

# Forecasting in R

## Exponential smoothing in ETS form



# Outline





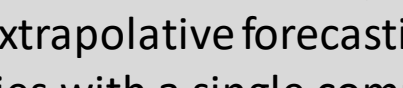





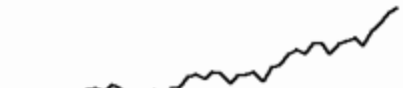
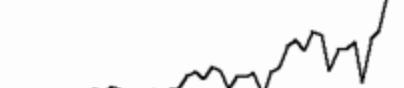



1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Introduction to ETS

- Different types of time series:

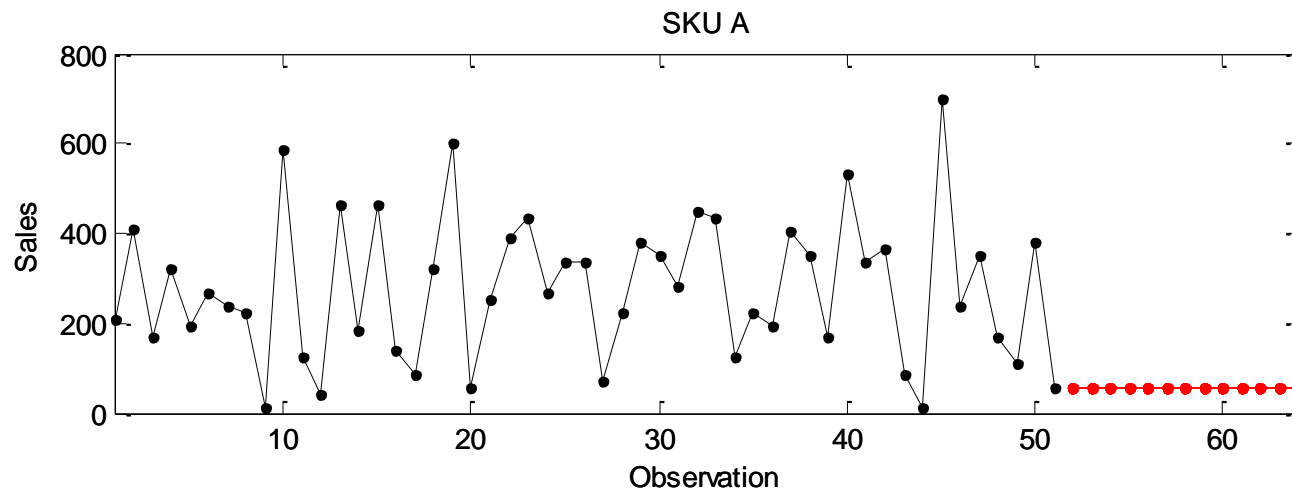
Trend	Seasonality		
	None	Additive	Multiplicative
None			
Additive			
Additive Damped			
Multiplicative			
Multiplicative Damped			

Let us understand the principles of extrapolative forecasting with series with a single component

# Naïve forecast

What is the simplest forecast you can think of for a time series?  
For example: what will the temperature be like in this room after 5 minutes?

$$\hat{y}_{t+1} = y_t$$



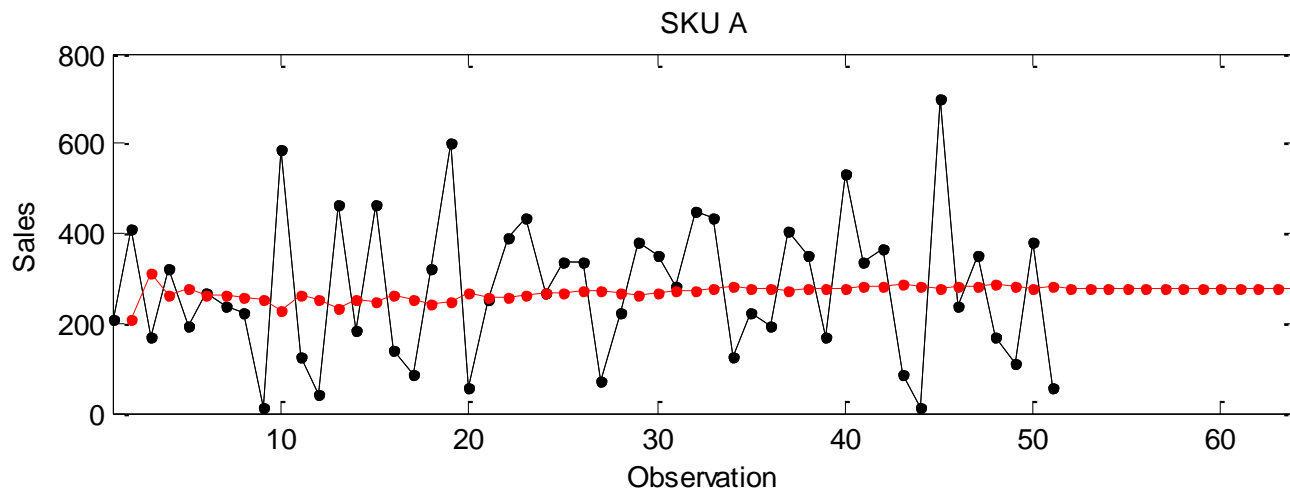
- The forecast is a straight line → always equal to the last observation.
- Is this a good forecast?

# Arithmetic mean

Another approach would be to calculate the average and use this as a forecast.

The average has long memory and the random movements of the noise will be cancelled out.

$$\hat{y}_{t+1} = \frac{1}{t} \sum_{i=1}^t y_i$$



Is this a good forecast?

Should the forecast be a straight line?

# Simple Moving Average

**Simple Moving Average** allows us to select the appropriate memory (length of the average).

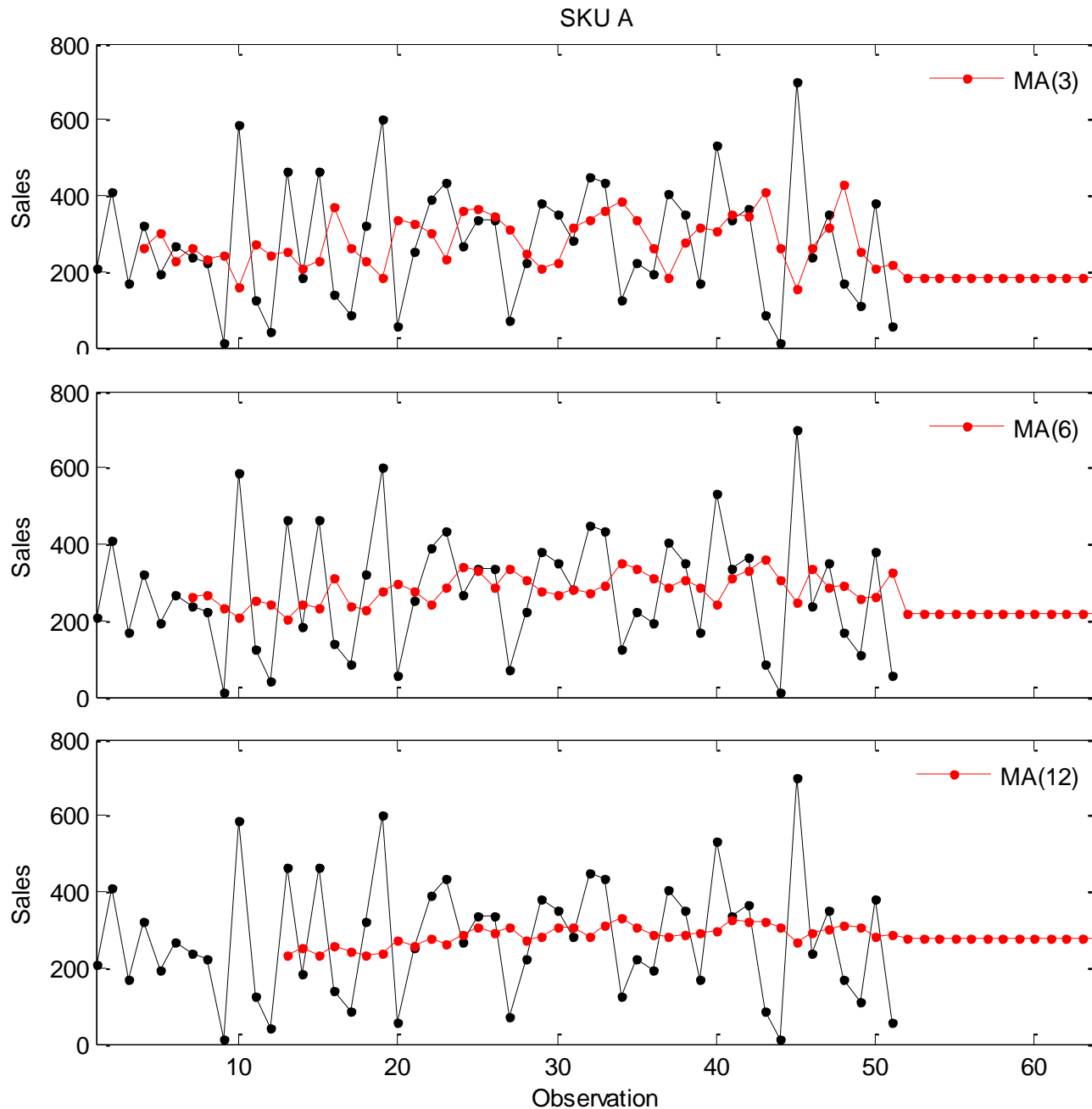
$$\hat{y}_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^t y_i$$

The **simple moving average**:

- Has a single parameter **k**. This controls the length of the moving average and it is also known as its order.
- Its variable length allows us to control how reactive we are to new information and how robust we are against noise.
- Gives equal importance to all **k** observations.



# Simple Moving Average

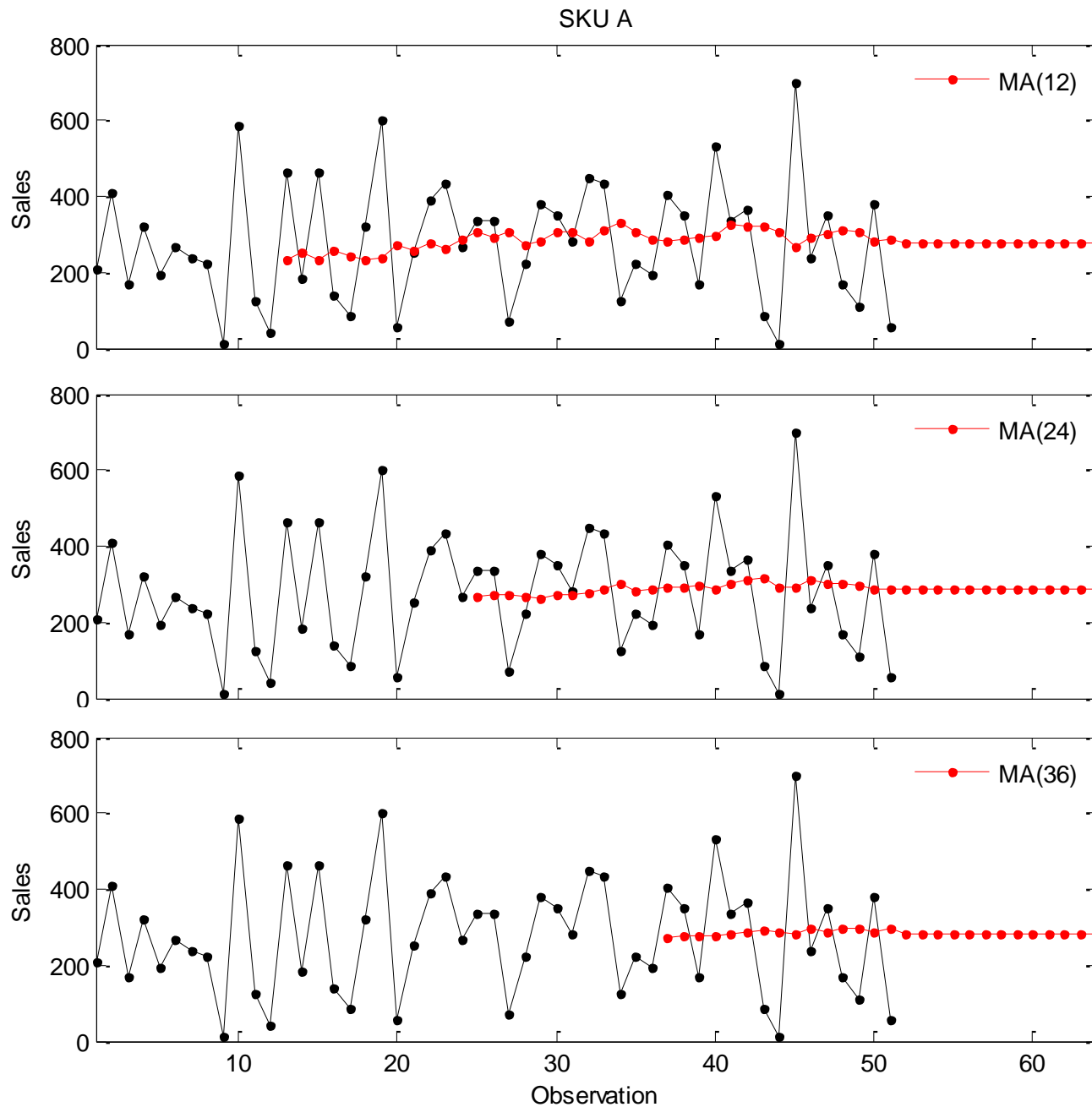


Which of the different length moving averages is the most appropriate for this SKU?

We choose the moving average gives us a smooth estimate of the level, here MA(12)



# Simple Moving Average



Which of the different length moving averages is the most appropriate for this SKU?

We do not need excessive moving average lengths. These will be far too insensitive to new information.

# Weighted Moving Average

We can overcome this limitation by allowing different weights for each observation in the average:

$$\hat{y}_{t+1} = \sum_{i=t-k+1}^t w_i y_i, \quad \text{w. r. t. } \sum_{i=1}^k w_i = 1$$

With the **weighted moving average**:

- We can control the length of the average and the importance of each observation
- All weights must add up to 100% or 1. Normally the older the observation the smaller the weight.
- Has  $k+1$  parameters, the length of the average and  $k$  weights.
- The number of weights makes it very challenging to use in practice.

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# The Exponential Smoothing Concept

Starting from the **weighted moving average** we can construct a heuristic to select the weights easily and consequently its order ( $k$ ).

<b>Data</b>	$y_t$	$y_{t-1}$	$y_{t-2}$	$y_{t-3}$	...
<b>Weights</b>	$W_t$	$W_{t-1}$	$W_{t-2}$	$W_{t-3}$	...

1. Make the more recent information more relevant, bigger weights
2. Remember! Weights must add up to 100% (or 1)

→ Take 50% for the first and then always take 50% of the remaining weight. (Sum of all weights  $\approx$  100%)

<b>Weights</b>	$W_t$	$W_{t-1}$	$W_{t-2}$	$W_{t-3}$	$W_{t-4}$	$W_{t-5}$	$W_{t-6}$
<b>Weights</b>	50%	25%	12.5%	6.25%	3.12%	1.56%	$\approx 0\%$

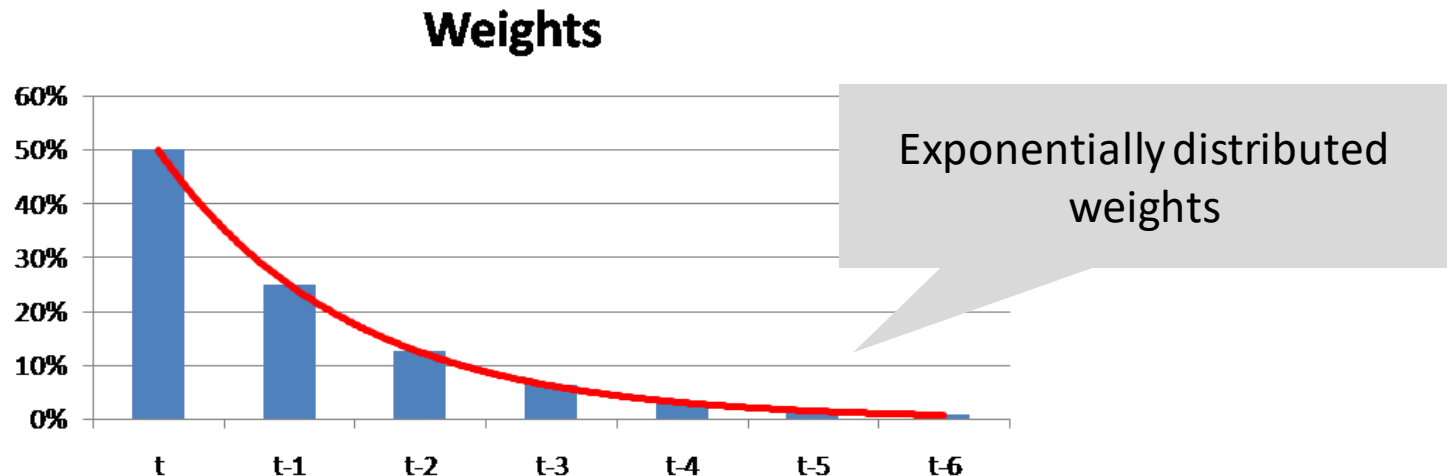
→ The length of the average is set automatically!

# The Exponential Smoothing Concept

Weights	$W_t$	$W_{t-1}$	$W_{t-2}$	$W_{t-3}$	$W_{t-4}$	$W_{t-5}$	$W_{t-6}$
	50%	25%	12.5%	6.25%	3.12%	1.56%	$\approx 0\%$

Only one parameter, the initial weight! Let this weight be **Alpha ( $\alpha$ )**...

$$\alpha(1-\alpha)^0 \quad \alpha(1-\alpha)^1 \quad \alpha(1-\alpha)^2 \quad \alpha(1-\alpha)^3 \quad \alpha(1-\alpha)^4 \quad \alpha(1-\alpha)^5 \quad \alpha(1-\alpha)^6$$



The exponential weighting scheme allows us to select reasonable weights and the length of the weighted moving average with a single parameter, the  $\alpha$ .

# The Exponential Smoothing Concept

$$\hat{y}_{t+1} = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha)^2 y_{t-2} + \alpha(1 - \alpha)^3 y_{t-3} + \dots$$

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha) \underbrace{(\alpha y_{t-1} + \alpha(1 - \alpha)y_{t-2} + \alpha(1 - \alpha)^2 y_{t-3} + \dots)}$$

What is this?

$$\hat{y}_t = \alpha y_{t-1} + \alpha(1 - \alpha)y_{t-2} + \alpha(1 - \alpha)^2 y_{t-3} + \dots$$

A simpler form of the model:

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

# Simple Exponential Smoothing

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

The parameter  $\alpha$ , is called **smoothing parameter** and is bounded between 0 and 1.

The exponential smoothing formula can be read as: the forecast is  $\alpha$  times the most recent observation and  $(1-\alpha)$  times all the previous information.

- A low  $\alpha$  implies that the forecast is mostly based on the previous information
- A high  $\alpha$  implies that the forecast is mostly based on the last information

Therefore the smoothing parameter  $\alpha$  controls how reactive is the forecast to new information.

This form was proposed by Brown (1956).

Much has changed since then...



