Comparing the accuracy of density forecasts from competing GARCH models

Shamiri, Ahmed and Shaari, Abu Hassan and Isa, Zaidi

National Univeristy of Malaysia

2008

Comparing the accuracy of density forecasts from competing GARCH models

A.Shamiri
Faculty of Science and Technology
Universiti Kebangsaan Malaysia

Abu Hassan Shaari Mohd Nor
School of Economics Studies
Faculty of Economics and Business
Universiti Kebangsaan Malaysia

Zaidi Isa
Faculty of Science and Technology
Universiti Kebangsaan Malaysia

Second draft Jun 2008

Abstract:

In this paper we introduce an analyzing procedure using the Kullback-Leibler information criteria (KLIC) as a statistical tool to evaluate and compare the predictive abilities of possibly misspecified density forecast models. The main advantage of this statistical tool is that we use the censored likelihood functions to compute the tail minimum of the KLIC, to compare the performance of a density forecast models in the tails. Use of KLIC is practically attractive as well as convenient, given its equivalent of the widely used LR test. We include an illustrative simulation to compare a set of distributions, including symmetric and asymmetric distribution, and a family of GARCH volatility models. Our results on simulated data show that the choice of the conditional distribution appears to be a more dominant factor in determining the adequacy and accuracy (quality) of density forecasts than the choice of volatility model.

Key words: D*ensity, conditional distribution, forecast accuracy, GARCH, KLIC*

1 Introduction

It is often argued that forecasts should be evaluated in an explicit decision context, that is, in terms of econometrics the consequences that would have resulted from using the forecasts to solve a sequence of decision problems. The incorporation of a specific loss function into the evaluation process would focus attention on the features of interest to the forecast user, perhaps

also showing the optimality of a particular forecast. In finance there is usually a more obvious profit and loss criterion, and there is a long tradition of forecast evaluation in the context of investment performance. This extends to volatility models but not yet to density forecasts (West et al 1993). Here there are relatively few results based on explicit loss functions. The basic result that a correct forecast is optimal regardless of the form of the loss function is extended from point forecasts to event probability forecasts by Granger and Pesaran (1996) and to density forecasts by Diebold et al. (1998). The latter authors also show that there is no ranking of sub-optimal density forecasts that holds for all loss functions. The problem of the choice of forecast would require the use of loss functions defined over the distance between forecast and actual densities. Therefore, the objective of density forecasters is to get close to the correct density in some sense, and practical evaluations are based on the same idea.

The issues described in those working papers stem from the fact that the prediction produced by a density forecasting model can rarely be compared to the true generating distribution in real world problems. Instead, only a single instance of the generating distribution, the actual outcome, is available to the forecaster to optimize and evaluate their model. Conventional diagnostics for evaluating point predictions such as the root-mean-squared error (RMSE) and others fail to assess probabilistic predictions. Furthermore, the ranking of different density forecasting models is difficult because a ranking depends on the loss function of the user (Diebold et al. 1998). For example, a user's loss function could be non-linear and/or asymmetric. In such cases the mean and variance of the forecast densities are not sufficient to rank predictive models. For example, a user with an asymmetric loss function would be particularly affected by the accuracy of a model's predictions of the skew in the conditional densities. Diebold et al. (1999) suggests that the problem of ranking density forecasts can be solved by assuming that the correct density is always preferred to an incorrect density forecast. Using the true density as a point of reference it is possible to rank densities relative to the true densities to determine the best models to use. Therefore, in the absence of a well defined loss function, the best model is the one that approximates the true density as well as possible. Diebold et al. (1998) go on to suggest the probability integral transform (PIT) as a suitable means of evaluating density forecasts in this way.

The research on evaluating each density forecast model has been very versatile since the seminal paper of Diebold et al. (1998), however there has been much less effort in comparing alternative density forecast models. Considering the recent empirical evidence on volatility clustering and asymmetry and heavy-tailed in financial return series, we believe that using a formal test in the context of density forecasts of a given model compared with alternative

distribution and volatility specifications, will contribute to the existing literature. Despite the burgeoning interest in and evaluation of volatility forecasts, a clear consensus on which distribution and/or volatility model specification to use has not yet been reached even for finance practitioners and risk professionals. As argued in (Poon & Granger 2003), "most of the volatility forecasting studies do not produce very conclusive results because only a subset of alternative models are compared, with a potential bias towards the method developed by the authors". It is further claimed that, lack of a uniform forecast evaluation technique makes volatility forecasting a difficult task. Being able to choose the most suitable volatility and distribution specifications is a more demanding task. In this paper we demonstrate that this gap can be filled by a rigorous density forecast comparison methodology.

