Comparing Density Forecast Models†

YONG BAO,¹ TAE-HWY LEE²* AND BURAK SALTOĞLU³
¹ University of Texas, San Antonio, Texas, USA
² University of California, Riverside, California USA
³ Bosphorous University, Istanbul, Turkey

ABSTRACT
In this paper we discuss how to compare various (possibly misspecified) density forecast models using the Kullback–Leibler information criterion (KLIC) of a candidate density forecast model with respect to the true density. The KLIC differential between a pair of competing models is the (predictive) log-likelihood ratio (LR) between the two models. Even though the true density is unknown, using the LR statistic amounts to comparing models with the KLIC as a loss function and thus enables us to assess which density forecast model can approximate the true density more closely. We also discuss how this KLIC is related to the KLIC based on the probability integral transform (PIT) in the framework of Diebold et al. (1998). While they are asymptotically equivalent, the PIT-based KLIC is best suited for evaluating the adequacy of each density forecast model and the original KLIC is best suited for comparing competing models. In an empirical study with the S&P500 and NASDAQ daily return series, we find strong evidence for rejecting the normal-GARCH benchmark model, in favor of the models that can capture skewness in the conditional distribution and asymmetry and long memory in the conditional variance. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS density forecast comparison; KLIC; predictive log-likelihood; reality check

INTRODUCTION

Forecasting densities has always been at the core of the finance and economics research agenda. For instance, most of the classical finance theories, such as asset pricing, portfolio selection and option valuation, aim to model the surrounding uncertainty via a parametric distribution function. Extracting information about market participants’ expectations from option prices can be considered another form of density forecasting exercise (e.g., Jackwerth and Rubinstein, 1996). Moreover, there has also been increasing interest in evaluating forecasting models of inflation, unemployment and output in terms of density forecasts (Clements and Smith, 2000). While the research on evaluating each density forecast model has been very versatile since the seminal paper of Diebold et al. (1998), there has

*Correspondence to: Tae-Hwy Lee, Department of Economics, University of California, Riverside, CA 92521, USA. E-mail: tae.lee@ucr.edu
†Previous versions of this paper have been circulated with the title, ‘A Test for Density Forecast Comparison with Applications to Risk Management’ since October 2003; see Bao et al. (2004).
been much less effort in comparing alternative density forecast models. Given the recent empirical evidence on volatility clustering and asymmetry and fat-tailedness in financial return series, a formal test of relative adequacy of a given model among alternative models will definitely fill a gap in the existing literature. Deciding on which distribution and/or volatility specification to use for a particular asset is a common task even for finance practitioners. For example, despite the existence of many volatility specifications, a consensus on which model is most appropriate has yet to be reached. As argued in Poon and 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.1

Following Diebold et al. (1998), it has become common practice to evaluate the adequacy of a density forecast model based on the probability integral transform (PIT) of the process with respect to the model’s density forecast: see Clements and Smith (2000), Berkowitz (2001), Freichs and Löffler (2003), Bauwens et al. (2004), among others. If the density forecast model is correctly specified, the PIT follows an IID uniform distribution on the unit interval and, equivalently, its inverse normal transform follows an IID normal distribution. We can therefore evaluate a density forecast model by examining the departure of the transformed PIT from this property (IID and normality). The departure can be quantified by the Kullback–Leibler (1951) information criterion, or KLIC, which is the expected logarithmic value of the likelihood ratio (LR) of the transformed PIT and the IID normal variate. Thus the LR statistic measures the distance of a candidate model to the unknown true model.

On the other hand, to compare competing density forecast models, we may not have to use the PIT. Instead, we propose using directly the KLIC divergence measure for the original series as suggested by Vuong (1989). Hence, we use the KLIC as a ‘loss’ function for comparing competing density forecast models. In the framework of White (2000), we formulate a test statistic in terms of the loss differentials between the competing models and the benchmark, to compare various density forecast models against a benchmark model. Therefore, using the KLIC as a loss function amounts to using the negative predictive log-likelihood function as a loss function (as the true density in the definition of KLIC cancels out in formulating the KLIC differential). In this framework, all the density forecast models can be misspecified. Even though the true density is unknown, our test compares models in terms of distances of these models to the true density, and thus enables us to assess which volatility and/or distribution specifications are statistically more appropriate to model the time series behavior of a return series.

As an application, in the empirical section we demonstrate how the proposed testing methodology can be used to assess density forecasts for financial asset returns. We show that, by using the test, it is possible to differentiate the relative importance of various volatility and distribution specifications. To this end, we conduct a density forecast comparison of 80 (= 10 × 8) models constructed from ten different distributions and eight different volatility models. Our empirical findings based on the daily S&P500 and NASDAQ return series confirm the recent evidence on financial return asymmetry and long memory in volatility. The results obtained in the paper may shed some light on how both the distribution and volatility specifications can be treated to provide a better forecasting practice.

1They claim that lack of a uniform forecast evaluation technique makes volatility forecasting a difficult task. They further state (p. 507), ‘However, it seems clear that one form of study that is included is conducted just to support a viewpoint that a particular method is useful. It might not have been submitted for publication if the required result had not been reached. This is one of the obvious weaknesses of a comparison such as this; the papers being prepared for different reasons, use different data sets, many kinds of assets, various intervals between readings, and a variety of evaluation techniques.’
Comparing Density Forecast Models

This paper is organized as follows. In the next section we develop our distance measure based on the out-of-sample conditional KLIC divergence measure between a candidate density forecast model and the true density forecast model. We also discuss the framework to evaluate a density forecast model based on the PIT. The third section shows how to compare competing density forecast models in the framework of White’s (2000) ‘reality check’ using the out-of-sample conditional KLIC distance measure as a loss function. In the fourth section we present our empirical study of the daily S&P500 and NASDAQ return series. Finally, the fifth section concludes. A more detailed description of the distribution and volatility models described in the fourth section is given in the Appendix.

FORECASTING DENSITY

In this section we discuss how to evaluate the quality of a density forecast model based on the KLIC measure. We consider two different ways to construct the KLIC, namely KLIC(y) and KLIC(x), the first based on the return process y and the second based on the double probability transforms denoted as x (the inverse normal transform of the return’s probability transform). We show that they are related but that they can be used in different contexts. KLIC(x) is a measure used in the density forecast evaluation literature, e.g., Diebold et al. (1998) and Berkowitz (2001). It would be difficult to use KLIC(y) for the purpose of evaluation due to the fact that the true density to form KLIC(y) is unknown. However, for comparing many competing density forecast models, we can use either KLIC(y) or KLIC(x), as both can measure the quality of a density forecast model. While they are related, KLIC(y) has several advantages for the comparison purpose (to be discussed below (‘Evaluating density forecast models’)). Both can be treated as a loss function in the framework of Diebold and Mariano (1995), West (1996), White (2000), and Giacomini and White (2003).

Set-up

Consider a financial return series \{y_t\}_{t=1}^T. This observed data on a univariate series is a realization of a stochastic process \(Y^T = \{Y_t : \Omega \rightarrow \mathbb{R}, \tau = 1, 2, \ldots, T\}\) on a complete probability space \((\Omega, F_T, P_0^T)\), where \(\Omega = \mathbb{R}^T = \mathbb{X}_0 \times \mathbb{R}\) and \(F_T = \mathcal{B}(\mathbb{R}^T)\) is the Borel \(\sigma\)-field generated by the open sets of \(\mathbb{R}^T\), and the joint probability measure \(P_0^T(B) = P_0[Y^T \in B], B \in \mathcal{B}(\mathbb{R}^T)\) completely describes the stochastic process. A sample of size \(T\) is denoted as \(y^T = (y_1, \ldots, y_T)\).

Let a \(\sigma\)-finite measure \(\nu^T\) on \(\mathcal{B}(\mathbb{R}^T)\) be given. Assume \(P_0^T(B)\) is absolutely continuous with respect to \(\nu^T\) for all \(T = 1, 2, \ldots\), so that there exists a measurable Radon-Nikodým density \(g^T(y^T) = dP_0^T/d\nu^T\), unique up to a set of zero measure-\(\nu^T\).

Following White (1994, see next subsection), we define a probability model \(P\) as a collection of distinct probability measures on the measurable space \((\Omega, F_T)\). A probability model \(P\) is said to be correctly specified for \(Y^T\) if \(P\) contains \(P_0^T\). Our goal is to evaluate and compare a set of parametric probability models \(\{P_\theta^T\}\), where \(P_\theta^T(B) = P_\theta(Y^T \in B)\). Suppose there exists a measurable Radon-Nikodým density \(f^T(y^T) = dP_\theta^T/d\nu^T\) for each \(\theta \in \Theta\), where \(\theta\) is a finite-dimensional vector of parameters and is assumed to be identified on \(\Theta\), a compact subset of \(\mathbb{R}^k\) (see White, 1994, Theorem 2.6).

---

1 Much of the forecasting literature deals with forecast evaluation, forecast comparison, and forecast combination. Our paper extends the density forecast evaluation of Diebold et al. (1998) to density forecast comparison. We note that Mitchell and Hall (2005) further extend Diebold et al. (1998), Bao et al. (2004), and this paper to density forecast combination.
In the context of forecasting, instead of the joint density \( g(y_t) \), we consider forecasting the conditional density of \( Y_t \), given the information \( F_{t-1} \) generated by \( Y_{t-1} \). Let \( q_i(y_t) = q_i(y_t|F_{t-1}) = g(y_t)/g(r_t(y_t)) \) for \( t = 2, 3, \ldots \) and \( q_i(y_t) = q_i(y_t|F_0) = g(y_t) = g(r_t(y_t)) \). Thus the goal is to forecast the (true, unknown) conditional density \( q_i(y_t) \).

