Regressions- och tidsserieanalys

Föreläsning 8 - Tidsserieanalys. Komponenter. Säsongsrensning med glidande medelvärden

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet

 mattiasvillani.com  @matvil  mattiasvillani

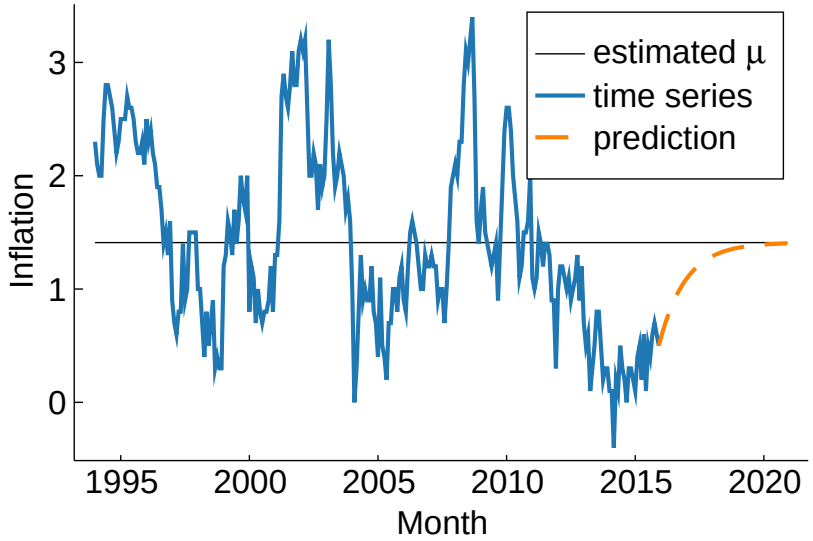
Översikt

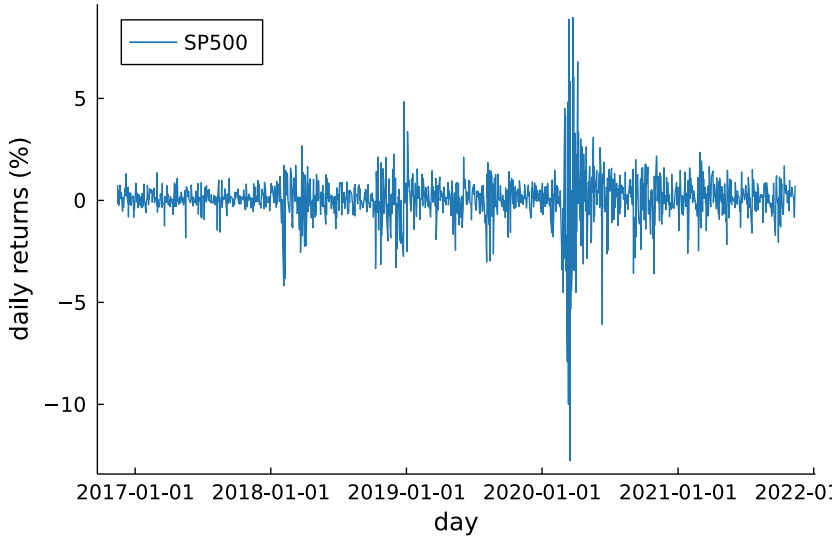
- Tidsserier
- Trendskattning - parametriska modeller
- Trendskattning - glidande medelvärden
- Säsongsrensning med glidande medelvärden
- Komponentsuppdelning av tidsserie.

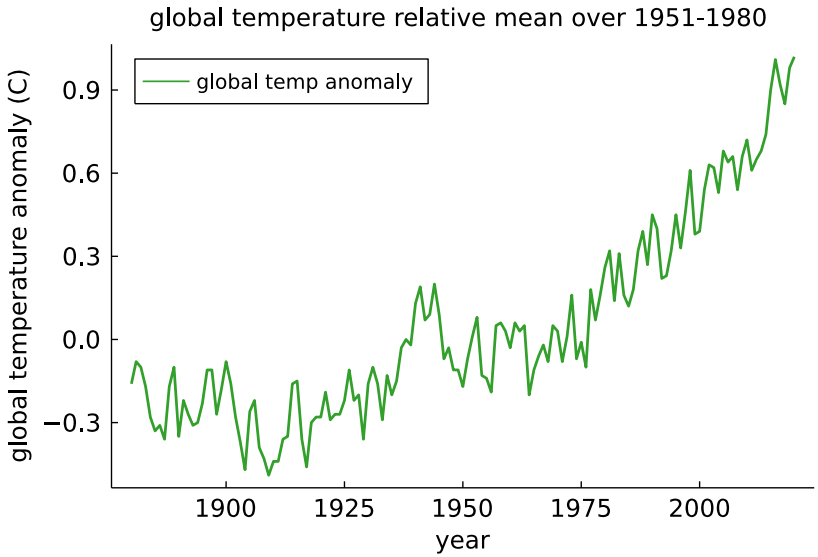
Tidsserier

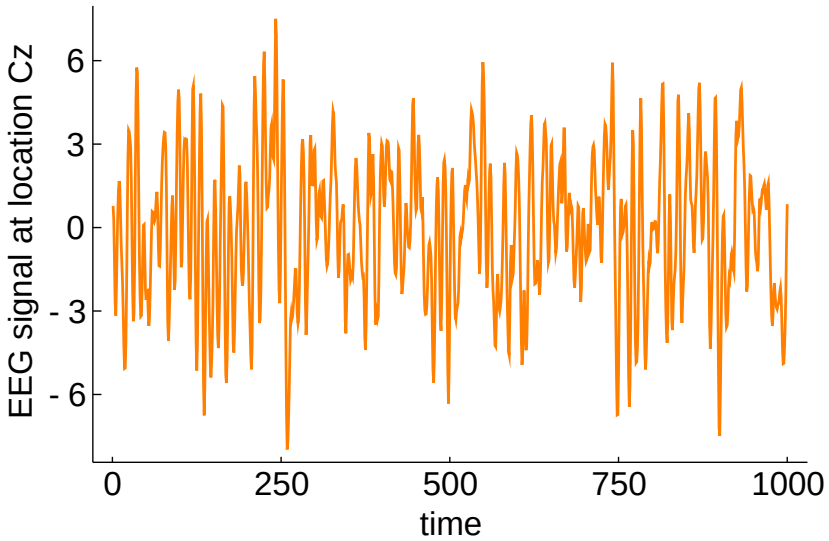
- **Tvärsnittsdata** data uppmätta vid en tidpunkt. Regression.
- **Tidsseriedata:** data uppmätta över **tid**. y_t, $t = 1, 2, ...$
- Mäts ofta vid tidpunkter med **likstora avstånd** (varje månad).
- Tidsserier är speciella:
  - **Trender**, **säsong**.
  - **Beroende observationer** över tid. Värdet igår y_{t-1} kan användas för att prediktera dagens värde y_t. **Autokorrelation**.
  - Kräver **speciella modeller** som tar hänsyn till beroenden.

Tidsserier

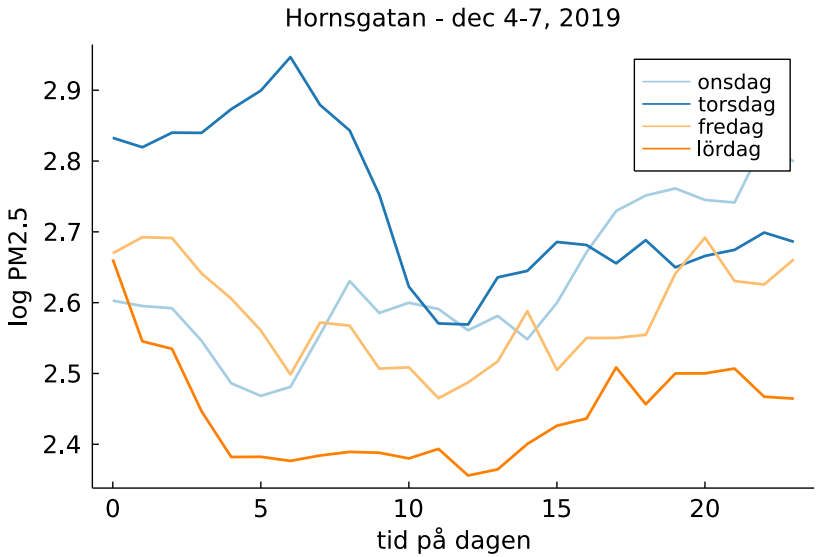

estimated μ
time series
prediction
Inflation
Month
3
2
1
0
1995
2000
2005
2010
2015
2020