# Simple Exponential Smoothing

We can interpret exponential smoothing in a different way:

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

$$\hat{y}_{t+1} = \alpha y_t + \hat{y}_t - \alpha \hat{y}_t$$

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

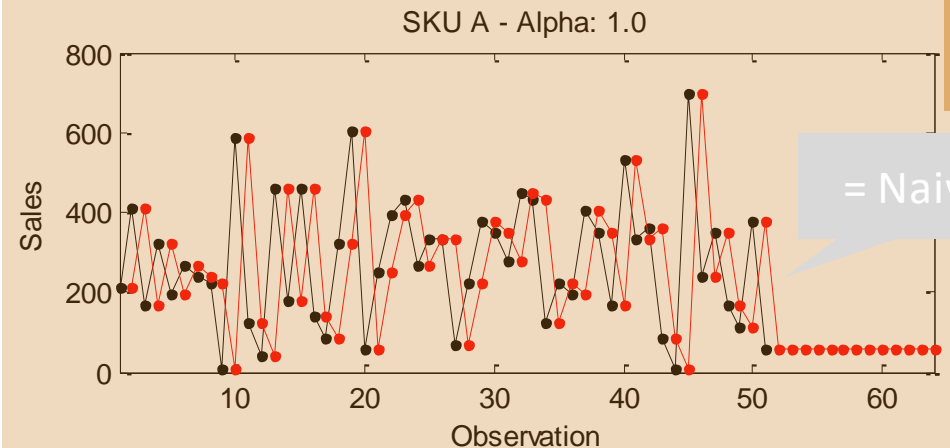
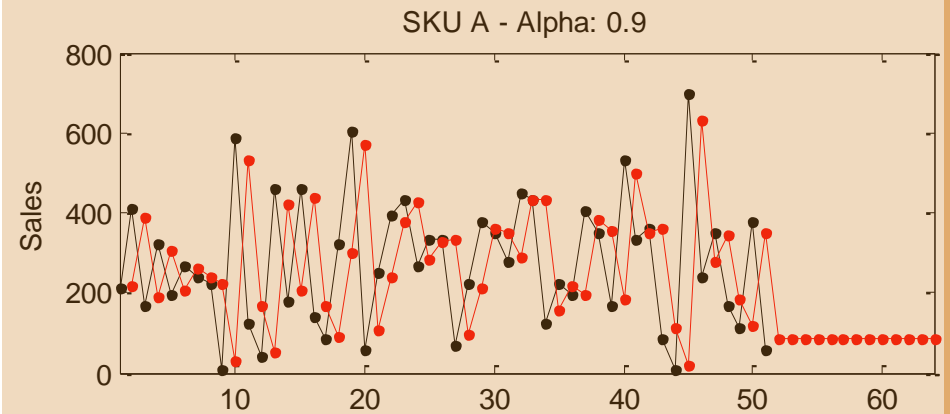
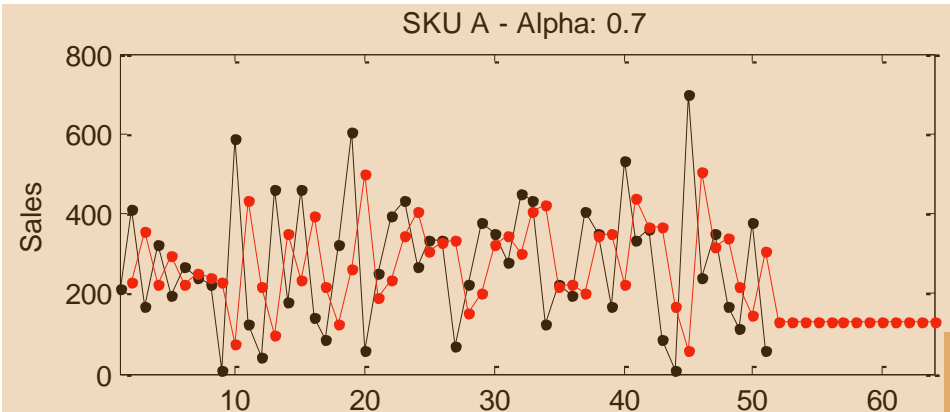
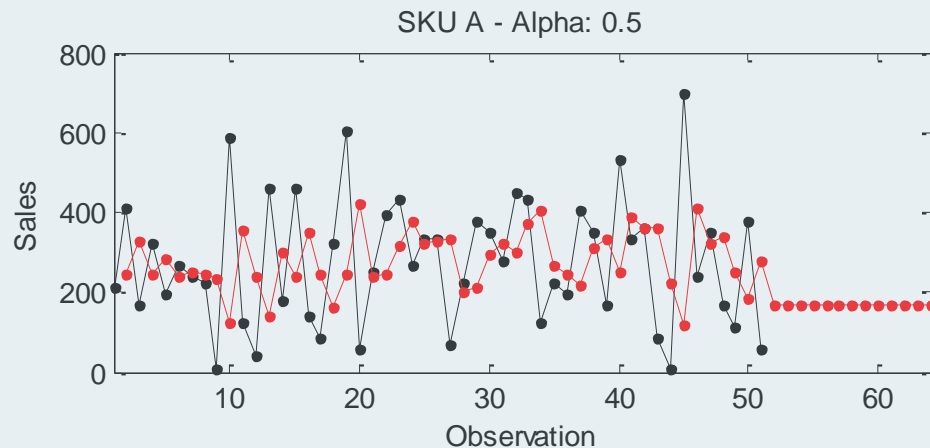
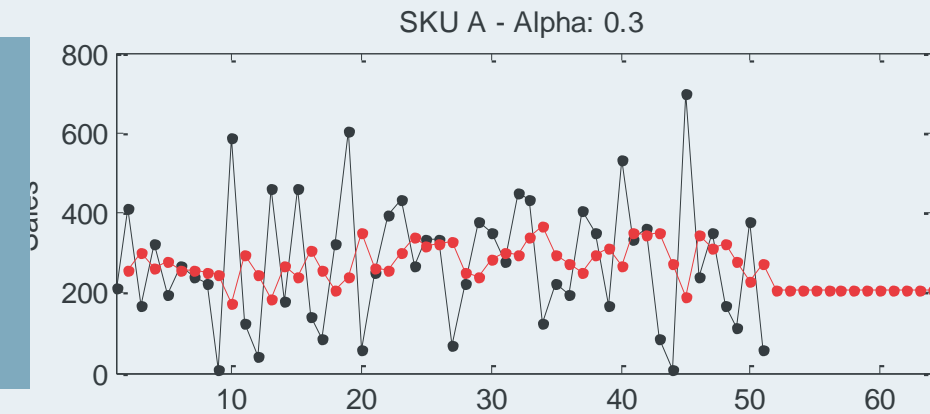
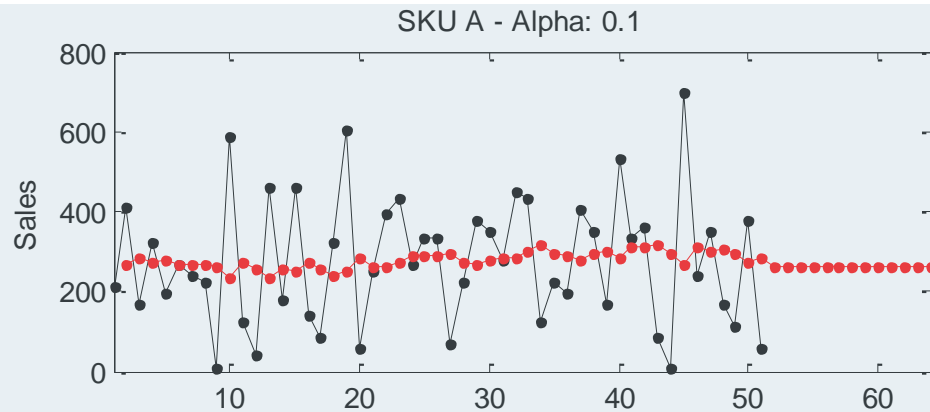
The difference between the Actuals and the Forecast is the forecast **error**.

$$\hat{y}_{t+1} = \hat{y}_t + \alpha e_t$$

This is known as the **error correction form** of exponential smoothing.

This interpretation says that the exponential smoothing forecast is the same as the previous forecast updated by alpha times the forecast error. This shows that  $\alpha$  controls how reactive exponential smoothing is to new information.

# Simple Exponential Smoothing

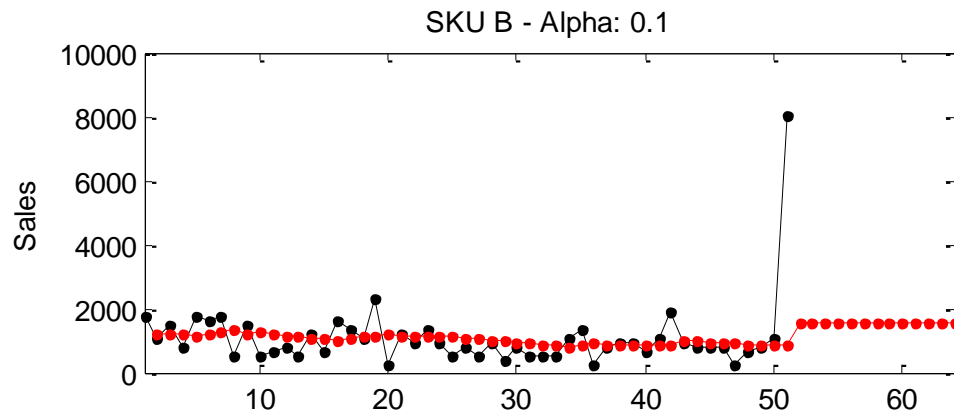


Noise is filtered

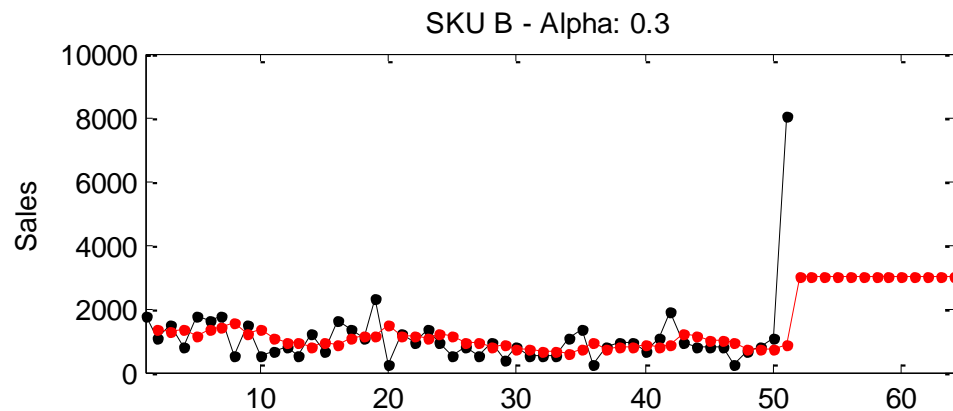
Noise is not filtered → Avoid

= Naive

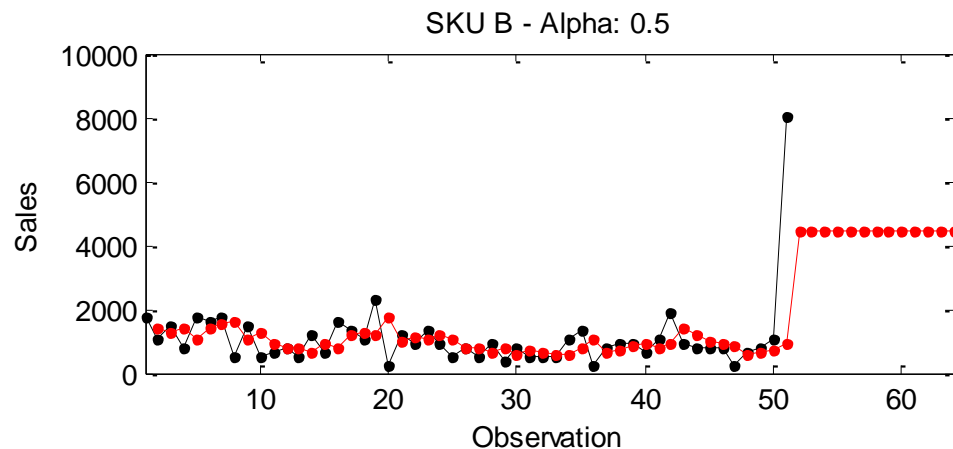
# Simple Exponential Smoothing



In the presence of high noise or outliers we need to use low values of alpha to make our forecasts more robust.

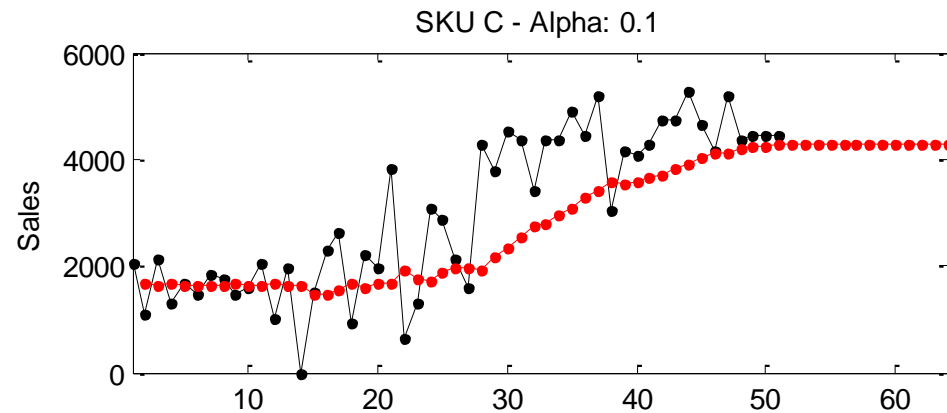


Here the outlier affects strongly our forecast.

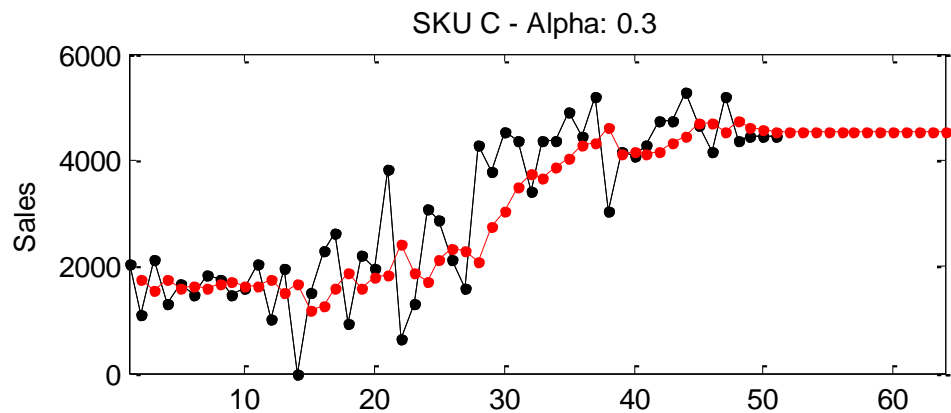


Here the outlier affects strongly our forecast.

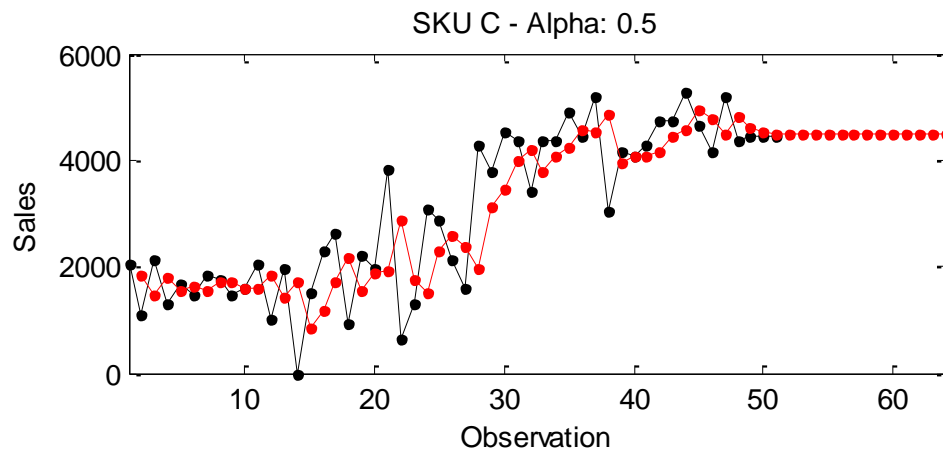
# Simple Exponential Smoothing



Very low alpha parameter makes our forecast too slow to adjust to the new level of sales.



Here the alpha achieves a good compromise between reactivity and robustness to noise.



Very high alpha parameter makes our forecast to react very fast, but now it does not filter out noise adequately.

# Simple Exponential Smoothing

We saw two ways to write the Single Exponential Smoothing method:

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \quad \longleftrightarrow \quad \hat{y}_{t+1} = \hat{y}_t + \alpha e_t$$

Keep in mind that SES is modelling the level of a time series, so we can write  $\hat{y}_{t+1} = l_t$

By shifting the indices by 1 period we can now write:

$$y_t = l_{t-1} + e_t \quad (1)$$

$$l_t = l_{t-1} + \alpha e_t \quad (2)$$

This will lead us to the so called **State Space Models**:

- Eq. (1) – the **observation equation**: says that the observed actuals are the result of some structure ( $l_t$ ) and noise ( $e_t$ ).
- Eq. (2) – the **state equation**: says that there is an unobserved process describing how the level of the time series evolves. For our case this is all the structure of the series.
- We could easily expand this model further to have more lines for other structural components → using this logic we can do some pretty neat stuff!
- Note that the SES **method** is not explicit that there is always some noise. Our **model** has the term  $e_t$  in Eq. (1). By modelling  $e_t$  we can describe the noise/uncertainty fully!

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Introduction to ETS

- Based on the time series decomposition we can have the pure additive model:

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$$

- And for the pure multiplicative one:

$$y_t = l_{t-1} b_{t-1} s_{t-m} \varepsilon_t$$

- And there are combinations between the two.











