Autoregressive Integrated Moving Average (ARIMA) Modeling

Introduction

A time series is the result of observing a magnitude over time (*see* **Time Series Analysis**). For instance, one can observe the number of defects in a production process every day, the thickness of a coating layer every hour, or the monthly demand for a product. An important class of time series is the stationary one. A time series is stationary if it is stable over time, so that the level of the series is constant, the variability is also constant, and the dependency relationships between successive values of the time series do not change over time.

Many real time series are not stationary but can be transformed to stationarity. Thus a nonstationary series $\{y_t\}$ that wanders over time without a fixed mean may have the property that its growth $x_t := y_t - y_{t-1}$ is a stationary time series. Then we say that the series $\{y_t\}$ is *integrated* and that the series $\{x_t\}$ is the *first difference* of $\{y_t\}$.

ARMA (auto-regressive moving average) models are useful to approximate the dynamics of many stationary time series, whereas ARIMA (autoregressive integrated moving average) models are useful for integrated time series. Figure 1 shows three time series generated by simulation using two ARMA models (a and b) and one ARIMA model (c). Simple inspection of these series shows that they have different dynamics because of their different autocorrelation structures (*see* **Autocorrelated Data**). The general form of ARMA and ARIMA models will be considered subsequently.

ARMA Models

A process $\{y_t\}$ is called stationary in the weak form if $E(y_t)$ and $\gamma_k := \mathrm{cov}(y_t, y_{t+k})$ for $k = 0, 1, \ldots$ are finite and do not depend on t. The sequence $\{\gamma_k\}$ is called the *autocovariance function* and $\rho_k := \gamma_k/\gamma_0$ is the *autocorrelation function* (*see* **Autocorrelation Function**). Wold [1] proved that any stationary process, $\{y_t\}$, without deterministic components can be written as

$$y_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots = \sum_{i=0}^{\infty} \psi_i a_{t-i} \qquad (\psi_0 = 1) \tag{1}$$

where $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ (*stationarity condition*) and $\{a_t\}$ is a zero-mean white noise process ($E(a_t a_{t+k}) = 0$ for $k \neq 0$) with variance $\sigma^2 := E(a_t^2)$. This is called the general, or the infinite **moving average** MA(∞), representation of a stationary process. Model (1) leads to $E(y_t) = 0$. If the process $\{y_t\}$ has deterministic mean $\{\mu_t\}$, the general representation applies to the zero-mean process $\{y_t - \mu_t\}$. The autocovariance function is $\gamma_k = E(y_t y_{t+k}) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$, for $k \geq 0$ (see also Appendix 1). Using the lag operator B such that $By_t = y_{t-1}$, the MA(∞) representation can formally be written as $y_t = \psi(B)a_t$, where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$. The white noise $\{a_t\}$ can often be assumed to be normally distributed.

If model (1) is invertible, it can be written in an alternative way that is called the infinite autoregressive AR(∞) representation of the process and is given by

$$y_t - \pi_1 y_{t-1} - \pi_2 y_{t-2} - \cdots = y_t - \sum_{i=1}^{\infty} \pi_i y_{t-i} = a_t \tag{2}$$

where $\sum_{i=1}^{\infty} \pi_i^2 < \infty$ (*invertibility condition*). Model (2) can formally be written as $\pi(B)y_t = a_t$, where $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$, so that $\pi(B)\psi(B) = 1$.

These two general representations have an infinite number of parameters. Approximations having a finite number of parameters are needed to model any finite time series $\mathbf{Y}' = (y_1, \ldots, y_n)$. The simplest approximation is to assume a finite number q of terms in equation (1), which leads to the q-order moving average MA(q) model $y_t = \theta_q(B)\, a_t$, where $\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ (minus signs are used customarily). For MA(q) models, $\gamma_0 = \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2) < \infty$, so that they are always stationary; moreover $\rho_k = 0$ for $k > q$. A second way to approximate the general form is to use a finite number p of terms in equation (2). This leads to the p-order autoregressive AR(p) model $\phi_p(B)y_t = a_t$, where $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$. An AR($p$) model is *stationary* if and only if all the roots of

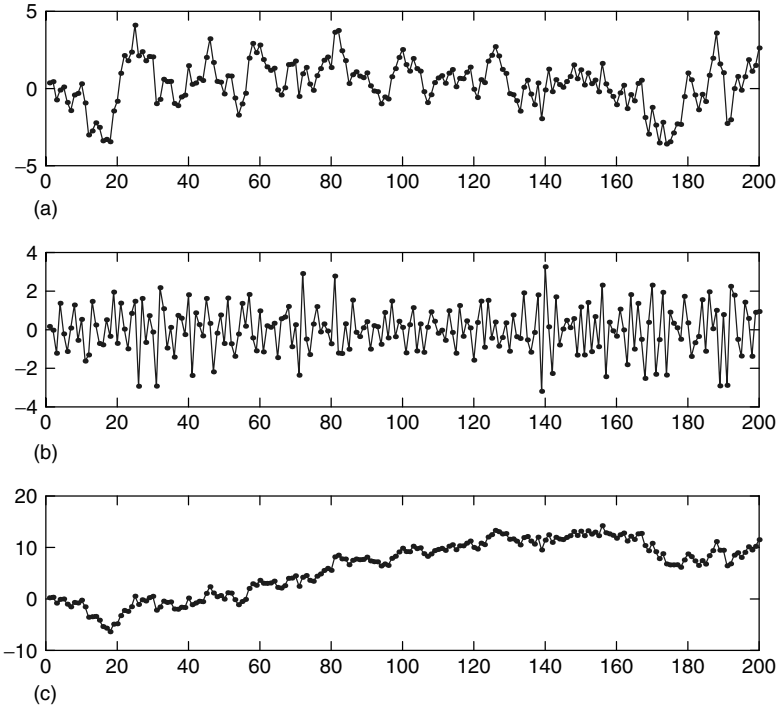

Figure 1 Time series generated using the AR(1) model $y_t - 0.8y_{t-1} = a_t$ (a), the MA(1) model $y_t = a_t - 0.8a_{t-1}$ (b), and the IMA(0, 1, 1) model $y_t - y_{t-1} = a_t - 0.2a_{t-1}$ (c), for a common simulated white noise series $(a_1, \ldots, a_{200})$ with zero mean and variance $\sigma^2 = 1$

the polynomial $\phi_p(B)$ are outside the unit circle, so that the model accepts the MA(∞) representation $y_t = \phi_p^{-1}(B)a_t$. Thus the AR(1) model $(1 - \phi B)y_t = a_t$ can be written as $y_t = (1 - \phi B)^{-1}a_t = (1 + \phi B + \phi^2 B^2 + \cdots)a_t$ if and only if $|\phi| < 1$, so that $\gamma_0 = \sigma^2(1 - \phi^2)^{-1} < \infty$; the autocorrelation function is $\rho_k = \phi^k$ implying a geometrical decay of $\rho_k \neq 0$, which is a feature common to all stationary AR(p) models.

