Analýza časových řad

Pavla Holubíková

Struktura otázky

1. Stacionární procesy
2. Autokovarianční funkce a její vlastnosti
3. Derivace a integrál náhodného procesu
4. Spektrální rozklad autokovariančních funkcí stacionárních procesů
5. Odhady středních hodnot a autokovariancí stacionárních náhodných procesů
6. Regresní modely globálního a lokálního trendu

Znaky:

1. Data jsou sbírána postupně v čase.
2. Záleží na pořadí.
3. Pozorování jsou vzájemně závislá.

Cíle:

1. Porozumět datům; podle jakého modelu vznikají nové hodnoty.
2. Předpovídat budoucí hodnoty na základě minulých.

Náhodný proces

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $T \subseteq \mathbb{R}$ indexová množina a $X : \Omega \to \mathbb{R}$ reálná funkce definovaná $\forall \omega \in \Omega$ a $\forall t \in T$. Jestliže $\forall t \in T$ je $X(\omega, t)$ borelovsky měřitelná funkce vzhledem k $\mathcal{A}$, pak tuto funkci nazýváme náhodným procesem. Značíme $\{X(\omega, t) : \omega \in \Omega, t \in T\} = \{X_t, t \in T\}$.

Borelovsky měřitelná funkce ($\mathcal{B}$ σ-algebra borelovských podmnožin)

$$\forall B \in \mathcal{B}, \forall t \in T : X^{-1}(B) = \{\omega \in \Omega : X(\omega, t) \in B\} \in \mathcal{A}.$$

$T \subseteq \mathbb{Z} \to$ proces s diskrétním časem (časová řada)

$T \in [t_1, t_2], -\infty \leq t_1 < t_2 \leq \infty \to$ proces se spojitým časem

X_t nabývají diskrétních hodnot $\to$ proces s diskrétními stavy

X_t nabývají hodnot z nějakého intervalu $\to$ proces se spojitými stavy

Systém distribučních funkcí

Konečně dimenzionální distribuční funkcí rozumíme funkci $F_{t_1,\dots,t_k}(x_1,\dots,x_k) = P(X_{t_1} \le x_1,\dots,X_{t_k} \le x_k), \forall k \in \mathbb{N},\ t_1,\dots,t_k \in T$.
Systém distribučních funkcí $\{F_{t_1,\dots,t_k} : k \in \mathbb{N},\ t_1,\dots,t_k \in T\}$ s vlastnostmi

- $\lim_{x_{k+1}\to\infty} F_{t_1,\dots,t_k,t_{k+1}}(x_1,\dots,x_k,x_{k+1}) = F_{t_1,\dots,t_k}(x_1,\dots,x_k)$,
- $F_{t_{i_1},\dots,t_{i_k}}(x_{i_k},\dots,x_{i_k}) = F_{t_1,\dots,t_k}(x_1,\dots,x_k)$ pro lib. permutaci $(i_1,\dots,i_k)$ čísel $(1,\dots,k)$

nazveme *konzistentní.*

Daniell-Kolmogorova věta

Nechť $\{F_{t_1,\dots,t_k} : k \in \mathbb{N},\ t_1,\dots,t_k \in T\}$ je konzistentní. Pak existuje $\{X_t, t \in T\}$ takový, že $\forall k \in \mathbb{N},\ t_1,\dots,t_k \in T$ je sdružená distribuční funkce $(X_{t_1},\dots,X_{t_k})$ právě $F_{t_1,\dots,t_k}$.

Procesy, jejichž konečně dimenzionální distribuce jsou všechny mnohorozměrné normální, se nazývají *gaussovské.*

Střední hodnota, autokovarianční funkce, ACF

- *Střední hodnota* NP je funkce

$$\mu_t = \mathrm{E} X_t, \ t \in \mathbb{Z}.$$

- *Autokovarianční funkce* NP je funkce

$$\gamma(s,t) = \mathrm{cov}(X_s, X_t) = \mathrm{E}[(X_s - \mu_s)(X_t - \mu_t)], \ s, t \in \mathbb{Z}.$$

- *Autokorelační funkce (ACF)* NP je funkce

$$\rho(s,t) = \mathrm{cor}(X_s, X_t) = \frac{\mathrm{cov}(X_s, X_t)}{\sqrt{\mathrm{var}\, X_s \mathrm{var}\, X_t}}, \ s, t \in \mathbb{Z}.$$

Píšeme:

$$\gamma(h) := \gamma(t, t+h) = \gamma(0, h)$$

a obdobně pro $\rho(h)$.