Therefore the main aim of this paper is to use and utilize the Kullback-Leibler Information Criterion (KLIC) as a unified test to evaluate, compare and assess which volatility model and/or distribution are statistically more appropriate to mimic the time series behavior of a return series. This generality follows from appreciation, that the (Berkowitz 2001) Likelihood Ratio (LR) test can be related to the KLIC (Bao et al. 2006), a well-respected measure of "distance" between two densities. As the true density is unknown, devising an equivalent LR evaluation test based on the PIT is computationally convenient. An extension to the aim of this paper is to modify the proposed test to compare the predictive abilities of alternative density forecast models in the tail area. For this purpose, a tail minimum KLIC discrepancy measure based on the censored likelihoods is used as a forecast loss function in the framework of (White 2000) and (Hansen 2001) reality check.

The structure of the remainder of this paper is as follows. We review the statistical evaluation of individual density forecasts using the PITs in section 2 and develop the distance measure based on the KLIC for candidate models in section 3. In section 4 we explain and discuss how the Berkowitz LR test can be re-interpreted as a test of whether the KLIC equals zero. Section 5 shows how the KLIC can be used to compare statistically the accuracy of two competing density forecasts applied to simulated data. Section 6 concludes the paper.

2 Probability Integral Transform

Statistical evaluations of real time density forecasts have recently begun to appear, although the key device, the probability integral transform, has a long history. The literature usually cites (Rosenblatt 1952) for the basic result, and the approach features in several expositions from different points of view, such as (Dawid 1984). For a sample of n one-step-ahead forecasts and the corresponding outcomes, the probability integral transform of the realized variables with respect to the forecast densities is defined as

$$z_t = \int_{-\infty}^{x_t} f_t(u)du = F_t(x_t); \quad (t = 1,...,T) \tag{1}$$

It is well known that if $f_t(.)$ coincides with the true density $g_t(.)$, then the sequence $\{z_t\}_{t=1}$ is *iid* U [0,1]. If the transformed time series $\{z_t\}$ is not *iid* U[0,1], then $f_t(.)$ is not an optimal density forecast model (Diebold et al. 1999). To describes the distribution, $q_t(z_t)$, of the probability integral transform. Let $g_t(x_t)$ be the true density of x_t, and let $f_t(x_t)$ be a density forecast of x_t, and let z_t be the probability integral transform of x_t with respect to $f_t(x_t)$. Then assuming that $\partial F_t^{-1}(z_t)/\partial z_t$ is continuous and non-zero over the support of x_t, z_t has unit interval with density;

$$q_t(z_t) = \left|\frac{\partial F_t^{-1}(z_t)}{\partial z_t}\right| g_t\left(F_t^{-1}(z_t)\right) = \frac{g_t\left(F_t^{-1}(z_t)\right)}{f_t\left(F_t^{-1}(z_t)\right)}$$

where $f_t(x_t) = \dfrac{\partial F_t(x_t)}{\partial x_t}$ and $x_t = F_t^{-1}(z_t)$. Therefore, in particular, a key fact; if $f_t(x_t) = g_t(x_t)$, then $z_t \in (0,1)$ and $q_t(z_t)$ is simply the U(0,1) density. This idea dates at least to Rosenblatt (1952). Therefore, a natural test of optimality of a density forecast model is to test the iid U[1,0] properties of the series $\{z_t\}$. Our task, however, is not to evaluate a single model, but to compare a battery of competing models. Since our objective, is to compare the out-of-sample predictive abilities among competing density forecast models. Suppose that, there are *l+1* models (k=0,1,…,*l*) in a set of competing models, possibly misspecified. To establish the notation with the model index k, let the density forecast model k (k=0,1,…,*l*) be denoted by $f_{k,t}(x)$. We used to sub-samples $\{z_t\}_{t=1}^{R}$ and $\{z_t\}_{t=R+1}^{T}$, the first sample to estimate the unknown parameters and the second sub-sample to check if the transformed PITs are iid N(0,1). That is, we first construct

$$z_{k,t} = \int_{-\infty}^{x_t} f_{k,t}(u)du = F_{k,t}(x_t); \quad (t = R+1,...,T) \tag{2}$$

where the inverse normal transform of the PIT is

$$z_{k,t}^{*} = \Phi^{-1} z_{k,t} \tag{3}$$

and $\Phi(.)$ is the CDF of the standard normal. In other words, testing the departure of $\left\{z_{k,t}^{*}\right\}_{t=1}^{T}$ from iid N(0,1) is equivalent to testing the distance of the forecasted density from the true –unknown- density. Consequently various single and joint test of U(0,1), N(0,1) and iid have been employed in empirical studies. These include Kolmogorov-Smirnov, Anderson-Darling and others as shown in section 5.

