



**MATEMATICKO-FYZIKÁLNÍ
FAKULTA**
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Michal Lain

Robustní odhady autokorelační funkce

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Šárka Hudecová, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční a pojistná matematika

Praha 2020

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Rád bych poděkoval vedoucí diplomové práce RNDr. Šárce Hudecové, Ph.D. za odborné vedení, cenné rady, vstřícnost a čas, které mi v průběhu zpracování diplomové práce věnovala.

Název práce: Robustní odhady autokorelační funkce

Autor: Bc. Michal Lain

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Autokorelační funkce je základním nástrojem zkoumání časových řad. Její klasický odhad je velmi náchylný na výskyt odlehlých pozorování, což může vést k zavádějícím výsledkům. Tato práce se zabývá robustními odhady autokorelační funkce, které jsou odolnější vůči odlehlým pozorováním než klasický odhad. Jsou zde uvedeny následující přístupy: metoda vynechání odlehlých pozorování z dat, nahrazení průměru mediánem, transformace dat, odhad jiného koeficientu, robustní odhad parciální autokorelační funkce či lineární regrese. Práce popisuje jejich použití, výhody a nevýhody a nutné předpoklady. Představené metody jsou také detailně porovnány v simulační studii. Práce obsahuje mimo jiné i aplikaci na reálná data z finanční oblasti.

Klíčová slova: autokorelační funkce, stacionární časová řada, odlehlá pozorování, robustní odhad

Title: Robust estimation of autocorrelation function

Author: Bc. Michal Lain

Department: Name of the department

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of probability and mathematical statistics

Abstract: The autocorrelation function is a basic tool for time series analysis. The classical estimation is very sensitive to outliers and can lead to misleading results. This thesis deals with robust estimations of the autocorrelation function, which is more resistant to the outliers than the classical estimation. There are presented following approaches: leaving out the outliers from the data, replacement the average with the median, data transformation, the estimation of another coefficient, robust estimation of the partial autocorrelation function or linear regression. The thesis describes the applicability of the presented methods, their advantages and disadvantages and necessary assumptions. All the approaches are compared in simulation study and applied to real financial data.

Keywords: autocorrelation function, stationary time serie, outliers, robust estimation

Obsah

Úvod	3
1 Základní pojmy	5
1.1 Autokorelační funkce	5
1.2 Parciální autokorelační funkce	8
1.3 Třídy časových řad	10
2 Klasický odhad autokorelační funkce	17
2.1 Metoda výpočtu a vlastnosti	17
2.2 Nevýhody klasického odhadu	21
3 Robustní statistika	27
3.1 Robustní odhad střední hodnoty	28
3.2 Robustní odhad polohy	29
3.3 Robustní odhad charakteristiky variability	31
3.4 Robustní statistika pro lineární regresi	33
3.5 Struktura odlehlosti pro časové řady	35
3.6 Míra robustnosti	39
4 Robustní odhad autokorelační funkce	40
4.1 Vynechávací metoda	40
4.1.1 Kroková vynechávací metoda	43
4.1.2 Odhad vynechávacího parametru	46
4.2 Mediánová metoda	47
4.3 Transformační metoda	49
4.4 Koeficientová metoda	50
4.5 Metoda odhadu pomocí robustního odhadu parciální autokorelační funkce	52
4.6 Regresní metoda	55
4.7 Porušení vlastností	60
4.7.1 Omezenost odhadu	60
4.7.2 Pozitivní semidefinitnost	62
5 Aplikace	64
5.1 Simulace	64
5.1.1 Aditivní symetrická odlehlá pozorování	64
5.1.2 Aditivní nesymetrická odlehlá pozorování	72
5.1.3 Inovační symetrická odlehlá pozorování	78
5.1.4 Inovační nesymetrická odlehlá pozorování	79
5.1.5 Souhrn simulací pro AR(1)	84
5.1.6 Souhrn simulací pro MA(4)	87
5.2 Reálná data	88
Závěr	91
Seznam použité literatury	93

Seznam obrázků	95
Seznam tabulek	97
Seznam použitých zkratk	98
A Dodatek	99
A.1 Yuleovy-Walkerovy rovnice	99
A.2 Příklady měř robustnosti	100
A.2.1 Influenční funkce	100
A.2.2 Bod selhání	101
A.2.3 Asymptotické vychýlení	101
A.2.4 Eficiency	102
A.2.5 Posouzení odlehlosti	102
A.3 Koeficienty korelace	103

Úvod

Autokorelační funkce je základním nástrojem pro analýzu časových řad. Její klasický odhad je velmi náchylný na odlehlá pozorování, což může vést k zavádějícím výsledkům. Tématem této práce je popsat a porovnat různé robustní odhady autokorelační funkce, které zmírní nebo zcela eliminují vliv odlehlých pozorování. V práci se budeme zabývat stacionárními časovými řadami. Vycházíme z článku (Dürre a kol., 2015), který se zabývá přehledem robustních metod pro odhad autokorelační funkce. Popíšeme některé z těchto metod a podíváme se podrobněji na chování a vlastnosti dle původních zdrojů. Doplníme také metody navržené v jiných zdrojích a jednu vlastní. Názorně ukážeme nevýhody klasického odhadu a uvedeme robustní základy, ze kterých některé metody vycházejí.

Nejdříve v první kapitole definujeme základní pojmy a vlastnosti související s autokorelační funkcí a časovými řadami. Zavedeme parciální autokorelační funkci a její souvislost s autokorelační funkcí. Poté uvedeme základní třídy časových řad a výpočet a tvar jejich autokorelačních funkcí.

V druhé kapitole zavedeme vzorec pro klasický odhad autokorelační funkce a odvodíme jeho základní vlastnosti. Poté se budeme zabývat odhadem parciální autokorelační funkce a jeho použitím pro odhad autokorelační funkce. Odvodíme limitní vliv odlehlého pozorování na klasický odhad a podložíme jej praktickými příklady.

Práce se zabývá použitím robustního odhadu, proto ve třetí kapitole uvedeme základy robustní statistiky, které poté budeme potřebovat. Definujeme si různé typy odlehlých pozorování. Zmíníme robustní odhady základních charakteristik časové řady. Nakonec zavedeme robustní odhad pro lineární regresi, kterou budeme v práci dále využívat.

Ve čtvrté kapitole popíšeme několik robustních metod, které k problému odlehlých pozorování budou přistupovat z různých úhlů pohledu. První bude *vynechávací metoda*, která vliv odlehlých pozorování eliminuje jejich vynecháním pro výpočet odhadu. Poté navrhne vlastní *krokovou vynechávací metodu*, která řeší problém s nesymetrickými odlehlými pozorováními. Nejsme si vědomi, že by daná metoda existovala v literatuře. Další je *mediánová metoda*, která nahrazuje průměr mediánem při výpočtu odhadu. Vliv odlehlých pozorování můžeme zmírnit transformací dat, ze které vychází *transformační metoda*. Odhad autokorelační funkce lze získat transformací odhadů jiných koeficientů, které jsou robustnější. Toho využijeme u *koeficientové metody*. Předposlední metodou je *odhad autokorelační funkce pomocí robustního odhadu parciální autokorelační funkce*. Poslední je *regresní metoda*, kde pomocí robustního odhadu směrnice regresní přímky získáme robustní odhad autokorelační funkce. Dále se budeme zabývat porušením některých obecných vlastností klasického odhadu, které mohou být robustní metodou porušeny. Všechny metody budou aplikovány na stejnou časovou řadu.

V závěru práce v páté kapitole odvozené metody mezi sebou porovnáme na základě simulací. Zohledníme různé typy odlehlých pozorování. Nakonec metody aplikujeme na reálná data.

V práci čerpáme jak z již zmíněného článku (Dürre a kol., 2015), tak z celé řady dalších odborných prací jako (Chang a Politis, 2014), (Chakhchoukh, 2010), (Mottonen a kol., 1999), (Chan a Wei, 1992) a další. Některé z těchto publikací

obsahovaly drobné překlepy nebo nepřesné značení, což uvádíme na pravou míru.

Přínos této práce vidíme především v popsaní různých robustních metod pro odhad autokorelační funkce a v jejich srovnání pomocí simulací. Také navrheme vlastní robustní metodu. Názorně ukážeme a popíšeme nevýhody klasického odhadu a smysl robustních metod.

1. Základní pojmy

V první kapitole se budeme zabývat zavedením potřebného aparátu, definujeme autokorelační funkci a parciální autokorelační funkci a pojmy s nimi související. Podíváme se na typy stacionarit časové řady. Uvedeme, jak přistupovat k časovým řadám, a jejich základní třídy.

Čerpáme z publikací (Cipra, 2008), (Prášková, 2004), (Brockwell a Davis, 2002) a (Dupač a Hušková, 2013).

1.1 Autokorelační funkce

Autokorelační funkce je základním nástrojem zkoumání časových řad. Dává nám určitou představu o struktuře závislosti mezi pozorováními různě vzdálenými v čase od sebe. Slouží mimo jiné jako užitečný nástroj pro rozpoznání vhodného modelu časové řady.

Pracujeme na *pravděpodobnostním prostoru*, který je běžně reprezentovaný trojicí (Ω, \mathcal{A}, P) . (Ω, \mathcal{A}) je měřitelný prostor, Ω je množina elementárních jevů, \mathcal{A} je příslušná σ -algebra a P je pravděpodobnostní míra na \mathcal{A} .

