Modeling Flat Stretches, Bursts, and Outliers in Time Series Using Mixture Transition Distribution Models

Nhu D. LE, R. Douglas MARTIN, and Adrian E. RAFTERY

The class of mixture transition distribution (MTD) time series models is extended to general non-Gaussian time series. In these models the conditional distribution of the current observation given the past is a mixture of conditional distributions given each one of the last p observations. They can capture non-Gaussian and nonlinear features such as flat stretches, bursts of activity, outliers, and changepoints in a single unified model class. They can also represent time series defined on arbitrary state spaces, univariate or multivariate, continuous, discrete or mixed, which need not even be Euclidean. They perform well in the usual case of Gaussian time series without obvious nonstandard behaviors. The models are simple, analytically tractable, easy to simulate, and readily estimated. The stationarity and autocorrelation properties of the models are derived. A simple EM algorithm is given and shown to work well for estimation. The models are applied to several real and simulated datasets with satisfactory results. They appear to capture the features of the data better than the best competing autoregressive integrated moving average (ARIMA) models.

KEY WORDS: Autocorrelation; Autoregressive integrated moving average model; EM algorithm; Mixture transition distribution; Non-Gaussian time series; Stationarity.

1. INTRODUCTION

Many time series exhibit non-Gaussian features such as flat stretches, bursts of activity, outliers, and changepoints. When the aim of the analysis is prediction or simulation, it is important to model such behavior explicitly. Alternative approaches in the presence of outliers and bursts are to use robust estimation procedures (Martin and Yohai 1986) or to remove outliers based on the use of diagnostics (Bruce and Martin 1989); other references on outliers in time series can be found in these papers. These approaches are valuable when the aim is estimation of structural features; but when the primary aim is prediction, it is important to obtain a full predictive distribution that takes into account the possibility of future flat stretches, bursts, and outliers. It is also possible to exploit the presence of flat stretches so as to obtain narrower and more precise predictive intervals. Similar remarks apply when the goal is simulation.

Here we consider a broad class of models, called *mixture transition distribution* (MTD) models, that can represent all of these features. The models are simple, analytically tractable, and easy to simulate and to estimate, and they appear capable of capturing a wide range of nonstandard behaviors seen in practice. In this article we focus on a small subclass consisting of stationary models involving mixtures of conditional Gaussian distributions, and we indicate how this may be extended in many directions.

As an example, consider Figure 1a, which shows a series of 300 consecutive hourly viscosity readings in a chemical process. This series clearly has frequent flat stretches during which the viscosity stayed virtually constant. One would expect this fact to be useful in prediction. This was series D of Box and Jenkins (1976), who chose to model it using an autoregressive AR(1) model. Figure 1b shows a simulated realization from the fitted AR(1) model. The AR(1) sample path does not look at all like the original data, and in particular fails to reproduce the flat stretches. Figure 1c shows a simulated realization from a fitted MTD model; it clearly reproduces the flat stretches, and it fits the data well in other ways as well.

For prediction, there is quite a lot to be gained by using an MTD model rather than an AR model in this example. The variability in the predictive distribution from the MTD model is 60% less than in that from the AR model, and the MTD-based predictive intervals also have empirical coverages much closer to their nominal coverages than do the AR-based predictive intervals. Further details of this example can be found in Section 5.

In Section 2 we describe the MTD model, and in Section 3 we obtain its stationarity and autocorrelation properties. In Section 4 we discuss estimation, in Section 5 we analyze some examples, and in Section 6 we outline several extensions.

2. THE MIXTURE TRANSITION DISTRIBUTION MODEL

Suppose that $\{Y_t: t \in N\}$ is a time series, a sequence of random variables taking values in an arbitrary space. Then $\{Y_t\}$ is generated by a *mixture transition distribution* (MTD) model if

$$F(y_t|\mathbf{y}^{t-1}) = \sum_{i=1}^{p} \alpha_i G_i(y_t|y_{t-i}). \tag{1}$$

In (1), $F(y_t|\mathbf{y}^{t-1})$ is the conditional cumulative distribution function (cdf) of Y_t given that $\mathbf{Y}^{t-1} = \mathbf{y}^{t-1}$, evaluated at y_t, where $\mathbf{Y}^s = (Y_1, \ldots, Y_s)^T$ and $\mathbf{y}^s = (y_1, \ldots, y_s)^T$. Also $\sum_{i=1}^{p} \alpha_i = 1, \alpha_i \geq 0, i = 1, \ldots, p$, and $G_i(\cdot|y)$ is a

Nhu D. Le is Research Scientist, Biometry Section, B.C. Cancer Agency, Vancouver B.C., Canada. Adrian E. Raftery is Professor of Statistics and Sociology, University of Washington, Seattle, WA 98195. R. Douglas Martin is Vice President of Technology, StatSci Division of MathSoft, and Professor of Statistics, University of Washington, Seattle, WA 98195. This research was supported by ONR contract N-00014-91-J-1074. Le's research was also partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors thank Don Percival, the editor, the associate editor, and two anonymous reviewers for their helpful comments.

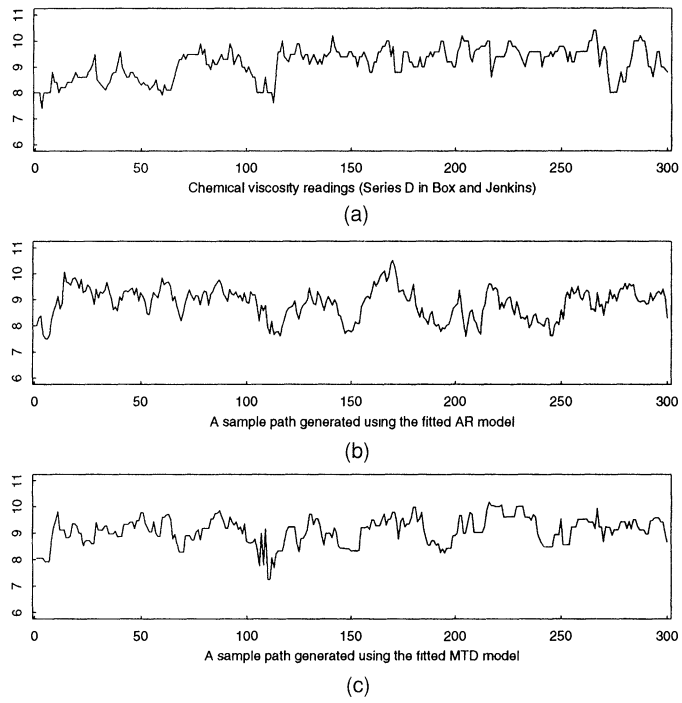

Figure 1. The Chemical Process Viscosity Readings and Simulated Sample Paths Generated Using the Fitted AR Model and the Fitted MTD Model. (a) Chemical viscosity readings (series D of Box and Jenkins); (b) a sample path generated using the fitted AR model; (c) a sample path generated using the fitted MTD model.

(conditional) cdf for each $i = 1, \ldots, p$ and for each value of y. $G_i(\cdot|\cdot)$ may be specified by a parameter $\boldsymbol{\theta}_i$.

The MTD model (1) was first introduced by Raftery (1985a,b) in the discrete case, as a model for high-order Markov chains. In that case it is far more parsimonious than the usual fully parameterized Markov chain, because it requires only one additional parameter for each extra lag. In the examples studied by Raftery (1985a), it fits the data very well. It is a discrete-valued analog to the usual AR(p) model in that the lagged bivariate distributions satisfy a system of matrix equations similar to the Yule–Walker equations, and in that the past values $Y_{t-1}, \ldots, Y_{t-p}$ do not interact in their effect on the conditional distribution of Y_t given the past. It is also physically motivated as a direct model for systems that tend to revert to previously occupied states. Martin and Raftery (1987) pointed out that the MTD model is defined for time series on arbitrary spaces and is not confined to the discrete-valued case.

Here we study the special case where the component conditional cdf's G_i are Gaussian, namely

$$G_i(y_t|y_{t-i}) = \Phi\left(\frac{y_t - \phi_i y_{t-i}}{\sigma_i}\right). \tag{2}$$

Although the conditional distribution of Y_t given the past is then a mixture of Gaussian distributions, we have found that the resulting model is able to model non-Gaussian sample path behavior such as outliers, bursts, and flat stretches, as we show by example in Section 5.

