Comparing Density Forecast Models*

Yong Bao†
University of Texas at San Antonio

Tae-Hwy Lee‡
Univeristy of California, Riverside and California Institute of Technology

Burak Saltoğlu§
Marmara University, Turkey

Oct 2006

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.

Key Words: Density forecast comparison, KLIC, Predictive log-likelihood, Reality check.

JEL Classification: C3, C5, G0.

*Previous versions of this paper have been circulated with the title, "A Test for Density Forecast Comparison with Applications to Risk Management." 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, Econometric Society San Diego Winter Meeting, NBER/NSF Dallas Time Series Conference, International Symposium on Forecasting, EC^2, 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.

†Department of Economics, University of Texas at San Antonio, San Antonio, TX 78249, U.S.A. Tel: +1 (210) 458-5303. Fax: +1 (210) 458-5837. Email: yong.bao@utsa.edu.

‡Corresponding author. Department of Economics, University of California, Riverside, CA 92521, U.S.A. Tel: +1 (951) 827-1509. Fax: +1 (951) 827-5685. Email: tae.lee@ucr.edu.

§Department of Economics, Marmara University, Istanbul, 81040, Turkey. Tel: +90 (216)3368487. Fax: +90 (216)3464356. Email: saltoglu@marmara.edu.tr.

1 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 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.

[1] They 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 evalution techniques."

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 ($= 10 \times 8$) 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.

This paper is organized as follows. In Section 2, 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. Section 3 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.[2] In Section 4 we present our empirical study of the daily S&P500 and NASDAQ return series. Finally, Section 5 concludes. More detailed description of the distribution and volatility models described in Section 4 is given in the appendix.

2 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

[2]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.

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 in Section 3). 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).

2.1 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 $\mathbf{Y}^T \equiv \{Y_\tau : \Omega \to \mathbb{R},\, \tau = 1, 2, \ldots, T\}$ on a complete probability space $(\Omega, \mathcal{F}_T, P_0^T)$, where $\Omega = \mathbb{R}^T \equiv \times_{\tau=1}^T \mathbb{R}$ and $\mathcal{F}_T = \mathcal{B}(\mathbb{R}^T)$ is the Borel σ-field generated by the open sets of $\mathbb{R}^T$, and the *joint* probability measure $P_0^T(B) \equiv P_0[\mathbf{Y}^T \in B]$, $B \in \mathcal{B}(\mathbb{R}^T)$ completely describes the stochastic process. A sample of size T is denoted as $\mathbf{y}^T \equiv (y_1, \ldots, y_T)'$.

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

Following White (1994, Section 2.2), we define a probability model $\mathcal{P}$ as a collection of distinct probability measures on the measurable space $(\Omega, \mathcal{F}_T)$. A probability model $\mathcal{P}$ is said to be correctly specified for $\mathbf{Y}^T$ if $\mathcal{P}$ contains P_0^T. Our goal is to evaluate and compare a set of parametric probability models $\{P_\theta^T\}$, where $P_\theta^T(B) \equiv P_\theta[Y^T \in B]$. Suppose there exists a measurable Radon-Nikodým density $f^T(\mathbf{y}^T) = dP_\theta^T/d\nu^T$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ is a finite-dimensional vector of parameters and is assumed to be identified on $\boldsymbol{\Theta}$, a compact subset of $\mathbb{R}^k$: see White (1994, Theorem 2.6).

In the context of forecasting, instead of the joint density $g^T(\mathbf{y}^T)$, we consider forecasting the *conditional* density of $\mathbf{Y}^t$, given the information $\mathcal{F}_{t-1}$ generated by $\mathbf{Y}^{t-1}$. Let $\varphi_t(y_t) \equiv \varphi_t(y_t|\mathcal{F}_{t-1}) \equiv g^t(\mathbf{y}^t)/g^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\varphi_1(y_1) \equiv \varphi_1(y_1|\mathcal{F}_0) \equiv g^1(\mathbf{y}^1) = g^1(y_1)$. Thus the goal is to forecast the (true, unknown) conditional density $\varphi_t(y_t)$.

For this, we use a one-step-ahead conditional density forecast model $\psi_t(y_t; \boldsymbol{\theta}) \equiv \psi_t(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) \equiv f^t(\mathbf{y}^t)/f^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\psi_1(y_1) \equiv \psi_1(y_1|\mathcal{F}_0) \equiv f^1(\mathbf{y}^1) = f^1(y_1)$. If $\psi_t(y_t; \boldsymbol{\theta}_0) = \varphi_t(y_t)$ almost surely for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, 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.

2.2 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,

$$\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}\right)=\mathbb{E}_{\varphi_t}[\ln\varphi_t\left(y_t\right)-\ln\psi_t\left(y_t;\boldsymbol{\theta}\right)],$$

where the expectation is with respect to the true conditional density $\varphi_t\left(\cdot|\mathcal{F}_{t-1}\right)$, $\mathbb{E}_{\varphi_t}\ln\varphi_t\left(y_t|\mathcal{F}_{t-1}\right)<\infty$, and $\mathbb{E}_{\varphi_t}\ln\psi_t\left(y_t|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)<\infty$. Following White (1982, 1994), we define the distance between a density model and the true density as the minimum of the KLIC

$$\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}_{t-1}^{*}\right)=\mathbb{E}_{\varphi_t}\left[\ln\varphi_t\left(y_t\right)-\ln\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^{*}\right)\right], \tag{1}$$

where $\boldsymbol{\theta}_{t-1}^{*}=\arg\min\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}\right)$ is the pseudo-true value of $\boldsymbol{\theta}$ (Sawa, 1978; White, 1982).[3] We assume that $\boldsymbol{\theta}_{t-1}^{*}$ is an interior point of $\boldsymbol{\Theta}$. The smaller this distance is, the closer the density forecast $\psi_t\left(\cdot|\mathcal{F}_{t-1};\boldsymbol{\theta}_{t-1}^{*}\right)$ is to the true density $\varphi_t\left(\cdot|\mathcal{F}_{t-1}\right)$.[4]

However, $\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}_{t-1}^{*}\right)$ is unknown since $\boldsymbol{\theta}_{t-1}^{*}$ is not observable. We need to estimate $\boldsymbol{\theta}_{t-1}^{*}$. 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 $\{y_{t-1},\ldots,y_{t-R}\}_{t=R+1}^{T}$ to estimate the unknown parameter vector $\boldsymbol{\theta}_{t-1}^{*}$ and denote the estimate as $\hat{\boldsymbol{\theta}}_{R,t-1}$ (where the subscript R denotes the size of the in-sample period).[5] Under some regularity conditions, we can consistently estimate

[3] We use the word "distance" loosely because KLIC does not satisfy some basic properties of a metric, i.e., $\mathbb{I}\left(\psi_1:\psi_2\right)\neq\mathbb{I}\left(\psi_2:\psi_1\right)$ 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 in Section 4), 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.

