Journal of Empirical Finance 10 (2003) 321–353

Journal of
EMPIRICAL
FINANCE

www.elsevier.com/locate/econbase

Realized volatility in the futures markets

Dimitrios D. Thomakos[1], Tao Wang*

Department of Economics, Florida International University, Florida, USA

Abstract

Using intraday returns on four futures contracts over a 5-year period, we calculate and analyze model-free measures of futures return volatility. We focus on the temporal characteristics and distributional properties of daily returns, return volatilities, (log of) standard deviations, standardized returns and pairwise correlations. The behavior of a number of tests for Gaussianity under long memory is explored via a simulation study. The simulation results indicate that tests of the "goodness-of-fit" variety are appropriate to use while the commonly employed Jarque–Bera test is severely oversized and its use is not recommended. We find that the standard deviations and the pairwise correlations exhibit long memory while the standardized returns are serially uncorrelated. We also find that the (unconditional) distributions of daily returns' volatility are leptokurtic and highly skewed to the right while the distributions of the standardized returns, the standard deviations and the pairwise correlations are statistically indistinguishable from the Gaussian distribution. Our results are consistent with that of Andersen et al. [Journal of the American Statistical Association 96 (2001a) 42; Journal of Financial Economics 61 (2001c) 43] on the time-series properties of realized volatility. The dynamic characteristics of the volatility series are modelled using fractionally integrated ARMA models.

JEL classification: C32; G1
Keywords: Realized volatility; High-frequency data; Long memory; Tests for normality; ARFIMA

1. Introduction

Accurate measures and good forecasts of asset return volatility are critical for the implementation and evaluation of asset pricing theories, portfolio choice and risk management. Asset return volatility can not be directly observed and is time-varying. As a result,

* Corresponding author. Tel.: +1-305-348-2682; fax: +1-305-348-1524.
E-mail addresses: thomakos@fiu.edu (D.D. Thomakos), wangtao@fiu.edu (T. Wang).
[1] Tel.: +1-305-348-6639.

0927-5398/02/$ - see front matter
doi:10.1016/S0927-5398(02)00052-X

economists have built increasingly sophisticated statistical models to capture the characteristics of financial markets' volatility. Previous studies focus on parametric ARCH and GARCH models, stochastic volatility models, implied volatility from certain option pricing models or direct indicators of volatility such as ex post squared or absolute returns. Partial surveys of the voluminous literature on these models are given by Bollerslev et al. (1994), Ghysels et al. (1996) and Campbell et al. (1997).

Recently, there has been renewed interest in obtaining improved daily volatility estimates by using high-frequency, intraday returns to construct daily "realized" or "integrated" volatility. Daily realized volatility is obtained as the sum of intraday squared returns. Using the theory of quadratic variation, Andersen and Bollerslev (1998) and Barndorff-Nielsen and Shephard (2001a) show that the realized volatility estimator, say v_t^2, is a consistent estimator of the actual volatility in the sense that $v_t^2 \to \sigma_t^2$, where σ_t^2 is the actual (but unobservable) volatility.[2] This is an interesting result for it is model-free and does not depend upon any particular form of σ_t^2.

In this paper, we construct and examine the distributional properties of daily futures realized volatility using high-frequency, intraday futures returns. We consider four of the most heavily traded futures contracts worldwide: the Deutsche Mark, the S&P500 index, US Bonds and the Eurodollar.[3]

We find that the logarithmic realized standard deviations exhibit long memory and approximate Gaussianity, while the standardized returns also exhibit approximate Gaussianity but are serially uncorrelated. Our analysis confirms previous work by Andersen et al. (2001a) (henceforth ABDL) and by Andersen et al. (2001c) (henceforth ABDE), who found similar results for daily foreign exchange return volatilities and daily stock return volatilities, respectively. Our approach differs, however, with the existing literature in several respects. First, rather than relying on descriptive statistics and graphs in investigating the distributional properties of the time series studied, as in ABDL (2001a) and ABDE (2001c), we perform a variety of formal statistical tests designed to capture the distributional properties of the volatility and return series. Second, we conduct a simulation study to examine the properties of these tests in the presence of long memory. Third, we model the realized volatility and logarithmic standard deviations using ARFIMA models and investigate the time-series properties of the series once long memory is removed.[4] Finally, we apply the methodology of realized volatility in the futures markets using a new dataset.

---

[2] Barndorff-Nielsen and Shephard (2001b) examined the finite-sample properties of the realized volatility estimator in a simulation study. Their experiments suggest that the realized volatility error can be substantial but the theoretical results are reliable for the realized logarithmic standard deviation.

[3] The concept of realized volatility has been used in financial economics for a number of years, but has been applied mainly to data with lower than intraday frequency. Hsieh (1991) uses 15-min equity returns to construct daily standard deviations. Campbell et al. (2001), Merton (1980), Poterba and Summers (1986), Schwert (1990) and French et al. (1987) used daily returns in constructing monthly stock return volatilities. However, little was known about the properties of the realized volatility estimator v_t^2 until recently.

[4] Our modeling of realized volatility can be viewed as a univariate complement to the work of Andersen et al. (2001b), who use a fractional vector autoregressive model (FVAR) to examine the joint behavior of three realized volatility series.

Our main motivation for applying a number of different tests for Gaussianity came from the observation that the commonly employed Jarque–Bera statistic (essentially a joint test for zero skewness and zero excess kurtosis) has been shown to over-reject the hypothesis of normality, especially in the presence of serially correlated observations. This result is frequently observed even when the empirical distribution function (edf) "looks" indistinguishable from the Gaussian distribution function.[5] In addition to the Jarque–Bera test, we use a number of normality tests, all of which are of the "goodness-of-fit" variety. They are the Anderson–Darling test, the Crámer–Von Misses test, the QQ test (the correlation coefficient between the theoretical quantiles of the standard normal and the order statistics of the sample) and the Kolmogorov–Smirnov test. Both the Jarque–Bera and the other normality statistics have asymptotic distributions that are valid under the assumption of independent observations. Since we find that our series can be well characterized as long-memory processes, we provide an illustrative simulation study about the properties of these tests when the data generating process exhibits long memory. Our simulation results complement the theoretical work of Beran and Gosch (1991), and clearly suggest the use of the "goodness-of-fit" tests for normality, as we find that the Jarque–Bera test is highly oversized (i.e., over-rejects normality). Additional details can be found in Appendix B.

To preview our results, we find that the distributions of the raw returns for all four futures contracts are highly non-Gaussian and leptokurtic. After normalizing the raw returns by the realized standard deviations, the standardized return distributions become statistically indistinguishable from the Gaussian distribution (the evidence on the Eurodollar futures contract is slightly mixed).The distributions of realized volatilities are all highly non-Gaussian and skewed to the right. However, the distributions of the logarithmic standard deviations for all the four futures contracts are approximately Gaussian. The distributions of the six daily covariances among the Deutsche Mark, the Eurodollar, the S&P500 and the T-bonds are highly non-Gaussian. Among the distributions of the daily correlations, we find evidence of Gaussianity only for the pairs Deutsche Mark/T-bonds, Eurodollar/T-bonds and S&P500/T-bonds. Our results suggest that all of the volatility measures can be characterized by slowly mean-reverting fractionally integrated processes with a degree of integration, d, estimated between 0.25 and 0.45. In addition, we find that asymmetric volatility, the asymmetry in the relationship between past negative and positive returns and future volatilities, is not statistically significant for the futures contracts studied except possibly for the S&P500 index. Finally, we model the realized volatility series using fractionally integrated ARMA (ARFIMA) models.

The remainder of the paper is organized as follows. The next section describes the data and volatility calculations. The temporal characteristics of the volatility and correlations are discussed in Section 3. Section 4 details the unconditional univariate return and volatility distributions. Section 5 discusses the leverage effects in the realized volatilities and realized correlations. The realized volatility modeling is discussed in Section 6. Section 7 offers some concluding remarks. Appendices A and B contain details about the

[5] In ABDL (2001a), even though the edfs of the logarithmic standard deviations, correlations and the daily standardized returns appear to be Gaussian, the Jarque–Bera test statistic rejected the normality hypothesis for all the series.

test statistics and the simulation results on the finite sample properties of the normality tests used in the paper.

2. Measurement and data

2.1. Data

In this section, we describe our dataset and discuss measurement issues. The T-bonds contract is traded on the Chicago Board of Trade (CBOT). The S&P500, the Deutsche Mark and the Eurodollar contracts are traded on the International Monetary Market unit of the Chicago Mercantile Exchange (CME).[6] These markets were chosen for two reasons. First, prices are available on a tick-by-tick basis and allows us to construct 5-min returns. The daily transaction record extends from 9:30 EST until 16:15 EST for the S&P500 contract, and from 8:20 EST until 15:00 EST for the Deutsche Mark, the Eurodollar and T-bonds contracts. There is a total of eighty-one 5-min returns for each day for the S&P500 contract, and a total of eighty 5-min returns for each day for the Deutsche Mark, the Eurodollar and T-bonds contracts. Second, all four contracts are heavily traded. The T-bonds contract, which calls for delivery of a U.S. Treasury bond with 15 or more years to maturity, is the most heavily traded long-term interest rate contract in the world until the G-bund took over the claim recently. The Eurodollar contract, which specifies cash delivery based on the 3-month London interbank offered rate, is the most heavily traded short-term interest rate futures contract. While the Eurodollar rate is the rate paid on offshore interbank deposits (to avoid insurance fees and regulations), it is a dollar rate dependent on domestic monetary conditions. The S&P500 futures index contract is one of the most heavily traded equity futures contract in the United States. Lastly, during our sample period, the Deutsche Mark futures contract was the most heavily traded currency futures contract on the CME. Our data are time and sale transaction prices, not bid-ask quotes, recorded by exchange personnel who observe the pits and post the most recent price.[7] Our data set begins January 3, 1995 and ends December 31, 1999 for the S&P500 and the Eurodollar contracts, a total of $T = 1261$ trading days for the S&P500, $T = 1258$ for the Eurodollar; begins October 2, 1995 and ends September 30, 1999 for T-bonds contract, a total of $T = 1003$ trading days; and begins January 3, 1995 and ends August 6, 1999 for the Deutsche Mark contract, a total of $T = 1157$ trading days.[8] In calculating the returns series, we use the nearby contracts to construct the continuous returns series for each futures contract.

Since the exchange does not record successive trades at the same price, the median length between price changes should be the upper bound for the median length of the times between trades. During the actual trading, market microstructure frictions, including price

[6] During the time period when the data are available, both the CME and CBOT have dual trading systems: trading in the pit and electronic trading, but in different hours. However, the trading volume in the pit dominates that in the electronic trading.

[7] The exchange observers record every change in price but do not separately record successive trades at the same price or trade volume.

[8] The four contracts do not necessarily trade at the same time. They differ in terms of their holiday schedules and trading hours. Detailed trading schedules are on the CME and CBOT websites.

discreteness, infrequent trading and measurement errors, affect the actual price record. In order to mitigate these effects while still sampling at a very high frequency, we use 5-min return horizon in the analysis. We briefly describe the reasons for selecting the 5-min interval and the construction of the 5-min return series in the following section.

2.2. Construction of realized futures volatility

Following ABDE (2001c) and ABDL (2001a), we construct the 5-min return series from the logarithmic differences between the prices recorded at or immediately before the corresponding 5-min marks. In practice, the selection of fixed-discrete-time return depends on the balance between the need of sampling at very high frequencies to approximate continuous-time models and the cost of market microstructure biases. Since tick-by-tick prices are generally available at unevenly spaced time points, it is well known that the uneven spacing of the observed prices and inherent bid-ask spreads may induce negative autocorrelation in the fixed-discrete-time return series.[9] Hence, the use of a fixed discrete time interval may allow dependence in the mean to systematically bias our volatility measures. Ultimately, the choice of fixed-interval return depends on market liquidity of the contract studied. Compared with the Dow Jones stocks which has a median intertrade duration varying from 7 s for Merck to 54 s for United Technology (ABDE, 2001c, p. 49), the median length of duration between price changes for our four futures contracts ranges from a low of 5 s for the S&P500 contract to a high of 44 s for the Eurodollar contract.[10] Therefore, the selection of 5-min returns is consistent with ABDE (2001c), who also use 5-min returns to construct their volatility series.

In order to remove the serial correlation induced by the uneven spacing of the observed prices and the bid-ask spread, we use an MA(1) model to "filter" the (demeaned) 5-min return series. The estimated innovations are then used in place of the demeaned returns.[11] Consistent with the spurious dependence which would be induced by the uneven spacing of the observed prices and inherent bid-ask spread, all estimated moving-average coefficients are negative, ranging from -0.18 for the Eurodollar contract to -0.02 for the Deutsche Mark contract.

