Chapter 1

Concepts in linear time series analysis

To understand the term "time series" and to follow the up-coming theory smoothly, it may be necessary to recall some fundamental facts.

A discrete time series is a set of time-ordered data $\{y_1, y_2, \dots y_n\}$ taken from observations of some phenomenon, usually at equally spaced time intervals ([9]). A short-hand for the above sequence is y_t, where the subscript $t = 1, \dots n$ is referred to as time, n denotes the length of the time series and y_t is assumed to be real. The main purpose of time series analysis is to understand the underlying mechanism that generates the observed data and, in turn, to forecast future values. We assume that the generating mechanism is probabilistic and that the observed series $\{y_1, y_2, \dots y_t, \dots\}$ is a realization of a stochastic process $\{Y_1, Y_2, \dots Y_t, \dots\}$, i.e., a sequence of random variables. For simplicity, in the following the term time series refers both to observed data and to stochastic process, and lower case notation is used.

Generally, when modelling typical (linear) time series one may encounter the following (classes of) components [9]:

- trend, the long-term component representing growth or decline over an extended period of time

- seasonal component, annually repeating pattern of changes constrained within the most natural periodicity

- cyclical component, a wavelike fluctuation around the trend

- residuals, usually stochastic remains after deterministic components removal

The residuals contain plethora of information and needs to be further analysed by means of so-called Box–Jenkins methodology that covers a large family of linear models such as autoregressive (AR), moving average (MA), integrated ARMA (ARIMA) and the like. All the above components can be picked up from a "noisy heap" in sequence or at once, usually by the ordinary least square (OLS) regression procedure, so that only a white noise is assumed to remain. However, the decomposition is not as simple as it may appear, there are many and different perspectives when dealing with data generating processes, for example such phenomenon as trend can be deterministic or stochastic and one possible way to define it (preferred also in this work) is through the context of autoregressive models. Therefore, the purpose of this chapter, mainly following [17] and [15], is to provide some useful concepts in linear modelling which definitely come handy later when describing extension to non-linear space and analysis of common features.

In section 1.1 we start our journey by the basics of Box-Jenkins methodology introducing the famous ARMA class of linear models, within context of which (nonseasonal, linear and non-trending univariate time series with constant variance) we can treat the concepts of empirical specification strategy as model identification, estimation and evaluation (section 1.2). These are generally useful or easily modifiable also in the case of contamination by the key features such as trend, seasonality and aberrant observations (outliers), which are focused on in the last three sections, consecutively.

1.1 Linear time series model

Assume the univariate time series of interest y_t that might be any geometric or physical variable observed for $t = 1, 2, \ldots n$. Let Ω_{t-1} denote the history or information set at time $t-1$, which contains all available information exploitable for forecasting future values $y_t, y_{t+1}, y_{t+2}, \ldots$. If Ω_{t-1} does not contain any of such information, the corresponding time series is usually called a white noise time series, hereafter denoted as ε_t, and is required to have a constant (unconditional) mean equal to zero and a constant (unconditional) variance σ^2 as well. More formally, white noise series can be defined by

$$\begin{aligned} \mathrm{E}[\varepsilon_t] &= 0, \\ \mathrm{E}[\varepsilon_t^2] &= \sigma^2, \\ \mathrm{E}[\varepsilon_t \varepsilon_s] &= 0, \qquad \forall\, s \neq t, \end{aligned} \tag{1.1}$$

where E stands for expectation operator.

So in general, any time series y_t can be thought of as being the sum of two parts: what can and what cannot be predicted using the knowledge from the past as gathered in Ω_{t-1}. That is, y_t can be decomposed as

$$y_t = \mathrm{E}[y_t|\Omega_{t-1}] + \varepsilon_t, \tag{1.2}$$

where $\mathrm{E}[\cdot|\cdot]$ denotes conditional expectations. A commonly applied model for the predictable component of y_t assume that it is a linear combination of p of its lagged values

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad t = 1, \ldots n, \tag{1.3}$$

where ϕ_i are unknown parameters. This simple model is called autoregressive model AR(p) or autoregression of order p and can be written in a more concise form as

$$\phi_p(L) y_t = \varepsilon_t, \tag{1.4}$$

using the lag operator L defined by $L^k y_t = y_{t-k}$ for any integer k, and

$$\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \tag{1.5}$$

which is so-called AR-polynomial in L of order p.

When p in the AR(p) model is large, one may try to approximate the AR-polynomial by a ratio of two polynomials which together involve a smaller number of parameters. The resultant model

$$\phi_p(L) y_t = \theta_q(L) \varepsilon_t \tag{1.6}$$

then is called autoregressive moving average model ARMA(p, q) with $\theta_q(L)$ being the polynomial of moving average model of order q. Sometimes it is efficient to cope only with MA(q) part of the model. However, it'll not be the case in this work and we will concentrate on AR model which is much more convenient for many practical purposes mainly because of the easiness of parameter estimation, diagnostic measures calculation and that it can be easily extended to allow for trending behaviour, seasonality, shifts and non-linearity.

A white noise as defined by (1.1) is a special case of a time series that is *covariance stationary*, which means it has constant mean and variance, and autocovariances depend only on time lag k, formally

$$\begin{aligned} \mathrm{E}[y_t] &= \mu, \\ \mathrm{E}[(y_t - \mu)^2] &= \gamma_0, \\ \mathrm{E}[(y_t - \mu)(y_{t-k} - \mu)] &= \gamma_k, \qquad \forall\, k = 1, 2, \ldots, \end{aligned} \tag{1.7}$$

where μ, γ_0, and γ_k are finite-valued numbers. Whether or not a time series y_t generated by an AR(p) model is covariance stationary is determined by the autoregressive parameters $\phi_1, \ldots, \phi_p$. For example, consider first-order autoregression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t \tag{1.8}$$

with intercept ϕ_0 included to describe a nonzero mean of y_t and rewrite it by recursive substitution as

$$y_t = \phi_1^t y_0 + \sum_{i=0}^{t-1} \phi_1^i \phi_0 + \sum_{i=0}^{t-1} \phi_1^i \varepsilon_{t-i}. \tag{1.9}$$

from which it follows that $\mathrm{E}[y_t] = \phi_1^t y_0 + \sum_{i=0}^{t-1} \phi_1^i \phi_0$. When $|\phi_1| < 1$, it holds that

$$\sum_{i=0}^{t-1} \phi_1^i = (1 - \phi_1^t)/(1 - \phi_1) < \infty \qquad \forall\, t \geq 0.$$

For t sufficiently large, $\mathrm{E}[y_t] = \mu = \frac{\phi_0}{1-\phi_1}$ which shows the relevance of condition $|\phi_1| < 1$ for stationarity. On the other hand, when $|\phi_1| > 1$ the time series is explosive, which is the feature that rarely occurs in practice. An interesting case concerns $\phi_1 = 1$ in which the effect of the past shocks remains the same as time increases. We will pay attention to this case in section 1.3. Similar conclusion holds from inspection of variance and autocovariances for an AR(1) time series.

To generalize the above results to AR(p), consider the *characteristic equation* of the AR(1) and AR(p) models, given by

$$1 - \phi_1 z = 0, \tag{1.10}$$

$$1 - \phi_1 z - \cdots - \phi_p z^p = 0, \tag{1.11}$$

respectively. The solution, or root, of (1.10) is $z = \phi_1^{-1}$. Hence, the condition that $|\phi_1|$ is less than 1 for time series y_t generated by AR(1) model to be stationary is equivalent to the condition that the root of (1.10) is larger than 1. The condition for covariance stationarity of y_t generated by AR(p) model then simply is that all p solutions of (1.11) are larger than 1 – or, rather, as the solutions can be complex numbers, that they are outside the unit circle. Notice that (1.11) can be rewritten as

$$(1 - \alpha_1 z)(1 - \alpha_2 z) \ldots (1 - \alpha_p z) = 0 \tag{1.12}$$

which shows that stationary condition is equivalent to the requirement that all α_i, $i = 1, \ldots p$, are inside the unit circle. When the largest of the α_is is equal to 1, $z = 1$ is a solution to (1.11). In this case we say that the AR(p) polynomial has a unit root.

