Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

September 26, 2017

Outline of the lecture

- Regression based methods, 3rd part:
  - Regression and exponential smoothing (Sec. 3.4)
  - Global and local trend models - an example (Sec. 3.6)
- Operators; the backward shift operator; sec. 4.5.

Predictions In Time Series

- Are model-based - one for all data (so far).
- What if no model fits our need (and data)?
- Methods that aren't model based in the usual sense applies.

Exponential smoothing

Given a forgetting factor $\lambda \in]0;1[$

$$\hat{\mu}_N = c \sum_{j=0}^{N-1} \lambda^j \, Y_{N-j} = c[\, Y_N + \lambda \, Y_{N-1} + \cdots + \lambda^{N-1} \, Y_1]$$

The constant c is chosen so that the weights sum to one, which implies that $c = (1-\lambda)/(1-\lambda^N)$.
When N is large $c \approx 1-\lambda$:

$$\begin{aligned}
\hat{\mu}_N &= (1-\lambda)[\, Y_N + \lambda \, Y_{N-1} + \cdots + \lambda^{N-1} \, Y_1] \\
&= (1-\lambda) \, Y_N + (1-\lambda)[\lambda \, Y_{N-1} + \cdots + \lambda^{N-1} \, Y_1] \\
&= (1-\lambda) \, Y_N + \lambda(1-\lambda)[\, Y_{N-1} + \cdots + \lambda^{N-2} \, Y_1] \\
&= (1-\lambda) \, Y_N + \lambda \hat{\mu}_{N-1}
\end{aligned}$$

Exponential Smoothing and prediction

Used as a prediction model:

$$\widehat{Y}_{N+\ell|N} = \hat{\mu}_N$$

Updating predictions with new observations:

$$\widehat{Y}_{N+\ell+1|N+1} = (1-\lambda)\, Y_{N+1} + \lambda\, \widehat{Y}_{N+\ell|N}$$

Simple Exponential Smoothing

For large N:

$$\hat{\mu}_{N+1} = (1 - \lambda)\, Y_{N+1} + \lambda \hat{\mu}_N$$

Definition (Simple Exponential Smoothing):

The sequence S_N defined by

$$S_N = (1 - \lambda)\, Y_N + \lambda S_{N-1}$$

is called the *simple exponential smoothing* or first order exponential smoothing of the time series Y.

- The smoothing constant $\alpha = 1 - \lambda$ (or the forgetting factor λ) determines how much the latest observation influence the prediction.

Choice of smoothing constant $\alpha = 1 - \lambda$

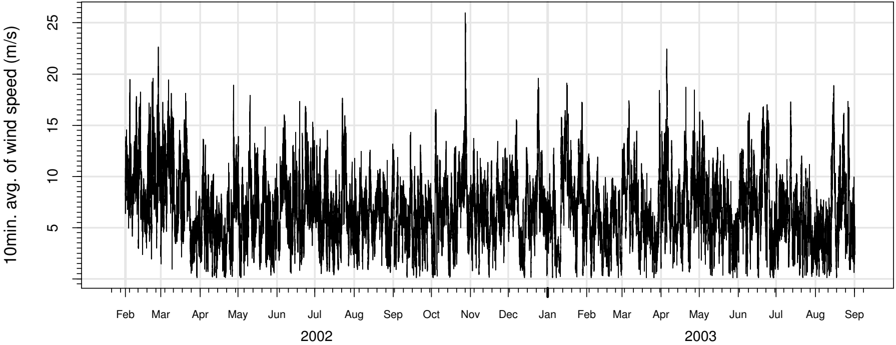
- Given a data set $t = 1, \ldots, N$ we construct

$$S(\alpha) = \sum_{t=1}^{N} (Y_t - \widehat{Y}_{t|t-l}(\alpha))^2 = \sum_{t=1}^{N} (Y_t - \hat{\mu}_{t-l}(\alpha))^2$$