3 The Distance Measure

The test for adequacy of a postulated distribution may be appropriately measured by Kullback Information Criterion (Kullback & Leibler 1951) divergence measure between two conditional densities, $D(g;f)=E\left[\ln g_t(x_t)-\ln f_t(x_t)\right]$, where the expectation is with respect to the true distribution. Following (Vuong 1989), we define the distance between a model and the true density as the minimum of KLCI

$$D_{KLCI}(g;f)=\int g_t(x_t)\ln\left\{\frac{g_t(x_t)}{f_t(x_t)}\right\}dx \text{ or} \tag{4}$$

$$D_{KLCI}(g;f)=E\left[\ln g_t(x_t)-\ln f_t(x_t)\right] \tag{5}$$

The smaller this distance the closer the density forecast is to the true density; $D(g;f)=0$ if and only if $g_t(x_t)=f_t(x_t)$.However, $D(g;f)$ is generally unknown, since we can not observe $g(.)$ and hence the expectation, it can be consistently estimated by

$$D_{KLCI}(g;f)=\frac{1}{T}\sum_{t=1}^{T}\left[\ln g_t(x_t)-\ln f_t(x_t)\right] \tag{6}$$

But we still do not know $g(.)$. The task of determining whether $g_t(x_t)=f_t(x_t)$ appear difficult, perhaps hopeless, because $g(.)$ is never observed, even after the fact. Moreover, and importantly, the true density $g(.)$ may exhibit structural change, as indicated by its time subscript. For this, we utilize the probability integral transform (PIT) of the actual realizations of the process with respect to the model's density forecast and hence to compare possibly misspecified models in terms of their distance to the true model.

4 Relating LR test to the KLIC

Re-interpreting the Berkowitz LR test as a test of whether the KLIC 'distance' between the true (unknown) density and the forecast density equals zero. Note the following equivalence (Berkowitz 2001):

$$\ln\left[g_t(x_t)/f_{k,t}(x_t)\right]=\ln\left[p_t(z_{k,t}^{*})/\phi_t(z_{k,t}^{*})\right] \tag{7}$$

where $p(.)$ is the unknown density of $z^*_{k,t}$, $\phi(.)$ is the standard normal density. In other words, testing the departure of $\{z^*_{k,t}\}^T_{t=1}$ from iid N(0,1) is equivalent to testing the distance of the forecasted density from the true –unknown- density $g_t(x_t)$. Along with (Bao et al. 2006), we believe that testing whether $p(.)$ is *iid* N(0,1) is both more convenient and more sensible than testing the distance between $g_t(x_t)$ and $f_{k,t}(x_t)$ since we do not know $g_t(x_t)$. To test the null hypothesis that $g_t(x_t) = f_{k,t}(x_t)$ we exploit the theoretical framework of West (1996) and White (2000). Consider the loss differential

$$d_t = \left[\ln g_t(x_t) - \ln f_{k,t}(x_t)\right] = \left[\ln p_t(z^*_{k,t}) - \ln \phi(z^*_{k,t})\right]; (t = 1,...,T) \tag{8}$$

the null hypothesis of the density forecast being correctly specified is then

$$H_0 : E(d_t) = 0 \Rightarrow D_{KLIC} = 0 \tag{9}$$

The sample mean $\bar{d}$ is defined as:

$$\bar{d} = D_{KLIC} = 1/T \sum_{t=1}^{T}\left[\ln p_t(z^*_{k,t}) - \ln \phi(z^*_{k,t})\right] \tag{10}$$

To test the hypothesis about $\bar{d}$ by a suitable central limit theorem we have the limiting distribution $\sqrt{T}\left(\bar{d} - E(d_t)\right) \to N(0,\Omega)$ where in general expression for the covariance matrix Ω is rather complicated because it allows for parameter uncertainty (West 1996). However, ignoring parameter uncertainty (which asymptotically we can as the sample size used to estimate the model's parameter grows relative to T; West (1996, Theorem 4.1)) Ω reduces to the long run covariance matrix associated with d_t or 2π the spectral density of $\left(d - E(d_t)\right)$ at frequency zero as is the case showed by (Diebold & Mariano 1995). This long run covariance matrix S_d is defined as $S_d = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j$, where $\gamma_j = E(d_t d_{t-j})$. Alternatively, to this asymptotic test, White (2000) suggested and justified using "bootstrap realty check" , a small sample test based on the bootstrap is called the "realty check p-value" for data snooping. This would involve re-sampling the test statistic $\bar{d} = D_{KLIC}$ by creating R bootstrap samples from $\{d_t\}^T_{t=1}$ accounting for dependence by using the so-called stationary bootstrap that resample using blocks of random length.