1.2 Model specification strategy

In this section a typical specification strategy for linear time series models is described. It holds also for nonlinear models in general, although sometimes there are differences in the statistical tools which should be used. The modelling sequence usually involves the following steps:

1. calculate certain statistics for a time series and compare with the theoretical values that would hold true if a certain model is adequate

2. estimate the parameters in the time series model suggested by the results in previous step

3. evaluate the model using diagnostic measures

4. respecify the model if necessary

5. use the model for descriptive or forecasting purposes

Model identification

If attention is restricted to linear ARMA(p, q) models, the main objective of the first step is to determine the appropriate orders p and q. This part of specification strategy is often called model identification. The most relevant statistics that may suggest the appropriate orders are contained in the *autocorrelation function* (ACF) and *partial autocorrelation function* (PACF). The ACF of a stationary time series is defined by

$$\rho_k = \gamma_k/\gamma_0, \qquad k = 1, 2, 3 \ldots, \tag{1.13}$$

where γ_k is the k-th order autocovariance of y_t defined in (1.7). The k-th order autocorrelation can be estimated by means of sample covariances as

$$\hat{\rho}_k = \frac{\frac{1}{n}\sum_{t=k+1}^{n}(y_t - \mu)(y_{t-k} - \mu)}{\frac{1}{n}\sum_{t=1}^{n}(y_t - \mu)^2}, \tag{1.14}$$

where μ is the sample mean of y_t. The k-th order partial autocorrelation can be interpreted as the correlation between y_t and y_{t-k} after accounting for the correlation by intermediate observations $y_{t-1}, \ldots y_{t-k+1}$. An easy way to obtain estimates of the partial autocorrelations is by estimating AR(k) models

$$y_t - \mu = \psi_1^{(k)}(y_{t-1} - \mu) + \cdots + \psi_k^{(k)}(y_{t-k} - \mu) + \upsilon_t, \tag{1.15}$$

for any values of k, where υ_t is not necessarily a white noise time series. The k-th order partial autocorrelation is given by the last coefficient estimate, $\hat{\psi}_k^{(k)}$.

If a time series is described most adequately by an ARMA(p, q) model, the orders p and q can be estimated by comparing the values of estimated (P)ACF with the theoretical values as implied by ARMA models for different p and q. However, the ACF and PACF are easy to interpret only for simple models, when the models become more complicated – say, an ARMA(4,3) – one needs considerable skill and experience to deduce the correct orders of this model based on (P)ACF only. Note that ACF is useful for identification of the order of a pure MA while PACF of the pure AR model.

An alternative specification strategy is to start off with a linear time series model, based on a rough guess using linear autocorrelation functions, and then, in a next step, to use diagnostic tests (performed on residuals) which have power against the alternative model of interest. In case two or more linear (also nonlinear) time series models pass relevant diagnostic tests, usually the final model *selection* is based on minimizing the value of certain criterion function. Whether the selection uses evaluation of in-sample fit or out-of-sample forecasting, it depends on one's concerns. Before we turn our attention to details, few words should be dedicated to estimation.

Estimation

The parameters in the AR(p) model

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{1.16}$$

can be estimated by ordinary least square (OLS) procedure. It can be shown that under relatively weak assumptions about the properties of the innovations ε_t (much weaker than (1.1) which we use here), the OLS estimates of the parameters are consistent and asymptotically normal, and that a standard t-statistic can be used to investigate the significance of ϕ_1 to ϕ_p. The mean μ of y_t can be estimated from $\hat{\mu} = \hat{\phi}_0/(1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p)$. Be aware of imposing the intercept ϕ_0 to be zero while μ is not, as it forces the estimate $(1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p)$ toward zero, and hence spuriously suggests the presence of a unit root. Other methods of AR as well as (AR)MA model parameters estimation can be found, e.g., in [3]. Finally, using the parameter estimates, the residual series ε_t can be constructed.

Diagnostic testing

There are various ways of checking if a model is satisfactory. The commonly used approach is to start with examining whether the residual series $\hat{\varepsilon}_t$ is approximately white noise, by testing whether its autocovariances - or autocorrelations - are equal to zero. There are three commonly applied methods to test for *residual autocorrelations*, all of which can also be considered (or modified) for nonlinear time series models. The first method is to look at individual elements of the sample ACF of the residuals

$$r_k(\hat{\varepsilon}) = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}, \tag{1.17}$$

for $k = 1, 2, \ldots$ to see if they lie between the $\pm 1.96/\sqrt{n}$ bounds that, assuming normality, correspond to 5% significance level. Given model adequacy, the population equivalents of $r_k(\hat{\varepsilon})$ are asymptotically uncorrelated and have variances approximately equal to $1/n$.

A second method amounts to testing for the joint significance of the first m residual autocorrelations. The test-statistic (developed by Ljung and Box, also referred to as portmanteau test-statistic) is given by

$$LB(m) = n(n+2) \sum_{k=1}^{m} \frac{r_k^2(\hat{\varepsilon})}{n-k} \tag{1.18}$$

and asymptotically follows a $\chi^2(m-p-q)$ distribution under the null hypothesis of no residual autocorrelation provided that m/n is small and m is moderately large. Despite this test may not have much power as shown by simulation studies, it is often used because of its ease of computation.

The third method follows the Lagrange Multiplier (LM) principle. To test an AR(p) model against an AR($p+r$) or an ARMA(p, r) model, we consider the auxiliary regression

$$\hat{\varepsilon}_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 \hat{\varepsilon}_{t-1} + \cdots + \beta_r \hat{\varepsilon}_{t-r} + \upsilon_t, \tag{1.19}$$

where $\hat{\varepsilon}_t$ are the residuals of AR(p) model with $\hat{\varepsilon}_t = 0$ for $t \leq 0$. The LM test-statistic which tests the significance of the parameters $\beta_1, \ldots \beta_r$ is calculated as nR^2, where R^2 is the (uncentred) coefficient of determination[1] from (1.19) and it is asymptotically $\chi^2(r)$ distributed under the null hypothesis that the AR(p) model is adequate. In small samples, the F-version of this LM test has better size and power properties.

[1]The square of the multiple correlation coefficient, called the coefficient of determination, is defined by $R^2 = 1 - \sum_{t=1}^{n} \hat{\varepsilon}_t^2 / \sum_{t=1}^{n} (y_t - \hat{\mu})^2$ and indicates the proportion of the variation in y_t "explained" by certain regression.

There are various other tests for checking the randomness of the residuals, for example the turning points test and difference-sign test, which can be easily implemented into one's hand-made algorithms. An observation y_i is a turning point if $(y_i - y_{i-1})(y_{i+1} - y_i) < 0$, in other words, it is a local extremum. Let n_{TP} denote the number of turning points. If the series is a realization of an identically and independently distributed random process, then n_{TP} has an asymptotic normal distribution with mean $\mu_{n_{TP}} = 2(n-2)/3$ and variance $\sigma^2_{n_{TP}} = (16n - 29)/90$. In the difference-sign test, the number of positive first differences $(y_i - y_{i-1} > 0)$ stands for the test-statistic which asymptotically follows normal distribution with mean $(n-1)/2$ and variance $(n+1)/12$. For proof, see [36].