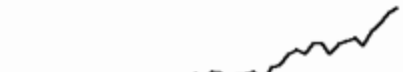
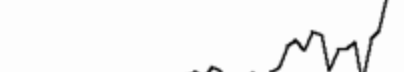



- For example, an ETS(M,A,M) model:

$$y_t = (l_{t-1} + b_{t-1}) s_{t-m} \varepsilon_t$$



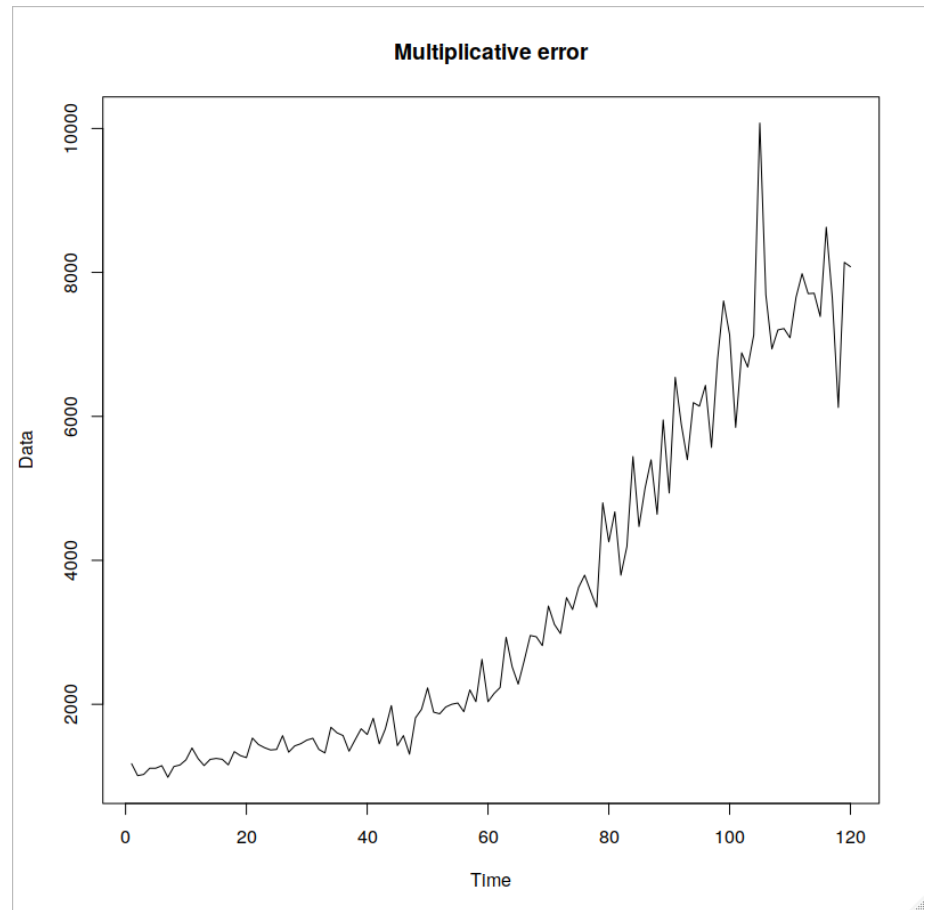
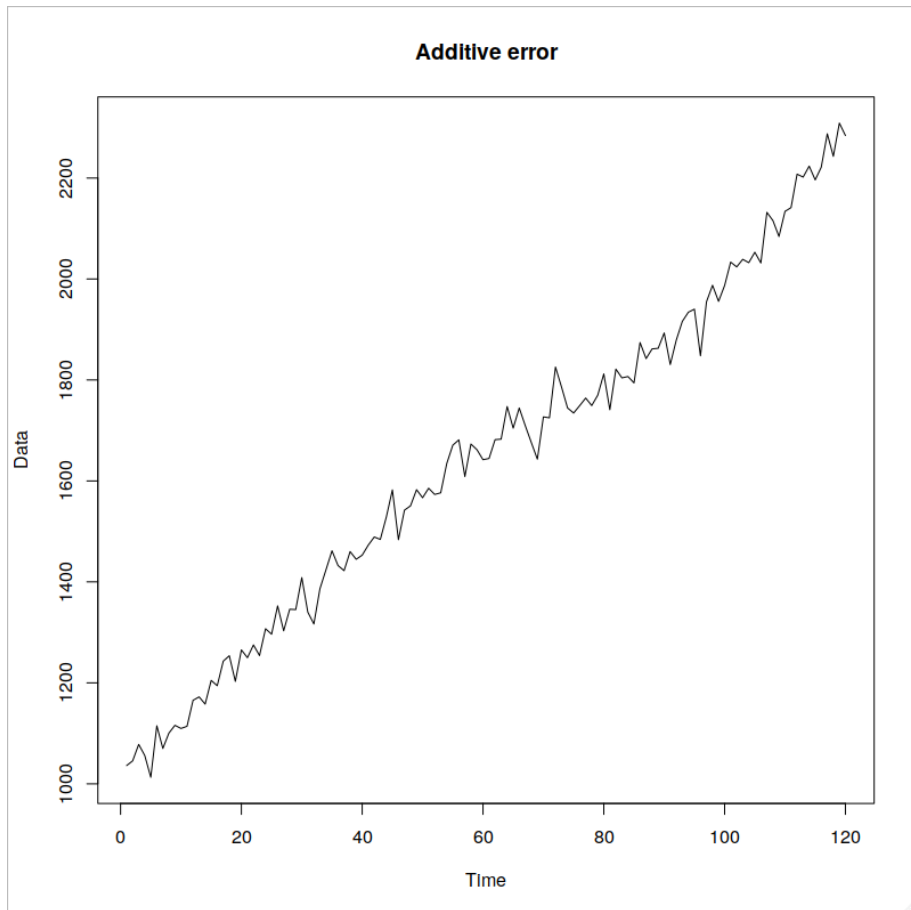
# Introduction to ETS

- Different types of components:

Trend	Seasonality		
	None "N"	Additive "A"	Multiplicative "M"
None "N"			
Additive "A"			
Additive Damped "Ad"			
Multiplicative "M"			
Multiplicative Damped "Md"			

# Introduction to ETS

- And two types of errors:



# Introduction to ETS

- ETS taxonomy includes:
  - 2 types of errors,
  - 5 types of trends,
  - 3 types of seasonality.
- Which gives us 30 models:
  - 6 pure additive models,
  - 6 pure multiplicative models,
  - 18 mixed models.

# Introduction to ETS

- All pure models make sense:
  - Additive assume that the variables can be positive, negative or zero;
  - Multiplicative ones assume that the response variable can only be positive.
- Not all mixed models are reasonable
  - For example, ETS(A,M,A) model:
$$y_t = l_{t-1}b_{t-1} + s_{t-m} + \varepsilon_t$$
  - Why?
- You can fit them and produce forecasts, but they break easily.

# Introduction to ETS

- The list of reasonable ETS models:
  - Additive error ( $\varepsilon_t = \epsilon_t$ ):

		Seasonal		
		"N"	A	M
Trend	"N"	$y_t = l_{t-1} + \epsilon_t$	$y_t = l_{t-1} + s_{t-m} + \epsilon_t$	-
	"A"	$y_t = l_{t-1} + b_{t-1} + \epsilon_t$	$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t$	-
	"Ad"	$y_t = l_{t-1} + \phi b_{t-1} + \epsilon_t$	$y_t = l_{t-1} + \phi b_{t-1} + s_{t-m} + \epsilon_t$	-
	"M"	-	-	-
	"Md"	-	-	-

- It is usually assumed that  $\epsilon_t \sim N(0, \sigma^2)$

# Introduction to ETS

- The list of reasonable ETS models:
  - Multiplicative error ( $\varepsilon_t = 1 + \epsilon_t$ ):

		Seasonal		
		"N"	A	M
Trend	"N"	$y_t = l_{t-1}(1 + \epsilon_t)$	$y_t = (l_{t-1} + s_{t-m})(1 + \epsilon_t)$	$y_t = l_{t-1}s_{t-m}(1 + \epsilon_t)$
	"A"	$y_t = (l_{t-1} + b_{t-1})(1 + \epsilon_t)$	$y_t = (l_{t-1} + b_{t-1} + s_{t-m})(1 + \epsilon_t)$	$y_t = (l_{t-1} + b_{t-1})s_{t-m}(1 + \epsilon_t)$
	"Ad"	$y_t = (l_{t-1} + \phi b_{t-1})(1 + \epsilon_t)$	$y_t = (l_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \epsilon_t)$	$y_t = (l_{t-1} + \phi b_{t-1})s_{t-m}(1 + \epsilon_t)$
	"M"	$y_t = l_{t-1}b_{t-1}(1 + \epsilon_t)$	-	$y_t = l_{t-1}b_{t-1}s_{t-m}(1 + \epsilon_t)$
	"Md"	$y_t = l_{t-1}b_{t-1}^\phi(1 + \epsilon_t)$	-	$y_t = l_{t-1}b_{t-1}^\phi s_{t-m}(1 + \epsilon_t)$

- Usual assumption is  $\epsilon_t \sim N(0, \sigma^2)$ , but in smooth it is  $1 + \epsilon_t \sim \log N(0, \sigma^2)$

# Introduction to ETS

- So far, we've discussed only one part of ETS model.
- It is called “measurement equation” and it shows how the data is formed.
  - For example, with local level model:  $y_t = l_{t-1} + \epsilon_t$
- But level, trend and seasonal components might change over time.
- So, there should be a mechanism for update of states.



# Introduction to ETS

- **Transition equation** – the equation that shows how the components change over time.
- For example, for ETS(A,N,N):

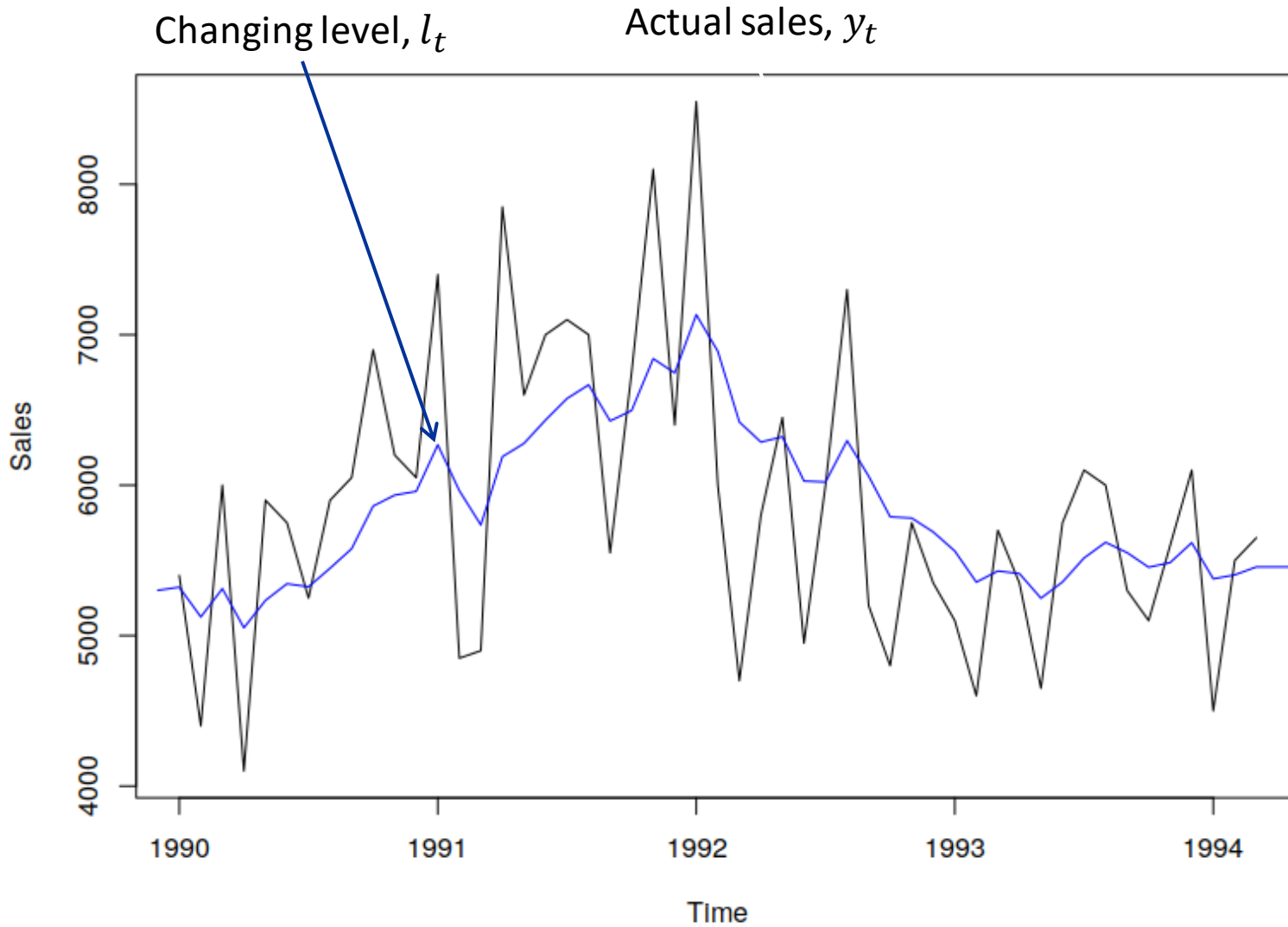
$$l_t = l_{t-1} + \alpha \epsilon_t$$

- Any ETS model consists of these two parts.
- So, ETS(A,N,N) can be represented as:

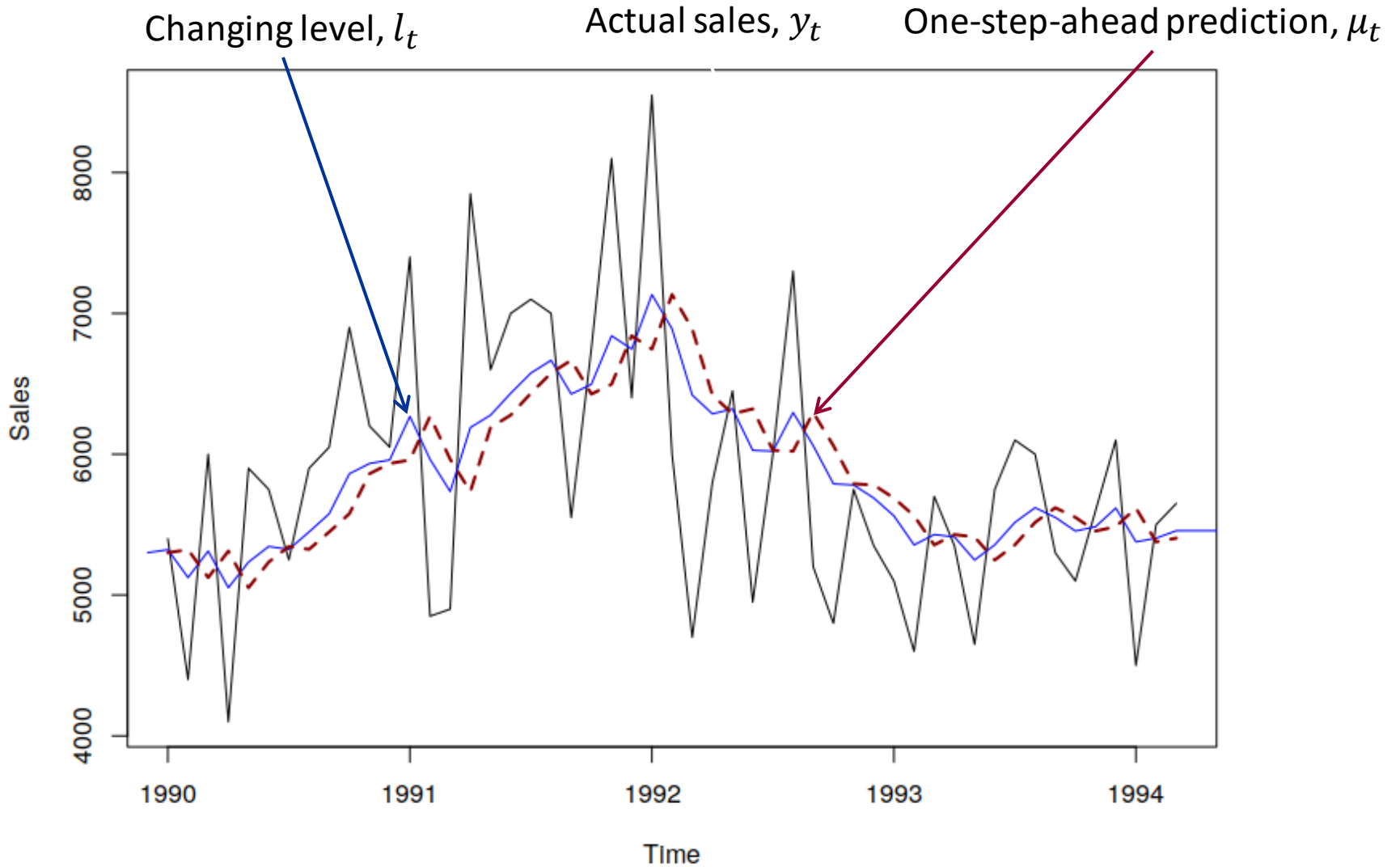
$$y_t = l_{t-1} + \epsilon_t$$

$$l_t = l_{t-1} + \alpha \epsilon_t$$

# Introduction to ETS



# Introduction to ETS



# Introduction to ETS

- In general pure additive model can be summarised as:

$$y_t = \mathbf{w}'\mathbf{v}_{t-1} + \epsilon_t$$
$$\mathbf{v}_t = \mathbf{F}\mathbf{v}_{t-1} + \mathbf{g}\epsilon_t$$

- $\mathbf{g}$  is the persistence vector... The rest is not important.
- See Hyndman et al. (2008) for details.
- Additional resources:
  - For pure additive models: <http://tiny.cc/znxc9y>
  - For pure multiplicative models: <http://tiny.cc/2oxc9y>
  - For the mixed ones: <http://tiny.cc/emxc9y>

# Introduction to ETS

- Why do we bother with ETS **model** and not just stick with **methods**?
- Models allow us:
  - producing point forecasts,
  - producing prediction intervals,
  - selecting the components (error / trend /seasonal),
  - adding explanatory variables (weather, promotions),
- + they can be estimated in a way, guaranteeing that the forecasts will be more stable.

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Local level model

- Local level model underlies SES.
- It can be:
  - either additive – ETS(A,N,N):

$$y_t = l_{t-1} + \epsilon_t$$

$$l_t = l_{t-1} + \alpha\epsilon_t$$

- or multiplicative – ETS(M,N,N):

$$y_t = l_{t-1}(1 + \epsilon_t)$$

$$l_t = l_{t-1}(1 + \alpha\epsilon_t)$$



# Local level model

- In the additive case:

$$y_t = l_{t-1} + \epsilon_t$$

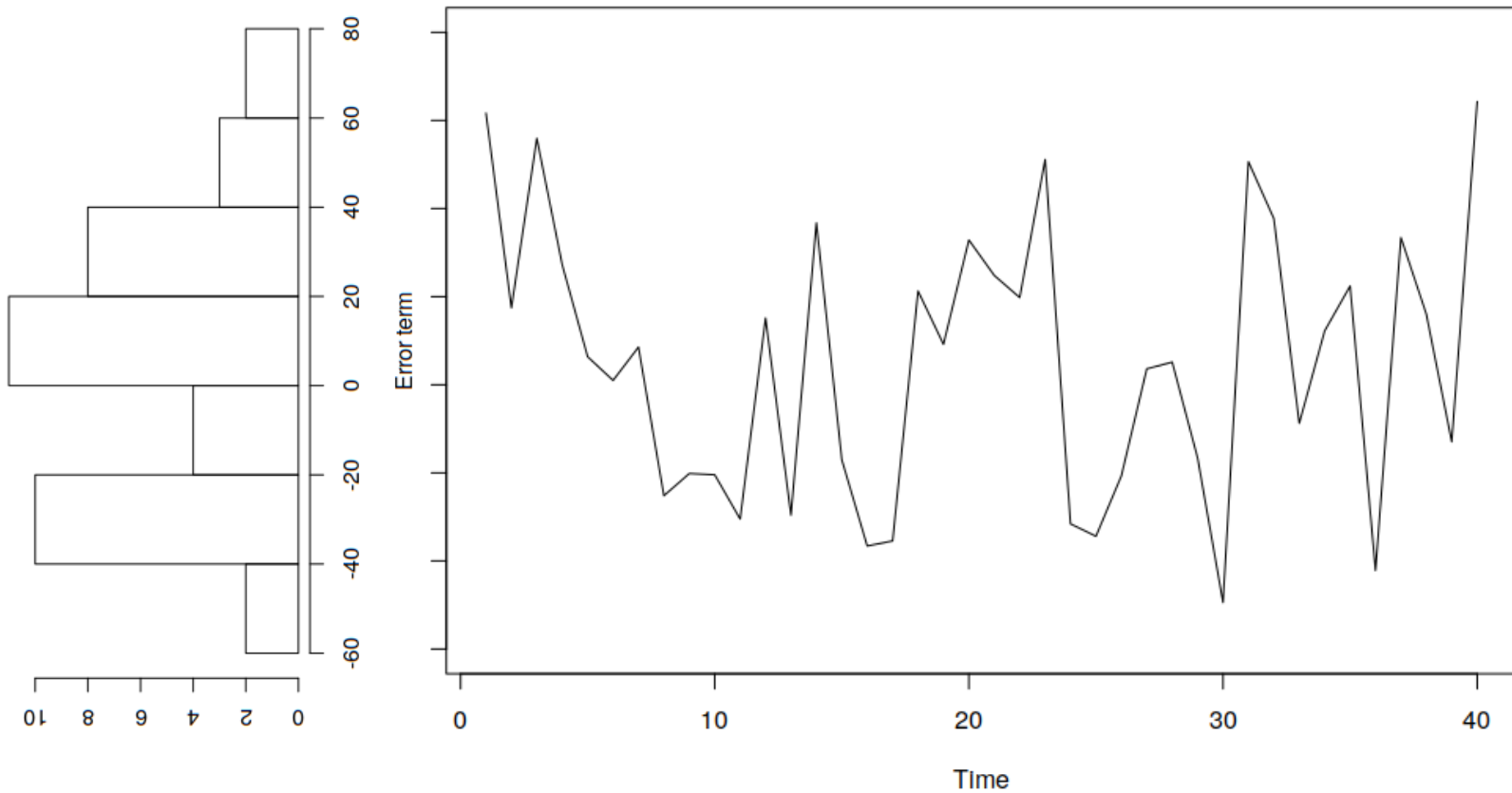
$$l_t = l_{t-1} + \alpha\epsilon_t$$

$$\epsilon_t \sim N(0, \sigma^2)$$

- The  $l_t$  represents the anticipated average demand in period  $t$  (e.g. average demand on ice cream);
- The  $\epsilon_t$  represents the unexpected demand (e.g. Nikos is in town);
- $\sigma$  is the size of the uncertainty about the demand;
- $\alpha$  is the rate of change of the level of demand;
- $\alpha\epsilon_t$  is the persistent effect on the level (e.g. Nikos goes out with his friends);