[5] There are different schemes to estimate $\boldsymbol{\theta}_{t-1}^{*}$, namely, recursive, fixed, and rolling schemes. That is, $\boldsymbol{\theta}_{t-1}^{*}$ can be estimated

$\boldsymbol{\theta}^*_{t-1}$ by $\hat{\boldsymbol{\theta}}_{R,t-1}$ by maximizing $R^{-1}\sum_{s=t-R+1}^{t}\ln\psi_s\left(y_s;\boldsymbol{\theta}\right)$: see White (1994, Theorem 2.12 and Theorem 3.4) for the sets of conditions for the existence and consistency of $\hat{\boldsymbol{\theta}}_{R,t-1}$.

Using $\{\hat{\boldsymbol{\theta}}_{R,t-1}\}_{t=R+1}^{T}$, we can obtain the out-of-sample estimate of $\mathbb{E}\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}^*_{t-1}\right)$ (where $\mathbb{E}(\cdot)$ is the unconditional expectation) by

$$\mathbb{I}_{R,n}(\varphi:\psi)\equiv\frac{1}{n}\sum_{t=R+1}^{T}\ln[\varphi_t(y_t)/\psi_t(y_t;\hat{\boldsymbol{\theta}}_{R,t-1})] \tag{2}$$

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

$$\mathbb{I}_{R,n}(\varphi:\psi)=\frac{1}{n}\sum_{t=R+1}^{T}\ln\left[\varphi_t(y_t)/\psi_t\left(y_t;\boldsymbol{\theta}^*_{t-1}\right)\right]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[\psi_t\left(y_t;\boldsymbol{\theta}^*_{t-1}\right)/\psi_t(y_t;\hat{\boldsymbol{\theta}}_{R,t-1})], \tag{3}$$

where the first term in $\mathbb{I}_{R,n}(\varphi:\psi)$ measures model misspecification (the distance between the optimal density $\varphi_t(y_t)$ and the model $\psi_t\left(y_t;\boldsymbol{\theta}^*_{t-1}\right)$) and the second term measures parameter estimation uncertainty due to the distance between $\boldsymbol{\theta}^*_{t-1}$ and $\hat{\boldsymbol{\theta}}_{R,t-1}$.[6]

2.3 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

$$u_t\equiv\int_{-\infty}^{y_t}\psi_t(y|\mathcal{F}_{t-1};\boldsymbol{\theta}^*_{t-1})dy. \tag{4}$$

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

Our purpose of invoking the PIT is to make $\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}^*_{t-1}\right)$ operational as $\varphi_t\left(\cdot\right)$ is unknown. We take a further transform

$$x_t\equiv\Phi^{-1}(u_t), \tag{5}$$

based on the whole subsample $\{y_{t-1},\ldots,y_1\}$, or a fixed sample $\{y_R,\ldots,y_1\}$, or a rolling sample $\{y_{t-1},\ldots,y_{t-R}\}$, 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

$$\ln\left[\varphi_t(y_t)/\psi_t\left(y_t;\boldsymbol{\theta}^*_{t-1}\right)\right] = \ln\left[p_t\left(x_t\right)/\phi\left(x_t\right)\right], \tag{6}$$

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

$$\hat{x}_t \equiv \Phi^{-1}\left(\int_{-\infty}^{y_t} \psi_t(y|\mathcal{F}_{t-1};\hat{\boldsymbol{\theta}}_{R,t-1})dy\right), \tag{7}$$

$t = R+1,\ldots,T$.

We have transformed the departure of $\psi_t(\cdot;\boldsymbol{\theta})$ from the unknown true density $\varphi_t(\cdot)$ 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 simpler: 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 $\psi_t(\cdot;\boldsymbol{\theta})$ there is no guarantee that the postulated $\psi_t(\cdot;\boldsymbol{\theta})$ will nest the unknown $\varphi_t(\cdot)$. We follow Berkowitz (2001) by specifying $\hat{x}_t$ as an AR(L) process

$$\hat{x}_t = \boldsymbol{\rho}' X_{t-1} + \sigma\eta_t, \tag{8}$$

where $X_{t-1} = (1, \hat{x}_{t-1}, \cdots, \hat{x}_{t-L})'$, $\boldsymbol{\rho}$ is an $(L+1)\times 1$ vector of parameters, and η_t is IID distributed and independent of X_{t-1}. In Berkowitz (2001), η_t is further assumed to be normally distributed. Clearly this is restrictive. A remedy for this is to specify a flexible alternative distribution for η_t, say $h(\eta_t;\boldsymbol{\vartheta}_\eta)$, where $\boldsymbol{\vartheta}_\eta$ is a vector of distribution parameters such that when $\boldsymbol{\vartheta}_\eta = \boldsymbol{\vartheta}^\dagger_\eta$, $h(\cdot;\boldsymbol{\vartheta}^\dagger_\eta)$ 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 $\boldsymbol{\vartheta} = (\boldsymbol{\rho}', \sigma, \boldsymbol{\vartheta}'_\eta)'$, say $\boldsymbol{\rho} = \mathbf{0}$, $\sigma = 1$, and $\boldsymbol{\vartheta}_\eta = \boldsymbol{\vartheta}^\dagger_\eta$. In particular, we assume that $\{\hat{x}_t\}_{t=R+1}^T$ follows the AR process (8) with η_t IID distributed with the semi-nonparametric (SNP) density function of order K (Gallant and Nychka, 1987),

$$h\left(\eta_t;\boldsymbol{\vartheta}_\eta\right) = \frac{\left(\sum_{k=0}^K r_k \eta_t^k\right)^2 \phi\left(\eta_t\right)}{\int_{-\infty}^{+\infty}\left(\sum_{k=0}^K r_k u^k\right)^2 \phi\left(u\right)du}, \tag{9}$$

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

$$p_t\left(\hat{x}_t;\boldsymbol{\vartheta}\right) = \frac{h\left(\frac{\left(\hat{x}_t - \boldsymbol{\rho}' X_{t-1}\right)}{\sigma};\boldsymbol{\vartheta}_\eta\right)}{\sigma}, \tag{10}$$

which degenerates to $\phi(\hat{x}_t) = \phi(\eta_t)$ by setting $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^\dagger \equiv (\mathbf{0}', 1, \mathbf{0}')'$. The parameter vector $\boldsymbol{\vartheta}$ can be estimated by $\hat{\boldsymbol{\vartheta}}_n = (\hat{\boldsymbol{\rho}}'_n, \hat{\sigma}_n, \hat{\boldsymbol{\vartheta}}'_{\eta n})'$ that maximizes $n^{-1}\sum_{t=R+1}^T \ln p_t(\hat{x}_t;\boldsymbol{\vartheta})$.

