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

Cees Diks†
CeNDEF, Amsterdam School of Economics
University of Amsterdam

Valentyn Panchenko‡
School of Economics
University of New South Wales

Dick van Dijk§
Econometric Institute
Erasmus University Rotterdam

November 29, 2008

Abstract

We propose new scoring rules based on partial likelihood for assessing the relative out-of-sample predictive accuracy of competing density forecasts over a specific region of interest, such as the left tail in financial risk management. By construction, existing scoring rules based on weighted likelihood or censored normal likelihood favor density forecasts with more probability mass in the given region, rendering predictive accuracy tests biased towards such densities. Our novel partial likelihood-based scoring rules do not suffer from this problem, as illustrated by means of Monte Carlo simulations and an empirical application to daily S&P 500 index returns.

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

JEL Classification: C12; C22; C52; C53

---

*We would like to thank 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) as well as seminar participants at Monash University, Queensland University of Technology, the University of Amsterdam, the University of New South Wales, and the Reserve Bank of Australia for providing useful comments and suggestions.

†Corresponding author: Center for Nonlinear Dynamics in Economics and Finance, Faculty of Economics and Business, 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@few.eur.nl

1 Introduction

The interest in density forecasts is rapidly expanding in both macroeconomics and finance. Undoubtedly this is due to the increased awareness that point forecasts are not very informative unless some indication of their uncertainty is provided, see Granger and Pesaran (2000) and Garratt *et al.* (2003) for discussions of this issue. Density forecasts, representing the future probability distribution of the random variable in question, provide the most complete measure of this uncertainty. Prominent macroeconomic applications are density forecasts of output growth and inflation obtained from a variety of sources, including statistical time series models (Clements and Smith, 2000), professional forecasters (Diebold *et al.*, 1999), and central banks and other institutions producing so-called 'fan charts' for these variables (Clements, 2004; Mitchell and Hall, 2005). In finance, density forecasts play a fundamental role in risk management as they form the basis for risk measures such as Value-at-Risk and Expected Shortfall, see Dowd (2005) and McNeil *et al.* (2005) for general overviews and Guidolin and Timmermann (2006) for a recent empirical application. In addition, density forecasts are starting to be used in other financial decision problems, such as derivative pricing (Campbell and Diebold, 2005; Taylor and Buizza, 2006) 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).

The increasing popularity of density forecasts has naturally led to the development of statistical tools for evaluating their accuracy. The techniques that have been proposed for this purpose can be classified into two groups. First, several approaches have been put forward for testing the quality of an individual density forecast, relative to the data-generating process. Following the seminal contribution of Diebold *et al.* (1998), the most prominent tests in this group are based on the probability integral transform (PIT) of Rosenblatt (1952). Under the null hypothesis that the density forecast is correctly specified, the PITs should be uniformly distributed, while for one-step ahead density forecasts

they also should be independent and identically distributed. Hence, Diebold *et al.* (1998) consider a Kolmogorov-Smirnov test for departure from uniformity of the empirical PITs and several tests for temporal dependence. Alternative test statistics based on the PITs are developed in Berkowitz (2001), Bai (2003), Bai and Ng (2005), Hong and Li (2005), Li and Tkacz (2006), and Corradi and Swanson (2006a), mainly to counter the problems caused by parameter uncertainty and the assumption of correct dynamic specification under the null hypothesis. We refer to Clements (2005) and Corradi and Swanson (2006c) for in-depth surveys on specification tests for univariate density forecasts. An extension of the PIT-based approach to the multivariate case is considered by Diebold *et al.* (1999), see also Clements and Smith (2002) for an application. For more details of multivariate PITs and goodness-of-fit tests based on these, see Breymann *et al.* (2003) and Berg and Bakken (2005), among others.

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

In many applications of density forecasts, we are mostly interested in a particular region of the density. Financial risk management is an example in case, where the main concern is obtaining an accurate description of the left tail of the distribution. Bao *et*

al. (2004) and Amisano and Giacomini (2007) suggest weighted likelihood ratio (LR) tests based on KLIC-type scoring rules, which may be used for evaluating and comparing density forecasts in a particular region. However, as mentioned by Corradi and Swanson (2006c) measuring the accuracy of density forecasts over a specific region cannot be done in a straightforward manner using the KLIC. The problem that occurs with KLIC-based scoring rules is that they favor density forecasts which have more probability mass in the region of interest, rendering the resulting tests biased towards such density forecasts.

In this paper we demonstrate that two density forecasts can be compared on a specific region of interest in a natural way by using scoring rules based on partial likelihood (Cox, 1975). We specifically introduce two different scoring rules based on partial likelihood. The first rule considers the value of the conditional likelihood, given that the actual observation lies in the region of interest. The second rule is based on the censored likelihood, where again the region of interest is used to determine if an observation is censored or not. We argue that these partial likelihood scoring rules behave better than KLIC-type rules, in that they always favor a correctly specified density forecast over an incorrect one. This is confirmed by our Monte Carlo simulations. Moreover, we find that the scoring rule based on the censored likelihood, which uses more of the relevant information present, performs better in all cases considered.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss conventional scoring rules based on the KLIC distance for evaluating density forecasts and point out the problem with the weighted versions of the resulting LR tests when these are used to focus on a particular region of the density. In Sections 3 and 4, we develop alternative scoring rules based on partial likelihood, and demonstrate that these do not suffer from this shortcoming. This is further illustrated by means of Monte Carlo simulation experiments in Section 5, where we assess the properties of tests of equal predictive accuracy of density forecasts with different scoring rules. We provide an empirical application concerning density forecasts for daily S&P 500 returns in Section 6, demonstrating the practical usefulness of our approach. Finally, we conclude in Section 7.

2 Scoring rules for evaluating density forecasts

Following Amisano and Giacomini (2007) we consider a stochastic process $\{Z_t : \Omega \rightarrow \mathbb{R}^{k+1}\}_{t=1}^{T}$, defined on a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and identify Z_t with $(Y_t, X_t')'$, where $Y_t : \Omega \rightarrow \mathbb{R}$ is the real valued random variable of interest and $X_t : \Omega \rightarrow R^k$ is a vector of predictors. The information set at time t is defined as $\mathcal{F}_t = \sigma(Z_1', \ldots, Z_t')'$. We consider the case where two competing forecast methods are available, each producing one-step ahead density forecasts, i.e. predictive densities of Y_{t+1}, based on $\mathcal{F}_t$. The competing density forecasts are denoted by the probability density functions (pdfs) $\hat{f}_t(Y_{t+1})$ and $\hat{g}_t(Y_{t+1})$, respectively. As in Amisano and Giacomini (2007), by 'forecast method' we mean the set of choices that the forecaster makes at the time of the prediction, including the variables X_t, the econometric model (if any), and the estimation method. The only requirement that we impose on the forecast methods is that the density forecasts depend only on the R most recent observations $Z_{t-R+1}, \ldots, Z_t$. Forecast methods of this type arise easily when model-based density forecasts are made and model parameters are estimated based on a moving window of R observations. This finite memory simplifies the asymptotic theory of tests of equal predictive accuracy considerably, see Giacomini and White (2006). To keep the exposition as simple as possible, in this paper we will be mainly concerned with the case of comparing 'fixed' predictive densities for i.i.d. processes.

Our interest lies in comparing the relative performance of $\hat{f}_t(Y_{t+1})$ and $\hat{g}_t(Y_{t+1})$, that is, assessing which of these densities comes closest to the true but unobserved density $p_t(Y_{t+1})$. One of the approaches that has been put forward for this purpose is based on scoring rules, which are commonly used in probability forecast evaluation, see Diebold and Lopez (1996). Lahiri and Wang (2007) provide an interesting application of several such rules to the evaluation of probability forecasts of gross domestic product (GDP) declines, that is, a rare event comparable to Value-at-Risk violations. In the current context, a scoring rule can be considered as a loss function depending on the density forecast and the actually observed data. The general idea is to assign a high score to a density forecast if an observation falls within a region with high probability, and a low score if it falls within a region with low probability. Given a sequence of density forecasts and corre-

sponding realizations of the time series variable Y_{t+1}, competing density forecasts may then be compared based on their average scores. Mitchell and Hall (2005), Amisano and Giacomini (2007), and Bao et al. (2004, 2007) focus on the logarithmic scoring rule

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

where y_{t+1} is the observed value of the variable of interest. Based on a sequence of P density forecasts and realizations for observations $R+1, \ldots, T \equiv R+P$, the density forecasts $\hat{f}_t$ and $\hat{g}_t$ can be ranked according to their average scores $P^{-1}\sum_{t=R}^{T-1} \log \hat{f}_t(y_{t+1})$ and $P^{-1}\sum_{t=R}^{T-1} \log \hat{g}_t(y_{t+1})$. The density forecast yielding the highest score would obviously be the preferred one. We may also test formally whether differences in average scores are statistically significant. Defining the score difference

$$\begin{aligned} d^l_{t+1} &= S^l(\hat{f}_t; y_{t+1}) - S^l(\hat{g}_t; y_{t+1}) \\ &= \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1}), \end{aligned}$$

the null hypothesis of equal scores is given by $H_0 : \mathsf{E}[d^l_{t+1}] = 0$. This may be tested by means of a Diebold and Mariano (1995) type statistic

$$\frac{\overline{d}^l}{\sqrt{\hat{\sigma}^2/P}} \xrightarrow{d} N(0,1), \tag{2}$$

where $\overline{d}^l$ is the sample average of the score difference, that is, $\overline{d}^l = P^{-1}\sum_{t=R}^{T-1} d^l_{t+1}$, and $\hat{\sigma}^2$ is a consistent estimator of the asymptotic variance of $\overline{d}^l_{t+1}$. Following Giacomini and White (2006), we focus on competing forecast methods rather than on competing models. This has the advantage that the test just described is still valid in case the density forecasts depend on estimates of unknown parameters, provided that a finite (rolling) window of past observations is used for parameter estimation.