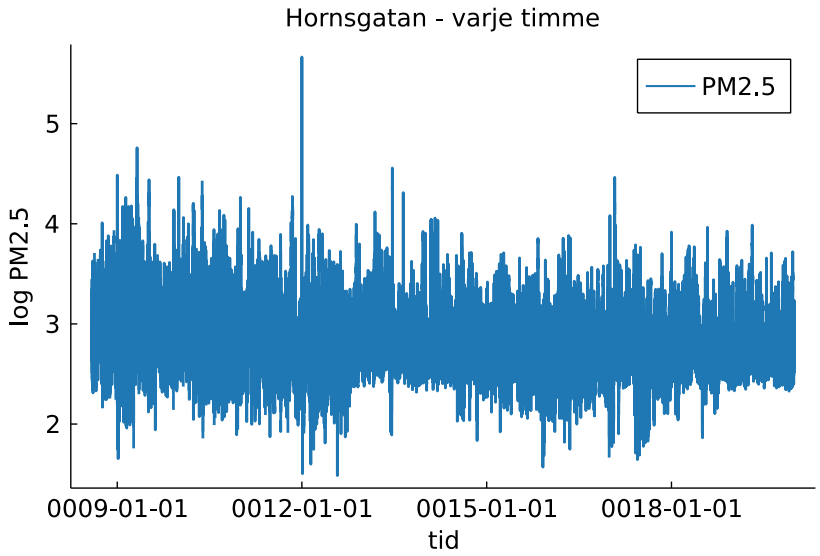

SP500
daily returns (%)
day
5
0
−5
−10
2017-01-01
2018-01-01
2019-01-01
2020-01-01
2021-01-01

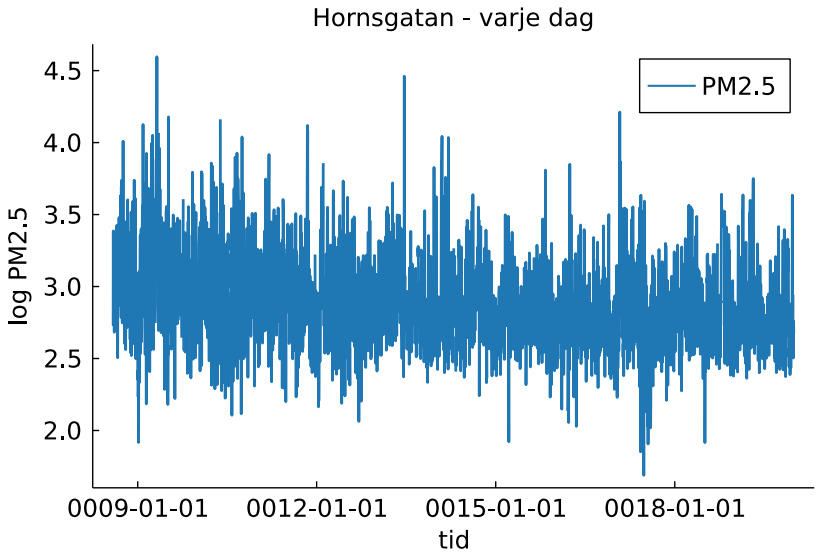

global temperature relative mean over 1951-1980
global temp anomaly
global temperature anomaly (C)
year
0.9
0.6
0.3
0.0
−0.3
1900
1925
1950
1975
2000

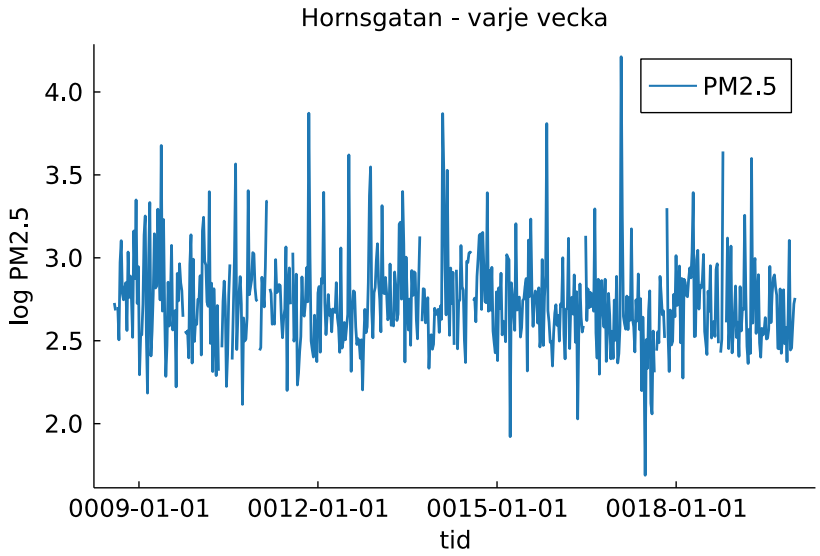

EEG signal at location Cz
time
6
3
0
- 3
- 6
0
250
500
750
1000

Miljöskadliga partiklar i luften på Hornsgatan







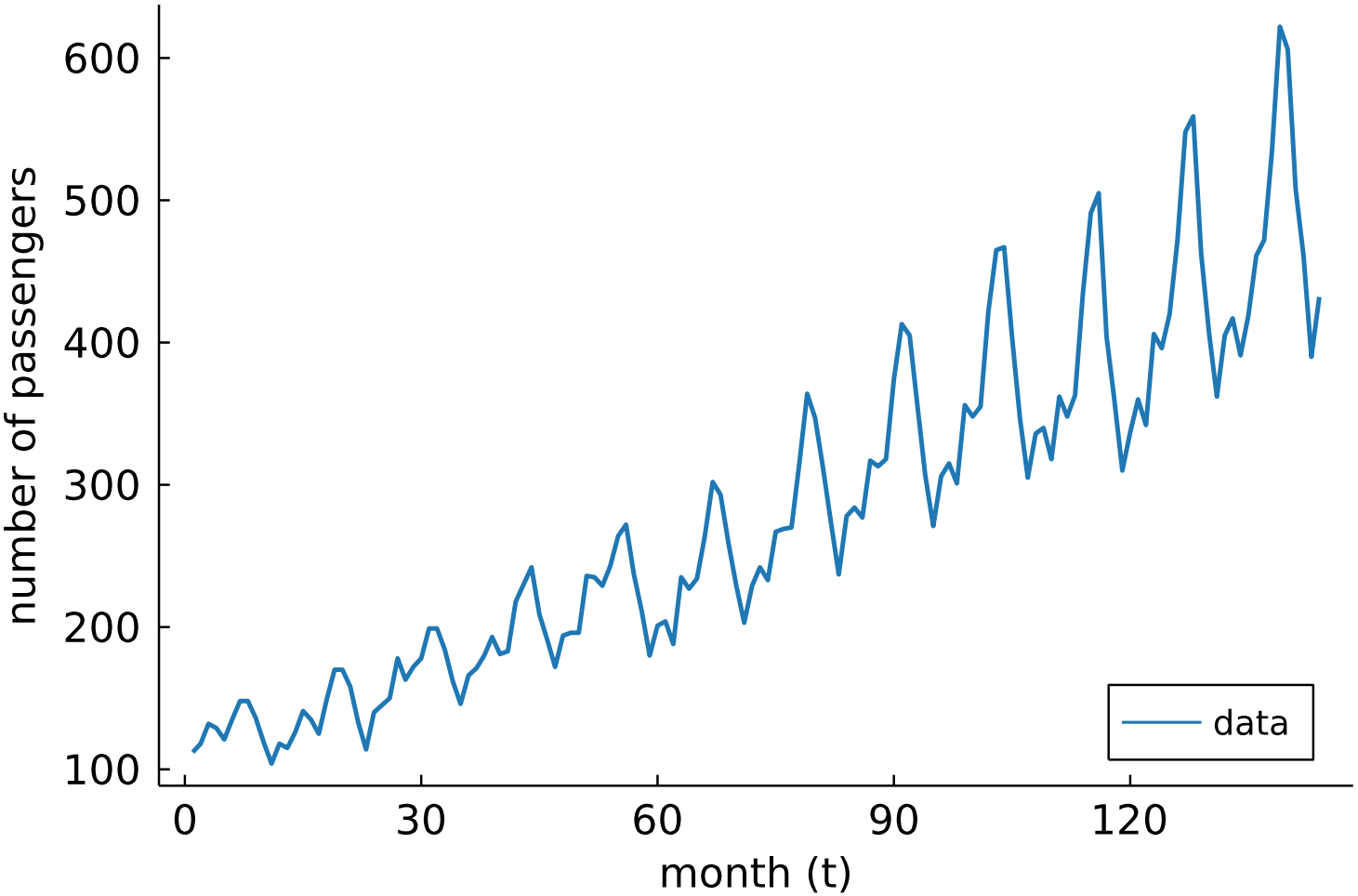


Airline passenger data

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
1960	417	391	419	461	472	535	622	606	508	461	390	432

