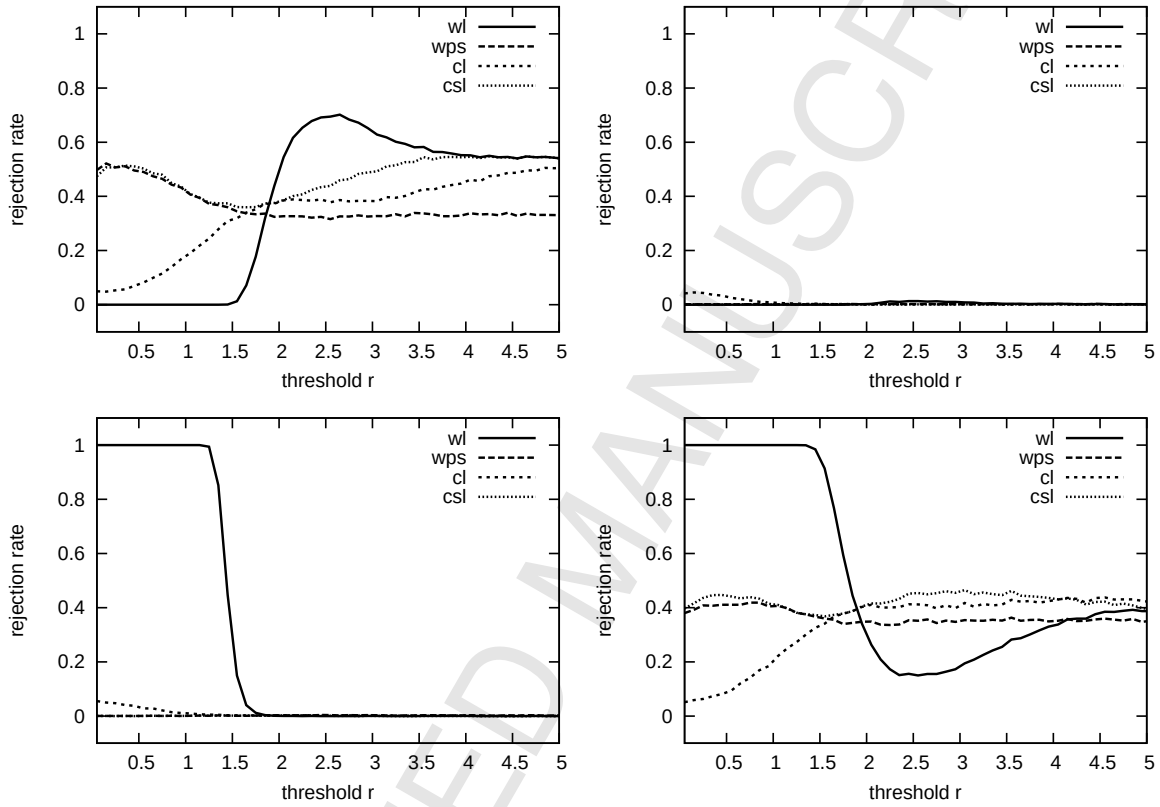

Figure 8: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (1) when using the weighted logarithmic (*wl*), the conditional likelihood (*cl*), and the censored likelihood (*csl*) scoring rules, under the weight function $w_t(y) = \mathrm{I}(-r \leq y \leq r)$ for $c = 200$ expected observations in the region of interest, 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 Student-$t(5)$ distribution and *vice versa*, respectively.

