3

Linear Innovations State Space Models

In Chap. 2, state space models were introduced for all 15 exponential smoothing methods. Six of these involved only linear relationships, and so are "linear innovations state space models." In this chapter, we consider linear innovations state space models, including the six linear models of Chap. 2, but also any other models of the same form. The advantage of working with the general framework is that estimation and prediction methods for the general model automatically apply to the six special cases in Chap. 2 and other cases conforming to its structure. There is no need to derive these results on a case by case basis.

The general linear innovations state space model is introduced in Sect. 3.1. Section 3.2 provides a simple algorithm for computing the one-step prediction errors (or innovations); it is this algorithm which makes innovations state space models so appealing. Some of the properties of the models, including stationarity and stability, are discussed in Sect. 3.3. In Sect. 3.4 we discuss some basic innovations state space models that were introduced briefly in Chap. 2. Interesting variations on these models are considered in Sect. 3.5.

3.1 The General Linear Innovations State Space Model

In a state space model, the observed time series variable y_t is supplemented by unobserved auxiliary variables called *states*. We represent these auxiliary variables in a single vector $\boldsymbol{x}_t$, which is called the *state vector*. The state vector is a parsimonious way of summarizing the past behavior of the time series y_t, and then using it to determine the effect of the past on the present and future behavior of the time series.

The general[1] *linear innovations state space model* is

$$y_t = \boldsymbol{w}'\boldsymbol{x}_{t-1} + \varepsilon_t, \tag{3.1a}$$

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{g}\varepsilon_t, \tag{3.1b}$$

where y_t denotes the observed value at time t and $\boldsymbol{x}_t$ is the state vector. This is a special case of the more general model (2.12). In exponential smoothing, the state vector contains information about the level, growth and seasonal patterns. For example, in a model with trend and seasonality, $\boldsymbol{x}_t = (\ell_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})'$.

From a mathematical perspective, the state variables are essentially redundant. In Chap. 11, it will be shown that the state variables contained in the state vector can be substituted out of the equations in which they occur to give a *reduced form* of the model. So why use state variables at all? They help us to define large complex models by first breaking them into smaller, more manageable parts, thus reducing the chance of model specification errors. Further, the components of the state vector enable us to gain a better understanding of the structure of the series, as can be seen from Table 2.1. In addition, this structure enables us to explore the need for each component separately and thereby to carry out a systematic search for the best model.

Equation (3.1a) is called the *measurement equation*. The term $\boldsymbol{w}'\boldsymbol{x}_{t-1}$ describes the effect of the past on y_t. The error term ε_t describes the unpredictable part of y_t. It is usually assumed to be from a Gaussian white noise process with variance σ^2. Because ε_t represents what is new and unpredictable, it is referred to as the *innovation*. The innovations are the only source of randomness for the observed time series, $\{y_t\}$.

Equation (3.1b) is known as the *transition equation*. It is a first-order recurrence relationship that describes how the state vectors evolve over time. $\boldsymbol{F}$ is the *transition matrix*. The term $\boldsymbol{F}\boldsymbol{x}_{t-1}$ shows the effect of the past on the current state $\boldsymbol{x}_t$. The term $\boldsymbol{g}\varepsilon_t$ shows the unpredictable change in $\boldsymbol{x}_t$. The vector $\boldsymbol{g}$ determines the extent of the effect of the innovation on the state. It is referred to as a *persistence vector*. The transition equation is the mechanism for creating the inter-temporal dependencies between the values of a time series.

The k-vectors $\boldsymbol{w}$ and $\boldsymbol{g}$ are fixed, and $\boldsymbol{F}$ is a fixed $k \times k$ matrix. These fixed components usually contain some parameters that need to be estimated.

The seed value $\boldsymbol{x}_0$ for the transition equation may be fixed or random. The process that generates the time series may have begun before period 1, but data for the earlier periods are not available. In this situation, the start-up time of the process is taken to be $-\infty$, and $\boldsymbol{x}_0$ must be random. We say that the *infinite start-up assumption* applies. This assumption is typically valid in the study of economic variables. An economy may have been operating for many centuries but an economic quantity may not have been measured until relatively recent times. Consideration of this case is deferred to Chap. 12.

[1] An even more general form is possible by allowing $\boldsymbol{w}$, $\boldsymbol{F}$ and $\boldsymbol{g}$ to vary with time, but that extension will not be considered here.

Alternatively, the process that generates a time series may have started at the beginning of period 1, and $\boldsymbol{x}_0$ is then fixed. In this case we say that the *finite start-up assumption* applies. For example, if y_t is the demand for an inventory item, the start-up time corresponds to the date at which the product is introduced. The theory presented in this and most subsequent chapters is based on the finite start-up assumption with fixed $\boldsymbol{x}_0$.

Upon further consideration, we see that even when a series has not been observed from the outset, we may choose to condition upon the state variables at time zero. We then employ the finite start-up assumption with fixed $\boldsymbol{x}_0$.

Model (3.1) is often called the *Gaussian innovations state space model* because it is defined in terms of innovations that follow a Gaussian distribution. It may be contrasted with alternative state space models, considered in Chap. 13, which involve different and uncorrelated sources of randomness in (3.1a) and (3.1b), rather than a single source of randomness (the innovations) in each case.

The probability density function for $\boldsymbol{y} = [y_1, \ldots, y_n]$ is a function of the innovations and has the relatively simple form

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{x}_0) &= \prod_{t=1}^{n} p(y_t \mid y_1, \ldots, y_{t-1}, \boldsymbol{x}_0) \\ &= \prod_{t=1}^{n} p(y_t \mid \boldsymbol{x}_{t-1}) \\ &= \prod_{t=1}^{n} p(\varepsilon_t). \end{aligned}$$

If we assume that the distribution is Gaussian, this expression becomes:

$$p(\boldsymbol{y} \mid \boldsymbol{x}_0) = (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2}\sum_{t=1}^{n} \varepsilon_t^2/\sigma^2\right). \tag{3.2}$$

This is easily evaluated provided we can compute the innovations $\{\varepsilon_t\}$. A simple expression for this computation is given in the next section.

