Chapter 2
Nonlinear Time Series Models

2.1 Some Probabilistic Aspects of Nonlinear Processes

2.1.1 Linear Representations and Linear Models

Assume that for $t \in \mathbb{Z}$, (Z_t) and (Z_t^*) are respectively uncorrelated and independent sequences of r.v's having identical marginal distribution $F(\cdot)$, with zero mean and variance $\sigma_Z^2 < \infty$. For any t, define the time series

$$Y_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i} \tag{2.1}$$

and

$$X_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}^*, \tag{2.2}$$

such that $\sum_{i=0}^{\infty} \psi_i^2 < \infty$, so that both Y_t and X_t are mean-square convergent, having finite variances. In the representation (2.2), specification of the marginal distribution for the independent r.v's (Z_t^*) is enough to specify the finite dimensional distributions of the output series X_t. Therefore (2.2) is a fully specified model for X_t. However, the specification of the marginal distribution for uncorrelated r.v's (Z_t) is not enough to specify fully the finite dimensional distributions of the process Y_t given in (2.1), unless Z_t is a Gaussian sequence. In this case, we can merely calculate uniquely the first two moments, namely the mean, the variance and the autocovariance function of the series Y_t. Therefore (2.1) is not a probabilistic model for Y_t, but can be called a representation. Since this representation uniquely specifies the second-order moments, we will call it the second-order representation for the time series Y_t. *Wold decomposition theorem* (see for example Brockwell and Davis 1991) shows that under fairly general conditions any stationary time series will have

K.F. Turkman et al., *Non-Linear Time Series*, DOI 10.1007/978-3-319-07028-5_2,

the causal linear representation (2.1), but these processes are not necessarily linear processes given in (2.2) and in fact, they can be highly nonlinear processes. Hence, Y_t in (2.1) is a representation for infinitely many time series, having the same (finite) second-order moments. On the other hand, if (Z_t) in (2.1) have marginal Normal distribution $N(0, \sigma^2)$, then they must also be independent. In this case, (2.1) and (2.2) are identical Gaussian processes. Uncorrelated versus independent innovations in (2.1) and (2.2) also have significant different effects on predictions. Y_{t+1} can be written as

$$\begin{aligned} Y_{t+1} &= \sum_{i=0}^{\infty} \psi_i Z_{t+1-i} \\ &= \psi_0 Z_{t+1} + \sum_{i=1}^{\infty} \psi_i Z_{t+1-i} \\ &= \psi_0 Z_{t+1} + \sum_{i=0}^{\infty} \psi_{i+1} Z_{t-i}. \end{aligned}$$

Similarly

$$X_{t+1} = \psi_0 Z^*_{t+1} + \sum_{i=0}^{\infty} \psi_{i+1} Z^*_{t-i}.$$

Let $\mathcal{B}_Z(t)$ be the σ-field generated by the r.v's $(Z_s, s \le t)$. The best mean-square predictor of Y_{t+1} in terms of $(Z_t, Z_{t-1}, \dots)$ is given by the conditional expectation

$$E(Y_{t+1}|\mathcal{B}_Z(t)) = \psi_0 E(Z_{t+1}|\mathcal{B}_Z(t)) + \sum_{i=0}^{\infty} \psi_{i+1} Z_{t-i}$$

with

$$E(Z_{t+1}|\mathcal{B}_Z(t)) = \int_x x d F_{Z_{t+1}|\mathcal{B}_Z(t)}(x),$$

where $F_{Z_{t+1}|\mathcal{B}_Z(t)}(x)$ is the distribution of Z_{t+1} conditional on $(Z_t, Z_{t-1}, \dots)$. Note that Z_{t+1} is not independent of $Z_t, Z_{t-1}, \dots$, hence, in general

$$F_{Z_{t+1}|\mathcal{B}_Z(t)}(x) \neq F_{Z_{t+1}}(x)$$

and

$$E(Z_{t+1}|\mathcal{B}_Z(t)) \neq 0.$$

In fact, this term will typically be complex, nonlinear function of $(Z_t, Z_{t-1}, \dots)$. Hence, the best predictor of Y_{t+1} in terms of $(Z_t, Z_{t-1}, \dots)$ will be a nonlinear function. If the process (2.1) is invertible, then the sigma fields generated respectively by $(Z_s, s \le t)$ and $(Y_s, s \le t)$ are identical, hence

$$E(Y_{t+1}|Y_s, s \le t) = E(Y_{t+1}|Z_s, s \le t),$$

so that $E(Y_{t+1}|Y_s, s \le t)$ in general is a nonlinear function of $(Y_s, s \le t)$. If (Z_t) have Normal distribution, then the best predictor of Y_{t+1} in this case is a linear function of the past observations $(Y(s), s \le t)$. On the other hand, $F_{Z^*_{t+1}|\mathcal{B}^*_Z(t)}(x) = F_{Z^*_{t+1}}(x)$ and $E(Z^*_{t+1}|\mathcal{B}^*_Z(t)) = 0$, so that

$$E(X_{t+1}|X_s, s \le t) = E(X_{t+1}|Z^*_s, s \le t),$$

is a linear function of $(X_s, s \le t)$, irrespective of the marginal distribution $F(\cdot)$ of Z_t.

In order to understand better the relation between best predictions and nonlinearity, we look at the geometric interpretation of predictions.

2.1.2 Linear and Nonlinear Optimal Predictions

Consider a probability space $(\Omega, \mathcal{F}, P)$ and the collection $\mathcal{C}$ of all r.v's defined on this space with zero-mean and finite second-order moments. For any elements X, Y of $\mathcal{C}$, define the inner product $<X, Y> = E(XY)$, so that the norm is given by $||X|| = \sqrt{E(X^2)}$. Thus, two elements X and Y of $\mathcal{C}$ are orthogonal iff $E(XY) = 0$, in which case we write $X \perp Y$. For simplicity in notation, we assume that the elements of $\mathcal{C}$ have zero means. Alternatively, rather than restricting the class to 0 mean r.v's, we can define the inner product on $\mathcal{C}$ as $<X, Y> = E(XY) - E(X)E(Y)$ and the norm $||X|| = \sqrt{E(X - E(X))^2}$, and the properties would still hold. The norm convergence of any sequence X_n is then given by

$$\lim_{n\to\infty} ||X_n - X||^2 = \lim_{n\to\infty} E|X_n - X|^2 = 0,$$

which is the usual mean-square convergence and we denote it by

$$X_n \overset{m.s}{\to} X.$$

Note that $X_n \overset{m.s}{\to} X$, iff

$$E(X_n - X_m)^2 \to 0,$$

as $m, n \to \infty$, in which case we call the sequence a Cauchy sequence.

If all sequences of $\mathcal{C}$ converge in mean-square, then $\mathcal{C}$ is complete and hence is a Hilbert space (e.g., Brockwell and Davis 1991). Let (X_n) be a stationary time series, such that $E(X_n^2) < \infty$. The norm convergence or mean-square convergence implies that if

$$X_n \overset{m.s}{\to} X$$

and

$$Y_n \overset{m.s}{\to} Y,$$

then

1. $E(X_n) \to E(X)$;
2. $E|X_n|^2 \to E|X|^2$, so that the variance of X_n converges to the variance of X;
3. $E(X_n Y_n) \to E(XY)$, so that the covariance and correlations between X_n and Y_n converge to the covariance and correlation between X and Y.

Now, let $\mathcal{C}_1$ be any closed subspace of $\mathcal{C}$. Then, from the *projection theorem*, for any $Y \in \mathcal{C}$, there is a unique element $\hat{X} = P_{\mathcal{C}_1} X \in \mathcal{C}_1$ such that

$$||Y - P_{\mathcal{C}_1} X||^2 = \inf_{X \in \mathcal{C}_1} ||Y - X||^2 = \inf_{X \in \mathcal{C}_1} E|Y - X|^2.$$

We know that the value of X which minimizes the mean-square error $E|Y - X|^2$ is given by $E(Y|X)$, so that the projection $P_{\mathcal{C}_1} X$ is the conditional expectation of Y given $\mathcal{C}_1$, and we denote it by $E_{\mathcal{C}_1}(Y)$. By the projection theorem $E_{\mathcal{C}_1}(Y)$ is a unique element X of $\mathcal{C}_1$ which satisfies

$$E(X E_{\mathcal{C}_1}(Y)) = E(XY),$$

for every $X \in \mathcal{C}_1$. We now define this conditional expectation in terms of multivariate r.v's in time series setting: let $(X_1, X_2, \ldots, X_n)$ be r.v's defined on $\{\Omega, \mathcal{F}, P\}$ and $Y \in \mathcal{C}$. Define the subspace $\mathcal{C}_1 = \mathcal{C}_1(X_1, X_2, \ldots, X_n)$ as the space of r.v's consisting of $X_1, X_2, \ldots, X_n$ and all other r.v's obtained by measurable transformations $f(X_1, X_2, \ldots, X_n)$. $\mathcal{C}_1$ is a closed subspace of $\mathcal{C}$. For any $Y \in \mathcal{C}$, let $P_{\mathcal{C}_1} Y = P_{\mathcal{C}_1(X_1,\ldots,X_n)} Y$ the projection of Y in $\mathcal{C}_1(X_1, \ldots, X_n)$. We define $P_{\mathcal{C}_1(X_1,\ldots,X_n)} Y = E_{\mathcal{C}_1(X_1,\ldots,X_n)}(Y)$ to be the conditional expectation of Y given $(X_1, \ldots, X_n)$. By the projection theorem, this conditional expectation is unique and can be obtained from the prediction equation

$$E(X E_{\mathcal{C}_1(X_1,X_2,\ldots,X_n)}(Y)) = E(XY), \tag{2.3}$$

for every element $X \in \mathcal{C}_1(X_1, \ldots, X_n)$. However, elements $X \in \mathcal{C}_1$ are in general nonlinear functions $f(X_1, \ldots, X_n)$ of $(X_1, \ldots, X_n)$ and therefore obtaining this unique conditional mean using the prediction equation (2.3) in general is

very difficult. However, there is one particular case, when this unique projection $P_{\mathcal{C}_1(X_1,X_2,\ldots,X_n)}X = E_{\mathcal{C}_1(X_1,X_2,\ldots,X_n)}(Y) = E_{X_1,\ldots,X_n}(Y)$ can be calculated with ease: Restrict $\mathcal{C}_1(X_1,\ldots,X_2)$ to be the closed span of $(X_1,\ldots,X_n)$ so that we only consider linear functions $f(X_1,\ldots,X_n) = \sum_{i=1}^n \alpha_i X_i$, and any element of $X \in \mathcal{C}_1(X_1,\ldots,X_n)$ is given by $X = \sum_{i=1}^n \alpha_i X_i$. In this case the optimal projection of any $Y \in \mathcal{C}$ into $\mathcal{C}_1$ is a linear function, and $\hat{Y} = P_{\mathcal{C}_1(X_1,\ldots,X_n)}(Y) = E(Y|X_1,\ldots,X_n) = \sum_{i=1}^n \alpha_i^* X_i$.

We call $\hat{Y}$ the best linear predictor for Y. This unique function can be obtained from the prediction equation by solving the set of equations

$$\sum_{i=1}^{n} \alpha_i^* E(X_i X_j) = E(YX_j), \tag{2.4}$$

for $j = 1, 2, \ldots, n$. However, the best linear predictor need not be the best predictor, since the best linear predictor is chosen within the closed span of $(X_1,\ldots,X_n)$,

$$\mathcal{C}_1 := \{X_1, X_2, \ldots, X_n \text{ and all linear functions of } (X_1,\ldots,X_n)\},$$

whereas the best predictor is chosen within the closed subspace

$$\mathcal{C}_1^* := \{X_1, X_2, \ldots, X_n \text{ and all measurable functions of } (X_1,\ldots,X_n)\}.$$

Clearly $\mathcal{C}_1 \subset \mathcal{C}_1^*$. The following definition is immediate.

Definition 2.1.1. A best linear prediction of Y in terms of a countable collection of r.v's $(X_t, t \in T)$ is defined to be the element of the closed span $\mathcal{C}_1$ of $(X_t, t \in T)$ which has the smallest mean-square distance from Y, and by the projection theorem is unique. On the other hand, the best predictor of Y in terms of the collection $(X_t, t \in T)$ is defined to be the element of the closed subspace $\mathcal{C}_1^*$ formed by all measurable functions of $(X_t, t \in T)$.

This definition will be extremely useful in discussing linear and nonlinear time series models. In general, $\mathcal{C}_1 \subsetneqq \mathcal{C}_1^*$ and $\mathcal{C}_1 = \mathcal{C}_1^*$, if $(Y, X_t, t \in T)$ have joint multivariate Normal distribution.

Example 2.1.1 (Brockwell and Davis 1991). Assume that $Y = X^2 + Z$, where X and Z are independent standard Normal r.v's. Let $\mathcal{C}^*(X)$ be the closed space formed by X and all measurable functions ϕ of X. By the projection theorem, the best mean-square predictor of Y in $\mathcal{C}^*(X)$ is the unique element $E_{\mathcal{C}(X)}(Y)$ of $\mathcal{C}(X)$, which satisfies

$$E(\phi(X)E_{\mathcal{C}(X)}(Y)) = E(\phi(X)Y).$$

$E_{\mathcal{C}^*(X)}(Y)$ is an element of $\mathcal{C}^*(X)$, so that $E_{\mathcal{C}^*(X)}(Y) = \phi^*(X)$ for some measurable function ϕ^* of X so that

$$E(\phi(X)\phi^*(X)) = E(\phi(X)Y)$$
$$= E(\phi(X)X^2) + E(\phi(X)Z),$$

since X and Z are independent, for any measurable function ϕ, $\phi(X)$ and Z are also independent, hence $E(\phi Z) = E(\phi)E(Z) = 0$. Now, the only measurable function ϕ^* of X which satisfies

$$E(\phi(X)\phi^*(X)) = E(\phi(X)X^2),$$

is $\phi^*(X) = X^2$, hence by the projection theorem, the best mean-square predictor $E_{C^*(X)}(Y)$ of Y is indeed the conditional expectation $E(Y|X) = X^2$ ($E(Z|X) = 0$ due to the independence of X and Z).

Now consider the best linear mean-square predictor of Y, that is, the best mean-square predictor of Y residing in $C(X)$, the closed span of X. Then $E_{C(X)}(Y) = aX + b$, satisfying

$$E[(aX + b)\phi(X)] = E[\phi(X)(X^2 + Z)],$$

for any $\phi(X)$ in $C(X)$. In particular, $\phi(X) = 1$ and $\phi(X) = X$ are in the closed span of X. Consequently from the prediction equations

$$< aX + b, 1 >=< y, 1 >= E(Y) = E(X^2) = 1,$$

and

$$< aX + b, X >= E(YX) = 0.$$

Solving for a and b gives $a = 0$ and $b = 1$, and the best linear predictor of Y in terms of X is given by $P_{C(X)}(X) = 1$. The prediction error of the best predictor is

$$||E(Y|X) - Y||^2 = E(Z^2) = 1,$$

whereas the prediction error of the best linear predictor is

$$||X^2 + Z - 1||^2 = E(X^2 + Z - 1)^2 = E(X^4) + E(Z^2) - 1 = 3.$$

Hence, the best linear predictor has three times as much prediction error as the best mean-square predictor, showing its clear inferior performance.

The above arguments can be applied to predict a future value of a time series. Consider a discrete parameter time series (X_t) defined on $(\Omega, \mathcal{F}, P)$, with zero mean and autocovariance function $\gamma(h)$. Consider the problem of best predictor of X_{n+1} in terms of $X_1, X_2, \ldots, X_n$. Clearly X_{n+1} and $X_1, \ldots, X_n$ are all elements of the Hilbert space with inner product $< X_i, X_{i+h} >:= E(X_i X_{i+h}) = \gamma(h)$, and norm $||X_i||^2 = \gamma(0)$. (Note that the mean is assumed to be zero, so that

$E(X_i X_{i+h}) = \gamma(h)$. Otherwise, we either study the series $X_t - E(X_t)$, or equivalently define the inner product to be

$$< X_i, X_{i+h} >= E((X_i - E(X_i))(X_{i+h} - E(X_{i+h}))) = \gamma(h).$$

Therefore, the assumption of zero mean is not restrictive.)

Consider the closed subspace $\mathcal{C}_1^*$ which includes the r.v's $X_1, \ldots, X_n$ and all measurable functions of $(X_1, \ldots, X_n)$. Clearly such closed subspace will include the closed span $\mathcal{C}_1$ of $(X_1, \ldots, X_n)$. From the projection theorem, the best predictor of X_{n+1} as a function of $(X_1, \ldots, X_n)$ is a unique element of $Y \in \mathcal{C}_1$ which has the smallest mean-square distance from X_{n+1}, that is a function $\hat{Y} = f(X_1, \ldots, X_n)$ such that

$$||X_{n+1} - \hat{Y}||^2 = \inf_{Y \in \mathcal{C}_1^*} E|X_{n+1} - Y|^2.$$

The projection theorem also says that $\hat{Y} = E_{\mathcal{C}_1^*}(X_{n+1}) = E(X_{n+1}|X_1, X_2, \ldots, X_n)$, can uniquely be obtained by solving the prediction equations

$$E(Y\hat{Y}) = E(YX_{n+1}),$$

for every $Y \in \mathcal{C}_1^*$. Since, Y is any (nonlinear) measurable function $f(X_1, \ldots, X_n)$, it is not easy to get the optimal predictor of X_{n+1} using the prediction equation (2.3). However, if we restrict ourselves to the closed span $\mathcal{C}_1$ of $(X_1, \ldots, X_n)$, we can solve the prediction equation to obtain the best projection of X_{n+1} into the closed span $\mathcal{C}_1$, namely the unique best linear mean-square predictor. In this case, all elements of $\mathcal{C}_1$ are of the form $Y = \sum_{i=1}^n \alpha_i X_i$, for some real numbers $\alpha_i, i = 1, \ldots, n$ therefore the best linear predictor of X_{n+1} is an element

$$\hat{X}_{n+1} = \sum_{i=1}^{n} \alpha_i^* X_i,$$

where α_i^* are obtained uniquely from the prediction equations given by

$$\sum_{i=1}^{n} \alpha_i^* E(X_i X_j) = E(X_{n+1} X_j),\ j = 1, 2, \ldots, n. \tag{2.5}$$

Writing $\boldsymbol{\alpha}^* := (\alpha_1^*, \ldots, \alpha_n^*)$, and

$$\boldsymbol{\Gamma}_n := \begin{pmatrix} \sigma^2 & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \sigma^2 & \cdots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \cdots & \cdots & \sigma^2 \end{pmatrix}$$

and $\boldsymbol{\gamma}_n := (\gamma(1), \ldots, \gamma(n))'$, we can write (2.5) as

$$\boldsymbol{\Gamma}_n \boldsymbol{\alpha}_n^* = \boldsymbol{\gamma}_n. \tag{2.6}$$

This system of equations will have a unique solution, provided $\boldsymbol{\Gamma}_n$ is not singular, which is satisfied when the function $\gamma(h)$ positive definite. If $\boldsymbol{\Gamma}_n$ is singular, then the best linear predictor of X_{n+1} will have infinitely many alternative representations in terms of $X_1, \ldots, X_n$.

Although simpler to calculate, best linear predictors often are inferior to best predictors, unless the relationship between X_{n+1} and $X_1, \ldots, X_n$ is linear; see Example 2.1.1. Note that if X_n is a Gaussian time series, then the conditional expectation $E(X_{n+1}|X_1, \ldots, X_n)$ is a linear function of $(X_1, \ldots, X_n)$ and in this case the best mean-square predictor and the best linear mean-square predictors coincide.