For this, we use a one-step-ahead conditional density forecast model \( \psi_i(y_t; \theta) = \psi_i(y_t|F_{t-1}; \theta) = f_i(y_t)/f_r(y_t) \) for \( t = 2, 3, \ldots \) and \( \psi_i(y_t) = \psi_i(y_t|F_0) = f_i(y_t) = f_r(y_t) \). If \( \psi_i(y_t; \theta) = q_i(y_t) \) almost surely for some \( \theta_i \in \Theta_i \), then the one-step-ahead density forecast is correctly specified, and it is said to be optimal because it dominates all other density forecasts for any loss functions (e.g., Diebold et al., 1998; Granger and Pesaran, 2000a, 2000b).

In practice, it is rarely the case that we can find an optimal model. As it is very likely that ‘the true distribution is in fact too complicated to be represented by a simple mathematical function’ (Sawa, 1978), all the models proposed by different researchers can be misspecified and thereby we regard each model as an approximation to the truth. Our task is then to investigate which density forecast model can approximate the true conditional density most closely. We have to first define a metric to measure the distance of a given model to the truth, and then compare different models in terms of this distance.

### Conditional KLIC for Density Forecast Models

The adequacy of a density forecast model can be measured by the conditional Kullback–Leibler (1951) information criterion (KLIC) divergence measure between two conditional densities:

\[
\mathbb{I}(\varphi : \psi, \theta) = \mathbb{E}_{\psi}[\ln \varphi_i(y_t) - \ln \varphi_i(y_t; \theta)]
\]

where the expectation is with respect to the true conditional density \( q_i(y_t|F_{t-1}) \), \( \mathbb{E}_{\psi} \ln q_i(y_t|F_{t-1}) < \infty \), and \( \mathbb{E}_{\psi} \ln \psi_i(y_t|F_{t-1}; \theta) < \infty \). Following White (1982, 1994), we define the distance between a density model and the true density as the minimum of the KLIC:

\[
\mathbb{I}(\varphi : \psi, \theta^*_{t+1}) = \mathbb{E}_{\psi}[\ln \varphi_i(y_t) - \ln \varphi_i(y_t; \theta^*_{t+1})]
\]

where \( \theta^*_{t+1} = \arg \min \mathbb{I}_i(\varphi : \psi, \theta) \) is the pseudo-true value of \( \theta \) (Sawa, 1978; White, 1982).\(^3\) We assume that \( \theta^*_{t+1} \) is an interior point of \( \Theta \). The smaller this distance is, the closer the density forecast \( \psi_i(y_t|F_{t-1}; \theta^*_{t+1}) \) is to the true density \( q_i(y_t|F_{t-1}) \).\(^4\)

However, \( \mathbb{I}_i(\varphi : \psi, \theta^*_{t+1}) \) is unknown since \( \theta^*_{t+1} \) is not observable. We need to estimate \( \theta^*_{t+1} \). As our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we split the data into two parts: one for estimation and the other for out-of-sample validation. At each period \( t \) in the out-of-sample period \( (t = R + 1, \ldots, T) \), we use the previous \( R \) rolling observations \( \{y_{t-1}, \ldots, y_{t-R}\}^{T}_{T+1} \) to estimate the unknown parameter vector \( \theta^*_{t+1} \).

\(^3\)We use the word ‘distance’ loosely because KLIC does not satisfy some basic properties of a metric, i.e., \( \mathbb{I}(\psi_i : \psi_i) = \mathbb{I}(\psi_i : \psi_i) \) and KLIC does not satisfy a triangle inequality. However, in this paper, as we will use the KLIC in comparing various competing models with a fixed benchmark model (i.e., the normal-GARCH model below (‘Application’)), KLIC can serve as a distance metric with respect to the fixed benchmark.

\(^4\)This motivates Vuong (1989) to design a model selection test for two competing models, while Amisano and Giacomini (2005) consider a weighted out-of-sample likelihood ratio test.
and denote the estimate as $\hat{q}_{R,t-1}$ (where the subscript $R$ denotes the size of the in-sample period). Under some regularity conditions, we can consistently estimate $q^*_{t-1}$ by $\hat{q}_{R,t-1}$ by maximizing

$$R^{-1} \sum_{i=R+1}^{T} \ln \psi_i(y_i; \theta)$$

see White (1994, Theorems 2.12 and 3.4) for the sets of conditions for the existence and consistency of $\hat{q}_{R,t-1}$.

Using $\{\hat{q}_{R,t-1}\}_{R+1}^{T}$, we can obtain the out-of-sample estimate of $E_q(\psi, \theta^*_{t-1})$ (where $E(\cdot)$ is the unconditional expectation)

$$I_{R,n}(\psi; \theta) = \frac{1}{n} \sum_{i=R+1}^{T} \ln \left[ \frac{\psi_i(y_i; \theta^*_{t-1})}{\psi_i(y_i; \hat{\theta}_{R,t-1})} \right]$$

where $n = T - R$ is the size of the out-of-sample period. Note that

$$I_{R,n}(\psi; \theta) = \frac{1}{n} \sum_{i=R+1}^{T} \ln \left[ \frac{\psi_i(y_i; \theta^*_{t-1})}{\psi_i(y_i; \theta^*_{t-1})} \right] + \frac{1}{n} \sum_{i=R+1}^{T} \ln \left[ \frac{\psi_i(y_i; \theta^*_{t-1})}{\psi_i(y_i; \hat{\theta}_{R,t-1})} \right]$$

where the first term in $I_{R,n}(\psi; \theta)$ measures model misspecification (the distance between the optimal density $q^*_{t-1}(y)$ and the model $\psi_i(y_i; \theta^*_{t-1})$) and the second term measures parameter estimation uncertainty due to the distance between $\theta^*_{t-1}$ and $\hat{\theta}_{R,t-1}$.

**Out-of-sample probability integral transform**

Alternatively, we may utilize an inverse normal transform of the PIT of the actual realization of the process with respect to the model’s density forecast. The PIT of the realization of the process with respect to the density forecast is defined as

$$u_t = \int_{-\infty}^{y_t} \psi_i(y | \mathcal{F}_{t-1}; \theta^*_{t-1}) dy$$

It is well known that if $\psi_i(y | \mathcal{F}_{t-1}; \theta^*_{t-1})$ coincides with the true density $q_i(y)$ almost surely, then the sequence $\{u_t\}$ is IID and uniform on the interval $[0, 1]$ (denoted $U[0, 1]$ henceforth). This provides a powerful approach to evaluating the quality of a density forecast model. Diebold et al. (1998) exploit and popularize this idea to check the adequacy of a density forecast model.

Our purpose of invoking the PIT is to make $I_{R,n}(\psi, \theta^*_{t-1})$ operational as $\psi_i(\cdot)$ is unknown. We take a further transform

---

5 There are different schemes to estimate $\theta^*_{t-1}$, namely, recursive, fixed, and rolling schemes. That is, $\theta^*_{t-1}$ can be estimated based on the whole subsample $\{y_1, \ldots, y_T\}$, or a fixed sample $\{y_R, \ldots, y_1\}$, or a rolling sample $\{y_{t-R}, \ldots, y_{t-1}\}$, respectively. See West and McCracken (1998, p. 819) for more discussion on the three forecasting schemes. While using the whole sample has the advantage of using more observations, we will use the rolling sample in our application because it may be more robust to possible parameter variation in the presence of potential structural breaks.

6 The effects of parameter estimation on prediction densities have been studied in recent literature (e.g., Pascual et al., 2001). In finance, Bawa et al. (1979) show that the predictive distribution of an asset return that is obtained by integrating the conditional distribution over the parameter space is different from the predictive distribution that is obtained when the parameters are treated as known. Also, Kandel and Stambaugh (1996), Barberis (2000) and Xia (2001), among others, explore the economic importance of estimation risk on predicting stock returns, selecting portfolios, hedging, long-horizon investment decisions, and market timing.
where \( \Phi(\cdot) \) is the standard normal CDF. If the sequence \( \{u_t\} \) is IID \( U[0, 1] \), then \( \{x_t\} \) is IID \( N(0, 1) \). More importantly, Berkowitz (2001, p. 467) shows that

\[
\ln \left[ \frac{\phi(y_t)}{\psi\left(y_t; \theta^*_{t-1}\right)} \right] = \ln \left[ \frac{p_t(x_t)}{\phi(x_t)} \right]
\]

where \( p_t(x_t) \) is the conditional density of \( x_t \), and \( \phi(x_t) \) is the standard normal density. To consider the effect of parameter estimation uncertainty, we define

\[
x_t = \Phi^{-1}(u_t)
\]

\[
t = R + 1, \ldots, T.
\]