Another property of the residuals which should be tested concerns the constancy of their variance. If this is indeed the case, the residuals are said to be homoscedastic, while if the variance changes they are called heteroscedastic. Neglecting *heteroscedasticity* of the residuals has potentially quite severe consequences. For example, even though the OLS estimates of the AR(MA) parameters are still consistent and asymptotically normal distributed, their variance-covariance matrix is no longer the usual one. Hence, ordinary t-statistic cannot be used to asses the significance of individual regressors in the model. Furthermore, other diagnostic tests, such as tests for nonlinearity (some of which will be discussed in chapter 2), are affected by heteroscedasticity as well, in the sense that their usual asymptotic distributions no longer apply. In particular, neglected heteroscedasticity can easily suggest spurious nonlinearity in the conditional mean. Finally, confidence intervals for forecasts, which are discussed in detail below, can no longer be computed in the usual manner. Several statistics for testing the null hypothesis of constant residual variance can be applied. Which test is used depends partly on whether or not one has a specific alternative in mind. As it happens much more often that no obvious alternative to homoscedasticity is available, a general test can be applied such as that of McLeod and Li, who compute the test-statistic in exactly the same way as the LB test (1.18), except that it tests for autocorrelation in the *squared* residuals. Heteroscedastic time series are treated by the class of models denoted as (Generalized) Autoregressive Conditional Heteroscedasticity, (G)ARCH, which are out of scope of this thesis, however, interested reader is referred to [3, 9, 17, 24] among many others.

The last but not the least among diagnostic tests is the testing for *normality* of the residuals. A usual assumption for the series ε_t is that its realizations are independent and identically distributed (as already mentioned

above) according to a normal distribution with zero mean and common variance σ^2. The notation for this assumption is $\varepsilon_t \sim \text{NID}(0, \sigma^2)$ and it adds Gaussianity to (1.1). Given this assumption, we can use standard tools to evaluate the parameter estimates and their t-ratios. Importantly, and relevant to the next chapter, if we erroneously consider a linear time series model while a nonlinear model would have been more appropriate, the estimated residuals from the linear model often are not NID. For the purpose of testing the assumption of NID, typically a $\chi^2(2)$ normality test is used which consists of a component for the skewness and for the kurtosis. Defining the jth moment of the estimated residuals as $\hat{m}_j = n^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_t^j$, the skewness of $\hat{\varepsilon}_t$ can be calculated as $\widehat{SK}_{\hat{\varepsilon}} = \hat{m}_3 / \sqrt{\hat{m}_2^3}$, and the kurtosis as $\widehat{K}_{\hat{\varepsilon}} = \hat{m}_4 / \hat{m}_2^2$. Because the normal distribution has skewness equal to 0 and kurtosis equal to 3, under the null hypothesis of normality (and no autocorrelation in $\hat{\varepsilon}_t$), the standardized skewness $\sqrt{n/6} \cdot \widehat{SK}_{\hat{\varepsilon}}$ and kurtosis $\sqrt{n/24} \cdot (\widehat{K}_{\hat{\varepsilon}} - 3)$ are independent and have an asymptotic N(0,1) distribution. A joint test for normality (the well-known Jarque-Bera test) is given by

$$JB = \frac{n}{6} \widehat{SK}_{\hat{\varepsilon}}^2 + \frac{n}{24} (\widehat{K}_{\hat{\varepsilon}} - 3)^2, \tag{1.20}$$

which has an asymptotic $\chi^2(2)$ distribution. Rejection of normality may indicate that there are outlying observations, that the error process is not homoscedastic, and/or that the data should better be described by a nonlinear time series model.

Model selection by in-sample fit

The previously discussed identification, estimation, and diagnostic stages can result in a set of tentatively useful models, in the sense that these models cannot be rejected using the above diagnostic measures. We may now want to select the best one by minimisation of some *information criterion* based on in-sample fit, although we may also opt to consider all models for out-of-sample forecasting in order to see, which one performs best on some previously unseen data (discussed later).

The standard coefficient of determination R^2 is not very useful for evaluating time series models as it is only a function of the parameter values. More appropriate model selection criteria are the information criteria put forward by Akaike and Schwarz. Both criteria compare the in-sample fit, which is measured be the residual variance, against the number of estimated parameters. Let k denote the total number of parameters in the ARMA model (i.e. $k = p + q + 1$) to be estimated, then the Akaike Information Criterion (AIC)

is given by

$$AIC(k) = n \ln \hat{\sigma}^2 + 2k \tag{1.21}$$

and Schwarz criterion (BIC), which originates from Bayesian arguments, is computed as

$$BIC(k) = n \ln \hat{\sigma}^2 + k \ln n, \tag{1.22}$$

where $\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_t^2$, with $\hat{\varepsilon}_t^2$ being the residuals from the ARMA model. The values of p and q that minimize $AIC(k)$ and/or $BIC(k)$ are selected as the appropriate orders for the ARMA model. The minimization is done by varying p and q such that k does not exceed certain upper bound which is set in advance. Because $\ln n > 2$ for $n > 8$, the BIC penalizes additional parameters more heavily than the AIC, that means the improvement caused by increasing the AR and/or MA orders needs to be quite substantial for the BIC to favour a more elaborate model. This has implications for the use of these criteria in evaluating nonlinear time series models, where sometimes quite a large number of parameters is needed to obtain only a slightly improved fit.

Out-of-sample forecasting and model selection

The other main purpose of specifying a statistical model for a time series y_t, besides describing certain of its features, is to forecast future values. Let $\hat{y}_{t+h|t}$ denote a forecast of y_{t+h} made at time t, which has an associated forecast (or prediction) error $e_{t+h|t}$,

$$e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}. \tag{1.23}$$

Obviously, many different forecasts $\hat{y}_{t+h|t}$ could be used to obtain an estimate of y_{t+h}. Analogous to the estimation of a time series model, where the parameters are chosen such that the residual variance is minimized, in forecasting it is often considered desirable to choose the forecast $\hat{y}_{t+h|t}$ which minimizes the squared prediction error (SPE)

$$SPE(h) = \mathrm{E}[e_{t+h|t}^2] \tag{1.24}$$

and which turns out to be the conditional expectation of y_{t+h} at time t, that is $\hat{y}_{t+h|t} = \mathrm{E}[y_{t+h}|\Omega_t]$.

Forecasts from AR models (or ARMA models in general) for different forecast horizons h can be obtained quite conveniently by using a recursive relationship. For example, given AR(2) model, the 1-step-ahead forecast at time t is $\hat{y}_{t+1|t} = \phi_1 y_t + \phi_2 y_{t-1}$. Notice that the parameters ϕ_1 and ϕ_2 are assumed to be known so that we do not explicitly introduce additional

uncertainty by considering $\hat{\phi}_i$ instead of ϕ_i. In general the optimal forecast ('optimal' in the squared prediction error sense) is

$$\hat{y}_{t+h|t} = \phi_1 \hat{y}_{t+h-1|t} + \phi_2 \hat{y}_{t+h-2|t} \tag{1.25}$$

with $\hat{y}_{t+i|t} = y_{t+i}$ for $i \leq 0$. To obtain expressions for the SPE from ARMA(p, q) model, it is convenient to rewrite the particular (stationary) model as an MA(∞) model, that is $y_t = \varepsilon_t + \eta_1 \varepsilon_{t-1} + \eta_2 \varepsilon_{t-2} + \ldots$, for which it holds that $SPE(h) = \sigma^2 \sum_{i=0}^{h-1} \eta_i^2$ with $\eta_0 = 1$. For the AR(2) model, e.g., it is easy to verify that $\eta_1 = \phi_1$ and $\eta_2 = \phi_1^2 + \phi_2$. Assuming normality, a 95% forecasting interval for y_{t+h} is bounded by $\hat{y}_{t+h|t} \pm 1.96 \cdot RSPE(h)$ where $RSPE(h)$ denotes the square root of $SPE(h)$. For most nonlinear time series models, the expressions for forecast error variances become much more complicated or even intractable analytically. In that case, one needs to rely on simulation techniques to construct confidence intervals for the forecasts $\hat{y}_{t+h|t}$.