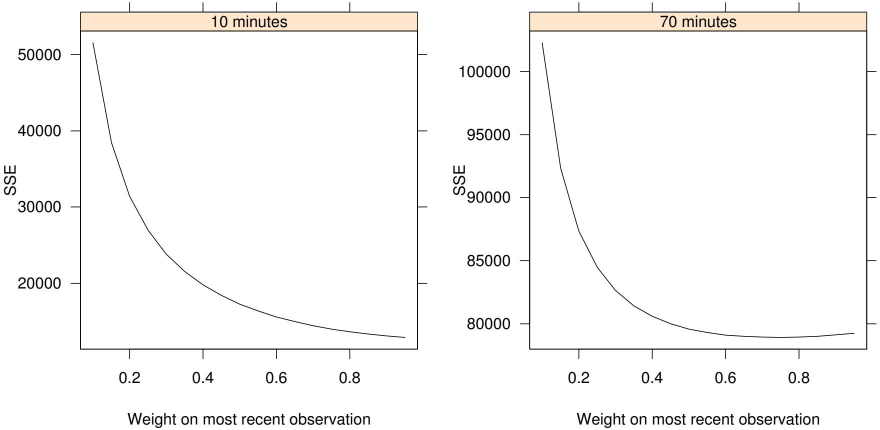
  The value minimizing $S(\alpha)$ is chosen.
- If the data set is large we eliminate the influence of the initial estimate by dropping the first part of the errors when evaluating $S(\alpha)$
- Keep in mind however what the smoothing is used for, and modify the criteria accordingly. The next slides show an example of this.

Example – wind speed 76 m a.g.l. at DTU Risø

- Measurements of wind speed every 10th minute
- Task: Forecast up to approximately 3 hours ahead using exponential smoothing

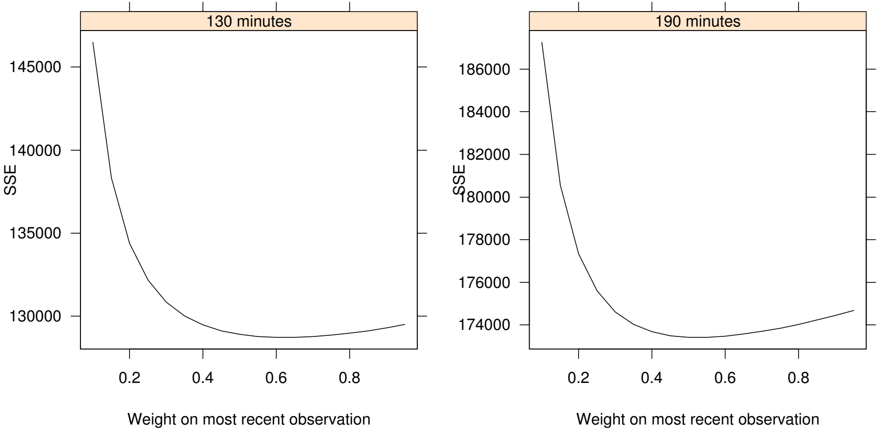


$S(\alpha)$ for horizons 10 and 70 minutes



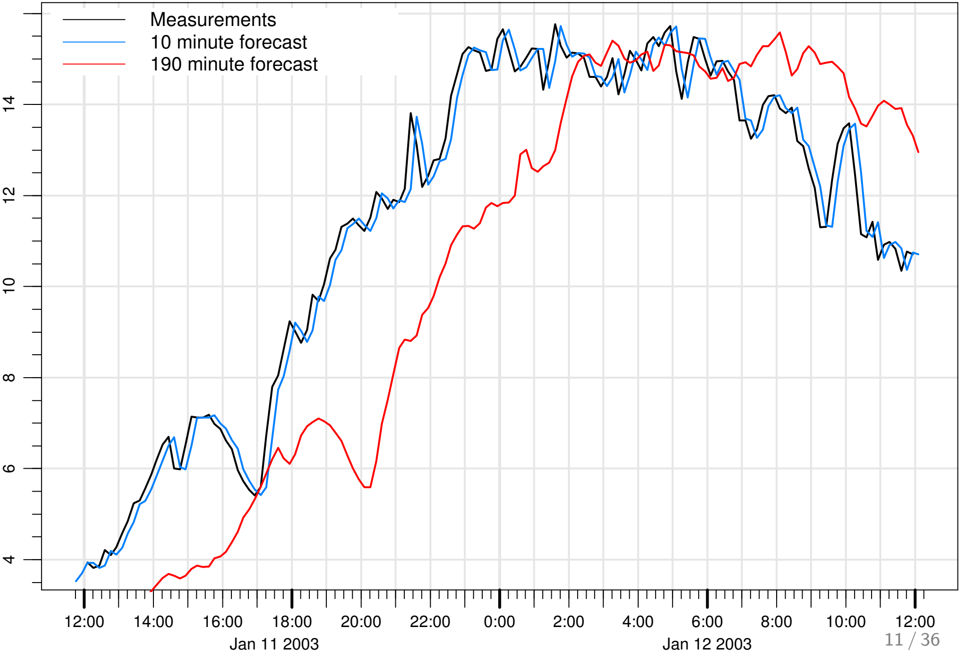
- 10 minutes (1-step): Use $\alpha = 0.95$ or higher
- 70 minutes (7-step): Use $\alpha \approx 0.7$