We have transformed the departure of \( y_t(\cdot; q) \) from the unknown true density \( \phi_t(\cdot) \) to the departure of \( p_t(\cdot) \) from IID \( N(0, 1) \). It may sound like a loop as we do not know \( p_t(\cdot) \) either. The truth about the transformed PIT \( \hat{x}_t \) is nevertheless quite simple: it should behave as IID \( N(0, 1) \) if the density forecast hits the true density. We can specify a flexible \( p_t(\cdot) \) to nest IID \( N(0, 1) \) as a special case, but when we specify \( y_t(\cdot; q) \) there is no guarantee that the postulated \( y_t(\cdot; q) \) will nest the unknown \( \phi_t(\cdot) \). We follow Berkowitz (2001) by specifying \( \hat{x}_t \) as an AR(\( L \)) process:

\[
\hat{x}_t = \rho\hat{X}_{t-1} + \sigma \eta_t
\]

where \( X_{t-1} = (1, \hat{x}_{t-1}, \ldots, \hat{x}_{t-L})' \), \( \rho \) is an \( (L + 1) \times 1 \) vector of parameters, and \( \eta_t \) is IID distributed and independent of \( X_{t-1} \). In Berkowitz (2001), \( \eta_t \) is further assumed to be normally distributed. Clearly this is restrictive. A remedy for this is to specify a flexible alternative distribution for \( \eta_t \), say \( h(\eta_t; \theta^*_t) \), where \( \theta_t \) is a vector of distribution parameters such that when \( \theta_t = \hat{\theta}^*_t \), \( h(\cdot; \theta^*_t) \) is IID \( N(0, 1) \). A test for IID \( N(0, 1) \) of \( \hat{x}_t \) can be constructed by testing elements of the parameter vector \( \theta = (\rho', \sigma, \theta^*_t)' \), say \( \rho = 0, \sigma = 1 \), and \( \hat{\theta}_t = \hat{\theta}^*_t \). In particular, we assume that \( \{\hat{x}_t\}_{t=R+1}^T \) follows the AR process (8) with \( \eta_t \) IID distributed with the semi-nonparametric (SNP) density function of order \( K \) (Gallant and Nychka, 1987):

\[
h(\eta; \theta^*_t) = \frac{\left( \sum_{k=0}^K r_k \eta_k^2 \right) \phi(\eta_t)}{\int_\mathbb{R} \left( \sum_{k=0}^K r_k u^k \right)^2 \phi(u)du}
\]

where \( r_0 = 1 \). Now \( \theta_t = (r_1, \ldots, r_K)' \). Setting \( r_k = 0 (k = 1, \ldots, K) \), \( h(\eta) = \phi(\eta) \). Given (8) and (9), the conditional density of \( \hat{x}_t \) is

\[
p_t(\hat{x}_t; \theta^*_t) = \frac{h\left(\hat{x}_t - \rho\hat{X}_{t-1}; \theta^*_t\right)}{\sigma}
\]
which degenerates to \( \phi(\hat{\xi}) = \phi(\eta) \) by setting \( \hat{\theta} = \hat{\theta}' = (\Theta', 1, \Theta')' \). The parameter vector \( \hat{\theta} \) can be estimated by \( \hat{\theta}_n = (\beta_n, \hat{\sigma}_n, \hat{\theta}^*_n) \) that maximizes \( n^{-1} \sum_{t=R+1}^T \ln p_t(\hat{x}_t; \hat{\theta}) \).

Then we can obtain the out-of-sample estimated KLIC based on \( \{ \hat{x}_t \}_{t=R+1}^T \) for measuring the departure of \( p(\cdot) \) from \( \phi(\cdot) \) as follows:

\[
\mathbb{I}_{R,n}(p; \phi) = \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{p_t(\hat{x}_t; \hat{\theta}_n)}{\phi(\hat{x}_t)} \right] 
= \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{p_t(\hat{x}_t; \hat{\theta}_n)}{p_t(\hat{x}_t; \Theta)} \right] + \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{p_t(\hat{x}_t; \Theta)}{p_t(x_t; \Theta)} \right] 
+ \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{p_t(x_t; \Theta)}{\phi(x_t)} \right] + \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{\phi(x_t)}{\phi(\hat{x}_t)} \right]
\]

(11)

where the last equality comes from (6). The third term measures the distance between \( q_t(y_t) \) and the forecast model \( \psi_t(y_t; \Theta^*_R) \) (model misspecification), the second and fourth terms measure the distance between \( \Theta^*_R \) and \( \hat{\Theta}_R \) (parameter estimation uncertainty), and the first term measures the sampling variation in the estimation of \( p_t(\hat{x}_t; \hat{\theta}_n) \).

**Evaluating density forecast models**

We use \( \mathbb{I}_{R,n}(p; \phi) \) to evaluate the adequacy of the density forecast model \( \psi_t(y_t; \Theta) \), because using \( \mathbb{I}_{R,n}(\phi; \psi) \) is infeasible due to the unknown \( q_t(y_t) \). On the other hand, they are related. From (2) and (11),

\[
\mathbb{I}_{R,n}(p; \phi) = \mathbb{I}_{R,n}(\phi; \psi) 
+ \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{p_t(\hat{x}_t; \Theta^*_R)}{p_t(\hat{x}_t; \Theta)} \right] + \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{p_t(\hat{x}_t; \Theta)}{p_t(x_t; \Theta)} \right] 
+ \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{\psi_t(y_t; \Theta^*_R)}{\psi_t(y_t; \Theta)} \right] + \frac{1}{n} \sum_{t=R+1}^T \ln \left[ \frac{\phi(x_t)}{\phi(\hat{x}_t)} \right]
\]

(12)

Note that \( \mathbb{I}_{R,n}(p; \phi) \to_{a.s.} \mathbb{I}_{R,n}(\phi; \psi) \) as \( n \to \infty \), \( R \to \infty \), since \( \hat{\Theta}_{R+1} \to_{a.s.} \Theta^*_R \) as \( R \to \infty \) (for all \( t = R + 1, \ldots, T \)) and \( \hat{\Theta}_R \to_{a.s.} \Theta^*_R \) as \( n \to \infty \).

Given the KLIC, \( \mathbb{I}_{R,n}(p; \phi) \), we can design an LR test statistic:

\[
LR_n = 2 \sum_{t=R+1}^T \ln \left[ \frac{p_t(\hat{x}_t; \hat{\theta}^*_n)}{\phi(\hat{x}_t)} \right] = 2 \times n \times \mathbb{I}_{R,n}(p; \phi)
\]

(13)

which is asymptotically \( \chi^2 \) with degrees of freedom \( (L + 1) + 1 + K \) if the model is correctly specified, or equivalently, if \( \hat{x}_t \) is IID \( N(0, 1) \), that is, if \( \rho = 0, \sigma = 1 \), and \( \hat{\theta}_n = (r_1, \ldots, r_K)' = \Theta' \). This can be regarded as a generalized version of the LR statistic of Berkowitz (2001).
COMPARING DENSITY FORECAST MODELS

Suppose there are \( l + 1 \) models \((j = 0, 1, \ldots, l)\) in the set of competing, possibly misspecified, models. To establish notation with model index \( j \), let the density forecast model \( j \) \((j = 0, 1, \ldots, l)\) be denoted by \( \psi^j(y; \hat{\theta}^j_{t-1}) \). Let the loss-differential between model 0 and model \( j \) at each time \( t \) be denoted as \( d_{j,t} \), and let \( \bar{d}_{j,n} = \frac{1}{n} \sum_{t=R+1}^T d_{j,t} \) be the average of \( d_{j,t} \) over the out-of-sample observations \( t = R + 1, \ldots, T \). Model comparison between model \( j \) and the benchmark model (model 0) can be conveniently formulated as the hypothesis testing of some suitable out-of-sample moment conditions on the loss-differential.

Loss functions

Since the KLIC measure takes on a smaller value when a model is closer to the truth, we can regard it as a loss function and use \( \sum_{R,R,n} \psi : \phi \) or \( \sum_{R,R,n} (p : \phi) \) to formulate the loss differential. For example, the out-of-sample average of the loss differential between model 0 and model \( j \) can be written as

\[
\bar{d}_{j,n}(x) = \sum_{R,R,n} (p^0 : \phi) - \sum_{R,R,n} (p^j : \phi)
\] (14)

However, for the purpose of comparing the density forecast models, it is better to use \( \sum_{R,R,n} (q : \psi) \) than \( \sum_{R,R,n} (p : \phi) \). Using \( \sum_{R,R,n} (q : \psi) \), we write the out-of-sample average of the loss-differential between model 0 and model \( j \) as follows:

\[
\bar{d}_{j,n}(y) = \sum_{R,R,n} (\psi^0 : \phi) - \sum_{R,R,n} (\psi^j : \phi)
\]

(15)

where \( d_{j}(y, \beta^j) = \ln[\psi^j(y; \theta^j_{R+1})/\psi^0(y; \hat{\theta}^0_{R+1})] \) and \( \beta^j = (\theta^0, \theta^j)' \).

For the purpose of comparing the density forecast models, several advantages of using \( \bar{d}_{j,n}(y) \) over \( \bar{d}_{j,n}(x) \) may be pointed out. One is that it is simpler as it does not involve the true density \( q_i(y) \) since it is cancelled out in forming the loss differential (15). Thus the loss differential \( \bar{d}_{j,n}(y) \) can be computed even though \( q_i(y) \) is unknown. Another advantage is that the effect of parameter estimation uncertainty is asymptotically negligible for the inference (we discuss more on this issue shortly—see footnote 7). A third advantage is rather obvious from observing the relationship in (12); that is, \( \sum_{R,R,n} (p : \phi) \) is more complicated than \( \sum_{R,R,n} (q : \psi) \).