data AirPassengers

Airline passenger data



Airline passenger data - linjär trend

- **Linjär trend**

$$y = a + b \cdot t$$

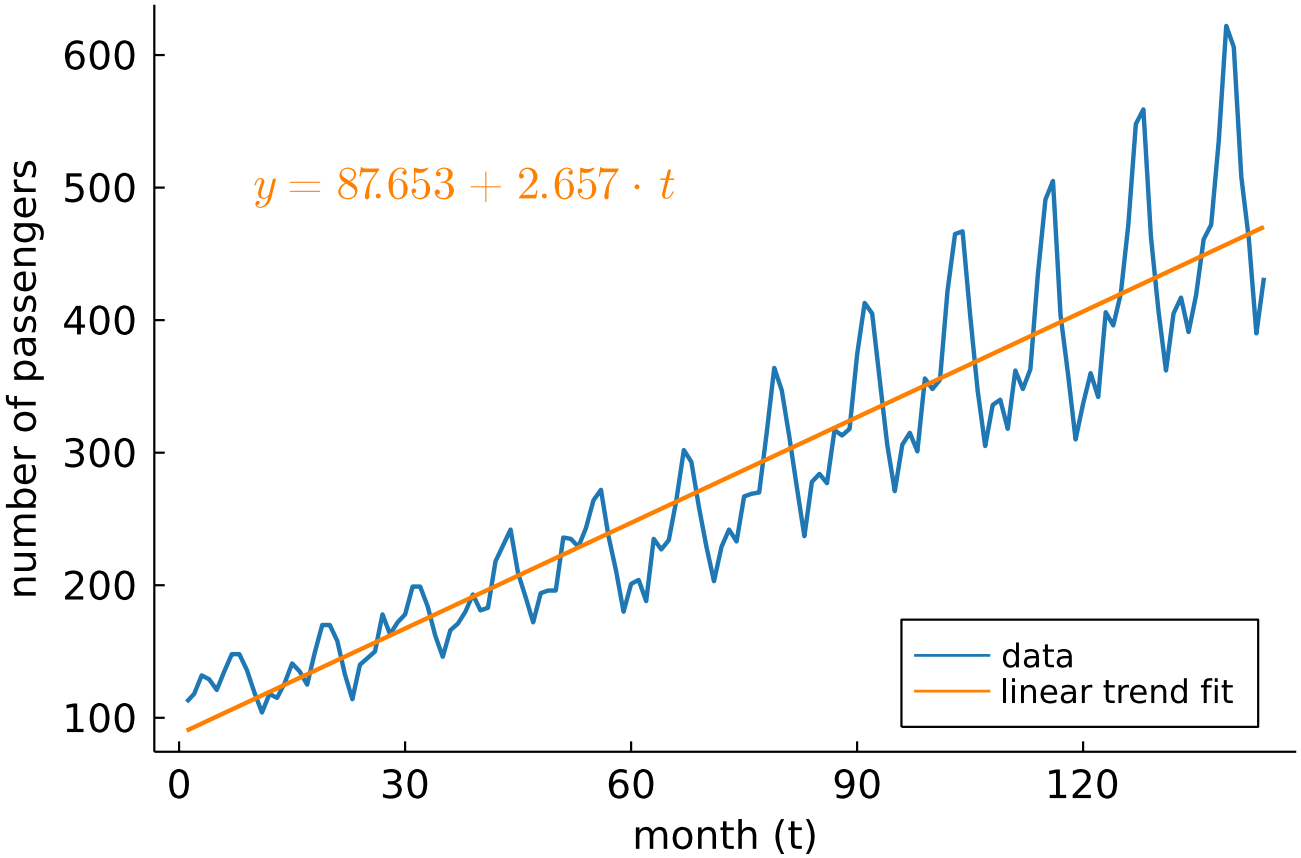
- **Minsta kvadrat**

```
passengers ~ 1 + time

Coefficients:
```

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	87.6528	7.71635	11.36	<1e-20	72.399	102.907
time	2.65718	0.0923325	28.78	<1e-60	2.47466	2.83971

Airline passenger data - linjär trend



- $R^2 = 0.853$.

Airline passenger data - kvadratisk trend

- **Kvadratisk trend**

$$y = a + b_1 \cdot t + b_2 \cdot t^2$$

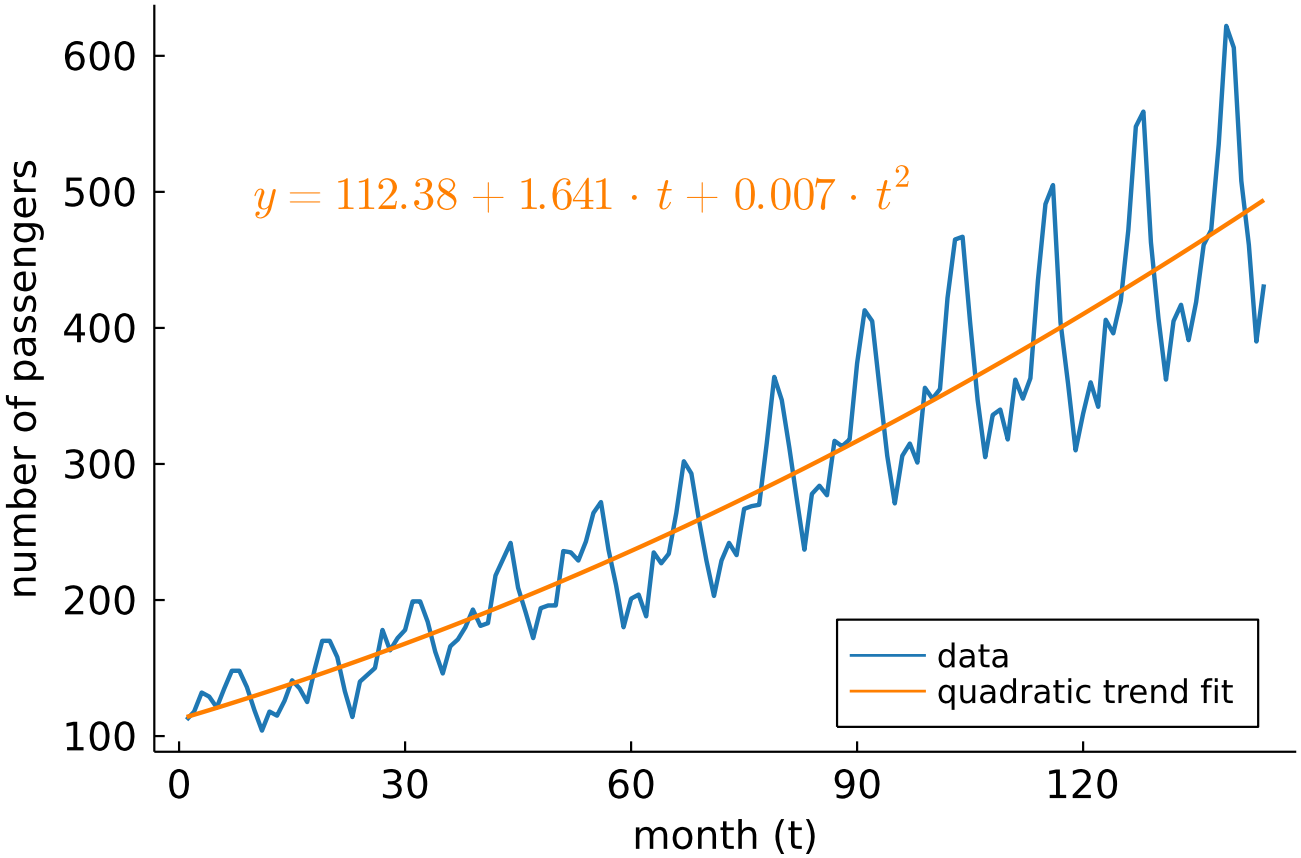
- Minsta kvadrat

```
passengers ~ 1 + time + :(time ^ 2)

Coefficients:
```

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	112.38	11.3841	9.87	<1e-17	89.8744	134.886
time	1.641	0.362473	4.53	<1e-04	0.92441	2.35758
time ^ 2	0.0070082	0.00242149	2.89	0.0044	0.00222108	0.0117953

Airline passenger data - kvadratisk trend



- $R^2 = 0.862$.

Airline passenger data - exponentiell trend

- **Exponentiell trend**

$$y = a \cdot b^t$$

- Skattas med minsta kvadrat genom att **logaritmera data**

$$\underbrace{\log y}_{\tilde{y}} = \underbrace{\log a}_{\tilde{a}} + \underbrace{\log b}_{\tilde{b}} \cdot t$$