$S(\alpha)$ for horizons 130 and 190 minutes



- 130 minutes (13-step): Use $\alpha \approx 0.6$
- 190 minutes (19-step): Use $\alpha \approx 0.5$

Example of forecasts with optimal α



From global to local trend models

Last week we worked with the global trend model

$$Y_{N+j} = \boldsymbol{f}^T(j)\boldsymbol{\theta} + \varepsilon_{N+j}$$

which was solved iteratively by

$$\begin{aligned}
\boldsymbol{F}_{N+1} &= \boldsymbol{F}_N + \boldsymbol{f}(-N)\boldsymbol{f}^T(-N) \\
\boldsymbol{h}_{N+1} &= \boldsymbol{L}^{-1}\boldsymbol{h}_N + \boldsymbol{f}(0)\,Y_{N+1} \\
\widehat{\boldsymbol{\theta}}_{N+1} &= \boldsymbol{F}_{N+1}^{-1}\boldsymbol{h}_{N+1}
\end{aligned}$$

Could we do that locally?

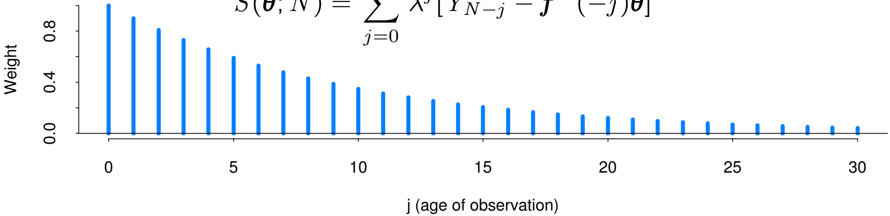
Local trend models

We forget old observations in an exponential manner:

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}; N)$$

where for $0 < \lambda < 1$

$$S(\boldsymbol{\theta}; N) = \sum_{j=0}^{N-1} \lambda^j [Y_{N-j} - \boldsymbol{f}^T(-j)\boldsymbol{\theta}]^2$$



WLS formulation

The criterion:

$$S(\boldsymbol{\theta}; N) = \sum_{j=0}^{N-1} \lambda^j [Y_{N-j} - \boldsymbol{f}^T(-j)\boldsymbol{\theta}]^2$$

can be written as:

$$\begin{bmatrix} Y_1 - \boldsymbol{f}^T(N-1)\boldsymbol{\theta} \\ Y_2 - \boldsymbol{f}^T(N-2)\boldsymbol{\theta} \\ \vdots \\ Y_N - \boldsymbol{f}^T(0)\boldsymbol{\theta} \end{bmatrix}^T \begin{bmatrix} \lambda^{N-1} & 0 & \cdots & 0 \\ 0 & \lambda^{N-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Y_1 - \boldsymbol{f}^T(N-1)\boldsymbol{\theta} \\ Y_2 - \boldsymbol{f}^T(N-2)\boldsymbol{\theta} \\ \vdots \\ Y_N - \boldsymbol{f}^T(0)\boldsymbol{\theta} \end{bmatrix}$$

which is a WLS criterion with $\boldsymbol{\Sigma} = \mathrm{diag}[1/\lambda^{N-1}, \ldots, 1/\lambda, 1]$

