ASYMMETRIC PREDICTIVE ABILITIES OF NONLINEAR MODELS FOR STOCK RETURNS: EVIDENCE FROM DENSITY FORECAST COMPARISON

Yong Bao and Tae-Hwy Lee

ABSTRACT

We investigate predictive abilities of nonlinear models for stock returns when density forecasts are evaluated and compared instead of the conditional mean point forecasts. The aim of this paper is to show whether the in-sample evidence of strong nonlinearity in mean may be exploited for out-of-sample prediction and whether a nonlinear model may beat the martingale model in out-of-sample prediction. We use the Kullback–Leibler Information Criterion (KLIC) divergence measure to characterize the extent of misspecification of a forecast model. The reality check test of White (2000) using the KLIC as a loss function is conducted to compare the out-of-sample performance of competing conditional mean models. In this framework, the KLIC measures not only model specification error but also parameter estimation error, and thus we treat both types of errors as loss. The conditional mean models we use for the daily closing S&P 500 index returns include the martingale difference,

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 41–62

ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20021-X

ARMA, STAR, SETAR, artificial neural network, and polynomial models. Our empirical findings suggest the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric in the sense that the right tails of the return series are predictable via many of the nonlinear models, while we find no such evidence for the left tails or the entire distribution.

1. INTRODUCTION

While there has been some evidence that financial returns may be predictable (see, e.g., Lo & MacKinlay, 1988; Wright, 2000), it is generally believed that financial returns are very close to a martingale difference sequence (MDS). The evidence against MDS is usually stronger from in-sample specification tests than from out-of-sample predictability tests using standard evaluation criteria such as the mean squared forecast error (MSFE) and mean absolute forecast error (MAFE).

In this paper, we investigate if this remains true when we evaluate forecasting models in terms of density forecasts instead of the conditional mean point forecasts using MSFE and MAFE. We examine if the evidence and its significance of the nonlinear predictability of financial returns depend on whether we use point forecast evaluation criteria (MSFE and MAFE) or we use the probability density forecasts. As Clements and Smith (2000, 2001) show, traditional measures such as MSFE may mask the superiority of nonlinear models, whose predictive abilities may be more evident through density forecast evaluation.

Motivated by the encouraging results of Clements and Smith (2000, 2001), we compare the density forecasts of various linear and nonlinear models for the conditional mean of the S&P 500 returns by using the method of Bao, Lee, and Saltoglu (2004, BLS henceforth), where the Kullback and Leibler's (1951) Information Criterion (KLIC) divergence measure is used for characterizing the extent of misspecification of a density forecast model. In BLS's framework, the KLIC captures not only model specification error but also parameter estimation error. To compare the performance of density forecast models in the tails of stock return distributions, we also follow BLS by using the censored likelihood functions to compute the tail minimum KLIC. The reality check test of White (2000) is then constructed using the KLIC as a loss function. We find that, for the entire distribution and the left tails, the S&P 500 daily closing returns are not predictable via various linear

or nonlinear models and the MDS model performs best for out-of-sample forecasting. However, from the right tail density forecast comparison of the S&P 500 data, we find, surprisingly, that the MDS model is dominated by many nonlinear models. This suggests that the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric.

This paper proceeds as follows. In Section 2, we examine the nature of the in-sample nonlinearity using the generalized spectral test of Hong (1999). Section 3 presents various linear and nonlinear models we use for the out-of-sample analysis. In Section 4, we compare these models for the S&P 500 return series employing the density forecast approach of BLS. Section 5 concludes. Throughout, we define $y_t = 100(\ln P_t - \ln P_{t-1})$, where P_t is the S&P 500 index at time t.

2. IN-SAMPLE TEST FOR MARTINGALE DIFFERENCE

We will first explore serial dependence (i.e., any departure from IID) in the S&P 500 returns using Hong's (1999) generalized spectrum. In particular, we are interested in finding significant and predictable nonlinearity in the conditional mean even when the returns are linearly unpredictable.

The basic idea is to transform a strictly stationary series y_t to $e^{\mathbf{i}uy_t}$ and consider the covariance function between the transformed variables $e^{\mathbf{i}uy_t}$ and $e^{\mathbf{i}vy_{t-|j|}}$

$$\sigma_j(u,v) \equiv \mathrm{cov}(e^{\mathbf{i}uy_t}, e^{\mathbf{i}vy_{t-|j|}}) \tag{1}$$

where $\mathbf{i} \equiv \sqrt{-1}$, $u, v \in (-\infty, \infty)$, and $j = 0, \pm 1, \ldots$. Suppose that $\{y_t\}_{t=1}^T$ has a marginal characteristic function $\varphi(u) \equiv \mathbb{E}(e^{\mathbf{i}uy_t})$ and a pairwise joint characteristic function $\varphi_j(u,v) \equiv \mathbb{E}(e^{\mathbf{i}(uy_t+vy_{t-|j|})})$. Straightforward algebra yields $\sigma_j(u,v) = \varphi_j(u,v) - \varphi(u)\varphi(v)$. Because $\varphi_j(u,v) = \varphi(u)\varphi(v)$ for all u, v if and only if y_t and $y_{t-|j|}$ are independent, $\sigma_j(u, v)$ can capture any type of pairwise serial dependence over various lags.

When $\sup_{u,v\in(-\infty,\infty)} \sum_{j=-\infty}^{\infty} |\sigma_j(u,v)| < \infty$, the Fourier transform of $\sigma_j(u, v)$ exists

$$f(\omega,u,v) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j(u,v) e^{-\mathbf{i}j\omega}, \quad \omega \in [-\pi, \pi] \tag{2}$$

Like $\sigma_j(u, v)$, $f(\omega, u, v)$ can capture all pairwise serial dependencies in $\{y_t\}$ over various lags. Hong (1999) calls $f(\omega, u, v)$ a "generalized spectral

density'' of $\{y_t\}$, and shows that $f(\omega, u, v)$ can be consistently estimated by

$$\hat{f}_n(\omega, u, v) \equiv \frac{1}{2\pi} \sum_{j=1-n}^{n-1} (1 - |j|/n)^{1/2} k(j/p) \hat{\sigma}_j(u, v) e^{-\mathbf{i}j\omega} \tag{3}$$

where $\hat{\sigma}_j(u, v) \equiv \hat{\varphi}_j(u, v) - \hat{\varphi}_j(u, 0)\hat{\varphi}_j(0, v)$ is the empirical generalized covariance, $\hat{\varphi}_j(u, v) \equiv (n - |j|)^{-1} \sum_{t=|j|+1}^{n} e^{\mathbf{i}(uy_{t+v}y_{t-|j|})}$ is the empirical pairwise characteristic function, $p \equiv p_n$ a bandwidth or lag order, and $k(\cdot)$ a kernel function or ''lag window''. Commonly used kernels include the Bartlett, Daniell, Parzen, and Quadratic–Spectral kernels.