We have found it useful to extend the MTD model defined by (1) and (2) so that it includes the standard AR(p) model as a special case. This leads to

$$F(y_t|\mathbf{y}^{t-1}) = \alpha_0\phi\left(\frac{y_t - \sum_{j=1}^{p}\phi_{0j}y_{t-j}}{\sigma_0}\right) + \sum_{i=1}^{p}\alpha_i\phi\left(\frac{y_t - \phi_i y_{t-i}}{\sigma_i}\right). \tag{3}$$

We refer to (3) as the Gaussian MTD, or GMTD, model. A further generalization, obtained by adding a further independent component in the mixture, allows us to model independent pure replacement-type outliers (Martin and Yohai 1985, 1986) directly. The model is then of the following form:

$$F(y_t|\mathbf{y}^{t-1}) = \alpha_0\Phi\left(\frac{y_t - \sum_j \phi_{0j}y_{t-j}}{\sigma_0}\right) + \sum_{i=1}^{p}\alpha_i\Phi\left(\frac{y_t - \phi_i y_{t-i}}{\sigma_i}\right) + \alpha_{p+1}\Phi\left(\frac{y_t}{\sigma_{p+1}}\right), \tag{4}$$

where σ_{p+1} is large and $\sum_{i=0}^{p+1}\alpha_i = 1$.

The GMTD model (4) has $(4p+3)$ independent parameters, which is relatively small given the modeling flexibility that it allows. In the examples that we have worked with, we found a special case, the *random walk GMTD,* to be useful. This is defined by (3) with

$$\begin{cases} \phi_1 = \phi_2 = \cdots = \phi_p = 1 \\ \sum_{i=1}^{p}\phi_{0i} = 1. \end{cases} \tag{5}$$

This generalizes the usual random walk and has $(3p+2)$ independent parameters.

There are potential near-nonidentifiability problems with the full GMTD model (3). For example, the conditional expectation $E(Y_t|Y^{t-1})$ involves the sums $(\phi_{0i} + \phi_i)$, but does not involve either ϕ_{0i} or ϕ_i separately. Estimation of ϕ_{0i} and ϕ_i separately depends on the shape of the conditional distribution and not on its location, and thus may not be very precise. The sums $(\phi_{0i} + \phi_i)$ are likely to be much better estimated than either term individually. However, for the purposes of prediction and simulation, these problems do not present major difficulties. The identifiability problem does not arise with the random walk GMTD model (5).

The GMTD model (3) can accommodate the situation where the series is reasonably well approximated by an AR model and has occasional bursts and outliers. In this case the main AR component of the series would be captured by the first term of the model (3), and the additional components such as outliers and bursts would be captured by the other terms. For instance, occasional outliers may be captured by a term in the model with a large variance σ_i^2 and a small α_i; bursts can be accommodated with a larger α_i or with a temporal dependent latent process driving the mixing process. The flat stretches can be captured by a random walk GMTD model with a very small variance.

3. PROPERTIES

We now give the stationarity and autocorrelation properties of a times series $\{Y_t\}$ governed by the GMTD model (3). The proofs are in the Appendix.

3.1 Stationarity

Theorem 1. A necessary and sufficient condition for the process Y_t to be stationary in the mean is that the roots $z_1, \ldots, z_p$ of the equation

$$1 - \sum_{i=1}^{p} (\alpha_0\phi_{0i} + \alpha_i\phi_i)z^{-i} = 0$$

all lie inside the unit circle.

For the case where there are only two mixture components, without the full AR component (i.e., $\alpha_0 = 0$ and $p = 2$), this condition gives a region of stationarity defined by

$$\alpha_1\phi_1 + \alpha_2\phi_2 < 1,$$
$$-\alpha_1\phi_1 + \alpha_2\phi_2 < 1,$$

and

$$-\alpha_2\phi_2 < 1.$$

This region is the triangular region displayed in Figure 2. The defining equations for the region of stationarity in this case are quite similar to those corresponding to the ordinary AR(2) process.

These stationarity conditions are derived for the process with mean zero. However, there are several ways to modify the GMTD model so that it has a nonzero mean and the stationarity conditions remain the same. One simple way is to introduce a constant into the standard AR(p) term of Equation (3), which then becomes

$$F(y_t|\mathbf{y}^{t-1}) = \alpha_0\Phi\left(\frac{y_t - \delta - \sum_{j=1}^{p}\phi_{0j}y_{t-j}}{\sigma_0}\right) + \sum_{i=1}^{p}\alpha_i\Phi\left(\frac{y_t - \phi_i y_{t-i}}{\sigma_i}\right).$$

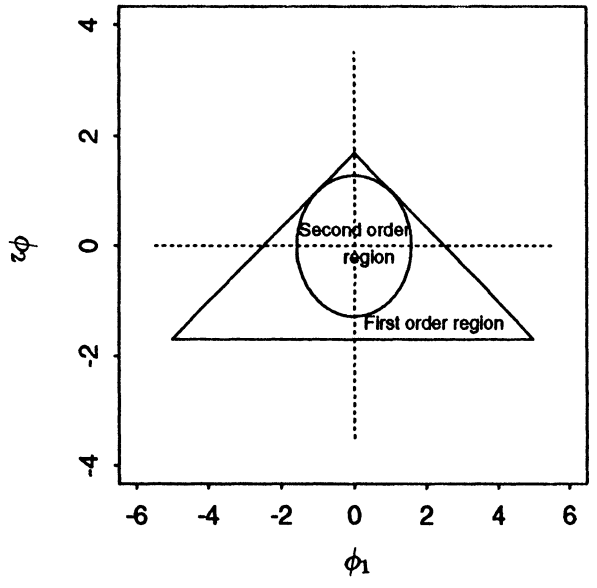

Figure 2. Stationarity Regions for Parameters of a MTD Model of Order 2.

Now suppose that the process Y_t is stationary in the mean. The conditions for second order and strict stationarity are quite difficult to obtain. However, when $\alpha_0 = 0$ (i.e., the process Y_t is defined by (1) and (2)), the conditions for second-order stationary can be derived as follows.

Theorem 2. Suppose that the process Y_t, defined by (1) and (2) with $\alpha_0 = 0$, is first-order stationary. A necessary and sufficient condition for the process to be second-order stationary is that the roots $z_1, \ldots, z_p$ of the equation

$$1 - \sum_{i=1}^{p}\alpha_i\phi_i^2 z^{-i} = 0$$

all lie inside the unit circle.

In the case where $p = 2$, this condition gives a region of stationarity defined by

$$\alpha_1\phi_1^2 + \alpha_2\phi_2^2 < 1.$$

This region is the ellipse displayed in Figure 2. Thus an extra condition is needed for the process to be second-order stationary. This is in contrast to the ordinary Gaussian AR model, which does not require any additional condition for second-order stationarity.

3.2 Autocorrelations

We now obtain the autocorrelations for the GMTD model (3). These autocorrelations satisfy a system of equations similar to the Yule–Walker equations. We then explicitly derive the range of autocorrelations for the case $\alpha_0 = 0$ and $p = 2$. This range is then compared with that of the ordinary AR(2) process.

Suppose that the process Y_t is second-order stationary and, without loss of generality, assume that its mean is zero. Let ρ_l be the lag-l autocorrelation. Then

$$\begin{aligned} E(Y_tY_{t-l}) &= E[E(Y_tY_{t-l}|Y^{t-1})] \\ &= E\left(Y_{t-l}\sum_{i=1}^{p}(\alpha_0\phi_{0i} + \alpha_i\phi_i)Y_{t-i}\right) \\ &= \sum_i (\alpha_0\phi_{0i} + \alpha_i\phi_i)E(Y_{t-l}Y_{t-i}). \end{aligned}$$

Because the process is second-order stationary, it follows that

$$\rho_l = \sum_i (\alpha_0\phi_{0i} + \alpha_i\phi_i)\rho_{|l-i|} \quad \text{for} \quad l = 1, \ldots, p. \tag{6}$$

These equations are similar to the Yule–Walker equations for the ordinary AR(p) process, where the coefficient

$\alpha_0\phi_{0\imath} + \alpha_\imath\phi_\imath$ is replaced by the lag-i coefficient of the AR(p) process.

