Partial Likelihood Ratio-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

January 22, 2008

Abstract

We propose and evaluate several new scoring rules based on (partial) likelihood ratios for comparing the out-of-sample accuracy of competing density forecasts. These scoring rules are particularly useful when the main interest lies in measuring the predictive accuracy over a specific region of the density, such as the left tail in financial risk management. By construction, conventional scoring rules based on KLIC or censored normal likelihood tend to favor density forecasts with more probability mass in the region of interest, rendering the resulting tests biased towards such densities. Our novel scoring rules based on partial likelihood 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 seminar participants at Queensland University of Technology, the University of New South Wales, the University of Amsterdam 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, of course 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.*, 1998), 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, 2008). Finally, it is 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 model for 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 was 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.* (2007). All statistics in this group compare the relative distance between the competing density forecasts and the true (but unobserved) density. 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. Berkowitz

(2001) 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 comparing two density forecasts on a specific region is possible using scoring rules based on partial likelihood. The main problem encountered when comparing two competing forecast densities on a given region is the presence of nuisance parameters. In this case there is an infinite dimensional nuisance parameter: the shape of the predictive density outside the region of interest. Following Cox (1975), we argue that partial likelihood can be used to annihilate the effect of nuisance parameters. To illustrate how this might be implemented, we introduce two possible scoring rules based on partial likelihood, and study their performance using Monte Carlo simulations as well as an empirical illustration. It turns out that the partial likelihood scoring rule that uses most of the available relevant information performs best 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 likelihoods, and demonstrate that these do not suffer from this problem. This is further illustrated by means of Monte Carlo simulation experiments in Section 5, where we assess the properties of tests of equal accuracy of density forecasts based on different scoring rules. An empirical application concerning density forecasts for daily S&P 500 returns in Section 6 demonstrates the practical usefulness of our approach. Finally, Section 7 concludes.

2 Scoring rules for evaluating density forecasts

Following Amisano and Giacomini (2007) we consider a stochastic process $\{Z_t : \Omega \longrightarrow \mathbb{R}^{s+1}\}_{t=1}^{T}$, defined on a complete probability space $(\Omega, \mathcal{F}, P)$, and identify Z_t with $(Y_t, X_t')'$, where $Y_t : \Omega \longrightarrow \mathbb{R}$ is the real valued random variable of interest and $X_t : \Omega \longrightarrow R^s$ 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 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) $\hat{f}_{m,t}(y)$ and $\hat{g}_{m,t}(y)$, respectively. The subscript m indicates that the density forecasts are assumed to depend only on $Z_{t-m+1}, \ldots, Z_t$. 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 the last m observations only. This finite memory simplifies the asymptotic theory of test statistics considerably. To keep the exposition as simple as possible, in this paper we will be mainly concerned with the simplest case of comparing 'fixed' predictive densities for i.i.d. processes. We therefore drop the subscript m from the predictive densities.

Our interest lies in comparing the relative performance $\hat{f}_t(Y_{t+1})$ 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. We refer to Lahiri and Wang (2007) for an interesting application of several such rules to the evaluation of probability forecasts of 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 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 corresponding realizations of the time series variable, competing density forecasts may then be compared based on their average scores. Mitchell and Hall (2005), Amisano and Giacomini (2007), and Bao *et al.* (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 n density forecasts and realizations for $t = m+1, \ldots, T \equiv m+n$, the density forecasts $\hat{f}_t$ and $\hat{g}_t$ can be ranked according to their average scores $n^{-1}\sum_{t=m+1}^{T} \log \hat{f}_t(y_{t+1})$ and $n^{-1}\sum_{t=m+1}^{T} \log \hat{g}_t(y_{t+1})$. 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

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

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

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

where $\overline{d}^l$ is the sample average of the score difference, that is, $\overline{d}^l = \frac{1}{n}\sum_{t=m+1}^{T} d_t^l$, and $\hat{\sigma}^2$ is a consistent estimator of the asymptotic variance of $\overline{d}_t^l$, see Mitchell and Hall (2005) and Bao *et al.* (2007) for details.

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.* (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

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

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 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.* (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.* (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 $p_t(y_{t+1})$ 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 $\hat{f}_t$. In other words, defining $z_{f,t+1} = \Phi^{-1}(\hat{F}_t(y_{t+1}))$ with $\hat{F}_t(y_{t+1}) = \int_0^{y_{t+1}} \hat{f}_t(y)dy$ and Φ the standard normal distribution function, it holds that

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

where q_t is the true conditional density of $z_{f,t+1}$ and ϕ is the standard normal density. Of course, in practice the density q_t is not known either, but if $\hat{f}_t$ is correctly specified, $\{z_{f,t+1}\}$ should behave as an i.i.d. standard normal sequence. As discussed in Bao *et al.* (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 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, while ignoring the remaining part of the distribution. 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 $n^{-1}\sum_{t=m+1}^{T} w(y_{t+1})\log \hat{f}_t(y_{t+1})$ and $n^{-1}\sum_{t=m+1}^{T} w(y_{t+1})\log \hat{g}_t(y_{t+1})$ for ranking two competing forecasts. Using the weighted score difference

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

the null hypothesis of equal weighted scores, $H_0 : \mathsf{E}[d_t^{wl}] = 0$, may be tested by means of a Diebold-Mariano type test statistic of the form (2), but using the sample average $\overline{d}^{wl} = n^{-1}\sum_{t=m+1}^{T} d_t^{wl}$ instead of $\overline{d}^{l}$ together with an estimate of the corresponding asymptotic variance of d_t^{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 the density region of interest. For evaluation of the left tail in risk management applications, for example, we may decide to 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. 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, note that if $\hat{g}_t(Y_{t+1}) > \hat{f}_t(Y_{t+1})$ 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^{wl} in (5) is never positive, and strictly negative for observations below the threshold value r, such that $\mathsf{E}[d_t^{wl}]$ is negative. Obviously, this can have far-reaching consequences when comparing models with different tail behavior. In particular, there will be cases where the fat-tailed distribution $\hat{g}$ is favored over the thin-tailed distribution $\hat{f}$, 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

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

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 positive 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 $y_{m+1}, \ldots, y_T$ of n 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 $w(y) = \mathrm{I}(y_{t+1} \leq r)$ 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 this is in favor of the fat-tailed Student-t density $g_t(y_{t+1})$.

