Forecasting Stationary Processes

3

An important goal of time series analysis is forecasting. In the following we will consider the problem of forecasting X_{T+h}, $h > 0$, given $\{X_T, \ldots, X_1\}$ where $\{X_t\}$ is a stationary stochastic process with known mean μ and known autocovariance function $\gamma(h)$. In practical applications μ and γ are unknown so that we must replace these entities by their estimates. These estimates can be obtained directly from the data as explained in Sect. 4.2 or indirectly by first estimating an appropriate ARMA model (see Chap. 5) and then inferring the corresponding autocovariance function using one of the methods explained in Sect. 2.4. Thus the forecasting problem is inherently linked to the problem of identifying an appropriate ARMA model from the data (see Deistler and Neusser 2012).

3.1 The Theory of Linear Least-Squares Forecasts

We restrict our discussion to linear *forecast functions*, also called *linear predictors*, $\mathbb{P}_T X_{T+h}$. Given observation from period 1 up to period T, these predictors take the form:

$$\mathbb{P}_T X_{T+h} = a_0 + a_1 X_T + \ldots + a_T X_1 = a_0 + \sum_{i=1}^{T} a_i X_{T+1-i}$$

with unknown coefficients $a_0, a_1, a_2, \ldots, a_T$. In principle, we should index these coefficients by T because they may change with every new observations. See the example of the MA(1) process in Sect. 3.1.2. In order not to overload the notation, we will omit this additional index.

In the Hilbert space of random variables with finite second moments the optimal forecast in the mean squared error sense is given by the conditional expectation

K. Neusser, *Time Series Econometrics*, Springer Texts in Business and Economics,
DOI 10.1007/978-3-319-32862-1_3

$\mathbb{E}(X_{T+h}|c, X_T, X_{T-1}, \ldots, X_1)$. However, having practical applications in mind, we restrict ourself to linear predictors for the following reasons[1]:

(i) The determination of the conditional expectation is usually very difficult because all possible functions must in principle be considered whereas linear predictors are easy to compute.
(ii) The coefficients of the optimal (in the sense of means squared errors) linear forecasting function depend only on the first two moments of the time series, i.e. on $\mathbb{E}X_t$ and $\gamma(j), j = 0, 1, \ldots, h + T - 1$.
(iii) In the case of Gaussian processes the conditional expectation coincides with the linear predictor.
(iv) The optimal predictor is linear when the process is a causal and invertible ARMA process even when Z_t follows an arbitrary distribution with finite variance (see Rosenblatt 2000, chapter 5).
(v) Practical experience has shown that even non-linear processes can be predicted accurately by linear predictors.

The coefficients $a_0, \ldots, a_T$ of the forecasting function are determined such that the mean squared errors are minimized. The use of mean squared errors as a criterion leads to a compact representation of the solution to the forecasting problem. It implies that over- and underestimation are treated equally. Thus, we have to solve the following minimization problem:

$$\begin{aligned} S = S(a_0, \ldots, a_T) &= \mathbb{E}\left(X_{T+h} - \mathbb{P}_T X_{T+h}\right)^2 \\ &= \mathbb{E}(X_{T+h} - a_0 - a_1 X_T - \ldots - a_T X_1)^2 \longrightarrow \min_{a_0, a_1, \ldots, a_T} \end{aligned}$$

As S is a quadratic function, the coefficients, a_j, $j = 0, 1, \ldots, T$, are uniquely determined by the so-called normal equations. These are obtained from the first order conditions of the minimization problem, i.e. from $\frac{\partial S}{\partial a_j} = 0$, $j = 0, 1, \ldots, T$:

$$\frac{\partial S}{\partial a_0} = \mathbb{E}\left(X_{T+h} - a_0 - \sum_{i=1}^{T} a_i X_{T+1-i}\right) = 0, \tag{3.1}$$

$$\frac{\partial S}{\partial a_j} = \mathbb{E}\left[\left(X_{T+h} - a_0 - \sum_{i=1}^{T} a_i X_{T+1-i}\right) X_{T+1-j}\right] = 0, \quad j = 1, \ldots, T. \tag{3.2}$$

The first equation can be rewritten as $a_0 = \mu - \sum_{i=1}^{T} a_i \mu$ so that the forecasting function becomes:

$$\mathbb{P}_T X_{T+h} = \mu + \sum_{i=1}^{T} a_i \left(X_{T+1-i} - \mu\right).$$

[1] Elliott and Timmermann (2008) provide a general overview of forecasting procedures and their evaluations.

The unconditional mean of the forecast error, $\mathbb{E}\left(X_{T+h} - \mathbb{P}_T X_{T+h}\right)$, is therefore equal to zero. This means that there is no bias, neither upward nor downward, in the forecasts. The forecasts correspond on average to the "true" value.

Inserting in the second normal equation the expression for $\mathbb{P}_T X_{T+h}$ from above, we get:

$$\mathbb{E}\left[\left(X_{T+h} - \mathbb{P}_T X_{T+h}\right) X_{T+1-j}\right] = 0, \qquad j = 1, 2, \ldots, T.$$

The forecast error is therefore uncorrelated with the available information represented by past observations. Thus, the forecast errors $X_{T+h} - \mathbb{P}_T X_{T+h}$ are orthogonal to $X_T, X_{T-1}, \ldots, X_1$. Geometrically speaking, the best linear forecast is obtained by finding the point in the linear subspace spanned by $\{X_T, X_{T-1}, \ldots, X_1\}$ which is closest to X_{T+h}. This point is found by projecting X_{T+h} on this linear subspace.[2]

The normal equations (3.1) and (3.2) can be rewritten in matrix notation as follows:

$$a_0 = \mu\left(1 - \sum_{i=1}^{T} a_i\right) \tag{3.3}$$

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(T-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(T-1) & \gamma(T-2) & \ldots & \gamma(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+T-1) \end{pmatrix}. \tag{3.4}$$

Denoting by ι, α_T and $\gamma_T(h)$ the vectors $(1, 1, \ldots, 1)'$, $(a_1, \ldots, a_T)'$ and $(\gamma(h), \ldots, \gamma(h+T-1))'$ and by $\Gamma_T = [\gamma(i-j)]_{i,j=1,\ldots,T}$ the symmetric $T \times T$ covariance matrix of $(X_1, \ldots, X_T)'$ the normal equations can be written compactly as:

$$a_0 = \mu\left(1 - \iota'\alpha_T\right) \tag{3.5}$$

$$\Gamma_T \alpha_T = \gamma_T(h). \tag{3.6}$$

Dividing the second equation by $\gamma(0)$, one obtains an equation in terms autocorrelations instead of autocovariances:

$$R_T \alpha_T = \rho_T(h), \tag{3.7}$$

where $R_T = \Gamma_T/\gamma(0)$ and $\rho_T(h) = (\rho(h), \ldots, \rho(h+T-1))'$. The coefficients of the forecasting function α_T are then obtained by inverting Γ_T, respectively R_T:

[2] Note the similarity of the forecast errors with the least-square residuals of a linear regression.

$$\alpha_T = \begin{pmatrix} a_1 \\ \vdots \\ a_T \end{pmatrix} = \Gamma_T^{-1}\gamma_T(h) = R_T^{-1}\rho_T(h).$$

A sufficient condition which ensures the invertibility of Γ_T, respectively R_T, is given by assuming $\gamma(0) > 0$ and $\lim_{h\to\infty}\gamma(h) = 0$.[3] The last condition is automatically satisfied for ARMA processes because $\gamma(h)$ converges even exponentially fast to zero (see Sect. 2.4).