The test statistics D_{KLIC} is proportional to the LR test of Berkowitz (2001), assuming normality of ε_t. In terms of(10), we follow Berkowitz (2001) by specifying $\left\{z_{k,t}^*\right\}_{t=1}^{T}$ as an AR(1) process

$$z_t^* = \rho Z_{t-1}^* + \varepsilon_t \tag{11}$$

where $Var(\varepsilon_t) = \sigma^2$, ρ is a vector of parameters, and ε_t is iid distributed. In Berkowitz (2001), ε_t is assumed to be normally distributed. Actually, if we specify $p(.)$ such as iid and normal, then our comparison based on the distance measure (10) will suffer the same criticism of the LR test of Berkowitz, as pointed out by (Clements & Smith 2000; Bao et al. 2006). A remedy to such criticism is to consider more general forms for $p_t(z_{k,t}^*)$. Bao et al.(2006) suggested the use of the seminonparametric (SNP) density of (Gallant & Nychka 1987) for ε_t in the AR process of the order *K*

$$p_t(\varepsilon_t;\theta) = \frac{\left[\sum_{k=0}^{K} R_K \varepsilon_t^K\right]^2 \phi(\varepsilon_t)}{\int_{-\infty}^{+\infty}\left[\sum_{k=0}^{K} R_K z_t^K\right]^2 \phi(z_t)\,dz} \tag{12}$$

A change of variables using the location-scale transformation, $y = R\varepsilon + \mu$, where R is an upper triangular matrix and μ is an M-vector. The change of variable formula applied to the location-scale transformation, the density of $z_{k,t}^*$ is

$$p_t(z_{k,t}^*) = \frac{p_t\left[(z_{k,t}^* - \rho Z_{t-1}^*)/\sigma\right]}{\sigma} \tag{13}$$

thus, the estimated minimum KLCI divergence measure is

$$D_{KLIC} = \frac{1}{T}\sum_{t=1}^{T}\left[\ln\frac{p_t\left[(z_{k,t}^* - \rho Z_{t-1}^*)/\sigma\right]}{\sigma} - \ln\phi(z_{k,t}^*)\right] \tag{14}$$

The LR test statistics of the adequacy of the density forecast model $f_{k,t}(.)$ in (Berkowitz 2001) is simply the above formula with $p(.) = \phi(.)$. Rather than evaluating the performance of the whole density we can also evaluate in any regions of particular interest. Risk managers and other practitioner in finance care more about the extreme values in the lower tail (larger loss) than about the values in other regions of the

distribution (small loss/gain). Therefore, a density forecast model that accurately predicts tail events, is of more interest in finance. For a complete evaluation of these forecasts, we need to integrate this approach with testing procedures applicable to the tails of the distribution. To do so, we can easily modify D_{KLIC} distance measure for the tail parts. We focus on the lower tails only. Therefore, we define

$$z_{k,t}^{*\tau} = \begin{cases} \Phi^{-1}(\alpha) \equiv \tau & \text{if } z_{k,t}^{*} \geq \tau \\ z_{k,t}^{*\tau} & \text{if } z_{k,t}^{*} < \tau \end{cases} \tag{15}$$

Let I(.) denote and indicator function that takes (1) if its argument is true and 0 otherwise, the distribution function for $z_{k,t}^{*\tau}$ can be constructed as

$$p_t^{\tau}(z_{k,t}^{*\tau}) = \left[1 - p\left(\frac{\tau - \rho Z_{t-1}^{*}}{\sigma}\right)\right]^{I(z_{k,t}^{*} \geq \tau)} \left[\frac{p\left[(z_{k,t}^{*} - \rho Z_{t-1}^{*})/\sigma\right]}{\sigma}\right]^{I(z_{k,t}^{*} < \tau)} \tag{16}$$

Therefore, the teal minimum D_{KLIC} divergence can be estimated analogously

$$D_{KLIC}^{\tau} = \frac{1}{T}\sum_{t=1}^{T}\left[\ln p_t^{\tau}(z_{k,t}^{*\tau}) - \ln \phi^{\tau}(z_{k,t}^{*\tau})\right] \tag{17}$$

where $\phi^{\tau}(z_{k,t}^{*\tau}) = \left[1 - \Phi(\tau)\right]^{1(z_{k,t}^{*} \geq \tau)} \left[\phi(z_{k,t}^{*})\right]^{1(z_{k,t}^{*} < \tau)}$