In the case $\alpha_0 = 0$ and $p = 2$, the equations (6) for $l = 1, 2$ become

$$\rho_1 = \alpha_1\phi_1 + \alpha_2\phi_2\rho_1 \tag{7}$$

and

$$\rho_2 = \alpha_1\phi_1\rho_1 + \alpha_2\phi_2. \tag{8}$$

The admissible region for the autocorrelations ρ_1 and ρ_2 is defined by

$$\alpha_1^2\rho_1^4 + (-2\alpha_1^2 + \alpha_1 + 1)\rho_1^2 - 2\rho_2\rho_1^2 + (1 - \alpha_1)\rho_2^2\rho_1^2 + \alpha_1\rho_2^2 - \alpha_1(1 - \alpha_1) < 0.$$

The boundary of this region satisfies the equation

$$\rho_2 = \frac{\rho_1^2 \pm \sqrt{\alpha_1^2(1 - \alpha_1)(1 - \rho_1^2)^3}}{(1 - \alpha_1)\rho_1^2 + \alpha_1}, \tag{9}$$

with $0 \le \alpha_1 \le 1$ and $-1 < \rho_1 < 1$. The derivation of the admissible region and its boundary defined by (9) are given in the Appendix.

The plot of ρ_1 against ρ_2, which represents the range of autocorrelations of the restricted GMTD model with $p = 2$ and $\alpha_0 = 0$, defined by (1) and (2), is displayed in Figure 3 and compared to the range of autocorrelations of the ordinary stationary AR(2) process. This shows that even when the full AR(2) component is not included, the range of autocorrelations allowed by the MTD model is almost as great as for the standard AR(2) process. Of course, for the full GMTD model (3), the range of possible autocorrelations is as great as that of the standard AR process.

4. ESTIMATION

4.1 An EM Estimation Algorithm

In this section we use the EM algorithm (Baum 1971; Dempster, Laird, and Rubin 1977; Good 1956; Hartley 1958) to estimate the parameters of the GMTD model (3). This provides a framework for estimating parameters in a mixture model that can be applied to the present problem as follows. Suppose that the observations $\mathbf{Y} = (y_1, \ldots, y_n)$ are generated from the GMTD model (3). Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ be the unobserved random variable, where $\mathbf{Z}_t$ is a $(p + 1)$-dimensional vector with component j equal to 1 if y_t comes from the jth component of the conditional cdf, and to zero otherwise. Then the distribution of $\mathbf{Z}_t \equiv (z_{t,0}, z_{t,1}, \ldots, z_{t,p})$ is

$$P[\mathbf{Z}_t = (1, 0, 0, \ldots, 0)] = \alpha_0,$$

$$P[\mathbf{Z}_t = (0, 1, 0, \ldots, 0)] = \alpha_1,$$

$$\vdots$$

$$P[\mathbf{Z}_t = (0, 0, 0, \ldots, 1)] = \alpha_p.$$

Let

$$\begin{aligned} \boldsymbol{\alpha} &= (\alpha_0, \ldots, \alpha_p), \\ \boldsymbol{\sigma} &= (\sigma_0, \ldots, \sigma_p), \\ \boldsymbol{\phi} &= (\phi_1, \ldots, \phi_p), \\ \boldsymbol{\phi}_0 &= (\phi_{01}, \ldots, \phi_{0p}), \end{aligned}$$

and

$$\mathbf{y}_p^{t-1} = (y_{t-1}, \ldots, y_{t-p})^T$$

and set

$$V(\boldsymbol{\phi}, \boldsymbol{\phi}_0, \boldsymbol{\sigma}) = \left(\log f\left(\frac{y_t - \boldsymbol{\phi}_0 \mathbf{y}_p^{t-1}}{\sigma_0} \right), \log f\left(\frac{y_t - \phi_1 y_{t-1}}{\sigma_1} \right), \ldots, \log f\left(\frac{y_t - \phi_p y_{t-p}}{\sigma_p} \right) \right)^T,$$

$$U(\boldsymbol{\alpha}) = (\log \alpha_0, \log \alpha_1, \ldots, \log \alpha_p)^T,$$

and

$$W(\boldsymbol{\sigma}) = (\log \sigma_0, \log \sigma_1, \ldots, \log \sigma_p)^T,$$

where f is the standard Gaussian density. Because Y and Z are independent and the $Z_\imath$ are independent of each other, the (conditional) log-likelihood for the complete data $X = (Y, Z)$ is

$$L(X, \boldsymbol{\phi}, \boldsymbol{\phi}_0, \boldsymbol{\sigma}, \boldsymbol{\alpha}) = \sum_{\imath=p+1}^{n} \mathbf{Z}_\imath V(\boldsymbol{\phi}, \boldsymbol{\phi}_0, \boldsymbol{\sigma}) + \sum_{\imath=p+1}^{n} \mathbf{Z}_\imath U(\boldsymbol{\alpha}) - \sum_{\imath=p+1}^{n} \mathbf{Z}_\imath W(\boldsymbol{\sigma}).$$

Here the log-likelihood function is evaluated conditionally on the first p observations. To ensure its comparability when

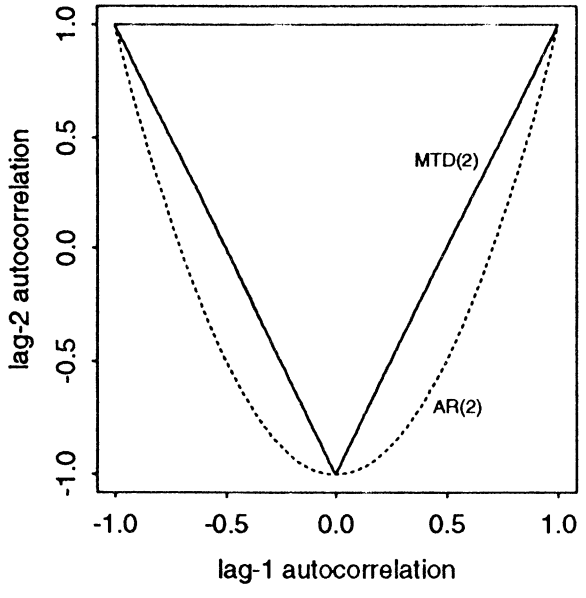

Figure 3. Range of Possible Autocorrelations for the Second-Order MTD Process Defined by Equations (1) and (2) With p = 2.

comparing different orders, we use the largest order for p. The iterative EM procedure for estimating the parameters by maximizing the log-likelihood function contains the following two steps:

E step: Suppose that $\boldsymbol{\phi}, \boldsymbol{\phi}_0, \boldsymbol{\sigma}$, and $\boldsymbol{\alpha}$ are known. The missing data $\mathbf{Z}$ are then replaced by their conditional expectations, conditioned on the parameters and on the observed data Y. In this case the conditional expectation of the jth component of $\mathbf{Z}_t$ is just the conditional probability that the observation y_t comes from the jth component of the mixture distribution, conditioned on the parameters and $\mathbf{Y}$. Denote the conditional expectation of the jth component of $\mathbf{Z}_t$ by $\tilde{z}_{t,j}$. Then the E step equations are as follows:

$$\tilde{z}_{t,j} = \frac{\alpha_j h_j(y_t)}{\alpha_0 h_0(y_t) + \sum_{l=1}^{p} \alpha_l h_l(y_t)} \qquad j = 1, \ldots, p,$$

and

$$\tilde{z}_{t,0} = \frac{\alpha_0 h_0(y_t)}{\alpha_0 h_0(y_t) + \sum_{l=1}^{p} \alpha_l h_l(y_t)},$$

where $h_j(y_t) = (1/\sigma_j) f((y_t - \phi_j y_{t-j})/\sigma_j), j = 1, \ldots, p$, and $h_0(y_t) = (1/\sigma_0) f((y_t - \sum_{l=1}^{p} \phi_{0l} y_{t-l})/\sigma_0)$.

M step: Suppose that the missing data are known. The estimates of the parameters $\boldsymbol{\phi}, \boldsymbol{\phi}_0, \boldsymbol{\sigma}$, and $\boldsymbol{\alpha}$ can then be obtained by maximizing the log-likelihood function L. The estimated parameters, $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}}_0, \hat{\boldsymbol{\sigma}}$, and $\hat{\boldsymbol{\alpha}}$, must satisfy the following M step equations:

$$\left.\frac{\partial L}{\partial \boldsymbol{\phi}}\right|_{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}}_0, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\alpha}}} = 0,$$

$$\left.\frac{\partial L}{\partial \boldsymbol{\phi}_0}\right|_{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}}_0, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\alpha}}} = 0,$$

$$\left.\frac{\partial L}{\partial \boldsymbol{\alpha}}\right|_{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}}_0, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\alpha}}} = 0,$$

and

$$\left.\frac{\partial L}{\partial \boldsymbol{\sigma}}\right|_{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}}_0, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\alpha}}} = 0.$$

The estimates of the parameters are then obtained by iterating these two steps until convergence. Wu (1983) has studied the convergence of the algorithm.