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

$$\begin{aligned}
\mathbb{I}_{R,n}(p:\phi) &\equiv \frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\hat{\boldsymbol{\vartheta}}_n)/\phi(\hat{x}_t)] \qquad (11)\\
&= \frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\hat{\boldsymbol{\vartheta}}_n)/p_t(\hat{x}_t;\boldsymbol{\vartheta})]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\boldsymbol{\vartheta})/p_t(x_t;\boldsymbol{\vartheta})]\\
&\quad+\frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(x_t;\boldsymbol{\vartheta})/\phi(x_t)]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[\phi_t(x_t)/\phi(\hat{x}_t)]\\
&= \frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\hat{\boldsymbol{\vartheta}}_n)/p_t(\hat{x}_t;\boldsymbol{\vartheta})]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\boldsymbol{\vartheta})/p_t(x_t;\boldsymbol{\vartheta})]\\
&\quad+\frac{1}{n}\sum_{t=R+1}^{T}\ln\left[\varphi_t(y_t)/\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^{*}\right)\right]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[\phi_t(x_t)/\phi(\hat{x}_t)],
\end{aligned}$$

where the last equality comes from (6). The third term measures the distance between $\varphi_t(y_t)$ and the forecast model $\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^{*}\right)$ (model misspecification), the second and fourth terms measure the distance between $\boldsymbol{\theta}_{t-1}^{*}$ and $\hat{\boldsymbol{\theta}}_{R,t-1}$ (parameter estimation uncertainty), and the first term measures the sampling variation in the estimation of $p_t(\hat{x}_t;\hat{\boldsymbol{\vartheta}}_n)$.

2.4 Evaluating Density Forecast Models

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

$$\begin{aligned}
\mathbb{I}_{R,n}(p:\phi) &= \mathbb{I}_{R,n}(\varphi:\psi) \qquad (12)\\
&\quad+\frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\hat{\boldsymbol{\vartheta}}_n)/p_t(\hat{x}_t;\boldsymbol{\vartheta})]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\boldsymbol{\vartheta})/p_t(x_t;\boldsymbol{\vartheta})]\\
&\quad-\frac{1}{n}\sum_{t=R+1}^{T}\ln[\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^{*}\right)/\psi_t(y_t;\hat{\boldsymbol{\theta}}_{R,t-1})]+\frac{1}{n}\sum_{t=R+1}^{T}\ln[\phi_t(x_t)/\phi(\hat{x}_t)].
\end{aligned}$$

Note that $\mathbb{I}_{R,n}(p:\phi)\to_{a.s.}\mathbb{I}_{R,n}(\varphi:\psi)$ as $n\to\infty, R\to\infty$, since $\hat{\boldsymbol{\theta}}_{R,t-1}\to_{a.s.}\boldsymbol{\theta}_{t-1}^{*}$ as $R\to\infty$ (for all $t=R+1,\ldots,T$) and $\hat{\boldsymbol{\vartheta}}_n\to_{a.s.}\boldsymbol{\vartheta}^{*}$ as $n\to\infty$.

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

$$LR_n \equiv 2\sum_{t=R+1}^{T}\ln[p_t(\hat{x}_t;\hat{\boldsymbol{\vartheta}}_n)/\phi(\hat{x}_t)] = 2\times n\times\mathbb{I}_{R,n}(p:\phi), \qquad (13)$$

which is asymptotically χ^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 $\boldsymbol{\rho}=\mathbf{0},\sigma=1$, and $\boldsymbol{\vartheta}_\eta=(r_1,\cdots,r_K)'=\mathbf{0}'$. This can be regarded as a generalized version of the LR statistic of Berkowitz (2001).

3 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 $\psi_t^j(y_t; \hat{\boldsymbol{\theta}}_{R,t-1}^j)$. 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.

3.1 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 $\mathbb{I}_{R,n}(\varphi : \psi)$ or $\mathbb{I}_{R,n}(p : \phi)$ 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

$$\bar{d}_{j,n}(\mathbf{x}) \equiv \mathbb{I}_{R,n}(p^0 : \phi) - \mathbb{I}_{R,n}(p^j : \phi). \tag{14}$$

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

$$\begin{aligned} \bar{d}_{j,n}(\mathbf{y}) &\equiv \mathbb{I}_{R,n}(\varphi : \psi^0) - \mathbb{I}_{R,n}(\varphi : \psi^j) \\ &= \frac{1}{n}\sum_{t=R+1}^{T} \ln[\psi_t^j(y_t; \hat{\boldsymbol{\theta}}_{R,t-1}^j)/\psi_t^0(y_t; \hat{\boldsymbol{\theta}}_{R,t-1}^0)] \\ &= \frac{1}{n}\sum_{t=R+1}^{T} d_{j,t}(y_t, \hat{\boldsymbol{\beta}}_{R,t-1}^j). \end{aligned} \tag{15}$$

where $d_{j,t}(y_t, \boldsymbol{\beta}^j) \equiv \ln[\psi_t^j(y_t; \boldsymbol{\theta}^j)/\psi_t^0(y_t; \boldsymbol{\theta}^0)]$ and $\boldsymbol{\beta}^j = (\boldsymbol{\theta}^{0\prime}, \boldsymbol{\theta}^{j\prime})'$.

For the purpose of comparing the density forecast models, several advantages of using $\bar{d}_{j,n}(\mathbf{y})$ over $\bar{d}_{j,n}(\mathbf{x})$ may be pointed out. One is that it is simpler as it does not involve the true density $\varphi_t(y_t)$ since it is cancelled out in forming the loss-differential (15). Thus the loss-differential $\bar{d}_{j,n}(\mathbf{y})$ can be computed even though $\varphi_t(y_t)$ 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, $\mathbb{I}_{R,n}(p : \phi)$ is more complicated than $\mathbb{I}_{R,n}(\varphi : \psi)$.

It is important to note that formulating $\bar{d}_{j,n}(\mathbf{y})$ amounts to using the negative predictive log-likelihood

as a loss function:

$$\bar{d}_{j,n}(\mathbf{y}) = \frac{1}{n}\sum_{t=R+1}^{T}[(-\ln\psi_t^0(y_t;\hat{\boldsymbol{\theta}}_{R,t-1}^0)) - (-\ln\psi_t^j(y_t;\hat{\boldsymbol{\theta}}_{R,t-1}^j))]. \tag{16}$$

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 $\{y_t\}$ (i.e., with $\bar{d}_{j,n}(\mathbf{y})$ in (15) as the loss-differential) should be more powerful compared with that based on $\{\hat{x}_t\}$ (i.e., with $\bar{d}_{j,n}(\mathbf{x})$ in (14) as the loss-differential) due to the estimation error associated with estimating the density of $\{\hat{x}_t\}_{t=R+1}^T$.