It is important to note that formulating \( \bar{d}_{j,n}(y) \) amounts to using the negative predictive log-likelihood as a loss function:

\[
\bar{d}_{j,n}(y) = \frac{1}{n} \sum_{R,R,n} [(-\ln \psi^j(y; \hat{\theta}^j_{R+1})) - (-\ln \psi^0(y; \hat{\theta}^0_{R+1}))]
\] (16)

This is closely related to the work of Amisano and Giacomini (2005) when comparing two density forecast models. In particular, they take the logarithmic scoring rule, which gives the log-likelihood ratio. In practice, we conjecture that the test based on \( \{y_i\} \) (i.e., with \( \bar{d}_{j,n}(y) \) in (15) as the loss differential) should be more powerful compared with that based on \( \{\hat{x}_i\} \) (i.e., with \( \bar{d}_{j,n}(x) \) in (14) as the loss differential) due to the estimation error associated with estimating the density of \( \{\hat{x}_i\}_{t=R+1}^T \).
Test statistic

When we compare model \( j \) with a benchmark (model 0), the null hypothesis is that model \( j \) is no better than the benchmark, \( H_0: \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) = 0 \) for \( j = 1, \ldots, l \). We can use the out-of-sample predictive ability tests of Diebold and Mariano (1995) by referring to the standard asymptotic result for \( \sqrt{n}d_{j,t} \). By a suitable central limit theorem, we have

\[
\sqrt{n}\left[ d_{j,n}(y) - \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) \right] \rightarrow N(0, \xi^2)
\]  

(17)

in distribution as \( n = n(T) \rightarrow \infty \) when \( T \rightarrow \infty \), where \( \xi^2 = \lim \text{var} \left( \sqrt{n}\left[ d_{j,n}(y) - \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) \right] \right) \). In general, the variance \( \xi^2 \) may be rather complicated because it depends on parameter estimation uncertainty (West, 1996). However, we note that one of the significant merits of using the KLIC as a loss function in comparing density forecast models is that parameter estimation uncertainty does not affect asymptotic inference. This is due to the use of the same objective function for the in-sample parameter estimation (minimizing the KLIC to get \( \hat{\theta} \)) and for the out-of-sample forecast validation (using the KLIC as a forecast evaluation criterion).

When we compare multiple models against a benchmark jointly, the null hypothesis of interest is that no model is better than the benchmark:

\[
H_0: \max_{1 \leq j \leq l} \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) \leq 0
\]  

(18)

This is a multiple hypothesis, the intersection of the one-sided individual hypotheses \( \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) \leq 0 \) for \( j = 1, \ldots, l \). The alternative is that \( H_0 \) is false, that is, there is at least one model that is superior to the benchmark. If the null hypothesis is rejected, there must be at least one model for which \( \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) \) is positive. Under appropriate conditions and defining \( \bar{d} \) to be an \( l \times 1 \) column vector by stacking \( d_{j,t}(y) \), then \( \sqrt{n}(\bar{d} - \mathbb{E}(\bar{d}^*)) \rightarrow N(0, \Omega) \) as \( n(T) \rightarrow \infty \) when \( T \rightarrow \infty \), for \( \Omega \) positive semi-definite, where \( \mathbb{E}\bar{d}^* = \mathbb{E}\bar{d}^* \) (assuming stationarity) for \( \bar{d}^* \) defined as an \( l \times 1 \) column vector obtained by stacking the \( d_{j,t}(y_t, \beta_{0j}^{*}) \). White’s (2000) test statistic for \( H_0 \) is formed as

\[
\bar{V}_n = \max_{1 \leq j \leq l} \sqrt{n}\left[ d_{j,n}(y) - \mathbb{E}d_{j,t}(y_t, \beta_{0j}^{*}) \right]
\]  

(19)

However, as the null limiting distribution of \( \bar{V}_n \) is unknown due to the presence of unknown \( \Omega \), White (2000) shows that the distribution of \( \sqrt{n}(\bar{d}^* - \bar{d}) \) converges to that of \( \sqrt{n}(\bar{d} - \mathbb{E}(\bar{d}^*)) \), where

---

7 As shown in West (1996, Theorem 4.1), when the derivative of the loss differential evaluated at \( \beta_{0j}^{*} = (\theta_{0j}^{*})' \) is zero, i.e., \( F = \mathbb{E}(\partial d_{j,t}(y_t, \beta_{0j}^{*})/\partial \theta_{0j}^{*}) = 0 \), under some proper regularity conditions, \( \xi^2 \) does not depend on parameter estimation uncertainty. We have this condition:

\[
F = \mathbb{E}(\partial d_{j,t}(y_t, \beta_{0j}^{*})/\partial \theta_{0j}^{*})
\]

\[
= \mathbb{E}(\partial \ln \psi(y_t; \theta_{0j}^{*})/\partial \theta_{0j}^{*})|/\partial \beta_{0j}^{*}) = 0
\]

because both terms inside the brackets are zero as \( \theta_{0j}^{*} \) maximizes \( \mathbb{E}\ln \psi(y_t; \theta_{0j}^{*}) \), which makes parameter estimation uncertainty due to the second term in (3) asymptotically negligible in \( \xi^2 \). The last line follows from the law of iterated expectations.

8 White (2000) does not require that \( \Omega \) be positive definite (which is required in West, 1996), but that \( \Omega \) be positive semi-definite (White 2000, p. 1105–1106). This allows for the case when some of the models under comparison are nested, as long as at least one of the competing models \( k = 1, \ldots, l \) is non-nested with respect to the benchmark.

Copyright © 2007 John Wiley & Sons, Ltd.  
DOI: 10.1002/for
\( \bar{d}^\dagger \) is obtained from the stationary bootstrap of Politis and Romano (1994). By the continuous mapping theorem this result extends to the maximal element of the vector \( \sqrt{n} (\bar{d}^\dagger - \bar{d}) \) so that the empirical distribution of

\[
\bar{V}_n^\dagger = \max_{1 \leq j \leq l} \sqrt{n} \left( \bar{d}_{j, n}(y) - \bar{d}_{j, n}(y) \right)
\]

may be used to compute the \( p \)-value of \( \bar{V}_n \). This bootstrap \( p \)-value for testing (18) is called the ‘reality check \( p \)-value’.

In practice, to implement the reality check test of White (2000), we set \( d_{j, n}(y) = 0 \). This setting makes the null hypothesis the least favorable to the alternative, because setting \( d_{j, n}(y) = 0 \) guarantees that the statistic satisfies the null hypothesis \( \mathbb{E}(\bar{d}_{j, n}(y)^\dagger - \bar{d}_{j, n}(y)) = 0 \) for all \( j \). Consequently, it renders a very conservative test. When a poor model is introduced, the reality check \( p \)-value for model \( j \) becomes very large and, depending on the variance of \( d_{j, n}(y) \), it may remain large even after the inclusion of better models. So we may regard White’s reality check \( p \)-value as an upper bound of the true \( p \)-value. Hansen (2005) considers the following modification to (20):

\[
\bar{V}_n^\dagger = \max_{1 \leq j \leq l} \sqrt{n} \left( \bar{d}_{j, n}(y)^\dagger - g(\bar{d}_{j, n}(y)) \right)
\]

where different \( g(\cdot) \) functions will produce different bootstrap distributions that are compatible with the null hypothesis. In this paper, we follow Hansen’s (2005) recommendation to set \( g(\cdot) \) as a function of the variance of \( d_{j, n}(y) \), i.e.

\[
g(d_{j, n}(y)) = \begin{cases} 
0 & \text{if } d_{j, n}(y) \leq -A_j \\
\bar{d}_{j, n}(y) & \text{if } d_{j, n}(y) > -A_j
\end{cases}
\]

where \( A_j = \frac{1}{T} n^{-1/4} \sqrt{\text{var}(n^{1/2} \bar{d}_{j, n}(y))} \) with the variance estimated from the bootstrap samples. When \( \mathbb{E}d_{j, n}(y), \beta_{\cdot j}^{(Y)} = 0 \) for all \( 1 \leq j \leq l \), then the reality check \( p \)-value of White (2000) will provide an asymptotically correct size. However, when some models are dominated by the benchmark model, i.e., \( \mathbb{E}d_{j, n}(y), \beta_{\cdot j}^{(Y)} < 0 \) for some \( 1 \leq j \leq l \), then the reality check \( p \)-value of White (2000) will make a conservative test. So, when bad models are included in the set of the competing models, White’s test tends to behave conservatively. Hansen’s (2005) modification is basically to remove the effects of those (very) bad models in the comparison.

**APPLICATION**

In this section, we apply the reality check test under the KLIC-based loss function to investigate the adequacy of various density forecast models for the daily S&P500 and NASDAQ return series. We consider the following model:

\[
y_t = \mu_t + \varepsilon_t = \mu_t + z_t \sigma_t
\]

where \( \mu_t = \mathbb{E}(y_t | F_{t-1}) \), \( \sigma_t^2 = \mathbb{E}(\varepsilon_t^2 | F_{t-1}) \), and \( z_t = \varepsilon_t / \sigma_t \). We assume that \( z_t \) is IID and \( \mu_t \) follows an MA(1) process without a constant term (see, e.g., Anderson et al., 2002; Hong and Lee, 2003). A density
Comparing Density Forecast Models

A forecast model based on (23) can be decomposed into two parts: specification of \( \sigma_t^2 \) and specification of the conditional distribution of \( \{ z_t \}_{t=1}^T \). We discuss technical details of the distribution and volatility models in the Appendix.

**DISTRIBUTION SPECIFICATIONS**

We can specify the distribution of the standardized residuals \( \{ z_t \}_{t=1}^T \) by a conditional density function \( f(z_t | \mathcal{F}_{t-1}) = f(z_t | \mathcal{F}_{t-1}; \theta) \) with a vector of distribution parameters \( \theta \). Given \( f(.|\cdot) \) and conditional volatility \( \sigma_t^2 = \sigma_t^2(\Theta) \) with a vector of volatility parameters \( \Theta \), the conditional distribution of \( y_t = \psi(y_t | \mathcal{F}_{t-1}; \Theta) = f(z_t) / \sigma_t \), where \( \psi(\cdot|\cdot) \) denotes the skewness and excess kurtosis parameters specified by a distribution, respectively. For symmetric distributions, \( \gamma_1 = 0 \).

