JOURNAL OF APPLIED ECONOMETRICS
*J. Appl. Econ.* **13**, 217–244 (1998)

# STYLIZED FACTS OF DAILY RETURN SERIES AND THE HIDDEN MARKOV MODEL

TOBIAS RYDÉN,[a] TIMO TERÄSVIRTA[b]* AND STEFAN ÅSBRINK[c]

[a]*Department of Mathematical Statistics, Lund University, Box 118, S-221 00 Lund, Sweden*
[b]*Department of Economic Statistics, Stockholm School of Economics, Box 6501, S-113 83 Stockholm, Sweden*
[c]*Trygg Hansa, Equities, S-106 26 Stockholm, Sweden*

## SUMMARY

In two recent papers, Granger and Ding (1995a,b) considered long return series that are first differences of logarithmed price series or price indices. They established a set of temporal and distributional properties for such series and suggested that the returns are well characterized by the double exponential distribution. The present paper shows that a mixture of normal variables with zero mean can generate series with most of the properties Granger and Ding singled out. In that case, the temporal higher-order dependence observed in return series may be described by a hidden Markov model. Such a model is estimated for ten subseries of the well-known S&P 500 return series of about 17,000 daily observations. It reproduces the stylized facts of Granger and Ding quite well, but the parameter estimates of the model sometimes vary considerably from one subseries to the next. The implications of these results are discussed. © 1998 John Wiley & Sons, Ltd.

## 1. INTRODUCTION

In two recent papers, Granger and Ding (1995a,b) considered long return series which are first differences of logarithmed price series. They established a few properties or stylized facts which seemed to hold for a large number of such series. The series were either daily observations or single transactions data. They also suggested a stochastic model that would generate series with such properties. Some of the properties were temporal, some other distributional. The temporal properties are as follows:

TP1: Returns $r_t$ are not autocorrelated (except possibly at lag one).

TP2: The autocorrelation functions of $|r_t|$ and $r_t^2$ decay slowly starting from the first autocorrelation, and $\mathrm{corr}(|r_t|, |r_{t-k}|) > \mathrm{corr}(r_t^2, r_{t-k}^2)$. The decay is much slower than the exponential rate of a stationary AR(1) or ARMA(1, $q$) model. The autocorrelations remain positive for very long lags.

TP3: $\mathrm{corr}(|r_t|, |r_{t-k}|) > \mathrm{corr}(|r_t|^\theta, |r_{t-k}|^\theta)$, $\theta \neq 1$. Autocorrelations of powers of absolute return are highest at power one. This effect is called the Taylor effect.

TP4: The observed autocorrelations of $\mathrm{sign}(r_t)$ are insignificant.

* Correspondence to: T. Teräsvirta, Department of Economic Statistics, Stockholm School of Economics, Box 6501, S-113 83, Stockholm, Sweden. E-mail: sttimo@hhs.se

Contract grant sponsor: Swedish Natural Science Research Council; Contract grant number: M-AA/MA 10538-303.
Contract grant sponsor: Swedish Council for Research in the Humanities and Social Sciences.
Contract grant sponsor: Tore Browaldh Foundation for Scientific Research; Contract grant number: T96541.

CCC 0883–7252/98/030217–28$17·50

*Received 24 June 1996*
*Revised 15 May 1997*

The distributional properties are as follows:

DP1: $|r_t|$ and sign($r_t$) are independent.
DP2: $|r_t|$ has the same mean and standard deviation.
DP3: The marginal distribution for $|r_t|$ is exponential (after outlier reduction).

An exponentially distributed stochastic variable $x_t$ has the following properties:

PED1: $E(x_t) = \sqrt{\text{Var}(x_t)}$; see DP2.
PED2: $E[x_t - Ex_t]^3/(\text{Var}(x_t))^{3/2} = 2$.
PED3: $E[x_t - E(x_t)]^4/(\text{Var}(x_t))^2 = 9$.

Although TP1 holds, the return series contain higher-order dependence. Granger and Ding (1995b) (henceforth abbreviated GD) considered the following model for the returns: $r_t = e_t h_t$ where $\{e_t\}$ is a sequence of i.i.d. double exponential variables with mean zero and unit variance. Furthermore, $h_t$ is an ARCH-type term in that it is a function of $|e_{t-k}|$, $k = 1, \ldots, q$; see also Ding, Granger, and Engle (1993), Granger and Ding (1995a), and Ding and Granger (1996). According to GD, a part of the idea was to model the distinctive shape of the return distribution near its centre and not just concentrate on the tails. A somewhat related paper discussing unconditional distributions for asset returns is Mittnik and Rachev (1993). These authors studied the usefulness of various stable distributions in modelling returns which they assumed independent. Of the alternative distributions they fitted to five years of S&P 500 daily returns in 1982–6 the Weibull distribution gave the best fit.

In this paper we present an alternative to the distribution theory of GD and discuss its properties in the light of an empirical example. Our starting point is the assumption that the marginal distribution of returns is a mixture of normal distributions. Other distributions could be considered, but normality is a convenient assumption to start with. We show that such a mixture allows data-generating processes capable of closely reproducing most of the distributional properties GD observed in the absolute returns. There also exists a computationally feasible solution to introducing higher-order temporal dependence in the process if the marginal distribution is a mixture of normal distributions. It consists of postulating the dependence in terms of the hidden Markov model (HMM) or Markov Switching Regime model of Lindgren (1978). For applications of this model to financial time series; see, for example, Tyssedal and Tjøstheim (1988), Hamilton (1988), Pagan and Schwert (1990), and Sola and Timmermann (1994). To find out how this idea works we apply the HMM to ten equally long subsets of the daily S&P 500 US stock price series. This series consists of 17,055 observations dating from 3 January 1928 to 30 April 1991. It is one of the series considered by GD and one which they found to have most of the properties listed above. Our paper thus has a different focus from GD, who concentrated on establishing the existence of the temporal and distributional properties listed above for a large number of series. We shall consider a small number of series which are in fact subsets of a very long single series and model them with the HMM. A major part of the interest lies in the properties of the estimated models. It turns out that at least in our case, the HMM is a very promising idea as far as reproducing the stylized facts of GD is concerned.

The plan of the paper is as follows. Section 2 contains preliminary considerations and Section 3 highlights properties of the hidden Markov model. Section 4 discusses parameter estimation, testing linearity against the HMM, selecting the number of regimes and the evaluation of estimated models. Evaluation also includes checking how well the models reproduce stylized facts

 

observed in the data. Section 5 is devoted to an application to the S&P 500 US stock index. Finally, Section 6 presents conclusions.

## 2. PRELIMINARY CONSIDERATIONS

As GD remarked, it is not difficult to find a model possessing at least some of the above-mentioned properties. Elaborating their example, suppose $r_t = e_t h_t$ where $\{e_t\}$ is a sequence of i.i.d. variables with zero mean independent of $h_t$ so that TP1 holds. If the distribution of $e_t$ is symmetric about zero then TP4 is also true. Assume $h_t = a > 0$ for $t = 1, 2, \ldots, T/2$, and $h_t = k^2 a$, for $t = T/2 + 1, \ldots, T$ ($T$ observations). In that case, TP2 holds approximately as well for sufficiently large $T$. This artificial but simple example provides a starting-point for our investigation.

Let $r_t = e_t h_t$ where $e_t = \text{sign}(r_t)$ and $h_t = |r_t|$. Assume furthermore that $r_t$, $t = 1, \ldots, T$, are drawn independently from one of two normal distributions $N_1 = \text{N}(0, 1)$ and $N_2 = \text{N}(0, \sigma^2)$. The probability of drawing from $N_1$ equals $p$. By construction, TP1 holds and because the normals are assumed to have zero means, TP4 is not violated in practice. We postpone a discussion of TP2 and TP3 until later and consider instead the properties of $h_t$ and compare them with PED1–3. To do that we need the expectation and the second, third, and fourth central moments of $h_t$. If we let $X_1 \sim N_1$ and $X_2 \sim N_2$ we have

$$Eh_t^\theta = pE|X_1|^\theta + (1-p)E|X_2|^\theta \qquad \theta = 1, 2, 3, 4 \tag{1}$$

from which $Eh_t$ and the central moments $\mu_\theta = E(h_t - Eh_t)^\theta$, $\theta = 2, 3, 4$, can be computed using normality of $X_i$, $i = 1, 2$. Skewness ($\mu_3/\mu_2^{3/2}$) and kurtosis ($\mu_4/\mu_2^2$) are functions of these central moments.

Using the first two moments, the skewness, and the kurtosis, we can see how well PED1–3 can be satisfied with our mixture of normals. Figure 1 depicts the combinations of $p$ and $\sigma$ which yield PED1. The figure also contains two dashed lines; one for which the ratio of the mean to the standard deviation is 0·8 and another corresponding to the combinations of $p$ and $\sigma$ for which the ratio equals 1·25. Consider the case $\sigma > 1$. It is seen that for large values of $p$ ($0{\cdot}5 \leqslant p \leqslant 0{\cdot}9$), $\sigma$ remains practically unchanged when $p$ is changed, whereas for large values of $\sigma$ ($\sigma \geqslant 5$), $p$ changes rather little when $\sigma$ is changed. If the ratio of the mean to the standard deviation is allowed to vary between 0·8 and 1·25, say, the set of combinations of $p$ and $\sigma$ satisfying this interval requirement is large. The other curves for $\sigma < 1$ form an inverted mirror image of the ones just discussed.

Figure 2 shows the corresponding values for the combinations of $p$ and $\sigma$ for which the skewness requirement PED2 holds. The shape of the curve resembles that for $\sigma > 1$ in Figure 1. The main difference is that the standard deviation starts increasing from about $p = 0{\cdot}6$ downwards, whereas the corresponding value of $p$ was about 0·4 in Figure 1. The dashed lines again show (log) symmetric deviations from PED2: one corresponding to $\mu_3/\mu_2^{3/2} = 1{\cdot}6$ and the other to $\mu_3/\mu_2^{3/2} = 2{\cdot}4$. For the range of $p$ for which $\sigma \approx 2{\cdot}5$, small deviations from this value (holding $p$ constant) cause relatively large changes in the skewness. On the other hand, for the range of $\sigma$ for which $p \approx 0{\cdot}6$ even fairly large deviations from this value (holding $\sigma$ constant) have only a minor effect on the skewness. Because of the symmetry, the curves for $\sigma < 1$ are not shown separately. Finally, the combinations of $p$ and $\sigma$ satisfying PED3 can be found in Figure 3. The dashed lines depict isoquants which satisfy $\mu_4/\mu_2^2 = 7{\cdot}2$ and $\mu_4/\mu_2^2 = 10{\cdot}8$, respectively. The general

 

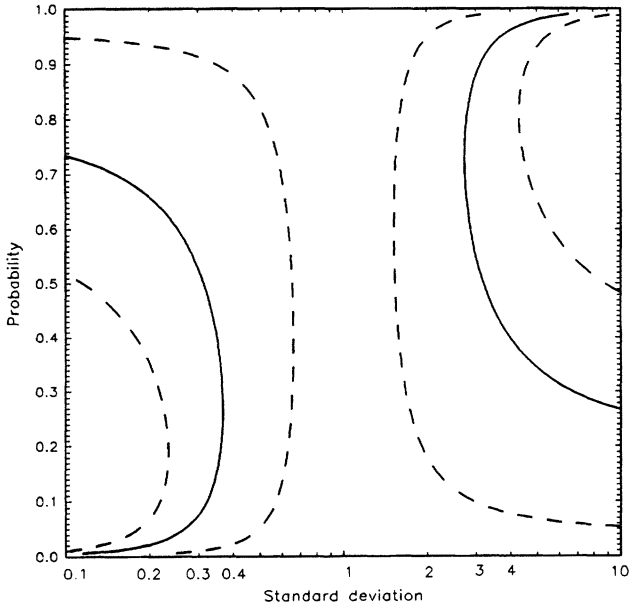


Figure 1. Combinations of $p$ and $\sigma$ which yield PED1: 'the standard deviation of the absolute returns equals the mean of the absolute returns' (solid line). Dashed lines indicate the combinations for which the mean/standard deviation ratio equals 0·8 and 1·25, respectively. The scale of the $\sigma$-axis is logarithmic