WLS solution

$$\widehat{\boldsymbol{\theta}}_N = (\boldsymbol{x}_N^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_N)^{-1} \boldsymbol{x}_N^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}$$

or

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_N &= \boldsymbol{F}_N^{-1} \boldsymbol{h}_N \\
\boldsymbol{F}_N &= \sum_{j=0}^{N-1} \lambda^j \boldsymbol{f}(-j) \boldsymbol{f}^T(-j) \\
\boldsymbol{h}_N &= \sum_{j=0}^{N-1} \lambda^j \boldsymbol{f}(-j)\, Y_{N-j}
\end{aligned}$$

Updating the estimates when Y_{N+1} is available

$$
\begin{aligned}
\boldsymbol{F}_{N+1} &= \boldsymbol{F}_N + \lambda^N \boldsymbol{f}(-N)\boldsymbol{f}^T(-N) \\
\boldsymbol{h}_{N+1} &= \lambda \boldsymbol{L}^{-1} \boldsymbol{h}_N + \boldsymbol{f}(0)\, Y_{N+1} \\
\widehat{\boldsymbol{\theta}}_{N+1} &= \boldsymbol{F}_{N+1}^{-1} \boldsymbol{h}_{N+1}
\end{aligned}
$$

As initial values we can use $\boldsymbol{h}_0 = \mathbf{0}$ and $\boldsymbol{F}_0 = \mathbf{0}$

For many functions $\lambda^N \boldsymbol{f}(-N)\boldsymbol{f}^T(-N) \to 0$ for $N \to \infty$ and we get the stationary result $\boldsymbol{F}_{N+1} = \boldsymbol{F}_N = \boldsymbol{F}$. Hence:

$$
\widehat{\boldsymbol{\theta}}_{N+1} = \boldsymbol{L}^T \widehat{\boldsymbol{\theta}}_N + \boldsymbol{F}^{-1} \boldsymbol{f}(0)[Y_{N+1} - \widehat{Y}_{N+1|N}]
$$

Variance estimation in local trend models

Define the *total memory*

$$T = \sum_{j=0}^{N-1} \lambda^j = \frac{1-\lambda^N}{1-\lambda}$$

T is a measure of how many observations estimation is essentially based upon.

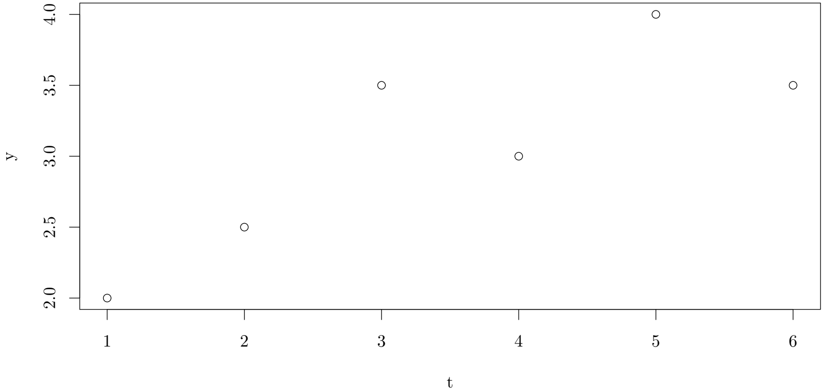
A variance estimator is therefore

$$\hat{\sigma}^2 = (\boldsymbol{Y} - \boldsymbol{x}_N \widehat{\boldsymbol{\theta}}_N)^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{x}_N \widehat{\boldsymbol{\theta}}_N)/(T-p), \quad T > p$$

Notice that the restriction on T is a restriction on λ. How do you interpret this?
(Note: This estimator is not in the book.)