The mean squared error or *variance of the forecast error* for the forecasting horizon h, $v_T(h)$, is given by:

$$\begin{aligned} v_T(h) &= \mathbb{E}\left(X_{T+h} - \mathbb{P}_T X_{T+h}\right)^2 \\ &= \gamma(0) - 2\sum_{i=1}^{T} a_i\gamma(h+i-1) + \sum_{i=1}^{T}\sum_{j=1}^{T} a_i\gamma(i-j)a_j \\ &= \gamma(0) - 2\alpha_T'\gamma_T(h) + \alpha_T'\Gamma_T\alpha_T \\ &= \gamma(0) - \alpha_T'\gamma_T(h), \end{aligned}$$

because $\Gamma_T\alpha_T = \gamma_T(h)$. Bracketing out $\gamma(0)$, one can write the mean squared forecast error as:

$$v_T(h) = \gamma(0)\left(1 - \alpha_T'\rho_T(h)\right). \tag{3.8}$$

Because the coefficients of the forecast function have to be recomputed with the arrival of every new observation, it is necessary to have a fast and reliable algorithm at hand. These numerical problems have been solved by the development of appropriate computer algorithms, like the Durbin-Levinson algorithm or the innovation algorithm (see Brockwell and Davis 1991, Chapter 5).

3.1.1 Forecasting with an AR(p) Process

Consider first the case of an AR(1) process:

$$X_t = \phi X_{t-1} + Z_t \qquad \text{with } |\phi| < 1 \text{ and } Z_t \sim \mathrm{WN}(0, \sigma^2).$$

The equation system (3.7) becomes:

[3] See Brockwell and Davis (1991, p. 167).

$$\begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \phi^h \\ \phi^{h+1} \\ \phi^{h+2} \\ \vdots \\ \phi^{h+T-1} \end{pmatrix}.$$

The guess-and-verify method immediately leads to the solution:

$$\alpha_T = (a_1, a_2, a_3, \dots, a_T)' = \left(\phi^h, 0, 0, \dots, 0\right)'.$$

We therefore get the following predictor:

$$\mathbb{P}_T X_{T+h} = \phi^h X_T.$$

The forecast therefore just depends on the last observation with the corresponding coefficient $a_1 = \phi^h$ being independent of T. All previous observations can be disregarded, they cannot improve the forecast further. To put it otherwise, all the useful information about X_{T+h} in the entire realization previous to X_T, i.e. in $\{X_T, X_{T-1}, \dots, X_1\}$, is contained in X_T.

The variance of the prediction error is given by

$$v_T(h) = \frac{1 - \phi^{2h}}{1 - \phi^2}\sigma^2.$$

For $h = 1$, the formula simplifies to σ^2 and for $h \to \infty$, $v_T(h) \to \frac{1}{1-\phi^2}\sigma^2$ the unconditional variance of X_t. Note also that the variance of the forecast error $v_T(h)$ increases with h.

The general case of an AR(p) process, $p > 1$, can be treated in the same way. The autocovariances follow a p-th order difference equation (see Sect. 2.4):

$$\gamma(j) = \phi_1\gamma(j-1) + \phi_2\gamma(j-2) + \dots + \phi_p\gamma(j-p).$$

Applying again the guess-and-verify method for the case $h = 1$ and assuming that $T > p$, the solution is given by $\alpha_T = \left(\phi_1, \phi_2, \dots, \phi_p, 0, \dots, 0\right)'$. Thus the one-step ahead predictor is

$$\mathbb{P}_T X_{T+1} = \phi_1 X_T + \phi_2 X_{T-1} + \dots + \phi_p X_{T+1-p}, \qquad T > p. \tag{3.9}$$

The one-step ahead forecast of an AR(p) process therefore depends only on the last p observations.

The above predictor can also be obtained in a different way. View for this purpose $\mathbb{P}_T$ as an *operator* with the following meaning: Take the linear least-squares forecast

with respect to the information $\{X_T, \ldots, X_1\}$. Apply this operator to the defining stochastic difference equation of the AR(p) process forwarded one period:

$$\mathbb{P}_T X_{T+1} = \mathbb{P}_T(\phi_1 X_T) + \mathbb{P}_T(\phi_2 X_{T-1}) + \ldots + \mathbb{P}_T\left(\phi_p X_{T+1-p}\right) + \mathbb{P}_T(Z_{T+1}).$$

In period T observations of $X_T, X_{T-1}, \ldots, X_1$ are known so that $\mathbb{P}_T X_{T-j} = X_{T-j}$, $j = 0, 1, \ldots, T-1$. Because $\{Z_t\}$ is a white noise process and because $\{X_t\}$ is a causal function with respect to $\{Z_t\}$, Z_{T+1} is uncorrelated with $X_T, \ldots, X_1$. This reasoning leads to the same predictor as in Eq. (3.9).

The forecasting functions for $h > 1$ can be obtained recursively by successively applying the forecast operator. Take, for example, the case $h = 2$:

$$\begin{aligned}
\mathbb{P}_T X_{T+2} &= \mathbb{P}_T(\phi_1 X_{T+1}) + \mathbb{P}_T(\phi_2 X_T) + \ldots + \mathbb{P}_T\left(\phi_p X_{T+2-p}\right) + \mathbb{P}_T(Z_{T+2}) \\
&= \phi_1\left(\phi_1 X_T + \phi_2 X_{T-1} + \ldots + \phi_p X_{T+1-p}\right) \\
&\qquad + \phi_2 X_T + \ldots + \phi_p X_{T+2-p} \\
&= \left(\phi_1^2 + \phi_2\right) X_T + \left(\phi_1\phi_2 + \phi_3\right) X_{T-1} + \ldots + \left(\phi_1\phi_{p-1} + \phi_p\right) X_{T+2-p} \\
&\qquad + \phi_1\phi_p X_{T+1-p}.
\end{aligned}$$

In this way forecasting functions for $h > 2$ can be obtained recursively.

Note that in the case of AR(p) processes the coefficient of the forecast function remain constant as long as $T > p$. Thus with each new observation it is not necessary to recompute the equation system and solve it again. This will be different in the case of MA processes. In practice, the parameters of the AR model are usually unknown and have therefore be replaced by some estimate. Section 14.2 investigates in a more general context how this substitution affects the results.