Given the model order, the information matrix can be obtained by numerical differentiation or by using the ECM algorithm (Meng and Rubin 1993), and used to obtain standard errors. We use the numerical method described by Tanner (1993) to obtain the standard errors of the estimates in the examples.

4.2 A Simulation Study

We now report the results of a simulation study designed to assess the performance of the EM estimation method. The simulation study is carried out with $p = 2$. There are 100 sample paths, each with 200 data points, generated by the GMTD model (3). The parameters of the model are chosen as follows:

$$\begin{aligned} \boldsymbol{\alpha} &= (\alpha_0, \alpha_1, \alpha_2) = (.4, .3, .3), \\ \boldsymbol{\sigma} &= (\sigma_0, \sigma_1, \sigma_2) = (1, 1, 5), \\ \boldsymbol{\phi} &= (\phi_1, \phi_2) = (-.7, .8), \end{aligned}$$

and

$$\boldsymbol{\phi}_0 = (\phi_{01}, \phi_{02}) = (.9, -.6).$$

For each sample path, the EM algorithm is used to estimate the parameters. In this case the E step equations remain the same as before and the M step equations become

$$\hat{\phi}_j = \frac{\sum_{t=p+1}^{n} \tilde{z}_{tj} y_t y_{t-j}}{\sum_t \tilde{z}_{tj} y_{t-j}^2}, \qquad j = 1, \ldots, p,$$

$$\hat{\alpha}_j = \frac{\sum_t \tilde{z}_{tj}}{\sum_t \sum_l \tilde{z}_{tl}}, \qquad j = 0, \ldots, p,$$

$$\hat{\sigma}_j = \left(\frac{\sum_t \tilde{z}_{tj} (y_t - \hat{\phi}_j y_{t-j})^2}{\sum_t \tilde{z}_{tj}} \right)^{1/2},$$

and

$$\hat{\sigma}_0 = \left(\frac{\sum_t \tilde{z}_{t0} (y_t - \hat{\boldsymbol{\phi}}_0 \mathbf{y}_p^{t-1})^2}{\sum_t \tilde{z}_{t0}} \right)^{1/2},$$

where the components of $\hat{\boldsymbol{\phi}}_0$ satisfy the following system of equations:

$$\sum_t \tilde{z}_{t0} y_t y_{t-i} = \sum_{j=1}^{p} \hat{\phi}_{0j} \sum_t \tilde{z}_{t0} y_{t-j} y_{t-i}, \qquad i = 1, \ldots, p.$$

The sample means and standard errors of the EM parameter estimates are shown in Table 1. The results indicate that the estimation method works well. The sample means are very close to the true values, and the sample standard errors are small.

5. EXAMPLES

5.1 Example 1: IBM Stock Prices

Figure 4a shows the daily IBM common stock closing price from May 17, 1961, to November 2, 1962 (369 observations); this is series B of Box and Jenkins (1976; hereafter

Table 1. Empirical Means and Standard Deviations (SD) of the EM Parameter Estimates for the Simulated Datasets

α	Mean (SD)	σ	Mean (SD)	ϕ	Mean (SD)	ϕ_0	Mean (SD)
.4	.40 (.05)	1	.98 (.13)	−.7	−.70 (.03)	.9	.90 (.02)
.3	.30 (.05)	1	.99 (.16)	.8	.79 (.04)	−.6	−.60 (.02)
.3	.30 (.05)	5	4.90 (.53)				

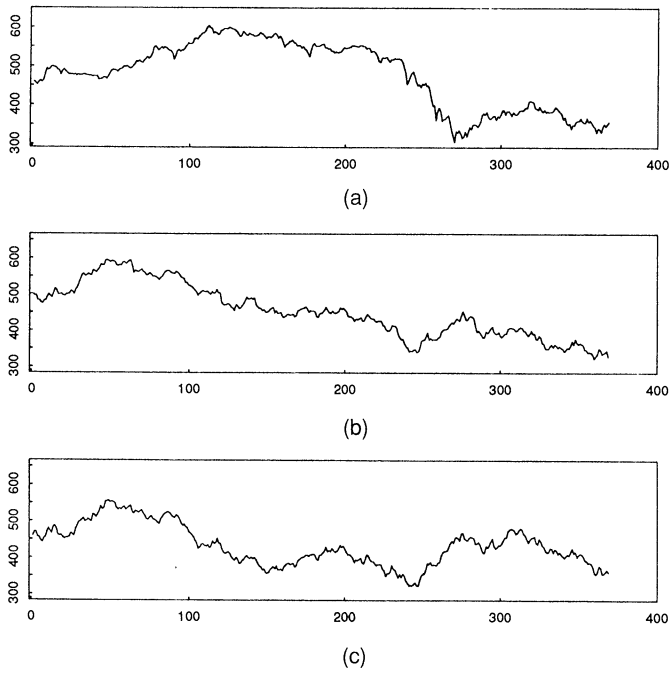

Figure 4. IBM Common Stock Daily Closing Price From May 17, 1961 (Day 1) to November 2, 1962 (Day 369). This is series B of Box and Jenkins (1976). (a) IBM daily common stock price (series B in BJ); (b) a sample path generated using the fitted MTD model; (c) a sample path generated using the fitted ARIMA model.

denoted by BJ). BJ identified an autoregressive integrated moving average model, ARIMA(0, 1, 1), as the best model for this series; namely,

$$y_t = y_{t-1} + \varepsilon_t - .09\varepsilon_{t-1},$$

where the $\{\varepsilon_t\}$ are independent $N(0, \sigma_\varepsilon^2)$ random variables with $\sigma_\varepsilon^2 = 52.2$. However, they found some evidence of model inadequacy, which could be due to a change in the moving average parameter (Box and Jenkins 1976), to variance shifts (Wichern, Miller, and Hsu 1976), or to nonlinearity (Tong 1990). Thus models other than the linear ARIMA model might perform well for these data.

To compare the rival, nonnested, models for this series and for those in other examples, we use the Bayes information criterion (BIC),

$$\text{BIC} = 2\log(\text{maximized likelihood}) - k\log(n),$$

where k is the number of independent parameters estimated. This was first derived by Schwarz (1978) in the case of independent observations; his results imply that in large samples, the BIC is approximately equal to minus twice the logarithm of the Bayes factor against the model of interest in favor of an appropriate baseline model. Thus the criterion favors models with large BIC values. In these examples the comparisons between the best ARIMA and MTD models are decisive and are not sensitive to the particular model-comparison procedure adopted. For example, the same conclusions would be reached using significance tests at the 5% or 1% levels, or using the Akaike information criterion (AIC) (Akaike 1973).

Table 2. Model Comparison for the IBM Stock Price Series

Model	*k*	*2 log-likelihood*	*BIC*
MTD	6	−2,431	−2,466
ARIMA	2	−2,494	−2,506
Difference	4	63	40

The best MTD model according to the BIC is the second-order random walk GMTD, defined by (5); namely,

$$F(y_t|\mathbf{y}^{t-1}) = \alpha_0\Phi\left(\frac{y_t - \phi y_{t-1} - (1-\phi)y_{t-2}}{\sigma_0}\right) + \alpha_1\Phi\left(\frac{y_t - y_{t-1}}{\sigma_1}\right) + \alpha_2\Phi\left(\frac{y_t - y_{t-2}}{\sigma_2}\right).$$

The estimated parameters are $\phi = 1.94$ with standard error $\text{SE}_\phi, = .28$,

$$(\alpha_0, \alpha_1, \alpha_2) = (.24, .69, .07),$$

and

$$(\sigma_0, \sigma_1, \sigma_2) = (6.38, 5.03, 11.23),$$

with

$$(\text{SE}_{\sigma_0}, \text{SE}_{\sigma_1}, \text{SE}_{\sigma_2}) = (1.06, .77, 3.41).$$

A comparison with the ARIMA(0, 1, 1) favored by BJ is shown in Table 2; the MTD model is strongly preferred.

A sample path generated using the fitted MTD model is displayed in Figure 4b. For comparison, we also generated a sample path for the fitted ARIMA model with the same sequence of observations for the white noise process; this path is displayed in Figure 4c. The plots indicate that both models are capable of producing sample paths that have qualitative characteristics similar to those of the original series. Both models have their estimated parameters near the boundary of stationarity.