3.2 Innovations and One-Step-Ahead Forecasts

If the value for $\boldsymbol{x}_0$ is known, the innovation ε_t is a one-step-ahead prediction error. This can be seen by applying (3.1a) and (3.1b) to obtain

$$\mathrm{E}(y_t \mid y_{t-1}, \ldots, y_1, \boldsymbol{x}_0) = \mathrm{E}(y_t \mid \boldsymbol{x}_{t-1}) = \boldsymbol{w}'\boldsymbol{x}_{t-1}.$$

Then the prediction of y_t, given the initial value $\boldsymbol{x}_0$ and observations $y_1, \ldots, y_{t-1}$, is $\boldsymbol{w}'\boldsymbol{x}_{t-1}$. If we denote the prediction by $\hat{y}_{t|t-1}$, the innovations can be computed recursively from the series values using the relationships

$$\hat{y}_{t|t-1} = \boldsymbol{w}'\boldsymbol{x}_{t-1}, \tag{3.3a}$$

$$\varepsilon_t = y_t - \hat{y}_{t|t-1}, \tag{3.3b}$$

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{g}\varepsilon_t. \tag{3.3c}$$

This transformation will be called *general exponential smoothing*. It was first outlined by Box and Jenkins (Box et al. 1994, pp. 176–180) in a much overlooked section of their book.

The forecasts obtained with this transformation are linear functions of past observations. To see this, first substitute (3.3a) and (3.3b) into (3.3c) to find

$$\boldsymbol{x}_t = \boldsymbol{D}\boldsymbol{x}_{t-1} + \boldsymbol{g}y_t, \tag{3.4}$$

where $\boldsymbol{D} = \boldsymbol{F} - \boldsymbol{g}\boldsymbol{w}'$. Then back-solve the recurrence relationship (3.4) to give

$$\boldsymbol{x}_t = \boldsymbol{D}^t\boldsymbol{x}_0 + \sum_{j=0}^{t-1} \boldsymbol{D}^j\boldsymbol{g}y_{t-j}. \tag{3.5}$$

This result indicates that the current state $\boldsymbol{x}_t$ is a linear function of the seed state $\boldsymbol{x}_0$ and past and present values of the time series. Finally, substitute (3.5), lagged by one period, into (3.3a) to give

$$\hat{y}_{t|t-1} = a_t + \sum_{j=1}^{t-1} c_j y_{t-j}, \tag{3.6}$$

where $a_t = \boldsymbol{w}'\boldsymbol{D}^{t-1}\boldsymbol{x}_0$ and $c_j = \boldsymbol{w}'\boldsymbol{D}^{j-1}\boldsymbol{g}$. Thus, the forecast is a linear function of the past observations and the seed state vector.

Equations (3.1), (3.3), and (3.4) demonstrate the beauty of the innovations approach. We may start from the state space model in (3.1) and generate the one-step-ahead forecasts directly using (3.3). When a new observation becomes available, the state vector is updated using (3.4), and the new one-step-ahead forecast is immediately available. As we shall see in Chap. 13, other approaches achieve the updating and the transition from model to forecast function with less transparency and considerably more effort.

3.3 Model Properties

3.3.1 Stability and Forecastability

When the forecasts of y_t are unaffected by observations in the distant past, we describe the model as *forecastable*. Specifically, a forecastable model has the properties

$$\sum_{j=1}^{\infty} |c_j| < \infty \qquad \text{and} \qquad \lim_{t\to\infty} a_t = a. \tag{3.7}$$

Our definition of forecastability allows the initial state $\boldsymbol{x}_0$ to have an ongoing effect on forecasts, but it prevents observations in the distant past having any effect. In most cases, $a = 0$, but not always; an example with $a \neq 0$ is given in Sect. 3.5.2.

A sufficient, but not necessary, condition for (3.7) to hold is that the eigenvalues of $\boldsymbol{D}$ lie inside the unit circle. In this case, $\boldsymbol{D}^j$ converges to a null matrix as j increases. This is known as the "stability condition" and such models are called *stable*. $\boldsymbol{D}$ is called the *discount matrix*. In a stable model, the coefficients of the observations in (3.6) decay exponentially. The exponential decline in the importance of past observations is a property that is closely associated with exponential smoothing.

It turns out that sometimes a_t converges to a constant and the coefficients $\{c_j\}$ converge to zero even when $\boldsymbol{D}$ has a unit root. In this case, the forecasts of y_t are unaffected by distant observations, while the forecasts of $\boldsymbol{x}_t$ may be affected by distant past observations even for large values of t. Thus, any stable model is also forecastable, but some forecastable models are not stable. Examples of unstable but forecastable models are given in Chap. 10. The stability condition on $\boldsymbol{D}$ is closely related to the invertibility restriction for ARIMA models; this is discussed in more detail in Chap. 11.

3.3.2 Stationarity

The other matrix that controls the model properties is the *transition matrix*, $\boldsymbol{F}$. If we iterate (3.1b), we obtain

$$\begin{aligned}\boldsymbol{x}_t &= \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{g}\varepsilon_t \\ &= \boldsymbol{F}^2\boldsymbol{x}_{t-2} + \boldsymbol{F}\boldsymbol{g}\varepsilon_{t-1} + \boldsymbol{g}\varepsilon_t \\ &\vdots \\ &= \boldsymbol{F}^t\boldsymbol{x}_0 + \sum_{j=0}^{t-1}\boldsymbol{F}^j\boldsymbol{g}\varepsilon_{t-j}.\end{aligned}$$

Substituting this result into (3.1a) gives

$$y_t = d_t + \sum_{j=0}^{t-1} k_j\varepsilon_{t-j}, \tag{3.8}$$

where $d_t = \boldsymbol{w}'\boldsymbol{F}^{t-1}\boldsymbol{x}_0$, $k_0 = 1$ and $k_j = \boldsymbol{w}'\boldsymbol{F}^{j-1}\boldsymbol{g}$ for $j = 1, 2, \ldots$. Thus, the observation is a linear function of the seed state $\boldsymbol{x}_0$ and past and present errors. Any linear innovations model may be represented in the form (3.8); this is an example of a finite Wold decomposition (Brockwell and Davis 1991, p. 180).