Let the (4×1) vector of raw, 5-min returns be denoted by $\boldsymbol{r}_{t+i\Delta,\Delta}$, where t stands for the tth day, i stands for ith 5-min mark and $1/\Delta$ stands for the number of 5-min returns during a trading day. The filtered series are then given by:

$$\hat{\varepsilon}_{t+i\Delta,\Delta} = [\boldsymbol{I} + \hat{\Theta}L]^{-1}(\boldsymbol{r}_{t+i\Delta,\Delta} - \bar{\boldsymbol{r}}_{\Delta}) \tag{1}$$

where $\hat{\Theta} = \mathrm{diag}(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$ is the (4×4) diagonal matrix of estimated MA coefficients and $\bar{\boldsymbol{r}}_{\Delta}$ is the estimated sample mean vector over all observations

$$\bar{\boldsymbol{r}}_{\Delta} = (\Delta/T) \sum_{t=1}^{T} \sum_{i=1}^{1/\Delta} \boldsymbol{r}_{t+i\Delta,\Delta}.$$

9 See, for example, Campbell et al. (1997).

10 The average number of price changes per day is 783 for the Deutsche Mark, 163 for the Eurodollar, 3366 for the SP500 and 1710 for T-bonds.

11 We also experimented with the use of unfiltered and linearly interpolated 5-min returns. The empirical results were very similar, and are available upon request.

The realized daily covariance matrix, denoted $\boldsymbol{C}_t$, for the four futures contracts is then computed as:

$$\boldsymbol{C}_t = \sum_{i=1}^{1/\Delta} \hat{\varepsilon}_{t+i\Delta,\Delta} \hat{\varepsilon}'_{t+i\Delta,\Delta} \tag{2}$$

where $t = 1, 2, \ldots, T$. T is 1261, 1258, 1003, 1157 for the S&P500, the Eurodollar, T-bonds and the Deutsche Mark contracts, respectively. $1/\Delta = 81$ or 80, depending on the contract. The realized volatilities for the four futures contracts are given by the diagonal elements of the realized daily covariance matrix denoted by v_{tj}^2, $j = 1, 2, 3, 4$. The corresponding daily logarithmic standard deviations are denoted as $lv_{tj} = \ln v_{tj}$. The realized covariances are given by the off-diagonal elements of the daily covariance matrix. For the pairs that have the same trading hour schedule, i.e., for the covariances between the Deutsche Mark, the Eurodollar and the T-bonds, the daily covariances are constructed by summing the cross products of the 5-min returns within the day. For the pairs that do not have the same trading schedule, i.e., between the S&P500 and other three contracts, the daily covariances are constructed by summing cross products of 5-min returns when there is an overlap in trading hours with the first three 5-min overlapped returns excluded.[12] Due to the effect of nonsynchronous trading between S&P500 and other three contracts, the correlations between the S&P500 and other three contracts could be potentially biased. There will be lagged response to news in the S&P500 futures market if news arrives after the Eurodollar, T-bonds and the Deutsche Mark futures markets open, but before the opening of the S&P500 futures. This lagged response will induce spurious covariances (Campbell et al., 1997, p. 85). Ederington and Lee (1993) show that "the price adjustment to a major news announcement occurs within the first minute, but volatility remains significantly higher than normal for roughly fifteen minutes...", providing a justification for the exclusion of the first three 5-min returns.

With the above in mind, the realized daily correlation matrix is computed as:

$$\boldsymbol{R}_t = \boldsymbol{V}_t^{-1/2} \boldsymbol{C}_t \boldsymbol{V}_t^{-1/2} \tag{3}$$

where $\boldsymbol{V}_t = \mathrm{diag}(v_{t1}^2, v_{t2}^2, v_{t3}^2, v_{t4}^2)$. By explicitly treating the volatilities and correlations as observable, we can now rely on conventional statistical procedures to characterize their distributional properties.

As a matter of definition, let $\boldsymbol{r}_t$ denote the vector of the daily unfiltered raw return series and $\bar{\mathbf{r}}$ denote the corresponding sample mean over all trading days. The vector of standardized daily returns is now given as:

$$z_t = \boldsymbol{V}_t^{-1/2} (\boldsymbol{r}_t - \bar{\boldsymbol{r}}) \tag{4}$$

3. Persistence and long memory

One of the empirical regularities of asset returns is the long-run temporal dependencies of return volatilities. The literature on volatility modeling has documented that such temporal

[12] We would like to thank the referee for this suggestion.

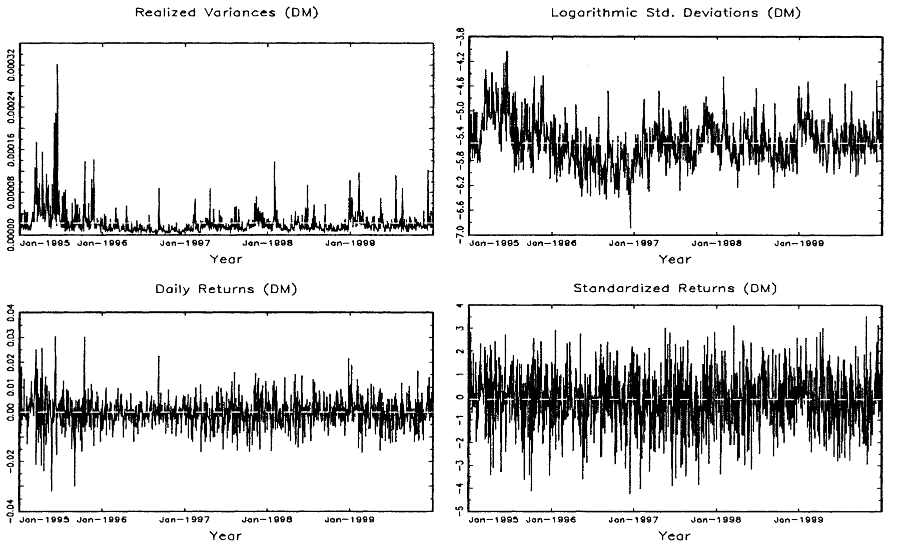

Fig. 1. Time series for the Deutsche Mark series.

dependencies are highly persistent. Fig. 1 plots the time series for the realized variance and realized logarithmic standard deviation (along with the return and the standardized return) for the Deutsche Mark futures contract.[13] The two volatility measures (top panel) appear positively serially correlated, consistent with the well documented volatility clustering effect reported in the ARCH, stochastic volatility and realized volatility literature.[14]

The following section explores the time-series properties of the logarithmic standard deviations, correlations and standardized returns in more detail.[15]

3.1. Logarithmic standard deviations

Table 1 lists a number of test statistics about the temporal correlation properties of the logarithmic standard deviations for all series. The fourth column summarizes the value of the standard Ljung–Box portmanteau test for the joint significance of the first 20 autocorrelations of the logarithmic realized daily standard deviations, lv_{tj}. The null hypothesis of no serial correlation is overwhelmingly rejected for all four futures contracts. The fifth column presents the value of the McLeod–Li test statistics; the test results are the same as the results from the Ljung–Box test.[16] The top two panels of Fig. 2 depict the

[13] Throughout the discussion, we only report figures for the Deutsche Mark contract. Figures for the other contracts are similar and available upon request.

[14] See Lamoureux and Lastrapes (1993), Kim and Kon (1994), ABDE (2001c) and ABDL(2001a).

[15] Note that we address the serial correlation characteristics of the time series first and then discuss their distributional properties. This is because the behavior of the test statistics used to assess the distributional properties depends on the temporal persistence of the series.

[16] This test has the same format and distribution as the Ljung–Box test but is applied to the squares of the series.

Table 1
Dynamic volatility dependence

	Daily standard returns $[(r_t - r)/v_t]$			log S.D. (log V_t)			Correlations			
	Q_{20}	Q_{20}^2	$\hat{d}_{SP}$	Q_{20}	Q_{20}^2	$\hat{d}_{SP}$	Pair	Q_{20}	Q_{20}^2	$\hat{d}_{SP}$
DM	29.73	1.58	− 0.0028 (0.0298)	2973	229	0.3508 (0.0298)	DM, ED	650	723	0.2892 (0.0327)
ED	21.66	11.92	− 0.0423 (0.0288)	5373	89	0.3949 (0.0288)	DM, US	636	936	0.2615 (0.0327)
SP	40.44	4.84	− 0.1137 (0.0288)	7458	164	0.4562 (0.0288)	ED, US	279	111	0.2443 (0.0327)
US	20.13	6.57	− 0.0576 (0.0315)	704	35	0.2598 (0.0315)	SP, DM	952	871	0.3313 (0.0316)
							SP, ED	4733	6109	0.3976 (0.0316)
							SP, US	9001	4818	0.5023 (0.0316)

(1) The sample covers the period from January 3, 1995 through December 31, 1999 for the S&P500 and Eurodollar contracts; October 2, 1995 through September 30, 1999 for the U.S. bond contract; January 3, 1995 through August 6, 1999 for the Deutsche Mark.
(2) DM stands for the Deutsche Mark, ED for the Eurodollar, SP for the S&P500 and US for the U.S. T-bonds.
(3) Q_{20} is the Ljung–Box portmanteau test for autocorrelation using 20 lags. The 5% critical value of the X_{20}^2 (0.05) distribution is 31.41. X_{20}^2 is the McLeod–Li portmanteau test for autocorrelation in the squares of the series using 20 lags.
(4) $\hat{d}_{SP}$ is the estimate of the fractional order using the Gaussian semiparametric estimator of Robinson (1995b). Standard errors are in parentheses.

autocorrelations for the two volatility measures regarding the Deutsche Mark futures. From the figure, we see that the autocorrelations are systematically above the conventional Bartlett 95% confidence error bands, confirming the results from the Ljung–Box and the McLeod–Li tests. Moreover, the two autocorrelations start around 0.4 and 0.5 and decay very slowly.

The low first-order autocorrelations, along with their slow decay, suggest that the logarithmic realized standard deviations do not contain a unit root but exhibit long memory. A number of studies argue that the long-run persistence in financial market volatility may be modeled by fractional integrated ARCH or stochastic volatility models (see, for example, Breidt et al., 1998; Robinson and Zaffaroni, 1998; ABDE, 2001c). The sixth column of Table 1 presents the estimates for the degree of fractional integration, or d, for the logarithmic standard deviations for all the four futures contracts. These estimates are obtained using the Gaussian semiparametric estimator of Robinson (1995b).[17] The estimates for the degree of fractional order for the four contracts are between 0.25 and 0.45, and are all statistically significantly different from both 0 and 1. Importantly, the parameter estimates d are more than two standard errors away from 0.5 (with the exception of the upper bound of the S&P500 confidence interval which is 0.5126, just outside of the stationary range) implying that the realized logarithmic standard deviations can be treated as covariance stationary and fractionally integrated.

[17] Similar results were obtained by using the log-periodogram regression method (see, for example, Geweke and Porter-Hudauk, 1983; Robinson, 1995a).

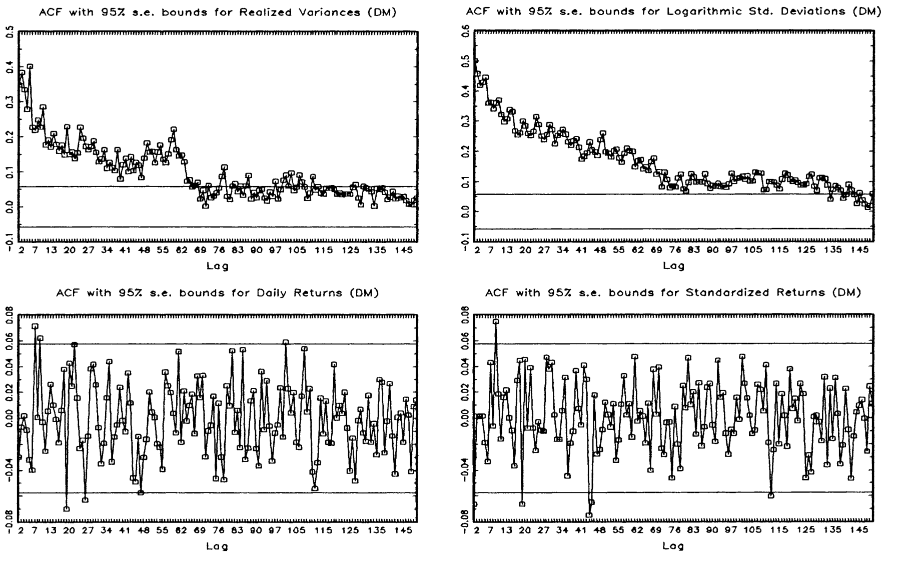

Fig. 2. Correlograms for the Deutsche Mark futures.