Definice 1 (Náhodný proces a časová řada). *Nechť (Ω, \mathcal{A}, P) je pravděpodobnostní prostor, $T \subset \mathbb{R}$ a (S, \mathcal{E}) je měřitelný prostor. Rodina náhodných veličin $\{X_t, t \in T\}$ definovaných na (Ω, \mathcal{A}, P) s hodnotami v S se nazývá náhodný proces. Pokud $T \subseteq \mathbb{Z}$ a náhodný proces je chronologicky uspořádaný v čase, pak mluvíme o časové řadě.*

Pomocí časových řad lze popsat různorodá data: roční míru nezaměstnanosti, měsíční prodejnost produktu či denní kurz akcií. Důležitou roli u časových řad hraje frekvence pozorování. Zda data pozorujeme ročně, jako je tomu například u míry nezaměstnanosti, nebo každou minutu, jako u kurzu akcií.

Realizací časové řady je *trajektorie*. Jedná se o funkci času $\{X_t(\omega), t \in T\}$ pro určitý elementární jev $\omega \in \Omega$. My pozorujeme právě *trajektorii* dané řady, která je značena jako $\{x_t, t \in T\}$.

V této práci se budeme zabývat reálnými časovými řadami, tedy $S = \mathbb{R}$. To, co definujeme či odvodíme pro časové řady, vychází z vlastností náhodných procesů. Časová řada je pouze specifickým náhodným procesem. Nyní se podíváme na střední hodnotu časové řady a jak se dle jejích vlastností řady dělí.

Definice 2 (Střední hodnota). *Nechť $\{X_t, t \in T\}$ je časová řada a $\mu_t = EX_t$ je střední hodnota řady v čase $t \in T$. Časová řada má konstantní střední hodnotu, pokud pro každé $t \in T$ platí $\mu_t = \mu \in \mathbb{R}$. Řadu nazveme centrovanou, pokud pro každé $t \in T$ platí $\mu_t = 0$.*

Definujeme *gaussovskou časovou řadu*.

Definice 3 (Gaussovská časová řada). *Časová řada $\{X_t, t \in T\}$ je gaussovská, pokud pro každé $n \in \mathbb{N}$ a $t_1, \dots, t_n \in T$ má vektor $(X_{t_1}, \dots, X_{t_n})$ n -rozměrné normální rozdělení.*

Potřebným nástrojem bude *uspořádaná verze* časové řady.

Definice 4 (Uspořádaná verze). *Nechť $\{X_t, t = 1, \dots, n\}$ je časová řada. Posloupanost $X_{(1)}, \dots, X_{(n)}$ udává její uspořádanou verzi, pokud splňuje*

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n-1)} \leq X_{(n)}.$$

$X_{(k)}$ udává k -tou pořádkovou statistiku a R_i je pořadí X_i v uspořádané verzi.

Definujeme *autokovarianční a autokorelační funkci*.

Definice 5 (Autokovarianční a autokorelační funkce). *Nechť $\{X_t, t \in T\}$ je časová řada se střední hodnotou $\mu_t = EX_t$ v čase $t \in T$, pro kterou platí, že pro každé $t \in T$ je $E|X_t|^2 < \infty$ (konečné druhé momenty). Funkce dvou proměnných definována na $T \times T$ předpisem*

$$\gamma(s, t) = \text{cov}(X_s, X_t) = E(X_s - \mu_s)(X_t - \mu_t), \quad s, t \in T$$

se nazývá autokovarianční funkce časové řady $\{X_t, t \in T\}$. Hodnota $\gamma(s, s)$ se nazývá rozptyl časové řady v čase s . Nechť dále $\{X_t, t \in T\}$ má kladný rozptyl pro každé $t \in T$. Funkce dvou proměnných definována na $T \times T$ předpisem

$$\rho(s, t) = \text{corr}(X_s, X_t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)}\sqrt{\gamma(t, t)}}, \quad s, t \in T$$

se nazývá autokorelační funkce časové řady $\{X_t, t \in T\}$.

Autokorelační funkce je normovanou verzí autokovarianční funkce.

Důležitou vlastností časové řady je *stacionarita*, rozlišujeme *striktní a slabou stacionaritu*. Právě slabá stacionarita hraje velmi důležitou roli při práci s autokorelačními funkcemi.

Definice 6 (Strikt ní stacionarita). *Časová řada $\{X_t, t \in T\}$ je striktně stacionární, pokud pro každé $n \in \mathbb{N}$, každé $x_1, \dots, x_n \in \mathbb{R}$, každé t_1, \dots, t_n a $h \in \mathbb{R}$ taková, že pro $k \in \{1, \dots, n\}$ je splněno $t_k \in T, t_k + h \in T$, platí*

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = F_{t_1+h, \dots, t_n+h}(x_1, \dots, x_n),$$

kde značíme F_{t_1, \dots, t_n} jako distribuční funkci sdruženého rozdělení náhodného vektoru $(X_{t_1}, \dots, X_{t_n})^\top$.

Tedy všechny náhodné veličiny X_t striktně stacionární časové řady mají stejné rozdělení a charakteristiky. Nemění se vzhledem ke změně času. Dále definujeme *slabou stacionaritu*.

Definice 7 (Slabá stacionarita). *Časová řada $\{X_t, t \in T\}$ je slabě stacionární, pokud pro každé $t \in T$ platí $E|X_t|^2 < \infty$ (konečné druhé momenty), má konstantní střední hodnotu, tedy pro každé $t \in T$ platí $\mu_t = \mu \in \mathbb{R}$, a její autokovarianční funkce $\gamma(s, t)$ je funkcí pouze rozdílu $s - t$.*

Dalším typem stacionarity je *kovarianční*.

Definice 8 (Kovarianční stacionarita). *Časová řada $\{X_t, t \in T\}$ je kovariančně stacionární, pokud pro každé $t \in T$ platí $E|X_t|^2 < \infty$ (konečné druhé momenty) a její autokovarianční funkce $\gamma(s, t)$ je funkcí pouze rozdílu $s - t$.*

Platí určité vztahy mezi striktní, slabou a kovarianční stacionaritou. Pokud je časová řada striktně stacionární s konečnými druhými momenty, pak je i slabě stacionární. Pokud je řada slabě stacionární a gaussovská, pak je i striktně stacionární. Každá slabě stacionární řada je současně kovariančně stacionární. Podrobnější popis nalezneme například v publikaci (Prášková, 2004).

Odvodíme vzorec pro autokorelační funkci slabě stacionární řady, který bude v souladu s definicí 5.

Věta 1 (Autokovarianční a autokorelační funkce slabě stacionární časové řady). *Nechť časová řada $\{X_t, t \in T\}$ je slabě stacionární. Pak její autokovarianční funkci $\gamma(s,t)$ lze chápat jako funkci jedné proměnné dle předpisu*

$$\gamma(t) := \gamma(t,0), \quad t \in T.$$

Autokorelační funkce $\rho(s,t)$ je dána předpisem

$$\rho(t) = \frac{\gamma(t)}{\gamma(0)}.$$

Důkaz. Z definice 7 slabé stacionarity autokovarianční funkce $\gamma(u,v)$ závisí pouze na rozdílu $u - v$. Pak existuje $t \in \mathbb{R}$, že platí $u - v = t$. Protože

$$u - v = t = t - 0,$$

platí $\gamma(u,v) = \gamma(t,0)$. Z předpokladu konstatní střední hodnoty dostáváme

$$\gamma(t) = \gamma(t,0) = E(X_t - \mu)(X_0 - \mu).$$

Vzorec pro autokorelační funkci dostáváme přímo z její definice a z odvození jednorozměrné autokovarianční funkce. □

V případě slabě stacionárních řad nám splývá značení pro autokovarianční funkci jako pro funkci jedné a dvou proměnných. O kolika proměnnou funkci se jedná, chápeme z kontextu. Obdobně pro autokorelační funkci. Poznamenejme, že někteří autoři zavádějí pro autokovarianční funkci jako funkci jedné proměnné například značení $\tilde{\gamma}(t)$. V této práci se budeme zabývat pouze slabě stacionárními časovými řadami, proto pracujeme s autokorelační funkcí výhradně jako s funkcí jedné proměnné. Nyní se podíváme na základní vlastnosti autokovarianční funkce.

Věta 2 (Základní vlastnosti). *Nechť $\{X_t, t \in T\}$ je slabě stacionární časová řada s autokovarianční funkcí $\gamma(t)$. Pak platí*

$$\begin{aligned} \gamma(0) &\geq 0, \\ \forall t \in T : |\gamma(t)| &\leq \gamma(0), \\ \forall t \in T : \gamma(-t) &= \gamma(t). \end{aligned}$$

Důkaz. První nerovnost plyne z vlastnosti rozptylu časové řady. Druhá nerovnost je důsledkem Cauchyho-Schwarzovy nerovnosti. Třetí nerovnost plyne ze symetrie kovariance, tedy $\text{cov}(X_t, X_{t+k}) = \text{cov}(X_{t+k}, X_t)$. □

Autokorelační funkce nabývá hodnot z intervalu $[-1,1]$, což plyne z věty 2. Velmi důležitou vlastností autokorelační funkce je *pozitivní semidefinitnost*.

Věta 3 (Pozitivní semidefinitnost). *Nechť $\{X_t, t \in T\}$ je slabě stacionární časová řada. Pak její autokovarianční funkce $\gamma(t)$ a autokorelační funkce $\rho(t)$ jsou pozitivně semidefinitní, tedy pro každé $n \in \mathbb{N}$, každé $c_1, \dots, c_n \in \mathbb{R}$, každé $t_1, \dots, t_n \in T$ taková, že $t_i - t_j \in T$, platí*

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j \gamma(t_i - t_j) \geq 0.$$

Obdobně pro autokorelační funkci $\rho(t)$.