The model is described as *stationary*[2] if

$$\sum_{j=0}^{\infty} |k_j| < \infty \qquad \text{and} \qquad \lim_{t \to \infty} d_t = d. \tag{3.9}$$

In such a model, the coefficients of the errors in (3.8) converge rapidly to zero, and the impact of the seed state vector diminishes over time.

We may then consider the limiting form of the model, corresponding to the infinite start-up assumption. Equation (3.8) becomes

$$y_t = d + \sum_{j=0}^{\infty} k_j \varepsilon_{t-j}.$$

This form is known as the Wold decomposition for a stationary series. It follows directly that $\mathrm{E}(y_t) = d$ and $\mathrm{V}(y_t) = \sigma^2 \sum_{j=0}^{\infty} k_j^2$.

A sufficient, but not necessary, condition for stationarity to hold is for the absolute value of each eigenvalue of $\boldsymbol{F}$ to lie strictly in the unit interval $(0, 1)$. Then $\boldsymbol{F}^j$ converges to a null matrix as j increases. As with the stability property, it turns out that sometimes d_t converges to a constant and the coefficients $\{k_j\}$ converge to zero even when $\boldsymbol{F}$ has a unit root. However, this does not occur with any of the models we consider, and so it will not be discussed further.

Stationarity is a rare property in exponential smoothing state space models. None of the models discussed in Chap. 2 are stationary. The six linear models described in that chapter have at least one unit root for the $\boldsymbol{F}$ matrix. However, it is possible to define stationary models in the exponential smoothing framework; an example of such a model is given in Sect. 3.5.1, where all of the transition equations involve damping.

3.4 Basic Special Cases

The linear innovations state space model effectively contains an infinite number of special cases that can potentially be used to model a time series; that is, to provide a stochastic approximation to the data generating process of a time series. However, in practice we use only a handful of special cases that possess the capacity to represent commonly occurring patterns such as trends, seasonality and business cycles. Many of these special cases were introduced in Chap. 2.

The simplest special cases are based on polynomial approximations of continuous real functions. A polynomial function can be used to approximate any real function in the neighborhood of a specified point (this is known

[2] The terminology "stationary" arises because the distribution of $(y_t, y_{t+1}, \ldots, y_{t+s})$ does not depend on time t when the initial state $\boldsymbol{x}_0$ is random.

as Taylor's theorem in real analysis). To demonstrate the idea, we temporarily take the liberty of representing the data by a continuous path, despite the fact that business and economic data are typically collected at discrete points of time.

The first special case to be considered, the local level model, is a zero-order polynomial approximation. As depicted in Fig. 3.1a, at any point along the data path, the values in the neighborhood of the point are approximated by a short flat line representing what is referred to as a local level. As its height changes over time, it is necessary to approximate the data path by many local levels. Thus, the local level effectively represents the state of a process generating a time series.

The gap between successive levels is treated as a random variable. Moreover, this random variable is assumed to have a Gaussian distribution that has a zero mean to ensure that the level is equally likely to go up or down.

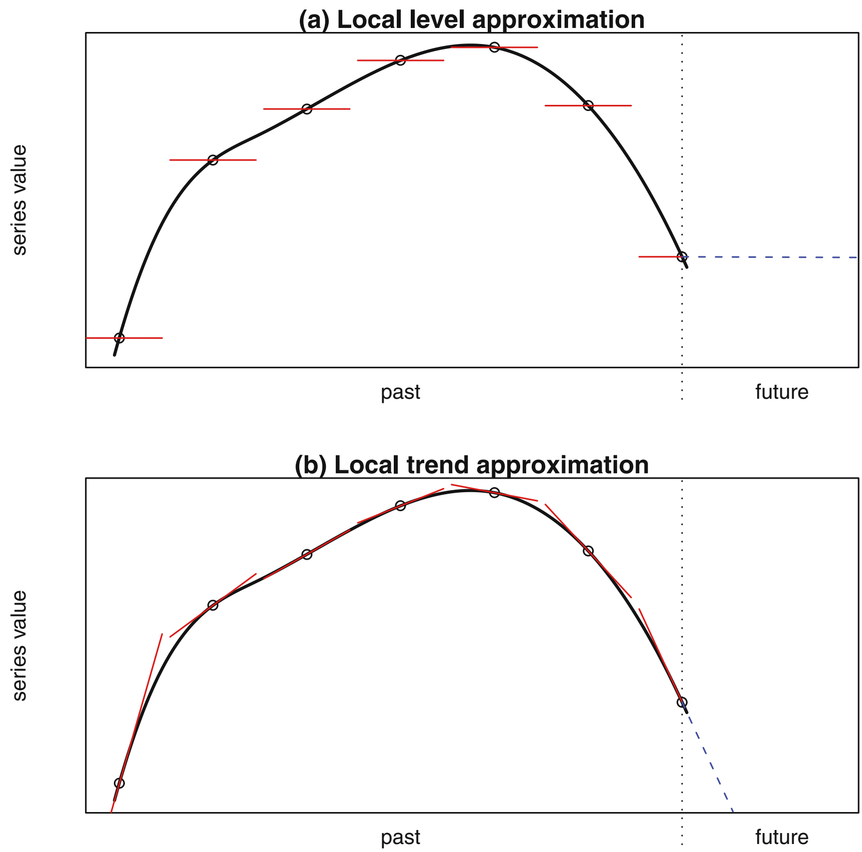

Fig. 3.1. Schematic representation of (**a**) a local level model; and (**b**) a local trend model.

The final local level is projected into the future to give predictions. As the approximation is only effective in a small neighborhood, predictions generated this way are only likely to be reliable in the shorter term.

The second special case involves a first-order polynomial approximation. At each point, the data path is approximated by a straight line. In the deterministic world of analysis, this line would be tangential to the data path at the selected point. In the stochastic world of time series data, it can only be said that the line has a similar height and a similar slope to the data path. Randomness means that the line is not exactly tangential. The approximating line changes over time, as depicted in Fig. 3.1b, to reflect the changing shape of the data path. The state of the process is now summarized by the level and the slope at each point of the path. The stochastic representation is based on the assumption that the gaps between successive slopes are Gaussian random variables with a zero mean. Note that the prediction is obtained by projecting the last linear approximation into the future.