Example 2.1.2 (Brockwell and Davis 1991). Consider the stationary discrete time series

$$X_t = A\cos(\omega t) + B\sin(\omega t),\ t \in \mathbb{Z}, \tag{2.7}$$

where $\omega \in (0, \pi)$ is a constant, A and B are uncorrelated r.v's with zero-mean and variance σ^2. The mean and the variance of the series are given respectively by $E(X_t) = 0$ and

$$V(X_t) = \cos^2(\omega t)V(A) + \sin^2(\omega t)V(B) = \sigma^2.$$

For any h

$$\begin{aligned}\gamma(h) &= E(X_t X_{t+h}) \\ &= \sigma^2(\cos(\omega t)\cos\omega(t+h) + \sin(\omega t)\sin\omega(t+h)) \\ &= \sigma^2\cos(\omega h),\end{aligned}$$

so that the time series (2.7) is second-order stationary. Now consider the best linear predictor of X_3 given by

$$\hat{X}_3 = \alpha_1 X_1 + \alpha_2 X_2.$$

From (2.6) it follows that

$$\begin{bmatrix} \sigma^2 & \gamma(1) \\ \gamma(1) & \sigma^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \end{bmatrix}. \tag{2.8}$$

Solving (2.8) for (α_1, α_2), we get $\alpha_1 = 2\cos(\omega)$, $\alpha_2 = -1$, so that the best linear predictor is given by $\hat{X}_3 = 2\cos(\omega)X_2 - X_1$. Note that the prediction error is

$$\begin{aligned} E(X_3 - \hat{X}_3)^2 &= E((X_3 - 2\cos(\omega)X_2 + X_1))^2 \\ &= E((X_3 - X_1)^2 - 4\cos(\omega)X_2(X_3 - X_1) + 4\cos^2(\omega)X_2^2) \\ &= 2\sigma^2 - 2\gamma(2) + 4\cos^2(\omega)\sigma^2 \\ &= 2\sigma^2(1 - \cos(2\omega)) + 4\cos^2(\omega)\sigma^2 \\ &= 0, \end{aligned}$$

since for any ω, $cos(2\omega) = 2\cos^2(\omega) - 1$. Hence X_3 is predicted from X_2 and X_1 without any error, which means that

$$X_3 \equiv 2\cos(\omega)X_2 - X_1.$$

Similarly, from stationarity

$$\hat{X}_4 = 2\cos(\omega)X_3 - X_2,$$

with a mean-square error 0. The projection theorem guarantees that there is a uniquely defined predictor $\hat{X}_4$. However, $\hat{X}_4$ has infinitely many linear representations in terms of X_1, X_2, X_3, but by the projection theorem they should give the same predictor. This is due to the fact that $(\alpha_1, \alpha_2, \alpha_3)$ in the representation $\hat{X}_4 = \sum_{i=1}^{3} \alpha_i X_i$ satisfies

$$\begin{bmatrix} \sigma^2 & \gamma(1) & \gamma(2) \\ \gamma(1) & \sigma^2 & \gamma(1) \\ \gamma(2) & \gamma(1) & \sigma^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \gamma(3) \end{bmatrix}. \tag{2.9}$$

However, the 3×3 matrix on the right and side of Eq. (2.9) is singular, giving infinitely many solutions for $(\alpha_1, \alpha_2, \alpha_3)$. It is easy to check that the determinant

$$|\Gamma_3| = \begin{vmatrix} 1 & \cos(\omega) & \cos(2\omega) \\ \cos(\omega) & 1 & \cos(\omega) \\ \cos(2\omega) & \cos(\omega) & 1 \end{vmatrix} = 0.$$

In fact, for any $h > 0$, the future values of the time series X_{t+h} given in (2.7) can be predicted with 0 mean-square error in terms of the linear combination of its observed values. Notice also that the time series (2.7), as well as its covariance function, is periodical with period 2π.

Definition 2.1.2. We call a time series deterministic, if for any $h > 0$, the optimal predictor of X_{t+h}, $\hat{X}_{t+h}$ can be predicted in terms of $(X_t, X_{t-1}, \dots)$ with zero prediction error.

2.1.3 Nonlinear Representations

In the previous section, we saw that if we are interested only in linear predictors due to its simplicity, then from (2.6), we only need to know the second-order moments to calculate the best linear predictor. Due to the Wold decomposition theorem, (2.1) is the most general model we can use for obtaining such linear predictions. However, we also see that unless the process is Gaussian, the best linear predictor is inferior to the best predictor which is a nonlinear function of the observed time series. Suppose that our time series is not Gaussian and we are not content with the best linear predictor. In this case, we will have to look beyond linear processes and second-order covariance structures. This situation is very common particularly in environmental sciences and economy.

The crucial restriction in the Wold decomposition theorem is that the linear representation is given in terms of an uncorrelated white noise process, so that this representation serves as a model only for the second order moments of the stationary process. Under what conditions, can we represent a (strictly) stationary process in terms of an independent and identically distributed input process (Z_t)? If this is possible, then we should be able to model the whole probability structure of the process in terms of this independent and identically distributed input process.

In Sect. 2.1.2, in order to obtain best linear predictor we looked at the Hilbert space generated by the closed span $\mathcal{C}_1$ of $(X_s, s \leq t)$, with the inner product $< X, Y >= Cov(X, Y)$. The members of this Hilbert space are made up of only the linear combinations of $(X_s, s \leq t)$ and their mean-square limits. The projection theorem then gave us the optimal linear predictors for X_{t+h} as projection of X_{t+h} in this closed span. If we want to extend these results to optimal (nonlinear) projections, we need to look for much more general setup. Now, consider again the set $(X_s, s \leq t)$ and consider the set of all r.v's with finite variance which are measurable with respect to this set, that is the set

$$\mathcal{C}_2 := \{Y = g(X_s), s \leq t : G \text{ measurable and } V(Y) \leq \infty\}.$$

This subspace is a Hilbert space and clearly contains the closed span of $(X_s, s \leq t)$. If we can find a closed orthogonal basis for this subspace, then any element Y of this subspace can be written as a linear combination of the orthogonal basis functions, and projection theorem will give us the optimal projection of X_{t+h} in terms of the elements of this subspace.

Definition 2.1.3. Hermite polynomials $H_n(x)$ of degree n are defined as

$$\int_{-\infty}^{\infty} H_n(x) H_m(x) \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx = n! I_{n,m}, \quad n, m = 0, 1, 2, \ldots \tag{2.10}$$

where

$$I_{n,m} = \begin{cases} 1, n = m; \\ 0, n \neq m. \end{cases}$$

These polynomials form a closed and complete orthogonal system in the Hilbert space $\mathcal{L}^2(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx)$ where the inner product is defined as

$$< f, g >= \int_{-\infty}^{\infty} f(x)g(x)\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx. \tag{2.11}$$

Hence, every Borel measurable function g such that

$$\int_{-\infty}^{\infty} g^2(x)\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx < \infty,$$

can be written as a linear combination (or as a limit) of these Hermite polynomials

$$g(x) = \lim_{N\to\infty} \sum_{n=0}^{N} \frac{g_n}{n!} H_n(x), \tag{2.12}$$

where, the coefficients g_n are given by

$$g_n = \int_{-\infty}^{\infty} g(x)H_n(x)\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx.$$

The convergence of (2.12) is in the mean-square sense

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} (g(x) - \sum_{n=0}^{N} \frac{g_n}{n!} H_n(x))^2 \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx = 0.$$

Hermite polynomials are given by

$$H_n(x) = (-1)^n \frac{1}{\sqrt{2\pi}} \exp(x^2/2) \frac{d^n}{dx^n} \exp(-x^2/2),$$

and they can also be calculated from the recursions

$$H_{n+1}(x) = xH_n(x) - \frac{d}{dx} H_n(x),$$

or

$$H_n(x) = xH_n(x) - nH_{n-1}(x).$$

The first five Hermite polynomials are given by

$$H_0(x) = 1$$
$$H_1(x) = x$$

$$H_2(x) = x^2 - 1$$
$$H_3(x) = x^3 - 3x$$
$$H_4(x) = x^4 - 6x^2 + 3.$$

Note that the inner product (2.11) is an integral with respect to the standard Gaussian density and hence the Hermite polynomials are orthogonal with respect to the standard normal probability distribution. Instead of Hermite polynomials, we can define Hermite functions

$$\psi_n(x) := \frac{1}{\sqrt{n!2^n\sqrt{2\pi}}} \exp(-x^2/2) H_n(x).$$

Hermite functions are normalized versions of the Hermite polynomials, therefore they form a closed and complete orthonormal basis for $\mathcal{L}^2(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx)$. The Hermite polynomials are orthogonal with respect to the standard Normal distribution, although it is possible to define Hermite polynomials which are orthogonal with respect to the Normal distribution $N(0, \sigma^2)$. The closed linear span of Hermite polynomials is the space of all polynomials, therefore any element of $\mathcal{L}^2(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx)$ can be written as a polynomial of finite- or infinite-order. Elements of $\mathcal{L}^2(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx)$ are deterministic functions. How can we pass from polynomial representation for deterministic functions to random functions? Consider now the simple case: let X be a standard Gaussian random variable and consider the set of all r.v's Y which are measurable functions of X with finite variances, that is the set

$$\mathcal{C}(X) := \{Y = g(X) : g \text{ measurable and } V(Y) < \infty\}.$$

Define the inner product $< Y_1, Y_2 >= Cov(Y_1, Y_2)$ on this set. This set forms a Hilbert space. The above results on Hermite polynomials immediately suggest the construction of the orthogonal base for this Hilbert space; let $H_n(X), n = 0, 1, 2, \ldots$ be r.v's, where $H_n(x)$ are Hermite polynomials defined in (2.10). Then any measurable function Y (of X) can be written as

$$Y = \sum_{n=0}^{\infty} \frac{g_n}{n!} H_n(X),$$

where

$$g_n = Cov(Y, H_n(X))$$

and

$$\lim_{N\to\infty} V(Y - \sum_{n=0}^{N} \frac{g_n}{n!} H_n(X)) = 0.$$

Note that if we restrict ourselves to the Hilbert space of the closed span generated by X, then any member of this space is a linear function of X. If we extend this space to include all measurable functions with finite variances, then the elements are again represented by a linear function, but this time a linear combination of (random and nonlinear) Hermite polynomials or simply polynomials of finite- or infinite-order.

Now let us introduce more complexity and start with a collection of standard Gaussian r.v's $(X_s, s \le t)$ and consider the space of all measurable functions defined on this collection with the usual inner product defined over it. Any element of this Hilbert space can be written as a linear combination of products of Hermite polynomials. Here we will not enter into details, which can be found in Terdik (1999). As an example, consider standard Gaussian r.v's $(X_1, X_2, \ldots, X_n)$ with covariances $r(i, j)$. The first five (random) Hermite polynomials which form the orthogonal basis for the Hilbert space of all measurable functions defined on $(X_1, X_2, \ldots, X_n)$ are given by

$$\begin{aligned}
H_0 &= 1\\
H_1(X_1) &= X_1\\
H_2(X_1, X_2) &= X_1X_2 - r(1,2)\\
H_3(X_1, X_2, X_3) &= X_1X_2X_3 - r(1,2)X_3 - r(1,3)X_2 - r(2,3)X_1\\
H_4(X_1, X_2, X_3, X_4) &= X_1X_2X_3X_4 - r(1,2)X_3X_4 - r(1,3)X_2X_4\\
&\quad - r(1,4)X_2X_3 - r(2,3)X_1X_4 - r(2,4)X_1X_3 - r(3,4)X_1X_2\\
&\quad + r(1,2)r(2,3) + r(1,3)r(2,4) + r(1,4)r(2,3).
\end{aligned}$$

Therefore any element of this Hilbert space can be represented as sums of products of polynomials given in the form

$$\sum_{p=0}^{\infty}\sum_{i_1=1}^{\infty}\cdots\sum_{i_p=1}^{\infty} a_{i_1 i_2 \cdots i_p} \prod_{v=1}^{p} X_{i_v},$$

with the convention $\prod_{v=1}^{0} X_{i_v} = 1$.

The following remarkable result due to Nisio (1960) extends this polynomial representation to any strictly stationary time series.

Definition 2.1.4. Let Z_t be independent, standard Gaussian r.v's. The polynomial representation

$$
\begin{aligned}
Y_t^{(m)} &= \sum_{p=1}^{m} \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} \cdots \sum_{i_m=-\infty}^{\infty} g_{i_1 i_2 \cdots i_m} \prod_{v=1}^{p} Z_{t-i_v} \\
&= \sum_{i_1=-\infty}^{\infty} g_{i_1} Z_{t-i_1} \\
&+ \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} g_{i_1 i_2} Z_{t-i_1} Z_{t-i_2} \\
&+ \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} \sum_{i_3=-\infty}^{\infty} g_{i_1 i_2 i_3} Z_{t-i_1} Z_{t-i_2} Z_{t-i_3} \\
&+ \cdots \\
&+ \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} \cdots \sum_{i_m=-\infty}^{\infty} g_{i_1 i_2 \cdots i_m} Z_{t-i_1} Z_{t-i_2} \cdots Z_{t-i_m},
\end{aligned}
$$

is called a Volterra series of order m. We will call

$$
Y_t = \sum_{p=1}^{\infty} \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} \cdots \sum_{i_p=-\infty}^{\infty} g_{i_1 i_2 \cdots i_p} \prod_{v=1}^{p} Z_{t-i_v}, \tag{2.13}
$$

the (infinite-order) Volterra series expansion.

Theorem 2.1.1 (Nisio 1960). *Let X_t be any strictly stationary time series. Then there exists a sequence of Volterra series $Y_t^{(m)}$ such that*

$$
\lim_{m \to \infty} Y_t^{(m)} \stackrel{d}{=} X_t,
$$

in the sense that for any n and for any $\theta_j, |j| \leq n$ as $m \to \infty$,

$$
|E \exp(i\theta_{-n} X_{-n} + \cdots + i\theta_n X_n) - E \exp(i\theta_{-n} Y_{-n}^{(m)} + \cdots + i\theta_n Y_n^{(m)})| \to 0.
$$

If further X_t is Gaussian, then X_t can be represented by

$$
X_t = \sum_{j=-\infty}^{\infty} g_j Z_{t-j}.
$$

The proof of the result above is beyond the scope of this book. However, we only mention that the proof is centered around first finding a polynomial representation for a uniformly bounded time series using Hermite polynomials and then extending the results to any time series using Slutsky type arguments. Although assumption of independence of the innovations Z_t is essential, normality is not essential.

One can define Hermite polynomials orthogonal with respect to any probability distribution and therefore the Volterra representation can be given in terms of any other distribution.

Nisio's theorem essentially says that although most stationary time series will have a linear representation in terms of uncorrelated innovations (Wold theorem), it will have very complicated, nonlinear representations in terms of the independent innovations. Therefore Nisio's theorem can be seen as the extension of the Wold decomposition theorem. While modeling with ARMA classes, we often require that the innovations are Gaussian, hence the modeling is restricted to the Volterra series of order 1. Note that (2.13) is a representation for the whole probability structure of the time series as contrast to the representation (2.1), which is representation for the covariance structure of the series. Let us give some examples to highlight this difference.

Example 2.1.3. Consider the process

$$X_t = Z_t + \alpha Z_{t-1} Z_{t-2}, \ t \in \mathbb{Z},$$

where (Z_t) is a zero-mean i.i.d. sequence with finite variance. It is easy to verify that X_t is covariance stationary with zero mean and constant variance and

$$Cov(X_t X_{t+h}) = 0.$$

Hence, X_t is an uncorrelated time series, whose correlation structure is equivalent to that of the independent innovation process Z_t. However, the probability structure of X_t is different from that of Z_t. For example,

$$E(X_t | X_{t-1}, X_{t-2}, \dots) = \alpha Z_{t-1} Z_{t-2},$$

whereas

$$E(Z_t | Z_{t-1}, \dots) = 0.$$

Hence, by looking at the second-order properties, we can decide that there is no structure in X_t to model, but certainly X_t has structure which should be studied by its higher-order moments. In fact, if Z_t are also Gaussian, then all cumulants higher then the second-order are zero. However, it is easy to check that the higher-order cumulants of X_t are not identically equal to zero.

Example 2.1.4 (All-pass models). The class of uncorrelated but not independent processes is quite rich. In fact, one can encounter uncorrelated but not independent linear processes. The class of all pass models (Andrews et al., 2006) is one example, which can be constructed within the ARMA class by choosing autoregressive and moving average polynomials in such a manner that the roots of the autoregressive polynomial are reciprocals of the roots of the moving average polynomial or vice-versa. Assume that $\phi_p(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ is a causal autoregressive

polynomial so that $\phi_p(z) \neq 0$ for $|z| \leq 1$. Define the moving average polynomial

$$\begin{aligned}\theta_p(z) &= \frac{z^p \phi_p(z^{-1})}{-\phi_p} \\ &= -(B^p - \phi_1 B^{p-1} - \cdots - \phi_p)/\phi_p,\end{aligned}$$

and consider the time series which satisfies the difference equation

$$\phi_p(B) X_t = \theta_p(B) Z_t,$$

where (Z_t) is an i.i.d. sequence with zero-mean and finite variance σ^2. The time series X_t has some interesting properties.

1. X_t is not invertible, but is causal.
2. The time series satisfies the difference equation

$$\begin{aligned}&X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} \\ &= Z_t + \frac{\phi_{p-1}}{-\phi_p} Z_{t-1} + \cdots + \frac{\phi_1}{\phi_p} Z_{t-p+1} - \frac{1}{\phi_p} Z_{t-p},\end{aligned}$$

   so that, when $p = 1$, and $|\phi_1| < 1$, first order all-pass model is given by

$$X_t - \phi_1 X_{t-1} = Z_t - \frac{1}{\phi_1} Z_{t-1}$$

   and the second-order all-pass model is given by

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t + \phi_1/\phi_2 Z_{t-1} - 1/\phi_2 Z_{t-2}.$$

3. The spectral density of X_t is given by the constant function

$$f(w) = \frac{\sigma^2}{\phi_p^2 2\pi},$$

   for every $w \in [-\pi, \pi]$, so that the X_t process is uncorrelated. Further, if Z_t are Gaussian, then X_t is an i.i.d. sequence with distribution $N(0, \phi_p^{-2}\sigma^2)$. However, if Z_t are not Gaussian, then for $p \geq 1$, X_t is not an independent sequence.
4. Since all-pass processes are uncorrelated but not independent, the usual second-order techniques based on autocorrelation and partial autocorrelation functions cannot identify an all-pass model, as these functions will report that the data have no structure. Inferential methods based on Gaussian likelihood or least squares do not give the desired results when fitting all-pass models. Instead, inferential techniques based on cumulants of order greater than two are often used; see Andrews et al. (2006) for details. The need for inferential methods

based on cumulants higher than two or approximate methods based on non-Gaussian likelihoods are quite universal while modeling nonlinear data.

Let us make a summary of the results:

1. $Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, Z_t uncorrelated r.v's is called a linear, causal representation.
2. $Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, Z_t i.i.d. r.v's is called a linear causal model.
3. If further, Z_t are Gaussian, then any linear representation is also a linear model and $Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ is called the Gaussian causal linear model.
4. (Almost) all non-deterministic, second-order stationary time series X_t have a unique linear representation in terms of uncorrelated innovations. In this case, moments of X_t and Y_t up to second-order coincide. However, moments of order higher then two, need not coincide, except when X_t is Gaussian.
5. (Almost) all strictly stationary time series X_t has a (infinite-order) Volterra series expansion

$$Y_t = \sum_{p=1}^{\infty} \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} \cdots \sum_{i_p=-\infty}^{\infty} g_{i_1 i_2 \cdots i_p} \prod_{v=1}^{p} Z_{t-i_v},$$

   for some i.i.d. innovation sequence Z_t.
6. Therefore, X_t has a linear causal model in terms of an i.i.d. innovation sequence Z_t iff it has a first order, one-sided Volterra series expansion, that is, iff X_t is a Gaussian process. Hence, the class of causal, linear models is not dense within the class of stationary time series.
7. If we want only the best linear predictors for future values of the time series, then we can work with linear causal representations, as we do not need information other than the second-order moments to obtain best linear predictors.
8. On the other hand, if we want the best predictor, then we need to look for models within the general class of Volterra series expansions.

Working with linear models, particularly with Gaussian linear model, is relatively simple, whereas working directly with the general, infinite order Volterra series is very difficult, if not impossible. For example, it not possible to give conditions of stationarity on the kernels $g_{i_1 i_2 \cdots i_p}$. Further, time series such as

$$X_t = Z_t + \alpha Z_{t-1} Z_{t-2},$$

or

$$X_t = Z_t + \alpha Z_{t-1}^2,$$

where Z_t is a sequence of independent r.v's, are not invertible (Granger and Andersen 1978). Hence, one would expect that Volterra series expansions have limited use as models for predicting future values, unless the input process (Z_t) is observable. Therefore, to model nonlinear data, we need to look for sub-classes of Volterra series expansions which are easier to study.

There are many ways a process can be nonlinear. Therefore, in order to come up with fairly general and useful classes of nonlinear models, we need to look at certain aspects of the probability structure of the processes to understand and describe the underlying nonlinear behavior. Since linear and nonlinear processes differ on moments higher than order two, particular emphasis has to be given to studying the higher moments and tails of the stationary distributions of the processes. We now look at certain aspects of nonlinear processes which may indicate how we should construct useful nonlinear models.