When $\{y_t\}$ is IID, $f(\omega, u, v)$ becomes a ''flat'' generalized spectrum:

$$f_0(\omega, u, v) \equiv \frac{1}{2\pi} \sigma_0(u, v), \quad \omega \in [-\pi, \pi]$$

Any deviation of $f(\omega, u, v)$ from the flat spectrum $f_0(\omega, u, v)$ is the evidence of serial dependence. Thus, to detect serial dependence, we can compare $\hat{f}_n(\omega, u, v)$ with the estimator

$$\hat{f}_0(\omega, u, v) \equiv \frac{1}{2\pi} \hat{\sigma}_0(u, v), \quad \omega \in [-\pi, \pi]$$

To explore the nature of serial dependence, one can compare the derivative estimators

$$\hat{f}_n^{(0,m,l)}(\omega, u, v) \equiv \tfrac{1}{2\pi} \sum_{j=1-n}^{n-1} (1 - |j|/n)^{1/2} k(j/p) \hat{\sigma}_j^{(m,l)}(u, v) e^{-\mathbf{i}j\omega}$$

$$\hat{f}_0^{(0,m,l)}(\omega, u, v) \equiv \tfrac{1}{2\pi} \hat{\sigma}_0^{(m,l)}(u, v)$$

where $\hat{\sigma}_j^{(m,l)}(u, v) \equiv \partial^{m+l} \hat{\sigma}_j(u, v) / \partial^m u \partial^l v$ for $m, l \geq 0$. Just as the characteristic function can be differentiated to generate various moments, generalized spectral derivatives can capture various specific aspects of serial dependence, thus providing information on possible types of serial dependence.

Hong (1999) proposes a class of tests based on the quadratic norm

$$\begin{aligned} Q\left(\hat{f}_n^{(0,m,l)}, \hat{f}_0^{(0,m,l)}\right) &\equiv \int \int_{-\pi}^{\pi} \left| \hat{f}_n^{(0,m,l)}(\omega, u, v) - \hat{f}_0^{(0,m,l)}(\omega, u, v) \right|^2 \mathrm{d}\omega \, \mathrm{d}W_1(u) \, \mathrm{d}W_2(v) \\ &= \frac{2}{\pi} \int \sum_{j=1}^{n-1} k^2(j/p)(1 - j/n) \left| \hat{\sigma}_j^{(m,l)}(u, v) \right|^2 \mathrm{d}W_1(u) \, \mathrm{d}W_2(v) \end{aligned}$$

where the second equality follows by Parseval's identity, and the unspecified integrals are taken over the support of $W_1(\cdot)$ and $W_2(\cdot)$, which are positive

and nondecreasing weighting functions that set weight about zero equally. The generalized spectral test statistic $M(m, l)$ is a standardized version of the quadratic norm. Given (m, l), $M(m, l)$ is asymptotically one-sided $N(0,1)$ under the null hypothesis of serial independence, and thus the upper-tailed asymptotic critical values are 1.65 and 2.33 at the 5% and 1% levels, respectively.

We may first choose $(m, l) = (0, 0)$ to check if there exists any type of serial dependence. Once generic serial dependence is discovered using $M(0,0)$, we may use various combinations of (m, l) to check specific types of serial dependence. For example, we can set $(m, l) = (1, 0)$ to check whether there exists serial dependence in mean. This checks whether $\mathbb{E}(y_t|y_{t-j}) = \mathbb{E}(y_t)$ for all $j > 0$, and so it is a suitable test for the MDS hypothesis. It can detect a wide range of deviations from MDS. To explore whether there exists linear dependence in mean, we can set $(m, l) = (1, 1)$. If $M(1,0)$ is significant but $M(1,1)$ is not, we can speculate that there may exist only nonlinear dependence in mean. We can go further to choose $(m, l) = (1, l)$ for $l = 2, 3, 4$, testing if $\mathrm{cov}(y_t, y_{t-j}^l) = 0$ for all $j > 0$. These essentially check whether there exist ARCH-in-mean, skewness-in-mean, and kurtosis-in-mean effects, which may arise from the existence of time-varying risk premium, asymmetry, and improper account of the concern over large losses, respectively. Table 1 lists a variety of spectral derivative tests and the types of dependence they can detect, together with the estimated $M(m, l)$ statistics.[1]

We now use the generalized spectral test to explore serial dependence of the daily S&P 500 closing return series, retrieved from *finance.yahoo.com*. They are from January 3, 1990 to June 30, 2003 ($T = 3403$).

The statistic $M(m, l)$ involves the choice of a bandwidth p in its computation, see Hong (1999, p. 1204). Hong proposes a data-driven method to choose p. This method still involves the choice of a preliminary bandwidth $\bar{p}$. Simulations in Hong (1999) show that the choice of $\bar{p}$ is less important than that of p. We consider $\bar{p}$ in the range 6–15 to examine the robustness of $M(m, l)$ with respect to the choice of $\bar{p}$. We use the Daniell kernel, which maximizes the asymptotic power of $M(m, l)$ over a class of kernels. We have also used the Bartlett, Parzen, and Quadratic–Spectral kernels, whose results are similar to those based on the Daniell kernel and are not reported in this paper.

Table 1 reports the values of $M(m, l)$ for $\bar{p} = 6, 9, 12, 15$. The results for various values of $\bar{p}$ are quite similar. $M(m, l)$ has an asymptotic one-sided $N(0,1)$ distribution, so the asymptotic critical value at the 5% level is 1.65. The $M(0,0)$ statistic suggests that the random walk hypothesis is strongly rejected. In contrast, the correlation test $M(1,1)$ is insignificant, implying

Table 1. Generalized Spectral Tests.

Test	Statistic $M(m, l)$	Test Function $\sigma_j^{(m,l)}(u,v)$	Preliminary Bandwidth $\bar{p}=6$	$\bar{p}=9$	$\bar{p}=12$	$\bar{p}=15$
IID	$M(0, 0)$	$\sigma_j = (u, v)$	51.02	58.23	63.75	67.85
MDS	$M(1, 0)$	$\text{cov}(y_t, e^{\mathbf{i}vy_{t-j}})$	17.40	18.28	18.67	19.04
Correlation	$M(1, 1)$	$\text{cov}(y_t, y_{t-j})$	−0.10	0.44	0.63	0.68
ARCH-in-mean	$M(1, 2)$	$\text{cov}(y_t, y_{t-j}^2)$	56.24	55.83	55.36	54.84
Skewness-in-mean	$M(1, 3)$	$\text{cov}(y_t, y_{t-j}^3)$	−0.11	−0.38	−0.50	−0.51
Kurtosis-in-mean	$M(1, 4)$	$\text{cov}(y_t, y_{t-j}^4)$	29.85	29.99	29.57	29.18
Nonlinear ARCH	$M(2, 0)$	$\text{cov}(y_t^2, e^{\mathbf{i}vy_{t-j}})$	62.15	70.75	76.71	81.30
Leverage	$M(2, 1)$	$\text{cov}(y_t^2, y_{t-j})$	9.25	8.57	8.52	8.57
Linear ARCH	$M(2, 2)$	$\text{cov}(y_t^2, y_{t-j}^2)$	172.87	182.41	188.53	193.64
Conditional skewness	$M(3, 0)$	$\text{cov}(y_t^3, e^{\mathbf{i}vy_{t-j}})$	7.63	6.98	6.64	6.36
Conditional skewness	$M(3, 3)$	$\text{cov}(y_t^3, y_{t-j}^3)$	27.82	26.66	26.69	26.83
Conditional kurtosis	$M(4, 0)$	$\text{cov}(y_t^4, e^{\mathbf{i}vy_{t-j}})$	17.16	18.17	19.12	20.10
Conditional kurtosis	$M(4, 4)$	$\text{cov}(y_t^4, y_{t-j}^4)$	35.56	35.22	35.22	35.25

Note: All generalized spectral test statistics $M(m, l)$ are asymptotically one-sided $N(0, 1)$ and thus upper-tailed asymptotic critical values are 1.65 and 2.33 at the 5% and 1% levels, respectively. $M(0, 0)$ is to check if there exists any type of serial dependence. $M(1, 0)$ is to check whether there exists serial dependence in mean. To explore whether there exists linear dependence in mean, we can set $(m, l) = (1, 1)$. If $M(1, 0)$ is significant but $M(1, 1)$ is not, we can speculate that there may exist only nonlinear dependence in mean. We choose $(m, l) = (1, l)$ with $l = 2, 3, 4$, to test if $\mathbb{E}(y_t|y_{t-j}^l) = 0$ for all $j>0$. The PN model is to exploit the nonlinear predictive evidence of polynomials found from $M(1, l)$ with $l = 2, 3, 4$.

that $\{y_t\}$ is an uncorrelated white noise. This, however, does not necessarily imply that $\{y_t\}$ is an MDS. Indeed, the martingale test $M(1,0)$ strongly rejects the martingale hypothesis as its statistic is above 17. This implies that the S&P 500 returns, though serially uncorrelated, has a nonzero mean conditional on its past history. Thus, suitable nonlinear time series models may be able to predict the future returns. The polynomial (PN) model (to be discussed in the next section) is to exploit the nonlinear predictive evidence of the lth power of returns, as indicated by the $M(1,l)$ statistics.