3.2. Correlations

Fig. 3 depicts the correlograms for the correlation pairs among the Deutsche Mark, Eurodollar, S&P500 and T-bonds futures. All panels indicate temporal dependencies. The Ljung–Box statistics for the joint significance of the first 20 autocorrelations, given in the

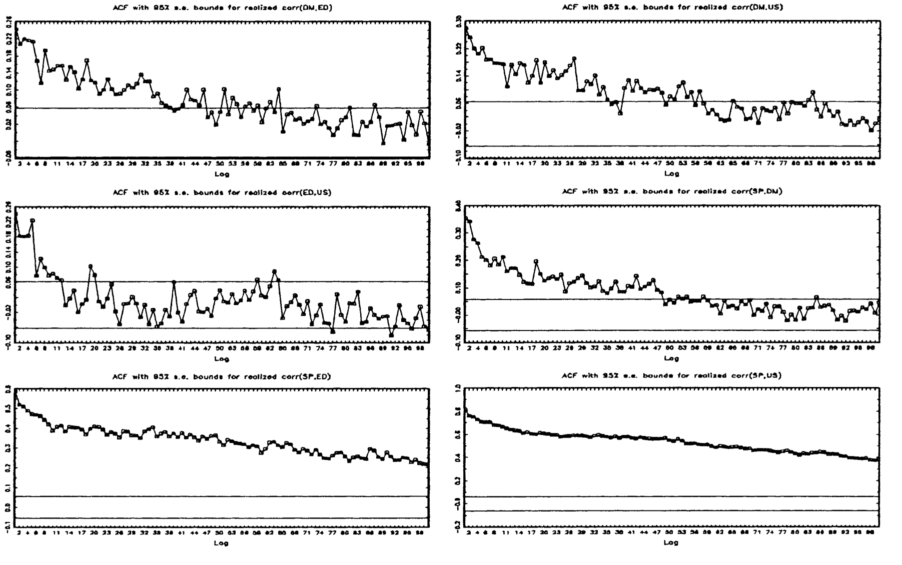

Fig. 3. Correlograms for the correlations.

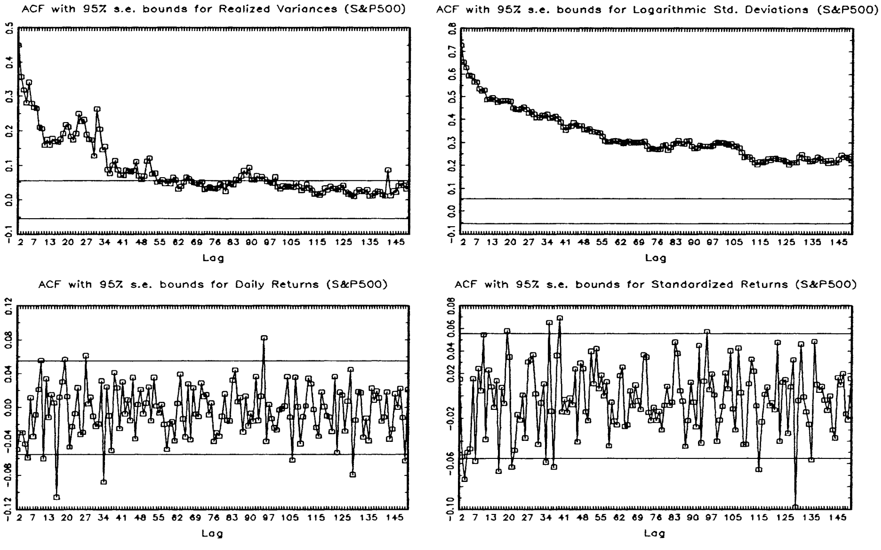

Fig. 4. Correlograms for the S&P500 futures.

rightmost columns of Table 1, confirm the visual results from the correlograms and reject the null hypothesis of no correlation. The results are similar for the McLeod–Li statistics. The correlograms for the correlation pairs between the S&P500 and the Eurodollar and the S&P500 and the T-bonds contracts start around 0.6 and 0.8, respectively, and decay very slowly. They are much more persistent than other pairs whose autocorrelations start around 0.3.[18] Our estimates for the fractional order *d* for the correlations are significantly different from 0 and 1 with values varying from 0.24 to 0.50.The realized correlations for all pairs, except for the pair between S&P500 and T-bonds, can be treated as covariance stationary as the parameter estimates for the fractional order *d* are more than 2 S.D. away from 0.5 (with the exception of the upperbound of the S&P500 and T-bond pair confidence interval which is 0.5642).

3.3. Standardized returns

The standardized returns series are, visually and statistically, serially uncorrelated. The results in the left panel of Table 1 indicate that the estimates of the fractional order *d* are all negative (indicating short memory or antipersistence) and statistically insignificant from zero at the conventional 5% level, except for the S&P500 series which has a *t*-ratio of -3.95. The two portmanteau tests for serial correlations have values less than the critical value of the $\chi^2_{20}(0.05)$ distribution, which is 31.41, again except for the S&P500 series whose Ljung–Box test value is 40.44. The McLeod–Li test does not indicate any significant correlation in the squares of the series. We note, however, that the correlogram

[18] We also experimented with excluding more 5-min returns in constructing daily correlations between the S&P500 and the other three futures returns. The results do not change.

values of the S&P500 series are inside the 95% critical values of $1.96T^{-1/2}$, indicating that the series has no serial correlation (Fig. 4). Therefore, there is strong evidence that all standardized returns series are serially uncorrelated.

In summary, the temporal characteristics of the realized logarithmic standard deviations, realized correlations and the standardized returns for our futures returns series are consistent with the findings from ABDE (2001c) for the Dow Jones stocks and ABDL (2001a) for the exchange rates. Both the realized logarithmic standard deviations and realized correlations can be adequately described by long-memory processes while the standardized returns series are serially uncorrelated.

4. Distributional properties

4.1. Returns and standardized returns

The summary statistics in Table 2 show that the daily raw returns analyzed have fatter tails than the normal distribution. Two futures returns (the Deutsche Mark and the

Table 2
Unconditional daily returns distributions

	Returns (r_t)				Standard returns $[(rt-r)/v_t]$			
	D-Mark	E-Dollar	S&P500	T-bonds	D-Mark	E-Dollar	S&P500	T-bonds
N	1157	1258	1261	1003	1157	1258	1261	1003
Mean	− 0.0002	2.62E− 5	0.0008	0.0001	− 0.1050	0.0248	0.2012	0.0516
S.D.	0.0062	0.0005	0.0109	0.0053	1.2805	0.7921	1.1146	1.0484
Skewness	0.1521	0.2966	− 0.5097	− 0.4824	− 0.1613	0.1644	0.2184	0.0130
Kurtosis	5.9134	10.5153	9.0817	4.9463	3.0147	3.1579	3.0521	2.5032
AD	7.5938[a]	28.6440[a]	12.2239[a]	2.4311[a]	0.6285	0.8174	0.8987	1.2203
CVM	1.3934[a]	5.2202[a]	2.1093[a]	0.3401[a]	0.0918	0.1521	0.1632	0.1887
JB	412.94[a]	2976.6[a]	1996.4[a]	196.61[a]	5.0164[b]	6.9657[b]	10.1583[a]	10.3125[a]
KS	1.8985[a]	4.2710[a]	2.4716[a]	1.1798[a]	0.8123	1.8729[a]	0.9413	0.7633
QQ	0.9836[a]	0.9467[a]	0.9695[a]	0.9891[a]	0.9986	0.9987	0.9984	0.9976[c]

(1) The sample covers the period from January 3, 1995 through December 31, 1999 for the S&P500 and Eurodollar contracts; October 2, 1995 through September 30, 1999 for the U.S. bond contract; January 3, 1995 through August 6, 1999 for the Deutsche Mark.

(2) N is the number of observations. All summary statistics are calculated from daily decimal (standardized) returns. E is 10 to the power of.

(3) AD, CVM, JB, KS and QQ are values of the normality test statistics of the Anderson–Darling, Crámer–Von Misses, Jarque–Bera, Kolmogorov–Smirnov and the QQ correlation coefficient between the sample and standard normal quantiles.

(4) The critical values derived under independence are: AD test: 1.929 (10%), 2.502 (5%) and 3.907 (1%). CVM test: 0.341 (10%), 0.458 (5%) and 0.744 (1%). JB test: 4.605 (10%), 5.992 (5%) and 9.210 (1%). KS test: 1.218 (10%), 1.350 (5%) and 1.628 (1%). QQ test: 0.998 (10%), 0.995 (5%) and 0.993 (1%).

(5) For the QQ test, reject the null hypothesis of normality if the table value is less than 0.995 at the 5% level, 0.993 at the 1% level.

[a] Significance at the 1% level.
[b] Significance at the 5% level.
[c] Significance at the 10% level.

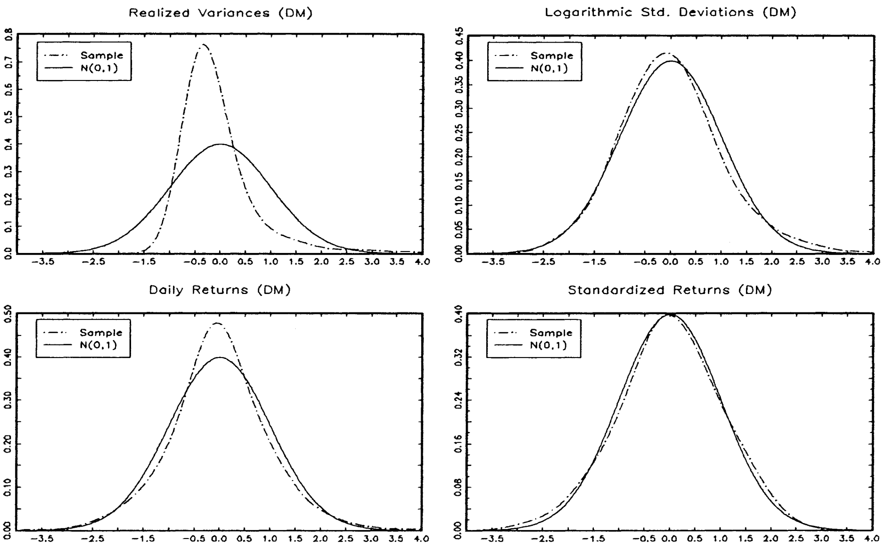

Fig. 5. Kernel density estimates for the Deutsche Mark futures.

Eurodollar) are skewed to the right while the other two (the S&P500 and T-bonds) are skewed to the left. The sample kurtosis coefficient for all daily returns is greater than 3, with T-bonds having the smallest estimate of about 4.9. The bottom left panel of Fig. 5 shows the kernel density estimate for the daily raw returns for the Deutsche Mark contract, along with a normal reference density.[19] The result is consistent with the existing literature that find that the distribution of the daily returns in the asset markets is not normal.

Also presented in Table 2 are the summary statistics for the standardized returns series. These are statistically unconditionally normally distributed for all the four futures contracts considered. The sample skewness for all four contracts is close to zero and the sample kurtosis is around 3. This is also evident from the bottom right panel of Fig. 5, which has the plot of the kernel density estimate for the (mean-zero and unit-variance) standardized returns for the Deutsche Mark contract, along with a normal reference density. The close approximation to the normal densities contrasts sharply to the leptokurtic distributions for the standardized daily returns that are typically obtained when relying on an estimate of the 1-day-ahead conditional variance from a parametric ARCH or stochastic volatility model (see, e.g., Bollerslev et al., 1994 for a general discussion). Instead, the summary statistics and figures for these futures returns are consistent with those for the standardized exchange rate returns and the standardized stock returns in ABDE (2001c) and ABDL (2001a), respectively.

In contrast to the above-mentioned papers, we formally test the normality hypothesis. We employ five statistics in testing whether standardized futures returns are normally distributed. These are the Anderson–Darling (AD), Crámer–Von Misses (CVM), Jarque–Bera

[19] All kernel density estimates are based on a Gaussian kernel with the bandwidth h_T determined by the standard formula $\hat{h}_T = 1.364 s T^{-1/5}$, where s is the sample standard deviation.

Table 3
Unconditional daily volatility distributions

	Volatilities (v_t^2)				log S.D. ($\log v_t$)			
	D-Mark	E-Dollar	S&P500	T-bonds	D-Mark	E-Dollar	S&P500	T-bonds
N	1157	1258	1261	1003	1157	1258	1261	1003
Mean	2.19e – 5	3.75e – 7	9.41e – 5	2.48e – 5	– 5.5133	– 7.6379	– 4.8539	– 5.4397
S.D.	2.41e – 5	5.77e – 7	1.49e – 4	2.83e – 5	0.3605	0.4570	0.4364	0.3419
Skewness	5.3815	8.1094	10.3267	8.1768	0.4415	0.4312	0.3411	0.6086
Kurtosis	48.1144	113.5945	172.2231	118.0011	3.8888	3.1138	3.4592	3.8306
AD	120.0047[a]	161.4947[a]	168.1104[a]	106.0506[a]	3.4085[b]	3.4944[b]	1.2202	3.3705[b]
CVM	22.3093[a]	30.6611[a]	31.6503[a]	19.6844[a]	0.5287[b]	0.4168[c]	0.1282	0.5202[b]
JB	103524[a]	654383[a]	1525808[a]	562196[a]	75.5421[a]	39.6350[a]	35.4981[a]	90.4839[a]
KS	7.1873[a]	9.7983[a]	9.8790[a]	7.3172[a]	1.4412[b]	1.6003[b]	0.7620	1.2772[c]
QQ	0.7464[a]	0.6735[a]	0.6313[a]	0.7020[a]	0.9930[c]	0.9921[a]	0.9958[c]	0.9901[a]

See Table 2 for the sample span and description of statistics.
[a] Significance at the 1% level.
[b] Significance at the 5% level.
[c] Significance at the 10% level.

