CHAPTER 4

Nonlinear Models and Their Applications

This chapter focuses on nonlinearity in financial data and nonlinear econometric models useful in analysis of financial time series. Consider a univariate time series x_t, which, for simplicity, is observed at equally spaced time points. We denote the observations by $\{x_t|t = 1, \ldots, T\}$, where T is the sample size. As stated in Chapter 2, a purely stochastic time series x_t is said to be linear if it can be written as

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}, \tag{4.1}$$

where μ is a constant, ψ_i are real numbers with $\psi_0 = 1$, and $\{a_t\}$ is a sequence of independent and identically distributed (iid) random variables with a well-defined distribution function. We assume that the distribution of a_t is continuous and $E(a_t) = 0$. In many cases, we further assume that $\text{Var}(a_t) = \sigma_a^2$ or, even stronger, that a_t is Gaussian. If $\sigma_a^2 \sum_{i=1}^{\infty} \psi_i^2 < \infty$, then x_t is weakly stationary (i.e., the first two moments of x_t are time invariant). The ARMA process of Chapter 2 is linear because it has an MA representation in Eq. (4.1). Any stochastic process that does not satisfy the condition of Eq. (4.1) is said to be nonlinear. The prior definition of nonlinearity is for purely stochastic time series. One may extend the definition by allowing the mean of x_t to be a linear function of some exogenous variables, including the time index and some periodic functions. But such a mean function can be handled easily by the methods discussed in Chapter 2, and we do not discuss it here. Mathematically, a purely stochastic time series model for x_t is a function of an iid sequence consisting of the current and past shocks—that is,

$$x_t = f(a_t, a_{t-1}, \ldots). \tag{4.2}$$

Analysis of Financial Time Series, Third Edition, By Ruey S. Tsay

The linear model in Eq. (4.1) says that $f(\cdot)$ is a linear function of its arguments. Any nonlinearity in $f(\cdot)$ results in a nonlinear model. The general nonlinear model in Eq. (4.2) is not directly applicable because it contains too many parameters.

To put nonlinear models available in the literature in a proper perspective, we write the model of x_t in terms of its conditional moments. Let F_{t-1} be the σ field generated by available information at time $t-1$ (inclusive). Typically, F_{t-1} denotes the collection of linear combinations of elements in $\{x_{t-1}, x_{t-2}, \ldots\}$ and $\{a_{t-1}, a_{t-2}, \ldots\}$. The conditional mean and variance of x_t given F_{t-1} are

$$\mu_t = E(x_t|F_{t-1}) \equiv g(F_{t-1}), \qquad \sigma_t^2 = \text{Var}(x_t|F_{t-1}) \equiv h(F_{t-1}), \tag{4.3}$$

where $g(\cdot)$ and $h(\cdot)$ are well-defined functions with $h(\cdot) > 0$. Thus, we restrict the model to

$$x_t = g(F_{t-1}) + \sqrt{h(F_{t-1})}\epsilon_t,$$

where $\epsilon_t = a_t/\sigma_t$ is a standardized shock (or innovation). For the linear series x_t in Eq. (4.3), $g(\cdot)$ is a linear function of elements of F_{t-1} and $h(\cdot) = \sigma_a^2$. The development of nonlinear models involves making extensions of the two equations in Eq. (4.3). If $g(\cdot)$ is nonlinear, x_t is said to be *nonlinear in mean*. If $h(\cdot)$ is time variant, then x_t is *nonlinear in variance*. The conditional heteroscedastic models of Chapter 3 are nonlinear in variance because their conditional variances σ_t^2 evolve over time. In fact, except for the GARCH-M models, in which μ_t depends on σ_t^2 and hence also evolves over time, all of the volatility models of Chapter 3 focus on modifications or extensions of the conditional variance equation in Eq. (4.3). Based on the well-known Wold decomposition, a weakly stationary and purely stochastic time series can be expressed as a linear function of uncorrelated shocks. For stationary volatility series, these shocks are uncorrelated but dependent. The models discussed in this chapter represent another extension to nonlinearity derived from modifying the conditional mean equation in Eq. (4.3).

Many nonlinear time series models have been proposed in the statistical literature, such as the bilinear models of Granger and Andersen (1978), the threshold autoregressive (TAR) model of Tong (1978), the state-dependent model of Priestley (1980), and the Markov switching model of Hamilton (1989). The basic idea underlying these nonlinear models is to let the conditional mean μ_t evolve over time according to some simple parametric nonlinear function. Recently, a number of nonlinear models have been proposed by making use of advances in computing facilities and computational methods. Examples of such extensions include the nonlinear state-space modeling of Carlin, Polson, and Stoffer (1992), the functional coefficient autoregressive model of Chen and Tsay (1993a), the nonlinear additive autoregressive model of Chen and Tsay (1993b), and the multivariate adaptive regression spline of Lewis and Stevens (1991). The basic idea of these extensions is either using simulation methods to describe the evolution of the conditional distribution of x_t or using data-driven methods to explore the nonlinear characteristics of a series. Finally, nonparametric and semiparametric methods such as kernel

regression and artificial neural networks have also been applied to explore the nonlinearity in a time series. We discuss some nonlinear models in Section 4.1 that are applicable to financial time series. The discussion includes some nonparametric and semiparametric methods.

Apart from the development of various nonlinear models, there is substantial interest in studying test statistics that can discriminate linear series from nonlinear ones. Both parametric and nonparametric tests are available. Most parametric tests employ either the Lagrange multiplier or likelihood ratio statistics. Nonparametric tests depend on either higher order spectra of x_t or the concept of dimension correlation developed for chaotic time series. We review some nonlinearity tests in Section 4.2. Sections 4.3 and 4.4 discuss modeling and forecasting of nonlinear models. Finally, an application of nonlinear models is given in Section 4.5.

4.1 NONLINEAR MODELS

Most nonlinear models developed in the statistical literature focus on the conditional mean equation in Eq. (4.3); see Priestley (1988) and Tong (1990) for summaries of nonlinear models. Our goal here is to introduce some nonlinear models that are applicable to financial time series.

4.1.1 Bilinear Model

The linear model in Eq. (4.1) is simply the first-order Taylor series expansion of the $f(\cdot)$ function in Eq. (4.2). As such, a natural extension to nonlinearity is to employ the second-order terms in the expansion to improve the approximation. This is the basic idea of bilinear models, which can be defined as

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} - \sum_{j=1}^{q} \theta_j a_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{s} \beta_{ij} x_{t-i} a_{t-j} + a_t, \tag{4.4}$$

where p, q, m, and s are nonnegative integers. This model was introduced by Granger and Andersen (1978) and has been widely investigated. Subba Rao and Gabr (1984) discuss some properties and applications of the model, and Liu and Brockwell (1988) study general bilinear models. Properties of bilinear models such as stationarity conditions are often derived by (a) putting the model in a state-space form (see Chapter 11) and (b) using the state transition equation to express the state as a product of past innovations and random coefficient vectors. A special generalization of the bilinear model in Eq. (4.4) has conditional heteroscedasticity. For example, consider the model

$$x_t = \mu + \sum_{i=1}^{s} \beta_i a_{t-i} a_t + a_t, \tag{4.5}$$

where $\{a_t\}$ is a white noise series. The first two conditional moments of x_t are

$$E(x_t|F_{t-1}) = \mu, \qquad \text{Var}(x_t|F_{t-1}) = \left(1 + \sum_{i=1}^{s} \beta_i a_{t-i}\right)^2 \sigma_a^2,$$

which are similar to that of the RCA or CHARMA model of Chapter 3.

Example 4.1. Consider the monthly simple returns of the CRSP equal-weighted index from January 1926 to December 2008 for 996 observations. Denote the series by R_t. The sample PACF of R_t shows significant partial autocorrelations at lags 1 and 3 so that an AR(3) model is used. The squared series of the AR(3) residuals suggests that the conditional heteroscedasticity might depend on lags 1, 3, and 8 of the residuals. Therefore, we employ the special bilinear model

$$R_t = \mu + \phi_1 R_{t-1} + \phi_3 R_{t-3} + (1 + \beta_1 a_{t-1} + \beta_3 a_{t-3})a_t$$

for the series, where $a_t = \beta_0 \epsilon_t$ with ϵ_t being an iid series with mean zero and variance 1. Note that lag 8 is omitted for simplicity. Assuming that the conditional distribution of a_t is normal, we use the conditional maximum-likelihood method and obtain the fitted model

$$\begin{aligned} R_t &= 0.0114 + 0.167R_{t-1} - 0.095R_{t-3} \\ &\quad + 0.071(1 + 0.377a_{t-1} - 0.646a_{t-3})\epsilon_t, \end{aligned} \tag{4.6}$$

where the standard errors of the parameters are, in the order of appearance, 0.0023, 0.032, 0.027, 0.002, 0.147, and 0.136, respectively. All estimates are significantly different from zero at the 5% level. Define

$$\hat{\epsilon}_t = \frac{R_t - 0.0114 - 0.167R_{t-1} + 0.095R_{t-3}}{0.071(1 + 0.377\hat{a}_{t-1} - 0.646\hat{a}_{t-3})},$$

where $\hat{\epsilon}_t = 0$ for $t \leq 3$, as the standardized residual series of the model. The sample ACF of $\hat{\epsilon}_t$ shows no significant serial correlations, but the series is not independent because the squared series $\hat{\epsilon}_t^2$ has significant serial correlations. The validity of model (4.6) deserves further investigation. For comparison, we also consider an AR(3)–ARCH(3) model for the series and obtain

$$\begin{aligned} R_t &= 0.013 + 0.223R_{t-1} + 0.006R_{t-2} - 0.013R_{t-3} + a_t, \\ \sigma_t^2 &= 0.002 + 0.185a_{t-1}^2 + 0.301a_{t-2}^2 + 0.197a_{t-3}^2, \end{aligned} \tag{4.7}$$

where all estimates but the coefficients of R_{t-2} and R_{t-3} are highly significant. The standardized residual series of the model shows no serial correlations, but the squared residuals show $Q(10) = 19.78$ with a p value of 0.031. Models (4.6) and (4.7) appear to be similar, but the latter seems to fit the data better. Further study shows that an AR(1)–GARCH(1,1) model fits the data well.

4.1.2 Threshold Autoregressive (TAR) Model

This model is motivated by several nonlinear characteristics commonly observed in practice such as asymmetry in declining and rising patterns of a process. It uses piecewise linear models to obtain a better approximation of the conditional mean equation. However, in contrast to the traditional piecewise linear model that allows for model changes to occur in the "time" space, the TAR model uses threshold space to improve linear approximation. Let us start with a simple 2-regime AR(1) model:

$$x_t = \begin{cases} -1.5x_{t-1} + a_t & \text{if } x_{t-1} < 0, \\ 0.5x_{t-1} + a_t & \text{if } x_{t-1} \geq 0, \end{cases} \tag{4.8}$$

where the a_t are iid $N(0, 1)$. Here the threshold variable is x_{t-1} so that the delay is 1, and the threshold is 0. Figure 4.1 shows the time plot of a simulated series of x_t with 200 observations. A horizontal line of zero is added to the plot, which illustrates several characteristics of TAR models. First, despite the coefficient -1.5 in the first regime, the process x_t is geometrically ergodic and stationary. In fact, the necessary and sufficient condition for model (4.8) to be geometrically ergodic is $\phi_1^{(1)} < 1$, $\phi_1^{(2)} < 1$, and $\phi_1^{(1)}\phi_1^{(2)} < 1$, where $\phi_1^{(i)}$ is the AR coefficient of regime i; see Petruccelli and Woolford (1984) and Chen and Tsay (1991). Ergodicity is an important concept in time series analysis. For example, the statistical theory

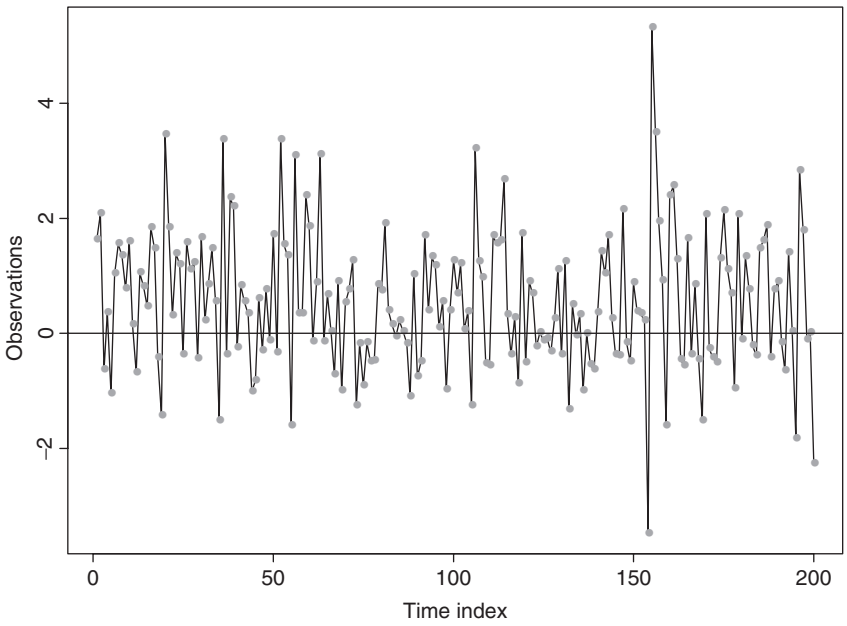

Figure 4.1 Time plot of simulated 2-regime TAR(1) series.

showing that the sample mean $\bar{x} = (\sum_{t=1}^{T} x_t)/T$ of x_t converges to the mean of x_t is referred to as the *ergodic theorem*, which can be regarded as the counterpart of the central limit theory for the iid case. Second, the series exhibits an asymmetric increasing and decreasing pattern. If x_{t-1} is negative, then x_t tends to switch to a positive value due to the negative and explosive coefficient -1.5. Yet when x_{t-1} is positive, it tends to take multiple time indexes for x_t to reduce to a negative value. Consequently, the time plot of x_t shows that regime 2 has more observations than regime 1, and the series contains large upward jumps when it becomes negative. The series is therefore not time reversible. Third, the model contains no constant terms, but $E(x_t)$ is not zero. The sample mean of the particular realization is 0.61 with a standard deviation of 0.07. In general, $E(x_t)$ is a weighted average of the conditional means of the two regimes, which are nonzero. The weight for each regime is simply the probability that x_t is in that regime under its stationary distribution. It is also clear from the discussion that, for a TAR model to have zero mean, nonzero constant terms in some of the regimes are needed. This is very different from a stationary linear model for which a nonzero constant implies that the mean of x_t is not zero.

A time series x_t is said to follow a k-regime self-exciting TAR (SETAR) model with threshold variable x_{t-d} if it satisfies

$$x_t = \phi_0^{(j)} + \phi_1^{(j)} x_{t-1} - \cdots - \phi_p^{(j)} x_{t-p} + a_t^{(j)}, \qquad \text{if} \quad \gamma_{j-1} \le x_{t-d} < \gamma_j, \quad (4.9)$$

where k and d are positive integers, $j = 1, \ldots, k$, γ_i are real numbers such that $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_{k-1} < \gamma_k = \infty$, the superscript (j) is used to signify the regime, and $\{a_t^{(j)}\}$ are iid sequences with mean 0 and variance σ_j^2 and are mutually independent for different j. The parameter d is referred to as the *delay parameter* and γ_j are the *thresholds*. Here it is understood that the AR models are different for different regimes; otherwise, the number of regimes can be reduced. Equation (4.9) says that a SETAR model is a piecewise linear AR model in the threshold space. It is similar in spirit to the usual piecewise linear models in regression analysis, where model changes occur in the order in which observations are taken. The SETAR model is nonlinear provided that $k > 1$.

Properties of general SETAR models are hard to obtain, but some of them can be found in Tong (1990), Chan (1993), Chan and Tsay (1998), and the references therein. In recent years, there is increasing interest in TAR models and their applications; see, for instance, Hansen (1997), Tsay (1998), and Montgomery et al. (1998). Tsay (1989) proposed a testing and modeling procedure for univariate SETAR models. The model in Eq. (4.9) can be generalized by using a threshold variable z_t that is measurable with respect to F_{t-1} (i.e., a function of elements of F_{t-1}). The main requirements are that z_t is stationary with a continuous distribution function over a compact subset of the real line and that z_{t-d} is known at time t. Such a generalized model is referred to as an *open-loop TAR model*.

Example 4.2. To demonstrate the application of TAR models, consider the U.S. monthly civilian unemployment rate, seasonally adjusted and measured in

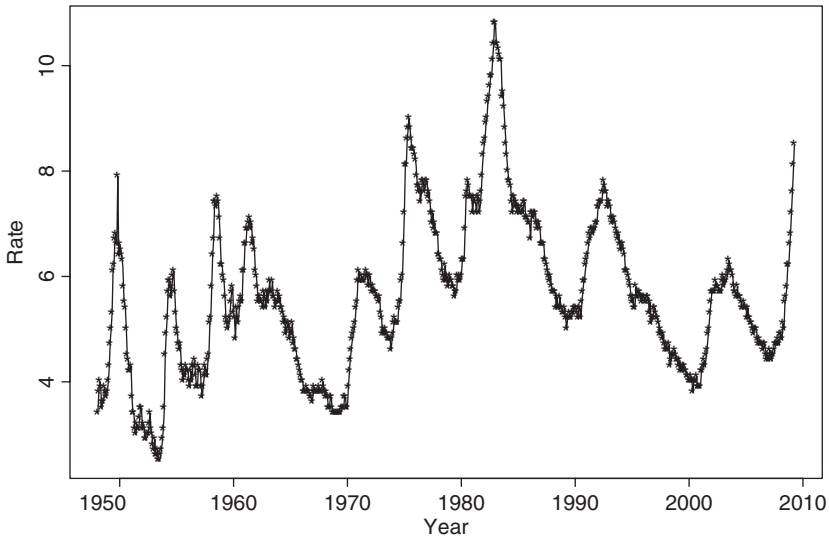

Figure 4.2 Time plot of monthly U.S. civilian unemployment rate, seasonally adjusted, from January 1948 to March 2009.

percentage, from January 1948 to March 2009 for 735 observations. The data are obtained from the Bureau of Labor Statistics, Department of Labor, and are shown in Figure 4.2. The plot shows two main characteristics of the data. First, there appears to be a slow but upward trend in the overall unemployment rate. Second, the unemployment rate tends to increase rapidly and decrease slowly. Thus, the series is not time reversible and may not be unit-root stationary, either.