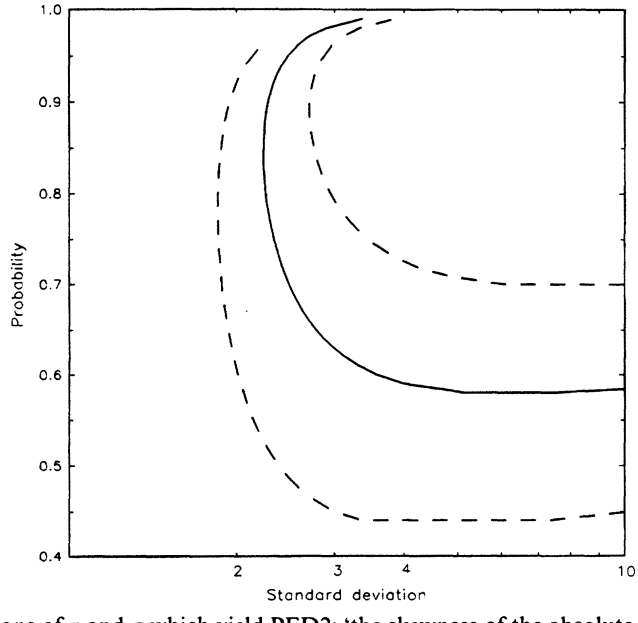


Figure 2. Combinations of $p$ and $\sigma$ which yield PED2: 'the skewness of the absolute returns = 2' (solid line) when $\sigma > 1$. Dashed lines indicate the combinations for which the skewness equals 1·6 and 2·5, respectively. The scale of the $\sigma$-axis is logarithmic

 

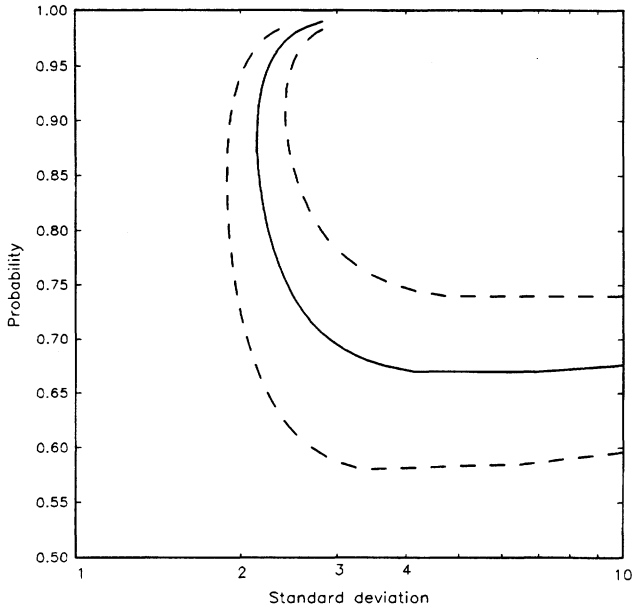


Figure 3. Combinations of $p$ and $\sigma$ which yield PED3: 'kurtosis of the absolute returns = 9' (solid line) when $\sigma > 1$. Dashed lines indicate the combinations for which the kurtosis equals 7·2 and 11·25, respectively. The scale of the $\sigma$-axis is logarithmic

shape of the curves is rather similar to those in Figure 2. A remarkable fact is that the steep descent in the isoquant in terms of $p$ occurs again for $\sigma \approx 2{\cdot}5$ as in Figure 2. On the other hand, for the kurtosis to equal nine at large values of $\sigma$, a larger value of $p$ ($0{\cdot}65 \leqslant p \leqslant 0{\cdot}7$) is now required.

In all, Figures 1–3 indicate that properties PED1–3 are not contradictory for a mixture of two normals. For $\sigma > 1$, with $\sigma$ near 2·5 and $p$ between 0·7 and 0·9 one comes close to satisfying these three stylized facts. It seems that PED1 has to be relaxed somewhat (the mean has to be allowed to be slightly larger than the standard deviation) if we at the same time want PED2 and PED3 to be satisfied. On the other hand, combinations with large $\sigma$ and small $p$ and vice versa are clearly unacceptable.

## 3. INTRODUCING DEPENDENCE

The above considerations rely on the normality of the components of the mixtures. If the idea of mixtures is to be applied to modelling returns then we have to introduce stochastic dependence between the observations. Of course, TP1 has to hold for $\{r_t\}$ and automatically does when $\{e_t\}$ is a sequence of i.i.d. variables with zero mean, but higher-order dependence has to be postulated. Thus we assume that there is dependence in $\{h_t\}$. Such dependence can be introduced by making the probability $p$ time-varying and conditional on its previous values. Following the suggestion of Lindgren (1978), define a random variable $S_t$ such that $\{S_t\}$ is a $d$-state first-order Markov chain. (We do not consider higher-order Markov chains here.) Thus $p_{ij} = P(S_t = j \mid S_{t-1} = i)$, $i$, $j = 1, \ldots, d$, are the transition probabilities for the process to move from state or regime $i$ at time $t-1$ to state $j$ at $t$. The discussion in the previous section would imply $d = 2$, but in the following

 

the number of states may exceed two. In the general case of $d$ states, the model for $r_t$ can be written as

$$r_t = \sum_{i=1}^{d} I(S_t = i)X_{it} \tag{2}$$

where $I(z)$ is the indicator function obtaining value unity if $z$ is true and zero otherwise. Furthermore, $X_{it}$, $i = 1, \ldots, d$, are $d$ independent normal variables with mean zero and variance $\sigma_i^2$. Model (2) in which $\{S_t\}$ obeys the transition probabilities $p_{ij}$, $i, j = 1, \ldots, d$, $\Sigma_{j=1}^{d} p_{ij} = 1$, $i = 1, \ldots, d$, is a special case of the hidden Markov model (HMM) or the Markov Switching Regime model of Lindgren (1978). Hamilton (1994, chapter 22) discussed many of the statistical properties of the HMM. We shall only remind the reader of those features of the model we shall require later on. Let $\mathbf{P} = \{p_{ij}\}$ be the $(d \times d)$ matrix of transition probabilities. Thus we have $\mathbf{P1} = \mathbf{1}$ where $\mathbf{1} = (1, 1, \ldots, 1)'$. This implies that unity is an eigenvalue of $\mathbf{P}$ and that $\mathbf{1}$ is the corresponding right eigenvector. The corresponding left eigenvector is $\mathbf{a}$: $\mathbf{aP} = \mathbf{a}$, normalized such that $\mathbf{a1} = 1$. Let $p_{ij}^n = P(S_t = j \mid S_{t-n} = i)$ be the $n$-step transition probability, $n \geqslant 1$, and $\mathbf{P}^{(n)} = \mathbf{P}^n = \{p_{ij}^n\}$ the corresponding transition matrix. If $\mathbf{P}$ is ergodic then $\mathbf{a} = (a_1, \ldots, a_d)$ is unique and consists of the $d$ unconditional probabilities $a_i$, $i = 1, \ldots, d$, of the process $\{S_t\}$ being in regime $i$ at any given time $t$. These probabilities are called stationary probabilities. Moreover,

$$\lim_{n \to \infty} \mathbf{P}^{(n)} = \mathbf{1a} \tag{3}$$

For example, for a two-state Markov chain

$$\mathbf{a}' = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} (1 - p_{22})/(2 - p_{11} - p_{22}) \\ (1 - p_{11})/(2 - p_{11} - p_{22}) \end{bmatrix} \tag{4}$$

Note that model (1) and thus the previous moment results hold for the two-state HMM when $p$ is replaced by $a_1$. To obtain expressions for covariances, let $g$ be a function of the random variable $r_t$ and assume that we want to calculate the covariance structure of $\{g(r_t)\}$. Furthermore, let $G_i = E\{g(r_t) \mid S_t = i\}$, i.e. the expected value of $g(r_t)$ given regime $i$. For instance, if $g(r_t) = r_t^2$ then $G_i$ equals the conditional second moment. Generally, the unconditional mean is

$$Eg(X_t) = \sum_{i=1}^{d} a_i G_i \tag{5}$$

where $d$, as before, is the number of states. Defining $\mathbf{G} = \text{diag}(G_1, \ldots, G_d)$, expression (5) can be written more compactly as

$$Eg(r_t) = \mathbf{aG1} \tag{6}$$

Furthermore,

$$E\{g(r_t)g(r_{t+n})\} = \sum_{i=1}^{d} \sum_{j=1}^{d} E\{g(r_t)g(r_{t+n}) \mid S_t = i, S_{t+n} = j\} \Pr(S_t = i, S_{t+n} = j)$$

 

where $n \geqslant 1$ and $\Pr(S_t = i, S_{t+n} = j) = \Pr(S_{t+n} = j \mid S_t = i)\Pr(S_t = i)$. But $\Pr(S_t = i) = a_i$ and $\Pr(S_{t+n} = j \mid S_t = i) = p_{ij}^n$. Also

$$E\{g(r_t)g(r_{t+n}) \mid S_t = i, S_{t+n} = j\} = E\{g(r_t) \mid S_t = i\}E\{g(r_{t+n}) \mid S_{t+n} = j\} = G_i G_j \tag{7}$$

so that

$$E\{g(r_t)g(r_{t+n})\} = \sum_{i=1}^{d}\sum_{j=1}^{d} G_i G_j a_i p_{ij}^n = \mathbf{aGP}^{(n)}\mathbf{G1}$$

Finally the covariance of $g(r_t)$ and $g(r_{t+n})$ is

$$\operatorname{cov}\{g(r_t), g(r_{t+n})\} = \mathbf{aGP}^{(n)}\mathbf{G1} - (\mathbf{aG1})^2 \tag{8}$$

for $n = 1, 2, \ldots$. The autocovariance (8) will be needed for estimating the autocorrelation function of $|r_t|$ for an HMM. For that purpose $g(r_t) = |r_t|$ in (8).

## 4. ESTIMATION, SPECIFICATION, AND EVALUATION

In this section we shall discuss a modelling cycle for HMMs consisting of specification, estimation, and evaluation of such models. Although this would be the natural ordering of stages in this cycle we consider estimation first. This is because parameter estimation is a necessary prerequisite for the specification technique we consider in this paper.

### 4.1. Estimation of Parameters

The most common approach to estimation of the parameters of an HMM is maximum likelihood which we too shall adopt. In this section we assume that the number of states, $d$, of the hidden Markov chain $\{S_t\}$ is fixed and known. In practice this is not the case, but we defer the problem of choosing $d$, i.e. the specification or the model selection problem, until the next section. Thus, the parameter vector $\theta$ to be estimated comprises the transition probabilities $p_{ij}$, $i, j = 1, \ldots, d$, and the (conditional) variances $\sigma_1^2, \ldots, \sigma_d^2$. Letting $f(\cdot; \sigma^2)$ denote the density of a normal variable with zero mean and variance $\sigma^2$, the likelihood of the model (i.e. the joint density of $r_1, \ldots, r_n$) can be written

$$L(\theta; r_1, \ldots, r_n) = \sum_{s_1=1}^{d} \cdots \sum_{s_n=1}^{d} a_{s_1} \prod_{t=2}^{n} p_{s_{t-1}, s_t} \prod_{t=1}^{n} f(r_t; \sigma_{s_t}^2) \tag{9}$$

where $a_j$, $j = 1, \ldots, d$, are the stationary probabilities for $\mathbf{P}$ as above and $n$ is the sample size. Evaluating the likelihood as it is expressed in equation (9), the total number of numerical operations increases exponentially as $n$ grows. To improve on this, introduce the diagonal matrix $\mathbf{F}(r; \theta) = \operatorname{diag}(f(r; \sigma_1^2), \ldots, f(r; \sigma_d^2))$ and note that $L$ may be written as

$$L(\theta; r_1, \ldots, r_n) = \mathbf{aF}(r_1; \theta)\mathbf{PF}(r_2; \theta)\mathbf{P} \ldots \mathbf{F}(r_{n-1}; \theta)\mathbf{PF}(r_n; \theta)\mathbf{1} \tag{10}$$

Here, the Markov property of $\{S_t\}$ is crucial. For any switching regime model which is not Markov, the computational complexity of evaluating $L$ grows exponentially with $n$. Evaluating $L$

 

this way, the computational complexity is only linear in $n$. Leroux (1992) showed that the maximum-likelihood estimate (MLE) of $\theta$ is consistent. Finding the MLE is not straightforward, though, since $L$, as a function of $\theta$, may have several local maxima in addition to the global maximum. Also note that one may always permute the numbering of the states without affecting the likelihood. We used the downhill-simplex algorithm (see, for example, Press *et al.*, 1989) to maximize $L$. In order to avoid reporting local maxima, the search routine was started at 200 randomly chosen points for $d = 2$ and at 500 randomly chosen points for $d = 3$.

