32

Time Series Modeling

Time series analysis aims to develop a *model*, which describes the time series in all its measurable features. This goes far beyond than merely determining statistical parameters from observed time series data (such as the variance, correlation, etc.) as described in Chap. 31. Estimators such as those appearing in Eq. 31.5 are examples of how *parameters* can be estimated which are subsequently used to *model* the stochastic process governing the time series (for example, a random walk with drift μ and volatility σ). To develop a model that is capable of simulation a time series with similar features is the principle goal of time series analysis. The object is thus to interpret a series of observed data points $\{X_t\}$, for example a historical price or volatility evolution (in this way acquiring a fundamental understanding of the process) and to *model* the processes underlying the observed historical evolution. In this sense, the historical sequence of data points is interpreted as just one *realization* of the time series process. The parameters of the process are then estimated from the available data and can subsequently be used in making *forecasts*, for example.

As much structure as possible should be extracted from a given data sequence and then transferred to the model. Let $\{\widehat{X}_t\}$ be the time series generated by the model process (called the *estimated* time series). The difference between this and the actually observed data points $\{X_t\}$ are called *residues* $\{X_t - \widehat{X}_t\}$. These should consist of only "noise", i.e., they should be unpredictable random numbers.

In order to be able to fit a time series model, the "raw data", i.e., the sequence of historical data points, must sometimes undergo a *pre-treatment*. In this procedure, *trends* and *seasonal components* are first eliminated and a change may

H.-P. Deutsch, M. W. Beinker, *Derivatives and Internal Models*, Finance and Capital Markets Series, https://doi.org/10.1007/978-3-030-22899-6_32

be made to the scale of the data, so that the resulting sequence is a *stationary time series*.[1] A stationary time series is characterized by the *time* invariance of its expectation, variance and covariance. In particular, the expectation and variance are constant. Without loss of generality, the expectation can be assumed to be zero since it can be eliminated during the pre-treatment through a *centering of the time series.* This is accomplished by subtracting the mean $\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$ from every data point in the time series $\{X_t\}$.

As just discussed, the *stationarity* of a time series implies $\mathrm{E}[X_t] = \mathrm{E}[X]\ \forall t$ and the autocovariance Eq. 31.14 becomes

$$\mathrm{cov}(X_{t+h}, X_t) = \mathrm{E}[X_{t+h}\, X_t] - \mathrm{E}[X]\,\mathrm{E}[X] = \mathrm{E}[X_{t+h}\, X_t]\ . \tag{32.1}$$

The final equality in the above equation holds if the time series has been centered in the pre-treatment. We will always assume this to be the case. Furthermore, the autocovariance and autocorrelation (just as the variance) are independent of t if the time series is stationary, and therefore depend only on the *time lag* h. We frequently write

$$\gamma(h) := \mathrm{cov}(X_{t+h}, X_t)$$

Likewise, if the time series is stationary we have $\varrho(t, h) = \varrho(h)$ in the autocorrelation Eq. 31.13. The following useful symmetry relations can be derived directly from the stationarity of the time series (this can be shown by substituting t with $t' = t - h$):

$$\gamma(-h) = \gamma(h) \quad , \qquad \varrho(-h) = \varrho(h)\ . \tag{32.2}$$

From definition 31.14, we can immediately obtain an estimate[2] for the autocorrelation and the autocovariance of a stationary data sequence

$$\widehat{\gamma}(h) = \widehat{\mathrm{cov}}(X_{t+h}, X_t) = \frac{1}{T}\sum_{t=1}^{T-h}(X_{t+h} - \langle X\rangle)(X_t - \langle X\rangle) \quad , \qquad \widehat{\varrho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \tag{32.3}$$

[1] More precisely, we are dealing with a *weakly stationary* time series in what follows.

[2] In the following material, we will distinguish the estimator of a parameter from the parameter itself with a "hat" notation.

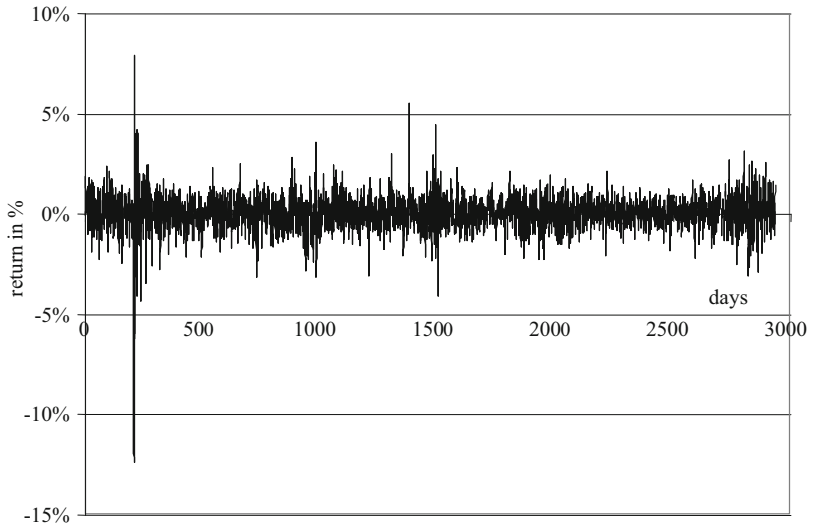

Fig. 32.1 Daily returns of the FTSE index as an example of a stationary data series. The crash in October 1987 is clearly visible

for $h \in \mathbb{N}_0$. The autocovariances (and autocorrelations) are usually computed for at most $h \leq 40$. Note that h has to be substantially smaller than T in all cases; the estimation is otherwise too inexact.[3]

Of course, we can fit *different* time series models to a stationary time series (after having undergone a pre-treatment if necessary) and then compare their goodness of fit and forecasting performance. Thus the following three general steps must be taken when modeling a given sequence of data points

1. Pre-treatment of the data sequence to generate a stationary series (elimination of trend and seasonal components, transformation of scale, etc.).
2. Estimation and/or fitting of the time series model and its parameters.
3. Evaluation of the goodness of fit and forecasting performance on the basis of which a decision is made as to whether the tested model should be accepted or a new model selected (step 2).

