2

Getting started

Although exponential smoothing methods have been around since the 1950s, a modeling framework incorporating stochastic models, likelihood calculations, prediction intervals, and procedures for model selection, was not developed until relatively recently with the work of Ord et al. (1997) and Hyndman et al. (2002). In these (and other) papers, a class of state space models has been developed that underlies all of the exponential smoothing methods.

In this chapter, we provide an introduction to the ideas underlying exponential smoothing and the associated state space models. Many of the details will be skipped over in this chapter, but will be covered in later chapters.

Figure 2.1 shows the four time series from Figure 1.1, along with point forecasts and 80% prediction intervals. These were all produced using exponential smoothing state space models. In each case, the particular models and all model parameters were chosen automatically with no intervention by the user. This demonstrates one very useful feature of state space models for exponential smoothing — they are easy to use in a completely automated way. In these cases, the models were able to handle data exhibiting a range of features including very little trend, strong trend, no seasonality, a seasonal pattern that stays constant, and a seasonal pattern with increasing variation as the level of the series increases.

2.1 Time series decomposition

It is common in business and economics to think of a time series as a combination of various components such as the trend (T), cycle (C), seasonal (S) and irregular or error (E) components. These can be defined as follows:

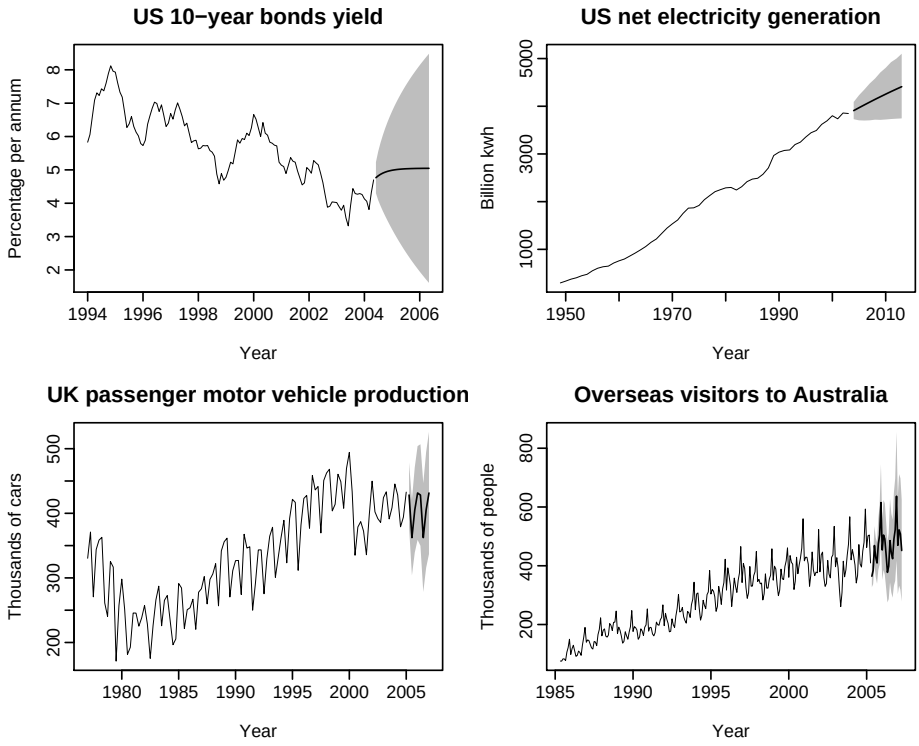

Fig. 2.1: *Four time series showing point forecasts and 80% prediction intervals obtained using exponential smoothing state space models.*

Trend (T):	the long-term direction of the series;
Seasonal(S):	a pattern that repeats with a known periodicity (e.g., 12 months per year, or 7 days per week);
Cycle (C):	a pattern that repeats with some regularity but with unknown and changing periodicity (e.g., a business cycle);
Irregular or Error (E):	the unpredictable component of the series.

In this monograph, we focus primarily upon the three components T, S and E. Any cyclic element will be subsumed within the trend component unless indicated otherwise.

These three components can be combined in a number of different ways. A purely additive model can be expressed as

$$y = T + S + E,$$

where the three components are added together to form the observed series. A purely multiplicative model is written as

$$y = T \times S \times E,$$

where the data are formed by the product of the three components. A *seasonally adjusted* series is then formed by extracting the seasonal component from the data, leaving only the trend and error components. In the additive model, the seasonally adjusted series is $y - S$, while in the multiplicative model, the seasonally adjusted series is y/S. See Makridakis et al. (1998, Chapter 4) for a detailed discussion of seasonal adjustment and time series decomposition.

Other combinations, apart from simple addition and multiplication, are also possible. For example,

$$y = (T + S) \times E$$

treats the irregular component as multiplicative but the other components as additive.[1]

2.2 Classification of exponential smoothing methods

In exponential smoothing, we always start with the trend component, which is itself a combination of a level term (ℓ) and a growth term (b). The level and growth can be combined in a number of ways, giving five future trend types. Let T_h denote the forecast trend over the next h time periods, and let ϕ denote a damping parameter $(0 < \phi < 1)$. Then the five trend types or growth patterns are as follows:

None:	$T_h = \ell$;
Additive:	$T_h = \ell + bh$;
Additive damped:	$T_h = \ell + (\phi + \phi^2 + \cdots + \phi^h)b$;
Multiplicative:	$T_h = \ell b^h$;
Multiplicative damped:	$T_h = \ell b^{(\phi+\phi^2+\cdots+\phi^h)}$.