Důkaz. Viz (Prášková, 2004, věta 1.6). □

Označme si pro každé $k \in \mathbb{N}$ matici

$$\mathbf{\Gamma}^{(k)} = \left(\Gamma_{i,j}^{(k)} \right)_{i,j=0,\dots,k-1}, \quad (1.1)$$

kteřá má na pozici (i, j) prvek $\gamma(i - j)$. Tato matice je dle věty 3 také pozitivně semidefinitní.

1.2 Parciální autokorelační funkce

Definujeme *parciální autokorelační funkci*, která je dalším užitečným nástrojem k popisu a zkoumání časové řady.

Definice 9 (Parciální autokorelační funkce). *Nechť $\{X_t, t \in T\}$ je reálná slabě stacionární časová řada. Parciální autokorelační funkce časové řady $\{X_t, t \in T\}$ je definována pro každé $t \in T$ předpisem*

$$\pi(t) = \begin{cases} \text{corr}(X_k, X_{k+1}), & t = 1, \\ \text{corr}(X_k - \tilde{X}_k, X_{k-t} - \tilde{X}_{k-t}), & t > 1, \end{cases}$$

kde

$$\begin{aligned} \tilde{X}_k &= \pi_{t-1,0} + \sum_{l=1}^{t-1} \pi_{t-1,l} X_{k-l}, \\ \tilde{X}_{k-t} &= \tilde{\pi}_{t-1,0} + \sum_{l=1}^{t-1} \tilde{\pi}_{t-1,l} X_{k-t+l}, \end{aligned}$$

kde $\pi_{t-1,0}, \dots, \pi_{t-1,t-1} \in \mathbb{R}$ jsou konstanty určené podmínkami

$$E(X_k - \tilde{X}_k)X_l = 0, \quad l = k - t + 1, \dots, k - 1.$$

Totéž platí pro \tilde{X}_{k-t} .

Tedy \tilde{X}_k je nejlepší lineární predikcí X_k pomocí $X_{k-1}, \dots, X_{k-t+1}$, 1 obdobně pro \tilde{X}_{k-t} . Označme

$$\begin{aligned} U_{t,k} &= X_k - \tilde{X}_k, \\ V_{t,k} &= X_{k-t} - \tilde{X}_{k-t}. \end{aligned}$$

Pak parciální autokorelační funkci lze přepsat do tvaru

$$\pi(t) = \begin{cases} \text{corr}(X_k, X_{k+1}), & t = 1, \\ \text{corr}(U_{t,k}, V_{t,k}), & t > 1. \end{cases}$$

Parciální autokorelační funkce udává korelaci mezi X_k a X_{k-t} po odstranění lineárního efektu pozorování $X_{k-1}, \dots, X_{k-t+1}$ a 1. V případě centrované časové řady se jedná pouze o efekt $X_{k-1}, \dots, X_{k-t+1}$ a koeficient $\pi_{t-1,0}$ se neuplatňuje.

Podíváme se na vztahy mezi autokorelační a parciální autokorelační funkcí.

Věta 4 (Vztah autokovarianční a parciální autokorelační funkce). *Nechť $\{X_t, t \in T\}$ je slabě stacionární časová řada s autokovarianční funkcí $\gamma(t)$ a parciální autokorelační funkcí $\pi(t)$. Pak platí*

$$\begin{aligned} \pi(0) &= 1 \\ \pi(t) &= \phi_{tt}, \quad t \geq 1, \end{aligned}$$

kde ϕ_{tt} je řešením

$$\mathbf{\Gamma}^{(t)} (\phi_{t1}, \dots, \phi_{tt})^\top = \boldsymbol{\gamma}^{(t)},$$

matice $\mathbf{\Gamma}^{(t)}$ je definována v (1.1), $\boldsymbol{\gamma}^{(t)} = (\gamma(1), \dots, \gamma(t))^\top$ a platí $\phi_{tk} = \pi_{t,k}$ pro $k = 1, \dots, t$.

Důkaz. Viz (Brockwell a kol., 1991, kapitola 3.4 a 5.1). □

Další vztah plyne z *Durbinova-Levinsonova algoritmu*, který nalezneme například v (Brockwell a Davis, 2002), (Dürre a kol., 2015) či (Masarotto, 1987). Tento vztah bude dále použit pro robustní odhad autokorelační funkce.

Věta 5 (Vztah autokorelační a parciální autokorelační funkce dle Durbinova-Levinsonova algoritmu). *Nechť $\{X_t, t \in T\}$ je centrovaná slabě stacionární časová řada s autokovarianční funkcí $\gamma(t)$ splňující $\lim_{t \rightarrow \infty} \gamma(t) = 0$ a $\gamma(0) > 0$ a parciální autokorelační funkcí $\pi(t)$. Pak platí vztah pro autokorelační a parciální autokorelační funkci dle Durbinova-Levinsonova algoritmu, rekurze je dána předpisem*

$$\pi(t) = \left(\rho(t) - \sum_{k=1}^{t-1} \pi_{t-1,k} \rho(t-k) \right) \nu_{t-1}^{-1}, \quad t \geq 2, \quad (1.2)$$

kde

$$\begin{pmatrix} \pi_{t,1} \\ \vdots \\ \pi_{t,t-1} \end{pmatrix} = \begin{pmatrix} \pi_{t-1,1} \\ \vdots \\ \pi_{t-1,t-1} \end{pmatrix} - \pi(t) \begin{pmatrix} \pi_{t-1,t-1} \\ \vdots \\ \pi_{t-1,1} \end{pmatrix}$$

$$\nu_t = \nu_{t-1} (1 - \pi(t)^2)$$

s počátečním nastavením $\pi(1) = \rho(1)$ a $\nu_0 = 1$ a platí $\pi_{t-1,t-1} = \pi(t-1)$.

Pro autokorelační funkci $\rho(t)$ platí vztah

$$\rho(t) = \sum_{k=1}^{t-1} \pi_{t-1,k} \rho(t-k) + \pi(t) \left(1 - \sum_{k=1}^{t-1} \pi_{t-1,k} \rho(k) \right). \quad (1.3)$$

Důkaz. Viz (Brockwell a Davis, 2002, kapitola 2.5.1) pro vztah (1.2). (Anderson, 1994, kapitola 5.4) a (Masarotto, 1987) pro vztah (1.3). □

1.3 Třídy časových řad

Podíváme se na třídy a strukturu časových řad. Vhodný postup práce s časovou řadou závisí na typu časové řady a účelu naší analýzy. Přehled metod a postupů nalezneme například v (Cipra, 2008).

První metodou je *dekompozice časové řady*. Rozloží řadu do speciálních složek, a to na trendovou Tr_t , sezónní Sz_t , cyklickou C_t a reziduální E_t složku. Časová řada je poté dána aditivním tvarem

$$X_t = Tr_t + Sz_t + C_t + E_t$$

nebo multiplikativním tvarem

$$X_t = Tr_t Sz_t C_t E_t.$$

Trendová, sezónní a cyklická složka jsou deterministickou funkcí času. Reziduální (chybová) složka je náhodná, je reprezentována náhodným procesem. Tato metoda pracuje v časové doméně. Základním nástrojem je regrese. Popis složek, jejich modelování, odhad a předpovědi nalezneme v (Cipra, 2008).

Další metodou je *spektrální analýza*. Ta pracuje ve spektrální doméně. Základním popisným nástrojem je spektrální hustota. Podrobnější popis nalezneme například v (Prášková, 2004).

Velmi užitečnou metodou je *Boxova-Jenkinsova metodologie*. Za základní kámen časové řady pokládá reziduální složku. Nahlíží na časovou řadu jako na stochastický model. Základním nástrojem je korelační analýza. Modely jsou velmi pružné a díky tomu jsou užitečným nástrojem pro zkoumání časových řad. Tato metoda je velmi oblíbeným nástrojem s mnoha aplikacemi.

Autokorelační a parciální autokorelační funkce slouží jako užitečný nástroj pro rozhodnutí o typu modelu. Důležité je například určit hodnotu t_0 , takzvaný *bod useknutí*, kdy pro $t > t_0$ je autokorelační funkce nulová. Obdobně pro parciální autokorelační funkci.

Definice 10 (Bod useknutí). *Nechť $\{X_t, t \in T\}$ je slabě stacionární časová řada s autokovarianční funkcí $\gamma(t)$. Bodem useknutí autokorelační funkce je nejmenší možné $t_0 \in N_0$ takové, že pro každé $t \in \mathbb{Z}$ splňující $|t| > t_0$ platí $\gamma(t) = 0$. Obdobně definujeme bod useknutí parciální autokorelační funkce.*

Samozřejmě autokovarianční a autokorelační funkce dané časové řady mají stejný bod useknutí, což plyne z jejich definice. V naší práci se zabýváme očištěnou časovou řadou o trendovou, sezónní a cyklickou složku. Pracujeme pouze s reziduální (chybovou) náhodnou složkou.

Nyní se budeme zabývat typy modelů dle *Boxovy-Jenkinsovy metodologie*, patří sem *klouzavé průměry* MA, *autoregresní modely* AR a *smíšené modely*

ARMA. Připomeňme, že pracujeme se slabě stacionární časovou řadou a platí pro všechna $t \in T$

$$\gamma(-t) = \gamma(t),$$

dle věty 2. Tato vlastnost nám usnadní práci a zápis, nadále se můžeme zabývat pouze $\{\gamma(t), t \geq 0\}$. Obdobně pro parciální autokorelační funkci. Dále budeme pracovat s $T = \mathbb{Z}$. Nejdříve si definujeme základní časovou řadu, která slouží jako konstrukční kámen pro další modely. Jedná se o *bílý šum*.

Definice 11 (Bílý šum). Bílý šum je posloupnost $\{Y_t, t \in \mathbb{Z}\}$ nekorelovaných náhodných veličin s nulovou střední hodnotou a konstantním rozptylem $0 < \sigma^2 < \infty$. Značíme $WN(0, \sigma^2)$.