Figure 32.1 shows the daily relative change (returns) of the FTSE Index taken from the daily data from Jan-01-1987 through Apr-01-1998 (2,935 days).

[3]The fact that only T appears in the denominator in Eq. 32.3 instead of $T - h$, as one might expect, guarantees that the estimator for the covariance matrix $[\hat{\gamma}(i-j)]_{i,j=1}^{T}$ is automatically positive definite.

This sequence of data points is defined as

$$X_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}} , \tag{32.4}$$

where $\{Y_t\}$ represents the original data sequence of FTSE values. The data set $\{X_t\}$ consists of 2,934 values. According to Eq. 30.9, the relative changes in Eq. 32.4 are approximately equal to the difference of the logarithms if the daily changes are sufficiently small:

$$X_t \approx \ln(Y_t) - \ln(Y_{t-1}) . \tag{32.5}$$

This is the first difference of the logarithm of the original sequence of FTSE index values. The above example represents a typical pre-treatment procedure performed on the data. Instead of the original data $\{Y_t\}$, which is by no means stationary (drift$\neq 0$ and variance increase with time as $\sim \sigma t$), we generate a stationary data sequence as in Eq. 32.5 through standard transformations in time series analysis. Specifically in our case, what is known as *Box-Cox scaling* (taking the logarithm of the original data) was performed and subsequently the first differences were calculated for the purpose of trend elimination. Stationary time series data like these are then used in the further analysis, in particular, when fitting a model to the data.

The above example should provide sufficient motivation for the pre-treatment of a time series. The interested reader is referred to Chap. 35 for further discussion of pre-treating time series data to generate stationary time series. We will assume from now on that the given time series have already been pre-treated, i.e., potential trends and seasonal components have already been eliminated and scaling transformations have already been performed appropriately, so that the resulting data sequences are stationary. Such a stationary time series is given by a sequence of random variables $\{X_t\}$, $t \in \mathbb{N}$.

32.1 Stationary Time Series and Autoregressive Models

This chapter introduces a basic approach in time series analysis employing a specific time series model, called autoregressive model. We then continue by extending the results to the case of a time-dependent variance (GARCH model) which finds application in modeling *volatility clustering* in financial

time series. This technique is widely used in modeling the *time evolution* of volatilities.

Rather than working under the idealized assumption of time-continuous processes, the processes modeled in this chapter are truly discrete in time. The discussion is geared to the needs of the user. We will forgo mathematical rigor and in most cases the proofs of results will not be given. Not taking these "shortcuts" would increase the expanse of this chapter considerably. However, the attempt will be made to provide thorough reasoning for all results presented.

A process for modeling a time series of stock prices, for example, has already been encountered in this text: the random walk. An important property of the random walk is the *Markov property*. Recall that the Markov property states that the next step in a random walk depends solely on its current value, but not on the values taken on at any previous times. If such a Markov process is unsatisfactory for modeling the properties of the time series under consideration, an obvious generalization would be to allow for the influence of past values of the process. Processes whose current values can be affected by values attained in the past are called *autoregressive*. In order to characterize these processes, we must first distinguish between the *unconditional* and *conditional* variance denoted by $\mathrm{var}[X_t]$ and $\mathrm{var}[X_t|X_{t-1}, \ldots, X_1]$, respectively. The *un*conditional variance is the variance we are familiar with from previous chapters, whereas the conditional variance is the variance of X_t *under the condition* that $X_{t-1}, \ldots, X_1$ have occurred. Analogously, we must differentiate between the *conditional* and *unconditional* expectation denoted by $\mathrm{E}[X_t]$ and $\mathrm{E}[X_t|X_{t-1}, \ldots, X_1]$, respectively, where the last is the expectation of X_t *under the condition* that $X_{t-1}, \ldots, X_1$ have occurred. There is no difference between the two when the process under consideration is independent of its history.

32.1.1 AR(p) Processes

Having made these preparatory remarks and definitions, we now want to consider processes whose current values are influenced by one or more of their predecessors. If, for example, the effect of the p previous values of a time series on the current value is *linear*, the process is referred to as an *autoregressive process of order* p, and denoted by AR(p). The general autoregressive process of pth order makes use of p process values in the past to generate a representation of

today's value, or explicitly

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots \phi_p X_{t-p} + \varepsilon_t \tag{32.6}$$
$$= \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t \,, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2) \,.$$

The changes ε_t here are independent of all previous time series values X_s, $s < t$, and thus represent an injection of truly new information into the process[4]. In particular, this means that $\mathrm{cov}[X_i, \varepsilon_j]$ is always zero. The conditional variance and conditional expectation of the process are

$$\mathrm{E}[X_t | X_{t-1}, \ldots, X_1] = \sum_{i=1}^{p} \phi_i X_{t-i} \tag{32.7}$$
$$\mathrm{var}[X_t | X_{t-1}, \ldots, X_1] = \mathrm{var}[\varepsilon_t] = \sigma^2 \,.$$

It can be shown that stationarity is guaranteed if the zeros z_k of the *characteristic polynomial*[5]

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \tag{32.8}$$

lie outside of the closed unit disk, i.e., when the norm $|z_k|$ is larger than 1 for all zeros z_k. In particular, if the process is stationary then the *un*conditional expectation and variance have the following properties: $\mathrm{E}[X_t] = \mathrm{E}[X_{t-i}]$ and $\mathrm{var}[X_t] = \mathrm{var}[X_{t-i}]$. Exploiting this, we can easily calculate explicit expressions for the *un*conditional expectation and variance. The unconditional expectation $\mathrm{E}[X_t]$ is

$$\mathrm{E}[X_t] = \mathrm{E}[\sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t] = \sum_{i=1}^{p} \phi_i \underbrace{\mathrm{E}[X_{t-i}]}_{\mathrm{E}[X_t]} + \underbrace{\mathrm{E}[\varepsilon_t]}_{0} = \mathrm{E}[X_t] \sum_{i=1}^{p} \phi_i \,.$$

In the first step we have simply used definition 32.6 for X_t. The second step is merely the linearity of the expectation operator. In the third step we have finally used the decisive properties of the process, namely stationarity of the

[4]The notation ε_t will always indicate independent, identically $\mathrm{N}(0, \sigma^2)$-distributed random variables. Another common definition is $\varepsilon_t \sim \mathrm{W}(0, \sigma^2)$, where W stands for *white noise*. This is a somewhat more general statement and is used in reference to random variables which are not normally distributed as well.