A damped trend method is appropriate when there is a trend in the time series, but one believes that the growth rate at the end of the historical data is unlikely to continue more than a short time into the future. The equations for damped trend do what the name indicates: dampen the trend as the length of the forecast horizon increases. This often improves the forecast accuracy, particularly at long lead-times.

Having chosen a trend component, we may introduce a seasonal component, either additively or multiplicatively. Finally, we include an error, either additively or multiplicatively. Historically, the nature of the error component has often been ignored, because the distinction between additive and multiplicative errors makes no difference to point forecasts.

If the error component is ignored, then we have the fifteen exponential smoothing methods given in the following table.

[1] See Hyndman (2004) for further discussion of the possible combinations of these components.

Trend Component		Seasonal Component N (None)	A (Additive)	M (Multiplicative)
N	(None)	N,N	N,A	N,M
A	(Additive)	A,N	A,A	A,M
A_d	(Additive damped)	A_d,N	A_d,A	A_d,M
M	(Multiplicative)	M,N	M,A	M,M
M_d	(Multiplicative damped)	M_d,N	M_d,A	M_d,M

This classification of methods originated with Pegels's (1969) taxonomy. This was later extended by Gardner (1985), modified by Hyndman et al. (2002), and extended again by Taylor (2003a), giving the fifteen methods in the above table.

Some of these methods are better known under other names. For example, cell (N,N) describes the simple exponential smoothing (or SES) method, cell (A,N) describes Holt's linear method, and cell (A_d,N) describes the damped trend method. Holt-Winters' additive method is given by cell (A,A), and Holt-Winters' multiplicative method is given by cell (A,M). The other cells correspond to less commonly used but analogous methods.

For each of the 15 methods in the above table, there are two possible state space models, one corresponding to a model with additive errors and the other to a model with multiplicative errors. If the same parameter values are used, these two models give equivalent point forecasts although different prediction intervals. Thus, there are 30 potential models described in this classification.

We are careful to distinguish exponential smoothing *methods* from the underlying state space *models*. An exponential smoothing method is an algorithm for producing point forecasts only. The underlying stochastic state space model gives the same point forecasts, but also provides a framework for computing prediction intervals and other properties. The models are described in Section 2.5, but first we introduce the much older point-forecasting equations.

2.3 Point forecasts for the best known methods

In this section, a simple introduction is provided to some of the best known exponential smoothing methods — simple exponential smoothing (N,N), Holt's linear method (A,N), the damped trend method (A_d,N) and Holt-Winters' seasonal method (A,A and A,M). We denote the observed time series by $y_1, y_2, \ldots, y_n$. A forecast of y_{t+h} based on all the data up to time t is denoted by $\hat{y}_{t+h|t}$. For one-step forecasts, we use the simpler notation $\hat{y}_{t+1} \equiv \hat{y}_{t+1|t}$. Usually, forecasts require some parameters to be estimated;

but for the sake of simplicity it will be assumed for now that the values of all relevant parameters are known.

2.3.1 Simple exponential smoothing (N,N method)

Suppose we have observed data up to and including time $t-1$, and we wish to forecast the next value of our time series, y_t. Our forecast is denoted by $\hat{y}_t$. When the observation y_t becomes available, the forecast error is found to be $y_t - \hat{y}_t$. The method of simple exponential smoothing[2], due to Brown's work in the mid-1950s and published in Brown (1959), takes the forecast for the previous period and adjusts it using the forecast error. That is, the forecast for the next period is

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t), \tag{2.1}$$

where α is a constant between 0 and 1.

It can be seen that the new forecast is simply the old forecast plus an adjustment for the error that occurred in the last forecast. When α has a value close to 1, the new forecast will include a substantial adjustment for the error in the previous forecast. Conversely, when α is close to 0, the new forecast will include very little adjustment.

Another way of writing (2.1) is

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t. \tag{2.2}$$

The forecast $\hat{y}_{t+1}$ is based on weighting the most recent observation y_t with a weight value α, and weighting the most recent forecast $\hat{y}_t$ with a weight of $1-\alpha$. Thus, it can be interpreted as a weighted average of the most recent forecast and the most recent observation.

The implications of exponential smoothing can be seen more easily if equation (2.2) is expanded by replacing $\hat{y}_t$ with its components as follows:

$$\begin{aligned}\hat{y}_{t+1} &= \alpha y_t + (1-\alpha)[\alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}] \\ &= \alpha y_t + \alpha(1-\alpha)y_{t-1} + (1-\alpha)^2\hat{y}_{t-1}.\end{aligned}$$

If this substitution process is repeated by replacing $\hat{y}_{t-1}$ with its components, $\hat{y}_{t-2}$ with its components, and so on, the result is

$$\begin{aligned}\hat{y}_{t+1} = {} & \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \alpha(1-\alpha)^3 y_{t-3} \\ & + \alpha(1-\alpha)^4 y_{t-4} + \cdots + \alpha(1-\alpha)^{t-1} y_1 + (1-\alpha)^t \hat{y}_1.\end{aligned} \tag{2.3}$$

So $\hat{y}_{t+1}$ represents a weighted moving average of all past observations with the weights decreasing exponentially, hence the name "exponential smoothing". We note that the weight of $\hat{y}_1$ may be quite large when α is small and the time series is relatively short. The choice of starting value then becomes

[2] This method is also sometimes known as "*single* exponential smoothing".

particularly important and is known as the "initialization problem", which we discuss in detail in Section 2.6.

For longer range forecasts, it is assumed that the forecast function is "flat". That is,

$$\hat{y}_{t+h|t} = \hat{y}_{t+1}, \qquad h = 2, 3, \ldots .$$

A flat forecast function is used because simple exponential smoothing works best for data that have no trend, seasonality, or other underlying patterns.