3.2 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, $\mathbb{H}_0 : \mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j}) \leq 0$ $(j = 1, \ldots, l)$. We can use the out-of-sample predictive ability tests of Diebold and Mariano (1995) by referring to the standard asymptotic result for $\sqrt{n}\bar{d}_{j,n}$. By a suitable central limit theorem, we have

$$\sqrt{n}[\bar{d}_{j,n}(\mathbf{y}) - \mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j})] \to N(0, \xi^2) \tag{17}$$

in distribution as $n \equiv n(T) \to \infty$ when $T \to \infty$, where $\xi^2 \equiv \lim_{n\to\infty} \mathrm{Var}(\sqrt{n}[\bar{d}_{j,n}(\mathbf{y}) - \mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j})])$. In general, the variance ξ^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 $\hat{\boldsymbol{\theta}}$) and for the out-of-sample forecast validation (using the KLIC as a forecast evaluation criterion).[7]

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

[7] As shown in West (1996, Theorem 4.1), when the derivative of the loss differential evaluated at $\boldsymbol{\beta}_{t-1}^{*j} = ((\boldsymbol{\theta}_{t-1}^{*0})', (\boldsymbol{\theta}_{t-1}^{*j})')'$ is zero, i.e., $F \equiv \mathbb{E}\partial d_{j,t}(y_t, \boldsymbol{\beta}^j)/\partial\boldsymbol{\beta}^j\big|_{\boldsymbol{\beta}_{t-1}^{*j}} = 0$, under some proper regularity conditions, ξ^2 does not depend on parameter estimation uncertainty. We have this condition

$$\begin{aligned} F &\equiv \mathbb{E}\partial d_{j,t}(y_t, \boldsymbol{\beta}^j)/\partial\boldsymbol{\beta}^j\big|_{\boldsymbol{\beta}_{t-1}^{*j}} \\ &= \mathbb{E}\partial \ln[\psi_t^j(y_t;\boldsymbol{\theta}^j)/\psi_t^0(y_t;\boldsymbol{\theta}^0)]/\partial\boldsymbol{\beta}^j\Big|_{\boldsymbol{\beta}_{t-1}^{*j}} \\ &= \mathbb{E}\left[\mathbb{E}_{\varphi_t}\partial\ln\psi_t^j(y_t;\boldsymbol{\theta}^j)/\partial\boldsymbol{\theta}^j\Big|_{\boldsymbol{\theta}_{t-1}^{*j}} - \mathbb{E}_{\varphi_t}\partial\ln\psi_t^0(y_t;\boldsymbol{\theta}^0)/\partial\boldsymbol{\theta}^0|_{\boldsymbol{\theta}_{t-1}^{*0}}\right] = 0 \end{aligned}$$

because both terms inside the brackets are zero as $\boldsymbol{\theta}_{t-1}^{*j}$ maximizes $\mathbb{E}_{\varphi_t}\ln\psi_t^j(y_t;\boldsymbol{\theta}^j)$, which makes parameter estimation uncertainty due to the second term in (3) asymptotically negligible in ξ^2. The last line follows from the law of iterated expectations.

model is better than the benchmark:

$$\mathbb{H}_0 : \max_{1\leq j\leq l} \mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j}) \leq 0. \tag{18}$$

This is a multiple hypothesis, the intersection of the one-sided individual hypotheses $\mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j}) \leq 0$ 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 $\mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j})$ is positive. Under appropriate conditions and defining $\bar{\mathbf{d}}$ to be an $l \times 1$ column vector by stacking $\bar{d}_{j,n}(\mathbf{y})$, then $\sqrt{n}(\bar{\mathbf{d}} - \mathbb{E}(\mathbf{d}^*)) \to N(0, \Omega)$ as $n(T) \to \infty$ when $T \to \infty$, for Ω positive semi-definite, where $\mathbb{E}\mathbf{d}^* = \mathbb{E}\mathbf{d}_t^*$(assuming stationarity) for $\mathbf{d}_t^*$ defined as an $l \times 1$ column vector obtained by stacking the $d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j})$.[8] White's (2000) test statistic for H_0 is formed as

$$\bar{V}_n \equiv \max_{1\leq j\leq l} \sqrt{n}[\bar{d}_{j,n}(\mathbf{y}) - \mathbb{E}d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j})]. \tag{19}$$

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

$$\bar{V}_n^\dagger = \max_{1\leq j\leq l} \sqrt{n}(\bar{d}_{j,n}(\mathbf{y})^\dagger - \bar{d}_{j,n}(\mathbf{y})) \tag{20}$$

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 $\bar{d}_{j,n}(\mathbf{y}) = 0$. This setting makes the null hypothesis the least favorable to the alternative, because setting $\bar{d}_{j,n}(\mathbf{y}) = 0$ guarantees that the statistic satisfies the null hypothesis $\mathbb{E}(\bar{d}_{j,n}(\mathbf{y})^\dagger - \bar{d}_{j,n}(\mathbf{y})) = 0$ 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 $\bar{d}_{j,n}(\mathbf{y})$, 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)

$$\bar{V}_n^\dagger = \max_{1\leq j\leq l} \sqrt{n}(\bar{d}_{j,n}(\mathbf{y})^\dagger - g(\bar{d}_{j,n}(\mathbf{y}))), \tag{21}$$

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

[8] White (2000) does not require that Ω be positive definite (which is required in West 1996), but that Ω 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 nonnested with respect to the benchmark.

variance of $\bar{d}_{j,n}(\mathbf{y})$, i.e.

$$g(\bar{d}_{j,n}(\mathbf{y})) = \begin{cases} 0 & \text{if } \bar{d}_{j,n}(\mathbf{y}) \leq -A_j \\ \bar{d}_{j,n}(\mathbf{y}) & \text{if } \bar{d}_{j,n}(\mathbf{y}) > -A_j \end{cases} \tag{22}$$

where $A_j = \frac{1}{4} n^{-1/4} \sqrt{Var(n^{1/2} \bar{d}_{j,n}(\mathbf{y}))}$ with the variance estimated from the bootstrap samples. When $\mathbb{E} d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j}) = 0$ 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., $\mathbb{E} d_{j,t}(y_t, \boldsymbol{\beta}_{t-1}^{*j}) < 0$ 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.

4 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

$$y_t = \mu_t + \varepsilon_t \equiv \mu_t + z_t \sigma_t, \tag{23}$$

where $\mu_t = \mathbb{E}(y_t | \mathcal{F}_{t-1})$, $\sigma_t^2 = \mathbb{E}(\varepsilon_t^2 | \mathcal{F}_{t-1})$, and $z_t \equiv \varepsilon_t / \sigma_t$. We assume that z_t is IID and μ_t follows an MA(1) process without a constant term (see, e.g., Anderson *et al.*, 2002; Hong and Lee, 2003). A density forecast model based on (23) can be decomposed into two parts: specification of σ_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.