(JB), Kolmogorov–Smirnov (KS) and the QQ correlation coefficient based on the correlation coefficient between the theoretical quantiles of the standard normal distribution and the order statistics of the sample.[20] Since the standardized returns were found to be serially uncorrelated, the usual critical values can be used when applying these tests. Based on the JB test in Table 2, the normality hypothesis is rejected for the standardized returns for all four futures contracts. The JB test statistics range from 5.0164 for the Deutsche Mark to 10.3125 for the T-bonds futures. None of the AD or CVM test statistics can reject the null hypothesis of normality for any of the futures standardized returns. The KS test cannot reject the normality hypothesis for the standardized returns for the Deutsche Mark, the S&P500 and the T-bonds contracts, but reject the hypothesis for the Eurodollar contract at the 1% level with the test statistic being equal to 1.8729. The QQ statistic rejects the null hypothesis for the T-bonds futures at the 10% level but not for the other three futures.[21] We emphasize, however, as all statistics will eventually reject the null hypothesis as the sample size increases, it is reasonable to emphasize the critical values at the 1% level instead of the 5% level. Therefore, based on the results from virtually all of the tests, the standardized futures returns studied here are Gaussian to a very good approximation. These results are also quite consistent with the descriptive statistics and density plots discussed above.

4.2. Volatilities and logarithmic standard deviations

Table 3 provides the same set of summary statistics for the unconditional distribution of the realized daily volatilities and the realized logarithmic standard deviations. The realized volatilities, the sum of the 5-min return squared, have means of 2.2e – 5, 3.8e – 7,

[20] Details on the five test statistics are given in Appendix A.
[21] The QQ plots corresponding to the returns and standardized returns for the Deutsche Mark contract are given in the bottom panel of Fig. 5. However, the QQ plot is a descriptive device and it is not a test for statistical inference. See also the discussion at the end of Appendix B.

9.4e − 5 and 2.5e − 5, respectively, for the Deutsche Mark, the Eurodollar, the S&P500 and the T-bonds futures contracts, implying an annualized standard deviation for the four futures contracts of approximately 7.4%, 1%, 15.4% and 7.9%, respectively. The standard deviations given in the third row also indicate that the realized daily volatilities fluctuate across different futures markets. Furthermore, the distributions of the realized volatilities are extremely skewed to the right and have much fatter tails than the normal distribution, as indicated in rows four and five of Table 3.[22]

Also shown in Table 3 are the distributional characteristics for the realized logarithmic standard deviations of the four futures contracts. The values of the sample kurtosis coefficient across all four contracts are reduced to about 3. In this case, the assumption of normality is clearly more appropriate. This point is also illustrated by the kernel density plot in the upper right panel of Fig. 5 for the Deutsche Mark futures contract, which graphs the unconditional sample distribution for the logarithmic standard deviation along with the standard normal density.

We employ the same five statistics in testing the normality hypothesis for the realized logarithmic standard deviations, as we did for the returns and standardized returns. As previously discussed, the logarithmic standard deviations exhibit strong temporal correlation in the form of long memory. It is possible that the presence of long memory can have adverse effects on the performance of the normality test statistics. Following the suggestion of the referee, we performed a simulation study (see the discussion in Appendix B) to examine the extent of these effects using sample sizes similar to those of our futures series. Our results (summarized in Table B.1) indicate that for the type of persistence exhibited by the logarithmic standard deviations, the use of the JB test is inappropriate as the test is oversized. When a series exhibits long memory and the sample size is between $T=1000$ and 1500, the JB statistic over-rejects the (true) null hypothesis while the other statistics do not. In particular, the JB test rejects more than 40% of the time the null hypothesis of Gaussianity when the nominal size used in the simulation is 5%. However, the other four normality statistics are largely unaffected by persistence in the data. In particular, we find that all of them are actually undersized when the fractional order is set to $d=0.3$ but they are about correctly sized when the fractional order is $d=0.4$ and $T=1200$ or 1500; for $d=0.4$ and $T=1000$, the tests are only mildly oversized, with an empirical rejection frequency of about 10%. Therefore, as the degree of fractional integration increases, a larger sample size is required in order for the tests to be correctly sized. The Crámer–Von Misses (CVM) test statistic has the best overall performance for $d=0.4$, in the sense of being less oversized than the other statistics for $T=1000$ (with an empirical rejection frequency of 8.5%) and being correctly sized for $T=1200$ and 1500 (with empirical rejection frequencies of 5.6% and 4.7%, respectively). The Kolmogorov–Smirnov (KS) test statistic appears to be the second best, while the Anderson–Darling (AD) test statistic can be ranked as third best. The QQ correlation coefficient test statistic is still undersized. Table B.2 gives the 95% and 99% empirical quantiles of all five test statistics, which can serve as rough guides in examining

[22] We do not present the summary statistics for the realized standard deviations since the realized standard deviations exhibit the same properties as the realized variances. The square root transformation of the variances to standard deviations reduces both the skewness and kurtosis estimates significantly. However, the distributions of the realized standard deviations are still skewed to the right and are significantly leptokurtic.

the changes on the tests' critical values as we move from independent observations to observations with long memory. For the JB test, these empirical quantiles are three to seven times larger than the conventional critical value from a chi-squared distribution with 2 *df*. For the other four tests, the empirical quantiles are quite close to those derived under the assumption of independent observations and, therefore, the latter critical values can be used in applications.

In summary, our simulation results suggest that when observations exhibit long memory, the use of the JB test statistic for assessing normality cannot be entertained as an appropriate choice. In contrast, all the other four test statistics appear to be largely unaffected by the presence of long memory and their use is thus recommended, at least when the sample sizes used for their construction are of the same (or larger) magnitude with the ones used in this study.

Not surprisingly, based on the statistics from the JB test in Table 3, the normality hypothesis is rejected for the logarithmic standard deviations for all four futures contracts. Neither the AD nor the CVM test statistics reject the normality hypothesis at the 1% level for the logarithmic standard deviations of the four futures contracts, but they do reject the normality hypothesis at the 5% level (CVM on Eurodollar is at the 10% level), except for the S&P500 futures where normality is not rejected. The results from the KS tests are similar to the results from both the AD and CVM test statistics. The null hypothesis of normality is not rejected at the 1% level for all the four futures contracts. Based on the QQ test, the normality hypothesis for the Deutsche Mark and the S&P500 cannot be rejected at the 1% level, but is rejected for the Eurodollar and T-bonds contracts (0.9930 is the 1% significance level for the QQ test when the sample size is T>300). However, since the critical values for the QQ test are really close to each other, it is difficult to be sure whether the test has rejected the null hypothesis or not. In practice, values of QQ over 0.99 are

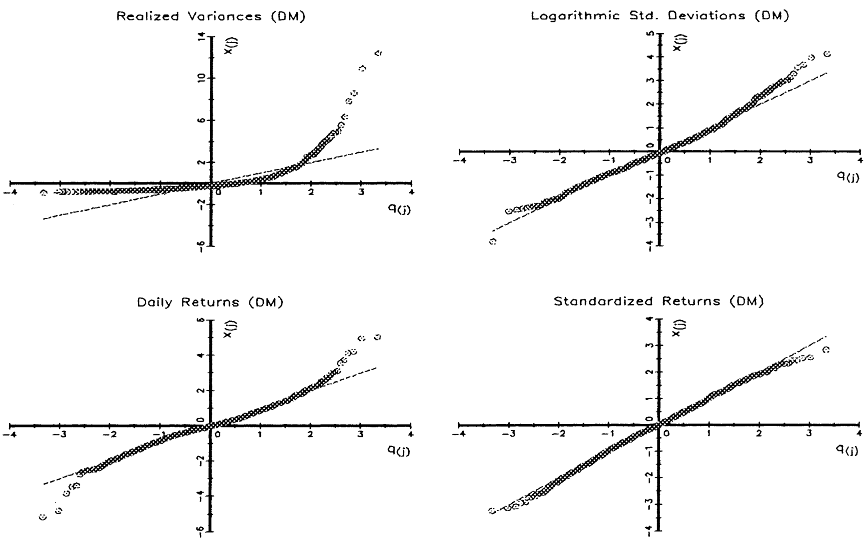

Fig. 6. QQ plots for the Deutsche Mark futures.

strong indications for normality. The QQ plot for the Deutsche Mark contract is presented in the upper panel of Fig. 6.[23]

The evidence on logarithmic standard deviations is consistent with Taylor (1986) and French et al. (1987), who find that the distribution of logarithmic monthly standard deviations constructed from the daily returns within the month is close to the Gaussian distribution. It is also consistent with ABDL (2001a) and ABDE (2001c), who find that realized daily foreign exchange rate volatilities and realized daily stock returns volatilities constructed from intraday high-frequency data are approximately log-normally distributed. Hence, the statistical evidence in Tables 1 and 2 imply that the unconditional distribution for the daily futures returns can be well approximated by a continuous lognormal–normal mixture, the Mixture-of-Distributions-Hypothesis proposed by Clark (1973).

4.3. Covariances and correlations

Table 4 has the summary statistics and normality tests for all pairs of daily realized covariances and correlations. These describe the relationships among one exchange rate, one equity market index, one long-term interest rate and one short-term interest rate.

All covariance pairs are heavily leptokurtic and either skewed to the right (Deutsche Mark, Eurodollar and T-bonds covariance pairs) or skewed to the left (S&P500 covariance pairs). The assumption of Gaussianity is strongly rejected for all six covariance pairs.

The six correlation pairs retain the same signs in their sample skewness coefficients as the covariance pairs. However, the sample kurtosis coefficients are now considerably smaller than the corresponding ones for the covariances. They vary from 2.43 to 3.88. The normality test statistics reject the hypothesis of Gaussianity for three correlation pairs, those between the Deutsche Mark and the Eurodollar, between the S&P500 and Eurodollar and between the S&P500 and the T-bonds, but can not reject normality for the other three correlation pairs. Fig. 7 presents the kernel density estimation for six correlation pairs, which conforms to the results in Table 4.

In summary, our results suggest that most of the realized correlations are approximately normally distributed; however, the evidence is not as strong as the evidence on realized logarithmic standard deviations.

5. Asymmetric volatility

Previous studies have documented the asymmetric relationship between equity volatility and returns: positive returns have a smaller impact on future volatility than have negative returns of the same absolute size. There are two popular explanations for the existence of asymmetric volatility. The leverage effect posits that a large negative return increases financial leverage, raising equity return volatility (Black, 1976; Christie, 1982). On the other hand, the volatility feedback or time-varying risk premia explanation argues that if the market risk premium is an increasing function of volatility, an anticipated

[23] See Appendix B for a discussion about deviations from the 45° line in a QQ plot in the presence of observations with long-memory.

Table 4
Unconditional daily covariance and correlation distributions covariances

	(DM, ED)	(DM, US)	(ED, US)	(SP, DM)	(SP, ED)	(SP, US)
Covariances						
Mean	3.33e − 8	9.61e − 10	1.43e − 6	− 5.50e − 6	2.34e − 7	2.88e − 6
S.D.	1.25e − 6	5.84e − 6	2.73e − 6	1.35e − 5	2.20e − 6	2.16e − 5
Skewness	17.901	0.7294	7.7048	− 7.0999	− 3.9545	− 2.1761
Kurtosis	486.28	29.253	84.545	96.1229	58.058	26.833
AD	169.337[a]	50.6356[a]	150.756[a]	134.521[a]	139.953[a]	54.419[a]
CVM	33.6012[a]	9.5156[a]	29.3353[a]	25.8687[a]	27.5334[a]	10.3144[a]
JB	11311578[a]	27250[a]	282658[a]	426298[a]	161531[a]	24359[a]
KS	8.7398[a]	4.8555[a]	9.1942[a]	7.8391[a]	8.7694[a]	5.2861[a]
QQ	0.5723[a]	0.8734[a]	0.6343[a]	0.7212[a]	0.7454[a]	0.8743[a]
Correlations						
Mean	0.0005	− 0.0136	0.4287	− 0.1309	0.1118	0.1830
S.D.	0.1673	0.1626	0.1922	0.1934	0.2743	0.3911
Skewness	0.4166	0.1325	0.0975	− 0.2721	− 0.6522	− 0.6717
Kurtosis	3.8824	2.6592	2.5568	3.1561	3.4225	2.4276
AD	4.3771[a]	0.7221	1.3846	1.5948	10.3781[a]	23.3647[a]
CVM	0.6975[b]	0.1022	0.2100	0.2378	1.7652[a]	4.0247[a]
JB	70.9401[a]	7.0965[b]	9.6203[a]	15.4029[a]	98.1569[a]	88.5005[a]
KS	1.7454[a]	0.8183	0.8652	1.0342	2.7748[a]	3.6639[a]
QQ	0.9924[a]	0.9973[c]	0.9976[c]	0.9971[c]	0.9858[a]	0.9658[a]

(1) DM stands for Deutsche Mark, ED stands for Eurodollar, SP for the S&P500 and US stands for the U.S. T-bonds.
(2) See Table 2 for the sample span and description of statistics.
[a] Significance at the 1% level.
[b] Significance at the 5% level.
[c] Significance at the 10% level.

increase in the volatility increases the risk premium, raising the required return on equity, leading to an immediate decline in equity price. Therefore, large negative returns increase future volatility more than positive returns due to a volatility feedback effect (Pindyck, 1984; French et al., 1987; Campbell and Hentschel, 1992; Bekaert and Wu, 2000). ABDE (2001c) examined the evidence on asymmetric volatility using realized volatilities on Dow Jones stocks. This section reexamines the relation between volatility and returns using the realized futures return volatility. While asymmetric volatility has been examined extensively for stock returns, it has not been examined closely for futures returns. Furthermore, since the relationships estimated here are for the futures volatility and returns, the leverage hypothesis can not be applied because there is no issue of financial leverage except possibly for the S&P500 futures.

