Scalable Data Engineering, Exercise 5

Henry Haustein

Task 1

(a) True, technically there is also a base part, but often this is included into the trend.

(b) False, the season mask is fixed in length.

(c) False, this does piece wise aggregate approximation. SAX represents segments with an alphabet.

(d) False, with very large time series this takes forever.

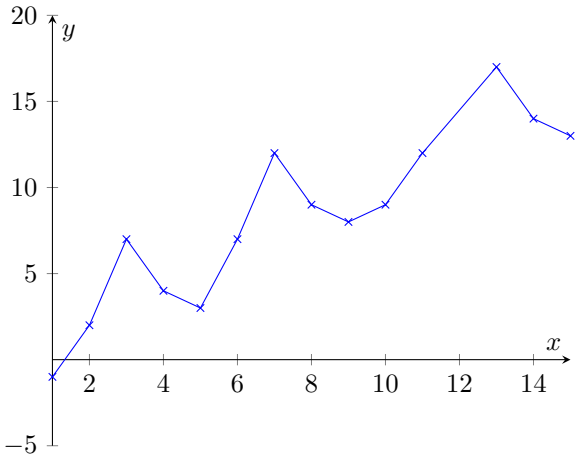
(e) False, for some forecasting methods you need the season length too.

(f) False, e.g. no one can predict the value of a share in the future.

(g) True.

Task 2

(a) Visualization



The local maxima are at 3, 8 and 13. This suggests a season length of 5. Running the following R code

```r
series = c(-1, 2, 7, 4, 3, 4, 7, 12, 9, 8, 9, 12, 17, 14, 13)
trend = filter(series, rep_len(1,5)/5)
detrend = series - trend
mat = matrix(detrend, ncol = 5, byrow = TRUE)
figure = apply(mat, 2, mean, na.rm = TRUE)
```

```
6  season = figure - mean(figure)
```

We get as `detrend: NA, NA, 4, -8.881784e-16, -2, -2, 0, 4, 0, -2, -2, -1.776357e-15, 4, NA, NA`. And as season mask: `-2, 0, 4, 4.440892e-16, -2`.

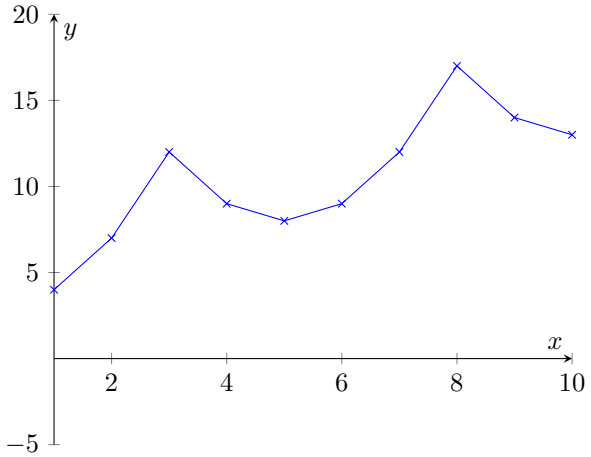
(b) I don't see any problem. Obviously one season has a length of 4, so running

```
1  series = c(-1, 2, 5, 4, 3, 6, 9, 8, 7, 10, 13, 12, 11, 14, 17,
       16)
2  decompose(ts(series, frequency = 4))
```

doesn't give any errors. All residuals are just 0.

Task 3

Let's plot the data first



We see $s = 5$. Then the predicted values are

x_t	**AR(1)**	**AR(2)**	**EGRV**	$\varepsilon^2_{AR(1)}$	$\varepsilon^2_{AR(2)}$	ε^2_{EGRV}
4	4	4	4	0	0	0
7	2	7	7	25	0	0
12	1	3	12	121	81	0
9	0,5	-0,3	9	72,25	86,49	0
8	0,25	-1,08	8	60,0625	82,4464	0
9	0,125	-0,558	9	78,765625	91,355364	0
12	0,0625	-0,0108	12	142,50390625	144,25931664	0
17	0,03125	0,16092	17	287,9384765625	283,5546152464	0
14	0,015625	0,099792	14	195,562744140625	193,215782443264	0
13	0,0078125	0,0115992	13	168,796936035156	168,698555341441	0

The AIC is

- for AR(1): 2305,760376

- for AR(2): 2264,040067
- for EGRV: 2

⇒ EGRV is for this time series the best forecast.