As already mentioned above, comparison of the forecast performance of two or more models under consideration may be a desired alternative to selecting a model according to measures of in-sample fit. Usually one then retains m observations to evaluate h-step-ahead forecasts from models which are fitted to the first n observations (thus the time series has length equal to $n + m$).

A simple check on the quality of forecasts concerns the percentage of m observations lying in the 95% forecast confidence interval. If there is less than 95 per cent within the interval, it is likely that the variance of the data is underestimated. Additionaly, a binomial test can be used to examine if the forecast errors are about equally often positive or negative. Rejection would point to under- or overestimation of the conditional mean, which is usually interpreted as that the deterministic component in the model such as mean and trend are not adequately specified. Another criteria are the mean squared prediction error (MSPE) and the mean absolute prediction error (MAPE)

$$MSPE = \frac{1}{m} \sum_{j=1}^{m} (\hat{y}_{n+j|n+j-h} - y_{n+j})^2 \tag{1.26}$$

$$MAPE = \frac{1}{m} \sum_{j=1}^{m} |\hat{y}_{n+j|n+j-h} - y_{n+j}| \tag{1.27}$$

Sometimes, if a time series display rather erratic behaviour with sudden exceptional values, it makes more sense to consider the median version of the above two criteria.

If we want to decide whether the SPEs or APEs of two alternative models A and B are significantly different, a simple procedure is to create the new variable (so-called loss differential)

$$d_j = g(e_{n+j|n+j-h,A}) - g(e_{n+j|n+j-h,B}), \qquad j = 1, 2, \dots m, \tag{1.28}$$

with the forecast errors $e_{\cdot,A}$ and $e_{\cdot,B}$ generated from models A and B, respectively, and $g(\cdot)$ being some specified loss function, e.g., $g(e) = e^2$ or $g(e) = |e|$ if the goal is to compare the SPEs or APEs, respectively. One possibility to test the null hypothesis that there is no qualitative difference between the forecasts from the two models is (according to [10]) to use the sign test-statistic $S' = \sum_{j=1}^{m} \mathrm{I}[d_j > 0]$ which has the binomial distribution with parameters m and $1/2$ under the null hypothesis. The indicator function $\mathrm{I}[A]$ equals 1 if the event A occurs and 0 otherwise. Significance may be assessed using a table of the cumulative binomial distribution. For large values of m, the studentized version of the sign test-statistic is (asymptotically) standard normal:

$$S = \frac{S' - m/2}{\sqrt{m/4}} = \frac{2}{\sqrt{m}} \sum_{j=1}^{m} \left(\mathrm{I}[d_j > 0] - \frac{m}{2} \right) \overset{a}{\sim} N(0,1). \tag{1.29}$$

Because the statistic S compares only the relative magnitude of the prediction errors, Diebold and Mariano [10] also developed a statistic which compares the absolute magnitudes by testing whether the average loss differential $\bar{d} = m^{-1} \sum_{j=1}^{m} d_j$ is significantly different from zero. The relevant test-statistic is given by

$$DM = \frac{\bar{d}}{\sqrt{\omega/m}} \overset{a}{\sim} N(0,1) \tag{1.30}$$

where ω/m is the asymptotic variance of $\bar{d}$ and ω is suggested to be estimated by an unweighted sum of the autocovariances of d_j, denoted as $\hat{\gamma}_i(d)$, as

$$\hat{\omega} = \gamma_0 + 2 \sum_{i=1}^{h-1} \hat{\gamma}_i(d) \qquad \text{with} \quad \hat{\gamma}_i(d) = \frac{1}{m} \sum_{j=i+1}^{m} (d_j - \bar{d})(d_{j-i} - \bar{d}), \tag{1.31}$$

assuming that h-step-ahead forecast exhibit dependence (or that forecast errors are serially correlated) up to the order $h-1$. As [10] reported, for moderately large samples the performance of their test was satisfactory in a wide range of situations, including contemporaneously correlated and autocorrelated forecast errors, and heavy-tailed as well as normal error distributions.

However, the test was found to be quite seriously over-sized for smaller number of (predicted) observations. Therefore Harvey, Leybourne and Newbold [25] proposed a modified version

$$MDM = \left(\frac{m + 1 - 2h + h(h-1)/m}{m} \right)^{\frac{1}{2}} DM \tag{1.32}$$

that corrects for the tendency of DM statistic to be over-sized in small samples, especially for greater horizons ($h \geq 2$). A further improvement brought by [25] lies in comparing MDM statistic with critical values from the Student's t-distribution with $(m-1)$ degrees of freedom, rather than from the standard normal. Finally, as both tests converge with $m \to \infty$, it's intuitively reasonable to use the modified test in practice.

In case one is more interested in accurate forecasts of the direction in which particular process is moving than in the exact magnitude of the change, then the so-called Directional Accuracy test is available, see [17] for details.

1.3 Trend

When looking at most of plots of the data with trending pattern such as position of a point moving in certain direction, the trend typically moves upwards. Although many practitioners would be able to indicate roughly what a trend is ("an upward moving pattern") a formal definition of a trend cannot be given otherwise than in the context of a model. In this thesis we mainly deal with trends within the framework of AR class of time series models. It is important to investigate the precise formulation of trend in a time series prior to putting effort in modelling and forecasting. Firstly, the trend will dominate long-run out-of-sample forecasts, secondly, trend makes time series to be non-stationary with no tendency of mean reversion (estimate of mean does not converge as n increases). Thirdly, the variance of the forecast errors increases with any new observation which implies that autocorrelation function can vary over time as well. To keep the summary statistics (mean, variance and autocovariances) to be interpretable, they should be constant over time.

Deterministic and stochastic

For investigating stationarity or trending behaviour, consider again the AR(p) model (1.3) or (1.4) with $\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$. Recall that the AR($p$) model is nonstationary if its characteristic equation has a unit root. The presence of unit root causes the autocorrelations to be varying over time and the

effect of shocks remain permanent. In that case, the AR polynomial can be factorized as

$$\phi_p(L) = \phi^*_{p-1}(L)(1-L), \tag{1.33}$$

where $\phi^*_{p-1}(L)$ is a lag-polynomial of order $p-1$ which has all roots outside the unit circle. Then the new variable $(1-L)y_t$ is described by a (covariance) stationary AR$(p-1)$ model. An important consequence of a unit root in the AR(p) polynomial is that the regressors for the nonzero mean and trend appears differently in models with and without unit root. For the sake of illustration, take the simple AR(1) model where y_t is considered in deviation from a possible mean and deterministic trend, that is

$$y_t - \mu - \delta t = \phi_1(y_{t-1} - \mu - \delta(t-1)) + \varepsilon_t, \tag{1.34}$$

written in shorter form as

$$y_t = \mu^* + \delta^* t + \phi_1 y_{t-1} + \varepsilon_t, \tag{1.35}$$

where $\mu^* = (1-\phi_1)\mu + \phi_1\delta$ and $\delta^* = (1-\phi_1)\delta$. Defining $z_t = y_t - \mu - \delta t$, we can solve (1.34) by recursively substituting lagged z_t values as $z_t = \phi_1^t z_0 + \sum_{i=1}^t \phi_1^{t-i}\varepsilon_i$ where z_0 is pre-sample starting value of z_t. When $|\phi_1| < 1$, the impact of z_0 decreases and the effect of shocks dies out in the long run (in other words, the shocks ε_t are transitory). Writing (1.34) as $\Delta_1 z_t = (\phi_1 - 1)z_{t-1} + \varepsilon_t$, where $\Delta_i = (1-L^i)$ denotes a differencing operator, positive and negative values of z_t correspond with y_t being larger or smaller than its (trending) mean $\mu + \delta t$, thus y_t displays so-called *mean- (or trend-) reverting behaviour*. Since the deterministic trend variable t is included in (1.35), the time series y_t is said to be *trend-stationary* and can be described by *deterministic trend* (DT) model.