Another way of writing this is to let $\ell_t = \hat{y}_{t+1}$. Then $\hat{y}_{t+h|t} = \ell_t$ and $\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$. The value of ℓ_t is a measure of the "level" of the series at time t. While this may seem a cumbersome way to express the method, it provides a basis for generalizing exponential smoothing to allow for trend and seasonality.

In order to calculate the forecasts using SES, we need to specify the initial value $\ell_0 = \hat{y}_1$ and the parameter value α. Traditionally (particularly in the pre-computer age), $\hat{y}_1$ was set to be equal to the first observation and α was specified to be a small number, often 0.2. However, there are now much better ways of selecting these parameters, which we describe in Section 2.6.

2.3.2 Holt's linear method (A,N method)

Holt (1957)[3] extended simple exponential smoothing to linear exponential smoothing to allow forecasting of data with trends. The forecast for Holt's linear exponential smoothing method is found using two smoothing constants, α and β^* (with values between 0 and 1), and three equations:

$$\text{Level:} \qquad \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}), \tag{2.4a}$$

$$\text{Growth:} \qquad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}, \tag{2.4b}$$

$$\text{Forecast:} \qquad \hat{y}_{t+h|t} = \ell_t + b_t h. \tag{2.4c}$$

Here ℓ_t denotes an estimate of the level of the series at time t and b_t denotes an estimate of the slope (or growth) of the series at time t. Note that b_t is a weighted average of the previous growth b_{t-1} and an estimate of growth based on the difference between successive levels. The reason we use β^* rather than β will become apparent when we introduce the state space models in Section 2.5.

In the special case where $\alpha = \beta^*$, Holt's method is equivalent to "Brown's double exponential smoothing" (Brown, 1959). Brown used a discounting argument to arrive at his forecasting equations, so $1-\alpha$ represents the common discount factor applied to both the level and trend components.

In Section 2.6 we describe how the procedure is initialized and how the parameters are estimated.

One interesting special case of this method occurs when $\beta^* = 0$. Then

[3] Reprinted as Holt (2004).

$$\begin{aligned} \text{Level:} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b), \\ \text{Forecast:} \quad & \hat{y}_{t+h|t} = \ell_t + bh. \end{aligned}$$

This method is known as "SES with drift", which is closely related to the "Theta method" of forecasting due to Assimakopoulos and Nikolopoulos (2000). The connection between these methods was demonstrated by Hyndman and Billah (2003).

2.3.3 Damped trend method (A_d,A method)

Gardner and McKenzie (1985) proposed a modification of Holt's linear method to allow the "damping" of trends. The equations for this method are:[4]

$$\text{Level:} \quad \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}), \tag{2.5a}$$

$$\text{Growth:} \quad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}, \tag{2.5b}$$

$$\text{Forecast:} \quad \hat{y}_{t+h|t} = \ell_t + (\phi + \phi^2 + \cdots + \phi^h) b_t. \tag{2.5c}$$

Thus, the growth for the one-step forecast of y_{t+1} is ϕb_t, and the growth is dampened by a factor of ϕ for each additional future time period. If $\phi = 1$, this method gives the same forecasts as Holt's linear method. For $0 < \phi < 1$, as $h \to \infty$ the forecasts approach an asymptote given by $\ell_t + \phi b_t/(1-\phi)$. We usually restrict $\phi > 0$ to avoid a negative coefficient being applied to b_{t-1} in (2.5b), and $\phi \leq 1$ to avoid b_t increasing exponentially.

2.3.4 Holt-Winters' trend and seasonality method

If the data have no trend or seasonal patterns, then simple exponential smoothing is appropriate. If the data exhibit a linear trend, then Holt's linear method (or the damped method) is appropriate. But if the data are seasonal, these methods on their own cannot handle the problem well.

Holt (1957) proposed a method for seasonal data. His method was studied by Winters (1960), and so now it is usually known as "Holt-Winters' method" (see Section 1.3).

Holt-Winters' method is based on three smoothing equations—one for the level, one for trend, and one for seasonality. It is similar to Holt's linear method, with one additional equation for dealing with seasonality. In fact, there are two different Holt-Winters' methods, depending on whether seasonality is modeled in an additive or multiplicative way.

[4] We use the same parameterization as Gardner and McKenzie (1985), which is slightly different from the parameterization proposed by Hyndman et al. (2002). This makes no difference to the value of the forecasts.

Multiplicative seasonality (A,M method)

The basic equations for Holt-Winters' multiplicative method are as follows:

$$\text{Level:} \quad \ell_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{2.6a}$$

$$\text{Growth:} \quad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \tag{2.6b}$$

$$\text{Seasonal:} \quad s_t = \gamma y_t/(\ell_{t-1} + b_{t-1}) + (1-\gamma)s_{t-m} \tag{2.6c}$$

$$\text{Forecast:} \quad \hat{y}_{t+h|t} = (\ell_t + b_t h)s_{t-m+h_m^+}, \tag{2.6d}$$

where m is the length of seasonality (e.g., number of months or quarters in a year), ℓ_t represents the level of the series, b_t denotes the growth, s_t is the seasonal component, $\hat{y}_{t+h|t}$ is the forecast for h periods ahead, and $h_m^+ = \big[(h-1) \bmod m\big] + 1$. The parameters ($\alpha$, β^* and γ) are usually restricted to lie between 0 and 1. See Section 2.6.2 for more details on restricting the values of the parameters. As with all exponential smoothing methods, we need initial values of the components and estimates of the parameter values. This is discussed in Section 2.6.