3.1.2 Forecasting with MA(q) Processes

The forecasting problem becomes more complicated in the case of MA(q) processes. In order to get a better understanding we analyze the case of a MA(1) process:

$$X_t = Z_t + \theta Z_{t-1} \qquad \text{with } |\theta| < 1 \text{ and } Z_t \sim \mathrm{WN}(0, \sigma^2).$$

Taking a forecast horizon of one period, i.e. $h = 1$, the equation system (3.7) in the case of a MA(1) process becomes:

$$\begin{pmatrix} 1 & \frac{\theta}{1+\theta^2} & 0 & \ldots & 0 \\ \frac{\theta}{1+\theta^2} & 1 & \frac{\theta}{1+\theta^2} & \ldots & 0 \\ 0 & \frac{\theta}{1+\theta^2} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.10}$$

Despite the fact that the equation system has a simple structure, the forecasting function will depend in general on all past observations of X_{T-j}, $0 \leq j \leq T$. We illustrate this point by a numerical example which will allow us to get a deeper understanding.

Suppose that we know the parameters of the MA(1) process to be $\theta = -0.9$ and $\sigma^2 = 1$. We start the forecasting exercise in period $T = 0$ and assume that, at this point in time, we have no observation at hand. The best forecast is therefore just the unconditional mean which in this example is zero. Thus, $\mathbb{P}_0 X_1 = 0$. The variance of the forecast error then is $\mathbb{V}(X_1 - \mathbb{P}_0 X_1) = v_0(1) = \sigma^2 + \theta^2\sigma^2 = 1.81$. This result is summarized in the first row of Table 3.1. In period 1, the realization of X_1 is observed. This information can be used and the forecasting function becomes $\mathbb{P}_1 X_2 = a_1 X_1$. The coefficient a_1 is found by solving the equation system (3.10) for $T = 1$. This gives $a_1 = \theta/(1 + \theta^2) = -0.4972$. The corresponding variance of the forecast error according to Eq. (3.8) is $\mathbb{V}(X_2 - \mathbb{P}_1 X_2) = v_1(1) = \gamma(0)(1 - \alpha_1'\rho_1(1)) = 1.81(1 - 0.4972 \times 0.4972) = 1.3625$. This value is lower compared to the previous forecast because additional information, the observation of the realization of X_1, is taken into account. Row 2 in Table 3.1 summarizes these results.

In period 2, not only X_1, but also X_2 is observed which allows us to base our forecast on both observations: $\mathbb{P}_2 X_3 = a_1 X_2 + a_2 X_1$. The coefficients can be found by solving the equation system (3.10) for $T = 2$. This amounts to solving the simultaneous equation system

$$\begin{pmatrix} 1 & \frac{\theta}{1+\theta^2} \\ \frac{\theta}{1+\theta^2} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \end{pmatrix}.$$

Inserting $\theta = -0.9$, the solution is $\alpha_2 = (a_1, a_2)' = (-0.6606, -0.3285)'$. The corresponding variance of the forecast error becomes

$$\begin{aligned} \mathbb{V}(X_3 - \mathbb{P}_2 X_3) &= v_2(1) = \gamma(0)(1 - \alpha_2'\rho_2(1)) \\ &= \gamma(0)\left(1 - \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \end{pmatrix}\right) \\ &= 1.81\left(1 - \begin{pmatrix} -0.6606 & -0.3285 \end{pmatrix} \begin{pmatrix} -0.4972 \\ 0 \end{pmatrix}\right) = 1.2155. \end{aligned}$$

These results are summarized in row 3 of Table 3.1.

In period 3, the realizations of X_1, X_2 and X_3 are known so that the forecast function becomes $\mathbb{P}_3 X_4 = a_1 X_3 + a_2 X_2 + a_3 X_1$. The coefficients can again be found by solving the equation system (3.10) for $T = 3$:

$$\begin{pmatrix} 1 & \frac{\theta}{1+\theta^2} & 0 \\ \frac{\theta}{1+\theta^2} & 1 & \frac{\theta}{1+\theta^2} \\ 0 & \frac{\theta}{1+\theta^2} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \\ 0 \end{pmatrix}.$$

Table 3.1 Forecast function for a MA(1) process with $\theta = -0.9$ and $\sigma^2 = 1$

Time	Forecasting function $\alpha_T = (a_1, a_2, \ldots, a_T)'$	Forecast error variance
$T = 0:$		$v_0(1) = 1.8100$
$T = 1:$	$\alpha_1 = (-0.4972)'$	$v_1(1) = 1.3625$
$T = 2:$	$\alpha_2 = (-0.6606, -0.3285)'$	$v_2(1) = 1.2155$
$T = 3:$	$\alpha_3 = (-0.7404, -0.4891, -0.2432)'$	$v_3(1) = 1.1436$
$T = 4:$	$\alpha_4 = (-0.7870, -0.5827, -0.3849, -0.1914)'$	$v_4(1) = 1.1017$
$\ldots$	$\ldots$	$\ldots$
$T = \infty:$	$\alpha_\infty = (-0.9000, -0.8100, -0.7290, \ldots)'$	$v_\infty(1) = 1$

For $\theta = -0.9$, the coefficients of the linear predictor are $\alpha_2 = (a_1, a_2, a_3)' = (-0.7404, -0.4891, -0.2432)'$. The corresponding variance of the forecast error becomes

$$\begin{aligned}
\mathbb{V}(X_4 - \mathbb{P}_3 X_4) = v_3(1) &= \gamma(0)(1 - \alpha_3' \rho_3(1)) \\
&= \gamma(0)\left(1 - \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \\ 0 \end{pmatrix}\right) \\
&= 1.81\left(1 - \begin{pmatrix} -0.7404 & -0.4891 & -0.2432 \end{pmatrix} \begin{pmatrix} -0.4972 \\ 0 \\ 0 \end{pmatrix}\right) \\
&= 1.1436.
\end{aligned}$$

These results are summarized in row 4 of Table 3.1. We can, of course, continue in this way and derive successively the forecast functions for $T = 4, 5, \ldots$.

From this exercise we can make several observations.

- In contrast to the AR process, every new information is used. The forecast $\mathbb{P}_T X_{T+1}$ depends on all available information, in particular on $X_T, X_{T-1}, \ldots, X_1$.
- The coefficients of the forecast function are not constant. They change as more and more information comes in.
- The importance of the new information can be "measured" by the last coefficient of α_T. These coefficients are termed *partial autocorrelations* (see Definition 3.2) and are of particular relevance as will be explained in Sect. 3.5. In our example they are $-0.4972, -0.3285, -0.2432$, and -0.1914.
- As more information becomes available, the variance of the forecast error (mean squared error) declines monotonically. It will converge to $\sigma^2 = 1$. The reason for this result can be explained as follows. Applying the forecasting operator to the defining MA(1) stochastic difference equation forwarded by one period gives: $\mathbb{P}_T X_{T+1} = \mathbb{P}_T Z_{T+1} + \theta \mathbb{P}_T Z_T = \theta \mathbb{P}_T Z_T$ with forecast error $X_{T+1} - \mathbb{P}_T X_{T+1} = Z_{T+1}$. As more and more observation become available, it

becomes better and better possible to recover the "true" value of the unobserved Z_T from the observations $X_T, X_{T-1}, \ldots, X_1$. As the process is invertible, in the limit it is possible to recover the value of Z_T exactly (almost surely). The only uncertainty remaining is with respect to Z_{T+1} which has a mean of zero and a variance of $\sigma^2 = 1$.