Because the sample autocorrelation function decays slowly, we employ the first differenced series $y_t = (1 - B)u_t$ in the analysis, where u_t is the monthly unemployment rate. Using univariate ARIMA models, we obtain the model

$$(1 - 1.13B + 0.27B^2)(1 - 0.51B^{12})y_t = (1 - 1.12B + 0.44B^2)(1 - 0.82B^{12})a_t, \tag{4.10}$$

where $\hat{\sigma}_a = 0.187$ and all estimates but the AR(2) coefficient are statistically significant at the 5% level. The t ratio of the estimate of AR(2) coefficient is -1.66. The residuals of model (4.10) give $Q(12) = 12.3$ and $Q(24) = 25.5$, respectively. The corresponding p values are 0.056 and 0.11, respectively, based on χ^2 distributions with 6 and 18 degrees of freedom. Thus, the fitted model adequately describes the serial dependence of the data. Note that the seasonal AR and MA coefficients are highly significant with standard error 0.049 and 0.035, respectively, even though the data were seasonally adjusted. The adequacy of seasonal adjustment deserves further study. Using model (4.10), we obtain the 1-step-ahead forecast of 8.8 for the April 2009 unemployment rate, which is close to the actual data of 8.9.

To model nonlinearity in the data, we employ TAR models and obtain the model

$$y_t = \begin{cases} 0.083y_{t-2} + 0.158y_{t-3} + 0.118y_{t-4} - 0.180y_{t-12} + a_{1t} & \text{if } y_{t-1} \le 0.1, \\ 0.421y_{t-2} + 0.239y_{t-3} - 0.127y_{t-12} + a_{2t} & \text{if } y_{t-1} > 0.1, \end{cases} \tag{4.11}$$

where the standard errors of a_{it} are 0.180 and 0.217, respectively, the standard errors of the AR parameters in regime 1 are 0.046, 0.043, 0.042, and 0.037, whereas those of the AR parameters in regime 2 are 0.054, 0.057, and 0.075, respectively. The number of data points in regimes 1 and 2 are 460 and 262, respectively. The standardized residuals of model (4.11) only shows some minor serial correlation at lag 12. Based on the fitted TAR model, the dynamic dependence in the data appears to be stronger when the change in monthly unemployment rate is greater than 0.1%. This is understandable because a substantial increase in the unemployment rate is indicative of weakening in the U.S. economy, and policy makers might be more inclined to take action to help the economy, which in turn may affect the dynamics of the unemployment rate series. Consequently, model (4.11) is capable of describing the time-varying dynamics of the U.S. unemployment rate.

The MA representation of model (4.10) is

$$\psi(B) \approx 1 + 0.01B + 0.18B^2 + 0.20B^3 + 0.18B^4 + 0.15B^5 + \cdots.$$

It is then not surprising to see that no y_{t-1} term appears in model (4.11).

As mentioned in Chapter 3, threshold models can be used in finance to handle the asymmetric responses in volatility between positive and negative returns. The models can also be used to study arbitrage tradings in index futures and cash prices; see Chapter 8 on multivariate time series analysis. Here we focus on volatility modeling and introduce an alternative approach to parameterization of TGARCH models. In some applications, this new general TGARCH model fares better than the GJR model of Chapter 3.

Example 4.3. Consider the daily log returns, in percentage and including dividends, of IBM stock from July 3, 1962, to December 31, 2003, for 10,446 observations. Figure 4.3 shows the time plot of the series, which is one of the longer return series analyzed in the book. The volatility seems to be larger in the latter years of the data. Because general TGARCH models are used in the analysis, we use the SCA package to perform estimation in this example.

If GARCH models of Chapter 3 are entertained, we obtain the following AR(2)–GARCH(1,1) model for the series:

$$\begin{aligned} r_t &= 0.062 - 0.024r_{t-2} + a_t, \qquad a_t = \sigma_t\epsilon_t, \\ \sigma_t^2 &= 0.037 + 0.077a_{t-1}^2 + 0.913\sigma_{t-1}^2, \end{aligned} \tag{4.12}$$

where r_t is the log return, $\{\epsilon_t\}$ is a Gaussian white noise sequence with mean zero and variance 1.0, the standard errors of the parameters in the mean equation

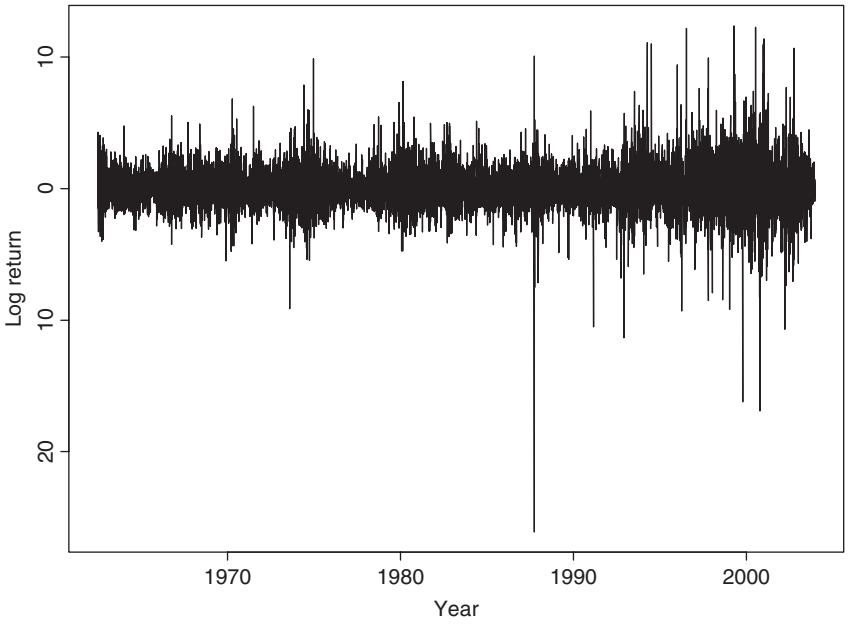

Figure 4.3 Time plot of daily log returns for IBM stock from July 3, 1962, to December 31, 2003.

are 0.015 and 0.010, and those of the volatility equation are 0.004, 0.003, and 0.003, respectively. All estimates are statistically significant at the 5% level. The Ljung–Box statistics of the standardized residuals give $Q(10) = 5.19(0.82)$ and $Q(20) = 24.38(0.18)$, where the number in parentheses denotes the p value obtained using χ^2_{m-1} distribution because of the estimated AR(2) coefficient. For the squared standardized residuals, we obtain $Q(10) = 11.67(0.31)$ and $Q(20) = 18.25(0.57)$. The model is adequate in modeling the serial dependence and conditional heteroscedasticity of the data. But the unconditional mean for r_t of model (4.12) is 0.060, which is substantially larger than the sample mean 0.039, indicating that the model might be misspecified.

Next, we employ the TGARCH model of Chapter 3 and obtain

$$\begin{aligned} r_t &= 0.014 - 0.028 r_{t-2} + a_t, \qquad a_t = \sigma_t \epsilon_t, \\ \sigma_t^2 &= 0.075 + 0.081 P_{t-1} a_{t-1}^2 + 0.157 N_{t-1} a_{t-1}^2 + 0.863 \sigma_{t-1}^2, \end{aligned} \tag{4.13}$$

where $P_{t-1} = 1 - N_{t-1}$, N_{t-1} is the indicator for negative a_{t-1} such that $N_{t-1} = 1$ if $a_{t-1} < 0$ and $= 0$ otherwise, the standard errors of the parameters in the mean equation are 0.013 and 0.009, and those of the volatility equation are 0.007, 0.008, 0.010, and 0.010, respectively. All estimates except the constant term of the mean equation are significant. Let $\tilde{a}_t$ be the standardized residuals of model (4.13). We have $Q(10) = 2.47(0.98)$ and $Q(20) = 25.90(0.13)$ for the $\{\tilde{a}_t\}$ series

and $Q(10) = 97.07(0.00)$ and $Q(20) = 170.3(0.00)$ for $\{\tilde{a}_t^2\}$. The model fails to describe the conditional heteroscedasticity of the data.

The idea of TAR models can be used to refine the prior TGARCH model by allowing for increased flexibility in modeling the asymmetric response in volatility. More specifically, we consider an AR(2)−TAR−GARCH(1,1) model for the series and obtain

$$\begin{aligned} r_t &= 0.033 - 0.023 r_{t-2} + a_t, \quad a_t = \sigma_t \epsilon_t, \\ \sigma_t^2 &= 0.075 + 0.041 a_{t-1}^2 + 0.903 \sigma_{t-1}^2 \\ &\quad + (0.030 a_{t-1}^2 + 0.062 \sigma_{t-1}^2) N_{t-1}, \end{aligned} \tag{4.14}$$

where N_{t-1} is defined in Eq. (4.13). All estimates in model (4.14) are significantly different from zero at the usual 1% level. Let $\hat{a}_t$ be the standardized residuals of model (4.14). We obtain $Q(10) = 6.09(0.73)$ and $Q(20) = 25.29(0.15)$ for $\{\hat{a}_t\}$ and $Q(10) = 13.54(0.20)$ and $Q(20) = 19.56(0.49)$ for $\{\hat{a}_t^2\}$. Thus, model (4.14) is adequate in modeling the serial correlation and conditional heteroscedasticity of the daily log returns of IBM stock considered. The unconditional mean return of model (4.14) is 0.033, which is much closer to the sample mean 0.039 than those implied by models (4.12) and (4.13). Comparing the two fitted TGARCH models, we see that the asymmetric behavior in daily IBM stock volatility is much stronger than what is allowed in a GJR model. Specifically, the coefficient of σ_{t-1}^2 also depends on the sign of a_{t-1}. Note that model (4.14) can be further refined by imposing the constraint that the sum of the coefficients of a_{t-1}^2 and σ_{t-1}^2 is one when $a_{t-1} < 0$.

Remark. A RATS program to estimate the AR(2)−TAR−GARCH(1,1) model used is given in Appendix A. The results might be slightly different from those of SCA given in the text. □

4.1.3 Smooth Transition AR (STAR) Model

A criticism of the SETAR model is that its conditional mean equation is not continuous. The thresholds $\{\gamma_j\}$ are the discontinuity points of the conditional mean function μ_t. In response to this criticism, smooth TAR models have been proposed; see Chan and Tong (1986) and Teräsvirta (1994) and the references therein. A time series x_t follows a 2-regime STAR(p) model if it satisfies

$$x_t = c_0 + \sum_{i=1}^{p} \phi_{0,i} x_{t-i} + F\left(\frac{x_{t-d} - \Delta}{s}\right)\left(c_1 + \sum_{i=1}^{p} \phi_{1,i} x_{t-i}\right) + a_t, \tag{4.15}$$

where d is the delay parameter, Δ and s are parameters representing the location and scale of model transition, and $F(\cdot)$ is a smooth transition function. In practice, $F(\cdot)$ often assumes one of three forms—namely, logistic, exponential, or a cumulative

distribution function. From Eq. (4.15) and with $0 \leq F(\cdot) \leq 1$, the conditional mean of a STAR model is a weighted linear combination between the following two equations:

$$\mu_{1t} = c_0 + \sum_{i=1}^{p} \phi_{0,i} x_{t-i},$$

$$\mu_{2t} = (c_0 + c_1) + \sum_{i=1}^{p} (\phi_{0,i} + \phi_{1,i}) x_{t-i}.$$

The weights are determined in a continuous manner by $F[(x_{t-d} - \Delta)/s]$. The prior two equations also determine properties of a STAR model. For instance, a prerequisite for the stationarity of a STAR model is that all zeros of both AR polynomials are outside the unit circle. An advantage of the STAR model over the TAR model is that the conditional mean function is differentiable. However, experience shows that the transition parameters Δ and s of a STAR model are hard to estimate. In particular, most empirical studies show that standard errors of the estimates of Δ and s are often quite large, resulting in t ratios of about 1.0; see Teräsvirta (1994). This uncertainty leads to various complications in interpreting an estimated STAR model.

Example 4.4. To illustrate the application of STAR models in financial time series analysis, we consider the monthly simple stock returns for Minnesota Mining and Manufacturing (3M) Company from February 1946 to December 2008. If ARCH models are entertained, we obtain the following ARCH(2) model:

$$R_t = 0.013 + a_t, \qquad a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = 0.003 + 0.088 a_{t-1}^2 + 0.109 a_{t-2}^2, \tag{4.16}$$

where standard errors of the estimates are 0.002, 0.0003, 0.047, and 0.050, respectively. As discussed before, such an ARCH model fails to show the asymmetric responses of stock volatility to positive and negative prior shocks. The STAR model provides a simple alternative that may overcome this difficulty. Applying STAR models to the monthly returns of 3M stock, we obtain the model

$$\begin{aligned} R_t &= 0.015 + a_t, \qquad a_t = \sigma_t \epsilon_t, \\ \sigma_t^2 &= (0.003 + 0.205 a_{t-1}^2 + 0.092 a_{t-2}^2) + \frac{0.001 - 0.239 a_{t-1}^2}{1 + \exp(-1000 a_{t-1})}, \end{aligned} \tag{4.17}$$

where the standard error of the constant term in the mean equation is 0.002 and the standard errors of the estimates in the volatility equation are 0.0002, 0.074, 0.043, 0.0004, and 0.080, respectively. The scale parameter 1000 of the logistic transition function is fixed a priori to simplify the estimation. This STAR model provides some support for asymmetric responses to positive and negative prior shocks. For

a large negative a_{t-1}, the volatility model approaches the ARCH(2) model

$$\sigma_t^2 = 0.003 + 0.205a_{t-1}^2 + 0.092a_{t-2}^2.$$

Yet for a large positive a_{t-1}, the volatility process behaves like the ARCH(2) model

$$\sigma_t^2 = 0.004 - 0.034a_{t-1}^2 + 0.092a_{t-2}^2.$$

The negative coefficient of a_{t-1}^2 in the prior model is counterintuitive, but the magnitude is small. As a matter of fact, for a large positive shock a_{t-1}, the ARCH effects appear to be weak even though the parameter estimates remain statistically significant. The results shown are obtained using the command `optim` in R. A RATS program for estimating the STAR model is given in Appendix A.

R Program for Estimating the STAR Model Used

```
> da=read.table("m-3m4608.txt",header=T)
> rtn=da[,2]
> source("star.R")
> par=c(.001,.002,.256,.141,.002,-.314)
> m2=optim(par,star,method=c("BFGS"),hessian=T)

# function to calculate the likelihood of a STAR model.
star <- function(par){
f = 0
T1=length(rtn)
h=c(1,1)
at=c(0,0)
for (t in 3:T1){
resi = rtn[t]-par[1]
at=c(at,resi)
sig=par[2]+par[3]*at[t-1]^2+par[4]*at[t-2]^2
sig1=par[5]+par[6]*at[t-1]^2
tt=sqrt(sig+sig1/(1+exp(-1000*at[t-1])))
h=c(h,tt)
x=resi/tt
f=f+log(tt)+0.5*x*x
}
f
}
```

4.1.4 Markov Switching Model

The idea of using probability switching in nonlinear time series analysis is discussed in Tong (1983). Using a similar idea, but emphasizing aperiodic transition between various states of an economy, Hamilton (1989) considers the Markov switching

autoregressive (MSA) model. Here the transition is driven by a hidden two-state Markov chain. A time series x_t follows an MSA model if it satisfies

$$x_t = \begin{cases} c_1 + \sum_{i=1}^{p} \phi_{1,i} x_{t-i} + a_{1t} & \text{if } s_t = 1, \\ c_2 + \sum_{i=1}^{p} \phi_{2,i} x_{t-i} + a_{2t} & \text{if } s_t = 2, \end{cases} \tag{4.18}$$

where s_t assumes values in {1,2} and is a first-order Markov chain with transition probabilities

$$P(s_t = 2|s_{t-1} = 1) = w_1, \qquad P(s_t = 1|s_{t-1} = 2) = w_2.$$

The innovational series $\{a_{1t}\}$ and $\{a_{2t}\}$ are sequences of iid random variables with mean zero and finite variance and are independent of each other. A small w_i means that the model tends to stay longer in state i. In fact, $1/w_i$ is the expected duration of the process to stay in state i. From the definition, an MSA model uses a hidden Markov chain to govern the transition from one conditional mean function to another. This is different from that of a SETAR model for which the transition is determined by a particular lagged variable. Consequently, a SETAR model uses a deterministic scheme to govern the model transition, whereas an MSA model uses a stochastic scheme. In practice, the stochastic nature of the states implies that one is never certain about which state x_t belongs to in an MSA model. When the sample size is large, one can use some filtering techniques to draw inference on the state of x_t. Yet as long as x_{t-d} is observed, the regime of x_t is known in a SETAR model. This difference has important practical implications in forecasting. For instance, forecasts of an MSA model are always a linear combination of forecasts produced by submodels of individual states. But those of a SETAR model only come from a single regime provided that x_{t-d} is observed. Forecasts of a SETAR model also become a linear combination of those produced by models of individual regimes when the forecast horizon exceeds the delay d. It is much harder to estimate an MSA model than other models because the states are not directly observable. Hamilton (1990) uses the EM algorithm, which is a statistical method iterating between taking expectation and maximization. McCulloch and Tsay (1994) consider a Markov chain Monte Carlo (MCMC) method to estimate a general MSA model. We discuss MCMC methods in Chapter 12.

