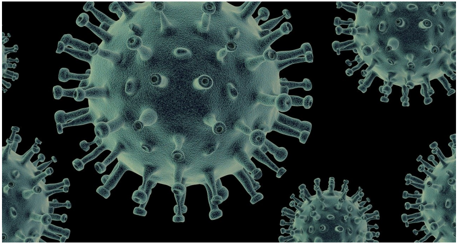
Classical Decomposition Model.

5 stycznia 2021



Classical Decomposition Model

Classical Decomposition Model - definition

The processes

$$X_t = \underbrace{m_t}_{\text{trend}} + \underbrace{s_t}_{\text{seasonality}} + \underbrace{Z_t}_{\text{random element}},$$

and we observe X_t for $t = 1, 2, \ldots, n$. Here

- m_t is a monotone or piecewise monotone deterministic function called *trend*;
- s_t is a periodical deterministic function called *trend*;
  - there is $T \in \mathbb{N}$ (assume it is known) such that $s_{t+T} = s_t$ for any $t \in \mathbb{N}$;
  - it holds:

    $$s_1 + s_2 + \ldots + s_T = 0;$$
- Z_t is a stationary ARMA(p,q) process with mean $EZ_t = 0$, i.e.

$$\begin{aligned} Z_t &= \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} \\ &+ \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}. \end{aligned}$$

Problems

- The main problem is to separate the trend, seasonality and residuals;
  - Preliminary estimation by *smoothing* the trend;
  - Estimation of seasonality;
  - Final estimation of trend;
- Having that we have to make analysis of random element:
  - By correlogram ACF, PACF, information criteria, and final estimation, fit the model ARMA(p,q) to the data;
  - Using ACF, PACF, Pierce-Box test and Ljung-Box test.

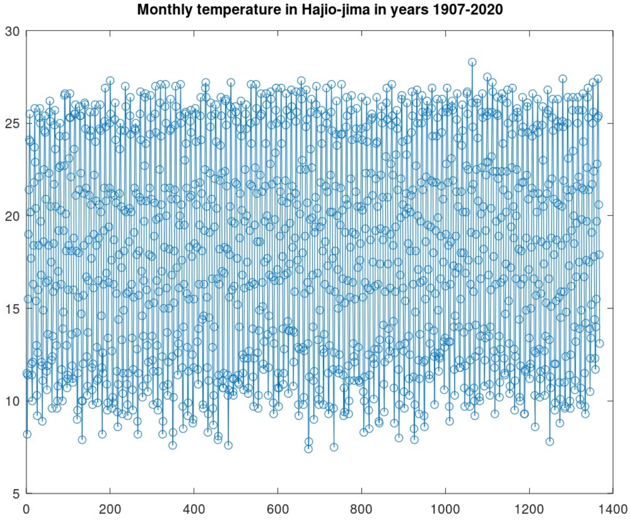
Smoothing - the reasons

Blurry vision

The long-term data is usually difficult to observe on the plot. For example, on the next slide we present the monthly average temperature in years in Hachijō-jima, the Japonise island in the Philippinea-See.

- The data of monthly temperature is denoted by o and the data is interpolated by the standard polygonal chain.
- As you will see the data is not illustrative:
  - We cannot see any direction of changing the temperature of monthly because we have to many data to observations (we have $2020 - 1906 = 114$ years each have 12 months and we have together $114 * 12 = 1368$ observations)
  - We cannot even formulate the hypothesis at a glance (see the picture in the next slide).

Monthly average temperature in years in Hachijō-jima



Smoothing

The purpose of smoothing is the preliminary estimation of the trend by the alignment of the data. We introduce two ways of smoothing.

- Smoothing with a finite moving average filter;
- Exponential smoothing.

Smoothing with a finite moving average filter

We define $\hat{m}_t^{MA}$ as follows:

- If T is an odd number in the form $T = 2r + 1\,(r \in \mathbb{N})$ then for $t \in [r+1, n-r]$ we define

$$\hat{m}_t^{MA} = \frac{1}{2r+1} \sum_{j=-r}^{r} X_{t-j}.$$

- If T is an even number in the form $T = 2r\,(r \in \mathbb{N})$ then for $t \in [r+1, n-r]$ we define

$$\hat{m}_t^{MA} = \frac{\frac{1}{2} X_{t-r} + \sum_{j=-r+1}^{r-1} X_{t-j} + \frac{1}{2} X_{t+r}}{2k}.$$

Then $\tilde{m}_t^{MA}$ is a preliminary estimator of m_t obtained by the *smoothing with a finite moving average filter*.