[Figure 1 about here.]

We emphasize that the problem discussed 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 $z_{f,t+1}$, as defined before, motivated by the fact that if $\hat{f}_t$ is correctly specified, $\{Z^*_{f,t+1}\}$ 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:

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

We argue that although for a correctly specified density forecast the sequence $\{z_{f,t+1}\}$ is i.i.d. normal, tests for the comparative accuracy of density forecasts based on (censored versions of) the normal log-likelihood may be biased, i.e. reject at a rate lower than the nominal size for certain alternatives. The reason is that, unlike in a standard test for normality, a particular forecast density $\hat{g}_t$ affects the sample distribution of the corresponding values $z_{g,t+1}$. It may then happen that a wrongly specified forecast density $\hat{g}_t$ receives a higher average score than the true conditional density. For instance, let $\hat{f}_t(y)$ and $\hat{g}_t(y)$ denote two competing densities, which happen to be the densities of the $N(0,1)$ and the

$N(0,2)$ distribution, respectively. Assume furthermore that the data Y_t are independent $N(0,1)$ random variables, such that $\hat{f}_t$ corresponds with the true conditional density. In this case the corresponding CDFs satisfy $|\widehat{F}_t(y)-\frac{1}{2}| > |\widehat{G}_t(y)-\frac{1}{2}|$ for all y, which implies $|z_{g,t+1}| < |z_{f,t+1}|$. In words, the incorrect density forecast has a larger dispersion than the correct one, such that its corresponding transforms $\{z_{g,t+1}\}$ are more concentrated around 0 than the transforms $\{z_{f,t+1}\}$. As a result, the incorrect density forecast $\hat{g}_t$ receives higher normal log-likelihood scores S^N than the correct density forecast $\hat{f}_t$, leading to biased test statistics.

Note that the bias described above only plays a role in testing density forecasts against alternative density forecasts. It is still possible to test a single given model for goodness-of-fit by testing for the i.i.d. standard normality of $\{z_{g,t+1}\}$ against non-standard normal or dependent alternatives. To illustrate the effect of the bias in our context, i.e. comparing *pairs* of density forecasts, we also include differences of Berkowitz-type scores in the simulation results presented below. Following Berkowitz (2001) we focus on the left tail by considering the censored variable $\tilde{z}_{f,t+1}$ defined as $\tilde{z}_{f,t+1} = z_{f,t+1}$ if $z_{f,t+1} < \Phi^{-1}(\alpha)$, and $\tilde{z}_{f,t+1} = \Phi^{-1}(\alpha)$ otherwise, for some $0 < \alpha < 1$. This allows us to associate a score equal to the censored normal log-likelihood (CNL) score function

$$\begin{aligned} S^{cN}(\hat{f}, y_{t+1}) &= \mathrm{I}(\widehat{F}_t(y_{t+1}) < \alpha)\log\phi(z_{f,t+1}) - \mathrm{I}(\widehat{F}_t(y_{t+1}) \geq \alpha)\log(1-\alpha) \\ &= w(z_{f,t+1})\log\phi(z_{f,t+1}) + (1-w(z_{f,t+1}))\log(1-\alpha), \end{aligned} \tag{7}$$

where $F_t(\cdot)$ denotes the CDF of f, and $w(z_{f,t+1}) = 1$ if $z_{f,t+1} < \Phi^{-1}(\alpha)$ and zero otherwise.

3 Scoring rules based on partial likelihood

The simple example in the previous section demonstrates that intuitively reasonable scoring rules can in fact favor the wrong model. 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. In the standard, unweighted case the log likelihood score $\log \hat{f}_t(Y_{t+1})$ is useful for measuring the divergence between the true and a candidate density, because the expectation

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

is maximized, under the constraint $\int_{-\infty}^{\infty} f_t(y_{t+1})dy = 1$, by $f_t(y_{t+1}) = p_t(y_{t+1})$, where $p_t(y_{t+1})$ is the true density of Y_{t+1}. To see this, consider any density $f_t(y_{t+1})$ different from $p_t(y_{t+1})$. Applying the inequality $\log x \leq x - 1$ to f_t/p_t, we obtain

$$\mathsf{E}_Y\left[\log\left(\frac{f_t(Y_{t+1})}{p_t(Y_{t+1})}\right)\right] \leq \mathsf{E}_Y\left[\frac{f_t(Y_{t+1})}{p_t(Y_{t+1})}\right] - 1 = \int_{-\infty}^{\infty} p_t(y)\frac{f_t(y)}{p_t(y)}dy - 1 \leq 0.$$

This shows that log-likelihood scores of different models 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 gives back the logarithmic score difference discussed before, $d_t^l = \log \hat{f}_t(y_{t+1}) - \log \hat{g}_t(y_{t+1})$.

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

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

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 associated with the observed values $V_{t+1} = \mathrm{I}(Y_{t+1} \le r) = v$ and subsequently $Y_{t+1} = y$, can thus be written as the product of two factors:

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

The likelihood is expressed as a product of the likelihood associated with 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. By deciding to disregard 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, which still can be interpreted as a true (albeit partial) likelihood.