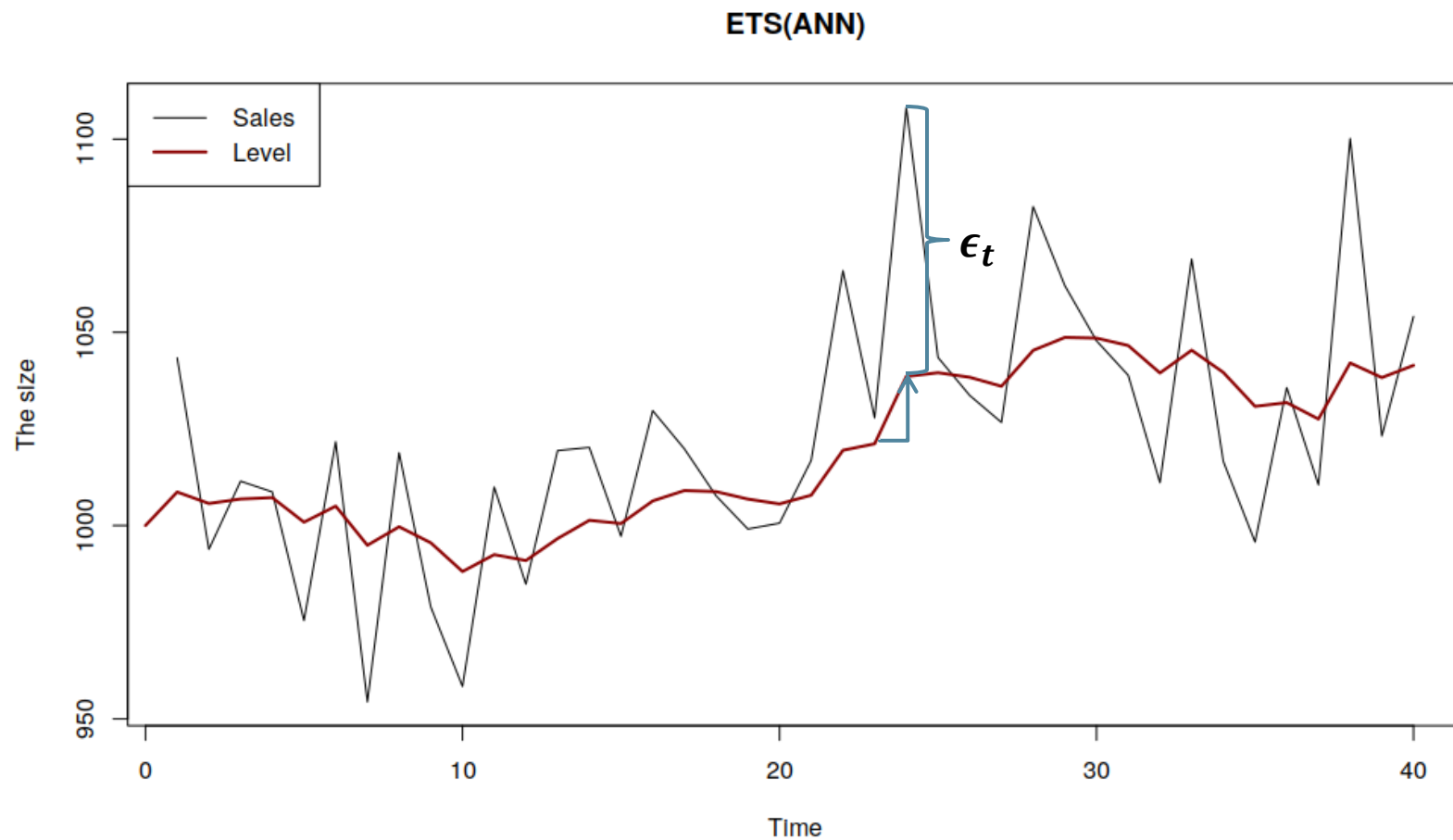
# Local level model

- An example.  $\sigma = 30$



# Local level model

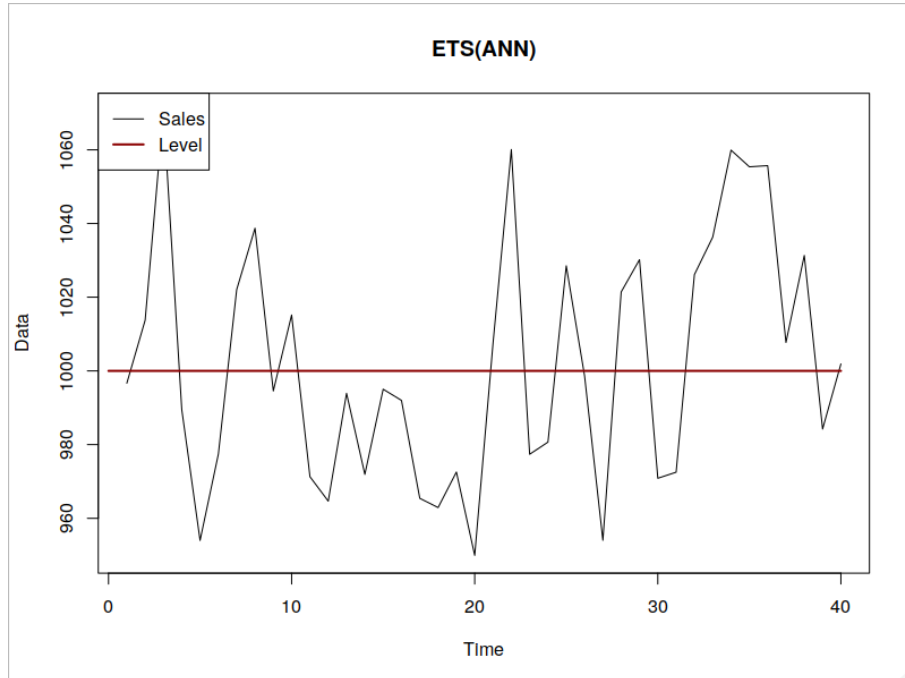
- An example with  $\alpha = 0.2$  and  $\sigma = 30$



# Local level model

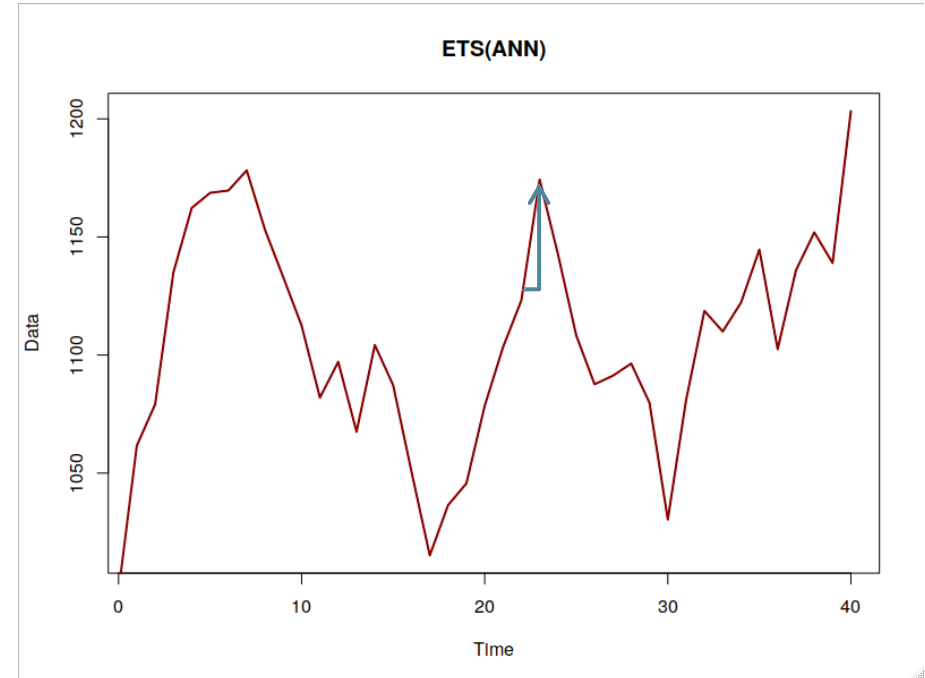
- Two cases of interest:

$$\alpha = 0$$



Global mean (global level)

$$\alpha = 1$$



Naïve (random walk)

# Local level model

- What should we do if the parameters are unknown?

- The model needs to be constructed:

$$\hat{y}_t = \hat{l}_{t-1}$$

$$e_t = y_t - \hat{y}_t$$

$$\hat{l}_t = \hat{l}_{t-1} + \hat{\alpha}e_t$$

- With some values of  $\hat{l}_0$  and  $\hat{\alpha}$ .
- Let's see how to do it in a small Excel exercise.

# Local level model

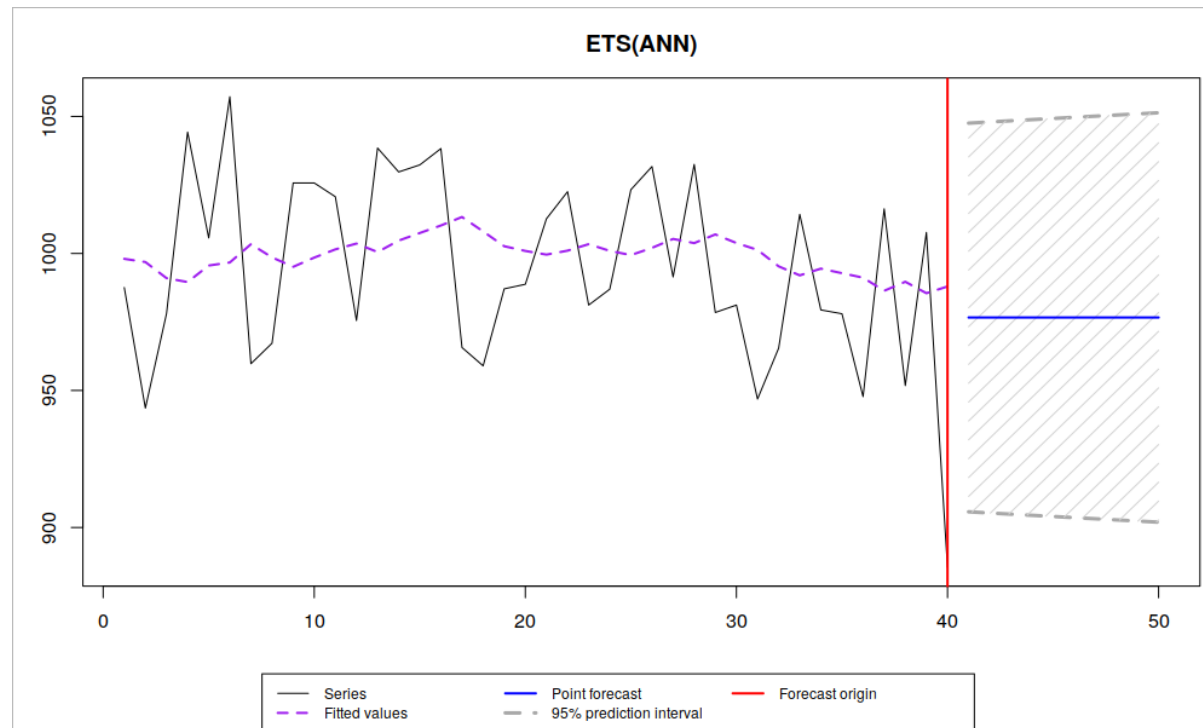
- One of the ways to estimate  $\hat{l}_0$  and  $\hat{\alpha}$  is by maximising the likelihood function.
- This function shows how likely it is that the data corresponds to the model with specific parameters.
- In case of ETS(A,N,N) maximising this likelihood is equivalent to minimising the Mean Squared Error.
- Why do we need it?
- Because:
  - the parameters of the model will be more stable;
  - we have a mechanism of model selection.

# Local level model

- The forecast is the straight line:

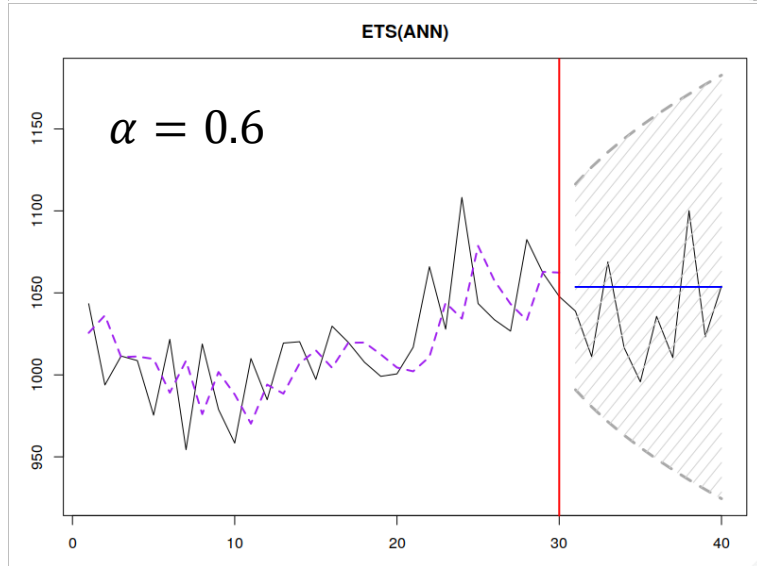
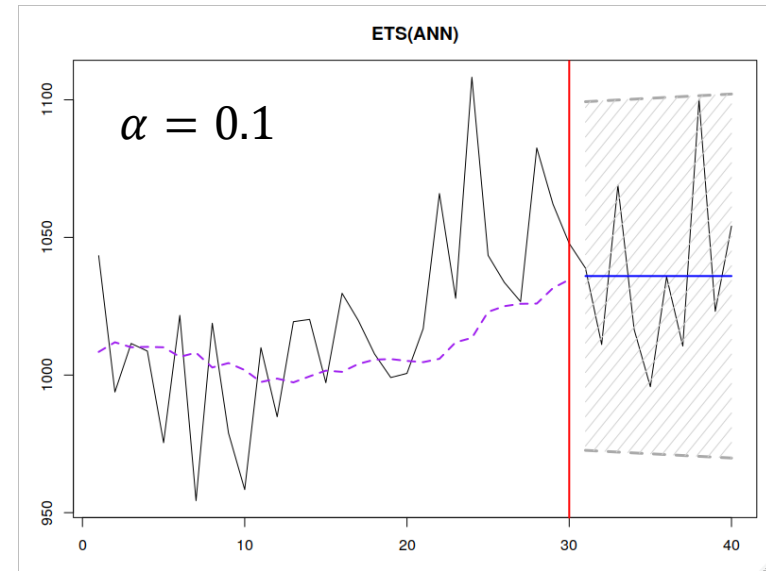
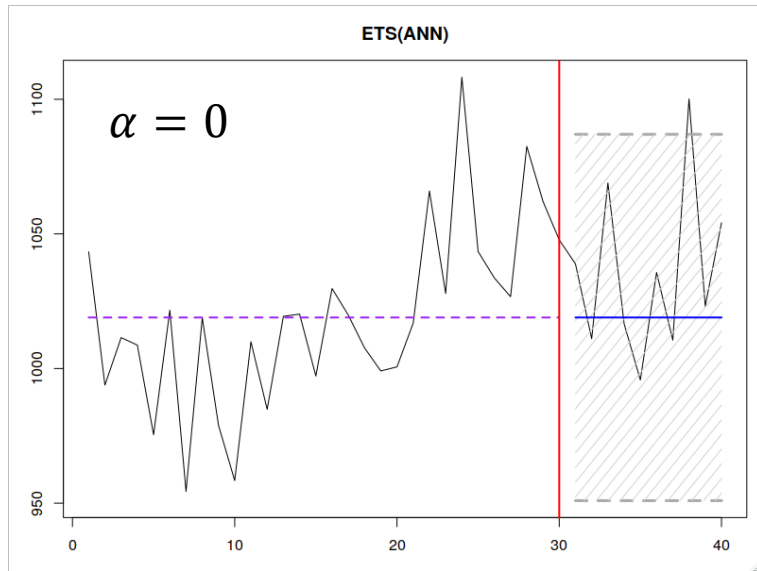
$$\hat{y}_{t+h} = l_t$$

- And we can construct prediction intervals based on  $\epsilon_t \sim N(0, \sigma^2)$

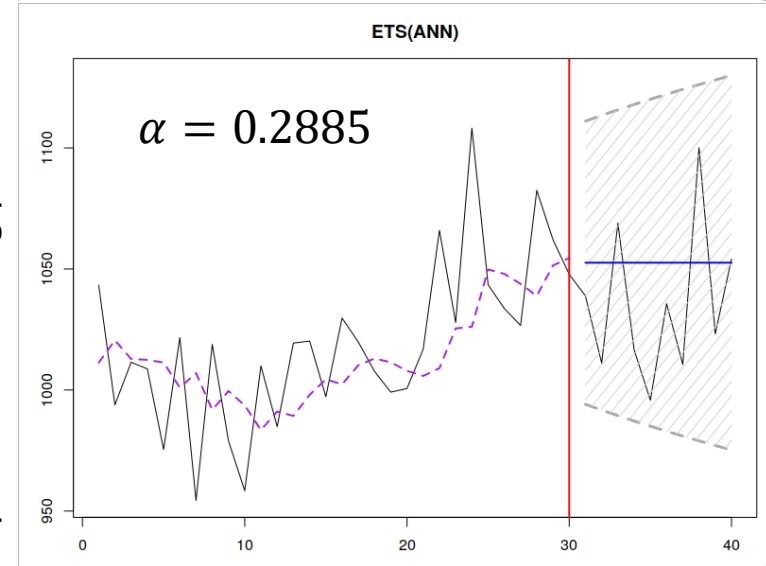


# Local level model

- An example with different values of  $\hat{\alpha}$ :



Optimal smoothing parameter





# Local level model

- Summarising:
  1.  $\alpha$  regulates the rate of change of the local level;
  2. The higher it is, the higher the responsiveness of the model;
  3. It also regulates the width of prediction interval;
  4. The higher  $\alpha$  means higher uncertainty, because of (2);
  5. We can optimise  $\alpha$ ;
  6. But optimal in-sample  $\neq$  optimal in the holdout.

# Local level model

- ETS(M,N,N) has properties similar to ETS(A,N,N):

$$\begin{aligned}y_t &= l_{t-1}(1 + \epsilon_t) \\l_t &= l_{t-1}(1 + \alpha\epsilon_t) \\1 + \epsilon_t &\sim \log N(0, \sigma^2)\end{aligned}$$

- It is estimated differently:

$$\begin{aligned}\hat{y}_t &= \hat{l}_{t-1} \\e_t &= \frac{y_t - \hat{y}_t}{\hat{y}_t} \\\hat{l}_t &= \hat{l}_{t-1}(1 + \hat{\alpha}e_t)\end{aligned}$$

- The forecast is the straight line again.
- But the prediction interval increases with the increase of level.