Smoothing with a finite moving average filter

Smoothing with a finite moving average filter removes the seasonality: for instance assume T is odd. Since $E(Z_t) = 0$ and $s_1 + s_2 + \ldots + s_T = 0$ we have

$$
\begin{aligned}
E(\hat{m}_t^{MA}) &= \frac{1}{2r+1} \sum_{j=-r}^{r} E(X_{t-j}) = \frac{\sum_{j=-r}^{r} m_{t-j} + s_{t-j}}{2r+1} \\
&= \frac{\sum_{j=-r}^{r} m_{t-j} + s_{t-j}}{2r+1} + \frac{\sum_{j=-r}^{r} s_{t-j}}{2r+1} \\
&= \frac{\sum_{j=-r}^{r} m_{t-j}}{2r+1} + \frac{s_1 + s_2 + \ldots + s_T}{2r+1} = \frac{\sum_{j=-r}^{r} m_{t-j}}{2r+1} \approx m_t .
\end{aligned}
$$

Smoothing with a finite moving average filter

Example of smoothing with the fictitious data

Let $(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = (3, 1, 2, 5, 2, 4, 0)$. Suppose that the period $T = 5$. Then $5 = 2r + 1$ for $r = 2$ and

$$\hat{m}_3^{MA} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} = \frac{3 + 1 + 2 + 5 + 2}{5} = \frac{13}{5},$$

$$\hat{m}_4^{MA} = \frac{X_2 + X_3 + X_4 + X_5 + X_6}{5} = \frac{1 + 2 + 5 + 2 + 4}{5} = \frac{14}{5},$$

$$\hat{m}_5^{MA} = \frac{X_3 + X_4 + X_5 + X_6 + X_7}{5} = \frac{2 + 5 + 2 + 4 + 0}{5} = \frac{13}{5},$$

and for $t = 1, 2, 6, 7$ the $\hat{m}_t^{MA}$ is undefined.

Smoothing with a finite moving average filter

Example of smoothing with the fictitious data

Let $(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = (3, 1, 2, 5, 2, 4, 0)$. Suppose that the period $T = 4$. Then $4 = 2r$ for $r = 2$ and

$$\hat{m}_3^{MA} = \frac{0.5X_1 + X_2 + X_3 + 0.5 * X_4}{4} = \frac{0.5 * 3 + 1 + 2 + 0.5 * 5}{4} = \frac{7}{4} = 1,$$

$$\hat{m}_4^{MA} = \frac{0.5X_2 + X_3 + X_4 + 0.5 * X_5}{4} = \frac{0.5 * 1 + 2 + 5 + 0.5 * 2}{4} = \frac{8.5}{4} = 1.7,$$

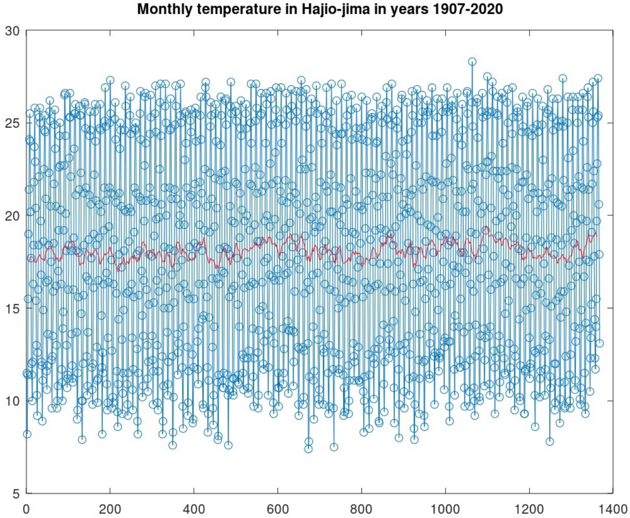
$$\hat{m}_5^{MA} = \frac{0.5X_3 + X_4 + X_5 + 0.5 * X_6}{4} = \frac{0.5 * 2 + 5 + 2 + 0.5 * 4}{4} = \frac{10}{4},$$

$$\hat{m}_6^{MA} = \frac{0.5X_4 + X_5 + X_6 + 0.5 * X_7}{4} = \frac{0.5 * 5 + 2 + 4 + 0.5 * 0}{4} = \frac{8.5}{7},$$

and for $t = 1, 2, 7$ the $\hat{m}_t^{MA}$ is undefined.