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

$$S^{cl}(\hat{f}; Y_{t+1}) = \mathrm{I}(Y_{t+1} \le r) \log(\hat{f}_t(Y_{t+1}) / \hat{F}_t(r)). \tag{8}$$

The main argument for using such a score function would be to evaluate models 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 model 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 similar scores to models 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 scheme 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

$$S^{csl}(\hat{f}; Y_{t+1}) = \mathrm{I}(Y_{t+1} \leq r) \log \hat{f}_t(Y_{t+1}) + \mathrm{I}(Y_{t+1} > r) \log(1 - \hat{F}_t(r)). \tag{9}$$

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 which cannot be observed above a certain threshold value (see Tobin, 1958). Note that this score differs from the Berkowitz score (7) in that it is evaluated directly in terms of Y_{t+1}, and not in terms of the (density forecast-dependent) normal transformed variable $z_{g,t+1}$. The effect of introducing a possible bias in favour of an incorrect predictive density, discussed around (6) and (7), does not occur here.

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 (8) or the censored likelihood score (9) in the same manner as before. That is, given a sample of density forecasts and corresponding realizations for n time periods $t = m + 1, \dots, T$, we may form the relative scores $d_t^{cl} = S^{cl}(\hat{f}; y_{t+1}) - S^{cl}(\hat{g}; y_{t+1})$ and $d_t^{csl} = S^{csl}(\hat{f}; y_{t+1}) - S^{csl}(\hat{g}; y_{t+1})$ 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 $1,000$ independent sequences of size $n = 2,000$ independent observations y_{t+1} from a standard normal distribution. For each sequence we compute the mean value of the weighted logarithmic scoring rule in (4), the censored normal likelihood in (7), the conditional likelihood in (8), and the censored likelihood in (9). 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 2 about here.]

Figure 2 shows the empirical CDF of the mean relative scores $\overline{d}^{\cdot}$ for the four rules considered. 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.

4 Generalizations to smooth weight functions

The two scoring rules discussed in the previous section focus on the case where V_{t+1} is a deterministic function of Y_{t+1}, that is, $V_{t+1} = \mathrm{I}(Y_{t+1} \leq r)$. 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 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 between 1% and 3%, the central part of the distribution may be of most interest. In the remainder of this section, we consider a possible generalization of the CL and CSL scoring rules in (8) and (9) to alternative weight functions.

As a straightforward approach we propose making V_{t+1} a random variable conditional on Y_{t+1}. The scoring rules that make use of the threshold weight function either include the second stage information or ignore it, depending on the value of Y_{t+1}. The obvious generalization would be to specify a weight function $w(y)$, satisfying $0 \leq w(y) \leq 1$, and interpret $w(Y_{t+1})$ as a conditional probability that the second-stage information is taken into account, that is,

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

Implemented as such, this recipe would lead to a *random* scoring rule, depending on the realization of the random variable V_{t+1}. However, since each scoring rule obtained in this way would be equally admissible, because it is a likelihood-based score conditional on the realization of V_{t+1}, we argue for averaging this score function over the conditional distribution of V_{t+1} given Y_{t+1}. The first stage information now corresponds to that involved in the subdivision of the data into two groups, that of elements for which $V_{t+1} = 0$ and 1. This first stage information in the generalized case is now not only random, as before, but even random conditional on Y_{t+1}.

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

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

where $P_{\hat{f}}(V_{t+1} = 1)$ is the probability that $V_{t+1} = 1$ under the assumption that Y_{t+1} has density $\hat{f}_t$. Given Y_{t+1} this is a random score function, depending on the realized values of V_{t+1}. Averaging over the conditional distribution of V_{t+1} given Y_{t+1} leads to $\mathsf{E}_{V_{t+1}|Y_{t+1};f}[\mathrm{I}(V_{t+1} = 1|Y_{t+1})] = P_f(V_{t+1} = 1|Y_{t+1}) = w(Y_{t+1})$, so that the score averaged over V_{t+1}, given Y_{t+1}, is

$$S(\hat{f}; Y_{t+1}) = w(Y_{t+1}) \log \left(\frac{\hat{f}_t(Y_{t+1})}{\int_{-\infty}^{\infty} \hat{f}_t(x) w(x) dx} \right). \tag{10}$$

It can be seen that this is a direct generalization of the scoring rule given in (8), 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 likelihood based on a single observation Y_{t+1} and given V_{t+1} is

$$\mathrm{I}(V_{t+1}=1)P_{\hat{f}}(V_{t+1}=1)\hat{f}_t(Y_{t+1}|V_{t+1}=1)+\mathrm{I}(V_{t+1}=0)P_{\hat{f}}(V_{t+1}=0).$$

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

$$S(\hat{f};Y_{t+1})=w(Y_{t+1})\log\hat{f}_t(Y_{t+1})+(1-w(Y_{t+1}))\log\left(1-\int_{-\infty}^{\infty}\hat{f}(x)w(x)\,\mathrm{d}x\right). \quad (11)$$

The choice $w(Y_{t+1})=\mathrm{I}(Y_{t+1}\le r)$ gives the scoring rule as given in (9).

Returning to the simulated example concerning the comparison of the normal and Student t density forecasts, we consider logistic weight functions of the form

$$w(y_{t+1})=1/(1+\exp(a(y_{t+1}-r))). \quad (12)$$

This is a sigmoidal function of y with center r and slope parameter a. Note that in the limit as $a\to\infty$, the threshold weight function $\mathrm{I}(y_{t+1}\le r)$ considered before is recovered. We fix the center at $r=-2$ and vary the slope parameter a among the values 1, 2, 5, and 10. The integrals $\int\hat{f}_t(y)w(y)dy$ and $\int\hat{g}_t(y)w(y)dy$ 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 3 about here.]