It is possible to move beyond linear functions to higher order polynomials with quadratic or cubic terms. However, these extensions are rarely used in practice. It is commonly thought that the randomness found in real time series typically swamps and hides the effects of curvature.

Another strategy that does often bear fruit is the search for periodic behavior in time series caused by seasonal effects. Ignoring growth for the moment, the level in a particular month may be closer to the level in the corresponding month in the previous year than to the level in the preceding month. This leads to seasonal state space models.

3.4.1 Local Level Model: ETS(A,N,N)

The simplest way to transmit the history of a process is through a single state, ℓ_t, called the level. The resulting state space model is defined by the equations

$$y_t = \ell_{t-1} + \varepsilon_t, \tag{3.10a}$$

$$\ell_t = \ell_{t-1} + \alpha\varepsilon_t, \tag{3.10b}$$

where $\varepsilon_t \sim \text{NID}(0, \sigma^2)$. It conforms to a state space structure with $x_t = \ell_t$, $w = 1$, $F = 1$ and $g = \alpha$. The values that are generated by this stochastic model are randomly scattered about the (local) levels as described in (3.10a). This is illustrated in Fig. 3.2 with a simulated series.

In demand applications, the level ℓ_{t-1} represents the anticipated demand for period t, and ε_t represents the unanticipated demand. Changes to the underlying level may be induced by changes in the customer base such as the arrival of new customers, or by new competitors entering the market. Changes like these transcend a single period and must affect the underlying level. It is assumed that the unanticipated demand includes a persistent and

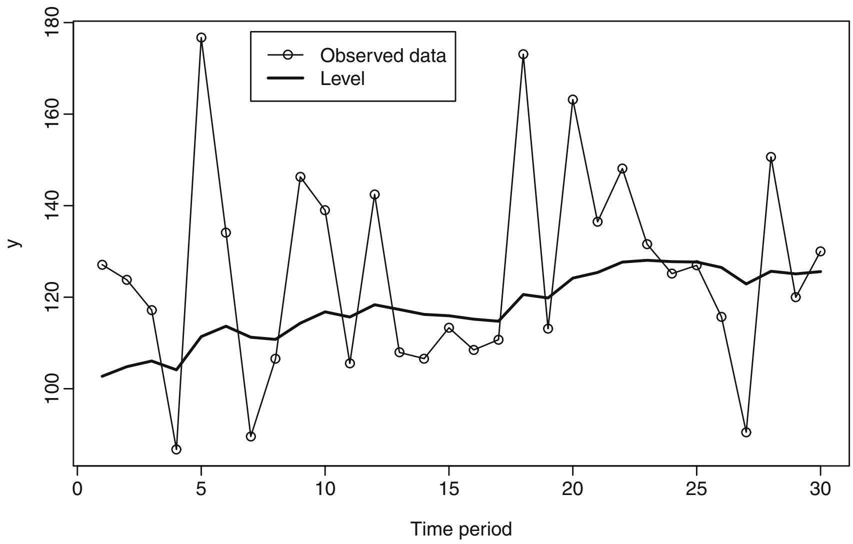

Fig. 3.2. Simulated series from the ETS(A,N,N) model. Here $\alpha = 0.1$ and $\sigma = 5$.

a temporary effect; $\alpha\varepsilon_t$ denotes the persistent effect, feeding through to future periods via the (local) levels governed by (3.10b).

The degree of change of successive levels is governed by the size of the smoothing parameter α. The cases where $\alpha = 0$ and $\alpha = 1$ are of special interest.

Case: $\alpha = 0$ The local levels do not change at all when $\alpha = 0$. Their common level is then referred to as the global level. Successive values of the series y_t are independently and identically distributed. Its moments do not change over time.

Case: $\alpha = 1$ The model reverts to a random walk $y_t = y_{t-1} + \varepsilon_t$. Successive values of the time series y_t are clearly dependent.

The special case of transformation (3.3) for model (3.10) is

$$\begin{aligned} \hat{y}_{t|t-1} &= \ell_{t-1}, \\ \varepsilon_t &= y_t - \ell_{t-1}, \\ \ell_t &= \ell_{t-1} + \alpha\varepsilon_t. \end{aligned}$$

It corresponds to simple exponential smoothing (Brown 1959), one of the most widely used methods of forecasting in business applications. It is a simple recursive scheme for calculating the innovations from the raw data. Equation (3.4) reduces to

$$\ell_t = (1 - \alpha)\ell_{t-1} + \alpha y_t. \tag{3.11}$$

The one-step-ahead predictions obtained from this scheme are linearly dependent on earlier series values. Equation (3.6) indicates that

$$\hat{y}_{t+1|t} = (1-\alpha)^t \ell_0 + \alpha \sum_{j=0}^{t-1} (1-\alpha)^j y_{t-j}. \tag{3.12}$$

This is a linear function of the data and seed level. Ignoring the first term (which is negligible for large values of t and $|1-\alpha| < 1$), the prediction $\hat{y}_{t|t-1}$ is an *exponentially weighted average* of past observations. The coefficients depend on the "discount factor" $1-\alpha$. If $|1-\alpha| < 1$, then the coefficients become smaller as j increases. That is, the stability condition is satisfied if and only if $0 < \alpha < 2$. The coefficients are positive if and only if $0 < (1-\alpha) < 1$, and (3.11) can then be interpreted as a weighted average of the past level ℓ_{t-1} and the current series value y_t. Thus, the prediction can only be interpreted as a weighted average if $0 < \alpha < 1$.

Consequently, there are two possible ranges for α that have been proposed: $0 < \alpha < 2$ on the basis of a stability argument, and $0 < \alpha < 1$ on the basis of an interpretation as a weighted average. The narrower range is widely used in practice.