The candidate distributions included in this paper are, of course, far from being complete. Distributions that have been used in the literature but which we do not use in this paper include the exponentially generalized beta distribution of type II (McDonald and Xu, 1995), the generalized \( t \) (McDonald and Newey, 1988) and its skewed generalization (Theodossiou, 1998), the semiparametric density (Engle and González-Rivera, 1991), the local density (Gourieroux and Jasiak, 2001), the generalized hyperbolic density (Eberlin and Keller, 1995), and the stable Pareto distribution (Panorska et al., 1995). In general, the cost of using these more general or complicated models is computational time and the sensitivity of estimation to starting values of parameters and to outliers in the data. Further, for some distributions, the second and higher moments do not even exist. Nonexistence of moments, especially the second moment, can cause some fundamental problems in pricing theory.

We consider five symmetric distributions: the standard normal, Student \( t \), generalized error, Laplace, and double Weibull distributions. The standard Gaussian normal density has long been used in the literature primarily due to its simplicity and its well-understood properties. However, the assumption of normality is quite dubious in that \( f(z) \) has a very restrictive shape with \( \gamma_1 = 0 \) and \( \gamma_2 = 0 \), which may not be consistent with the actual data. The Student \( t \) distribution has fatter tails compared with the standard normal distribution. Nelson (1991) uses the generalized error distribution (GED) to model the distribution of stock market returns. The GED is still symmetric, but is quite flexible in the tails. The Laplace distribution (also known as the double-exponential distribution), a special case of the GED, is used in Mittnik and Rachev (1993) to model unconditional distributions of asset returns. The double Weibull distribution, which generalizes the Laplace distribution, but differs from the GED, is also proposed in Mittnik and Rachev (1993).

While the GED and double Weibull distribution allow for considerable flexibility for \( \gamma_2 \), they cannot model the skewness observed in many financial series. In this paper, we include five flexible parametric distributions which can model both skewness and excess kurtosis, namely the skewed \( t \), inverse hyperbolic sine, mixture of normals, double gamma, and Gram–Charlier/Edgeworth–Sargan densities. The skewed \( t \) is proposed by Fernández and Steel (1998) as a four-parameter skewed Student distribution, where the four parameters specifying the location, dispersion, skewness and kurtosis have meaningful interpretations. The inverse hyperbolic sine (IHS) transformation with two parameters is used in Hansen et al. (2000) and Choi (2001) to model asymmetric and fat-tailed distributions. Another distribution which can exhibit skewness and excess kurtosis is the mixture of normal distributions; see Venkataraman (1997) and Wang (2001) for its application. For simplicity, in this paper we focus on a mixture of two normals. The double gamma...
distribution is another candidate distribution, which was first proposed by Knight et al. (1995). The Gram–Charlier/Edgeworth–Sargan density (to be simply denoted as Sargan henceforth) has also been used in the literature: see, e.g., Mauleón and Perote (1995) and Jondeau and Rockinger (2001). We follow Jondeau and Rockinger (2001) and use the type-B density and adopt their algorithm to implement the positiveness constraint.

**Volatility specifications**

The conditional variance $\sigma_t^2$ can be specified with various volatility models. We can specify it non-parametrically, parametrically (the GARCH family), or through some stochastic volatility model. For comparison of these different volatility specifications, see Poon and Granger (2003) and references therein. In this paper, we focus on the GARCH family.

The ARCH model of Engle (1982) and its GARCH generalization by Bollerslev (1986) can capture a salient feature of financial series: volatility persistence. Note that the GARCH model implies an exponential decay in the autocorrelation of the conditional variance. Empirical findings suggest that a shock in volatility seems to have long memory. This gives rise to the fractionally integrated GARCH model (FIGARCH) of Bollerslev and Mikkelsen (1996), where a fractional parameter $d$ controls the rate of hyperbolic decay in the autocorrelation of the conditional variance. A generalization of FIGARCH is the hyperbolic GARCH (HYGARCH) model of Davidson (2004). The component GARCH (CGARCH) model of Ding and Granger (1996) and Engle and Lee (1999) is also capable of capturing slow decay in the second moments.

The above GARCH-family models are symmetric in the sense that positive and negative shocks have the same effects on the conditional variance. Asymmetric GARCH models include the exponential GARCH (EGARCH) model of Nelson (1991), the GJR-GARCH of Glosten et al. (1993), the threshold GARCH (TGARCH) of Zakoian (1994), the smooth transition GARCH (STGARCH) of González-Rivera (1998), and the asymmetric power ARCH (APARCH) of Ding et al. (1993). Note that the APARCH model nests ARCH, GARCH, GJR, and TGARCH. If we combine the idea of fractional integration with asymmetric GARCH models, we can easily have the HYAPARCH, for example.

**Empirical results**

In this section, we use two datasets to compare the 80 density forecast models described in the previous subsections (ten different distribution models and eight different volatility models). The two datasets are the daily S&P500 and NASDAQ return series, retrieved from finance.yahoo.com and CRSP.9 They are from 3 January 1990 to 30 June 2003 ($T = 3403$). We split the sample into two parts (roughly into two halves): one for in-sample estimation of size $R = 1703$ and another for out-of-sample density forecasts of size $n = 1700$. We use a rolling-sample scheme. That is, the first density forecast is based on observations 1 through $R$ (3 January 1990 to 24 September 1996), the second density forecast is based on observations 2 through $R + 1$ (4 January 1990 to 25 September 1996), and so on. We fix the lag of the conditional variance and the lag of the squared errors in all the volatility models to be both 1.

---

9 There were a few missing observations in the NASDAQ series from finance.yahoo.com, which were checked against the CRSP data provided by Canlin Li. Other than these few observations, the Nasdaq series from the two sources were consistent. Regarding the week following 11 September, 2001, we treat it as a long holiday.
Table I. Evaluating density forecast models

<table>
<thead>
<tr>
<th></th>
<th>GARCH</th>
<th>GJR</th>
<th>APARCH</th>
<th>EGARCH</th>
<th>STGARCH</th>
<th>HYGARCH</th>
<th>HYAPARCH</th>
<th>CGARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;P 500 Normal</td>
<td>0.0277</td>
<td>0.0246</td>
<td>0.0193</td>
<td>0.0211</td>
<td>0.0308</td>
<td>0.0291</td>
<td>0.0192</td>
<td>0.0228</td>
</tr>
<tr>
<td>Student t 0.0001</td>
<td>0.0104</td>
<td>0.0108</td>
<td>0.0090</td>
<td>0.0104</td>
<td>0.0133</td>
<td>0.0103</td>
<td>0.0087</td>
<td>0.0089</td>
</tr>
<tr>
<td>GED</td>
<td>0.0147</td>
<td>0.0144</td>
<td>0.0125</td>
<td>0.0141</td>
<td>0.0182</td>
<td>0.0144</td>
<td>0.0123</td>
<td>0.0133</td>
</tr>
<tr>
<td>Laplace</td>
<td>0.0365</td>
<td>0.0444</td>
<td>0.0409</td>
<td>0.0429</td>
<td>0.0425</td>
<td>0.0331</td>
<td>0.0406</td>
<td>0.0359</td>
</tr>
<tr>
<td>Double Weibull</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Skewed t</td>
<td>0.0581</td>
<td>0.0373</td>
<td>0.0392</td>
<td>0.0358</td>
<td>0.0405</td>
<td>0.0584</td>
<td>0.0394</td>
<td>0.0665</td>
</tr>
<tr>
<td>IHS</td>
<td>0.0095</td>
<td>0.0122</td>
<td>0.0112</td>
<td>0.0127</td>
<td>0.0130</td>
<td>0.0094</td>
<td>0.0109</td>
<td>0.0084</td>
</tr>
<tr>
<td>Mixture</td>
<td>0.0068</td>
<td>0.0097</td>
<td>0.0086</td>
<td>0.0098</td>
<td>0.0101</td>
<td>0.0065</td>
<td>0.0083</td>
<td>0.0056</td>
</tr>
<tr>
<td>Double gamma</td>
<td>0.0111</td>
<td>0.0108</td>
<td>0.0089</td>
<td>0.0103</td>
<td>0.0182</td>
<td>0.0084</td>
<td>0.0081</td>
<td>0.0109</td>
</tr>
<tr>
<td>Sargan</td>
<td>0.0136</td>
<td>0.0148</td>
<td>0.0108</td>
<td>0.0172</td>
<td>0.0142</td>
<td>0.0109</td>
<td>0.0094</td>
<td>0.0094</td>
</tr>
<tr>
<td>NASDAQ Normal</td>
<td>0.0270</td>
<td>0.0318</td>
<td>0.0303</td>
<td>0.0313</td>
<td>0.0327</td>
<td>0.0269</td>
<td>0.0302</td>
<td>0.0213</td>
</tr>
<tr>
<td>Student t 0.0000</td>
<td>0.0192</td>
<td>0.0243</td>
<td>0.0226</td>
<td>0.0217</td>
<td>0.0213</td>
<td>0.0187</td>
<td>0.0225</td>
<td>0.0167</td>
</tr>
<tr>
<td>GED</td>
<td>0.0243</td>
<td>0.0273</td>
<td>0.0259</td>
<td>0.0262</td>
<td>0.0254</td>
<td>0.0226</td>
<td>0.0258</td>
<td>0.0205</td>
</tr>
<tr>
<td>Laplace</td>
<td>0.0599</td>
<td>0.0665</td>
<td>0.0584</td>
<td>0.0593</td>
<td>0.0619</td>
<td>0.0536</td>
<td>0.0586</td>
<td>0.0564</td>
</tr>
<tr>
<td>Double Weibull</td>
<td>0.0310</td>
<td>0.0346</td>
<td>0.0265</td>
<td>0.0244</td>
<td>0.0267</td>
<td>0.0226</td>
<td>0.0240</td>
<td>0.0191</td>
</tr>
<tr>
<td>Skewed t</td>
<td>0.0268</td>
<td>0.0206</td>
<td>0.0198</td>
<td>0.0196</td>
<td>0.0220</td>
<td>0.0276</td>
<td>0.0200</td>
<td>0.0304</td>
</tr>
<tr>
<td>IHS</td>
<td>0.0222</td>
<td>0.0218</td>
<td>0.0198</td>
<td>0.0205</td>
<td>0.0216</td>
<td>0.0189</td>
<td>0.0203</td>
<td>0.0161</td>
</tr>
<tr>
<td>Mixture</td>
<td>0.0159</td>
<td>0.0112</td>
<td>0.0110</td>
<td>0.0104</td>
<td>0.0174</td>
<td>0.0167</td>
<td>0.0115</td>
<td>0.0153</td>
</tr>
<tr>
<td>Double gamma</td>
<td>0.0231</td>
<td>0.0240</td>
<td>0.0232</td>
<td>0.0224</td>
<td>0.0253</td>
<td>0.0212</td>
<td>0.0213</td>
<td>0.0213</td>
</tr>
<tr>
<td>Sargan</td>
<td>0.0282</td>
<td>0.0372</td>
<td>0.0328</td>
<td>0.0042</td>
<td>0.0098</td>
<td>0.0103</td>
<td>0.0099</td>
<td>0.0294</td>
</tr>
</tbody>
</table>