Figure 3 shows the empirical CDFs of the scores 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 $w(y_{t+1})$ in (12) 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. To understand why the score distributions become more similar for smaller values of a, note that for small a both scoring rules converge to the unconditional likelihood (up to a constant factor 2). In the limit as $a \to 0$, $w(y_{t+1})$ 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 $\int w(y)\hat{f}_t(y)dy = \int w(y)\hat{g}_t(y)dy \to \frac{1}{2}$. Consequently, both relative scores d_t^{cl} and d_t^{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 present the results of a Monte Carlo analysis to examine the properties of the Diebold-Mariano type test statistic as given in (2) based on the various scoring rules discussed before. We use a HAC-estimator for the asymptotic variance $\hat{\sigma}^2$ of the relative score $d_t^{\cdot}$, that is $\hat{\sigma}_n^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{K-1} a_k\hat{\gamma}_k$, where $\hat{\gamma}_k$ denotes the lag-k sample covariance of the sequence $\{d_t^{\cdot}\}_{t=m+1}^{T}$ and a_k are the Bartlett weights $a_k = 1 - k/K$ with $K = \lfloor n^{-1/4} \rfloor$, where $n = T - m$ is the sample size. Under the null hypothesis of equal predictive ability the test statistic is asymptotically standard normally distributed. Note that the concept of equal predictive ability depends on the (sequence of) weight functions used. We report one-sided size and powers to indicate which of the pair of competing predictive densities has better predictive ability according to each of the tests.

[Figure 4 about here.]

To assess the size of tests a case is required with two competing predictive densities that are both 'equally incorrect'. However, whether or not the null hypothesis of equal predictive ability holds (for a given notion of predictive ability) depends on the weight function used. Therefore it is impossible to construct an example with two different predictive densities for which the predictive ability is identical, regardless of the weight function.

In order to still evaluate the size of the tests, we include simulations for an i.i.d. standard normal sequence, with two competing predictive distributions, the $N(-0.2, 1)$ and the $N(0.2, 1)$ distribution, respectively. At least in the case without weighting these both have equally poor predictive densities by symmetry, under all notions of predictive ability which, like 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).

The top two panels of Fig. 4 display the rejection rates (at nominal significance level 5%) obtained with tests based on the weighted logarithmic scoring rule in (4), the censored normal likelihood in (7), the conditional likelihood in (8), and the censored likelihood in (9). As before, we use the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r)$ for the WL scoring rule score, while the CNL score is used with $\alpha = \Phi(r)$, where $\Phi(\cdot)$ represents the standard normal CDF. Rejection rates (based on 1000 replications) are shown as a function of the threshold value r, for sample size $n = 500$. For large values of r, when practically the entire distribution is taken into account by all four scoring rules, we can interpret the rejection rates as the actual sizes of the four tests. As expected, for large values of the threshold r the differences between the tests disappear. The results show that the rejection rates of all tests are close to the nominal value of 5% for large r. As noted above, whether or not the null hypothesis holds, depends on the measure of predictive ability (the mean score) as well as on the weight function. Therefore we generally cannot interpret the observed rejection rates for smaller values of r as type I error rates.

In case one of the competing density forecasts is correct, it will be the best predictive density regardless of the weight function so that it is possible to assess the power of the tests. The last four panels of Fig. 4 summarize the observed rejection rates (again for sample size $n = 500$, based on 1000 replications) against the standard normal and standardized $t(5)$ distribution, for data that were drawn from the standard normal (center row) and the standardized $t(5)$ distribution (bottom row). The left-hand-side panels correspond to rejections of the null of equal predictive ability against superior predictive ability of the standard normal density, and the right-hand-side panels against superior predictive ability of the standardized $t(5)$ distribution. Hence, the center left and bottom right panels report true power (rejections in favor of the correct density), while the center right and bottom left panels report spurious power (rejections in favor of the incorrect density). It can be observed that the power of some scoring rules (in particular the weighted logarithmic scoring rule) depends strongly on the threshold parameter r. This will be discussed in some more detail below. It also can be observed that our censored likelihood scoring rule, although it uses more information than the conditional likelihood scoring rule, does not give rise to a uniformly larger power. Clearly, having a power as large as possible is important for practical applications. However, spurious power is undesirable, as it may lead to a false indication of statistical evidence in favor of the wrong model. The results suggest that both the WL and the CNL scoring rules, and to a lesser extent our CSL scoring rule, suffer from spurious power. In particular for the i.i.d. standardized $t(5)$ process, the test based on the censored likelihood exhibits large spurious power for small values of the threshold parameter r.

[Figure 5 about here.]

We next investigate how these results change if the sample size increases. After all, the scores may have very skew distributions, which may lead to poor asymptotic approximations to finite sample distributions. In fact it is not even known if the standard asymptotic results apply, since no first and second moment conditions of the scoring rules have been verified. We consider this beyond the scope of the present paper. Results for sample size $n = 2000$ are reported in Fig. 5. Most of the observations made for sample size $n = 500$ still apply, apart from the fact that now the censored likelihood scoring rule does display uniformly larger power than the conditional likelihood scoring rule, and that the spurious rejection rate of the CSL scoring rule for small r has decreased, while that of the WL and CNL-type scores has increased. Although future analytical work should confirm this, it suggest that the spurious power associated with the censored likelihood scoring rule may disappear asymptotically, at least for fixed r.

[Figure 6 about here.]

As a final note on these Monte Carlo simulation results, we wish to briefly discuss the non-monotonous nature of the power curves of the test based on the WL scoring rule.

Fig. 6 shows the mean score $E[d_t^{wl}]$ as a function of the threshold r, for i.i.d. standard normal data, using relative scores of the standard normal versus the standardized $t(5)$ density. This mean was obtained by numerical integration. It can be observed that the mean changes sign several times, in exact accordance with the patterns in the center left panels of Figs 4 and Figs 5. Whenever the mean score is positive the associated test has high power, while it has high spurious power for negative mean scores.

[Figure 7 about here.]

As mentioned before, in some cases the central part of the distribution may of primary interest, for instance to policymakers aiming to keep inflation between 1 and 3%. To study how tests behave when they are being based on the central part of the distribution, we perform the same set of simulation experiments using the weight function $I_{(-r,r)}(y)$ in the various scoring rules. Only part of these results are included here; full details are available upon request. Fig. 7 shows rejection rates obtained for an i.i.d. standard normal process, 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 models has better predictive ability, for sample size $n = 2000$. The left panel shows rejection rates for testing the null hypothesis against better predictive performance of the (correct) $N(0,1)$ density, while the right panel shows the spurious power obtained when testing the null 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, while the spurious power for the tests based on the partial likelihood scores remains smaller than the nominal level (5%).