2.1.4 Sensitive Dependence on Initial Conditions, Lyapunov Exponents

The most striking feature of nonlinear processes is the strong dependence on initial conditions and the noise amplification. Let us start with deterministic difference equations, representing some dynamic system in discrete time. Suppose that $x_n = f(x_{n-1})$ defines a deterministic difference equation, for some function f. Starting from the initial condition x_0, let

$$x_n = f^{(n)}(x_0) = f(f(\cdots(f(x_0))))$$

be the value of the system after n iterations. Now let us disturb the initial starting value x_0 by a small number δ_0 to $x_0 + \delta_0$. We would be interested in the impact of this initial disturbance on the dynamic system after n iterations, namely

$$\delta_n = f^{(n)}(x_0 + \delta_0) - f^{(n)}(x_0),$$

and in particular, we may be interested in the limit as $n \to \infty$ and $\delta_0 \to 0$. If f is a linear function so that

$$x_n = \alpha x_{n-1} + \beta,$$

then it is easy to verify that

$$x_n = \alpha^n x_0 + \beta(\alpha^{n-1} + \alpha^{n-2} + \cdots + 1),$$

so that

$$\delta_n = \alpha^n \delta_0$$

and

$$\frac{f^{(n)}(x_0 + \delta_0) - f^{(n)}(x_0)}{f^{(n)}(x_0)} = O(\delta_0).$$

On the other hand, consider the logistic difference equation

$$x_{n+1} = \alpha x_n(1 - x_n). \tag{2.14}$$

Here, α is called the *driving parameter*. Now start with an initial value $x_0 \in (0, 1)$. This difference equation has a peculiar behavior for different values of α. If $\alpha \in [0, 3)$, then as $n \to \infty$, the difference equation converges to a single number. When $\alpha = 3.0$, then x_n no longer converges but oscillates between two values. As α is increased, in the limit x_n oscillates between increasingly different numbers, and for $\alpha > 3.57$ the sample path behavior of $x_1, \ldots, x_n$ is chaotic, resembling a sample path of a random process. This chaotic behavior is due to the sensitive dependence of the difference equation on the initial value x_0 for increasing values of the parameter α. In fact, when $\alpha = 4$, this difference equation has an analytical solution

$$x_n = \sin^2(2^n \beta\pi),$$

where $\beta \in [0, 1)$ is a function of the initial value x_0. When $x_0 \in [0, 1]$ then β is almost surely an irrational number which will have different dyadic representation for each iteration n causing a chaotic behavior of the sample path $x_1, \ldots, x_n$. This chaotic behavior caused by the dependence on the initial condition is quite common for nonlinear difference equations and measuring this dependence on initial conditions may give a degree of nonlinearity that exists in a difference equation (Fig. 2.1).

Lyapunov exponent λ of a dynamic system is a quantity that characterizes this dependence on the initial conditions through the relationship

$$\delta_n \sim e^{n\lambda}\delta_0. \tag{2.15}$$

One can give an heuristic argument for the definition of the Lyapunov exponent. Assume that $x_n = f(x_{n-1})$ and f is everywhere differentiable. Then using first-order Taylor series approximation,

$$\begin{aligned} \delta_n &= f^{(n)}(x_0) - f^{(n)}(x_0 + \delta_0) \\ &\sim \delta_0 \frac{d}{dx} f^{(n)}(x_0). \end{aligned}$$

Here,

$$\frac{d}{dx} f^{(n)}(x_0) = \frac{d}{dx} f^{(n)}(x + \delta_0),$$

calculated at $x = x_0$. Since

$$f^{(n)}(x) = f(f(\cdots(f(x)))),$$

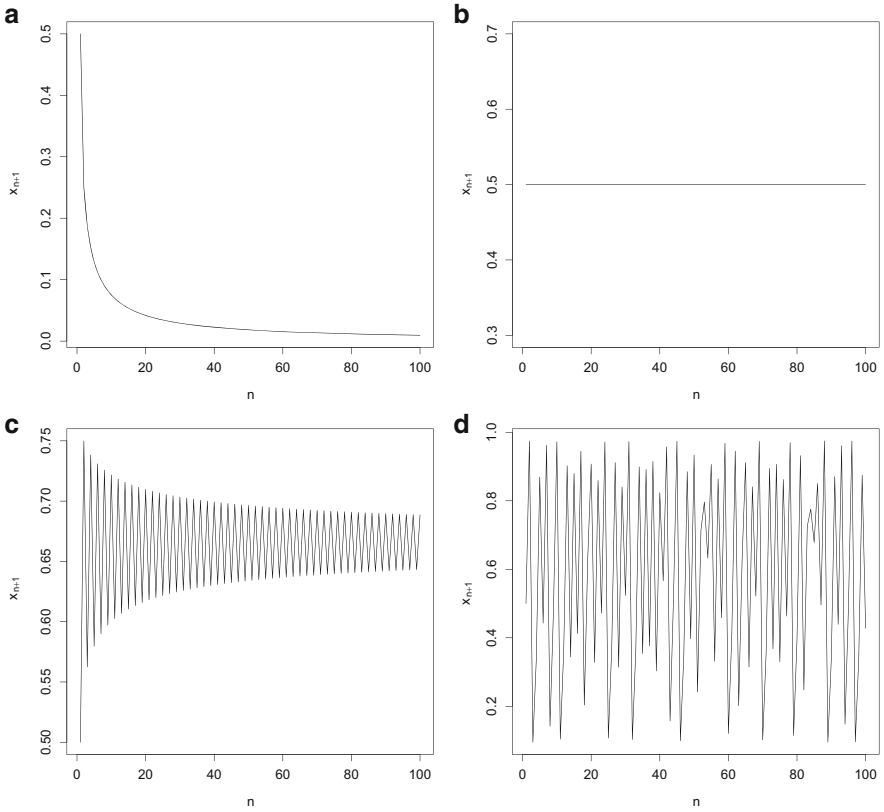

Fig. 2.1 Simulated samples of 100 observations from the logistic difference equation; (**a**) $\alpha = 1.0$; (**b**) $\alpha = 2.0$ with $x_0 = 0.5$; (**c**) $\alpha = 3.0$; and (**d**) $\alpha = 3.9$

by the chain rule

$$\frac{d}{dx} f^{(n)}(x_0) \sim \exp(n \log \frac{d}{dx} f(x_0)),$$

assuming that each of the factors $\frac{d}{dx} f(x_n) \sim \frac{d}{dx} f(x_0)$ have comparable sizes. Therefore it is reasonable to consider

$$\lambda = \lim_{n \to \infty} \frac{1}{n} \ln |\frac{d}{dx} f^{(n)}(x_0)|,$$

as an indicator of the degree of dependence on the initial conditions, or equivalently, as an indicator of the degree of nonlinearity through the relationship (2.15). When

$\lambda < 0$, the dynamic system is called dissipative or non-conservative. Such a dynamic system exhibits asymptotic stability, typically resulting from damped harmonic oscillations. When $\lambda = 0$, the system is called conservative and is said to be Lyapunov stable. The case $\lambda > 0$ corresponds to an unstable system, resulting in chaotic sample paths.

Quantifying the degree of dependence on initial conditions or equivalently quantifying the degree of nonlinearity of stochastic difference equations representing dynamic random systems needs more attention. This is due to the fact that the system in each iteration is perturbed by a random noise. Since each sample path of the dynamic system will have different realizations of random shocks, it makes sense to consider the divergence of expected values (ensemble average) of these sample paths.

Example 2.1.5. Consider the stochastic difference equation

$$X_{n+1} = A_n X_n + B_n, \ n \geq 0. \tag{2.16}$$

Here, for each n, A_n and B_n are dependent scalar r.v's, but the pair (A_n, B_n) is an i.i.d. sequence. The stochastic difference equation in (2.16) and its multivariate versions, where X_n, B_n are random vectors in $\mathbb{R}^d$ and A_n are $d \times d$ matrices, often appear as basis for studying many different forms of nonlinear time series and will be revisited in future chapters. Starting from X_0 and upon n iterations the process will be in the state (Brandt 1986)

$$X_n = \sum_{j=0}^{n-1} \left(\prod_{i=n-j}^{n-1} A_i \right) B_{n-j-1} + \left(\prod_{i=0}^{n-1} A_i \right) X_0.$$

Conditional on the two initial values $X_0 = x_0$, $X_0 = x_0 + \delta_0$, we can quantify the deviation in the sample paths with

$$\begin{aligned} \delta_n &= f^{(n)}(x_0) - f^{(n)}(x_0 + \delta_0) \\ &= \left(\prod_{i=0}^{n-1} A_i \right) \delta_0. \end{aligned}$$

Note that δ_n is a random variable. Lyapunov exponent λ can now be defined as the expected deviation on the sample paths upon n iteration, conditional on two initial realizations $X_0 = x_0$, $X_0 = x_0 + \delta_0$ through the relation

$$E(\delta_n) = e^{n\lambda} \delta_0,$$

in which case

$$\begin{aligned}\lambda &= E[\frac{1}{n}\log(|A_0||A_1|\cdots|A_n|)] \\ &= E\log(A_0).\end{aligned}$$

(Here, $|A_i|$ rather than A_i appear in the expression to insure that $\frac{d}{dx}f(x_n) = A_n$ have comparable sizes and contributions). Brandt (1986) also shows that if the process is dissipative, that is, $\lambda = E\log|A_0| < 0$ and $E(\log|B_0|)^+ < \infty$ then

$$\lim_{n\to\infty} X_n = \sum_{j=0}^{\infty}\left(\prod_{i=n-j}^{n-1} A_i\right) B_{n-j-1},$$

is the unique stationary solution of (2.16). If $\lambda > 0$, one would expect a chaotic behavior, without a stationary limit for the difference equation. Hence, the existence of stationary solutions for the stochastic difference equation given in (2.16) depends on the degree of dependence on the initial conditions.

Example 2.1.6 (Fan and Yao 2003). Consider again the dynamic system defined by the deterministic difference equation $x_t = f(x_{t-1})$, where f is an everywhere differentiable function. But now we disturb the dynamic system at each iteration by a small i.i.d. noise Z_t, resulting in the stochastic difference equation

$$X_t = f(X_{t-1}) + Z_t.$$

The process that satisfies this difference equation is called the first order nonlinear autoregressive model of order 1 (NLA(1)). In order to facilitate arguments, assume further that Z_t are independent of $(X_s, s < t)$. It may be interesting to know how much these additive noises affect the variation in this process after n steps. Again, let us consider two sample paths of this process, starting from $X_0 = x_0$ and $X_0 = x_0 + \delta_0$ and look at how much (on average) these two sample paths diverge after n iterations. Note that, if f is a linear function, then with any uncorrelated noise with finite variance, the divergence between these two sample paths would be of order $O(\delta_0)$. Let $f^{(n)} = f(f(\cdots f(x)))$ be the n fold composition of f. Then by the arguments given in Fan and Yao (2003) which are based on iterative Taylor series expansions,

$$X_n = f^{(n)}(X_0) + \sum_{j=1}^{n-1}\prod_{k=j}^{n-1} f'(X_{n-k})Z_{n-j+1} + Z_n. \tag{2.17}$$

In general the derivatives $f'(X_{n-k})$ are functions of $Z_{n-k}, Z_{n-k-1}, \ldots Z_1$. However, if we assume that the random shocks are of small order, that is $|Z_n| < \eta < 1$ almost surely for every n, then by (2.17) for any fixed n,

$f(X_n) \sim f^{(n)}(X_0) + O(\eta)$, and this can be used as a second-order approximation in the arguments of the derivates to give

$$f(X_n) \sim f^{(n)}(X_0) + \sum_{j=1}^{n-1} \prod_{k=j}^{n-1} f'(f^k(X_0)) Z_{n-j} + Z_n + O(\eta). \tag{2.18}$$

Let $\sigma_n^2(x_0) := V(X_n | X_0 = x_0)$ be the variance of the process after n iterations. Then from (2.18),

$$\sigma_n^2(x_0) = (1 + \sum_{j=1}^{n-1} \prod_{k=j}^{n-1} f'(f^{(k)}(x_0)))^2 \sigma^2 + o(\eta).$$

Hence, even when the shocks Z_t are almost surely small, the variance of the process after n iterations is amplified by a quantity $(1 + \sum_{j=1}^{n-1} \prod_{k=j}^{n-1} f'(f^{(k)}(x_0)))^2$, which may be quite significant.

2.1.5 *Limit Cycles*

We have seen that the logistic difference equation given in (2.14) can have very different sample paths, from a constant to total chaotic behavior depending on the value of its parameter α. The region $\alpha \in [3.0, 3.7)$ is interesting, as the sample paths oscillate among a finite number of states. This type of limiting behavior is quite common in deterministic and stochastic dynamic systems, particularly involving population dynamics and is called limit cycle. Typically, dynamics of a population depends on many internal and external factors, the size of the population being one of these factors. As the population increases in size over passing a critical threshold, typically this has a negative influence on the reproductive and survival capacities of the population, lowering its growth rate. Moreover, as the population size goes down, these capacities tend to increase, increasing its growth rate. Hence, under equilibrium conditions the sample paths of a population size will show limit cycles, switching at random epochs. For example, consider the deterministic difference equation

$$N_{t+1} = N_t \frac{b}{(1 + aN_t)^c},$$

often used for modeling annual plant population. Here a, b and c are parameters of the model. The parameter a does not affect the dynamics of the model, whereas the parameter c has a very strong effect on the dynamics. Again, this deterministic difference equation will have very different sample path properties, depending basically on the values of the parameters b and c. For example, when $c = 1$, for

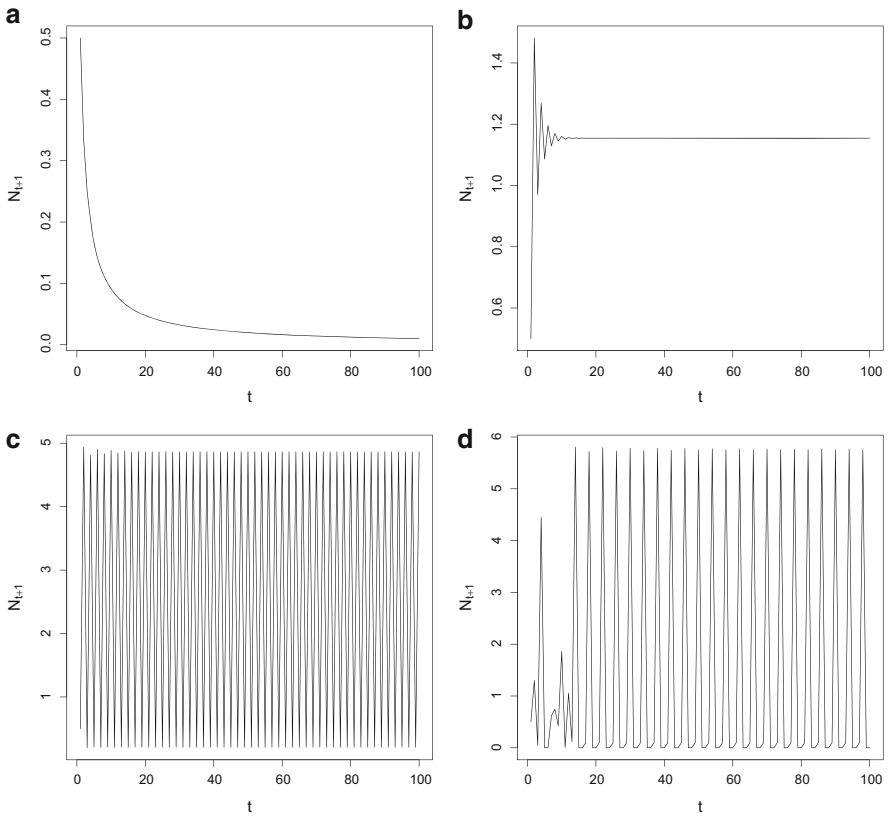

Fig. 2.2 Limit cycles of simulated samples of 100 observations (**a**) $a = 1, b = 1, c = 1$; (**b**) $a = 1, b = 10, c = 3$; (**c**) $a = 1, b = 50, c = 4$; and (**d**) $a = 1, b = 150, c = 10$

any value of b, the sample paths will converge monotonically to a constant, whereas when $c > 2$ and $b = 10$, the sample paths will show damped oscillations, finally converging to a constant. This sample path behavior then starts getting ever more erratic as b and c increase. For values of $b = 50$ and $c > 3.5$, the sample paths oscillate between fixed number of population sizes, and this behavior is called the stable limit cycles. Ultimately, for $b > 100$ and $c > 5$, the sample paths behave in a chaotic way. The limit cycles generated by several samples of size $n = 500$ are presented in Fig. 2.2. In random dynamic systems, stable limit cycle behavior can manifest itself in many different ways. For example, rather than switching between fixed number of values, the process can switch between different linear models at random epochs, depending on internal or external factors, resulting in many different piecewise linear models such as threshold models. These models will be discussed in Sect. 2.2.1.

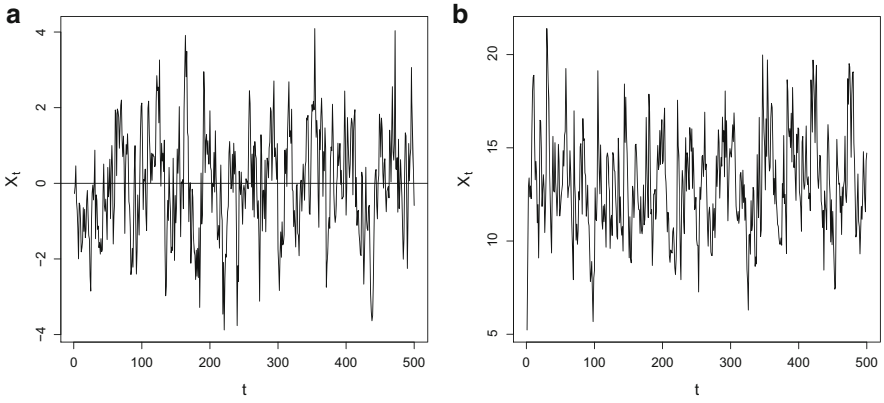

Fig. 2.3 Simulated Gaussian AR(1) model, $n = 500$ and parameter 0.7 (**a**). In (**b**) the same model with Gamma(4,1) residuals

2.1.6 Time Reversibility

A process X_t is time reversible if

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_n}, X_{t_{n-1}}, \ldots, X_{t_1}),$$

for every n and $t_1, \ldots, t_n$. Gaussian processes are time reversible, and except for few special cases, non-Gaussian processes are time irreversible. In general, if a stationary time series is stationary and time reversible then for every k, kth order cumulants satisfy

$$C(-u_1, -u_2, \ldots, -u_k) = C(u_1, u_2, \ldots, u_k).$$

This a very strong condition and it is very unlikely that there will be many non-Gaussian processes that satisfy this condition. Therefore, time irreversibility must be a rule among nonlinear processes. Note that if X_t is a stationary time series and $Y_t = h(X_t)$ is a one-to-one transformation, then Y_t is time reversible if and only if X_t is time reversible. Therefore, fitting Gaussian time series models to transformed data cannot be a adequate method of dealing with nonlinearity. In other words, in most cases, we cannot get rid of nonlinearity by transformation of the data. The simplest way of checking reversibility is by plotting the data. In general, for a reversible stationary time series, the plots of $x_n, x_{n-1}, \ldots, x_1$ and $x_1, x_2 \ldots, x_n$ should look the same. Similarly, since time irreversibility is a characteristic of Gaussian data rather than linearity, a time series which is non-Gaussian should be treated as irreversible (see Fig. 2.3).

2.1.7 *Invertibility*

When studying stochastic difference equations of the general form

$$X_t = f(\{X_s, Z_s, s \leq t\}),$$

representing a dynamic system, we restrict our study to relationships (X_t, Z_t) in which X_t is measurable with respect to $(Z_s, s < t)$. These restrictions are called the conditions of stationarity. Typically in these difference equations, what is observed is the time series X_t, and the innovations are unobserved. However, almost all statistical properties of the relationship in (X_t, Z_t) are given in terms of the innovations Z_t. Therefore, in order to be able to make inference and predictions on the dynamic system, the residuals should be recovered from the observations $x_s, s < t$. The set of conditions which guarantee this possibility are called the conditions of invertibility, under which the innovations Z_t are measurable with respect to $(X_s, s < t)$. Note that, conditions of invertibility and stationarity are joint properties of the processes (X_t, Z_t), rather than being a property of X_t alone. Unfortunately these conditions, particularly conditions of invertibility, are not so easy to obtain for general nonlinear difference equations, except for some special cases. We will look at these conditions for specific cases, whenever possible.