The impact of various values of α may be discerned from Fig. 3.3. It shows simulated time series from an ETS(A,N,N) model with $\ell_0 = 100$ and $\sigma = 5$ for various values of α. The same random number stream from a Gaussian distribution was used for the three series, so that any perceived differences can be attributed entirely to changes in α. For the case $\alpha = 0.1$, the underlying level is reasonably stable. The plot has a jagged appearance because there is a tendency for the series to switch direction between successive observations. This is a consequence of the fact, shown in Chap. 11, that successive first-differences of the series, Δy_t and Δy_{t-1}, are negatively correlated when α is restricted to the interval $(0,1)$. When $\alpha = 0.5$, the underlying level displays a much greater tendency to change. There is still a tendency for successive observations to move in opposite directions. In the case $\alpha = 1.5$, there is an even greater tendency for the underlying level to change. However, the series is much smoother. This reflects the fact, also established in Chap. 11, that successive first-differences of the series are positively correlated for cases where α lies in the interval $(1,2)$.

3.4.2 Local Trend Model: ETS(A,A,N)

The local level model can be augmented by a growth rate b_t to give

$$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t, \tag{3.13a}$$
$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t, \tag{3.13b}$$
$$b_t = b_{t-1} + \beta\varepsilon_t, \tag{3.13c}$$

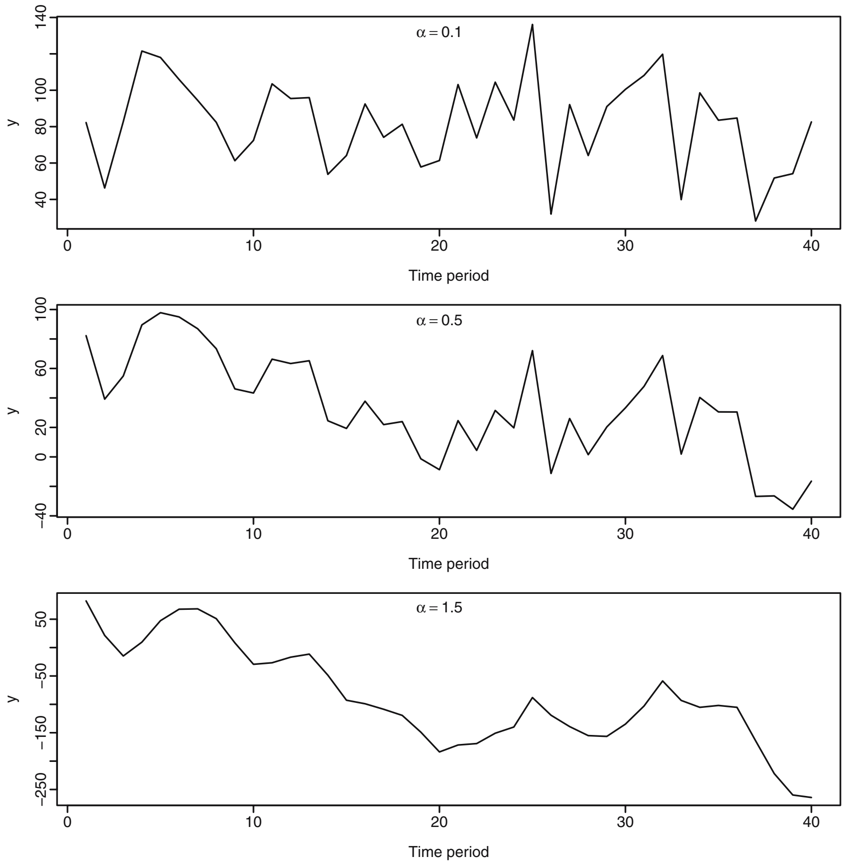

Fig. 3.3. Comparison of simulated time series from a local level model. Here $\sigma = 5$.

where there are now two smoothing parameters α and β. The growth rate (or slope) b_t may be positive, zero or negative. Model (3.13) has a state space structure with

$$\boldsymbol{x}_t = \begin{bmatrix}\ell_t \ b_t\end{bmatrix}', \quad \boldsymbol{w} = \begin{bmatrix}1 \ 1\end{bmatrix}', \quad \boldsymbol{F} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \quad \text{and} \quad \boldsymbol{g} = \begin{bmatrix}\alpha \ \beta\end{bmatrix}'.$$

The size of the smoothing parameters reflects the impact of the innovations on the level and growth rate. Figure 3.4 shows simulated values from the model for different settings of the smoothing parameters. When $\beta = 0$, the growth rate is constant over time. If, in addition, $\alpha = 0$, the level changes at a constant rate over time. That is, there is no random change in the level or growth. This case will be called a *global trend*. The constant growth rate is sometimes interpreted as a long-term growth rate. For other values of the smoothing parameters, the growth rate follows a random walk over time. As