Global and Local Trend Models - an Example

6 observations ($N = 6$):



Global Linear Trend:

$$Y_{N+j} = \theta_0 + \theta_1 j + \varepsilon_{N+j} \Rightarrow \boldsymbol{f}(j) = \begin{pmatrix} 1 & j \end{pmatrix}^T$$

Linear Model form:

$$\begin{pmatrix} 2.0 \\ 2.5 \\ 3.5 \\ 3.0 \\ 4.0 \\ 3.5 \end{pmatrix} = \begin{pmatrix} 1 & -5 \\ 1 & -4 \\ 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \Leftrightarrow \boldsymbol{y} = \boldsymbol{x}_6 \theta + \varepsilon$$

Global linear trend: Estimation

$$
\begin{aligned}
\boldsymbol{F}_6 &= \boldsymbol{x}_6^T \boldsymbol{x}_6 = & \begin{pmatrix} 6 & -15 \\ -15 & 55 \end{pmatrix} \\
\boldsymbol{h}_6 &= \boldsymbol{x}_6^T \boldsymbol{y} = & \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} \\
\widehat{\boldsymbol{\theta}}_6 &= \boldsymbol{F}_6^{-1} \boldsymbol{h}_6 = & \begin{pmatrix} 0.5238 & 0.1429 \\ 0.1429 & 0.0571 \end{pmatrix} \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} = \begin{pmatrix} 3.905 \\ 0.329 \end{pmatrix}
\end{aligned}
$$

Global linear trend: Prediction

Linear predictor:

$$\widehat{Y}_{6+\ell|6} = f(\ell)^T \widehat{\theta}_6 = 3.905 + 0.328\ell$$

LS-estimate for σ^2:

$$\widehat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{x}_6\widehat{\theta}_6)^T(\boldsymbol{y} - \boldsymbol{x}_6\widehat{\theta}_6)/(6-2) = 0.453^2$$

Prediction error:

$$\begin{aligned} \varepsilon_6(\ell) &= Y_{6+\ell} - \widehat{Y}_{6+\ell|6} \\ \widehat{\mathrm{Var}}(\varepsilon_6(\ell)) &= \widehat{\sigma}^2 \left(1 + \boldsymbol{f}^T(\ell)\boldsymbol{F}_6^{-1}\boldsymbol{f}(\ell)\right) \end{aligned}$$

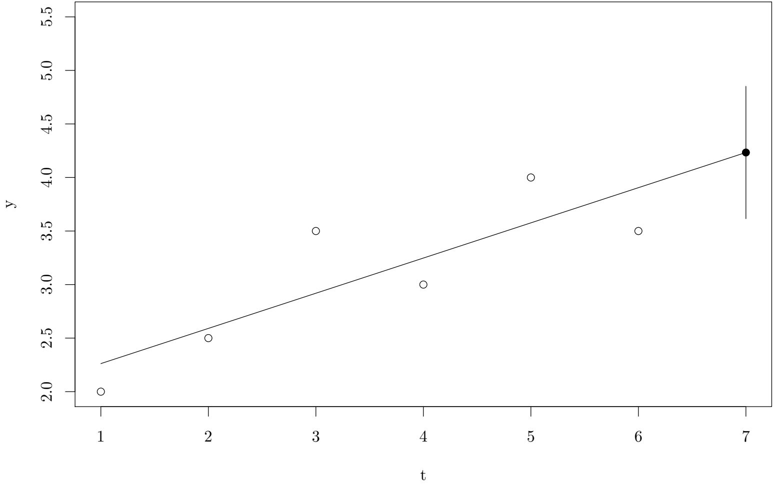
For example,

$$\widehat{Y}_{7|6} = 4.234 \text{ with } \widehat{\mathrm{Var}}(\varepsilon_6(1)) = 0.619^2.$$

90% prediction interval:

$$\widehat{Y}_{7|6} \pm t_{0.05}(6-2)\sqrt{\widehat{\mathrm{Var}}(\varepsilon_6(1))} = 4.234 \pm 1.320$$