Equation (2.6c) is slightly different from the usual Holt-Winters' equations such as those in Makridakis et al. (1998) or Bowerman et al. (2005). These authors replace (2.6c) with

$$s_t = \gamma y_t/\ell_t + (1-\gamma)s_{t-m}.$$

The modification given in (2.6c) was proposed by Ord et al. (1997) to make the state space formulation simpler. It is equivalent to Archibald's (1990) variation of Holt-Winters' method. The modification makes a small but usually negligible difference to the forecasts.

Additive seasonality (A,A method)

The seasonal component in Holt-Winters' method may also be treated additively, although in practice this seems to be less commonly used. The basic equations for Holt-Winters' additive method are as follows:

$$\text{Level:} \quad \ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{2.7a}$$

$$\text{Growth:} \quad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \tag{2.7b}$$

$$\text{Seasonal:} \quad s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \tag{2.7c}$$

$$\text{Forecast:} \quad \hat{y}_{t+h|t} = \ell_t + b_t h + s_{t-m+h_m^+}. \tag{2.7d}$$

The second of these equations is identical to (2.6b). The only differences in the other equations are that the seasonal indices are now added and subtracted instead of taking products and ratios.

As with the multiplicative model, the usual equation given in textbooks for the seasonal term is slightly different from (2.7c). Most books use

$$s_t = \gamma^*(y_t - \ell_t) + (1 - \gamma^*)s_{t-m}.$$

If ℓ_t is substituted using (2.7a), we obtain

$$s_t = \gamma^*(1-\alpha)(y_t - \ell_{t-1} - b_{t-1}) + [1 - \gamma^*(1-\alpha)]s_{t-m}.$$

Thus, we obtain identical forecasts using this approach by replacing γ in (2.7c) with $\gamma^*(1-\alpha)$.

2.4 Point forecasts for all methods

Table 2.1 gives recursive formulae for computing point forecasts h periods ahead for all of the exponential smoothing methods. In each case, ℓ_t denotes the series level at time t, b_t denotes the slope at time t, s_t denotes the seasonal component of the series at time t, and m denotes the number of seasons in a year; α, β^*, γ and ϕ are constants, and $\phi_h = \phi + \phi^2 + \cdots + \phi^h$.

Some interesting special cases can be obtained by setting the smoothing parameters to extreme values. For example, if $\alpha = 0$, the level is constant over time; if $\beta^* = 0$, the slope is constant over time; and if $\gamma = 0$, the seasonal pattern is constant over time. At the other extreme, naïve forecasts (i.e., $\hat{y}_{t+h|t} = y_t$ for all h) are obtained using the (N,N) method with $\alpha = 1$. Finally, the additive and multiplicative trend methods are special cases of their damped counterparts obtained by letting $\phi = 1$.

2.5 State space models

We now introduce the state space models that underlie exponential smoothing methods. For each method, there are two models—a model with additive errors and a model with multiplicative errors. The point forecasts for the two models are identical (provided the same parameter values are used), but their prediction intervals will differ.

To distinguish the models with additive and multiplicative errors, we add an extra letter to the front of the method notation. The triplet (E,T,S) refers to the three components: error, trend and seasonality. So the model ETS(A,A,N) has additive errors, additive trend and no seasonality—in other words, this is Holt's linear method with additive errors. Similarly, ETS(M,M_d,M) refers to a model with multiplicative errors, a damped multiplicative trend and multiplicative seasonality. The notation ETS($\cdot$,$\cdot$,$\cdot$) helps in remembering the order in which the components are specified. ETS can also be considered an abbreviation of **E**xponen**T**ial **S**moothing.

Once a model is specified, we can study the probability distribution of future values of the series and find, for example, the conditional mean of a future observation given knowledge of the past. We denote this as

Trend	Seasonal		
	N	**A**	**M**
N	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/\ell_{t-1}) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t$	$\hat{y}_{t+h\|t} = \ell_t + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = \ell_t s_{t-m+h_m^+}$
A	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1} - b_{t-1})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t + hb_t$	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t-m+h_m^+}$
$\mathbf{A_d}$	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1} - \phi b_{t-1})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t-m+h_m^+}$
M	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$
	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1}b_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1}b_{t-1})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t b_t^h$	$\hat{y}_{t+h\|t} = \ell_t b_t^h + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = \ell_t b_t^h s_{t-m+h_m^+}$
$\mathbf{M_d}$	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$
	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$
		$s_t = \gamma(y_t - \ell_{t-1}b_{t-1}^{\phi}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1}b_{t-1}^{\phi})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h}$	$\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} s_{t-m+h_m^+}$

Table 2.1: *Formulae for recursive calculations and point forecasts. In each case, ℓ_t denotes the series level at time t, b_t denotes the slope at time t, s_t denotes the seasonal component of the series at time t, and m denotes the number of seasons in a year; α, β^*, γ and ϕ are constants, $\phi_h = \phi + \phi^2 + \cdots + \phi^h$ and $h_m^+ = [(h-1) \bmod m] + 1$.*

$\mu_{t+h|t} = \mathrm{E}(y_{t+h} \mid \boldsymbol{x}_t)$, where $\boldsymbol{x}_t$ contains the unobserved components such as ℓ_t, b_t and s_t. For $h = 1$ we use $\mu_{t+1} \equiv \mu_{t+1|t}$ as a shorthand notation. For most models, these conditional means will be identical to the point forecasts given earlier, so that $\mu_{t+h|t} = \hat{y}_{t+h|t}$. However, for other models (those with multiplicative trend or multiplicative seasonality), the conditional mean and the point forecast will differ slightly for $h \geq 2$.