Stacionární procesy

Řekneme, že $\{X_t, t \in T\}$ je *striktně stacionární*, pokud je sdružená distribuční funkce $(X_{t_1}, \ldots, X_{t_k})$ stejná jako sdružená distribuce $(X_{t_1+h}, \ldots, X_{t_k+h}), \forall k \in \mathbb{N}$, $t_1, \ldots, t_k \in T$, $h \in \mathbb{Z}$, tj. $F_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = F_{t_1+h,\ldots,t_k+h}(x_1, \ldots, x_k)$.

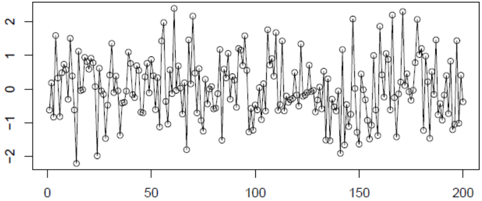
Řekneme, že $\{X_t, t \in T\}$ je *(slabě) stacionární*, pokud je μ_t konstantní v čase a $\gamma(s,t)$ závisí pouze na rozdílu $s - t$.

Striktní stacionarita $\Rightarrow$ *slabá stacionarita*

Pro gaussovský NP: *striktní stacionarita* $=$ *slabá stacionarita*

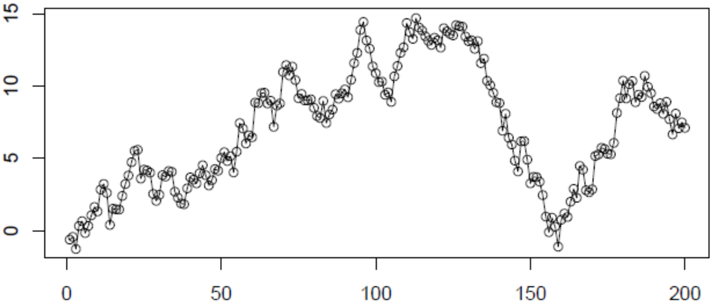
Příklady

- *Bílý šum*: $\{\varepsilon_t\} \sim WN(0, \sigma^2)$,
  kde $\{\varepsilon_t, t \in \mathbb{Z}\}$ je posloupnost náhodných veličin:
  $\mathrm{E}\varepsilon_t = 0$, $\mathrm{var}\varepsilon_t = \sigma^2$, $\mathrm{cov}(\varepsilon_s, \varepsilon_t) = 0$ pro $s \neq t$
  stacionární

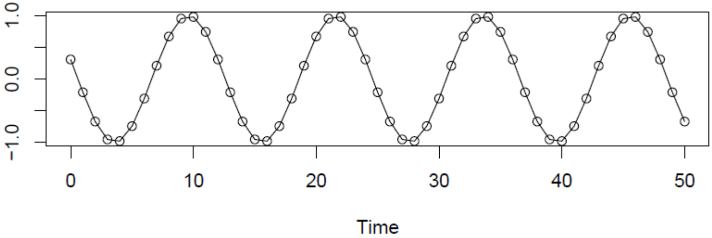


- *Klouzavé průměry*: $\{X_t, t \in \mathbb{Z}\}$,
  kde $X_t = (\varepsilon_t + \varepsilon_{t-1})/2$ pro $\{\varepsilon_t\} \sim WN(0, \sigma^2)$,
  stacionární, neboť $\mu_t = 0$,
  $\gamma(t, t) = 0.5\sigma^2$, $\gamma(t, t+1) = 0.25\sigma^2$, $\gamma(t, t+h) = 0$ pro $|h| > 1$.

- *Náhodná procházka*: $\{X_t, t \in \mathbb{Z}\}$,
  kde $X_t = \sum_{i=1}^{t} \varepsilon_i$ pro $\{\varepsilon_t\} \sim WN(0, \sigma^2)$,
  nestacionární, neboť $\mu_t = 0$, pro $s < t$
  $\gamma(t, t) = t\sigma^2$, $\gamma(s, t) = \min(s, t)\sigma^2$.