Global linear trend: Estimation - global linear trend

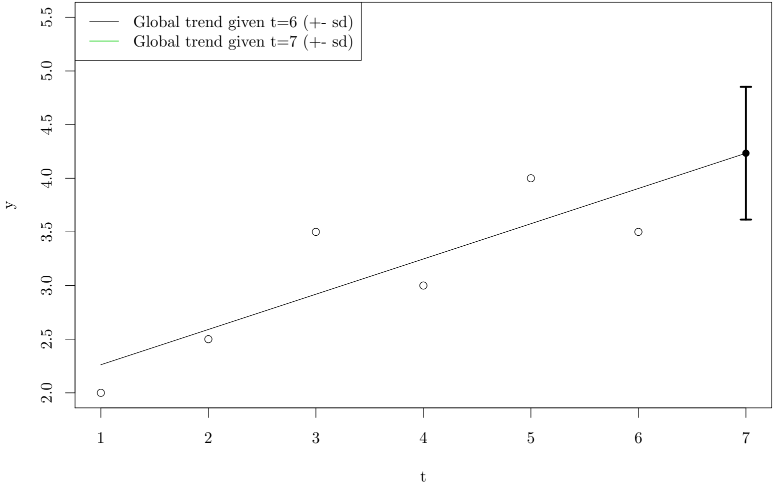


Global linear trend: Updating the parameters

- New observation: $y_7 = 3.5$.

$$
\begin{aligned}
\boldsymbol{F}_7 &= \boldsymbol{F}_6 + \boldsymbol{f}^T(-6)\boldsymbol{f}(-6) \\
&= \begin{pmatrix} 6 & -15 \\ -15 & 55 \end{pmatrix} + \begin{pmatrix} 1 & -6 \end{pmatrix} \begin{pmatrix} 1 \\ -6 \end{pmatrix} = \begin{pmatrix} 7 & -21 \\ -21 & 91 \end{pmatrix}, \\
\boldsymbol{h}_7 &= L^{-1}\boldsymbol{h}_6 + \boldsymbol{f}(0)\boldsymbol{y}_7 \\
&= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} 3.5 = \begin{pmatrix} 22 \\ -59 \end{pmatrix}, \\
\widehat{\boldsymbol{\theta}}_7 &= \begin{pmatrix} 0.4643 & 0.1071 \\ 0.1071 & 0.0357 \end{pmatrix} \begin{pmatrix} 22 \\ -59 \end{pmatrix} = \begin{pmatrix} 3.896 \\ 0.250 \end{pmatrix}.
\end{aligned}
$$

Global linear trend: Updating - global linear trend

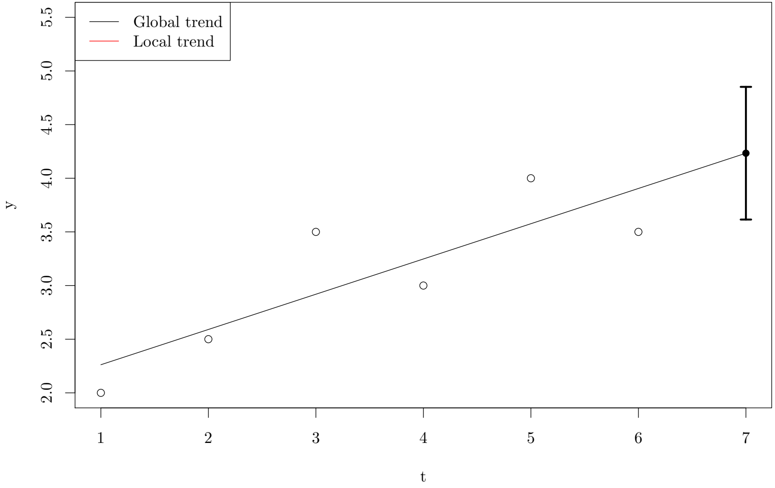


Local linear trend: Estimation

- Forgetting factor $\lambda = 0.9$. Linear model unchanged.