5.1. Logarithmic standard deviations

We first look for asymmetric effects on the realized logarithmic standard deviations. In order to control for the strong serial correlation discussed in Section 3, it is important that the dynamic dependencies are properly modeled. Towards this end, we use the fractionally differenced logarithmic standard deviations as our dependent variable where the fractional

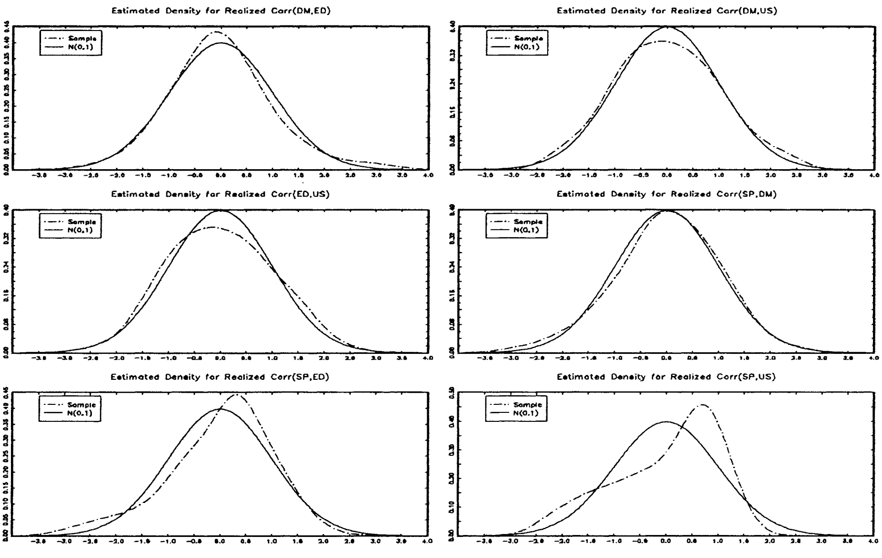

Fig. 7. Kernel density estimates for the correlations.

order d is fixed at the $\hat{d}_{\mathrm{SP}}$ estimates reported in Table 4. We then estimate the following equation using least squares:

$$(1-L)^{\hat{d}_{\mathrm{SP},j}} lv_{tj} = \alpha_j + \beta_j (1-L)^{\hat{d}_{\mathrm{SP},j}} lv_{t-1j} + \gamma_j r_{t-1,j} + \delta_j r_{t-1,j} I(r_{t-1,j} < 0) + u_{tj} \qquad (5)$$

where $I(\cdot)$ is the indicator function, j stands for one of the four futures contracts, lv_{tj} is the logarithmic realized standard deviation at time t, lv_{t-1j} is the lagged logarithmic standard deviation and $r_{t-1,j}$ is the lagged return. Our specification is consistent with the literature on the effects of lagged returns and is similar to that in Duffee (1995), Engle and Ng (1993) and Nelson (1991), but differs from that in ABDE (2001c).[24] The coefficient on the interacting term δ_j is the one that captures any asymmetries. If asymmetries are indeed present, we would expect δ_j to be negative and larger (in absolute magnitude) than the return coefficient γ_j.

Table 5 reports the regression estimates with their standard errors for all four contracts. The γ_j coefficient from the Deutsche Mark regression is positive and statistically significant while the coefficient from the T-bond regression is negative and statistically significant. The γ_j coefficients are not statistically significant for the Eurodollar and S&P500 regressions. As for the asymmetry coefficient δ_j, the estimates are equal to -7.2101, 1.4085, -12.7742 and 19.6546 for the Deutsche Mark, Euro Dollar, S&P500 and T-bond, respectively, and only that of S&P500 is statistically significant at the 1% level. The result on S&P500 futures contract is consistent with the asymmetric volatility literature, indicating asymmetry in the influence of past negative and positive stock index returns on current stock index volatility. Conversely, the results on the three other futures

[24] We thank the referee for suggesting to include the lagged filtered logarithmic standard deviations in the estimation as the filtered series still exhibit some serial correlation (see Section 6).

Table 5
News impact functions for log S.D.

	Coefficient estimates and S.E.			
	α_j	β_j	γ_j	δ_j
D-Mark	– 0.4575 (0.0304)***	0.0311 (0.0466)	6.1779 (3.7322)*	– 7.2101 (6.4429)
E-Dollar	– 0.0044 (0.0003)***	– 0.0429 (0.0359)	– 0.4825 (0.4648)	1.4085 (0.7229)*
S&P500	– 0.2284 (0.0157)***	– 0.1091 (0.0326)***	0.3258 (1.2440)	– 12.7742 (1.9891)***
T-Bonds	– 0.7343 (0.0575)***	0.1388 (0.0519)***	– 12.6609 (6.1446)**	19.6546 (10.6464)*

(1) Least squares coefficient estimates for the regression
$(1-\mathrm{L})^{d_j} lv_{tj} = \alpha_j + \beta_j (1-\mathrm{L})^{d_j} lv_{t-1,j} + \gamma_j r_{t-1,j} + \delta_j r_{t-1,j} I(r_{t-1,j} < 0)_{+utj}$
where $\hat{d}_{j\mathrm{SP}}$ is fixed at the estimate of the fractional order from Table 4, $I(\cdot)$ is the indicator function and $r_{t-1,j}$ is the lag of the standardized return of the jth contract. Standard errors are in the parentheses. Newey–West heteroscedasticity and autocorrelation consistent standard errors are in parentheses (using 20 lags).
(2) * Significance at the 10% level. ** Significance at the 5% level. *** Significance at the 1% level.

contracts do not support the existence of asymmetric volatility, i.e., negative shocks having a larger effect on volatility than positive shocks of the same magnitude.

Our results are consistent with that of ABDE (2001c) who found that the economic importance of asymmetric effects was marginal. Fig. 8 displays the scatterplots for the logarithmic standard deviations against the lagged standardized returns of the four futures contracts. The asymmetric effect is clearly not apparent except for the S&P500 series where the news impact curve is more steeply sloped to the left of the origin. Similar results have been reported in Tauchen et al. (1996).

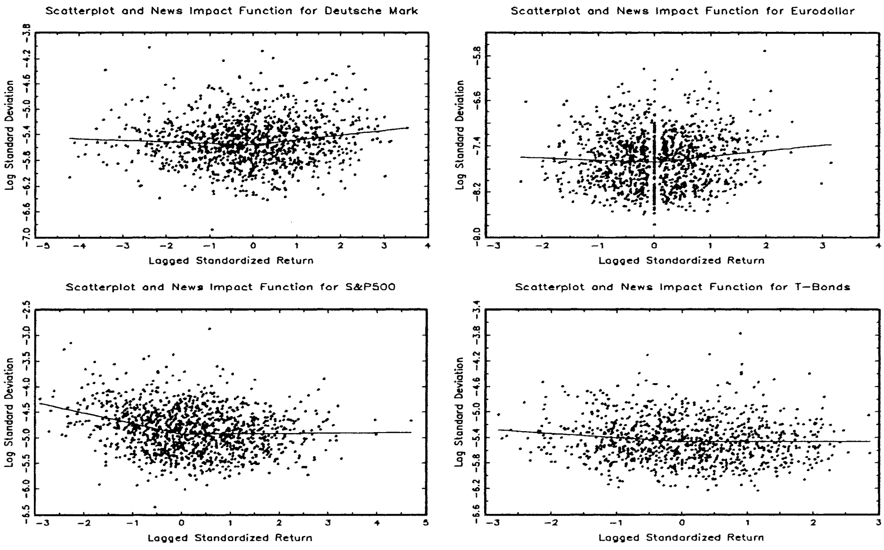

Fig. 8. News impact functions for daily logarithmic standard deviations.

5.2. Correlations

Previous literature has found statistically significant asymmetries in the covariances or correlations for returns on stock portfolios (Kroner and Ng, 1998; Ang and Chen, 2002), as well as Dow Jones stock returns (ABDE, 2001). In this section, we look for evidence on asymmetries on six futures correlations between the Deutsche Mark, the Eurodollar, the S&P500 and the T-bonds futures. The relevant equation used to look for evidence of asymmetric volatility in the realized correlations is given by:

$$(1-L)^{\hat{d}_{\mathrm{SP},ij}}\mathrm{corr}_{t,ij} = \alpha_{ij} + \theta_{ij}z_{t-1,ij} + \beta_{ij}w_{t-1,ij} + \gamma_{ij}w_{t-1,ij}I^1_{t-1,ij} + \delta_{ij}w_{t-1,ij}I^2_{t-1,ij} + u_{t,ij} \tag{6}$$

where $\hat{d}_{\mathrm{SP},ij}$ are the estimates of the fractional order from Table 1, $w_{t-1,ij} = r_{t-1,i} + r_{t-1,j}$, $z_{t-1,ij} = (1-L)^{d_i}lv_{t-1,i} + (1-L)^{d_j}lv_{t-1,j}$, and $I^1_{t-1,ij} = I(r_{t-1,i}\cdot r_{t-1,j} > 0)$ and $I_{t-1,ij}^{2} = I(r_{t-1,i}<0, r_{t-1,j}<0)$ are two indicator functions. $r_{t-1,i}$ and $r_{t-1,j}$ are the lagged returns. The estimation results are summarized in Table 6. Note that δ_{ij} gives us the measure on asymmetry with a negative δ_{ij} indicating asymmetric effects of lagged returns on correlations. While we do not adopt the specification in ABDE (2001c) who used the lagged logarithmic standard deviation as explanatory variable to test for asymmetry, our specification is in line with the literature and similar to their specification in spirit, and can

Table 6
News impact functions for correlations

Pair	Coefficient estimates and S.E.				
	α_{ij}	θ_{ij}	β_{ij}	γ_{ij}	δ_{ij}
DM, ED	0.0032	0.0047	0.0011	4.6269	− 3.1883
	(0.0078)	(0.0055)	(0.0048)	(1.8104)***	(2.4036)
DM, US	0.0104	0.0083	0.0550	4.3698	− 2.6171
	(0.0131)	(0.0061)	(0.0039)***	(1.4008)***	(2.0452)
ED, US	0.0347	− 0.0240	− 7.4764	10.1606	− 4.9530
	(0.0135)***	(0.0057)***	(2.9247)***	(3.6449)***	(2.9436)*
SP, DM	− 0.0063	0.0021	1.9003	− 1.9410	1.8413
	(0.0057)	(0.0057)	(0.6064)***	(0.6063)***	(0.6982)***
SP, ED	− 0.0217	− 0.0269	2.6281	− 0.7466	− 5.0499
	(0.0064)***	(0.0062)***	(0.5754)***	(1.0746)	(1.5188)***
SP, US	− 0.0438	− 0.0232	3.4122	− 1.3111	− 5.1173
	(0.0103)***	(0.0061)***	(0.7402)***	(1.0518)	(1.2231)***

(1) Least squares coefficient estimates for the regression. Standard errors are in the parentheses.
$(1-\mathrm{L})^{\hat{d}_{ij\mathrm{SP}}}\mathrm{corr}_{t,ij} = \alpha_{ij} + \theta_{ij}z_{t-1,ij} + \beta_{\mathrm{ij}}\mathrm{w}_{t-1,ij} + \gamma_{ij}w_{t-1,ij}I^1_{t-1,ij} + \delta_{ij}w_t - 1,ijI^2_{t-1,ij} + u_{t,ij}$
where $w_{t-1,ij} = \mathrm{r}_{t-1,i} + r_{t-1,j}$, $z_{t-1i,j} = (1-L)^{d_i}lv_{t-1,i} + (1-L)^{d_j}lv_{t-1,j}$,
and where $I^1_{t-1,ij} = I(r_{t-1,i}\cdot r_{t-1,j} > 0)$ and $I^2_{t-1,ij} = I(r_{t-1,i} < 0, r_{t-1,j} < 0)$ are two indicator functions.
(2) Newey–West heteroscedasticity and autocorrelation consistent standard errors in parentheses (using 1 lag).
(3) Significance at the 5% level.
* Significance at the 10% level.
** Significance at the 5% level.
*** Significance at the 1% level.

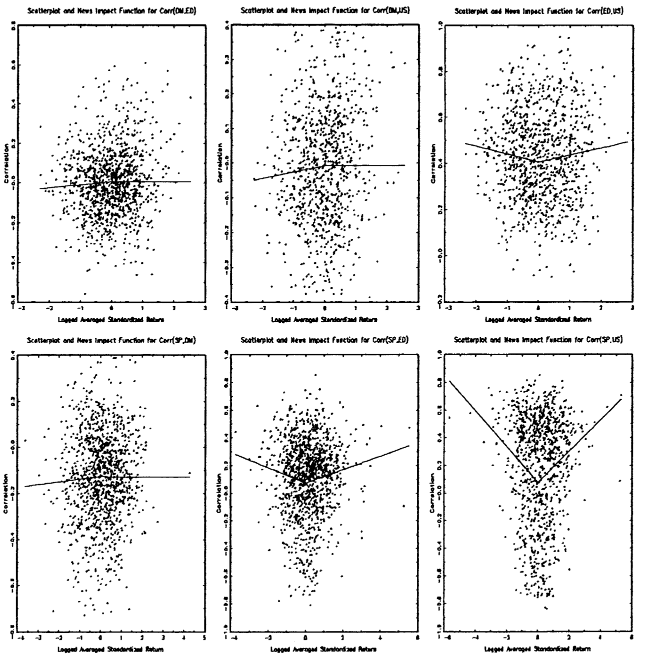

Fig. 9. News impact functions for daily correlations.

clearly shed some light on whether negative lagged returns have asymmetric effects on realized correlations.