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

$$\text{KLIC}(\hat{f}_t) = \int p_t(y_{t+1}) \log\left(\frac{p_t(y_{t+1})}{\hat{f}_t(y_{t+1})}\right) \mathrm{d}y_{t+1} = \mathsf{E}[\log p_t(Y_{t+1}) - \log \hat{f}_t(Y_{t+1})]. \tag{3}$$

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

It is useful to note that both Mitchell and Hall (2005) and Bao et al. (2004, 2007) employ the same approach for testing the null hypothesis of correct specification of an individual density forecast, that is, H_0 : KLIC($\hat{f}_t$) $= 0$. The problem that the true density p_t in (3) is unknown then is circumvented by using the result established by Berkowitz (2001) that the KLIC of $\hat{f}_t$ is equal to the KLIC of the density of the inverse normal transform of the PIT of the density forecast $\hat{f}_t$. 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 KLIC takes the same functional form before and after the inverse normal transform of $\{y_{t+1}\}$, which is essentially a consequence of the general invariance of the KLIC under invertible measurable coordinate transformations. Of course, in practice the density q_t is not known either, but if $\hat{f}_t$ is correctly specified, $\{z_{\hat{f},t+1}\}$ should behave as an i.i.d. standard normal sequence. As discussed in Bao et al. (2004, 2007), q_t may be estimated by means of a flexible density function to obtain an estimate of the KLIC, which then allows testing for departures of q_t from the standard normal. Finally, we note that the KLIC has also been used by Mitchell and Hall (2005) and Hall and Mitchell (2007) for combining density forecasts.

2.1 Weighted scoring rules

In empirical applications of density forecasting it frequently occurs that a particular region of the density is of most interest. For example, in risk management applications such as

Value-at-Risk and Expected Shortfall estimation, an accurate description of the left tail of the distribution obviously is of crucial importance. In that case, 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. MODIFICATION HERE: We wish to emphasize that the approach considered here differs from evaluation of point forecasts for conditional quantile considered e.g. in Giacomini and Komunjer (2005). END Within the framework of scoring rules, this may be achieved by introducing a weight function $w(y_{t+1})$ to obtain a *weighted* scoring rule, see Franses and van Dijk (2003) for a similar idea in the context of testing equal predictive accuracy of point forecasts. For example, Amisano and Giacomini (2007) suggest the weighted logarithmic (WL) scoring rule

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

to assess the quality of the density forecast $\hat{f}_t$, together with the weighted average scores $P^{-1} \sum_{t=R}^{T-1} w(y_{t+1}) \log \hat{f}_t(y_{t+1})$ and $P^{-1} \sum_{t=R}^{T-1} w(y_{t+1}) \log \hat{g}_t(y_{t+1})$ for ranking two competing forecasts. Using the weighted score difference

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

the null hypothesis of equal weighted scores, $H_0 : \mathsf{E}[d_{t+1}^{wl}] = 0$, may be tested by means of a Diebold-Mariano type test statistic of the form (2), but using the sample average $\overline{d}^{wl} = P^{-1} \sum_{t=R}^{T-1} d_{t+1}^{wl}$ instead of $\overline{d}^{l}$ together with an estimate of the corresponding asymptotic variance of d_{t+1}^{wl}. From the discussion above, it follows that an alternative interpretation of the resulting statistic is to say that it tests equality of the weighted KLICs of $\hat{f}_t$ and $\hat{g}_t$.

The weight function $w(y_{t+1})$ should be positive and bounded but may otherwise be chosen arbitrarily to focus on a particular density region of interest. For evaluation of the left tail in risk management applications, for example, we may use the 'threshold' weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$, where $\mathrm{I}(A) = 1$ if the event A occurs and zero otherwise, for some value r. MODIFICATION HERE: The values of r can be chosen in practice on the basis of set policies, e.g. inflation targets, target portfolio Value-at-Risk, etc. If no policy guidelines are available r can be deduced from the data using, e.g., the inverse of the empirical CDF. END However, we are then confronted with the

problem pointed out by Corradi and Swanson (2006c) that measuring the accuracy of density forecasts over a specific region cannot be done in a straightforward manner using the KLIC or log scoring rule. In this particular case the weighted logarithmic score may be biased towards fat-tailed densities. To understand why this occurs, consider the situation where $\hat{g}_t(Y_{t+1}) > \hat{f}_t(Y_{t+1})$ for all Y_{t+1} smaller than some given value y^*, say. Using $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$ with $r < y^*$ 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, there will be cases where the fat-tailed distribution $\hat{g}_t$ is favored over the thin-tailed distribution $\hat{f}_t$, even if the latter is the true distribution from which the data are drawn.

The following example illustrates the issue at hand. Suppose we wish to compare the accuracy of two density forecasts for Y_{t+1}, one being the standard normal distribution with pdf

$$\hat{f}_t(y_{t+1}) = (2\pi)^{-\frac{1}{2}} \exp(-y_{t+1}^2/2),$$

and the other being the (fat-tailed) Student-t distribution with ν degrees of freedom, standardized to unit variance, with pdf

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

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

[Figure 1 about here.]

We emphasize that the problem sketched above is not limited to the logarithmic scoring rule but occurs more generally. For example, Berkowitz (2001) advocates the use of the inverse normal transform $z_{\hat{f},t+1}$, as defined before, motivated by the fact that if $\hat{f}_t$ is correctly specified, these should be an i.i.d. standard normal sequence. Taking the standard normal log-likelihood of the transformed data leads to the following scoring rule:

$$S^N(\hat{f}_t; y_{t+1}) = \log \phi(z_{\hat{f},t+1}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}z^2_{\hat{f},t+1}. \tag{6}$$

Although for a correctly specified density forecast the sequence $\{z_{\hat{f},t+1}\}$ is i.i.d. normal, tests for the comparative accuracy of density forecasts in a particular region based on the normal log-likelihood may be biased towards incorrect alternatives. A weighted normal scoring rule of the form

$$S^{wN}(\hat{f}_t; y_{t+1}) = w(y_{t+1}) \log \phi(z_{\hat{f},t+1}) \tag{7}$$

also would not guarantee that a correctly specified density forecast is preferred over an incorrectly specified density forecast. Figure 2 illustrates this point for standard normal and standardized Student-$t(5)$ distributions, denoted as $\hat{f}_t$ and $\hat{g}_t$ as before. When focusing on the left tail by using the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$, it is seen that $|z_{\hat{f},t+1}| > |z_{\hat{g},t+1}|$ for all values of y less than $y^* \approx -2$. As for the weighted logarithmic scoring rule (4), it then follows that if the threshold $r < y^*$ the relative score $d^{wN}_{t+1} = \frac{1}{2}w(y_{t+1})(z^2_{\hat{g},t+1} - z^2_{\hat{f},t+1})$ is strictly negative for observations y_{t+1} in the left tail and zero otherwise, such that the test based on the average score difference $\overline{d}^{wN}$ is biased against the Student-t alternative.

[Figure 2 about here.]

Berkowitz (2001) proposes a different way to use the scoring rule based on the normal transform for evaluation of the accuracy of density forecasts in the left tail, based on the idea of censoring. Specifically, for a given quantile $0 < \alpha < 1$, define the censored normal log-likelihood (CNL) score function

$$\begin{aligned} S^{cN}(\hat{f}_t, y_{t+1}) &= \mathrm{I}(\widehat{F}_t(y_{t+1}) < \alpha) \log \phi(z_{\hat{f},t+1}) + \mathrm{I}(\widehat{F}_t(y_{t+1}) \geq \alpha) \log(1-\alpha) \\ &= w(z_{\hat{f},t+1}) \log \phi(z_{\hat{f},t+1}) + (1 - w(z_{\hat{f},t+1})) \log(1-\alpha), \end{aligned} \tag{8}$$

where $w(z_{\hat{f},t+1}) = \mathrm{I}(z_{\hat{f},t+1} < \Phi^{-1}(\alpha))$, which is equivalent to $\mathrm{I}(\widehat{F}_t(y_{t+1}) < \alpha)$. The CNL scoring rule evaluates the shape of the density forecast for the region below the αth quantile and the frequency with which this region is visited. The latter form of (8) can also be used to focus on other regions of interest by using a weight function $w(Y_{t+1})$ equal to one in the region of interest, together with $\alpha_{\widehat{f},t} = \mathsf{P}_{\hat{f},t}(w(Y_{t+1}) = 1)$. One can arrange for the weight functions of competing density forecasts to coincide by allowing α to depend on the predictive density at hand. Particularly, in the case of the threshold weight function the same threshold r can be used for two competing density forecasts by using two different values of α, equal to the left exceedance probability of the value r for the respective density forecast.