$$
\begin{aligned}
\boldsymbol{F}_6 &= \sum_{j=0}^{5} \lambda^j \boldsymbol{f}(-j)\boldsymbol{f}^T(-j) = \begin{pmatrix} 4.6856 & -10.284 \\ -10.284 & 35.961 \end{pmatrix} \\
\boldsymbol{h}_6 &= \sum_{j=0}^{5} \lambda^j \boldsymbol{f}(-j)\, Y_{6-j} = \begin{pmatrix} 14.902 \\ -28.580 \end{pmatrix} \\
\widehat{\boldsymbol{\theta}}_6 &= \boldsymbol{F}_6^{-1}\boldsymbol{h}_6 = \begin{pmatrix} 0.573 & 0.164 \\ 0.164 & 0.075 \end{pmatrix} \begin{pmatrix} 14.902 \\ -28.580 \end{pmatrix} = \begin{pmatrix} 3.85 \\ 0.308 \end{pmatrix} \\
\hat{\sigma}^2 &= (\boldsymbol{Y} - \boldsymbol{x}_6\widehat{\boldsymbol{\theta}}_6)^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{x}_6\widehat{\boldsymbol{\theta}}_6)/(T-2) = 0.496^2 \\
\widehat{\mathrm{Var}}(\varepsilon_6(1)) &= \widehat{\sigma}^2 \left(1 + \boldsymbol{f}^T(1)\boldsymbol{F}_6^{-1}\boldsymbol{f}(1)\right) = 0.697^2
\end{aligned}
$$

Local linear trend: Predicting for $t = 7$



Local linear trend: Estimating $\hat{\sigma}^2$

- We can use the WLS estimator for $\hat{\theta}_N$
- but not for $\hat{\sigma}^2$!?!
- Reason: Local trend models assume that ϵ_t are i.i.d.
- The proposed estimator:

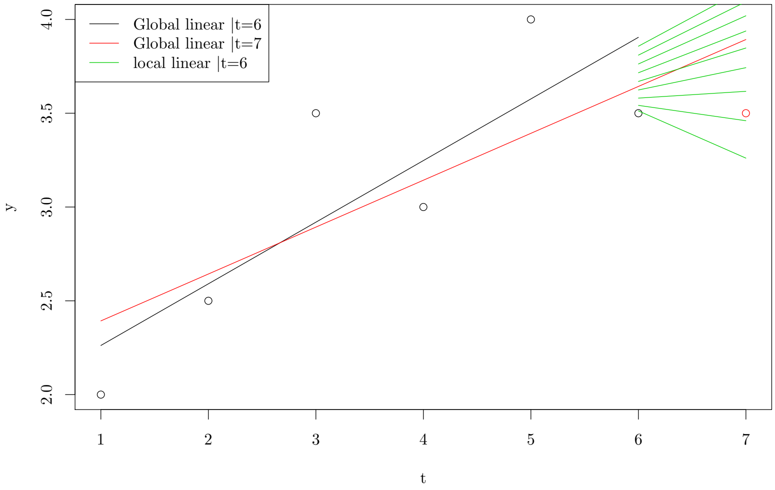
$$\hat{\sigma_N}^2 = (\boldsymbol{Y} - \boldsymbol{x}_N \widehat{\boldsymbol{\theta}}_N)^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{x}_N \widehat{\boldsymbol{\theta}}_N)/(T-p), \quad T > p$$

  provides a local estimate.
- A global estimator is given as:

$$\hat{\sigma_N}^2 = \frac{1}{N-n} \sum_{j=n+1}^{N} \frac{(Y_j - \hat{Y}_{j|j-1})^2}{1 + \boldsymbol{f}^T(1) \boldsymbol{F}_{j-1}^{-1} \boldsymbol{f}(1)}$$

- Note that the first n predictions are ignored to stabilize the estimator
- Note that the prediction errors are normed.

Local linear trend: Estimation



Which of the (green) local trend models have the highest λ?

Operators; The backwards shift operator B

- An operator A is (here) a function of a time series $\{x_t\}$ (or a stochastic proces $\{X_t\}$).
- Application of an operator on a time series $\{x_t\}$ yields a new time series $\{Ax_t\}$. Likewise of a stochastic process $\{AX_t\}$.
- Most important operator for us: The backwards shift operator B : $Bx_t = x_{t-1}$. Obviously, $B^j x_t = x_{t-j}$.
- All other operators we shall consider in this lecture may be expressed in terms of B.

The forward shift F and difference ∇

The forward shift operator

- $Fx_t = x_{t+1}$; $F^j x_t = x_{t+j}$;
- Obviously, combining a forward and backward shift yields the identity operator 1, ie. F and B are each others inverse: $B^{-1} = F$ and $F^{-1} = B$.