6 Empirical illustration

We compare the empirical performance of the different scoring rules in the context of the out-of-sample evaluation of two forecasting models for daily stock returns. We consider Standard and Poor's 500 index log-returns $X_t = \ln(P_t/P_{t-1})$, 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 May 31, 2007, giving a total of 6900 observations (source: Datastream).

The evaluation of the two models (specified below) is based on their out-of-sample predictive densities, using a rolling window scheme for parameter estimation, similar to Giacomini and White (2006). The estimation sample size m is set to $m = 1000$ observations. We focus on one-step-ahead density forecasts, such that the number of out-of-sample observations is equal to $n = 5900$. For comparing the accuracy of the density forecasts we use the Diebold-Mariano type test in based on the weighted logarithmic scoring rule in (4), the censored normal likelihood in (7), the conditional likelihood in (8), and the censored likelihood in (9). We concentrate on the left tail of the distribution that is of special interest in risk management bu using the threshold weight function $w(y_{t+1}) = \mathrm{I}(y_{t+1} \leq r_t)$ for the WL scoring rule score, while the CNL score is used with $\alpha = \Phi(r_t)$. We use two different time-varying thresholds r_t that are determined as the one-day Value-at-Risk estimates at the 95% and 99% level, assuming a normal distribution with constant mean and variance, which are set equal to the corresponding sample moments of the relevant estimation window comprising the previous $m = 1000$ observations. The same thresholds are used in the CL and CSL scoring rules.

For illustrative purposes we define two models in such a way that one of the models is superior to the other. This can be achieved by specifying a general, or an unrestricted model and restrict one of its parameters to a fixed value, arriving at a restricted model. Using the unrestricted model as a benchmark we expect no better predictive performance of the restricted model. The unrestricted model for the returns is specified as AR(5) process for the conditional mean and GARCH(1,1) process for the conditional variance. The model is based on a standardized Student-t distribution with ν degrees of freedom and unit variance, i.e.

$$\begin{aligned} X_t &= \mu_t + \varepsilon_t, \\ \mu_t &= \rho_0 + \sum_{\ell=1}^{5} \rho_\ell X_{t-\ell}, \\ \varepsilon_t &= \sqrt{h_t}\eta_t, \text{ with } \eta_t \sim \text{ i.i.d. stand. t}_\nu h_t \qquad &= c + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}. \end{aligned}$$

Note that the number of the degrees of freedom ν is a parameter that is to be estimated.

In the restricted model we fix the number of degrees of freedom to $\nu = 3$. Both the restricted model and the unrestricted model are estimated using the maximum likelihood criterion. The estimate of the parameter ν in the unrestricted model varies from 6 to 11.

We apply different scoring rules to select a better model on the basis of one-step-ahead out-of-sample prediction. The predictive score of the restricted model are subtracted from the scores of the unrestricted model. We expect this difference to be non-negative indicating better predictive abilities of the unrestricted model. Table 1 shows the resulting average of score differences and the standardized average score differences, that is average of scores differences divided by their sample variance. The sample variance is computed using HAC estimator to account for serial dependence. For both values of the threshold r the scoring rules of Amisano and Giacomini, and Berkowitz suggest superior predictive abilities of the restricted model, while the threshold weight function type I, i.e. the conditional likelihood and type II, i.e. the likelihood based on censored observations, suggest no improvement of the predictive abilities of the restricted model. The results of the latter two scoring rules are consistent with our intuition that a restrictive model should not have a higher predictive ability.

[Table 1 about here.]

The restricted model would be selected if we would use the scoring rules of Amisano and Giacomini and Berkowitz. We try to assess the consequences of selecting the restricted model in favour of the unrestricted model in terms of risk management. For these two models we compute popular measures of portfolio risk used by practitioners, i.e. the predicted one-day 5% and 1% VaR and the predicted one day 5% and 1% expected shortfall. To assess the accuracy of the VaR prediction, we compute the empirical frequency of observing a return lower then the predicted -VaR for a particular day. The procedure is based on out-of-sample observations using a rolling window scheme. In the case of the correct 5% and 1% VaR prediction we expect the frequency to be close the the nominal level of 0.05 and 0.01 respectively.

[Table 2 about here.]

Table 2, shows the average predicted values of the VaR, the empirical frequency of observing a return lower then the VaR, and the expected shortfall. The average VaR and the expected shortfall are higher for the restricted model. The empirical frequency of observing a return lower then the predicted -VaR is close to the nominal level for the unrestricted model and is lower than the nominal for the restricted model. This means that in the restricted model the extreme events occur less frequently than predicted. Thus, using a restricted models leads to over-prediction of risk. This, in turn, could result in suboptimal asset allocation within a portfolio.

7 Conclusions

In this paper we have developed scoring rules based on partial likelihood functions for comparing the out-of-sample accuracy of competing density forecasts. It was shown that these scoring rules are particularly useful when the main interest lies in measuring the accuracy of density forecasts over a specific region of the density, 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, as by construction they tend to favor density forecasts which have more probability mass in the region of interest, rendering the resulting tests biased towards such density forecasts. 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 give rise to spurious rejections due to the possible bias in favor of an incorrect model. The simulations results also showed that this phenomenon is strongly reduced for the new scoring rules, and where present, diminishes considerably 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 model selection 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 models, resulting in rather different 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). Testing skewness, kurtosis and normality in time series data. *Journal of Business and Economic Statistics*, **23,** 49–61.

Bao, Y., Lee, T.-H. and Saltoğlu, B. (2007). A test for density forecast comparison with applications to risk management. *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. Norwegian Computing Center, Oslo, Norway. Report number SAMBA/41/05.

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

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

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

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,** number 5, 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*, *Amsterdam* (eds G. Elliott, C.W.J. Granger and A. Timmermann), pp. 197–284. Elsevier.