When $\phi_1 = 1$, there is no mean-reverting behaviour (since $\Delta_1 z_t = \varepsilon_t$) and (1.35) becomes

$$y_t = \delta + y_{t-1} + \varepsilon_t, \tag{1.36}$$

where the trend variable has disappeared. This model is called *random walk* with *drift* δ. Recursive substitution results in

$$y_t = y_0 + \delta t + \sum_{i=1}^{t} \varepsilon_i, \tag{1.37}$$

where the partial sum time series $S_t = \sum_{i=1}^t \varepsilon_i$ is called the *stochastic trend*. Hence, when y_t can be described by (1.36) it has a deterministic trend and a stochastic trend. In order to avoid confusion, when $\phi_1 = 1$, y_t is said

to be described by a stochastic trend (ST) model. Note that shocks have permanent effect.

When ε_t in (1.34) is replaced by $\eta_t = [\phi_{p-1}(L)]^{-1}\varepsilon_t$, where $\phi_{p-1}(L)$ does not contain the component $(1-L)$, all the above results continue to hold. Hence, when a AR(p) polynomial can be decomposed as $\phi_{p-1}(L)(1-L)$, the time series y_t has a stochastic trend. A time series with a stochastic trend can be made stationary by applying the differencing filter Δ_1, therefore in this case y_t is sometimes called *difference-stationary*.

Once again to sum up, we may see from (1.37) that stochastic trend can be accompanied by (any) deterministic trend component. However, the key difference between data generated by (1.35) and (1.36) is that ST series can deviate from this trend for lengthy periods of time because it lacks any mean-reverting forces.

The time series that requires first differencing to remove the stochastic trend is called a time series that is *integrated* of order 1, and denoted $I(1)$. The meaning of the name becomes clear from expansion $y_t = \Delta_1 y_t + y_{t-1} = \Delta_1 y_t + \Delta_1 y_{t-1} + y_{t-2} = \cdots = \Delta_1 y_t + \Delta_1 y_{t-1} + \cdots + \Delta_1 y_{t-k} + y_{t-k-1}$, where $\Delta_1 y_t = y_t - y_{t-1}$. The effect of y_{t-k-1} dies out with large k, so the y_t is obtained by successive summation ("integration") of the mixed process $\Delta_1 y_t$. Further on, $I(2)$ time series needs the Δ_1 filter twice to become stationary (contains two unit roots), etc. In the family of time series linear models there is a class of models named ARIMA(p,d,q), where "I" stands for integrated of order d, to handle such a data that are stationary only after d-th differentiation, i.e $\Delta_1^d y_t$.

Testing for stochastic trend

In order to select between ST and DT model for a given empirical time series y_t, there exists a wide variety of methods. These methods either pay close attention to the $(1-L)$ component in the AR(p) model for y_t or to the relative importance of the stochastic trend component $\sum_{i=1}^{t}\varepsilon_i$. The first set of methods is called tests for unit roots, the second set is stationarity tests.

In order to *test for a unit root*, Dickey and Fuller proposed a simple approach based on the idea that for AR(p) time series with unit root, the sum of AR parameters equals 1. To test the empirical validity of such a parameter restriction, it is useful to decompose the AR polynomial as

$$\phi_p(L) = (1 - \phi_1 - \cdots - \phi_p)L^i + \phi^*_{p-1}(L)(1-L), \tag{1.38}$$

which holds for any $i \in \{1, 2, \ldots p\}$. For illustration, when setting $i = 1$, the AR(2) polynomial is $(1 - \phi_1 L - \phi_2 L^2) = (1 - \phi_1 - \phi_2)L + (\phi^*_0 - \phi^*_1 L)(1-L)$

with $\phi_0^* = 1$ and $\phi_1^* = -\phi_2$. Hence AR(2) model can be rewritten as

$$\phi_{p-1}^*(L)\Delta_1 y_t = (\phi_1 + \phi_2 - 1)y_{t-1} + \varepsilon_t, \tag{1.39}$$

with $\phi_{p-1}^*(L) = (1 - \phi_1^* L)$, or - going further - as

$$\Delta_1 y_t = (\phi_1 + \phi_2 - 1)y_{t-1} + \phi_1^* \Delta_1 y_{t-1} + \varepsilon_t.$$

When $\phi_1 + \phi_2 - 1$ equals zero, (1.39) collapses to an AR(1) model for $\Delta_1 y_t$, in other words it becomes an ARI(1,1) model.

Based on (1.38), the so-called *Augmented Dickey-Fuller* (ADF) test focuses on the statistical relevance of y_{t-1} in the auxiliary regression

$$\Delta_1 y_t = \rho y_{t-1} + \phi_1^* \Delta_1 y_{t-1} + \cdots + \phi_{p-1}^* \Delta_1 y_{t-(p-1)} + \varepsilon_t. \tag{1.40}$$

The null hypothesis is $\rho = 0$ and the relevant alternative is $\rho < 0$, resulting in one-sided test-statistic. For ρ it is the t-test-statistic[2] $t(\hat{\rho})$ commonly referred to as ADF test-statistic. It has nonstandard asymptotic distribution and the critical values have to be obtained through Monte Carlo simulation. Some of the critical values (at 5% significance level) are displayed in Table 1.1. The null hypothesis of unit root is rejected when $t(\hat{\rho})$ is lower than critical value. When the order p in the AR model is selected through sequential t-tests on the ϕ_{p-1}^* to ϕ_1^* parameters in (1.40) (or via an application of AIC or BIC), the same critical values can be used.

Comparing (1.36) with (1.35) we see that the parameter μ for the mean is not identified under the null hypothesis of a unit root, but only under the alternative one. In general, it appears best to include a mean and linear trend in the ADF regression to make the test independent of nuisance parameters. The ADF regression (1.40) then becomes

$$\Delta_1 y_t = \mu^{**} + \delta^{**} t + \rho y_{t-1} + \phi_1^* \Delta_1 y_{t-1} + \cdots + \phi_{p-1}^* \Delta_1 y_{t-(p-1)} + \varepsilon_t. \tag{1.41}$$

Under the null hypothesis not only ρ but also δ^{**} is zero. There exists a joint F-test for $\rho = \delta^{**} = 0$, and in the case of no trend, for $\rho = \mu^{**} = 0$, however a common practice procedure is to test $\rho = 0$ in (1.41) and to consider critical values depending on the type of deterministic regressors included. From Table 1.1 it is clear that these critical values shift to the left. Intuitively, if the data are generated by a random walk model, the inclusion of a trend biases the estimate for ρ away from zero, and hence we need even larger values of the test-statistic to reject the null hypothesis. It was shown (see

[2]Well-known as $t(\hat{\rho}) = \frac{\hat{\rho}}{SE(\hat{\rho})}$ where SE denotes standard error.

reference in [15]) that erroneously neglecting deterministic terms is worse than including redundant variables.

The overall conclusion is that ADF test result should be evaluated with care, in the sense that in case of doubt, we may be better off assuming possible adequacy of both the DT and ST model, and to see which of the two does a better job in out-of-sample forecasting. Further confidence in the empirical outcomes is also obtained when the ADF test results appear robust to changes in the sample size, outliers, additional lags and the inclusion or exclusion of deterministic components.