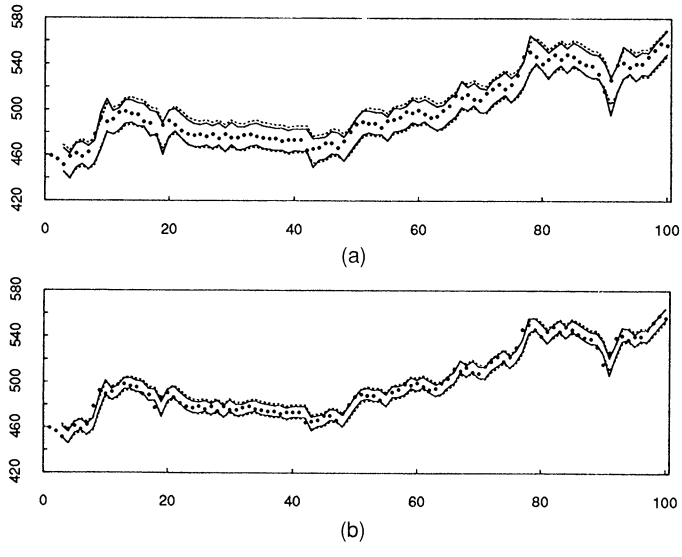

Figure 5. 90% (a) and 60% (b) Predictive Intervals for the IBM Stock Price Series. The dots denote the original observations. The dashed lines are for the ARIMA-based predictive intervals and the solid lines are for the GMTD-based predictive intervals.

Table 3. Mean Squared Widths of the $(1-\alpha)\%$ PI's for the IBM Stock Price Series

$(1-\alpha)\%$	.90	.80	.70	.60	.50
MTD	519.1	304.1	184.1	117.6	69.5
ARIMA	569.1	346.2	225.8	148.9	95.6

The predictive intervals (PI's) for the series are constructed with one-step ahead prediction using both the fitted ARIMA and MTD models. The 90% and 60% PI's are shown in Figure 5.

For easy visual examination, only the first 100 points are plotted. The plots show that the MTD model produces smaller PI's than those produced by the ARIMA model used by BJ. To examine this point further, we show the mean squared widths of the PI's in Table 3. The results confirm that the ARIMA-based PI's are longer on average than the corresponding GMTD-based PI's for each nominal coverage. The empirical coverages of these PI's (namely, the percentages of the data that fall within them) are displayed in Table 4. The result shows that the PI's constructed using the GMTD model are better than those constructed using the ARIMA model in the sense that the GMTD-based predictive intervals have empirical coverages closer to the nominal coverages than the ARIMA-based predictive intervals, especially for the lower nominal coverages.

The normal probability plots of the MTD-based and ARIMA-based prediction residuals are shown in Figure 6. Both plots indicate that the prediction residuals have distributions with tails heavier than that of the Gaussian distribution. For this reason, the ARIMA-based PI's, which assume normality for the prediction residuals, are longer than they are supposed to be and consequently overestimate the true coverage. This mild, but important, non-Gaussianity is better captured by the GMTD model, explaining its superiority for this series. Box, Jenkins, and Reinsel (1994, pp. 317–318) also considered the possibility that the inadequacy of the ARIMA(0, 1, 1) might be accounted for in part by a change in variance.

5.2 Example 2: Chemical Process Viscosity Series

Figure 1a shows series D from BJ, a series of consecutive hourly viscosity readings from a chemical process. This series exhibits frequent flat stretches, and it is clear from Figures 1b and 1c that these are captured better by the MTD model than by the preferred ARIMA model of BJ; namely, the AR(1) model

$$y_t = .87y_{t-1} + 1.17 + \varepsilon_t,$$

Table 4. Empirical Coverages of the $(1-\alpha)\%$ PI's for the IBM Stock Price Series

$(1-\alpha)\%$	.90	.80	.70	.60	.50
MTD	.91	.81	.71	.62	.51
ARIMA	.90	.84	.78	.73	.56

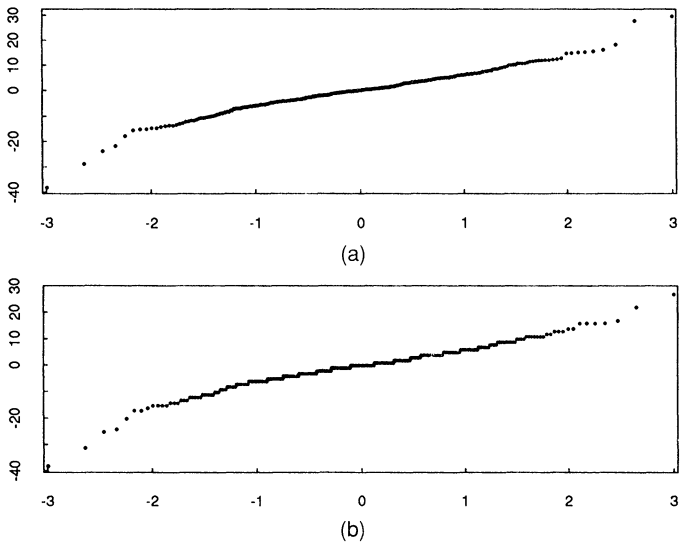

Figure 6. Normal Probability Plots for the MTD-Based (a) and ARIMA-Based (b) Prediction Residuals for the IBM Stock Price Series.

where $\sigma_\varepsilon^2 = .09$.

The preferred MTD model according to BIC is the second-order random walk GMTD model defined by (5). This is the same model that was selected in Example 1. The estimated parameters are $\phi = 1.06$ with $\mathrm{SE}_\phi = .04$,

$$(\alpha_0, \alpha_1, \alpha_2) = (.63, .25, .12),$$

and

$$(\sigma_0, \sigma_1, \sigma_2) = (.1195, .0014, .0989),$$

with

$$(\mathrm{SE}_{\sigma_0}, \mathrm{SE}_{\sigma_1}, \mathrm{SE}_{\sigma_2}) = (.002, .0004, .005).$$

A comparison with the AR(1) model favored by BJ is presented in Table 5. Again, the preference for the MTD model is strong.

The 90% and 60% PI's for the first 100 observations are shown in Figure 7. The 90% PI's produced by the MTD model are quite similar to those produced by the AR model, but the 60% GMTD-based PI's are smaller than the corresponding AR-based PI's. Table 6 shows that although the AR-based 90% PI's are slightly narrower on average than the corresponding MTD-based PI's, the other AR-based PI's are wider than the corresponding MTD-based PI's. The empirical coverages of these PI's are displayed in Table 7. The results indicate that both models produce similar results for the 70%–90% PI's. Although the 50% and 60% GMTD-based PI's underestimate the coverages slightly, they are quite close to the nominal coverages. On the other hand,

Table 5. Model Comparison for the Chemical Process Viscosity Series

Model	k	2 log-likelihood	BIC
MTD	7	−84	−124
AR	3	−134	−151
Difference	4	50	27

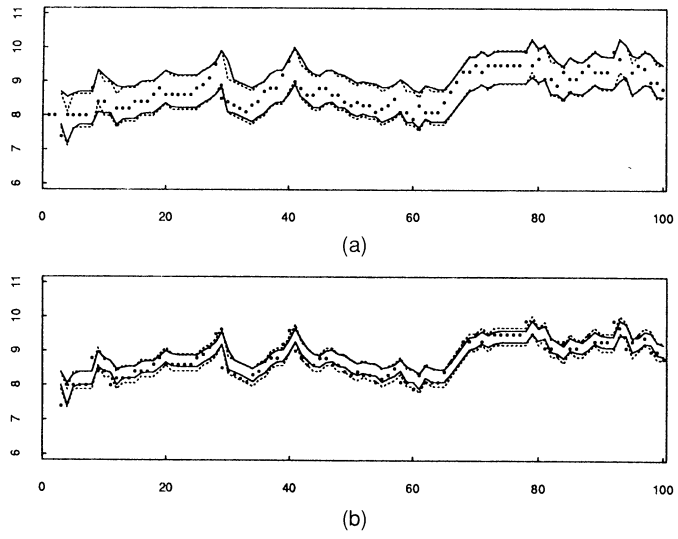

Figure 7. 90% (a) and 60% (b) Predictive Intervals for the Chemical Process Viscosity Series. The dots denote the original observations. The dashed lines are for the AR-based predictive intervals, and the solid lines are for the GMTD-based predictive intervals.

the corresponding AR-based PI's overestimate the nominal coverages by a much bigger margin. Overall, the MTD model is preferred for these data.