The last column in Table 6 indicates that estimates on δ_{ij} are all negative except for the pair between the Deutsche Mark and the S&P500. The estimates for δ_{ij} are statistically significant at the 1% level for the correlations pairs related with S&P500, but not for the other correlation pairs. Fig. 9 plots the daily realized correlations against the (pairwise) average lagged standardized returns for the six correlation pairs. As with the realized logarithmic standard deviations, the systematic influence of the asymmetric effect is not very apparent.

6. Realized volatility modeling

In this section, we present a limited analysis and relevant results from ARFIMA modeling of the realized volatilities and the logarithmic standard deviations.[25] The main objective of this analysis is to examine, in some detail, the temporal correlation that remains in the series after accounting for long-range dependence. This is of interest if, for example, one wants to perform out-of-sample forecasting using the realized volatility series. However, as our purpose in this paper is not to examine forecasting performance of different models for futures volatility, we simply present in-sample estimation results. Nevertheless, our model is similar to that in ABDL (2001b), which uses a multivariate

[25] See Appendices A and B for details about ARFIMA models.

Table 7
ARFIMA models for realized volatility and logarithmic standard deviations

	Realized volatility	
	Coefficient estimates	$L \times 10^{-3}$
D-Mark	$(1-L)^{0.314} v_t^2 = \hat{\epsilon}_t + 0.156\hat{\epsilon}_{t-1} - 0.158\hat{\epsilon}_{t-5}$	10.833
E-Dollar	$(1-L)^{0.322} v_t^2 = \hat{\epsilon}_t + 0.252\hat{\epsilon}_{t-1}$	16.365
S&P500	$(1-L)^{0.302} v_t^2 = \hat{\epsilon}_t - 0.109\hat{\epsilon}_{t-5}$	9.521
T-bonds	$(1-L)^{0.118} v_t^2 = \hat{\epsilon}_t - 0.220\hat{\epsilon}_{t-19}$	9.118
	Logarithmic standard deviations	
	Coefficient estimates	$L \times 10^{-3}$
D-Mark	$(1-L)^{0.275} lv_t = 0.782 lv_{t-5} + \hat{\epsilon}_t + 0.691\hat{\epsilon}_{t-5}$	− 0.167
E-Dollar	$(1-L)^{0.301} lv_t = 0.7902 lv_{t-5} + \hat{\epsilon}_t + 0.692\hat{\epsilon}_{t-5}$	− 0.384
S&P500	$(1-L)^{0.444} lv_t = 1.272 lv_{t-1} + 1.118 lv_{t-5} + \hat{\epsilon}_t + 1.285\hat{\epsilon}_{t-1} + 1.119\hat{\epsilon}_{t-5}$	− 0.159
T-Bond	$(1-L)^{0.503} lv_t = 0.297 lv_{t-1} + 0.152 lv_{t-5} + \hat{\epsilon}_t + 0.663\hat{\epsilon}_{t-1} + 0.126\hat{\epsilon}_{t-5}$	− 0.432

(1) ϵ_t is assumed iid $(0,\sigma^2)$. (2) Estimation by maximum likelihood in the frequency domain. (3) All estimated coefficients significant at the 1% level. (4) L is the value of the maximized log-likelihood function. (5) Fractional order coefficient jointly estimated with ARMA coefficients.

VAR to model and forecast realized exchange rate volatility. Using a new dataset on futures contracts, we present a univariate ARFIMA modeling of the realized futures volatility. Thus, our results complement the results in ABDL (2001b).

Using the fractional order estimates $\hat{d}_{\mathrm{SP}}$ of Table 1, we first fractionally differenced the realized volatilities and logarithmic standard deviation series to obtain $\hat{u}_t = (1-L)^{\hat{d}_{\mathrm{SP}}} y_{tj}$, with $y_{tj} = v_{tj}^2$, or $y_{tj} = lv_{tj}$. Then, we used traditional identification techniques to determine an appropriate ARMA(p,q) model for the demeaned series. Finally, we reestimated (using maximum likelihood (ML) in the frequency domain) all parameters of the ARFIMA model simultaneously, that is, the autoregressive and moving average parameters as well as the fractional order parameter d. Our estimation results are summarized in Table 7.[26]

In general, the ARFIMA models for the realized volatilities are simpler and more parsimonious than their counterparts for the logarithmic standard deviations. The ARFIMA estimates for the fractional order are, in most cases, in close agreement with the Gaussian semiparametric estimates of Section 4.

For the S&P500 series and the Deutsche Mark series, simple ARFIMA (0, d, 5) models are adequate for the realized volatility series. For the S&P500 series, the first four MA coefficients were insignificant and were, thus, restricted to zero; for the Deutsche Mark series, the second, third and fourth MA coefficients were insignificant and were also restricted to zero. The equations reported in Table 7 are for this final model.[27] Note that

[26] The results from the use of ML for estimating the ARFIMA models are very robust with the underlying distributional properties of the volatility series. As is well known (see, for example, Beran, 1994; Brockwell and Davis, 1991), Gaussian ML estimation is appropriate (under certain conditions) for Gaussian and non-Gaussian data alike. We alternatively estimated our ARFIMA models by conditional Least Squares in the time domain and the results where practically identical to the ML estimation results.

[27] This MA(5) model provides better fit than a simple MA(1) model in terms of the value of the maximized log-likelihood and the Akaike order selection criterion. A similar comment applies to all models, which were cross-checked with simpler formulations.

both series have a significant coefficient at the weekly lag and, thus, the model exhibits some cyclical behavior.[28] The estimates of the fractional order from this model were $\hat{d}=0.302$ for the S&P500 series and $\hat{d}=0.341$ for the Deutsche Mark series (compared to the $\hat{d}_{\mathrm{SP}}=0.314$ and $\hat{d}_{\mathrm{SP}}=0.313$ estimates obtained using the Gaussian semiparametric estimator). For the Eurodollar series, the best fit was provided from a simple ARFIMA(0, d, 1) model. The estimate of the fractional order was higher than the corresponding estimate using the Gaussian semiparametric estimator ($\hat{d}=0.322$ versus $\hat{d}_{\mathrm{SP}}=0.237$). Finally, for the T-bonds fractionally differenced series, the correlogram and partial correlogram suggested that an ARFIMA(0, d, 19) was appropriate with the first eighteen coefficients being restricted to zero. Again, the model exhibits cyclical behavior, now at the monthly rather than at the weekly lag. The estimate of the fractional order here was $\hat{d}=0.118$ compared to $\hat{d}_{\mathrm{SP}}=0.147$ obtained from the Gaussian semiparametric estimator.

The results for the logarithmic standard deviation series are somewhat different. Here, an ARFIMA(5, d, 5) was appropriate (with some restrictions) for all four series. The correlogram and partial correlogram indicated strong cyclical behavior at multiples of the weekly lag (at lags 5, 10, 15, 20, etc.). After some experimentation, we found out that the ARFIMA(5, d, 5) with the first four AR and MA terms restricted to zero (for the Deutsche Mark and the Eurodollar series) or the second, third and fourth AR and MA terms restricted to zero (for the S&P500 and T-bonds series) produced sensible results and retained most of the cyclical behavior, as indicated in the correlograms. For three out of the four series, the estimates of the fractional order were smaller compared to their Gaussian semiparametric estimator counterparts. For the S&P500 series, the estimate from the ARFIMA model was $\hat{d}=0.444$ compared to $\hat{d}_{\mathrm{SP}}=0.456$; for the Deutsche Mark series, the estimate from the ARFIMA model was $\hat{d}=0.275$ compared to $\hat{d}_{\mathrm{SP}}=0.351$; for the Eurodollar series, the estimate from the ARFIMA model was $\hat{d}=0.301$ compared to $\hat{d}_{\mathrm{SP}}=0.395$. Finally, for the T-bonds series, the ARFIMA estimate of d was $\hat{d}=0.503$ compared to $\hat{d}_{\mathrm{SP}}=0.260$. However, we view this last estimate with some caution: its standard error produces a rather large 95% confidence interval [0.084,0.922] that contains many values in the nonstationary region ($d>0.5$).[29]

All ARFIMA modes examined in this section indicate that even after removing long-range dependences, the series still have significant cyclical correlation concentrated mainly at the weekly lag. This correlation is stronger for the logarithmic standard deviations than it is for the realized variance series, as the ARFIMA models for the former series include autoregressive terms while the ARFIMA models for the latter series do not. These results are interesting as they suggest predictability of futures volatility. A useful extension of the current work is to compare the forecasting performance of different time-series models applied to our futures series.

---

[28] As suggested by the referee, we also investigated whether the cyclicality mentioned could have been an artifact of the day-of-week effect. We removed the day-of-week effect by a simple regression on a constant and four daily dummies and used the residuals from this regression to identify the ARFIMA models. However, the day-of-week effect was so small (small coefficients and R^2) that the same models as those reported in the text were obtained.

[29] All other 95% confidence intervals of the fractional order estimates, whether from the ARFIMA models or the Gaussian semiparametric estimator, had upper bounds strictly less than 0.5. Thus, there is strong evidence for long memory in our futures series.

7. Conclusion

In this paper, we constructed and examined the statistical properties of daily realized volatilities for four of the most heavily traded futures contracts using intraday returns over a 5-year period. We focus on the temporal characteristics and the distributional properties of daily returns, return volatilities and (log of) standard deviations, standardized returns and pairwise correlations. We also explored the behavior of a number of tests for Gaussianity under long memory via a simulation study. The simulation results indicate that tests of the "goodness-of-fit" variety are appropriate to use while the commonly employed Jarque–Bera test is severely oversized and its use is not recommended. It was found that the standard deviations and the pairwise correlations exhibit long memory while the standardized returns are serially uncorrelated. It was also found that the (unconditional) distributions of daily returns' volatility are leptokurtic and highly skewed to the right while the distributions of the standardized returns, of the standard deviations and of the pairwise correlations are statistically indistinguishable from the Gaussian distribution. Our analysis suggests that there is only limited evidence of asymmetric volatility for our futures return series.

Our overall findings are consistent with existing work on realized volatility such as in ABDE (2001c) and in ABDL (2001a,b), and as such, they extend the current literature on model-free estimates of realized volatility and correlation using high-frequency, intraday futures data. Two interesting extensions of the current work, in the context of our futures data, are the examination of the forecasting performance of models based on realized volatility (complementing and extending the work of Bollerslev and Wright, 2001; ABDL, 2001b) and the effects of news announcements on the temporal and distributional characteristics of the volatility and correlation series. Both of these extensions are currently pursued by the authors.

Acknowledgements

We would like to thank Geert Bekaert, Russell Chuderewicz and an anonymous referee for comments and suggestions that improved the quality of the paper. Financial assistance from the Grants-in-Aid Program of the College of Arts and Sciences at Florida International University is gratefully acknowledged. All remaining errors are ours.

Appendix A. Methodology and test statistics

Let $\{x_t(\omega){:}\Omega \times \mathbb{N}_+ \rightarrow \mathbb{R}\}$ be a discrete-time, real-valued time series. As usual, Ω denotes the sample space, $\mathbb{N}_+=\{1, 2, \ldots\}$ denotes the field of natural numbers, and $\mathbb{R}$ denotes the field of real numbers. Based on a sample realization $\{x_t\}_t^T = 1$, we are interested in testing the distributional and temporal correlation properties of the time series. This appendix contains a brief description of the test statistics used in the main body of the paper and the simulations. Throughout the appendix, we use $\bar{x} = T^{-1} \sum_{t=1}^{T} x_t$ to denote the sample mean, and $\hat{\sigma} = [T^{-1} \sum_{t=1}^{T} (x_t - \bar{x})^2]^{1/2}$ to denote the sample standard deviation of the realization.

The hypothesized underlying distribution that we test for in the paper is the standard normal. There is a variety of statistics available for assessing the assumption of normality. The normality statistics we employ are standard. We note, however, that all of them are, strictly speaking, valid only under the assumption that the observations are independent and identically distributed. This fact compromises inferences made using these statistics since most of our examined time series exhibit a high degree of temporal correlation. To address this issue, we have performed the simulations reported in Appendix B.