Note: The first row for each combination of volatility and distribution gives the estimated KLIC \( I_{KL}(p, f) \) in equation (11) based on the AR(L)-SNP(K) model for the transformed PIT as shown in equations (8) and (10), where the orders of \( L \) and \( K \) are chosen by the AIC criteria for \( L = 0, \ldots, 5 \) and \( K = 0, \ldots, 8 \). The second row gives the \( p \)-value associated with the LR statistic in equation (13) for each model, testing the null hypothesis that the density forecast model is optimal, i.e., the transformed PIT is IID \( N(0,1) \).

As all the models are possibly misspecified, for each combination of volatility and conditional distribution, we first evaluate the adequacy of each density forecast model. Table I reports \( I_{KL}(p, f) \) (first row) for each model and the associated \( p \)-value (second row) of the LR statistic (13) for testing the optimality of a density forecast model, where \( L \) and \( K \) in the AR(L)-SNP(K) specification are chosen by the Akaike information criterion (AIC). From Table I, we observe the following.

Firstly, when we look for the model that gives the minimum value of \( I_{KL}(p, f) \), i.e., the transformed PIT is IID \( N(0,1) \), we find that for the S&P500 data, given any volatility, the mixture distribution fares best, followed by IHS, and then double gamma. The common prop-
ertainty of these distributions is that they aim to model the potential fat-tailedness and skewness. More specifically, mixture-CGARCH obtains the least KLIC measure followed by mixture-HYGARCH. On the other hand, the normal, Laplace and skewed \( t \) distributions deliver the highest KLIC statistics, indicating their relatively poor performance. For the NASDAQ, we obtain a somewhat similar picture. Here both the Sargan and mixture densities appear to be the two most successful distributions, while the normal and Laplace distributions are among the worst.

Secondly, given a volatility model, the lowest KLIC comes from one of the five skewed distributions. In particular, for the S&P500 data, the mixture normal distribution seems to do very well, while for the NASDAQ data both the mixture normal and Sargan densities fare well. Given a distribution model, the lowest KLIC comes from either an asymmetric volatility model (e.g., EGARCH) or a long-memory volatility model (e.g., CGARCH), or an asymmetric and long-memory volatility model (e.g., HYAPARCH). This indicates that symmetric and short-memory volatility models may not be adequate to approximate the conditional variance of return series.

Table II reports the values of the negative predictive log-likelihood, 

\[-n^{-1}\sum_{t=1}^{T}\ln p(y_t; \hat{\theta}_{k,t-1})\]

The conclusions we can draw from Table II are consistent with those from Table I—there is further evidence for skewness in the conditional distribution, and evidence of asymmetry and long memory in the conditional volatility. As in Table I, the mixture and IHS distributions are the best alternatives. We also observe that overall model ranking based on KLIC(\( x \)) is consistent with that based on the negative predictive log-likelihood of \( y \), in the sense that a low (high) KLIC(\( x \)) is often associated with a low (high) negative predictive log-likelihood.

Also reported in Table II are the reality check \( p \)-values, where the benchmark is the Normal-GARCH model, the most commonly used combination in empirical studies. The \( p_1 \) and \( p_2 \) refer to

Table II. Comparing density forecast models

<table>
<thead>
<tr>
<th></th>
<th>GARCH</th>
<th>GJR</th>
<th>APARCH</th>
<th>EGARCH</th>
<th>STGARCH</th>
<th>HYGARCH</th>
<th>HYAPARCH</th>
<th>CGARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S&amp;P 500</strong></td>
<td>( p_1 = 0.007 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>1.6347</td>
<td>1.6225</td>
<td>1.6135</td>
<td>1.6134</td>
<td>1.6459</td>
<td>1.6421</td>
<td>1.6410</td>
<td>1.6397</td>
</tr>
<tr>
<td>Student ( t )</td>
<td>1.6288</td>
<td>1.6121</td>
<td>1.6067</td>
<td>1.6070</td>
<td>1.6309</td>
<td>1.6276</td>
<td>1.6063</td>
<td>1.6255</td>
</tr>
<tr>
<td>GED</td>
<td>1.6310</td>
<td>1.6166</td>
<td>1.6107</td>
<td>1.6107</td>
<td>1.6364</td>
<td>1.6317</td>
<td>1.6103</td>
<td>1.6304</td>
</tr>
<tr>
<td>Laplace</td>
<td>1.6564</td>
<td>1.6451</td>
<td>1.6405</td>
<td>1.6406</td>
<td>1.6602</td>
<td>1.6549</td>
<td>1.6406</td>
<td>1.6545</td>
</tr>
<tr>
<td>Double Weibull</td>
<td>1.6231</td>
<td>1.6209</td>
<td>1.6181</td>
<td>1.6116</td>
<td>1.6346</td>
<td>1.6299</td>
<td>1.6111</td>
<td>1.6292</td>
</tr>
<tr>
<td>Skewed ( t )</td>
<td>1.6280</td>
<td>1.6109</td>
<td>1.6061</td>
<td>1.6065</td>
<td>1.6301</td>
<td>1.6267</td>
<td>1.6057</td>
<td>1.6249</td>
</tr>
<tr>
<td>IHS</td>
<td>1.6281</td>
<td>1.6107</td>
<td>1.6060</td>
<td>1.6065</td>
<td>1.6301</td>
<td>1.6268</td>
<td>1.6057</td>
<td>1.6251</td>
</tr>
<tr>
<td>Mixture</td>
<td>1.6303</td>
<td>1.6084</td>
<td>1.6030</td>
<td>1.6042</td>
<td>1.6336</td>
<td>1.6257</td>
<td>1.6044</td>
<td>1.6274</td>
</tr>
<tr>
<td>Double gamma</td>
<td>1.6424</td>
<td>1.6320</td>
<td>1.6140</td>
<td>1.6224</td>
<td>1.6368</td>
<td>1.6392</td>
<td>1.6157</td>
<td>1.6324</td>
</tr>
<tr>
<td>Sargan</td>
<td>1.6365</td>
<td>1.6178</td>
<td>1.6108</td>
<td>1.6114</td>
<td>1.6388</td>
<td>1.6409</td>
<td>1.6149</td>
<td>1.6309</td>
</tr>
</tbody>
</table>

| **NASDAQ**     | \( p_1 = 0.145 \) |      |        |        |         |         |          |        |
| Normal         | 2.0298 | 2.0220 | 2.0191 | 2.0174 | 2.0309 | 2.0270 | 2.0184 | 2.0222 |
| Student \( t \) | 2.0220 | 2.0162 | 2.0140 | 2.0126 | 2.0235 | 2.0210 | 2.0132 | 2.0153 |
| GED            | 2.0263 | 2.0192 | 2.0169 | 2.0156 | 2.0275 | 2.0246 | 2.0162 | 2.0198 |
| Laplace        | 2.0641 | 2.0597 | 2.0554 | 2.0544 | 2.0673 | 2.0600 | 2.0550 | 2.0617 |
| Double Weibull | 2.0506 | 2.0319 | 2.0341 | 2.0207 | 2.0336 | 2.0306 | 2.0214 | 2.0222 |
| Skewed \( t \) | 2.0190 | 2.0119 | 2.0094 | 2.0080 | 2.0207 | 2.0177 | 2.0087 | 2.0125 |
| IHS            | 2.0531 | 2.0347 | 2.0366 | 2.0524 | 2.0544 | 2.0534 | 2.0423 | 2.0138 |
| Mixture        | 2.0246 | 2.0139 | 2.0087 | 2.0092 | 2.0268 | 2.0205 | 2.0530 | 2.0226 |
| Double gamma   | 2.0480 | 2.0576 | 2.0388 | 2.0593 | 2.0640 | 2.0608 | 2.0515 | 2.0419 |
| Sargan         | 2.0653 | 2.0497 | 2.0524 | 2.0117 | 2.0273 | 2.0234 | 2.0133 | 2.0207 |

Note: Each cell gives the negative predictive log-likelihood \(-n^{-1}\sum_{t=1}^{T}\ln p(y_t; \hat{\theta}_{k,t-1})\) for each model given the distribution and volatility combination. For comparison we take the normal-GARCH combination as a benchmark model, with which the remaining 79 models are compared using the negative predictive log-likelihood as a loss, where \( p_1 \) and \( p_2 \) refer to the reality check \( p \)-values of White (2000) and Hansen (2005), respectively.
Comparing Density Forecast Models

the reality check p-values of White (2000) and Hansen (2005), respectively. For the S&P500 data, with the p-value of 0.007, the null hypothesis that no competing model is superior to the benchmark is strongly rejected. For the NASDAQ data, the null hypothesis may be rejected at the 10% level with Hansen’s p-value of 0.097. Hence, the evidence that some of the competing density forecast models considered in the comparison dominate the normal-GARCH model is strong for the S&P500 returns and moderate for the NASDAQ returns. Overall, this is consistent with our conclusion that skewness in the conditional distribution and asymmetric and long memory in the conditional volatility are two salient features of financial return series.