5.3 Example 3: A Simulated Series

We now consider a simulated series that exhibits bursts of activity. The series consisting of 300 points displayed in Figure 8a is generated using model (3) with $\alpha_0 = 0, p = 3$ where

$$(\phi_1, \phi_2, \phi_3) = (.3, .3, -2.5),$$
$$(\alpha_1, \alpha_2, \alpha_3) = (.4, .4, .2),$$

and

$$(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 5).$$

According to the BIC, the best MTD model is the third-order MTD model with estimated parameters

$$(\phi_1, \phi_2, \phi_3) = (.3, .3, -2.38),$$
$$(\alpha_1, \alpha_2, \alpha_3) = (.37, .42, .21),$$

and

$$(\sigma_1, \sigma_2, \sigma_3) = (.89, .95, 4.75),$$

with $(\mathrm{SE}_{\phi_1}, \mathrm{SE}_{\phi_2}, \mathrm{SE}_{\phi_3}) = (.008, .006, .06)$ and $(\mathrm{SE}_{\sigma_1}, \mathrm{SE}_{\sigma_2}, \mathrm{SE}_{\sigma_3}) = (.12, .11, .46)$. The best AR model for this series is of order 5. Table 8 indicates that the MTD model is strongly preferred, as one would expect.

We also generated a sample path for the fitted AR model, which is shown in Figure 8b. This shows that the simulated AR sample path does not have the main characteristic of the series being analyzed—bursts of activity. This comparison is carried out purely to show that the AR model is not adequate in this situation, but in fairness it must be remembered that the AR model is not designed to perform well in such circumstances.

Table 6. Mean Squared Widths of the $(1 - \alpha)$% PI's for the Chemical Process Viscosity Series

$(1-\alpha)$%	.90	.80	.70	.60	.50
MTD	1.03	.54	.28	.14	.06
AR	.97	.59	.39	.26	.16

Table 7. Empirical Coverages of the $(1 - \alpha)$% PI's for the Chemical Process Viscosity Series

$(1-\alpha)$%	.90	.80	.70	.60	.50
MTD	.91	.80	.67	.56	.48
AR	.91	.79	.73	.68	.59

Tables 9 and 10 show that the MTD-based PI's are much better than the corresponding AR-based PI's, and the GMTD-based PI's are much narrower. Also, the GMTD-based PI's have very good empirical coverages, whereas the AR-based PI's overestimate the nominal coverages very badly.

6. EXTENSIONS AND GENERALIZATIONS

The general MTD model (1) can be used to model quite general non-Gaussian time series, even when it takes values in a non-Euclidean space. The model was first introduced for discrete data by Raftery (1985a) and was studied further by Adke and Deshmukh (1988), Kwok (1989), Li and Kwok (1989), Raftery (1985b), and Raftery and Taveré (1994). It is also appropriate for sequences of counts and success proportions; models for the Poisson and binomial case were proposed by Raftery (1985b).

The MTD model may also be used to represent positive-valued time series; for example, exponential, gamma, or Weibull. This model seems to avoid the problems associated with previous efforts in this direction, such as those of Gaver and Lewis (1980) and Lawrence and Lewis (1985) (see Raftery 1985c).

The models can also fit time series with both discrete

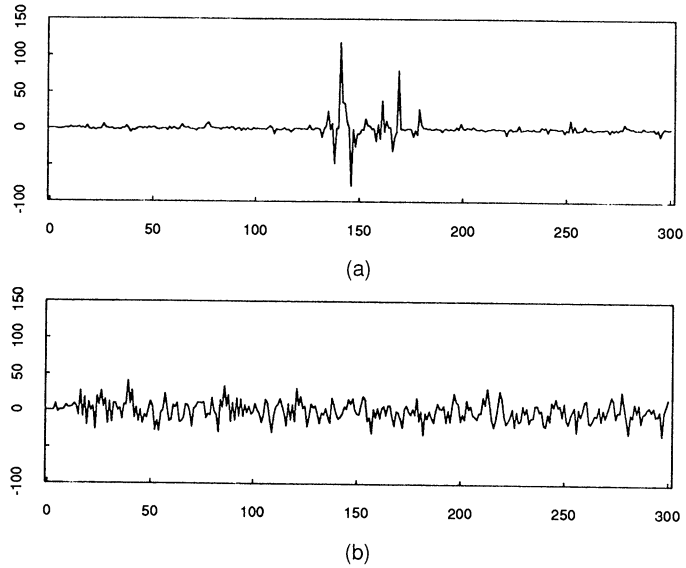

Figure 8. Simulated Series With Bursts (Example 3) (a), and a Simulated AR(5) Sample Path (b).

Table 8. Model Comparison for Example 3

Model	k	2 log-likelihood	BIC
GMTD	8	−1,239	−1,284
AR	5	−2,198	−2,226
Difference	3	959	942

and continuous components in a natural way. One important application of this is rainfall series (Stern and Coe 1984). It lends itself naturally to the analysis of time series of angles and has been used successfully by Craig (1989) and Raftery and Tavaré (1994) to model time series of wind directions. It could also be used to model time series of continuous proportions, taking values in the simplex. (For a review of the analysis of independent data of this kind, see Aitchison 1986.)

Another extension consists of allowing the random variables $\{\mathbf{Z}_t\}$ defined in Section 4.1 to follow a Markov chain rather than to be independent as in the present formulation. The variable $\mathbf{Z}_t$ determines which component of the conditional distribution of Y_t is chosen at time t. When this is a Markov chain, $\{Y_t\}$ is a hidden Markov process and so belongs to a class of models that have proven useful in research on speech recognition and image reconstruction (Levinson, Rabiner, and Sondhi 1983). The parameter estimates can be obtained in this case using the *forward–backward* algorithm of Baum (1971) (see Le, Leroux, and Puterman 1992).

It may be possible to generalize the model further by allowing the $G_i(\cdot|\cdot)$ in Equation (1) to be nonparametric and to estimate them from the data. When both the α_i and the G_i are unknown, this is a semiparametric estimation problem (Bickel, Klaasen, Rartov, and Wellner 1993; Wellner 1985). It may be possible to avoid the apparent nonidentifiability problem by using an iterative estimation scheme, switching between estimation of the α_i and the G_i.

Most of our discussion has related to short-memory stationary time series. However, the general MTD framework also may be used to model both long-memory stationary time series and nonstationary time series. Long-memory stationary time series are characterized by the fact that the variance of the sample mean declines more slowly than $O(1/n)$, that the autocorrelation function declines asymptotically much more slowly than in the short-memory case (hyperbolically rather than exponentially), and that the spectrum is unbounded at zero. The long-memory property can be captured by an *inifinite-lag* MTD model, defined by (1) with $p = \infty$, where the α_i and the G_i are suitably parameterized. One possible choice for the α_i is

$$\alpha_i = \frac{d(1-d)\cdots(i-1-d)}{i!}. \tag{10}$$

Table 9. Mean Squared Widths of the $(1-\alpha)$% PI's for Example 3

$(1-\alpha)$%	.90	.80	.70	.60	.50
MTD	350	298	240	198	23
AR	1,084	658	430	284	182

Table 10. The Empirical Coverages of the $(1-\alpha)$% PI's for Example 3

$(1-\alpha)$%	.90	.80	.70	.60	.50
MTD	.89	.77	.70	.61	.49
AR	.95	.94	.92	.91	.88

These are the same as the π weights for the fractionally differenced ARIMA$(0, d, 0)$ process (Granger and Joyeux 1980; Hosking 1981), and so this can be regarded as a non-Gaussian generalization of the fractional differencing model. This may be generalized to allow for both long-memory and short-memory dependence by setting the α_i equal to the π weights for the fractionally differenced ARIMA(p, d, q) process. Mehran (1989) has also discussed infinite-lag MTD models, although in a different context and only for discrete-valued time series.

A natural way of modeling nonstationary time series is to use state-space models. Kitagawa (1987) has described quite a general framework for non-Gaussian state-space modeling. This is based on the idea of an unobserved *state* of the process, x_t, which may be vector valued, and evolves in a Markovian fashion according to a *state equation* that specifies the transition density $p(x_t|x_{t-1})$. Conditionally on the state x_t, the observations are independent with a conditional density $p(y_t|x_t)$, specified by the *observation equation.* Kitagawa (1987) showed how the resulting filtering, updating, and smoothing equations may be solved by numerical integration; this becomes extremely demanding computationally as the dimension of the state increases. Martin and Raftery (1987) showed how the problem can be simplified if the state equation has a form similar to the MTD equation (1).

7. DISCUSSION

We have discussed a class of time series models, the MTD models, that can capture many non-Gaussian features, including flat stretches, bursts, and outlier stretches, as well as much more general non-Gaussian behavior, in a single unified model framework. Raftery (1994) has shown that they can also be used to represent changepoint-like behavior.