The test $M(2,0)$ shows possibly nonlinear time-varying volatility, and the linear ARCH test $M(2,2)$ indicates very strong linear ARCH effects. We also observe that the leverage effect ($M(2,1)$) is significant and there exist significant conditional skewness as evidenced from $M(3,0)$ and $M(3,3)$, and large conditional kurtosis as evidenced from $M(4,0)$ and $M(4,4)$.

It is important to explore the implications of these in-sample findings of nonlinearity in the conditional mean. The fact that the S&P 500 return series

is not an MDS implies it may be predictable in the conditional mean. In the next section, we will use various linear and nonlinear time series models to examine this issue.

3. CONDITIONAL MEAN MODELS

Let $\mathscr{F}_{t-1}$ be the information set containing information about the process $\{y_t\}$ up to and including time $t-1$. Since our interest is to investigate the predictability of stock returns in the conditional mean $\mu_t = \mathbb{E}(y_t|\mathscr{F}_{t-1})$, we assume that y_t is conditionally normally distributed and the conditional variance $\sigma_t^2 = \mathbb{E}(\varepsilon_t^2|\mathscr{F}_{t-1})$ follows a GARCH(1,1) process $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$, where $\varepsilon_t = y_t - \mu_t$. We consider the following nine models for μ_t in three classes:

(i) the MDS model

$$\text{MDS} \qquad y_t = \varepsilon_t$$

(ii) four linear autoregressive moving average (ARMA) models

Constant	$y_t = a_0 + \varepsilon_t$
MA(1)	$y_t = a_0 + b_1\varepsilon_{t-1} + \varepsilon_t$
ARMA(1, 1)	$y_t = a_0 + a_1 y_{t-1} + b_1\varepsilon_{t-1} + \varepsilon_t$
AR(2)	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$

(iii) four nonlinear models, namely, the polynomial (PN), neural network (NN), self-exciting transition autoregressive (SETAR), and smooth transition autoregressive (STAR) models,

PN(2, 4)	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \sum_{j=1}^{2}\sum_{i=2}^{4} a_{ij} y_{t-j}^i + \varepsilon_t$
NN(2, 5)	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \sum_{i=1}^{5} \delta_i G(\gamma_{0i} + \sum_{j=1}^{2} \gamma_{ji} y_{t-j}) + \varepsilon_t$
SETAR	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + (b_0 + b_1 y_{t-1} + b_2 y_{t-2})\mathbf{1}(y_{t-1} > c) + \varepsilon_t$
STAR	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + (b_0 + b_1 y_{t-1} + b_2 y_{t-2}) G(\gamma(y_{t-1} - c)) + \varepsilon_t$

where $G(z) = 1/(1 + e^{-z})$ is a logistic function and $\mathbf{1}(\cdot)$ denotes an indicator function that takes 1 if its argument is true and 0 otherwise. Note that the four nonlinear models nest the AR(2) model.

All the above models have been used in the literature, with apparently mixed results on the predictability of stock returns. Hong and Lee (2003) use the AR, PN, and NN models. McMillan (2001) and Kanas (2003) find evidence supporting the NN and STAR models, while Bradley and Jansen (2004) find no evidence for the STAR model. We note that these authors use the MSFE criterion for out-of-sample forecasting evaluation. Racine (2001) finds no predictability evidence using the NN model. Anderson, Benzoni, and Lund (2002) use the MA model for estimation.

The results from the generalized spectral test reported in Table 1 suggest that $\mathbb{E}(y_t|\mathcal{F}_{t-1})$ is time-varying in a nonlinear manner, because $M(1,1)$ is insignificant but $M(1,0)$ is significant. Also, we note that the $M(1,l)$ statistics are significant with $l = 2, 4$ but not with $l = 1, 3$, indicating volatility and tail observations may have some predictive power for the returns but not the skewness of the return distribution. The PN model is to exploit this nonlinear predictive evidence of the lth order power of the lagged returns.

4. OUT-OF-SAMPLE TEST FOR MARTINGALE DIFFERENCE

We now examine if the in-sample evidence of nonlinear predictability of the S&P 500 returns from the generalized spectral test in the previous section may be carried over to the out-of-sample forecasting. While the in-sample generalized spectral test does not involve parameter estimation of a particular nonlinear model, the out-of-sample test requires the estimation of model parameters since it is based on some particular choice of nonlinear models. Despite the strong nonlinearity found in the conditional mean from the in-sample tests, model uncertainty and parameter uncertainty usually make the out-of-sample results much weaker than the in-sample nonlinear evidence, see Meese and Rogoff (1983).

Given model uncertainty, econometricians tend to search for a proper model over a large set of candidate models. This can easily cause the problem of data snooping, see Lo and MacKinlay (1990). In order to take care of the data snooping bias, we follow the method of White (2000) in our comparison of multiple competing models. Moreover, we use the KLIC

measure as our loss function in the comparison. As emphasized by BLS (2004), the KLIC measure captures both the loss due to model specification error and the loss due to parameter estimation error. It is important to note that in comparing forecasting models we treat parameter estimation error as a loss. Now, we briefly discuss the BLS test.[2]

4.1. The BLS Test

Suppose that $\{y_t\}$ has a true, unknown conditional density function $f(y_t) \equiv f(y_t|\mathscr{F}_{t-1})$. Let $\psi(y_t;\boldsymbol{\theta}) = \psi(y_t|\mathscr{F}_{t-1};\boldsymbol{\theta})$ be a one-step-ahead conditional density forecast model with parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a finite-dimensional vector of parameters in a compact parameter space $\boldsymbol{\Theta}$. If $\psi(\cdot;\boldsymbol{\theta}_0) = f(\cdot)$ for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then the one-step-ahead density forecast is correctly specified and hence optimal, as it dominates all other density forecasts for any loss function (e.g., Diebold, Gunther, & Tay, 1998; Granger & Pesaran, 2000a, b). As our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we consider two subsamples $\{y_t\}_{t=1}^{R}$ and $\{y_t\}_{t=R+1}^{T}$: we use the first sample to estimate the unknown parameter vector $\boldsymbol{\theta}$ and the second sample to compare the out-of-sample density forecasts.