Although it is perfectly valid to assess the quality of an individual density forecast in an absolute sense based on (8), see Berkowitz (2001), this scoring rule should be used with caution when testing the relative accuracy of competing density forecasts, as considered by Bao *et al.* (2004). Like the WL scoring rule, the CNL scoring rule with the threshold weight function has a tendency to favor predictive densities with more probability mass in the left tail. Suppose $\hat{G}_t(y_{t+1})$ is fat-tailed on the left whereas $\hat{F}_t(y_{t+1})$ is not. For values of y_{t+1} sufficiently far in the left tail it then holds that $\hat{F}_t(y_{t+1}) < \hat{G}_t(y_{t+1}) < \frac{1}{2}$. Because of the asymptote of $\Phi^{-1}(s)$ at $s = 0$, $z_{\hat{f},t+1} = \Phi^{-1}\left(\hat{F}_t(y_{t+1})\right)$ will be much larger in absolute value than $z_{\hat{g},t+1} = \Phi^{-1}\left(\hat{G}_t(y_{t+1})\right)$, so that $\phi(z_{\hat{f},t+1}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}z^2_{\hat{f},t+1} \ll -\frac{1}{2}\log(2\pi) - \frac{1}{2}z^2_{\hat{g},t+1} = \phi(z_{\hat{g},t+1})$. That is, a much higher score is assigned to the fat-tailed predictive density $\hat{g}_t$ than to $\hat{f}_t$.

3 Scoring rules based on partial likelihood

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

date density $\hat{f}_t$ and the true density p_t, because, under the constraint $\int_{-\infty}^{\infty} \hat{f}_t(y_{t+1})\,\mathrm{d}y = 1$, the expectation

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

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

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

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

This shows that log-likelihood scores of different forecast methods can be compared in a meaningful way, provided that the densities under consideration are properly normalized to have unit total probability. The quality of a normalized density forecast $\hat{f}_t$ can therefore be quantified by the average score $\mathsf{E}_Y[\log \hat{f}_t(Y_{t+1})]$. If $p_t(y)$ is the true conditional density of Y_{t+1}, the KLIC is nonnegative and defines a divergence between the true and an approximate distribution. If the true data generating process is unknown, we can still use KLIC differences to measure the relative performance of two competing densities, which renders the logarithmic score difference discussed before, $d^l_{t+1} = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$.

The implication from the above is that likelihood-based scoring rules may still be used to assess the (relative) accuracy of density forecasts in a particular region of the distributions, as long as the scoring rules correspond to (possibly partial) likelihood functions. In the specific case of the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$ we can break down the observation of Y_{t+1} in two stages. First, it is revealed whether Y_{t+1} is smaller than the threshold value r or not. We introduce the random variable V_{t+1} to denote the outcome of this first stage experiment, defining it as

$$V_{t+1} = \begin{cases} 1 & \text{if } Y_{t+1} \leq r, \\ 0 & \text{if } Y_{t+1} > r. \end{cases} \tag{9}$$

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

However, this is not made explicit in the subsequent notation to keep the exposition as simple as possible.

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

$$(F(r))^v(1-F(r))^{1-v}\left[\frac{f(y)}{1-F(r)}\mathrm{I}(v=0)+\frac{f(y)}{F(r)}\mathrm{I}(v=1)\right].$$

By disregarding either the information revealed by V_{t+1} or $Y_{t+1}|v$ (possibly depending on the first-stage outcome V_{t+1}) we can construct various partial likelihood functions. This enables us to formulate several scoring rules that may be viewed as weighted likelihood scores. As these still can be interpreted as a true (albeit partial) likelihood, they can be used for comparing the predictive accuracy of different density forecasts. In the following, we discuss two specific scoring rules as examples.

Conditional likelihood scoring rule For a given density forecast $\hat{f}_t$, if we decide to ignore information from the first stage and use the information revealed in the second stage only if it turns out that $V_{t+1} = 1$ (that is, if Y_t is a tail event), we obtain the conditional likelihood (CL) score function

$$S^{cl}(\hat{f}_t; y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)\log(\hat{f}_t(y_{t+1})/\hat{F}_t(r)). \tag{10}$$

The main argument for using such a score function would be to evaluate density forecasts based on their behavior in the left tail (values less than or equal to r). However, due to the normalization of the total tail probability we lose information of the original density forecast on how often tail observations actually occur. This is because the information regarding this frequency is revealed only by the first-stage experiment, which we have explicitly ignored here. As a result, the conditional likelihood scoring rule attaches

comparable scores to density forecasts that have similar tail shapes, but completely different tail probabilities. This tail probability is obviously relevant for risk management purposes, in particular for Value-at-Risk evaluation. Hence, the following scoring rule takes into account the tail behavior as well as the relative frequency with which the tail is visited.

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

$$S^{csl}(\hat{f}_t; y_{t+1}) = \mathrm{I}(y_{t+1} \leq r) \log \hat{f}_t(y_{t+1}) + \mathrm{I}(y_{t+1} > r) \log(1 - \hat{F}_t(r)). \tag{11}$$

This scoring rule uses the information of the first stage (essentially information regarding the CDF $\hat{F}_t(y)$ at $y = r$) but apart from that ignores the shape of $f_t(y)$ for values above the threshold value r. In that sense this scoring rule is similar to that used in the Tobit model for normally distributed random variables that cannot be observed above a certain threshold value (see Tobin, 1958). To compare the censored likelihood score with the censored normal score of (8), consider the case $\alpha = \hat{F}_t(r)$ in the latter. The indicator fuctions, as well as the second terms, in both scoring rules then coincide exactly, while the only difference between the scoring rules is that $\log \hat{f}_t(y_{t+1})$ occurs in the CSL rule in place where the CNL rule has $\log \phi(z_{\hat{f},t+1})$. It is the latter term that gave rise to the (too) high scores in the left tail to the standardized $t(5)$-distribution in case the correct distribution is the standard normal.

We may test the null hypothesis of equal performance of two density forecasts $\hat{f}_t(y)$ and $\hat{g}_t(y)$ based on the conditional likelihood score (10) or the censored likelihood score (11) in the same manner as before. That is, given a sample of density forecasts and corresponding realizations for P time periods $R + 1, \ldots, T$, we may form the relative scores $d^{cl}_{t+1} = S^{cl}(\hat{f}_t; y_{t+1}) - S^{cl}(\hat{g}_t; y_{t+1})$ and $d^{csl}_{t+1} = S^{csl}(\hat{f}_t; y_{t+1}) - S^{csl}(\hat{g}_t; y_{t+1})$ and use these as the basis for computing a Diebold-Mariano type test statistic of the form given in (2).

We revisit the example from the previous section in order to illustrate the properties of the various scoring rules and the associated tests for comparing the accuracy of competing

density forecasts. We generate $10,000$ independent series of $P = 2000$ independent observations y_{t+1} from a standard normal distribution. For each sequence we compute the mean value of the weighted logarithmic scoring rule in (4), the censored normal likelihood in (8), the conditional likelihood in (10), and the censored likelihood in (11). For the WL scoring rule score we use the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$, where the threshold is fixed at $r = -2.5$. The CNL score is used with $\alpha = \Phi(r)$, where $\Phi(\cdot)$ represents the standard normal CDF. The threshold value $r = -2.5$ is also used for the CL and CSL scores. Each scoring rule is computed for the (correct) standard normal density $\hat{f}_t$ and the standardized Student-t density $\hat{g}_t$ with five degrees of freedom.

[Figure 3 about here.]

Figure 3 shows the empirical CDF of the mean relative scores $\overline{d}^*$, where $*$ is *wl*, *cnl*, *cl* or *csl*. The average WL and CNL scores take almost exclusively negative values, which means that for the weight function considered, on average they attach a lower score to the correct normal distribution than to the Student-t distribution, leading to a bias in the corresponding test statistic towards the incorrect, fat-tailed distribution. The two scoring rules based on partial likelihood both correctly favor the true normal density. The scores of the censored likelihood rule appear to be better at detecting the inadequacy of the Student-t distribution, in the sense that its relative scores stochastically dominate those based on the conditional likelihood.

We close this section by noting that the validity of the CL and CSL scoring rules in (10) and (11) does not depend on the particular definition of V_{t+1} (or weight function $w(Y_{t+1})$) used. The two scoring rules discussed above focus on the case where $V_{t+1} = \mathrm{I}(Y_{t+1} \leq r)$. This step function is the analogue of the threshold weight function $w(Y_{t+1})$ used in the introductory example which motivated our approach. This weight function seems an obvious choice in risk management applications, as the left tail behavior of the density forecast then is of most concern. In other empirical applications of density forecasting, however, the focus may be on a different region of the distribution, leading to alternative weight functions. For example, for monetary policymakers aiming to keep inflation within a certain range, the central part of the distribution may be of most interest, suggesting to

define V_{t+1} as $V_{t+1} = \mathrm{I}(r_l \leq Y_{t+1} \leq r_u)$ for given lower and upper bounds r_l and r_u.

4 Smooth weight functions and generalized scoring rules

The conditional and censored likelihood scoring rules in in (10) and (11) with the particular definitions of V_{t+1} and corresponding weight functions $w(Y_{t+1})$ strictly define a precise region for which the density forecasts are evaluated. This is appropriate in case it is perfectly obvious which specific region is of interest. In practice this may not be so clear-cut, and it may be desirable to define a certain region with less stringent bounds. For example, instead of the threshold weight function $w(Y_{t+1}) = \mathrm{I}(Y_{t+1} \leq r)$, we may consider a logistic weight function of the form

$$w(Y_{t+1}) = 1/(1 + \exp(a(Y_{t+1} - r))) \qquad \text{with } a > 0. \tag{12}$$

This sigmoidal function changes monotonically from 1 to 0 as Y_{t+1} increases, with $w(r) = \frac{1}{2}$ and the slope parameter a determining the speed of the transition. Note that in the limit as $a \to \infty$, the threshold weight function $\mathrm{I}(Y_{t+1} \leq r)$ is recovered. In this section we demonstrate how the partial likelihood scoring rules can be generalized to alternative weight functions, including smooth functions such as (12).