4.1 Distribution Specifications

We can specify the distribution of the standardized residuals $\{z_t\}_{t=1}^T$ by a conditional density function $f(z) = f(z|\mathcal{F}_{t-1}) = f(z|\mathcal{F}_{t-1}; \boldsymbol{\theta}_d)$ with a vector of distribution parameters $\boldsymbol{\theta}_d$. Given $f(\cdot)$ and conditional volatility $\sigma_t^2 = \sigma_t^2(\boldsymbol{\theta}_v)$ with a vector of volatility parameters $\boldsymbol{\theta}_v$, the conditional distribution of y_t is $\psi_t(y_t|\mathcal{F}_{t-1}) = \psi_t(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) = f(z_t)/\sigma_t$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_d', \boldsymbol{\theta}_v')'$ and $\boldsymbol{\theta}_\mu$ is the MA parameter. Throughout, we use γ_1 and γ_2, if they exist, to denote 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 Paretian 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 $\phi(z)$ 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 γ_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.

4.2 Volatility Specifications

The conditional variance σ_t^2 can be specified with various volatility models. We can specify it nonparametrically, 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.

4.3 Empirical Results

In this section, we use two data sets to compare the 80 density forecast models described in the previous subsections (ten different distribution models and eight different volatility models). The two data sets are the daily S&P500 and NASDAQ return series, retrieved from *finance.yahoo.com* and CRSP.[9] They are from January 3, 1990 to June 30, 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

[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 September 11, 2001, we treat it as a long holiday.

(January 3, 1990 to September 24, 1996), the second density forecast is based on observations 2 through $R+1$ (January 4, 1990 to September 25, 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.

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 1 reports $\mathbb{I}_{R,n}(p:\phi)$ (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 1, we observe the following.

Firstly, when we look for the model that gives the minimum value of $\mathbb{I}_{R,n}(p:\phi)$, i.e., the model that best approximates the true density forecast, 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 property 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 enough to approximate the conditional variance of return series.

Table 2 reports the values of the *negative* predictive log-likelihood, $-n^{-1}\sum_{t=R+1}^{T}\ln\psi_t^j(y_t;\hat{\boldsymbol{\theta}}_{R,t-1}^j)$. The conclusions we can draw from Table 2 are consistent with those from Table 1 – there is further evidence for skewness in the conditional distribution, and evidence of asymmetry and long-memory in the conditional volatility. As in Table 1, 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 2 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 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.

5 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 and 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 (2004), 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.

References

Amisano, G. and R. Giacomini. 2005. Comparing density forecasts via weighted likelihood ratio tests. Working paper, University of California, Los Angeles.

Anderson, T. G., L. Benzoni, and J. Lund. 2002. An empirical investigation of continuous-time equity return models. *Journal of Finance* 57: 1239-1284.

Bao, Y., T.-H. Lee, and B. Saltoğlu. 2004. A test for density forecast comparison with applications to risk management. Working paper, Department of Economics, UC Riverside.

Barberis, N. 2000. Investing for the long run when returns are predictable. *Journal of Finance* 55: 225-264.

Bauwens, L., P. Giot, J. Grammig, and D. Veredas. 2004. A comparison of financial duration models via density forecasts. *International Journal of Forecasting* 20: 589-609.

Bawa, V. S., S. J. Brown, and R. W. Klein. 1979. *Estimation Risk and Optimal Portfolio Choice.* North-Holland: New York.

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

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307-327.

Bollerslev, T. and H. O. Mikkelsen. 1996. Modelling and pricing long memory in stock market volatility. *Journal of Econometrics* 73: 151-184.

Choi, P. 2001. Estimation of value at risk using Johnson S_U-normal distribution. Working paper, Texas A&M University.

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: 255-276.

Corradi, V. and N.R. Swanson. 2004. Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, forthcoming.

Davidson, J. 2004. Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *Journal of Business and Economics Statistics* 22: 16-29.

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

Diebold, F. X. and R. Mariano. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13: 253-265.

Ding, Z. and C. W. J. Granger. 1996. Modeling volatility persistence of speculative returns: a new approach. *Journal of Econometrics* 73: 185-215.

Ding, Z., C. W. J. Granger, and R. F. Engle. 1993. A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1: 83-106.

Engle, R. F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50: 987-1008.

Engle, R. F. and G. González-Rivera. 1991. Semiparametric ARCH models. *Journal of Business and Economic Statistics* 9: 345-360.

Engle, R. F. and G. Lee. 1999. A long-run and short-run component model of stock return volatility. In *Cointegration, Causality, and Forecasting*, R. F. Engle, and H. White (eds.), Oxford University Press.

Eberlin, E. and U. Keller. 1995. Hyperbolic distributions in finance. *Bernoulli* 1: 281-299.

Fernández, C. and M. Steel. 1998. On Bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association* 93: 359-371.

Freichs, H. and G. Löffler. 2003. Evaluating credit risk models using loss density forecasts. *Journal of Risk* 5: 1-23.

Gallant, A. R. and D. W. Nychka. 1987. Semi-nonparametric maximum likelihood estimation. *Econometrica* 55: 363-390.

Giacomini, R., and H. White. 2003. Conditional tests for predictive ability. Working paper, University of California, San Diego.

Glosten, L., R. Jagannathan, and D. Runkle. 1993. Relationship between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48: 1779-1801.

González-Rivera, G. 1998. Smooth-transition GARCH models. *Studies in Nonlinear Dynamics and Econometrics* 3: 61-78.

Gourieroux, C. and J. Jasiak. 2001. Local likelihood density estimation and value at risk. Working paper. CREST and CEPREMAP and York University.

Granger, C. W. J. and M. H. Pesaran. 2000a. A decision theoretic approach to forecasting evaluation. In *Statistics and Finance: An Interface*, W. S. Chan, W. K. Li, and Howell Tong (eds.), London: Imperial College Press.

Granger, C. W. J. and M. H. Pesaran. 2000b. Economic and statistical measures of forecast accuracy. *Journal of Forecasting* 19: 537-560.

Hansen, C. B., J. B. McDoanld, and P. Theodossiou. 2000. Some flexible parametric models for skewed and leptokurtic data. Working paper, Brigham Young University and Rutgers University.

Hansen, P. R. 2005. A test for superior predictive ability. *Journal of Business and Economic Statistics* 23: 365-380.

Hong, Y. and T.-H. Lee. 2003. Diagnostic checking for adequacy of nonlinear time series models. *Econometric Theory* 19: 1065-1121.

Jackwerth, J. C. and M. Rubinstein. 1996. Recovering probability distributions from option prices. *Journal of Finance* 51: 1611-1631.

Jondeau, E. and M. Rockinger. 2001. Gram-Charlier densities. *Journal of Economic Dynamics and Control* 25: 1457-1483.

Kandel, S. and R. F. Stambaugh. 1996. On the predictability of stock returns: An asset-allocation perspective. *Journal of Finance* 51: 385-424.

Knight, J. L., S. E. Satchell, and K. C. Tran. 1995. Statistical modelling of asymmetric risk in asset returns. *Applied Mathematical Finance* 2: 155-172.

Kullback, L. and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22: 79-86.

Mauleón, I. and J. Perote. 1995. Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to Student's *t*. *European Journal of Finance* 6: 225-239.