# Local level model

- How many parameters do we need to estimate in ETS(A,N,N)?
- Three:
  - $\hat{l}_0$ ,  $\hat{\alpha}$  and  $\hat{\sigma}^2$ .
- No matter how we try, there will always be:
  - An uncertainty in the error term  $e_t$ ,
  - An uncertainty in the estimation of parameters  $\hat{l}_0$ ,  $\hat{\alpha}$  and  $\hat{\sigma}^2$ ,
  - An uncertainty about the selected model.

# Local level model

- An exercise in R, using smooth package:
  - ETS(A,N,N) with different values of parameters;
  - ETS(A,N,N) with the optimal smoothing parameter;
  - Similarly, with ETS(M,N,N).

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Local trend model

- Are there any other components in time series?
- Why not add a trend component, ETS(A,A,N) :

ETS(A,A,N)

$$\begin{aligned}y_t &= l_{t-1} + b_{t-1} + \epsilon_t \\l_t &= l_{t-1} + b_{t-1} + \alpha \epsilon_t \\b_t &= b_{t-1} + \beta \epsilon_t \\\epsilon_t &\sim N(0, \sigma^2)\end{aligned}$$

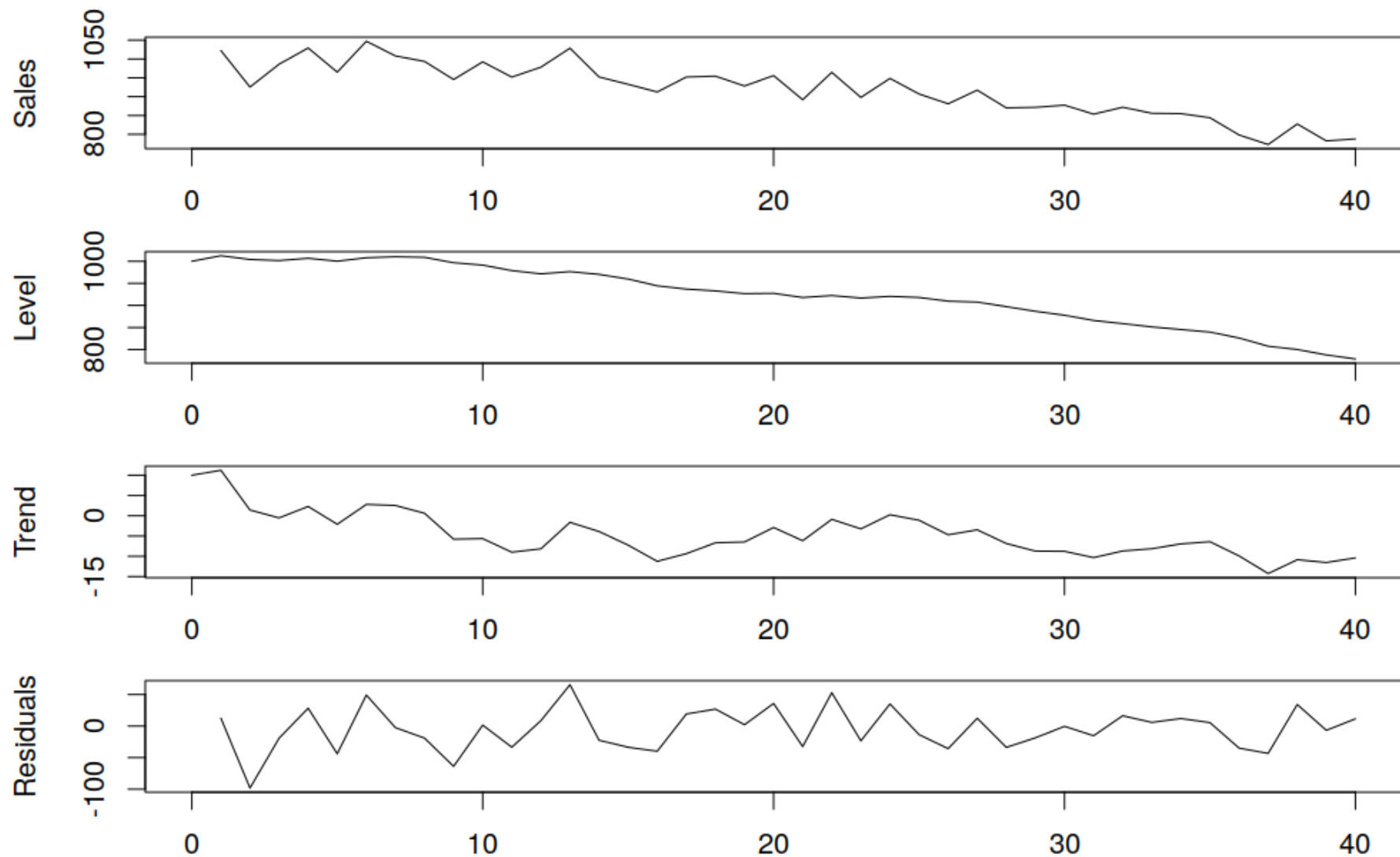
ETS(A,N,N)

$$\begin{aligned}y_t &= l_{t-1} + \epsilon_t \\l_t &= l_{t-1} + \alpha \epsilon_t \\\epsilon_t &\sim N(0, \sigma^2)\end{aligned}$$

- The mechanism is similar to ETS(A,N,N).
- This model underlies “Holt’s method”.
- But now we also update the trend.

# Local trend model

- Decomposition of time series due to ETS(A,A,N):



# Local trend model

- ETS(A,A,N) :

$$y_t = l_{t-1} + b_{t-1} + \epsilon_t$$

$$l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t$$

$$b_t = b_{t-1} + \beta\epsilon_t$$

- $\alpha$  has the same property as in ETS(A,N,N).
- $\beta$  defines the rate of change of the trend:
  - $\beta = 0$ ,  $b_t = b_{t-1}$ , the trend is constant;
  - $\beta = 1$ ,  $b_t = b_{t-1} + \epsilon_t$ , the trend is changing rapidly.



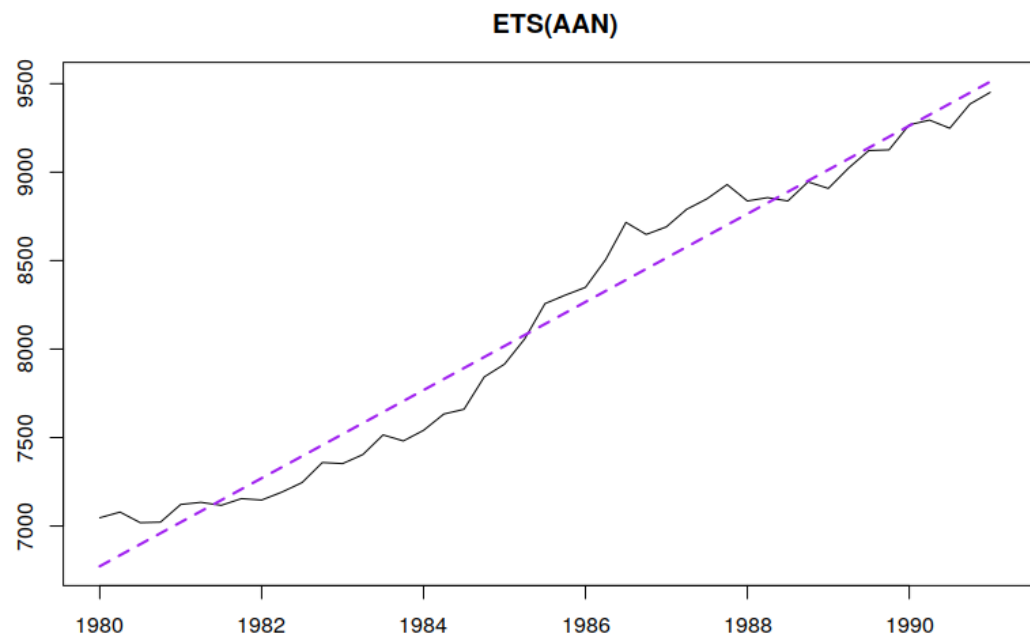
# Local trend model

- If both  $\alpha = 0$  and  $\beta = 0$ , then we have a deterministic trend:

$$y_t = l_{t-1} + b_{t-1} + \epsilon_t$$

$$l_t = l_{t-1} + b_{t-1}$$

$$b_t = b_{t-1} = b_0$$

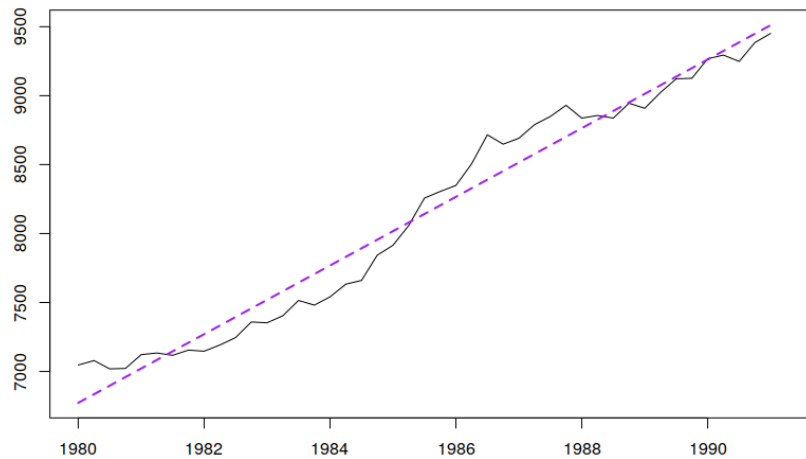


# Local trend model

- But it's difficult to select the appropriate parameters manually:

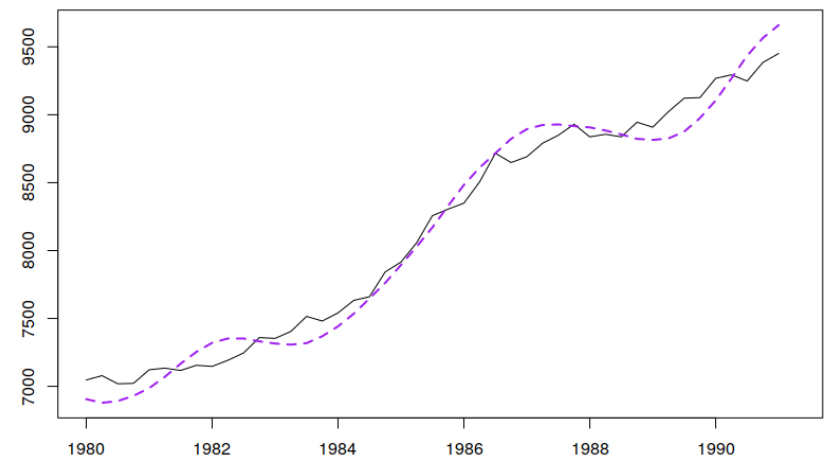
$$\alpha = 0, \beta = 0$$

ETS(AAN)



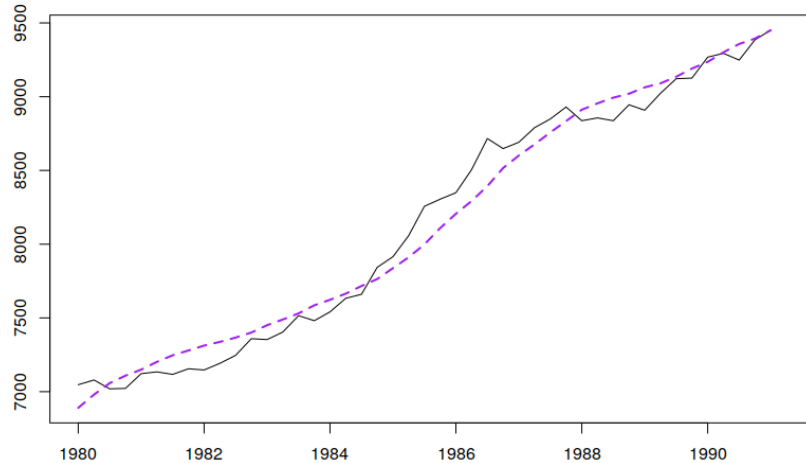
$$\alpha = 0, \beta = 0.2$$

ETS(AAN)



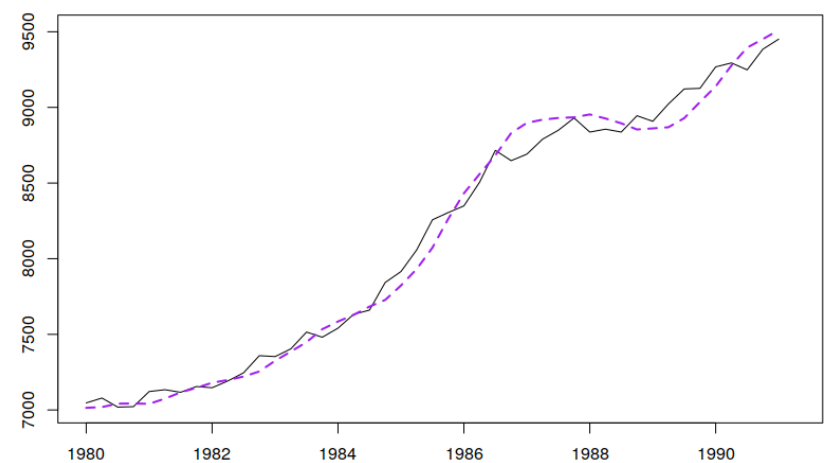
$$\alpha = 0.2, \beta = 0$$

ETS(AAN)



$$\alpha = 0.2, \beta = 0.2$$

ETS(AAN)



# Local trend model

- The model can be constructed using:

$$\hat{y}_t = \hat{l}_{t-1} + \hat{b}_{t-1}$$

$$e_t = y_t - \hat{y}_t$$

$$\hat{l}_t = \hat{l}_{t-1} + \hat{b}_{t-1} + \hat{\alpha}e_t$$

$$\hat{b}_t = \hat{b}_{t-1} + \hat{\beta}e_t$$

- We can use likelihood,
  - the estimation procedure is the same as in ETS(A,N,N).
- MS Excel exercise?

- The forecast is a line:

$$\hat{y}_{t+h} = \hat{l}_t + h\hat{b}_t$$

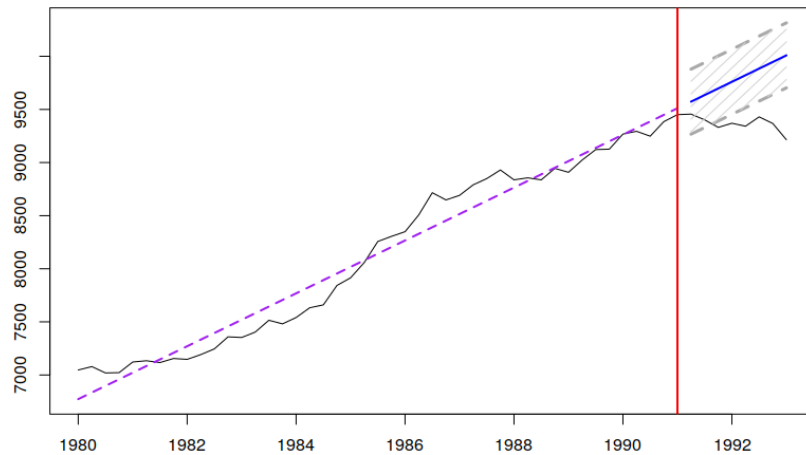
- The width of intervals changes with the change of both smoothing parameters.

# Local trend model

- The influence of parameters on forecasts:

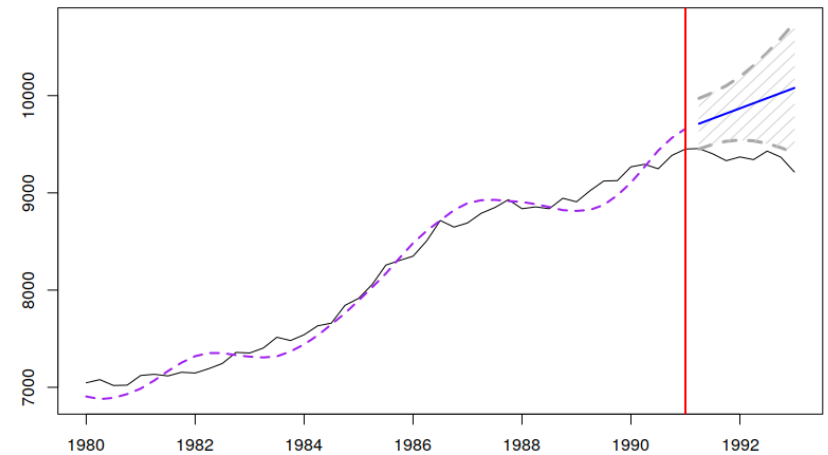
$$\alpha = 0, \beta = 0$$

ETS(AAN)



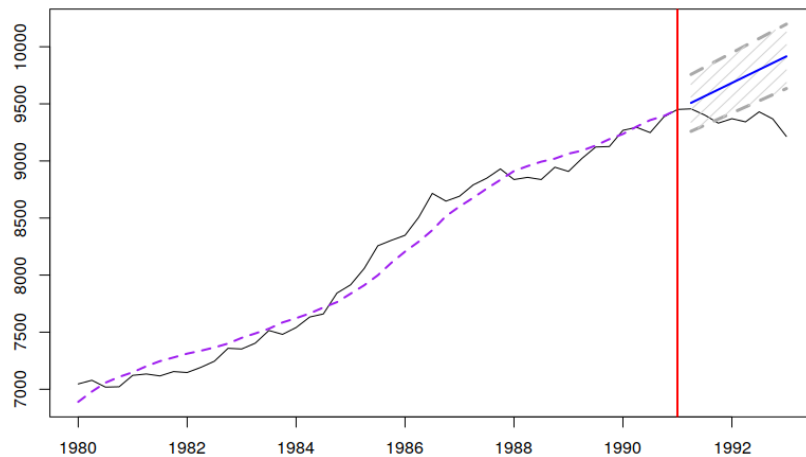
$$\alpha = 0, \beta = 0.2$$

ETS(AAN)



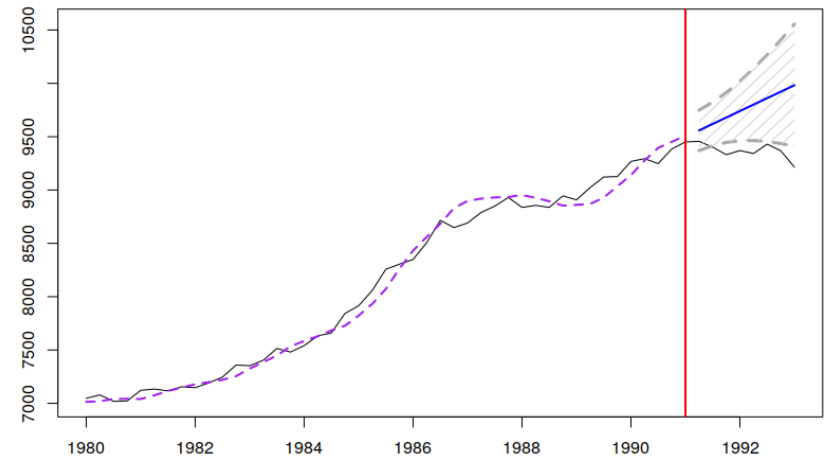
$$\alpha = 0.2, \beta = 0$$

ETS(AAN)