The next slide illustrates the application of smoothing with the real data of temperature in Hachijō-jima. This time the seasonality is $T = 12$ (12 months is the year) with $r = 6$

Monthly average temperature (1907-2020) in Hachijō-jima with the *moving average trend* (in red)



Let $\alpha \in [0, 1]$. We define $\hat{m}_t^E$ recursively as follows:

$$\hat{m}_t^E = \begin{cases} X_1 & \textit{if} \quad t = 1 \\ \alpha X_t + (1-\alpha)\hat{m}_{t-1}^E & \textit{if} \quad t \geq 2. \end{cases}$$

Then $\tilde{m}_t^E$ is a preliminary estimator of m_t obtained by the *exponentia smoothing*.

Remark

Unlike in case of $\hat{m}_t^{MA}$, the exponential smoothing is defined for exery t. Furthermore, for $t > 1$ we can express $\hat{m}_t^E$ explicitly as follows:

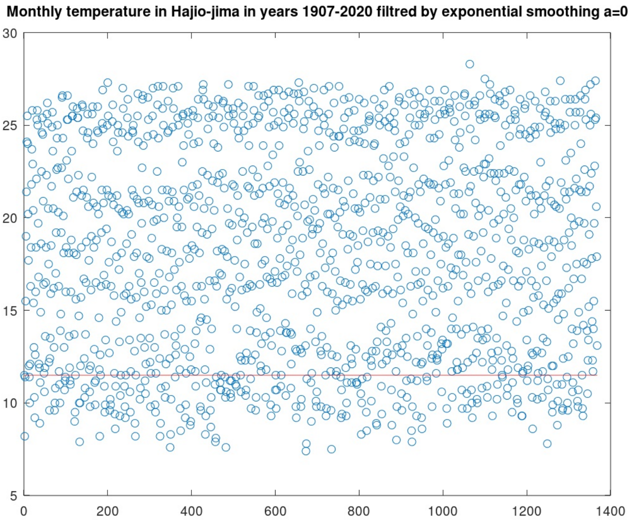
$$\hat{m}_t^E = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{t-1} X_1.$$

Remark

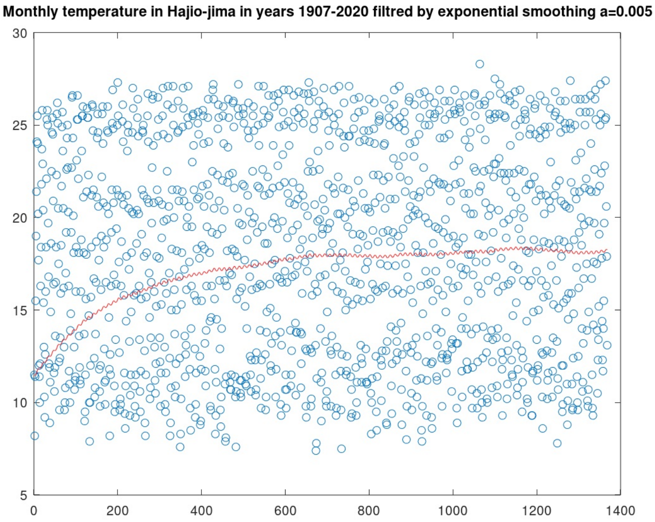
For extreme values $\alpha \in \{0, 1\}$ the exponential smoothing $\hat{m}_t^E$ are not illustrative.