In practice, it is rarely the case that we can find an optimal model as all the models can be possibly misspecified. Our task is then to investigate which model can approximate the true model 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. The adequacy of a postulated distribution may be appropriately measured by the KLIC divergence measure between $f(\cdot)$ and $\psi(\cdot)$: $I(f:\psi,\boldsymbol{\theta}) = \mathbb{E}\left[\ln f(y_t) - \ln\psi(y_t;\boldsymbol{\theta})\right]$, where the expectation is with respect to the true distribution. Following Vuong (1989), we define the distance between a model and the true density as the minimum of the KLIC

$$I(f:\psi,\boldsymbol{\theta}^*) = \mathbb{E}\left[\ln f(y_t) - \ln\psi(y_t;\boldsymbol{\theta}^*)\right] \tag{4}$$

and $\boldsymbol{\theta}^*$ is the pseudo-true value of $\boldsymbol{\theta}$, the parameter value that gives the minimum $I(f:\psi,\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$ (e.g., White, 1982). The smaller this distance, the closer the model $\psi(\cdot;\boldsymbol{\theta})$ is to the true density. Thus, we can use this measure to compare the relative distance of a battery of competing models to the true model $f_t(\cdot)$. However, $I(f:\psi,\boldsymbol{\theta}^*)$ is generally unknown, since we cannot observe $f(\cdot)$ and thereby we can not evaluate the expectation in (4). Under some regularity conditions, it can nevertheless be shown that

$\mathbb{E}\left[\ln f(y_t) - \ln\psi(y_t;\boldsymbol{\theta}^*)\right]$ can be consistently estimated by

$$\hat{I}(f:\psi) = \frac{1}{n}\sum_{t=R+1}^{T}\left[\ln f(y_t) - \ln\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})\right] \tag{5}$$

where $n = T-R$ and $\hat{\boldsymbol{\theta}}_{t-1}$ is consistently estimated from a rolling sample $\{y_{t-1},\ldots,y_{t-R}\}$ of size R. But we still do not know $f(\cdot)$. For this, we utilize the equivalence relationship between $\ln[f(y_t)/\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})]$ and the log-likelihood ratio of the inverse normal transform of the probability integral transform (PIT) of the actual realizations of the process with respect to the models' density forecast. This equivalence relationship enables us to consistently estimate $I(f:\psi,\boldsymbol{\theta}^*)$ and hence to conduct the out-of-sample comparison of possibly misspecified models in terms of their distance to the true model.

The (out-of-sample) PIT of the realization of the process with respect to the model's density forecast is defined as

$$u_t = \int_{-\infty}^{y_t}\psi(y;\hat{\boldsymbol{\theta}}_{t-1})\mathrm{d}y, \quad t = R+1,\ldots,T \tag{6}$$

It is well known that if $\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})$ coincides with the true density $f(y_t)$ for all t, then the sequence $\{u_t\}_{t=R+1}^{T}$ is IID and uniform on the interval [0, 1] ($U[0,1]$ henceforth). This provides a powerful approach to evaluating the quality of a density forecast model. Our task, however, is not to evaluate a single model, but to compare a battery of competing models. Our purpose of utilizing the PITs is to exploit the following equivalence between $\ln[f(y_t)/\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})]$ and the log likelihood ratio of the transformed PIT and hence to construct the distance measure. The inverse normal transform of the PIT is

$$x_t = \Phi^{-1}(u_t) \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. If the sequence $\{u_t\}_{t=R+1}^{T}$ is IID $U[0,1]$ then $\{x_t\}_{t=R+1}^{T}$ is IID standard normal $N(0,1)$ (IID $N(0,1)$ henceforth). More importantly, Berkowitz (2001, Proposition 2, p. 467) shows that

$$\ln\left[f(y_t)/\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})\right] = \ln\left[p(x_t)/\phi(x_t)\right] \tag{8}$$

where $p(\cdot)$ is the density of x_t and $\phi(\cdot)$ the standard normal density. Therefore, the distance of a density forecast model to the unknown true

model can be equivalently estimated by the departure of $\{x_t\}_{t=R+1}^{T}$ from IID $N(0, 1)$,

$$\tilde{I}(f : \psi) = \frac{1}{n}\sum_{t=R+1}^{T}[\ln p(x_t) - \ln \phi(x_t)] \tag{9}$$

We transform the departure of $\psi(\cdot;\boldsymbol{\theta})$ from $f(\cdot)$ to the departure of $p(\cdot)$ from IID $N(0, 1)$. To specify the departure from IID $N(0, 1)$, we want $p(\cdot)$ to be as flexible as possible to reflect the true distribution of $\{x_t\}_{t=R+1}^{T}$ and at the same time it can be IID $N(0, 1)$ if the density forecast model coincides with the true model. We follow Berkowitz (2001) by specifying $\{x_t\}_{t=R+1}^{T}$ as an AR(L) process

$$x_t = \boldsymbol{\rho}' X_{t-1} + \sigma\eta_t \tag{10}$$

where $X_{t-1} = (1, x_{t-1}, \ldots, x_{t-L})'$, $\boldsymbol{\rho}$ is an $(L+1)\times 1$ vector of parameters, and η_t IID distributed. We specify a flexible distribution for η_t, say, $p(\eta_t; \boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ is a vector of distribution parameters such that when $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ for some $\boldsymbol{\gamma}^*$ in the parameter space, $p(\eta_t; \boldsymbol{\gamma}^*)$ is IID $N(0, 1)$. A test for IID $N(0, 1)$ of $\{x_t\}_{t=R+1}^{T}$ per se can be constructed by testing elements of the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\rho}', \sigma, \boldsymbol{\gamma}')'$, say, $\boldsymbol{\rho} = \mathbf{0}$, $\sigma = 1$, and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. We assume the semi-nonparametric (SNP) density of order K of Gallant and Nychka (1987) for η_t

$$p(\eta_t; \boldsymbol{\gamma}) = \frac{\left(\sum_{k=0}^{K}\gamma_k \eta_t^k\right)\phi(\eta_t)}{\int_{-\infty}^{+\infty}\left(\sum_{k=0}^{K}\gamma_k u^k\right)^2 \phi(u)\mathrm{d}u}, \tag{11}$$

where $\gamma_0 = 1$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_K)'$. Setting $\gamma_k = 0$ $(k = 1, \ldots, K)$ gives $p(\eta_t) = \phi(\eta_t)$. Given (10) and (11), the density of x_t is

$$p(x_t; \boldsymbol{\beta}) = \frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma; \boldsymbol{\gamma}\right]}{\sigma},$$

which degenerates into IID $N(0,1)$ by setting $\boldsymbol{\beta} = \boldsymbol{\beta}^* = (\mathbf{0}', 1, \mathbf{0}')'$. Then $\tilde{I}(\varphi : \psi)$ as defined in (9) can be estimated by

$$\tilde{I}(f : \psi; \boldsymbol{\beta}) = \frac{1}{n}\sum_{t=R+1}^{T}\left[\ln\frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma; \boldsymbol{\gamma}\right]}{\sigma} - \ln\phi(x_t)\right].$$

The likelihood ratio test statistic of the adequacy of the density forecast model $\psi(\cdot;\boldsymbol{\theta})$ in Berkowitz (2001) is simply the above formula with $p(\cdot) = \phi(\cdot)$. As $\boldsymbol{\beta}$ is unknown, we estimate $\tilde{I}(\varphi : \psi)$ by

$$\tilde{I}(f : \psi; \hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{t=R+1}^{T}\left[\ln\frac{p\left[(x_t - \hat{\boldsymbol{\rho}}_n' X_{t-1})/\hat{\sigma}_n; \hat{\boldsymbol{\gamma}}_n\right]}{\hat{\sigma}_n} - \ln\phi(x_t)\right], \tag{12}$$

where $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\rho}}_n', \hat{\sigma}_n, \boldsymbol{\gamma}_n')'$ is the maximum likelihood estimator that maximizes $n^{-1}\sum_{t=R+1}^{T} \ln p(x_t; \boldsymbol{\beta})$.