Cox, D. (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 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.D. (1998). Evaluating density forecasts of inflation: The survey of professional forecasters. In *Cointegration, Causality, and Forecasting: A Festschrift in Honor of C.W.J. Granger*, *Oxford* (eds R.F. Engle and H. White), pp. 76–90. 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 macroeconometric modelling: An application to the UK economy. *Journal of the American Statistical Association*, **98,** 829–838.

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

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. (2008). 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 spot 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 forecast for GDP declines. 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., 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.

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 for relationships with limited dependent variables. *Econometrica*, **26,** 24–36.

White, H. (2000). A reality check for data snooping. *Econometrica*, **68,** 1097–1128.

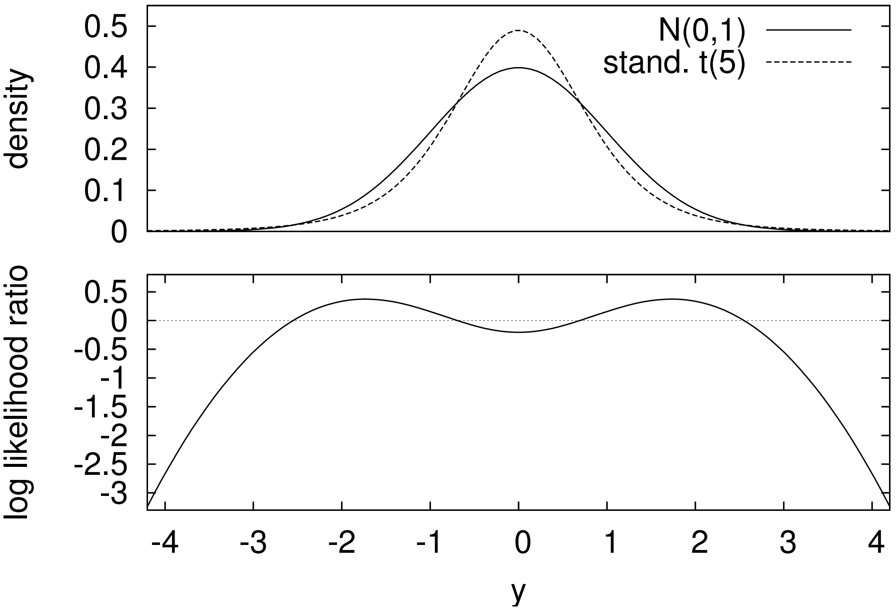

Figure 1: *Probability density functions of the $N(0, 1)$ distribution and standardized $t(5)$ distribution (upper panel) and corresponding log-likelihood scores of the $N(0, 1)$ density relative to the standardized $t(5)$ density (lower panel).*

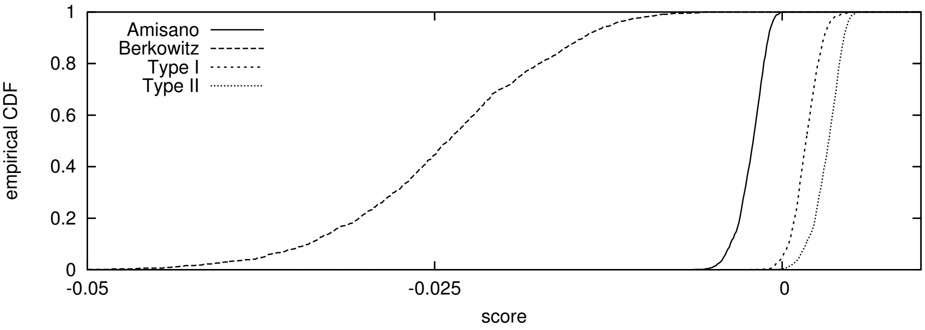

Figure 2: *Empirical CDFs of mean scores (sample size* 2000*) under the threshold weighting function* $w(y) = I(y \leq r)$*. Number of replications:* 1000*.*

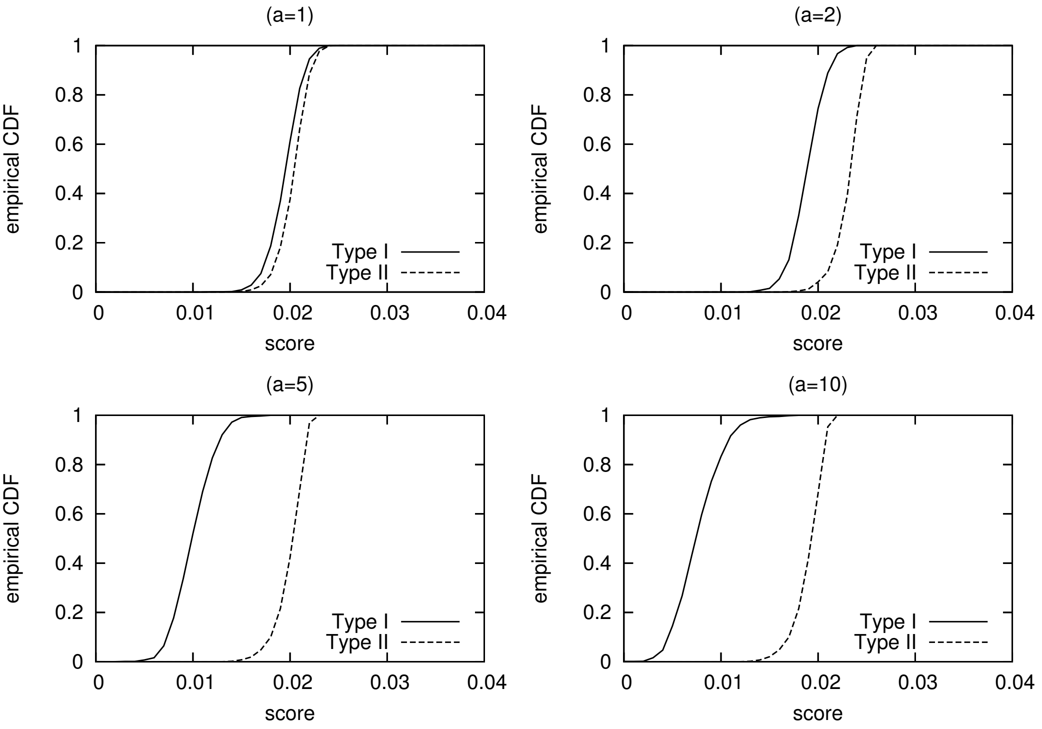

Figure 3: *Empirical CDFs of scores under the two smooth weighting schemes for various values of the slope parameter a of the weight function.*