- If $\alpha = 0$ then $\hat{m}_t^E = X_1$ for all t, hence is constant;
- If $\alpha = 1$ then $\hat{m}_t^E = X_t$ for all t, hence is coincides with data;
- Only for $\alpha \in (0, 1)$ the estimator $\hat{m}_t^E$ may approximate m_t, the choice depends on the problem;
- The next slides illustrates exponential smoothing with distinct $\alpha \in (0, 1)$ for monthly average temperature in years 1907-2020 in Hachijō-jima

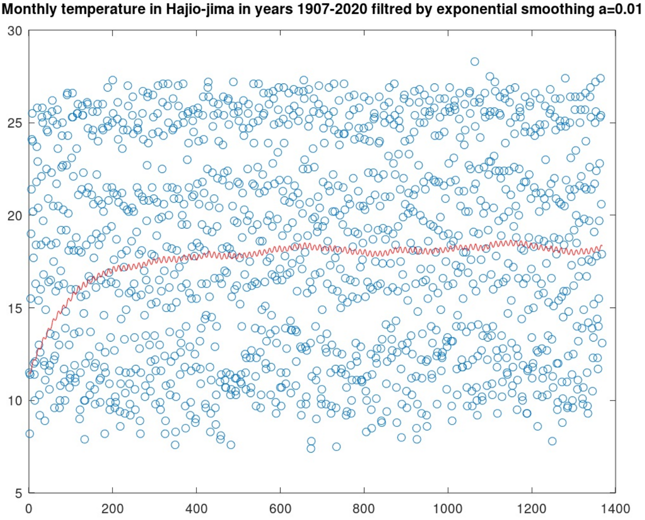
Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0$) (in red)



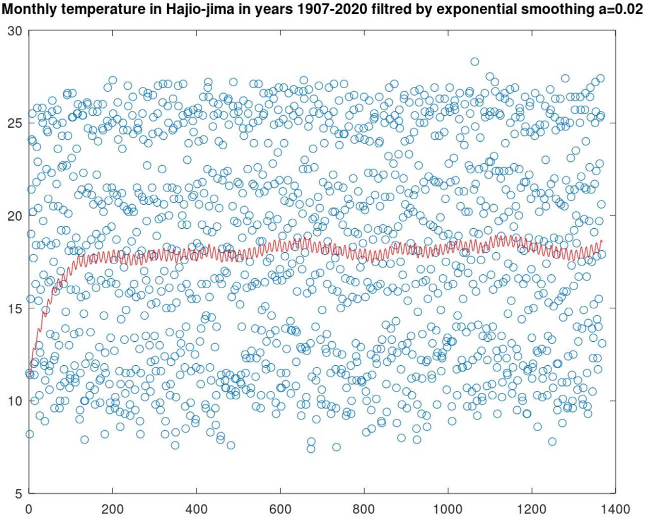
Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0.005$) (in red)



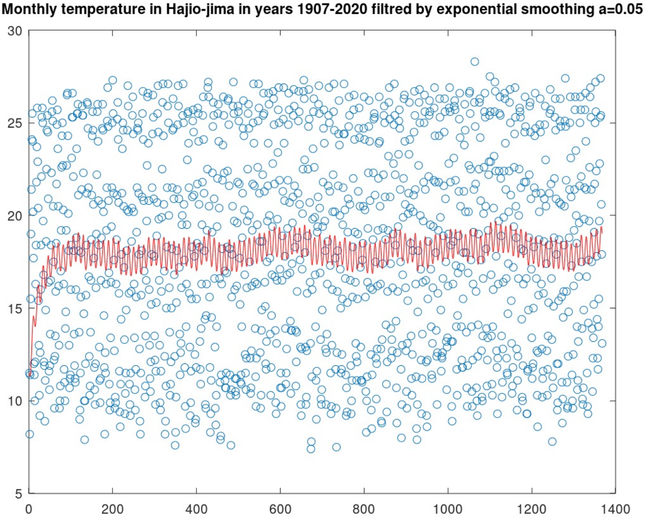
Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0.01$) (in red)



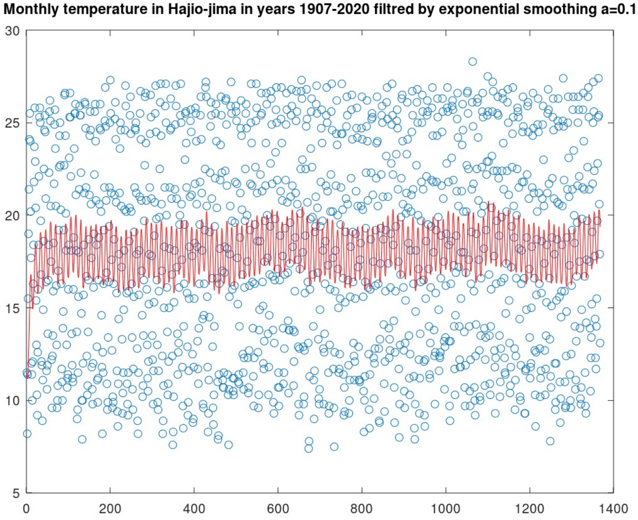
Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0.02$) (in red)