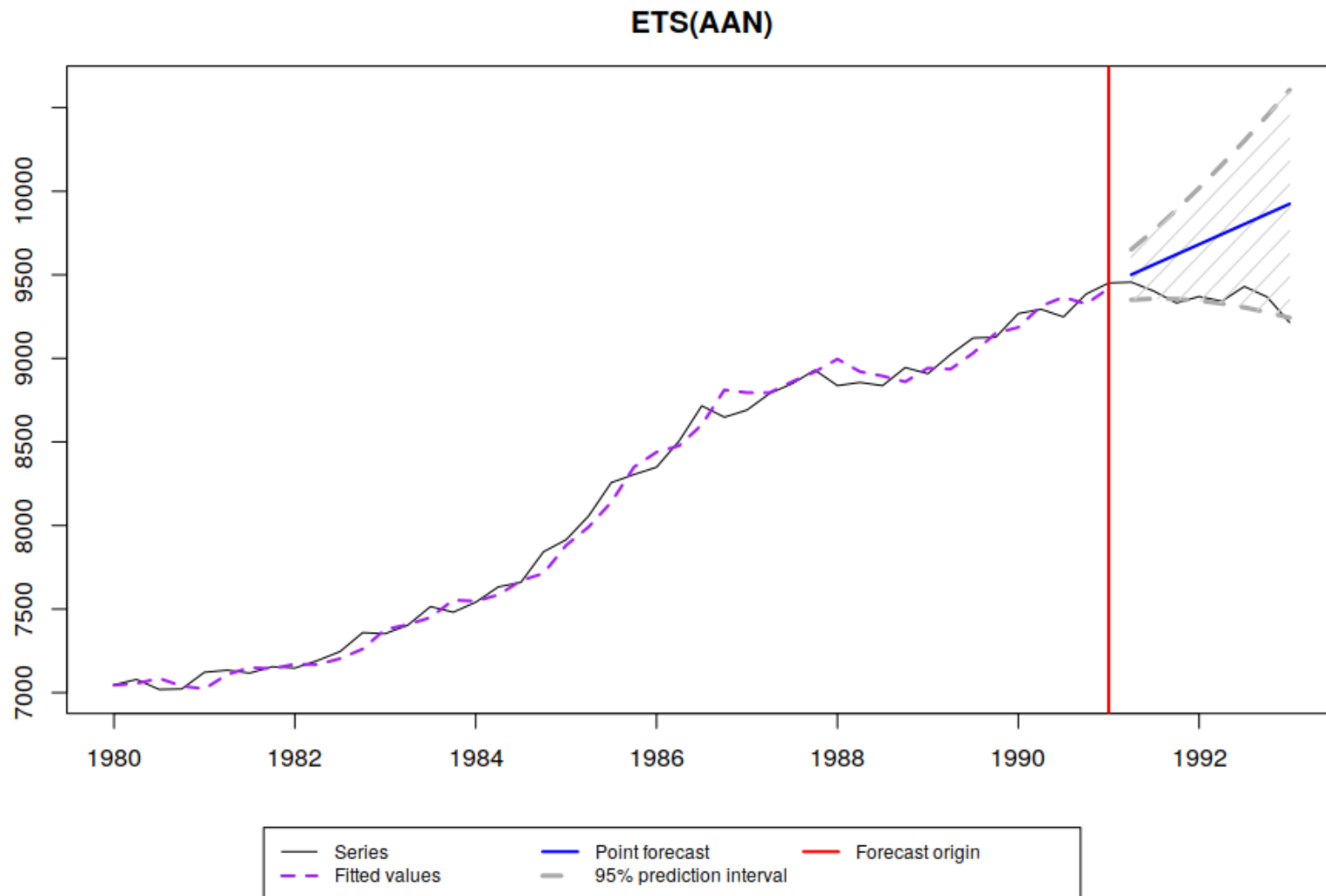
$$\alpha = 0.2, \beta = 0.2$$

ETS(AAN)



# Local trend model

- An example with optimal parameters:

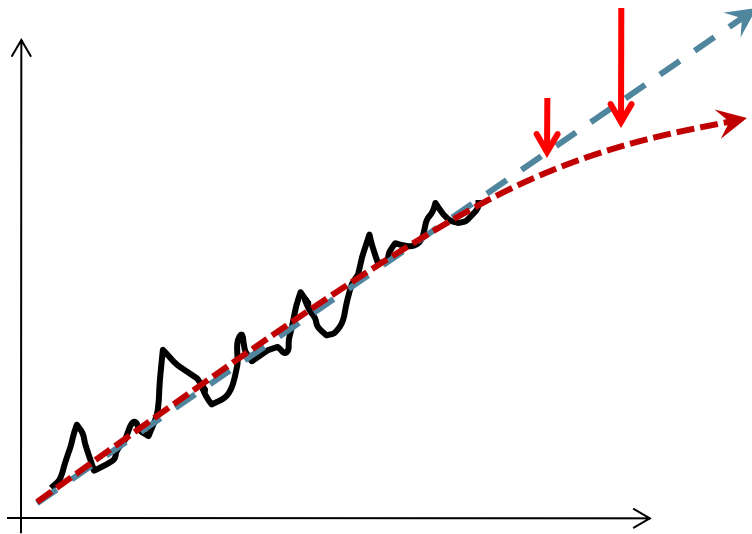


# Local trend model

- How many parameters do we need to estimate in ETS(A,A,N)?
- Five:
  - $\hat{l}_0, \hat{b}_0, \hat{\alpha}, \hat{\beta}$  and  $\hat{\sigma}^2$ .
- ETS(M,A,N) is similar, but assumes a different error term.
  - What does it imply?

# Damped trend model

- Is it reasonable to assume that sales will go up / down forever?



How about competition?  
How about the size of the market?

Pressure curbs the trend  
(or maybe the structure is just like that!)

This is called a **damped trend**

- $ETS(A,Ad,N)$

# Damped trend model

ETS(A,Ad,N)

$$\begin{aligned}y_t &= l_{t-1} + \phi b_{t-1} + \epsilon_t \\l_t &= l_{t-1} + \phi b_{t-1} + \alpha \epsilon_t \\b_t &= \phi b_{t-1} + \beta \epsilon_t \\\epsilon_t &\sim N(0, \sigma^2)\end{aligned}$$

ETS(A,A,N)

$$\begin{aligned}y_t &= l_{t-1} + b_{t-1} + \epsilon_t \quad l_t = \\& l_{t-1} + b_{t-1} + \alpha \epsilon_t \quad b_t = \\& b_{t-1} + \beta \epsilon_t \\\epsilon_t &\sim N(0, \sigma^2)\end{aligned}$$

- The construction is similar to ETS(A,A,N).
- The forecast from an estimated ETS(A,Ad,N):

$$\hat{y}_{t+h} = \hat{l}_t + \hat{b}_t \sum_{j=1}^h \hat{\phi}^j$$

- $\phi$  is a dampening parameter.

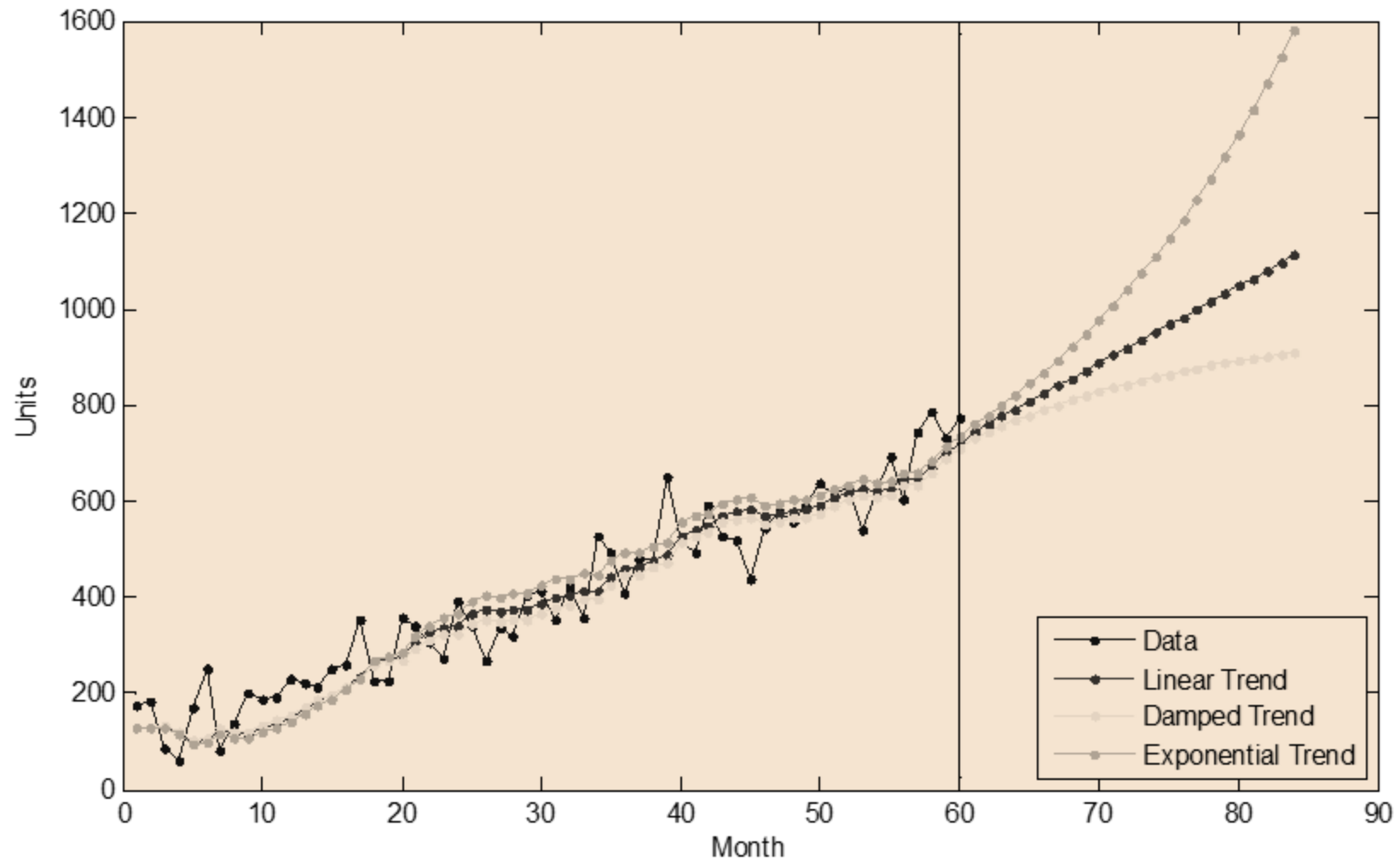


# Damped trend model

$$\hat{y}_{t+h} = \hat{l}_t + \hat{b}_t \sum_{j=1}^h \hat{\phi}^j$$

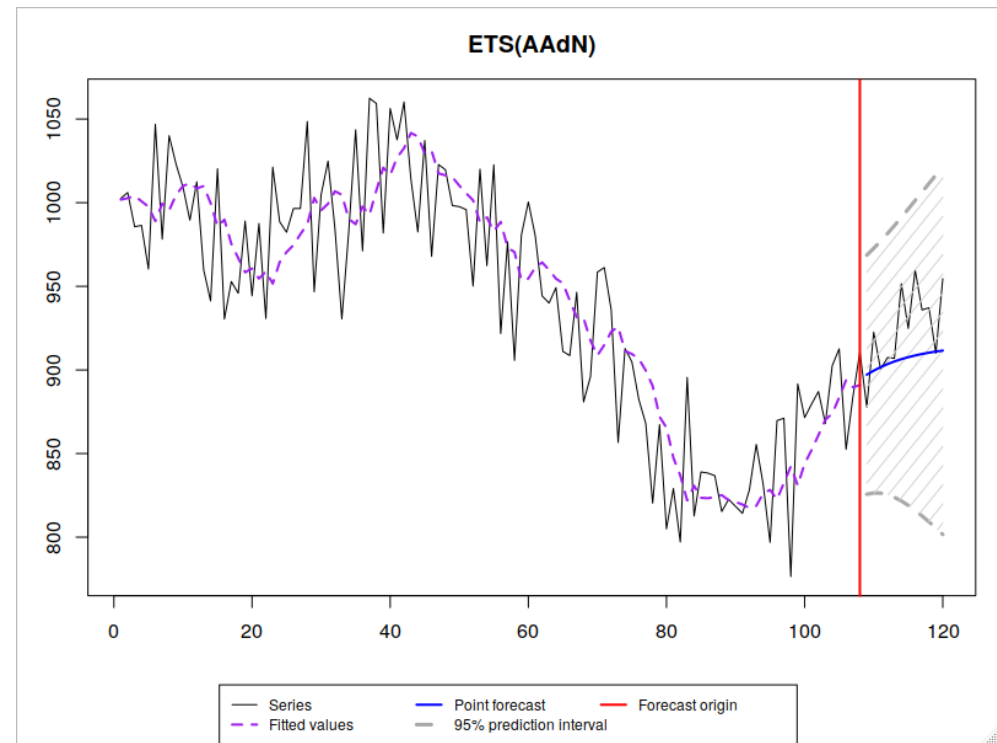
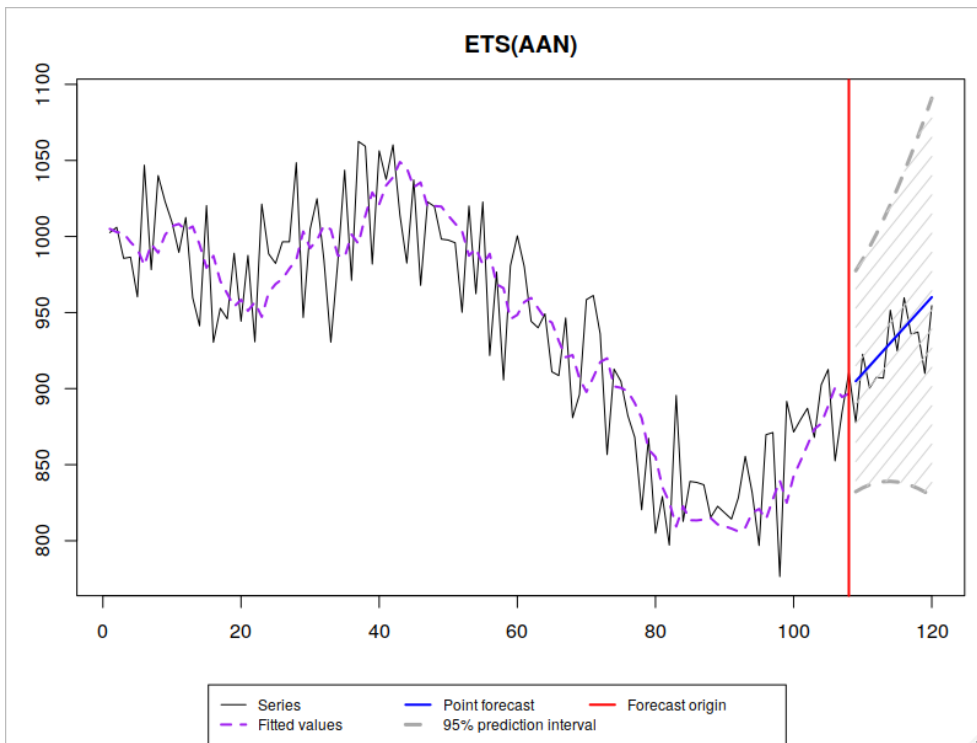
- Different values of  $\phi$ :
  - If  $\phi = 0$ , then the model reverts to ETS(A,N,N);
  - If  $\phi = 1$ , then it becomes ETS(A,A,N);
  - $0 < \phi < 1$ , then we have dampening;
  - $\phi > 1$ , exponential growth.

# Damped trend model



# Damped trend model

- An example



# Multiplicative trend model

- What about multiplicative trends?
- An example of ETS(M,Md,N):

$$y_t = l_{t-1} b_{t-1}^{\phi} (1 + \epsilon_t) \longrightarrow \log y_t = \log l_{t-1} + \phi \log b_{t-1} + \log(1 + \epsilon_t)$$

$$l_t = l_{t-1} b_{t-1}^{\phi} (1 + \alpha \epsilon_t)$$

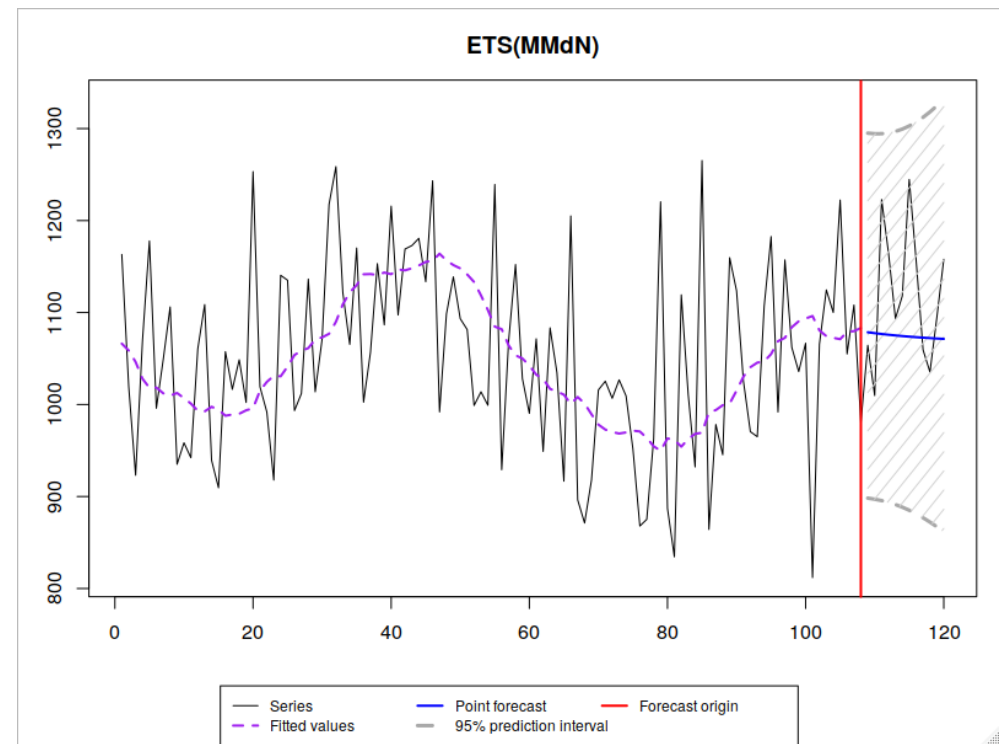
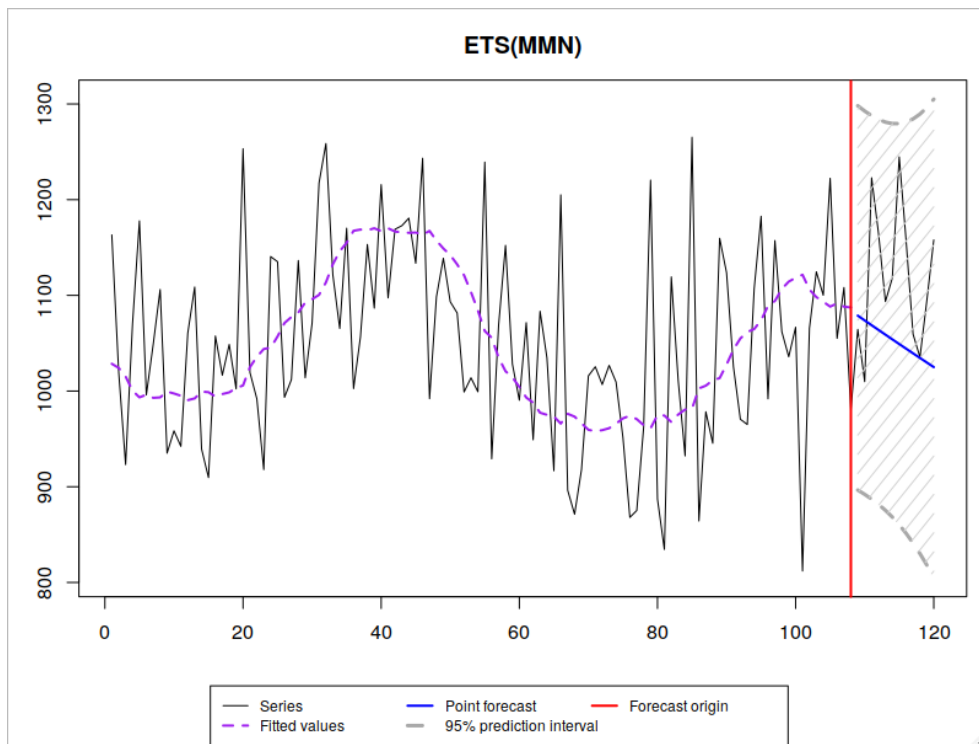
$$b_t = b_{t-1}^{\phi} (1 + \beta \epsilon_t)$$

$$1 + \epsilon_t \sim \log N(0, \sigma^2)$$

- It always produces exponential trajectories;
- The dampening parameter  $\phi$  slows them down.