McCulloch and Tsay (1993) generalize the MSA model in Eq. (4.18) by letting the transition probabilities w_1 and w_2 be logistic, or probit, functions of some explanatory variables available at time $t - 1$. Chen, McCulloch, and Tsay (1997) use the idea of Markov switching as a tool to perform model comparison and selection between nonnested nonlinear time series models (e.g., comparing bilinear and SETAR models). Each competing model is represented by a state. This approach to select a model is a generalization of the odds ratio commonly used in Bayesian analysis. Finally, the MSA model can easily be generalized to the case of more than two states. The computational intensity involved increases rapidly, however. For more discussions of Markov switching models in econometrics, see Hamilton (1994, Chapter 22).

Example 4.5. Consider the growth rate, in percentages, of the U.S. quarterly real gross national product (GNP) from the second quarter of 1947 to the first quarter of 1991. The data are seasonally adjusted and shown in Figure 4.4, where a horizontal line of zero growth is also given. It is reassuring to see that a majority of the growth rates are positive. This series has been widely used in nonlinear analysis of economic time series. Tiao and Tsay (1994) and Potter (1995) use TAR models, whereas Hamilton (1989) and McCulloch and Tsay (1994) employ Markov switching models.

Employing the MSA model in Eq. (4.18) with $p = 4$ and using a Markov chain Monte Carlo method, which is discussed in Chapter 12, McCulloch and Tsay (1994) obtain the estimates shown in Table 4.1. The results have several interesting findings. First, the mean growth rate of the marginal model for state 1 is $0.909/(1 - 0.265 - 0.029 + 0.126 + 0.11) = 0.965$ and that of state 2 is $-0.42/(1 - 0.216 - 0.628 + 0.073 + 0.097) = -1.288$. Thus, state 1 corresponds to quarters with positive growth, or expansion periods, whereas state 2 consists of quarters with negative growth, or a contraction period. Second, the relatively large posterior standard deviations of the parameters in state 2 reflect that there are few observations in that state. This is expected as Figure 4.4 shows few quarters with negative growth. Third, the transition probabilities appear to be different for different states. The estimates indicate that it is more likely for the U.S. GNP to get out of a contraction period than to jump into one -0.286 versus 0.118. Fourth, treating $1/w_i$ as the expected duration for the process to stay in state i, we see that the expected durations for

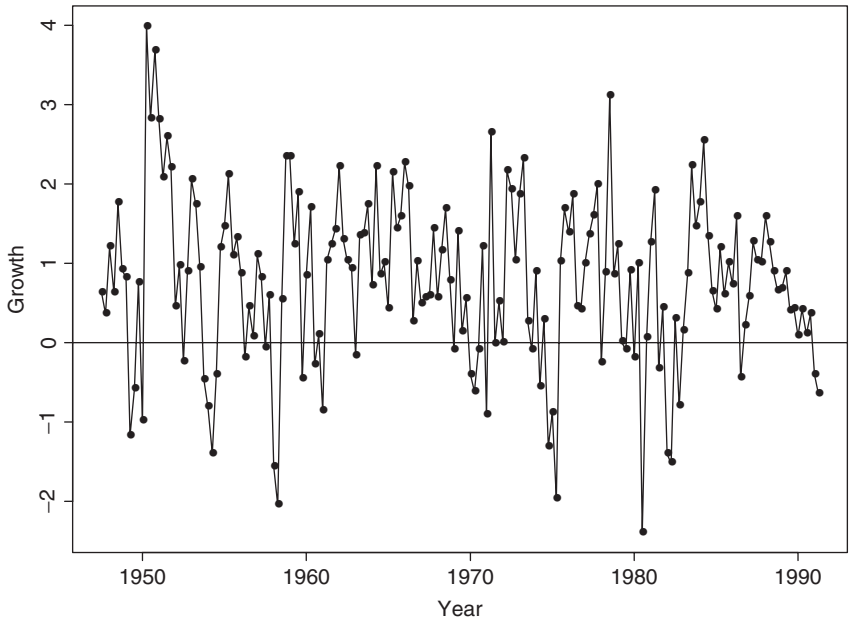

Figure 4.4 Time plot of growth rate of U.S. quarterly real GNP from 1947.II to 1991.I. Data are seasonally adjusted and in percentages.

TABLE 4.1 Estimation Results of Markov Switching Model with $p = 4$ for Growth Rate of U.S. Quarterly Real GNP, Seasonally Adjusted[a]

Parameter	c_i	ϕ_1	ϕ_2	ϕ_3	ϕ_4	σ_i	w_i
				State 1			
Estimate	0.909	0.265	0.029	−0.126	−0.110	0.816	0.118
Standard Error	0.202	0.113	0.126	0.103	0.109	0.125	0.053
				State 2			
Estimate	−0.420	0.216	0.628	−0.073	−0.097	1.017	0.286
Standard Error	0.324	0.347	0.377	0.364	0.404	0.293	0.064

[a]The estimates and their standard errors are posterior means and standard errors of a Gibbs sampling with 5000 iterations.

a contraction period and an expansion period are approximately 3.69 and 11.31 quarters. Thus, on average, a contraction in the U.S. economy lasts about a year, whereas an expansion can last for 3 years. Finally, the estimated AR coefficients of x_{t-2} differ substantially between the two states, indicating that the dynamics of the U.S. economy are different between expansion and contraction periods.

4.1.5 Nonparametric Methods

In some financial applications, we may not have sufficient knowledge to prespecify the nonlinear structure between two variables Y and X. In other applications, we may wish to take advantage of the advances in computing facilities and computational methods to explore the functional relationship between Y and X. These considerations lead to the use of nonparametric methods and techniques. Nonparametric methods, however, are not without cost. They are highly data dependent and can easily result in overfitting. Our goal here is to introduce some nonparametric methods for financial applications and some nonlinear models that make use of nonparametric methods and techniques. The nonparametric methods discussed include kernel regression, local least-squares estimation, and neural network.

The essence of nonparametric methods is *smoothing*. Consider two financial variables Y and X, which are related by

$$Y_t = m(X_t) + a_t, \tag{4.19}$$

where $m(\cdot)$ is an arbitrary, smooth, but unknown function and $\{a_t\}$ is a white noise sequence. We wish to estimate the nonlinear function $m(\cdot)$ from the data. For simplicity, consider the problem of estimating $m(\cdot)$ at a particular date for which $X = x$. That is, we are interested in estimating $m(x)$. Suppose that at $X = x$ we have repeated independent observations $y_1, \ldots, y_T$. Then the data become

$$y_t = m(x) + a_t, \qquad t = 1, \ldots, T.$$

Taking the average of the data, we have

$$\frac{\sum_{t=1}^{T} y_t}{T} = m(x) + \frac{\sum_{t=1}^{T} a_t}{T}.$$

By the law of large numbers, the average of the shocks converges to zero as T increases. Therefore, the average $\bar{y} = (\sum_{t=1}^{T} y_t)/T$ is a consistent estimate of $m(x)$. That the average $\bar{y}$ provides a consistent estimate of $m(x)$ or, alternatively, that the average of shocks converges to zero shows the power of smoothing.

In financial time series, we do not have repeated observations available at $X = x$. What we observed are $\{(y_t, x_t)\}$ for $t = 1, \ldots, T$. But if the function $m(\cdot)$ is sufficiently smooth, then the value of Y_t for which $X_t \approx x$ continues to provide accurate approximation of $m(x)$. The value of Y_t for which X_t is far away from x provides less accurate approximation for $m(x)$. As a compromise, one can use a weighted average of y_t instead of the simple average to estimate $m(x)$. The weight should be larger for those Y_t with X_t close to x and smaller for those Y_t with X_t far away from x. Mathematically, the estimate of $m(x)$ for a given x can be written as

$$\hat{m}(x) = \frac{1}{T} \sum_{t=1}^{T} w_t(x) y_t, \tag{4.20}$$

where the weights $w_t(x)$ are larger for those y_t with x_t close to x and smaller for those y_t with x_t far away from x. In Eq. (4.20), we assume that the weights sum to T. One can treat $1/T$ as part of the weights and make the weights sum to one.

From Eq. (4.20), the estimate $\hat{m}(x)$ is simply a *local weighted average* with weights determined by two factors. The first factor is the distance measure (i.e., the distance between x_t and x). The second factor is the assignment of weight for a given distance. Different ways to determine the distance between x_t and x and to assign the weight using the distance give rise to different nonparametric methods. In what follows, we discuss the commonly used kernel regression and local linear regression methods.

Kernel Regression

Kernel regression is perhaps the most commonly used nonparametric method in smoothing. The weights here are determined by a *kernel*, which is typically a probability density function, is denoted by $K(x)$, and satisfies

$$K(x) \geq 0, \qquad \int K(z)\, dz = 1.$$

However, to increase the flexibility in distance measure, one often rescales the kernel using a variable $h > 0$, which is referred to as the *bandwidth*. The rescaled kernel becomes

$$K_h(x) = \frac{1}{h} K(x/h), \qquad \int K_h(z)\, dz = 1. \tag{4.21}$$

The weight function can now be defined as

$$w_t(x) = \frac{K_h(x - x_t)}{\sum_{t=1}^{T} K_h(x - x_t)}, \tag{4.22}$$

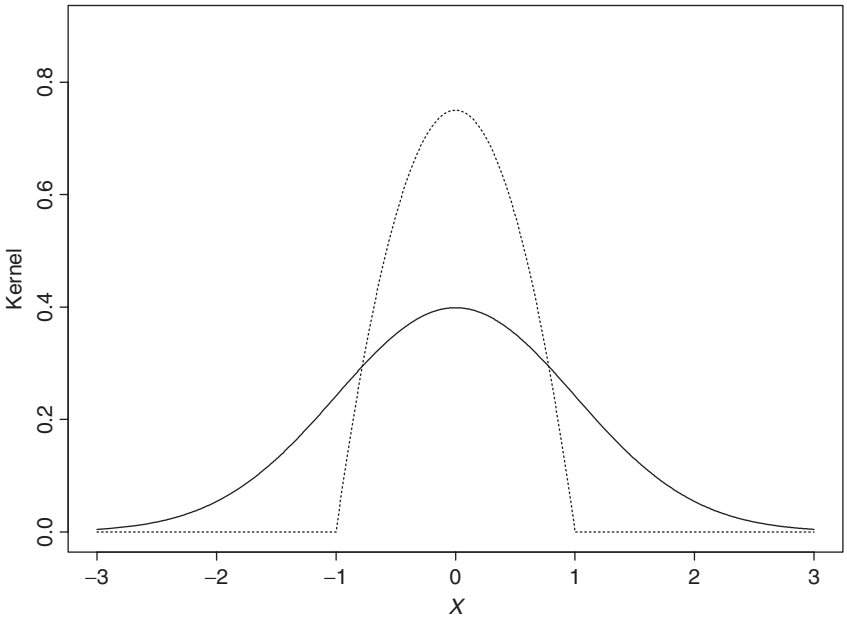

Figure 4.5 Standard normal kernel (solid line) and Epanechnikov kernel (dashed line) with bandwidth $h = 1$.

where the denominator is a normalization constant that makes the smoother adaptive to the local intensity of the X variable and ensures the weights sum to one. Plugging Eq. (4.22) into the smoothing formula (4.20), we have the well-known Nadaraya–Watson kernel estimator

$$\hat{m}(x) = \sum_{t=1}^{T} w_t(x) y_t = \frac{\sum_{t=1}^{T} K_h(x - x_t) y_t}{\sum_{t=1}^{T} K_h(x - x_t)}; \tag{4.23}$$

see Nadaraya (1964) and Watson (1964). In practice, many choices are available for the kernel $K(x)$. However, theoretical and practical considerations lead to a few choices, including the Gaussian kernel

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right)$$

and the *Epanechnikov* kernel (Epanechnikov, 1969)

$$K_h(x) = \frac{0.75}{h}\left(1 - \frac{x^2}{h^2}\right) I\left(\left|\frac{x}{h}\right| \leq 1\right),$$

where $I(A)$ is an indicator such that $I(A) = 1$ if A holds and $I(A) = 0$ otherwise. Figure 4.5 shows the Gaussian and Epanechnikov kernels for $h = 1$.

To understand the role played by the bandwidth h, we evaluate the Nadaraya–Watson estimator with the Epanechnikov kernel at the observed values $\{x_t\}$ and consider two extremes. First, if $h \to 0$, then

$$\hat{m}(x_t) \to \frac{K_h(0)y_t}{K_h(0)} = y_t,$$

indicating that small bandwidths reproduce the data. Second, if $h \to \infty$, then

$$\hat{m}(x_t) \to \frac{\sum_{t=1}^{T} K_h(0)y_t}{\sum_{t=1}^{T} K_h(0)} = \frac{1}{T}\sum_{t=1}^{T} y_t = \bar{y},$$

suggesting that large bandwidths lead to an oversmoothed curve—the sample mean. In general, the bandwidth function h acts as follows. If h is very small, then the weights focus on a few observations that are in the neighborhood around each x_t. If h is very large, then the weights will spread over a larger neighborhood of x_t. Consequently, the choice of h plays an important role in kernel regression. This is the well-known problem of bandwidth selection in kernel regression.

Bandwidth Selection

There are several approaches for bandwidth selection; see Härdle (1990) and Fan and Yao (2003). The first approach is the plug-in method, which is based on the asymptotic expansion of the mean integrated squared error (MISE) for kernel smoothers

$$\text{MISE} = E\int_{-\infty}^{\infty} [\hat{m}(x) - m(x)]^2\, dx,$$

where $m(\cdot)$ is the true function. The quantity $E[\hat{m}(x) - m(x)]^2$ of the MISE is a pointwise measure of the mean squared error (MSE) of $\hat{m}(x)$ evaluated at x. Under some regularity conditions, one can derive the *optimal bandwidth* that minimizes the MISE. The optimal bandwidth typically depends on several unknown quantities that must be estimated from the data with some preliminary smoothing. Several iterations are often needed to obtain a reasonable estimate of the optimal bandwidth. In practice, the choice of preliminary smoothing can become a problem. Fan and Yao (2003) give a normal reference bandwidth selector as

$$\hat{h}_{\text{opt}} = \begin{cases} 1.06sT^{-1/5} & \text{for the Gaussian kernel,} \\ 2.34sT^{-1/5} & \text{for the Epanechnikov kernel,} \end{cases}$$

where s is the sample standard error of the independent variable, which is assumed to be stationary.

The second approach to bandwidth selection is the leave-one-out *cross validation*. First, one observation (x_j, y_j) is left out. The remaining $T-1$ data points are used to obtain the following smoother at x_j:

$$\hat{m}_{h,j}(x_j) = \frac{1}{T-1}\sum_{t\neq j} w_t(x_j) y_t,$$

which is an estimate of y_j, where the weights $w_t(x_j)$ sum to $T-1$. Second, perform step 1 for $j = 1, \ldots, T$ and define the function

$$\text{CV}(h) = \frac{1}{T}\sum_{j=1}^{T}[y_j - \hat{m}_{h,j}(x_j)]^2 W(x_j),$$

where $W(\cdot)$ is a nonnegative weight function satisfying $\sum_{j=1}^{n} W(x_j) = T$, that can be used to down-weight the boundary points if necessary. Decreasing the weights assigned to data points close to the boundary is needed because those points often have fewer neighboring observations. The function CV(h) is called the cross-validation function because it validates the ability of the smoother to predict $\{y_t\}_{t=1}^{T}$. One chooses the bandwidth h that minimizes the CV$(\cdot)$ function.

Local Linear Regression Method

Assume that the second derivative of $m(\cdot)$ in model (4.19) exists and is continuous at x, where x is a given point in the support of $m(\cdot)$. Denote the data available by $\{(y_t, x_t)\}_{t=1}^{T}$. The local linear regression method to nonparametric regression is to find a and b that minimize

$$L(a, b) = \sum_{t=1}^{T}[y_t - a - b(x - x_t)]^2 K_h(x - x_t), \tag{4.24}$$

where $K_h(\cdot)$ is a kernel function defined in Eq. (4.21) and h is a bandwidth. Denote the resulting value of a by $\hat{a}$. The estimate of $m(x)$ is then defined as $\hat{a}$. In practice, x assumes an observed value of the independent variable. The estimate $\hat{b}$ can be used as an estimate of the first derivative of $m(\cdot)$ evaluated at x.