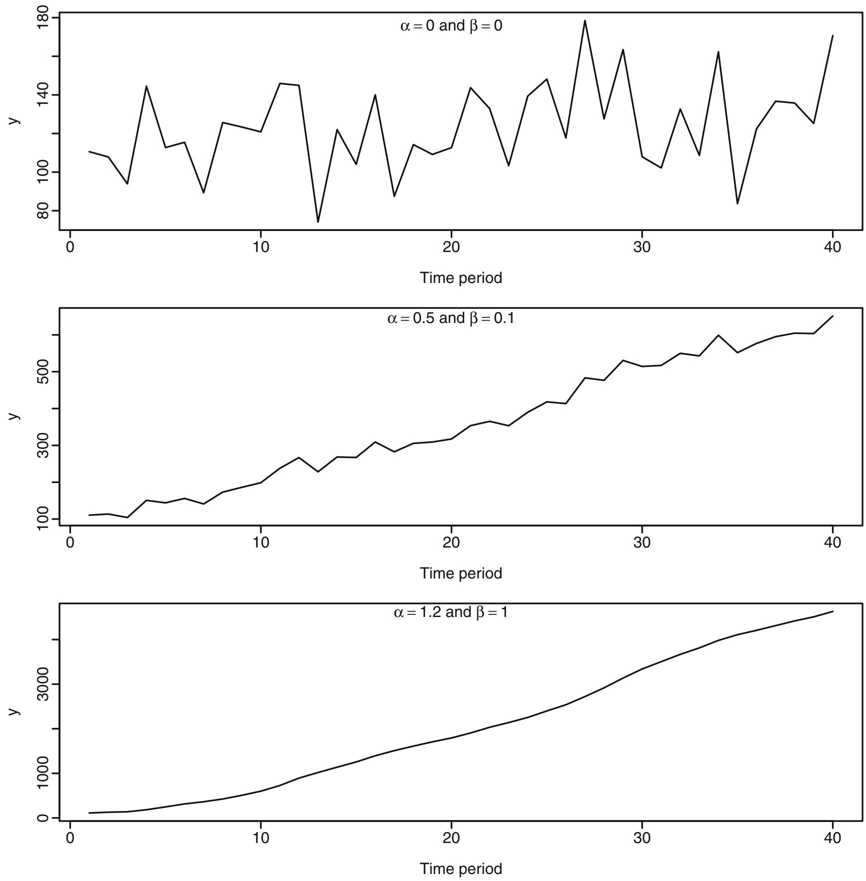

Fig. 3.4. Comparison of simulated time series from a local trend model. Here $\sigma = 5$.

the smoothing parameters increase in size, there is a tendency for the series to become smoother.

For this model, the transformation (3.3) of series values into innovations becomes

$$\begin{aligned}\hat{y}_{t|t-1} &= \ell_{t-1} + b_{t-1},\\ \varepsilon_t &= y_t - \hat{y}_{t|t-1},\\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t,\\ b_t &= b_{t-1} + \beta\varepsilon_t.\end{aligned}$$

This corresponds to Holt's linear exponential smoothing (Holt 1957). An equivalent system of equations is

$$\hat{y}_{t|t-1} = \ell_{t-1} + b_{t-1}, \tag{3.14a}$$

$$\varepsilon_t = y_t - \hat{y}_{t|t-1}, \tag{3.14b}$$

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}), \tag{3.14c}$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}, \tag{3.14d}$$

where $\beta^* = \beta/\alpha$. The term $\ell_t - \ell_{t-1}$ is often interpreted as the "actual growth" as distinct from the predicted growth b_{t-1}.

Equations (3.14c) and (3.14d) may be interpreted as weighted averages if $0 < \alpha < 1$ and $0 < \beta^* < 1$, or equivalently, if $0 < \alpha < 1$ and $0 < \beta < \alpha$. These restrictions are commonly applied in practice. Alternatively, it can be shown (see Chap. 10) that the model is stable (i.e., the discount matrix $\boldsymbol{D}^j$ converges to $\mathbf{0}$ as j increases) when $\alpha > 0$, $\beta > 0$ and $2\alpha + \beta < 4$. This provides a much larger parameter region than is usually allowed.

3.4.3 Local Additive Seasonal Model: ETS(A,A,A)

For time series that exhibit seasonal patterns, the local trend model can be augmented by seasonal effects, denoted by s_t. Often the structure of the seasonal pattern changes over time in response to changes in tastes and technology. For example, electricity demand used to peak in winter, but in some locations it now peaks in summer due to the growing prevalence of air conditioning. Thus, the formulae used to represent the seasonal effects should allow for the possibility of changing seasonal patterns. The ETS(A,A,A) model is

$$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t, \tag{3.15a}$$

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t, \tag{3.15b}$$

$$b_t = b_{t-1} + \beta\varepsilon_t, \tag{3.15c}$$

$$s_t = s_{t-m} + \gamma\varepsilon_t. \tag{3.15d}$$

This model corresponds to the first-order state space model where

$$\boldsymbol{w}' = \begin{bmatrix}1 \; 1 \; 0 \cdots 0 \; 1\end{bmatrix},$$

$$\boldsymbol{x}_t = \begin{bmatrix} \ell_t \\ b_t \\ s_t \\ s_{t-1} \\ \vdots \\ s_{t-m+1} \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{g} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Careful inspection of model (3.15) shows that the level and seasonal terms are confounded. If an arbitrary quantity δ is added to the seasonal elements and subtracted from the level, the following equations are obtained

$$\begin{aligned}
y_t &= (\ell_{t-1} - \delta) + b_{t-1} + (s_{t-m} + \delta) + \varepsilon_t, \\
\ell_t - \delta &= \ell_{t-1} - \delta + b_{t-1} + \alpha\varepsilon_t, \\
b_t &= b_{t-1} + \beta\varepsilon_t, \\
(s_t + \delta) &= (s_{t-m} + \delta) + \gamma\varepsilon_t,
\end{aligned}$$

which is equivalent to (3.15). To avoid this problem, it is desirable to constrain the seasonal component so that any sequence $\{s_t, s_{t+1}, \ldots, s_{t+m-1}\}$ sums to zero (or at least has mean zero). The seasonal components are said to be *normalized* when this condition is true. Normalization of seasonal factors involves a subtle modification of the model and will be addressed in Chap. 8. In the meantime, we can readily impose the constraint that the seasonal factors in the initial state $\boldsymbol{x}_0$ must sum to zero. This means that the seasonal components start off being normalized, although there is nothing to constrain them from drifting away from zero over time.

The transformation from series values to prediction errors can be shown to be

$$\begin{aligned}
\hat{y}_{t|t-1} &= \ell_{t-1} + b_{t-1} + s_{t-m}, \\
\varepsilon_t &= y_t - \hat{y}_{t|t-1}, \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t, \\
b_t &= b_{t-1} + \beta\varepsilon_t, \\
s_t &= s_{t-m} + \gamma\varepsilon_t.
\end{aligned}$$

This corresponds to a commonly used additive version of seasonal exponential smoothing (Winters 1960). An equivalent form of these transition equations is

$$\hat{y}_{t|t-1} = \ell_{t-1} + b_{t-1} + s_{t-m}, \tag{3.16a}$$

$$\varepsilon_t = y_t - \hat{y}_{t|t-1}, \tag{3.16b}$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}), \tag{3.16c}$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}, \tag{3.16d}$$

$$s_t = \gamma^*(y_t - \ell_t) + (1-\gamma^*)s_{t-m}, \tag{3.16e}$$

where the series value is deseasonalized in the trend equations and detrended in the seasonal equation, $\beta^* = \beta/\alpha$ and $\gamma^* = \gamma/(1-\alpha)$. Equations (3.16c–e) can be interpreted as weighted averages, in which case the natural parametric restrictions are that each of α, β^* and γ lie in the $(0,1)$ interval. Equivalently, $0 < \alpha < 1$, $0 < \beta < \alpha$ and $0 < \gamma < 1-\alpha$. However, a consideration of the properties of the discount matrix $\boldsymbol{D}$ leads to a different parameter region; this will be discussed in Chap. 10.