The test procedure for unit roots compares ST model with DT model. When the null hypothesis of a unit root can not be rejected, the ST model is preferred over the DT model. Contrary to that ST model can be sometimes of most importance as it assumes permanent effects of shocks, in other occasions, however, we may be interested more in DT model hypothesis. A test that takes (trend) stationarity as the null hypothesis is called KPSS test (after Kwiatkovsky, Phillips, Schmidt and Shin). It focuses on the partial sum series $\hat{S}_t = \sum_{i=1}^{t} \hat{e}_i$, where the relevant $\hat{e}_t$ are obtained from an auxiliary regression like $y_t = \hat{\mu} + \hat{\delta}t + \hat{e}_t$. The test-statistic of interest (based on LM-type test) is

$$LM = \frac{1}{n^2 s^2(l)} \sum_{t=1}^{n} \hat{S}_t^2, \tag{1.42}$$

where the scaling factor $s^2(l)$ (so-called long-run variance of $\hat{e}_t$) can be estimated as

$$\hat{s}^2(l) = \frac{1}{n} \sum_{t=1}^{n} \hat{e}_t^2 + \frac{2}{n} \sum_{j=1}^{l} w(j,l) \sum_{t=j+1}^{n} \hat{e}_t \hat{e}_{t-j}, \tag{1.43}$$

where the weights can be of form $w(j,l) = 1 - j/(l+1)$ and the value of l is usually set at $l = \sqrt{n}$. The null hypothesis of (trend) stationarity is rejected when LM exceeds the (asymptotic) critical value given in Table 1.1. The test is one sided.

1.4 Seasonality

When empirical time series originated in nature are observed in some sub-day or sub-year time steps (such es every hour or month), it is often the case that the time series display a seasonal pattern. Similar to the feature of a trend, where definition of a trend depends on the model used to described the trend, there does not exist a very precise definition of seasonality. We may often refer to seasonality when observations follow the same more-or-less smooth

Table 1.1: Critical values for ADF and KPSS tests at 5% significance level

test	deterministic terms	sample size		
		100	500	∞
ADF	none	-1.95	-1.95	-1.95
	constant	-2.89	-2.87	-2.86
	constant and trend	-3.45	-3.42	-3.41
KPSS	constant			0.46
	constant and trend			0.18

(possibly sine-shaped) pattern every S time steps or, on the other hand, when observations in certain seasons display strikingly different features to those in other seasons. So if the seasonal component of the time series can be suitably estimated by a mathematical curve like sinusoid, an AR time series model will be accompanied by a pair of deterministic terms, such like in

$$y_t = \beta_1 sin\left(\frac{2\pi t}{S}\right) + \beta_2 cos\left(\frac{2\pi t}{S}\right) + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad (1.44)$$

where S is number of seasons, or period of seasonal variation. However, many times such a deterministic behaviour is not the case and the model can be of (less restrictive) form

$$y_t = \phi_{0,1} D_{1,t} + \cdots + \phi_{0,S} D_{S,t} + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad (1.45)$$

where $D_{S,t}$ is a seasonal dummy variable. If we denote the number of seasonal cycles in the sample period as N, the number of observations will be $n = SN$ and the dummy variable is defined as

$$D_{s,t} = \begin{cases} 1 & \text{if } t = (T-1)S + s, \\ 0 & \text{otherwise,} \end{cases} \qquad (1.46)$$

with $s = 1, 2, \ldots S$ and $T = 1, 2, \ldots N$. In other words, $D_{s,t}$ takes value 1 in season s and 0 in other seasons. If we consider monthly observations gathered for, e.g, 10 years, then $S = 12$ and $n = 12 \times 10 = 120$. Anyway, varying intercept $\phi_{0,s}$ in (1.45) allows the mean to vary across different seasons. Note that $\mu_s = \phi_{0,s}/(1 - \phi_1 - \cdots - \phi_p)$. If seasonal variation is approximately deterministic, one will find that the estimated means $\hat{\mu}_s \neq \hat{\mu}$, where $\hat{\mu}$ is the estimated mean from an AR(p) model with a single intercept.

Obviously, seasonal feature of a time series can be also of stochastic nature. In the following, we briefly outline two commonly considered models for such seasonal data. The first one assumes that seasonal variation appears in the lag structure while the second model assumes seasonal variation in ARMA parameters.

Loosely speaking, if *seasonal variation* appears *in lags*, a model (e.g. (1.45)) contains y_{t-S}, y_{t-2S}, and so on[3]. Moreover, if the AR parameters in the model are such that the differencing filter Δ_S is required to transform y_t to stationarity, a time series are said to be seasonally integrated. Writing $\Delta_S = (1 - L^S)$ and solving the equation $(1 - z^S) = 0$ or $\exp(Si\phi) = 1$ for z or ϕ, the solutions are equal to 1 and $\cos(2\pi k/S) + i\sin(2\pi k/S)$ for $k = 1, 2 \ldots S-1$ (compare to (1.10)). This amounts to S different solutions, which all lie on the unit circle. The first solution 1 is called nonseasonal unit root and the $S-1$ other solutions are called seasonal unit roots. When a time series has seasonal unit roots, shocks change seasonal pattern permanently.

The alternative seasonal model is a *periodic autoregression* (PAR), which extends a non-periodic AR model by allowing its autoregressive parameters to vary with seasons. In other words, the PAR model assumes that the observations in each of the seasons can be described by a different model. A PAR(p) model can be written as

$$y_t = \mu_s + \phi_{1,s} y_{t-1} + \cdots + \phi_{p,s} y_{t-p} + \varepsilon_t, \tag{1.47}$$

or $\phi_{p,s}(L) y_t = \mu_s + \varepsilon_t$, with $\mu_s = \sum_{s=1}^{S} \phi_{0,s} D_{s,t}$ and $s = 1, 2, \ldots S$. The ε_t is assumed to be standard white noise with constant variance σ^2, although it may also be allowed to have seasonal variances σ_s^2. Since some $\phi_{i,s}$ parameters ($i = 1, 2, \ldots p$) can take zero values, the order p is the maximum of all p_s, where p_s denotes the AR order per season s. There are at least two approaches to modeling PAR time series. The first is to investigate the possible usefulness of periodic models via checking the properties of estimated residuals from non-periodic ones. The second approach is simply to test the estimated parameters of PAR model for periodic variation.

There are several tests for seasonal unit roots available for both seasonal models, however as it is far behind the scope of this thesis, an interested reader is referred to [15]. Details of seasonal adjustment methods (if a removal is desired) are discussed in [9, 15].

[3]This, in fact, implies the inclusion of another member of ARMA family of models, so-called Seasonal ARMA. For completeness, pure SAR(P) is given a form $\phi_P(L^S) y_t = \varepsilon_t$ whilst the full SARMA$(p, q)(P, Q)$, which contains both (stationary) non-seasonal and seasonal part, will be $\phi_p(L)\phi_P(L^S) y_t = \phi_q(L)\theta_Q(L^S)\varepsilon_t$ ([3]), where AR and MA polynomials are built similarly to (1.5).

1.5 Outliers

In most geodetic time series we are quite likely to find (few or plenty of) observations that can be viewed as aberrant. An obvious question then pops up concerning whether an aberrant observation somehow belongs to the time series, in the sense that it is a part of data generating process, or that it should be viewed as a measurement error. Hence, when modelling linear (or nonlinear) data, it is important to study the presence of aberrant observations and their effects on modelling and forecasting. In the following, let's take a look on the three most common representatives, that are additive outlier, innovative outlier and permanent level shift.