[5]This polynomial plays a central role in the theory of time series.

expectation and randomness of the residues. The result is therefore

$$\mathrm{E}[X_t]\left(1-\sum_{i=1}^{p}\phi_i\right)=0$$

This implies that the unconditional expectation must be zero since stationarity guarantees that the sum of the ϕ_i is *not* equal to one.[6]

The unconditional variance can be computed using similar arguments

$$\begin{aligned}\mathrm{var}[X_t] &= \mathrm{var}[\sum_{i=1}^{p}\phi_i X_{t-i}\ +\varepsilon_t]\\ &= \sum_{i,j=1}^{p}\phi_i\phi_j\mathrm{cov}[X_{t-i},X_{t-j}]+\sum_{i=1}^{p}\phi_i\mathrm{cov}[X_{t-i},\varepsilon_t]+\mathrm{var}[\varepsilon_t]\\ &= \sum_{i,j=1}^{p}\phi_i\phi_j\mathrm{cov}[X_{t-i},X_{t-j}]+0+\sigma^2\\ &= \mathrm{var}[X_t]\sum_{i,j=1}^{p}\phi_i\phi_j\varrho(i-j)+\sigma^2\ ,\end{aligned}$$

where we used Eq. A.17 and—in the last step—definition 31.13 for stationary processes. Solving for $\mathrm{var}[X_t]$ yields immediately

$$\mathrm{var}[X_t]=\frac{\sigma^2}{1-\sum_{i,j=1}^{p}\phi_i\phi_j\varrho(i-j)}\ . \tag{32.9}$$

An expression for the autocorrelation function ϱ of the process can be obtained by multiplying both sides of Eq. 32.6 by X_{t-h} and taking the expectation. Here stationarity is used in form of Eqs. 32.2 and 32.1:

$$\begin{aligned}\varrho(h)=\varrho(-h)&=\frac{\mathrm{cov}(X_{t-h},X_t)}{\mathrm{cov}(X_t,X_t)}=\frac{\mathrm{E}(X_{t-h},X_t)}{\mathrm{E}(X_t^2)}\\ &=\frac{1}{\mathrm{E}(X_t^2)}\mathrm{E}(X_{t-h},\sum_{i=1}^{p}\phi_i X_{t-i}\ +\varepsilon_t)\end{aligned}$$

[6] This can be shown using the characteristic polynomial.

$$= \frac{1}{\mathrm{E}(X_t^2)} \sum_{i=1}^{p} \phi_i \mathrm{E}(X_{t-h}, X_{t-i}) + \frac{1}{\mathrm{E}(X_t^2)} \underbrace{\mathrm{E}(X_{t-h}, \varepsilon_t)}_{0}$$

$$= \sum_{i=1}^{p} \phi_i \frac{\mathrm{E}(X_{t-h+i}, X_t)}{\mathrm{E}(X_t^2)} = \sum_{i=1}^{p} \phi_i \frac{\mathrm{E}(X_{t-(h-i)}, X_t)}{\mathrm{E}(X_t^2)} ,$$

and thus

$$\varrho(h) = \sum_{i=1}^{p} \phi_i \varrho(h-i) . \tag{32.10}$$

These are the *Yule-Walker equations* for the autocorrelations ϱ. The autocorrelations can thus be computed recursively by setting the initial condition $\varrho(0) = 1$. Consider the following example of an AR(2) process:

$$\varrho(1) = \phi_1 \varrho(1-1) + \phi_2 \varrho(1-2) = \phi_1 1 + \phi_2 \varrho(1) \Rightarrow \quad \varrho(1) = \frac{\phi_1}{1-\phi_2}$$

$$\varrho(2) = \phi_1 \varrho(1) + \phi_2 \varrho(0) = \frac{\phi_1^2}{1-\phi_2} + \phi_2 \text{ , and so on.}$$

Here, the symmetry indicated in Eq. 32.2 was used together with Eq. 32.10. Substituting these autocorrelations into Eq. 32.9 finally yields the unconditional variance of an AR(2) process:

$$\mathrm{var}[X_t] = \frac{\sigma^2}{1 - \phi_1^2 \varrho(1-1) - \phi_1 \phi_2 \varrho(1-2) - \phi_2 \phi_1 \varrho(2-1) - \phi_2^2 \varrho(2-2)}$$

$$= \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \varrho(1)} = \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1^2 \phi_2/(1-\phi_2)}$$

In practice, however, the autocorrelations should be computed from the original data series itself with the aid of Eq. 32.3, instead of from the coefficients ϕ_i, $i = 1, 2, \ldots, p$ which themselves are only estimated values.

The Autoregressive Process of First Order

We now consider the most simple case, namely $p = 1$. Explicitly, the *autoregressive process of first order AR(1)* is defined as

$$X_t = \phi X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2) \ . \tag{32.11}$$

The stationarity condition for this process implies that $|\phi| < 1$ since Eq. 32.8 states simply that

$$1 - \phi z = 0 \ \text{ for some } z \text{ where } |z| > 1 \ .$$

The conditional variance and conditional expectation of the process are

$$\begin{aligned} \mathrm{E}[X_t | X_{t-1}, \ldots, X_1] &= \phi X_{t-1} \\ \mathrm{var}[X_t | X_{t-1}, \ldots, X_1] &= \mathrm{var}[\varepsilon_t] = \sigma^2 \ . \end{aligned}$$

The *un*conditional expectation is equal to zero as was shown above to hold for general AR(p) processes. The *un*conditional variance can be calculated as

$$\begin{aligned} \mathrm{var}[X_t] &= \mathrm{var}[\phi X_{t-1} + \varepsilon_t] \\ &= \phi^2 \mathrm{var}[X_{t-1}] + \phi \mathrm{cov}[X_{t-1}, \varepsilon_t] + \mathrm{var}[\varepsilon_t] \\ &= \phi^2 \mathrm{var}[X_t] + 0 + \sigma^2 \implies \\ \mathrm{var}[X_t] &= \frac{\sigma^2}{1 - \phi^2} . \ . \end{aligned} \tag{32.12}$$