McDonald, J. B. and W. Newey. 1988. Partially adaptive estimation of regression models via the generalized *t* Distribution. *Econometric Theory* 4: 428-457.

McDonald, J. B. and Y. J. Xu. 1995. A generalization of the beta distribution with applications. *Journal of Econometrics* 66: 133-152.

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

Mittnik, S. and T. Rachev. 1993. Modeling asset returns with alternative stable models. *Econometric Reviews* 12: 261-330.

Nelson, D. B. 1991. Conditional heteroscedasticity in asset returns: a new approach. *Econometrica* 59: 347-370.

Panorska, A. K., S. Mittnik, and S. T. Rachev. 1995. Stable GARCH models for financial time series. *Applied Mathematics Letters* 8: 33-37.

Pascual, L., J. Romo, and E. Ruiz. 2001. Effects of parameter estimation on prediction densities: a bootstrap approach. *International Journal of Forecasting* 17: 83-103.

Politis, D. N. and J. P. Romano. 1994. The stationary bootstrap. *Journal of American Statistical Association* 89: 1303-1313.

Poon, S. and C. W. J. Granger. 2003. Forecasting volatility in financial markets. *Journal of Economic Literature* 41: 478-539.

Sarno, L. and G. Valente. 2004. Comparing the accuracy of density forecasts from competing models. *Journal of Forecasting* 23: 541-557.

Sawa, T. 1978. Information criteria for discriminating among alternative regression models. *Econometrica* 46: 1273-1291.

Theodossiou, P. 1998. Financial data and skewed generalized t distribution. *Management Science* 44: 1650-1661.

Venkataraman, S. 1997. Value at risk for a mixture of normal distributions: the use of quasi-Bayesian estimation techniques. *Economic Perspectives*, March, 2-13, Federal Reserve Bank of Chicago.

Vuong, Q. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307-333.

Wang, J. 2001. Generating daily changes in market variables using a multivariate mixture of normal distributions, 283-289. In *Proceedings of the 2001 Winter Simulation Conference,* B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer (eds.), Arlington, Virginia.

West, K. 1996. Asymptotic inference about prediction ability. *Econometrica* 64: 1067-1084.

West, K. D. and M. W. McCracken. 1998. Regression-based tests of predictive ability. *International Economic Review* 39: 817-840.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1-25.

White, H. 1994. *Estimation, Inference, and Specification Analysis*. Cambridge: Cambridge University Press.

White, H. 2000. A reality check for data snooping. *Econometrica* 68: 1097-1128.

Xia, Y. 2001. Learning about predictability: The effects of parameter uncertainty on dynamic asset allocation. *Journal of Finance* 56: 205-246.

Zakoian, J. 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control* 18: 931-955.

APPENDIX A: Parametric Distributions

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

1. Normal
$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right),\ \gamma_2 = 0.$$

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

3. GED
$$f(z;\nu) = \frac{\nu \exp\left(-0.5\left|z/\lambda\right|^{\nu}\right)}{\lambda 2^{(1+1/\nu)}\Gamma(1/\nu)},\ \nu>0,\ \lambda = \sqrt{2^{(-2/\nu)}\Gamma(1/\nu)/\Gamma(3/\nu)},\ \gamma_2 = \frac{\Gamma(1/v)\Gamma(5/v)}{\left[\Gamma(3/v)\right]^2}.$$

4. Laplace
$$f(z) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}\left|z\right|\right),\ \gamma_2 = 3.$$

5. Double Weibull
$$f(z;a) = \frac{a}{2\sigma}\left|\frac{z}{\sigma}\right|^{a-1} \exp\left(-\left|\frac{z}{\sigma}\right|^{a}\right),\ a>0,\ \sigma = \sqrt{\frac{1}{\Gamma\left(\frac{a+2}{a}\right)}},\ \gamma_2 = \frac{\Gamma\left(\frac{4+a}{a}\right)}{\left[\Gamma\left(\frac{2+a}{a}\right)\right]^2} - 3.$$

6. Skewed t
$$\begin{aligned}
f(z;\xi,\nu) &= \begin{cases} \frac{2s}{\xi+\xi^{-1}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi(\nu-2)}} \left(1+\frac{\xi^2(sz+m)^2}{\nu-2}\right)^{-\frac{\nu+1}{2}} & \text{if } z \le -\frac{m}{s} \\ \frac{2s}{\xi+\xi^{-1}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi(\nu-2)}} \left(1+\frac{\xi^{-2}(sz+m)^2}{\nu-2}\right)^{-\frac{\nu+1}{2}} & \text{if } z > -\frac{m}{s} \end{cases}, \\
\gamma_1 &= \frac{(\nu-2)^{3/2}\left(\xi^2-1\right)\left(\xi^4+1\right)\Gamma\left[(\nu-3)/2\right]}{\sqrt{\pi}\xi^3 s^3 \Gamma(v/2)} - \frac{m\left(m^2+3s\right)}{s^3}, \\
\gamma_2 &= \frac{3\left(\xi^5+\xi^{-5}\right)(\nu-2)}{\left(\xi+\xi^{-1}\right)(\nu-4)s^4} + \frac{3m^2\left(m^2+2s\right)}{s^4} \\
&\quad - \frac{4m(\nu-2)^{3/2}\left(\xi^2-1\right)\left(\xi^4+1\right)\Gamma\left[(\nu-3)/2\right]}{\sqrt{\pi}\xi^3 s^4 \Gamma(v/2)},
\end{aligned}$$

   where $\xi>0$, $\nu>2$, $I = \mathbf{1}(z \ge -m/s)$, $m = \Gamma\left[(\nu-1)/2\right]\sqrt{(\nu-2)/\pi}\,(\xi-1/\xi)/\Gamma(\nu/2)$, $s = \sqrt{(\xi^2+1/\xi^2-1)-m^2}$, γ_1 exists if $\nu>3$ and γ_2 exists if $\nu>4$.

7. Inverse Hyperbolic Sine
$$\begin{aligned}
f(z;\lambda,\delta) &= \frac{s}{\sqrt{2\pi\left[(zs+\mu)^2+1\right]\delta^2}} \exp\left(-\frac{\left[\sinh^{-1}(zs+\mu)-\lambda\right]^2}{2\delta^2}\right), \\
\gamma_1 &= \frac{\omega^{1/2}(\omega-1)^2\left[\omega(\omega+2)\sinh(3\lambda)+3\sinh(\lambda)\right]}{4s^3}, \\
\gamma_2 &= \frac{(\omega-1)^2\left[\omega^2\left(\omega^4+2\omega^3+3\omega^2-3\right)\cosh(4\lambda)+4\omega^2(\omega+2)\cosh(2\lambda)+3(2\omega+1)\right]}{8s^4} \\
&\quad -3,
\end{aligned}$$

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

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

10. Mixture of Normals

$$\begin{aligned} f(z; p_1, \mu_1, \sigma_1) &= p_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[\frac{-(z-\mu_1)^2}{2\sigma_1^2}\right] + p_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[\frac{-(z-\mu_2)^2}{2\sigma_2^2}\right], \\ \gamma_1 &= p_1\left(3\mu_1\sigma_1^2 + \mu_1^3\right) + p_2\left(3\mu_2\sigma_2^2 + \mu_2^3\right), \\ \gamma_2 &= p_1\left(6\mu_1^2\sigma_1^2 + \mu_1^4 + 3\sigma_1^4\right) + p_2\left(6\mu_2^2\sigma_2^2 + \mu_2^4 + 3\sigma_2^4\right) - 3, \end{aligned}$$

where p_2, μ_2, and σ_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 $\sum_{i=1}^2 p_i\left(\mu_i^2 + \sigma_i^2\right) = 1$.