To check the performance of a density forecast model in certain regions of the distribution, we can easily modify our distance measure tailored for the tail parts only. For the lower (left) tail, we define the censored random variable

$$x_t^\tau = \begin{cases} x_t & \text{if } x_t < \tau \\ \Phi^{-1}(\alpha) \equiv \tau & \text{if } x_t \geq \tau. \end{cases} \tag{13}$$

For example, $\tau = -1.645$ for $\alpha = 0.05$, the left 5% tail. As before, we consider an AR model (10) with η_t distributed as in (11). Then the censored random variable x_t^τ has the distribution function

$$p^\tau(x_t^\tau; \boldsymbol{\beta}) = \left[\frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma\right]}{\sigma}\right]^{\mathbf{1}(x_t < \tau)} \left[1 - P\left(\frac{\tau - \boldsymbol{\rho}' X_{t-1}}{\sigma}; \boldsymbol{\gamma}\right)\right]^{\mathbf{1}(x_t \geq \tau)}, \tag{14}$$

in which $P(\cdot; \boldsymbol{\gamma})$ is the CDF of the SNP density function and is calculated in the way as discussed in BLS. Given $p^\tau(x_t^\tau; \boldsymbol{\beta})$, the (left) tail minimum KLIC divergence measure can be estimated analogously

$$\tilde{I}^\tau(f : \psi; \hat{\boldsymbol{\beta}}_n^\tau) = \frac{1}{n}\sum_{t=R+1}^{T} \left[\ln p^\tau(x_t^\tau; \hat{\boldsymbol{\beta}}_n^\tau) - \ln \phi^\tau(x_t^\tau)\right], \tag{15}$$

where $\phi^\tau(x_t^\tau) = [1 - \Phi(\tau)]^{\mathbf{1}(x_t \geq \tau)}[\phi(x_t)]^{\mathbf{1}(x_t < \tau)}$ and $\hat{\boldsymbol{\beta}}_n^\tau$ maximizes $n^{-1}\sum_{t=R+1}^{T} \ln p^\tau(x_t^\tau; \boldsymbol{\beta})$.

For the upper (right) tail we define the censored random variable

$$x_t^\tau = \begin{cases} \Phi^{-1}(\alpha) \equiv \tau & \text{if } x_t \leq \tau \\ x_t & \text{if } x_t > \tau \end{cases} \tag{16}$$

For example, $\tau = 1.645$ for $\alpha = 0.95$, the right 5% tail. Then the censored random variable x_t^τ has the distribution function

$$p^\tau(x_t^\tau; \boldsymbol{\beta}) = \left[1 - P\left(\frac{\tau - \boldsymbol{\rho}' X_{t-1}}{\sigma}; \boldsymbol{\gamma}\right)\right]^{\mathbf{1}(x_t \leq \tau)} \left[\frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma\right]}{\sigma}\right]^{\mathbf{1}(x_t > \tau)}, \tag{17}$$

and the (right) tail minimum KLIC divergence measure can be estimated by (15) with $p^\tau(x_t^\tau; \boldsymbol{\beta})$ given by (17).

Therefore, given (6) and (7), we are able to estimate the minimum distance measure (4) by (12) (or its tail counterpart by (15)), which is proposed by BLS as a loss function to compare the out-of-sample predictive abilities of a set of $l+1$ competing models, each of which can be possibly misspecified. To

establish the notation with model indexed by k ($k = 0, 1, \ldots, l$), let the density forecast model k be denoted by $\psi_k(y_t;\ \boldsymbol{\theta}_k)$. Model comparison between a single model (model k) and the benchmark model (model 0) can be conveniently formulated as hypothesis testing of some suitable moment conditions. Consider constructing the loss differential

$$d_k = d_k(\psi_0, \psi_k) = \left[\ln f(y_t) - \ln\psi_0(y_t; \boldsymbol{\theta}_0^*)\right] - \left[\ln f(y_t) - \ln\psi_k(y_t; \boldsymbol{\theta}_k^*)\right] \quad (18)$$

where $1 \le k \le l$. Note that $\mathbb{E}(d_k) = I(f : \psi_0; \boldsymbol{\theta}_0^*) - I(f : \psi_k; \boldsymbol{\theta}_k^*)$ is the difference in the minimum KLIC of models 0 and k. When we compare multiple l models against a benchmark jointly, the null hypothesis of interest is that the best model is no better than the benchmark

$$\mathbb{H}_0 : \max_{1\le k\le l} \mathbb{E}(d_k) \le 0 \quad (19)$$

To implement this test, we follow White (2000) to bootstrap the following test statistic

$$\bar{V}_n \equiv \max_{1\le k\le l} \sqrt{n}\left[\bar{d}_{k,n} - \mathbb{E}(d_k)\right] \quad (20)$$

where $\bar{d}_{k,n} = \tilde{I}(f : \psi_0; \hat{\boldsymbol{\beta}}_{0,n}) - \tilde{I}(f : \psi_k; \hat{\boldsymbol{\beta}}_{k,n})$, and $\tilde{I}(f : \psi_0; \hat{\boldsymbol{\beta}}_{0,n})$ and $\tilde{I}(f : \psi_k; \hat{\boldsymbol{\beta}}_{k,n})$ are estimated by (12) for models 0 and k, with the normal-inversed PIT $\{x_t\}_{t=R+1}^{T}$ constructed using $\hat{\boldsymbol{\theta}}_{0,t-1}$ and $\hat{\boldsymbol{\theta}}_{k,t-1}$ estimated by a rolling-sample scheme, respectively. A merit of using the KLIC-based loss function for comparing forecasting models is that $\bar{d}_{k,n}$ incorporates both model specification error and parameter estimation error (note that x_t is constructed using $\hat{\boldsymbol{\theta}}$ rather than $\boldsymbol{\theta}^*$) as argued by BLS (2004).

To obtain the p-value for $\bar{V}_n$ White (2000) suggests using the stationary bootstrap of Politis and Romano (1994). This bootstrap p-value for testing $\mathbb{H}_0$ is called the "reality check p-value" for data snooping. As discussed in Hansen (2001), White's reality check p-value may be considered as an upper bound of the true p-value, since it sets $\mathbb{E}(d_k) = 0$. Hansen (2001) considers a modified reality check test, which removes the "bad" models from the comparison and thereby improves the size and the power of the test. The reality check to compare the performance of density forecast models in the tails can be implemented analogously.

4.2. Results of the BLS Test

We split the sample used in Section 2 into two parts (roughly into two halves): one for in-sample estimation of size $R = 1703$ and another for out-of-sample density forecast of size $n = 1700$. We use a rolling-sample scheme.

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

The results are presented in Tables 2–4, with each table computing the statistics with different ways of selecting L and K. We present the reality check results with the whole density (100% with $\alpha = 1.00$), three left tails (10%, 5%, 1% with $\alpha = 0.10, 0.05, 0.01$), and three right tails (10%, 5%, 1% with $\alpha = 0.90, 0.95, 0.99$). With the AR($L$)–SNP($K$) model as specified in (10) and (11), we need to choose L and K. In Table 2, we fix $L = 3$ and $K = 5$. We minimize the Akaike information criteria (AIC) in Table 3, and the Schwarz information criteria (SIC) in Table 4, for the selection of L and K from the sets of $\{0, 1, 2, 3\}$ for L and $\{0, 1, \ldots, 8\}$ for K.