Bílý šum reprezentuje slabě stacionární časovou řadu s nulovou střední hodnotou a konstantním rozptylem. Náhodné veličiny jsou nekorelované. Často slouží k modelování chybové složky pro složitější struktury. Běžně se používá bílý šum, který je stejně rozdělený. Speciálním příkladem je posloupnost stejně rozdělených a nezávislých náhodných veličin s nulovou střední hodnotou a kladným konečným rozptylem. Nyní se podíváme na tvar autokorelační funkce.

Věta 6 (Autokorelační funkce bílého šumu). Necht $\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $WN(0, \sigma^2)$. Pak autokovarianční funkce je dána předpisem

$$\gamma(t) = \begin{cases} \sigma^2, & t = 0, \\ 0, & t \neq 0. \end{cases}$$

Autokorelační funkce je dána předpisem

$$\rho(t) = \begin{cases} 1, & t = 0, \\ 0, & t \neq 0. \end{cases}$$

Důkaz. Vzorec plyne ze slabé stacionarity bílého šumu a definice autokorelační a autokovarianční funkce. □

Nyní se podíváme na model *klouzavých průměrů* MA.

Definice 12 (Model klouzavých průměrů $MA(q)$). Posloupnost klouzavých průměrů řádu q je náhodná posloupnost $\{X_t, t \in \mathbb{Z}\}$ definovaná předpisem

$$X_t = b_0 Y_t + b_1 Y_{t-1} + \dots + b_q Y_{t-q}, \quad t \in \mathbb{Z},$$

kde $\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $WN(0, \sigma^2)$, $b_0, b_1, \dots, b_q \in \mathbb{R}$ jsou konstanty a $b_0 \neq 0, b_q \neq 0$. Posloupnost značíme $MA(q)$.

Z definice klouzavého průměru plyne, že již zde pozorujeme určitou korelovanost mezi náhodnými veličinami, které jsou dostatečně blízko sebe. Korelovanost je zapříčiněna sdílením prvků bílého šumu. Pro dostatečně rozestoupené náhodné veličiny X_t a X_{t+k} pro $k > q$ zaniká korelovanost.

Dalším pojmem je *kauzální lineární řada*.

Definice 13 (Kauzální lineární řada $MA(\infty)$). Kauzální lineární řada je náhodná posloupnost $\{X_t, t \in \mathbb{Z}\}$ definovaná předpisem

$$X_t = \sum_{k=0}^{\infty} b_k Y_{t-k}, \quad t \in \mathbb{Z},$$

kde $\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $WN(0, \sigma^2)$ a $\{b_i, i \in \mathbb{N}_0\}$ je posloupnost reálných koeficientů splňující podmínku $\sum_{k=0}^{\infty} |b_k| < \infty$. Posloupnost značíme $MA(\infty)$.

Nekonečnou sumu v definici 13 chápeme jako limitu v pravděpodobnosti

$$\sum_{k=0}^n b_k Y_{t-k} \xrightarrow[n \rightarrow \infty]{P} \sum_{k=0}^{\infty} b_k Y_{t-k}, \quad t \in \mathbb{Z},$$

Nyní se podíváme na autokorelační funkci.

Věta 7 (Autokorelační funkce $MA(q)$ a kauzální časové řady). Necht $\{X_t, t \in \mathbb{Z}\}$ je $MA(q)$ tvořený bílým šumem $WN(0, \sigma^2)$. Pak pro $t \in \mathbb{N}_0$ je autokovarianční funkce dána předpisem

$$\gamma(t) = \begin{cases} \sigma^2 \sum_{k=0}^{q-t} b_{k+t} b_k, & 0 \leq t \leq q, \\ 0, & q < |t|. \end{cases}$$

Autokorelační funkce je dána předpisem

$$\rho(t) = \begin{cases} \frac{\sum_{k=0}^{q-t} b_{k+t} b_k}{\sum_{k=0}^q b_k^2}, & -q \leq t \leq 0, \\ 0, & q < |t|. \end{cases}$$

Pro kauzální lineární řadu $MA(\infty)$ tvořenou bílým šumem $WN(0, \sigma^2)$ je autokovarianční funkce dána předpisem

$$\gamma(t) = \sigma^2 \sum_{k=0}^{\infty} b_{k+t} b_k, \quad 0 \leq t.$$

Autokorelační funkce je dána předpisem

$$\rho(t) = \frac{\sum_{k=0}^{\infty} b_{k+t} b_k}{\sum_{k=0}^{\infty} b_k^2}, \quad 0 \leq t.$$

Důkaz. Viz (Prášková, 2004, věta 5.1, věta 5.4 a kapitola 5.1). □

Bod useknutí autokorelační funkce pro klouzavý průměr řádu q je $t_0 = q$. Dle (Cipra, 2008) parciální autokorelační funkce nemá bod useknutí.

Složitější strukturou jsou *autoregresní modely AR*.

Definice 14 (Autoregresní model $AR(p)$). Autoregresní posloupnost řádu p je náhodná posloupnost $\{X_t, t \in \mathbb{Z}\}$ definovaná předpisem

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + Y_t, \quad t \in \mathbb{Z},$$

kde $\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $WN(0, \sigma^2)$, $\varphi_1, \dots, \varphi_p \in \mathbb{R}$ jsou konstanty a $\varphi_p \neq 0$. Posloupnost značíme $AR(p)$.

Nyní se podíváme na autokorelační funkci.

Věta 8 (Autokorelační funkce $AR(p)$). *Necht $\{X_t, t \in \mathbb{Z}\}$ je $AR(p)$ tvořený bílým šumem $WN(0, \sigma^2)$ a všechny kořeny polynomu $1 - \varphi_1 z - \dots - \varphi_p z^p$ leží mimo jednotkový kruh v \mathbb{C} . Pak $\{X_t, t \in \mathbb{Z}\}$ je slabě stacionární a lze jej reprezentovat jako kauzální lineární řadu*

$$X_t = \sum_{k=0}^{\infty} b_k Y_{t-k}, \quad t \in \mathbb{Z},$$

s autokovarianční a autokorelační funkcí danou předpisem ve větě 7. Koeficienty b_i jsou definovány vztahem

$$\sum_{k=0}^{\infty} b_k z^k = \frac{1}{1 - \varphi_1 z - \dots - \varphi_p z^p}, \quad |z| \leq 1 \quad (1.4)$$

Důkaz. Viz (Prášková, 2004, věta 5.5). □

Náhodná veličina X_t autoregresního náhodného procesu závisí na minulých náhodných veličinách vzhledem k času. Tím pádem všechny autokorelace jsou nenulové. Bod useknutí autokorelační funkce pro autoregresní model řádu p neexistuje. Ze struktury modelu a definice plyne, že parciální autokorelační funkce má bod useknutí $t_0 = p$, viz také (Cipra, 2008).

Poznamenejme, že některá literatura či software pracují s polynomem z věty 8 ve tvaru $z^p - \varphi_1 z^{p-1} - \dots - \varphi_p$, pak nás zajímá, zda kořeny jsou v jednotkovém kruhu v \mathbb{C} . Tento polynom má pouze převrácené kořeny než polynom z věty 8, proto je jedno jaký postup zvolíme. Při volbě způsobu výpočtu je pouze potřeba dodržet pravidlo, kdy nás zajímají kořeny mimo a uvnitř jednotkového kruhu.

Dle věty 8 lze autokovarianční funkci spočítat jako pro kauzální lineární proces. Prvně je třeba spočítat koeficienty b_i v kauzálním lineárním tvaru. K tomu se používá rozklad na parciální zlomky. Další možností je řešení systému rovnic, který získáme díky rovnosti (1.4) z věty 8.

$$\begin{aligned} b_0 &= 1, \\ b_1 - b_0 \varphi_1 &= 0, \\ &\vdots \\ b_j - b_{j-1} \varphi_1 - \dots - b_{j-p} \varphi_p &= 0, \quad j = p, p+1, \dots \end{aligned}$$

Poslední předpis nám dává systém homogenních diferenčních rovnic řádu p s počátečními podmínkami na b_0, \dots, b_{p-1} dané soustavou výše. Způsob řešení diferenčních rovnic nalezneme například v (Prágerová, 1971).

V praxi používanější způsob výpočtu autokovarianční a autokorelační funkce jsou Yuleovy-Walkerovy rovnice, jejichž popis nalezneme například v (Prášková, 2004). Základní popis je i v dodatku A.1.

Obecnou strukturou jsou *smíšené procesy* ARMA.

Definice 15 (Smíšený proces $ARMA(p, q)$). *Smíšený proces řádu p a q je náhodná posloupnost $\{X_t, t \in \mathbb{Z}\}$ definovaná předpisem*

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + Y_t + \theta_1 Y_{t-1} + \dots + \theta_q Y_{t-q}, \quad t \in \mathbb{Z},$$

kde $\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $WN(0, \sigma^2)$, $\varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q \in \mathbb{R}$ jsou konstanty a $\varphi_p \neq 0, \theta_q \neq 0$. Posloupnost značíme $ARMA(p, q)$.

Smíšené procesy jsou kombinací klouzavých průměrů a autoregresních modelů. Tvoří velmi obecný a užitečný nástroj pro popisování časových řad. Nyní se podíváme na autokorelační funkci.