A closely related approach to compare density forecasts statistically have been proposed by (Corradi & Swanson 2004). Their approach is to compare the cumulative distribution function (CDF) of the unknown density to the empirical distribution (EDF) of the known density. The authors showed that, the distance of the unknown density to the true density is measured by the mean square error of the CDF and the EDF, integrated out over the domain of the series. But rather than rely on the PIT's –our case- they estimate the true density or CDF empirically. In our opinion, benefits from based on the PIT's and not relying on estimation of $g_t(x_t)$. The relationship in (7) enable us to map the misspecification of a model to the deviation of $\left\{z_{k,t}^{*}\right\}_{t=1}^{T}$ from iid N(0,1). The equivalence relationship (7) also tells us that LR statistics based on the transformed PIT's is actually an estimate of the KLIC divergence measure between the model and the true distribution. In additional, we believe these tests – over a specific region- which forecast will be more accurate at a future date rather than, as with the unconditional tests.

5 Applications to Simulated Data

Before proceeding to apply our density forecast evaluation methods to real data, it is useful to examine their efficacy on simulated data, for which we know the true data generating process.

We propose a simple and general simulation approach that examine both model adequacy and forecast accuracy. Returns are generated that match the statistical features of a financial asset, hence we examine data simulated from a realistic fat-tailed distribution with GARCH process designed to mimic high-frequency financial asset return data (Bollerslev 1986). Specifically, we use a GARCH(1,1) data generating process, the conditional density of which is a standardized Student's-t with six degrees of freedom,

$$y_t = \left(h_t\right)^{1/2}\left[v/(v-2)\right]^{-1/2} t_v$$

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}$$

We create a simulated series of length 8000, chosen to mimic the sample sizes typical of high-frequency financial data, and we choose the parameters in accordance with those typically obtained when fitting GARCH models to high-frequency financial asset returns. Given starting values $\alpha_0 = 0.01$, $\alpha_1 = 0.13$, $\beta_1 = 0.86$, the simulated data is plotted in Figure 1. The persistence in conditional volatility is visually obvious. Then we estimate the GARCH parameters using the standard GARCH optimization technique on the first 4000 observations, the remaining 4000 observation are used for out-of-sample forecast.

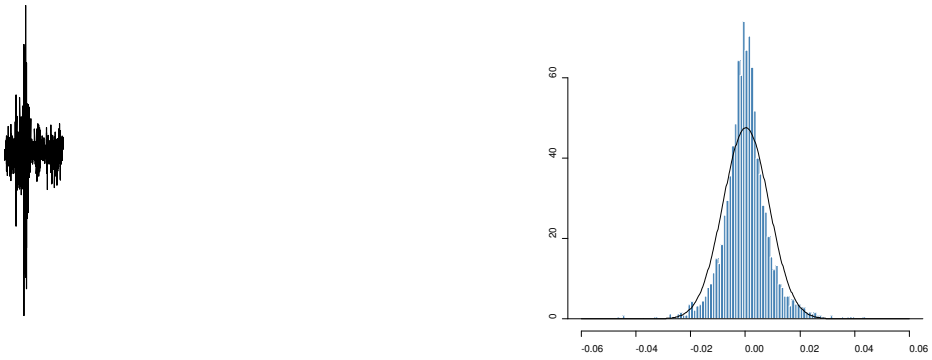
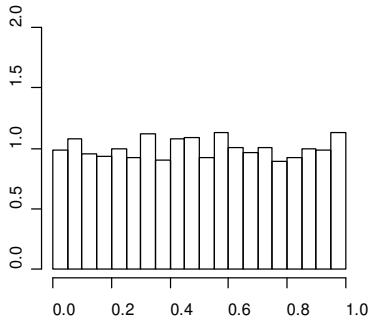


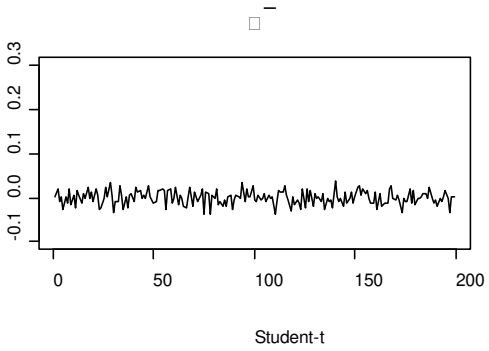
Figure 1. Simulated GARCH(1,1)-t, the gray region indicates
data points used for out-of-sample forecast evaluation.