# Multiplicative trend model

- An example:



# Local trend model

- An exercise in R, using smooth package:
  - `ETS(A,A,N);`
  - `ETS(A,Ad,N);`
  - `ETS(M,Md,N).`

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Trend seasonal model

- Now we can formulate a more complicated model.

- We start with ETS(A,A,A): ETS(A,A,N)

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t \quad y_t = l_{t-1} + b_{t-1} + \epsilon_t \quad l_t =$$

$$l_t = l_{t-1} + b_{t-1} + \alpha \epsilon_t \quad l_{t-1} + b_{t-1} + \alpha \epsilon_t \quad b_t =$$

$$b_t = b_{t-1} + \beta \epsilon_t \quad b_{t-1} + \beta \epsilon_t$$

$$s_t = s_{t-m} + \gamma \epsilon_t$$

$$\epsilon_t \sim N(0, \sigma^2) \quad \epsilon_t \sim N(0, \sigma^2)$$

- Almost the same as ETS(A,A,N).
- $\gamma$  now regulates the rate of change for the seasonal component.
- The forecast is produced as:

$$\hat{y}_{t+h} = \hat{l}_t + h\hat{b}_t + \hat{s}_{t-m+h}$$

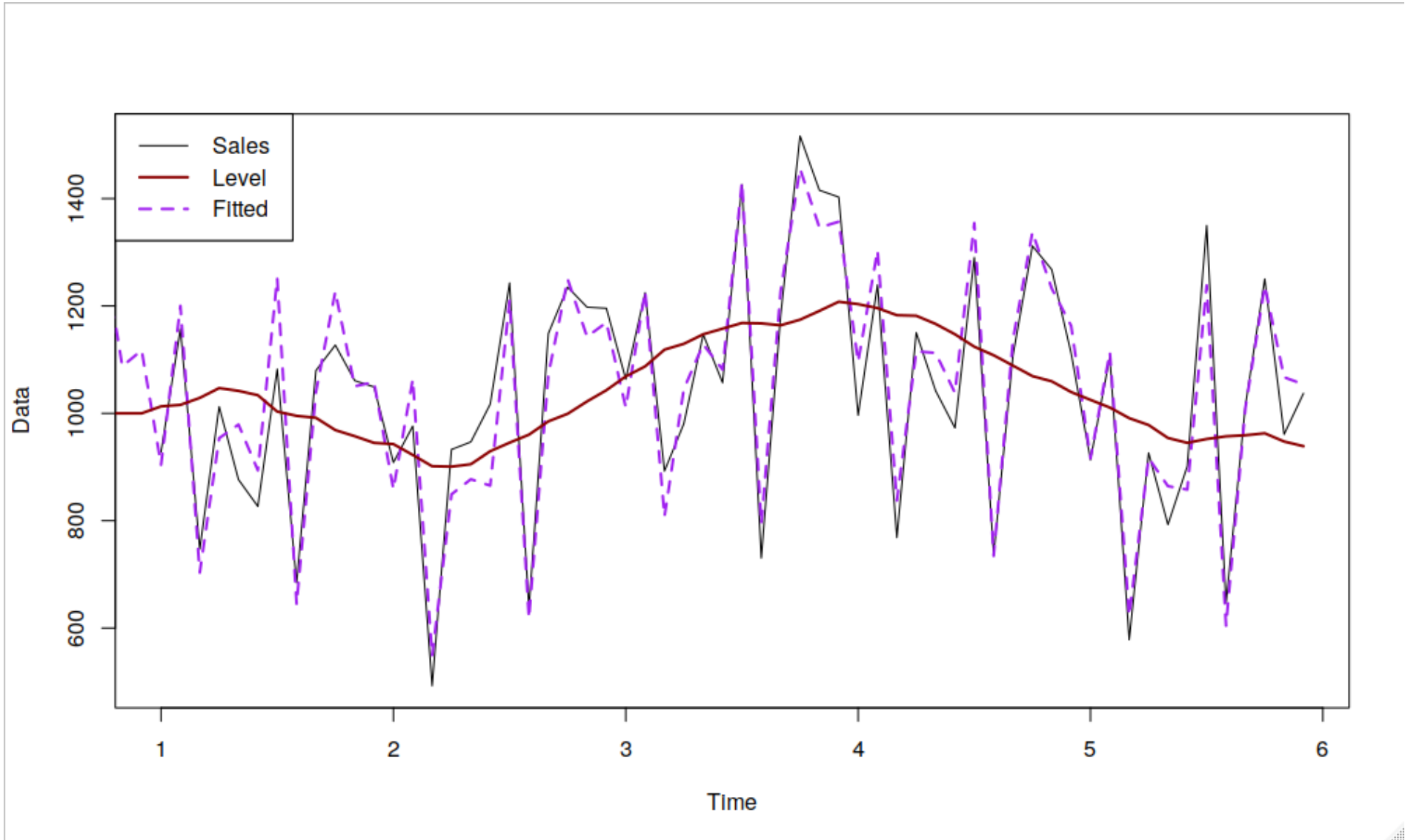


# Trend seasonal model

- The construction is similar to ETS(A,A,N), with a new equation for the seasonal component.
- The model underlies “Holt-Winters method”.
- How many parameters do we have in the trend seasonal model?
- $6 + m$ :
  - $\hat{l}_0, \hat{b}_0,$
  - $\hat{\alpha}, \hat{\beta}, \hat{\gamma},$
  - $m$  seasonal indices  $s_1, s_2, \dots, s_m,$
  - and  $\hat{\sigma}^2$ .

# Trend seasonal model

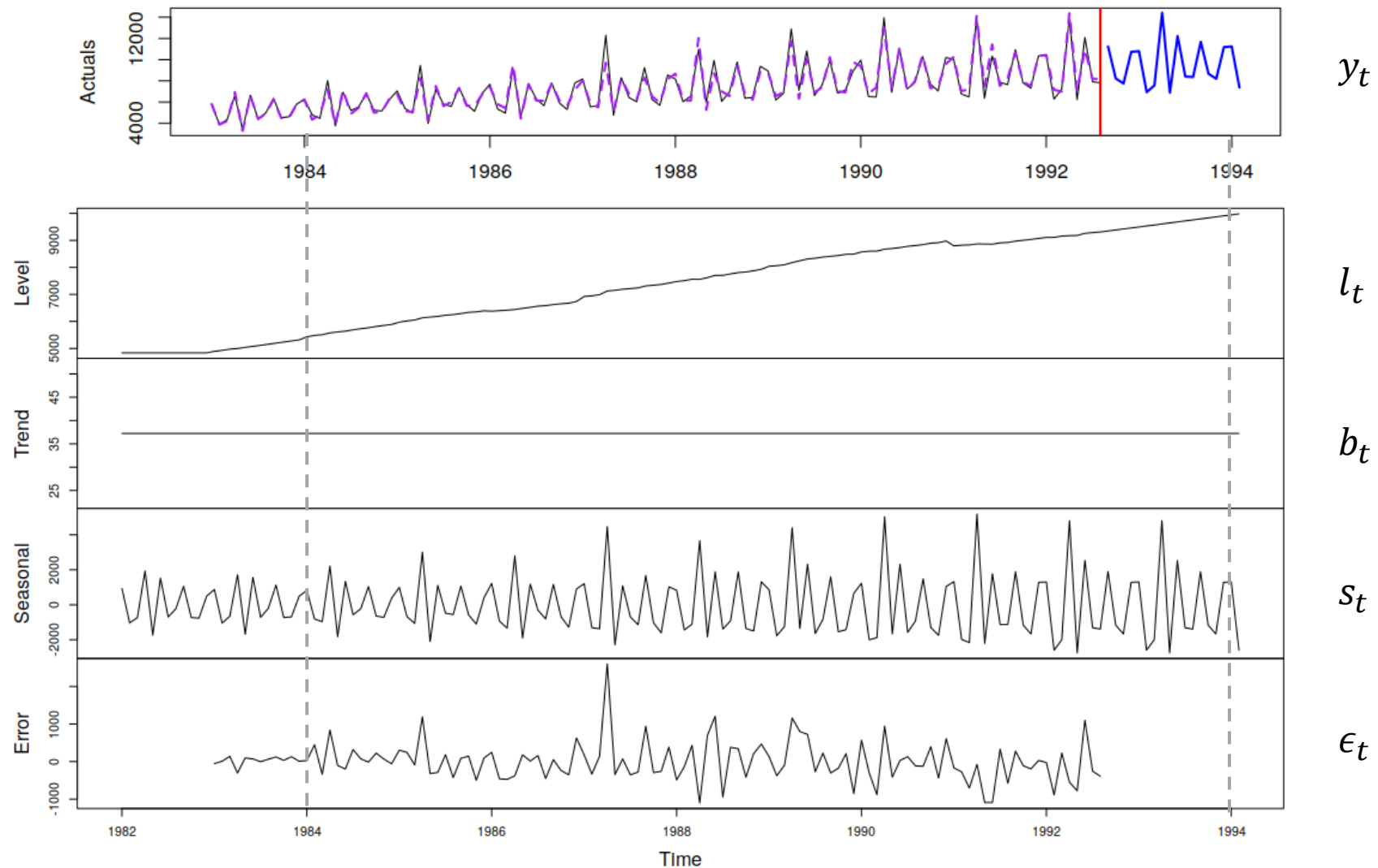
- An example with ETS(A,A,A):



# Trend seasonal model

- A series can be decomposed based on ETS(A,A,A):

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t$$



# Trend seasonal model

- Let's make things even more complicated...

- ETS(A,Ad,A):

$$y_t = l_{t-1} + \phi b_{t-1} + s_{t-m} + \epsilon_t$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha \epsilon_t$$

$$b_t = \phi b_{t-1} + \beta \epsilon_t$$

$$s_t = s_{t-m} + \gamma \epsilon_t$$

$$\epsilon_t \sim N(0, \sigma^2)$$

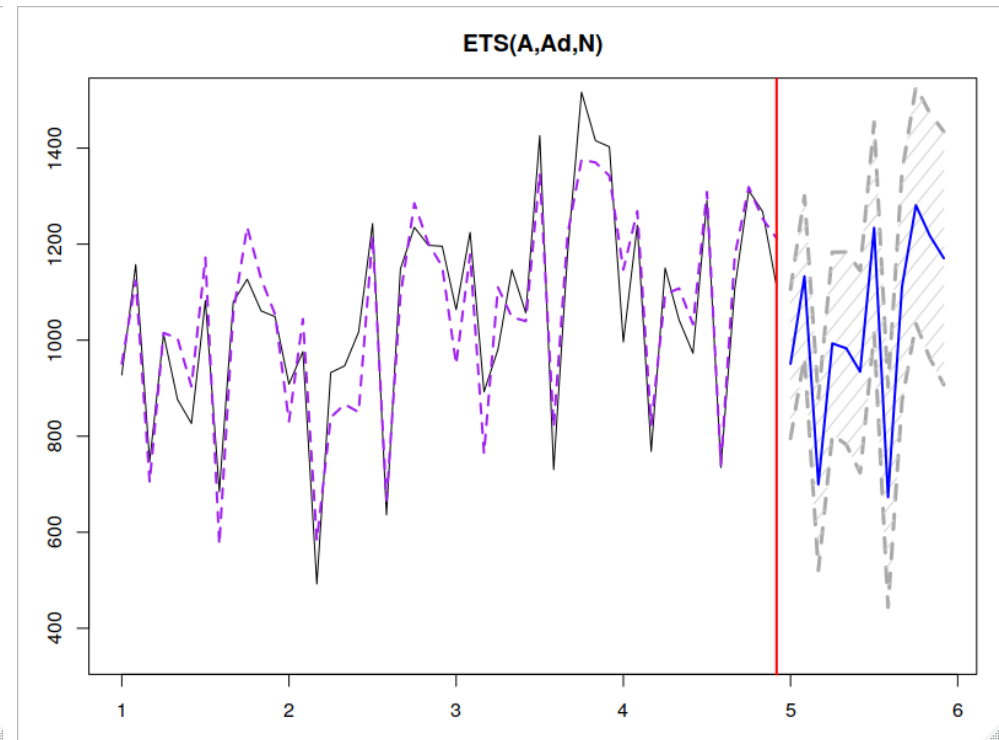
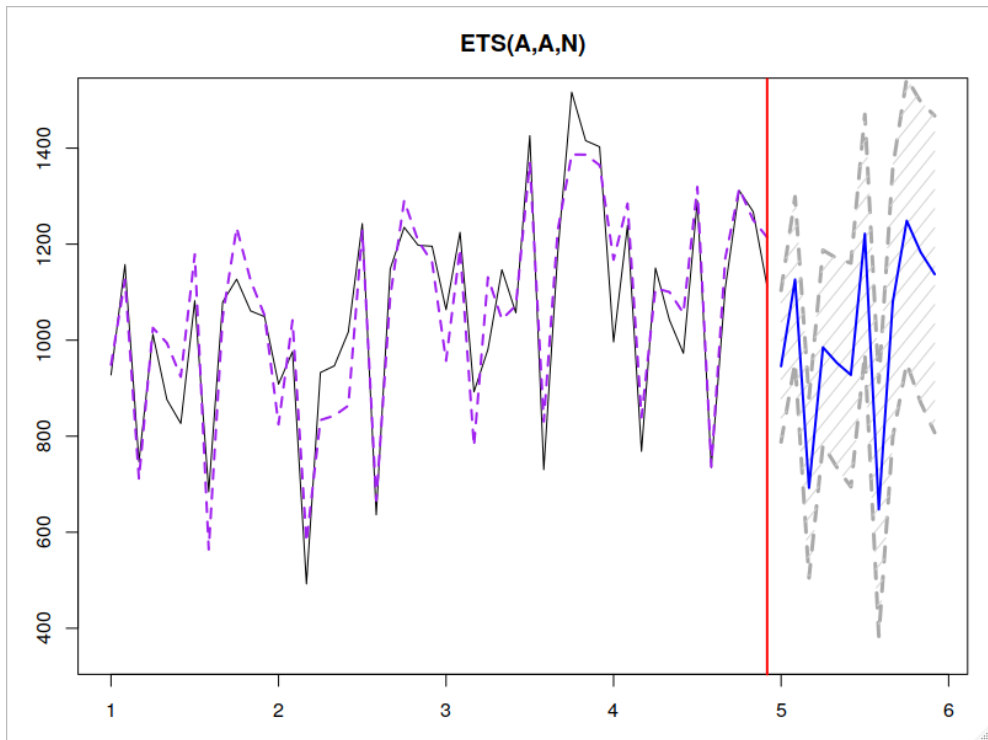
- Similar to ETS(A,A,A), but with an additional parameter.

# Trend seasonal model

- An example for the two models:

$$\alpha = 0.421, \beta = 0.022, \gamma = 0.003$$

$$\alpha = 0.405, \beta = 0.002, \gamma = 0.000, \phi = 0.989$$



# Trend seasonal model

- The pure multiplicative model, ETS(M,M,M):

$$y_t = l_{t-1} b_{t-1} s_{t-m} (1 + \epsilon_t)$$

$$l_t = l_{t-1} b_{t-1} (1 + \alpha \epsilon_t)$$

$$b_t = b_{t-1} (1 + \beta \epsilon_t)$$

$$s_t = s_{t-m} (1 + \gamma \epsilon_t)$$

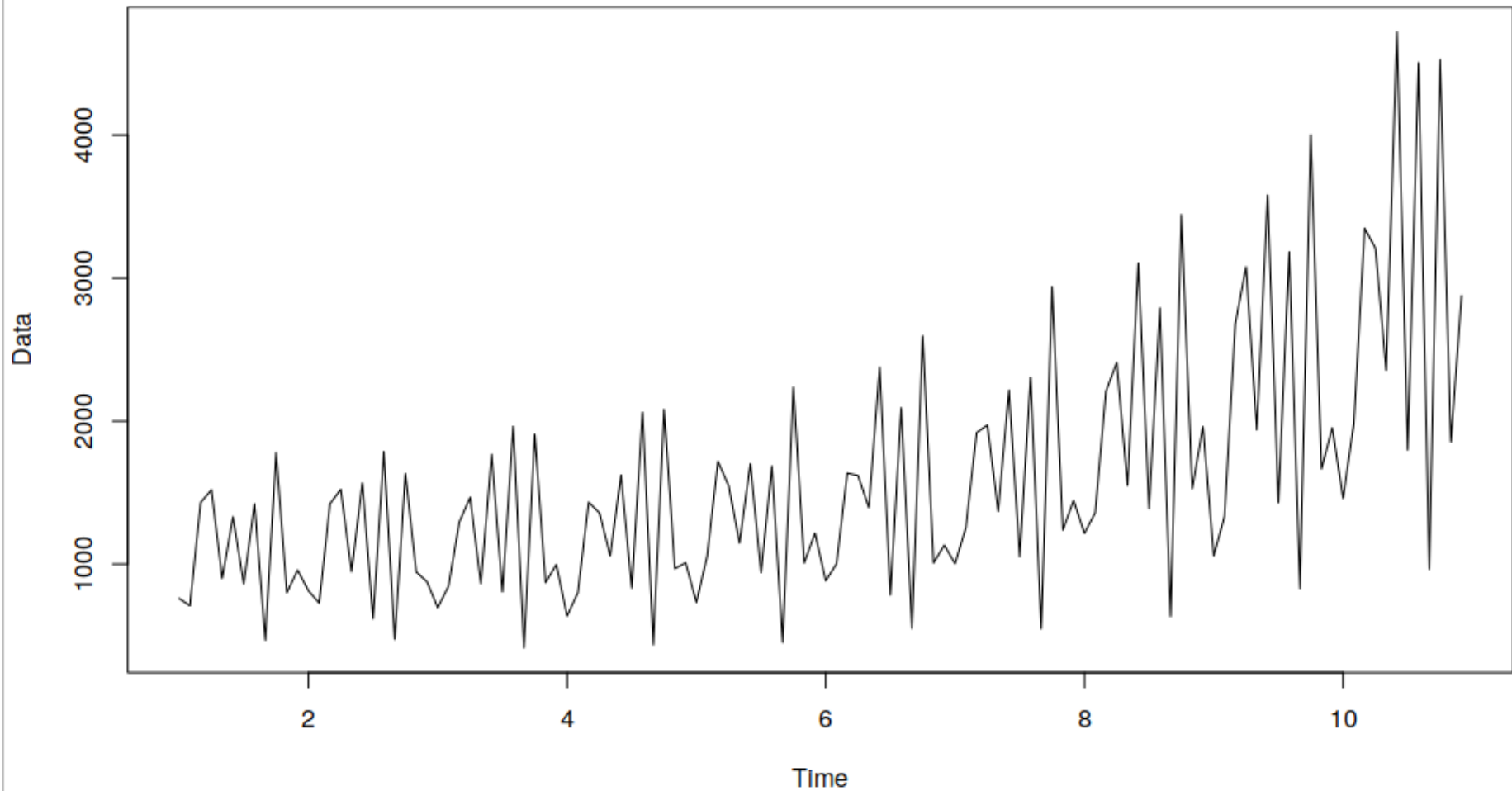
$$1 + \epsilon_t \sim \log N(0, \sigma^2)$$