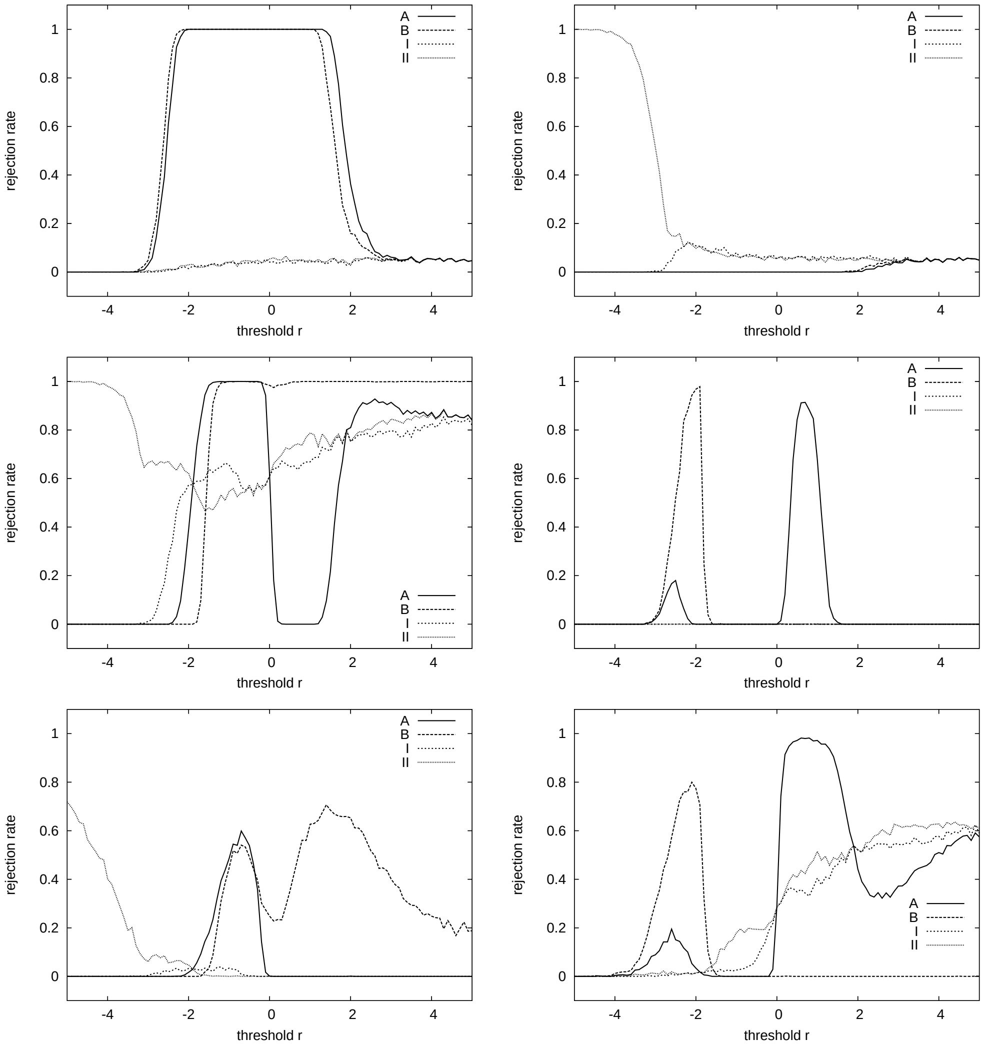

Figure 4: *One-sided rejection rates (nominal size* 5%*) based on sample size* 500, 1000 *replications, for testing the null of equal predictive ability. Top row: true DGP: i.i.d. standard normal, test against better predictive ability of the* $N(-0.2, 1)$ *distribution (top left) and of the* $N(0.2, 1)$ *distribution right (top right). Center row: true DGP: i.i.d. standard normal, test against better predictive ability of the standard normal distribution (center left) and of the* $t(5)$ *distribution (center right). Bottom row: true DGP i.i.d. standardized* $t(5)$*, test against better predictive ability of the standard normal distribution (bottom left) and of the standardized* $t(5)$ *distribution (bottom right).*