In mixture models and HMMs, the likelihood may under some circumstances be unbounded as a function of the parameters. Consider, for example, the case of an HMM whose conditional 'output distributions' are normal with common mean but with different variances. We may then make the likelihood arbitrarily large by setting the mean equal to an arbitrary observation and letting one of the (conditional) variances tend to zero. The conditional output distributions of our HMMs have fixed means, however, whence the likelihood is bounded (unless at least one observation is exactly zero).

### 4.2. Testing Linearity and Selecting the Number of States

The HMM is a non-linear model with the property that it is not identified if the number of states in reality is less than postulated. Many non-linear time series models share a similar property (see, for example, Granger and Teräsvirta, 1993, chapter 6). For a general discussion of this identification problem, see Davies (1977). In the case of an unidentified HMM, some of the transition probabilities are nuisance parameters, and consistent estimation of the parameters is not possible. An important special case is the one in which one postulates an HMM with two states but the observations originate from just a single normal distribution. This means that the conditional and unconditional variances of $r_t$ are equal. Thus testing linearity (constant conditional variance) is at least as important in our case as it is, for instance, when the alternative to that hypothesis is an ARCH model. In fact it is even more important here because fitting an HMM to data generated by a linear model does not lead to consistent parameter estimates.

Likelihood ratio tests and penalized likelihood criteria are two standard procedures for making the 'best' choice among a sequence of nested classes of models, and both of these may be applied to HMMs. Let $H_d$ denote the class of HMMs for which the hidden Markov chain $\{S_t\}$ has $d$ states, and for which the conditional distribution of $r_t$ given $S_t = i$ is normal with zero mean. Then $H_d$ is characterized by $d(d-1)$ transition probabilities and $d$ conditional variances, $m_d = d^2$ parameters in all. Moreover, these classes are nested, i.e. $H_d \subseteq H_{d+1}$, in the sense that for each model in $H_d$, there exists a model in $H_{d+1}$ that induces the same distribution for $\{r_t\}$ (in fact there are infinitely many such models).

The use of the penalized likelihood criterion may be interpreted as a likelihood ratio test when two nested models are compared with each other. Let $\hat{\theta}_n^d$ be the maximum-likelihood estimator of $\theta$ over $H_d$. A penalized likelihood estimator selects the class $H_{\hat{d}}$, where $\hat{d}$ maximizes

$$\log L(\hat{\theta}_n^d; r_1, \ldots, r_n) - \omega_{n,d} \tag{11}$$

over $d = 1, 2, 3, \ldots$. Here $\omega_{n,d}$ is a penalty term, satisfying $\omega_{n,d} \geqslant 0$ and $\omega_{n,d} < \omega_{n,d+1}$ for each $n$, and thus preventing an overly large model from being selected. The two most common choices

 

for $\omega_{n,d}$ are $\omega_{n,d} = m_d$ (AIC, Akaike Information Criterion) and $\omega_{n,d} = (1/2)m_d \log n$ (BIC, Bayesian Information Criterion). Define the LR statistic as

$$LR_n^d = 2\{\log L(\hat{\theta}_n^{d+1}; r_1, \ldots, r_n) - \log L(\hat{\theta}_n^d; r_1, \ldots, r_n)\}$$

and, according to equation (11), we select $H_{d+1}$ if $LR_n^d > 2(\omega_{n,d+1} - \omega_{n,d})$. Thus the penalties define the critical value of the test. However, in the present case the corresponding size remains unknown because $LR_n^d$ does not have the customary asymptotic $\chi^2$-distribution with $m_{d+1} - m_d$ degrees of freedom when the null hypothesis 'the true parameter $\theta_0$ is in $H_d$' holds and $n \to \infty$. This is because of the lack of identifiability of the HMM under this null hypothesis when the alternative is that $\theta_0$ is in $H_{d+1}$.

Short of the limiting distribution, we approximated it by bootstrap techniques. We preferred a parametric approach, i.e. resampling from the distribution induced by the ML-estimate $\hat{\theta}_n^d$. McLachlan (1987) did the same to test the number of components in a normal mixture. Furthermore, we did not resample $n$ observations, but only $m$, where $m$ is a function of $n$ such that $m/n \to 0$ as $n \to \infty$. The reason for this is that in order for the 'full' bootstrap to work, some kind of smoothness (differentiability) condition on the mapping $\theta \to F_\theta^n$, where $F_\theta^n$ is the law of $LR_n^d$ under $\theta$, is needed; see Bickel and Freedman (1981, p. 1200) in the non-parametric setting. Such a condition is difficult to verify. By performing an '$m$-out-of-$n$' bootstrap, we can trade smoothness of this mapping for smoothness of the mapping $\theta \to P_\theta^n$, where $P_\theta^n$ is the law of $r_1, \ldots, r_n$ under $\theta$; cf. Bickel, Götze, and Van Zwet (1997), in particular the proof of their Theorem 3. In the context of HMMs, this idea is due to Peter Bickel (personal communication) and may be characterized as follows. For an HMM with a smooth parameterization, $\theta \to P_\theta^n$, such as ours, the proof of Lemma 3.14 in Bickel and Ritov (1996) shows that $d_H^2(P_\theta^m, P_{\theta^0}^m) = O(m \mid \theta - \theta^0 \mid^2)$. Hence, since $\| P_\theta^m - P_{\theta^0}^m \| \leqslant 2d_H(P_\theta^m, P_{\theta^0}^m)$ (Le Cam, 1986, p. 47),

$$\| P_\theta^m - P_{\theta^0}^m \| = O(\sqrt{m} \mid \theta - \theta^0 \mid)$$

where $d_H$ and $\| \cdot \|$ denote Hellinger distance and variational norm, respectively. If $\theta^0 \in H_d$ (that is, if the null hypothesis is true), then the ML-estimate $\hat{\theta}_n^d$ is $\sqrt{n}$-consistent, i.e. $n^\beta(\hat{\theta}_n^d - \theta^0) \to 0$ in $P_{\theta^0}$-probability for each $\beta < 1/2$ (see Bickel, Ritov, and Rydén, 1997). Note that $\hat{\theta}_n^d$ is computed from $n$ observations ($n > m$), so that the size of the bootstrap replicates is smaller than the full sample size.

Thus, if $m = n^\beta$ with $\beta < 1$,

$$\| P_{\hat{\theta}_n^d}^m - P_{\theta^0}^m \| \to 0$$

in $P_{\theta^0}$-probability. In particular,

$$\sup_{A \subseteq \mathbf{R}} \mid P_{\hat{\theta}_n^d}^m (LR_m^d \in A) - P_{\theta^0}^m (LR_m^d \in A) \mid \to 0$$

in $P_{\theta^0}$-probability, so that the $m$-out-of-$n$ parametric bootstrap consistently approximates the distribution of the $LR$-statistic. In practice, a number of bootstrap replicates, say $K$, of sample size $m$ are simulated from the distribution $P_{\hat{\theta}_n^d}$, and the corresponding $LR$-statistics $LR_{m,1}^d, \ldots, LR_{m,K}^d$ are computed. If $k$ of these statistics exceed the observed $LR$-statistic (computed from $m$ observations), then the $p$-value of the test is estimated by $(k+1)/(K+1)$. For the results reported below, we used $K = 50$.

### 4.3. Evaluation of the Model

Most time-series models, for example the autoregressive, ARCH, and GARCH models, are defined in terms of an i.i.d. sequence of innovations. After estimating the parameters of such a model, one may try to reconstruct these innovations by estimating them: this yields a sequence of residuals. The model may then be evaluated by checking whether or not the residuals are at least approximately i.i.d. with the prescribed marginal distribution. The HMMs under study may also be constructed in terms of an innovation sequence. This is done by letting $\{e_t\}$ be an i.i.d. sequence of standard normal random variables and defining $r_t = \sigma_{s_t} e_t$. It is then immediate that $\{r_t\}$ is an HMM of the form given above. Contrary to many other time-series models, however, $\{e_t\}$ cannot be reconstructed from $\{r_t\}$ even if the true parameter were known. This is true since the hidden Markov chain cannot be reconstructed exactly, only conditional probabilities of the type $P(S_t = i \mid r_1, \ldots, r_n)$ may be computed. Accordingly, HMMs must be evaluated by other means.

In this paper we consider a goodness-of-fit test of Milhøj (1981) that focuses on the spectral properties of a zero mean stationary process. If $\{x_t\}$ is such a process, then Milhøj's statistic is

$$W = \left[\frac{2\pi}{n}\sum_{t=1}^{n-1}\{I(\omega_t)/g(\hat{\theta}, \omega_t)\}^2\right] \Big/ \left[\frac{2\pi}{n}\sum_{t=1}^{n-1}\{I(\omega_t)/g(\hat{\theta}, \omega_t)\}\right]^2 \tag{12}$$

where $\omega_t = 2\pi t/n$, and $I(\omega)$ is the periodogram,

$$I(\omega) = \frac{2\pi}{n}\left|\sum_{t=0}^{n-1} x_t \, \mathrm{e}^{-\mathrm{i}\omega t}\right|^2$$

Furthermore, $g(\theta, \omega)$ is the spectral density of the process, and $\hat{\theta}$ is a parameter estimate. A large value of $W$ indicates that the model is not adequate. In order to apply the test statistic (12) we have to derive the spectral density of the HMM. This is done in the Appendix. The Milhøj test is applied to the series of absolute returns. Applying it to the original series is useless because the series themselves are uncorrelated (our HMMs generate series with exactly this property), and hence have constant spectra. To obtain series with zero mean we subtracted the sample mean of $|r_t|$ from each subseries.

Milhøj (1981) derived the asymptotic distribution of $W$ as $n \to \infty$ for processes that can be expanded in a Wold decomposition with i.i.d. innovations. For an HMM it is not known, however, whether or not such an expansion is valid, because Wold's theorem guarantees only the validity of an expansion based on uncorrelated innovations. For this reason, we approximate the distribution of $W$ by the parametric bootstrap as in the previous section. This time the size $m$ of the bootstrap replicates is set to $n$, as this test in general may have relatively low power. Saikkonen (1984) showed that the asymptotic relative efficiency (ARE) of Milhøj's test with respect to the usual portmanteau statistic for testing the hypothesis of no autocorrelation in the error sequence of an ARMA model equals zero. The test can therefore be expected to be less powerful than the time-domain portmanteau test even for finite sample sizes and global alternatives.

In the present situation the portmanteau test is not available because the innovations cannot be reconstructed. Nevertheless, the above ARE result may serve as a general indication suggesting that in many cases Milhøj's statistic may not be very powerful. Reducing the replicate sample size would thus decrease an already low power. Although a completely formal justification for this

 

test does not exist it may in any case be viewed as an indicator of poor fit. Also in this case, 50 bootstrap replicates were simulated, and the test results are presented in Section 5.3. Finding out how well the estimated models reproduce the stylized facts discussed in the Introduction constitutes another way of evaluating them. Details of this approach will be discussed in connection with the empirical modelling results in Section 5.4.