The out-of-sample average KLIC loss (denoted as "Loss" in tables) as well as the reality check p-value of White (2000) (denoted as "p_1") and the modified reality check p-value of Hansen (2001) (denoted as "p_2") are presented in these tables. In comparing the models using the reality check tests, we regard each model as a benchmark model and it is compared with the remaining eight models. We set the number of bootstraps for the reality check to be 1,000 and the mean block length to be 4 for the stationary bootstrap of Politis and Romano (1994). The estimated out-of-sample KLIC (12) and its censored versions as defined from (15) with different values of τ (each corresponding to different α) are reported in Tables 2–4. Note that in these tables a small value of the out-of-sample average loss and a large reality check p-value indicate that the corresponding model is a good density forecast model, as we fail to reject the null hypothesis that the other remaining eight models is no better than that model.

As our purpose is to test for the MDS property of the S&P 500 returns in terms of out-of-sample forecasts, we examine the performance of the MDS model as the benchmark (Model 0 with $k = 0$) in comparison with the remaining eight models indexed by $k = 1, \ldots, l$ ($l = 8$). The eight competing models are Constant, MA, ARMA, AR, PN, NN, SETAR, and STAR.

Table 2 shows the BLS test results computed using $L = 3$ and $K = 5$. First, comparing the entire return density forecasts with $\alpha = 100\%$, the KLIC loss value for the MDS model is $\tilde{I}(f : \psi_0; \hat{\boldsymbol{\beta}}_{0,n}) = 0.0132$, that is the smallest loss, much smaller than those of the other eight models. The large reality check p-values for the MDS model as the benchmark ($p_1 = 0.982$ and $p_2 = 0.852$) indicate that none of the other eight models are better than the MDS model, confirming the efficient market hypothesis that the S&P 500 returns may not be predictable using linear or nonlinear models.

Table 2. Reality Check Results Based on AR(3)-SNP(5).

Tail	Model	Left Tail					Right Tail				
		Loss	p_1	p_2	L	K	Loss	p_1	p_2	L	K
	MDS	0.0132	0.982	0.852	3	5					
	Constant	0.0233	0.370	0.370	3	5					
	MA	0.0237	0.357	0.357	3	5					
	ARMA	0.0238	0.357	0.357	3	5					
100%	AR	0.0171	0.547	0.486	3	5					
	PN	0.0346	0.071	0.071	3	5					
	NN	0.0239	0.355	0.355	3	5					
	SETAR	0.0243	0.354	0.354	3	5					
	STAR	0.0238	0.354	0.354	3	5					
	MDS	0.0419	1.000	0.730	3	5	0.0409	0.000	0.000	3	5
	Constant	0.0467	0.000	0.000	3	5	0.0359	0.415	0.394	3	5
	MA	0.0472	0.000	0.000	3	5	0.0352	0.794	0.686	3	5
	ARMA	0.0468	0.000	0.000	3	5	0.0359	0.442	0.397	3	5
10%	AR	0.0469	0.000	0.000	3	5	0.0363	0.309	0.298	3	5
	PN	0.0466	0.000	0.000	3	5	0.0348	0.627	0.615	3	5
	NN	0.0469	0.000	0.000	3	5	0.0365	0.269	0.261	3	5
	SETAR	0.0476	0.000	0.000	3	5	0.0360	0.396	0.372	3	5
	STAR	0.0470	0.000	0.000	3	5	0.0366	0.243	0.238	3	5
	MDS	0.0199	1.000	0.671	3	5	0.0229	0.000	0.000	3	5
	Constant	0.0218	0.002	0.002	3	5	0.0205	0.040	0.040	3	5
	MA	0.0217	0.003	0.003	3	5	0.0206	0.039	0.039	3	5
	ARMA	0.0217	0.003	0.003	3	5	0.0206	0.040	0.040	3	5
5%	AR	0.0211	0.023	0.000	3	5	0.0208	0.031	0.031	3	5
	PN	0.0226	0.000	0.000	3	5	0.0170	0.981	0.518	2	5
	NN	0.0211	0.026	0.000	3	5	0.0210	0.022	0.022	3	5
	SETAR	0.0211	0.040	0.004	3	5	0.0213	0.019	0.019	3	5
	STAR	0.0213	0.017	0.000	3	5	0.0208	0.029	0.029	3	5
	MDS	0.0068	1.000	0.992	3	5	0.0073	0.042	0.042	3	5
	Constant	0.0074	0.148	0.148	3	5	0.0068	0.172	0.172	3	5
	MA	0.0074	0.163	0.163	3	5	0.0068	0.167	0.167	3	5
	ARMA	0.0074	0.178	0.178	3	5	0.0068	0.160	0.160	3	5
1%	AR	0.0075	0.132	0.132	3	5	0.0068	0.151	0.151	3	5
	PN	0.0074	0.240	0.240	3	5	0.0058	0.932	0.914	3	5
	NN	0.0075	0.128	0.128	3	5	0.0068	0.150	0.150	3	5
	SETAR	0.0076	0.064	0.064	3	5	0.0068	0.135	0.135	3	5
	STAR	0.0074	0.181	0.181	3	5	0.0068	0.169	0.169	3	5

Note: "Loss" refers to is the out-of-sample averaged loss based on the KLIC measure; "p_1" and "p_2" are the reality check p-values of White's (2000) and Hansen's (2001) tests, respectively, where each model is regarded as a benchmark model and is compared with the remaining eight models. We use an AR(3)–SNP(5) model for the transformed PIT $\{x_t\}$. We retrieve the S&P 500 returns series from finance.yahoo.com. The sample observations are from January 3, 1990 to June 30, 2003 ($T = 3303$), the in-sample observations are from January 3, 1990 to September 24, 1996 ($R = 1703$), and the out-of-sample observations are from September 25, 1996 to June 30, 2003 ($n = 1700$).

Table 3. Reality Check Results Based on Minimum AIC.