The AR(p) and MA(q) models can be combined to form the ARMA(p, q) model $\phi_p(B)y_t = \theta_q(B)a_t$. ARMA($p$, q) models are *stationary* if and only if all the roots of $\phi_p(B)$ are outside the unit circle, and they are *invertible* if and only if all the roots of $\theta_q(B)$ are outside the unit circle. Only invertible models yield time series *forecasts* directly (see the section titled "Forecasting Time Series").

ARIMA Models

ARMA(p, q) models can be extended to model the dynamics of integrated time series. A nonstationary process $\{y_t\}$ whose dth difference $\{(1 - B)^d y_t\}$ is stationary is called an *integrated* process of order $d > 0$. Using the operator $\nabla = 1 - B$, the general ARIMA(p, d, q) equation is

$$\phi_p(B)\nabla^d(y_t - \mu_t) = \theta_q(B)a_t \tag{3}$$

where $\phi_p(B)$ and $\theta_q(B)$ have no common roots and are invertible (i.e., all their roots are outside the unit circle), and $\{\mu_t\}$ is a deterministic trend. In model (3), $\{\nabla^d \mu_t\}$ is the mean of the differenced series $\{\nabla^d y_t\}$ and is estimable, but the d integrating constants in $\{\mu_t\}$ cannot be estimated. Important examples of this type of models are the random walk, $y_t = y_{t-1} + a_t$, also called ARIMA(0, 1, 0) or I(1), and the IMA(0, 1, 1) model $y_t = y_{t-1} + a_t - \theta a_{t-1}$.

Under model (3), $\{\nabla^d(y_t - \mu_t)\}$ is a zero-mean stationary process. The mean and variance of y_t conditioned on given past values $\{y_0, y_{-1}, \ldots\}$ are finite. Thus for a random walk $y_t = y_{t-1} + a_t$ and $t > 0$, one has $E(y_t|y_0) = y_0$ whereas $\text{var}(y_t|y_0) = \sigma^2 t$ increases indefinitely as t increases. Also, for $k > 0$,

$\text{cov}(y_t, y_{t+k}) = \sigma^2 t \geq 0$, so that the autocorrelation coefficient $\rho(y_t, y_{t+k}) = (1 + k/t)^{-1/2}$ decreases with k, which is a property of integrated processes. The function $V(k) = \frac{1}{2}E(y_k - y_0)^2 = \gamma_0(1 - \rho_k)$ is called the *variogram*. The quotient $V(k)/V(1)$ is called [2] the *standardized variogram* (SV) and is a dimensionless function of k. The SV of ARMA models is $(1 - \rho_k)/(1 - \rho_1)$ and is always bounded. Unlike the autocovariance function, the SV of ARIMA(p, 1, q) models always exists and equals $k + 2\sum_{i=1}^{k-1}(k - i)\text{corr}(\nabla y_t, \nabla y_{t-i})$; for the IMA(0, 1, 1) model, it is $1 + (k - 1)(1 - \theta)^2/(1 + \theta^2)$ and increases linearly with k.

Applications

ARIMA models are important for generating forecasts and providing understanding in all kinds of time series problems from economics to health care applications [2–9]. In particular, in quality and reliability, they are important in process monitoring if observations are correlated, designing schemes for process adjustment, monitoring a reliability system over time, forecasting time series, estimating missing values, finding **outliers** and atypical events, understanding the effects of changes in a system, and so on.

Forecasting Time Series

ARIMA models can be used to forecast time series. Forecasting with ARIMA models is very simple when $q = 0$. Using the notation $\phi_p(B)\nabla^d = 1 - \varphi_1 B - \cdots - \varphi_{p+d}B^{p+d}$ and assuming, for simplicity, that $\mu_t = 0$ in equation (3), the minimum mean squared error (MMSE) forecast of y_t given $(y_{t-1}, \ldots, y_{t-p-d})$ is $\widehat{y}_t = \varphi_1 y_{t-1} + \cdots + \varphi_{p+d} y_{t-p-d}$. The forecast error is a_t. When $q > 0$, however, the forecasts depend on previously observed forecast errors. Thus for given observations $(y_{t-1}, \ldots, y_{t-p-d})$, and computed forecast errors $(a_{t-1}, \ldots, a_{t-q})$, the MMSE forecast of y_t is

$$\widehat{y}_t = \varphi_1 y_{t-1} + \cdots + \varphi_{p+d} y_{t-p-d} - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad (4)$$

and the forecast error is a_t. The quick recursion (4) is numerically stable when used iteratively if and only if $\theta_q(B)$ is *invertible*.

For a finite series $\mathbf{Y}' = (y_1, \ldots, y_n)$, equation (4) is exact only asymptotically and has to be initialized. Suppose a series $\mathbf{Y}$ having zero mean and **covariance** matrix $\mathbf{\Omega} := \text{cov}(\mathbf{Y})$ has to be forecast. This problem can be formulated as successively finding uncorrelated (orthogonal) forecast errors $\mathbf{E}' = (e_1, \ldots, e_n)$ with zero mean and covariance matrix $\mathbf{D} := \text{cov}(\mathbf{E}) = \text{diag}(D_{11}, \ldots, D_{nn})$. This can be solved by applying Gram–Schmidt orthogonalization to $\mathbf{Y}$. Then $\widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{E}$. Starting with $e_1 = y_1$ (so that $D_{11} = \Omega_{11}$ and $\widehat{y}_1 = 0$) and using $e_t = y_t - \widehat{y}_t = y_t - \sum_{j=1}^{t-1}\lambda_{tj}e_j$, one gets $\text{cov}(y_t, e_i) = \text{cov}\left(e_t + \sum_{j=1}^{t-1}\lambda_{tj}e_j, e_i\right) = \lambda_{ti}\text{var}(e_i)$ and $\text{cov}(y_t, e_t) = \text{var}(e_t) = D_{tt}$. Then $\lambda_{ti} = \text{cov}(y_t, e_i)/\text{cov}(y_i, e_i)$. The required covariances are found using the recursion