## 5. APPLICATION TO THE S&P 500 US STOCK INDEX

### 5.1. Testing Linearity and Selecting the Number of States

We apply our model to the S&P 500 US stock price index. The series we have is the same as in GD, consisting of 17,055 daily observations from 3 January 1928 to 30 April 1991. We split the series into 10 subseries with 1700 observations in each, omitting the first 55 observations. The mean is subtracted from the series and this is done for each subsample before we start the analysis. The subseries are lettered from A to J in chronological order. Statistics on the subseries can be found in Table I. Each subseries has been tested for linearity (meaning the null was constant conditional variance) by bootstrap as described in Section 4.2. As the procedure is time consuming, the number of bootstrap replications of length $m = 800$ was restricted to 50. In Table II it is seen that the null hypothesis of a single regime is rejected for all subseries at the significance level 0.02. Since it is well known that the return series usually have a time-varying conditional variance, this is hardly a surprising result. In testing two regimes against three, the results in Table III are more variable. For three series, E, F, and H, the null hypothesis is not rejected at the 0.05 significance level. For those we consider a two-regime HMM to be the final one.

GD found that the series they considered only satisfied the distributional properties PED2 and PED3 after outlier reduction. As we were interested in seeing how well we could reproduce the stylized facts that GD reported, we also wanted to model outlier-reduced series. Following GD, we 'cleaned' the series by replacing every observation $r_t$ outside the interval $\bar{r}_t \pm 4\hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard deviation, by the limit of the interval and carried out the tests for the transformed series. We did this separately for each subseries, whereas GD did it for the whole (17,055 observations) series. Thus we shrank a somewhat greater number of observations towards the sample mean than did GD. The outlier reduction also enabled us to investigate the

Table I. Statistics concerning the absolute return series estimated directly from the S&P 500 subseries (D) and from the estimated hidden Markov model (M), original observations. The estimated models are those found by specification tests

| | Subseries | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J |
| Mean/standard deviation ratio | | | | | | | | | | |
| D | 0·94 | 1·03 | 0·83 | 0·94 | 1·04 | 0·98 | 1·04 | 1·13 | 1·16 | 0·79 |
| M | 0·95 | 1·09 | 0·86 | 0·97 | 1·10 | 1·03 | 0·49 | 1·14 | 1·18 | 0·93 |
| Skewness | | | | | | | | | | |
| D | 2·53 | 2·20 | 4·22 | 2·93 | 3·13 | 3·74 | 2·25 | 1·78 | 1·62 | 9·99 |
| M | 2·86 | 1·89 | 3·32 | 2·66 | 1·90 | 2·62 | 30·93 | 1·70 | 1·47 | 7·26 |
| Kurtosis | | | | | | | | | | |
| D | 12·8 | 10·9 | 33·7 | 17·8 | 27·1 | 33·7 | 12·4 | 7·83 | 7·17 | 208 |
| M | 16·4 | 8·24 | 20·9 | 15·5 | 8·23 | 14·0 | 1357 | 7·00 | 5·80 | 117 |

Table II. Maximal values of the observed *LR* statistics of a two-regime HMM and those of the simulated *LR* statistics for the original subseries A–J when testing linearity against a two-regime HMM using a parametric bootstrap, the number of times ($k$) when the simulated likelihood ratio exceeds the observed one and the $p$-value of the test

| | Obs. *LR* | Max. sim *LR*[a] | $k$ | $(k+1)/(50+1)$ |
|---|---|---|---|---|
| A | 413·14 | 6·77 | 0 | 0·02 |
| B | 55·65 | 6·77 | 0 | 0·02 |
| C | 377·77 | 6·77 | 0 | 0·02 |
| D | 180·35 | 6·77 | 0 | 0·02 |
| E | 60·28 | 6·77 | 0 | 0·02 |
| F | 45·22 | 6·77 | 0 | 0·02 |
| G | 211·97 | 6·77 | 0 | 0·02 |
| H | 149·76 | 6·77 | 0 | 0·02 |
| I | 46·18 | 6·77 | 0 | 0·02 |
| J | 558·95 | 6·77 | 0 | 0·02 |

[a] All simulated maximal *LR* statistics agree just because the same seed was used for the random number generator throughout.

Table III. Maximal values of the observed *LR* statistics of a three-regime HMM and those of the simulated *LR* statistics for the original subseries A–J when testing a two-regime HMM against a three-regime HMM using a parametric bootstrap, the number of times ($k$) when the simulated likelihood ratio exceeds the observed one and the $p$-value of the test

| | Obs. *LR* | Max. sim *LR* | $k$ | $(k+1)/(50+1)$ |
|---|---|---|---|---|
| A | 78·30 | 5·04 | 0 | 0·02 |
| B | 16·87 | 6·82 | 0 | 0·02 |
| C | 58·88 | 12·90 | 0 | 0·02 |
| D | 28·43 | 11·02 | 0 | 0·02 |
| E | 3·65 | 6·71 | 12 | 0·25 |
| F | 1·65 | 7·56 | 21 | 0·43 |
| G | 15·74 | 7·84 | 0 | 0·02 |
| H | 1·13 | 7·14 | 19 | 0·39 |
| I | 8·80 | 5·78 | 0 | 0·02 |
| J | 68·26 | 6·09 | 0 | 0·02 |

role and significance of the outlying, and thus rare, observations on the estimation results. An across-the-board reduction would have affected early subseries more than the more recent ones, which we did not want. Table IV contains statistics for the cleaned series. The test results appear in Table V and are rather similar to those in Table II. The linearity tests all reject the null hypothesis. When testing two regimes against three, Table VI shows that the null hypothesis is not rejected for subseries E. In addition, for series F and H the maximum value of the likelihood of the three-regime model remains below that of the two-regime one despite 500 starts, which is taken to imply that a two-regime model is adequate. (To ensure that the test results are just not numerical artefacts we set the tolerance limit of the optimization routine so low that any further tightening of the stopping rule only had a negligible effect on the LR statistic.) Cleaning series J means removing the crash of October 1987, but the test result suggests a three-regime model for this series all the same. Series B is a borderline case, and we consider the three-regime HMM to be the final one for it. The estimation results themselves will be discussed in the next section.

 

Table IV. Statistics concerning the absolute return series estimated directly from the S&P 500 subseries (D) and from the estimated hidden Markov model (M), outlier-reduced observations. The estimated models are those found by specification tests

| | Subseries | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J |
| Mean/standard deviation ratio | | | | | | | | | | |
| D | 0·97 | 1·06 | 0·94 | 1·00 | 1·10 | 1·07 | 1·07 | 1·14 | 1·17 | 1·02 |
| M | 0·98 | 1·09 | 0·97 | 1·00 | 1·14 | 1·10 | 1·09 | 1·14 | 1·20 | 1·05 |
| Skewness | | | | | | | | | | |
| D | 2·05 | 1·89 | 2·28 | 2·08 | 1·78 | 2·02 | 1·83 | 1·64 | 1·49 | 2·18 |
| M | 2·42 | 1·60 | 2·25 | 2·13 | 1·63 | 2·03 | 1·80 | 1·67 | 1·40 | 2·14 |
| Kurtosis | | | | | | | | | | |
| D | 8·14 | 7·53 | 9·93 | 8·76 | 7·29 | 9·06 | 7·29 | 6·61 | 6·00 | 10·2 |
| M | 12·0 | 6·10 | 10·4 | 9·76 | 6·60 | 9·44 | 7·30 | 6·89 | 5·64 | 9·97 |

Table V. Maximal values of the observed *LR* statistics of a two-regime HMM and those of the simulated *LR* statistics for the outlier-reduced subseries A–J when testing linearity against a two-regime HMM using a parametric bootstrap, the number of times ($k$) when the simulated likelihood ratio exceeds the observed one and the $p$-value of the test

| | Obs. *LR* | Max. sim *LR*[a] | $k$ | $(k+1)/(50+1)$ |
|---|---|---|---|---|
| A | 252·67 | 6·77 | 0 | 0·02 |
| B | 33·40 | 6·77 | 0 | 0·02 |
| C | 141·37 | 6·77 | 0 | 0·02 |
| D | 69·84 | 6·77 | 0 | 0·02 |
| E | 41·10 | 6·77 | 0 | 0·02 |
| F | 24·34 | 6·77 | 0 | 0·02 |
| G | 168·44 | 6·77 | 0 | 0·02 |
| H | 131·05 | 6·77 | 0 | 0·02 |
| I | 34·75 | 6·77 | 0 | 0·02 |
| J | 177·99 | 6·77 | 0 | 0·02 |

[a] All simulated maximal *LR* statistics agree just because the same seed was used for the random number generator throughout.

Table VI. Maximal values of the observed *LR* statistics of a three-regime HMM and those of the simulated *LR* statistics for the outlier-reduced subseries A–J when testing a two-regime HMM against a three-regime HMM using a parametric bootstrap, the number of times ($k$) when the simulated likelihood ratio exceeds the observed one and the $p$-value of the test

| | Obs. *LR* | Max. sim *LR* | $k$ | $(k+1)/(50+1)$ |
|---|---|---|---|---|
| A | 13·44 | 8·51 | 0 | 0·02 |
| B | 5·34 | 6·15 | 1 | 0·04 |
| C | 49·85 | 17·09 | 0 | 0·02 |
| D | 29·84 | 11·65 | 0 | 0·02 |
| E | 1·23 | 14·55 | 29 | 0·57 |
| F | −0·031 | | | |
| G | 8·24 | 6·98 | 0 | 0·02 |
| H | −0·12 | | | |
| I | 6·90 | 6·34 | 0 | 0·02 |
| J | 27·98 | 6·49 | 0 | 0·02 |

The logical next step in the specification sequence would be to test three regimes against four for series for which a three-regime model was accepted. This would have involved estimating four-regime models and a four-regime bootstrap. Since the amount of computing time necessary for doing this would have been prohibitive we had to abstain from such an extension.

### 5.2. Estimation Results

We report results of both original and outlier-reduced data. First, consider the HMMs estimated from the original data. Note that the regimes are ordered by increasing standard deviations. Table VII contains the estimated transition probabilities and standard deviations for its two or three normal distributions, for the ten original subseries. It also contains the stationary probabilities of the process being in a given state. There exist some similarities between the models. Of the two-regime models, E and F (calling the models by the same name as the series) are similar in the sense that they have one rather persistent regime: $\hat{p}_{11} > 0{\cdot}96$, for both models,

Table VII. Estimated transition probabilities, standard deviations of normal distributions of different regimes and stationary probability estimates for models A–J from original observations

| Model | Transition probabilities | | | Standard deviations | Stationary probabilities |
|---|---|---|---|---|---|
| A | 0·966 | 0·034 | 0·000 | 0·0091 | 0·248 |
| | 0·013 | 0·960 | 0·027 | 0·0216 | 0·672 |
| | 0·000 | 0·230 | 0·770 | 0·0601 | 0·079 |
| B | 0·225 | 0·759 | 0·016 | 0·0074 | 0·227 |
| | 0·299 | 0·695 | 0·006 | 0·0126 | 0·578 |
| | 0·016 | 0·021 | 0·963 | 0·0245 | 0·195 |
| C | 0·375 | 0·612 | 0·013 | 0·0028 | 0·351 |
| | 0·374 | 0·612 | 0·014 | 0·0082 | 0·586 |
| | 0·000 | 0·196 | 0·804 | 0·0250 | 0·063 |
| D | 0·457 | 0·541 | 0·002 | 0·0034 | 0·348 |
| | 0·312 | 0·670 | 0·018 | 0·0087 | 0·601 |
| | 0·019 | 0·210 | 0·771 | 0·0223 | 0·051 |
| E | 0·964 | 0·036 | | 0·0049 | 0·750 |
| | 0·108 | 0·892 | | 0·0109 | 0·250 |
| F | 0·987 | 0·013 | | 0·0049 | 0·884 |
| | 0·098 | 0·902 | | 0·0139 | 0·116 |
| G | 0·970 | 0·029 | 0·001 | 0·0039 | 0·600 |
| | 0·043 | 0·955 | 0·002 | 0·0087 | 0·399 |
| | 0·361 | 0·332 | 0·307 | 0·2372 | 0·001 |
| H | 0·995 | 0·005 | | 0·0063 | 0·691 |
| | 0·011 | 0·989 | | 0·0126 | 0·309 |
| I | 0·473 | 0·510 | 0·017 | 0·0055 | 0·315 |
| | 0·404 | 0·586 | 0·010 | 0·0087 | 0·390 |
| | 0·028 | 0·004 | 0·968 | 0·0125 | 0·295 |
| J | 0·992 | 0·006 | 0·002 | 0·0078 | 0·861 |
| | 0·048 | 0·930 | 0·022 | 0·0150 | 0·133 |
| | 0·018 | 0·759 | 0·223 | 0·0745 | 0·006 |

 

whereas the higher regime is 'semi-persistent', $\hat{p}_{22} \approx 0{\cdot}90$. The three-regime models are less similar to each other. The estimation results seem heavily dependent on outlying observations. Models A, G and perhaps J have two persistent states, B and I have one and C and D none. For models A, C, D, G and J, the regime with the highest standard deviation is clearly an outlier regime with $\hat{p}_{33} \leqslant 0{\cdot}8$ and a low stationary probability. These findings accord with the fact that the absolute return series A, C, D, and J have high kurtosis. On the other hand, the kurtosis of series G is relatively low. Compared to $\hat{\sigma}_3 = 0{\cdot}237$, an extremely high value, this is a first indication of model G misrepresenting the data quite badly.