Tail	Model	Left Tail					Right Tail				
		Loss	p_1	p_2	L	K	Loss	p_1	p_2	L	K
	MDS	0.0247	0.719	0.719	3	8					
	Constant	0.0231	0.938	0.933	3	3					
	MA	0.0266	0.534	0.534	3	7					
	ARMA	0.0236	0.874	0.862	3	3					
100%	AR	0.0266	0.533	0.533	3	7					
	PN	0.0374	0.238	0.238	1	8					
	NN	0.0238	0.853	0.845	3	3					
	SETAR	0.0266	0.539	0.539	3	7					
	STAR	0.0268	0.532	0.532	3	8					
	MDS	0.0583	1.000	0.767	3	8	0.0658	0.050	0.050	3	8
	Constant	0.0652	0.064	0.064	2	8	0.0593	0.717	0.675	3	8
	MA	0.0651	0.070	0.070	3	8	0.0582	0.960	0.947	3	8
	ARMA	0.0647	0.078	0.078	3	8	0.0588	0.839	0.813	3	8
10%	AR	0.0649	0.075	0.075	3	8	0.0594	0.701	0.655	3	8
	PN	0.0650	0.076	0.076	2	8	0.0628	0.239	0.239	3	8
	NN	0.0649	0.075	0.075	3	8	0.0597	0.619	0.571	3	8
	SETAR	0.0653	0.068	0.068	3	8	0.0584	0.845	0.808	3	8
	STAR	0.0653	0.063	0.063	3	8	0.0591	0.775	0.737	3	8
	MDS	0.0315	1.000	0.973	2	8	0.0390	0.333	0.092	3	8
	Constant	0.0334	0.628	0.473	2	7	0.0359	0.852	0.665	3	8
	MA	0.0339	0.468	0.320	2	8	0.0359	0.917	0.840	3	8
	ARMA	0.0340	0.425	0.266	3	8	0.0359	0.926	0.865	3	8
5%	AR	0.0336	0.495	0.340	3	8	0.0360	0.929	0.885	3	8
	PN	0.0356	0.256	0.129	3	8	0.0515	0.006	0.006	2	7
	NN	0.0337	0.487	0.327	3	8	0.0362	0.841	0.741	3	8
	SETAR	0.0334	0.529	0.386	3	8	0.0357	0.913	0.753	3	8
	STAR	0.0705	0.000	0.000	1	8	0.0359	0.908	0.826	3	8
	MDS	0.0098	0.973	0.939	3	7	0.0122	0.008	0.008	3	8
	Constant	0.0113	0.321	0.111	2	7	0.0113	0.042	0.042	3	8
	MA	0.0114	0.322	0.131	2	7	0.0116	0.018	0.018	3	8
	ARMA	0.0114	0.330	0.137	2	7	0.0118	0.019	0.019	3	8
1%	AR	0.0116	0.279	0.108	2	7	0.0118	0.020	0.020	3	8
	PN	0.0116	0.292	0.109	2	7	0.0092	0.989	0.564	3	7
	NN	0.0116	0.279	0.107	2	7	0.0118	0.020	0.020	3	8
	SETAR	0.0113	0.314	0.109	2	7	0.0112	0.060	0.060	3	7
	STAR	0.0229	0.000	0.000	2	8	0.0117	0.022	0.022	3	8

Note: "Loss" refers to is the out-of-sample averaged loss based on the KLIC measure; "p_1" and "p_2" are the reality check p-values of White's (2000) and Hansen's (2001) tests, respectively, where each model is regarded as a benchmark model and is compared with the remaining eight models; "L" and "K" are the AR and SNP orders, respectively, chosen by the minimum AIC criterion in the AR(L)–SNP(K) models, $L = 0, \ldots, 3$, $K = 0, \ldots, 8$ for the transformed PIT $\{x_t\}$. We retrieve the S&P 500 returns series from finance.yahoo.com. The sample observations are from January 3, 1990 to June 30, 2003 ($T = 3303$), the in-sample observations are from January 3, 1990 to September 24, 1996 ($R = 1703$), and the out-of-sample observations are from September 25, 1996 to June 30, 2003 ($n = 1700$).

Table 4. Reality Check Results Based on Minimum SIC.

Tail	Model	Left Tail					Right Tail				
		Loss	p_1	p_2	L	K	Loss	p_1	p_2	L	K
	MDS	0.0196	0.950	0.950	1	3					
	Constant	0.0211	0.628	0.628	1	3					
	MA	0.0216	0.562	0.562	1	3					
	ARMA	0.0215	0.570	0.570	1	3					
100%	AR	0.0215	0.580	0.580	1	3					
	PN	0.0342	0.157	0.157	2	4					
	NN	0.0214	0.599	0.599	1	3					
	SETAR	0.0221	0.538	0.538	1	3					
	STAR	0.0213	0.596	0.596	1	3					
	MDS	0.0583	1.000	0.767	3	8	0.0658	0.050	0.050	3	8
	Constant	0.0652	0.064	0.064	2	8	0.0593	0.717	0.675	3	8
	MA	0.0651	0.070	0.070	3	8	0.0582	0.960	0.947	3	8
	ARMA	0.0647	0.078	0.078	3	8	0.0588	0.839	0.813	3	8
10%	AR	0.0649	0.075	0.075	3	8	0.0594	0.701	0.655	3	8
	PN	0.0650	0.076	0.076	2	8	0.0628	0.239	0.239	3	8
	NN	0.0649	0.075	0.075	3	8	0.0597	0.619	0.571	3	8
	SETAR	0.0653	0.068	0.068	3	8	0.0584	0.845	0.808	3	8
	STAR	0.0653	0.063	0.063	3	8	0.0591	0.775	0.737	3	8
	MDS	0.0306	0.991	0.871	2	7	0.0390	0.333	0.092	3	8
	Constant	0.0334	0.384	0.213	2	7	0.0359	0.852	0.665	3	8
	MA	0.0328	0.416	0.241	3	7	0.0359	0.917	0.840	3	8
	ARMA	0.0328	0.411	0.238	3	7	0.0359	0.926	0.865	3	8
5%	AR	0.0324	0.495	0.318	3	7	0.0360	0.929	0.885	3	8
	PN	0.0343	0.250	0.098	3	7	0.0515	0.006	0.006	2	7
	NN	0.0325	0.489	0.311	3	7	0.0362	0.841	0.741	3	8
	SETAR	0.0323	0.522	0.357	3	7	0.0357	0.913	0.753	3	8
	STAR	0.0705	0.000	0.000	1	8	0.0359	0.908	0.826	3	8
	MDS	0.0000	1.000	0.544	3	1	0.0016	0.077	0.077	3	2
	Constant	0.0043	0.000	0.000	3	3	0.0014	0.090	0.090	3	2
	MA	0.0044	0.000	0.000	3	3	0.0000	1.000	0.999	3	1
	ARMA	0.0044	0.000	0.000	3	3	0.0000	1.000	0.999	3	1
1%	AR	0.0045	0.000	0.000	3	3	0.0000	1.000	0.997	3	1
	PN	0.0044	0.000	0.000	3	3	0.0014	0.109	0.109	3	2
	NN	0.0045	0.000	0.000	3	3	0.0000	0.852	0.718	3	1
	SETAR	0.0045	0.000	0.000	3	3	0.0015	0.095	0.095	3	2
	STAR	0.0229	0.000	0.000	2	8	0.0000	0.910	0.757	3	1

Note: "Loss" refers to is the out-of-sample averaged loss based on the KLIC measure; "p_1" and "p_2" are the reality checks p-values of White's (2000) and Hansen's (2001) tests, respectively, where each model is regarded as a benchmark model and is compared with the remaining eight models; "L" and "K" are the AR and SNP orders, respectively, chosen by the minimum SIC criterion in the AR(L)–SNP(K) models, $L = 0, \ldots, 3$, $K = 0, \ldots, 8$, for the transformed PIT $\{x_t\}$. We retrieve the S&P 500 returns series from finance.yahoo.com. The sample observations are from January 3, 1990 to June 30, 2003 ($T = 3303$), the in-sample observations are from January 3, 1990 to September 24, 1996 ($R = 1703$), and the out-of-sample observations are from September 25, 1996 to June 30, 2003 ($n = 1700$).

Next, comparing the left tails with $\alpha = 10\%, 5\%, 1\%$, we find the results are similar to those for the entire distribution. That is, the MDS model has the smallest KLIC loss values for these left tails, much smaller than those of the other eight models. The reality check p-values for the MDS model as the benchmark are all very close to one, indicating that none of the eight models are better than the MDS, confirming the efficient market hypothesis in the left tails of the S&P 500 returns.

Interestingly and somewhat surprisingly, when we look at the right tails with $\alpha = 90\%, 95\%, 99\%$ (i.e., the right 10%, 5%, and 1% tails, respectively), the results are exactly the opposite to those for the left tails. That is, the KLIC loss values for the MDS model for all of these three right tails are the largest, larger than those of the other eight competing models. The reality check p-values with the MDS model as the benchmark are zero or very close to zero, indicating that some of the other eight models are significantly better than the MDS model, hence rejecting the efficient market hypothesis in the right tails of the S&P 500 return density. This implies that the S&P 500 returns may be more predictable when the market goes up than when it goes down, during the sample period from January 3, 1990 to June 30, 2003. To our knowledge, this empirical finding is new to the literature, obtained as a benefit of using the BLS method that permits evaluation and comparison of the density forecasts on a particular area of the return density.