Věta 9 (Autokorelační funkce $ARMA(p, q)$). *Nechť $\{X_t, t \in \mathbb{Z}\}$ je $ARMA(p, q)$ tvořený bílým šumem $WN(0, \sigma^2)$, všechny kořeny polynomu $1 - \varphi_1 z - \dots - \varphi_p z^p$ leží mimo jednotkový kruh v \mathbb{C} a polynomy $1 - \varphi_1 z - \dots - \varphi_p z^p$ a $1 + \theta_1 z + \dots + \theta_q z^q$ nemají žádné společné kořeny. Pak $\{X_t, t \in \mathbb{Z}\}$ je slabě stacionární a lze jej reprezentovat jako kauzální lineární řadu*

$$X_t = \sum_{k=0}^{\infty} b_k Y_{t-k}, \quad t \in \mathbb{Z},$$

s autokovarianční a autokorelační funkcí danou předpisem ve větě 7. Koeficienty b_i jsou definovány vztahem

$$\sum_{k=0}^{\infty} b_k z^k = \frac{1 + \theta_1 z + \dots + \theta_q z^q}{1 - \varphi_1 z - \dots - \varphi_p z^p}, \quad |z| \leq 1.$$

Důkaz. Viz (Prášková, 2004, věta 5.6). □

Opět používáme přepis do kauzálního lineárního tvaru. Věta 9 nám dává vztah, který lze použít pro výpočet. V praxi se opět používají spíše Yuleovy-Walkerovy rovnice.

Autokorelační funkce smíšeného procesu řádu p a q nemá bod useknutí dle vět 8 a 7. Dle (Cipra, 2008) ani parciální autokorelační funkce nemá bod useknutí.

Tabulka 1.1 převzatá z (Cipra, 2008, tabulka 10.3.1) shrnuje identifikaci modelu podle chování autokorelační a parciální autokorelační funkce.

Proces	Autokorelační f.	Parciál. autokor. f.
MA(q)	$t_0 = q$	t_0 neexistuje omezena křivkou U
AR(p)	t_0 neexistuje ve tvaru křivky U	$t_0 = p$
ARMA(p, q)	t_0 neexistuje ve tvaru křivky U po prvních $q - p$ hodnotách	t_0 neexistuje omezena křivkou U po prvních $p - q$ hodnotách

Tabulka 1.1: Chování autokorelační a autokovarianční funkce k identifikaci modelu. U je křivkou ve tvaru lineární kombinace klesajících geometrických posloupností a sinusoid s geometricky klesajícími amplitudami.

Pro praktická data se na základě odhadu autokorelační a parciální autokorelační funkce snažíme zjistit bod useknutí a na základě toho identifikovat model. Proto potřebujeme kvalitní odhad autokorelační funkce a právě tímto se bude

zabývat tato práce. Z průběhu autokorelační funkce se dají vypočítat i další informace o časové řadě, například periodicitu či trend. O tyto informace můžeme špatným odhadem přijít.

Výše zmíněné časové řady definujeme jako centrované, s nulovou střední hodnotou. Samozřejmě je možné pracovat s řadou s nenulovou střední hodnotou, vyžadujeme pouze její konstantnost. ARMA(p, q) řada by poté byla ve tvaru

$$X_t - \mu = \varphi_1(X_{t-1} - \mu) + \dots + \varphi_p(X_{t-p} - \mu) + Y_t + \theta_1 Y_{t-1} + \dots + \theta_q Y_{t-q}, \quad t \in \mathbb{Z},$$

kde μ je střední hodnota řady.

Uvedeme několik příkladů na výpočet autokorelační funkce pro různé časové řady.

Příklad. Podíváme se na konkrétní tvar autokorelačních funkcí výše definovaných modelů. Prvně budeme pracovat s klouzavým průměrem MA(2) ve tvaru

$$X_t = b_0 Y_t + b_1 Y_{t-1} + b_2 Y_{t-2}, \quad t \in \mathbb{Z},$$

kde $b_0, b_1, b_2 \in \mathbb{R}$ jsou koeficienty. Autokorelační funkci spočítáme snadno dle věty 7 a dostaneme

$$\rho(t) = \begin{cases} \frac{b_0^2 + b_1^2 + b_2^2}{b_0^2 + b_1^2 + b_2^2} = 1, & t = 0 \\ \frac{b_0 b_1 + b_0 b_2}{b_0^2 + b_1^2 + b_2^2}, & t = 1, -1 \\ \frac{b_0 b_2}{b_0^2 + b_1^2 + b_2^2}, & t = 2, -2 \\ \frac{0}{b_0^2 + b_1^2 + b_2^2} = 0, & |t| > 2. \end{cases}$$

Příklad. Nyní se podíváme na autoregresní model AR(1)

$$X_t = \varphi_1 X_{t-1} + Y_t, \quad t \in \mathbb{Z}.$$

Předpokládejme, že $|\varphi_1| < 1$, aby se jednalo o slabě stacionární řadu dle věty 8. Převedeme zápis procesu do kauzálního tvaru pomocí soustavy rovnic

$$\begin{aligned} b_0 &= 1, \\ b_k - \varphi_1 b_{k-1} &= 0, \quad k \geq 1. \end{aligned}$$

Dostaneme počáteční podmínky

$$\begin{aligned} b_0 &= 1, \\ b_1 &= \varphi_1. \end{aligned}$$

Pracujeme s diferenční rovnicí

$$z - \varphi_1 = 0.$$

Získáme kořen $z = \varphi_1$. Řešení je ve tvaru $b_k = a_1 \varphi_1^k$ pro $k \geq 0$. Z počátečních podmínek vyplývá $a_1 = 1$. Pak kauzální tvar je

$$X_t = \sum_{k=0}^{\infty} \varphi_1^k Y_{t-k}, \quad t \in \mathbb{Z}.$$

Dle věty 8 dostaneme autokorelační funkci

$$\rho(t) = \frac{\sum_{k=0}^{\infty} \varphi_1^k \varphi_1^{k+t}}{\sum_{k=0}^{\infty} \varphi_1^{2k}} = \varphi_1^t, \quad t \geq 0. \quad (1.5)$$

Pro $t \leq 0$ použijeme vztah $\rho(t) = \rho(-t)$.

Příklad. Jako poslední uvedeme smíšený model ARMA(1,1)

$$X_t = \varphi_1 X_{t-1} + Y_t + \theta_1 Y_{t-1}, \quad t \in \mathbb{Z}.$$

Předpokládejme, že $|\varphi_1| < 1$ a $\varphi_1 \neq -\theta_1$, aby se jednalo o slabě stacionární řadu dle věty 9. Sestavíme Yuleovy-Walkerovy rovnice dle (Prášková, 2004) pomocí přenosobení X_t a X_{t-1} a aplikací střední hodnoty na obě strany.

$$\begin{aligned} EX_t^2 - \varphi_1 EX_t X_{t-1} &= EX_t Y_t + \theta_1 EX_t Y_{t-1}, \\ EX_t X_{t-1} - \varphi_1 EX_{t-1}^2 &= EX_{t-1} Y_t + \theta_1 EX_{t-1} Y_{t-1}, \\ EX_t X_{t-k} - \varphi_1 EX_{t-1} X_{t-k} &= EX_{t-k} Y_t + \theta_1 EX_{t-k} Y_{t-1}, \quad k \geq 2. \end{aligned}$$

Platí, že $EX_{t-k} Y_t = 0$ pro $k \geq 1$ a $EX_{t-k} Y_{t-1} = 0$ pro $k \geq 2$. Nyní uděláme totéž ale přenosobením Y_t a Y_{t-1} .

$$\begin{aligned} EX_t Y_t - \varphi_1 EX_{t-1} Y_t &= EY_t^2 + \theta_1 EY_{t-1} Y_t, \\ EX_t Y_{t-1} - \varphi_1 EX_{t-1} Y_{t-1} &= EY_t Y_{t-1} + \theta_1 EY_{t-1}^2. \end{aligned}$$

Z této soustavy dostaneme vztahy

$$\begin{aligned} EX_t Y_t &= \sigma^2, \\ EX_t Y_{t-1} &= \sigma^2(\varphi_1 + \theta_1), \end{aligned}$$

kde σ^2 je rozptyl bílého šumu z definice 15. Dostaneme soustavu rovnic autokovariančních funkcí

$$\begin{aligned} \gamma(0) - \varphi_1 \gamma(1) &= \sigma^2(1 + \varphi_1 \theta_1 + \theta_1^2), \\ \gamma(1) - \varphi_1 \gamma(0) &= \sigma^2 \theta_1, \\ \gamma(k) - \varphi_1 \gamma(k-1) &= 0, \quad k \geq 2. \end{aligned}$$

Poslední rovnici lze použít pro rekurentní předpis autokorelační funkce

$$\gamma(k) = \varphi_1 \gamma(k-1), \quad k \geq 2.$$

Ze soustavy výše dostaneme počáteční podmínky

$$\begin{aligned} \gamma(0) &= \sigma^2 \frac{1 + \theta^2 + 2\varphi_1 \theta_1}{1 - \varphi_1^2}, \\ \gamma(1) &= \sigma^2 \theta_1 + \varphi_1 \gamma(0), \end{aligned}$$

a dle rekurentního předpisu dostaneme jako řešení autokorelační funkce

$$\gamma(k) = \varphi_1 \gamma(k-1) = \varphi_1^{k-1} \gamma(1), \quad k \geq 2.$$

Pro $t \leq 0$ použijeme vztah $\rho(t) = \rho(-t)$.

2. Klasický odhad autokorelační funkce

V této kapitole se budeme zabývat klasickým odhadem autokorelační funkce a způsoby odhadu parciální autokorelační funkce. Uvedeme některé jeho vlastnosti. Na modelovém příkladu ukážeme jeho nevýhody.