The results of modelling the outlier-reduced series in Table VIII are somewhat different from those obtained with the original data although similarities do exist. The two-regime models E, F, and H are similar to the ones obtained from the original data because these subseries do not contain large outliers (compare the kurtosis with and without outlier reduction in Tables I and IV). In models E and F, the higher regime is slightly more persistent than before. The three-regime model for series G now has two persistent regimes ($\hat{p}_{ii} > 0{\cdot}94$, $i = 1, 3$) and a rather rarely visited middle regime. The estimate $\hat{\sigma}_3$ is not a cause of worry. The same is true for model A ($\hat{p}_{ii} > 0{\cdot}98$, $i = 1, 2$). For other series modelled with three-regime HMMs, reducing the outliers also makes a difference. In particular, models I and J have changed completely. Model J for the data with the October 1987 outliers reduced contains only a single seemingly high-persistence regime and it is the highest one. However, the estimated probability of entering that state is extremely low so that the state is still more of an outlier regime than anything else. Most of the time the process is visiting the two lowest regimes. The lowest state in model I now seems an outlier regime and the remaining two are persistent ones. We may conclude that the outlier reduction may not have an impact on the number of states selected, but it does seem to change the interpretation of many of the three-regime models. The number of pure 'outlier-determined' models is less than in the case of original observations but at least three regimes are still required for the same series as before.

It is clear from Table VII that the estimated relationship is not fully stable over time. Nevertheless, some regularities between adjacent models can be found in the results. First, the three-regime models C and D are rather alike. Second, the two-regime models E, F, and H that follow have a fairly similar pattern. The estimated standard deviations of the regimes in different models are reasonably close to each other. Model G in between is different but then it fails the evaluation tests. Thus, although the results are not stable overall, the same pattern seems to prevail at least over a number of subperiods. This fact is even more visible in Table VIII, based on outlier-reduced data as model G is now quite well in line with E, F, and H. This indicates that the obtained results are not completely spurious.

### 5.3. Evaluation: Test Results

As mentioned above, testing three regimes against four is not a practical idea because of the amount of computation involved. Nevertheless, the goodness of fit of three-regime HMMs can be investigated using the test of Milhøj (1981) as discussed in Section 4.3. The results for the models based on the original data appear in Table IX. The test rejects models A and G at the 0·05 significance level. We shall see later on that most evaluation checks also reject model G. The fit of model C is a borderline case ($\alpha = 0{\cdot}10$). These three models are ones with at least one outlier regime. The other three-regime models pass the test. When the outlier-reduced series are concerned (see Table X) four of the five three-regime HMMs pass the test, while model G is rejected.

Table VIII. Estimated transition probabilities, standard deviations of normal distributions of different regimes and stationary probability estimates for models A–J from outlier-reduced observations

| Model | Transition probabilities | | | Standard deviations | Stationary probabilities |
|---|---|---|---|---|---|
| A | 0·986 | 0·014 | 0·000 | 0·0106 | 0·475 |
| | 0·016 | 0·848 | 0·135 | 0·0220 | 0·421 |
| | 0·000 | 0·549 | 0·451 | 0·0452 | 0·104 |
| B | 0·677 | 0·313 | 0·010 | 0·0074 | 0·347 |
| | 0·430 | 0·523 | 0·047 | 0·0137 | 0·261 |
| | 0·000 | 0·040 | 0·960 | 0·0212 | 0·392 |
| C | 0·262 | 0·679 | 0·058 | 0·0026 | 0·254 |
| | 0·300 | 0·700 | 0·000 | 0·0073 | 0·598 |
| | 0·058 | 0·042 | 0·900 | 0·0163 | 0·148 |
| D | 0·281 | 0·605 | 0·114 | 0·0031 | 0·273 |
| | 0·329 | 0·671 | 0·000 | 0·0077 | 0·583 |
| | 0·030 | 0·187 | 0·783 | 0·0160 | 0·144 |
| E | 0·963 | 0·037 | | 0·0046 | 0·648 |
| | 0·069 | 0·931 | | 0·0092 | 0·352 |
| F | 0·985 | 0·015 | | 0·0047 | 0·822 |
| | 0·070 | 0·930 | | 0·0107 | 0·178 |
| G | 0·953 | 0·029 | 0·018 | 0·0040 | 0·647 |
| | 0·416 | 0·340 | 0·244 | 0·0068 | 0·029 |
| | 0·057 | 0·000 | 0·943 | 0·0090 | 0·324 |
| H | 0·995 | 0·005 | | 0·0062 | 0·686 |
| | 0·011 | 0·989 | | 0·0124 | 0·314 |
| I | 0·158 | 0·681 | 0·161 | 0·0032 | 0·058 |
| | 0·061 | 0·939 | 0·000 | 0·0076 | 0·664 |
| | 0·033 | 0·001 | 0·966 | 0·0121 | 0·278 |
| J | 0·564 | 0·427 | 0·008 | 0·0060 | 0·613 |
| | 0·895 | 0·105 | 0·000 | 0·0115 | 0·293 |
| | 0·056 | 0·000 | 0·944 | 0·0190 | 0·094 |

To gain extra insight in the functioning of the test we also carried it out for the two-regime models of all series A–J. For the original data, the results are found in Table XI. The Milhøj test rejects models C and G at $\alpha = 0{\cdot}02$, whereas D is a borderline case. These results agree with those of the bootstrapped likelihood ratio test. On the other hand, the two-regime models of series A, B, I, and J are accepted although the likelihood ratio test rejects them. As to the outlier-reduced series, it is seen from Table XII that the two-regime model is rejected for series C, D, and G. For model D the significance level is 0·06, for the other two models 0·02. The likelihood ratio test also rejects the two-regime model for these series. Furthermore, both tests accept the two-regime model for series E, F, and H. Again there exist series (A, B, I, and J) for which the likelihood ratio test rejects the null hypothesis of two states, whereas the Milhøj test does not. A tentative conclusion thus is that the goodness-of-fit test may be less powerful than the likelihood ratio specification test. This result is not surprising since the alternative hypothesis in the Milhøj test is more general than that in the likelihood ratio test.

 

Table IX. $p$-values of the Milhøj goodness-of-fit test for estimated HMMs with three regimes, original observations; $k$ = number of times the simulated value of the statistic exceeds the value computed from the estimated model

| Model | $k$ | $(k+1)/(50+1)$ |
|---|---|---|
| A | 1 | 0·04 |
| B | 19 | 0·39 |
| C | 4 | 0·10 |
| D | 18 | 0·37 |
| G | 0 | 0·02 |
| I | 12 | 0·25 |
| J | 45 | 0·90 |

Table X. $p$-values of the Milhøj goodness-of-fit test for estimated HMMs with three regimes, outlier-reduced observations; $k$ = number of times the simulated value of the statistic exceeds the value computed from the estimated model

| Model | $k$ | $(k+1)/(50+1)$ |
|---|---|---|
| A | 5 | 0·12 |
| B | 4 | 0·10 |
| C | 5 | 0·12 |
| D | 13 | 0·27 |
| G | 0 | 0·02 |
| I | 12 | 0·25 |
| J | 50 | 1·00 |

Table XI. $p$-values of the Milhøj goodness-of-fit test for estimated HMMs with two regimes, original observations; $k$ = number of times the simulated value of the statistic exceeds the value computed from the estimated model

| Model | $k$ | $(k+1)/(50+1)$ |
|---|---|---|
| A | 8 | 0·18 |
| B | 16 | 0·33 |
| C | 0 | 0·02 |
| D | 4 | 0·10 |
| E | 9 | 0·20 |
| F | 24 | 0·49 |
| G | 0 | 0·02 |
| H | 19 | 0·39 |
| I | 11 | 0·24 |
| J | 38 | 0·76 |

Table XII. $p$-values of the Milhøj goodness-of-fit test for estimated HMMs with two regimes, outlier-reduced observations; $k$ = number of times the simulated value of the statistic exceeds the value computed from the estimated model

| Model | $k$ | $(k+1)/(50+1)$ |
|---|---|---|
| A | 7 | 0·16 |
| B | 19 | 0·39 |
| C | 0 | 0·02 |
| D | 2 | 0·06 |
| E | 17 | 0·35 |
| F | 17 | 0·35 |
| G | 0 | 0·02 |
| H | 18 | 0·37 |
| I | 10 | 0·22 |
| J | 48 | 0·96 |

### 5.4. Evaluation: Stylized Facts

Another, more informal way of checking the results is to see how well the models reproduce stylized facts in the data. Table I contains the mean/standard deviation ratio, skewness and kurtosis of absolute returns from the original data for all the subseries and their estimates from corresponding models. The latter were computed from equation (6) by using maximum likelihood estimates of the stationary probabilities and standard deviations. The mean/standard deviation ratio is reproduced rather well, model G being an exception. Given Figure 1 (for a two-regime HMM) this may not be surprising because a large number of combinations of standard deviation and stationary probability yield ratios close to unity, and that is where the ratios observed from the data lie. As to the skewness and kurtosis, a clear tendency emerges. If these values are relatively small in the data, the models also yield small estimates, and they increase when the corresponding subseries display higher values. The only clear failure is model G. The other models reproduce the stylized facts reasonably well. The kurtosis of series J estimated from the model remains far below that estimated directly from the data, but on the other hand the latter estimate is remarkably large.

Figure 4 shows the autocorrelation functions of absolute returns for the first 50 lags for the series A–J and the corresponding models. It is seen that the autocorrelations estimated directly from the series decay very slowly as Granger and Ding (1995a) and GD noted. The autocorrelation function estimated from the model typically shows somewhat faster decay, but it does often capture the general tendency reasonably well. Models C, D, and again G are perhaps the largest exceptions to this observation. It should be noted that the HMM can only produce series with exponentially decaying autocorrelation functions. As for TP2, the model thus seems doomed from the start. Nevertheless, the exponential decay in an HMM may also be quite slow for appropriate values of the transition probabilities. Our results show, however, that given all the other properties of the data the HMM have to satisfy, the maximum likelihood estimated models cannot capture the behaviour of the empirical autocorrelation functions in a satisfactory fashion.