$$\text{cov}(y_t, e_i) = \text{cov}\left(y_t, y_i - \sum_{j=1}^{i-1}\lambda_{ij}e_j\right) = \Omega_{ti} - \sum_{j=1}^{i-1}\lambda_{ij}\text{cov}(y_t, e_j) \qquad (5)$$

for $t = 1, \ldots, n$ and $i = 1, \ldots, t$. Hence $\mathbf{Y} = \mathbf{L}_\Omega\mathbf{E}$, where $\mathbf{L}_\Omega$ is the lower triangular matrix with elements λ_{tj} for $t = 1, \ldots, n$ and $j = 1, \ldots, t$ (note that $\lambda_{tt} = 1$), and $\mathbf{\Omega} = \mathbf{L}_\Omega\mathbf{D}\mathbf{L}'_\Omega$.

If the series $\mathbf{Y}$ is generated by the ARMA(p, q) model $\phi_p(B)(y_t - \mu_t) = \theta_q(B)a_t$, $\mathbf{\Omega}$ has elements $\Omega_{ij} = \gamma_{|i-j|}$. Then, it is advantageous to apply the recursion (5) to the working series $\mathbf{W}' = (w_1, \ldots, w_n)$ defined by $w_t = y_t - \mu_t$ for $t = 1, \ldots, p$ and $w_t = \phi_p(B)(y_t - \mu_t)$ for $t = p + 1, \ldots, n$, because $\mathbf{\Omega}$ is replaced by the *band* matrix $\mathbf{\Theta} := \text{cov}(\mathbf{W})$ (see Appendix 2). Then $\mathbf{W} = \mathbf{L}_\Theta\mathbf{E}$; $\mathbf{E} = \mathbf{W} - \widehat{\mathbf{W}} = \mathbf{Y} - \widehat{\mathbf{Y}}$; and $\mathbf{\Theta} = \mathbf{L}_\Theta\mathbf{D}\mathbf{L}'_\Theta$, where $\mathbf{L}_\Theta$ is a lower triangular band matrix with the same lower bandwidth as $\mathbf{\Theta}$. The forecast errors $\{e_t\}$ converge to the noise $\{a_t\}$ and $\lambda_{t,t-j}$ converges to $-\theta_j$ as t increases, so that one can switch from equation (5) to equation (4) as soon as equation (4) becomes sufficiently accurate. Note also that $\mathbf{W} = \mathbf{\Phi}(\mathbf{Y} - \boldsymbol{\mu})$, where $\mathbf{\Phi}$ is the lower triangular band matrix with bandwidth p having ones in the diagonal; elements $\Phi_{i,i-k} = -\phi_k$ for $i = p + 1, \ldots, n$ and $k = 1, \ldots, p$; and zeros elsewhere. Then $\mathbf{\Theta} = \mathbf{\Phi}\mathbf{\Omega}\mathbf{\Phi}'$ and $\mathbf{L}_\Theta = \mathbf{\Phi}\mathbf{L}_\Omega$.

This method can also be applied to ARIMA models provided that limiting formulas are used.

For example, for ARMA(p, q) models with constant mean μ, the MMSE forecast of y_2 given y_1 is $\widehat{y}_2 = \mu + \rho_1(y_1 - \mu)$ and the variance of the forecast error is $D_{22} = \gamma_0(1 - \rho_1^2)$. Hence, for ARIMA($p$, 1, q) models, $\widehat{y}_2 = y_1$ with forecast error variance $\lim_{\rho_1 \to 1} D_{22} = \text{var}(\nabla y_t)$.

MMSE forecasts $\widehat{y}_{t|t-k}$ of y_t at origin $t-k$ for any lead time $k > 1$ are found replacing equation (4) by $\widehat{y}_{t|t-k} = \sum_{i=1}^{k-1} \varphi_i \widehat{y}_{t-i|t-k} + \sum_{i=k}^{p+d} \varphi_i y_{t-i} - \sum_{i=k}^{q} \theta_i a_{t-i}$; the forecast error is $e_{t|t-k} = y_t - \widehat{y}_{t|t-k}$ and $\text{var}(e_{t|t-k}) = \sigma^2 \sum_{i=0}^{k-1} \psi_i^2$. Equation (5) can also be adapted to forecast y_t at origin $t-k$ by taking $\mathbf{Y}' = (y_1, \ldots, y_{t-k}, y_t)$. Thus $\widehat{y}_{t|t-k} = \sum_{j=1}^{t-k} \lambda_{tj} e_j$ and $\text{var}(e_{t|t-k}) = \sum_{j=t-k+1}^{t} \lambda_{tj}^2 D_{jj}$. For ARMA($p$, q) models with constant μ, $\widehat{y}_{k+1|1} = \mu + \rho_k(y_1 - \mu)$ and $\text{var}(e_{k+1|1}) = \gamma_0(1 - \rho_k^2)$. For ARIMA($p$, 1, q) models, $\widehat{y}_{k+1|1} = y_1$ and $\text{var}(e_{k+1|1}) = k\text{var}(\nabla y_t) + 2\sum_{i=1}^{k-1}(k-i)\text{cov}(\nabla y_t, \nabla y_{t-i})$.

Figure 2 shows the last 11 observations in the time series of Figure 1 together with the MMSE forecasts $\widehat{y}_{t+k|t}$ of y_{t+k} at origin $t = 200$ for lead times $k = 1, \ldots, 5$ and corresponding probability limits for each individual y_{t+k} at ± 2 standard deviations.

Process Control: Process Monitoring and Process Adjustment

Quality control is often implemented by the complementary use of process monitoring and process adjustment [2, 5]. Process *monitoring* is a part of statistical process control (SPC) and assumes that the process has already been brought to a state of stationary control by using techniques such as the standardization of materials, machines, and methods. The purpose of monitoring is to *guarantee* that the process stays in its stationary state. Hence stationary time series models are used. Process *adjustment* is primarily used when current efforts to bring the process to a state of stationary control have failed. The purpose is to determine when and by how much the process should be adjusted to attain stationarity. This can be implemented by sequentially forecasting the deviation from the target and compensating for this deviation by adjusting an input variable when and as required (*see* **Feedback Control**). Nonstationary and stationary time series models are used to describe the dynamics of the unadjusted and adjusted processes, respectively.