2.5.1 State space models for Holt's linear method

We illustrate the ideas using Holt's linear method.

Additive error model: ETS(A,A,N)

Let $\mu_t = \hat{y}_t = \ell_{t-1} + b_{t-1}$ denote the one-step forecast of y_t assuming we know the values of all parameters. Also let $\varepsilon_t = y_t - \mu_t$ denote the one-step forecast error at time t. From (2.4c), we find that

$$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t, \tag{2.8}$$

and using (2.4a) and (2.4b) we can write

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t \tag{2.9}$$
$$b_t = b_{t-1} + \beta^*(\ell_t - \ell_{t-1} - b_{t-1}) = b_{t-1} + \alpha\beta^*\varepsilon_t. \tag{2.10}$$

We simplify the last expression by setting $\beta = \alpha\beta^*$. The three equations above constitute a state space model underlying Holt's method. We can write it in standard state space notation by defining the state vector as $\boldsymbol{x}_t = (\ell_t, b_t)'$ and expressing (2.8)–(2.10) as

$$y_t = [1 \;\; 1]\, \boldsymbol{x}_{t-1} + \varepsilon_t \tag{2.11a}$$
$$\boldsymbol{x}_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \boldsymbol{x}_{t-1} + \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \varepsilon_t. \tag{2.11b}$$

The model is fully specified once we state the distribution of the error term ε_t. Usually we assume that these are independent and identically distributed, following a Gaussian distribution with mean 0 and variance σ^2, which we write as $\varepsilon_t \sim \mathrm{NID}(0, \sigma^2)$.

Multiplicative error model: ETS(M,A,N)

A model with multiplicative error can be derived similarly, by first setting $\varepsilon_t = (y_t - \mu_t)/\mu_t$, so that ε_t is a relative error. Then, following a similar approach to that for additive errors, we find

$$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$$

$$\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$$
$$b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t,$$

or

$$y_t = [1\ \ 1]\,\boldsymbol{x}_{t-1}(1 + \varepsilon_t)$$
$$\boldsymbol{x}_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \boldsymbol{x}_{t-1} + [1\ \ 1]\,\boldsymbol{x}_{t-1} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \varepsilon_t.$$

Again we assume that $\varepsilon_t \sim \text{NID}(0, \sigma^2)$.

Of course, this is a nonlinear state space model, which is usually considered difficult to handle in estimating and forecasting. However, that is one of the many advantages of the innovations form of state space models — we can still compute forecasts, the likelihood and prediction intervals for this nonlinear model with no more effort than is required for the additive error model.

2.5.2 State space models for all exponential smoothing methods

We now give the state space models for all 30 exponential smoothing variations. The general model involves a state vector $\boldsymbol{x}_t = (\ell_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})'$ and state space equations of the form

$$y_t = w(\boldsymbol{x}_{t-1}) + r(\boldsymbol{x}_{t-1})\varepsilon_t \tag{2.12a}$$
$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + g(\boldsymbol{x}_{t-1})\varepsilon_t, \tag{2.12b}$$

where $\{\varepsilon_t\}$ is a Gaussian white noise process with variance σ^2, and $\mu_t = w(\boldsymbol{x}_{t-1})$. The model with additive errors has $r(\boldsymbol{x}_{t-1}) = 1$, so that $y_t = \mu_t + \varepsilon_t$. The model with multiplicative errors has $r(\boldsymbol{x}_{t-1}) = \mu_t$, so that $y_t = \mu_t(1 + \varepsilon_t)$. Thus, $\varepsilon_t = (y_t - \mu_t)/\mu_t$ is the relative error for the multiplicative model. The models are not unique. Clearly, any value of $r(\boldsymbol{x}_{t-1})$ will lead to identical point forecasts for y_t.

Each of the methods in Table 2.1 can be written in the form (2.12a) and (2.12b). The underlying equations for the additive error models are given in Table 2.2. We use $\beta = \alpha\beta^*$ to simplify the notation. Multiplicative error models are obtained by replacing ε_t with $\mu_t\varepsilon_t$ in the equations of Table 2.2. The resulting multiplicative error equations are given in Table 2.3.

Some of the combinations of trend, seasonality and error can occasionally lead to numerical difficulties; specifically, any model equation that requires division by a state component could involve division by zero. This is a problem for models with additive errors and either multiplicative trend or multiplicative seasonality, as well as the model with multiplicative errors, multiplicative trend and additive seasonality. These models should therefore be used with caution. The properties of these models are discussed in Chapter 15.

Trend	Seasonal		
	N	**A**	**M**
N	$\mu_t = \ell_{t-1}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$\mu_t = \ell_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathbf{A_d}$	$\mu_t = \ell_{t-1} + \phi b_{t-1}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$\mu_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$
M	$\mu_t = \ell_{t-1}b_{t-1}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$\mu_t = \ell_{t-1}b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
$\mathbf{M_d}$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1}^{\phi})$

Table 2.2: *State space equations for each additive error model in the classification.*

Trend	Seasonal		
	N	**A**	**M**
N	$\mu_t = \ell_{t-1}$ $\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t)$	$\mu_t = \ell_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = \ell_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
A	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
$\mathbf{A_d}$	$\mu_t = \ell_{t-1} + \phi b_{t-1}$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$\mu_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
M	$\mu_t = \ell_{t-1}b_{t-1}$ $\ell_t = \ell_{t-1}b_{t-1}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}(1+\beta\varepsilon_t)$	$\mu_t = \ell_{t-1}b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}(1+\beta\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
$\mathbf{M_d}$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1+\beta\varepsilon_t)$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1+\beta\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$

Table 2.3: *State space equations for each multiplicative error model in the classification.*

The multiplicative error models are useful when the data are strictly positive, but are not numerically stable when the data contain zeros or negative values. So when the time series is not strictly positive, only the six fully additive models may be applied.