This can be achieved by not taking V_{t+1} as a deterministic function of Y_{t+1} as in (9) but instead allowing V_{t+1} to take the value 1 with probability $w(Y_{t+1})$ and 0 otherwise, that is,

$$V_{t+1}|Y_{t+1} = \begin{cases} 1 & \text{with probability } w(Y_{t+1}), \\ 0 & \text{with probability } 1 - w(Y_{t+1}), \end{cases} \tag{13}$$

so that V_{t+1} given Y_{t+1} is a Bernoulli random variable with success probability $w(Y_{t+1})$. In this more general setting the two-stage experiment can still be thought of as observing only V_{t+1} in the first stage and Y_{t+1} in the second stage. The (partial) likelihoods based on the first and/or second stage information then allow the construction of more general versions of the CL and CSL scoring rules, involving the weight function $w(Y_{t+1})$. To see this, recall that the CL and CSL scoring rules based on the threshold weight function either include the likelihood from the second stage experiment or ignore it, depending on whether V_{t+1} is one or zero, respectively. Applying the same recipe in this more general case would lead

to a *random* scoring rule, depending on the realization of the random variable V_{t+1} given Y_{t+1}. Nevertheless, being likelihood-based scores, these random scores *could* be used for density forecast comparison, provided that the same realizations of V_{t+1} are used for both density forecasts under consideration. Random scoring rules would obviously not be very practical, but it is important to notice their validity from an partial likelihood point of view. In practice we propose to integrate out the random variation by averaging the random scores over the conditional distribution of V_{t+1} given Y_{t+1}, which is independent of the density forecast. The following clarifies how this leads to generalized CL and CSL scoring rules.

Generalized conditional likelihood scoring rule For the first scoring rule, where only the conditional likelihood of Y_{t+1} given $V_{t+1} = 1$ is used and no other information on the realized values of V_{t+1}, the likelihood given V_{t+1} is

$$\mathrm{I}(V_{t+1} = 1|Y_{t+1}) \log \left(\frac{\hat{f}_t(Y_{t+1})}{\mathsf{P}_{\hat{f}}(V_{t+1} = 1)} \right),$$

where $\mathsf{P}_{\hat{f}}(V_{t+1} = 1)$ is the probability that $V_{t+1} = 1$ under the assumption that Y_{t+1} has density $\hat{f}_t$. This is a random score function, in the sense that it depends on the random variable V_{t+1}. Averaging over the conditional distribution of V_{t+1} given Y_{t+1} leads to $\mathsf{E}_{V_{t+1}|Y_{t+1};f}[\mathrm{I}(V_{t+1} = 1|Y_{t+1})] = \mathsf{P}_{\hat{f}}(V_{t+1} = 1|Y_{t+1}) = w(Y_{t+1})$, so that the score averaged over V_{t+1}, given Y_{t+1}, is

$$S(\hat{f}_t; Y_{t+1}) = w(Y_{t+1}) \log \left(\frac{\hat{f}_t(Y_{t+1})}{\int_{-\infty}^{\infty} \hat{f}_t(x) w(x) \, \mathrm{d}x} \right). \tag{14}$$

It can be seen that this is a direct generalization of the CL scoring rule given in (10), which is obtained by choosing $w(Y_{t+1}) = \mathrm{I}(Y_{t+1} \leq r)$.

Generalized censored likelihood scoring rule As mentioned before, the conditional likelihood scoring rule is based on the conditional likelihood of the second stage experiment only. The censored likelihood scoring rule also includes the information revealed by the realized value of V_{t+1}, that is, the first stage experiment. The log likelihood based on

Y_{t+1} and V_{t+1} is

$$\mathrm{I}(V_{t+1} = 1) \log \hat{f}_t(Y_{t+1}) + \mathrm{I}(V_{t+1} = 0) \log \left(\int_{-\infty}^{\infty} \hat{f}(x)(1 - w(x)) \, \mathrm{d}x \right),$$

which, after averaging over V_{t+1} given Y_{t+1} gives the scoring rule

$$S(\hat{f}_t; Y_{t+1}) = w(Y_{t+1}) \log \hat{f}_t(Y_{t+1}) + (1 - w(Y_{t+1})) \log \left(1 - \int_{-\infty}^{\infty} \hat{f}(x) w(x) \, \mathrm{d}x \right). \quad (15)$$

The choice $w(Y_{t+1}) = \mathrm{I}(Y_{t+1} \leq r)$ gives the CSL scoring rule as given in (11).

Returning to the simulated example concerning the comparison of the normal and Student-t density forecasts, we consider logistic weight functions as given in (12). We fix the center at $r = -2.5$ and vary the slope parameter a among the values 1, 2, 5, and 10. The integrals $\int \hat{f}_t(y) w(y) \, \mathrm{d}y$ and $\int \hat{g}_t(y) w(y) \, \mathrm{d}y$ for the threshold weight function were determined numerically with the CDF routines from the GNU Scientific Library. For other weight functions the integrals were determined numerically by averaging $w(y_{t+1})$ over a large number (10^6) of simulated random variables y_{t+1} with density $\hat{f}_t$ and $\hat{g}_t$, respectively.

[Figure 4 about here.]

Figure 4 shows the empirical CDFs of the mean relative scores $\overline{d}^*$ obtained with the conditional likelihood and censored likelihood scoring rules for the different values of a. It can be observed that for the smoothest weight function considered ($a = 1$) the two score distributions are very similar. The difference between the scores increases as a becomes larger. For $a = 10$, the logistic weight function is already very close to the threshold weight function $\mathrm{I}(y_{t+1} \leq r)$, such that for larger values of a essentially the same score distributions are obtained. The score distributions become more similar for smaller values of a because, as $a \to 0$, $w(y_{t+1})$ in (12) converges to a constant equal to $\frac{1}{2}$ for all values of y_{t+1}, so that $w(y_{t+1}) - (1 - w(y_{t+1})) \to 0$, and moreover $\int w(y) \hat{f}_t(y) \, \mathrm{d}y = \int w(y) \hat{g}_t(y) \, \mathrm{d}y \to \frac{1}{2}$. Consequently, both scoring rules converge to the unconditional likelihood (up to a constant factor 2) and the relative scores d_{t+1}^{cl} and d_{t+1}^{csl} have the limit

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

5 Monte Carlo simulations

In this section we examine the implications of using the weighted logarithmic scoring rule in (4), the censored normal likelihood in (8), the conditional likelihood in (10), and the censored likelihood in (11) for constructing a test of equal predictive ability of two competing density forecasts. Specifically, we consider the size and power properties of the Diebold-Mariano type test as given in (2). The null hypothesis states that the two competing density forecasts have equal expected scores, or

$$H_0 : \mathsf{E}[d^*_{t+1}] = 0,$$

under scoring rule $*$, where $*$ is either *wl*, *cnl*, *cl* or *csl*. We focus on one-sided rejection rates to highlight the fact that some of the scoring rules may favor a wrongly specified density forecast over a correctly specified one. Throughout we use a HAC-estimator for the asymptotic variance of the relative score d^*_{t+1}, that is $\hat{\sigma}^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{K-1} a_k \hat{\gamma}_k$, where $\hat{\gamma}_k$ denotes the lag-k sample covariance of the sequence $\{d^*_{t+1}\}_{t=R}^{T-1}$ and a_k are the Bartlett weights $a_k = 1 - k/K$ with $K = \lfloor P^{-1/4} \rfloor$, where $P = T - R$ is the sample size.

5.1 Size

In order to assess the size properties of the tests a case is required with two competing predictive densities that are both 'equally (in)correct'. However, whether or not the null hypothesis of equal predictive ability holds depends on the weight function $w(y_{t+1})$ that is used in the scoring rules. This complicates the simulation design, also given the fact that we would like to examine how the behavior of the tests depends on the specific settings of the weight function. For the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$ it appears to be impossible to construct an example with two different density forecasts having identical predictive ability regardless of the value of r. We therefore evaluate the size of the tests by focusing on the central part of the distribution using the weight function $w(y) = \mathrm{I}(-r \leq y \leq r)$. As mentioned before, in some cases this region of the distribution may be of primary interest, for instance to monetary policymakers aiming to keep inflation between certain lower and upper bounds. The data generating process (DGP) is taken to be an i.i.d. standard normal distribution, while the two competing density forecasts are normal

distributions with means equal to -0.2 and 0.2, and variance equal to 1. In this case, independent of the value of r the competing density forecasts have equal predictive ability, as the scoring rules considered here are invariant under a simultaneous reflection about zero of all densities of interest (the true conditional density as well as the two competing density forecasts under consideration). In addition, it turns out that for this combination of DGP and predictive densities, the relative scores d^*_{t+1} for the WL, CL and CSL rules based on $w(y) = \mathrm{I}(-r \leq y \leq r)$ are identical. For this weight function the CNL rule takes the form of the last expression in (8), with $\alpha = \mathsf{P}_t(Y_{t+1} = 1) = \mathsf{P}(-r \leq Y_{t+1} \leq r) = \Phi(r) - \Phi(-r)$.