Under the least-squares theory, Eq. (4.24) is a weighted least-squares problem and one can derive a closed-form solution for a. Specifically, taking the partial derivatives of $L(a, b)$ with respect to both a and b and equating the derivatives to zero, we have a system of two equations with two unknowns:

$$\sum_{t=1}^{T} K_h(x - x_t) y_t = a\sum_{t=1}^{T} K_h(x - x_t) + b\sum_{t=1}^{T}(x - x_t) K_h(x - x_t),$$

$$\sum_{t=1}^{T} y_t(x - x_t) K_h(x - x_t) = a\sum_{t=1}^{T}(x - x_t) K_h(x - x_t) + b\sum_{t=1}^{T}(x - x_t)^2 K_h(x - x_t).$$

Define

$$s_{T,\ell} = \sum_{t=1}^{T} K_h(x - x_t)(x - x_t)^{\ell}, \qquad \ell = 0, 1, 2.$$

The prior system of equations becomes

$$\begin{bmatrix} s_{T,0} & s_{T,1} \\ s_{T,1} & s_{T,2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} K_h(x - x_t) y_t \\ \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) y_t \end{bmatrix}.$$

Consequently, we have

$$\hat{a} = \frac{s_{T,2} \sum_{t=1}^{T} K_h(x - x_t) y_t - s_{T,1} \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) y_t}{s_{T,0} s_{T,2} - s_{T,1}^2}.$$

The numerator and denominator of the prior fraction can be further simplified as

$$\begin{aligned} s_{T,2} \sum_{t=1}^{T} K_h(x - x_t) y_t - s_{T,1} \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) y_t & \\ = \sum_{t=1}^{T} \{K_h(x - x_t)[s_{T,2} - (x - x_t) s_{T,1}]\} y_t. & \\ s_{T,0} s_{T,2} - s_{T,1}^2 = \sum_{t=1}^{T} K_h(x - x_t) s_{T,2} - \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) s_{T,1} & \\ = \sum_{t=1}^{T} K_h(x - x_t)[s_{T,2} - (x - x_t) s_{T,1}]. & \end{aligned}$$

In summary, we have

$$\hat{a} = \frac{\sum_{t=1}^{T} w_t y_t}{\sum_{t=1}^{T} w_t}, \tag{4.25}$$

where w_t is defined as

$$w_t = K_h(x - x_t)[s_{T,2} - (x - x_t) s_{T,1}].$$

In practice, to avoid possible zero in the denominator, we use the following $\hat{m}(x)$ to estimate $m(x)$:

$$\hat{m}(x) = \frac{\sum_{t=1}^{T} w_t y_t}{\sum_{t=1}^{T} w_t + 1/T^2}. \tag{4.26}$$

Notice that a nice feature of Eq. (4.26) is that the weight w_t satisfies

$$\sum_{t=1}^{T}(x - x_t)w_t = 0.$$

Also, if one assumes that $m(\cdot)$ of Eq. (4.19) has the first derivative and finds the minimizer of

$$\sum_{t=1}^{T}(y_t - a)^2 K_h(x - x_t),$$

then the resulting estimator is the Nadaraya–Watson estimator mentioned earlier. In general, if one assumes that $m(x)$ has a bounded kth derivative, then one can replace the linear polynomial in Eq. (4.24) by a $(k-1)$-order polynomial. We refer to the estimator in Eq. (4.26) as the local linear regression smoother. Fan (1993) shows that, under some regularity conditions, the local linear regression estimator has some important sampling properties. The selection of bandwidth can be carried out via the same methods as before.

Time Series Application

In time series analysis, the explanatory variables are often the lagged values of the series. Consider the simple case of a single explanatory variable. Here model (4.19) becomes

$$x_t = m(x_{t-1}) + a_t,$$

and the kernel regression and local linear regression method discussed before are directly applicable. When multiple explanatory variables exist, some modifications are needed to implement the nonparametric methods. For the kernel regression, one can use a multivariate kernel such as a multivariate normal density function with a prespecified covariance matrix:

$$K_h(\boldsymbol{x}) = \frac{1}{(h\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2h^2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right),$$

where p is the number of explanatory variables and $\boldsymbol{\Sigma}$ is a prespecified positive-definite matrix. Alternatively, one can use the product of univariate kernel functions as a multivariate kernel—for example,

$$K_h(\boldsymbol{x}) = \prod_{i=1}^{p} \frac{0.75}{h_i}\left(1 - \frac{x_i^2}{h_i^2}\right) I\left(\left|\frac{x_i}{h_i}\right| < 1\right).$$

This latter approach is simple, but it overlooks the relationship between the explanatory variables.

Example 4.6. To illustrate the application of nonparametric methods in finance, consider the weekly 3-month Treasury bill secondary market rate from 1970 to 1997 for 1461 observations. The data are obtained from the Federal Reserve Bank of St. Louis and are shown in Figure 4.6. This series has been used in the literature as an example of estimating stochastic diffusion equations using discretely observed data. See references in Chapter 6. Here we consider a simple model

$$y_t = \mu(x_{t-1})\,dt + \sigma(x_{t-1})\,dw_t,$$

where x_t is the 3-month Treasury bill rate, $y_t = x_t - x_{t-1}$, w_t is a standard Brownian motion, and $\mu(\cdot)$ and $\sigma(\cdot)$ are smooth functions of x_{t-1}, and apply the local smoothing function `lowess` of R or S-Plus to obtain nonparametric estimates of $\mu(\cdot)$ and $\sigma(\cdot)$; see Cleveland (1979). For simplicity, we use $|y_t|$ as a proxy of the volatility of x_t.

For the simple model considered, $\mu(x_{t-1})$ is the conditional mean of y_t given x_{t-1}, that is, $\mu(x_{t-1}) = E(y_t|x_{t-1})$. Figure 4.7(a) shows the scatterplot of $y(t)$ versus x_{t-1}. The plot also contains the local smooth estimate of $\mu(x_{t-1})$ obtained by `lowess` of R or S-Plus. The estimate is essentially zero. However, to better understand the estimate, Figure 4.7(b) shows the estimate $\hat{\mu}(x_{t-1})$ on a finer scale. It is interesting to see that $\hat{\mu}(x_{t-1})$ is positive when x_{t-1} is small but becomes negative when x_{t-1} is large. This is in agreement with the common sense that

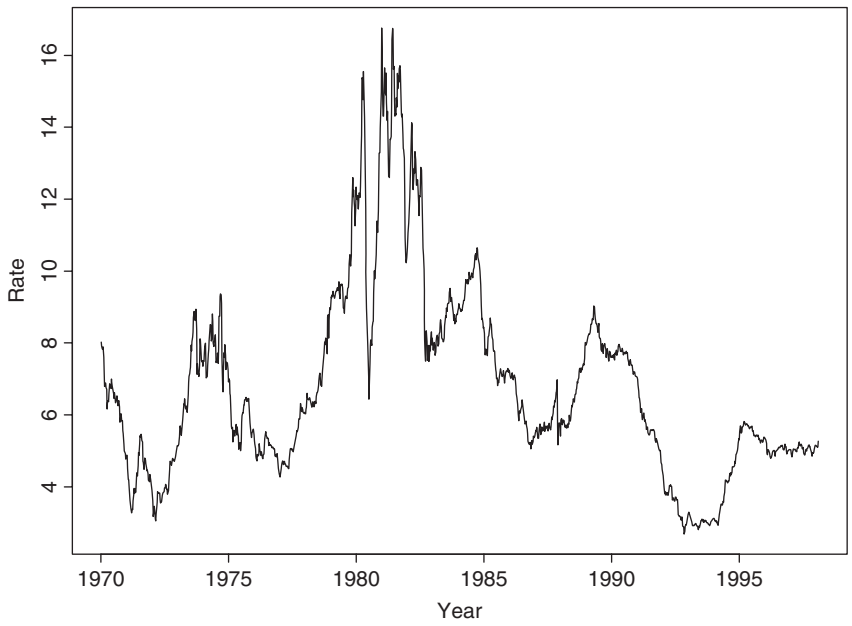

Figure 4.6 Time plot of U.S. weekly 3-month Treasury bill rate in secondary market from 1970 to 1997.

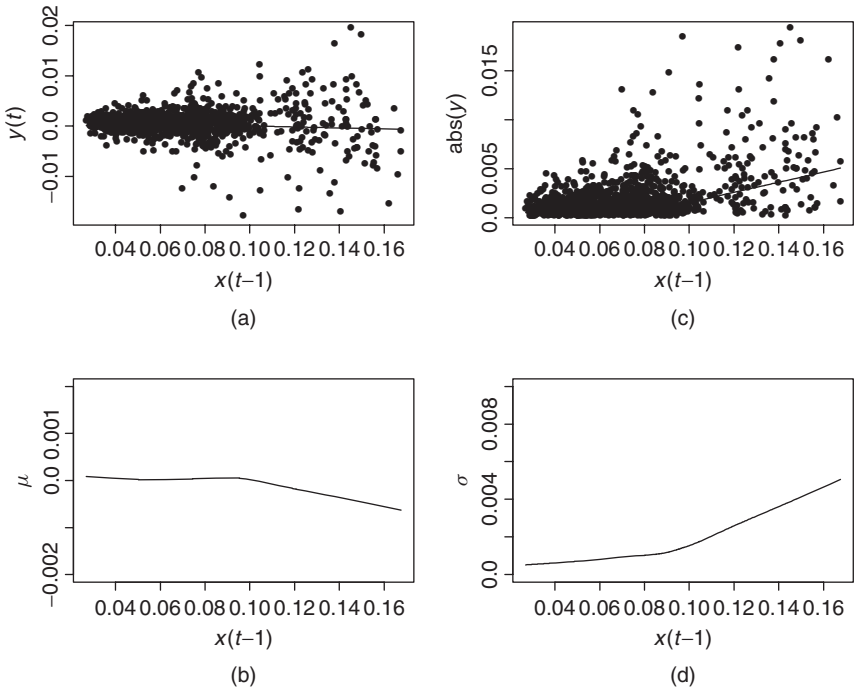

Figure 4.7 Estimation of conditional mean and volatility of weekly 3-month Treasury bill rate via a local smoothing method: (a) y_t vs. x_{t-1}, where $y_t = x_t - x_{t-1}$ and x_t is interest rate; (b) estimate of $\mu(x_{t-1})$; (c) $|y_t|$ vs. x_{t-1}; and (d) estimate of $\sigma(x_{t-1})$.

when the interest rate is high, it is expected to come down, and when the rate is low, it is expected to increase. Figure 4.7(c) shows the scatterplot of $|y(t)|$ versus x_{t-1} and the estimate of $\hat{\sigma}(x_{t-1})$ via `lowess`. The plot confirms that the higher the interest rate, the larger the volatility. Figure 4.7(d) shows the estimate $\hat{\sigma}(x_{t-1})$ on a finer scale. Clearly, the volatility is an increasing function of x_{t-1} and the slope seems to accelerate when x_{t-1} is approaching 10%. This example demonstrates that simple nonparametric methods can be helpful in understanding the dynamic structure of a financial time series.

R and S-Plus Commands Used in Example 4.6

```
> z1=read.table('w-3mtbs7097.txt',header=T)
> x=z1[4,1:1460]/100
> y=(z1[4,2:1461]-z1[4,1:1460])/100
> par(mfcol=c(2,2))
> plot(x,y,pch='*',xlab='x(t-1)',ylab='y(t)')
> lines(lowess(x,y))
> title(main='(a) y(t) vs x(t-1)')
> fit=lowess(x,y)
```

```
> plot(fit$x,fit$y,xlab='x(t-1)',ylab='mu',type='l',
+ ylim=c(-.002,.002))
> title(main='(b) Estimate of mu(.)')
> plot(x,abs(y),pch='*',xlab='x(t-1)',ylab='abs(y)')
> lines(lowess(x,abs(y)))
> title(main='(c) abs(y) vs x(t-1)')
> fit2=lowess(x,abs(y))
> plot(fit2$x,fit2$y,type='l',xlab='x(t-1)',ylab='sigma',
+ ylim=c(0,.01))
> title(main='(d) Estimate of sigma(.)')
```

The following nonlinear models are derived with the help of nonparametric methods.

4.1.6 Functional Coefficient AR Model

Recent advances in nonparametric techniques enable researchers to relax parametric constraints in proposing nonlinear models. In some cases, nonparametric methods are used in a preliminary study to help select a parametric nonlinear model. This is the approach taken by Chen and Tsay (1993a) in proposing the functional coefficient autoregressive (FAR) model that can be written as

$$x_t = f_1(X_{t-1})x_{t-1} + \cdots + f_p(X_{t-1})x_{t-p} + a_t, \tag{4.27}$$

where $X_{t-1} = (x_{t-1}, \ldots, x_{t-k})'$ is a vector of lagged values of x_t. If necessary, X_{t-1} may also include other explanatory variables available at time $t-1$. The functions $f_i(\cdot)$ of Eq. (4.27) are assumed to be continuous, even twice differentiable, almost surely with respect to their arguments. Most of the nonlinear models discussed before are special cases of the FAR model. In application, one can use nonparametric methods such as kernel regression or local linear regression to estimate the functional coefficients $f_i(\cdot)$, especially when the dimension of X_{t-1} is low (e.g., X_{t-1} is a scalar). Recently, Cai, Fan, and Yao (2000) applied the local linear regression method to estimate $f_i(\cdot)$ and showed that substantial improvements in 1-step-ahead forecasts can be achieved by using FAR models.

4.1.7 Nonlinear Additive AR Model

A major difficulty in applying nonparametric methods to nonlinear time series analysis is the "curse of dimensionality." Consider a general nonlinear AR(p) process $x_t = f(x_{t-1}, \ldots, x_{t-p}) + a_t$. A direct application of nonparametric methods to estimate $f(\cdot)$ would require p-dimensional smoothing, which is hard to do when p is large, especially if the number of data points is not large. A simple, yet effective way to overcome this difficulty is to entertain an additive model that only requires lower dimensional smoothing. A time series x_t follows a nonlinear additive AR

(NAAR) model if

$$x_t = f_0(t) + \sum_{i=1}^{p} f_i(x_{t-i}) + a_t, \tag{4.28}$$

where the $f_i(\cdot)$ are continuous functions almost surely. Because each function $f_i(\cdot)$ has a single argument, it can be estimated nonparametrically using one-dimensional smoothing techniques and hence avoids the curse of dimensionality. In application, an iterative estimation method that estimates $f_i(\cdot)$ nonparametrically conditioned on estimates of $f_j(\cdot)$ for all $j \neq i$ is used to estimate a NAAR model; see Chen and Tsay (1993b) for further details and examples of NAAR models.

The additivity assumption is rather restrictive and needs to be examined carefully in application. Chen, Liu, and Tsay (1995) consider test statistics for checking the additivity assumption.

4.1.8 Nonlinear State-Space Model

Making using of recent advances in MCMC methods (Gelfand and Smith, 1990), Carlin, Polson, and Stoffer (1992) propose a Monte Carlo approach for nonlinear state-space modeling. The model considered is

$$S_t = f_t(S_{t-1}) + u_t, \qquad x_t = g_t(S_t) + v_t, \tag{4.29}$$

where S_t is the state vector, $f_t(\cdot)$ and $g_t(\cdot)$ are known functions depending on some unknown parameters, $\{u_t\}$ is a sequence of iid multivariate random vectors with zero mean and nonnegative definite covariance matrix $\mathbf{\Sigma}_u$, $\{v_t\}$ is a sequence of iid random variables with mean zero and variance σ_v^2, and $\{u_t\}$ is independent of $\{v_t\}$. Monte Carlo techniques are employed to handle the nonlinear evolution of the state transition equation because the whole conditional distribution function of S_t given S_{t-1} is needed for a nonlinear system. Other numerical smoothing methods for nonlinear time series analysis have been considered by Kitagawa (1998) and the references therein. MCMC methods (or computing-intensive numerical methods) are powerful tools for nonlinear time series analysis. Their potential has not been fully explored. However, the assumption of knowing $f_t(\cdot)$ and $g_t(\cdot)$ in model (4.29) may hinder practical use of the proposed method. A possible solution to overcome this limitation is to use nonparametric methods such as the analyses considered in FAR and NAAR models to specify $f_t(\cdot)$ and $g_t(\cdot)$ before using nonlinear state-space models.

4.1.9 Neural Networks

A popular topic in modern data analysis is neural networks, which can be classified as a semiparametric method. The literature on neural networks is enormous, and its application spreads over many scientific areas with varying degrees of success;

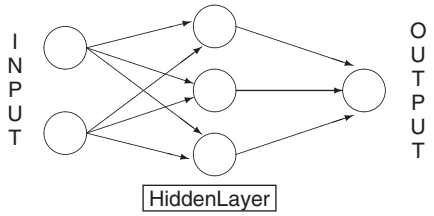

Figure 4.8 Feed-forward neural network with one hidden layer for univariate time series analysis.

see Section 2 of Ripley (1993) for a list of applications and Section 10 for remarks concerning its application in finance. Cheng and Titterington (1994) provide information on neural networks from a statistical viewpoint. In this subsection, we focus solely on the *feed-forward* neural networks in which inputs are connected to one or more *neurons*, or *nodes*, in the input layer, and these nodes are connected forward to further layers until they reach the output layer. Figure 4.8 shows an example of a simple feed-forward network for univariate time series analysis with one hidden layer. The input layer has two nodes, and the hidden layer has three. The input nodes are connected forward to each and every node in the hidden layer, and these hidden nodes are connected to the single node in the output layer. We call the network a 2–3–1 feed-forward network. More complicated neural networks, including those with feedback connections, have been proposed in the literature, but the feed-forward networks are most relevant to our study.