The simplest statistics employed for assessing normality are the sample measures of skewness and excess kurtosis; their theoretical values for the standard normal distribution are both zero. Let $\hat{\mathcal{S}}_T$ be the estimator of the sample skewness and $\hat{\mathcal{K}}_T$ be the estimator of sample kurtosis. We have that:

$$\hat{\mathcal{S}}_T = \frac{1}{T\hat{\sigma}^3} \sum_{t=1}^{T} (x_t - \bar{x})^3, \quad \hat{\mathcal{K}}_T = \frac{1}{T\hat{\sigma}^4} \sum_{t=1}^{T} (x_t - \bar{x})^4 \tag{7}$$

Under the assumption of independent and normally distributed observations, we have that $\mathrm{E}(\hat{\mathcal{S}}_T)=0$ and $\mathrm{Var}(\hat{\mathcal{S}}_T)=6/T$ for the sample skewness, and that $\mathrm{E}(\hat{\mathcal{K}}_T)=3$ and $\mathrm{Var}(\hat{\mathcal{K}}_T)=24/T$. We can, thus, test for zero skewness and zero excess kurtosis separately using z-type tests that are $T \to \infty$, normally distributed. It can be shown that these z-tests are independent and usually they are combined into a single statistic (see, for example, Bera and Jarque, 1982) that is then used as a test for normality. This statistic is given by:

$$\mathcal{JB} = T\left[\frac{\hat{\mathcal{S}}_T^2}{6} + \frac{(\hat{\mathcal{K}}_T - 3)^2}{24}\right] \sim \chi^2_{(2)}, \; T \to \infty \tag{8}$$

The $\mathcal{JB}$ statistic has been know to frequently reject normality, especially in the presence of serially correlated observations, and to have a weak record of finite sample performance in general.

A potentially more powerful approach, even in the presence of correlated observations, would be to use a variety of other statistics that do not depend on sample moments. In the simulations, we reported results from four such statistics: the quantile–quantile correlation coefficient, the Kolmogorov–Smirnov statistic, the Anderson–Darling statistic and the Crámer–Von Misses statistic. We describe each one in turn.

The quantile–quantile correlation coefficient is the correlation coefficient between the theoretical quantiles of the standard normal distribution and the order statistics of the sample (see, for example, Johnson and Wichern, 1998, p. 189). Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq z_{(\mathrm{T})}$ denote the sample order statistics, and let $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(\mathrm{T})}$ denote the sample order statistics from the standardized observations, $z_{(j)}=(x_{(j)} - \bar{x})/\hat{\sigma}$. Assuming that the sample values are all distinct, there are exactly j observations less than or equal to $z_{(\mathrm{j})}$. Approximate the sample proportion j/T by $\hat{q}_{(j)}=(j-0.5)/T$ and let $q(j)$ denote the corresponding quantile of the standard normal distribution; $q(j)$ is then obtained by the relationship:

$$\int_{-\infty}^{q(j)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z = \hat{q}_{(j)} \tag{9}$$

A normality statistic is then based on the sample correlation coefficient between the standardized sample quantiles $z_{(j)}$ and the theoretical quantiles $q_{(j)}$:

$$\mathcal{QQ} = \frac{\sum_{j=1}^{T}(z_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\hat{\sigma}\hat{\sigma}_Q} \tag{10}$$

where $\bar{q}$ and $\hat{\sigma}_Q$ are the mean and standard deviation of the normal quantiles. For a sample as large as the one used in the paper (T>1000), the 1% critical value for the correlation coefficient should be over 0.993, under independence. Formally, if $\mathcal{QQ} < 0.993$, the assumption of normality is to be rejected. However, values of $\mathcal{QQ}$ over 0.99 are strong indications for normality.

Another powerful test for normality is the Kolmogorov–Smirnov test. This is the well-known nonparametric procedure that involves a comparison between the empirical distribution function and the corresponding theoretical distribution function, which in our analysis is the standard normal. Let $\mathcal{W}=\{w_1, w_2, \ldots, w_N\}$, $N \geq T$, denote an appropriate grid of ordered values, where possibly $w_N > z_{(T)}$ and $w_1 < z_{(1)}$.[30] Let $\hat{F}_T(w)$ denote the empirical distribution function defined as $\hat{F}_T(w) = N^{-1}\Sigma_{j=1}^{N} I(w_1,w)$, where $I(w_1,w)$ is the indicator function for the interval $[w_1,w]$. Let $G(w)$ denote the theoretical distribution function. The Kolmogorov–Smirnov statistic is given by:

$$\mathcal{KS} = \sqrt{T}\sup_{w \in \mathscr{W}} \mid \hat{F}_T(w) - G(w) \mid \tag{11}$$

The test rejects the null hypothesis H_0: $x_t \sim G(\cdot)$, $t=1, 2, ..., T$, if the observed value of the statistic exceeds the appropriate critical value, i.e., if $\mathcal{KS} > c_\alpha$. Under independence, the 5% and 1% critical values are given by $c_{0.05} = 1.350$ and $c_{0.01} = 1.628$, respectively.

The Anderson–Darling statistic and the Crámer–Von Misses statistic are similar in nature to the Kolmogorov–Smirnov statistic. They are also based on comparing the sample order statistics with the theoretical distribution being tested. Consider the sample order statistics for the standardized deviates $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(T)}$ and let $\Phi(\cdot)$ to denote the standard normal cumulative density function and $\tau_t = t$ to denote the observation index. The Anderson–Darling statistic is then defined as:

$$\mathcal{AD} = -T - \frac{1}{T}\sum_{t=1}^{T}(2\tau_t - 1)\{\ln[\Phi(z_{(t)})] + \ln[1 - \Phi(z_{(T-t+1)})]\} \tag{12}$$

Under independence, the 5% critical value is given by $c_{0.05} = 2.502$.

The Crámer–Von Misses statistic is defined next as:

$$\mathcal{CVM} = \frac{1}{12T} + \sum_{t=1}^{T}\left\{\Phi(z_{(t)}) - \frac{2\tau_t - 1}{2T}\right\}^2 \tag{13}$$

Under independence, the 5% critical value is given by $c_{0.05} = 0.458$.

[30] This grid includes all the sample standardized values but is finer.

For assessing the serial correlation of our series, we have used standard portmanteau tests. Let $\hat{\rho}(j)$ be the jth sample autocorrelation for the series x_t. The Ljung–Box portmanteau test for autocorrelation of order k is given by:

$$Q_h = T(T+2)\sum_{j=1}^{k}\frac{\hat{\rho}(j)}{T-j} \sim \chi^2_{(k)}, \quad T \to \infty \tag{14}$$

Substituting in the above expression the sample autocorrelations of the square of the series, x_t^2, we obtain the McLeod–Li Q_h^2 statistic. In addition, a simple plot of the sample autocorrelation function was enough to reveal the patterns of correlation for the series under study.

One of the testable hypothesis of this paper is that the time series under study are well approximated by long-memory processes. The stochastic process x_t is said to exhibit long memory if its autocorrelation function $\rho(j)$ behaves like $\rho(j) \sim Ch^{2d-1}$ as $j \to \infty$, for some $C \neq 0$ and $d \in (-0.5,0.5)$. Long-memory processes are stationary for values of $d \in (-0.5,0.5)$; if d exceeds 0.5, then the process is nonstationary. The usual range of interest for d is, however, $d \in (0,0.5)$. In the main body of the paper, we consider one class of long-memory processes, fractionally integrated ARMA processes, ARFIMA for short. We also use the ARFIMA class in modeling the realized volatility series.

For the ARFIMA class of long-memory processes, it can be shown that if x_t is such a process it can always be written as:

$$x_t = \psi(L)u_t \equiv \sum_{j=0}^{\infty}\psi_j u_{t-j} \tag{15}$$

where u_t is some stationary and invertible ARMÁ(p,q) process, say $\phi(L)u_t = \theta(L)\epsilon_t$, with $\epsilon_t \sim \text{iid}(0,\sigma^2)$ and where

$$\psi_j = \prod_{0<k\leq j}\frac{k-1+d}{k}, \quad j = 0,\ 1,\ 2,\ \ldots \tag{16}$$

The filter $\psi(L) = \Sigma_{j=0}^{\infty}\psi_j L^j$ defines the fractional difference operator as $[\psi(L)]^{-1} = (1-L)^d$.

Alternatively, using the above representation, an ARFIMA process has a spectral density $f(\lambda)$:

$$f(\lambda) \sim \lambda^{-2d}\frac{\sigma^2}{2\pi}\left[\frac{\theta(1)}{\phi(1)}\right]^2, \quad \lambda \to \infty \tag{17}$$

In estimating the order of fractional integration d, we used the Gaussian semiparametric estimator given in Robinson (1995b). This estimator is obtained as:

$$\hat{d} = \text{argmin}_{d\in(-0.5,0.5)}R(d) \equiv \log\left\{\frac{1}{M}\sum_{j-1}^{M}P_u(\lambda_j)\lambda_j^{2d}\right\} - 2d\frac{1}{M}\sum_{j=1}^{M}\log\lambda_j \tag{18}$$

where P_u is the periodogram of the filtered series $u=(1-L)^d x_t$, and M indicates the number of Fourier frequencies, $\lambda_j = 2\pi j/T$, to consider when minimizing $R(d)$, that satisfies $M \leq T/2$.

The asymptotic standard error for the estimator $\hat{d}$ is simply $1/(2\sqrt{M})$. Finally, when we estimated all parameters of an ARFIMA model jointly (i.e., the AR and MA parameters and the fractional order d), we used maximum likelihood in the frequency domain.

Appendix B. Simulations on the Performance of Normality Tests

In this section, we present results from a simulation study about the properties of the normality tests used in the paper.[31] In particular, we are interested in the properties of these tests when the data generating process (DGP) is Gaussian but exhibits long memory and the sample size is fairly large. These are cases of practical significance for the relevant literature, where the shape of estimated densities "looks" Gaussian and the samples commonly employed are in excess of $T = 1000$. We present results for the empirical size (empirical rejection frequencies) at the nominal 5% level of significance, as well as on the 95% and 99% empirical quantiles obtained through simulations; complete results about the quantiles are available upon request. Note that the critical values used to obtain the empirical rejection frequencies are those implied by independent observations. The simulation results indicate that these critical values work quite well for all normality tests except for the one based on sample skewness and sample kurtosis (the Jarque–Bera test). While our simulations are limited, our results compliment those of Beran and Gosch (1991) about the behavior of goodness of fit tests under long memory (see also Beran, 1994, Chap. 10). These authors have shown, both theoretically and using simulations, that the behavior of normality test statistics can be adversely affected by the presence of long memory in the observations, especially when a distribution is tested under a simple hypothesis (specified mean and standard deviation). However, they have also shown that these test statistics have good size properties when the distribution is tested under a composite hypothesis (estimated mean and standard deviation).[32] Our results are useful because (a) we explore the behavior of the test statistics for the case where large samples are used, (b) we explore the behavior of the Jarque–Bera test and Crámer–Von Misses tests (which were not examined by Beran and Gosch, 1991) and show that the Jarque–Bera test over-rejects normality and (c) provide additional support for the empirical results of our paper.

We only consider one type of DGP, fractional Gaussian ARIMA (ARFIMA) models.[33] Some details about the DGP can be found in Appendix A. The ARFIMA models were simulated using the algorithm of Davies and Harte (1987), implemented in GAUSS, along with the program's random number generator. The algorithm was adapted to the autocovariance function of the ARFIMA model. Results were obtained for three different sample sizes $T = \{1000, 1200, 1500\}$, that roughly correspond to the sample sizes of the futures series used in the paper. In all simulations, the standard deviation of the Gaussian

[31] We would like to thank the referee for suggesting the simulation analysis.

[32] In particular, for the case of simple hypothesis, the tests' size converges to 1, while for the case of a composite hypothesis, the tests' size converges to some $\alpha < 1$.

[33] We also tried another type of DGP, fractional Gaussian noise, the results are similar and available upon request.

deviates was set to $\sigma = 1$; our results are robust to different values of σ. Since for most of our future series, the estimates of the fractional order were similar in magnitude, the fractional order in the simulations was set to $d=\{0.3,0.4\}$.The coefficients of the ARFIMA models were kept to a minimum of two, with one AR and one MA coefficient both set to 0.5. Again, in view of our empirical results on ARFIMA model, this is an acceptable and empirically relevant compromise. Experimenting with higher AR and MA orders did not significantly affect the simulation results. In all experiments, we performed 10,000 replications.

The results in Table B.1 are very informative. When the fractional order is $d = 0.3$, all tests except the Jarque–Bera test are undersized, with empirical rejection frequencies between 0.3% and 3.1%. The Jarque–Bera test is grossly oversized with an empirical rejection frequency above 37%. Note that the empirical size of the Jarque–Bera test increases as the sample size increases. As the fractional order increases from $d = 0.3$ to 0.4, all tests become mildly oversized at $T = 1000$, with empirical rejection frequencies ranging from 8.5% to 11.3%. The Jarque–Bera test is again grossly oversized with an empirical rejection frequency above 50%. However, for $T = 1200$ and 1500, the other tests become correctly sized with empirical rejection frequencies being between 4.7% and 7.2%, save for the QQ correlation coefficient statistic which remains undersized. It appears that the test statistic with the best overall performance is the Crámer–Von Misses (CVM), with the Kolmogorov–Smirnov (KS) test being second best and the Anderson–Darling test being third best.