Process monitoring and process adjustment require that data should be taken routinely at equally spaced periods of time separated by the so-called unit sampling interval. The determination of the optimal length of this sampling interval is important. It may also be appropriate to change the length of the sampling interval depending on the recent behavior of the process. Hence it is useful that the models used for process control are *invariant* to changes in the length of the sampling interval (only the model parameters may change).

In *process monitoring*, stationary ARMA models that are invariant to the length of the sampling interval are used. The basic type of these models is the white noise $y_t = \mu + a_t$, where the observed process $\{y_t\}$ has mean μ, which should be equal to the target value, and the white noise series $\{a_t\}$ is a disturbance. Alternative invariant models are the ARMA(1, 1) $y_t - \mu = \phi(y_{t-1} - \mu) + a_t - \theta a_{t-1}$ with $0 < \phi < 1$, and the AR(1) with $\theta = 0$.

In *process adjustment*, nonstationary ARIMA models that are invariant to the length of the sampling interval are used. The basic type of these models is the random walk $y_t = y_{t-1} + a_t$, used by Taguchi *et al.* [10] to model drifting observed processes. Box and Luceño [2] consider that the random walk is often too nonstationary for process adjustment applications and use the IMA(0, 1, 1) model instead. Luceño *et al.* [11] and [5] use the IMA(0, 1, 1) with trend $y_t = \beta + y_{t-1} + a_t - \theta a_{t-1}$, which is also invariant and has parameters θ and β to deal with different degrees of nonstationarity and the trend of unadjusted processes.

Fitting ARIMA Models

The approach suggested by Box and Jenkins [3, 12] is often used to fit ARIMA models to finite time series. It has three steps: model identification, model estimation, and model diagnostic.

Model Identification

Suppose that a series $\mathbf{Y}' = (y_1, \ldots, y_n)$ has been observed. The first step is to decide whether it is a stationary or an integrated time series. This decision is often made by visual inspection of plots of the time series and its autocorrelation function, as well as similar plots for the first and second difference of the series. It is also possible to use unit root tests of the type introduced in [13]. The following estimators are applied

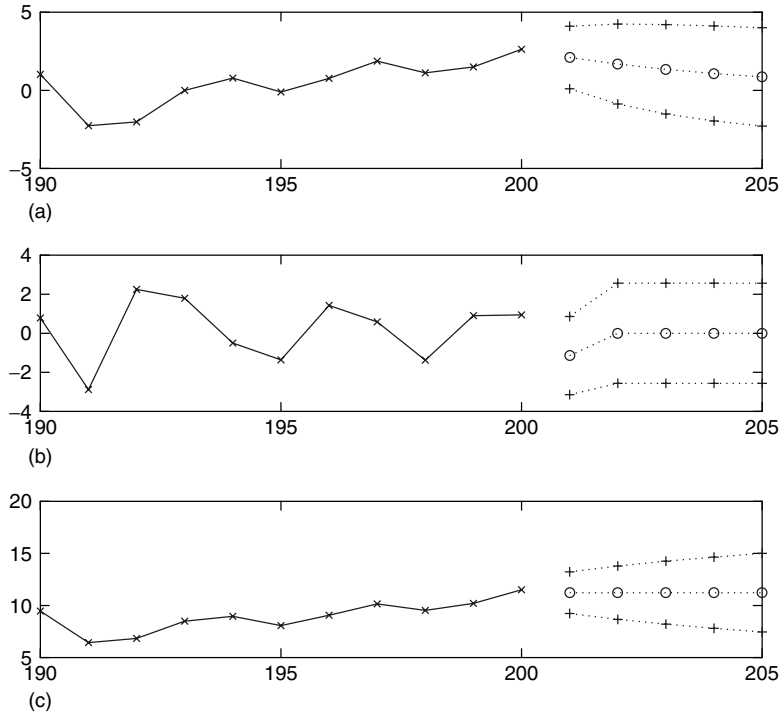

Figure 2 Observations in the time series of Figure 1 for $t = 190, \ldots, 200$ ($\times$) along with their MMSE forecasts for $t = 201, \ldots, 205$ ($\circ$) with ± 2 standard deviation limits ($+$)

to the series itself and its first differences: $\overline{y} = n^{-1} \sum_{t=1}^{n} y_t$ for the mean, $\widehat{\gamma}_0 = n^{-1} \sum_{t=1}^{n} (y_t - \overline{y})^2$ for the variance,

$$\widehat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \overline{y})(y_{t+k} - \overline{y}) \qquad (k \geq 0) \qquad (6)$$

for the autocovariances, and $r_k = \widehat{\gamma}_k / \widehat{\gamma}_0$ for the autocorrelations. If the sample autocorrelation function of the original or differenced series has only its first q coefficients significantly different from zero, one can tentatively assume an MA(q) model for that series. Otherwise, one should entertain AR, ARMA, or ARIMA models. Figure 3 shows the sample autocorrelation functions of the three time series in Figure 1. The sequence $\{r_k\}$ decreases approximately geometrically for the AR(1) series; only r_1 is clearly significantly different from zero for the MA(1) series; whereas the sequence $\{r_k\}$ decreases very slowly for the IMA(0, 1, 1) series, indicating the need to differentiate this time series.

Many statistical tools have been derived to identify the orders p, d, and q of a model: the partial autocorrelation function [3], the extended autocorrelation function [14], the inverse autocorrelation function [15, 16], and so on. However, because the estimation of ARMA models is nowadays fast and reliable, we recommend that all possible models compatible with the observed autocorrelations be estimated and that a model selection criterion be used to choose from among them (see the sections titled "Model Estimation by Maximum Likelihood", "Joint Estimation of Missing Values and Model Parameters", and "Model Diagnostic").