Feed-Forward Neural Networks

A neural network processes information from one layer to the next by an "activation function." Consider a feed-forward network with one hidden layer. The jth node in the hidden layer is defined as

$$h_j = f_j\left(\alpha_{0j} + \sum_{i \to j} w_{ij} x_i\right), \tag{4.30}$$

where x_i is the value of the ith input node, $f_j(\cdot)$ is an activation function typically taken to be the logistic function

$$f_j(z) = \frac{\exp(z)}{1 + \exp(z)},$$

α_{0j} is called the bias, the summation $i \to j$ means summing over all input nodes feeding to j, and w_{ij} are the weights. For illustration, the jth node of the hidden layer of the 2–3–1 feed-forward network in Figure 4.8 is

$$h_j = \frac{\exp(\alpha_{0j} + w_{1j}x_1 + w_{2j}x_2)}{1 + \exp(\alpha_{0j} + w_{1j}x_1 + w_{2j}x_2)}, \qquad j = 1, 2, 3. \tag{4.31}$$

For the output layer, the node is defined as

$$o = f_o\left(\alpha_{0o} + \sum_{j\to o} w_{jo}h_j\right), \tag{4.32}$$

where the activation function $f_o(\cdot)$ is either linear or a Heaviside function. If $f_o(\cdot)$ is linear, then

$$o = \alpha_{0o} + \sum_{j=1}^{k} w_{jo}h_j,$$

where k is the number of nodes in the hidden layer. By a Heaviside function, we mean $f_o(z) = 1$ if $z > 0$ and $f_o(z) = 0$ otherwise. A neuron with a Heaviside function is called a *threshold neuron*, with 1 denoting that the neuron fires its message. For example, the output of the 2–3–1 network in Figure 4.8 is

$$o = \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3,$$

if the activation function is linear; it is

$$o = \begin{cases} 1 & \text{if } \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3 > 0, \\ 0 & \text{if } \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3 \le 0, \end{cases}$$

if $f_o(\cdot)$ is a Heaviside function.

Combining the layers, the output of a feed-forward neural network can be written as

$$o = f_o\left[\alpha_{0o} + \sum_{j\to o} w_{jo}f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij}x_i\right)\right]. \tag{4.33}$$

If one also allows for direct connections from the input layer to the output layer, then the network becomes

$$o = f_o\left[\alpha_{0o} + \sum_{i\to o}\alpha_{io}x_i + \sum_{j\to o} w_{jo}f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij}x_i\right)\right], \tag{4.34}$$

where the first summation is summing over the input nodes. When the activation function of the output layer is linear, the direct connections from the input nodes to the output node represent a linear function between the inputs and output. Consequently, in this particular case model (4.34) is a generalization of linear models.

For the 2–3–1 network in Figure 4.8, if the output activation function is linear, then Eq. (4.33) becomes

$$o = \alpha_{0o} + \sum_{j=1}^{3} w_{jo} h_j,$$

where h_j is given in Eq. (4.31). The network thus has 13 parameters. If Eq. (4.34) is used, then the network becomes

$$o = \alpha_{0o} + \sum_{i=1}^{2} \alpha_{io} x_i + \sum_{j=1}^{3} w_{jo} h_j,$$

where again h_j is given in Eq. (4.31). The number of parameters of the network increases to 15.

We refer to the function in Eq. (4.33) or (4.34) as a semiparametric function because its functional form is known, but the number of nodes and their biases and weights are unknown. The direct connections from the input layer to the output layer in Eq. (4.34) mean that the network can skip the hidden layer. We refer to such a network as a *skip-layer* feed-forward network.

Feed-forward networks are known as *multilayer percetrons* in the neural network literature. They can approximate any continuous function uniformly on compact sets by increasing the number of nodes in the hidden layer; see Hornik, Stinchcombe, and White (1989), Hornik (1993), and Chen and Chen (1995). This property of neural networks is the universal approximation property of the multilayer percetrons. In short, feed-forward neural networks with a hidden layer can be seen as a way to parameterize a general continuous nonlinear function.

Training and Forecasting

Application of neural networks involves two steps. The first step is to *train* the network (i.e., to build a network, including determining the number of nodes and estimating their biases and weights). The second step is inference, especially forecasting. The data are often divided into two nonoverlapping subsamples in the training stage. The first subsample is used to estimate the parameters of a given feed-forward neural network. The network so built is then used in the second subsample to perform forecasting and compute its forecasting accuracy. By comparing the forecasting performance, one selects the network that outperforms the others as the "best" network for making inference. This is the idea of cross validation widely used in statistical model selection. Other model selection methods are also available.

In a time series application, let $\{(r_t, \boldsymbol{x}_t)|t = 1, \ldots, T\}$ be the available data for network training, where $\boldsymbol{x}_t$ denotes the vector of inputs and r_t is the series of interest (e.g., log returns of an asset). For a given network, let o_t be the output of

the network with input $\boldsymbol{x}_t$; see Eq. (4.34). Training a neural network amounts to choosing its biases and weights to minimize some fitting criterion—for example, the least squares

$$S^2 = \sum_{t=1}^{T} (r_t - o_t)^2.$$

This is a nonlinear estimation problem that can be solved by several iterative methods. To ensure the smoothness of the fitted function, some additional constraints can be added to the prior minimization problem. In the neural network literature, the *back propagation* (BP) learning algorithm is a popular method for network training. The BP method, introduced by Bryson and Ho (1969), works backward starting with the output layer and uses a gradient rule to modify the biases and weights iteratively. Appendix 2A of Ripley (1993) provides a derivation of back propagation. Once a feed-forward neural network is built, it can be used to compute forecasts in the forecasting subsample.

Example 4.7. To illustrate applications of the neural network in finance, we consider the monthly log returns, in percentages and including dividends, for IBM stock from January 1926 to December 1999. We divide the data into two subsamples. The first subsample consisting of returns from January 1926 to December 1997 for 864 observations is used for modeling. Using model (4.34) with three inputs and two nodes in the hidden layer, we obtain a 3–2–1 network for the series. The three inputs are r_{t-1}, r_{t-2}, and r_{t-3} and the biases and weights are given next:

$$\hat{r}_t = 3.22 - 1.81 f_1(\boldsymbol{r}_{t-1}) - 2.28 f_2(\boldsymbol{r}_{t-1}) - 0.09 r_{t-1} - 0.05 r_{t-2} - 0.12 r_{t-3}, \tag{4.35}$$

where $\boldsymbol{r}_{t-1} = (r_{t-1}, r_{t-2}, r_{t-3})$ and the two logistic functions are

$$f_1(\boldsymbol{r}_{t-1}) = \frac{\exp(-8.34 - 18.97 r_{t-1} + 2.17 r_{t-2} - 19.17 r_{t-3})}{1 + \exp(-8.34 - 18.97 r_{t-1} + 2.17 r_{t-2} - 19.17 r_{t-3})},$$
$$f_2(\boldsymbol{r}_{t-1}) = \frac{\exp(39.25 - 22.17 r_{t-1} - 17.34 r_{t-2} - 5.98 r_{t-3})}{1 + \exp(39.25 - 22.17 r_{t-1} - 17.34 r_{t-2} - 5.98 r_{t-3})}.$$

The standard error of the residuals for the prior model is 6.56. For comparison, we also built an AR model for the data and obtained

$$r_t = 1.101 + 0.077 r_{t-1} + a_t, \qquad \sigma_a = 6.61. \tag{4.36}$$

The residual standard error is slightly greater than that of the feed-forward model in Eq. (4.35).

Forecast Comparison

The monthly returns of IBM stock in 1998 and 1999 form the second subsample and are used to evaluate the out-of-sample forecasting performance of neural networks. As a benchmark for comparison, we use the sample mean of r_t in the first subsample as the 1-step-ahead forecast for all the monthly returns in the second subsample. This corresponds to assuming that the log monthly price of IBM stock follows a random walk with drift. The mean squared forecast error (MSFE) of this benchmark model is 91.85. For the AR(1) model in Eq. (4.36), the MSFE of 1-step-ahead forecasts is 91.70. Thus, the AR(1) model slightly outperforms the benchmark. For the 3–2–1 feed-forward network in Eq. (4.35), the MSFE is 91.74, which is essentially the same as that of the AR(1) model.

Remark. The estimation of feed-forward networks is done by using the `nnet` package of S-Plus with default starting weights; see Venables and Ripley (1999) for more information. Our limited experience shows that the estimation results vary. For the IBM stock returns used in Example 4.7, the out-of-sample MSE for a 3–2–1 network can be as low as 89.46 and as high as 93.65. If we change the number of nodes in the hidden layer, the range for the MSE becomes even wider. The S-Plus commands used in Example 4.7 are given in Appendix B. □

Example 4.8. Nice features of the feed-forward network include its flexibility and wide applicability. For illustration, we use the network with a Heaviside activation function for the output layer to forecast the direction of price movement for IBM stock considered in Example 4.7. Define a direction variable as

$$d_t = \begin{cases} 1 & \text{if } r_t \geq 0, \\ 0 & \text{if } r_t < 0. \end{cases}$$

We use eight input nodes consisting of the first four lagged values of both r_t and d_t and four nodes in the hidden layer to build an 8–4–1 feed-forward network for d_t in the first subsample. The resulting network is then used to compute the 1-step-ahead probability of an "upward movement" (i.e., a positive return) for the following month in the second subsample. Figure 4.9 shows a typical output of probability forecasts and the actual directions in the second subsample with the latter denoted by circles. A horizontal line of 0.5 is added to the plot. If we take a rigid approach by letting $\hat{d}_t = 1$ if the probability forecast is greater than or equal to 0.5 and $\hat{d}_t = 0$ otherwise, then the neural network has a successful rate of 0.58. The success rate of the network varies substantially from one estimation to another, and the network uses 49 parameters. To gain more insight, we did a simulation study of running the 8–4–1 feed-forward network 500 times and computed the number of errors in predicting the upward and downward movement using the same method as before. The mean and median of errors over the 500 runs are 11.28 and 11, respectively, whereas the maximum and minimum number of errors are 18 and 4.

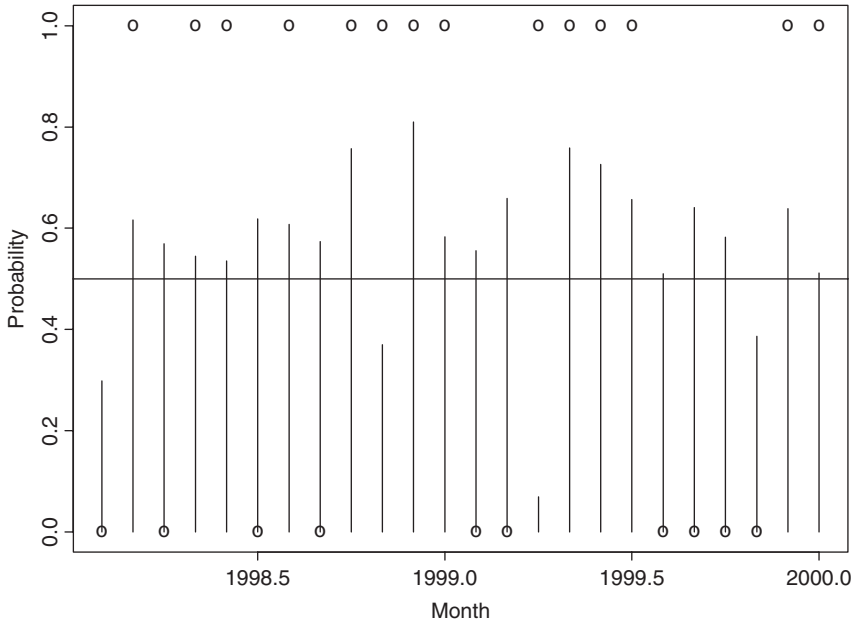

Figure 4.9 One-step-ahead probability forecasts for positive monthly return for IBM stock using an 8–4–1 feed-forward neural network. Forecasting period is from January 1998 to December 1999.

For comparison, we also did a simulation with 500 runs using a random walk with drift—that is,

$$\hat{d}_t = \begin{cases} 1 & \text{if } \hat{r}_t = 1.19 + \epsilon_t \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where 1.19 is the average monthly log return for IBM stock from January 1926 to December 1997 and $\{\epsilon_t\}$ is a sequence of iid $N(0, 1)$ random variables. The mean and median of the number of forecast errors become 10.53 and 11, whereas the maximum and minimum number of errors are 17 and 5, respectively. Figure 4.10 shows the histograms of the number of forecast errors for the two simulations. The results show that the 8–4–1 feed-forward neural network does not outperform the simple model that assumes a random walk with drift for the monthly log price of IBM stock.

4.2 NONLINEARITY TESTS

In this section, we discuss some nonlinearity tests available in the literature that have decent power against the nonlinear models considered in Section 4.1. The tests discussed include both parametric and nonparametric statistics. The Ljung–Box

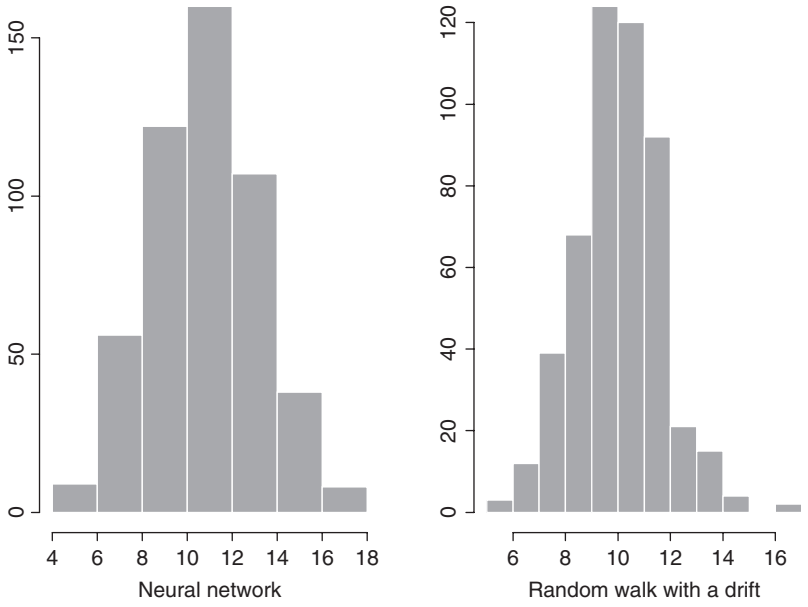

Figure 4.10 Histograms of number of forecasting errors for directional movements of monthly log returns of IBM stock. Forecasting period is from January 1998 to December 1999.

statistics of squared residuals, the bispectral test, and the Brock, Dechert, and Scheinkman (BDS) test are nonparametric methods. The RESET test (Ramsey, 1969), the F tests of Tsay (1986, 1989), and other Lagrange multiplier and likelihood ratio tests depend on specific parametric functions. Because nonlinearity may occur in many ways, there exists no single test that dominates the others in detecting nonlinearity.

4.2.1 Nonparametric Tests

Under the null hypothesis of linearity, residuals of a properly specified linear model should be independent. Any violation of independence in the residuals indicates inadequacy of the entertained model, including the linearity assumption. This is the basic idea behind various nonlinearity tests. In particular, some of the nonlinearity tests are designed to check for possible violation in quadratic forms of the underlying time series.

Q-Statistic of Squared Residuals

McLeod and Li (1983) apply the Ljung–Box statistics to the squared residuals of an ARMA(p, q) model to check for model inadequacy. The test statistic is

$$Q(m) = T(T+2)\sum_{i=1}^{m} \frac{\hat{\rho}_i^2(a_t^2)}{T-i},$$

where T is the sample size, m is a properly chosen number of autocorrelations used in the test, a_t denotes the residual series, and $\hat{\rho}_i(a_t^2)$ is the lag-i ACF of a_t^2. If the entertained linear model is adequate, $Q(m)$ is asymptotically a chi-squared random variable with $m - p - q$ degrees of freedom. As mentioned in Chapter 3, the prior Q-statistic is useful in detecting conditional heteroscedasticity of a_t and is asymptotically equivalent to the Lagrange multiplier test statistic of Engle (1982) for ARCH models; see Section 3.4.3. The null hypothesis of the test is $H_0 : \beta_1 = \cdots = \beta_m = 0$, where β_i is the coefficient of a_{t-i}^2 in the linear regression

$$a_t^2 = \beta_0 + \beta_1 a_{t-1}^2 + \cdots + \beta_m a_{t-m}^2 + e_t$$

for $t = m + 1, \ldots, T$. Because the statistic is computed from residuals (not directly from the observed returns), the number of degrees of freedom is $m - p - q$.

Bispectral Test

This test can be used to test for linearity and Gaussianity. It depends on the result that a properly normalized bispectrum of a linear time series is constant over all frequencies and that the constant is zero under normality. The bispectrum of a time series is the Fourier transform of its third-order moments. For a stationary time series x_t in Eq. (4.1), the third-order moment is defined as

$$c(u, v) = g \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+u} \psi_{k+v}, \tag{4.37}$$

where u and v are integers, $g = E(a_t^3)$, $\psi_0 = 1$, and $\psi_k = 0$ for $k < 0$. Taking Fourier transforms of Eq. (4.37), we have

$$b_3(w_1, w_2) = \frac{g}{4\pi^2} \Gamma[-(w_1 + w_2)] \Gamma(w_1) \Gamma(w_2), \tag{4.38}$$

where $\Gamma(w) = \sum_{u=0}^{\infty} \psi_u \exp(-i w u)$ with $i = \sqrt{-1}$, and w_i are frequencies. Yet the spectral density function of x_t is given by

$$p(w) = \frac{\sigma_a^2}{2\pi} |\Gamma(w)|^2,$$

where w denotes the frequency. Consequently, the function

$$b(w_1, w_2) = \frac{|b_3(w_1, w_2)|^2}{p(w_1) p(w_2) p(w_1 + w_2)} = \text{constant for all } (w_1, w_2). \tag{4.39}$$

The bispectrum test makes use of the property in Eq. (4.39). Basically, it estimates the function $b(w_1, w_2)$ in Eq. (4.39) over a suitably chosen grid of points and applies a test statistic similar to Hotelling's T^2 statistic to check the constancy of $b(w_1, w_2)$. For a linear Gaussian series, $E(a_t^3) = g = 0$ so that the bispectrum is zero for all frequencies (w_1, w_2). For further details of the bispectral test, see Priestley (1988), Subba Rao and Gabr (1984), and Hinich (1982). Limited experience shows that the test has decent power when the sample size is large.