While the simulations conducted are limited in range of fractional order exponents, it is clear from the results that the use of the Jarque–Bera test for assessing normality in large financial samples with long memory will lead to over-rejection, even when normality is indeed present. The use of other tests of normality is recommended in that instance, all of which are either correctly sized or only mildly over/undersized at the sample sizes examined. Our results about the 95% and 99% empirical quantiles (in Table B.2) indicate that the use of conventional critical values, derived under the assumption of independent observations, is appropriate for the AD, CVM and KS test statistics, with a preference given to the CVM test.

Table B.1
Empirical sizes of normality tests for ARFIMA (1, d, 1)

Test statistic	$d = 0.3$			$d = 0.4$		
	$T = 1000$	$T = 1200$	$T = 1500$	$T = 1000$	$T = 1200$	$T = 1500$
AD	0.0170	0.0180	0.0310	0.1130	0.0720	0.0630
CVM	0.0130	0.0140	0.0180	0.0850	0.0560	0.0470
JB	0.3730	0.4230	0.4130	0.5330	0.4810	0.4710
KS	0.0190	0.0240	0.0280	0.1100	0.0640	0.0520
QQ	0.0180	0.0120	0.0030	0.0940	0.0340	0.0200

(1) The nominal size of the simulations is 5%. Rejection is based on 5% critical values under independence.
(2) AD, CVM, JB, KS and QQ stand for the Anderson–Darling, Crámer–Von Misses, Jarque–Bera, Kolmogorov–Smirnov and the QQ correlation coefficient test statistics.
(3) The parameters of the ARFIMA (1, d, 1)model are: RA(1) coefficient is 0.5, MA(1) coefficient is 0.5. The standard deviation of the error term is 1. The number of replications is 10,000.

Table B.2
Empirical quantiles of normality tests for ARFIMA (1, d, 1)

	$d=0.3$			$d=0.4$		
	$T=1000$	$T=1200$	$T=1500$	$T=1000$	$T=1200$	$T=1500$
Empirical 95% quantiles						
AD	1.9233	2.0061	2.1606	3.2336	2.8837	2.6811
CVM	0.3193	0.3301	0.3363	0.5549	0.4810	0.4441
JB	19.0386	19.8660	22.9837	33.4778	25.9609	29.3804
KS	1.2252	1.2253	1.2177	1.5485	1.4004	1.3603
QQ	0.9995	0.9995	0.9996	0.9994	0.9995	0.9996
Empirical 99% quantiles						
AD	2.7842	3.0525	3.0517	5.3939	4.4361	4.1899
CVM	0.4657	0.4982	0.4932	0.8926	0.7757	0.6963
JB	40.4986	38.3574	39.7461	67.7439	56.5086	73.9772
KS	1.4257	1.4402	1.4707	1.8732	1.8617	1.6005
QQ	0.9993	0.9994	0.9995	0.9991	0.9993	0.9995

The parameters of the ARFIMA (1, d, 1) model are the same as in Table B.1.

An interesting point, raised by the referee, is the tail behavior of the observations used to compute and graph a normal QQ plot. For example, in Fig. 6, there appears to be a significant tail deviation from the 45° line, for the logarithmic standard deviations (even though the normality tests accept the hypothesis of normality at the 1% level). Can it be the case that these large deviations are due to the presence of long memory, even if the underlying observations are Gaussian? We explored this issue using the same simulation methodology as above. Let $\zeta_i = \max_{1 \leq t \leq T} |z_{(t),i} - q_{(t)}|$ denote the maximum absolute deviation between a sample order statistic and the corresponding quantile from the standard normal distribution in the ith simulation run. We examined the sampling distribution of ζ_i over $i=1, 2, \ldots, R$, where R is the number of simulations. In each simulation run, we computed two ζ_i values, one obtained from iid standard normal data, denoted ζ_i^{IID}, and one obtained from data using the ARFIMA model above, denoted ζ_i^{LM}.

Table B.3 has the means and standard deviations for the two ζ_i series, computed over $R=1000$ replications[34], and the same three sample sizes as above $T=\{1000,1200,1500\}$. It is clear that the sampling distribution of the ζ_i^{LM} series is skewed to the right, compared with the sampling distribution of the ζ_i^{IID} series, as the mean of ζ_i^{LM} is always larger and statistically significant from the mean of ζ_i^{IID}. It, therefore, appears that the presence of long memory in the original observations will lead to more and larger tail deviations in the QQ plot even when the observations are normally distributed. Note that this is much more pronounced for the smaller sample of $T=1000$. Notice that even for a sample with iid observations, deviations from the 45° line are to be expected as well.

To further establish the differences induced by long memory in the ζ_i series, we used the well known result (see, for example, Leadbetter et al., 1983) that the distribution of extreme value statistics belong to one (and only one) of three families of distributions. For

[34] Note that the number of replications becomes the sample size for examining the distribution of ζ_i. Since the replications are formed by independent draws, our sample size is $R=1000$ observations of iid data. Results using a larger R were practically identical.

Table B.3
Means, S.D. and shape exponent for ζ_i series

	$d=0.3$			$d=0.4$		
	$T=1000$	$T=1200$	$T=1500$	$T=1000$	$T=1200$	$T=1500$
Mean of ζ_i^{LM}	0.8645	0.4897	0.4805	0.6001	0.5475	0.5137
Mean of ζ_i^{IID}	0.4338	0.4287	0.4276	0.4275	0.4386	0.4262
S.D. of ζ_i^{LM}	0.3884	0.2062	0.2147	0.3019	0.1954	0.2048
S.D. of ζ_i^{IID}	0.1936	0.1874	0.1855	0.1904	0.1982	0.1818
$\hat{a}$ of ζ_i^{LM}	1.4934	0.8509	0.8297	0.9756	0.9712	0.8999
$\hat{a}$ of ζ_i^{IID}	0.7715	0.7654	0.7622	0.7569	0.7675	0.7612

(1) $\zeta_i=\max_{1\leq t\leq T}|z_{(t),i}-q_{(t)}|$ denotes the maximum absolute deviation between a sample order statistic and the corresponding quantile from the standard normal distribution in the ith replication run. ζ_i^{LM} stands for deviations obtained from series with long memory and ζ_i^{IID} stands for deviations obtained from iid standard normal series.
(2) Mean is the sample mean, S.D. is the sample standard deviation of the ζ_i series computed over $R=1000$ replications.
(3) $\hat{a}$ is the ML estimate of the shape exponent of the type II extreme value distribution for the ζ_i series. The log-likelihood function is $\mathscr{L}(a)=R\ln a-(-a+1)\,\Sigma_{i=1}^{R}\ln\zeta_i-\Sigma_{i=1}^{R}\zeta_i^{-a}$.

the ζ_i series, we have that $\zeta_i>0$ for all i and, therefore, is appropriate to consider the type II distribution with cdf given by:

$$\Pr(\zeta\leq x)\equiv G(x;a)=\exp(-x^{-a}) \tag{19}$$

where α is the shape parameter of the distribution. We estimated the shape exponent by maximum likelihood for all the ζ_i series that we generated with the above simulations. The results, given in the last row of Table B.3. The ζ_i^{LM} series have higher shape exponent estimates than the ζ_i^{IID} series, and this shifts the cdf up on the vertical axis. Therefore, there is always a higher chance for a larger tail deviation in a normal QQ plot when the underlying data exhibit long memory. Consider, for example, the point on the horizontal axis, where $x=\mu$, the mean of the distribution. Take the corresponding estimates for the mean from Table B.3 and consider the case where $d=0.4$ and $T=1200$.We have that $G(0.5475;0.9712)=0.1661$ for the long-memory series, and $G(0.4386;0.7675)=0.1522$ for the iid series, a 9.13% difference. Similar differences can be found for $d=0.3$ and $T=1000$ and 1500.

In summary, the pronounced tail deviation from the 45° line on the QQ plot is, in large part, due to the presence of long memory in the sample. However, even in iid samples, some deviations are also expected. Therefore, in assessing normality, one needs to examine the results from the normality tests and the QQ plot jointly.

References

Andersen, T.G., Bollerslev, T., 1998. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. International Economic Review 39, 885–905.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001a. The distribution of exchange rate volatility. Journal of the American Statistical Association 96, 42–55.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001b. Modeling and forecasting realized volatility. NBER Working Paper #8160, March.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H., 2001c. The distribution of stock return volatility. Journal of Financial Economics 61, 43–76.

Ang, A., Chen, J., 2000. Asymmetric correlations of equity portfolios. Manuscript, Graduate School of Business, Columbia University.

Barndorff-Niesen, O.E., Shephard, N., 2001a. Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. Journal of the Royal Statistical Society, Series B 63, 167–241.

Barndorff-Niesen, O.E., Shephard, N., 2001b. How accurate is the asymptotic approximation to the distribution of realized volatility? Mimeo.

Bekaert, G., Wu, G., 2000. Asymmetric volatility and risk in equity markets. Review of Financial Studies 13, 1–42.

Bera, A., Jarque, C., 1982. Model specification tests: a simultaneous approach. Journal of Econometrics 20, 59–82.

Beran, J., 1994. Statistics for Long Memory Processes. CRC Press, Boca Raton, Florida.

Beran, J., Gosch, S., 1991. Slowly decaying correlations, testing normality, nuisance parameters. Journal of the American Statistical Association 86, 785–791.

Black, F., 1976. Studies of stock market volatility changes. Proceedings of the American Statistical Association, Business and Economic Statistics Section, 177–181.

Bollerslev, T., Wright, J.H., 2001. Volatility forecasting, high-frequency data, and frequency domain inference. Review of Economics and Statistics 83, 596–602.

Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. In: Engle, R.F., McFadden, D.Handbook of Econometrics, vol. IV. North-Holland, Amsterdam, pp. 2959–3038.

Breidt, F.J., Crato, N., de Lima, P., 1998. On the detection and estimation of long-memory in stochastic volatility. Journal of Econometrics 83, 325–348.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, 2nd edn., Springer Verlag, New York.

Campbell, J.Y., Hentschel, L., 1992. No news is good news: an asymmetric model of changing volatility in stock returns. Journal of Financial Economics 31, 281–318.

Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton Univ. Press, Princeton, New Jersey.

Campbell, J.Y., Lettau, M., Malkiel, B.G., Xu, Y., 2001. Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. Journal of Finance 56, 1–43.

Christie, A.A., 1982. The stochastic behavior of common stock variances: value, leverage and interest rate effects. Journal of Financial Economics 10, 407–432.

Clark, P.K., 1973. A subordinated stochastic process model with finite variance for speculative prices. Econometrica 41, 135–155.

Davies, R.B., Harte, D.S., 1987. Tests for hurst effect. Biometrika 74, 95–101.

Duffee, G.R., 1995. Stock returns and volatility: a firm level analysis. Journal of Financial Economics 37, 399–420.

Ederington, L., Lee, J.H., 1993. How markets process information: news releases and volatility. Journal of Finance, 1160–1191. (September).

Engle, R.F., Ng, V.K., 1993. Measuring and testing the impact of news on volatility. Journal of Finance 48, 1749–1778.

French, K.R., Schwert, G.W., Stambaugh, R.F., 1987. Expected stock returns and volatility. Journal of Financial Economics 19, 3–29.

Geweke, J., Porter-Hudauk, S., 1993. The estimation and application of long memory time series models. Journal of Time Series Analysis 4, 221–238.

Ghysels, E., Harvey, A., Renault, E., 1996. Stochastic volatility. In: Maddala, G.S. (Ed.), Handbook of Statistics, vol. 14. North-Holland, Amsterdam, 119–191.

Hsieh, D.A., 1991. Chaos and nonlinear dynamics: application to financial markets. Journal of Finance 46, 1839–1877.

Johnson, R.A., Wichern, D.W., 1998. Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River, New Jersey.

Kim, D., Kon, S.J., 1994. Alternative models for the conditional heteroscedasticity of stock returns. Journal of Business 67, 563–598.

Kroner, K.F., Ng, V.K., 1998. Modeling asymmetric comovements of asset returns. Review of Financial Studies 11, 817–844.

Lamoureux, C.G., Lastrapes, W.D., 1993. Forecasting stock returns variances: towards understanding stochastic implied volatility. Review of Financial Studies 6, 293–326.

Leadbetter, M.R., Lindgren, G., Rootzen, H., 1983. Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag, New York.

Merton, R.C., 1980. On estimating the expected return on the market: an exploratory investigation. Journal of Financial Economics 8, 323–361.

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

Pindyck, R.S., 1984. Risk, inflation, and the stock market. American Economic Review 74, 334–351.

Poterba, J., Summers, L., 1986. The persistence of volatility and stock market fluctuations. American Economic Review 76, 1142–1151.

Robinson, P.M., 1995a. Log-periodogram regression of time series with long-range dependence. Annals of Statistics 23, 1048–1072.

Robinson, P.M., 1995b. Gaussian semiparametric estimation of long range dependence. Annals of Statistics 23, 1630–1661.

Robinson, P.M., Zaffaroni, P., 1998. Nonlinear time series with long memory: a model for stochastic volatility. Journal of Statistical Planning and Inference 68, 359–371.

Schwert, G.W., 1990. Stock market volatility. Financial Analyst Journal 46, 23–34.

Tauchen, G.E., Zhang, H., Liu, M., 1996. Volume, volatility and leverage: a dynamic analysis. Journal of Econometrics 74, 177–208.

Taylor, S.J., 1986. Modeling Financial Time Series. Wiley, Chichester.