2.2 A Selection of Nonlinear Time Series Models

According to Tjøstheim (1994), nonlinear models can be broadly classified into the following categories;

1. Parametric models
   - Parametric models for the conditional mean
   - Parametric models for the conditional variance
   - Mixed parametric models for the conditional mean and variance
   - Generalized state space models
2. Semiparametric and nonparametric models

The above classification is by no means exhaustive and mutually exclusive. In its most general form, a nonlinear model can be written as a stochastic difference equation

$$X_t = f(X_{t-1}, \ldots, X_{t-p}, Z_t, Z_{t-1}, \ldots, Z_{t-q}, \boldsymbol{\theta}), \tag{2.19}$$

for some integers p and q, model parameters $\boldsymbol{\theta}$ and some measurable function f which renders a stationary causal solution. Such general representation contains

both nonlinear conditional mean and conditional variance components for X_t in terms of the past values. Often, it is easier to look at a simpler class of models

$$\begin{aligned} X_t &= f(X_{t-1}, \ldots, X_{t-p_1}, Z_{t-1}, \ldots, Z_{t-q_1}, \boldsymbol{\theta}_1) \\ &\quad + g(X_{t-1}, \ldots, X_{t-p_2}, Z_{t-1}, \ldots, Z_{t-q_2}, \boldsymbol{\theta}_2) Z_t, \end{aligned} \tag{2.20}$$

for measurable functions f and g, separating the nonlinear models for the conditional mean and the conditional variance components. Taking g as a constant, for various combinations of f functions, we get subclasses of nonlinear models for the conditional mean, whereas taking f constant and for various combinations of g, we get subclasses of models for the conditional variance. The general class (2.19) can be classified as the class of mixed models, although Tjøstheim (1994) classifies (2.20) as the class of mixed models.

In this chapter, we give a brief description of some of these models.

2.2.1 *Parametric Models for the Conditional Mean*

These models represent the conditional mean function of the process X_t as a nonlinear function of the past observations, keeping the conditional variance constant. Hence, an appropriate general model is given as

$$X_t = f(\mathcal{F}_{t-1}, \boldsymbol{\theta}) + Z_t.$$

Here, the function f has a known parametric form, $\mathcal{F}_{t-1}$ is the sigma-field generated by X_t up to time $t-1$, Z_t is an i.i.d. sequence and $\boldsymbol{\theta}$ is an unknown parameter vector to be estimated. In some cases, the function f may also depend on other external processes. Several different forms of f give different classes of nonlinear models. One important subclass is the regime models or regime switching models. Models in this class are typically made up of several piecewise linear processes and the generating process switches from one linear model to another, depending on the value of an indicator. This indicator may be a random variable, such as the delayed value of the series itself, or it can be the value of a different, possibly latent process. Depending on the parameter values, such piecewise linear regime models are stationary but nonlinear, in the sense that they cannot be represented in the form (2.1). This class of models include threshold models, first introduced by Tong (1990), and later enriched by other classes of similar nature. The fundamental reason for introducing such classes of models is the need to model random cyclic behavior that exists in many time series; see Sect. 2.1.5 for further details. As we will see, the class of bilinear processes, which is by far the most general class of nonlinear models, in the sense that they form a dense subset of the Volterra expansions, cannot generate limit cycles (e.g., Tong 1990) and therefore the threshold models have gained importance on their own right in modeling time series. For general treatment

of regime models see Hamilton (2008), Granger and Teräsvirta (1993), and Franses and Van Dijk (2000). Regime models may also switch at deterministic but unknown times, in which case the process will be linear but not stationary. Such models are called segmented time series (e.g., Davis et al. 2008). We now look at some of the regime models.

Threshold Autoregressive (TAR) and Self-Exciting Threshold (SETAR) Models

The basic idea behind this class is as follows: we start with a linear model for X_t and allow the parameters of the model to vary according to the values of a finite number of past values of X_t, or a finite number of past values of an associated series Y_t. Hence, such regime models in general can be written as

$$X_t = \begin{cases} a_0^{(1)} + \sum_{i=1}^{p_1} a_i^{(1)} X_{t-i} + Z_t, \text{ if } Y_t \le r \\ \\ a_0^{(2)} + \sum_{i=1}^{p_2} a_i^{(2)} X_{t-i} + Z_t, \text{ if } Y_t > r \end{cases}, \tag{2.21}$$

where r is the threshold and Y_t is a switching process which can be a latent or an observable process, determining which regime describes the process in a certain moment of time. Such processes are called Threshold autoregressive (TAR) models. When the switching process is the time series itself observed at a certain lag, we have the SETAR sub-class. In its simplest form a first-order SETAR is given as

$$X_t = \begin{cases} a_1 X_{t-1} + Z_t, & \text{if } X_{t-1} \in A^{(1)} \\ a_2 X_{t-1} + Z_t, & \text{if } X_{t-1} \in A^{(2)} \end{cases},$$

where $A^{(i)}$ are some regions. Typically, these regions are intervals such as $A^{(1)} = \{X_{t-1} \le r\}$, and $A^{(2)} = \{X_{t-1} > r\}$, for some threshold r. We can generalize this class of models to

$$X_t = \sum_{j=1}^{p} a_{ij} X_{t-j} + Z_t^{(i)}, \ \ (X_{t-1}, \ldots, X_{t-p}) \in A^{(i)}, \ \ i = 1, 2, \ldots, l. \tag{2.22}$$

having different error structures in each segment. Note that when $l = 1$, the first-order threshold model can be seen as a piecewise linear approximation to the general nonlinear first order model

$$X_t = f(X_{t-1}) + Z_t,$$

whereas the pth order model in (2.22) is a linear piecewise approximation to the general nonlinear equation

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-p}) + Z_t.$$

In practice it is not feasible to fit a model of the form (2.22) with a large p, since the identification of the threshold regions would involve search in a p-dimensional space. A sub-class of the form

$$X_t = a_0^{(i)} + \sum_{j=1}^{p} a_{ij} X_{t-j} + Z_t^{(i)}, \quad X_{t-d} \in A^{(i)},$$

where $A^{(i)}$ is in $\mathbb{R}$ can be considered, thus simplifying the identification of such models. These models can still be extended to include cases when switching between sets of parameters is determined by the past values of a different process Y_t, extending the TAR model given in (2.21)

$$X_t = a_{0i} + \sum_{j=1}^{m_i} a_{ij} X_{t-j} + \sum_{j=1}^{l_i} b_{ij} Y_{t-j} + Z_t^{(i)}, \quad Y_{t-d} \in A^{(j)}.$$

Such models are known to be very useful, particularly in modeling data which shows random cyclic movements.

Smooth Threshold Autoregressive (STAR) Models

As mentioned above, TAR/SETAR models should be used when the process to be modeled shifts from one regime to another abruptly. However, if the transition is gradual, then the STAR models are more appropriate. A two-regime STAR(p) model is defined by Chan and Tong (1986) as follows:

$$X_t = c_0 + \sum_{i=1}^{p} a_{0,i} X_{t-i} + G\left(\frac{X_{t-d} - a}{b}\right)\left(c_1 + \sum_{i=1}^{p} a_{1,i} X_{t-i}\right) + Z_t,$$

where d is the delay parameter, a and b represent the location and scale parameters of G, respectively. The transition function G, that enables the transition between one regime to the other, is a smooth, continuous and monotonically increasing function, satisfying the inequality $0 < G(z) < 1$. Two subclasses of STAR models are the exponential and logistic STAR models, when the function G respectively is given by the expressions

$$G(z) = 1 - \exp[-a(z-b)^2], \ a > 0,$$

$$G(z) = \frac{1}{1 + \exp[-a(z-b)]}, \ a > 0.$$

Chan and Tong (1986) give an alternative STAR model with Gaussian smooth transition function

$$G(z) = \Phi\big[a(z-b)\big],$$

where $\Phi(\cdot)$ the cdf of the standard Normal distribution. The parameter b can be regarded as the threshold and a controls how fast and how abrupt the model shifts from one regime to another (see e.g., Zivot and Wang 2006).

Markov Switching AutoRegressive (MAR) Models

This class of models was developed by Hamilton (1989), based on ideas previously proposed by Goldfeld and Quandt (1973). Let S_t be a discrete first-order homogeneous Markov chain with state space $S = \{0, 1, \ldots, k\}$. Each member of S corresponds to a regime. Let $P(S_t = j \mid S_{t-1} = i) = p_{ij}$ be the transition matrix given by

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}.$$

Each state, at time t, has an associated probability given by $\pi_t := (P_1, P_2, \ldots, P_k)$, where $\pi = P'\pi$. A k-regime MAR model is given as

$$X_t = \mu_{S_t} + \mathbf{X}_{t-1}\boldsymbol{\theta}_{S_t} + Z_t,$$

where $\mathbf{X}_{t-1} := (X_{t-1}, X_{t-2}, \ldots, X_{t-p})$, and μ_{S_t}, $\boldsymbol{\theta}_{S_t}$ are the model parameters that switch between k different values according to the latent Markov chain. Z_t is assumed to be a Gaussian sequence with mean zero and the variance can be taken as constant, or may switch between k different values depending on the realization of S_t. A classical application of a two-state MAR model to the US GNP time series is given in Hamilton (1989).

Random Coefficient Models

Sometimes it may be useful to introduce random regime switch into the model parameters, giving rise to a different class of models. A simple model within this class is the first order AR model

$$X_t = \psi_t X_{t-1} + Z_t,$$

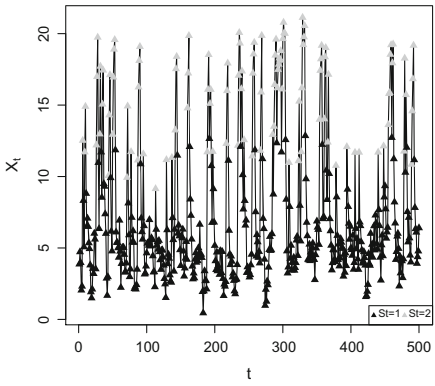

Fig. 2.4 Sample path of size $n = 500$ of the model (2.23)

where ψ_t is a homogeneous Markov chain with a finite space and transition probabilities p_{ij}, for example taking values a_1 and a_2. In this case, the process X_t will alternate between the two processes

$$X_t = a_1 X_{t-1} + Z_t,$$

and

$$X_t = a_2 X_{t-1} + Z_t,$$

according to the transition probabilities (p_{ij}) for $i, j = 1, 2$. Smoother changes in the parameter can be modeled by state space type model

$$X_t = \psi_t X_{t-1} + Z_t,$$

$$\psi_t = a\psi_{t-1} + \nu_t,$$

where Z_t and ν_t are independent i.i.d. sequences. In general, a regime model will take the form (X_t, S_t), where, (S_t) is a latent process, typically a homogeneous Markov chain with a finite state space, such that at any time t, X_t conditional on $S_t = j$ follows a linear model ARMA(p_j, q_j). Hence the process will alternate among various linear models in accordance with the transient behavior of the unobserved process S_t. The estimation, identification and diagnostics for these models are complicated although not impossible, due to the fact that the process S_t is not observed. Note that if the residual process is made to depend on the unobserved Markov chain then the variance of the process also changes from one regime to another. In Fig. 2.4, we have a sample path of the process

$$X_t = C^{(S_t)} + 0.5X_{t-1} + Z_t, \tag{2.23}$$

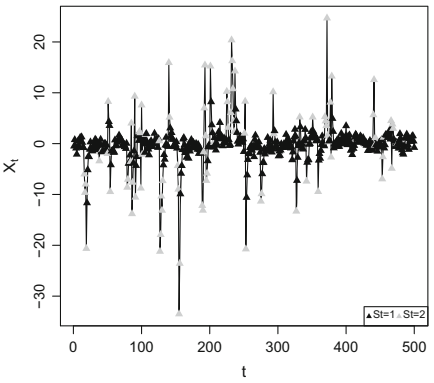

Fig. 2.5 Sample path of size $n = 500$ of the model (2.24)

where Z_t are i.i.d. N(0,1) and

$$C^{(S_t)} = \begin{cases} 2 & \text{if } S_t = 1 \\ 10 & \text{if } S_t = 2 \end{cases},$$

where S_t is a homogeneous Markov chain with transition matrix

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}.$$

For this matrix, the stationarity limit distribution is given by $\pi = (0.8, 0.2)$. A sample path of the extended process

$$X_t = 0.5X_{t-1} + Z_t^{(S_t)}, \tag{2.24}$$

where

$$Z_t^{(S_t)} = \begin{cases} N(0,1) & \text{if } S_t = 1 \\ N(0,10) & \text{if } S_t = 2 \end{cases},$$

is presented in Fig. 2.5.

Segmented Time Series

Davis et al. (2008) introduced a broad class of nonlinear and non-stationary time series segmented into several pieces. Each segment is assumed to be a stationary time series modeled by a parametric class of time series, whereas the number and the locations of the break points or the segments are treated as unknown model parameters. Thus the observed time series y_t is assumed to be generated by a time series Y_t of the form

$$Y_t = X_{t,j},\ \tau_{j-1} \le t < \tau_j,\ j = 1, \ldots, m,$$

where $\tau_j, j = 1, \ldots, m$ are the break-points or segments of unknown number m and each segmented $X_{t,j}$ are stationary times series independent of each other. Typically, these time series can be AR(p_j) or GARCH(p_j, q_j) models. In a more general form, each segment may be composed of time series having different state space representations. The model choice, namely the identification of the number of segments and their respective locations, as well as the order of the models in each segment is then performed by using a genetic algorithm. Simulation results indicate that these models perform well and the availability of software to fit these models makes this class of models a good candidate for regime models. The main difference between segmented time series and threshold models is that, whereas in threshold models, the transition between models is triggered by lagged values of the time series, in segmented time series, the changes occur at specified time points. Examples of time series models studied by Davis et al. (2008) include:

- *Segmented AR process:*
  In this case, each segment $X_{t,j}$ is assume to be an AR(p_j) process given by

  $$X_{t,j} = a_{1j} X_{t-1,j} + \cdots + a_{p_j,j} X_{t-p_j,j} + Z_{t,j},$$

  where $Z_{t,j} \sim \text{WN}(0, \sigma_j^2)$. Here the (unknown) parameters of the model are $\boldsymbol{\theta}_j := (p_j, \boldsymbol{a}_j, \sigma_j^2)$.[1]
- *Segmented GARCH(p_j, q_j) process:*
  In this case, each segment Y_j is modeled by a GARCH(p_j, q_j) given by

  $$X_{t,j} = \sigma_{t,j} Z_t,$$

  where $Z_t \sim \text{WN}(0, 1)$ and

  $$\sigma_{t,j}^2 = a_{j,0} + a_{j,1} X_{t-1,j}^2 + \cdots + a_{j,p_j} X_{t-p_j,j}^2 + b_{j,1} \sigma_{t-1,j}^2 + \cdots + b_{j,q_j} \sigma_{t-q_j,j}^2,$$
  $$\tau_{j-1} \le t < \tau_j,$$

  where to satisfy stationarity of each segment, the parameters are restricted by $a_{0,j} > 0$, $b_{0,j} \ge 0$ and

  $$\sum_{i=1}^{p_j} a_{i,j} + \sum_{i=1}^{q_j} b_{i,j} < 1.$$

  Here, the model parameters are given by $\boldsymbol{\theta}_j := (p_j, q_j, \boldsymbol{a}_j, \boldsymbol{b}_j)$.

[1] A discrete counterpart of conventional segmented AR processes, based on the thinning operator in (1.5), was proposed by Kashikar et al. (2013).

- *Segmented state space models:*
  In this case the jth segment $Y_{t,j}$ has a state space representation given by the equations

$$p_j(y_t|\mathbf{x}_{j,t}) = p_j(y_t|x_{t,j}, \mathbf{x}_{t-1,j}, \mathbf{y}_{t-1}), \tau_{j-1} \leq t \leq \tau_j$$

  and the state process $X_{t,j}$ follows a AR(p_j) process.

2.2.2 Exponential Autoregressive Models

Consider for example, the second-order autoregressive model

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + Z_t,$$

where a_1, a_2 instead of being constants, are functions of X_{t-1}. Specifically, assume that they are exponential functions of X_{t-1}^2 taking the form

$$a_1 = \phi_1 + \pi_1 \exp(-\gamma X_{t-1}^2),$$

$$a_2 = \phi_2 + \pi_2 \exp(-\gamma X_{t-1}^2).$$

Such a model then is called second-order EAR(2) model. Note that for large $|X_{t-1}|$, $a_1 \sim \phi_1, a_2 \sim \phi_2$, whereas for small $|X_{t-1}|$, $a_1 \sim \phi_1 + \pi_1, a_2 \sim \phi_2 + \pi_2$, so that the EAR model behaves like the threshold AR model where the coefficients change smoothly between two extreme parameter values. EAR models are capable of producing amplitude dependent frequency effect, limit cycles and jump phenomena; see Tong (1990) or Priestley (1981). The coefficients a_1, a_2 can be defined as a function of X_{t-1} in different ways to assure smooth transitions. For example, in the case of EAR(1) model, a_1 can be parameterized as

$$a_1 = \theta_1 X_{t-1} + \theta_2 X_{t-1}\{[1 + \exp(\theta_3(X_{t-1} - \theta_4))]^{-1} - 1/2\}, \quad \theta_3 > 0,$$

in which case the model is called logistic exponential model. These models can be generalized to the form

$$X_t := \boldsymbol{a}' \boldsymbol{X}_{t-1} + \sum_{i=1}^{k} b_i \phi_i(\boldsymbol{\gamma}_i' \boldsymbol{X}_{t-1}) + Z_t,$$

where $\boldsymbol{X}_{t-1} := (X_{t-1}, X_{t-2}, \ldots, X_{t-p})$, and $\boldsymbol{a}' := (a_1, \ldots, a_p)$, $\boldsymbol{\gamma}' := (\gamma_1, \ldots, \gamma_p)$ are p-dimensional parameter vectors and $\phi_i(\cdot)$ are known specific functions. Although such models are used, it is evident that there would be problems of estimation as the parameter space increases.

2.2.3 Polynomial-Type Models

One can also use nonlinear regression type models based on polynomials of the form

$$X_t = \sum_{i=1}^{k} a_i X_{t-1}^i + Z_t.$$

More general polynomial models can be devised by introducing terms depending on $X_{t-2}, X_{t-3}, \ldots, X_{t-p}$ and cross terms. These models are not very much used due to the feedback of X_t into itself, causing explosive behavior.

2.2.4 Bilinear Models

The process X_t is said to be a bilinear process $\mathrm{BL}(p, q, m, l)$ if it satisfies the difference equation

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{l} b_{ij} X_{t-i} Z_{t-j} + Z_t. \tag{2.25}$$

The conditional mean of the process (2.25) is given by

$$E(X_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{p} \phi_j x_{t-j} + \sum_{j=1}^{q} \theta_j z_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{l} b_{ij} x_{t-i} z_{t-j},$$

whereas the conditional variance is given by $V(X_t|\mathcal{F}_{t-1}) = \sigma_Z^2$. Hence the bilinear model given in (2.25) represents the nonlinear dynamics present in the mean. This class obviously can be extended to include cross terms of $(X_{t-1}, \ldots, X_{t-m})$ with Z_t resulting in models

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{m} \sum_{j=0}^{l} b_{ij} X_{t-i} Z_{t-j} + Z_t. \tag{2.26}$$

In this case, $V(X_t|\mathcal{F}_{t-1})$ will also be a function of passed values of the series, therefore bilinear models described in (2.26) fall in the class of mixed models for the conditional mean and variance.

The class of bilinear models plays an important role in modeling nonlinearity for various reasons. The class is an obvious generalization of $\mathrm{ARMA}(p, r)$ models resulting in nonlinear conditional mean. Under fairly general conditions, bilinear processes approximate finite order Volterra series expansions to any desired order

of accuracy over finite time intervals (Brockett 1976). Due to Nisio's theorem, Volterra series expansion are a dense class within the class of nonlinear time series, therefore, under fairly general conditions, bilinear processes are also a dense class within nonlinear processes, approximating any nonlinear process to a desired level of accuracy. However, it is well known that bilinear processes cannot capture random cyclic movements, such as limit cycles and jump phenomena. The class is fairly well-studied, and conditions for the existence of unique and stationary solutions are known. Although identification, estimation and diagnostic techniques are available, much of the work on the class remains to be completed. Volterra series expansions and bilinear processes are often used in the control theory and are somewhat different from the context within which they are used in time series. In the control theory, the output X_t, as well as the input process Z_t are observable, making the probabilistic structure simple. For example conditional on the passed values of Z_t, the process X_t is linear, and conditional on the passed values of X_t, the process Z_t is also linear. In the time series context, the input random process Z_t is not observed and unfortunately, the lack of verifiable conditions for invertibility (except for very simple bilinear processes) limits the use of these processes as models. Bilinear processes are capable of producing sudden bursts of large values and hence are suitable for modeling time series showing heavy tailed phenomena.