$$\tilde{y} = \tilde{a} + \tilde{b} \cdot t$$

$$\tilde{a} = \log a$$

$$\tilde{b} = \log b$$

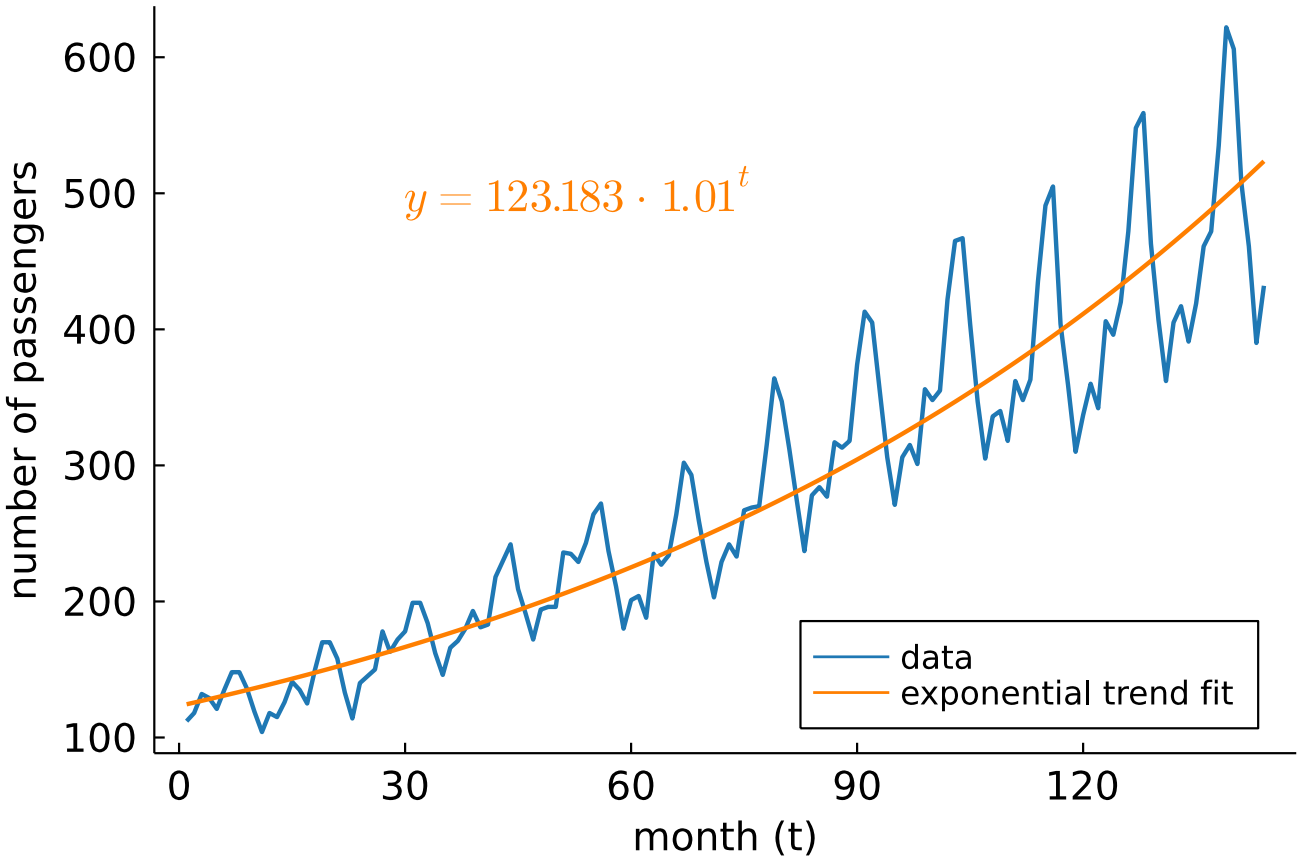
```
logpassengers ~ 1 + time

Coefficients:
```

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	2.09055	0.0101165	206.65	<1e-99	2.07055	2.11055
time	0.00436396	0.000121052	36.05	<1e-72	0.00412466	0.00460325

- $a = 10^{\tilde{a}} = 10^{2.09055} \approx 123.183$
- $b = 10^{\tilde{b}} = 10^{0.00436396} \approx 1.010.$

Airline passenger data - exponentiell trend



- $R^2 = 0.902$ för logarimerade data. Kan inte jämföras med tidigare modeller!

Airline passenger data - exponentiell trend

```
logpassengers ~ 1 + time

Coefficients:
```

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	2.09055	0.0101165	206.65	<1e-99	2.07055	2.11055
time	0.00436396	0.000121052	36.05	<1e-72	0.00412466	0.00460325

- Approximativt (n=144) 95% konfidensintervall för $\tilde{b}$

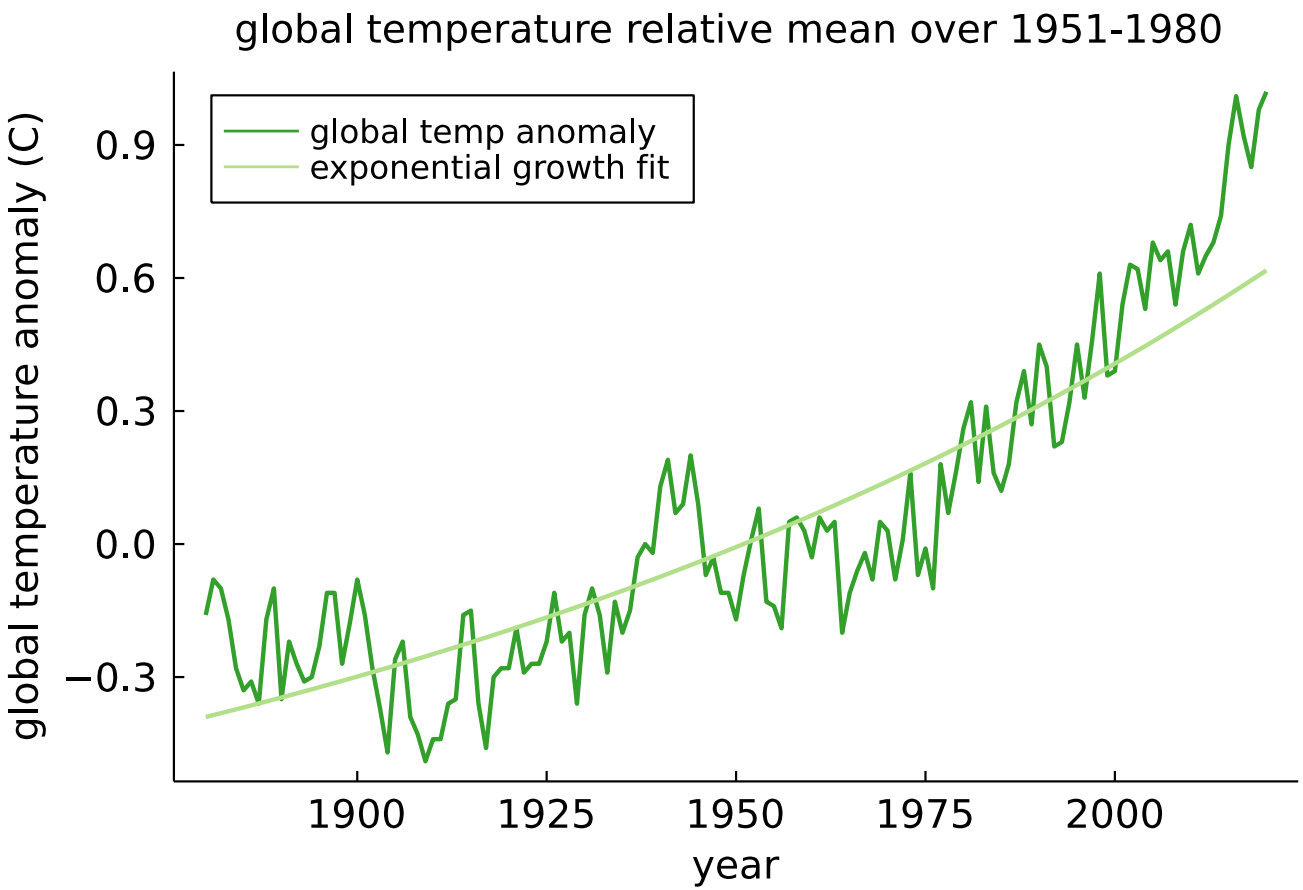
  $$0.00436396 \pm 1.96 \cdot 0.0001211052 = (0.004126594, 0.00460133)$$

- Approximativt (n=144) 95% konfidensintervall för b genom att anti-logga gränserna

  $$(10^{0.004126594}, 10^{0.00460133}) \approx (1.0095, 1.0107)$$

  dvs mellan 0.95% och 1.07% ökning per månad.
- 1.07% ökning per månad blir $1.0107^{12} \approx 1.1362$, dvs ca 13.62% ökning per år.

Global temperatur - exponentiell trend



SUdatasets globaltemp

- $R^2 = 0.764$ för logaritmerade data.