CONCLUSIONS

In this paper we consider a framework to compare density forecast models using the KLIC of a candidate density forecast model with respect to the true density as a loss function. We show that using the KLIC as a loss function amounts to using the (negative) predictive log-likelihood function as a loss function. Even though the true density is unknown, our test is in fact to compare models based on the KLIC distances of these models to the true density, which thus enables us to assess which density forecast model can better approximate the true density. While they are asymptotically equivalent, the PIT-based KLIC is best suited for evaluating a single density forecast model and the KLIC based on the return series is best suited for comparing multiple models.

There have been several efforts in constructing statistical methods to compare density forecast models in the literature. Compared to related tests by Sarno and Valente (2004), using the nonparametric kernel-based density, and by Corradi and Swanson (2006), using the Kolmogorov statistics, our approach using the likelihoods is both computationally and conceptually more straightforward. We also show that one of the significant merits of using the KLIC as a loss function in comparing density forecast models is that parameter estimation uncertainty does not complicate asymptotic inference due to the use of the same objective function for the in-sample parameter estimation and the out-of-sample forecast validation.

Our empirical findings based on the daily S&P500 and NASDAQ return series confirm the recent evidence on skewness in financial return distributions. We also find strong evidence for supporting asymmetry and long memory in the conditional volatility.

The method discussed in this paper can also be used to compare alternative credit risk models, to compare density forecasts of asset portfolios, and to compare simulated densities from alternative economic models.

APPENDIX A: PARAMETRIC DISTRIBUTIONS

In the following, \( f(z; \theta) \) with parameter(s) \( \theta \), if any, is the standardized density function of the standardized residuals \( \{z_t\} \).

1. Normal

\[
f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \gamma_z = 0
\]
2. Student $t$

$$f(z; v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi(v-2)}} \frac{1}{\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{z^2}{v-2}\right)^{-\frac{v+1}{2}}, \quad v > 2, \quad \gamma_2 = \frac{6}{v-4}, \quad v > 4$$

3. GED

$$f(z; v) = \frac{v}{\lambda^2} \frac{\exp(-0.5|z/\lambda|)}{(v-2)^{(v-1)/2} \Gamma(1/v)} \gamma_2 = \frac{(1/v)\Gamma(5/v)}{[\Gamma(3/v)]^2}$$

4. Laplace

$$f(z) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|z|), \quad \gamma_2 = 3$$

5. Double Weibull

$$f(z; a) = \frac{a}{2\sigma} \left|\frac{z}{\sigma}\right|^{a-1} \exp\left(-\left|\frac{z}{\sigma}\right|^a\right), \quad a > 0, \quad \sigma = \frac{1}{\Gamma\left(\frac{a+2}{a}\right)}, \quad \gamma_2 = \frac{(a+4)/a}{\gamma(2+a)/a} - 3$$

6. Skewed $t$

$$f(z; \xi, v) = \begin{cases} \frac{2s}{\xi + \xi^{-1}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{\pi(v-2)}} \left(1 + \frac{\xi^2(sz+m)^2}{v-2}\right)^{-\frac{v+1}{2}} & \text{if } z \leq -\frac{m}{s}, \\ \frac{2s}{\xi + \xi^{-1}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{\pi(v-2)}} \left(1 + \frac{\xi^{-2}(sz+m)^2}{v-2}\right)^{-\frac{v+1}{2}} & \text{if } z > -\frac{m}{s} \end{cases}$$

$$\gamma_1 = \frac{(v-2)^{3/2}(\xi^2-1)(\xi^4+1)\Gamma[(v-3)/2]}{\sqrt{\pi}\xi^3s^3\Gamma(v/2)} - \frac{m(m^2+3s)}{s^3},$$

$$\gamma_2 = \frac{3(\xi^5+\xi^{-5})(v-2)}{(\xi^2+\xi^{-2})(v-4)s^4} + \frac{3m^2(m^2+2s)}{s^4} - \frac{4m(v-2)^{3/2}(\xi^2-1)(\xi^4+1)\Gamma[(v-3)/2]}{\sqrt{\pi}\xi^3s^4\Gamma(v/2)}$$
where $\xi > 0$, $v > 2$, $I = 1$ ($z \geq -m/s$), $m = \Gamma((v-1)/2)\sqrt{(v-2)/\pi} (\xi - 1/\xi)/\Gamma(v/2)$, $s = \sqrt{(\xi^2 + 1/\xi^2 - 1)m^2}$, $\gamma_1$ exists if $v > 3$ and $\gamma_2$ exists if $v > 4$.

7. Inverse hyperbolic sine

$$f(z; \lambda, \delta) = \frac{s}{\sqrt{2\pi[(z\lambda + \mu)^2 + 1]^2}} \exp\left(-\frac{[\sinh^{-1}(z\lambda + \mu) - \lambda]^2}{2\delta^2}\right),$$

$$\gamma_1 = \frac{\omega^{v/2}(\omega - 1)^2[\omega(\omega + 2)\sinh(3\lambda) + 3\sinh(\lambda)]}{4s^3},$$

$$\gamma_2 = \frac{(\omega - 1)^2[\omega^2(\omega^4 + 2\omega^3 + 3\omega^2 - 3)\cosh(4\lambda) + 4\omega^2(\omega + 2)\cosh(2\lambda) + 3(\omega + 1)]}{8s^4} - 3$$

where $\delta > 0$, $\mu = \omega^{v/2} \sinh(\lambda)$, $s = \sqrt{(w-1)[w\cosh(2\lambda) + 1]/2}$, $\omega = \exp(\delta^2)$. Note that

$$\sinh(x) = \frac{e^x - e^{-x}}{2}, \sinh^{-1}(x) = \ln(x + \sqrt{x^2 + 1}), \cosh(x) = \frac{e^x + e^{-x}}{2}$$

10. Mixture of normals

$$f(z; \alpha_1, \mu_1, \sigma_1) = p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(z - \mu_1)^2}{2\sigma_1^2}\right) + p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(z - \mu_2)^2}{2\sigma_2^2}\right),$$

$$\gamma_1 = p_1(3\mu_1\sigma_1^2 + \mu_1^3) + p_2(3\mu_2\sigma_2^2 + \mu_2^3),$$

$$\gamma_2 = p_1(6\mu_1^2\sigma_1^2 + \mu_1^4) + p_2(6\mu_2^2\sigma_2^2 + \mu_2^4 + 3\sigma_2^4) - 3$$

where $p_2$, $\mu_2$, and $\sigma_2$ are determined through the constraints $p_1 + p_2 = 1$, $p_1 \geq 0$, $p_2 \geq 0$, $\sum_{i=1}^{2} p_i \mu_i = 0$, and $\sum_{i=1}^{2} p_i (\mu_i^2 + \sigma_i^2) = 1$

11. Double gamma

$$f(z; \alpha_1, \alpha_2, \rho) = \begin{cases} 
\frac{(1 - \rho)\lambda_1^\rho}{\Gamma(\alpha_1)} |z|^{\alpha_1 - 1} \exp(-\lambda_1|z|) & \text{if } z \leq 0, \\
\frac{\rho\lambda_2^{\rho}}{\Gamma(\alpha_2)} z^{\alpha_2 - 1} \exp(-\lambda_2z) & \text{if } z > 0 
\end{cases}$$

$$\gamma_1 = \frac{\rho\Gamma(3 + \alpha_2)}{\lambda_2^2 \Gamma(\alpha_2)} - \frac{(1 - \rho)\Gamma(3 + \alpha_1)}{\lambda_1^2 \Gamma(\alpha_1)},$$

$$\gamma_2 = \frac{\rho\Gamma(4 + \alpha_2)}{\lambda_2^4 \Gamma(\alpha_2)} - \frac{(1 - \rho)\Gamma(4 + \alpha_1)}{\lambda_1^4 \Gamma(\alpha_1)} - 3$$

where $1 > \rho > 0$, $\alpha_1 > 0$, $\alpha_2 > 0$,
\[ \lambda_1 = \frac{p\alpha_1\lambda_2}{(1-p)\alpha_2}, \quad \lambda_2 = \sqrt{\frac{\alpha_2(1-p)[(\alpha_1+1)\alpha_2(1-p)]}{p\alpha_1} + (\alpha_2+1)} \]

12. Gram–Charlier/Edgeworth–Sargan density

\[ f_j(z; \gamma_1, \gamma_2) = \left[ 1 + \frac{\gamma_1}{6} H_3(z) + \frac{\gamma_2}{24} H_4(z) \right] \phi(z) \]

where \( 4 \geq \gamma_2 \geq 0, \gamma_1' \geq \gamma_1 \geq \gamma_1' \), \( H_3(z) = z^3 - 3z, \quad H_4(z) = z^4 - 6z^2 + 3. \)

APPENDIX B: VOLATILITY SPECIFICATION

In the following, in some cases we use \( \alpha(L) = \sum p_i \alpha_j L^i, \beta(L) = \sum q_i \beta_j L^i \), where \( L \) is the back-shift operator. Similarly, \( \phi(L) \) is a polynomial in \( L \).