Recursively constructing future values via Eq. 32.11 starting from X_t yields

$$\begin{aligned} X_{t+h} &= \phi X_{t+h-1} + \varepsilon_{t+h} \\ &= \phi^2 X_{t+h-2} + \phi \varepsilon_{t+h-1} + \varepsilon_{t+h} \\ &\ldots \\ &= \phi^h X_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i} \end{aligned}$$

The autocovariance of the AR(1) thus becomes explicitly

$$\begin{aligned}\operatorname{cov}(X_{t+h}, X_t) &= \operatorname{cov}(\phi^h X_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}, X_t) \\ &= \phi^h \operatorname{cov}(X_t, X_t) + \sum_{i=0}^{h-1} \phi^i \underbrace{\operatorname{cov}(\varepsilon_{t+h-i}, X_t)}_{0} \\ &= \phi^h \operatorname{var}[X_t] \\ &= \phi^h \frac{\sigma^2}{1-\phi^2} \,.\end{aligned}$$

The auto*correlation* is therefore simply ϕ^h, and as such is an exponentially decreasing function of h. The same result can of course be obtained from the Yule-Walker equations

$$\varrho(h) = \sum_{i=1}^{1} \phi_i \varrho(h-i) = \phi\varrho(h-1) = \phi^2\varrho(h-2) = \cdots = \phi^h \underbrace{\varrho(h-h)}_{1} \,.$$

It is worthwhile to consider a *random walk* from this point of view. A (one-dimensional) random walk is by definition constructed by adding an independent, identically distributed random variable (*iid*, for short) with variance σ^2 to the last value of attained in the walk. The random walk can thus be written as

$$X_t = X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2) \,.$$

It follows from this definition that the conditional variance of the random walk is σ^2 and the expectation equals zero. The random walk corresponds to an AR(1) process with $\phi = 1$. This contradicts the stationarity criterion $|\phi| < 1$! The random walk is therefore a *non*-stationary AR(1) process. The non-stationarity can be seen explicitly by considering the *un*conditional variance:

$$\begin{aligned}\operatorname{var}[X_t] &= \operatorname{var}[X_{t-1} + \varepsilon_t] \\ &= \operatorname{var}[X_{t-1}] + \operatorname{cov}[X_{t-1}, \varepsilon_t] + \operatorname{var}[\varepsilon_t] \\ &= \operatorname{var}[X_{t-1}] + 0 + \sigma^2 \,.\end{aligned}$$

Thus, for all $\sigma \neq 0$ we have $\text{var}[X_{t-1}] \neq \text{var}[X_t]$, i.e., the process *cannot* be stationary. Therefore we cannot obtain a closed from expression similar to Eq. 32.12 for the unconditional variance (this can also be seen from the fact that if $\phi = 1$, Eq. 32.12 would imply a division by zero). The unconditional variance can, however, be determined recursively

$$\text{var}[X_t] = k\sigma^2 + \text{var}[X_{t-k}] \; .$$

Assuming from the outset that a value $X_{t=0}$ is known (and because it is known, has zero variance) we obtain the well-known property of the random walk

$$\text{var}[X_t] = t\sigma^2 \; .$$

The variance is thus time dependent; this is a further indication that the random walk is not stationary. Since the variance is linear in the time variable, the standard deviation is proportional to the square root of time. This is the well-known *square root law* for scaling the volatility with time.

Another special case of an AR(1) process is *white noise* which has an expectation equal to zero and constant variance. It is defined by

$$X_t = \varepsilon_t \; .$$

The random variables$\{\varepsilon_t\}$ are *iid* random variables with variance σ^2. This corresponds to the AR(1) process with $\phi = 0$. The stationarity criterion $|\phi| < 1$ is satisfied and the above results for the stationary AR(1) process can be applied with $\phi = 0$, for example, $\text{cov}(X_{t+h}, X_t) = 0$ and $\text{var}(X_t) = \sigma^2$.

32.1.2 Univariate GARCH(p, q) Processes

The conditional variance of the AR(p) processes introduced above was always a constant function of time; in each case it was equal to the variance of ε_t. This, however, is not usually the case for financial time series. Take, for example, the returns of the FTSE data set in Fig. 32.1. It is clear to see that the variance of the data sequence is not constant as a function of time. On the contrary, the process goes through both calm and quite volatile periods. It is much more probable that large price swings will occur close to other large price swings than to small swings. This behavior is typical of financial time series and is referred to as *volatility clustering* or simply *clustering*. A process which is capable of modeling such behavior is the *GARCH(p, q) process*, which will

be introduced below. The decisive difference between GARCH and AR(p) processes is that not only past values of X_t are used in the construction of a GARCH process, but past values of the *variance* enter into the construction as well. The GARCH(p, q) process is defined as

$$X_t = \sqrt{H_t}\varepsilon_t \quad \text{with} \quad H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2, \quad \varepsilon_t \sim \mathrm{N}(0,1)\ , \tag{32.13}$$

where the $\{\varepsilon_t\}$ are *iid* standard normally distributed. The $\{\varepsilon_t\}$ are independent of X_t. Therefore, the time series$\{X_t\}$ is nothing else than white noise $\{\varepsilon_t\}$ with a time-dependent variance which is determined by the coefficients $\{H_t\}$. These H_t take into consideration the past values of the time series *and* the variance. If the $\{X_t\}$ are large (distant from the equilibrium value which is in this case zero as $\mathrm{E}[\varepsilon_t] = 0$), then so is $\{H_t\}$. For small values $\{X_t\}$ the opposite holds. In this way, clustering can be modeled. The order q indicates how many past values of the time series $\{X_t\}$ influence the current value H_t. Correspondingly, p is the number of past values of the variance itself which affects the current value of H_t. In order to ensure that the variance is positive, the parameters must satisfy the following conditions:

$$\begin{aligned} \alpha_0 &\geq 0 \\ \beta_1 &\geq 0 \\ \sum_{j=0}^{k} \alpha_{j+1}\beta_1^{k-j} &\geq 0, \quad k = 0, \ldots, q-1\ . \end{aligned} \tag{32.14}$$