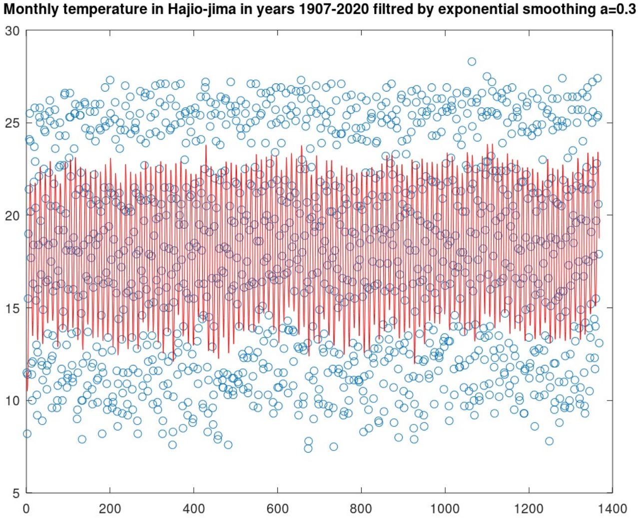
Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0.05$) (in red)



Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0.1$) (in red)



Monthly average temperature (1907-2020) in Hachijō-jima with the *exponential smoothing* ($\alpha = 0.3$) (in red)

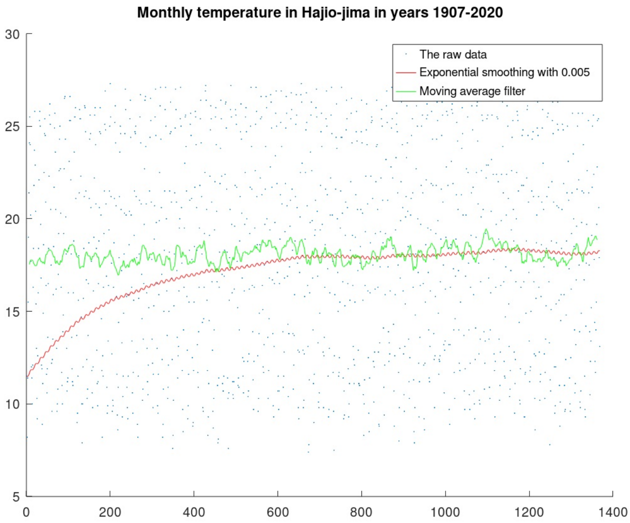


Moving average or exponential smoothing

Remarks

- The smoothing with moving average filter pays equal attention to all observations $X_{r+1}, X_{r+2}, \ldots, X_{n-r}$ and less attention to *far past* observation $X_1, X_2, ..., X_r$ and *far recent* $(X_{n-r+1}, X_{n-r+2}, \ldots, X_n)$;
- The exponential smoothing pays the greatest attention to the first observation X_1 and pays decreasing attention from the fast past to the fast recent observations;
- The choice of smoothing depends on us and the hypothesis, as well as on the illustrative property of the plot. In case of Hachijō-jima, the most illustrative plots are for moving average filter and the exponential smoothing with $\alpha = 0.005$. Both plots differ
- The next slides shows how the smoothness may provide different outputs

Moving average or exponential smoothing



Comparing of plots

Observe that

- The estimators $\hat{m}_t^{MA}$ (moving average) and $\hat{m}_t^{E}$ significantly differ for months $t = 1, 2, \ldots, 670$ which correspond July 1962, and behave similarly for $t > 670$.
- This time the plot $\hat{m}_t^{E}$ is smoother than $\hat{m}_t^{MA}$;