The bilinear process BL(p, q, m, l) given in (2.26) can be written in the form (Resnick and Van den Berg 2000)

$$\mathbf{X}_t = \mathbf{A}_{t-1}\mathbf{X}_{t-1} + \mathbf{B}_t,$$

where

$$\mathbf{B}_t = \mathbf{\Theta}\mathbf{Z}_t,$$

$\mathbf{\Theta}$ is a $p \times (1 + q)$ matrix given by

$$\mathbf{\Theta} = \begin{pmatrix} 1 & \theta_1 & \dots & \theta_{q-1} & \theta_q \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \dots & \ddots & \dots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix},$$

$\mathbf{A}_{t-1}$ is a $p \times p$ matrix given by

$$\mathbf{A}_{t-1} = \begin{pmatrix} \phi_1 + \sum_{j=1}^{l} b_{1j} Z_{t-j} & \phi_2 + \sum_{j=1}^{l} b_{2j} Z_{t-j} & \dots \dots & \phi_p + \sum_{j=1}^{l} b_{pj} Z_{t-j} \\ 1 & 0 & \dots\ 0 & 0 \\ \vdots & \vdots & \ddots\ \vdots & \vdots \\ 0 & 0 & \dots\ 1 & 0 \end{pmatrix}$$

and $\mathbf{X}$ and $\mathbf{Z}_t$ are respectively $(p-1)\times 1$ and $q\times 1$ column vectors

$$\mathbf{X}_t := (X_t, X_{t-1}, \ldots, X_{t-p+1})',$$

$$\mathbf{Z}_t := (Z_t, Z_{t-1}, \ldots, Z_{t-q})'.$$

The general bilinear model (2.25) can also be written in an equivalent state space form with the observation equation

$$X_t = \mathbf{H}'\boldsymbol{W}_{t-1} + Z_t,$$

and the state equation

$$\boldsymbol{W}_t = \mathbf{A}_t\boldsymbol{W}_{t-1} + \mathbf{C}_t.$$

Here, the state vector $\boldsymbol{W}_t$ is a Markov chain, and $\mathbf{A}_t$, $\mathbf{C}_t$ are random matrices, depending on the specific form of the general bilinear process given in (2.25). The general form is quite complicated (e.g., Fan and Yao 2003) but simpler bilinear models can conveniently be written in this form. For example the model $\mathrm{BL}(p, 0, p, 1)$ given by

$$X_t = \sum_{j=1}^{p}\phi_j X_{t-j} + \sum_{i=1}^{p} b_{i1}X_{t-i}Z_{t-1} + Z_t,$$

can be written in the vector state space form (e.g., Priestley 1981)

$$\mathbf{X}_t = \mathbf{H}\boldsymbol{W}_t + \mathbf{C}Z_t,$$

$$\boldsymbol{W}_t = (\mathbf{A} + \mathbf{B}W_t)\boldsymbol{W}_{t-1} + (\mathbf{A} + \mathbf{B}Z_t)\mathbf{C}Z_t,$$

where

$$\boldsymbol{W}_t := (X_t, X_{t-1}, \ldots, X_{t-p})',$$

$$\mathbf{H}_{1\times p} := (1, 0, \ldots, 0),$$

$$\mathbf{C}_{p\times 1} := (1, 0, \ldots, 0)',$$

$$\mathbf{A}_{p\times p} := \begin{pmatrix} \phi_1 & \phi_2 & \cdots & a_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$\mathbf{B}_{p\times p} := \begin{pmatrix} b_{11} & b_{21} & \dots & b_{p1} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Note that $\boldsymbol{W}_{t-1}$ is independent of the coefficient $(\mathbf{A} + \mathbf{B}Z_t)$ and the error $(\mathbf{A} + \mathbf{B}Z_t)\mathbf{C}Z_t$, and is a Markov chain. Note also that the pair $(\mathbf{A}+\mathbf{B}Z_t, (\mathbf{A}+\mathbf{B}Z_t)\mathbf{C}Z_t)$, forms an i.i.d. sequence of random matrices, but the components of this pair are not independent of each other. This state space representation, due to its Markovian nature facilitates the study of the probabilistic properties of the process. For example, if the process $\boldsymbol{W}_t$ is stationary, then so is X_t. Due to its Markovian structure, it is relatively easy to study the conditions under which $\boldsymbol{W}_t$ is stationary; see Meyn and Tweedie (2009) for the study of probabilistic properties of Markov processes.

If we solve the difference equation given by

$$\boldsymbol{W}_t = \mathbf{A}_t \boldsymbol{W}_{t-1} + \mathbf{C}_t,$$

where $(\mathbf{A}_t, \mathbf{C}_t)$ is an i.i.d. sequence of random matrices, iteratively n times, the partial solution for $\boldsymbol{W}_t$ is given by

$$\boldsymbol{W}_t = \prod_{i=0}^{n-1} \mathbf{A}_{t-i} \boldsymbol{W}_{t-n} + \sum_{j=1}^{n-1} \prod_{i=1}^{j-1} \mathbf{A}_{t-i}\mathbf{C}_{t-j},$$

so that the convergence in probability

$$\sum_{j=1}^{n-1} \prod_{i=1}^{j-1} \mathbf{A}_{t-i}\mathbf{C}_{t-j} \to 0,$$

is a sufficient condition for the existence of a stationary solution. For example, consider the simple bilinear process

$$X_t = aX_{t-1} + bX_{t-1}Z_{t-1} + Z_t.$$

Solving iteratively for X_t, upon n iterations we get

$$\begin{aligned} X_t = &\prod_{i=1}^{n}(a + bZ_{t-i})X_{t-n} \\ &+ \sum_{j=1}^{n-1} \prod_{i=1}^{j}(a + bZ_{t-i})Z_{t-j}, \end{aligned}$$

and, if in probability

$$\prod_{j=1}^{n}(a + bZ_{t-j}) \to 0, \tag{2.27}$$

then

$$X_t = \sum_{j=1}^{\infty}\prod_{i=1}^{j}(a + bZ_{t-i})Z_{t-j}. \tag{2.28}$$

A sufficient condition for (2.27) is given by Pham and Tran (1981). If Z_t are i.i.d. zero-mean r.v's with $E(Z_t^2) = \sigma^2$ and $E(Z_t^4) < \infty$, then (2.27) converges in mean-square if $a^2 + b^2\sigma^2 < 1$, in which case, (2.28) is the unique stationary solution. This is also a sufficient condition for invertibility. However, it is far from being a necessary condition for stationarity and invertibility. Note that (2.28) is a moving average representation

$$X_t = \sum_{j=1}^{\infty}\theta_j Z_{t-j},$$

with random coefficients

$$\theta_j = \prod_{i=1}^{j}(a + bZ_{t-i}).$$

Therefore, one would expect that the second-order properties of this process resembles that of a linear process. Indeed, assuming $\sigma^2 = 1$ simple calculations show that

$$\begin{aligned}\mu = E(X_t) &= \sum_{j=1}^{\infty} a^{j-1}b\sigma^2 \\ &= \frac{b}{1-a},\end{aligned}$$

$$E(X_t^2) = \frac{1+2b^2}{1-b^2},$$

$$\gamma(1) = E(X_t X_{t-1}) = 2b^2,$$

and for $k \geq 2$,

$$\gamma(k) = a\gamma(k-1).$$

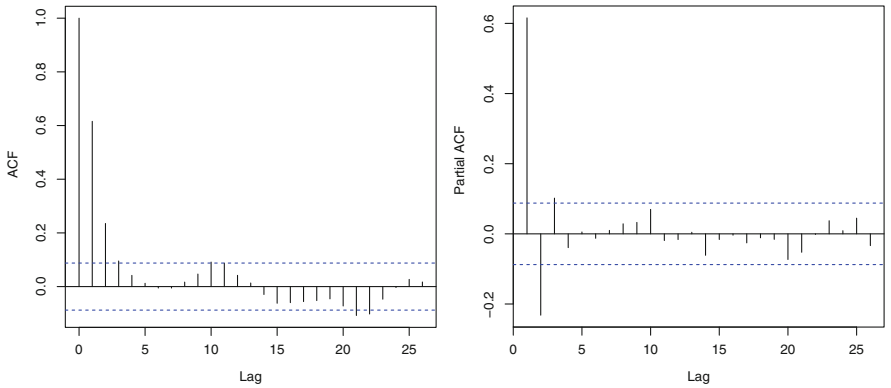

Fig. 2.6 ACF and PACF of the bilinear model (2.29)

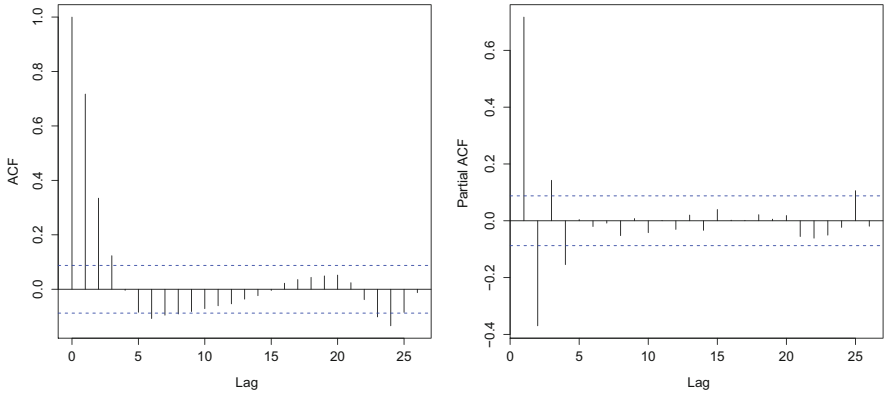

Fig. 2.7 ACF and PACF of the linear model (2.30)

Note that this is exactly the covariance structure of a linear MA(1, 1) process. Similarly, the autocovariance function of a BL(p, q, m, l) process behaves like the autocovariance function of the process MA(p, q_0) where $q_0 := \max(q, l)$; see Fan and Yao (2003) for details. It is clear once again that one cannot differentiate a nonlinear model from a linear model by studying only the second-order properties. In Figs. 2.6 and 2.7 below, we give respectively the autocorrelation and partial autocorrelation functions based on a data of dimension 500, simulated from the models

$$X_t = 0.5X_{t-1} + 0.6X_{t-1}Z_{t-1} + Z_t, \tag{2.29}$$

and

$$X_t = 0.5X_{t-1} + 0.6Z_{t-1} + Z_t \tag{2.30}$$

with i.i.d. standard Normal innovations.

Unfortunately, like for most nonlinear processes, the condition of invertibility which is crucial for estimation and prediction, is not well understood and cannot be checked except for some simple bilinear processes (see Chap. 4, for some empirical ways of checking invertibility). Therefore, although bilinear processes have desired properties as models, their use in practice is quite restricted.

Here we note a fundamental difference between the bilinear and threshold models. Threshold models, as in the case of bilinear models, can be put in the state space representation

$$X_t = \mathbf{H}\boldsymbol{W}_t,$$

$$\boldsymbol{W}_t = \mathbf{A}^{(i)}\boldsymbol{W}_{t-1} + \mathbf{B}^{(i)}Z_t^{(i)},$$

for some properly chosen state vector $\boldsymbol{W}_t$, and constant matrices $\mathbf{A}^{(i)}$,$\mathbf{B}^{(i)}$ and regions $\mathbb{R}^{(i)}$. However, the essential difference between threshold models and bilinear models is that whereas in bilinear processes the nonlinearity is introduced by the cross terms $Z_{t-i}X_{t-j}$, in the threshold models the relation between $\boldsymbol{W}_t$ and $\boldsymbol{W}_{t-1}$ is nonlinear (that is a nonlinear function of $X_s, s < t$), with residuals still entering the model linearly. This difference has strong influence on the type of nonlinear behavior. Bilinear processes, due to this cross terms, in general, are capable of producing extreme observations but cannot produce limit cycle behavior, hence each class has its own use in modeling different nonlinear phenomena.

2.2.5 *Parametric Models for the Conditional Variance*

These models are special case of the representation (2.20) with $f \equiv 0$ and are based on modeling the function g in different forms. A useful conceptual division of these models can be made as

1. Observation-driven models and
2. Parameter-driven models.

Observation-Driven Models

Lets assume that for each t, the time series satisfies

$$X_t|\sigma_t^2 \sim N(0, \sigma_t^2).$$

The observation-driven models are based on representing σ_t^2 as a function of lagged values of X_t taking the general form

$$X_t = g(\mathcal{F}_{t-1}, \boldsymbol{\theta}_2) Z_t,$$

giving rise to the rich classes of ARCH and GARCH models. Since

$$V(X_t|\mathcal{F}_{t-1}) = g^2(\mathcal{F}_{t-1}, \boldsymbol{\theta}_2)\sigma_Z^2,$$

it is customary to represent the function g by σ_t.

Since the seminal paper of Engle (1982) traditional time series tools such as the ARMA models for the mean have been extended to essentially analogous models for the variance. Autoregressive conditional heteroscedasticity (ARCH) models are now widely used to describe and forecast changes in volatility of financial time series. For a survey of ARCH-type models and various extensions, see Bollerslev et al. (1992, 1994), Pagan (1996), Palm (1996), Shephard (1996), Berkes et al. (2003), Bauwens et al. (2006), Silvennoinen and Teräsvirta (2009), and Teräsvirta (2009). According to Engle (2004) the original idea was to find a model to assess the validity of the conjecture of Friedman (1977) that the unpredictability of inflation was a primary cause of business cycles. Uncertainty due to this unpredictability would affect the investor's behavior. Pursuing this idea requires a model which characterizes the time dynamics of this uncertainty.

Financial time series, such as relative returns of stock indices, share prices and foreign exchange rates, often show the following features (usually referred to as *stylized facts*):

- The sample mean of the data is close to zero whereas the sample variance is of the order 10^{-4} or smaller;
- Exceedances of high/low thresholds tend to occur in clusters. This property indicates that there exists dependence in the tails;
- Return data exhibit heavy-tailed marginal distributions;
- The sample autocovariance function of such data is statistically insignificant at all lags (with a possible exception of the first lag), whereas the sample autocovariance function of the absolute values or the squares of the time series are different from zero for a large number of lags and stay almost constant and positive for large lags.
- As one increases the time scale on which returns are calculated, their distribution looks more and more a Gaussian. This means that the *peakedness* around zero and the *heavy-tailedness* of the empirical distribution turn into bell shapedness.

The list above is far from being complete. An exhaustive analysis of stylized facts can be found in Cont (2001).

Most models for financial time series (and in particular for return data) used in practice to accommodate such features are given in the multiplicative form

$$X_t = \sigma_t Z_t, \ t \in \mathbb{Z}, \tag{2.31}$$

where (Z_t) forms an i.i.d. sequence of real-valued innovations or noise variables with zero mean and unit variance, (σ_t) is a stochastic process such that σ_t and

Z_t are independent for fixed t. In general, (σ_t) and (X_t) are assumed to be strictly stationary. Motivation for considering this particular choice of a simple multiplicative model comes from the fact that (a) in practice, the direction of price changes is well modelled by the sign of Z_t, whereas σ_t provides a good description of the order of magnitude of this change; and (b) the volatility σ_t^2 represents the conditional variance of X_t given σ_t.

Engle (1982) suggested the following simple model for the volatility σ_t:

$$\sigma_t^2 = a_0 + a_1 X_{t-1}^2, \ t \in \mathbb{Z}, \tag{2.32}$$

for positive constants a_0 and a_1. Equations (2.31) and (2.32) define an *AutoRegressive Conditionally Heteroscedastic model of order one* (in short ARCH(1)). For example, assume Z_t to be an i.i.d. Gaussian white noise distribution. Then the distribution of tomorrow's return X_{t+1}, conditionally on today's return X_t, has Normal distribution with zero mean and variance $a_0 + a_1 X_t^2$. This allows one to give a *distributional forecast* of X_{t+1} given X_t. The ARCH(1) fit to real-life data can be improved by introducing the ARCH(p) model, with $p \in \mathbb{N}$, where σ_t obeys the recursive equation

$$\sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i X_{t-i}^2, \ t \in \mathbb{Z}, \tag{2.33}$$

with $a_0 > 0, a_1, \ldots, a_{p-1} \geq 0$ and $a_p > 0$. A major improvement upon the expression in (2.33) was achieved by Bollerslev (1986) and Taylor (1986), independently of each other, who introduced the Generalized ARCH (GARCH) models of order p and q. In this model, the conditional variance is also a linear function of its own lags and takes the form

$$\begin{aligned} \sigma_t^2 &= a_0 + \sum_{i=1}^{p} a_i X_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2, \\ &:= a_0 + a(B) X_t^2 + b(B) \sigma_t^2, \ t \in \mathbb{Z}, \end{aligned} \tag{2.34}$$

with $a_0 > 0, a_1, \ldots, a_{p-1} \geq 0, a_p > 0, b_1, \ldots, b_{q-1} \geq 0$ and $b_q > 0$. The requirement that all the coefficients are non-negative ensures that σ_t^2 is also non-negative. The most popular GARCH model in applications has been the GARCH(1, 1) model, with $p = q = 1$ in (2.34).

The family of GARCH models has been generalized and extended in various directions in order to accommodate different features often exhibited by financial time series. One possible generalization of the GARCH models is the so-called ARCH(∞) sequences defined as follows:

Definition 2.2.1. A random sequence (Y_t) is said to satisfy ARCH(∞) equations if there exists a sequence of i.i.d. non-negative r.v's (η_t) such that

$$Y_t = \zeta_t \eta_t, \ t \in \mathbb{Z}, \tag{2.35}$$

and

$$\zeta_t = a_0^* + \sum_{i=1}^{\infty} a_i^* Y_{t-i}, \tag{2.36}$$

with $a_i^* \geq 0$ for $i = 0, 1 \ldots$.

The general framework leading to the model in (2.35) and (2.36) traces back to Robinson (1991). This class of models include, among others, the classical squared ARCH(∞) model, that is the model in (2.31) and (2.33) with $p = \infty$ and $Y_t = X_t^2$, $\zeta_t = \sigma_t^2$, $\eta_t = Z_t^2$ and the coefficients $a_0^* = a_0$ and $a_i^* = a_i$ for $i = 1, \ldots, p$; or the squared GARCH(1, 1) with $Y_t = X_t^2$, $\zeta_t = \sigma_t^2$, $\eta_t = Z_t^2$, $a_0^* = a_0/(1-b_1)$, and $a_i^* = b_1^{i-1} a_1$.

On the other hand, several extensions of the GARCH models aim at accommodating asymmetric response of the volatility for positive and negative shocks. Giving heed to this problem, Ding et al. (1993) proposed the Asymmetric Power ARCH of order (p, q), in short APARCH(p, q), model defined as

$$\sigma_t^\delta = \omega + \sum_{i=1}^{p} a_i (|X_{t-i}| - \gamma_i X_{t-i})^\delta + \sum_{j=1}^{q} b_j \sigma_{t-j}^\delta,$$

where $\omega > 0$, $a_i \geq 0$, $b_j \geq 0$, $\delta \geq 0$ represents the parameter for the power term, and $-1 < \gamma_i < 1$ is the leverage parameter. This model allows detecting asymmetric responses of the volatility for positive or negative shocks. If $\gamma_i > 0$, negative shocks have stronger impact on volatility than positive shocks, as would be expected in the analysis of financial time series. If $\gamma_i < 0$, the reverse happens. The APARCH model includes as special cases the GARCH(p, q) model, the Taylor/Schwert GARCH in standard deviation model (Schwert 1989, 1990; Taylor 1986), the GJR-GARCH model (Glosten et al. 1993), the TARCH model (Rabemananjara and Zakoïan 1993; Zakoïan 1994), the NARCH models (Higgins and Bera 1992) and the log-ARCH model (Geweke 1986; Pantula 1986).