Table IV contains the mean/standard deviation ratio, skewness and kurtosis of absolute returns for the outlier-reduced series and the estimated models. All models, that for series G included, reproduce the stylized facts quite well. Model A underestimates both the skewness and kurtosis, but in general, the outlier reduction has a remarkable effect on the results. It is clear

 

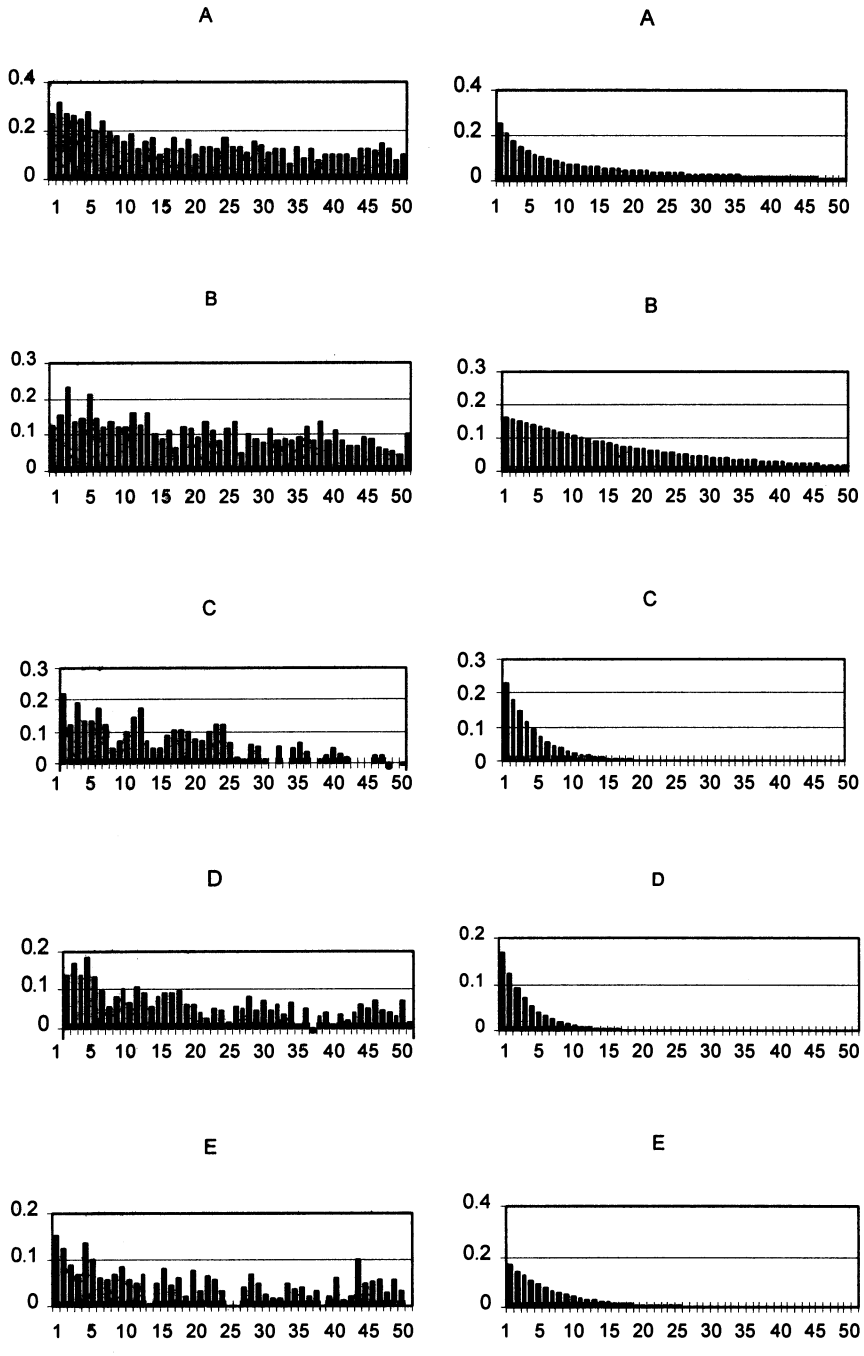


Figure 4a

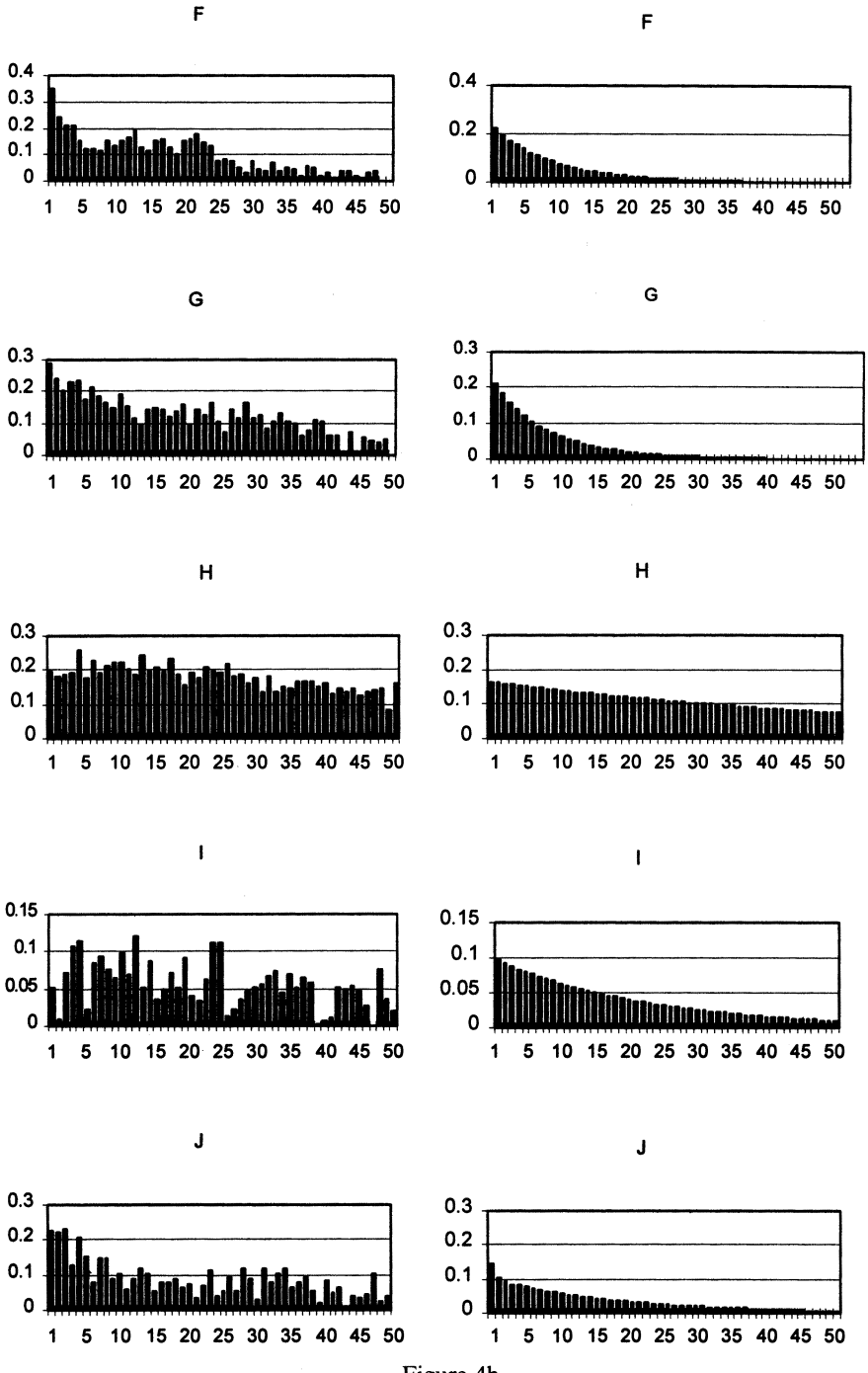


Figure 4b

Figure 4. Autocorrelation functions of absolute returns for subseries A–J estimated from the original observations (left-hand panels) and from the hidden Markov models (right-hand panels)

 

Table XIII. Values of $\theta$ maximizing the first-order autocorrelation of $|r_t|^\theta$ estimated from series A–J and the corresponding HMMs, original observations

| Period | Data | Model |
|---|---|---|
| A | 1·4 | 1·0 |
| B | 1·4 | 1·4 |
| C | 1·4 | 1·4 |
| D | 1·0 | 1·6 |
| E | 1·2 | 1·2 |
| F | 1·6 | 1·4 |
| G | 1·4 | 0·8 |
| H | 1·2 | 1·2 |
| I | 1·8 | 1·4 |
| J | 1·2 | 1·0 |

from this comparison that the HMM is sensitive to outlying observations but then this means that the model faithfully reflects the properties of the data. As to the autocorrelation functions the broad picture is as before: the decay of the autocorrelations in the models is faster than in reality. The graphs are therefore not shown separately. The slow decay seems the most difficult stylized fact to reproduce with an HMM.

The remaining stylized fact of GD to check is the Taylor effect (TP3). For this we need autocorrelation estimates from the models. After estimating an HMM we obtain conditional estimates of $E|X|^\theta$ using

$$E|X|^\theta = \mu_\theta = \sqrt{\frac{2}{\pi\sigma^2}} \int_0^\infty y^\theta \, \mathrm{e}^{-y^2/2\sigma^2} \, \mathrm{d}y$$
$$= \frac{(2\sigma^2)^{\theta/2}}{\sqrt{\pi}} \Gamma((\theta + 1)/2) \tag{13}$$

where $\Gamma(\cdot)$ is the Gamma function. These are computed for each regime in turn by substituting the estimated variance of the corresponding normal distribution for $\sigma^2$ in expression (13). To obtain the autocorrelation function, the results are inserted while defining $g(r_t) = |r_t|^\theta$. This and

Table XIV. Values of $\theta$ maximizing the first-order autocorrelation of $|r_t|^\theta$ estimated from series A–J and the corresponding HMMs, outlier-reduced observations

| Period | Data | Model |
|---|---|---|
| A | 1·6 | 0·8 |
| B | 1·6 | 0·8 |
| C | 2·0 | 1·4 |
| D | 1·4 | 1·4 |
| E | 1·4 | 1·2 |
| F | 2·0 | 1·4 |
| G | 1·8 | 1·0 |
| H | 1·4 | 1·2 |
| I | 2·0 | 1·6 |
| J | 2·0 | 1·8 |