This implies that $\alpha_1 \geq 0$ always holds, the other α_i however, may be negative. Furthermore, the time series $\{X_t\}$ should be (weakly) stationary to prevent it from "drifting away". The following condition is sufficient to guarantee this stationarity:

$$\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1\ . \tag{32.15}$$

The two most important properties of this process pertain to the conditional expectation and the conditional variance

$$\mathrm{E}[X_t|X_{t-1},\ldots,X_1] = 0 \quad \text{and} \tag{32.16}$$
$$\mathrm{var}[X_t|X_{t-1},\ldots,X_1] = H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

The first equation holds because $\mathrm{E}[\varepsilon_t] = 0$, the second because $\mathrm{var}[\varepsilon_t] = 1$. The H_t are thus the conditional variances of the process. The conditional expectation (under the condition that all X up to time $t-1$ are known) of H_t is simply H_t itself since no stochastic variable ε appears in Eq. 32.13 where H_t is defined, and thus

$$\mathrm{E}[H_t|X_{t-1},\ldots,X_1] = H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2 \,. \tag{32.17}$$

H is thus always known one time step in advance of X. This may seem trivial but will be quite useful in Sect. 33.2 when making volatility forecasts.

The *un*conditional variance is by definition

$$\begin{aligned}\mathrm{var}[X_t] &= \mathrm{E}[X_t^2] - \mathrm{E}[X_t]^2 \\ &= \mathrm{E}[H_t\varepsilon_t^2] - \mathrm{E}[\sqrt{H_t}\varepsilon_t]^2 \\ &= \mathrm{E}[H_t]\mathrm{E}[\varepsilon_t^2] - (\mathrm{E}[\sqrt{H_t}]\mathrm{E}[\varepsilon_t])^2 \,,\end{aligned}$$

where in the last step we have made use of the fact that $\{\varepsilon_t\}$ are uncorrelated with $\{H_t\}$. Furthermore, since the $\{\varepsilon_t\}$ are *iid* $\mathrm{N}(0,1)$ distributed

$$\mathrm{E}[\varepsilon_t] = 0 \quad \text{and} \quad \mathrm{E}[\varepsilon_t^2] = \mathrm{E}[\varepsilon_t^2] - 0 = \mathrm{E}[\varepsilon_t^2] - (\mathrm{E}[\varepsilon_t])^2 = \mathrm{var}[\varepsilon_t] = 1$$

and therefore

$$\mathrm{var}[X_t] = \mathrm{E}[H_t] = \alpha_0 + \sum_{i=1}^{p} \beta_i \mathrm{E}[H_{t-i}] + \sum_{j=1}^{q} \alpha_j \mathrm{E}[X_{t-j}^2] \,.$$

Just as the ε_t, the X_t have zero expectation, which also implies that $\mathrm{E}[X_t^2] = \mathrm{var}[X_t]$. Under consideration of this relation and the stationarity (constant

variance), all of the expectations involving squared terms in the above equation can be written as the variance of X_t:

$$\mathrm{E}[X_{t-j}^2] = \mathrm{var}[X_{t-j}] = \mathrm{var}[X_t]$$
$$\mathrm{E}[H_{t-i}] = \mathrm{var}[X_{t-j}] = \mathrm{var}[X_t] \; .$$

This leads to the following equation for the unconditional variance:

$$\mathrm{var}[X_t] = \alpha_0 + \sum_{i=1}^{p} \beta_i \mathrm{var}[X_t] + \sum_{j=1}^{q} \alpha_j \mathrm{var}[X_t] \Longleftrightarrow$$
$$\mathrm{var}[X_t] = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j} =: \widetilde{\alpha}_0 \; . \tag{32.18}$$

This unconditional variance can of course also be estimated from the observed data, i.e., as usual through the computation of the empirical variance estimator over a large number of realizations of $\{X_t\}$.

The GARCH(p, q) process can be expressed in terms of the unconditional variance $\widetilde{\alpha}_0$ as follows:

$$\begin{aligned} H_t &= \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2 \\ &= \alpha_0 \frac{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j} + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2 \\ &= \widetilde{\alpha}_0 - \widetilde{\alpha}_0 \sum_{i=1}^{q} \alpha_i - \widetilde{\alpha}_0 \sum_{j=1}^{p} \beta_j + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2 \\ &= \widetilde{\alpha}_0 + \sum_{i=1}^{p} \beta_i (H_{t-i} - \widetilde{\alpha}_0) + \sum_{j=1}^{q} \alpha_j (X_{t-j}^2 - \widetilde{\alpha}_0) \; . \end{aligned} \tag{32.19}$$

The conditional variance H_t can thus be interpreted as the unconditional variance $\widetilde{\alpha}_0$ plus the sum of the distances from this unconditional variance. If all α_j and β_i are greater than zero (which is always the case for a GARCH(1,1) process), this form of the conditional variance has another interpretation:

The β_i terms cause a kind of *persistence* of the variance which serves to model the *volatility clustering* phenomenon: the greater H_{t-i} becomes in comparison to the long-term expectation $\widetilde{\alpha}_0$ (the unconditional variance), the greater the positive contribution of these terms to H_t; the H_t tend to get even larger. Conversely, for values of H_{t-i} which are smaller than $\widetilde{\alpha}_0$ the contribution of these terms become negative and thus H_t will tend to get even smaller.

The terms involving α_i describe the *reaction* of the volatility to the process itself. Values X_{t-j}^2 larger than $\widetilde{\alpha}_0$ favor a growth in the variance; values X_{t-j}^2 smaller than $\widetilde{\alpha}_0$ result in a negative contribution and thus favor a decline in the variance. If the process itself describes a price *change*, as is common in the financial world, this is precisely the effect that strong price changes tend to induce growth in volatility.

Overall, these properties lead us to expect that GARCH models are indeed an appropriate choice for modeling certain phenomena observed in the financial markets (in particular, volatility clustering and the reaction of the volatility to price changes). In practice, we often set $p = 1$ and even $q = 1$. It has been shown that significantly better results are not achieved with larger values of p and q and thus the number of parameters to be estimated would be unnecessarily increased.