1. GARCH\((p, q)\)

\[ \sigma_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2, \]
\[ \omega > 0, \alpha_i \geq 0, \beta_j \geq 0, \sum_{i=1}^p \beta_i + \sum_{j=1}^q \alpha_j < 1 \]

2. GJR\((p, q)\)

\[ \sigma_t^2 = \omega + \sum_{i=1}^q (\alpha_i \varepsilon_{t-i}^2 + \phi_i S_{L,i} \varepsilon_{t-i}^2) + \sum_{j=1}^p \beta_j \sigma_{t-j}^2, \]
\[ S_{L,i} = \begin{cases} 1 & \text{if } \varepsilon_{t-i} < 0 \\ 0 & \text{if } \varepsilon_{t-i} \geq 0 \end{cases} \]
\[ \omega > 0, \beta_j > 0, \alpha_i + \phi_i > 0 \]

3. APARCH\((p, q)\)

\[ \sigma_t^\delta = \omega + \sum_{i=1}^q \alpha_i (|\varepsilon_{t-i}| - \phi_i \varepsilon_{t-i})^\delta + \sum_{j=1}^p \beta_j \sigma_{t-j}^\delta, \]
\[ \omega > 0, \beta_j > 0, \alpha_j > 0, 1 > \phi_i > -1, \delta > 0 \]

4. EGARCH\((p, q)\)

\[ \ln \sigma_t^2 = \omega + \left[ 1 - \beta(L) \right] \left[ 1 + \alpha(L) \phi_1 \varepsilon_{t-1} + \phi_2 (|\varepsilon_{t-1}| - E[Z_{t-1}])] \right] \]

5. STGARCH\((p, q)\)

\[
\sigma_i^2 = \omega + \sum_{j=1}^{p} \beta_j \sigma_{i-d}^2 \sum_{j=1}^{q} \left[ \alpha_{ij} \varepsilon_{i-j}^2 + \alpha_{2j} \varepsilon_{i-j}^2 F(\varepsilon_{i-d}) \right]
\]

\[
F(\varepsilon_{i-d}) = \frac{1}{1 + \exp(\varphi \varepsilon_{i-d})} - \frac{1}{2},
\]

\(\omega > 0, \beta_j > 0, \alpha_{ij} > \frac{1}{2} \alpha_{2j} > 0, \varphi > 0\)

6. FIGARCH\((p, d, q)\)/HYGARCH\((p, d, q)\)

Bollerslev and Mikkelsen (1996) introduce a FIGARCH model as follows (by setting \(k = 1\), where \(k\) is an additional parameter for its generalization HYGARCH)

\[
\sigma_i^2 = \omega \left[ 1 - \beta(L) \right]^d \left( 1 - \phi(L) \right) \left( 1 + \kappa \left( 1 - L^d - 1 \right) \right) \varepsilon_i^2
\]

\[
= \omega \left[ 1 - \beta(L) \right]^d \left( 1 - \phi(L) \right) \left( \lambda(L) \right) \varepsilon_i^2
\]

If we rewrite \(\phi(L) = 1 - \phi^*(L)\) (note that the definition of \(\phi(L) = (1 - \alpha(L) - \beta(L)) (1 - L)^{-1}\) implies that it has a constant term), we have

\[
\lambda(L) = 1 - \beta(L) - (1 - \phi^*(L)) \left( 1 + \kappa \left( 1 - L^d - 1 \right) \right)
\]

\[
= 1 - \beta(L) - \kappa \left( 1 - L^d - 1 \right) + \phi^*(L) \kappa \left( 1 - L^d - 1 \right)
\]

\[
= -\beta(L) + \phi^*(L) - \kappa \left( 1 - L^d - 1 \right) + \phi^*(L) \kappa \left( 1 - L^d - 1 \right)
\]

where

\[
(1 - L^d - 1) = \sum_{k=1}^{\infty} \delta_k L^k, \quad \delta_k = \begin{cases} \frac{-d}{(k-1-d)} \delta_{k-1}, & k = 1 \\ \delta_{k-1}, & k \geq 2 \end{cases}
\]

Hence we have

\[
\lambda(L) = -\beta(L) + \phi^*(L) - \kappa \sum_{k=1}^{\infty} \delta_k L^k + \kappa \phi^*(L) \sum_{k=1}^{\infty} \delta_k L^k
\]

\[
= -\beta(L) + \phi^*(L) - \kappa \sum_{k=1}^{\infty} \delta_k L^k + \kappa \sum_{k=1}^{\infty} \delta_k^* L^k
\]

where

\[
\delta_k^* = \begin{cases} 0 & k = 1 \\ \sum_{j=1}^{k-1} \phi_j^* \delta_{k-j} & k = 2, \ldots, q + 1 \\ \sum_{j=1}^{q} \phi_j^* \delta_{k-j} & k > q + 1 \end{cases}
\]
Therefore, we can define a HYGARCH\((p, d, q)\) model as

\[
\sigma_t^2 = \omega + \beta(L)\sigma_t^2 + \left[ \phi^*(L) - \beta(L) + \sum_{i=1}^q \pi_i L^i \right] \varepsilon_t^2,
\]

\[
\pi_k = \kappa\left( \delta_k^\theta - \delta_i \right), \quad \beta(L) = \sum_{i=1}^q \beta_i L^i, \quad \phi^*(L) = \sum_{i=1}^q \phi_i^* L^i
\]

The parameters are \(\theta = (\omega, \beta', \phi^*, \kappa, d')\). For a HYGARCH\((1, d, 1)\) model, the following constraints are sufficient for positiveness of \(\sigma_t^2\): \(\phi^* - \beta + \kappa d \geq 0, \phi^* \leq 0, \beta \geq 0, \omega \geq 0\).

7. FIAPARCH\((p, d, q)\)/HYAPARCH\((p, d, q)\)

\[
\sigma_t^2 = \omega + (1 - [1 - \beta(L)]^{-1}) \phi(L) (1 + \kappa [(1 - L)^d - 1]) (|\varepsilon_t| - \varphi \varepsilon_t)^d
\]

which can be rewritten as

\[
\sigma_t^2 = \omega^* + \beta(L)\sigma_t^2 + \left[ \phi^*(L) - \beta(L) + \sum_{i=1}^q \pi_i L^i \right] (|\varepsilon_t| - \varphi \varepsilon_t)^d
\]

where \(\phi^*(L), \beta(L), \) and \(\pi_k\) are defined in FIGARCH/HYGARCH. The parameters are \(\theta = (\omega^*, \beta', \phi^*, \varphi, \kappa, d)\). For a HYAPARCH\((1, d, 1)\) model, the following constraints are sufficient for positiveness of \(\sigma_t^2\): \(\varphi > -1, \delta > 0, \phi^* - \beta + \kappa d \geq 0, \phi^* \leq 0, \beta \geq 0, \omega^* \geq 0\).

8. CGARCH\((1,1)\)

\[
\sigma_t^2 = \Omega_t + \beta(\sigma_{t-1}^2 - \Omega_{t-1}) + \alpha(\varepsilon_{t-1}^2 - \Omega_{t-1}^2)
\]

\[
\Omega_t = \omega + \rho \Omega_{t-1} + \phi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2),
\]

\(1 > \rho > \beta + \alpha, \beta > \phi > 0, \alpha > 0, \omega > 0\)

where \(\Omega_t\) is the long-run volatility.

ACKNOWLEDGEMENT

We would like to thank the editors (Terence Mills and Allan Timmermann), a referee, Jin-Seo Cho, Alexei V. Egorov, Eric Ghysels, Lutz Kilian, Essie Maasoumi, Norm Swanson, Hal White, and Ximing Wu, as well as seminar participants at the 13th annual conference of the Midwest Econometrics Group (Columbia 2003), Econometric Society Winter Meeting (San Diego 2004), NBER/NSF Time Series Conference (Dallas 2004), International Symposium on Forecasting (San Antonio 2005), EC\textsuperscript{2} (Istanbul 2005), University of British Columbia, and Université Catholique de Louvain (CORE), for useful discussions and comments. We also thank Canlin Li for providing the CRSP data. All remaining errors are our own.
REFERENCES


Authors' biographies:

Yong Bao is Assistant Professor of Economics at the University of Texas, San Antonio. He earned his PhD in Economics in 2004 from the University of California, Riverside. His major research fields are time series analysis and finite sample econometrics.

Tae-Hwy Lee is Professor of Economics at the University of California, Riverside. He earned a PhD in Economics (1990) from UCSD. His research interests are time series econometrics and financial econometrics. He is currently Visiting Professor of Economics at California Institute of Technology for 2005–2006.

Burak Saltoğlu is Professor of Economics at Marmara University, Turkey. He earned a PhD in Economics (1996) from the University of Essex, UK. His major research fields are financial econometrics and financial risk management.

Authors' addresses:

Yong Bao, Department of Economics, University of Texas, San Antonio, TX 78249, USA.

Tae-Hwy Lee, Department of Economics, University of California, Riverside, CA 92521, USA.

Burak Saltoğlu, Department of Economics, Bosphorous University, 34342 Bebek, Istanbul, Turkey.