Model Estimation by Maximum Likelihood

Let $\mathbf{Y}' = (y_1, \ldots, y_n)$ be a series generated by a stationary ARMA(p, q) model $\phi_p(B)(y_t - \mu_t) = \theta_q(B)a_t$, so that $\boldsymbol{\mu}' := E(\mathbf{Y}') = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\Omega} := \mathrm{cov}(\mathbf{Y})$. Let $\mathbf{W} := \boldsymbol{\Phi}(\mathbf{Y} - \boldsymbol{\mu})$ be the working series of the section titled "Forecasting Time Series", so that $\boldsymbol{\Theta} := \mathrm{cov}(\mathbf{W}) = \boldsymbol{\Phi}\boldsymbol{\Omega}\boldsymbol{\Phi}'$, $\mathbf{W} = \mathbf{L}_{\Theta}\mathbf{E}$,

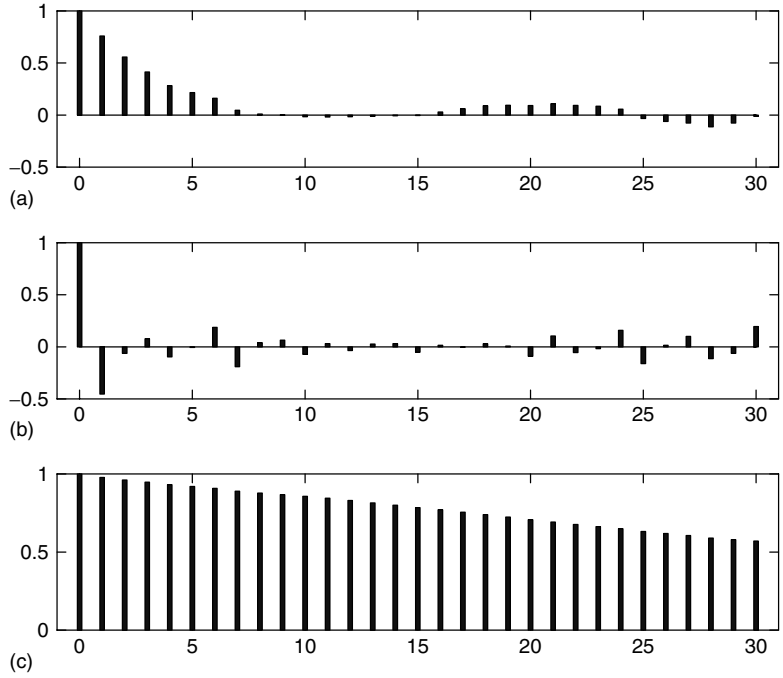

Figure 3 Sample autocorrelation functions for each of the three time series in Figure 1

$\boldsymbol{\Theta} = \mathbf{L}_\Theta \mathbf{D} \mathbf{L}_\Theta'$, $|\boldsymbol{\Omega}| = |\boldsymbol{\Theta}| = |\mathbf{D}|$, $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$, and $\mathbf{L}_\Theta$ are band matrices, and $\mathbf{D}$ is diagonal. For normal $\{a_t\}$, the likelihood function of the parameters $\boldsymbol{\phi}' = (\phi_1, \ldots, \phi_p)$, $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_q)$, $\boldsymbol{\mu}$, and σ^2 given $\mathbf{Y}$ is

$$\begin{aligned} L(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\mu}, \sigma^2 | \mathbf{Y}) &= (2\pi)^{-n/2} |\boldsymbol{\Omega}|^{-1/2} \\ &\quad \times \exp\{-(\mathbf{Y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{Y} - \boldsymbol{\mu})/2\} \\ &= (2\pi)^{-n/2} |\boldsymbol{\Theta}|^{-1/2} \\ &\quad \times \exp(-\mathbf{W}' \boldsymbol{\Theta}^{-1} \mathbf{W}/2) \\ &= (2\pi)^{-n/2} |\mathbf{D}|^{-1/2} \exp(-\mathbf{E}' \mathbf{D}^{-1} \mathbf{E}/2) \end{aligned} \tag{7}$$

where $|\mathbf{D}| = \prod_{i=1}^n D_{ii}$ and $\mathbf{E}'\mathbf{D}^{-1}\mathbf{E} = \sum_{i=1}^n e_i^2 / D_{ii}$. Maximum-likelihood estimates (MLEs) can be found by minimizing $-\ln[L\{\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\mu}(\boldsymbol{\beta}), \sigma^2 | \mathbf{Y}\}]$, where $\boldsymbol{\mu}$ is parameterized as $\boldsymbol{\mu}(\boldsymbol{\beta})$ (*see* **Maximum Likelihood**).

For AR(p) models, $\boldsymbol{\Theta}$ equals the identity matrix excepting its first p rows and columns (and $\boldsymbol{\Omega}^{-1}$ has bandwidth p); hence the MLE of $\boldsymbol{\phi}'$ is close to its minimum sum of squared error estimate. For an AR(1) model, $|\boldsymbol{\Theta}|^{-1} = \sigma^{-2n}(1 - \phi^2)$ and $\mathbf{W}'\boldsymbol{\Theta}^{-1}\mathbf{W}\sigma^2 = C_0(1 + \phi^2) - 2\phi C_1 \quad -\phi^2 \{(y_1 - \mu_1)^2 + (y_n - \mu_n)^2)\}$, where $C_k = \sum_{t=1}^{n-k} (y_t - \mu_t)(y_{t+k} - \mu_{t+k})$; hence disregarding the effects of $|\boldsymbol{\Theta}|$, y_1, and y_n, the estimate $\widehat{\phi} = C_1/C_0$ results. For MA(q) models, $\boldsymbol{\Phi}$ is the identity matrix, $\boldsymbol{\Omega} = \boldsymbol{\Theta}$ has bandwidth q, but the log-likelihood function may be far from quadratic. The MLEs for the parameters of the three time series in Figure 1 are $(\hat{\phi} = 0.765, \hat{\sigma}^2 = 0.927)$, $(\hat{\theta} = 0.838, \hat{\sigma}^2 = 0.925)$, and $(\hat{\theta} = 0.197, \hat{\sigma}^2 = 0.934)$, respectively.