32.1.3 Simulation of GARCH Processes

One of the examples to be found in the Excel workbook GARCH.XLSX from the download section [50] is the simulation of a *GARCH(1,1) process*. The first simulated value X_1 of the time series is obtained, according to Eq. 32.13, from a realization of a standard normal random variable followed by multiplication of this number by $\sqrt{H_1}$. Subsequently, H_2 is computed from the values now known at time $t = 1$. Then, a realization X_2 is generated from a standard normal random variable and multiplied by $\sqrt{H_2}$. This procedure is repeated until the end of the time series is reached.

In order to generate a GARCH(p, q) process, q values of the X process

$$X_{-q+1}, X_{-q+2}, \ldots, X_0$$

and p values of the conditional variance

$$H_{-p+1}, H_{-p+2}, \ldots, H_0$$

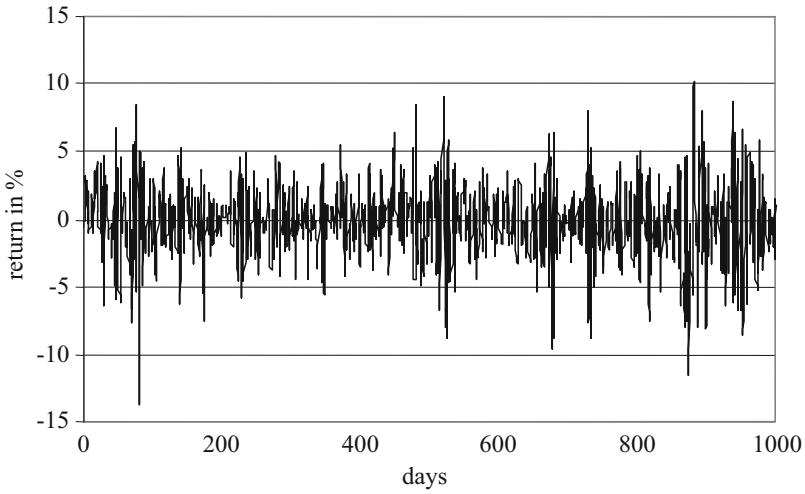

Fig. 32.2 Simulated GARCH(1,1) process. The first 100 values have not been used. Clustering can clearly be observed

must be given in order to be able to compute the first conditional variance H_1 as indicated in Eq. 32.13. The choice of these initial values is not unique but the orders of magnitude of the time series values and the variances should at least be correct. The unconditional expectation $\mathrm{E}[X_t]$ and the unconditional variance $\mathrm{var}[X_t]$ are therefore good candidates for this choice. The first values of the generated time series should then be rejected (often, 50 values are sufficient), since they still include the above "initial conditions". After taking several steps, realizations of the desired GARCH process can be generated. Figure 32.2 illustrates a simulated GARCH process.

Such simulated time series can be implemented to test optimization methods which have the objective of "finding" parameters from the simulated data series which have been previously used for the simulation. After all, if a data set is given (real or simulated), the parameters of a model have to be determined. Methods for doing this are the subject of the next section.

32.2 Calibration of Time Series Models

All of the time series models introduced above include parameters which may be varied for the purpose of fitting the model "optimally" to the time series data. We represent these parameters as a parameter vector θ. For an AR(p) process, the free parameters are the ϕ_i and σ^2 while the GARCH(p, q) has

the free parameters α_i and β_i. Thus

$$\begin{aligned} \theta &= (\phi_1, \phi_2, \ldots, \phi_p, \sigma^2) && \text{for AR}(p) \\ \theta &= (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \beta_2, \ldots \beta_p) && \text{for GARCH}(p, q)\ . \end{aligned}$$

A widely used estimation procedure for the determination of unknown parameters in statistics is the *maximum likelihood estimator*. This procedure selects the parameter values which maximize the likelihood of the model being correct. These are just the parameter values which maximize the *probability* (called the *likelihood*) that the values observed will be realized by the assumed model. Using the model, the probability is expressed as a function of the parameters θ. Then this probability function is maximized by varying the parameter values. The parameter values for which the probability function attains a maximum corresponds to a "best fit" of the model to the given data sequence. They are the most probable parameter values given the information available (i.e., given the available time series). This procedure will now be performed explicitly for both an AR(p) and a GARCH(p,q) process.

32.2.1 Parameter Estimation for AR(p) Processes

The likelihood for the AR(p) process is obtained as follows: from Eqs. 32.6 and 32.7 we can see that if we assume an AR(p) process with a parameter vector

$$\theta = (\phi_1, \phi_2, \ldots, \phi_p, \sigma^2)$$

then X_t has the normal distribution

$$\mathrm{N}(\sum_{i=1}^{p} \phi_i X_{t-i}, \sigma^2)$$

The conditional probability for one single observed value of X_t (also called the *conditional likelihood* of X_t) is thus

$$\begin{aligned} &L_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}) \\ &\quad = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 \right\}\ . \end{aligned}$$

The total likelihood for all T measured data points is in consequence of the independence of ε_t simply a product of all conditional likelihoods:

$$L_\theta(X_1, X_2, \ldots, X_T) = \prod_{t=1}^{T} L_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p})$$
$$= \frac{1}{(2\pi)^{T/2}\,\sigma^T} \prod_{t=1}^{T} \exp\left\{ -\frac{1}{2\sigma^2} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 \right\} .$$

Observe that for the likelihoods of the first data points X_t where $t < p+1$, a further p data points $\{X_0, X_{-1}, \ldots, X_{-p+1}\}$ are required in advance. The extent of the data sequence needed is thus a data set encompassing $T + p$ data points.