3.5 Variations on the Common Models

A number of variations on the basic models of the previous section can be helpful in some applications.

3.5.1 Damped Level Model

One feature of the models in the framework described in Chap. 2 is that the mean and variance are *local* properties. We may define these moments given the initial conditions, but they do not converge to a stable value as t increases without limit. In other words, the models are all nonstationary; the $\boldsymbol{F}$ matrix has at least one unit root in every case. However, it is possible to describe analogous models that are stationary.

Consider the damped local level model

$$\begin{aligned} y_t &= \mu + \phi\ell_{t-1} + \varepsilon_t, \\ \ell_t &= \phi\ell_{t-1} + \alpha\varepsilon_t. \end{aligned}$$

The transition matrix is simply $F = \phi$, which has no roots greater than one provided $|\phi| < 1$. Thus, the model is stationary for $|\phi| < 1$.

The discount matrix is $D = \phi - \alpha$. Thus, the model is stable provided $|\phi - \alpha| < 1$, or equivalently, $\phi - 1 < \alpha < \phi + 1$.

We may eliminate the state variable to arrive at

$$y_t = \mu + \phi^t\ell_0 + \varepsilon_t + \alpha[\phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots + \phi^{t-1}\varepsilon_1].$$

When $|\phi| < 1$, the mean and variance approach finite limits as $t \to \infty$:

$$\begin{aligned} \mathrm{E}(y_t \mid \ell_0) &= \mu + \phi^t\ell_0 && \to \mu, \\ \mathrm{V}(y_t \mid \ell_0) &= \sigma^2\left[1 + \frac{\alpha^2\phi^2(1-\phi^{2t-2})}{1-\phi^2}\right] && \to \sigma^2\left[1 + \frac{\alpha^2\phi^2}{1-\phi^2}\right]. \end{aligned}$$

Thus, whenever $|\phi| < 1$, the mean reverts to the stable value μ and the variance remains finite. When the series has an infinite past, the limiting values are the unconditional mean and variance. Such stationary series play a major role in the development of Auto Regressive Integrated Moving Average (ARIMA) models, as we shall see in Chap. 11.

There are two reasons why our treatment of mean reversion (or stationarity) is so brief. First, the use of a finite start-up assumption means that stationarity is not needed in order to define the likelihood function. Second, stationary series are relatively uncommon in business and economic applications. Nevertheless, our estimation procedures (Chap. 5) allow mean reverting processes to be fitted if required.

3.5.2 Local Level Model with Drift

A local trend model allows the growth rate to change stochastically over time. If $\beta = 0$, the growth rate is constant and equal to a value that will be denoted by b. The local level model then reduces to

$$y_t = \ell_{t-1} + b + \varepsilon_t,$$
$$\ell_t = b + \ell_{t-1} + \alpha\varepsilon_t,$$

where $\varepsilon_t \sim \text{NID}(0, \sigma^2)$. It is called a "local level model with drift" and has a state space structure with

$$\boldsymbol{x}_t = \begin{bmatrix}\ell_t \; b\end{bmatrix}', \quad \boldsymbol{w} = \begin{bmatrix}1 \; 1\end{bmatrix}', \quad \boldsymbol{F} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \quad \text{and} \quad \boldsymbol{g} = \begin{bmatrix}\alpha \; 0\end{bmatrix}'.$$

This model can be applicable to economic time series that display an upward (or downward) drift. It is sometimes preferred for longer term forecasting because projections are made with the average growth that has occurred throughout the sample rather than a local growth rate, which essentially represents the growth rate that prevails towards the end of the sample.

The discount matrix for this model is

$$\boldsymbol{D} = \begin{bmatrix}1-\alpha & 1-\alpha\\ 0 & 1\end{bmatrix},$$

which has eigenvalues of 1 and $1 - \alpha$. Thus, the model is not stable as $\boldsymbol{D}^j$ does not converge to $\mathbf{0}$. It is, however, forecastable, provided $0 < \alpha < 2$. The model is also forecastable when $\alpha = 0$, as it then reduces to the linear regression model $y_t = \ell_0 + bt + \varepsilon_t$. Discussion of this type of discount matrix will occur in Chap. 10.

The local level model with drift is also known as "simple exponential smoothing with drift." Hyndman and Billah (2003) showed that this model is equivalent to the "Theta method" of Assimakopoulos and Nikolopoulos (2000) with b equal to half the slope of a linear regression of the observed data against their time of observation.

3.5.3 Damped Trend Model: ETS(A,A$_d$,N)

Another possibility is to take the local trend model and dampen its growth rate with a factor ϕ in the region $0 \leq \phi < 1$. The resulting model is

$$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t,$$
$$\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t,$$
$$b_t = \phi b_{t-1} + \beta\varepsilon_t.$$

The characteristics of the damped local trend model are compatible with features observed in many business and economic time series. It sometimes yields better forecasts than the local trend model. Note that the local trend model is a special case where $\phi = 1$.

The ETS(A,A_d,N) model performs remarkably well when forecasting real data (Fildes 1992).