2.1 Metoda výpočtu a vlastnosti

Připomeňme, že v celé práci pracujeme s časovou řadou $\{X_t, t \in \mathbb{Z}\}$, která je slabě stacionární dle definice 7. Její autokovarianční funkce je γ a autokorelační funkce je ρ . Dle věty 2 se můžeme zabývat pouze $\{\gamma(t), t \geq 0\}$. Obdobně pro parciální autokorelační funkci.

Klasický odhad autokovarianční funkce je dán předpisem

$$\hat{\gamma}(t) = \frac{1}{n} \sum_{k=1}^{n-t} (X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n), \quad t = 0, 1, \dots, n-1, \quad (2.1)$$

kde

$$\bar{X}_n = \frac{1}{n} \sum_{t=1}^n X_t,$$

a n je délka časové řady. Připomeňme, že platí $\hat{\gamma}(t) = \hat{\gamma}(-t)$ pro $t < 0$.

Klasickým odhadem autokorelační funkce je

$$\hat{\rho}(t) = \frac{\hat{\gamma}(t)}{\hat{\gamma}(0)} = \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2}, \quad (2.2)$$

pokud $\hat{\gamma}(0) = \frac{1}{n} \sum_{t=1}^n (X_t - \bar{X}_n)^2 > 0$

V situacích, kdy $t \geq n$, není odhad definovaný. V případě, kdy například $t = n-2$, máme velmi malé množství měření, ze kterých bychom odhad počítali. Takový odhad by byl nespolehlivý. V literatuře (Brockwell a Davis, 2002) je doporučený minimální rozsah $n = 50$ a odhadovat autokovarianční a autokorelační funkci pouze pro $t \leq \frac{n}{4}$. Malé množství měření může být vážným problémem pro data s roční frekvencí měření.

Klasický odhad není nestranný pro $t \geq 1$ dle (Brockwell a Davis, 2002, kapitola 2.4.2), tedy $E\hat{\rho}(t) \neq \rho(t)$. Klasický odhad je konzistentní, což dokážeme v následující větě 10.

Věta 10 (Konzistence). *Nechť $\{X_t, t \in \mathbb{Z}\}$ je slabě stacionární časová řada s konstantní střední hodnotou μ , autokorelační funkcí $\rho(t)$, autokovarianční funkcí $\gamma(t)$ splňující*

$$\gamma(n) \xrightarrow{n \rightarrow \infty} 0,$$

a platí, že pro každé $t, k \in \mathbb{Z}$ je $E|X_t|^4 < \infty$ a

$$E(X_k X_{k+t} - EX_k X_{k+t})(X_{k+n} X_{k+t+n} - EX_{k+n} X_{k+t+n}) \xrightarrow{n \rightarrow \infty} 0,$$

Pak klasický odhad $\hat{\rho}(t)$ je konzistentní.

Důkaz. Pro $t \in \mathbb{Z}$ pevně určené

$$\begin{aligned}
\hat{\rho}(t) &= \frac{\hat{\gamma}(t)}{\hat{\gamma}(0)} \\
&= \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} \\
&= \frac{n-t}{n} \frac{\frac{1}{n-t} \sum_{k=1}^{n-t} (X_k X_{k+t} - \bar{X}_n X_k - \bar{X}_n X_{k+t} + \bar{X}_n^2)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} \\
&= \frac{n-t}{n} \frac{\frac{1}{n-t} \sum_{k=1}^{n-t} (X_k X_{k+t}) - \bar{X}_n \frac{1}{n-t} \sum_{k=1}^{n-t} (X_k) - \bar{X}_n \frac{1}{n-t} \sum_{k=1}^{n-t} (X_{k+t}) + \bar{X}_n^2}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} \\
&\xrightarrow{P}_{n \rightarrow \infty} 1 \cdot \frac{EX_0 X_t - \mu^2 - \mu^2 + \mu^2}{E(X_0 - \mu)^2} \\
&= \frac{EX_0 X_t - \mu^2}{E(X_0 - \mu)^2} \\
&= \frac{E(X_t - \mu)(X_0 - \mu)}{E(X_0 - \mu)^2} \\
&= \rho(t).
\end{aligned}$$

Bavíme se zde o konvergenci v pravděpodobnosti. Konvergenci dostaneme z následujících vztahů

$$\begin{aligned}
\bar{X}_n &\xrightarrow{P}_{n \rightarrow \infty} \mu, \\
\frac{1}{n-t} \sum_{k=1}^{n-t} X_k X_{k+t} &\xrightarrow{P}_{n \rightarrow \infty} EX_0 X_t, \\
\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 &= \frac{1}{n} \sum_{k=1}^n (X_k^2) - 2\bar{X}_n \frac{1}{n} \sum_{k=1}^n (X_k) + \bar{X}_n^2 \xrightarrow{P}_{n \rightarrow \infty} \\
&\xrightarrow{P}_{n \rightarrow \infty} EX_0^2 - 2\mu^2 + \mu^2 = E(X_0 - \mu)^2,
\end{aligned}$$

které nám plynou z ergodické věty (Brockwell a kol., 1991, věta 7.1.1) (podrobně komentováno v (Prášková, 2004, kapitola 14.1)), pokud $\gamma(n) \xrightarrow{n \rightarrow \infty} 0$, pak $\text{var} \bar{X}_t \xrightarrow{n \rightarrow \infty} 0$ a dostaneme $\bar{X}_n \xrightarrow{P}_{n \rightarrow \infty} \mu$. Obdobně pro výsledek $EX_0 X_t$. \square

Podíváme se na aproximaci rozdělení vektoru $(\hat{\rho}(1), \dots, \hat{\rho}(k))^\top$ z (Brockwell a kol., 1991).

Věta 11 (Bartlettova formule). *Nechť $\{X_t, t \in \mathbb{Z}\}$ je slabě stacionární časová řada daná předpisem*

$$X_t - \mu = \sum_{k=-\infty}^{\infty} b_k Y_{t-k},$$

kde $\mu = EX_t$, $\{Y_t, t \in \mathbb{Z}\}$ jsou stejně rozdělené a nezávislé náhodé veličiny s nulovou střední hodnotou a konstantním rozptylem a platí, že $\sum_{k=-\infty}^{\infty} |b_k| < \infty$ a $EY_t^4 < \infty$. Označme $\boldsymbol{\rho}^{(t)} = (\rho(1), \dots, \rho(t))^\top$.

Pak vektor $\hat{\boldsymbol{\rho}}^{(t)} = (\hat{\rho}(1), \dots, \hat{\rho}(t))^\top$ pro každé $t \geq 1$ konverguje v distribuci

$$\sqrt{n}(\hat{\boldsymbol{\rho}}^{(t)} - \boldsymbol{\rho}^{(t)}) \xrightarrow[n \rightarrow \infty]{D} N_t(\mathbf{0}_t, \mathbf{W}),$$

kde $\mathbf{0}_t$ je vektor 0 velikosti t , \mathbf{W} je matice $t \times t$, jejíž (i, j) prvek je dán tzv. Bartlettovou formulí

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

což lze přepsat do tvaru

$$w_{i,j} = \sum_{k=1}^{\infty} [\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)][\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)]$$

Důkaz. Viz (Brockwell a kol., 1991, věty 7.2.1 a 7.2.2 a kapitola 7.3). □

Předpoklad o konečném čtvrtém momentu $EY_t^4 < \infty$ lze nahradit podmínkou $\sum_{k=-\infty}^{\infty} b_k^2 |k| < \infty$ dle (Brockwell a kol., 1991, věty 7.2.2 a kapitola 7.3).

Pro $\{X_t, t \in \mathbb{Z}\}$ časovou řadu stejně rozdělených a nezávislých náhodných veličin s nulovou střední hodnotou a konstantním rozptylem je

$$w_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Toto plyne z $\rho(t) = 0$ pro $|t| \geq 1$, což dosadíme do předpisu $w_{i,j}$. Poté dostaneme, že složky vektoru $\hat{\boldsymbol{\rho}}^{(t)}$ jsou asymptoticky nezávislé.

Dle věty 11 jsme schopni sestavit asymptotický interval spolehlivosti pro $\rho(t)$ pro $|t| \geq 1$. Podíváme se na příklad s bílým šumem. Předpokládejme, že $\{X_t, t = 1, \dots, n\}$ je stejně rozdělený bílý šum $WN(0, \sigma^2)$. Pak lze sestavit $1 - \alpha$ asymptotický interval spolehlivosti $\rho(t)$ pro $|t| \geq 1$ ve tvaru

$$\left(\frac{-1}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \frac{1}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right),$$

kde $u_{1-\frac{\alpha}{2}}$ je $1 - \frac{\alpha}{2}$ kvantil normovaného normálního rozdělení. Tento pás kreslí některé softwary, například R, automaticky do odhadu autokorelační funkce.

Dále se budeme zabývat odhadem parciální autokorelační funkce. Větu 4 použijeme k odhadu parciální autokorelační funkce, ke kterému použijeme odhad autokovarianční funkce. Uvažujme naměřená data x_1, \dots, x_n o rozsahu n . Předpokládejme, že pro každé $i, j \in \{1, \dots, n\}, i \neq j$ je $x_i \neq x_j$. Pak odhad parciální autokorelační funkce je

$$\begin{aligned} \hat{\pi}(0) &= 1 \\ \hat{\pi}(t) &= \hat{\phi}_{tt}, \quad t \geq 1, \end{aligned}$$

kde $\hat{\phi}_{tt}$ je řešením

$$\hat{\mathbf{\Gamma}}^{(t)} \left(\hat{\phi}_{t1}, \dots, \hat{\phi}_{tt} \right)^\top = \hat{\boldsymbol{\gamma}}^{(t)},$$

matice $\hat{\mathbf{\Gamma}}^{(t)} = [\hat{\gamma}(i-j)]_{i,j=0}^{t-1}$ a $\hat{\boldsymbol{\gamma}}^{(t)} = (\hat{\gamma}(1), \dots, \hat{\gamma}(t))^\top$.