[Figure 5 about here.]

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

It is useful to note that the rejection rates of the CNL scoring rule converge to those based on the WL/CL/CSL scoring rules for large values of r, when practically the entire distribution is taken into account. This is due to the fact that the 'unconditional' scoring rules (which are obtained as $r \to \infty$) all coincide as they become equal to scores based on the unconditional normal likelihood.

5.2 Power

We evaluate the power of the test statistics by performing simulation experiments where one of the competing density forecasts is correct, in the sense that it corresponds with the underlying DGP. In that case the true density always is the best possible one, regardless of the region for which the densities are evaluated, that is, regardless of the weight function used in the scoring rules. Given that our main focus in this paper has been on comparing

density forecasts in the left tail, in these experiments we return to the threshold weight function $w(y) = \mathrm{I}(y \leq r)$ for the WL, CL and CSL rules. For each value of r considered, the CNL score is used with $\alpha = \Phi(r)$, where $\Phi(\cdot)$ represents the standard normal CDF.

[Figure 6 about here.]

[Figure 7 about here.]

Figures 6 and 7 plot the observed rejection rates for sample sizes $P = 500$ and 2000, respectively (again based on $10,000$ replications), for data drawn from the standard normal distribution (top row) or the standardized Student-$t(5)$ distribution (bottom row). In both cases, the null hypothesis being tested is equal predictive accuracy of standard normal and standardized $t(5)$ density forecasts. The left (right) panels in these Figures show rejection rates (at nominal significance level 5%) against superior predictive ability of the standard normal (standardized $t(5)$) distribution, as a function of the threshold parameter r. Hence, the top left and bottom right panels report true power (rejections in favor of the correct density), while the top right and bottom left panels report spurious power (rejections in favor of the incorrect density).

Several interesting conclusions emerge from these graphs. First, for large values of the threshold r, the tests based on the WL, CL and CSL scoring rules behave similarly (recall that they become identical when $r \to \infty$) and achieve rejection rates against the correct alternative of around 80% for $P = 500$ and nearly 100% for $P = 2000$. When the threshold r is fairly large, the CNL-based test has even better power properties for the standard normal DGP, with rejection rates close to 100% already for $P = 500$. By contrast, it has no power whatsoever in case of the standardized Student-t DGP. The latter occurs because in this case the normal scoring rule in (6) (which is the limiting case of the CNL-rule when $r \to \infty$) has the same expected value for both density forecasts, such that the null hypothesis of equal predictive accuracy actually holds true. To see this, let $\hat{f}_t(y)$ and $\hat{g}_t(y)$ denote the $N(0,1)$ and standardized $t(5)$ density forecasts, respectively, with corresponding CDFs $\hat{F}_t$ and $\hat{G}_t$ and inverse normal transforms $Z_{f,t+1}$ and $Z_{g,t+1}$. It then follows that

$$\mathsf{E}_t(Z_{f,t+1}^2) = \int_{-\infty}^{\infty} (y_{t+1})^2 \, \hat{g}_t(y_{t+1}) \, \mathrm{d}y_{t+1} = \mathrm{Var}_t(Y_{t+1}) = 1,$$

and

$$\mathsf{E}_t(Z_{g,t+1}^2) = \int_{-\infty}^{\infty} \left(\Phi^{-1}(\hat{G}_t(y_{t+1}))\right)^2 \hat{g}_t(y_{t+1})\,\mathrm{d}y_{t+1} = \int_0^1 \left(\Phi^{-1}(u)\right)\right)^2 \mathrm{d}u = 1,$$

such that $\mathsf{E}[d_{t+1}^{cN}] = 0$ when r becomes large.

[Figure 8 about here.]

Second, the power of the weighted logarithmic scoring rule depends strongly on the threshold parameter r. For the normal DGP, for example, the test has excellent power for values of r larger than 2 and between -2 and 0, but for other threshold values the rejection rates against the correct alternative drop to zero. In fact, for these regions of threshold values, we observe substantial rejection rates against the incorrect alternative of superior predictive ability of the Student-t density. Comparing Figures 6 and 7 shows that this is not a small sample problem. In fact, the spurious power increases as P becomes larger. To understand the non-monotonous nature of these power curves, we use numerical integration to obtain the mean relative score $\mathsf{E}[d_{t+1}^{wl}]$ for various values of the threshold r for i.i.d. standard normal data. The results are shown in Figure Figure 8. It can be observed that the mean changes sign several times, in exact accordance with the patterns in the top panels of Figures 6 and 7. Whenever the mean score difference (computed as the score of the standard normal minus the score of the standardized $t(5)$ density) is positive the associated test has high power, while it has high spurious power for negative mean scores. Similar considerations also explain the spurious power that is found for the CNL scoring rule. Note that in case of the normal DGP, the rejection rates against the incorrect alternative are particularly high for those values of r that are most relevant when the interest is in comparing the predictive accuracy for the left tail of the distribution, suggesting that the CNL-rule is not suitable for this purpose.

Third, we find that the tests based on our partial likelihood scoring rules have reasonable power when a considerable part of the distribution is taken into account. For positive threshold values, rejection rates against the correct alternative are between 50 and 80% for $P = 500$ and close to 100% for the larger sample size $P = 2000$. For the CL-based test, power declines as r becomes smaller, due to the reduced number of observations

falling in the relevant region that is taken into consideration. The power of the CSL-based statistic remains higher, in particular for the normal DGP, suggesting that the additional information concerning the actual coverage probability of the left tail region helps to distinguish between the competing density forecasts. Perhaps even more importantly, we find that the partial likelihood scores do not suffer from spurious power. For both the CL and CSL rules, the rejection rates against the incorrect alternative remain below the nominal significance level of 5%. The only exception occurs for the i.i.d. standardized $t(5)$ DGP, where the CSL-based exhibits spurious power for small values of the threshold parameter r when $P = 500$. Comparing the bottom-left panels of Figures 6 and 7 suggests that this is a small sample problem though, as the rejection rates decline considerably when increasing the number of forecasts to $P = 2000$.

[Figure 9 about here.]

Finally, to study the power properties of the tests when they are used to compare density forecasts on the central part of the distribution, we perform the same simulation experiments but using the weight function $w(y) = \mathrm{I}(-r \leq y \leq r)$ in the various scoring rules. Only a small part of these results are included here; full details are available upon request. Figure 9 shows rejection rates obtained for an i.i.d. standard normal DGP, when we test the null of equal predictive ability of the $N(0,1)$ and standardized $t(5)$ distributions against the alternative that either of these density forecasts has better predictive ability, for sample size $P = 500$. The left panel shows rejection rates against better predictive performance of the (correct) $N(0,1)$ density, while the right panel shows the spurious power, that is rejection rates against better predictive performance of the (incorrect) standardized $t(5)$ distribution. Clearly, all tests have high power, provided that the observations from a sufficiently wide interval $(-r, r)$ are taken into account. It can also be observed that the tests based on WL and CNL scores suffer from a large spurious power even for quite reasonable values of r, while the spurious power for the tests based on the partial likelihood scores remains smaller than the nominal level (5%).

6 Empirical illustration

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

For illustrative purposes we define two forecast methods based on GARCH models in such a way that *a priori* one of the methods is expected to be superior to the other. Examining a large variety of GARCH specifications for forecasting daily US stock index returns, Bao *et al.* (2007) conclude that the accuracy of density forecasts depends more on the choice of the distribution of the standardized innovations than on the choice of the volatility specification. Therefore, we differentiate our forecast methods in terms of the innovation distribution, while keeping identical specifications for the conditional mean and the conditional variance. We consider an AR(5) model 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 has been that GARCH models with a normal distribution for η_t are not able to fully account for the kurtosis observed in stock returns. We therefore concentrate on leptokurtic distributions for the standardized innovations. Specifically, for one forecast method the distribution of 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 returns, using a rolling window scheme for parameter estimation. The width of the estimation window is set to $R = 2000$ observations, so that the number of out-of-sample observations is equal to $P = 5115$. For comparing the density forecasts' accuracy we use the Diebold-Mariano type test based on the weighted logarithmic scoring rule in (4), the censored normal likelihood in (8), the conditional likelihood in (10), and the censored likelihood in (11). We concentrate on the left tail of the distribution by using the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r_t)$ for the WL, CL and CSL scoring rules. We consider two time-varying thresholds r_t, that are determined as the one-day Value-at-Risk estimates at the 95% and 99% level based on the corresponding quantiles of the empirical CDF of the return observations in the relevant estimation window. For the CNL scoring rule in (8) we use the corresponding values $\alpha = 0.05$ and 0.01, respectively. 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 1 shows the average score differences $\overline{d}^*$ with the accompanying tests of equal predictive accuracy as in (2), where we use a HAC estimator for the asymptotic variance $\hat{\sigma}^2$ 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 both choices of the threshold r_t the WL and CNL scoring rules suggest superior predictive ability of the forecast method based on Laplace innovations. By contrast, the CL scoring rule suggests that the performance of the GARCH-t density forecasts is superior. The CSL scoring rule points towards the same conclusion as the CL rule, although the evidence for better predictive ability of the GARCH-t specification is somewhat weaker. In the remainder of this section we seek to understand the reasons for these conflicting results, and explore the consequences of selecting either forecast method for risk management purposes. In addition, this allows us to obtain circumstantial evidence that shows which of the two competing forecast methods is most appropriate.