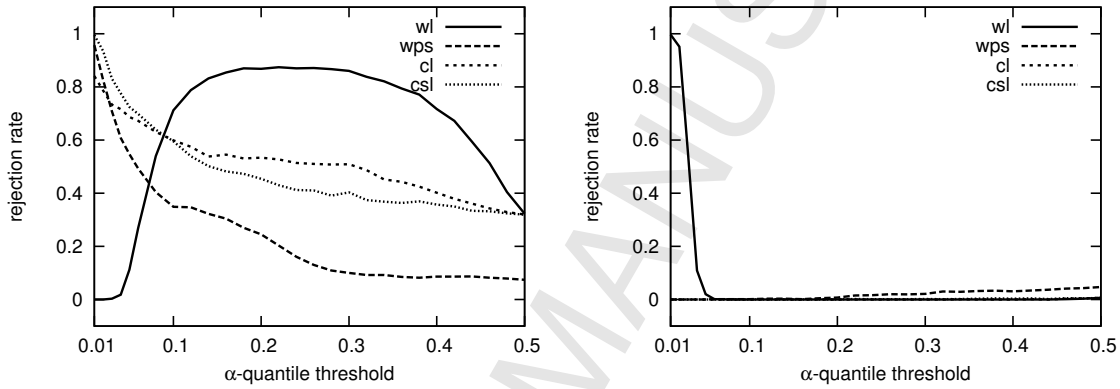

Figure 9: One-sided rejection rates (at nominal significance level 5%) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in (1) when using the weighted logarithmic (*wl*), the weighted probability *wps*, the conditional likelihood (*cl*), and the censored likelihood (*csl*) scoring rules for $c = 40$ expected observations in the region of interest, based on 1,000 replications. The data follow an GARCH(1)-process with the standard normal innovations. The competing model is based on the standardized Student-$t(5)$ innovations. The graphs on the left and right show rejection rates against better predictive ability of the forecast based on the standard normal innovations compared to the forecast based on standardized Student-$t(5)$ innovations and *vice versa*, respectively.

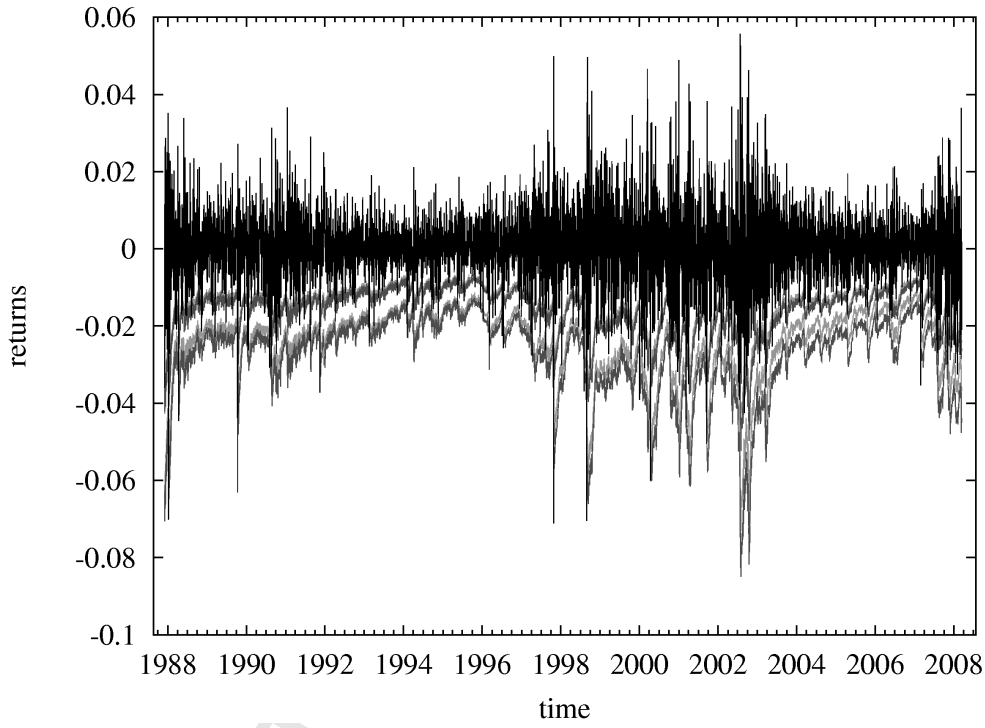

Figure 10: Daily S&P 500 log-returns (black) for the period December 2, 1987 – March 14, 2008 and out-of-sample 95% and 99% VaR forecasts derived from the AR(5)-GARCH(1,1) specification using Student-t innovations (light gray) and Laplace innovations (dark gray).