3.1.3 Forecasting from the Infinite Past

The forecasting function based on the *infinitely remote past* is of particular theoretical interest. Thereby we look at the problem of finding the optimal linear forecast of X_{T+1} given $X_T, X_{T-1}, \ldots, X_1, X_0, X_{-1}, \ldots$ taking the mean squared error again as the criterion function. The corresponding forecasting function (predictor) will be denoted by $\widetilde{\mathbb{P}}_T X_{T+h}$, $h > 0$.

Noting that the MA(1) process with $|\theta| < 1$ is invertible, we have

$$Z_t = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \ldots$$

We can therefore write X_{t+1} as

$$X_{t+1} = Z_{t+1} + \theta \underbrace{\left(X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \ldots\right)}_{Z_t}$$

The predictor of X_{T+1} from the infinite past, $\widetilde{\mathbb{P}}_T$, is then given by:

$$\widetilde{\mathbb{P}}_T X_{T+1} = \theta \left(X_T - \theta X_{T-1} + \theta^2 X_{T-2} - \ldots\right)$$

where the mean squared forecasting error is

$$v_\infty(1) = \mathbb{E}\left(X_{T+1} - \widetilde{\mathbb{P}}_T X_{T+1}\right)^2 = \sigma^2.$$

Applying this result to our example gives:

$$\widetilde{\mathbb{P}}_T X_{T+1} = -0.9 X_T - 0.81 X_{T-1} - 0.729 X_{T-2} - \ldots$$

with $v_\infty(1) = 1$. See last row in Table 3.1.

Example of an ARMA(1,1) Process

Consider now the case of a causal and invertible ARMA(1,1) process $\{X_t\}$:

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1},$$

where $|\phi| < 1$, $|\theta| < 1$ and $Z_t \sim \text{WN}(0, \sigma^2)$. Because $\{X_t\}$ is causal and invertible with respect to $\{Z_t\}$,

$$X_{T+1} = Z_{T+1} + (\phi + \theta) \sum_{j=0}^{\infty} \phi^j Z_{T-j},$$

$$Z_{T+1} = X_{T+1} - (\phi + \theta) \sum_{j=0}^{\infty} (-\theta)^j X_{T-j}.$$

Applying the forecast operator $\widetilde{\mathbb{P}}_T$ to the second equation and noting that $\widetilde{\mathbb{P}}_T Z_{T+1} = 0$, one obtains the following one-step ahead predictor

$$\widetilde{\mathbb{P}}_T X_{T+1} = (\phi + \theta) \sum_{j=0}^{\infty} (-\theta)^j X_{T-j}.$$

Applying the forecast operator to the first equation, we obtain

$$\widetilde{\mathbb{P}}_T X_{T+1} = (\phi + \theta) \sum_{j=0}^{\infty} (\phi)^j Z_{T-j}.$$

This implies that the one-step ahead prediction error is equal to $X_{T+1} - \widetilde{\mathbb{P}}_T X_{T+1} = Z_{T+1}$ and that the mean squared forecasting error of the one-step ahead predictor given the infinite past is equal to $\mathbb{E} Z_{T+1}^2 = \sigma^2$.

3.2 The Wold Decomposition Theorem

The *Wold Decomposition theorem* is essential for the theoretical understanding of stationary stochastic processes. It shows that *any* stationary process can essentially be represented as a linear combination of current and past forecast errors. Before we can state the theorem precisely, we have to introduce the following definition.

Definition 3.1 (Deterministic Process)**.** *A stationary stochastic process $\{X_t\}$ is called (purely)* deterministic *or (purely)* singular *if and only if it can be forecasted exactly from the infinite past. More precisely, if and only if*

$$\sigma^2 = \mathbb{E}\left(X_{t+1} - \widetilde{\mathbb{P}}_t X_{t+1}\right)^2 = 0 \quad \textit{for all } t \in \mathbb{Z}$$

where $\widetilde{\mathbb{P}}_t X_{t+1}$ denotes the best linear forecast of X_{t+1} given its infinite past, i.e. given $\{X_t, X_{t-1}, \ldots\}$.

The most important class of deterministic processes are the *harmonic processes*. These processes are characterized by finite or infinite sums of sine and cosine functions with stochastic amplitude.[4] A simple example of a harmonic process is given by

[4] More about harmonic processes can be found in Sect. 6.2.

$$X_t = A\cos(\omega t) + B\sin(\omega t) \quad \text{with } \omega \in (0, \pi).$$

Thereby, A and B denote two uncorrelated random variables with mean zero and finite variance. One can check that X_t satisfies the deterministic difference equation

$$X_t = (2\cos\omega)X_{t-1} - X_{t-2}.$$

Thus, X_t can be forecasted exactly from its past. In this example the last two observations are sufficient. We are now in a position to state the Wold Decomposition Theorem.

Theorem 3.1 (Wold Decomposition)**.** *Every stationary stochastic process $\{X_t\}$ with mean zero and finite positive variance can be represented as*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t = \Psi(\mathrm{L})Z_t + V_t, \tag{3.11}$$

where

(i) $Z_t = X_t - \widetilde{\mathbb{P}}_{t-1}X_t = \widetilde{\mathbb{P}}_t Z_t$;
(ii) $Z_t \sim \mathrm{WN}(0, \sigma^2)$ *with* $\sigma^2 = \mathbb{E}\left(X_{t+1} - \widetilde{\mathbb{P}}_t X_{t+1}\right)^2 > 0$;
(iii) $\psi_0 = 1$ *and* $\sum_{j=0}^{\infty} \psi_j^2 < \infty$;
(iv) $\{V_t\}$ *is deterministic;*
(v) $\mathbb{E}(Z_t V_s) = 0$ *for all* $t, s \in \mathbb{Z}$.

The sequences $\{\psi_j\}$, $\{Z_t\}$, and $\{V_t\}$ are uniquely determined by (3.11).

Proof. The proof, although insightful, requires some knowledge about Hilbert spaces which is beyond the scope of this book. A rigorous proof can be found in Brockwell and Davis (1991, Section 5.7).

It is nevertheless instructive to give an intuition of the proof. Following the MA(1) example from the previous section, we start in period 0 and assume that no information is available. Thus, the best forecast $\mathbb{P}_0 X_1$ is zero so that trivially

$$X_1 = X_1 - \mathbb{P}_0 X_1 = Z_1.$$

Starting with $X_1 = Z_1, X_2, X_3, \ldots$ can then be constructed recursively:

$$\begin{aligned}
X_2 &= X_2 - \mathbb{P}_1 X_2 + \mathbb{P}_1 X_2 = Z_2 + a_1^{(1)} X_1 = Z_2 + a_1^{(1)} Z_1 \\
X_3 &= X_3 - \mathbb{P}_2 X_3 + \mathbb{P}_2 X_3 = Z_3 + a_1^{(2)} X_2 + a_2^{(2)} X_1 \\
&= Z_3 + a_1^{(2)} Z_2 + \left(a_1^{(2)} a_1^{(1)} + a_2^{(2)}\right) Z_1 \\
X_4 &= X_4 - \mathbb{P}_3 X_4 + \mathbb{P}_3 X_4 = Z_4 + a_1^{(3)} X_3 + a_2^{(3)} X_2 + a_3^{(3)} X_1
\end{aligned}$$

$$= Z_4 + a_1^{(3)} Z_3 + \left(a_1^{(3)} a_1^{(2)} + a_2^{(3)}\right) Z_2 + \left(a_1^{(3)} a_1^{(2)} a_1^{(1)} + a_1^{(3)} a_2^{(2)} + a_2^{(3)} a_1^{(1)} + a_3^{(3)}\right) Z_1$$