11. Double Gamma

$$\begin{aligned} f(z; \alpha_1, \alpha_2, p) &= \begin{cases} \frac{(1-p)\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)} |z|^{(\alpha_1-1)} \exp(-\lambda_1 |z|) & \text{if } z \leq 0 \\ \frac{p\lambda_2^{\alpha_2}}{\Gamma(\alpha_2)} z^{(\alpha_2-1)} \exp(-\lambda_2 z) & \text{if } z > 0 \end{cases}, \\ \gamma_1 &= \frac{p\Gamma(3+\alpha_2)}{\lambda_2^3\Gamma(\alpha_2)} - \frac{(1-p)\Gamma(3+\alpha_1)}{\lambda_1^3\Gamma(\alpha_1)}, \\ \gamma_2 &= \frac{p\Gamma(4+\alpha_2)}{\lambda_2^4\Gamma(\alpha_2)} - \frac{(1-p)\Gamma(4+\alpha_1)}{\lambda_1^4\Gamma(\alpha_1)} - 3, \end{aligned}$$

where $1 > p > 0$, $\alpha_1 > 0$, $\alpha_2 > 0$,

$$\lambda_1 = \frac{p\alpha_1\lambda_2}{(1-p)\alpha_2},\ \lambda_2 = \sqrt{\alpha_2(1-p)\left[\frac{(\alpha_1+1)\alpha_2(1-p)}{p\alpha_1} + (\alpha_2+1)\right]}.$$

12. Gram-Charlier / Edgeworth-Sargan Density

$$f_4(z; \gamma_1, \gamma_2) = \left[1 + \frac{\gamma_1}{6}H_3(z) + \frac{\gamma_2}{24}H_4(z)\right]\phi(z),$$

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

APPENDIX B: Volatility Specification

In the following, in some cases we use $\alpha(L) = \sum_{i=1}^{p} \alpha_i L^i$, $\beta(L) = \sum_{j=1}^{q} \beta_j L^j$, where L is the backshift operator. Similarly, $\phi(L)$ is a polynomial in L.

1. GARCH(p, q)

$$\begin{aligned} \sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \\ \omega &> 0,\ \alpha_i \geq 0,\ \beta_j \geq 0,\ \sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1. \end{aligned}$$

2. GJR(p, q)

$$\begin{aligned} \sigma_t^2 &= \omega + \sum_{i=1}^{q} \left(\alpha_i \varepsilon_{t-i}^2 + \varphi_i S_{t-i} \varepsilon_{t-i}^2 \right) + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \\ S_{t-i} &= \begin{cases} 1 & \text{if } \varepsilon_{t-i} < 0 \\ 0 & \text{if } \varepsilon_{t-i} \geq 0 \end{cases}, \\ \omega &> 0,\ \beta_j > 0,\ \alpha_i + \varphi_i > 0. \end{aligned}$$

3. APARCH(p, q)

$$\begin{aligned} \sigma_t^{\delta} &= \omega + \sum_{i=1}^{q} \alpha_i \left(|\varepsilon_{t-i}| - \varphi_i \varepsilon_{t-i} \right)^{\delta} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta}, \\ \omega &> 0,\ \beta_i > 0,\ \alpha_j > 0,\ 1 > \varphi_i > -1,\ \delta > 0. \end{aligned}$$

4. EGARCH(p, q)

$$\ln \sigma_t^2 = \omega + \left[1 - \beta(L) \right]^{-1} \left[1 + \alpha(L) \right] \left[\varphi_1 z_{t-1} + \varphi_2 \left(|z_{t-1}| - E\,|z_{t-1}| \right) \right].$$

5. STGARCH(p, q)

$$\begin{aligned} \sigma_t^2 &= \omega + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \left[\alpha_{1j} \varepsilon_{t-j}^2 + \alpha_{2j} \varepsilon_{t-j}^2 F(\varepsilon_{t-d}) \right], \\ F(\varepsilon_{t-d}) &= \frac{1}{1 + \exp(\varphi \varepsilon_{t-d})} - \frac{1}{2}, \\ \omega &> 0,\ \beta_i > 0,\ \alpha_{1j} > \frac{1}{2} \alpha_{2j} > 0,\ \varphi > 0. \end{aligned}$$

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

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

$$\begin{aligned} \sigma_t^2 &= \omega \left[1 - \beta(L) \right]^{-1} + \left(1 - \left[1 - \beta(L) \right]^{-1} \phi(L) \left\{ 1 + \kappa \left[(1 - L)^d - 1 \right] \right\} \right) \varepsilon_t^2 \\ &= \omega \left[1 - \beta(L) \right]^{-1} + \left[1 - \beta(L) \right]^{-1} \lambda(L)\, \varepsilon_t^2. \end{aligned}$$

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

$$\begin{aligned}\lambda(L) &= 1-\beta(L)-(1-\phi^*(L))\left\{1+\kappa\left[(1-L)^d-1\right]\right\} \\ &= 1-\beta(L)-1-\kappa\left[(1-L)^d-1\right]+\phi^*(L)+\phi^*(L)\kappa\left[(1-L)^d-1\right] \\ &= -\beta(L)+\phi^*(L)-\kappa\left[(1-L)^d-1\right]+\phi^*(L)\kappa\left[(1-L)^d-1\right],\end{aligned}$$

where

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

Hence we have

$$\begin{aligned}\lambda(L) &= -\beta(L)+\phi^*(L)-\kappa\sum_{k=1}^{\infty}\delta_k L^k+\kappa\phi^*(L)\sum_{k=1}^{\infty}\delta_k L^k \\ &= -\beta(L)+\phi^*(L)-\kappa\sum_{k=1}^{\infty}\delta_k L^k+\kappa\sum_{k=1}^{\infty}\delta_k^* L^k,\end{aligned}$$

where

$$\delta_k^* = \begin{cases} 0 & k=1 \\ \sum_{j=1}^{k-1}\phi_j^*\delta_{k-j} & k=2,\cdots,q+1 \\ \sum_{j=1}^{q}\phi_j^*\delta_{k-j} & k>q+1 \end{cases}.$$