BDS Statistic

Brock, Dechert, and Scheinkman (1987) propose a test statistic, commonly referred to as the *BDS test*, to detect the iid assumption of a time series. The statistic is, therefore, different from other test statistics discussed because the latter mainly focus on either the second- or third-order properties of x_t. The basic idea of the BDS test is to make use of a "correlation integral" popular in chaotic time series analysis. Given a k-dimensional time series X_t and observations $\{X_t\}_{t=1}^{T_k}$, define the correlation integral as

$$C_k(\delta) = \lim_{T_k \to \infty} \frac{2}{T_k(T_k - 1)} \sum_{i<j} I_\delta(X_i, X_j), \tag{4.40}$$

where $I_\delta(u, v)$ is an indicator variable that equals one if $\|u - v\| < \delta$, and zero otherwise, where $\|\cdot\|$ is the supnorm. The correlation integral measures the fraction of data pairs of $\{X_t\}$ that are within a distance of δ from each other. Consider next a time series x_t. Construct k-dimensional vectors $X_t^k = (x_t, x_{t+1}, \ldots, x_{t+k-1})'$, which are called k *histories*. The idea of the BDS test is as follows. Treat a k history as a point in the k-dimensional space. If $\{x_t\}_{t=1}^{T}$ are indeed iid random variables, then the k-histories $\{X_t\}_{t=1}^{T_k}$ should show no pattern in the k-dimensional space. Consequently, the correlation integrals should satisfy the relation $C_k(\delta) = [C_1(\delta)]^k$. Any departure from the prior relation suggests that x_t are not iid. As a simple, but informative example, consider a sequence of iid random variables from the uniform distribution over [0, 1]. Let $[a, b]$ be a subinterval of [0, 1] and consider the "2-history" (x_t, x_{t+1}), which represents a point in the two-dimensional space. Under the iid assumption, the expected number of 2-histories in the subspace $[a, b] \times [a, b]$ should equal the square of the expected number of x_t in $[a, b]$. This idea can be formally examined by using sample counterparts of correlation integrals. Define

$$C_\ell(\delta, T) = \frac{2}{T_k(T_k - 1)} \sum_{i<j} I_\delta(X_i^*, X_j^*), \qquad \ell = 1, k,$$

where $T_\ell = T - \ell + 1$ and $X_i^* = x_i$ if $\ell = 1$ and $X_i^* = X_i^k$ if $\ell = k$. Under the null hypothesis that $\{x_t\}$ are iid with a nondegenerated distribution function $F(\cdot)$, Brock, Dechert, and Scheinkman (1987) show that

$$C_k(\delta, T) \to [C_1(\delta)]^k \quad \text{with probability 1,} \quad \text{as} \quad T \to \infty$$

for any fixed k and δ. Furthermore, the statistic $\sqrt{T}\{C_k(\delta, T) - [C_1(\delta, T)]^k\}$ is asymptotically distributed as normal with mean zero and variance:

$$\sigma_k^2(\delta) = 4\left(N^k + 2\sum_{j=1}^{k-1} N^{k-j} C^{2j} + (k-1)^2 C^{2k} - k^2 N C^{2k-2}\right),$$

where $C = \int [F(z+\delta) - F(z-\delta)]\, dF(z)$ and $N = \int [F(z+\delta) - F(z-\delta)]^2\, dF(z)$. Note that $C_1(\delta, T)$ is a consistent estimate of C, and N can be consistently estimated by

$$N(\delta, T) = \frac{6}{T_k(T_k-1)(T_k-2)} \sum_{t<s<u} I_\delta(x_t, x_s) I_\delta(x_s, x_u).$$

The BDS test statistic is then defined as

$$D_k(\delta, T) = \frac{\sqrt{T}\{C_k(\delta, T) - [C_1(\delta, T)]^k\}}{\sigma_k(\delta, T)}, \tag{4.41}$$

where $\sigma_k(\delta, T)$ is obtained from $\sigma_k(\delta)$ when C and N are replaced by $C_1(\delta, T)$ and $N(\delta, T)$, respectively. This test statistic has a standard normal limiting distribution. For further discussion and examples of applying the BDS test, see Hsieh (1989) and Brock, Hsieh, and LeBaron (1991). In application, one should remove linear dependence, if any, from the data before applying the BDS test. The test may be sensitive to the choices of δ and k, especially when k is large.

4.2.2 Parametric Tests

Turning to parametric tests, we consider the RESET test of Ramsey (1969) and its generalizations. We also discuss some test statistics for detecting threshold nonlinearity. To simplify the notation, we use vectors and matrices in the discussion. If necessary, readers may consult Appendix A of Chapter 8 for a brief review on vectors and matrices.

The RESET Test

Ramsey (1969) proposes a specification test for linear least-squares regression analysis. The test is referred to as a RESET test and is readily applicable to linear AR models. Consider the linear AR(p) model

$$x_t = \boldsymbol{X}_{t-1}'\boldsymbol{\phi} + a_t, \tag{4.42}$$

where $\boldsymbol{X}_{t-1} = (1, x_{t-1}, \ldots, x_{t-p})'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)'$. The first step of the RESET test is to obtain the least-squares estimate $\widehat{\boldsymbol{\phi}}$ of Eq. (4.42) and compute the fit $\hat{x}_t = \boldsymbol{X}_{t-1}'\widehat{\boldsymbol{\phi}}$, the residual $\hat{a}_t = x_t - \hat{x}_t$, and the sum of squared residuals $\text{SSR}_0 = \sum_{t=p+1}^{T} \hat{a}_t^2$, where T is the sample size. In the second step, consider the linear regression

$$\hat{a}_t = \boldsymbol{X}_{t-1}'\boldsymbol{\alpha}_1 + \boldsymbol{M}_{t-1}'\boldsymbol{\alpha}_2 + v_t, \tag{4.43}$$

where $\boldsymbol{M}_{t-1} = (\hat{x}_t^2, \ldots, \hat{x}_t^{s+1})'$ for some $s \geq 1$, and compute the least-squares residuals

$$\hat{v}_t = \hat{a}_t - \boldsymbol{X}_{t-1}'\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{M}_{t-1}'\widehat{\boldsymbol{\alpha}}_2$$

and the sum of squared residuals $\text{SSR}_1 = \sum_{t=p+1}^{T} \hat{v}_t^2$ of the regression. The basic idea of the RESET test is that if the linear AR(p) model in Eq. (4.42) is adequate, then $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ of Eq. (4.43) should be zero. This can be tested by the usual F statistic of Eq. (4.43) given by

$$F = \frac{(\text{SSR}_0 - \text{SSR}_1)/g}{\text{SSR}_1/(T - p - g)} \quad \text{with} \quad g = s + p + 1, \tag{4.44}$$

which, under the linearity and normality assumption, has an F distribution with degrees of freedom g and $T - p - g$.

Remark. Because $\hat{x}_t^k$ for $k = 2, \ldots, s + 1$ tend to be highly correlated with $\boldsymbol{X}_{t-1}$ and among themselves, principal components of $\boldsymbol{M}_{t-1}$ that are not co-linear with $\boldsymbol{X}_{t-1}$ are often used in fitting Eq. (4.43). Principal component analysis is a statistical tool for dimension reduction; see Chapter 8 for more information. □

Keenan (1985) proposes a nonlinearity test for time series that uses $\hat{x}_t^2$ only and modifies the second step of the RESET test to avoid multicollinearity between $\hat{x}_t^2$ and $\boldsymbol{X}_{t-1}$. Specifically, the linear regression (4.43) is divided into two steps. In step 2(a), one removes linear dependence of $\hat{x}_t^2$ on $\boldsymbol{X}_{t-1}$ by fitting the regression

$$\hat{x}_t^2 = \boldsymbol{X}_{t-1}'\boldsymbol{\beta} + u_t$$

and obtaining the residual $\hat{u}_t = \hat{x}_t^2 - \boldsymbol{X}_{t-1}\widehat{\boldsymbol{\beta}}$. In step 2(b), consider the linear regression

$$\hat{a}_t = \hat{u}_t\alpha + v_t,$$

and obtain the sum of squared residuals $\text{SSR}_1 = \sum_{t=p+1}^{T}(\hat{a}_t - \hat{u}_t\hat{\alpha})^2 = \sum_{t=p+1}^{T} \hat{v}_t^2$ to test the null hypothesis $\alpha = 0$.

The F Test

To improve the power of Keenan's test and the RESET test, Tsay (1986) uses a different choice of the regressor $\boldsymbol{M}_{t-1}$. Specifically, he suggests using $\boldsymbol{M}_{t-1} = \text{vech}(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}')$, where vech($\boldsymbol{A}$) denotes the half-stacking vector of the matrix $\boldsymbol{A}$ using elements on and below the diagonal only; see Appendix B of Chapter 8 for more information about the operator. For example, if $p = 2$, then $\boldsymbol{M}_{t-1} = (x_{t-1}^2, x_{t-1}x_{t-2}, x_{t-2}^2)'$. The dimension of $\boldsymbol{M}_{t-1}$ is $p(p + 1)/2$ for an AR(p) model. In practice, the test is simply the usual partial F statistic for testing $\boldsymbol{\alpha} = 0$ in the linear least-squares regression

$$x_t = \boldsymbol{X}_{t-1}'\boldsymbol{\phi} + \boldsymbol{M}_{t-1}'\boldsymbol{\alpha} + e_t,$$

where e_t denotes the error term. Under the assumption that x_t is a linear AR(p) process, the partial F statistic follows an F distribution with degrees of freedom

g and $T - p - g - 1$, where $g = p(p+1)/2$. We refer to this F test as the *Ori-F test*. Luukkonen, Saikkonen, and Teräsvirta (1988) further extend the test by augmenting $\boldsymbol{M}_{t-1}$ with cubic terms x_{t-i}^3 for $i = 1, \ldots, p$.

Threshold Test

When the alternative model under study is a SETAR model, one can derive specific test statistics to increase the power of the test. One of the specific tests is the likelihood ratio statistic. This test, however, encounters the difficulty of undefined parameters under the null hypothesis of linearity because the threshold is undefined for a linear AR process. Another specific test seeks to transform testing threshold nonlinearity into detecting model changes. It is then interesting to discuss the differences between these two specific tests for threshold nonlinearity.

To simplify the discussion, let us consider the simple case that the alternative model is a 2-regime SETAR model with threshold variable x_{t-d}. The null hypothesis H_0: x_t follows the linear AR(p) model

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + a_t, \tag{4.45}$$

whereas the alternative hypothesis H_a: x_t follows the SETAR model

$$x_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p} \phi_i^{(1)} x_{t-i} + a_{1t} & \text{if } x_{t-d} < r_1, \\ \phi_0^{(2)} + \sum_{i=1}^{p} \phi_i^{(2)} x_{t-i} + a_{2t} & \text{if } x_{t-d} \geq r_1, \end{cases} \tag{4.46}$$

where r_1 is the threshold. For a given realization $\{x_t\}_{t=1}^{T}$ and assuming normality, let $l_0(\hat{\boldsymbol{\phi}}, \hat{\sigma}_a^2)$ be the log-likelihood function evaluated at the maximum-likelihood estimates of $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_p)'$ and σ_a^2. This is easy to compute. The likelihood function under the alternative is also easy to compute if the threshold r_1 is given. Let $l_1(r_1; \hat{\boldsymbol{\phi}}_1, \hat{\sigma}_1^2; \hat{\boldsymbol{\phi}}_2, \hat{\sigma}_2^2)$ be the log-likelihood function evaluated at the maximum-likelihood estimates of $\boldsymbol{\phi}_i = (\phi_0^{(i)}, \ldots, \phi_p^{(i)})'$ and σ_i^2 conditioned on knowing the threshold r_1. The log-likelihood ratio $l(r_1)$ defined as

$$l(r_1) = l_1(r_1; \hat{\boldsymbol{\phi}}_1, \hat{\sigma}_1^2; \hat{\boldsymbol{\phi}}_2, \hat{\sigma}_2^2) - l_0(\hat{\boldsymbol{\phi}}, \hat{\sigma}_a^2)$$

is then a function of the threshold r_1, which is unknown. Yet under the null hypothesis, there is no threshold and r_1 is not defined. The parameter r_1 is referred to as a *nuisance parameter* under the null hypothesis. Consequently, the asymptotic distribution of the likelihood ratio is very different from that of the conventional likelihood ratio statistics. See Chan (1991) for further details and critical values of the test. A common approach is to use $l_{\max} = \sup_{v < r_1 < u} l(r_1)$ as the test statistic, where v and u are prespecified lower and upper bounds of the threshold. Davis (1987) and Andrews and Ploberger (1994) provide further discussion on hypothesis testing involving nuisance parameters under the null hypothesis. Simulation is often used to obtain empirical critical values of the test statistic $l_{\max}$, which depends on

the choices of v and u. The average of $l(r_1)$ over $r_1 \in [v, u]$ is also considered by Andrews and Ploberger as a test statistic.

Tsay (1989) makes use of arranged autoregression and recursive estimation to derive an alternative test for threshold nonlinearity. The arranged autoregression seeks to transfer the SETAR model under the alternative hypothesis H_a into a model change problem with the threshold r_1 serving as the change point. To see this, the SETAR model in Eq. (4.46) says that x_t follows essentially two linear models depending on whether $x_{t-d} < r_1$ or $x_{t-d} \geq r_1$. For a realization $\{x_t\}_{t=1}^{T}$, x_{t-d} can assume values $\{x_1, \ldots, x_{T-d}\}$. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(T-d)}$ be the ordered statistics of $\{x_t\}_{t=1}^{T-d}$ (i.e., arranging the observations in increasing order). The SETAR model can then be written as

$$x_{(j)+d} = \beta_0 + \sum_{i=1}^{p} \beta_i x_{(j)+d-i} + a_{(j)+d}, \quad j = 1, \ldots, T-d, \tag{4.47}$$

where $\beta_i = \phi_i^{(1)}$ if $x_{(j)} < r_1$ and $\beta_i = \phi_i^{(2)}$ if $x_{(j)} \geq r_1$. Consequently, the threshold r_1 is a change point for the linear regression in Eq. (4.47), and we refer to Eq. (4.47) as an arranged autoregression (in increasing order of the threshold x_{t-d}). Note that the arranged autoregression in (4.47) does not alter the dynamic dependence of x_t on x_{t-i} for $i = 1, \ldots, p$ because $x_{(j)+d}$ still depends on $x_{(j)+d-i}$ for $i = 1, \ldots, p$. What is done is simply to present the SETAR model in the threshold space instead of in the time space. That is, the equation with a smaller x_{t-d} appears before that with a larger x_{t-d}. The threshold test of Tsay (1989) is obtained as follows.

- *Step 1*. Fit Eq. (4.47) using $j = 1, \ldots, m$, where m is a prespecified positive integer (e.g., 30). Denote the least-squares estimates of β_i by $\hat{\beta}_{i,m}$, where m denotes the number of data points used in estimation.
- *Step 2*. Compute the predictive residual

$$\hat{a}_{(m+1)+d} = x_{(m+1)+d} - \hat{\beta}_{0,m} - \sum_{i=1}^{p} \hat{\beta}_{i,m} x_{(m+1)+d-i}$$

  and its standard error. Let $\hat{e}_{(m+1)+d}$ be the standardized predictive residual.
- *Step 3*. Use the recursive least-squares method to update the least-squares estimates to $\hat{\beta}_{i,m+1}$ by incorporating the new data point $x_{(m+1)+d}$.
- *Step 4*. Repeat steps 2 and 3 until all data points are processed.
- *Step 5*. Consider the linear regression of the standardized predictive residual

$$\hat{e}_{(m+j)+d} = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{(m+j)+d-i} + v_t, \qquad j = 1, \ldots, T-d-m \tag{4.48}$$

  and compute the usual F statistic for testing $\alpha_i = 0$ in Eq. (4.48) for $i = 0, \ldots, p$. Under the null hypothesis that x_t follows a linear AR(p) model,

the F ratio has a limiting F distribution with degrees of freedom $p + 1$ and $T - d - m - p$.

We refer to the earlier F test as a *TAR-F test*. The idea behind the test is that under the null hypothesis there is no model change in the arranged autoregression in Eq. (4.47) so that the standardized predictive residuals should be close to iid with mean zero and variance 1. In this case, they should have no correlations with the regressors $x_{(m+j)+d-i}$. For further details including formulas for a recursive least-squares method and some simulation study on performance of the TAR-F test, see Tsay (1989). The TAR-F test avoids the problem of nuisance parameters encountered by the likelihood ratio test. It does not require knowing the threshold r_1. It simply tests that the predictive residuals have no correlations with regressors if the null hypothesis holds. Therefore, the test does not depend on knowing the number of regimes in the alternative model. Yet the TAR-F test is not as powerful as the likelihood ratio test if the true model is indeed a 2-regime SETAR model with a known innovational distribution.