Maximizing this likelihood through the variation of the parameters $\phi_1, \phi_2, \ldots, \phi_p$ and σ^2, we obtain the parameters $\{\phi_1, \phi_2, \ldots, \phi_p, \sigma^2\}$ which, under the given model assumptions,[7] actually maximizes the (model) probability that the observed realization $\{X_t\}$ will actually appear. It is, however, simpler to maximize the *logarithm* of the likelihood (because of the size of the terms involved and the fact that sums are more easily dealt with than products). Since the logarithm function is strictly monotone increasing, the maximum of the likelihood function is attained for the same parameter values as the maximum of the logarithm of the likelihood function. The *log-likelihood* function for the AR(p) process is given by

$$\mathcal{L}_\theta = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 .$$

$\phi_1, \phi_2, \ldots, \phi_p$ appear only in the last expression (the sum), which appears with a negative sign in the log-likelihood function. The values of $\phi_1, \phi_2, \ldots, \phi_p$ which maximize the log-likelihood function therefore minimize the expression

$$\sum_{t=1}^{T} \left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \right]^2 \tag{32.20}$$

[7] The model assumption is that the time series was generated by an AR(p) process.

This, however, is just a sum of the quadratic deviations. The desired parameter estimates $\{\widehat{\phi}_1, \widehat{\phi}_2, \ldots, \widehat{\phi}_p\}$ are thus the solution to a least squares problem. The $\widehat{\phi}_t$ can thus be determined independently from the variance σ^2. The estimation of the variance is obtained from simple calculus by taking the derivative of the log-likelihood function with respect to σ^2 and setting the resulting value equal to zero (after substituting the optimal ϕ_i, namely the $\widehat{\phi}_i$):

$$\begin{aligned}\frac{\partial \mathcal{L}_\theta}{\partial \sigma^2} &= -\frac{T}{2}\frac{\partial \ln(2\pi\sigma^2)}{\partial \sigma^2} - \frac{\partial}{\partial \sigma^2}\left(\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[X_t - \sum_{i=1}^{p}\widehat{\phi}_i X_{t-i}\right]^2\right) \\ &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T}\left[X_t - \sum_{i=1}^{p}\widehat{\phi}_i X_{t-i}\right]^2 \stackrel{!}{=} 0\,.\end{aligned}$$

The optimal estimate for σ^2 becomes

$$\widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left[X_t - \sum_{i=1}^{p}\widehat{\phi}_i X_{t-i}\right]^2 . \tag{32.21}$$

For example, the maximum likelihood estimator for ϕ_1 in the AR(1) process in Eq. 32.11, obtained by minimizing the expression in 32.20, can be determined through the following computation:

$$\begin{aligned}0 &= \frac{\partial}{\partial \phi_1}\sum_{t=1}^{T}[X_t - \phi_1 X_{t-1}]^2 = 2\sum_{t=1}^{T}(X_t - \phi_1 X_{t-1})(-X_{t-1}) \\ &= 2\phi_1\sum_{j=1}^{T}X_{j-1}^2 - 2\sum_{t=1}^{T}X_{t-1}X_t \Longrightarrow \\ \widehat{\phi}_1 &= \frac{\sum_{t=1}^{T}X_{t-1}X_t}{\sum_{j=1}^{T}X_{j-1}^2}\,.\end{aligned}$$

Substituting this into Eq. 32.21 yields the maximum likelihood estimator for σ^2

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{T}\sum_{t=1}^{T}\left[X_t - \widehat{\phi}_1 X_{t-1}\right]^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[X_t - X_{t-1}\frac{\sum_{i=1}^{T} X_{i-1}\,X_i}{\sum_{j=1}^{T} X_{j-1}^2}\right]^2 .\end{aligned}$$

32.2.2 Parameter Estimation for GARCH(p, q) Processes

The likelihood for the GARCH(p, q) process is obtained as follows: from Eqs. 32.13 and 32.16 we see that

$$X_t|\{X_{t-1}, \ldots, X_{t-q}, H_{t-1}, \ldots, H_{t-p}\} \sim \mathrm{N}(0, H_t) .$$

This implies that, given the information $\{X_{t-1}, \ldots, X_{t-q}, H_{t-1}, \ldots, H_{t-p}\}$, X_t is normally distributed according to $\mathrm{N}(0, H_t)$. The conditional likelihood for one single observation X_t is then

$$L_\theta(X_t|\{X_{t-1}, \ldots, X_{t-q}, H_{t-1}, \ldots, H_{t-p}\}) = \frac{1}{\sqrt{2\pi H_t}} e^{-X_t^2/2H_t}$$

where

$$H_t = \alpha_0 + \sum_{i=1}^{p} \beta_i H_{t-i} + \sum_{j=1}^{q} \alpha_j X_{t-j}^2$$

and with a parameter vector

$$\theta = (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p) .$$

The overall likelihood of all observations together is, in consequence of the independence of $\{\varepsilon_t\}$, merely the product

$$L_\theta = \prod_{t=1}^{T} L_\theta(X_t|\{X_{t-q}, \ldots, X_{t-1}, H_{t-p}, \ldots, H_{t-1}\}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi H_t}} e^{-X_t^2/2H_t} .$$

Observe that for the likelihood of the first data point X_1 further data points $\{X_0,\ X_{-1}, \ldots, X_{-q+1}, H_0, H_{-1}, \ldots, H_{-p+1}\}$ are required in advance. The total required data sequence $\{X_t\}$ thus encompasses $T + q$ data points. If $T + q$ observations of X_t are available, the first are required as information in advance, the remaining T are included in the likelihood function as observed data. In addition the values $\{H_0, H_{-1}, \ldots, H_{-p+1}\}$ are required as information in advance. In choosing the size of T it is necessary to make a compromise between the exactness of the estimator (T is chosen to be as large as possible) and the time scale with which the market mechanisms change (T is chosen to be as small as possible).