Alternativní odhad parciální autokorelační funkce dle (Makhoul, 1981) získáme jako

$$\hat{\pi}(t) = 2 \frac{\sum_{k=t+1}^n \hat{U}_{t,k} \hat{V}_{t,k}}{\sum_{k=t+1}^n (\hat{U}_{t,k}^2 + \hat{V}_{t,k}^2)}, \quad (2.3)$$

kde $\hat{U}_{t,k}$ a $\hat{V}_{t,k}$ jsou odhady $U_{t,k}$ a $V_{t,k}$ z definice 9. K tomu je potřeba spočítat odhady \tilde{X}_k a \tilde{X}_{k-t} . Odhad $\hat{\tilde{X}}_k$ pro \tilde{X}_k je dán předpisem

$$\hat{\tilde{X}}_k = \hat{\pi}_{t-1,0} + \sum_{l=1}^{t-1} \hat{\pi}_{t-1,l} X_{k-l},$$

kde $\hat{\pi}_{t-1,0}, \dots, \hat{\pi}_{t-1,t-1} \in \mathbb{R}$ jsou konstanty určené podmínkou

$$(\hat{\pi}_{t-1,0}, \dots, \hat{\pi}_{t-1,t-1}) = \arg \min_{\pi_{t-1,0}, \dots, \pi_{t-1,t-1} \in \mathbb{R}} \sum_{k=1}^n (X_k - \tilde{X}_k)^2.$$

Tyto odhady koeficientů poté použijeme i pro

$$\hat{\tilde{X}}_{k-t} = \hat{\pi}_{t-1,0} + \sum_{l=1}^{t-1} \hat{\pi}_{t-1,l} X_{k-t+l}.$$

Pro centrovanou časovou řadu je $\hat{\pi}_{t-1,0} = 0$. Pak dle věty 5 lze použít odhad parciální autokorelační funkce k odhadu autokorelační funkce rekurentním vztahem

$$\hat{\rho}(t) = \sum_{k=1}^{t-1} \hat{\pi}_{t-1,k} \hat{\rho}(t-k) + \hat{\pi}(t) \left(1 - \sum_{k=1}^{t-1} \hat{\pi}_{t-1,k} \hat{\rho}(k) \right),$$

a k odhadu koeficientů použijeme vztah

$$\begin{pmatrix} \hat{\pi}_{t,1} \\ \vdots \\ \hat{\pi}_{t,t-1} \end{pmatrix} = \begin{pmatrix} \hat{\pi}_{t-1,1} \\ \vdots \\ \hat{\pi}_{t-1,t-1} \end{pmatrix} - \hat{\pi}(t) \begin{pmatrix} \hat{\pi}_{t-1,t-1} \\ \vdots \\ \hat{\pi}_{t-1,1} \end{pmatrix}$$

s počáteční podmínkou $\hat{\pi}(1) = \hat{\rho}(1)$ a platí $\hat{\pi}_{t,t} = \hat{\pi}(t)$. Odhady se provádějí iterativně. V prvním kroku máme $\hat{\rho}(1) = \hat{\pi}(1) = \hat{\pi}_{1,1}$, které použijeme pro odhady $\hat{U}_{2,k}$ a $\hat{V}_{2,k}$. Ty poté použijeme ke spočtení $\hat{\pi}(2)$. Poté řešíme rovnost

$$\hat{\pi}_{2,1} = \hat{\pi}_{1,1} - \hat{\pi}(2) \hat{\pi}_{1,1}.$$

V druhém kroku použijeme odhady $\hat{\pi}_{2,1}$ a $\hat{\pi}_{2,2} = \hat{\pi}(2)$ k odhadům $\hat{U}_{3,k}$ a $\hat{V}_{3,k}$ a získání $\hat{\pi}(3)$. Poté řešíme systém rovnic

$$\begin{pmatrix} \hat{\pi}_{3,1} \\ \hat{\pi}_{3,2} \end{pmatrix} = \begin{pmatrix} \hat{\pi}_{2,1} \\ \hat{\pi}_{2,2} \end{pmatrix} - \hat{\pi}(3) \begin{pmatrix} \hat{\pi}_{2,2} \\ \hat{\pi}_{2,1} \end{pmatrix},$$

který nám dá odhady koeficientů do dalšího kroku algoritmu, ve kterém budeme pracovat s $\hat{\pi}_{3,1}$, $\hat{\pi}_{3,2}$ a $\hat{\pi}_{3,3} = \hat{\pi}(3)$.

2.2 Nevýhody klasického odhadu

Podíváme se na chování klasického odhadu pro centrovanou časovou řadu X_1, \dots, X_n s jedním odlehlým pozorováním. Pro centrovanou časovou řadu lze zavést zkrácený zápis odhadu autokorelační funkce

$$\tilde{\rho}(t) = \frac{\sum_{k=1}^{n-t} X_k X_{k+t}}{\sum_{k=1}^n X_k^2}.$$

Budeme pracovat s realizacemi $\{x_t, t = 1, \dots, n\}$. Bez újmy na obecnosti řekneme, že odlehlé pozorování je x_n , pro názornost přeznačíme na $x_n = K$, kde $K \in \mathbb{R}$. Pak odhad autokorelační funkce dle zkráceného zápisu je

$$\hat{\rho}(t) = \frac{\sum_{k=1}^{n-t-1} x_k x_{k+t}}{\sum_{k=1}^{n-1} x_k^2 + K^2} + \frac{x_{n-t} K}{\sum_{k=1}^{n-1} x_k^2 + K^2}, \quad t = 0, 1, \dots, n-1.$$

Vidíme, že velké hodnoty K zapříčiní, že ostatní složky výpočtu jsou zanedbatelné a záleží pouze na chování K . Spočítáme limitu pro K jdoucí do ∞ pro pevné $t \geq 1$

$$\lim_{K \rightarrow \infty} \hat{\rho}(t) = \lim_{K \rightarrow \infty} \frac{\sum_{k=1}^{n-t-1} x_k x_{k+t}}{\sum_{k=1}^{n-1} x_k^2 + K^2} + \frac{x_{n-t} K}{\sum_{k=1}^{n-1} x_k^2 + K^2} = 0. \quad (2.4)$$

Čímž přicházíme o informaci o jakékoliv korelovanosti mezi daty a výsledný odhad je zcela zavádějící.

Nyní se podíváme na působení odlehlých pozorování na odhad autokorelační funkce. Nasimulujeme si data pro MA(2) a AR(1).

Příklad. Nejdříve budeme pracovat s klouzavým průměrem MA(2) ve tvaru

$$X_t = Y_t + 0.9Y_{t-1} + 0.6Y_{t-2}, \quad t \in \mathbb{Z}.$$

Autokorelační funkci spočítáme dle věty 7 a dostaneme

$$\rho(t) = \begin{cases} = 1, & t = 0 \\ \doteq 0.664, & t = 1, -1 \\ \doteq 0.276, & t = 2, -2 \\ = 0, & |t| > 2. \end{cases}$$

Vygenerujeme si dle tohoto předpisu časovou řadu, pro ni spočítáme klasický odhad autokorelační funkce a porovnáme se skutečnými hodnotami autokorelační funkce výše spočítanými.

Dále budeme řešit vložení odlehlých pozorování do časové řady. Prvním příkladem jsou aditivní odlehlá pozorování. Odlehlost je způsobena změnou hodnoty X_t , nikoliv odlehlostí chybové složky Y_t . Tedy se jedná o izolovanou odlehlost bez vlivu na ostatní hodnoty X_t . Druhým případem jsou inovační odlehlá pozorování. Odlehlost je v chybové složce Y_t , takže má vliv na více pozorování X_t . Způsoby odlehlosti se budeme zabývat později. Opět spočítáme odhad autokorelační funkce pro řady s odlehlými pozorováními a porovnáme se skutečnými hodnotami.

Výstup nalezneme na obrázku 2.1. Na obrázku 2.1c je časová řada s aditivními odlehlými pozorováními, která ovlivňuje svojí odlehlostí pouze jedno pozorování

časové řady. Naproti tomu na obrázku 2.1e s inovačními odlehlými pozorováními vidíme, že efekt odlehlosti se přenáší i na několik dalších pozorování. Odhad autokorelační funkce s aditivními odlehlými pozorováními je významně horší než odhad pro řadu bez odlehlých pozorování. V tomto konkrétním příkladě pro inovační odlehlá pozorování dostáváme srovnatelný odhad autokorelační funkce jako pro řadu bez odlehlých pozorování. Poznamenejme, že v praxi obecně nevíme, o jaký typ odlehlosti se jedná.

Nyní se podíváme na autoregresní model AR(1)

$$X_t = 0.8X_{t-1} + Y_t, \quad t \in \mathbb{Z}.$$

Dle předchozího příkladu (1.5) máme autokorelační funkci

$$\rho(t) = 0.8^{|t|}, \quad t \in \mathbb{Z}.$$

Vygenerujeme časovou řadu a postupujeme jako výše. Opět se zabýváme aditivními a inovačními odlehlými pozorováními. U aditivních pozorování nejdříve vygenerujeme celou časovou řadu a až poté změníme individuální hodnoty X_t . Tím zajistíme, aby se odlehlost nepřenesla na další členy řady vzhledem k AR struktuře. Inovační odlehlost je opět vytvořena odlehlostí v chybové složce Y_t .