Through this section we demonstrate the utilization of the PITs and examine the usefulness of density forecast evaluation methods in assessing three density forecasts with two volatility models particularly GARCH(1,1) and EGARCH(1,1).

Then, we evaluate forecasts that are based on the correctly specified volatility model GARCH estimated under three assumptions; (a) incorrect assumption that the conditional density is

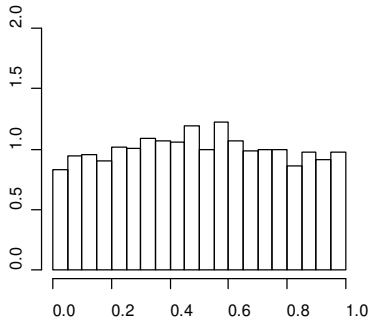
Normal, (b) correct assumption that the conditional density is fat-tailed (student-t) and (c) incorrect assumption that the conditional density is fat and skewed (Skewed-t). Figure 2 shows the resulting PIT histograms form forecasting with GARCH(1,1) combined with three distributions and randomly sampled from the data generating distribution $t_{(6)}$ were used to determine each PIT histogram. First we evaluate GARCH-Normal, the histograms display peaks at either end and a hump in the middle, they have the butterfly shape, indicating the departure form U(0,1). To evaluate whether z is iid, the correlograms are obtained, figure 3 indicate that the N(0,1) forecasts shows no evidence of neglected conditional volatility, as expected, that the conditional GARCH-N(0,1) model delivers consistent estimates of the conditional variance parameters, in spite of the fact that the conditional density is misspecified (Bollerslev & Wooldridge, 1992).

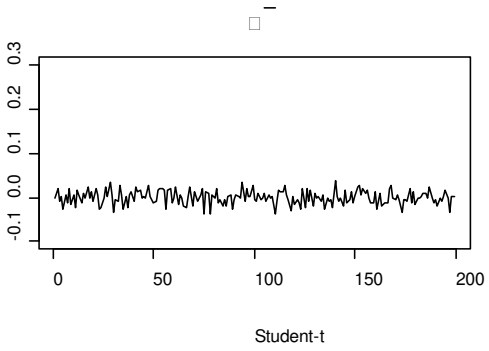



-0.1
0.0
0.1
0.2
0.3
0
50
100
150
200
Student-t

that GARCH is an adequate volatility model and student-t makes an accurate density forecast model.

In additional to modeling the simulated data with GARCH(1,1) model, we suggest to model the data with a more sophisticated misspecified volatility model. The EGARCH model is well known, and extensively used in the literature, therefore, we use EGARCH(1,1) model as a misspecified volatility model combined with three density forecasts particularly, Normal, student-t and Skewed-t. The PIT histogram for the density forecasts are shown in Figure 4. The results in Figure 4 with EGARCH model are similar to the results drawn from GARCH in spite of the fact that the volatility model is misspecified. The correlograms in Figure 5 also remain good.




-0.1
0.0
0.1
0.2
0.3
0
50
100
150
200
Student-t

whether the maximum difference between the empirical CDF of the $\{z_{k,t}\}_{t=1}^{T}$ are significantly different from the theoretical uniform CDF. The Kolmogorov-Smirnov D test statistic and the other test offers significant evidence against the null hypothesis of uniformity. Moreover, GARCH-t with a smallest A^2 score of (0.697) has comparable statistical consistency with the other density forecast models. Anderson-Darling test is known as a quadratic test because it is based upon a weighted square of the vertical distance between the empirical stepwise density function and target cumulative density function. It differs from the well known Kolmogorov-Smirnov test, which finds the maximum vertical distance between the empirical and target density.

Table 1. Uniform distributional tests $U(0,1)$ of probability integral transform (z) from GARCH(1,1) model

Test Statistic	D	A^2	W^2	U^2	V
Normal	0.084406	72.49528	12.58512	12.57397	0.164574
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Student-t	0.010265	0.696606	0.075248	0.069698	0.018855
	(0.7913)	(0.5617)	(0.7198)	(0.4972)	(0.5416)
Skewed-t	0.159017	19647.37	32.23230	21.45346	0.243267
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

While, the results in Table 2, on the misspecified volatility model (EGARCH) has a strong rejection of the null hypothesis for both Normal and Skewed-t density forecast, there are mixture of results with Student-t density forecast. The failure to reject uniformity for the EGARCH specification was probably due to a lack of statistical power, rather than to the good fit of the density forecast to the actual density. However, GARCH-t still holds smallest A^2 score of (0.697) which make it statistically consistence compare to (3.814) with EGARCH-t.