- The amplitude of seasonality will increase with the increase of the level...
- The forecast:

$$\hat{y}_{t+h} = \hat{l}_t \hat{b}_t^h \hat{s}_{t-m+h}$$

# Trend seasonal model

ETS(MMM)



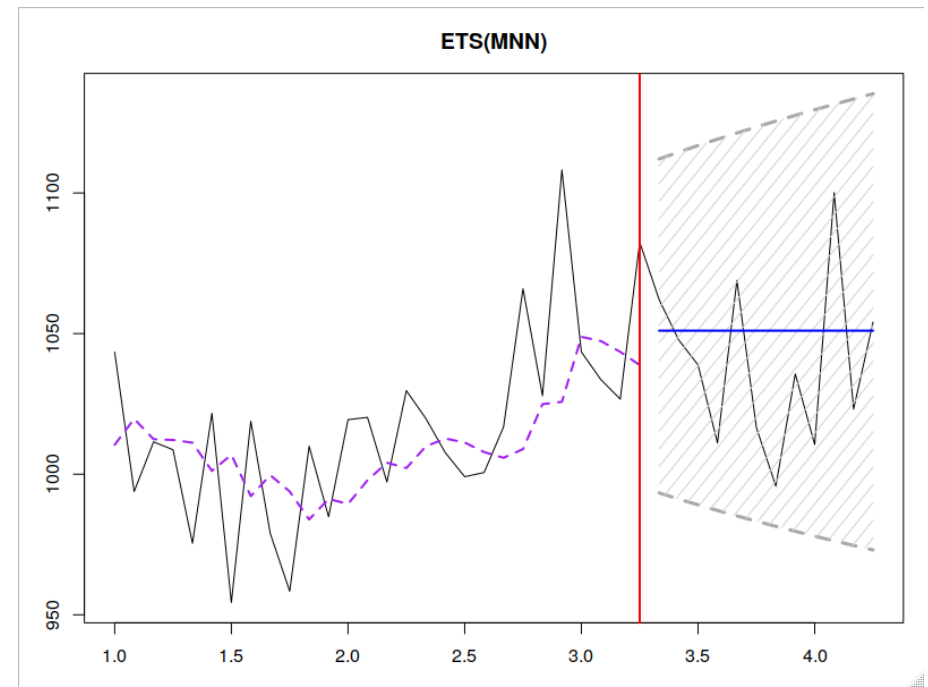
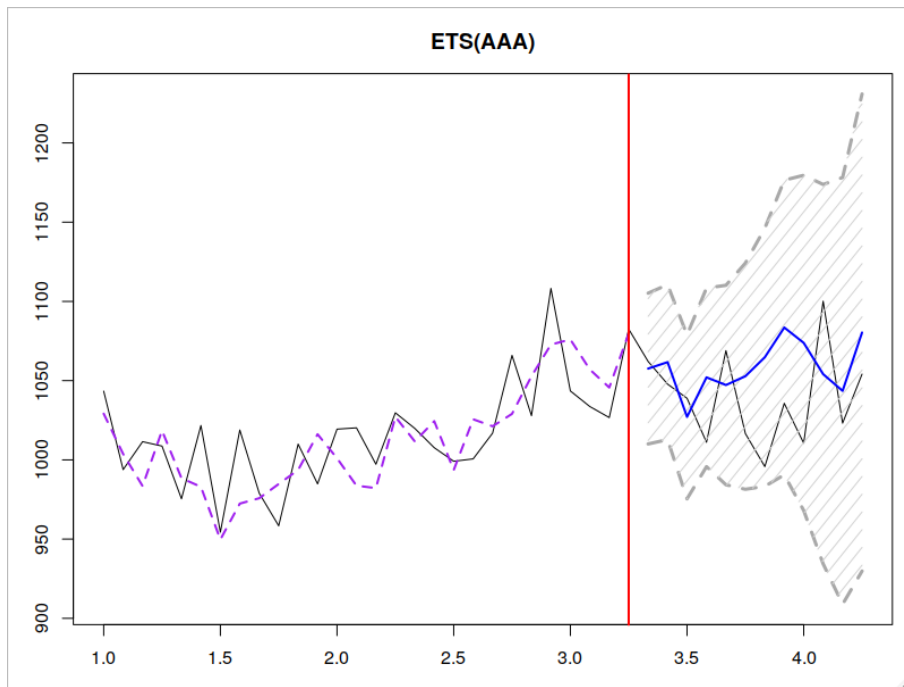
# Trend seasonal model

- There are many other options for the trend seasonal model;
- The number of parameters to estimate in all of them is high;
- They might fit very well to many different series...
- ...sometimes overfitting the data



# Trend seasonal model

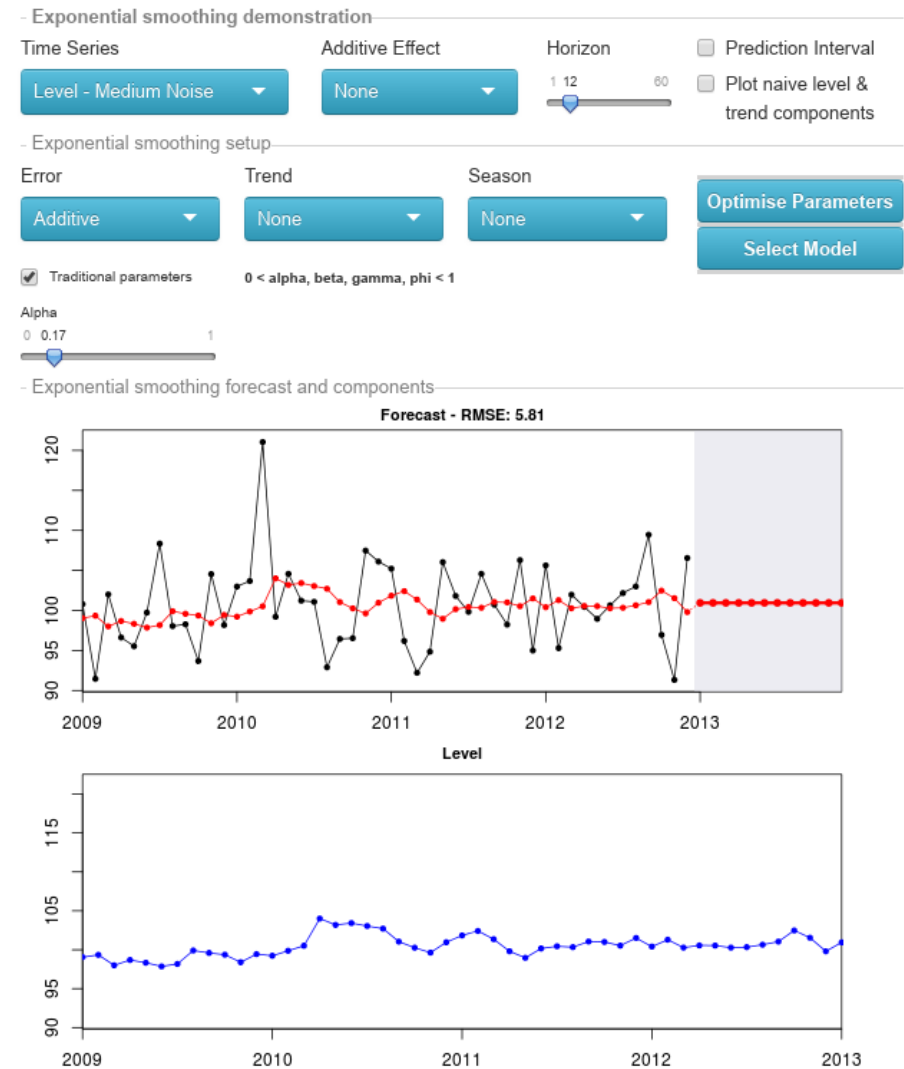
- An example with the local level data.
- Which of these makes more sense?



- The left one overfits the data.

# Trend seasonal model

- An exercise in R:
  - ETS(A,A,A), ETS(A,Ad,A), ETS(M,M,M),
  - ETS(X,X,X), ETS(Y,Y,Y), ETS(Z,Z,Z).
- Additional exercise – working with ETS:
  - <https://kourentzes.com/forecasting/2014/10/30/exponential-smoothing-demo/>



Exponential smoothing demo using the [forecast](#) package for R. Nikolaos Kourentzes, 2014

# Outline

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.

# Estimation of ETS

- Remember the pure additive ETS model?

$$y_t = \mathbf{w}'\mathbf{v}_{t-1} + \epsilon_t$$
$$\mathbf{v}_t = \mathbf{F}\mathbf{v}_{t-1} + \mathbf{g}\epsilon_t$$

- How can we estimate it?
- We use the assumption that  $\epsilon_t \sim N(0, \sigma^2)$
- Based on this assumption we can derive a likelihood function, using pdf of normal distribution:

$$f(y_t|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}}$$

# Estimation of ETS

$$f(y_t|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}}$$

- We can use that in order to calculate for each observation:

$$L(\boldsymbol{\theta}|y_t) = f(y_t|\boldsymbol{\theta})$$

- We are measuring the chance that each observation is from normal distribution.
- Then we use a summary value for the whole sample based on that:

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{t=1}^T f(y_t|\boldsymbol{\theta})$$

# Estimation of ETS

Usually the logarithm is used in order to linearise likelihood.

- Log-likelihood:

$$\ell(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{t=1}^T \log(f(y_t|\boldsymbol{\theta}))$$

- For the normal distribution we will have:

$$\ell(\boldsymbol{\theta}|\mathbf{Y}) = -\frac{T}{2} \log(2\pi\sigma^2) - \sum_{t=1}^T \frac{\epsilon_t^2}{2\sigma^2}$$

# Estimation of ETS

It can be shown that the maximum of that likelihood is achieved, when the variance of the error is estimated as:

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T e_t^2$$

What is this?

Inserting this in the log-likelihood we obtain so called “concentrated” log-likelihood:

$$\ell(\boldsymbol{\theta}, \hat{\sigma}^2 | \mathbf{Y}) = -\frac{T}{2} (\log(2\pi e) + \log(\hat{\sigma}^2))$$

# Estimation of ETS

$$\ell(\boldsymbol{\theta}, \hat{\sigma}^2 | \mathbf{Y}) = -\frac{n}{2} (\log(2\pi e) + \log(\hat{\sigma}^2))$$

- Maximum of this likelihood is achieved, when  $\log(\hat{\sigma}^2)$  is minimised.
- In turn  $\log(\hat{\sigma}^2)$  is minimised, when  $\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T e_t^2$  is minimised.
- So the likelihood of normal distribution is maximised, when MSE is minimised.
- OLS estimates correspond to Maximum Likelihood Estimates (MLE) when we use normal distribution assumption.



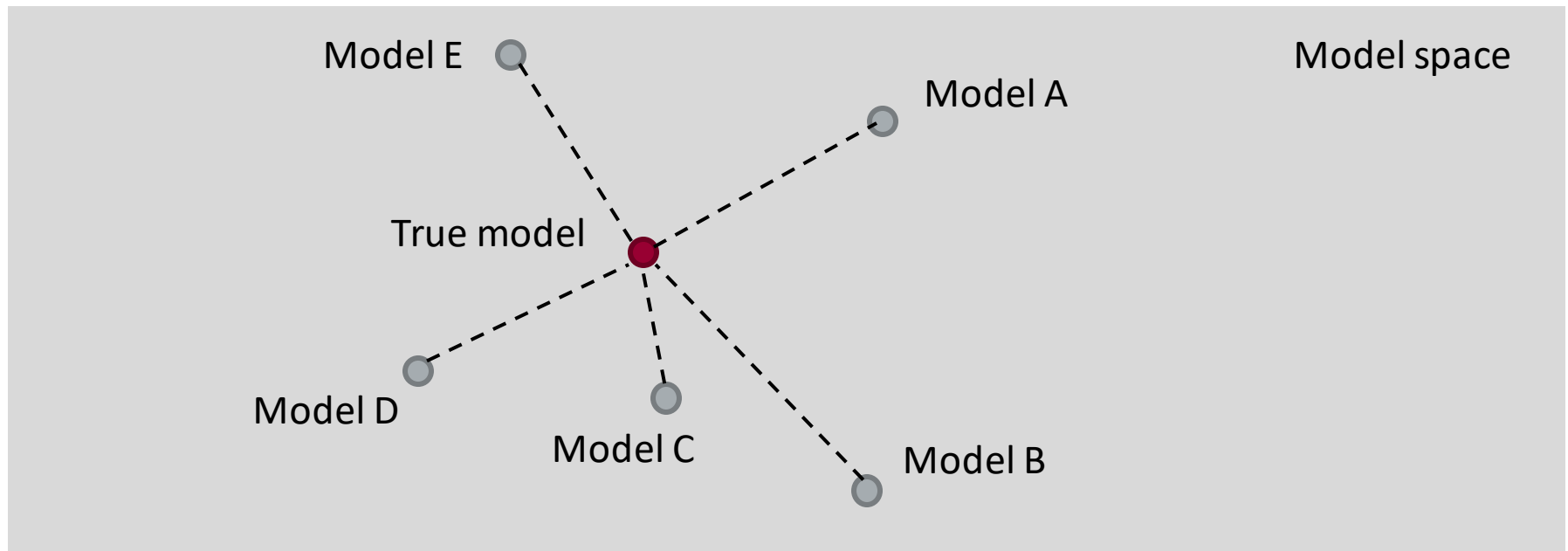
# Estimation of ETS

- Why is MLE useful?
- Likelihood has good statistical properties:
  - MLE of parameters are consistent and efficient.
- Likelihood can be used in calculation of information criteria. Thus, model selection is possible.
- What about multiplicative models?
  - The approach is similar, but the likelihood function is different.

# Information criteria

Can we measure distance between the true model and our model?

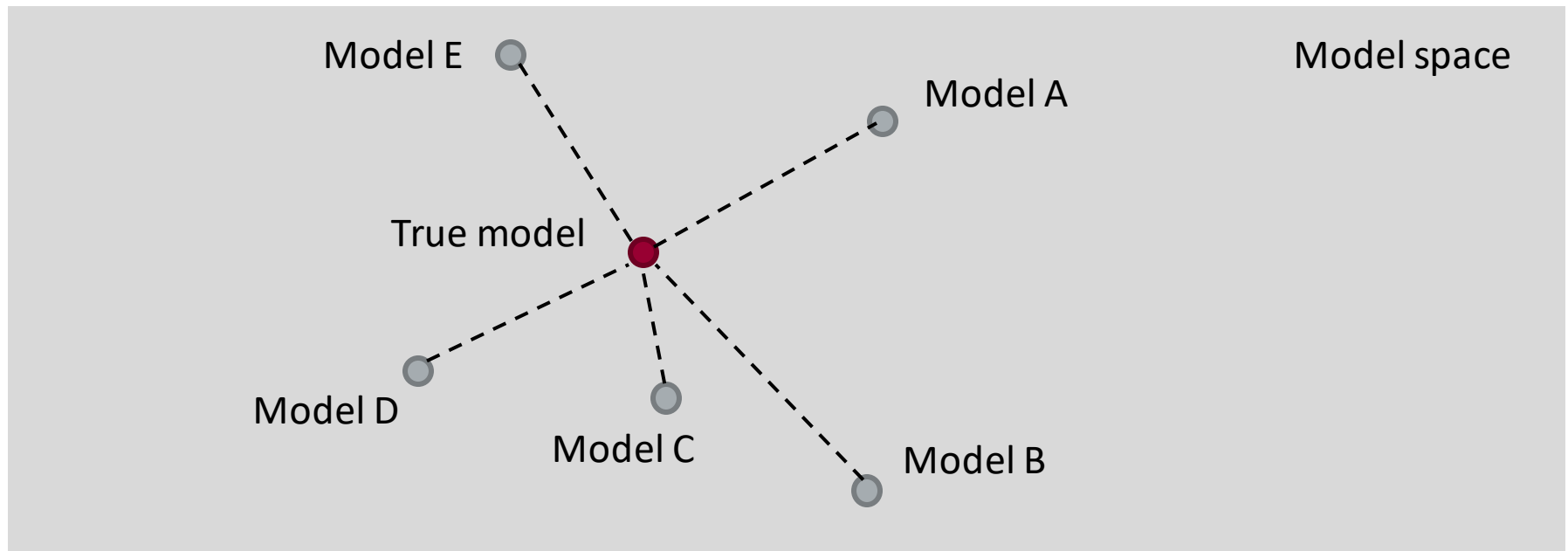
- Yes, if we know the truth:



# Information criteria

What makes the model closer to the true one?

- The ETS components,
- The transformation of the variable,
- The estimates of parameters.



# Information criteria

We can measure distances using likelihood function!

- Distance measure:

$$d = \ell(\text{true model}) - \ell(\text{our model})$$

- We can order models by this distance:

1.  $d_C,$

2.  $d_D,$

3.  $d_E,$

4. ...

# Information criteria

- If the true model is fixed, then we don't care about its likelihood and can compare distances differently.

Instead of:

$$d = \ell(\text{true model}) - \ell(\text{our model})$$

- We use:

$$d = -\ell(\text{model})$$

- This preserves the order of models:
  - The closer the model is to the true one, the lower  $d$  is.

# Model selection in ETS

- But we need to take into account that we estimated the models in sample and correct the bias of likelihood:

$$AIC = 2k - 2\ell(\text{model})$$

- where  $\ell$  is the log-likelihood value and  $k$  is the number of **all** the estimated parameters
- There are other ICs:
  - AICc – corrected for the sample size AIC;
  - BIC – Bayesian IC (aka Schwartz IC);
  - ...

Assumes normal distribution,  
Used by default  
in R functions.

# Model selection in ETS

So, in the ETS framework, we can:

1. fit all the possible models,
  2. calculate their likelihoods,
  3. calculate the number of parameters (including  $\hat{\sigma}^2$ ),
  4. calculate values of the selected IC,
  5. select the model that has the lowest IC.
- This is what all the ETS functions in R do by default.

# Summary

1. Forecasting level series;
2. Simple Exponential Smoothing;
3. Introduction to ETS;
4. Local level model;
5. Local trend model;
6. Trend seasonal model;
7. Model estimation and selection.



# Summary

- Packages and functions in R:
  - forecast package:
    - ets() – basic ETS with 19 models;
    - bats(), tbats() – models for multiple frequencies.
  - fable package:
    - ETS() – similar to ets() from forecast.
      - 19 models, only additive trend;
  - smooth package:
    - es() – more flexible ETS:
      - 30 models,
      - different loss functions,
      - allows including explanatory variables.

# ETS models... final word

- There are different types of bounds for smoothing parameters:
  - Usual:  $\alpha \in [0,1], \beta \in [0, \alpha], \gamma \in [0, 1 - \alpha]$
  - Admissible...
- Admissible bounds make sure that the model puts more weight on the recent observations than on the older ones.
  - E.g. for ETS(A,N,N)  $\alpha \in [0,2]$
  - What is the meaning of  $\alpha \in (1,2]$  and what happens with the model?

# Lab Session 7

Use ETS model to produce forecast for 42 days:

- Split daily time series into training and test set
- Specify and train training set using the following models:
  - single exponential smoothing
  - holt-winter
  - automatic ETS()
- use glance, tidy and report functions to extract information from trained models
- report forecast accuracy
- which model is more accurate?

# Thank you!

## Thank you for your attention!

### Questions?

Ivan Svetunkov

[i.svetunkov@lancaster.ac.uk](mailto:i.svetunkov@lancaster.ac.uk)



[@iSvetunkov](https://twitter.com/iSvetunkov)

<https://forecasting.svetunkov.ru>

Full or partial reproduction of the slides is not permitted without author's consent. Please contact [i.svetunkov@lancaster.ac.uk](mailto:i.svetunkov@lancaster.ac.uk) for more information.

Marketing Analytics  
& Forecasting



Lancaster University  
Management School