In case of an *additive outlier* (AO), the data point is aberrant due to a cause outside the intrinsic nature environment that generates the time series data. Given a y_t, additive outliers cannot be predicted using the historical information set Ω_{t-1}. More formally, if τ denotes the time of outlier occurrence, then

$$y_t = x_t + \omega \, \mathrm{I}[t = \tau], \qquad t = 1, \dots n \tag{1.48}$$

where $\mathrm{I}[\cdot]$ is the usual indicator variable yielding 1 or 0, the time series x_t is uncontaminated but unobserved, while y_t is the observed variable, and the size of AO is denoted by ω. In practice, time τ may be unknown. When we apply OLS to estimate the parameters in, for instance, an AR(1) model for y_t, a neglected AO will have a downward-biasing effect on $\hat{\phi}_1$ (in absolute value). Also, AOs yield large values of skewness and kurtosis because the two observations at time τ and $\tau + 1$ cannot be properly predicted by the model. Finally, the estimated standard error for AR parameter will increase with increasing ω.

Another important type of outlier is the *innovative outlier* (IO), where the outlier occurs in the noise process. Within an ARMA model framework it can be found as $\phi_p(L)y_t = \theta_q(L)(\varepsilon_t + \omega \, \mathrm{I}[t = \tau])$, or more illustratively within AR(1) model

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \omega \, \mathrm{I}[t = \tau]. \tag{1.49}$$

In case the IO is neglected, the forecast error associated with the optimal 1-step-ahead forecast equals $\varepsilon_{\tau,\tau-1} = \varepsilon_\tau + \omega$, expectation of which does not equal to zero and, hence, the predictor for y_τ is biased. However, in contrast to AO, the predictor $\hat{y}_{\tau+1}$ has no bias and the OLS estimate $\hat{\phi}_1$ will be influenced in a much lesser amount.

When ϕ_1 in (1.49) equals 1, IO at time τ can result in a permanent change in the level of a time series. An alternative description of such a *level shift*

in case of an AR model, which does not require that $\phi_1 = 1$, is given by

$$\phi_p(L)y_t = \phi_0 + \omega\, \mathrm{I}[t \geq \tau] + \varepsilon_t, \tag{1.50}$$

where the mean of y_t shifts from $\phi_0/(1 - \phi_1 - \cdots - \phi_p)$ in the first part of the sample to $(\phi_0 + \omega)/(1 - \phi_1 - \cdots - \phi_p)$ in the second part.

In practice, the possible presence of aberrant observations is often indicated by a large value of the JB normality test, see (1.20). There are also other methods, that can be viewed as diagnostic checks for model accuracy, and search over all possible τ for the presence of some type of aberrant data. An alternative approach to guarding against the influence of outliers is to use robust estimation methods to obtain unbiased estimates of a time series model parameters. When a time series seems to have many aberrant data, it is possible that a univariate time series model such as ARMA does not yield a good description of data. In fact, approximating a nonlinear time series model with a linear model may result in many large residuals. Furthermore, outliers may reflect the fact that a multivariate time series model or an AR model with exogenous variables may be more appropriate.

1.6 Multivariate modelling

Univariate time series autoregressive models can be very useful for out-of-sample forecasting and descriptive analysis, however, their empirical specification may be hampered by many outliers and structural shifts, which in turn may be attributed to one or more other variables. It is then desirable to consider, e.g, $\phi_p(L)y_t = \beta x_t + \varepsilon_t$ where β measures the effect of x_t on y_t at time t. Hence, if the estimated residuals do not show typical aberrant data, including only a single variable can substantially reduce the number of parameters since no additional descriptive measures for outliers and structural breaks are needed. Now, the above model brings new questions, mainly whether the x_t is to be included by its present value or with a time lag, i.e. x_{t-k} for some integer k. This time lag will surely depend also on the sampling interval of the data. Another question concerns the reverse causality, whether also x_t cannot somehow depends on current and/or past y_t. When we want to take all possible relations between variables into account, it seems sensible to construct a model for a vector of time series instead of constructing models for all individual series, even in case we are not certain about which variables are exogenous and which endogenous. Such a general (unrestricted) multiple time series model can be a useful starting point of analysis, at least because a static regression model like $y_t = \alpha x_t + u_t$ may lead to spurious inference.

In simple case, the general multivariate model can be of form

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 x_{t-1} + \varepsilon_{y,t} \\ x_t &= \phi_3 y_{t-1} + \phi_4 x_{t-1} + \varepsilon_{x,t} \end{aligned} \quad \text{or} \quad \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_3 & \phi_4 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{y,t} \\ \varepsilon_{x,t} \end{bmatrix}, \quad (1.51)$$

which is called vector autoregression of order 1, VAR(1), since on the right-hand side it includes only y_t and x_t variables with one time lag. Terms $\varepsilon_{y,t}$, $\varepsilon_{x,t}$ stand for corresponding white noise series. If possible simultaneous effects (instead of lagged versions) of x_t on y_t and vice versa are to be allowed, and retaining the same parameter notation, then a so-called dynamic simultaneous model rise up from (1.51),

$$\begin{bmatrix} 1 & -\phi_2 \\ -\phi_3 & 1 \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_4 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{y,t} \\ \varepsilon_{x,t} \end{bmatrix}, \quad (1.52)$$

which after simple rearrangement can be written in form of a VAR(1) model. In fact, any simultaneous equation model with one or more lagged endogenous variables leads to a VAR model.

For practical purposes, the VAR model is often the most useful (particularly for analysing stochastic trends) and in the following we therefore discuss some of its aspects to be used in later chapters. Now, consider a general VAR(p) model abbreviated as

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t \quad (1.53)$$

or

$$\boldsymbol{\Phi}_p(L)\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t, \qquad \text{with} \quad \boldsymbol{\Phi}_p(L) = \boldsymbol{I}_k - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p,$$

where

$$\boldsymbol{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix}, \quad \boldsymbol{\Phi}_i = \begin{bmatrix} \phi_{11,i} & \cdots & \phi_{1k,i} \\ \phi_{21,i} & \cdots & \phi_{2k,i} \\ \vdots & \ddots & \vdots \\ \phi_{k1,i} & \cdots & \phi_{kk,i} \end{bmatrix}, \quad \boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{k,t} \end{bmatrix},$$

$\boldsymbol{y}_t$ is $(k \times 1)$ vector of endogenous time series, $\boldsymbol{\Phi}_i$ is $(k \times k)$ matrix of AR parameters corresponding to i-th lag, and $\boldsymbol{I}_k$ is $(k \times k)$ identity matrix. Vector $\boldsymbol{\varepsilon}_t$ of individually white noise series $\varepsilon_{1,t}$ to $\varepsilon_{k,t}$ follows multinormal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ not necessarily equals $\sigma^2 \boldsymbol{I}_k$ or $\text{diag}(\sigma_1^2, \ldots \sigma_k^2)$. This means that the individual $\varepsilon_{i,t}$ series are uncorrelated with their own past and with the past of the other

$\varepsilon_{j,t}$ variables ($i \neq j$), but that there can be contemporaneous correlation between the error series.

VAR(p) model is said to be stable, or corresponding vector $\boldsymbol{y}_t$ series is stationary, if all solutions to

$$|\boldsymbol{\Phi}_p(z)| = 0, \tag{1.54}$$

lie outside the unit circle. Operator $|\cdot|$ denotes determinant of a matrix. When one or more solutions to (1.54) lie on unit circle, the VAR(p) model contains unit roots. Although testing hypotheses of the unit root presence in VAR polynomial is difficult, there exist some useful methods based on cointegration techniques. To obtain a preliminary and tentative impression of stationarity, we can calculate the eigenvalues of $\sum_{i=1}^{p} \boldsymbol{\Phi}_i$ to see if these are close to unity, which may indicate unit roots in VAR model. Note that finding a unit root in bivariate VAR(1) model means a presence of unit root in each of the nested univariate AR models, hence it seems that they have the unit root in common. The phenomenon of having such a common feature is called a cointegration and will be discussed in chapter 3 in more detail.