The difference operator

- $\nabla x_t = x_t - x_{t-1} = \mathbf{1}x_t - Bx_t = (1 - B)x_t$.
- Thus: $\nabla = 1 - B$.

The summation S

$$\begin{aligned} Sx_t &= x_t + x_{t-1} + x_{t-2} + \dots \\ &= x_t + Bx_t + B^2 x_t \dots \\ &= (1 + B + B^2 + \dots) x_t \end{aligned}$$

- Summation, then difference (remember $Sx_t = x_t + Sx_{t-1}$)
$$\nabla Sx_t = Sx_t - Sx_{t-1} = x_t + Sx_{t-1} - Sx_{t-1} = x_t$$

- Difference, then summation
$$\begin{aligned} S\nabla x_t &= (1 + B + B^2 \dots) x_t - (1 + B + B^2 \dots) x_{t-1} \\ &= (1 + B + B^2 \dots) x_t - (B + B^2 \dots) x_t = x_t \end{aligned}$$

- So ∇ and S are each others inverse:
$$\nabla^{-1} = \frac{1}{1 - B} = 1 + B + B^2 + \dots = S$$

Properties of B, F, ∇ and S

- The operators are all linear, ie.

$$H[\lambda x_t + (1-\lambda)y_t] = \lambda H[x_t] + (1-\lambda)H[y_t]$$

- The operators may be combined into new operators:
  The power series

$$a(z) = \sum_{i=0}^{\infty} a_i z^i$$

  defines a new operator from an operator H by linear combinations:

$$a(H) = \sum_{i=0}^{\infty} a_i H^i$$

Examples of combined operators

- ∇^{-1}:

$$\frac{1}{1-z} = \sum_{i=0}^{\infty} z^i \quad \text{so} \quad \nabla^{-1} = \frac{1}{1-B} = \sum_{i=0}^{\infty} B^i = S$$

- Operator polynomial of order q:

$$\theta(z) = \sum_{i=0}^{q} \theta_i z^i$$

ie. $\theta_i = 0$ for $i > q$.

$$\theta(B) = (1 + \theta_1 B + \cdots + \theta_q B^q)$$

where θ_0 is chosen to be 1

The Cauchy product (discrete convolution)

The equation

$$\{\lambda_i\} * \{\psi_i\} = \{\pi_i\}$$

means that

$$\begin{aligned}
\pi_0 &= \lambda_0\psi_0 \\
\pi_1 &= \lambda_1\psi_0 + \lambda_0\psi_1 \\
&\vdots \\
\pi_i &= \lambda_i\psi_0 + \lambda_{i-1}\psi_1 + \ldots + \lambda_0\psi_i \\
&\vdots
\end{aligned}$$

Multiplying combined operators

Theorem 4.13

- For the operator H the following operators are given:

$$\lambda(H) = \sum_{i=0}^{\infty} \lambda_i H^i, \quad \psi(H) = \sum_{i=0}^{\infty} \psi_i H^i, \quad \pi(H) = \sum_{i=0}^{\infty} \pi_i H^i$$

  such that $\lambda(H)\psi(H) = \pi(H)$.

- Then λ, ψ, π satisfies the equation

$$\{\lambda_i\} * \{\psi_i\} = \{\pi_i\}.$$

Highlights

- Local trend model: $S(\boldsymbol{\theta}; N) = \sum_{j=0}^{N-1} \lambda^j [Y_{N-j} - \boldsymbol{f}^T(-j)\boldsymbol{\theta}]^2$
- Iterative updates

$$\begin{aligned} \boldsymbol{F}_N &= \sum_{j=0}^{N-1} \lambda^j \boldsymbol{f}(-j)\boldsymbol{f}^T(-j) \\ \boldsymbol{h}_N &= \sum_{j=0}^{N-1} \lambda^j \boldsymbol{f}(-j) Y_{N-j} \\ \widehat{\boldsymbol{\theta}}_N &= \boldsymbol{F}_N^{-1} \boldsymbol{h}_N \end{aligned}$$

- Backwards shift operator: $BX_t = X_{t-1}$
- Difference operator: $\nabla X_t = X_t - X_{t-1}$