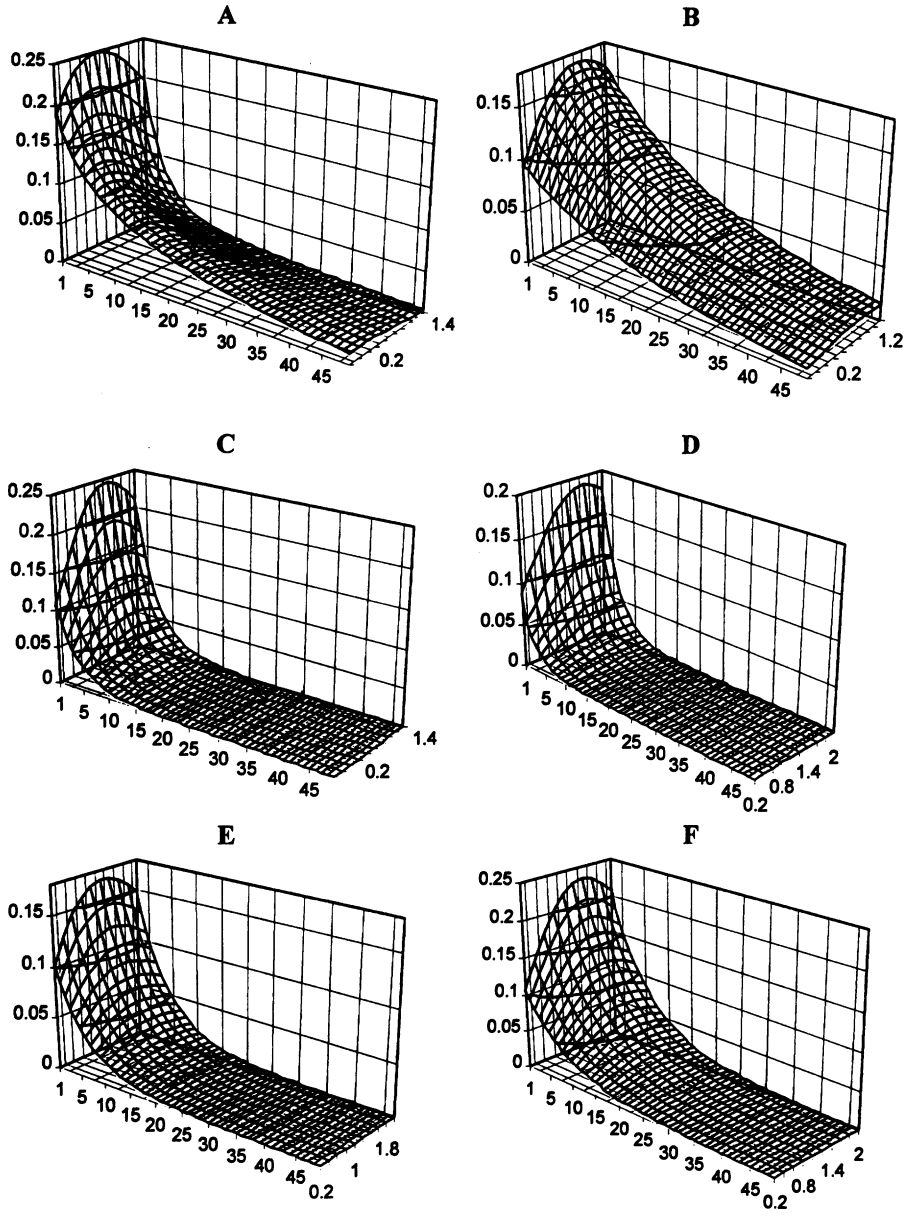


Figure 5a

the use of estimated stationary and transition probabilities yield the autocovariances defined in matrix form in expression (8). An estimate of the second moment of $|r_t|^{\theta}$ needed in the autocorrelation function is obtained through expression (6) by setting $g(r_t) = |r_t|^{2\theta}$. Tables XIII and XIV contain the values which maximize the first-order autocorrelation for each model over the range $\theta = 0{\cdot}2, 0{\cdot}4, \ldots, 2{\cdot}0$. As Figures 5 and 6 indicate, the autocorrelation functions estimated

 

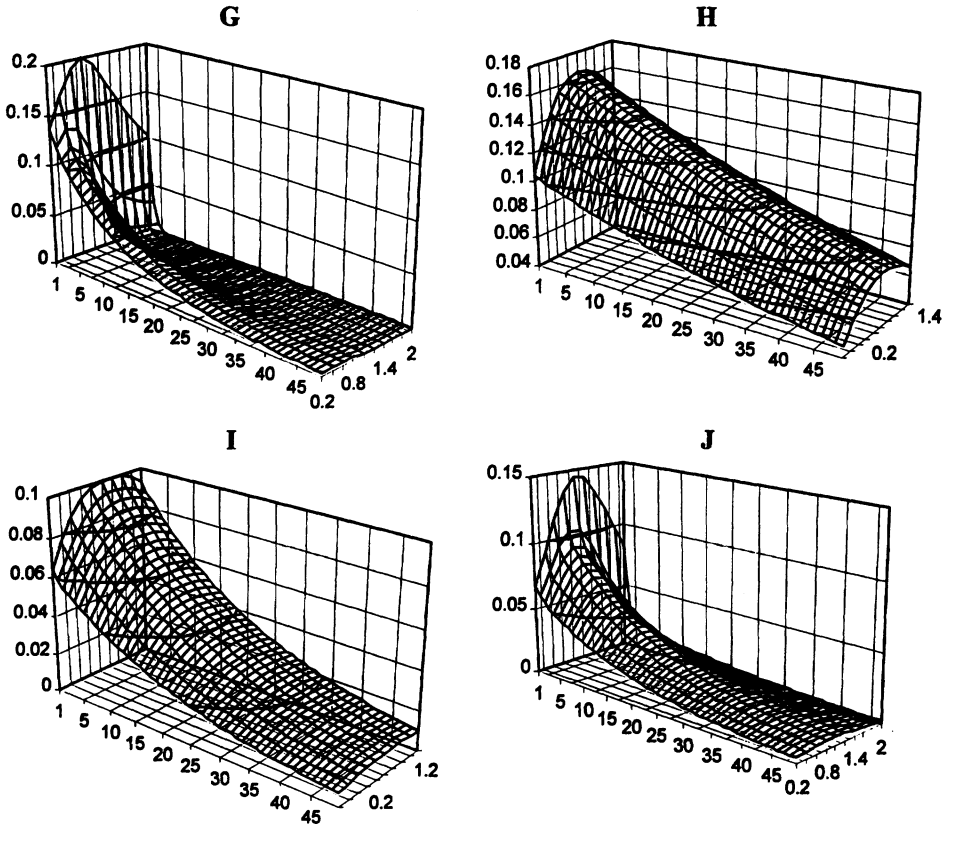


Figure 5b

Figure 5. Autocorrelation functions for transformed subseries $|r_t^{\theta}|$, $\theta = 0{\cdot}2, 0{\cdot}4, \ldots, 2{\cdot}0$, estimated from the hidden Markov models A–J, original data

from the models have a ridge appearing for $\theta < 2$ for both the original and the outlier-reduced data. The maximum first-order autocorrelations are obtained for $0{\cdot}8 \leqslant \theta \leqslant 1{\cdot}8$ (see Tables XIII and XIV). The maximizing values of $\theta$ agree reasonably well with those obtained by estimating the autocorrelations directly from the subseries A–J. Another interesting feature in Table XIV is that the outlier reduction seems to weaken the Taylor effect in the data as the autocorrelation-maximizing $\theta$ systematically increases with this transformation. For the transformed subseries C, F, I, and J the maximizing $\theta$ even has the value 2·0.

## 6. CONCLUSIONS

We have shown that in modelling returns, a mixture of normals is capable of characterizing stylized facts that Granger and Ding (1995a) and GD found in a large number of high-frequency series. Furthermore, higher-order dependence present in those series can be conveniently modelled using the hidden Markov model. The one stylized fact that cannot be generated by this model, at least not easily by ML estimation, is the very slowly decaying autocorrelation function for the absolute

 

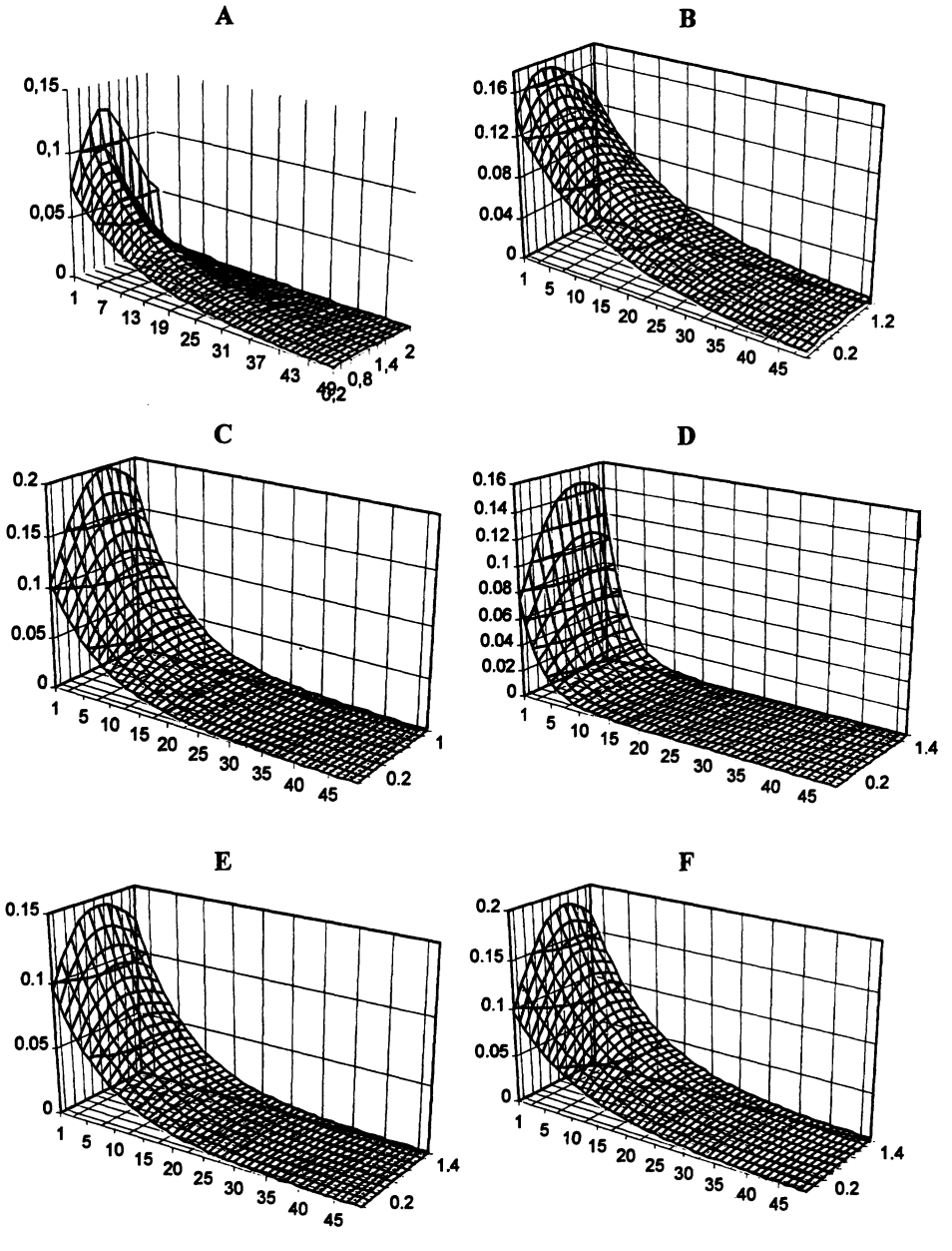


Figure 6a

returns. While Ding and Granger (1996) proposed a model which reproduces this property, that paper is not explicitly concerned about the distributional properties of the absolute returns.

There exist many other studies applying the HMM to economic series. Our study differs from many of them in the sense that the number of regimes has not been fixed in advance. Because of

 

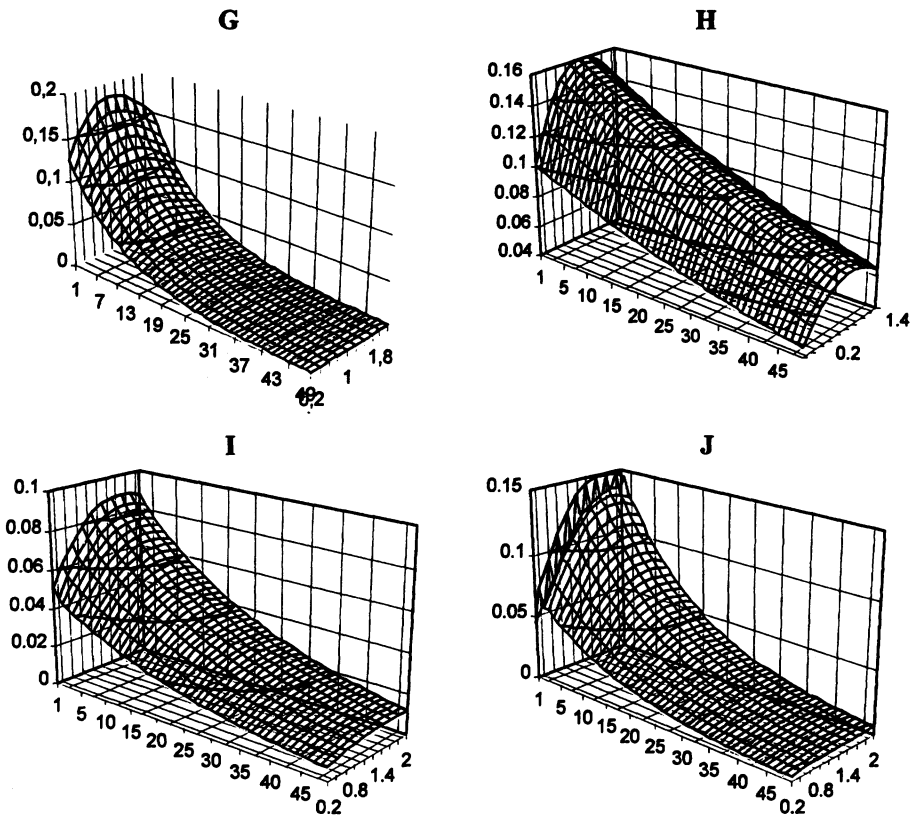


Figure 6b

Figure 6. Autocorrelation functions for transformed subseries $|r_t|^{\theta}$, $\theta = 0{\cdot}2, 0{\cdot}4, \ldots, 2{\cdot}0$, estimated from the hidden Markov models A–J, outlier-reduced data

the identification problem present in the form of an overparameterized HMM we select the model by testing from specific to general and thus seek to minimize the risk of fitting unidentified models to data. We believe that this should be a standard approach when modelling with the HMM.