Joint Estimation of Missing Values and Model Parameters

Suppose that series $\mathbf{Y}' = (y_1, \ldots, y_n)$ has m missing values (*see* **Missing Data and Imputation**) $\mathbf{y}_u' = (y_{u_1}, \ldots, y_{u_m})$, at positions $1 < u_1 < \cdots < u_m < n$, having means $\boldsymbol{\mu}_u' = (\mu_{u_1}, \ldots, \mu_{u_m})$. Let $\mathbf{M}'$ be the $m \times n$ matrix with elements $M'_{i,u_i} = 1$, $i = 1, \ldots, m$, and zeros elsewhere; and $\mathbf{B}'$ be the $(n - m) \times n$ matrix found eliminating the rows of $\mathbf{M}'$ from the $n \times n$ identity matrix. The MMSE estimator $\widehat{\mathbf{y}}_u$ of the missing values is the conditional mean of $\mathbf{y}_u$ given the observations $\mathbf{B}'\mathbf{Y}$,

that is,

$$\widehat{\mathbf{y}}_u = \boldsymbol{\mu}_u + (\mathbf{M}'\boldsymbol{\Omega}\mathbf{B})(\mathbf{B}'\boldsymbol{\Omega}\mathbf{B})^{-1}\mathbf{B}'(\mathbf{Y} - \boldsymbol{\mu})$$
$$= \boldsymbol{\mu}_u - (\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M})^{-1}(\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{B})\mathbf{B}'(\mathbf{Y} - \boldsymbol{\mu}) \quad (8)$$

The two formulas in equation (8) are equivalent, but the latter (using $\boldsymbol{\Omega}^{-1}$) is usually much faster than the former (using $\boldsymbol{\Omega}$) as $m \ll n - m$ [17–19]. Moreover, $\text{cov}(\mathbf{y}_u) = \mathbf{M}'\boldsymbol{\Omega}\mathbf{M}$, $\text{cov}(\mathbf{y}_u - \widehat{\mathbf{y}}_u) = (\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M})^{-1}$, and $\text{cov}(\widehat{\mathbf{y}}_u) = \mathbf{M}'\boldsymbol{\Omega}\mathbf{M} - (\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M})^{-1}$. For AR($p$) models with $\mu = 0$, equation (8) yields $\widehat{y}_u = -\sum_{k=1}^{p} \alpha_k(y_{u-k} + y_{u+k})/\alpha_0$, if $m = 1$ and $p < u \le n - p$, where $\alpha_k = \sum_{j=0}^{p-k} \phi_j\phi_{j+k}$ and $\phi_0 = -1$. The sequence $\{\alpha_k/\alpha_0\}$ is called the *inverse autocorrelation function* of the AR(p) model. Moreover, equation (8) accounts for end effects if $u \le p$ or $u > n - p$.

Defining $\widehat{\mathbf{Y}} = \mathbf{Y} + \mathbf{M}(\widehat{\mathbf{y}}_u - \mathbf{y}_u)$ and using $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Phi}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Phi} = \boldsymbol{\Phi}'\mathbf{L}_\Theta'^{-1}\mathbf{D}^{-1}\mathbf{L}_\Theta^{-1}\boldsymbol{\Phi}$, one can efficiently obtain [17–20] MLEs of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, and σ^2 maximizing

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\mu}, \sigma^2|\mathbf{B}'\mathbf{Y})$$
$$= (2\pi)^{-(n-m)/2}(|\boldsymbol{\Omega}||\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M}|)^{-1/2}$$
$$\times \exp\{-(\widehat{\mathbf{Y}} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\widehat{\mathbf{Y}} - \boldsymbol{\mu})/2\} \quad (9)$$

This procedure can also be applied to integrated series replacing $(\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M})^{-1}(\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{B})$ by $(\mathbf{M}_d'\boldsymbol{\Omega}_d^{-1}\mathbf{M}_d)^{-1}$ $(\mathbf{M}_d'\boldsymbol{\Omega}_d^{-1}\mathbf{B}_d)$ in equation (8), where $\boldsymbol{\Omega}_d := \text{cov}(\mathbf{Y}_d)$, and $\mathbf{Y}_d'$, $\mathbf{M}_d'$, and $\mathbf{B}_d'$ are obtained differencing d times the rows of $\mathbf{Y}'$, $\mathbf{M}'$, and $\mathbf{B}'$. Thus, for an ARIMA (p, 1, q), if y_3 is missing, $\mathbf{Y}_1$ has two unknown values ($y_3 - y_2$ and $y_4 - y_3$) that require only one row of length $n - 1$ in $\mathbf{M}_1'$, namely, $(0, 1, -1, 0, \ldots, 0)$, the first difference of $(0, 0, 1, 0, 0, \ldots .0)$. Moreover, $\mathbf{B}_1\mathbf{B}'\mathbf{Y} = (y_2 - y_1, -y_2, y_4, y_5 - y_4, \ldots)'$. Thus, for a random walk, $\widehat{y}_3 = (y_2 + y_4)/2$.

Model Diagnostic

This stage is designed to check if the **residuals** of the estimated models behave approximately as white noise series and to select the best models among those verifying this condition. If a model is correct, the autocorrelation coefficients $\widehat{r}_k$ of its residuals $(\widehat{a}_1, \ldots, \widehat{a}_n)$ should asymptotically be normal variables with zero mean and variance $1/n$. Ljung and Box [21] propose the test statistic $Q(h) = n(n + 2)\sum_{k=1}^{h} \widehat{r}_k^2/(n - k)$, which, for large n, follows a χ^2 distribution with $h - \nu$ **degrees of freedom**, where ν is the number of parameters $\{\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\mu}\}$ estimated in equation (7).

Peña and Rodriguez [22] propose a test statistic that can be up to 50% more powerful than the one proposed by Ljung and Box. The new statistic is $D_h = -n(h + 1)^{-1} \ln|\widehat{\mathbf{R}}_h|$, where $\mathbf{R}_h$ is an $h \times h$ matrix of residual autocorrelations with elements $\widehat{r}_{|i-j|}$. The asymptotic distribution of D_h is gamma with mean $h/2 - \nu$ and variance $h(2h + 1)/(3h + 3) - 2\nu$.

The selection among the fitted models whose residuals behave approximately as white noise series can be based on *model selection criteria*. Akaike [23, 24] proposes the selection of the model that leads to the smallest out-of-sample forecast error and shows that this is equivalent to minimizing $AIC = n \ln \widehat{\sigma}_{MV}^2 + 2\nu$, where $\widehat{\sigma}_{MV}^2$ is the MLE of the residual variance. Using a Bayesian approach, Schwarz [25] proposes a criterion that selects the model with highest posterior probability by minimizing $BIC = n \ln \widehat{\sigma}_{MV}^2 + \nu \ln n$. See [26] for a comparison of the performance of these criteria for model selection.

Atypical Values

Time series data are often subject to outliers or discordant observations (*see* **Outliers**). Their study has been approached from two points of view. The first is the *diagnostic approach*, in which outliers are identified by using the residuals of the estimated model and their effect is tested afterward. A model incorporating the identified outliers is proposed, and the outlier effects and model parameters are estimated jointly. The second is the *robust approach*, in which the estimation method is modified so that the estimators are insensitive to the presence of outliers. Outliers can be easily identified and tested using these robust estimators. Both methodologies complement each other, and ideas from one approach can be used to improve the other.