4.2.3 Applications

In this subsection, we apply some of the nonlinearity tests discussed previously to five time series. For a real financial time series, an AR model is used to remove any serial correlation in the data, and the tests apply to the residual series of the model. The five series employed are as follows:

1. r_{1t}: A simulated series of iid $N(0, 1)$ with 500 observations.
2. r_{2t}: A simulated series of iid Student-t distribution with 6 degrees of freedom. The sample size is 500.
3. a_{3t}: The residual series of monthly log returns of CRSP equal-weighted index from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.180B + 0.099B^3 - 0.105B^9)r_{3t} = 0.0086 + a_{3t}.$$

4. a_{4t}: The residual series of monthly log returns of CRSP value-weighted index from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.098B + 0.111B^3 - 0.088B^5)r_{4t} = 0.0078 + a_{4t}.$$

5. a_{5t}: The residual series of monthly log returns of IBM stock from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.077B)r_{5t} = 0.011 + a_{5t}.$$

Table 4.2 shows the results of the nonlinearity test. For the simulated series and IBM returns, the F tests are based on an AR(6) model. For the index returns, the

TABLE 4.2 Nonlinearity Tests for Simulated Series and Some Log Stock Returns[a]

	Q	Q	BDS($\delta = 1.5\hat{\sigma}_a$)			
Data	(5)	(10)	2	3	4	5
$N(0,1)$	3.2	6.5	−0.32	−0.14	−0.15	−0.33
t_6	0.9	1.7	−0.87	−1.18	−1.56	−1.71
ln(ew)	2.9	4.9	9.94	11.72	12.83	13.65
ln(vw)	1.0	9.8	8.61	9.88	10.70	11.29
ln(ibm)	0.6	7.1	4.96	6.09	6.68	6.82
		$d = 1$	BDS($\delta = \hat{\sigma}_a$)			
Data	Ori-F	TAR-F	2	3	4	5
$N(0,1)$	1.13	0.87	−0.77	−0.71	−1.04	−1.27
t_6	0.69	0.81	−0.35	−0.76	−1.25	−1.49
ln(ew)	5.05	6.77	10.01	11.85	13.14	14.45
ln(vw)	4.95	6.85	7.01	7.83	8.64	9.53
ln(ibm)	1.32	1.51	3.82	4.70	5.45	5.72

[a]The sample size of simulated series is 500 and that of stock returns is 864. The BDS test uses $k = 2, \ldots, 5$.

AR order is the same as the model given earlier. For the BDS test, we chose $\delta = \hat{\sigma}_a$ and $\delta = 1.5\hat{\sigma}_a$ with $k = 2, \ldots, 5$. Also given in the table are the Ljung–Box statistics that confirm no serial correlation in the residual series before applying nonlinearity tests. Compared with their asymptotic critical values, the BDS test and F tests are insignificant at the 5% level for the simulated series. However, the BDS tests are highly significant for the real financial time series. The F tests also show significant results for the index returns, but they fail to suggest nonlinearity in the IBM log returns. In summary, the tests confirm that the simulated series are linear and suggest that the stock returns are nonlinear.

4.3 MODELING

Nonlinear time series modeling necessarily involves subjective judgment. However, there are some general guidelines to follow. It starts with building an adequate linear model on which nonlinearity tests are based. For financial time series, the Ljung–Box statistics and Engle's test are commonly used to detect conditional heteroscedasticity. For general series, other tests of Section 4.2 apply. If nonlinearity is statistically significant, then one chooses a class of nonlinear models to entertain. The selection here may depend on the experience of the analyst and the substantive matter of the problem under study. For volatility models, the order of an ARCH process can often be determined by checking the partial autocorrelation function of the squared series. For GARCH and EGARCH models, only lower orders such as (1,1), (1,2), and (2,1) are considered in most applications.

Higher order models are hard to estimate and understand. For TAR models, one may use the procedures given in Tong (1990) and Tsay (1989, 1998) to build an adequate model. When the sample size is sufficiently large, one may apply nonparametric techniques to explore the nonlinear feature of the data and choose a proper nonlinear model accordingly; see Chen and Tsay (1993a) and Cai, Fan, and Yao (2000). The MARS procedure of Lewis and Stevens (1991) can also be used to explore the dynamic structure of the data. Finally, information criteria such as the Akaike information criterion (Akaike, 1974) and the generalized odd ratios in Chen, McCulloch, and Tsay (1997) can be used to discriminate between competing nonlinear models. The chosen model should be carefully checked before it is used for prediction.

4.4 FORECASTING

Unlike the linear model, there exist no closed-form formulas to compute forecasts of most nonlinear models when the forecast horizon is greater than 1. We use parametric bootstraps to compute nonlinear forecasts. It is understood that the model used in forecasting has been rigorously checked and is judged to be adequate for the series under study. By a model, we mean the dynamic structure and innovational distributions. In some cases, we may treat the estimated parameters as given.

4.4.1 Parametric Bootstrap

Let T be the forecast origin and ℓ be the forecast horizon ($\ell > 0$). That is, we are at time index T and interested in forecasting $x_{T+\ell}$. The parametric bootstrap considered computes realizations $x_{T+1}, \ldots, X_{T+\ell}$ sequentially by (a) drawing a new innovation from the specified innovational distribution of the model, and (b) computing x_{T+i} using the model, data, and previous forecasts $x_{T+1}, \ldots, x_{T+i-1}$. This results in a realization for $x_{T+\ell}$. The procedure is repeated M times to obtain M realizations of $x_{T+\ell}$ denoted by $\{x_{T+\ell}^{(j)}\}_{j=1}^{M}$. The point forecast of $x_{T+\ell}$ is then the sample average of $x_{T+\ell}^{(j)}$. Let the forecast be $x_T(\ell)$. We used $M = 3000$ in some applications and the results seem fine. The realizations $\{x_{T+\ell}^{(j)}\}_{j=1}^{M}$ can also be used to obtain an empirical distribution of $x_{T+\ell}$. We make use of this empirical distribution later to evaluate forecasting performance.

4.4.2 Forecasting Evaluation

There are many ways to evaluate the forecasting performance of a model, ranging from directional measures to magnitude measures to distributional measures. A directional measure considers the future direction (up or down) implied by the model. Predicting that tomorrow's S&P 500 index will go up or down is an example of directional forecasts that are of practical interest. Predicting the year-end value of the daily S&P 500 index belongs to the case of magnitude measure. Finally,

assessing the likelihood that the daily S&P 500 index will go up 10% or more between now and the year end requires knowing the future conditional probability distribution of the index. Evaluating the accuracy of such an assessment needs a distributional measure.

In practice, the available data set is divided into two subsamples. The first subsample of the data is used to build a nonlinear model, and the second subsample is used to evaluate the forecasting performance of the model. We refer to the two subsamples of data as *estimation* and *forecasting subsamples*. In some studies, a rolling forecasting procedure is used in which a new data point is moved from the forecasting subsample into the estimation subsample as the forecast origin advances. In what follows, we briefly discuss some measures of forecasting performance that are commonly used in the literature. Keep in mind, however, that there exists no widely accepted single measure to compare models. A utility function based on the objective of the forecast might be needed to better understand the comparison.

Directional Measure

A typical measure here is to use a 2×2 contingency table that summarizes the number of "hits" and "misses" of the model in predicting ups and downs of $x_{T+\ell}$ in the forecasting subsample. Specifically, the contingency table is given as

Actual	Predicted		
	Up	Down	
Up	m_{11}	m_{12}	m_{10}
Down	m_{21}	m_{22}	m_{20}
	m_{01}	m_{02}	m

where m is the total number of ℓ-step-ahead forecasts in the forecasting subsample, m_{11} is the number of "hits" in predicting upward movements, m_{21} is the number of "misses" in predicting downward movements of the market, and so on. Larger values in m_{11} and m_{22} indicate better forecasts. The test statistic

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(m_{ij} - m_{i0}m_{0j}/m)^2}{m_{i0}m_{0j}/m}$$

can then be used to evaluate the performance of the model. A large χ^2 signifies that the model outperforms the chance of random choice. Under some mild conditions, χ^2 has an asymptotic chi-squared distribution with 1 degree of freedom. For further discussion of this measure, see Dahl and Hylleberg (1999).

For illustration of the directional measure, consider the 1-step-ahead probability forecasts of the 8–4–1 feed-forward neural network shown in Figure 4.9. The 2×2 table of "hits" and "misses" of the network is

Actual	Predicted		
	Up	Down	
Up	12	2	14
Down	8	2	10
	20	4	24

The table shows that the network predicts the upward movement well, but fares poorly in forecasting the downward movement of the stock. The chi-squared statistic of the table is 0.137 with a p value of 0.71. Consequently, the network does not significantly outperform a random-walk model with equal probabilities for "upward" and "downward" movements.

Magnitude Measure

Three statistics are commonly used to measure performance of point forecasts. They are the mean squared error (MSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE). For ℓ-step-ahead forecasts, these measures are defined as

$$\text{MSE}(\ell) = \frac{1}{m}\sum_{j=0}^{m-1}[x_{T+\ell+j} - x_{T+j}(\ell)]^2, \tag{4.49}$$

$$\text{MAD}(\ell) = \frac{1}{m}\sum_{j=0}^{m-1}|x_{T+\ell+j} - x_{T+j}(\ell)|, \tag{4.50}$$

$$\text{MAPE}(\ell) = \frac{1}{m}\sum_{j=0}^{m-1}\left|\frac{x_{T+j}(\ell)}{x_{T+j+\ell}} - 1\right|, \tag{4.51}$$

where m is the number of ℓ-step-ahead forecasts available in the forecasting subsample. In application, one often chooses one of the above three measures, and the model with the smallest magnitude on that measure is regarded as the best ℓ-step-ahead forecasting model. It is possible that different ℓ may result in selecting different models. The measures also have other limitations in model comparison; see, for instance, Clements and Hendry (1993).

Distributional Measure

Practitioners recently began to assess forecasting performance of a model using its predictive distributions. Strictly speaking, a predictive distribution incorporates parameter uncertainty in forecasts. We call it *conditional predictive distribution* if the parameters are treated as fixed. The empirical distribution of $x_{T+\ell}$ obtained by the parametric bootstrap is a conditional predictive distribution. This empirical distribution is often used to compute a distributional measure. Let $u_T(\ell)$ be the percentile of the observed $x_{T+\ell}$ in the prior empirical distribution. We then have

a set of m percentiles $\{u_{T+j}(\ell)\}_{j=0}^{m-1}$, where again m is the number of ℓ-step-ahead forecasts in the forecasting subsample. If the model entertained is adequate, $\{u_{T+j}(\ell)\}$ should be a random sample from the uniform distribution on [0, 1]. For a sufficiently large m, one can compute the Kolmogorov–Smirnov statistic of $\{u_{T+j}(\ell)\}$ with respect to uniform [0, 1]. The statistic can be used for both model checking and forecasting comparison.

4.5 APPLICATION

In this section, we illustrate nonlinear time series models by analyzing the quarterly U.S. civilian unemployment rate, seasonally adjusted, from 1948 to 1993. This series was analyzed in detail by Montgomery et al. (1998). We repeat some of the analyses here using nonlinear models. Figure 4.11 shows the time plot of the data. Well-known characteristics of the series include that (a) it tends to move countercyclically with U.S. business cycles, and (b) the rate rises quickly but decays slowly. The latter characteristic suggests that the dynamic structure of the series is nonlinear.

Denote the series by x_t and let $\Delta x_t = x_t - x_{t-1}$ be the change in unemployment rate. The linear model

$$(1 - 0.31B^4)(1 - 0.65B)\Delta x_t = (1 - 0.78B^4)a_t, \qquad \hat{\sigma}_a^2 = 0.090 \tag{4.52}$$

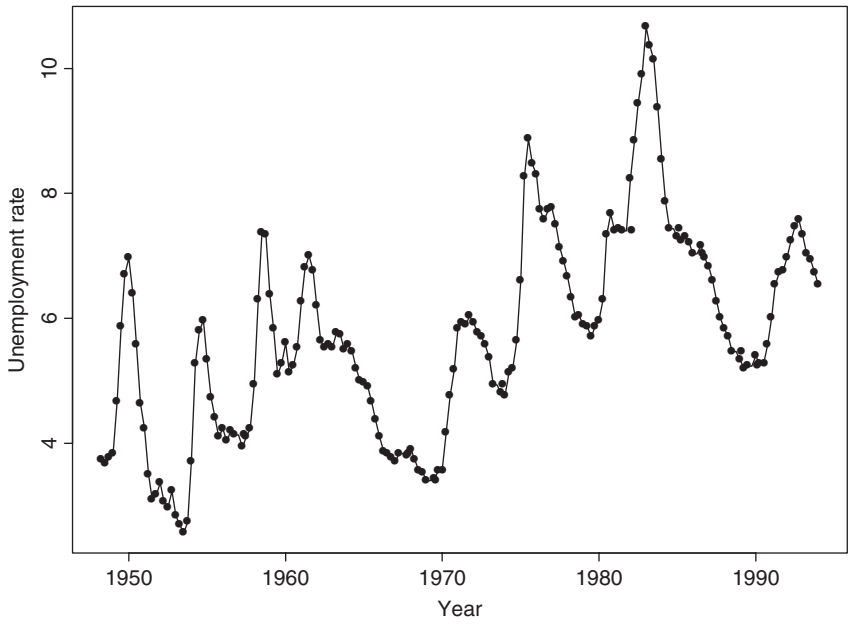

Figure 4.11 Time plot of U.S. quarterly unemployment rate, seasonally adjusted, from 1948 to 1993.

was built by Montgomery et al. (1998), where the standard errors of the three coefficients are 0.11, 0.06, and 0.07, respectively. This is a seasonal model even though the data were seasonally adjusted. It indicates that the seasonal adjustment procedure used did not successfully remove the seasonality. This model is used as a benchmark model for forecasting comparison.

To test for nonlinearity, we apply some of the nonlinearity tests of Section 4.2 with an AR(5) model for the differenced series Δx_t. The results are given in Table 4.3. All of the tests reject the linearity assumption. In fact, the linearity assumption is rejected for all AR(p) models we applied, where $p = 2, \ldots, 10$.

Using a modeling procedure similar to that of Tsay (1989), Montgomery et al. (1998) build the following TAR model for the Δx_t series:

$$\Delta x_t = \begin{cases} 0.01 + 0.73\Delta x_{t-1} + 0.10\Delta x_{t-2} + a_{1t} & \text{if } \Delta x_{t-2} \le 0.1, \\ 0.18 + 0.80\Delta x_{t-1} - 0.56\Delta x_{t-2} + a_{2t} & \text{otherwise.} \end{cases} \tag{4.53}$$

The sample variances of a_{1t} and a_{2t} are 0.76 and 0.165, respectively, the standard errors of the three coefficients of regime 1 are 0.03, 0.10, and 0.12, respectively, and those of regime 2 are 0.09, 0.1, and 0.16. This model says that the change in the U.S. quarterly unemployment rate, Δx_t, behaves like a piecewise linear model in the reference space of $x_{t-2} - x_{t-3}$ with threshold 0.1. Intuitively, the model implies that the dynamics of unemployment act differently depending on the recent change in the unemployment rate. In the first regime, the unemployment rate has had either a decrease or a minor increase. Here the economy should be stable, and essentially the change in the rate follows a simple AR(1) model because the lag-2 coefficient is insignificant. In the second regime, there is a substantial jump in the unemployment rate (0.1 or larger). This typically corresponds to the contraction phase in the business cycle. It is also the period during which government interventions and industrial restructuring are likely to occur. Here Δx_t follows an AR(2) model with a positive constant, indicating an upward trend in x_t. The AR(2) polynomial contains two complex characteristic roots, which indicate possible cyclical behavior in Δx_t. Consequently, the chance of having a turning point in x_t increases, suggesting that the period of large increases in x_t should be short. This implies that the contraction phases in the U.S. economy tend to be shorter than the expansion phases.

TABLE 4.3 Nonlinearity Test for Changes in the U.S. Quarterly Unemployment Rate: 1948.II–1993.IV[a]

Type	Ori-F	LST	TAR(1)	TAR(2)	TAR(3)	TAR(4)
Test	2.80	2.83	2.41	2.16	2.84	2.98
p Value	.0007	.0002	.0298	.0500	.0121	.0088

[a] An AR(5) model was used in the tests, where LST denotes the test of Luukkonen et al. (1988) and TAR(d) means threshold test with delay d.

Applying a Markov chain Monte Carlo method, Montgomery et al. (1998) obtain the following Markov switching model for Δx_t:

$$\Delta x_t = \begin{cases} -0.07 + 0.38\Delta x_{t-1} - 0.05\Delta x_{t-2} + \epsilon_{1t} & \text{if } s_t = 1, \\ 0.16 + 0.86\Delta x_{t-1} - 0.38\Delta x_{t-2} + \epsilon_{2t} & \text{if } s_t = 2. \end{cases} \tag{4.54}$$

The conditional means of Δx_t are -0.10 for $s_t = 1$ and 0.31 for $s_t = 2$. Thus, the first state represents the expansionary periods in the economy, and the second state represents the contractions. The sample variances of ϵ_{1t} and ϵ_{2t} are 0.031 and 0.192, respectively. The standard errors of the three parameters in state $s_t = 1$ are 0.03, 0.14, and 0.11, and those of state $s_t = 2$ are 0.04, 0.13, and 0.14, respectively. The state transition probabilities are $P(s_t = 2|s_{t-1} = 1) = 0.084(0.060)$ and $P(s_t = 1|s_{t-1} = 2) = 0.126(0.053)$, where the number in parentheses is the corresponding standard error. This model implies that in the second state the unemployment rate x_t has an upward trend with an AR(2) polynomial possessing complex characteristic roots. This feature of the model is similar to the second regime of the TAR model in Eq. (4.53). In the first state, the unemployment rate x_t has a slightly decreasing trend with a much weaker autoregressive structure.