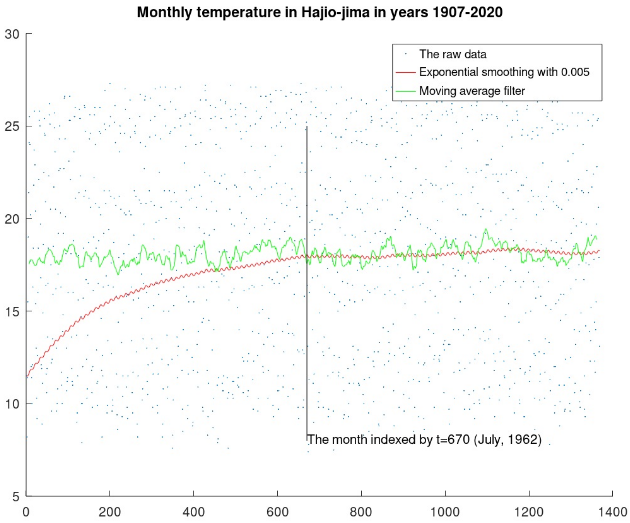
Different smoothing, different hypothesis

Hypothesis at a glance

According to this plot both estimators we have preliminary results:

- The estimator $\hat{m}_t^E$ detects increasing tendency from January 1907 to july 1962, but cannot detect any significant monotone tendency after this date;
- The estimator $\hat{m}_t^{MA}$ cannot detect any significant monotone tendency from entire period. See the next slide.

Different smoothing, different hypothesis



Further steps

Regardless of the smoothing method, the estimator cannot be final because of the following reasons:

- Both $\hat{m}_t^E$ and $\hat{m}_t^{MA}$ are useless for forecasting (extrapolation) since provide no formula;
- The eye may mislead and because of that, only the statistical and economical methods may arbitrate what the trend is;
- Further analysis are necessarily:
    - Estimation of the seasonality;
    - Final estimator of trend;
    - Forecasting;

Estimation of of the seasonality based on moving average filter.

Suppose that the seasonality is T and let $r = \left\lfloor \frac{T}{2} \right\rfloor$. For t such that $\hat{m}_t^{MA}$ is defined $t = r+1, \dots, n-r$ we define:

$$Y_t^{MA} := X_t - \hat{m}_t^{MA}.$$

For any $t = 1, 2, \dots, T$ we define

$$w_t := \frac{\sum_{j \in C_t^{MA}} Y_j^{MA}}{\# C_t^{MA}} \quad \text{where } C_t^{MA} := \{j \in \mathbb{N} : r+1 \leq t + j * T \leq n - r\},$$

and the estimator of seasonality we define as

$$\hat{s}_t := w_t - \frac{w_1 + w_2 + \ldots + w_T}{T}$$

for $t = 1, 2, \dots, T$ and for $t > T$ we replicate the values at every T such that

$$\hat{s}_t = \hat{s}_{t \bmod T}.$$

Estimation of of the seasonality based on exponential smoothing.

Suppose that the seasonality is T and let $r = \left\lfloor \frac{T}{2} \right\rfloor$. For t such that $\hat{m}_t^{MA}$ is defined $t = r+1, \dots, n-r$ we define:

$$Y_t^E := X_t - \hat{m}_t^E.$$

For any $t = 1, 2, \dots, T$ we define

$$w_t := \frac{\sum_{j \in C_t^E} Y_j^E}{\# C_t^E} \quad \text{where } C_t^E := \{j \in \mathbb{N} : 1 \leq t + j * T \leq n\},$$

and the estimator of seasonality we define as

$$\hat{s}_t := w_t - \frac{w_1 + w_2 + \dots + w_T}{T}$$

for $t = 1, 2, \dots, T$ and for $t > T$ we replicate the values at every T such that

$$\hat{s}_t = \hat{s}_{t \bmod T} = \hat{s}_{t - T\left\lfloor \frac{t}{T} \right\rfloor}.$$

Estimation of of the seasonality.