Other models that can capture some of these features have also been proposed. For example, bilinear time series models (Subba Rao 1984) can represent bursts quite well. Threshold models (Tong 1990) can deal with flat stretches. The literature on Bayesian forecasting and recursive updating provides ways of dealing with some of these features that also involve mixtures, although they arise in a quite different way (Alspach and Sorenson 1972; Denby and Martin 1979; Harrison and Stevens 1976; Martin 1979; Martin and Yohai 1985; Smith and West 1983; Taplin and Raftery 1994; West 1986). A useful aspect of the MTD model class is that it can represent a wide range of such behaviors in the context of the same model.

The conditional mean predictor for the MTD model (1) is

$$E[Y_t|\mathbf{Y}^{t-1} = y^{t-1}] = \sum_{i=1}^{p} \alpha_i g_i(y_{t-i}),$$

where $g_i(y) = \int x\, dG_i(x|y)$. This is an ACE-type predictor for time series (Owen 1983); see Breiman and Friedman (1985) for ACE modeling in regression. This conditional expectation is the same as that for the generalized additive model (Hastie and Tibshirani 1986), but here the error distribution is a mixture of conditional distributions. So generalized additive models are quite different from what we propose here.

APPENDIX A: PROOFS OF THEOREMS IN SECTION 3

The theorems in Section 3 are proven by using a result due to Beněs (1967). Roughly, Beněs showed that for a Markov process Y_t to have a finite invariant measure, it is necessary and sufficient that there exist a moment of Y_t, say of order k, that is finite for all t. The existence of a finite invariant measure would ensure the stationarity of the process, assuming that the process started somewhere very far in the past or started with the initial distribution corresponding to the invariant measure. That is, the process is strictly stationary with a finite kth order moment. First, we state the results obtained by Beněs (1967) in the following theorem.

Theorem 3 (Beněs 1967). Let y_t be a Markov process on a locally compact metric space $(\mathcal{E}, \rho)$ with a countable base. Let $\mathcal{B}$ be the σ algebra generated by the open sets, and for $A \in \mathcal{B}, t > 0$, set

$$P(t, y, A) = \Pr\{y_t \in A | y_0 = y\}.$$

Assume that for fixed t and A, $P(t, y, A)$ is continuous in y, that for an ε neighborhood $S_\varepsilon(y)$ about y, $P(t, y, S_\varepsilon(y)) \to 1$ as $t \to 0$, and that $P(t, y, \mathcal{E}) \to 1$. Let C^+ be the strictly positive cone of the Banach space $ca(\mathcal{E}, \mathcal{B})$ consisting of the countably additive set functions on $\mathcal{B}$ with variation norm. Let $U_t, t \geq 0$ be the semigroup defined for a measure μ by the formula

$$U_t\mu(A) = \int_{\mathcal{E}} P(t, y, A)\mu\,(dy),$$

and let (f, μ) denote the action of a linear functional on μ. A $\mathcal{B}$-measurable function f on $\mathcal{E}$ is a "moment" iff $f \geq 0$ and $\inf_{y \in \mathcal{E} - K_n} f(y) \to \infty$ as some sequence of compact set $K_n \uparrow \mathcal{E}$. Then the following conditions are equivalent:

- y_t has a positive finite invariant measure.
- there exists $\mu \in C^+$ and there exists moment g, with $\sup_{t \geq 0}(g, U_t\mu) < \infty$.

Using this result, the proofs of the theorems go as follows. The notation used here is the same as in Section 3.

Proof of Theorem 1

Let μ_t be the mean of Y_t. Then by the model (3),

$$\begin{aligned}\mu_t &= E(Y_t) = EE(Y_t|Y_{t-1}, \ldots, Y_{t-p}) \\ &= E\sum_{i=1}^{p} (\alpha_0\phi_{0i} + \alpha_i\phi_i)Y_{t-i} = \sum_{i=1}^{p} (\alpha_0\phi_{0i} + \alpha_i\phi_i)\mu_{t-i}.\end{aligned}$$

The necessary and sufficient condition for this homogeneous difference equation to have a stable solution, which is finite and independent of t, is that the roots $z_1, \ldots, z_p$ of the equation

$$1 - \sum_{i=1}^{p} (\alpha_0\phi_{0i} + \alpha_i\phi_i)z^{-i} = 0$$

all lie inside the unit circle (Goldberg 1958). Let μ be the measure corresponding to the distribution of the process at time zero (i.e., finite mixture of Gaussian distributions), and hence $\mu \in C^+$. Thus by the Beněs result, there exists a finite invariant measure for the process Y_t.

Proof of Theorem 2

Let $g_t(y_t)$ be the marginal density of Y_t. By (1) and (2), it can be written as

$$g_t(y_t) = \sum_{i=1}^{p} \alpha_i \int h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i},$$

where $h_i(y_t) = (1/\sigma_i) f((y_t - \phi_i y_{t-i})/\sigma_i)$. Thus

$$\begin{aligned}E(Y_t)^2 &= \int (y_t)^2 \sum_{i=1}^{p} \alpha_i \int h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i}\, dy_t \\ &= \sum_{i=1}^{p} \alpha_i \int\int (y_t - \phi_i y_{t-i} + \phi_i y_{t-i})^2 \\ &\quad \times h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i}\, dy_t \\ &= \sum_{i=1}^{p} \alpha_i \int\int (y_t - \phi_i y_{t-i})^2 h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i}\, dy_t \\ &\quad + \sum_{i=1}^{p} \alpha_i \int\int (\phi_i y_{t-i})^2 h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i}\, dy_t.\end{aligned}$$

The first term on the right side is then

$$\begin{aligned}&= \sum_{i=1}^{p} \alpha_i \sigma_i^2 \int\int \left(\frac{y_t - \phi_i y_{t-i}}{\sigma_i}\right)^2 h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i}\, dy_t \\ &= \sum_{i=1}^{p} \alpha_i \sigma_i^2 = C,\end{aligned}$$

which does not depend on t. The second term is

$$\begin{aligned}&= \sum_{i=1}^{p} \alpha_i (\phi_i)^2 \int\int (y_{t-i})^2 h_i(y_t) g_{t-i}(y_{t-i})\, dy_{t-i}\, dy_t \\ &= \sum_{i=1}^{p} \alpha_i (\phi_i)^2 E(Y_{t-i})^2.\end{aligned}$$

Thus

$$E(Y_t)^2 = \sum_{i=1}^{p} \alpha_i (\phi_i)^2 E(Y_{t-i})^2 + C.$$

The necessary and sufficient condition for a nonhomogeneous difference equation to have a stable solution that is finite and independent of t is that the roots $z_1, \ldots, z_p$ of the equation

$$1 - \sum_{i=1}^{p} \alpha_i (\phi_i)^2 z^{-i} = 0$$

all lie inside the unit circle (Goldberg 1958). Let μ be the measure corresponding to the distribution of the process at time zero (i.e., finite mixture of Gaussian distributions), and hence $\mu \in C^+$. Thus by the Beněs result, there exists a finite invariant measure for the process Y_t.

APPENDIX B: RANGE OF AUTOCORRELATIONS

This is the derivation of Equation (9). By Equations (7) and (8), we have

$$\rho_1 = \frac{\alpha_1\phi_1}{1 - \alpha_2\phi_2} \tag{B.1}$$

and

$$\rho_2 = \alpha_1\phi_1\rho_1 + \alpha_2\phi_2. \tag{B.2}$$

Solving the system (B.1)–(B.2) yields

$$\alpha_1\phi_1 = \frac{\rho_1 - \rho_2\rho_1}{1-\rho_1^2} \tag{B.3}$$

and

$$\alpha_2\phi_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2}. \tag{B.4}$$

Because the process $\{y_t\}$ is second-order stationary, by Theorem 2 we have

$$\alpha_1\phi_1^2 + \alpha_2\phi_2^2 < 1$$

or, equivalently,

$$\frac{(\alpha_1\phi_1)^2}{\alpha_1} + \frac{(\alpha_2\phi_2)^2}{\alpha_2} < 1. \tag{B.5}$$