Moreover, evidence of long memory and persistence (accordingly to the most common definition of long memory: autocovariance function, $\gamma(k)$, decaying at the hypergeometric rate k^{2d-1}, with $0 < d < 0.5$) has been documented in many fields in economics, including volatility of financial series and trading intensity in financial durations data. Baillie et al. (1996) proposed the Fractionally IGARCH(p, d, q), or FIGARCH(p, d, q), in order to accommodate long memory in volatility. The authors started by writing the GARCH(p, q) process as an ARMA(m, p) in X_t^2

$$(1 - a(B) - b(B))X_t^2 = \omega + (1 - b(B))v_t,$$

where $m = \max\{p, q\}$ and $\nu_t = X_t^2 - \sigma_t^2$. When the autoregressive lag polynomial $\phi(B) := 1 - a(B) - b(B)$ contains a unit root, the GARCH(p, q) process is said to be integrated in variance (Engle and Bollerslev 1986). The Integrated GARCH(p, q) or IGARCH(p, q) class of models is given by

$$\phi(B)(1 - B)X_t^2 = \omega + (1 - b(B))\nu_t.$$

The FIGARCH(p, d, q) class of models is simply obtained by allowing the differencing operator in the above equation to take non-integer values, that is

$$\phi(B)(1 - B)^d X_t^2 = \omega + (1 - b(B))\nu_t,$$

with $b(B)$ and $\phi(B)$ representing lag polynomials having all their roots lying outside the unit circle. The fractional differencing parameter is denoted as d. The fractional differencing operator $(1 - B)^d$ is most conveniently expressed as

$$(1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k.$$

After rearrangement, the FIGARCH(p, d, q) model can be represented as

$$\sigma_t^2 = \frac{\omega}{1 - b(1)} + \lambda(B)X_t^2, \tag{2.37}$$

where

$$\lambda(B) = 1 - \phi(B)(1 - B)^d(1 - b(B))^{-1} = \sum_{i=1}^{\infty} \lambda_i B^i. \tag{2.38}$$

Here, $\lambda(1) = 1$ for every d, with $\lambda_i \geq 0$, for $i = 1, 2, \ldots$, so that the FIGARCH(p, d, q) model is well-defined and the conditional variance is positive for all t. Conrad and Haag (2006) obtained two sets of sufficient conditions for the conditional variance of the FIGARCH process to be non-negative almost surely. Nonetheless, general conditions are difficult to establish. The simplest version of the FIGARCH(p, d, q) model, which appears to be particularly useful in practice, is the FIGARCH$(1, d, 1)$ for which the volatility σ_t^2 takes the form as in (2.37) with $b(B) = b_1 B$ and $\phi(B) = \phi_1 B$ with $|b_1| < 1$. Necessary and sufficient conditions for the non-negativity of the conditional variance for the FIGARCH$(1, d, 1)$ were obtained by Conrad and Haag (2006). The FIGARCH model has the property that for high lags, say k, the distributed lag coefficients are $\lambda_k \simeq ck^{-d-1}$, with c a positive constant. This implies that the conditional variance can be expressed as a distributed lag of past squared returns with coefficients that decay at a slow, that is hyperbolic, rate which is consistent with the long memory property. Davidson (2004) proposed an alternative definition of the persistence properties

of the FIGARCH process in terms of the hyperbolic memory, aiming to make the distinction of the FIGARCH model from the geometric memory cases represented by the GARCH and IGARCH processes more precise.

The statistical properties of the general FIGARCH(p, d, q) process, however, remain unestablished. For example, conditions for the existence of a stationary solution as well as the source of long memory on volatility are not known. For example, Mikosch and Starica (2004) and Granger and Hyung (2004) advocated that spurious long memory can be detected from time series exhibiting structural breaks. As solution to this problem, Baillie and Morana (2009) proposed the Adaptative FIGARCH model, or A-FIGARCH model in short, which simultaneously accounts for long memory and incorporates a deterministic time-varying intercept which allows for breaks, cycles and changes in drift. The A-FIGARCH (p, d, q, k) model can be derived from the FIGARCH (p, d, q) in (2.37) by letting the intercept ω to be time-varying, that is

$$\sigma_t^2 = \omega_t + [1 - \phi(B)(1 - B)^d(1 - b(B))^{-1}]X_t^2,$$

or

$$\sigma_t^2 = \omega_t + \lambda(B)X_t^2,$$

with

$$\omega_t = \omega_0 + \sum_{j=1}^{k}[\gamma_j \sin(2\pi jt/T) + \delta_j \cos(2\pi jt/T)].$$

In practice, k is a small integer often taken as $k = 1$ or 2. An immediate advantage of this model is that it does not require pretesting to determine either the number of structural break points or their locations. Furthermore, this model does not require any smooth transition between volatility regimes. Note that the inclusion of the time-varying intercept component implies that the A-FIGARCH process is neither ergodic nor strictly stationary.

The FIAPARCH(p, d, q) model of Tse (1998) is a special case of (2.31) with

$$\sigma_t^\delta = \frac{\omega}{1 - \beta(B)} + \lambda(B)g(X_t), \tag{2.39}$$

where $g(X_t) = (|X_t| - \gamma X_t)^\delta$ with $|\gamma| < 1$ and $\delta \geq 0$, and $\lambda(B)$ defined as in (2.38) for every $0 < d < 1$, with $\lambda_i \geq 0$, for $i \in \mathbb{N}$, and $\omega > 0$. Furthermore, in order to allow for long memory, the fractional differencing parameter d is constrained to lie in the interval $0 < d < 1/2$. The FIAPARCH model nests two major classes of ARCH-type models: the APARCH and the FIGARCH models of Ding et al. (1993) and Baillie et al. (1996), respectively. When $d = 0$ the process reduces to the APARCH(p, q) model, whereas for $\gamma = 0$ and $\delta = 2$ the process reduces

to the FIGARCH(p, d, q) model. Conrad et al. (2008) pointed out some advantages of the FIAPARCH(p, d, q) class of models, namely (a) it allows for an asymmetric response of volatility to positive and negative shocks, thus being able to traduce the leverage effect. If $\gamma > 0$, negative shocks have stronger impact on volatility than positive shocks as would be expected in the analysis of financial time series. If $\gamma < 0$, the reverse happens; (b) in this particular class of models, it is the data that determines the power of returns for which the predictable structure in the volatility pattern is the strongest, and (c) the models are able to accommodate long memory in volatility, depending on the differencing parameter d.

The simplest version of the FIAPARCH(p, d, q) model, which appears to be particularly useful in practice, is the FIAPARCH$(1, d, 1)$ for which the volatility σ_t takes the form as in (2.39) with $\beta(B) = \beta B$ and $\phi(B) = \phi B$. Necessary and sufficient conditions for the non-negativity of the conditional variance for the FIAPARCH$(1, d, 1)$ resembles the ones obtained by Conrad and Haag (2006) for the FIGARCH$(1, d, 1)$ model.

Volatility, asymmetry and long memory may also be captured using various extensions of the model introduced by Tse (1998) and Davidson (2004) among others. For example, Diongue and Guégan (2007) introduced the so-called seasonal hyperbolic APARCH, in short S-HY-APARCH, model where

$$[1 - b(B)]\sigma_t^{\delta} = \omega + \{\phi(B)[1 - \tau(1 - (1 - B^s)^d)]\}g(X_t). \tag{2.40}$$

The parameter $\tau \geq 0$ permits to eliminate the non-stationarity of the process. Moreover, by assuming that the roots of $[1 - b(B)] = 0$ lie outside the unit circle, the conditional variance in (2.40) can be expressed as

$$\sigma_t^{\delta} = \frac{\omega}{1 - b(1)} + \{1 - a(B)(1 - b(B))^{-1}(1 - \tau(1 - (1 - B^s)^d))\}g(X_t).$$

Another popular class of GARCH-type models is the Exponential GARCH, EGARCH in short. Nelson (1991) introduced it in order to overcome some disadvantages exhibited by the GARCH models, namely (a) parameter restrictions that are often violated by estimated coefficients; (b) asymmetric responses of shocks; and (c) interpreting whether shocks to conditional variance persist or not is difficult in GARCH models, since the usual norms measuring persistence often do not agree. The family of EGARCH(p, q) models can be defined as in (2.31) with

$$\ln(\sigma_t^2) = a_0 + \sum_{i=1}^{p} a_i g(Z_{t-i}) + \sum_{j=1}^{q} b_j \ln(\sigma_{t-j}^2). \tag{2.41}$$

For example, setting $g(Z_t) = \theta Z_t + \gamma(|Z_t| - E|Z_t|)$ with non-zero θ and γ in (2.41), we get the EGARCH model of Nelson (1991). Moreover, if in (2.31) and (2.41) we set $g(Z_t) = \theta_i \ln(Z_t^2)$, for $i = 1, \ldots, p$, then we get the logarithmic GARCH (LGARCH) model proposed by Geweke (1986) and Pantula (1986).

As a final class of GARCH-type processes, we mention the model introduced by Liu (2009) which is a generalization of the first-order GARCH processes family introduced in He and Teräsvirta (1999) and further developed by Ling and McAleer (2002). These authors defined the following general class for the GARCH(1, 1) model. Assume that in (2.31), σ_t is modeled by

$$\sigma_t^{\delta} = g(Z_{t-1}) + c(Z_{t-1})\sigma_{t-1}^{\delta}, \ t \in \mathbb{Z},$$

where $\delta > 0$, (Z_t) is a sequence of i.i.d. non-degenerate r.v's with mean zero. Further, it is assumed that Z_t is independent of $X_{t-1}, X_{t-2}, \ldots$, and $g(\cdot)$ is a positive function whereas $c(\cdot)$ is a non-negative function. This family of GARCH processes includes the GARCH(1, 1) model of Bollerslev (1986), the absolute value GARCH(1, 1) model of Taylor (1986) and of Schwert (1989), the nonlinear GARCH(1, 1) model of Engle (1990), the asymmetric GJR-GARCH(1, 1) model of Ding et al. (1993), the TARCH model (Rabemananjara and Zakoïan 1993; Zakoïan 1994), the 4NLGMACH(1, 1) model of Yang and Bewley (1995), the generalized quadratic ARCH(1, 1) model of Sentana (1995), and the volatility switching GARCH(1, 1) model of Fornari and Mele (1997).

Liu (2009) extends He and Teräsvirta (1999) results by allowing for an influence of higher-order past errors and conditional variances on the current conditional variance. Specifically, Liu model for σ_t stands as follows:

$$\sigma_t^{\delta} = g(Z_{t-1}, \ldots, Z_{t-s}) + \sum_{k=1}^{r} c_k(Z_{t-k})\sigma_{t-k}^{\delta}, \ t \in \mathbb{Z},$$

where $g(Z, t, s) = g(Z_{t-1}, \ldots, Z_{t-s})$ is a strictly positive function and $c_k(\cdot)$, $k = 1, \ldots, r$, all are nonnegative functions. This new family of GARCH processes includes:

1. The GARCH(p, q) model of Bollerslev (1986) for $\delta = 2$, $g(Z, t, s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + a_k Z_{t-k}^2$ for $k = 1, \ldots, r$ with $r = \max\{p, q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.
2. The absolute value GARCH(1, 1) model of Taylor (1986) and of Schwert (1989) for $\delta = 1$, $g(Z, t, s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + a_k|Z_{t-k}|$ for $k = 1, \ldots, r$ with $r = \max\{p, q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.
3. The volatility switching GARCH(1, 1) model of Fornari and Mele (1997) for $\delta = 2$, $g(Z, t, s) = a_0 + \sum_{k=1}^{s} \gamma_k \text{sgn}(Z_{t-k})$, $c_k(Z_{t-k}) = b_k + a_k Z_{t-k}^2$ for $k = 1, \ldots, r$ with $r = \max\{p, q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.
4. The nonlinear GARCH(p, q) model of Engle (1990).

   (a) Case $\delta = 1$: $g(Z, t, s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + a_k(1 - 2\eta \text{sgn}(Z_{t-k}) + \eta^2)|Z_{t-k}|$ for $k = 1, \ldots, r$ with $r = \max\{p, q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.

(b) Case $\delta = 2$: $g(Z,t,s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + a_k(1 - 2\eta\text{sgn}(Z_{t-k}) + \eta^2)Z_{t-k}^2$ for $k = 1,\ldots,r$ with $r = \max\{p,q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.

5. The GJR-GARCH(p,q) model of Glosten et al. (1993) for $\delta = 2$ $g(Z,t,s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + (a_k\omega_k I(Z_{t-k}))Z_{t-k}^2$ where $I(Z_{t-k}) = 1$ if $Z_{t-k} < 0$ and $I(Z_{t-k}) = 0$ otherwise, for $k = 1,\ldots,r$ with $r = \max\{p,q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.
6. The APARCH(p,q) model of Ding et al. (1993) for $\delta > 0$, $g(Z,t,s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + a_k(1 - 2\eta\text{sgn}(Z_{t-k}) + \eta^2)|Z_{t-k}|^\delta$ for $k = 1,\ldots,r$ with $r = \max\{p,q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively.
7. The threshold GARCH(p,q) model for $\delta > 0$, $g(Z,t,s) \equiv a_0$, $c_k(Z_{t-k}) = b_k + (a_{1k}(1 - I(Z_{t-k})) + a_{2k}I(Z_{t-k}))|Z_{t-k}|^\delta$ for $k = 1,\ldots,r$ with $r = \max\{p,q\}$, $a_i = 0$ and $b_j = 0$ for $i > p$ and $j > q$, respectively. Note that this is generalization of the models introduced by Zakoïan (1994), Hwang and Woo (2001), and Hwang and Basawa (2004).
8. The 4NLGMACH$(1,1)$ model of Yang and Bewley (1995) for $\delta = 2$, $g(Z,t,s) = a_0 + \sum_{k=1}^{s} a_{1k}(Z_{t-k} - d_k)^2 + a_{2k}(Z_{t-1} - d_k)^4$, $c_k(Z_{t-k}) = b_k$ for $k = 1,\ldots,r$. As pointed out by Liu (2009) this is a generalization of the family of moving-average conditional heteroskedasticity models proposed by Yang and Bewley (1995).
9. The first-order GARCH model of He and Teräsvirta (1999) with $r = 1$ and $s = 1$

We refer the reader to Andersen et al. (2009) for the recent developments and applications of this class of models.

Parameter-Driven Models

Parameter driven models for conditional variance are based on representing the variance of the process by a latent stochastic component. A simple example is the log-normal stochastic variance or volatility model

$$X_t|W_t \sim N(0, \exp(W_t)),$$

$$W_{t+1} = \gamma_0 + \gamma_1 W_t + \nu_t,$$

where $\nu_t \sim i.i.d.\ N(0,\sigma^2)$. Here W_t is not observed but can be estimated using the observations. These models lack analytic one-step ahead forecast densities and they need to be approximated through numerical methods. However they extend to higher dimensions and have continuous time analogs; see Sect. 2.2.7 for an extended treatment of parameter-driven models. Recent advances in hierarchical modeling techniques and simulation-based inferential methods make these generalized state space models very attractive.

2.2.6 Mixed Models for the Conditional Mean and Variance

The objective behind these models is to join models for the conditional mean and conditional variance given in the previous sections under a single model. In its simplest form, these composite models can be given as

$$X_t = f(\mathcal{F}_{t-1}, \boldsymbol{\theta}_1) + g(\mathcal{F}_{t-1}, \boldsymbol{\theta}_2) Z_t,$$

Note that

$$E(X_t|\mathcal{F}_{t-1}) = f(\mathcal{F}_{t-1}, \boldsymbol{\theta}_1),$$

$$V(X_t|\mathcal{F}_{t-1}) = g(\mathcal{F}_{t-1}, \boldsymbol{\theta}_2)^2 V(Z_t).$$

Here, the function f and g can be chosen in accordance with the partial models for the conditional mean and variance, discussed in the previous sections. In the simplest case, the conditional mean can be modeled by a linear ARMA model, whereas the conditional variance can be modeled by a GARCH model. Typically, first the model for the conditional mean is fitted, then the conditional variance model is fitted to the residuals from this model. This is the standard procedure in fitting GARCH models.

However, models of the type

$$\begin{aligned} X_t &= \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{m} \sum_{j=0}^{l} b_{ij} X_{t-i} Z_{t-j} + Z_t \\ &= \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{l} b_{ij} X_{t-i} Z_{t-j} \\ &+ \sum_{i=1}^{m} b_{i0} X_{t-i} Z_t + Z_t. \end{aligned}$$

give rise to richer and more complex structures. For example, the model

$$X_t = aX_{t-1} + bX_{t-1}Z_{t-1} + cX_{t-1}Z_t + Z_t,$$

includes nonlinear dynamics both in the mean and the variance, since

$$E(X_t|\mathcal{F}_{t-1}) = ax_{t-1} + bx_{t-1}z_{t-1},$$

and

$$V(X_t|\mathcal{F}_{t-1}) = (1 + cx_{t-1}^2)\sigma^2.$$

Alternatively, we can consider bilinear models given in (2.25), whose innovations are generated by a GARCH model. For example the model

$$X_t = \sum_{i=1}^{r} \sum_{j=1}^{s} b_{ij} X_{t-i} Z_{t-j} + Z_t, \tag{2.42}$$

where $i > j$, and the innovations Z_t are generated by the ARCH(q) process will represent nonlinear dynamics both in the mean and the variance.

The fundamental difference between GARCH and bilinear models is that whereas for GARCH models $E(X_t|\mathcal{F}_{t-1}) = 0$, and $V(X_t|\mathcal{F}_{t-1}) = h_t\sigma_t^2$, for bilinear processes given by (2.25), $E(X_t|\mathcal{F}_{t-1})$ has a nonlinear structure and $V(X_t|\mathcal{F}_{t-1})$ is constant. However, both classes of models can have similar unconditional moments. Often, upon fitting an adequate linear model for the conditional mean, the presence of linear dependence in the squared residuals is tested and this test is used as an indication for the presence of GARCH or bilinear type nonlinear structures in the series. However, these tests cannot provide a guidance in choosing the specific model for the series.

Mixed models of the type described above are quite rich in representing nonlinear dynamics and are seemingly attractive, but conditions of stationarity and invertibility are very difficult if not impossible to verify. Also as described above, there are problems with model identification, thus making these classes of models difficult to manage in practice.

Finite-Order Volterra Series

Infinite-order convergent Volterra series representation is the most general nonlinear representation for stationary time series. This suggests using the finite-order Volterra series

$$\begin{aligned} Y_t^{(m)} = & \sum_{i_1=0}^{k_1} g_{i_1} Z_{t-i_1} \\ & + \sum_{i_1=0}^{k_2} \sum_{i_2=0}^{k_3} g_{i_1 i_2} Z_{t-i_1} Z_{t-i_2} \\ & + \sum_{i_1=0}^{k_4} \sum_{i_2=0}^{k_5} \sum_{i_3=0}^{k_6} g_{i_1 i_2 i_3} Z_{t-i_1} Z_{t-i_2} Z_{t-i_3} \\ & + \cdots \\ & + \sum_{i_1=0}^{k_7} \sum_{i_2=0}^{k_8} \cdots \sum_{i_m=0}^{k_{2m}} g_{i_1 i_2 \cdots i_m} Z_{t-i_1} Z_{t-i_2} \cdots Z_{t-i_m}, \end{aligned}$$

as a parametric model. Finite-order Volterra series are used as flexible models for input-output systems where the input process, as well as the output process, are observable. In these models, often $m = 2$, so that second-order approximations are used; see, for example, Mathews and Sicuranza (2000). Within the univariate time series context, it is possible to identify the order m of the series using tail index estimation; see Sect. 4.3 for details. Conditional least square method can be used for parameter estimation. However, the innovation process Z_t is not observed and Granger and Andersen (1978) argue that processes of the form

$$X_t = Z_t + aZ_{t-1}Z_{t-2},$$

where Z_t are i.i.d. r.v's cannot be invertible. Therefore, finite Volterra series as models have limited practical value since they cannot be used for forecasting; see Sect. 4.2 for further discussion on invertibility.

2.2.7 Generalized State Space Models

Although all arguments given in this section can be extended to multivariate time series, for the sake of ease in notation, we will consider only univariate time series. A state space model for a linear time series Y_t consists of two equations, denoted by the observation and the state equations, which are given by

$$Y_t = \mathbf{H}_t\mathbf{X}_t + U_t, t = 1, 2, \ldots \tag{2.43}$$

$$\mathbf{X}_{t+1} = \mathbf{G}_t\mathbf{X}_t + \mathbf{V}_t, t = 1, 2, \ldots. \tag{2.44}$$

Here, in the first equation, $\mathbf{H}_t$ is a sequence of matrices whose elements are (constant) parameters and observations Y_t are written as a linear function of the unobserved (latent) v-dimensional state vector $\mathbf{X}_t$, plus a white noise U_t. The second equation determines the evolution of the state process in time in terms of the previous state. Here, $\mathbf{G}_t$ is a sequence of $v \times v$ matrices of parameters and $\mathbf{V}_t$ is a v dimensional white noise process, uncorrelated with U_t. In the simplest case, when Y_t is univariate, we may model the observations as

$$Y_t = m_t + Z_t,$$

where the state process m_t is the mean of the process, and Z_t is white noise. The latent mean m_t of the process can be modeled as a simple random walk

$$m_t = m_{t-1} + v_t.$$

It is also possible to add further structure to the model for m_t. The Kalman recursions allow a unified approach to prediction and estimation for state-space models; see Brockwell and Davis (1996) and West and Harrison (1997). Fundamental

assumptions behind the state-space representation and the consequent Kalman recursions are linearity and the normality of the error structures. When these assumptions are no longer valid, then observation and state equations are given by

$$Y_t = f(\mathbf{X}_{t-1}, U_t),$$

$$\mathbf{X}_t = g(\mathbf{X}_{t-1}, \mathbf{V}_t),$$

for some nonlinear functions f and g and white noise processes U_t and $\mathbf{V}_t$ independent of each other. Except for some special cases, satisfactory treatment of such a system of difference equations is not possible, and it is more advantageous to work directly with conditional distributions (or densities if they exist) which represent the probability structure of the system. In general terms, these equations are represented by two conditional densities $p(y_t|\mathbf{x}_t, \boldsymbol{\theta})$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ is the vector of all model parameters of this state space representation. This general state space structure can take several forms depending on different sets of further assumptions on these densities, which we examine below. Typically, there are two sets of fundamental assumptions to facilitate mathematical tractability of these state space models. In parameter-driven models, observations $\mathbf{Y}_t$ are assumed to be independent, conditional on the realization of the state vector $\mathbf{X}_t$, and that the state process $\mathbf{X}_t$ is assumed to be a (latent) Markov process.