...

where $a_j^{(t-1)}, j = 1, \ldots, t-1$, denote the coefficients of the forecast function for X_t based on $X_{t-1}, \ldots, X_1$. This shows how X_t unfolds into the sum of forecast errors. The stationarity of $\{X_t\}$ ensures that the coefficients of Z_j converge, as t goes to infinity, to ψ_j which are independent of t. □

Every stationary stochastic process is thus representable as the sum of a moving-average of infinite order and a (purely) deterministic process.[5] The weights of the infinite moving average are thereby normalized such that $\psi_0 = 1$. In addition, the coefficients ψ_j are square summable. This property is less strong than absolute summability which is required for a causal representation (see Definition 2.2).[6] The process $\{Z_t\}$ is a white noise process with positive variance $\sigma^2 > 0$. The Z_t's are called *innovations* as they represent the one-period ahead forecast errors based on the infinite past, i.e $Z_t = X_t - \widetilde{\mathbb{P}}_{t-1} X_t$. Z_t is the additional information revealed from the t-th observation. Thus, the Wold Decomposition Theorem serves as a justification for the use of causal ARMA models. In this instance, the deterministic component $\{V_t\}$ vanishes.

The second part of Property (i) further means that the innovation process $\{Z_t\}$ is *fundamental* with respect to $\{X_t\}$, i.e. that Z_t lies in the linear space spanned by $\{X_t, X_{t-1}, X_{t-2}, \ldots\}$ or that $Z_t = \mathbb{P}_t Z_t$. This implies that $\Psi(\mathrm{L})$ must be invertible and that Z_t can be perfectly (almost surely) recovered from the observations of $X_t, X_{t-1}, \ldots$ Finally, property (v) says that the two components $\{Z_t\}$ and $\{V_t\}$ are uncorrelated with each other at all leads and lags. Thus, in essence, the Wold Decomposition Theorem states that every stationary stochastic process can be uniquely decomposed into a weighted sum of current and past forecast errors plus a deterministic process.

Although the Wold Decomposition is very appealing from a theoretical perspective, it is not directly implementable in practice because it requires the estimation of infinitely many parameters $(\psi_1, \psi_2, \ldots)$. This is impossible with only a finite amount of observations. It is therefore necessary to place some assumptions on $(\psi_1, \psi_2, \ldots)$. One possibility is to assume that $\{X_t\}$ is a causal ARMA process and

[5] The Wold Decomposition corresponds to the decomposition of the spectral distribution function of F into the sum of F_Z and F_V (see Sect. 6.2). Thereby the spectral distribution function F_Z has spectral density $f_Z(\lambda) = \frac{\sigma^2}{2\pi} |\Psi(e^{-\imath\lambda})|^2$.

[6] The series $\psi_j = 1/j$, for example, is square summable, but not absolutely summable.

to recover the ψ_j's from the causal representation. This amounts to say that $\Psi(\mathrm{L})$ is a rational polynomial which means that

$$\Psi(\mathrm{L}) = \frac{\Theta(\mathrm{L})}{\Phi(\mathrm{L})} = \frac{1 + \theta_1 \mathrm{L} + \theta_2 \mathrm{L}^2 + \ldots + \theta_q \mathrm{L}^q}{1 - \phi_1 \mathrm{L} - \phi_2 \mathrm{L}^2 - \ldots - \phi_p \mathrm{L}^p}.$$

Thus, the process is characterized by only a finite number, $p + q$, of parameters. Another possibility is to place restrictions on the smoothness of the spectrum (see Chap. 6).

The Wold Decomposition Theorem has several implications which are presented in the following remarks.

Remark 3.1. In the case of ARMA processes, the purely deterministic part $\{V_t\}$ can be disregarded so that the process is represented only by a weighted sum of current and past innovations. Processes with this property are called *purely non-deterministic*, *linearly regular*, or *regular* for short. Moreover, it can be shown that every regular process $\{X_t\}$ can be approximated arbitrarily well by an ARMA process $\{X_t^{(\mathrm{ARMA})}\}$ meaning that

$$\sup_{t \in \mathbb{Z}} \mathbb{E}\left(X_t - X_t^{(\mathrm{ARMA})}\right)^2$$

can be made arbitrarily small. The proof of these results can be found in Hannan and Deistler (1988, Chapter 1).

Remark 3.2. The process $\{Z_t\}$ is white noise, but not necessarily Gaussian. In particular, $\{Z_t\}$ need not be independently and identically distributed (IID). Thus, $\mathbb{E}(Z_{t+1}|X_t, X_{t-1}, \ldots)$ need not be equal to zero although $\widetilde{\mathbb{P}}_t Z_{t+1} = 0$. The reason is that $\widetilde{\mathbb{P}}_t Z_{t+1}$ is only the best linear forecast function, whereas $\mathbb{E}(Z_{t+1}|X_t, X_{t-1}, \ldots)$ is the best forecast function among all linear and non-linear functions. Examples of processes which are white noise, but not IID, are GARCH processes discussed in Chap. 8.

Remark 3.3. The innovations $\{Z_t\}$ may not correspond to the "true" shocks of the underlying economic system. In this case, the shocks to the economic system cannot be recovered from the Wold Decomposition. Thus, they are not fundamental with respect to $\{X_t\}$. Suppose, as a simple example, that $\{X_t\}$ is generated by a noninvertible MA(1) process:

$$X_t = U_t + \theta U_{t-1}, \qquad U_t \sim \mathrm{WN}(0, \sigma^2) \quad \text{and } |\theta| > 1.$$

This generates an impulse response function with respect to the true shocks of the system equal to $(1, \theta, 0, \ldots)$. The above mechanism can, however, not be the Wold Decomposition because the noninvertibility implies that U_t cannot be recovered from the observation of $\{X_t\}$. As shown in the introduction, there is an

observationally equivalent MA(1) process, i.e. a process which generates the same ACF. Based on the computation in Sect. 1.5, this MA(1) process is

$$X_t = Z_t + \tilde{\theta} Z_{t-1}, \qquad Z_t \sim \mathrm{WN}(0, \tilde{\sigma}^2),$$

with $\tilde{\theta} = \theta^{-1}$ and $\tilde{\sigma}^2 = \frac{1+\theta^2}{1+\theta^{-2}}\sigma^2$. This is already the Wold Decomposition. The impulse response function for this process is $(1, \theta^{-1}, 0, \ldots)$ which is different from the original system. As $|\tilde{\theta}| = |\theta^{-1}| < 1$, the innovations $\{Z_t\}$ can be recovered from the observations as $Z_t = \sum_{j=0}^{\infty} (-\tilde{\theta})^j X_{t-j}$, but they do not correspond to the shocks of the system $\{U_t\}$. Hansen and Sargent (1991), Quah (1990), and Lippi and Reichlin (1993) among others provide a deeper discussion and present additional more interesting economic examples.