Fox [27] defines additive outliers (AOs) and innovative outliers (IOs) in time series and proposes the use of maximum-likelihood ratio tests for detecting them. An AO corresponds to an external error or exogenous change of the observed value of the time series at a particular time point T; that is, instead of observing the series $\{y_t\}$, a new series $\{z_t\}$ is observed which differs from $\{y_t\}$ by the magnitude

ω_A of the AO only at the time T of occurrence of the outlier. The model for the observed series is $z_t = \omega_A I_t^{(T)} + \Psi(B)a_t$, where $I_t^{(T)} = 0$ for $t \neq T$, $I_T^{(T)} = 1$, and $\Psi(B) = \nabla^{-d}\phi_p^{-1}(B)\theta_q(B)$ is the ARIMA model for the outlier-free time series.

An AO can be interpreted as a measurement error at time T, $1 \leq T \leq n$, or as an impulse effect due to exogenous causes. Thus if the original series is the output from a system, an AO may correspond to an unexpected event that happens at T, such as a strike, an accident, or a breakdown, which modifies the output of the system at T, without further effects on the future values of the series. AOs can have a very serious effect on the properties of the observed series [9]. In SPC, AOs are monitored using Shewhart or CUmulative SCORE (CUSCORE) charts (*see* **Control Charts, Overview**).

The second type of outliers introduced in [27], IO, can be generated by some internal change or endogenous effect on the time series which appears as an outlier on the process noise. The model for an IO is built by adding an impulse effect to the noise of the original process, that is, $z_t = \Psi(B)(\omega_I I_t^{(T)} + a_t)$, where ω_I is the outlier size. IOs are expected to produce small effects on the sample autocorrelation function and the parameter estimates.

A third important type of outlier in time series is a level shift, that is, a modification of the local mean or level of the process starting at time T and continuing afterward. The model for the observed series is $z_t = \omega_L S_t^{(T)} + \Psi(B)a_t$, where $S_t^{(T)} = 0$ for $t < T$, $S_t^{(T)} = 1$ for $t \geq T$ (a step function), and ω_L is the magnitude of the shift. In SPC, level shifts are monitored using CUmulative SUM (CUSUM), exponentially weighted moving average (EWMA), or CUSCORE charts (*see* **Exponentially Weighted Moving Average (EWMA) Control Chart**; **Cumulative Sum (CUSUM) Chart**).

Several procedures for outlier detection have been proposed [28–33] after the seminal work of Chang *et al.* [34]; see [9] for a revision of some of these procedures.

Model Strengths and Weaknesses

Some strengths of ARIMA models are (a) they provide a very flexible family of models, which is closed to model addition; (b) the properties of the model are well understood and simple to use; (c) their short-term forecasts are difficult to beat with more complex models; and (d) they can be easily generalized (e.g., they can generate time irreversible series using nonnormal noise).

Some weaknesses of ARIMA models are (a) they provide little help for long term forecasting; (b) exponential smoothing is easier to use and can often provide convenient approximations to reality and forecasts nearly as good as those of full *fitted* ARIMA approximations; and (c) the framework of ARIMA models may be inconvenient to handle (e.g., most models are not invariant to changes in the sampling interval, which is important for process control; or to temporal aggregation, which is important in economy).

Related Topics

ARIMA models can be generalized to incorporate long memory (ARIMA with fractional differencing or auto-regressive fractionally integrated moving average (ARFIMA) models [35, 36]); conditional **heteroscedasticity** (generalized auto-regressive conditional heteroscedasticity (GARCH) models [37]); nonlinearity (e.g., threshold AR models [38, 39]); and information from related simultaneous series (vector ARMA, ARIMA and cointegrated models [8, 40–44]).

The analysis of time series presented here is called time-domain analysis. Time series can also be analyzed using the frequency-domain approach in which Fourier transforms of the series and its autocovariance function are used. Bayesian models can also be used.

An important alternative to ARIMA models are the so-called state space or structural dynamic models [45–48]. In these models, the time series $\{y_t\}$ is represented as a linear function $y_t = h_t'\alpha_t + a_t$ of some unknown state parameters α_t, which are assumed to follow the equation $\alpha_t = A_t\alpha_{t-1} + u_t$, with appropriate noise series $\{a_t\}$ and $\{u_t\}$. The estimation and forecasting of these models are based on the *Kalman filter algorithm* [45, 46, 49].

Appendix 1: Evaluation of Some Useful Covariances

The autocovariance function $\gamma_k := \mathrm{cov}(y_t, y_{t+k})$ may be computed efficiently using the formula

$\sum_{i=0}^{p} \phi_i \gamma_{|i-j|} = \sum_{i=0}^{q} \theta_i \gamma_{ay}(j-i)$, where $\phi_0 = \theta_0 = -1$ and $\gamma_{ay}(-k) := \text{cov}(a_t, y_{t+k}) = \psi_k \sigma^2$ for $k \geq 0$. For $j = 0, \ldots, p$, the formula provides a linear system having solution $\gamma_0, \ldots, \gamma_p$; for $j > p$, it can be used recursively. The covariances $\gamma_{ay}(-k)$ are null for $k < 0$ and satisfy the recursion $\gamma_{ay}(-k) = -\theta_k \sigma^2 + \sum_{i=1}^{p} \phi_i \gamma_{ay}(-k+i)$ for $k \geq 0$.

Appendix 2: Covariance Matrix of Working Series W

The covariance matrix $\mathbf{\Theta}$ of the working series $\mathbf{W}' = (w_1, \ldots, w_n)$ defined in the section titled "Forecasting Time Series" and used in the section titled "Model Estimation by Maximum Likelihood" is a band matrix having bandwidth $\max(p-1, q)$ in their p first rows and q thereafter, because $w_t = y_t$ for $t = 1, \ldots, p$ and $w_t = \theta_q(B)a_t$ for $t = p+1, \ldots, n$. Their elements are $\Theta_{ij} = \gamma_{|i-j|}$ for $i, j = 1, \ldots, p$; $\Theta_{ij} = \Theta_{ji} = -\sum_{k=1}^{q} \theta_k \gamma_{ay}(j-i-k)$ for $i = 1, \ldots, p$ and $j = p+1, \ldots, n$, where γ_{ay} is given in Appendix 1; and $\Theta_{ij} = \sigma^2 \sum_{k=|j-i|}^{q} \theta_k \theta_{k-|j-i|}$ for $i, j = p+1, \ldots, n$, where $\theta_0 = -1$.