Maximizing this likelihood function by allowing the parameter values in θ to vary, we obtain the parameters which, under the model assumption (a GARCH(p, q) process), maximize the probability of a realization of the market values $\{X_t\}$ observed. It is again easier to work with the log-likelihood function in determining this maximum. Since the log function is strictly monotone increasing, the maximum of the likelihood and the log-likelihood function is attained at the same parameter point. The log-likelihood for the GARCH(p, q) process is given by

$$\begin{aligned}\mathcal{L}_\theta &= \sum_{t=1}^{T} \ln L_\theta(X_t | \{X_{t-q}, \ldots, X_{t-1}, H_{t-p}, \ldots, H_{t-1}\}) \\ &= \sum_{t=1}^{T} \ln\left(\frac{1}{\sqrt{2\pi H_t}} e^{-X_t^2/2H_t}\right) \\ &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln(H_t) - \frac{1}{2}\sum_{t=1}^{T}\frac{X_t^2}{H_t} \end{aligned} \tag{32.22}$$

where

$$H_t = H_t(\theta) = \alpha_0 + \sum_{j=1}^{p} \beta_j H_{t-j} + \sum_{k=1}^{q} \alpha_k X_{t-k}^2 \,.$$

This is the function which must now be maximized through the variation of the parameter vector θ. The space of valid parameters θ is limited by the constraints stated in Eqs. 32.14 and 32.15. This represents an additional difficulty for the optimization. The optimization is quite difficult because (as opposed to the AR(1) process) maximizing the likelihood function cannot be

computed analytically but must be accomplished by means of a *numerical optimization* procedure. As the function to be maximized has multiple local maxima, a complex "likelihood surface" further complicates the optimization process since *local* optimization methods, such as gradient methods, are unsuitable if the initial value is not well chosen, i.e., if it does not lie close to the global maximum. A suitable algorithm for finding a *global* maximum in such a situation is *simulated annealing*.

32.2.3 Simulated Annealing

Simulated annealing is a numerical algorithm used to find a *global* minimum or maximum of a given function. Its construction is motivated by an effect observed in physics, namely cooling. The cooling of a physical body results in its moving through decreasing energy states traveling a path ending in a state of *minimum energy*. The simulated annealing algorithm attempts to imitate this process. The function whose minimum is to be found thus corresponds to the energy of the physical body.

As a physical body cools, the temperature T declines resulting in a steady loss of energy. The body is composed of billions of atoms which all make a contribution to its total energy. This being the case, there are a multitude of possible energy states with a multitude of *local* energy minima. If the temperature declines very *slowly*, the body surprisingly finds its *global* minimum (for example, the atoms in the body may assume a characteristic lattice configuration). A simple approach to this can be taken from *thermodynamics*: the probability of a body being in a state with energy E when the temperature of the body is T is proportional to the *Boltzmann factor,* $\exp(-E/kT)$:

$$P(E) \sim \exp\left(-\frac{E}{kT}\right) ,$$

where k is a thermodynamic constant, the *Boltzmann constant*. It follows that a higher energy state can be attained at a certain temperature though the probability of such an event declines with a decline in temperature. In this way, "unfavorable" energy states *can* be attained and thus the system *can* escape from local energy minima. However, if the temperature drops too quickly, the body remains in a so-called meta-stable state and cannot reach its global energy minimum.[8] It is therefore of utmost importance to cool the body *slowly*.

[8] Physicist speak in such cases of "frustrated" systems. An example of such a frustrated system is glass.

This strategy observed in nature is now to be simulated on a computer. In order to replicate the natural scenario, a configuration space (the domain of possible values of the pertinent parameters θ) must be defined. This might be a connected set but could also consist of discrete values (*combinatorial optimization*). In addition, a mechanism is required governing the transition from one configuration to another. And finally, we need a scheme for the cooling process controlling the decline in "temperature" T ($T_0 \to T_1 \to \cdots \to T_n \to \cdots$). The last two points mentioned are of particular importance; the change-of-configuration mechanism determine how efficient the configuration space is sampled while the second of the above requirements serves to realize the "slow cooling".

For each temperature the parameter sequence forms a *Markov chain*. Each new test point θ is accepted with the probability[9]

$$P = \min\left\{e^{-[f(\theta_p)-f(\theta_{p-1})]/T}, 1\right\}$$

where θ_{p-1} represents the previously accepted parameter configuration. The function f is the function to be minimized for each specific problem and is, for example, the (negative) log-likelihood function from Eq. 32.22. This function corresponds to the energy function in physics.

After having traveled a certain number of steps in the Markov chain, the temperature declines according to some mechanism which could for instance be as simple as

$$T_n = \alpha T_{n-1} \quad (0 < \alpha < 1) \,.$$

A new Markov chain is then started. The starting point for the new chain is the end point of the previous chain. In a concrete optimization, the temperature is naturally not to be understood in the physical sense; it is merely an abstract parameter directing the course of the optimization by controlling the transition probability in the Markov chain. However, we choose to retain the designations temperature or cooling scheme as a reminder of the procedure's origin. Figure 32.3 shows a schematic representation of the algorithm.

Simulated annealing is demonstrated in the Excel workbook GARCH.XLSX by means of a VBA program. The algorithm in the workbook is used to fit the

[9]The minimum function is only required since a probability can be at most equal to one.

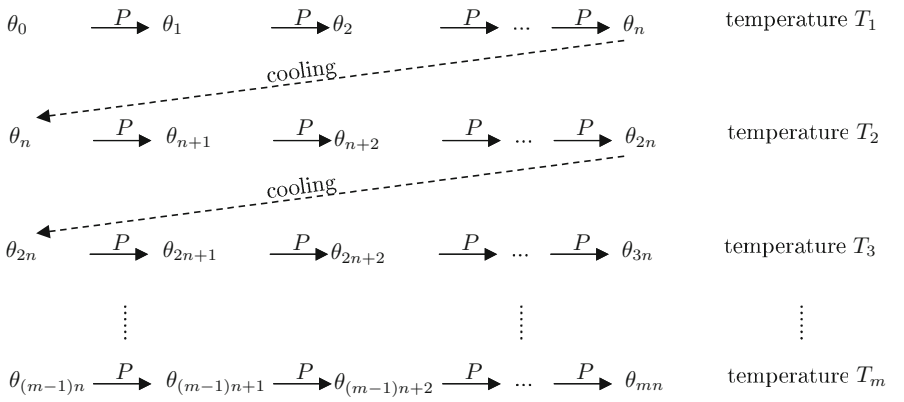

Fig. 32.3 Simulated annealing using m Markov chains with n steps in each chain. If the cooling is slow enough and m and n are large enough, then θ_{mn} is a good approximation of the parameter vector necessary to achieve the global minimum of the function f

parameters of a GARCH(1,1) process making use of the first 400 points of a given (simulated) data set. No emphasis is placed on the speed of computation since our object is to demonstrate the fundamental principles as clearly as possible.