3.3 Exponential Smoothing

Besides the method of least-squares forecasting *exponential smoothing* can often be seen as a valid alternative. This method views X_t as a function of time:

$$X_t = f(t; \beta) + \varepsilon_t,$$

whereby $f(t; \beta)$ typically represents a polynomial in t with coefficients β. The above equation is similar to a regression model with error term ε_t. This error term is usually specified as a white noise process $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$.

Consider first the simplest case where X_t just moves randomly around a fixed mean β. This corresponds to the case where $f(t; \beta)$ is a polynomial of degree zero:

$$X_t = \beta + \varepsilon_t.$$

If β is known then $\mathbb{P}_T X_{T+h}$, the forecast of X_{T+h} given the observations $X_T, \ldots, X_1$, clearly is β. If, however, β is unknown, we can substitute β by $\overline{X}_T$, the average of the observations:

$$\widehat{\mathbb{P}}_T X_{T+h} = \hat{\beta} = \overline{X}_T = \frac{1}{T} \sum_{t=1}^{T} X_t,$$

where "$\hat{}$" means that the model parameter β has been replaced by its estimate. The one-period ahead forecast function can then be rewritten as follows:

$$\begin{aligned} \widehat{\mathbb{P}}_T X_{T+1} &= \frac{T-1}{T} \widehat{\mathbb{P}}_{T-1} X_T + \frac{1}{T} X_T \\ &= \widehat{\mathbb{P}}_{T-1} X_T + \frac{1}{T} \left(X_T - \widehat{\mathbb{P}}_{T-1} X_T \right). \end{aligned}$$

The first equation represents the forecast for $T+1$ as a linear combination of the forecast for T and of the last additional information, i.e. the last observation. The weight given to the last observation is equal to $1/T$ because we assumed that the mean remains constant and because the contribution of one observation to the mean is $1/T$. The second equation represents the forecast for $T+1$ as the forecast for T plus a correction term which is proportional to the last forecast error. One advantage of this second representation is that the computation of the new forecast, i.e. the forecast for $T+1$, only depends on the forecast for T and the additional observation. In this way the storage requirements are minimized.

In many applications, the mean does not remain constant, but is a slowly moving function of time. In this case it is no longer meaningful to give each observation the same weight. Instead, it seems plausible to weigh the more recent observation higher than the older ones. A simple idea is to let the weights decline exponentially which leads to the following forecast function:

$$\mathbb{P}_T X_{T+1} = \frac{1-\omega}{1-\omega^T} \sum_{t=0}^{T-1} \omega^t X_{T-t} \qquad \text{with } |\omega| < 1.$$

ω thereby acts like a discount factor which controls the rate at which agents forget information. $1-\omega$ is often called the smoothing parameter. The value of ω should depend on the speed at which the mean changes. In case when the mean changes only slowly, ω should be large so that all observations are almost equally weighted; in case when the mean changes rapidly, ω should be small so that only the most recent observations are taken into account. The normalizing constant $\frac{1-\omega}{1-\omega^T}$ ensures that the weights sum up to one. For large T the term ω^T can be disregarded so that one obtains the following forecasting function based on *simple exponential smoothing*:

$$\begin{aligned}
\mathbb{P}_T X_{T+1} &= (1-\omega)\left[X_T + \omega X_{T-1} + \omega^2 X_{T-2} + \ldots\right] \\
&= (1-\omega)X_T + \omega\, \mathbb{P}_{T-1} X_T \\
&= \mathbb{P}_{T-1} X_T + (1-\omega)\left(X_T - \mathbb{P}_{T-1} X_T\right).
\end{aligned}$$

In the economics literature this forecasting method is called *adaptive expectation*. Similar to the model with constant mean, the new forecast is a weighted average between the old forecast and the last (newest) observation, respectively between the previous forecast and a term proportional to the last forecast error.

One important advantage of adaptive forecasting methods is that they can be computed recursively. Starting with value S_0, the following values can be computed as follows:

$$\begin{aligned}
\mathbb{P}_0 X_1 &= S_0 \\
\mathbb{P}_1 X_2 &= \omega \mathbb{P}_0 X_1 + (1-\omega) X_1
\end{aligned}$$

$$\mathbb{P}_2 X_3 = \omega \mathbb{P}_1 X_2 + (1-\omega) X_2$$
$$\dots$$
$$\mathbb{P}_T X_{T+1} = \omega \mathbb{P}_{T-1} X_T + (1-\omega) X_T.$$

Thereby S_0 has to be determined. Because

$$\mathbb{P}_T X_{T+1} = (1-\omega)\left[X_T + \omega X_{T-1} + \dots + \omega^{T-1}, X_1\right] + \omega^T S_0,$$

the effect of the starting value declines exponentially with time. In practice, we can take $S_0 = X_1$ or $S_0 = \overline{X}_T$. The discount factor ω is usually set a priori to be a number between 0.7 and 0.95. It is, however, possible to determine ω optimally by choosing a value which minimizes the mean squared one-period forecast error:

$$\sum_{t=1}^{T} (X_t - \mathbb{P}_{t-1} X_t)^2 \longrightarrow \min_{|\omega|<1}.$$

From a theoretical perspective one can ask the question for which class of models exponential smoothing represents the optimal procedure. Muth (1960) showed that this class of models is given by

$$\Delta X_t = X_t - X_{t-1} = Z_t - \omega Z_{t-1}.$$

Note that the process generated by the above equation is no longer stationary. This has to be expected as the exponential smoothing assumes a non-constant mean. Despite the fact that this class seems rather restrictive at first, practice has shown that it delivers reasonable forecasts, especially in situations when it becomes costly to specify a particular model.[7] Additional results and more general exponential smoothing methods can be found in Abraham and Ledolter (1983) and Mertens and Rässler (2005).

3.4 Exercises

Exercise 3.4.1. *Compute the linear least-squares predictor $\mathbb{P}_T X_{T+h}$, $T > 2$, and the mean squared error $v_T(h)$, $h = 1, 2, 3$, if $\{X_t\}$ is given by the AR(2) process*

$$X_t = 1.3X_{t-1} - 0.4X_{t-2} + Z_t \qquad \textit{with} \quad Z_t \sim \mathrm{WN}(0, 2).$$

To which values do $\mathbb{P}_T X_{T+h}$ and $v_T(h)$ converge for h going to infinity?

[7] This happens, for example, when many, perhaps thousands of time series have to be forecasted in a real time situation.

Exercise 3.4.2. *Compute the linear least-squares predictor $\mathbb{P}_T(X_{T+1})$ and the mean squared error $v_T(1)$, $T = 0, 1, 2, 3$, if $\{X_t\}$ is given by the MA(1) process*

$$X_t = Z_t + 0.8Z_{t-1} \qquad \textit{with} \quad Z_t \sim \mathrm{WN}(0, 2).$$

To which values do $\mathbb{P}_T X_{T+h}$ and $v_T(h)$ converge for h going to infinity?