3.5.4 Seasonal Model Based only on Seasonal Levels

If there is no trend in a time series with a seasonal pattern, the ETS(A,N,A) model can be simplified to a model that has a different level in each season. A model for a series with m periods per annum is

$$y_t = \ell_{t-m} + \varepsilon_t, \tag{3.17a}$$
$$\ell_t = \ell_{t-m} + \alpha\varepsilon_t. \tag{3.17b}$$

It conforms to a state space model where

$$\boldsymbol{w}' = \begin{bmatrix} 0\, 0 \cdots 1 \end{bmatrix},$$

$$\boldsymbol{x}_t = \begin{bmatrix} \ell_t \\ \ell_{t-1} \\ \vdots \\ \ell_{t-m+1} \end{bmatrix} \quad \boldsymbol{F} = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{g} = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The weighted average requirement is satisfied if $0 < \alpha < 1$. Because there is no link between the observations other than those m periods apart, we may consider the m sub-models separately. It follows directly that the model is stable when all the sub-models are stable, which is true provided $0 < \alpha < 2$.

3.5.5 Parsimonious Local Seasonal Model

The problem with the seasonal models (3.15) and (3.17) is that they potentially involve a large number of states, and the initial seed state $\boldsymbol{x}_0$ contains a set of parameters that need to be estimated. Modeling weekly demand data, for example, would entail 51 independent seed values for the seasonal recurrence relationships. Estimation of the seed values then makes relatively high demands on computational facilities. Furthermore, the resulting predictions may not be as robust as those from more parsimonious representations.

To emphasize the possibility of a more parsimonious approach, consider the case of a product with monthly sales that peak in December for Christmas, but which tend to be the same, on average, in the months of January to November. There are then essentially two seasonal components, one for the months of January to November, and a second for December. There is no need for 12 separate monthly components.

We require a seasonal model in a form that allows a reduced number of seasonal components. First, redefine m to denote the number of seasonal components, as distinct from the number of seasons per year. In the above example, $m = 2$ instead of 12. An m-vector $\boldsymbol{z}_t$ indicates which seasonal component applies in period t. If seasonal component j applies in period t, then the element $z_{tj} = 1$ and all other elements equal 0. It is assumed that the typical seasonal component j has its own level, which in period t is denoted by ℓ_{tj}. The levels are collected into an m-vector denoted by $\boldsymbol{\ell}_t$. Then the model is

$$y_t = \boldsymbol{z}_t'\boldsymbol{\ell}_{t-1} + b_{t-1} + \varepsilon_t, \tag{3.18a}$$

$$\boldsymbol{\ell}_t = \boldsymbol{\ell}_{t-1} + \mathbf{1}b_{t-1} + (\mathbf{1}\alpha + \boldsymbol{z}_t\gamma)\varepsilon_t, \tag{3.18b}$$

$$b_t = b_{t-1} + \beta\varepsilon_t, \tag{3.18c}$$

where $\mathbf{1}$ represents an m-vector of ones. The term $\boldsymbol{z}_t'\boldsymbol{\ell}_{t-1}$ picks out the level of the seasonal component relevant to period t. The term $\mathbf{1}b_{t-1}$ ensures that each level is adjusted by the same growth rate. It is assumed that the random change has a common effect and an idiosyncratic effect. The term $\mathbf{1}\alpha\varepsilon_t$ represents the common effect, and the term $\boldsymbol{z}_t\beta\varepsilon_t$ is the adjustment to the seasonal component associated with period t.

This model must be coupled with a method that searches systematically for months that possess common seasonal components. We discuss this problem in Chap. 14. In the special case where no common components are found (e.g., $m = 12$ for monthly data), the above model is then equivalent to the seasonal model in Sect. 3.4.3. If, in addition, there is no growth, the model is equivalent to the seasonal level model in Sect. 3.5.4.

Model (3.18) is easily adapted to handle multiple seasonal patterns. For example, daily demand may be influenced by a trading cycle that repeats itself every week, in addition to a seasonal pattern that repeats itself annually. Extensions of this kind are also considered in Chap. 14.

An important point to note is that this seasonal model does not conform to the general form (3.1), because the $\boldsymbol{g}$ and $\boldsymbol{w}$ vectors are time-dependent. A more general time-varying model must be used instead.

3.5.6 Composite Models

Two different models can be used as basic building blocks to yield even larger models. Suppose two basic innovations state space models indexed by $i = 1, 2$ are given by

$$y_t = \boldsymbol{w}_i'\boldsymbol{x}_{i,t-1} + \varepsilon_{it},$$

$$\boldsymbol{x}_{it} = \boldsymbol{F}_i\boldsymbol{x}_{i,t-1} + \boldsymbol{g}_i\varepsilon_{it},$$

where $\varepsilon_{it} \sim \text{NID}(0, v_i)$. A new model can be formed by combining them as follows:

$$y_t = \boldsymbol{w}_1' \boldsymbol{x}_{1,t-1} + \boldsymbol{w}_2' \boldsymbol{x}_{2,t-1} + \varepsilon_t,$$

$$\begin{bmatrix} \boldsymbol{x}_{1t} \\ \boldsymbol{x}_{2t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{1,t-1} \\ \boldsymbol{x}_{2,t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \end{bmatrix} \varepsilon_t.$$

For example, the local trend model (3.13) in Sect. 3.4.2 and the seasonal model (3.17) in Sect. 3.5.4 can be combined using this principle. To avoid conflict with respect to the levels, the ℓ_t in the seasonal model (3.17) is replaced by s_t. The resulting model is the local additive seasonal model (3.15) in Sect. 3.4.3.

3.6 Exercises

Exercise 3.1. Consider the local level model ETS(A,N,N). Show that the process is forecastable and stationary when $\alpha = 0$ but that neither property holds when $\alpha = 2$.

Exercise 3.2. Consider the local level model with drift, defined in Sect. 3.5.2. Define the detrended variable $z_{1t} = y_t - bt$ and the differenced variable $z_{2t} = y_t - y_{t-1}$. Show that both of these processes are stable provided $0 < \alpha < 2$ but that only z_{2t} is stationary.

Exercise 3.3. Consider the local level model ETS(A,N,N). Show that the mean and variance for $y_t|\ell_0$ are ℓ_0 and $\ell_0{}^2(1 + (t-1)\alpha^2)$ respectively.

Exercise 3.4. For the damped trend model ETS(A,A$_\text{d}$,N), find the discount matrix $\boldsymbol{D}$ and its eigenvalues.