Forecasting Performance

A rolling procedure was used by Montgomery et al. (1998) to forecast the unemployment rate x_t. The procedure works as follows:

1. Begin with forecast origin $T = 83$, corresponding to 1968.II, which was used in the literature to monitor the performance of various econometric models in forecasting unemployment rate. Estimate the linear, TAR, and MSA models using the data from 1948.I to the forecast origin (inclusive).
2. Perform 1-quarter to 5-quarter ahead forecasts and compute the forecast errors of each model. Forecasts of nonlinear models used are computed by using the parametric bootstrap method of Section 4.4.
3. Advance the forecast origin by 1 and repeat the estimation and forecasting processes until all data are employed.
4. Use MSE and mean forecast error to compare performance of the models.

Table 4.4 shows the relative MSE of forecasts and mean forecast errors for the linear model in Eq. (4.52), the TAR model in Eq. (4.53), and the MSA model in Eq. (4.54), using the linear model as a benchmark. The comparisons are based on overall performance as well as the status of the U.S. economy at the forecast origin. From the table, we make the following observations:

1. For the overall comparison, the TAR model and the linear model are very close in MSE, but the TAR model has smaller biases. Yet the MSA model has the highest MSE and smallest biases.

TABLE 4.4 Out-of-Sample Forecast Comparison among Linear, TAR, and MSA Models for the U.S. Quarterly Unemployment Rate[a]

	Relative MSE of Forecast				
Model	1-step	2-step	3-step	4-step	5-step
			Overall Comparison		
Linear	1.00	1.00	1.00	1.00	1.00
TAR	1.00	1.04	0.99	0.98	1.03
MSA	1.19	1.39	1.40	1.45	1.61
MSE	0.08	0.31	0.67	1.13	1.54
			Forecast Origins in Economic Contractions		
Linear	1.00	1.00	1.00	1.00	1.00
TAR	0.85	0.91	0.83	0.72	0.72
MSA	0.97	1.03	0.96	0.86	1.02
MSE	0.22	0.97	2.14	3.38	3.46
			Forecast Origins in Economic Expansions		
Linear	1.00	1.00	1.00	1.00	1.00
TAR	1.06	1.13	1.10	1.15	1.17
MSA	1.31	1.64	1.73	1.84	1.87
MSE	0.06	0.21	0.45	0.78	1.24
	Mean of Forecast Errors				
Model	1-step	2-step	3-step	4-step	5-step
			Overall Comparison		
Linear	0.03	0.09	0.17	0.25	0.33
TAR	−0.10	−0.02	−0.03	−0.03	−0.01
MSA	0.00	−0.02	−0.04	−0.07	−0.12
			Forecast Origins in Economic Contractions		
Linear	0.31	0.68	1.08	1.41	1.38
TAR	0.24	0.56	0.87	1.01	0.86
MSA	0.20	0.41	0.57	0.52	0.14
			Forecast Origins in Economic Expansions		
Linear	−0.01	0.00	0.03	0.08	0.17
TAR	−0.05	−0.11	−0.17	−0.19	−0.14
MSA	−0.03	−0.08	−0.13	−0.17	−0.16

[a]The starting forecast origin is 1968.II, where the row marked by MSE shows the MSE of the benchmark linear model.

2. For forecast origins in economic contractions, the TAR model shows improvements over the linear model both in MSE and bias. The MSA model also shows some improvement over the linear model, but the improvement is not as large as that of the TAR model.
3. For forecast origins in economic expansions, the linear model outperforms both nonlinear models.

The results suggest that the contributions of nonlinear models over linear ones in forecasting the U.S. quarterly unemployment rate are mainly in the periods when the U.S. economy is in contraction. This is not surprising because, as mentioned before, it is during the economic contractions that government interventions and industrial restructuring are most likely to occur. These external events could introduce nonlinearity in the U.S. unemployment rate. Intuitively, such improvements are important because it is during the contractions that people pay more attention to economic forecasts.

APPENDIX A: SOME RATS PROGRAMS FOR NONLINEAR VOLATILITY MODELS

Program Used to Estimate an AR(2)–TAR–GARCH(1,1) Model for Daily Log Returns of IBM Stock

Assume that the data file is `d-ibmln03.txt`.

```
all 0 10446:1
open data d-ibmln03.txt
data(org=obs) / rt
set h = 0.0
nonlin mu p2 a0 a1 b1 a2 b2
frml at = rt(t)-mu-p2*rt(t-2)
frml gvar = a0 + a1*at(t-1)**2+b1*h(t-1) $
            + % if(at(t-1) < 0,a2*at(t-1)**2+b2*h(t-1),0)
frml garchln = -0.5*log(h(t)=gvar(t))-0.5*at(t)**2/h(t)
smpl 4 10446
compute mu = 0.03, p2 = -0.03
compute a0 = 0.07, a1 = 0.05, a2 = 0.05, b1 = 0.85, b2 = 0.05
maximize(method=simplex,iterations=10) garchln
smpl 4  10446
maximize(method=bhhh,recursive,iterations=150) garchln
set fv = gvar(t)
set resid = at(t)/sqrt(fv(t))
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

Program Used to Estimate a Smooth TAR Model for the Monthly Simple Returns of 3M Stock

The data file is `m-3m4608.txt`.

```
all 0 755:1
open data m-3m4608.txt
data(org=obs) / date mmm
set h = 0.0
nonlin a0 a1 a2  a00 a11 mu
frml at = mmm(t) - mu
frml var1 =  a0+a1*at(t-1)**2+a2*at(t-2)**2
frml var2 = a00+a11*at(t-1)**2
frml gvar = var1(t)+var2(t)/(1.0+exp(-at(t-1)*1000.0))
frml garchlog = -0.5*log(h(t)=gvar(t))-0.5*at(t)**2/h(t)
smpl 3 623
compute a0 = .01, a1 = 0.2, a2 = 0.1
compute a00 = .01, a11 = -.2, mu = 0.02
maximize(method=bhhh,recursive,iterations=150) garchlog
set fv = gvar(t)
set resid = at(t)/sqrt(fv(t))
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

APPENDIX B: R AND S-PLUS COMMANDS FOR NEURAL NETWORK

The following commands are used in R or S-Plus to build the 3−2−1 skip-layer feed-forward network of Example 4.7. A line starting with # denotes a comment. The data file is `m-ibmln.txt`. The library used is `nnet`.

```
# load the data into R or S-Plus workspace.
x_scan(file='m-ibmln.txt')
# select the output: r(t)
y_x[4:864]
# obtain the input variables: r(t-1), r(t-2), and r(t-3)
ibm.x_cbind(x[3:863]_,x[2:862],x[1:861])
# build a 3-2-1 network with skip layer connections
# and linear output.
ibm.nn_nnet(ibm.x,y,size=2,linout=T,skip=T,maxit=10000,
decay=1e-2,reltol=1e-7,abstol=1e-7,range=1.0)
# print the summary results of the network
summary(ibm.nn)
# compute \& print the residual sum of squares.
sse_sum((y-predict(ibm.nn,ibm.x))^2)
print(sse)
```

```
#eigen(nnet.Hess(ibm.nn,ibm.x,y),T)$values
# setup the input variables in the forecasting subsample
ibm.p_cbind(x[864:887],x[863:886],x[862:885])
# compute the forecasts
yh_predict(ibm.nn,ibm.p)
# The observed returns in the forecasting subsample
yo_x[865:888]
# compute \& print the sum of squares of forecast errors
ssfe_sum((yo-yh)^2)
print(ssfe)
# quit S-Plus or R
q()
```

EXERCISES

4.1. Consider the daily simple returns of Johnson & Johnson stock from January 1998 to December 2008. The data are in the file `d-jnj9808.txt` or can be obtained from CRSP. Convert the returns into log returns in percentage. (a) Build a GJR model for the log return series. Write down the fitted model. Is the leverage effect significant at the 1% level? (b) Build a general threshold volatility model for the log return series. (c) Compare the two TGARCH models.

4.2. Consider the monthly simple returns of General Electric (GE) stock from January 1926 to December 2008 with 996 observations. You may download the data from CRSP or use the file `m-ge2608.txt` on the Web. Convert the returns into log returns in percentages. Build a TGARCH model with GED innovations for the series using a_{t-1} as the threshold variable with zero threshold, where a_{t-1} is the shock at time $t-1$. Write down the fitted model. Is the leverage effect significant at the 5% level?

4.3. Suppose that the monthly log returns of GE stock, measured in percentages, follow a smooth threshold IGARCH(1,1) model. For the sampling period from January 1926 to December 2008, the fitted model is

$$r_t = 1.14 + a_t, \qquad a_t = \sigma_t \epsilon_t$$

$$\sigma_t^2 = 0.119 a_{t-1}^2 + 0.881\sigma_{t-1}^2 + \frac{1}{1+\exp(-10a_{t-1})}(4.276 - 0.084\sigma_{t-1}^2),$$

where all of the estimates are highly significant, the coefficient 10 in the exponent is fixed a priori to simplify the estimation, and $\{\epsilon_t\}$ are iid $N(0, 1)$. Assume that $a_{996} = -5.06$ and $\sigma_{996}^2 = 50.5$. What is the 1-step-ahead volatility forecast $\widehat{\sigma}_{996}(1)$? Suppose instead that $a_{996} = 5.06$. What is the 1-step-ahead volatility forecast $\widehat{\sigma}_{996}(1)$?

4.4. Suppose that the monthly log returns, in percentages, of a stock follow the following Markov switching model:

$$r_t = 1.25 + a_t, \qquad a_t = \sigma_t \epsilon_t,$$
$$\sigma_t^2 = \begin{cases} 0.10a_{t-1}^2 + 0.93\sigma_{t-1}^2 & \text{if } s_t = 1, \\ 4.24 + 0.10a_{t-1}^2 + 0.78\sigma_{t-1}^2 & \text{if } s_t = 2, \end{cases}$$

where the transition probabilities are

$$P(s_t = 2|s_{t-1} = 1) = 0.15, \qquad P(s_t = 1|s_{t-1} = 2) = 0.05.$$

Suppose that $a_{100} = 6.0$, $\sigma_{100}^2 = 50.0$, and $s_{100} = 2$ with probability 1.0. What is the 1-step-ahead volatility forecast at the forecast origin $t = 100$? Also, if the probability of $s_{100} = 2$ is reduced to 0.8, what is the 1-step-ahead volatility forecast at the forecast origin $t = 100$?

4.5. Consider the monthly simple returns of GE stock from January 1926 to December 2008. Use the last three years of data for forecasting evaluation.

(a) Using lagged returns $r_{t-1}, r_{t-2}, r_{t-3}$ as input, build a 3–2–1 feed-forward network to forecast 1-step-ahead returns. Calculate the mean squared error of forecasts.

(b) Again, use lagged returns $r_{t-1}, r_{t-2}, r_{t-3}$ and their signs (directions) to build a 6–5–1 feed-forward network to forecast the 1-step ahead direction of GE stock price movement with 1 denoting upward movement. Calculate the mean squared error of forecasts.
Note: Let `rtn` denote a time series in R or S-Plus. To create a direction variable for `rtn`, use the command

```
drtn = ifelse(rtn > 0, 1, 0)
```

4.6. Because of the existence of inverted yield curves in the term structure of interest rates, the spread of interest rates should be nonlinear. To verify this, consider the weekly U.S. interest rates of (a) Treasury 1-year constant maturity rate and (b) Treasury 3-year constant maturity rate. As in Chapter 2, denote the two interest rates by r_{1t} and r_{3t}, respectively, and the data span is from January 5, 1962, to April 10, 2009. The data are in files `w-gs3yr.txt` and `w-gs1yr.txt` on the Web and can be obtained from the Federal Reserve Bank of St. Louis.

(a) Let $s_t = r_{3t} - r_{1t}$ be the spread in log interest rates. Is $\{s_t\}$ linear? Perform some nonlinearity tests and draw the conclusion using the 5% significance level.

(b) Let $s_t^* = (r_{3t} - r_{3,t-1}) - (r_{1t} - r_{1,t-1}) = s_t - s_{t-1}$ be the change in interest rate spread. Is $\{s_t^*\}$ linear? Perform some nonlinearity tests and draw the conclusion using the 5% significance level.

(c) Build a threshold model for the s_t series and check the fitted model.

(d) Build a threshold model for the s_t^* series and check the fitted model.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**: 716–723.

Andrews, D. W. K. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* **62**: 1383–1414.

Brock, W., Dechert, W. D., and Scheinkman, J. (1987). A test for independence based on the correlation dimension. Working paper, Department of Economics, University of Wisconsin, Madison.

Brock, W., Hsieh, D. A., and LeBaron, B. (1991). *Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence*. MIT Press, Cambridge, MA.

Bryson, A. E. and Ho, Y. C. (1969). *Applied Optimal Control*. Blaisdell, New York

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**: 941–956.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association* **87**: 493–500.

Chan, K. S. (1991). Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society Series B* **53**: 691–696.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of Statistics* **21**: 520–533.

Chan, K. S. and Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis* **7**: 179–190.

Chan, K. S. and Tsay, R. S. (1998). Limiting properties of the conditional least squares estimator of a continuous TAR model. *Biometrika* **85**: 413–426.

Chen, C., McCulloch, R. E., and Tsay, R. S. (1997). A unified approach to estimating and modeling univariate linear and nonlinear time series. *Statistica Sinica* **7**: 451–472.

Chen, R. and Tsay, R. S. (1991). On the ergodicity of TAR(1) processes. *Annals of Applied Probability* **1**: 613–634.

Chen, R. and Tsay, R. S. (1993a). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**: 298–308.

Chen, R. and Tsay, R. S. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association* **88**: 955–967.

Chen, R., Liu, J., and Tsay, R. S. (1995). Additivity tests for nonlinear autoregressive models. *Biometrika* (1995) **82**: 369–383.

Chen, T. and Chen, H. (1995). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks* **6**: 911–917.

Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science* **9**: 2–54.

Clements, M. P. and Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting* **12**: 617–637.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**: 829–836.

Dahl, C. M. and Hylleberg, S. (1999). *Specifying nonlinear econometric models by flexible regression models and relative forecast performance*. Working paper, Department of Economics, University of Aarhus, Denmark.

Davis, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**: 33–43.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations. *Econometrica* **50**: 987–1007.

Epanechnikov, V. (1969). Nonparametric estimates of a multivariate probability density. *Theory of Probability and Its Applications* **14**: 153–158.

Fan, J. (1993). Local linear regression smoother and their minimax efficiencies. *Annals of Statistics* **21**: 196–216.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409.

Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoek and Ruprecht, Gottingen.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**: 357–384.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics* **45**: 39–70.

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

Hansen, B. E. (1997). Inference in TAR models. *Studies in Nonlinear Dynamics and Econometrics* **1**: 119–131.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, New York.

Hinich, M. (1982). Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis* **3**: 169–176.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks* **6**: 1069–1072.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2**: 359–366.

Hsieh, D. A. (1989). Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business* **62**: 339–368.

Keenan, D. M. (1985). A Tukey non-additivity-type test for time series nonlinearity. *Biometrika* **72**: 39–44.

Kitagawa, G. (1998). A self-organizing state space model. *Journal of the American Statistical Association* **93**: 1203–1215.

Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression spline (MARS). *Journal of the American Statistical Association* **86**: 864–877.

Liu, J. and Brockwell, P. J. (1988). On the general bilinear time-series model. *Journal of Applied Probability* **25**: 553–564.

Luukkonen, R., Saikkonen, P., and Teräsvirta (1988). Testing linearity against smooth transition autoregressive models. *Biometrika* **75**: 491–499.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* **88**: 968–978.

McCulloch, R. E. and Tsay, R. S. (1994). Statistical inference of macroeconomic time series via Markov switching models. *Journal of Time Series Analysis* **15**: 523–539.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**: 269–273.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998). Forecasting the U.S. unemployment rate, *Journal of the American Statistical Association* **93**: 478–493.

Nadaraya, E. A. (1964). On estimating regression. *Theory and Probability Application* **10**: 186–190.

Petruccelli, J. and Woolford, S. W. (1984). A threshold AR(1) model. *Journal of Applied Probability* **21**: 270–286.

Potter, S. M. (1995). A nonlinear approach to U.S. GNP. *Journal of Applied Econometrics* **10**: 109–125.

Priestley, M. B. (1980). State-dependent models: A general approach to nonlinear time series analysis. *Journal of Time Series Analysis* **1**: 47–71.

Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*, Academic Press, London, UK.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society Series B* **31**: 350–371.

Ripley, B. D. (1993). Statistical aspects of neural networks. In O. E. Barndorff-Nielsen, J. L. Jensen, and W. S. Kendall (eds.). *Networks and Chaos—Statistical and Probabilistic Aspects*, pp. 40–123. Chapman and Hall, London, UK.

Subba Rao, T. and Gabr, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Lecture Notes in Statistics, vol. **24**. Springer, New York.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**: 208–218.

Tiao, G. C. and Tsay, R. S. (1994). Some advances in nonlinear and adaptive modeling in time series. *Journal of Forecasting* **13**: 109–131.

Tong, H. (1978). On a threshold model. In C. H. Chen (ed.). *Pattern Recognition and Signal Processing*. Sijhoff & Noordhoff, Amsterdam.

Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, Springer, New York.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press, Oxford, UK.

Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika* **73**: 461–466.

Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association* **84**: 231–240.

Tsay, R. S. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* **93**: 1188–1202.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, 3rd edn. Springer, New York.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A* **26**: 359–372.