Similar to univariate AR, the construction of VAR models involves several specification steps. First, an initial value of p needs to be specified, next we should estimate parameters and investigate the properties of estimated residuals, and finally select between several values of p. For practical purposes, multivariate extensions to ACFs are not very straightforward to interpret since it involves a large system of serial and cross-equation correlations, therefore we usually fit a set of VAR models with orders $1, 2, \ldots p_{max}$ for some value of p_{max}, and then evaluate whether one or more of these models fit well to data. The parameters can be estimated using OLS *per equation*, which gives consistent and efficient estimates. When estimated, some of the VAR parameters may seem insignificant. At this stage, however, it is not sensible to restrict these parameters to zero, unless we are confident about the stationarity of component series of $\boldsymbol{y}_t$, as the t-ratios of the estimates are not distributed as standard normal in case of stochastic trends presence. Model selection is similar to univariate case (though, in practice, performed before examination of the estimated residuals), useful model selection criteria are the multivariate extensions to AIC and BIC given by

$$AIC(p) = n \ln |\hat{\boldsymbol{\Sigma}}_p| + 2k^2 p, \tag{1.55}$$

$$BIC(p) = n \ln |\hat{\boldsymbol{\Sigma}}_p| + \ln(n) k^2 p, \tag{1.56}$$

respectively, where $|\hat{\boldsymbol{\Sigma}}_p|$ denotes the determinant of the residual covariance matrix for the VAR(p) model. These criteria performs well even in case

of unit root contamination. Next step is diagnostic checking on estimated residuals. As mentioned above, this is no easy matter because, besides the white noise properties of individual $\hat{\boldsymbol{\varepsilon}}_t$ series, we must also check whether there are no systematic patterns across current $\hat{\varepsilon}_{i,t}$ and lagged $\hat{\varepsilon}_{j,t}$, for $i \neq j$. Multivariate extensions to the portmanteau and LM tests for detecting serial correlations can be used, however, only to give a warning of dramatical model misspecification, since their power may not be very large. Additional diagnostic tests are available and discussed in the work of several authors, see references in [15].

After the whole investigation it may come out, that some variable is exogenous to key parameters in the model. Imposing exogeneity can imply a reduction of the estimation demands and also improve precision in forecasting. In this context, one can encounter a term *Granger causality* in works of many authors in a sense that it allows us to draw inference on the dynamic impact of one variable on another. In concept of forecastability, a variable (e.g.) $y_{2,t}$ is said to be Granger-non-causal for (e.g.) $y_{1,t}$ if $\mathrm{E}[y_{1,t}|\Omega_{1,t-1}, \Omega_{2,t-1}] = \mathrm{E}[y_{1,t}|\Omega_{1,t-1}]$, that is, the past of $y_{2,t}$ does not help in forecasting $y_{1,t}$. In bivariate AR(1) model this implies that $\phi_{12,1} = 0$.

Forecasting from a VAR(p) is a straightforward extension of forecasting from an AR(p). Consider (stationary) VAR(p) process at time $(t+h)$ rewritten into VMA(∞) or multivariate *Wold representation* [4]

$$\boldsymbol{y}_{t+h} = \boldsymbol{\varepsilon}_{t+h} + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t+h-1} + \cdots + \boldsymbol{\Psi}_{h-1}\boldsymbol{\varepsilon}_{t+1} + \boldsymbol{\Psi}_h\boldsymbol{\varepsilon}_t + \cdots = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_i\boldsymbol{\varepsilon}_{t+h-i}, \quad (1.57)$$

$\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then the optimal h-step-ahead forecast of $\boldsymbol{y}_{t+h}$ at time t is

$$\hat{\boldsymbol{y}}_{t+h|t} = \mathrm{E}[\boldsymbol{y}_{t+h}|\boldsymbol{\Omega}_t] = \boldsymbol{\Psi}_h\boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_{h+1}\boldsymbol{\varepsilon}_{t-1} + \ldots \quad (1.58)$$

where $\boldsymbol{\Omega}_t$ denotes history of $\boldsymbol{y}_t$ up to and including the observation at time t. The forecast (prediction) error is given by

$$\boldsymbol{e}_{t+h|t} = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\varepsilon}_{t+h} + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t+h-1} + \cdots + \boldsymbol{\Psi}_{h-1}\boldsymbol{\varepsilon}_{t+1} \quad (1.59)$$

and it's covariance matrix (denoted as squared prediction error)

$$\boldsymbol{SPE}(h) = \mathrm{E}[\boldsymbol{e}_{t+h|t}\boldsymbol{e}'_{t+h|t}] = \sum_{i=0}^{h-1} \boldsymbol{\Psi}_i\boldsymbol{\Sigma}\boldsymbol{\Psi}'_i. \quad (1.60)$$

[4] From stationary VAR(p) model $\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t$ we may write $\boldsymbol{y}_t = \boldsymbol{\Phi}(L)^{-1}\boldsymbol{\varepsilon}_t = \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t = \sum_{i=0}^{\infty}\boldsymbol{\Psi}_i\boldsymbol{\varepsilon}_{t-i}$, where obviously $\boldsymbol{\Psi}_0 = \boldsymbol{I}$ and $\lim_{i\to\infty}\boldsymbol{\Psi}_i = 0$. The Wold coefficients $\boldsymbol{\Psi}_i$ may be determined from the VAR coefficients $\boldsymbol{\Phi}_i$ by solving $\boldsymbol{\Phi}(L)\boldsymbol{\Psi}(L) = \boldsymbol{I}$, which implies $\boldsymbol{\Psi}_1 = \boldsymbol{\Phi}_1$, $\boldsymbol{\Psi}_2 = \boldsymbol{\Phi}_1\boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_2$, generally $\boldsymbol{\Psi}_i = \boldsymbol{\Phi}_1\boldsymbol{\Psi}_{i-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Psi}_{i-p}$.

If $\boldsymbol{\varepsilon}_t$ follows normal distribution and $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}_s$ are uncorrelated for $t \neq s$, then confidence interval for h-step-ahead forecast of j-th process in $\boldsymbol{y}_t$ $(j = 1, \ldots k)$ is bounded by $\hat{y}_{j,t+h|t} \pm u_{1-\alpha/2}\sigma_j(h)$, where α denotes significance level, u_q stands for q-th quantile of standard normal distribution and $\sigma_j(h)$ is square root of j-th diagonal element of $\boldsymbol{SPE}(h)$. Once estimates of VAR(p) model parameters are available, forecast for horizon h may be computed using a "chain rule"

$$\hat{\boldsymbol{y}}_{t+h|t} = \hat{\boldsymbol{\Phi}}_1\hat{\boldsymbol{y}}_{t+h-1|t} + \cdots + \hat{\boldsymbol{\Phi}}_p\hat{\boldsymbol{y}}_{t+h-p|t}, \tag{1.61}$$

where $\hat{\boldsymbol{y}}_{t+i|t} = \boldsymbol{y}_{t+i}$ for $i \leq 0$.

To compare rival empirical models we may consider the determinant or the trace of the $\boldsymbol{SPE}(h)$ matrices , since the forecast errors for $y_{j,t}$ are also affected by the forecasts for the other $k-1$ variables in $\boldsymbol{y}_t$.

In practice, having m additional observations at disposal, two models, say A and B, can be efficiently compared by setting

$$d_j = g(\boldsymbol{e}_{t+j|t+j-h,A}) - g(\boldsymbol{e}_{t+j|t+j-h,B}) \quad j = 1, 2, \ldots m \tag{1.62}$$

with $g(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{a}$, instead of univariate version (1.28), and applying (modified) Diebold-Mariano test (1.32).