Table 2. Uniform distributional tests $U(0,1)$ of probability integral transform (z) from EGARCH(1,1) model

Test Statistic	D	A^2	W^2	U^2	V
Normal	0.250635	444.7028	95.45883	31.86551	0.252287
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Student-t	0.021530	3.814787	0.606202	0.599987	0.041103
	(0.0483)	(0.0107)	(0.0216)	(0.0000)	(0.0001)
Skewed-t	0.408080	445868.5	257.9255	49.48348	0.408080
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Therefore, the issue of model selection is a critical one. However, a potential problem arises in the application of PIT approach; it is possible that an incorrect density model could have a uniform z series. (Hamill 2000), a uniform z series is a necessary but not sufficient criterion for determining that the model is reliable. And it becomes more problematic when using more than one volatility model combined with variant density forecasts. Therefore, a visualize graph provides a good indicator for uniformity test in the case of single volatility model combined with variant density forecasts. On the other hand, goodness-of-fit tests for uniformity could be miss-leading due to their weakness and lack of statistical power. In this case the accuracy distance measure (KLIC) should highlight the existence of an incorrectly specified model. However, it is important to be aware of this point. By claiming that most of the testing procedures outlined above are not powerful enough, we first apply a transformation to the PIT series and then obtain the distance measure with White's (2000) and Hansen (2001) reality check-*p*-values as shown in Table 3.

Table 3. The D_{KLIC} distance measure and the Reality Check-*p*- values

Panel A

GARCH model	Normal	Student-t	Skewed-t
100%	0.02764370	0.00006246	0.002468316
	0.1001	1.0000	0.3068
	0.0000	1.0000	0.2112
10%	0.017791262	0.001133953	0.0170185444
	0.0000	0.9492	0.5109
	0.0000	0.9067	0.3702
5%	0.012090814	0.000750008	0.00110854240
	0.0000	1.0000	0.8709
	0.0000	0.9984	0.7290
1%	0.008046276	0.002452157	0.0024593474
	0.0000	0.9021	0.8999
	0.0000	0.8112	0.8099

Panel B

EGARCH model	Normal	Student-t	Skewed-t
100%	0.040761623	0.0000946779	0.002468316
	0.0020	1.0000	0.2968
	0.0000	0.9653	0.0152
10%	0.015568908	0.001169720	0.017043230
	0.0004	0.8952	0.5002
	0.0000	0.8089	0.3028

5%	0.013651425	0.0007609941	0.0011095617
	0.0000	0.9900	0.7987
	0.0000	0.9209	0.7091
1%	0.009330717	0.002649095	0.002597212
	0.0000	0.9001	0.7979
	0.0000	0.7864	0.6942

Critical values:
at 10% (LR=19.81, D_{KLIC} = 0.00247625)
at 5% (LR=22.36, D_{KLIC} =0.002795)
at 1% (LR=27.69, D_{KLIC} =0.00346125)

To evaluate the adequacy of each model in Table 3 we use KLIC measure to compare the different density forecast models. Noting that KLIC loss D_{KLIC} is related to the LR, we may use the LR test statistic to assess the adequacy of each single model. Therefore, to evaluate the adequacy of each density forecast, the LR is to test the null hypothesis that $\{z_{k,t}\}_{t=1}^{T}$ follows iid N(0,1). The critical values for the LR statistic are 19.81 (at 10%), 22.36 (at 5%) and 27.69(at 1%). Thus, the critical values for D_{KLIC} are the critical values of LR divided by 2(n-3), which are 0.0025 (at 10%), 0.0028 (at 5%) and 0.0035 (at 1%) for n=4000. Therefore, if the value of D_{KLIC} reported in the table is greater than say (0.0028), then the model can be rejected as an adequate density forecast model. Note that, a smaller value of D_{KLIC} the first number in each cell indicates a lower sample loss and hence a better density forecast model from a pair of volatility model and distribution. While, the larger reality check-*p*- value indicate the better density forecast model corresponding to the cell, as we fail to reject the null hypothesis that the other 5 competing models is no better than this model. In general, a low D_{KLIC} should parallel with high reality check-*p*- value, however this relationship is not perfect since the testing not only depends on the point value of the loss differential, but also it depends on the variance.