Regardless on the smoothing method

- First we compute $\hat{s}_t$ for $t = 1, 2, \ldots, T$ by average of the values of Y_j^{MA} or Y_j^E where j is taken over all all j such that j after dividing by T yields the remainder is t (j=T);
- Having $\hat{s}_t$ for $t = 1, 2, \ldots, T$, for any $\tau > T$ such that $\tau = t + jT$ for $j \in \mathbb{N}$, we define $\hat{s}_\tau$ as follows

$$\hat{s}_\tau := \hat{s}_{t+jT} = \hat{s}_t.$$

- We can write the estimator of seasonality in the sequence form:

$$(\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T, \hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T, \hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T, \ldots).$$

Final estimator of the trend - linear polynomial regression

Having estimator $\hat{s}_t$ of s_t we put:

$$Y_t := X_t - \hat{s}_t \approx m_t + Z_t.$$

We can approximate (t, Y_t) by the standard polynomial regression

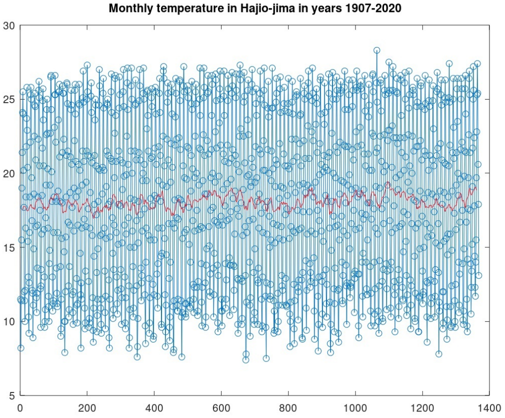
$$Y_t = a_0 + a_1 t + \ldots + a_k t^k + Z_t$$

and estimator of $(a_0, a_1, \ldots, a_k)$ is denoted as usual $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_k)$ as the solution of minimization problem

$$\min_{a_0, a_1, \ldots, a_k} \sum_{t=1}^{n} (Y_t - a_0 - a_1 t - \ldots - a_k t^k)^2.$$

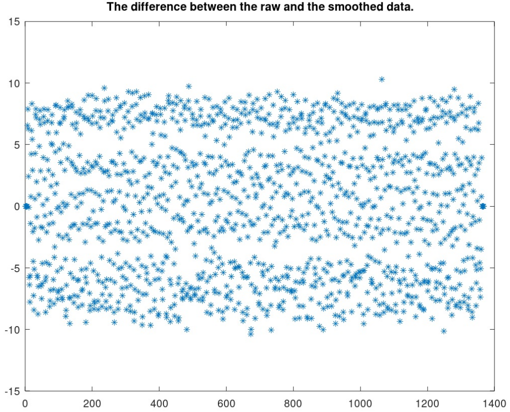
Monthly average temperature (1907-2020) in Hachijō-jima -countinue

The figure below shows the raw data X_t together with the trend estimated preliminarily by smoothing with moving average $\hat{m}_t^{MA}$.



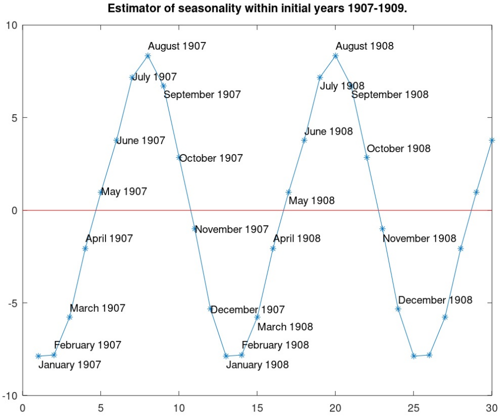
Monthly average temperature (1907-2020) in Hachijō-jima -countinue

The estimation of the seasonality is based on $X_t - \hat{m}_t^{MA}$ whose plot is below.



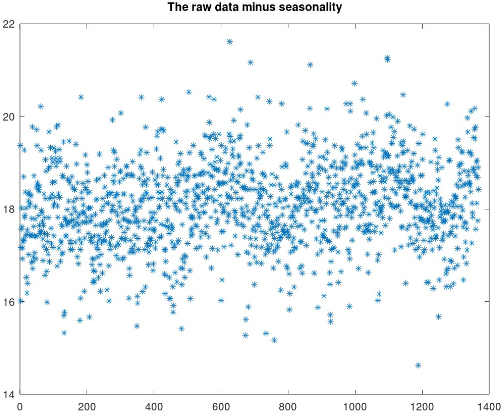
Monthly average temperature (1907-2020) in Hachijō-jima -countinue

The estimation of the seasonality is below.