Trendskattning genom glidande medelvärden

- 3-punkts (centrerat) **glidande medelvärde** med **lika vikter**:

$$M_t = (y_{t-1} + y_t + y_{t+1})/3$$

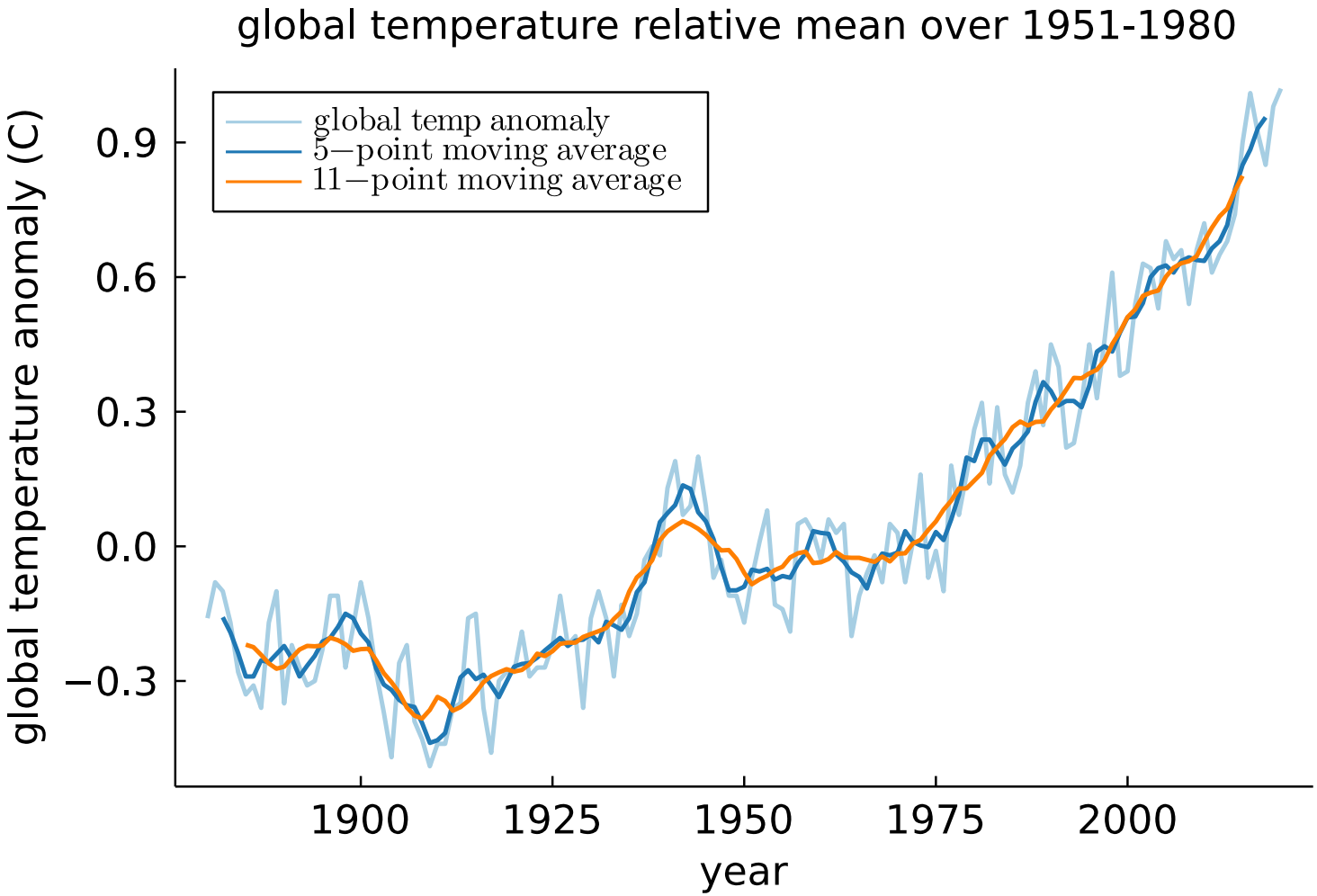
- 3-punkts **glidande medelvärde** med **olika vikter**:

$$M_t = w_{-1}y_{t-1} + w_0 y_t + w_1 y_{t+1}$$

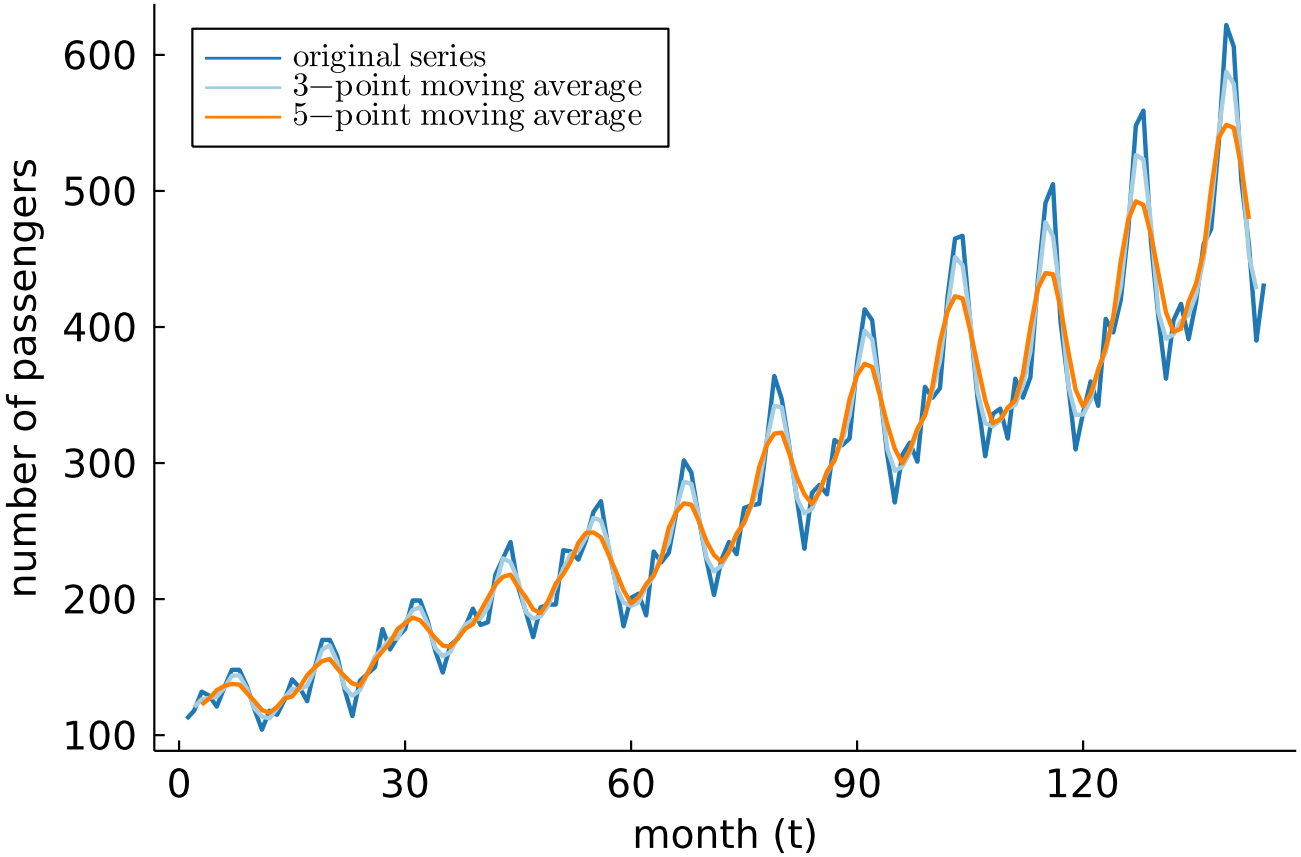
- Notera att vikterna måste summera till 1.
- $2r+1$-**punkts glidande medelvärde**

$$M_t = \sum_{s=-r}^{r} w_s y_{t+s}$$

Trendskattning genom glidande medelvärden



Airline passenger data - glidande medelvärden



Trendskattning - glidande medelvärden - säsong

- Kvartalsdata (ex: $t = \text{Kvartal3}$):