In observation-driven models, again observations $\mathbf{Y}_t$ are assumed to be independent conditional on the realizations of the state vector $\mathbf{X}_t$, but rather then assuming a Markovian structure for the state vector, a model is specified directly for $\mathbf{X}_t$ conditional on $\mathbf{Y}_{t-1}$ through the conditional density $p(\mathbf{x}_t|\mathbf{y}_{t-1}, \boldsymbol{\theta})$. These two types of models show fundamental differences, particularly in inferential methods. Parameter driven models, otherwise known as hidden Markov models, are particularly suitable for Bayesian hierarchical modeling and simulation-based inferential techniques. Due to some awkward integrals and updating equations, classical likelihood and least squares methods are not particularly suitable for these models. On the other hand, observation-driven models do not involve such updating equations and difficult integrations and hence permit straight forward likelihood and least square methods. However, it is very difficult to verify stationarity conditions for the observation-driven models; see Brockwell and Davis (1996) for detailed comparison of these models. Here we give a brief summary of these models.

Parameter-Driven Models

For simplicity, assume that we have univariate time series Y_t and the corresponding univariate state X_t. Let $\mathbf{Y}_{t-1} := (Y_{t-1}, Y_{t-2}, \dots)$ and $\mathbf{X}_{t-1} := (X_{t-1}, X_{t-2}, \dots)$. Instead of the linear equations (2.43) and (2.44), we define the observation and state equations in terms of the conditional densities, assuming they exist, in the following manner:

Assume that Y_t conditional on X_t, is independent of $(\mathbf{X}_{t-1}, \mathbf{Y}_{t-1})$, so that the density of Y_t conditional on $(\mathbf{X}_t, \mathbf{Y}_{t-1})$ can be written as

$$p(y_t|x_t, \mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\theta}) = p(y_t|x_t, \boldsymbol{\theta}). \tag{2.45}$$

We also assume that X_{t+1} conditional on X_t is independent of $(\mathbf{X}_{t-1}, \mathbf{Y}_t)$ so that we can write

$$p(x_{t+1}|x_t, \mathbf{x}_{t-1}, \mathbf{y}_t, \boldsymbol{\theta}) = p(x_{t+1}|x_t, \boldsymbol{\theta}). \tag{2.46}$$

For linear Gaussian state space equations, (2.43), (2.44), and the conditional densities (2.45), (2.46) represent the same probability model, with $p(y_t|x_t, \boldsymbol{\theta})$ and $p(x_{t+1}|x_t, \boldsymbol{\theta})$ being normal densities. The joint density of the n observations $\mathbf{Y}_n$ and the state $\mathbf{X}_n$ at each time point $t = 1, \ldots, n$ can be written as

$$\begin{aligned} p(\mathbf{y}_t, \mathbf{x}_t|\boldsymbol{\theta}) &= p(y_n|x_n, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \boldsymbol{\theta}) p(x_n, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}|\boldsymbol{\theta}) \\ &= p(y_n|x_n, \boldsymbol{\theta}) p(x_n|\mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \boldsymbol{\theta}) p(\mathbf{x}_{n-1}, \mathbf{y}_{n-1}|\boldsymbol{\theta}) \\ &\vdots \\ &= \left(\prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta})\right) \left(\prod_{i=2}^{n} p(x_i|x_{i-1}, \boldsymbol{\theta})\right) p(x_1). \end{aligned}$$

Note that

$$p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}) = \prod_i p(y_i|x_i, \boldsymbol{\theta}),$$

hence observations are independent, conditional on the state of the process, and the time series Y_t inherits the dependence structure of the state process X_t, which is often called the latent process. Note also that from (2.46), the state X_t is a Markov process. These are indeed strong assumptions but are necessary to bring in some mathematical tractability to nonlinear, non Gaussian structures.

Conditional densities $p(x_t|\mathbf{y}_t, \boldsymbol{\theta})$ and $p(y_{t+1}|\mathbf{y}_t, \boldsymbol{\theta})$ are particularly relevant in the study of the system from which one can calculate the conditional expectations $E(X_t|\mathbf{y}_t)$ and $E(Y_{t+1}|\mathbf{y}_t)$. The former and the latter conditional expectations are the best predictors for X_t and Y_{t+1} in terms of the observation $\mathbf{y}_t$ and are respectively called the filtering and prediction problem. With the above assumptions and using the Bayes' Theorem

$$\begin{aligned} p(x_t|\mathbf{y}_t, \boldsymbol{\theta}) &= \frac{p(x_t, y_t, \mathbf{y}_{t-1}|\boldsymbol{\theta})}{p(\mathbf{y}_t|\boldsymbol{\theta})} \\ &= \frac{p(y_t|x_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) p(x_t|\mathbf{y}_{t-1}, \boldsymbol{\theta}) p(\mathbf{y}_{t-1}|\boldsymbol{\theta})}{p(y_t|\mathbf{y}_{t-1}, \boldsymbol{\theta}) p(\mathbf{y}_{t-1}|\boldsymbol{\theta})} \\ &= \frac{p(y_t|x_t, \boldsymbol{\theta}) p(x_t|\mathbf{y}_{t-1}, \boldsymbol{\theta})}{p(y_t|\mathbf{y}_{t-1}, \boldsymbol{\theta})}. \end{aligned} \tag{2.47}$$

Here, the conditional density $p(x_t|\mathbf{y}_{t-1}, \boldsymbol{\theta})$ has to be calculated from the integral

$$p(x_t|\mathbf{y}_{t-1}, \boldsymbol{\theta}) = \int p(x_t, x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\theta})dx_{t-1}$$
$$= \int p(x_t|x_{t-1}, \boldsymbol{\theta})p(x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\theta})dx_{t-1}. \quad (2.48)$$

For non-Gaussian and nonlinear processes, this updating equation for X_t in terms of the state equations $p(x_t|x_{t-1}, \boldsymbol{\theta})$ is not immediately available and can be computationally complicated; hence $p(x_t|\mathbf{y}_t, \boldsymbol{\theta})$ in (2.47) does not admit closed form expression.

In order to solve recursive relation in t, one assumes that $p(x_1|\mathbf{y}_0, \boldsymbol{\theta}) = p(x_1|\boldsymbol{\theta})$. The density $p(x_{t+1}|\mathbf{y}_t, \boldsymbol{\theta})$ and the corresponding conditional expectation give the prediction for the future value of the state equation, whereas the predictions for the future observation $\hat{y}_{t+1}$ can be obtained as the expected value of the conditional density $p(y_{t+1}|\mathbf{y}_t, \boldsymbol{\theta})$. In the classical approach, where $\boldsymbol{\theta}$ are unknown but fixed model parameters to be estimated from data, the unknown parameters are substituted by their estimates $\hat{\boldsymbol{\theta}}$ and the *plug-in* predictions are obtained from $p(y_{t+1}|\mathbf{y}_t, \hat{\boldsymbol{\theta}})$. In the Bayesian context the parameters are r.v's and the predictions are obtained from the predictive density $p(y_{t+1}|\mathbf{y}_t)$ through the relationship

$$p(y_{t+1}|\mathbf{y}_t) = \int p(y_{t+1}|\mathbf{y}_t, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_t)d\boldsymbol{\theta}$$

and

$$p(y_{t+1}|\mathbf{y}_t, \boldsymbol{\theta}) = \int p(y_{t+1}|x_{t+1}, \boldsymbol{\theta})p(x_{t+1}|\mathbf{y}_t, \boldsymbol{\theta})dx_{t+1}.$$

The key expression for the classical and the Bayesian inferential methods is the likelihood function $L(\boldsymbol{\theta}|\mathbf{y}_t)$ which can be computed from the relation

$$L(\boldsymbol{\theta}|\mathbf{y}_t) = \int \cdots \int p(\mathbf{x}_t|\boldsymbol{\theta})p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})dx_1 \cdots dx_n$$
$$= \int \cdots \int p(x_1|\boldsymbol{\theta}) \prod_{i=2}^{n} p(x_i|x_{i-1}, \boldsymbol{\theta})p(y_i|x_i, \boldsymbol{\theta})dx_1 \cdots dx_n. \quad (2.49)$$

The computation of the likelihood given in (2.49) requires the computation of n-dimensional integrals. Except for few special cases, calculation of such integrals are very difficult. Thus one relies on approximate solutions based on numerical methods on Monte Carlo methods. Recent advances in Bayesian simulation-based inferential methods and composite likelihood methods permit efficient simulation based estimation techniques and approximations. In a Bayesian hierarchical setup, upon defining a prior density $p(\boldsymbol{\theta})$ for the (random) model parameters $\boldsymbol{\theta}$,

Bayesian inference relies on the joint density $p(\boldsymbol{\theta}, \mathbf{x}_t|\mathbf{y}_t)$ which is proportional to $p(\mathbf{x}_t, \mathbf{y}_t|\boldsymbol{\theta})p(\boldsymbol{\theta})$. This joint density does not have closed form expressions, and Monte Carlo methods, in particular recent sequential Monte Carlo methods and particle filters (see, e.g. Andrieu et al. 2010) provide a flexible computational framework to carry out inference for these data sets with complex time dependence structures. In Sect. 4.5 we give a very brief introduction to these simulation-based methods. In Sect. 4.4.3 we also give a brief introduction to composite likelihood methods which are used as alternative pseudo-likelihood method for the observation-driven generalized state space models.

However, it may be possible to escape from such computational difficulties by specifying in (2.47) a model for $p(x_t|\mathbf{y}_t, \boldsymbol{\theta})$, thus eliminating the need for the updating Eq. (2.48). This strategy simplifies inference for generalized state-space models and the resulting models are called the observation-driven models.

Observation-Driven Models

In observation-driven models, the observation equation is the same as in parameter-driven models; namely it is assumed that

$$p(y_t|x_t, \mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\theta}) = p(y_t|x_t, \boldsymbol{\theta}). \tag{2.50}$$

However, the representation of the state is done through the densities

$$p(x_t|\mathbf{y}_{t-1}, \boldsymbol{\theta}), t = 1, 2, \ldots. \tag{2.51}$$

Here, the updating equation for the state

$$p(x_t|x_{t-1}, \boldsymbol{\theta}),$$

is not specified, since the conditional density of the state vector given the data $p(x_t|\mathbf{y}_t, \boldsymbol{\theta})$ and the predictive density can be directly calculated from (2.47) and (2.48) respectively, with the estimated value of the parameter $\hat{\boldsymbol{\theta}}$. Within the Bayesian framework, when $\boldsymbol{\theta}$ is random with prior specification $p(\boldsymbol{\theta})$, this predictive density is calculated from

$$p(y_{t+1}|\mathbf{y}_t) = \int p(y_{t+1}|x_{t+1}, \boldsymbol{\theta})p(x_{t+1}|\boldsymbol{\theta}, \mathbf{y}_t)p(\boldsymbol{\theta}|\mathbf{y}_t)dx_{t+1}d\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta}|\mathbf{y}_t)$ is the posterior density.

The state equation (2.51) without specifying precisely how x_t translates from x_{t-1}, simplifies the calculation of the posterior and the predictive distributions, but observations are no longer Markovian and y_t depend on the whole $\mathbf{y}_t$ rather than y_{t-1}, so that

$$p(y_1, \ldots, y_n|\boldsymbol{\theta}) = \prod_{t=1}^{n} p(y_t|\mathbf{y}_{t-1}, \boldsymbol{\theta}). \tag{2.52}$$

The lack of Markovian property particularly makes it more difficult to verify stationarity conditions. Also, due to the lack of Markovian structure, observation-driven models are not suitable for Bayesian hierarchical modeling. The specification given by (2.50) and (2.51) is not unique, in the sense that it can hold for two different state equations having different transitions, resulting in the same likelihood (2.52) for the data. This model miss-specification can be overcome by assuming that

$$p(x_{t+1}|\mathbf{x}_t, \mathbf{y}_t) = p(x_{t+1}|\mathbf{y}_t),$$

that is assuming that x_t, conditional on $\mathbf{y}_{t-1}$, is independent of $\mathbf{x}_{t-1}$. In this case

$$\begin{aligned} p(\mathbf{x}_n, \mathbf{y}_n) &= p(y_n|x_n)p(x_n|\mathbf{y}_{n-1})p(\mathbf{x}_{n-1}, \mathbf{y}_{n-1}) \\ &\vdots \\ &= \prod_{t=1}^{n} p(y_n|x_n)p(x_t|\mathbf{y}_{t-1}). \end{aligned}$$

We give an example to highlight the difference between the two modeling strategies.

Example: State Space Models for Count Data

In this example, we follow Brockwell and Davis (1996) and Davis et al. (2003a). Assume that Y_t is a time series of counts. Let $\mathcal{F}_{t-1}$ be the σ-field generated by the observation $(Y_s, s \leq t_1)$, and let $\boldsymbol{W}_t$ be a vector of explanatory variables with dimension p, observed at time t.

1. **Parameter-driven model**:
   We assume that observations, conditional on the intensity function λ_t, are independent, having the observation equation

   $$Y_t|\lambda_t \sim \text{Po}(\lambda_t),$$

   so that the likelihood for the data $y_1, \ldots, y_n$ conditional on the realization of the state process or the intensity process λ_t is given by

   $$p(y_1, \ldots, y_n|\lambda_1, \ldots, \lambda_n) = \prod_{t=1}^{n} \frac{e^{-\lambda_t}(\lambda_t)^{y_t}}{y_t!}. \tag{2.53}$$

   The dependence structure is then introduced into the model through the state equation (or the link function)

$$\log \lambda_t = \boldsymbol{\beta}' \boldsymbol{W}_t + U_t, \tag{2.54}$$

where $\boldsymbol{\beta}$ is a p dimensional vector of regression coefficients and U_t a latent time-dependent process. In the simplest case, U_t is assumed to follow an AR(1) process of the form

$$U_t = \phi U_{t-1} + Z_t,$$

where (Z_t) is a sequence of i.i.d. $N(0, \sigma^2)$, independent of the Y_t process. In this case, the state equation can equivalently be written in terms of the conditional density $p(\lambda_t|\lambda_{t-1})$ by

$$p(\lambda_t|\lambda_{t-1}) \sim N(\mu_t, \sigma^2),$$

where

$$\mu_t = \boldsymbol{\beta}' \boldsymbol{W}_t + \phi(\lambda_{t-1} - \boldsymbol{\beta}' \boldsymbol{W}_{t-1}).$$

The above model expressed in terms of Eqs. (2.53) and (2.54) can be implemented in Bayesian context as a hierarchical model upon defining appropriately the prior specifications of parameters and hyper-parameters. The posterior density $p(\lambda_t|\mathcal{F}_{t-1})$ and the predictive density $p(Y_{t+1}|\mathcal{F}_t)$, as well as the posterior densities of all other model parameters can be obtained by applying proper simulation-based inferential methods; see Brockwell and Davis (1996) for an alternative Monte Carlo-based estimation method. Implementation of this model, using standard maximum likelihood estimation is not straightforward and can be difficult, since the closed form for the unconditional likelihood $p(y_1, \ldots, y_n)$ is obtained by integrating the conditional likelihood (2.53) with respect to the joint density of $p(\lambda_1 \ldots, \lambda_t)$. Brockwell and Davis (1996) suggest a simulation-based estimation based on the EM algorithm.

2. **Observation-driven model for the counts:**
Assume for the time-being that there are no explanatory variables available in modeling the counts and that only information available are the counts themselves $(y_1, \ldots, y_n)$. In this case, the observation-driven model can be written as

$$Y_t|\lambda_t \sim \text{Po}(\lambda_t),$$

where λ_t is written as a positive function of the observations $y_{t-1}, \ldots, y_1$. The class of INGARCH(p, q) processes is constructed by assuming a specific linear function for λ_t, where

$$\lambda_t = \mu + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, \tag{2.55}$$

where $\mu > 0, a_i \geq 0, b_j \geq 0$ for every $i = 1, \ldots, p$ and $j = 1, \ldots, q$, so that λ_t is strictly positive for every t. If we further assume that all the roots of the polynomial $A(B) = 1 - \sum_{i=1}^{p} a_i B^i$ lie outside the unit circle (for non-negative a_i this is equivalent to the condition $\sum_{i=1}^{p} a_i < 1$), then λ_t can be written in terms of the $(Y_s, s < t)$ as

$$\lambda_t = A^{-1}(B)\mu + \sum_{j=1}^{\infty} \pi_j Y_{t-j}.$$

Note that (Y_t) are no longer conditionally independent and the joint density of $(Y_1, \ldots, Y_n)$ is written as

$$p(y_1, \ldots, y_n) = \prod_{t=1}^{n} p(y_t | \mathcal{F}_{t-1}) p(y_1),$$

where

$$p(y_t | \mathcal{F}_{t-1}) = p(y_t | \lambda_t).$$

INGARCH(p, q) processes are restricted, so that the state equation λ_t can be written as a strictly positive, linear function of the observations. Such restriction simplifies the conditions of existence of stationary solutions, as well as estimation procedures. For example, the INGARCH(p, q) process defined above with $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$, is strictly stationary with finite second-order moments (see Ferland et al. 2006, for the case $p = q = 1$ and Weiß 2009 for the general one). The classical (conditional) likelihood-based inference is also relatively easy. Set $\boldsymbol{\beta} := (a_1, \ldots, a_p, b_1, \ldots, b_q)$, then the conditional log-likelihood is written in the form

$$L(\boldsymbol{\beta} | \mathbf{y}_n) = \sum_{t=\max(p,q)}^{n} [-\lambda_t(\boldsymbol{\beta}) + y_t \log \lambda_t(\boldsymbol{\beta}) - \log y_t!],$$

from which we get the score function

$$\frac{\partial L(\boldsymbol{\beta} | \mathbf{y}_n)}{\partial \boldsymbol{\beta}} = \left(\frac{\partial L(\boldsymbol{\beta} | \mathbf{y}_n)}{\partial \beta_i}, i = 1, \ldots, p + q\right),$$

where

$$\frac{\partial L(\boldsymbol{\beta} | \mathbf{y}_n)}{\partial \beta_i} = \sum_{t=\max(p,q)}^{n} \frac{\partial \lambda_t(\boldsymbol{\beta})}{\partial \beta_i} \left(\frac{y_t}{\lambda_t(\boldsymbol{\beta})} - 1\right).$$

The elements of the Hessian matrix $H_n(\boldsymbol{\beta})$ are calculated as

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_i} = \sum_{t=\max(p,q)}^{n} \left[\frac{\partial^2 \lambda_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_i} \left(\frac{y_t}{\lambda_t(\boldsymbol{\beta})} - 1 \right) - \frac{y_t}{\lambda_t^2(\boldsymbol{\beta})} \frac{\partial \lambda_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} \right] \tag{2.56}$$

from which a proper numerical optimization method can be constructed.

Extensions of (2.55) for the simplest case $p = q = 1$ have been recently proposed by Fokianos et al. (2009) by considering a more general representation for λ_t, namely

$$\lambda_t = f(\lambda_{t-1}) + g(Y_{t-1}), \ t \geq 1, \tag{2.57}$$

where $f, g : \mathbb{R}^+ \to \mathbb{R}^+$ are known functions up to an unknown finite dimensional parameter vector. The initial values Y_0 and λ_0 are assumed to be fixed. Special models for λ_t in (2.57) include the model in (2.55) upon defining $f(x) = c + dx$ and $g(x) = bx$ with $c, d, g > 0$ and $x \geq 0$, and the so-called exponential autoregressive model with

$$\lambda_t = (a_1 + c_1 \exp\{-\gamma_1 \lambda_{t-1}\})\lambda_{t-1} + bY_{t-1}.$$

Fokianos et al. (2009) proved that under geometric ergodicity the maximum likelihood estimators of the parameters are asymptotically Gaussian in the linear model (2.55); see also Tjøstheim (2012) and Fokianos (2011) for further details.