The point forecasts given earlier are easily obtained from these models by iterating the equations (2.12) for $t = n+1, n+2, \ldots, n+h$, and setting $\varepsilon_{n+j} = 0$ for $j = 1, \ldots, h$. In most cases (notable exceptions being models with multiplicative seasonality or multiplicative trend for $h \geq 2$), the point forecasts can be shown to be equal to $\mu_{t+h|t} = \mathrm{E}(y_{t+h} \mid \boldsymbol{x}_t)$, the conditional expectation of the corresponding state space model.

The models also provide a means of obtaining prediction intervals. In the case of the linear models, where the prediction distributions are Gaussian, we can derive the conditional variance $v_{t+h|t} = \mathrm{V}(y_{t+h} \mid \boldsymbol{x}_t)$ and obtain prediction intervals accordingly. This approach also works for many of the nonlinear models, as we show in Chapter 6.

A more direct approach that works for all of the models is to simply simulate many future sample paths, conditional on the last estimate of the state vector, $\boldsymbol{x}_t$. Then prediction intervals can be obtained from the percentiles of the simulated sample paths. Point forecasts can also be obtained in this way by taking the average of the simulated values at each future time period. One advantage of this approach is that we generate an estimate of the complete predictive distribution, which is especially useful in applications such as inventory planning, where the expected costs depend on the whole distribution.

2.6 Initialization and estimation

In order to use these models for forecasting, we need to specify the type of model to be used (model selection), the value of $\boldsymbol{x}_0$ (initialization), and the values of the parameters α, β, γ and ϕ (estimation). In this section, we discuss initialization and estimation, leaving model selection to Section 2.8.

2.6.1 Initialization

Traditionally, the initial values $\boldsymbol{x}_0$ are specified using ad hoc values, or via a heuristic scheme. The following heuristic scheme, based on Hyndman et al. (2002), seems to work very well.

- **Initial seasonal component:** For seasonal data, compute a $2 \times m$ moving average through the first few years of data. Denote this by $\{f_t\}$, $t = m/2+1, m/2+2, \ldots$. For additive seasonality, detrend the data to obtain $y_t - f_t$; for multiplicative seasonality, detrend the data to obtain y_t / f_t.
  Compute initial seasonal indices, $s_{-m+1}, \ldots, s_0$, by averaging the detrended data for each season. Normalize these seasonal indices so that

they add to zero for additive seasonality, and add to m for multiplicative seasonality.

- **Initial level component:** For seasonal data, compute a linear trend using linear regression on the first 10 seasonally adjusted values (using the seasonal indices obtained above) against a time variable $t = 1, \dots, 10$. For non-seasonal data, compute a linear trend on the first 10 observations against a time variable $t = 1, \dots, 10$. Then set ℓ_0 to be the intercept of the trend.
- **Initial growth component:** For additive trend, set b_0 to be the slope of the trend. For multiplicative trend, set $b_0 = 1 + b/a$, where a denotes the intercept and b denotes the slope of the fitted trend.

These initial states are then refined by estimating them along with the parameters, as described below.

2.6.2 Estimation

It is easy to compute the likelihood of the innovations state space model (2.12), and so obtain maximum likelihood estimates. In Chapter 5, we show that

$$\mathcal{L}^*(\boldsymbol{\theta}, \boldsymbol{x}_0) = n \log\Big(\sum_{t=1}^{n} \varepsilon_t^2\Big) + 2\sum_{t=1}^{n} \log|r(\boldsymbol{x}_{t-1})|$$

is equal to twice the negative logarithm of the likelihood function (with constant terms eliminated), conditional on the parameters $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \phi)'$ and the initial states $\boldsymbol{x}_0 = (\ell_0, b_0, s_0, s_{-1}, \dots, s_{-m+1})'$, where n is the number of observations. This is easily computed by simply using the recursive equations in Table 2.1. Unlike state space models with multiple sources of error, we do not need to use the Kalman filter to compute the likelihood.

The parameters $\boldsymbol{\theta}$ and the initial states $\boldsymbol{x}_0$ can be estimated by minimizing $\mathcal{L}^*$. Alternatively, estimates can be obtained by minimizing the one-step mean squared error (MSE), minimizing the residual variance σ^2, or via some other criterion for measuring forecast error. Whichever criterion is used, we usually begin the optimization with $\boldsymbol{x}_0$ obtained from the heuristic scheme above and $\boldsymbol{\theta} = (0.1, 0.01, 0.01, 0.99)'$.

There have been several suggestions for restricting the parameter space of α, β and γ. The traditional approach is to ensure that the various equations can be interpreted as weighted averages, thus requiring α, $\beta^* = \beta/\alpha$, $\gamma^* = \gamma/(1-\alpha)$ and ϕ to all lie within $(0, 1)$. This suggests

$$0 < \alpha < 1, \qquad 0 < \beta < \alpha, \qquad 0 < \gamma < 1 - \alpha, \qquad \text{and} \qquad 0 < \phi < 1.$$

However, we shall see in Chapter 10 that these restrictions are usually stricter than necessary (although in a few cases they are not restrictive enough).

We also constrain the initial states $\boldsymbol{x}_0$ so that the seasonal indices add to zero for additive seasonality, and add to m for multiplicative seasonality.