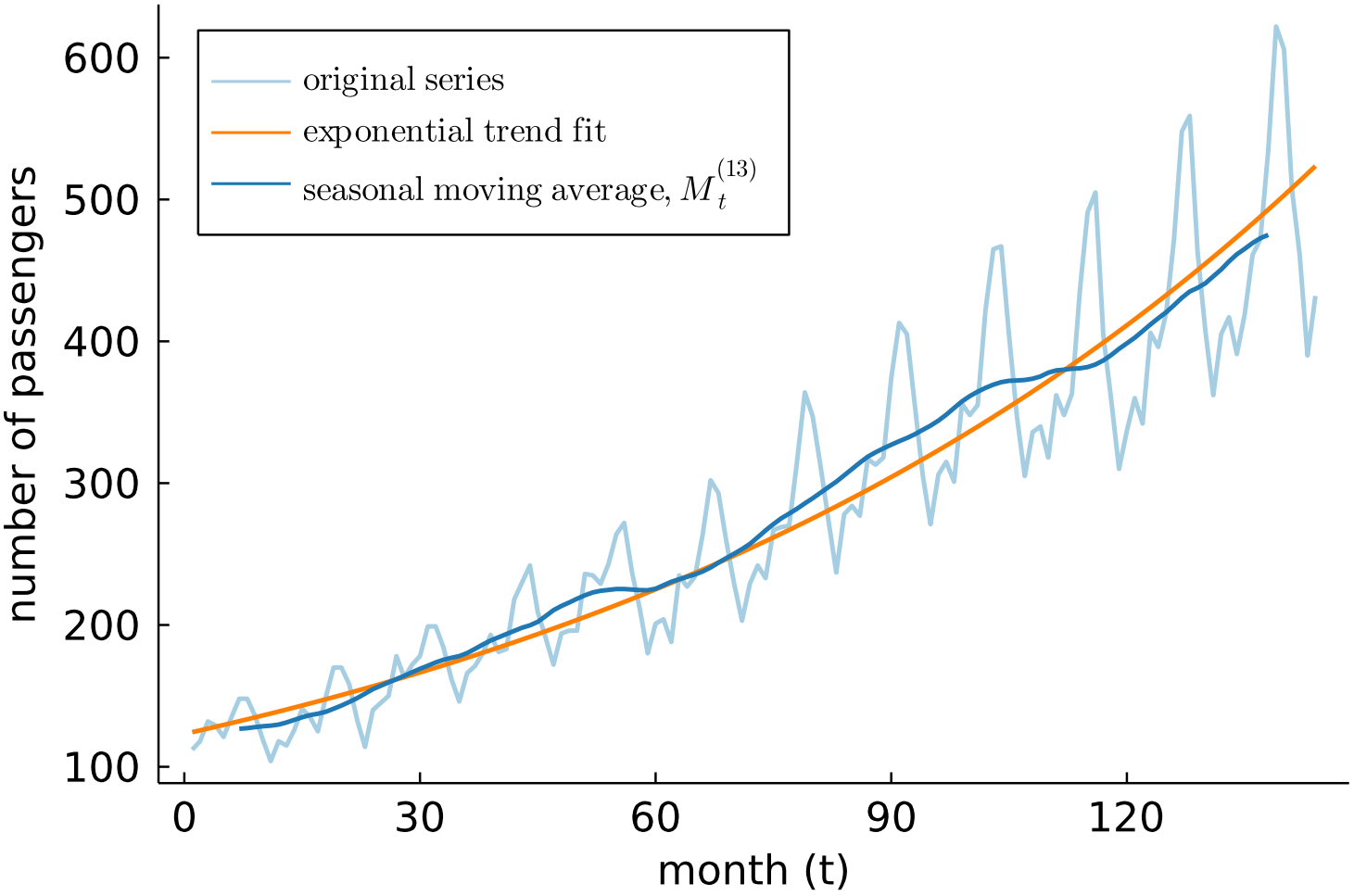
$$M_t^{(5)} = \left(\underbrace{y_{t-2}}_{\text{Kv1}} + 2\underbrace{y_{t-1}}_{\text{Kv2}} + 2\underbrace{y_t}_{\text{Kv3}} + 2\underbrace{y_{t+1}}_{\text{Kv4}} + \underbrace{y_{t+2}}_{\text{Kv1}} \right) /8$$

- Månadsdata (ex: $t = \text{juni}$):

$$M_t^{(13)} = \left(\underbrace{y_{t-6}}_{\text{dec}} + 2\underbrace{y_{t-5}}_{\text{jan}} + \ldots + 2\underbrace{y_t}_{\text{juni}} + \ldots + 2\underbrace{y_{t+5}}_{\text{nov}} + \underbrace{y_{t+6}}_{\text{dec}} \right) /24$$

´

Trendskattning - glidande medelvärden - säsong



Komponentsuppdelning

- En tidsserie kan delas upp i komponenter:
  - **Trend variation** (T)
  - **Cyklisk variation** (C)
  - **Säsongsvariation** (S)
  - **Slumpkomponent** (E)
- **Additiv modell**

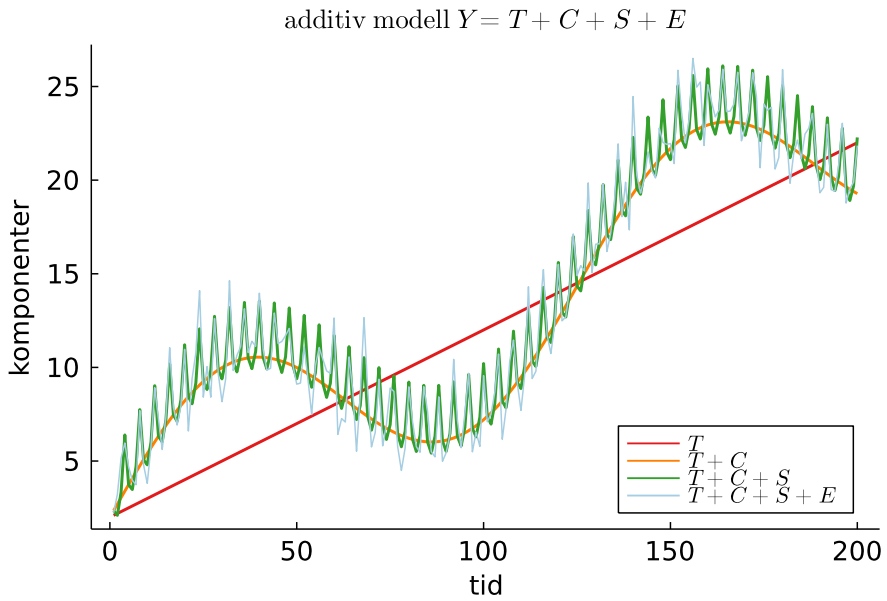
$$y_t = T_t + C_t + S_t + E_t$$

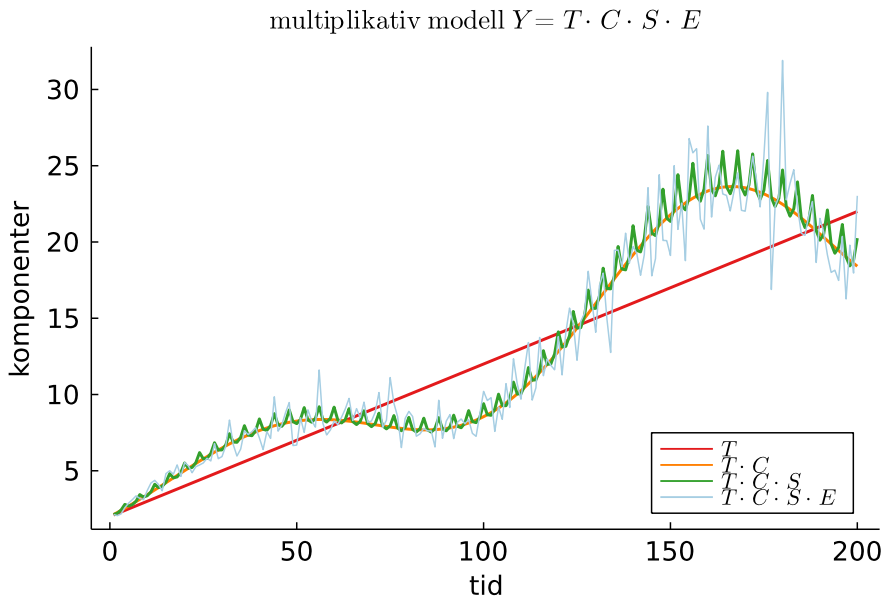
- Säsongseffekten är **visst värde över/under trend**, t ex decemberförsäljningen är 200 tkr högre i december.
- **Multiplikativ modell**

$$y_t = T_t \cdot C_t \cdot S_t \cdot E_t$$

- Säsongseffekten är **visst procent över/under trend**, t ex decemberförsäljningen är 18% högre i december.

Additiv vs multiplikativ uppdelning





Komponentsuppdelning - additiv modell

- Additiv model utan cyklisk komponent:

$$y_t = T_t + S_t + E_t$$

- Steg 1: **Bedöm modelltypen** genom att plotta tidsserien: **additiv** eller **multiplikativ**? Vilken **trendmodell**?
- Steg 2: Skatta **trendkomponenten** $\hat{T}_t$.
  T ex parametrisk modell eller glidande medelvärde (ev säsongsvariant).
- Steg 3: **Rensa bort trenden**: $y_t - \hat{T}_t \approx S_t + E_t$
- Steg 4: Skatta säsongskomponenten genom att beräkna medelvärden av $y_t - \hat{T}_t$ för varje säsong separat.

Airline - additiv med glidande säsongsmedel

månad	tidsserie y_t	trend $\hat{T}_t$	grov säsong $y_t - \hat{T}_t$	säsong S^+	säsongsjust. $y_t - S^+$
1949-01-01	112	·	·	-24.749	136.749
1949-02-01	118	·	·	-36.188	154.188
1949-03-01	132	·	·	-2.241	134.241
1949-04-01	129	·	·	-8.037	137.037
1949-05-01	121	·	·	-4.506	125.506
1949-06-01	135	·	·	35.403	99.597
1949-07-01	148	126.792	21.208	63.831	84.169
1949-08-01	148	127.25	20.75	62.823	85.177
1949-09-01	136	127.958	8.042	16.520	119.480
1949-10-01	119	128.583	-9.583	-20.643	139.643
1949-11-01	104	129.0	-25.0	-53.594	157.594
1949-12-01	118	129.75	-11.75	-28.620	146.620
1950-01-01	115	131.25	-16.25	-24.749	139.749
1950-02-01	126	133.083	-7.083	-36.188	162.188
1950-03-01	141	134.917	6.083	-2.241	143.241
1950-04-01	135	136.417	-1.417	-8.037	143.037
1950-05-01	125	137.417	-12.417	-4.506	129.506
1950-06-01	149	138.75	10.25	35.403	113.597
1950-07-01	170	140.917	29.083	63.831	106.169
1950-08-01	170	143.167	26.833	62.823	107.177
1950-09-01	158	145.708	12.292	16.520	141.480
1950-10-01	133	148.417	-15.417	-20.643	153.643
1950-11-01	114	151.542	-37.542	-53.594	167.594
1950-12-01	140	154.708	-14.708	-28.620	168.620
1951-01-01	145	157.125	-12.125	-24.749	169.749
1951-02-01	150	159.542	-9.542	-36.188	186.188
1951-03-01	178	161.833	16.167	-2.241	180.241
1951-04-01	163	164.125	-1.125	-8.037	171.037
1951-05-01	172	166.667	5.333	-4.506	176.506
.	.	.	.	.	.