In Table 3 the results for 100%, 10% , 5% and 1% tails of the simulated data are presented in Panels A and B. As expected that the best density forecast model is GARCH-t and out perform the rest of the density models, for 100% distribution GARCH-t has the lowest D_{KLIC} = 0.00006246 with White's *p*-value=1 and Hansen adjusted *p*-value=1.. Turning to 10% tail, Student-t with GARCH specification produces the best performance with D_{KLIC} = 0.001133953. Similar results are obtained for 5% and 1% tail distributions, which support the true

data generating process of GARCH(1,1)-t. None of the other distribution and volatility model produce adequate density forecast

6 Conclusions

In recent years, there has been increasing concern among researchers, practitioners and regulators over how to evaluate models of financial risk. This paper has analyzed and used the Kullback-Leibler Information Criterion (KLIC) as a unified statistical tool to evaluate, and compare density forecasts. Computation of the KLIC is facilitated by exploiting its relationship with the well-known Berkowitz LR test for the evaluation of individual density forecasts based on the PITs. To compare the performance of density forecast models in the tails, we also use a censored LR statistics to estimate the tail minimum D_{KLIC} .

The testing framework on the simulated data is flexible and intuitive. Moreover, the D_{KLIC} testing approach appears to deliver extremely good power. Our findings based on the simulated data confirm that successful density forecast depends much more heavily on the choice of distributional model than the choice of volatility model.

Reference:

BAO, Y., T. H. LEE and B. SALTOGLU (2006): Comparing Density Forecast Models. *Journal of Forecasting 26(3), 203-225*.

BERKOWITZ, J. (2001): Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics* 19(4), 465-474.

BOLLERSLEV, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31(3), 307–327.

BOLLERSLEV, T. and J. WOOLDRIDGE (1992): Quasi maximum likelihood estimation and inference in dynamic models with time varying covariances *Econometric Reviews* 5.

CLEMENTS, M. P. and J. SMITH (2000): Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting* 19(4), 255-276.

CORRADI, V. and N. R. SWANSON (2004): "Predictive Density Evaluation, in the Handbook of Economic Forecasting, G. Elliott, CWJ Granger and A. Timmermann ed. s," in.: North Holland Press, Amsterdam.

D'AGOSTINO, R. B., M. A. STEPHENS and R. B. D'AGOSTINO (1986): *Goodness-Of-Fit-Techniques*: Marcel Dekker.

DIEBOLD, F. X., T. A. GUNTHER and A. S. TAY (1998): Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review* 39(4), 863-883.

DIEBOLD, F. X. and R. S. MARIANO (1995): Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13(3), 253-263.

DIEBOLD, F. X., A. S. TAY and K. WALLIS (1999): "Evaluating Density Forecasts of Ination: the Survey of Professional Forecasts," in.: Festschrift in Honor of CWJ Granger, Oxford: Oxford University Press, 76-90.

FERNANDEZ-VILLAVERDE, J. and J. RUBIO-RAMÍREZ (2004): Comparing Dynamic Equilibrium Models to Data: A Bayesian Approach. *Journal of Econometrics* 123(1), 153–187.
GALLANT, A. R. and D. W. NYCHKA (1987): Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica* 55(2), 363-390.
GRANGER, C. W. J. and M. H. PESARAN (1996): *A Decision-theoretic Approach to Forecast Evaluation*: University of Cambridge, Department of Applied Economics.
HAMILL, T. M. (2000): Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review* 129(3), 550-560.
HANSEN, P. R. (2001): An unbiased and powerful test for superior predictive ability. *Manuscript, Brown University*.
KULLBACK, S. and R. A. LEIBLER (1951): On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1), 79-86.
LI, F. and G. TKACZ (2001): *Evaluating Linear and Non-linear Time-varying Forecast-combination Methods*: Bank of Canada.
NOCETI, P., J. SMITH and S. HODGES (2003): An evaluation of tests of distributional forecasts. *Journal of Forecasting* 22(6-7), 447-455.
POON, S. H. and C. W. J. GRANGER (2003): Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature* 41(2), 478-539.
ROSENBLATT, M. (1952): Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics* 23(3), 470-472.
SARNO, L. and G. VALENTE (2004): Comparing the accuracy of density forecasts from competing models. *Journal of Forecasting* 23(8), 541-557.
TAY, A. S. and K. F. WALLIS (2000): Density forecasting: a survey. *Journal of Forecasting* 19(4), 235-254.
VUONG, Q. H. (1989): Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* 57(2), 307-333.
WEST, K. D. (1996): Asymptotic Inference about Predictive Ability. *Econometrica* 64(5), 1067-1084.
WEST, K. D., H. J. EDISON and D. CHO (1993): A utility-based comparison of some models of exchange rate volatility. *Journal of International Economics* 35(1-2), 23–46.
WHITE, H. (2000): A Reality Check for Data Snooping. *Econometrica* 68(5), 1097-1126.