[Table 1 about here.]

For most estimation windows, the degrees of freedom parameter in the Student-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 and CNL scoring rules indicate superior predictive ability of the Laplace distribution simply because this density has more probability mass in the region of interest, that is, the problem that motivated our analysis in the first place. To see this from a slightly different perspective, we compute one-day 95% and 99% Value-at-Risk (VaR) and Expected Shortfall (ES) estimates as implied by the two forecast methods. The $100 \times (1-\alpha)\%$ Value-at-Risk 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 Expected Shortfall 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 and, thus, imply 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 2.

[Figure 10 about here.]

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 95% and 99% VaR violations, which should be close to 0.05 and 0.01, respectively, if the innovation distribution is correctly specified. We compute the likelihood ratio (LR) test of correct unconditional coverage (CUC) suggested by Christoffersen (1998) to determine whether the empirical violation frequencies differ significantly from these nominal levels. Additionally, we use Christoffersen's (1998) LR tests of independence of VaR violations (IND) and for correct conditional coverage (CCC). Define the indicator variables $\mathbb{I}_{\hat{f},t+1}(y_{t+1} \leq \mathrm{VaR}_{\hat{f},t}(\alpha))$ for $\alpha = 0.05$ and 0.01, which take the value 1 if the condition in brackets is satisfied and 0 otherwise. Independence of the VaR violations is tested against a first-order Markov alternative, that is, the null hypothesis is given by $H_0 : \mathsf{E}(\mathbb{I}_{\hat{f},t+1}|\mathbb{I}_{\hat{f},t}) = \mathsf{E}(\mathbb{I}_{\hat{f},t+1})$. 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 Expected Shortfall estimates $\mathrm{ES}_{\hat{f},t}(\alpha)$ we employ the test suggested by McNeil and Frey (2000). For every return y_t that falls below the $\mathrm{VaR}_{\hat{f},t}(\alpha)$ estimate, define the standardized 'residual' $e_{t+1} = (y_{t+1} - \mathrm{ES}_{\hat{f},t}(\alpha))/h_{t+1}$, where h_{t+1} is the conditional volatility forecast obtained from the corresponding GARCH model. Under the null of correct specification, the expected value of e_{t+1} is equal to zero, which can easily be assessed by means of a two-sided t-test with HAC variance estimator.

[Table 2 about here.]

The results reported in Table 2 show that the empirical VaR violation frequencies are very close to the nominal levels for the Student-t innovation distribution. For the Laplace distribution, they are considerably lower. This is confirmed by the CUC test, which convincingly rejects the null of correct unconditional coverage for the Laplace distribution but not for the Student-t distribution. The null hypothesis of independence is not rejected in any of the cases at the 5% significance level. Finally, the McNeil and Frey (2000) test does not reject the adequacy of the 95% ES estimates for either of the two distributions, but it does for the 99% ES estimates based on the Laplace innovation distribution. In sum, the

VaR and ES estimates suggest that the Student-t distribution is more appropriate than the Laplace distribution, confirming the density forecast evaluation results obtained with the scoring rules based on partial likelihood. In terms of risk management, using the GARCH-Laplace forecast method would lead to larger estimates of risk than the GARCH-t forecast method. This, in turn, could result in suboptimal asset allocation and 'over-hedging'.

7 Conclusions

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

Monte Carlo simulations were used to demonstrate that the conventional scoring rules may indeed give rise to spurious rejections due to the possible bias in favor of an incorrect density forecast. The simulation results also showed that this phenomenon is virtually non-existent for the new scoring rules, and where present, diminishes quickly upon increasing the sample size.

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

References

Amisano, G. and Giacomini, R. (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 Ng, S. (2005). Tests for skewness, kurtosis, and normality of time series data. *Journal of Business and Economic Statistics*, **23,** 49–60.

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

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

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

Berg, D. and Bakken, H. (2005). A goodness-of-fit test for copulae based on the probability integral transform. Technical report number SAMBA/41/05, Norwegian Computing Center.

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

Bollerslev, Tim (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, **69,** number 3, 542–547.

Breymann, W., Dias, A. and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, **3,** 1–14.

Campbell, S.D. and Diebold, F.X. (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*. New York: Palgrave-Macmillan.

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

Clements, M.P. and Smith, J. (2002). Evaluating multivariate forecast densities: A comparison of two approaches. *International Journal of Forecasting*, **18,** 397–407.

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

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

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

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

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62,** 269–276.

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

Diebold, F.X., Hahn, J. and Tay, A.S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics*, **81,** 661–673.

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

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

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

Dowd, K. (2005). *Measuring Market Risk*, 2 edn. Chicester: John Wiley & Sons.

Egorov, A.V., Hong, Y. and Li, H. (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 van Dijk, D. (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., Lee, K., Pesaran, M.H. and Shin, Y. (2003). Forecast uncertainties in macroeconomic modelling: An application to the UK economy. *Journal of the American Statistical Association*, **98,** 829–838.

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

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

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

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

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

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

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

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

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

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

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

Perez-Quiros, G. and Timmermann, A. (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 Wohar, M.E. (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 Valente, G. (2004). Comparing the accuracy of density forecasts from competing models. *Journal of Forecasting*, **23,** 541–557.

Sarno, L. and Valente, G. (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 Buizza, R. (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.

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

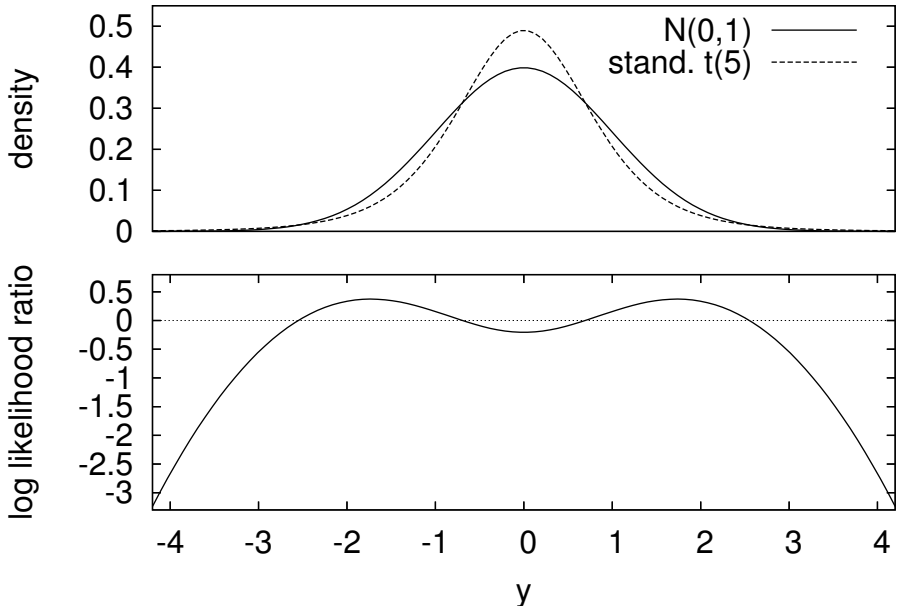

Figure 1: *Probability density functions of the standard normal distribution* $\hat{f}_t(y_{t+1})$ *and standardized Student-*$t(5)$ *distribution* $\hat{g}_t(y_{t+1})$ *(upper panel) and corresponding relative log-likelihood scores* $\log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$ *(lower panel).*

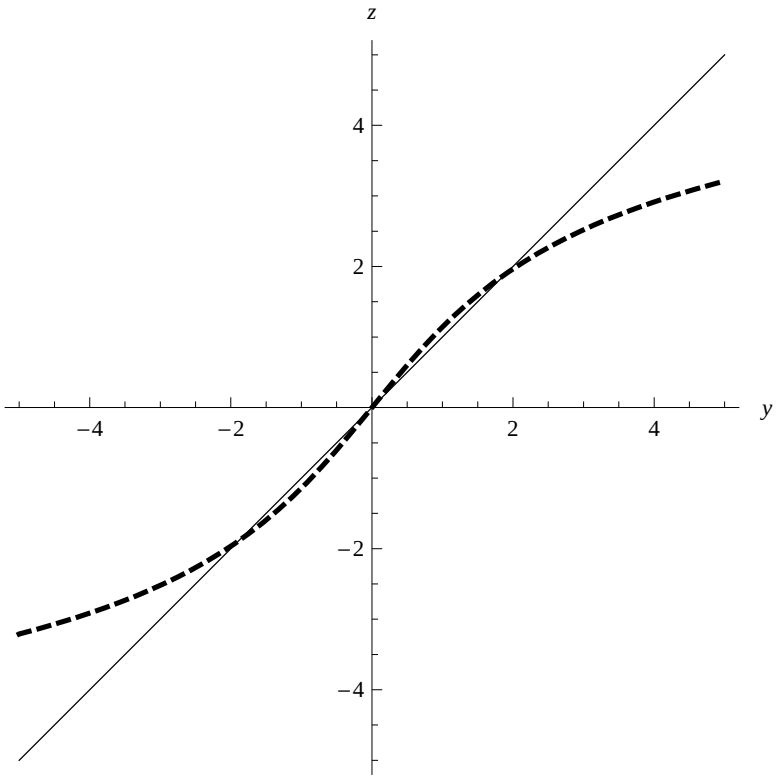

Figure 2: *Inverse normal transformations* $z_{\hat{f},t+1} = \Phi^{-1}(\hat{F}_t(y_{t+1}))$ *of the probability integral transforms of the standard normal distribution (solid line) and the standardized Student-*$t(5)$ *distribution (dashed line).*