- *Náhodná kosinová vlna* $\{X_t, t \in \mathbb{Z}\}$, kde $X_t = a\cos[2\pi(ft + \phi)]$ pro amplitudu a, frekvenci f (např. $f = 1/12$), náhodnou fázi $\phi \in [0, 1]$ (tj. z nějakého rozdělení, $Rs([0, 1])$)
  stacionární, neboť $\mathrm{E}X_t = 0$,
  $\gamma(s, t) = \frac{a^2}{2}\cos[2\pi f(s - t)]$.



Vlastnosti autokovarianční funkce

$\gamma(s,t) = \text{cov}(X_s, X_t),\ s, t \in T$

- pro NP s konečnými druhými momenty je pozitivně semidefinitní: $\forall k \in \mathbb{N},\ t_1, \dots, t_k \in T,\ c_1, \dots, c_k \in \mathbb{R}$

$$0 \leq \text{var}\left(\sum_{i=1}^{k} c_i X_{t_i}\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} c_i c_j \gamma(t_i, t_j)$$

- Naopak: Pro pozitivně semidefinitní, konečnou fci g na $T \times T$ $\exists$ NP: g je autokovarianční fce
- Platí:

$$\gamma(t,t) = \gamma(0) \geq 0 \quad \text{a} \quad |\gamma(s,t)| \leq \sqrt{\gamma(t,t) \cdot \gamma(s,s)}$$

- $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

Spojitost NP

Řekneme, že $\{X_t, t \in T\}$ je *spojitý podle středu v bodě* $t_0 \in T$, jestliže při $t \to t_0$ konvergují X_t k X_{t_0} podle kvadratického středu, tj.

$$\mathrm{E}|X_t - X_{t_0}|^2 \to 0 \quad \text{pro} \quad t \to t_0.$$

Píšeme („limit in the mean“)

$$\underset{t \to t_0}{\mathrm{l.i.m.}}\, X_t = X_{t_0}.$$

Je-li $\{X_t, t \in T\}$ spojitý v každém bodě množiny T je *spojitý*.

Kritérium spojitosti

NP $\{X_t, t \in T\}$ je spojitý $\Leftrightarrow$ jeho $\gamma(s,t)$ je spojitá v bodech (s,t), pro něž $s = t$.

Derivace NP

Řekneme, že $\{X_t, t \in T\}$ má *v bodě* $t_0 \in T$ *derivaci* X'_{t_0}, jestliže pro $t_0 + h \in T$ platí

$$\underset{h \to 0}{\text{l.i.m.}} \frac{X_{t_{0+h}} - X_{t_0}}{h} = X'_{t_0}$$

Má-li NP derivaci ve všech bodech T říkáme, že NP má *derivaci*.

Integrál NP

Nechť $T = (a, b)$ je konečný nedegenerovaný interval, $[T_1, T_2] \subset T$, kde $T_1 < T_2$. Pomocí *dělení* D_n intervalu $[T_1, T_2]$ takového, že $T_1 = t_0 < t_1 < \cdots < t_n = T_2$ vytvoříme *součet* $I_n = \sum_{i=1}^{n} X_{t_i}(t_i - t_{i-1})$. Označme $\Delta_n = \max_{1\le i\le n}(t_i - t_{i-1})$ *normu dělení* D_n.

Jestliže posloupnost I_n konverguje podle kvadratického středu k náh. veličině I pro $n \to \infty$ pro lib. dělení D_n takové, že $\Delta_n \to 0$, pak se I nazývá *Riemannův integrál* $\{X_t, t \in T\}$ na $[T_1, T_2]$. Píšeme $I = \int_{T_1}^{T_2} X_t \,\mathrm{d}t$.

Kritérium existence integrálu

Nechť $\{X_t, t \in T\}$ je NP s $\mathrm{E}X_t = 0$, $\gamma(s, t)$ a konečnými 2. momenty. Pak R.I. $\int_{T_1}^{T_2} X_t \,\mathrm{d}t \,\exists \Leftrightarrow \exists$ R.I. $\int_{T_1}^{T_2}\int_{T_1}^{T_2} \gamma(s, t) \,\mathrm{d}s \,\mathrm{d}t$.