Substituting (B.3) and (B.4) into (B.5) yields

$$\frac{1}{\alpha_1}\left(\frac{\rho_1 - \rho_2\rho_1}{1-\rho_1^2}\right)^2 + \frac{1}{\alpha_2}\left(\frac{\rho_2 - \rho_1^2}{1-\rho_1^2}\right)^2 < 1,$$

$$\alpha_2(\rho_1 - \rho_2\rho_1)^2 + \alpha_1(\rho_2 - \rho_1^2)^2 - \alpha_1\alpha_2(1-\rho_1^2)^2 < 0,$$

and

$$(1-\alpha_1)(\rho_1 - \rho_2\rho_1)^2 + \alpha_1(\rho_2 - \rho_1^2)^2 - \alpha_1(1-\alpha_1)(1-\rho_1^2)^2 < 0,$$

since $\alpha_1 + \alpha_2 = 1$. Hence, rearranging this equation gives the range of autocorrelations defined by

$$\alpha_1^2\rho_1^4 + (-2\alpha_1^2 + \alpha_1 + 1)\rho_1^2 - 2\rho_2\rho_1^2 + (1-\alpha_1)\rho_2^2\rho_1^2 + \alpha_1\rho_2^2 - \alpha_1(1-\alpha_1) < 0, \tag{B.6}$$

where

$$|\rho_1| \le 1,$$

$$|\rho_2| \le 1,$$

and

$$|\alpha_1| \le 1.$$

Thus the boundary of the region defined by (B.6) satisfies the following equation for fixed α_1 and ϕ_1:

$$a\rho_2^2 + 2b\rho_2 + c = 0,$$

where

$$a = (1-\alpha_1)\rho_1^2 + \alpha_1,$$

$$b = -\rho_1^2,$$

and

$$c = \alpha_1^2\rho_1^4 + (-2\alpha_1^2 + \alpha_1 + 1)\rho_1^2 - \alpha_1(1-\alpha_1).$$

Because

$$b^2 - ac = (\alpha_1^2 - \alpha_1^3)(1-\rho_1^2)^3 \ge 0,$$

the solutions of the foregoing quadratic equation satisfy

$$\rho_2 = \frac{\rho_1^2 \pm \sqrt{\alpha_1^2(1-\alpha_1)(1-\rho_1^2)^3}}{(1-\alpha_1)\rho_1^2 + \alpha_1}.$$

This is Equation (9) in Section 3.

[Received December 1993. Revised April 1996.]

REFERENCES

Adke, S. R., and Deshmukh, S. R. (1988), "Limit Distribution of a High-Order Markov Chain," *Journal of the Royal Statistical Society*, Ser. B, 50, 105–108.

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, London: Chapman and Hall.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *The Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csake, Hungary: Akademiai Kiado, pp. 267–281.

Alspach, D. L., and Sorenson, H. W. (1972), "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Transactions on Automatic Control*, 17, 439–448.

Baum, L. E. (1971), "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," in *Inequalities, III: Proceedings of a Symposium*, ed. Q. Shisha, New York: Academic Press.

Beneš, V. E. (1967), "Existence of Finite Invariant Measures for Markov Processes," *Proceedings of the American Mathematical Society*, 18, 1058–1061.

Bickel, P. J., Klaassen, C. A., Rartov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation in Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis Forecasting and Control*, (2nd ed.), San Francisco: Holden-Day.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994), *Time Series Analysis Forecasting and Control* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–619.

Bruce, A. G., and Martin, R. D. (1989), "Leave-k-Out Diagnostics for Time Series" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 51, 363–424.

Craig, P. C. (1989), "Time Series Analysis for Directional Data," unpublished Ph.D. thesis, Trinity College.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Denby, L., and Martin, R. D. (1979), "Robust Estimation of the First-Order Autoregressive Parameter," *Journal of the American Statistical Association*, 74, 140–146.

Gaver, D. P., and Lewis, P. A. W. (1980), "First-Order Autoregressive Gamme Sequences and Point Processes," *Advanced Applications in Probability*, 12, 727–745.

Goldberg, S. (1958), *Difference Equations*, New York: Wiley.

Good, I. J. (1956), "On the Estimation of Small Frequencies in Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 18, 113–124.

Granger, C. W. J., and Joyeux, R. (1980), "An Introduction to Long-Memory Time Series Models and Fractional Differening," *Journal of Time Series Analysis*, 1, 15–29.

Harrison, P. J., and Stevens, C. F. (1976), "Bayesian Forecasting" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 1–41.

Hartley, H. O. (1958), "Maximum Likelihood Estimation From Incomplete Data," *Biometrics*, 14, 174–194.

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–310.

Hosking, J. R. M. (1981), "Fractional Differencing," *Biometrika*, 68, 165–176.

Kitagawa, G. (1987), "Non-Gaussian State-Space Modeling of Nonstationary Time Series," *Journal of the American Statistical Association*, 82, 1032–1041.

Kwok, M. (1989), "Some Results on Higher-Order Markov Chain Models," unpublished M. Phil. thesis, University of Hong Kong.

Lawrance, A. J., and Lewis, P. A. W. (1985), "Modelling and Residual Analysis of Nonlinear Autoregressive Time Series in Exponential Variables," *Journal of the Royal Statistical Society*, Ser. B, 47, 165–202.

Le, N. D., Leroux, B. L., and Puterman, M. L. (1992), "Exact Likelihood Evaluation in a Markov Mixture Model for Time Series of Seizure Counts," *Biometrics*, 48, 317–323.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983), "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *The Bell System Technical Journal*, 62, 1035–1074.

Li, W. K., and Kwok, M. (1989), "Estimation and Testing for Higher-Order Markov Chain Models," unpublished manuscript, University of Hong Kong, Dept. of Statistics.

Martin, R. D. (1979), "Approximate Conditional-Mean–Type Smoothers and Interpolators," in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Berlin: Springer, pp. 117–143.

Martin, R. D., and Raftery, A. E. (1987), "Robustness, Computation, and Non-Euclidean Models," *Journal of the American Statistical Association*, 82, 1044–1050.

Martin, R. D., and Yohai, V. J. (1985), "Robustness in Time Series and Estimating ARMA Models," in *Handbook of Statistics 5*, eds. E. J. Hannan, P. R. Krishnaiah, and M. M. Rao, New York: Elsevier, pp. 119–155.

——— (1986), "Influence Functionals for Time Series," *The Annals of Statistics*, 14, 781–855.

Mehran, F. (1989), "Analysis of Discrete Longitudinal Data: Infinite-Lag Markov Models," in *Statistical Data Analysis and Inference*, ed. Y. Dodge, Amsterdam: Elsevier, pp. 533–541.

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Owen, A. (1983), "Optimal Transformations for Autoregressive Time Series Models," Technical Report 20, Stanford University, Dept. of Statistics.

Raftery, A. E. (1985a), "A Model for High-Order Markov Chains," *Journal of the Royal Statistical Society*, Ser. B, 47, 528–539.

——— (1985b), "A New Model for Discrete-Valued Time Series: Autocorrelations and Extensions," *Rassegna di Metodi Statistici ed Applicazioni*, 4, 149–162.

——— (1985c), Discussion of "Modelling and Residual Analysis of Nonlinear Time Series in Exponential Variables," by A. J. Lawrance and P. A. W. Lewis (1985), *Journal of the Royal Statistical Society*, Ser. B, 47, 185–187.

——— (1994), "Change Point and Change Curve Modeling in Stochastic Processes and Spatial Statistics," *Journal of Applied Statistical Science*, 1, 403–424.

Raftery, A. E., and Tavaré, S. (1994), "Estimation and Modeling Repeated Patterns in the Mixture Transition Distribution (MTD) Model for High-Order Markov Chains," *Applied Statistics*, 43, 179–199.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Smith, A. F. M., and West, M. (1983), "Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter," *Biometrics*, 39, 867–878.

Stern, R. D., and Coe, R. (1984), "A Model-Fitting Analysis of Daily Rainfall Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 147, 1–34.

Subba Rao, T., and Gabr, M. M. (1984), *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, New York: Springer-Verlag.

Tanner, M. A. (1993), *Tools for Statistical Inference*, New York: Springer-Verlag.

Taplin, R., and Raftery, A. E. (1994), "Analysis of Agricultural Field Trials in the Presence of Outliers and Fertility Jumps," *Biometrics*, 50, 764–781.

Tong, H. (1990), *Non-Linear Time Series*, New York: Oxford University Press.

Wellner, J. A. (1985), "Semiparametric Models: Progress and Problems" (with discussion), in *Proceedings of the 45th Session of the International Statistical Institute*, Amsterdam.

West, M. (1986), "Bayesian Model Monitoring," *Journal of the Royal Statistical Society*, Ser. B, 48, 70–78.

Wichern, D. W., Miller, R. B., and Hsu, D. (1976), "Changes of Variance in First-Order Autoregressive Time Series Models—With an Application," *Applied Statistics*, 25, 248–256.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.