Acknowledgments

Luceño and Peña acknowledge the support of the Spanish dirección general de investigación (DGI) under grants MTM2005-00 287 and SEJ2004-03 303, respectively, and thank George Box and two editors for helpful comments.

References

[1] Wold, H.O. (1938). *A Study in the Analysis of Stationary Time Series*, Almquist and Wicksell, Uppsala.

[2] Box, G.E.P. & Luceño, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*, John Wiley & Sons, New York.

[3] Box, G.E.P. & Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

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

[5] Luceño, A. (2003). Dead-band adjustment schemes for on-line feedback quality control, in *Handbook of Statistics-Statistics in Industry*, R., Khattree & C.R., Rao, eds, Elsevier Science B. V., Vol. 22, pp. 695–727.

[6] Pandit, S.M. & Wu, S.M. (1983). *Time Series and System Analysis with Applications*, John Wiley & Sons, New York.

[7] Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models*, John Wiley & Sons, New York.

[8] Shumway, R.H. & Stoffer, D.S. (2000). *Time Series Analysis and its Applications*, Springer-Verlag, New York.

[9] Peña, D., Tiao, G.C. & Tsay, R.S. (2001). *A Course in Time Series Analysis*, John Wiley & Sons, New York.

[10] Taguchi, G., Elsayed, E.A. & Hsiang, T. (1989). *Quality Engineering in Production Systems*, McGraw-Hill, New York.

[11] Luceño, A., González, F.J. & Puig-Pey, J. (1996). Computing optimal adjustment schemes for the general tool-wear problem, *Journal of Statistical Computation and Simulation* **54**, 87–113.

[12] Newbold, P. (1981). Some recent developments in time series analysis, *International Statistical Review* **49**, 53–66.

[13] Dickey, D.A. & Fuller, W.A. (1981). Likelihood ratio statistics for autoregressive processes with a unit root, *Econometrica* **49**, 1057–1072.

[14] Tsay, R.S. & Tiao, G.C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function of stationary and nonstationary ARMA models, *Journal of the American Statistical Association* **79**, 84–96.

[15] Cleveland, W.S. (1972). The inverse autocorrelations of a time series and their applications, *Technometrics* **14**, 277–293.

[16] Chatfield, C. (1980). Inverse autocorrelations, *Journal of the Royal Statistical Society. Series A* **142**, 363–377.

[17] Ljung, G.M. (1989). A note on the estimation of missing values in time series, *Communications in Statistics Part B* **18**, 459–465.

[18] Peña, D. & Maravall, A. (1991). Interpolation, outliers and inverse autocorrelations, *Communications in Statistics Part A* **20**, 3175–3186.

[19] Luceño, A. (1997). Estimation of missing values in possibly partially nonstationary vector time series, *Biometrika* **84**, 495–499.

[20] Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics* **22**, 389–396.

[21] Ljung, G.M. & Box, G.E.P. (1978). On a measure of lack of fit in time series models, *Biometrika* **65**, 297–303.

[22] Peña, D. & Rodríguez, J. (2006). The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series, *Journal of Statistical Planning and Inference* **136**, 2706–2708.

[23] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Proceedings of the 2nd International Symposium on Information Theory*, Akademiai Kiadó, Budapest, pp. 267–281.

[24] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**, 203–217.

[25] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**, 461–464.
[26] McQuarrie, A.D.R. & Tsai, C.L. (1998). *Regression and Time Series Model Selection*, World Scientific, Singapore.
[27] Fox, A.J. (1972). Outliers in time series, *Journal of the Royal Statistical Society. Series B* **34**, 350–363.
[28] Tsay, R.S. (1988). Outliers, level shifts, and variance changes in time series, *Journal of Forecasting* **7**, 1–20.
[29] Wu, L.S.-Y., Ravishanker, N. & Hosking, J.R.M. (1993). Reallocation outliers in time series, *Applied Statistics* **42**, 301–313.
[30] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition, John Wiley & Sons, Chichester.
[31] Chen, C. & Liu, L.M. (1993). Forecasting time series with outliers, *Journal of Forecasting* **12**, 13–35.
[32] Luceño, A. (1998). Detecting possibly non-consecutive outliers in industrial time series, *Journal of the Royal Statistical Society B* **60**, 295–310.
[33] Tsay, R.S., Peña, D. & Pankratz, A. (2000). Outliers in multivariate time series, *Biometrika* **87**, 789–804.
[34] Chang, I., Tiao, G.C. & Chen, C. (1988). Estimation of time series parameters in the presence of outliers, *Technometrics* **30**, 193–204.
[35] Hosking, J.R.M. (1981). Fractional differencing, *Biometrika* **68**, 165–176.
[36] Luceño, A. (1996). A fast likelihood approximation for vector general linear processes with long series: Application to fractional differencing, *Biometrika* **83**, 603–614.
[37] Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations, *Econometrica* **50**, 987–1007.
[38] Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*, Springer-Verlag, New York.
[39] Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*, Oxford University Press, Oxford.
[40] Hannan, E.J. (1970). *Multiple Time Series*, John Wiley & Sons, New York.
[41] Wei, W.W.S. (1990). *Time Series Analysis*, Addison-Wesley, Redwood.
[42] Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*, Springer-Verlag, New York.
[43] Reinsel, G.C. (1993). *Elements of Multivariate Time Series Analysis*, Springer-Verlag, New York.
[44] Engle, R.F. & Granger, C.W.J. (1987). Co-integration and error correction: representation, estimation and testing, *Econometrica* **55**, 251–276.
[45] Kalman, R.E. (1960). A New approach to linear filtering and prediction problems, *Transactions of the ASME, Series D, Journal of Basic Engineering* **82**, 35–45.
[46] Kalman, R.E. & Bucy, R.S. (1961). New results in linear filtering and prediction theory, *Transactions of the ASME, Series D, Journal of Basic Engineering* **83**, 95–107.
[47] Harvey, A.C. (1989). *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press.
[48] West, M. & Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York.
[49] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer-Verlag, New York.

Related Article

Brownian Motion; **Correlation**; **Covariance**; **Exponentially Weighted Moving Average (EWMA)**; **Markov Processes**; **Moving Averages**; **Normal Distribution**.

ALBERTO LUCEÑO AND DANIEL PEÑA