Spektrální rozklad $\gamma(t)$

Předpoklad: $\{X_t, t \in \mathbb{Z}\}$ stacionární, $\mathrm{E}X_t = 0$, konečné 2. momenty
$\Rightarrow \gamma(t)$ jako (nespočetný) součet harmonických fcí s různými frekvencemi a amplitudami

Herglotzova věta

Je-li $\{X_t, t \in \mathbb{Z}\}$ stacionární posloupnost, pak se dá její $\gamma(t)$ vyjádřit jako

$$\gamma(t) = \int_{-\pi}^{\pi} \mathrm{e}^{it\lambda}\,\mathrm{d}F(\lambda). \tag{1}$$

Bochnerova věta

Je-li $\{X_t, t \in \mathbb{R}\}$ stacionární NP a spojitý podle středu, pak se dá $\gamma(t)$ vyjádřit jako

$$\gamma(t) = \int_{-\infty}^{\infty} \mathrm{e}^{it\lambda}\,\mathrm{d}F(\lambda). \tag{2}$$

$F(\lambda)$ neklesající, zprava spojitá a jediná:

$$F(-\pi) = 0 \ \text{ a } \ F(\pi) = \gamma(0), \quad \text{příp.} \quad F(-\infty) = 0 \ \text{ a } \ F(\infty) = \gamma(0)$$

Spektrální rozklad, distribuční funkce a hustota

Vzorce (1) a (2) se nazývají *spektrální rozklad* $\gamma(t)$.
Fce $F(\lambda)$ se nazývá *spektrální distribuční funkce*.
Je-li $F(\lambda)$ absolutně spojitá, pak existuje *spektrální hustota* $f(\lambda)$: pro stacionární posloupnosti (NP) platí

$$F(\lambda) = \int_{-\pi}^{\lambda} f(x)\,\mathrm{d}x, \quad \text{příp.} \quad F(\lambda) = \int_{-\infty}^{\lambda} f(x)\,\mathrm{d}x$$

Spektrální hustota $f(\lambda)$ reálného spojitého stacionárního NP nebo reálné stacionární posloupnosti je sudá fce v tom smyslu, že pro ni platí $f(\lambda) = f(-\lambda)$ s.v. vzhledem k Lebesgueově míře.

Existence spektrální hustoty

K existenci $f(\lambda)$ stacionární posloupnosti stačí, aby platilo

$$\sum_{t=-\infty}^{\infty} |\gamma(t)| < \infty.$$

K existenci $f(\lambda)$ stacionárního NP stačí, aby platilo

$$\int_{-\infty}^{\infty} |\gamma(t)| \, \mathrm{d}t < \infty.$$

Existuje-li $f(\lambda)$ stacionární posloupnosti a má-li konečnou variaci na $[-\pi, \pi]$, pak $f(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \mathrm{e}^{-it\lambda} \gamma(t)$ s.v. vzhledem k Lebesgueově míře (tj. ve všech bodech spojitosti).

Existuje-li $f(\lambda)$ spojitého stacionárního NP a je-li $\int_{-\infty}^{\infty} |\gamma(t)| \, \mathrm{d}t < \infty$, pak $f(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-it\lambda} \gamma(t) \mathrm{d}t$.

Odhady středních hodnot stacionárních NP

Předpokládejme, že $\mathrm{E}X_t = \mu$, $\forall t \in T$. Jako odhad zvolíme výběrový průměr

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t.$$

- nestranný odhad: $\mathrm{E}\hat{\mu} = \mathrm{E}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) = \mu$
- míra kvality odhadu pomocí rozptylu

$$\begin{aligned}\mathrm{var}\,\hat{\mu} = \mathrm{var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) &= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{i=1}^{n}\mathrm{cov}(X_t, X_i) = \\ &= \frac{1}{n^2}\{n\gamma(0) + 2(n-1)\gamma(1) + \cdots + 2\gamma(n-1)\} \\ &= \frac{\gamma(0)}{n}\left\{1 + 2\sum_{i=1}^{n-1}\frac{n-i}{n}\rho(i)\right\}\end{aligned}$$

(WN: $\mathrm{var}\,\hat{\mu} = \frac{\gamma(0)}{n}$ $\Rightarrow$ $2\sum_{i=1}^{n-1}\frac{n-i}{n}\rho(i)$ korekční člen, odráží vliv autokorelace)

Konzistentnost

Nechť $\{X_t, t \in T\}$ je stacionární s μ a $\gamma(h) \to 0$ pro $h \to \infty$. Pak $\mathrm{var}\overline{X}_n \to 0$ pro $n \to \infty$.