Fitting an HMM to subseries of the long S&P 500 series reveals a potential problem: very exceptional observations have a tendency of being allocated in a separate regime. Having said that we would like to remind the reader of the fact that there is some stability in the parameter estimates over time so that the results do not appear completely spurious. The HMM based on a mixture of normal distributions thus seems an interesting alternative to the double exponential error distribution Granger and Ding (1995a) and GD used to characterize high-frequency return series.

## APPENDIX

In this appendix we derive the spectral density of an HMM. To that end, let $\{(S_t, X_t)\}$ be an HMM, i.e. $\{S_t\}$ is the hidden Markov chain and $\{X_t\}$ is the observable process. The transition probability matrix $\mathbf{P}$ of $\{S_t\}$ is assumed to be ergodic. Let also $\mathbf{M}_1$ and $\mathbf{M}_2$ be the diagonal

 

matrices with entries $E\{X_t \mid S_t = i\}$ and $E\{X_t^2 \mid S_t = i\}$, respectively. In our case with $X_t = |r_t|$, $\mathbf{M}_1 = \text{diag}(\sqrt{[2/\pi]}\sigma_i)$ and $\mathbf{M}_2 = \text{diag}(\sigma_i^2)$, respectively. We can then write $E\{X_t\} = \mathbf{aM}_1\mathbf{1}$ and $E\{X_t^2\} = \mathbf{aM}_2\mathbf{1}$, where $\mathbf{a}$ is the vector of stationary state probabilities for $\mathbf{P}$ and $\mathbf{1}$ is a column vector of ones. Thus,

$$\text{Var}(X_t) = \mathbf{aM}_2\mathbf{1} - (\mathbf{aM}_1\mathbf{1})^2$$

and as shown in Section 3,

$$\text{cov}(X_0, X_t) = \mathbf{aM}_1(\mathbf{P}^t - \mathbf{1a})\mathbf{M}_1\mathbf{1}$$

for $t \geqslant 1$. The spectral density of $\{X_t\}$ is defined by

$$g(\omega) = \frac{1}{2\pi}\sum_{t=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t}\,\text{cov}(X_0, X_t)$$

**Theorem 1** For the HMM defined above,

$$2\pi g(\omega) = \mathbf{aM}_2\mathbf{1} - (\mathbf{aM}_1\mathbf{1})^2 + 2\mathbf{aM}_1(\mathbf{B}\cos\omega - \mathbf{B}^2)(\mathbf{I} - 2\mathbf{B}\cos\omega + \mathbf{B}^2)^{-1}\mathbf{M}_1\mathbf{1}$$

where $\mathbf{B} = \mathbf{P} - \mathbf{1a}$.

For the proof, we need the following lemma.

**Lemma 1** For each $t = 1, 2, 3, \ldots$, $\mathbf{P}^t - \mathbf{1a} = (\mathbf{P} - \mathbf{1a})^t$.

*Proof.* This follows easily by induction.

*Proof of Theorem 1.* Since $\mathbf{P}$ is ergodic, it has a single eigenvalue of modulus one, namely $\lambda = 1$. Thus, the spectral radius of $\mathbf{B}$ is less than one, and the same holds true for $\mathbf{B}\exp(\mathrm{i}\omega t)$. This fact and Lemma 1 imply that

$$\sum_{t=1}^{\infty}\mathrm{e}^{-\mathrm{i}\omega t}\,\text{cov}(X_0, X_t) = \mathbf{aM}_1\left\{\sum_{t=1}^{\infty}(\mathbf{B}\,\mathrm{e}^{-\mathrm{i}\omega})^t\right\}\mathbf{M}_1\mathbf{1} = \mathbf{aM}_1\{(\mathbf{I} - \mathbf{B}\,\mathrm{e}^{-\mathrm{i}\omega})^{-1} - \mathbf{I}\}\mathbf{M}_1\mathbf{1}.$$

Similarly,

$$\sum_{t=-\infty}^{-1}\mathrm{e}^{-\mathrm{i}\omega t}\,\text{cov}(X_0, X_t) = \mathbf{aM}_1\{(\mathbf{I} - \mathbf{B}\,\mathrm{e}^{\mathrm{i}\omega})^{-1} - \mathbf{I}\}\mathbf{M}_1\mathbf{1}$$

whence

$$2\pi g(\omega) = \text{Var}(X_t) + \mathbf{aM}_1\{(\mathbf{I} - \mathbf{B}\,\mathrm{e}^{-\mathrm{i}\omega})^{-1} + (\mathbf{I} - \mathbf{B}\,\mathrm{e}^{\mathrm{i}\omega})^{-1} - 2\mathbf{I}\}\mathbf{M}_1\mathbf{1}$$

Now,

$$\begin{aligned}(\mathbf{I} - \mathbf{B}\,\mathrm{e}^{-\mathrm{i}\omega})^{-1} &= (\mathbf{I} - \mathbf{B}\,\mathrm{e}^{\mathrm{i}\omega})\{(\mathbf{I} - \mathbf{B}\,\mathrm{e}^{-\mathrm{i}\omega})(\mathbf{I} - \mathbf{B}\,\mathrm{e}^{\mathrm{i}\omega})\}^{-1}\\ &= (\mathbf{I} - \mathbf{B}\,\mathrm{e}^{\mathrm{i}\omega})(\mathbf{I} - 2\mathbf{B}\cos\omega + \mathbf{B}^2)^{-1}\end{aligned}$$

 

so that

$$(\mathbf{I} - \mathbf{B}\,e^{-i\omega})^{-1} + (\mathbf{I} - \mathbf{B}\,e^{i\omega})^{-1} = (\mathbf{I} - \mathbf{B}\,e^{i\omega} + \mathbf{I} - \mathbf{B}\,e^{-i\omega})(\mathbf{I} - 2\mathbf{B}\cos\omega + \mathbf{B}^2)^{-1}$$
$$= (2\mathbf{I} - 2\mathbf{B}\cos\omega)(\mathbf{I} - 2\mathbf{B}\cos\omega + \mathbf{B}^2)^{-1}$$

and

$$(\mathbf{I} - \mathbf{B}\,e^{-i\omega})^{-1} + (\mathbf{I} - \mathbf{B}\,e^{i\omega})^{-1} - 2\mathbf{I} = (2\mathbf{B}\cos\omega - 2\mathbf{B}^2)(\mathbf{I} - 2\mathbf{B}\cos\omega + \mathbf{B}^2)^{-1}$$

The proof is complete.

ACKNOWLEDGEMENTS

The research of TR was supported by the Swedish Natural Science Research Council (contract No. M-AA/MA 10538-303), whereas TT received support from the Swedish Council for Research in the Humanities and Social Sciences and SÅ from the Tore Browaldh Foundation for Scientific Research (contract No. T96541). Versions of the paper have been presented at the 14th International Symposium on Forecasting, Istanbul, June 1996, the NBER/NSF Time Series Seminar, Rotterdam, October 1996, and in seminars at the Central Bank of Norway (Oslo), CREST-INSEE (Malinvaud Seminar, Paris), GREMAQ-Université de Sciences Sociales (Toulouse), Humboldt-Universität zu Berlin, Institute for Advanced Studies (Vienna) and Swedish School of Economics (Helsinki). Comments from participants, Christian Robert in particular, and two referees are gratefully acknowledged. John Geweke (associate editor) has contributed with several useful suggestions. Furthermore, we wish to thank Clive Granger for inspiration, Tony Hall and Gabriela Mundaca for useful remarks, and William Schwert for the S&P 500 daily series. The responsibility of any errors or shortcomings in the paper remains ours.

REFERENCES

Bickel, P. J. and D. A. Freedman (1981), 'Some asymptotic theory for the bootstrap', *Annals of Statistics*, **9**, 1196–1217.

Bickel, P. J. and Y. Ritov (1996), 'Inference in hidden Markov models: Local asymptotic normality in the stationary case', *Bernoulli*, **2**, 199–228.

Bickel, P. J., F. Götze and W. R. van Zwet (1997), 'Resampling fewer than *n* observations: gains, losses, and remedies for losses', *Statistica Sinica*, **7**, 1–31.

Bickel, P. J., Y. Ritov and T. Rydén (1997), 'Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models', preprint.

Davies, R. B. (1977), 'Hypothesis testing when a nuisance parameter is present only under the alternative', *Biometrika*, **64**, 247–54.

Ding, Z. and C. W. J. Granger (1996), 'Modeling volatility persistence of speculative returns: A new approach', *Journal of Econometrics*, **73**, 185–215.

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

Granger, C. W. J. and Z. Ding (1995a), 'Some properties of absolute returns. An alternative measure of risk', *Annales d'économie et de statistique*, **40**, 67–91.

Granger, C. W. J. and Z. Ding (1995b), 'Stylized facts on the temporal and distributional properties of daily data from speculative markets', Department of Economics, University of California, San Diego, unpublished paper.

 

Granger, C. W. J. and T. Teräsvirta (1993), *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

Hamilton, J. D. (1988), 'Rational-expectations econometric analysis of changes in regime: an investigation of the term structure of interest rates', *Journal of Economic Dynamics and Control*, **12**, 385–423.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Leroux, B. G. (1992), 'Maximum-likelihood estimation for hidden Markov models', *Stochastic Processes and Their Applications*, **40**, 127–43.

Lindgren, G. (1978), 'Markov regime models for mixed distributions and switching regressions', *Scandinavian Journal of Statistics*, **5**, 81–91.

McLachlan, G. J. (1987), 'On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture', *Applied Statistics*, **36**, 318–24.

Milhøj, A. (1981), 'A test of fit in time series models', *Biometrika*, **68**, 177–87.

Mittnik, S. and S. T. Rachev (1993), 'Modelling asset returns with alternative stable distributions', *Econometric Reviews*, **12**, 261–330.

Pagan, A. R. and G. W. Schwert (1990), 'Alternative models for conditional stock volatility', *Journal of Econometrics*, **45**, 267–90.

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling (1989), *Numerical Recipes*, Cambridge University Press, Cambridge.

Saikkonen, P. (1984), 'Asymptotic relative efficiency of some tests of fit in time series models', *Journal of Time Series Analysis*, **4**, 69–78.

Sola, M. and A. Timmermann (1994), 'Fitting the moments: A comparison of ARCH and regime switching models for daily stock returns', London Business School, Centre for Economic Forecasting, Discussion Paper No. DP 6-94.

Tyssedal, J. S. and D. Tjøstheim (1988), 'An autoregressive model with suddenly changing parameters and an application to stock market prices', *Applied Statistics*, **37**, 353–69.

 