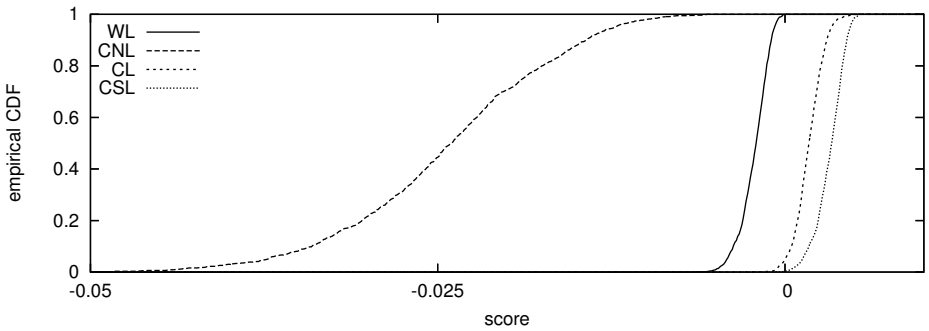

Figure 3: *Empirical CDFs of mean relative scores* $\overline{d}^*$ *for the weighted logarithmic (WL) scoring rule in* (4)*, the censored normal likelihood (CNL) in* (8)*, the conditional likelihood (CL) in* (10)*, and the censored likelihood (CSL) in* (11) *for series of* $P = 2000$ *independent observations from a standard normal distribution. The scoring rules are based on the threshold weight function* $w(y) = I(y \leq r)$ *with* $r = -2.5$*. 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.*

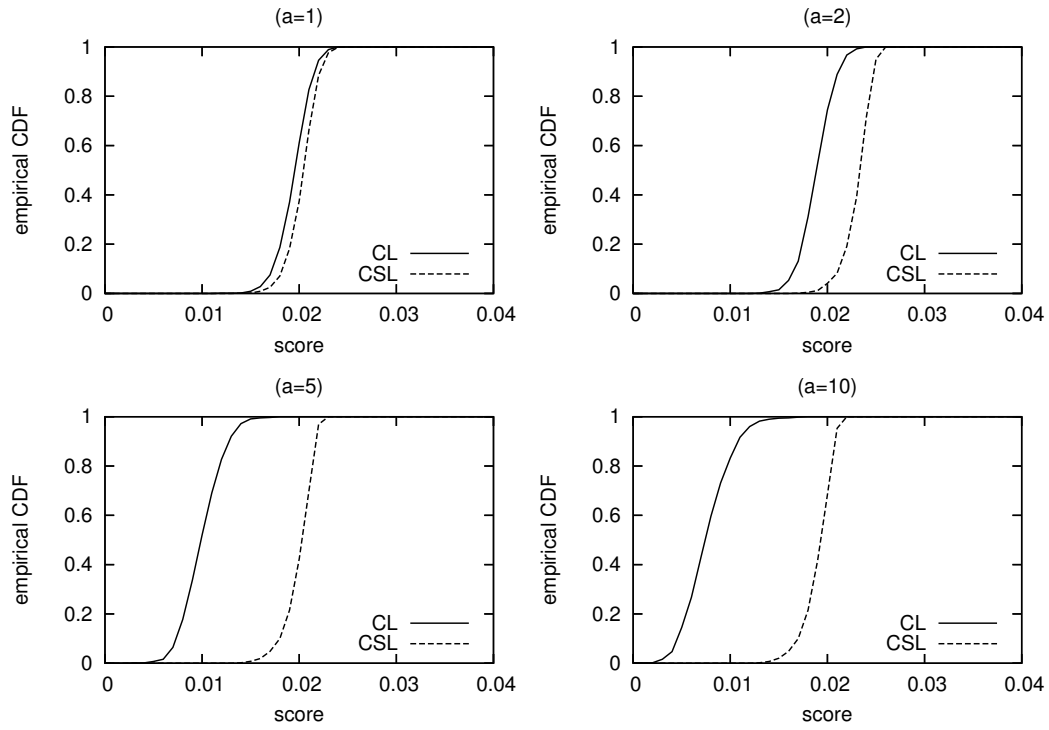

Figure 4: *Empirical CDFs of mean relative scores* $\overline{d}^*$ *for the generalized conditional likelihood (CL) and censored likelihood (CSL) scoring rules for series of* $P = 2000$ *independent observations from a standard normal distribution. The scoring rules are based on the logistic weight function* $w(y)$ *defined in* (12) *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 5: *One-sided rejection rates of the Diebold-Mariano type test statistic of equal predictive accuracy defined in* (2) *when using the weighted logarithmic (WL), the censored normal likelihood (CNL), the conditional likelihood (CL), and the censored likelihood (CSL) scoring rules, under the weight function* $w(y) = I(-r \leq y \leq r)$ *for sample size* $P = 500$*, 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.*

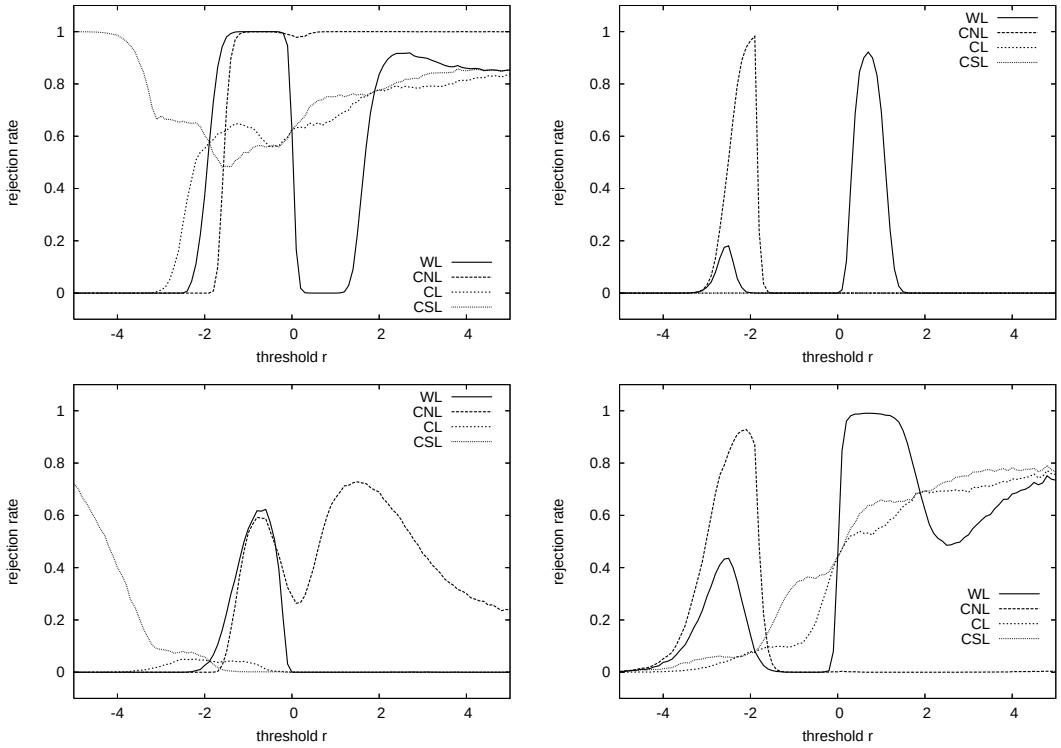

Figure 6: *One-sided rejection rates (at nominal significance level* 5%*) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in* (2) *when using the weighted logarithmic (WL), the censored normal likelihood (CNL), the conditional likelihood (CL), and the censored likelihood (CSL) scoring rules, under the threshold weight function* $w(y) = I(y \leq r)$ *for sample size* $P = 500$*, based on 10,000 replications. For the graphs in the top and bottom rows, the DGP is i.i.d. standard normal and i.i.d. standardized* $t(5)$*, respectively. The test compares the predictive accuracy of the standard normal and the standardized* $t(5)$ *distributions. The graphs in the left (right) panels show rejection rates against superior predictive ability of the standard normal (standardized* $t(5)$*) distribution, as a function of the threshold parameter* r*.*