Aproximované rozdělení

Nechť $\{X_t, t \in T\}$ je stacionární. Pak za určitých podmínek má $\overline{X}_n$ přibližně (pro velká n) normální rozdělení:

$$\overline{X}_n \xrightarrow{d} N\left(\mu, \frac{1}{n}\sum_{h-\infty}^{\infty} \gamma(h)\right)$$

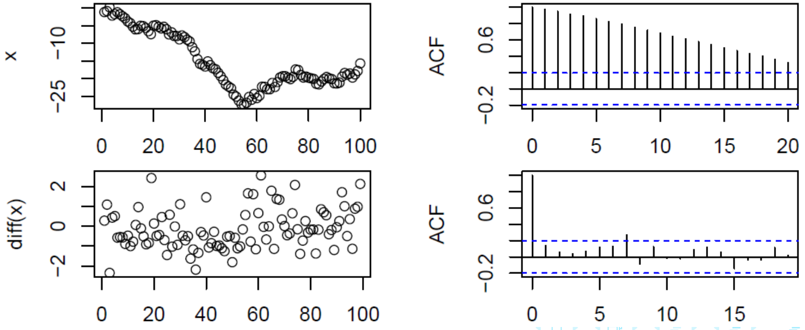
Odhady autokovariance stacionárních NP

Odhad $\gamma(h) = \mathrm{E}(X_t - \mu)(X_{t+h} - \mu)$ (příp. $n-h$ namísto n):

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-h} (X_i - \overline{X})(X_{i+h} - \overline{X}).$$

Odhad $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$:

$$\hat{\rho}(h) = \frac{\gamma(\hat{h})}{\gamma(\hat{0})} = \frac{\sum_{i=1}^{n-h} (X_i - \overline{X})(X_{i+h} - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})}.$$



Ozn. $\hat{c}(h) = (\hat{\gamma}(0), \ldots, \hat{\gamma}(h))'$ a $\hat{r}(h) = (\hat{\rho}(1), \ldots, \hat{\rho}(h))'$

Je-li $\{X_t, t \in T\}$ stacionární s konečnými 4. momenty, pak má $\hat{c}(h)$ za určitých podmínek pro lib. pevné h přibližně (pro $n \to \infty$) normální rozdělení:

$$\sqrt{n}\,(\hat{c}(h) - c(h)) \xrightarrow{d} N_{h+1}(0, V).$$

Mnohdy lepší studovat závislost bez ohledu na měřítko $\to$ ACF ($\rho()$= fce $\gamma()$ $\Rightarrow$ *delta metoda*)

Je-li $\{X_t, t \in T\}$ stacionární s konečnými 4. momenty, pak má $\hat{r}(h)$ za určitých podmínek pro lib. pevné h přibližně (pro $n \to \infty$) normální rozdělení:

$$\sqrt{n}\,(\hat{r}(h) - r(h)) \xrightarrow{d} N_h(0, W),$$

kde $W = \{w_{ij}\}_{i,j=1,\ldots,h}$ a

$$w_{ij} = \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) + 2\rho(i)\rho(k)\}\{\rho(k+j) + \rho(k-j) + 2\rho(j)\rho(k)\}.$$