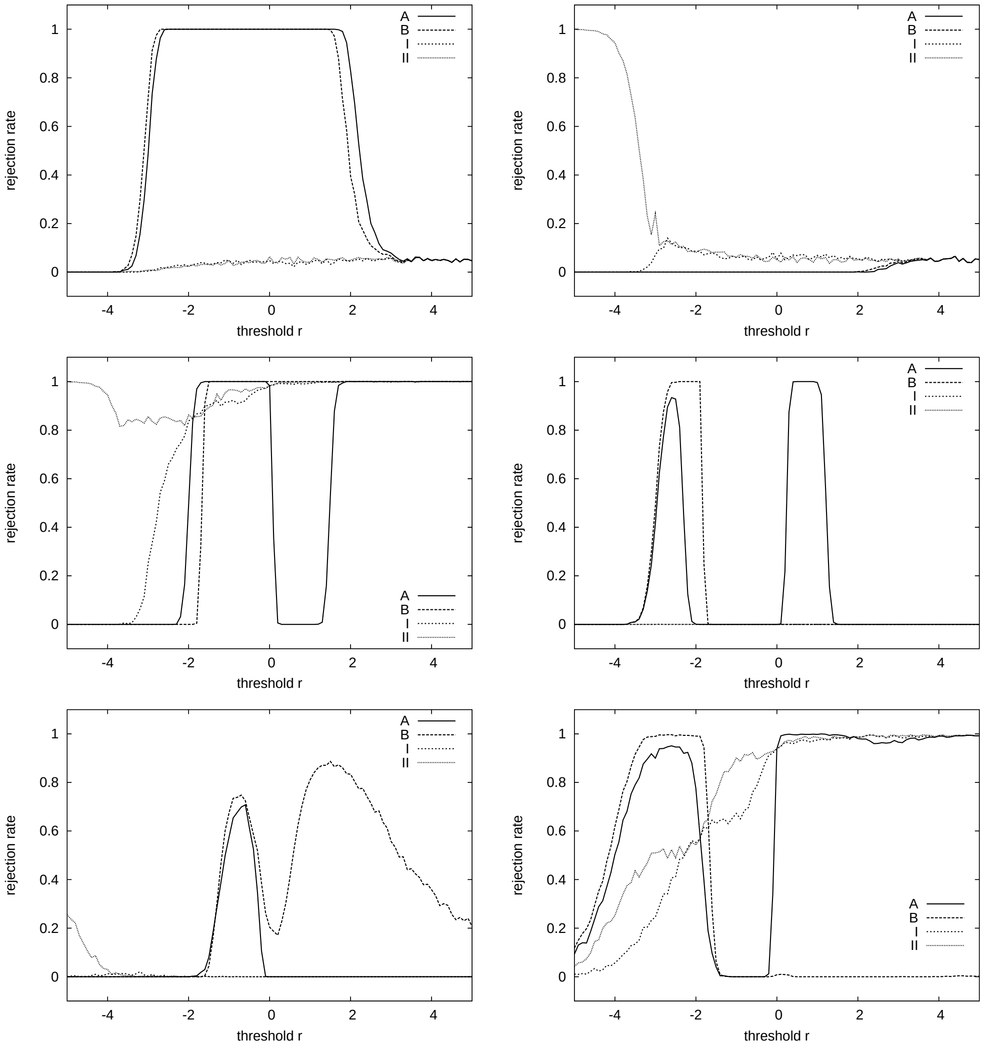

Figure 5: *Sample size* 2000. *Other details as for Fig. 4.*

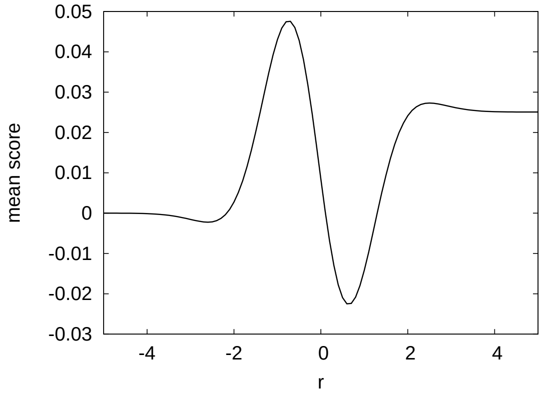

Figure 6: *Mean WL score $E[d_t^{wl}]$ as a function of the threshold value r, for the standard normal DGP.*

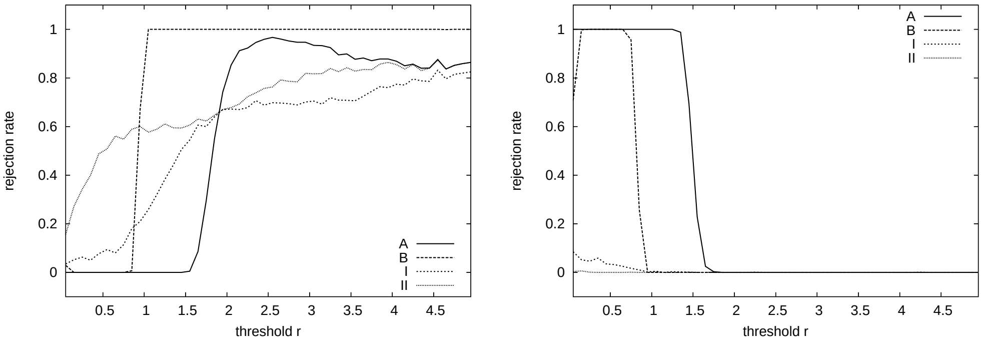

Figure 7: *Sample size* 2000*. Power (left panel) and spurious power (right panel) for an i.i.d.* $N(0,1)$ *process with competing densities* $N(0,1)$ *and standardized* $t(5)$*, with weight function* $I_{(-r,r)}(y)$*, as a function of* r*.*

scoring rule	$r_t = \Phi^{-1}(0.05)$		$r = \Phi^{-1}(0.01)$	
	$\bar{\delta}$ scores	standardized $\bar{\delta}$	$\bar{\delta}$ scores	standardized $\bar{\delta}$
WL	-0.0058	-4.1105	-0.0066	-4.4724
CNL	-0.0199	-8.4132	-0.0202	-8.5945
CL	0.0011	0.8569	-0.0004	-0.3549
CSL	0.0026	1.8118	0.0017	1.2457

Table 1: *Performance of different scoring rules assessing the predictive abilities of the unrestricted model over the restricted model for different values of threshold r computed as inverse of the normal CDF, Φ, parameters of which are estimated from the data.*

Measure	Unrestricted model	Restricted model
Average 5% VaR	0.0148	0.0160
Frequency of $(X_t < -5\% \text{ VaR}_t)$	0.0524	0.0425
Average 5% Expected Shortfall	0.0206	0.0263
Average 1% VaR	0.0238	0.0308
Frequency of $(X_t < -1\% \text{ VaR}_t)$	0.0115	0.0049
Average 1% Expected Shortfall	0.0303	0.0475

Table 2: *Predicted risk measures using unrestricted and restricted model.*