2.7 Assessing forecast accuracy

The issue of measuring the accuracy of forecasts from different methods has been the subject of much attention. We summarize some of the approaches here. A more thorough discussion is given by Hyndman and Koehler (2006).

There are three possible ways the forecasts can have arisen:

1. The forecasts may be computed from a common base time, and be of varying forecast horizons. That is, we may compute out-of-sample forecasts $\hat{y}_{n+1|n}, \ldots, \hat{y}_{n+h|n}$ based on data from times $t = 1, \ldots, n$. When $h = 1$, we write $\hat{y}_{n+1} \equiv \hat{y}_{n+1|n}$.
2. The forecasts may be from varying base times, and be of a consistent forecast horizon. That is, we may compute forecasts $\hat{y}_{1+h|1}, \ldots, \hat{y}_{m+h|m}$ where each $\hat{y}_{j+h|j}$ is based on data from times $t = 1, \ldots, j$.
3. We may wish to compare the accuracy of methods between many series at a single forecast horizon. That is, we compute a single $\hat{y}_{n+h|n}$ based on data from times $t = 1, \ldots, n$ for each of m different series.

While these are very different situations, measuring forecast accuracy is the same in each case.

The measures defined below are described for one-step-ahead forecasts; the extension to h-steps-ahead is immediate in each case and raises no new questions of principle.

2.7.1 Scale-dependent errors

The one-step-ahead forecast error is simply $e_t = y_t - \hat{y}_t$, regardless of how the forecast was produced. Similarly the h-step-ahead forecast error is $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$. This is on the same scale as the data. Accuracy measures that are based on e_t are therefore scale-dependent.

The two most commonly used scale-dependent measures are based on the absolute error or squared errors:

$$\text{Mean Absolute Error (MAE)} = \text{mean}(|e_t|)$$
$$\text{Mean Squared Error (MSE)} = \text{mean}(e_t^2).$$

When comparing forecast methods on a single series, we prefer the MAE as it is easy to understand and compute. However, it cannot be used to make comparisons between series as it makes no sense to compare accuracy on different scales.

2.7.2 Percentage errors

The percentage error is given by $p_t = 100e_t/y_t$. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance between different data sets. The most commonly used measure is:

$$\text{Mean Absolute Percentage Error (MAPE)} = \text{mean}(|p_t|)$$

Measures based on percentage errors have the disadvantage of being infinite or undefined if $y_t = 0$ for any t in the period of interest, and having an extremely skewed distribution when any y_t is close to zero. Another problem with percentage errors that is often overlooked is that they assume a meaningful zero. For example, a percentage error makes no sense when measuring the accuracy of temperature forecasts on the Fahrenheit or Celsius scales.

They also have the disadvantage that they put a heavier penalty on positive errors than on negative errors. This observation led to the use of the so-called "symmetric" MAPE proposed by Makridakis (1993), which was used in the M3 competition (Makridakis and Hibon, 2000). It is defined by

$$\text{Symmetric Mean Absolute Percentage Error (sMAPE)} = \text{mean}(200|y_t - \hat{y}_t|/(y_t + \hat{y}_t))$$

However, if y_t is zero, $\hat{y}_t$ is also likely to be close to zero. Thus, the measure still involves division by a number close to zero. Also, the value of sMAPE can be negative, so it is not really a measure of "absolute percentage errors" at all.

2.7.3 Scaled errors

The MASE was proposed by Hyndman and Koehler (2006) as a generally applicable measure of forecast accuracy. They proposed scaling the errors based on the *in-sample* MAE from the naïve forecast method. Thus, a scaled error is defined as

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}$$

which is independent of the scale of the data. A scaled error is less than one if it arises from a better forecast than the average one-step naïve forecast computed in-sample. Conversely, it is greater than one if the forecast is worse than the average one-step naïve forecast computed in-sample.

The Mean Absolute Scaled Error is simply

$$\text{MASE} = \text{mean}(|q_t|).$$

The *in-sample* MAE is used in the denominator as it is always available and effectively scales the errors. In contrast, the out-of-sample MAE for the naïve method can be based on very few observations and is therefore more variable. For some data sets, it can even be zero. Consequently, the in-sample MAE is preferable in the denominator.

The MASE can be used to compare forecast methods on a single series, and to compare forecast accuracy between series as it is scale-free. It is the only available method which can be used in all circumstances.

2.8 Model selection

The forecast accuracy measures described in the previous section can be used for selecting a model for a given set of data, provided the errors are computed from data in a hold-out set and not from the same data as were used for model estimation. However, there are often too few out-of-sample errors to draw reliable conclusions. Consequently, a penalized method based on in-sample fit is usually better.

One such method is via a penalized likelihood such as Akaike's Information Criterion:

$$\text{AIC} = \mathcal{L}^*(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_0) + 2q,$$

where q is the number of parameters in $\boldsymbol{\theta}$ plus the number of free states in $\boldsymbol{x}_0$, and $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{x}}_0$ denote the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{x}_0$. (In computing the AIC, we also require that the state space model has no redundant states—see Section 10.1, p.157.) We select the model that minimizes the AIC amongst all of the models that are appropriate for the data.

The AIC also provides a method for selecting between the additive and multiplicative error models. Point forecasts from the two models are identical, so that standard forecast accuracy measures such as the MSE or MAPE are unable to select between the error types. The AIC is able to select between the error types because it is based on likelihood rather than one-step forecasts.

Obviously, other model selection criteria (such as the BIC) could also be used in a similar manner. Model selection is explored in more detail in Chapter 7.

2.8.1 Automatic forecasting

We combine the preceding ideas to obtain a robust and widely applicable automatic forecasting algorithm. The steps involved are summarized below.