Časové řady

Popisujeme pomocí

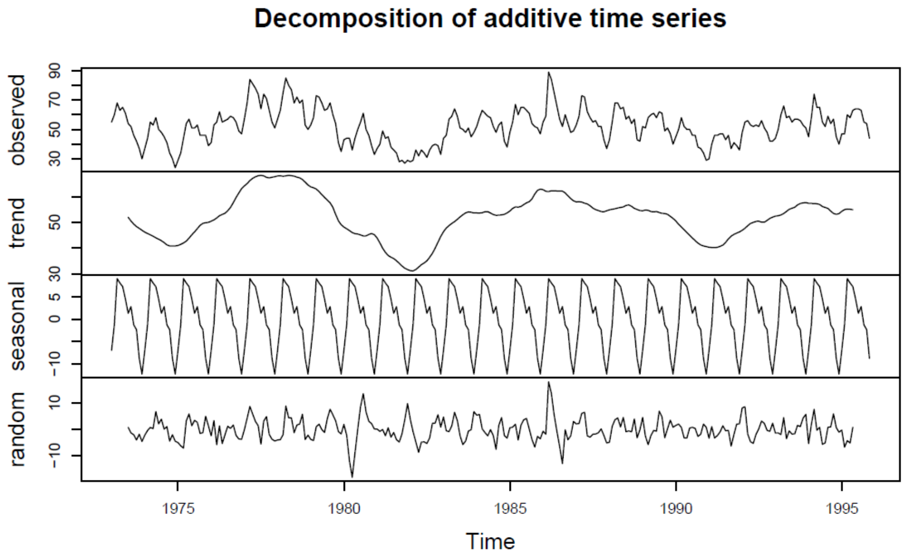
$$\begin{aligned} X_t &= \mathrm{DeterministickaCast}_t + \mathrm{StochastickaCast}_t \\ &= \{\mathrm{Trend}_t + \mathrm{Sezonnost}_t\} + \{\mathrm{PredikovatelnaVariabilita}_t + \varepsilon_t\} \\ &= T_t + S_t + E_t \end{aligned}$$

tj. *aditivní model* (příp. $X_t = T_t \cdot S_t \cdot E_t$ *multiplikativní*, který lze převést na aditivní zlogaritmováním)

- Trend – odráží dlouhodobé působení vlivů
- Sezonnost – popisuje periodické změny (roční období. . .)
- Náhodné fluktuace – modeluje vlivy, které působí nepravidelně

Klasická dekompozice časových řad

- tj. rozklad $X_t = T_t + S_t + E_t$ (v případě aditivního modelu)
- založeno na regresní analýze
- T_t se mění v čase flexibilně, S_t konstantní
- Postup:
  1. Odhad T_t
  2. „Odtrendování“ $X_t - T_t$
  3. Odhad sezonnosti S_t z $X_t - T_t$
  4. Výpočet reziduální složky $E_t = X_t - T_t - S_t$
- Pro odhady složek se využívají metody:
  - parametrické – lineární modely, harmonické LRM
  - neparametrické – metoda klouzavých průměrů, jádrové vyhlazování
- Jiné metody – STL dekompozice, exponenciální vyhlazování, Holt-Winter (včetně predikce)


Decomposition of additive time series
observed
trend
seasonal
random
90
70
50
30
50
30
5
0
−10
10
−10
1975
1980
1985
1990
1995
Time

Regresní modely můžeme dělit na modely:

- *globálního trendu* – využití všech $t \in T$
- *lokálního trendu* – volba určitého počtu $t \in T$, tj. intervalu

Uvažujme $X_t = \mu_t + U_t$, kde $\mu_t = \mathrm{E} X_t$ a rezidua $U_t : \mathrm{E}\, U_t = 0$.
Trend můžeme uvažovat jako polynom (LRM), v praxi buď:

- lineární $\mu_t = \beta_0 + \beta_1 t$, nebo
- kvadratický $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$.

Vhodné je zahrnout i sezonní složku (s sezón, $k \in \mathbb{N}_0$)

$$\mu_t = \beta_0 + \beta_1 t + \sum_{j=2}^{s} \alpha_j 1_{[t=ks+j]}.$$

Metoda klouzavých průměrů

Uvažujme $X_t = T_t + E_t$.
Trend můžeme modelovat také neparametricky pomocí klouzavých průměrů:

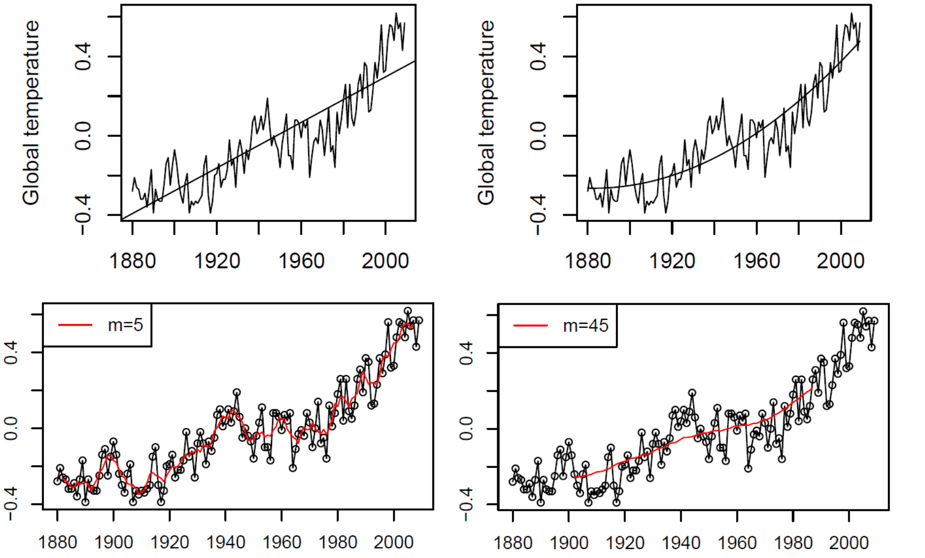
- tzn. průměrujeme hodnoty vždy na určitém intervalu
- řádu $m = 2k + 1$

$$\widehat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} X_{t+j}$$

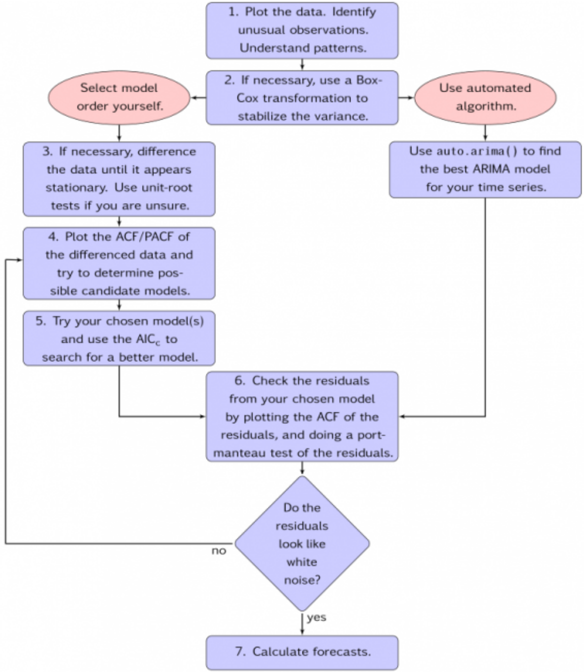
- řádu $m = 2k$

$$\widehat{T}_t = \frac{1}{2m} X_{t-k} + \frac{1}{m} X_{t-k+1} + \cdots + \frac{1}{m} X_{t+k-1} + \frac{1}{2m} X_{t+k}$$

- s rostoucím m se odhadnutá křivka vyhlazuje! (tj. šířka vyhlazovacího okénka)



Global temperature
0.4
0.0
−0.4
1880
1920
1960
2000
Global temperature
0.4
0.0
−0.4
1880
1920
1960
2000
m=5
0.4
0.0
−0.4
1880
1900
1920
1940
1960
1980
2000
m=45
0.4
0.0
−0.4
1880
1900
1920
1940
1960
1980
2000


1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AIC_c to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

ARMA modely

- Box-Jenkinsonova metodologie
- pro stacionární NP
- spojení
  - autoregresních NP (AR) a
  - NP klouzavých průměrů (MA) tj. autoregressive–moving-average models
- ARMA(p, q) : $X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$
- integrovaná ARMA = ARIMA, tj. ARMA(p, d, q)
- sezonní ARMA = SARMA, tj. ARMA$(p, q)(P, Q)_s$
- SARIMA, tj. ARMA$(p, d, q)(P, D, Q)_s$
  Např. ARMA$(0, 1, 1)(0, 1, 1)_{12}$

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_1^* \varepsilon_{t-12} + \theta_1 \theta_1^* \varepsilon_{t-13}$$

Literatura

- Přednášky: M9121 Časové řady I
- Pravděpodobně i přednášky M0122 Časové řady II ¯_(ツ)_/¯
- Skripta: https://www.math.muni.cz/~vondra/uvm/vystupy/KA1/M5201/M5201.pdf