As the right tail results imply, the S&P 500 returns in the right tails are predictable via some of the eight linear and nonlinear competing models considered in this paper. To see the nature of the nonlinearity in mean, we compare the KLIC loss values of these models. It can be seen that the PN model has the smallest loss values for all the three right tails. The reality check p-values show that the PN model (as a benchmark) is not dominated by any of the remaining models for the three 10%, 5%, and 1% right tails. The other three nonlinear models (NN, SETAR, and STAR) are often worse than the linear models in terms of out-of-sample density forecasts. We note that, while PN is the best model for the right tails, it is the worst model for forecasting the entire density ($\alpha = 100\%$).

Summing up, the significant in-sample evidence from the generalized spectral statistics $M(1, 2)$ and $M(1, 4)$ reported in Table 1 suggests that the squared lagged return and the fourth order power of the lagged returns (i.e. the conditional kurtosis representing the influence of the tail observations) have a predictive power for the returns. The out-of-sample evidence adds that this predictability of the squared lagged return and the fourth order power of the lagged returns is *asymmetric*, i.e., significant only when the market goes up.

Table 3 shows the BLS test results computed with L and K chosen by the AIC. The results of Table 3 are similar to those of Table 2. Comparing the entire distribution with $\alpha = 100\%$, the large reality check p-values for the MDS model as the benchmark ($p_1 = 0.719$ and $p_2 = 0.719$) indicate that none of the other eight models are better than the MDS model (although the KLIC loss value for the MDS model is not the smallest). This confirms the market efficiency hypothesis that the S&P500 returns may not be predictable using linear or nonlinear models. The results for the left tails with $\alpha = 10\%$, 5%, 1% are also comparable to those in Table 2. That is, the KLIC loss values for the MDS model for these left tails are the smallest. The reality check p-values with the MDS model as the benchmark are all close to one ($p_1 = 1.000$, 1.000, 0.973 and $p_2 = 0.767$, 0.973, 0.939), indicating that none of the other eight models are better than the MDS model, again confirming the market efficiency hypothesis in the left tails of the S&P 500 returns. The right tail results are different from those for the left tails, as in Table 2. The MDS model for the right tails is generally worse than the other eight models. The reality check p-values of Hansen (2001) for the MDS model as the benchmark are very small ($p_2 = 0.050$, 0.092 and 0.008) for the three right tails, indicating that some of the eight models are significantly better than the MDS model. This shows that the S&P 500 returns may be more predictable when the market goes up than when it goes down – the nonlinear predictability is asymmetric.

Table 4 shows the BLS test results computed with L and K chosen by the SIC. The results are very similar to those of Tables 2 and 3. Comparing the entire distribution with $\alpha = 100\%$, the KLIC loss value for the benchmark MDS model is the smallest with the large reality check p-values ($p_1 = 0.950$ and $p_2 = 0.950$). The results for the left tails with $\alpha = 10\%$, 5%, 1% are consistent with those in Tables 2 and 3. That is, the KLIC loss values for the MDS model for these left tails are the smallest. The reality check p-values with the MDS as the benchmark are all close to one ($p_1 = 1.000$, 0.991, 1.000 and $p_2 = 0.767$, 0.871, 0.544), indicating that none of the other eight models are better than the MDS model, again confirming the market efficiency hypothesis in the left tails. The story for the right tails is different from that for the left tails, as in Tables 2 and 3. The MDS model for the right tails is generally worse than the other eight models. The reality check p-values of Hansen (2001) for the MDS model as the benchmark are very small ($p_2 = 0.050$, 0.092, and 0.077) for the three right tails, indicating that some of the linear and nonlinear models are significantly better than the MDS model. This shows that the S&P 500 returns may be more predictable when the market goes up than when it goes down.

5. CONCLUSIONS

In this paper, we examine nonlinearity in the conditional mean of the daily closing S&P 500 returns. We first examine the nature of the nonlinearity using the in-sample test of Hong (1999), where it is found that there are strong nonlinear predictable components in the S&P 500 returns. We then explore the out-of-sample nonlinear predictive ability of various linear and nonlinear models. The evidence for out-of-sample nonlinear predictability is quite weak in the literature, and it is generally accepted that stock returns follow a martingale. While most papers in the literature use MSFE, MAFE, or some economic measures (e.g., wealth or returns) to evaluate nonlinear models, we use the density forecast approach in this paper.

We find that for the entire distribution the S&P 500 daily closing returns are not predictable, and various nonlinear models we examine are no better than the MDS model. For the left tails, the returns are not predictable. The MDS model is the best in the density forecast comparison. For the right tails, however, the S&P 500 daily closing returns are found to be predictable by using some linear and nonlinear models. Hence, the out-of-sample nonlinear predictability of the S&P 1500 daily closing returns is asymmetric. These findings are robust to the choice of L and K in computation of the BLS statistics.

We note that the asymmetry in the nonlinear predictability which we have found is with regards to the two tails of the return distribution. Our results may not imply that the bull market is more predictable than the bear market because stock prices can fall (left tail) or rise (right tail) under both market conditions. Nonlinear models that incorporate the asymmetric tail behavior as well as the bull and bear market regimes would be an interesting model to examine, which we leave for future research.

NOTES

1. We would like to thank Yongmiao Hong for sharing his GAUSS code for the generalized spectral tests.
2. The GAUSS code is available upon request.

ACKNOWLEDGEMENT

We would like to thank the co-editor Tom Fomby and an anonymous referee for helpful comments. All remaining errors are our own.

REFERENCES

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

Bao, Y., Lee, T.-H., & Saltoglu, B. (2004). A test for density forecast comparison with applications to risk management. Working paper, University of California, Riverside.

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

Bradley, M. D., & Jansen, D. W. (2004). Forecasting with a nonlinear dynamic model of stock returns and industrial production. *International Journal of Forecasting*, *20*, 321–342.

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

Clements, M. P., & Smith, J. (2001). Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance*, *20*, 133–148.

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

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*, 363–390.

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

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

Hansen, P. R. (2001). An unbiased and powerful test for superior predictive ability. Working paper, Stanford University.

Hong, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association*, *84*, 1201–1220.

Hong, Y., & Lee, T.-H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, *85*(4), 1048–1062.

Kanas, A. (2003). Non-linear forecasts of stock returns. *Journal of Forecasting*, *22*, 299–315.

Kullback, L., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, *1*, 41–66.

Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, *3*, 175–208.

McMillan, D. G. (2001). Nonlinear predictability of stock market returns: Evidence from nonparametric and threshold Models. *International Review of Economics and Finance*, *10*, 353–368.

Meese, R. A., & Rogoff, K. (1983). The out of sample failure of empirical exchange rate models: Sampling error or misspecification. In: J. Frenkel (Ed.), *Exchange rates and international economics*. Chicago: University of Chicago Press.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*, 1303–1313.

Racine, J. (2001). On the nonlinear predictability of stock returns using financial and economic variables. *Journal of Business and Economic Statistics*, *19*(3), 380–382.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

Wright, J. H. (2000). Alternative variance-ratio tests using ranks and signs. *Journal of Business and Economic Statistics*, *18*, 1–9.