If we have explanatory variables to account for the variations in the latent intensity λ_t of the counts, then the statistical and probabilistic properties of the model get more complicated. In this case, in order to satisfy the positivity of the intensity process λ_t, we model $\log \lambda_t$ by a linear function, giving rise to

$$Y_t | \mathcal{F}_{t-1} \sim \text{Po}(\lambda_t),$$

$$\log \lambda_t = \boldsymbol{\beta}' \boldsymbol{W}_t + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}.$$

In this case, it is not clear under what conditions this process may be stationary. For example, the simpler process with

$$\log \lambda_t = \boldsymbol{\beta}' \boldsymbol{W}_t + \sum_{j=1}^{q} b_j Y_{t-j},$$

cannot be stationary unless some normalization is applied to the observations. Davis et al. (2003a) suggest using the model

$$\log \lambda_t = \boldsymbol{\beta}' \boldsymbol{W}_t + \sum_{j=1}^{q} \theta_i Z_{t-j},$$

where

$$Z_t = \frac{Y_t - \lambda_t}{\lambda_t^{\eta}}, \ \eta \geq 0.$$

Note that Y_t is not Markov process, but the intensity process λ_t is pth-order Markov. Existence of a stationary solution depends on the value of η. For example, for the simpler first-order model and assuming that $\boldsymbol{\beta}' \boldsymbol{W}_t = \boldsymbol{\beta}'$,

$$\log \lambda_t = \boldsymbol{\beta}' + \frac{Y_{t-1} - \lambda_{t-1}}{\lambda_t^{\eta}}.$$

Davis et al. (2003a) proved the existence of a stationary solution for $\eta \in [1/2, 1]$, showing that this solution is unique when $\eta = 1$.

Estimation of the parameters using likelihood is relatively easy and the likelihood is maximized by using the Newton-Raphson method; see Davis et al. (2003b) for details.

In Chap. 5, we will study alternative models for integer-valued time series, which have linear representations similar to ARMA models but are constructed with thinning operations.

2.2.8 *Max-Stable Moving Average Processes*

Max-stable moving average processes are introduced as models for heavy-tailed data by Davis and Resnick (1993). This class is defined as follows: X_t is said to be max-stable moving average process if

$$X_t = \bigvee_{j=0}^{\infty} \psi_j Z_{t-j},$$

where $\bigvee_j \psi_j \equiv \max_j \psi_j$ and (Z_t) are i.i.d. r.v's with distribution $\exp[-\sigma z^{-1}]$. Analogous finite parameter version of these models are also defined. The reason why the authors suggest such classes for modeling heavy-tailed data is that their sample paths very much resemble the sample paths of corresponding linear models formed from the same residuals, and the predictions and estimation of parameters for these models can be done by an optimality criterion which minimizes the probability of large errors, that is likely to give better fit to sudden burst. The optimal predictor can be explicitly written for several models. However, since second-order moments

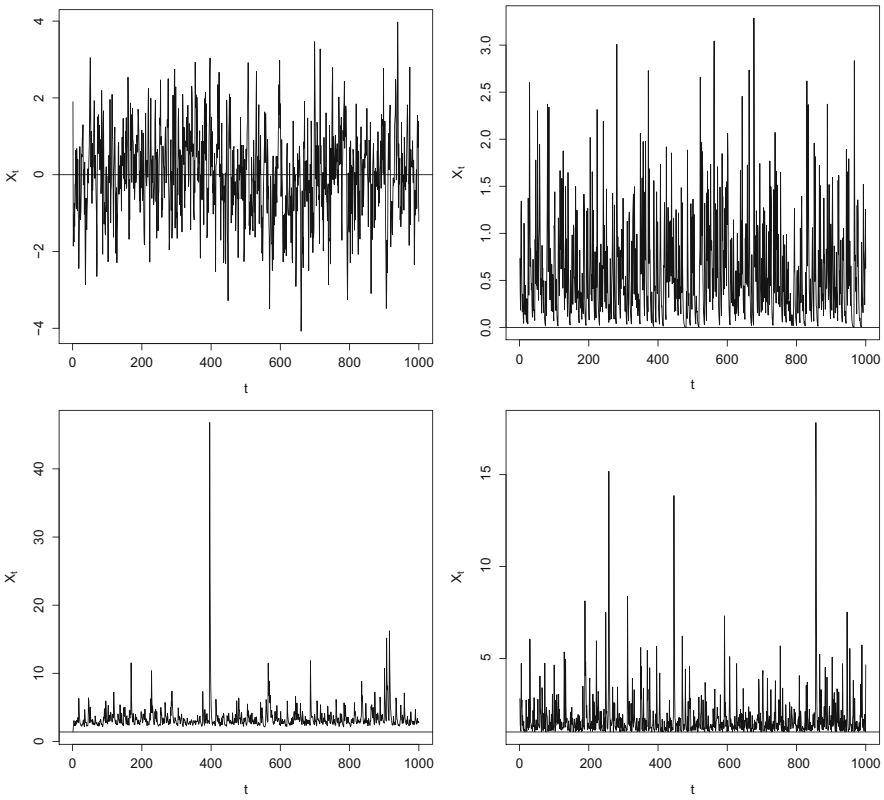

Fig. 2.8 Sample paths – AR(1) and max-stable models with N(0,1) errors (*top row*) and Pareto($\alpha = 2.5$) errors (*bottom row*)

cannot be used for identification and estimation, such classes are not very frequently used in practice. Figure 2.8 shows the sample path of $n = 1{,}000$ observations generated from an AR(1), $X_t = 0.5X_{t-1} + Z_t$, and the corresponding max-stable process $X_t = \max(0.5X_{t-1}, Z_t)$, where Z_t is a N(0, 1) sequence. The same models are represented in the bottom row for Pareto($\alpha = 2.5$) residuals.

2.2.9 Nonparametric Methods

In the class of parametric models, the main emphasis was on building parametric models for the conditional mean and variance of the process, either separately or jointly. If the emphasis is on prediction rather than on explaining how these conditional means and variances change in time, then a plausible alternative is to estimate them using nonparametric methods. This would be quite flexible, since one

is not restricted by a specific parametric model. The most common way is to use kernel estimators. For example, for a given time series (X_t) the conditional mean

$$M(x_1, \ldots, x_p) := E(X_t | X_{t-1} = x_1, \ldots, X_{t-p} = x_p),$$

can be estimated by

$$\hat{M}(x_1, \ldots, x_p) = \frac{(n-p)^{-1} \sum_{t=p+1}^{n} X_t \prod_{i=1}^{p} K_h(X_{t-i} - x_i)}{(n-p+1)^{-1} \sum_{t=p+1}^{n+1} \prod_{i=1}^{p} K_h(X_{t-i} - x_i)},$$

where

$$K_h(x) = \frac{1}{h} K(\frac{x}{h})$$

and K representing a kernel function. A similar expression can be obtained for the conditional variance. The drawback with these models is that one needs large sample sizes for a reasonable fit and the sample needs to increase drastically with the increase in p, a typical case of the curse of dimension. The curse of dimension can be reduced by simplifying the model. In the simplest case, one can model the conditional mean function by

$$M(x_1, \ldots, x_p) = \sum_j f_j(x_p),$$

where the $f_j(\cdot)$ are unknown functions to be estimated. Each of these functions are one-dimensional, thus simplifying the problem. A similar model can be written for the conditional variance. Such additive models can further be extended to include linear combinations of past values. Such models are known as projection pursuit models. In general, these models are taken from regression context and adopted to the time series context. Further models and references can be found in Tjøstheim (1994) and Gao (2007). Another possibility is to use splines in estimating these conditional means and variances. We refer the reader to Hardle et al. (1997) for a general review of nonparametric methods in time series analysis.

Bayesian Nonparametric Methods

Bayesian nonparametric methods have been one of the fastest growing topics in statistics. Bayesian methods inherently are likelihood based; therefore they need specification of a parametric model. Indeed, what is usually called nonparametric Bayesian method, in fact corresponds to models with priors defined on an infinite dimensional parameter space. Suppose that X_t is a stochastic process with probability measure F defined through its finite dimensional distributions. In ordinary Bayesian inferential methods, one assumes a parametric model for the probability

distribution and expresses prior belief on the parameters and then the inference concentrates on deriving the posterior distribution of the model parameters and the predictive distribution for the future values of the process. In nonparametric Bayesian methods, no parametric form is assumed for the probability distribution; instead prior beliefs are assigned to the probability distributions (i.e. models) which are now random elements themselves belonging to some measure space. Hence consider the probability space $(\Omega, \mathcal{B}, P)$ where the random variable (or the stochastic process) resides. Typically $\Omega = \mathbb{R}^d$, $\mathcal{B}$ the Borel σ-algebra over Ω and P is the probability measure of the random variable. Assume that P is a random measure residing in a space of probability measures $(\mathcal{P}, \mathcal{C}, \mathcal{Q})$ so that the probability measures P of the random variable (or the stochastic process) is a simple element of $\mathcal{P}$. Often $(\Omega, \mathcal{B}, P)$ is called the base space and $(\mathcal{P}, \mathcal{C}, \mathcal{Q})$ the distributional space. The Dirichlet process is a probability measure on $(\mathcal{P}, \mathcal{C})$ and is often used as the prior distribution for the random measure P. A Dirichlet process (DP) is defined by a concentration parameter α and base distribution P_0. Random measure P is said to follow a DP prior if for any measurable partition $(A_1, A_2, \ldots, A_k)$ of the sample space of the random variable, the vector $(P(A_1), \ldots, P(A_k))$ has a Dirichlet distribution with parameters $(\alpha P_0(A_1), \ldots, \alpha P_0(A_k))$. The DP is centered at P_0, so that $E(P(A)) = P_0(A)$ for any measurable set $A \in \mathcal{B}$. The inferential problem is then given in terms of a hierarchical representation. For example, in the simplest form, when the observed data are i.i.d. with common marginal distribution F, the hierarchical model is given as

$$x_1, x_2, \ldots, x_n | F \sim \text{ i.i.d. } F,$$

$$F | \alpha, F_0 \sim DP(\alpha, F_0),$$

whereas the classical Bayesian parametric modeling paradigm would result in the following hierarchical representation;

$$x_1, x_2, \ldots, x_n | \boldsymbol{\theta} \sim \text{ i.i.d. } F(x|\boldsymbol{\theta}),$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}),$$

where, $\pi(\boldsymbol{\theta})$ is the prior distribution of the model parameters $\boldsymbol{\theta}$. The difference in these alternative approaches is evident.

For time-dependent data, the specification of the hierarchical model needs the notion of dependent Dirichlet processes and is beyond the scope of this book. We refer the reader to Rodriguez (2007) and Hjort et al. (2010) for excellent accounts of Bayesian nonparametric modeling.

In Sect. 4.5 we will give a detailed summary of Bayesian inferential methods for nonlinear time series based on parametric likelihood methods.

References

Andersen TG, Davis RA, Kreib J-P, Mikosch T (eds) (2009) Handbook of financial time series. Springer, Berlin/Heidelberg

Andrews B, Davis RA, Breidt FJ (2006) Maximum likelihood estimation for all-pass time series models. J Multivar Anal 97:1638–1659

Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods. J R Stat Soc B 72:269–342

Baillie RT, Morana C (2009) Modelling long memory and structural breaks in conditional variances: an adaptive FIGARCH approach. J Econ Dyn Control 33:1577–1592

Baillie RT, Bollerslev T, Mikkelsen HO (1996) Fractionally integrated generalized autoregressive conditional heteroskedasticity. J Econom 74:3–30

Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH models: a survey. J Appl Econom 21:79–109

Berkes I, Horváth L, Kokoszka P (2003) GARCH processes: structure and estimation. Bernoulli 9:201–227

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31:307–327

Bollerslev T, Chou RY, Kroner KF (1992) ARCH modelling in finance: a review of the theory and empirical evidence. J Econom 52:5–59

Bollerslev T, Engle RF, Nelson DB (1994) ARCH models. In: Engle RF, McFadden DL (eds) Handbook of econometrics. North Holland, Amsterdam, pp 2959–3038

Brandt A (1986) The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. Adv Appl Probab 18:211–220

Brockett RW (1976) Non-linear systems and differential geometry. Automatica 12:167–176

Brockwell PJ, Davis RA (1991) Time series: theory and methods. Springer, New York

Brockwell PJ, Davis RA (1996) Introduction to time series and forecasting. Springer, New York

Chan KS, Tong H (1986) On estimating thresholds in autoregressive models. J Time Ser Anal 7:179–190

Conrad C, Haag BR (2006) Inequality constraints in the fractionally integrated GARCH model. J Financ Econom 4:413–449

Conrad C, Karanasos M, Zeng N (2008) Multivariate fractionally integrated APARCH modeling of stock market volatility: a multi-country study. Discussion paper no. 472, University of Heidelberg

Cont R (2001) Empirical properties of asset returns: stylized facts and statistical issues. Quant Finance 1:223–236

Davidson JEH (2004) Conditional heteroskedasticity models and a new model. J Bus Econom Stat 22:16–29

Davis RA, Resnick SI (1993) Prediction of stationary max-stable processes. Ann Appl Probab 3:497–525

Davis RA, Dunsmuir WTM, Streett SB (2003a) Observation-driven models for Poisson counts. Biometrika 90:777–790

Davis RA, Dunsmuir WTM, Streett SB (2003b) Maximum likelihood estimation for an observation driven model for Poisson counts. Methodol Comput Appl Probab 7:149–159

Davis RA, Lee TCM, Rodriguez-Yam GA (2008) Break detection for a class of nonlinear time series models. J Time Ser Anal 29:834–867

Ding Z, Granger CWJ, Engle RF (1993) A long memory property of stock market returns and a new model. J Empir Finance 1:83–106

Diongue AK, Guégan D (2007) The stationary seasonal hyperbolic asymmetric power ARCH model. Stat Probab Lett 77:1158–1164

Engle RF (1982) Autoregressive conditional heteroskedascity with estimates of the United Kingdom inflation. Econometrica 50:987–1008

Engle RF (1990) Discussion: stock market volatility and the crash of 87. Rev Financ Stud 3: 103–106
Engle RF (2004) Nobel lecture. Risk and volatility: econometric models and financial practice. Am Econ Rev 94:405–420
Engle RF, Bollerslev T (1986) Modelling the persistence of conditional variances. Econom Rev 5:1–50
Fan J, Yao Q (2003) Nonlinear time series. Springer, New York
Ferland R, Latour A, Oraichi D (2006) Integer-valued GARCH processes. J Time Ser Anal 27: 923–942
Fokianos K (2011) Some recent progress in count time series. Stat Pap 45:49–58
Fokianos K, Rahbek A, Tjøstheim D (2009) Poisson autoregression. J Am Stat Assoc 104:1430–1439
Fornari F, Mele A (1997) Sign- and volatility-switching ARCH models: theory and applications to international stock markets. J Appl Econom 12:49–65
Franses PH, van Dijk D (2000) Non-Linear Time Series Models in Empirical Finance. Cambridge University Press, New York
Friedman M (1977) Nobel lecture: inflation and unemployment. J Polit Econ 85:451–472
Gao J (2007) Nonlinear time series: semiparametric and nonparametric methods. Chapman and Hall, Boca Raton
Geweke J (1986) Modeling the persistence of conditional variances: a comment. Econom Rev 5:57–61
Glosten L, Jagannathan R, Runkle D (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks. J Finance 48:1779–1801
Goldfeld SM, Quandt R (1973) The estimation of structural shifts by switching regressions. Ann Econ Soc Meas 2:475–85
Granger CWJ, Andersen A (1978) On the invertibility of time series models. Stoch Process Appl 8:87–92
Granger CWJ, Hyung N (2004) Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. J Empir Finance 11:399–421
Granger CWJ, Teräsvirta T (1993) Modelling nonlinear economic relationships. Oxford University Press, New York
Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
Hamilton JD (2008) Regime switching models. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, Basingstoke/New York
Härdle W, Lütkepohl H, Chen R (1997) A review of nonparametric time series analysis. Int. Stat. Rev. 65:49–72
He C, Teräsvirta T (1999) Properties of moments of a family of GARCH processes. J Econom 92:173–192
Higgins ML, Bera AK (1992) A class of nonlinear ARCH models. Int Econ Rev 33:137–158
Hjort N, Holmes C, Mueller P, Walker S (2010) Bayesian Nonparametrics: Principles and Practice. Cambridge University Press, Cambridge
Hwang SY, Basawa IV (2004) Stationarity and moment structure for Box-Cox transformed threshold GARCH(1, 1) processes. Stat Probab Lett 68:209–220
Hwang SY, Woo MJ (2001) Threshold ARCH(1) processes: asymptotic inference. Stat Probab Lett 53:11–20
Kashikar AS, Rohan N, Ramanathan TV (2013) Integer autoregressive models with structural breaks. J. Appl. Statist. 40:2653–2669
Ling S, McAleer M (2002) Stationary and the existence of moments of a family of GARCH processes. J Econom 106:109–117
Liu J-C (2009) Stationarity of a family of GARCH processes. Econom J 12:436–446
Mathews VJ, Sicuranza GL (2000) Polynomial signal processing. Wiley, New York
Meyn S, Tweedie RL (2009) Markov chains and stochastic stability. Cambridge University Press, Cambridge

Mikosch T, Starica C (2004) Changes of structure in financial time series and the GARCH model. REVSTAT 2:41–73
Nelson DB (1991) Conditional heteroskedasticity in asset returns: a new approach. Econometrica 2:347–370
Nisio M (1960) On polynomial approximation for strictly stationary processes. J. Math. Soc. Japan 12:207–226
Pagan A (1996) The econometrics of financial markets. J Empir Finance 3:15–102
Palm F (1996) GARCH models of volatility. In: Rao CR, Maddala GS (eds) Handbook of statistics, vol 14. North Holland, Amsterdam, pp 209–240
Pantula SG (1986) Modeling the persistence of conditional variances: a comment. Econom Rev 5:71–74
Pham DT, Tran TL (1981) On the first-order bilinear time series model. J Appl Probab 18:617–627
Priestley MB (1981) Spectral analysis and time series. Academic, London
Rabemananjara R, Zakoïan JM (1993) Threshold ARCH models and asymmetries in volatility. J Appl Econom 8:31–49
Resnick SI, Van den Berg E (2000) Sample correlation behaviour for the heavy tailed general bilinear process. Stoch Models 16:233–258
Robinson PM (1991) Testing for strong serial correlation and dynamic conditional heteroskedasity in multiple regression. J Econom 47:67–78
Rodriguez A (2007) Some advances in Bayesian nonparametric modeling. Unpublished doctoral thesis, Institute of Statistics and Decision Science, Duke University
Schwert GW (1989) Why does stock market volatility change over time? J Finance 45:1129–1155
Schwert GW (1990) Stock volatility and the crash of '87. Rev Financ Stud 3:77–102
Sentana E (1995) Quadratic ARCH models. Rev Econ Stud 62:639–661
Shephard N (1996) Statistical aspects of ARCH and stochastic volatility. In: Cox DR, Barndorff-Nielsen OE (eds) Likelihood, time series with econometric and other applications. Chapman and Hall, London
Silvennoinen A, Teräsvirta T (2009) Multivariate GARCH models. In: Andersen TG, Davis RA, Kreiss J-P, Mikosch T (eds) Handbook of financial time series. Springer, New York, pp 201–229
Taylor S (1986) Modeling financial time series. Wiley, New York
Teräsvirta T (2009) An introduction to univariate GARCH models. In: Andersen TG, Davis RA, Kreiss J-P, Mikosch T (eds) Handbook of financial time series. Springer, New York, pp 17–42
Terdik G (1999) Bilinear stochastic models and related problems of nonlinear time series analysis. Springer, New York
Tjøstheim D (1994) Non-linear time series: a selective review. Scand J Stat 21:97–130
Tjøstheim D (2012) Some recent theory for autoregressive count time series. Test 21:413–438. (With discussion)
Tong H (1990) Non-linear time series. Oxford Science Publications, Oxford
Tse Y (1998) The conditional heteroskedascity of the Yen-Dollar exchange rate. J Appl Econom 13:49–55
Weiß CH (2009) Modelling time series of counts with overdispersion. Stat Methods Appl 18: 507–519
West M, Harrison J (1997) Bayesian forecasting and dynamic models. Springer, New York
Yang M, Bewley R (1995) Moving average conditional heteroskedastic processes. Econ Lett 49:367–372
Zakoïan JM (1994) Threshold heteroskedastic models. J Econ Dyn Control 18:931–955
Zivot E, Wang J (2006) Modeling financial time series with S-PLUS. Springer, New York

http://www.springer.com/978-3-319-07027-8

Non-Linear Time Series
Extreme Events and Integer Value Problems
Turkman, K.; Scotto, M.G.; de Zea Bermudez, P.
2014, XII, 245 p. 41 illus., Hardcover
ISBN: 978-3-319-07027-8