Monthly average temperature (1907-2020) in Hachijō-jima -countinue

The seasonality is a distorting parameter, and we fit the line to the plot of $Y_t = X_t - \hat{s}_t$.



Monthly average temperature (1907-2020) in Hachijō-jima -countinue

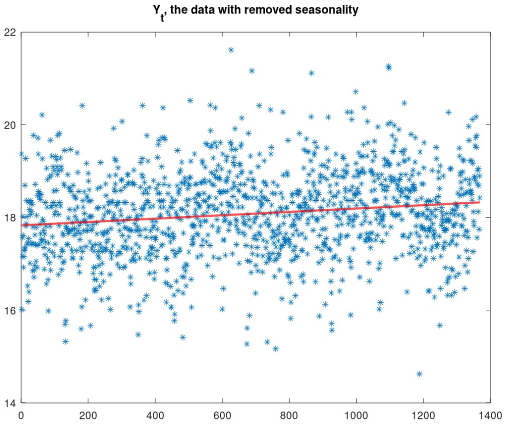
Results.

- We have $Y_t = X_t - \hat{s}_t \approx m_t + Z_t$.
- Assuming $m_t = a_0 + a_1 * t$ we estimate a_0 and a_1 by the least squared method.
- By calculations we obtain $a_1 = 0.00036232$, $a_0 = 17.83152175$ hence the trend has the following formula:

$$\hat{m}_t = 17.83152175 + 0.00036232 * t.$$

- In the next slide we see the estimator of the trend on the background of data with the removed seasonality.

Monthly average temperature (1907-2020) in Hachijō-jima -countinue



Monthly average temperature (1907-2020) in Hachijō-jima -countinue

Now we present the raw data X_t with the plot of $\hat{m}_t + \hat{s}_t$.

Residuals analysis.

Having $\hat{m}_t$ and $\hat{s}_t$ we determine the residuals as the approximation of Z_t

$$\hat{Z}_t = X_t - \hat{m}_t - \hat{s}_t.$$

We proceed further:

- Using t-test and Wald test we verify the significances of of variables in the model (in our case of $\hat{m}_t = \hat{a}_0 + \hat{a}_1 * t$ we verify whether $\hat{a}_1$ is significant). If not, remove this variable and find a new model without it.
- Test whether $\hat{Z}_t$ is stationary ARMA(p,q) using the current procedure:
  - Find corellogram i.e. ACF, PACF and identify model ARMA(p,q);
  - Find parameters of ARMA(p,q);
  - Find residuals of the model ARMA(p,q), $\hat{\epsilon}_t$, and using ACF,PACF again as well as Ljung-Box, Pierce -Box test verify $\hat{\epsilon}_t$ is white noise;
  - Verify whether $\hat{\epsilon}_t$ has a normal distribution (Jarque-Bera, Shapiro-Wilk, etc.).

Final task on laboratory.

Having $\hat{m}_t$ and $\hat{s}_t$ we determine the residuals as the approximation of Z_t

$$\hat{Z}_t = X_t - \hat{m}_t - \hat{s}_t.$$

We proceed further:

- Find a real data from internet which can be viewed as a time series: e.g. stock, currency exchange, average month temperature in selected place, population in selected city, town, region or country, daily infected people by COVID-19;
- For this data find the classical decomposition model

  $$X_t = m_t + s_t + Z_t$$

  as in this lecture, where m_t is a polynomial trend;
  - perform a smoothing of $\hat{m}_t$ by moving average $\hat{m}_t^{MA}$ or exponential smoothing $\hat{m}_t^{E}$;
  - find the estimator of seasonality;
  - using the least squared method find the final estimation of trend m_t in the polynomial form.

Final task on laboratory.

- Consider residuals from the model as a new time series $Z_t \approx \hat{Z}_t := X_t - \hat{m}_t - \hat{s}_t$ type ARMA(p,q)
  - using ACF and PACF find lags p,q;
  - for such p, q, using the maximal likelihood method find estimators of parameters $\hat{\phi}_1, \dots, \hat{\phi}_p$ and $\hat{\theta}_1, \dots, \hat{\theta}_q$;
  - Verify whether the residuals

$$\hat{\epsilon}_t = Z_t - \sum_{j=1}^{p} \hat{\phi}_j Z_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{\epsilon}_{t-j}$$

    form the white noise