Exercise 3.4.3. *Suppose that you observe $\{X_t\}$ for the two periods $t = 1$ and $t = 3$, but not for $t = 2$.*

(i) *Compute the linear least-squares forecast for X_2 if*

$$X_t = \phi X_{t-1} + Z_t \qquad \textit{with} \quad |\phi| < 1 \quad \textit{and} \quad Z_t \sim \mathrm{WN}(0, 4)$$

Compute the mean squared error for this forecast.

(ii) *Assume now that $\{X_t\}$ is the MA(1) process*

$$X_t = Z_t + \theta Z_{t-1} \qquad \textit{with} \quad Z_t \sim \mathrm{WN}(0, 4).$$

Compute the mean squared error for the forecast of X_2.

Exercise 3.4.4. *Let*

$$X_t = A\cos(\omega t) + B\sin(\omega t)$$

with A and B being two uncorrelated random variables with mean zero and finite variance. Show that $\{X_t\}$ satisfies the deterministic difference equation:

$$X_t = (2\cos\omega)X_{t-1} - X_{t-2}.$$

3.5 The Partial Autocorrelation Function

Consider again the problem of forecasting X_{T+1} from observations $X_T, X_{T-1}, \ldots, X_2, X_1$. Denoting, as before, the best linear predictor by $\mathbb{P}_T X_{T+1} = a_1 X_T + a_2 X_{T-1} + a_{T-1} X_2 + a_T X_1$, we can express X_{T+1} as

$$X_{T+1} = \mathbb{P}_T X_{T+1} + Z_{T+1} = a_1 X_T + a_2 X_{T-1} + a_{T-1} X_2 + a_T X_1 + Z_{T+1}$$

where Z_{T+1} denotes the forecast error which is uncorrelated with $X_T, \ldots, X_1$. We can now ask the question whether X_1 contributes to the forecast of X_{T+1} *controlling* for $X_T, X_{T-2}, \ldots, X_2$ or, equivalently, whether a_T is equal to zero. Thus, a_T can be viewed as a measure of the importance of the additional information provided by X_1. It is referred to as the *partial autocorrelation*. In the case of an AR(p) process, the whole information useful for forecasting X_{T+1}, $T > p$, is incorporated in the

last p observations so that $a_T = 0$. In the case of the MA process, the observations on $X_T, \ldots, X_1$ can be used to retrieve the unobserved $Z_T, Z_{T-1} \ldots, Z_{t-q+1}$. As Z_t is an infinite weighted sum of past X_t's, every new observation contributes to the recovering of the Z_t's. Thus, the partial autocorrelation a_T is not zero. Taking T successively equal to 0, 1, 2, etc. we get the partial autocorrelation function (PACF).

We can, however, interpret the above equation as a regression equation. From the Frisch-Lovell-Waugh Theorem (See Davidson and MacKinnon 1993), we can obtain a_T by a two-stage procedure. Project (regress) in a first stage X_{T+1} on $X_T, \ldots, X_2$ and take the residual. Similarly, project (regress) X_1 on $X_T, \ldots, X_2$ and take the residual. The coefficient a_T is then obtained by projecting (regressing) the first residual on the second. Stationarity implies that this is nothing but the correlation coefficient between the two residuals.

3.5.1 Definition

The above intuition suggests two equivalent definitions of the partial autocorrelation function (PACF).

Definition 3.2 (Partial Autocorrelation Function I)**.** *The* partial autocorrelation function *(PACF)* $\alpha(h), h = 0, 1, 2, \ldots$*, of a stationary process is defined as follows:*

$$\begin{aligned} \alpha(0) &= 1 \\ \alpha(h) &= a_h, \qquad h = 1, 2, \ldots, \end{aligned}$$

where a_h *denotes the last element of the vector* $\alpha_h = \Gamma_h^{-1}\gamma_h(1) = R_h^{-1}\rho_h(1)$ *(see Sect. 3.1 and Eq. (3.7)).*

Definition 3.3 (Partial Autocorrelation Function II)**.** *The* partial autocorrelation function *(PACF)* $\alpha(h), h = 0, 1, 2, \ldots$*, of a stationary process is defined as follows:*

$$\begin{aligned} \alpha(0) &= 1 \\ \alpha(1) &= \operatorname{corr}(X_2, X_1) = \rho(1) \\ \alpha(h) &= \operatorname{corr}\left[X_{h+1} - \mathbb{P}\left(X_{h+1}|1, X_2, \ldots, X_h\right), X_1 - \mathbb{P}\left(X_1|1, X_2, \ldots, X_h\right)\right], \end{aligned}$$

where $\mathbb{P}\left(X_{h+1}|1, X_2, \ldots, X_h\right)$ *and* $\mathbb{P}\left(X_1|1, X_2, \ldots, X_h\right)$ *denote the best, in the sense mean squared forecast errors, linear forecasts of* X_{h+1}*, respectively* X_1 *given* $\{1, X_2, \ldots, X_h\}$.

Remark 3.4. If $\{X_t\}$ has a mean of zero, then the constant in the projection operator can be omitted.

The first definition implies that the partial autocorrelations are determined from the coefficients of the forecasting function which are themselves functions

of the autocorrelation coefficients. It is therefore possible to express the partial autocorrelations as a function of the autocorrelations. More specifically, the partial autocorrelation functions can be computed recursively from the autocorrelation function according to the Durbin-Levinson algorithm (Durbin 1960):

$$\begin{aligned}
\alpha(0) &= 1 \\
\alpha(1) &= a_{11} = \rho(1) \\
\alpha(2) &= a_{22} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} \\
&\dots \\
\alpha(h) &= a_{hh} = \frac{\rho(h) - \sum_{j=1}^{h-1} a_{h-1,j}\rho_{h-j}}{1 - \sum_{j=1}^{h-1} a_{h-1,j}\rho_j},
\end{aligned}$$

where $a_{h,j} = a_{h-1,j} - a_{hh}a_{h-1,h-j}$ for $j = 1, 2, \dots, h-1$.

Autoregressive Processes

The idea of the PACF can be well illustrated in the case of an AR(1) process

$$X_t = \phi X_{t-1} + Z_t \qquad \text{with } 0 < |\phi| < 1 \text{ and } Z_t \sim \text{WN}(0, \sigma^2).$$

As shown in Chap. 2, X_t and X_{t-2} are correlated with each other despite the fact that there is no direct relationship between the two. The correlation is obtained "indirectly" because X_t is correlated with X_{t-1} which is itself correlated with X_{t-2}. Because both correlation are equal to ϕ, the correlation between X_t and X_{t-2} is equal to $\rho(2) = \phi^2$. The ACF therefore accounts for all correlation, including the indirect ones. The partial autocorrelation on the other hand only accounts for the direct relationships. In the case of the AR(1) process, there is only an indirect relation between X_t and X_{t-h} for $h \geq 2$, thus the PACF is zero.

Based on the results in Sect. 3.1 for the AR(1) process, the definition 3.2 of the PACF implies:

$$\begin{aligned}
\alpha_1 &= \phi && \Rightarrow \alpha(1) = \rho(1) = \phi, \\
\alpha_2 &= (\phi, 0)' && \Rightarrow \alpha(2) = 0, \\
\alpha_3 &= (\phi, 0, 0)' && \Rightarrow \alpha(3) = 0.
\end{aligned}$$

The partial autocorrelation function of an AR(1) process is therefore equal to zero for $h \geq 2$.