1. For each series, apply all models that are appropriate, optimizing the parameters of the model in each case.
2. Select the best of the models according to the AIC.
3. Produce point forecasts using the best model (with optimized parameters) for as many steps ahead as required.
4. Obtain prediction intervals[5] for the best model either using the analytical results, or by simulating future sample paths for $\{y_{n+1}, \dots, y_{n+h}\}$ and finding the $\alpha/2$ and $1-\alpha/2$ percentiles of the simulated data at each forecasting horizon. If simulation is used, the sample paths may be generated using the Gaussian distribution for errors (parametric bootstrap) or using the resampled errors (ordinary bootstrap).

[5] The calculation of prediction intervals is discussed in Chapter 6.

This algorithm resulted in the forecasts shown in Figure 2.1. The models chosen were:

- ETS(A,A_d,N) for monthly US 10-year bond yields ($\alpha = 0.99$, $\beta = 0.12$, $\phi = 0.80$, $\ell_0 = 5.30$, $b_0 = 0.71$);
- ETS(M,M_d,N) for annual US net electricity generation ($\alpha = 0.99$, $\beta = 0.01$, $\phi = 0.97$, $\ell_0 = 262.5$, $b_0 = 1.12$);
- ETS(A,N,A) for quarterly UK passenger vehicle production ($\alpha = 0.61$, $\gamma = 0.01$, $\ell_0 = 343.4$, $s_{-3} = 24.99$, $s_{-2} = 21.40$, $s_{-1} = -44.96$, $s_0 = -1.42$);
- ETS(M,A,M) for monthly Australian overseas visitors ($\alpha = 0.57$, $\beta = 0.01$, $\gamma = 0.19$, $\ell_0 = 86.2$, $b_0 = 2.66$, $s_{-11} = 0.851$, $s_{-10} = 0.844$, $s_{-9} = 0.985$, $s_{-8} = 0.924$, $s_{-7} = 0.822$, $s_{-6} = 1.006$, $s_{-5} = 1.101$, $s_{-4} = 1.369$, $s_{-3} = 0.975$, $s_{-2} = 1.078$, $s_{-1} = 1.087$, $s_0 = 0.958$).

Although there is a lot of computation involved, it can be handled remarkably quickly on modern computers. The forecasts shown in Figure 2.1 took a few seconds on a standard PC.

Hyndman et al. (2002) applied this automatic forecasting strategy to the M-competition data (Makridakis et al., 1982) and IJF-M3 competition data (Makridakis and Hibon, 2000), and demonstrated that the methodology is particularly good at short term forecasts (up to about 6 periods ahead), and especially for seasonal short-term series (beating all other methods in the competition for these series).

2.9 Exercises

Exercise 2.1 *Consider the innovations state space model* (2.12). *Equations* (2.12a) *and* (2.12b) *are called the* measurement equation *and* transition equation *respectively.*

a. *For the ETS(A,A_d,N) model, write the measurement equation and transition equations with a separate equation for each of the two states (level and growth).*
b. *For the ETS (A,A_d,N) model write the measurement and transition equations in matrix form, defining $\boldsymbol{x}_t$, $w(\boldsymbol{x}_{t-1})$, $r(\boldsymbol{x}_{t-1})$, $f(\boldsymbol{x}_{t-1})$, and $g(\boldsymbol{x}_{t-1})$. See Section 2.5.1 for an example based on the ETS(A,A,N) model.*
c. *Repeat parts a and b for the ETS(A,A,A) model.*
d. *Repeat parts a and b for the ETS(M,A_d,N) model.*
e. *Repeat parts a and b for the ETS(M,A_d,A) model.*

Exercise 2.2 *Use the innovations state space model, including the assumptions about ε_t, to derive the specified point forecast,*

$$\hat{y}_{t+h|t} = \mu_{t+h|t} = \mathrm{E}(y_{t+h} \mid \boldsymbol{x}_t),$$

and variance of the forecast error,

$$v_{t+h|t} = \mathrm{V}(y_{t+h} \mid \boldsymbol{x}_t),$$

for the following models.

a. *For ETS(A,N,N), show* $\hat{y}_{t+h|t} = \ell_t$ *and* $v_{t+h|t} = \sigma^2[1 + (h-1)\alpha^2]$.
b. *For ETS(A,A,N), show* $\hat{y}_{t+h|t} = \ell_t + hb_t$ *and*

$$v_{t+h|t} = \sigma^2\Big[1 + \sum_{j=1}^{h-1} (\alpha + \beta j)^2\Big].$$

c. *For ETS(M,N,N), show* $\hat{y}_{t+h|t} = \ell_t$, $v_{t+1|t} = \ell_t^2\sigma^2$, *and*

$$v_{t+2|t} = \ell_t^2\left[(1+\alpha^2\sigma^2)(1+\sigma^2) - 1\right].$$

Exercise 2.3 *Use* **R** *to reproduce the results in Section 2.8.1 for each of the four time series: US 10-year bond yields, US net electricity, UK passenger vehicle production, and Australian overseas visitors. The data sets are named* `bonds`, `usnetelec`, `ukcars` *and* `visitors` *respectively. The* `ets()` *function found in the* `forecast` *package can be used to specify the model or to automatically select a model.*

Exercise 2.4 *Using the results of Exercise 2.3, use* **R** *to reproduce the results in Figure 2.1 for point forecasts and prediction intervals for each of the four time series. The* `forecast()` *function in the* `forecast` *package can be used to produce the point forecasts and prediction intervals for each model found in Exercise 2.3.*