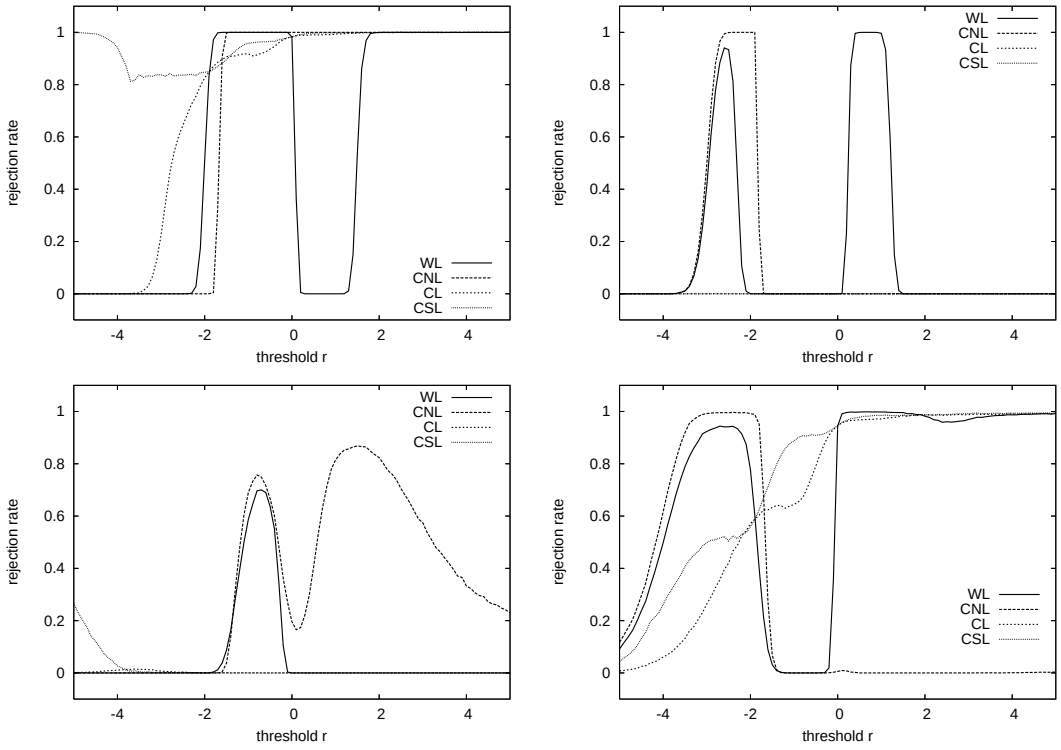

Figure 7: *One-sided rejection rates (at nominal significance level* 5%*) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in* (2) *when using the weighted logarithmic (WL), the censored normal likelihood (CNL), the conditional likelihood (CL), and the censored likelihood (CSL) scoring rules, under the threshold weight function* $w(y) = I(y \leq r)$ *for sample size* $P = 2000$*, based on 10,000 replications. Specifications of the simulation experiments are identical to Figure 6.*

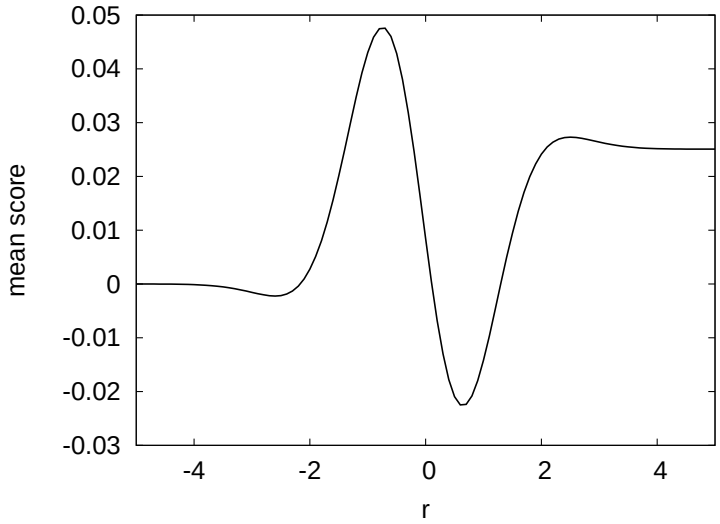

Figure 8: *Mean relative WL score* $E[d_{t+1}^{wl}]$ *with threshold weight function* $w(y) = I(y \leq r)$ *for the standard normal versus the standardized* $t(5)$ *density as a function of the threshold value* r*, for the standard normal DGP.*

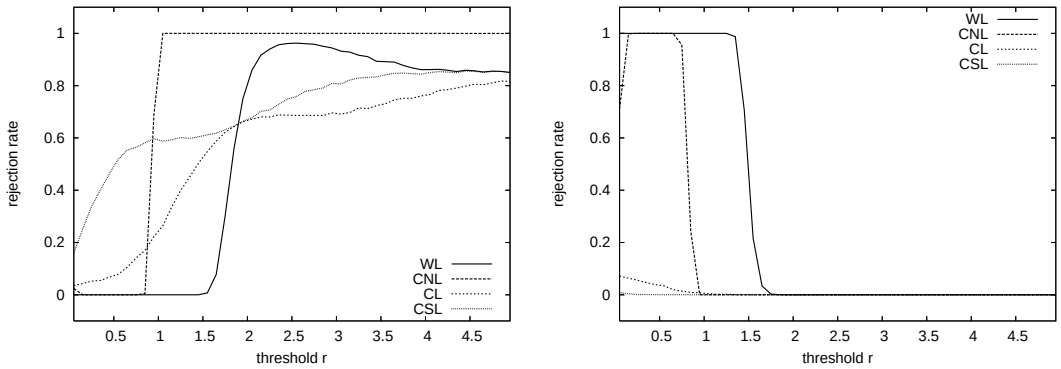

Figure 9: *One-sided rejection rates (at nominal significance level* 5%*) of the Diebold-Mariano type test statistic of equal predictive accuracy defined in* (2) *when using the weighted logarithmic (WL), the censored normal likelihood (CNL), the conditional likelihood (CL), and the censored likelihood (CSL) scoring rules, under the weight function* $w(y) = I(-r \leq y \leq r)$ *for sample size* $P = 500$*, based on 10,000 replications. The DGP is i.i.d. standard normal. The graphs on the left and right show rejection rates against better predictive ability of the standard normal distribution compared to the standardized* $t(5)$ *distribution and vice versa.*

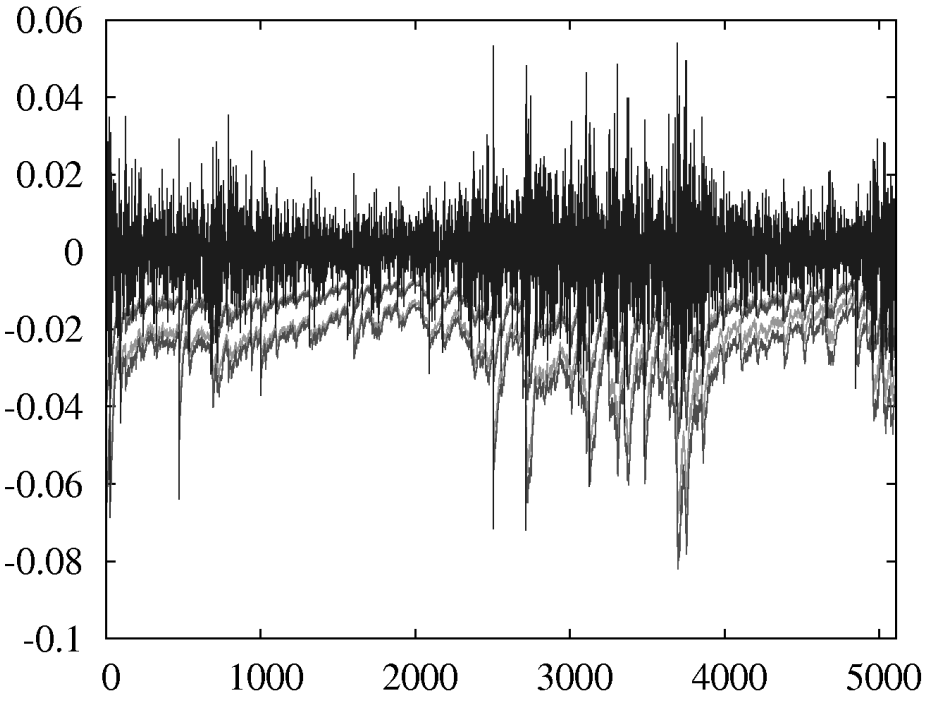

Figure 10: *Daily S&P 500 log-returns (black) for the period December 2, 1987 – March 14, 2008 and out-of-sample 95% and 99% VaR forecasts derived from the AR(5)-GARCH(1,1) specification using Student-t innovations (light gray) and Laplace innovations (dark gray).*

Table 1: *Average score differences and tests of equal predictive accuracy*

Scoring rule	$\alpha = 0.05$		$\alpha = 0.01$	
	$\overline{d}^*$	Test	$\overline{d}^*$	Test
WL	-0.0053	-4.820	-0.0032	-3.835
CNL	-0.0081	-5.269	-0.0068	-4.382
CL	0.0016	2.328	0.0008	1.819
CSL	0.0016	1.537	0.0012	1.373

Note: The table presents the average score difference $\overline{d}^*$ for the weighted logarithmic (WL) scoring rule in (4), the censored normal likelihood (CNL) in (8), the conditional likelihood (CL) in (10), and the censored likelihood (CSL) in (11). The WL, CL and CSL scoring rules are based on the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r_t)$, where r_t is the α-th quantile of the empirical (in-sample) CDF, where $\alpha = 0.01$ or 0.05. These values for α are also used for the CNL scoring rule. The score difference d_{t+1} is computed for density forecasts obtained from an AR(5)-GARCH(1,1) model with (standardized) Student-$t(\nu)$ innovations relative to the same model but with Laplace innovations, for daily S&P500 returns over the evaluation period December 2, 1987 – March 14, 2008.

Table 2: *Value-at-Risk and Expected Shortfall characteristics*

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

Note: The average VaRs reported are the observed average 5% and 1% quantiles of the density forecasts based on the GARCH model with $t(\nu)$ and Laplace innovations, respectively. The coverages correspond with the observed fraction of returns below the respective VaRs, which ideally would coincide with the nominal rate 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 expected shortfalls (equal to the conditional mean return, given a realization below the predicted VaR) based on the different density forecasts. The bottom two rows report McNeil-Frey test statistics and corresponding p-values for evaluating the expected shortfall estimates $\mathrm{ES}_{\hat{f},t}(\alpha)$.