This logic can be easily generalized. The PACF of a causal AR(p) process is equal to zero for $h > p$, i.e. $\alpha(h) = 0$ for $h > p$. This property characterizes an AR(p) process as shown in the next section.

Moving-Average Processes

Consider now the case of an invertible MA process. For this process we have:

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \qquad \Rightarrow \qquad X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + Z_t.$$

X_t is therefore "directly" correlated with each X_{t-h}, $h = 1, 2, \ldots$. Consequently, the PACF is never exactly equal to zero, but converges exponentially to zero. This convergence can be monotonic or oscillating.

Take the MA(1) process as an illustration:

$$X_t = Z_t + \theta Z_{t-1} \qquad \text{with } |\theta| < 1 \text{ and } Z_t \sim \text{WN}(0, \sigma^2).$$

The computations in Sect. 3.1.2 showed that

$$\alpha_1 = \frac{\theta}{1+\theta^2} \qquad \Rightarrow \alpha(1) = \rho(1) = \frac{\theta}{1+\theta^2},$$

$$\alpha_2 = \left(\frac{\theta(1+\theta^2)}{1+\theta^2+\theta^4}, \frac{-\theta^2}{1+\theta^2+\theta^4} \right)' \qquad \Rightarrow \alpha(2) = \frac{-\theta^2}{1+\theta^2+\theta^4}.$$

Thus we get for the MA(1) process:

$$\alpha(h) = -\frac{(-\theta)^h}{1+\theta^2+\ldots+\theta^{2h}} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}$$

3.5.2 Interpretation of ACF and PACF

The ACF and the PACF are two important tools to determining the nature of the underlying mechanism of a stochastic process. In particular, they can be used to determine the orders of the underlying AR, respectively MA processes. The analysis of ACF and PACF to identify appropriate models is know as the Box-Jenkins methodology (Box and Jenkins 1976). Table 3.2 summarizes the properties of both tools for the case of a causal AR and an invertible MA process.

If $\{X_t\}$ is a causal and invertible ARMA(p,q) process, we have the following properties. As shown in Sect. 2.4, the ACF is characterized for $h > \max\{p, q+1\}$ by the homogeneous difference equation $\rho(h) = \phi_1\rho(h-1)+\ldots+\phi_p\rho(h-p)$. Causality implies that the roots of the characteristic equation are all inside the unit circle. The autocorrelation coefficients therefore decline exponentially to zero. Whether this convergence is monotonic or oscillating depends on the signs of the roots. The PACF starts to decline to zero for $h > p$. Thereby the coefficients of the PACF exhibit the same behavior as the autocorrelation coefficients of $\theta^{-1}(\text{L})X_t$.

Table 3.2 Properties of the ACF and the PACF

Processes	ACF	PACF
AR(p)	Declines exponentially (monotonically or oscillating) to zero	$\alpha(h) = 0$ for $h > p$
MA(q)	$\rho(h) = 0$ for $h > q$	Declines exponentially (monotonically or oscillating) to zero

3.6 Exercises

Exercise 3.6.1. *Assign the ACF and the PACF from Fig. 3.1 to the following processes:*

$$
\begin{aligned}
X_t &= Z_t, \\
X_t &= 0.9X_{t-1} + Z_t, \\
X_t &= Z_t + 0.8Z_{t-1}, \\
X_t &= 0.9X_{t-1} + Z_t + 0.8Z_{t-1}
\end{aligned}
$$

with $Z_t \sim \mathrm{WN}(0, \sigma^2)$.

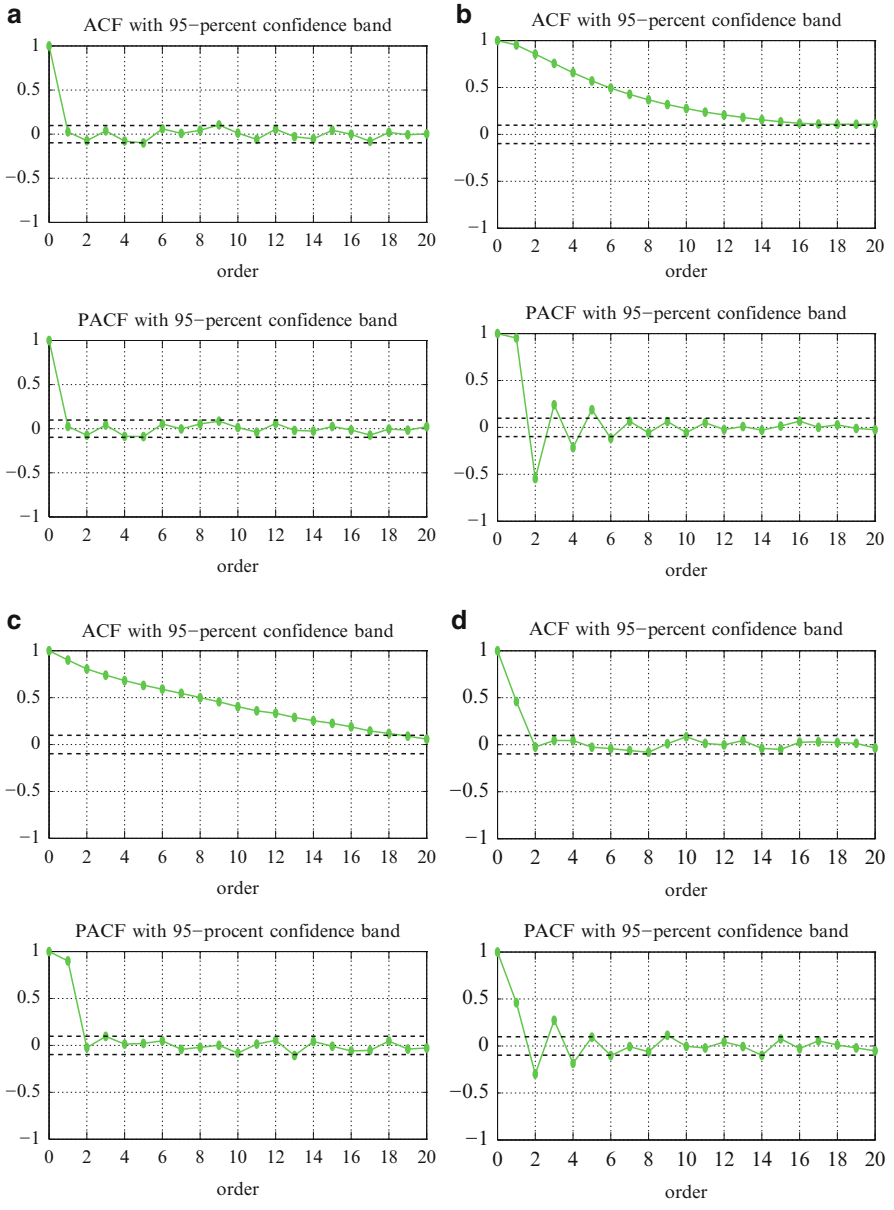

Fig. 3.1 Autocorrelation and partial autocorrelation functions. (**a**) Process 1. (**b**) Process 2. (**c**) Process 3. (**d**) Process 4