Skattning av säsongskomponenten

- Steg 4: **Skatta säsongskomponenten**. Ex kvartalsdata:

$$\bar{S}_1 = \frac{\sum_{\text{alla t som är kvartal 1}} (y_t - \hat{T}_t)}{\text{antal kvartal 1 observationer}}$$

$$\bar{S}_2 = \frac{\sum_{\text{alla t som är kvartal 2}} (y_t - \hat{T}_t)}{\text{antal kvartal 2 observationer}}$$

$$\bar{S}_3 = \frac{\sum_{\text{alla t som är kvartal 3}} (y_t - \hat{T}_t)}{\text{antal kvartal 3 observationer}}$$

$$\bar{S}_4 = \frac{\sum_{\text{alla t som är kvartal 4}} (y_t - \hat{T}_t)}{\text{antal kvartal 4 observationer}}$$

- Steg 5: **Korrigera säsongen** så summan av säsongskomponenterna är noll:

$$S_i^+ = \bar{S}_i - \frac{\bar{S}_1 + \bar{S}_2 + \bar{S}_3 + \bar{S}_4}{4}$$

Skattning av säsongskomponenten

- Steg 6: **Rensa bort säsongen** genom att:
  - dra av S_1^+ från alla observationer i kvartal 1
  - dra av S_2^+ från alla observationer i kvartal 2, osv

$$y_t - \hat{T}_t - S_{i_t}^+ \approx E_t$$

där i_t är säsongen vid tidpunkt t. T ex $i_7 = 2$ om tidpunkt $t = 7$ är i kvartal 2.

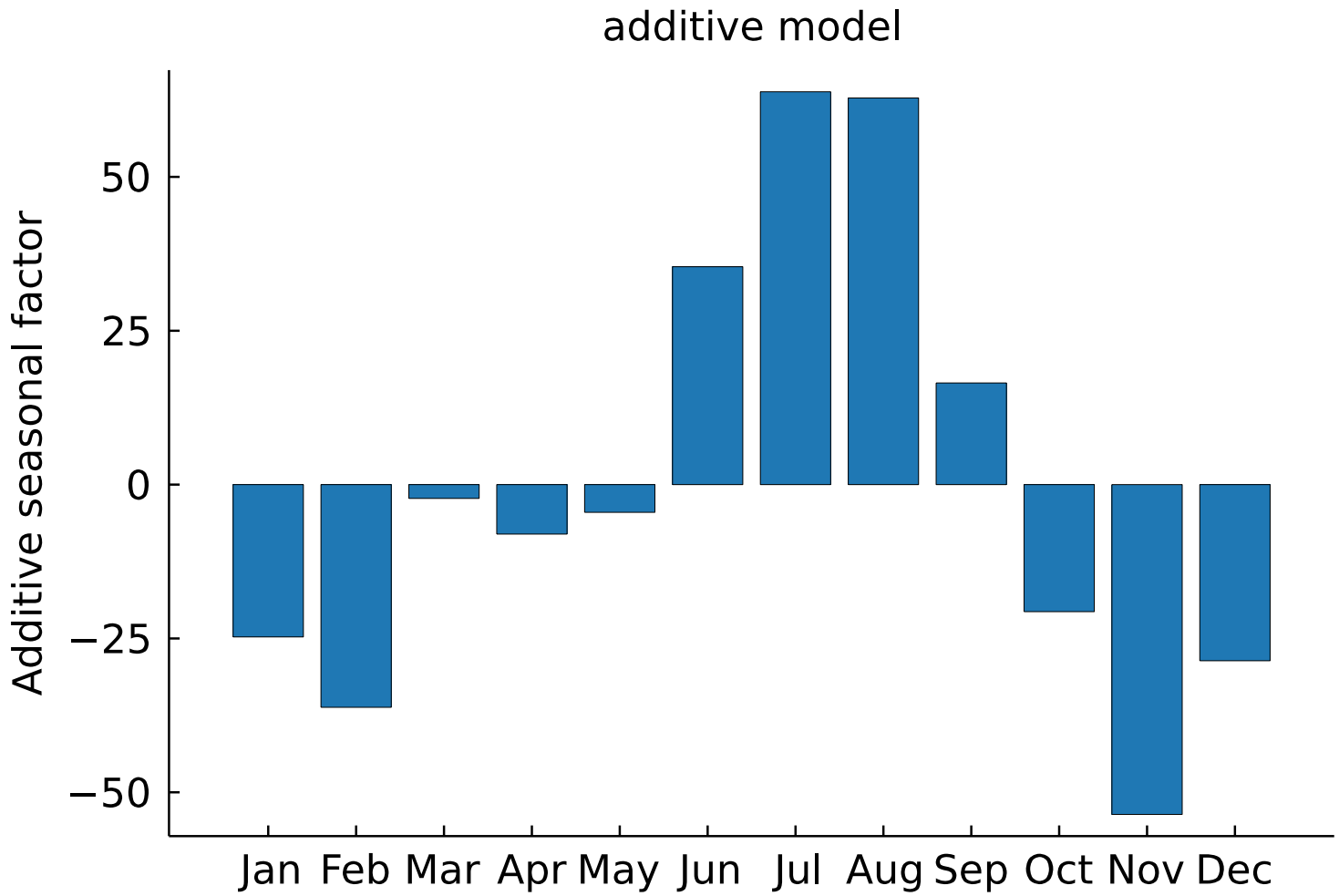
- **Multiplikativ modell - Variant 1**: logga för göra additiv

$$\log y_t = \log T_t + \log C_t + \log S_t + \log E_t = \tilde{T}_t + \tilde{C}_t + \tilde{S}_t + \tilde{E}_t$$

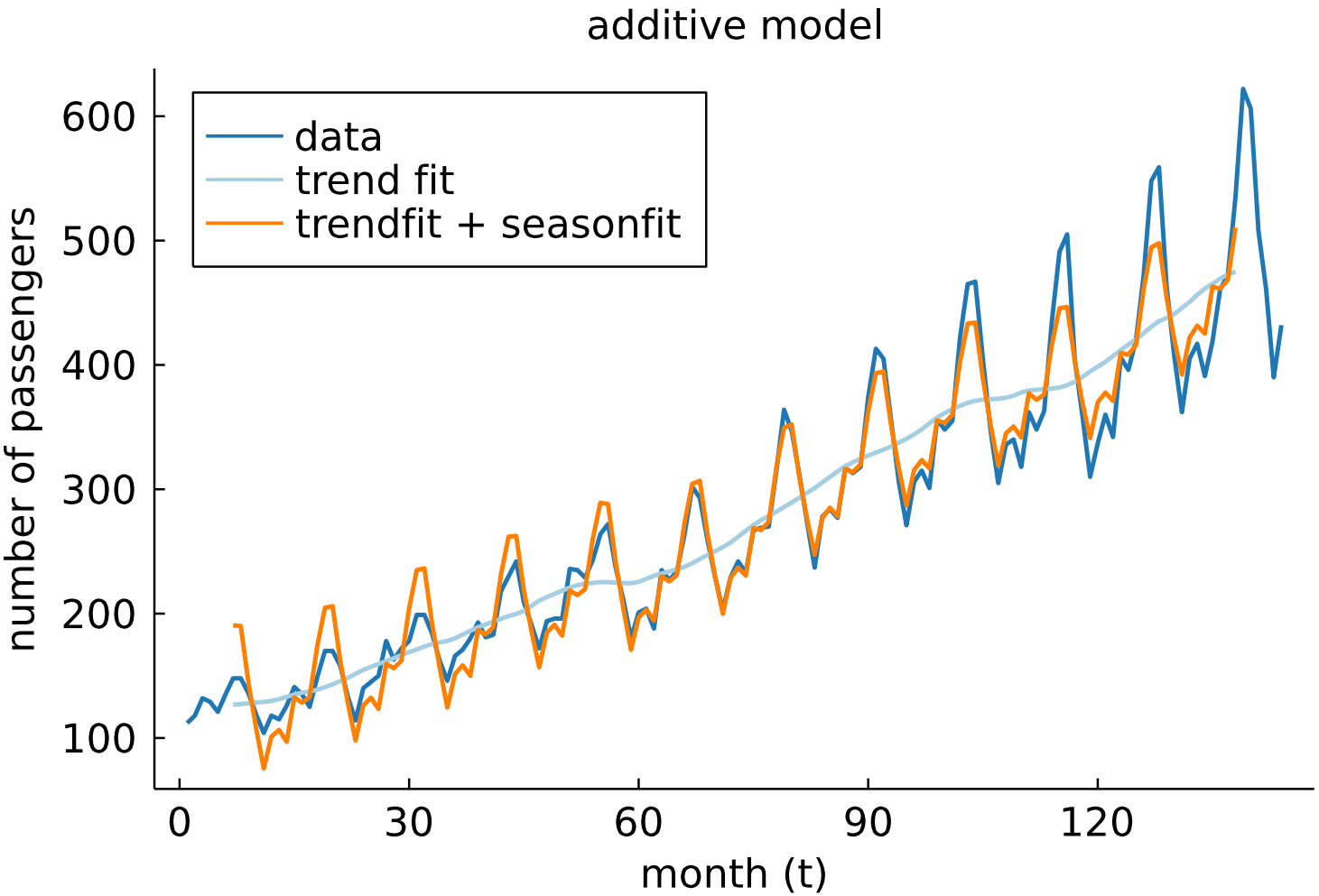
- **Multiplikativ modell - Variant 2**: uppdelning på orginalskala. Dividera istället för subtrahera för att rensa, ex:

$$\frac{y_t}{\hat{T}_t} \approx S_t \cdot E_t$$

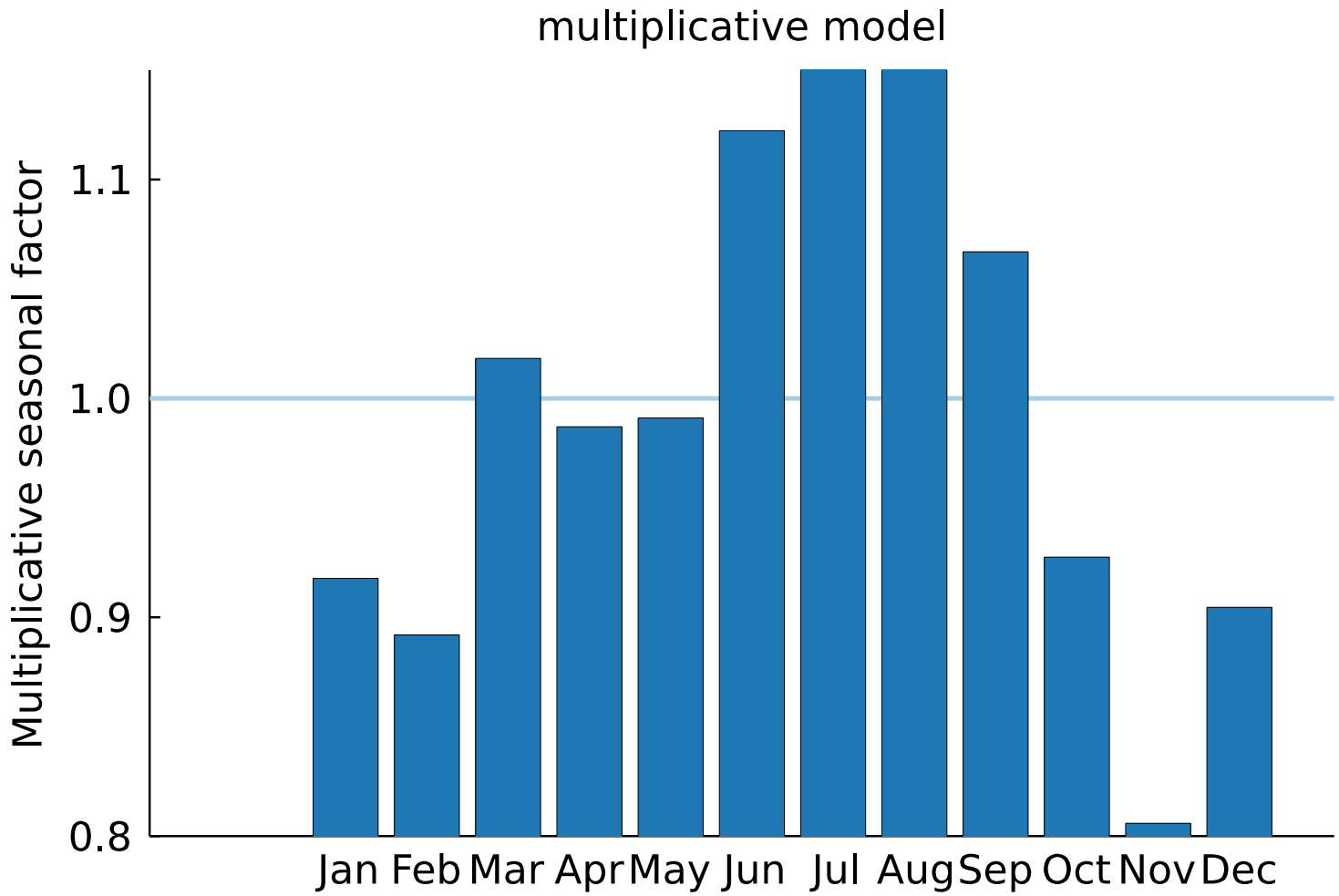
Airline passenger data - säsongskomponent S_i^+



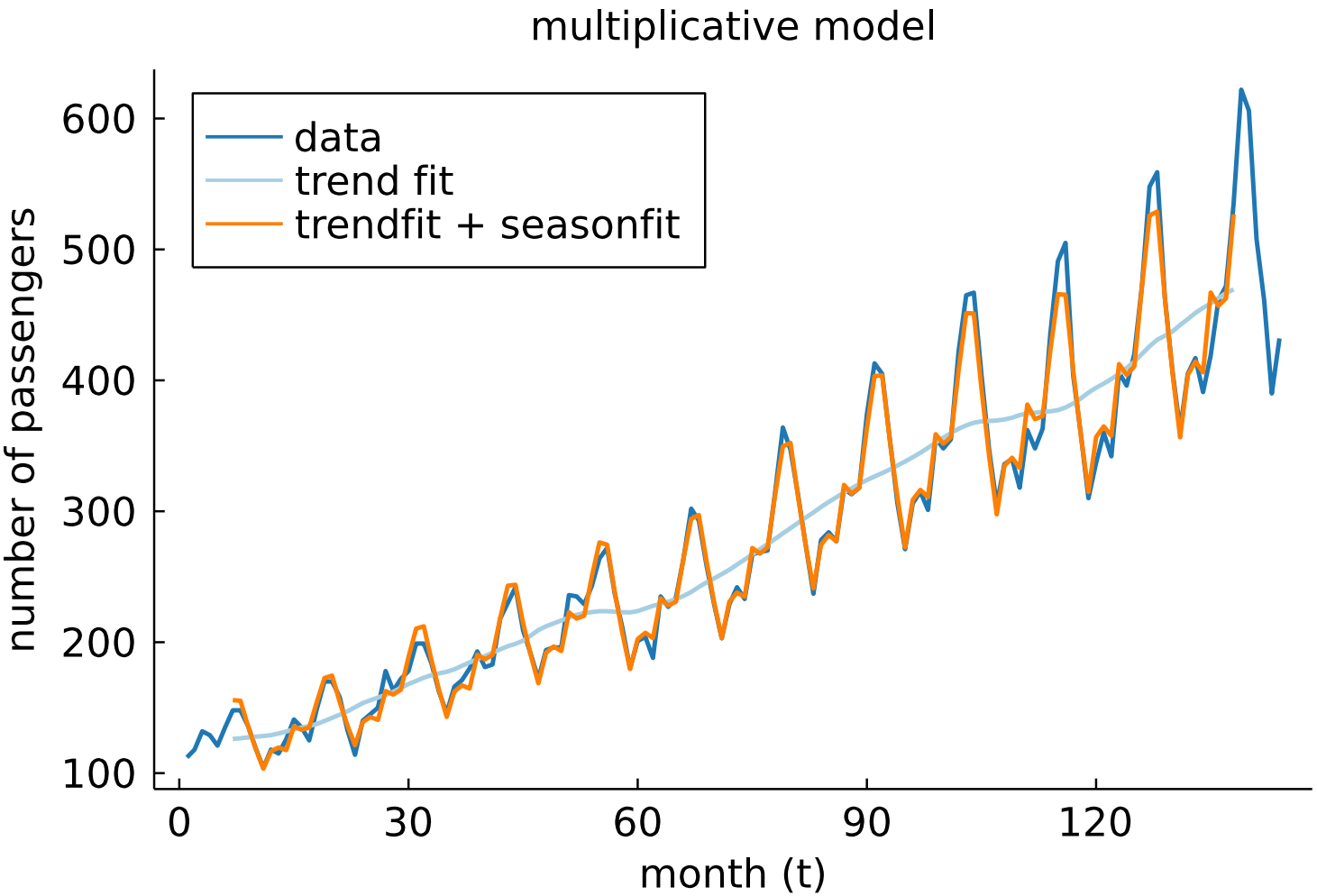
Airline passenger data - komponentanpassning



Airline passenger data - säsongskomponent S_i^+



Airline passenger data - komponentanpassning



Airline passenger data - komponentanpassning