Therefore, we can define a HYGARCH(p,d,q) model as

$$\begin{aligned}\sigma_t^2 &= \omega+\beta(L)\sigma_t^2+\left[\phi^*(L)-\beta(L)+\sum_{k=1}^{\infty}\pi_k L^k\right]\varepsilon_t^2, \\ \pi_k &= \kappa(\delta_k^*-\delta_k),\ \beta(L)=\sum_{i=1}^{p}\beta_i L^i,\ \phi^*(L)=\sum_{j=1}^{q}\phi_j^* L^j.\end{aligned}$$

The parameters are $\boldsymbol{\theta}_v = \left(\omega, \boldsymbol{\beta}, \boldsymbol{\phi}^{*\prime}, \kappa, d\right)'$. For a HYGARCH$(1,d,1)$ model, the following constraints are sufficient for positiveness of σ_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)

$$\sigma_t^{\delta} = \omega+\left(1-[1-\beta(L)]^{-1}\phi(L)\left\{1+\kappa\left[(1-L)^d-1\right]\right\}\right)(|\varepsilon_t|-\varphi\varepsilon_t)^{\delta},$$

which can be rewritten as

$$\sigma_t^{\delta} = \omega^*+\beta(L)\sigma_t^{\delta}+\left[\phi^*(L)-\beta(L)+\sum_{k=1}^{\infty}\pi_k L^k\right](|\varepsilon_t|-\varphi\varepsilon_t)^{\delta},$$

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

8. CGARCH(1,1)

$$\begin{aligned}\sigma_t^2 &= \Omega_t+\beta\left(\sigma_{t-1}^2-\Omega_{t-i}\right)+\alpha\left(\varepsilon_{t-1}^2-\Omega_{t-1}\right) \\ \Omega_t &= \omega+\rho\Omega_{t-1}+\phi\left(\varepsilon_{t-1}^2-\sigma_{t-1}^2\right), \\ 1 &> \rho>\beta+\alpha>0,\ \beta>\phi>0,\ \alpha>0,\ \omega>0,\end{aligned}$$

where Ω_t is the long-run volatility.

Table 1: Evaluating Density Forecast Models

		GARCH	GJR	APARCH	EGARCH	STGARCH	HYGARCH	HYAPARCH	CGARCH
S&P 500	Normal	0.0277	0.0246	0.0193	0.0211	0.0308	0.0291	0.0192	0.0228
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Student *t*	0.0104	0.0108	0.0090	0.0104	0.0133	0.0103	0.0087	0.0089
		0.0001	0.0000	0.0001	0.0000	0.0000	0.0001	0.0001	0.0008
	GED	0.0147	0.0144	0.0125	0.0141	0.0182	0.0144	0.0123	0.0133
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Laplace	0.0363	0.0444	0.0409	0.0429	0.0425	0.0331	0.0406	0.0359
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Double Weibull	0.0166	0.0160	0.0163	0.0147	0.0190	0.0142	0.0129	0.0134
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Skewed *t*	0.0581	0.0373	0.0392	0.0358	0.0405	0.0584	0.0394	0.0665
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	IHS	0.0095	0.0122	0.0112	0.0127	0.0130	0.0094	0.0109	0.0084
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	Mixture	0.0068	0.0097	0.0086	0.0098	0.0101	0.0065	0.0083	0.0056
		0.0017	0.0000	0.0001	0.0000	0.0000	0.0025	0.0002	0.0084
	Double Gamma	0.0111	0.0108	0.0089	0.0103	0.0182	0.0084	0.0081	0.0109
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0001	0.0000
	Sargan	0.0136	0.0148	0.0108	0.0172	0.0142	0.0115	0.0094	0.0094
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0004
NASDAQ	Normal	0.0270	0.0318	0.0303	0.0313	0.0327	0.0269	0.0302	0.0213
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Student *t*	0.0192	0.0243	0.0226	0.0217	0.0213	0.0187	0.0225	0.0167
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	GED	0.0243	0.0273	0.0259	0.0262	0.0254	0.0226	0.0258	0.0205
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Laplace	0.0599	0.0665	0.0584	0.0593	0.0619	0.0536	0.0586	0.0564
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Double Weibull	0.0310	0.0346	0.0265	0.0244	0.0267	0.0226	0.0240	0.0191
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Skewed *t*	0.0268	0.0206	0.0198	0.0196	0.0220	0.0276	0.0200	0.0304
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	IHS	0.0222	0.0218	0.0198	0.0205	0.0216	0.0189	0.0203	0.0161
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Mixture	0.0159	0.0112	0.0110	0.0104	0.0174	0.0167	0.0115	0.0153
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Double Gamma	0.0231	0.0240	0.0232	0.0224	0.0253	0.0211	0.0213	0.0213
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Sargan	0.0282	0.0372	0.0328	0.0042	0.0098	0.0103	0.0099	0.0294
		0.0000	0.0000	0.0000	0.0149	0.0000	0.0001	0.0001	0.0000

Note: The first row for each combination of volatility and distribution gives the estimated KLIC $I_{R,n}$ (p:ϕ) 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 chose by the AIC criteria for L = 0,…,5 and K = 0,…,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).

Table 2: Comparing Density Forecast Models

			GARCH	GJR	APARCH	EGARCH	STGARCH	HYGARCH	HYAPARCH	CGARCH
S&P 500	p_1=0.007	Normal	1.6437	1.6225	1.6135	1.6134	1.6459	1.6421	1.6140	1.6397
	p_2=0.007	Student *t*	1.6288	1.6121	1.6067	1.6070	1.6309	1.6276	1.6063	1.6255
		GED	1.6330	1.6166	1.6107	1.6107	1.6346	1.6317	1.6103	1.6304
		Laplace	1.6564	1.6451	1.6405	1.6406	1.6602	1.6549	1.6406	1.6545
		Double Weibull	1.6368	1.6209	1.6181	1.6116	1.6346	1.6299	1.6111	1.6292
		Skewed *t*	1.6280	1.6109	1.6061	1.6065	1.6301	1.6267	1.6057	1.6249
		IHS	1.6281	1.6107	1.6060	1.6065	1.6301	1.6268	1.6057	1.6251
		Mixture	1.6303	1.6084	1.6030	1.6042	1.6336	1.6257	1.6044	1.6274
		Double Gamma	1.6424	1.6320	1.6140	1.6224	1.6368	1.6392	1.6157	1.6324
		Sargan	1.6365	1.6178	1.6108	1.6114	1.6388	1.6409	1.6149	1.6309
NASDAQ	p_1=0.145	Normal	2.0298	2.0220	2.0191	2.0174	2.0309	2.0270	2.0184	2.0222
	p_2=0.097	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 $n^{-1}\sum_{t=R+1}^{T} -\ln \psi_t^j(y_t;\hat{\theta}_{R,t-1}^j)$ 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.