Výstup nalezneme na obrázku 2.2. Chování aditivních pozorování je stejné jako u MA řady. U inovačních odlehlých pozorování je odlehlost znatelnější než pro MA řadu. Dle AR struktury se efekt odlehlosti přenáší na více pozorování řady než v MA struktuře a trvá déle, než se stane nevýznamným. Opět získáváme velmi špatný odhad autokorelační funkce pro aditivní odlehlá pozorování oproti řadě bez odlehlých pozorování. V tomto konkrétním příkladě pro inovační odlehlá pozorování dostáváme srovnatelný odhad autokorelační funkce jako pro řadu bez odlehlých pozorování.

Tyto poznatky nám ukazují vážnost nerozpoznání a neošetření odlehlých pozorování na odhad autokorelační funkce. Za tímto účelem byly navrženy robustní metody, které jsou hlavním cílem této práce.

Příklad. Nyní se podíváme na situaci, kdy bílý šum bude pocházet ze Studentova t -rozdělení se 3 stupni volnosti t_3 , které má těžké chvosty. Ty mohou zapříčinit, že klasický odhad je zavádějící. Právě těžké chvosty jsou jedním z důvodů proč se zavádějí robustní metody.

Pracujeme s MA(2)

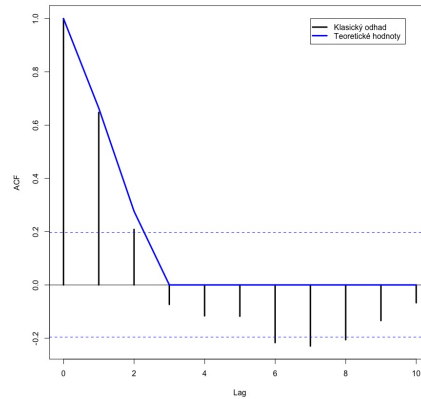
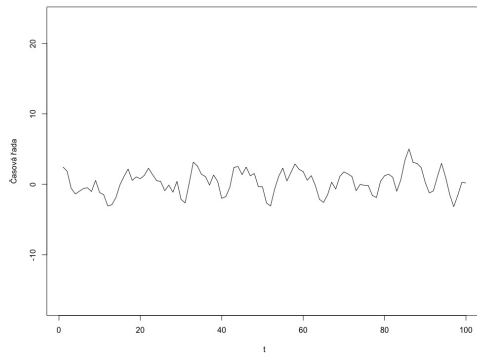
$$X_t = Y_t + 0.9Y_{t-1} + 0.6Y_{t-2}, \quad t \in \mathbb{Z},$$

s autokorelační funkcí

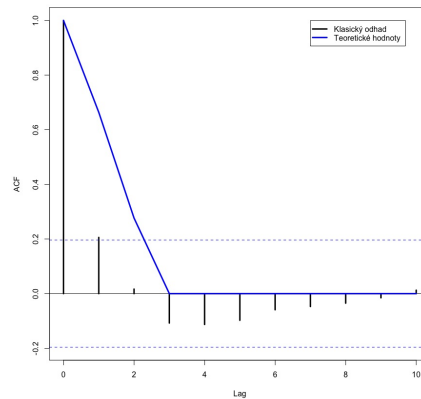
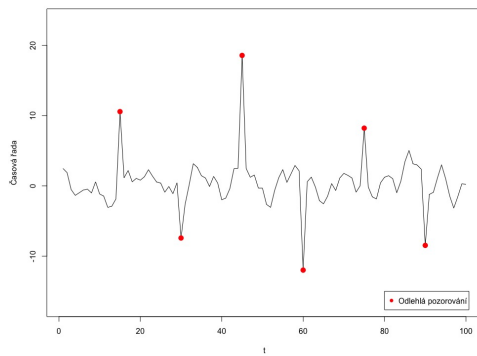
$$\rho(t) = \begin{cases} = 1, & t = 0 \\ \doteq 0.664, & t = 1, -1 \\ \doteq 0.276, & t = 2, -2 \\ = 0, & |t| > 2, \end{cases}$$

a AR(1)

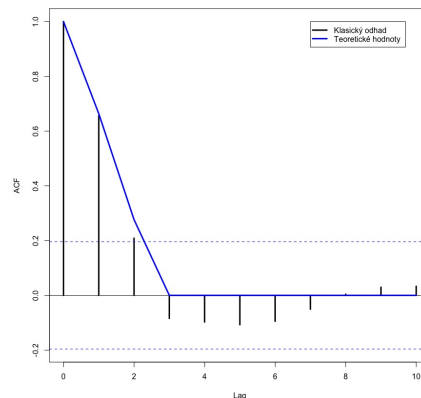
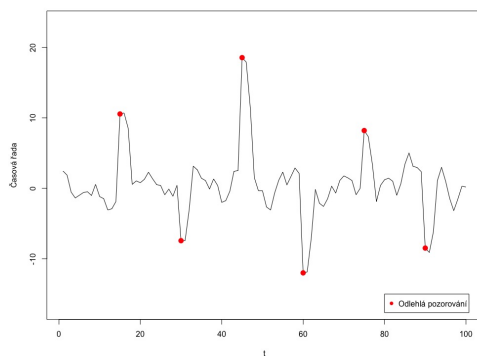
$$X_t = 0.8X_{t-1} + Y_t, \quad t \in \mathbb{Z}.$$



(a) MA(2) bez jakýchkoliv odlehlých pozorování (b) Klasický odhad ACF bez odlehlých pozorování

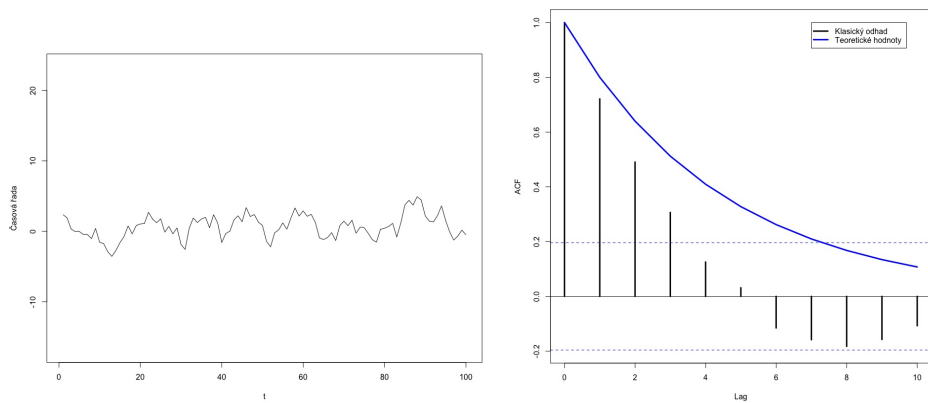


(c) MA(2) s aditivními odlehlými pozorováními (d) Klasický odhad ACF pro aditivní odlehlá pozorování

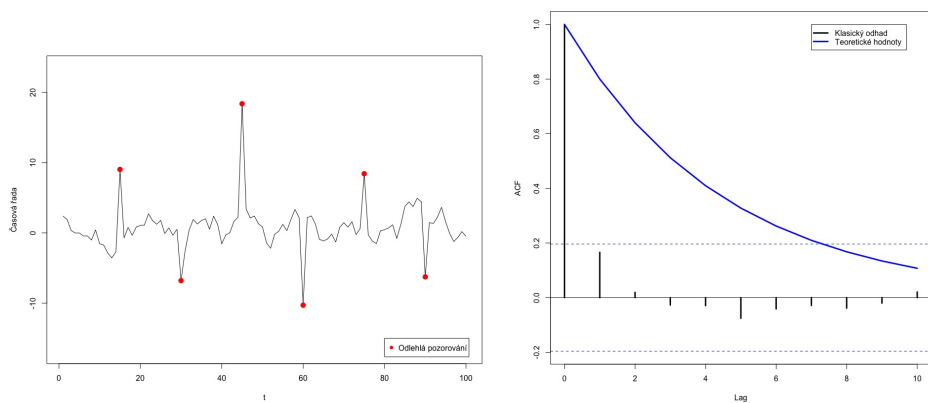


(e) MA(2) s inovačními odlehlými pozorováními (f) Klasický odhad ACF pro inovační odlehlá pozorování

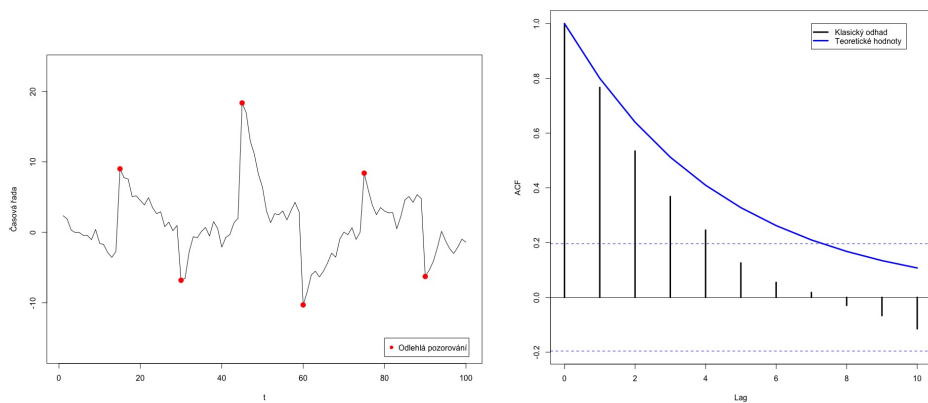
Obrázek 2.1: Porovnání klasického odhadu s a bez odlehlých pozorování pro MA(2). Časová řada bez odlehlých pozorování je $X_t = Y_t + 0.9Y_{t-1} + 0.6Y_{t-2}$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.



(a) AR(1) bez jakýchkoliv odlehlých pozorování (b) Klasický odhad ACF bez odlehlých pozorování

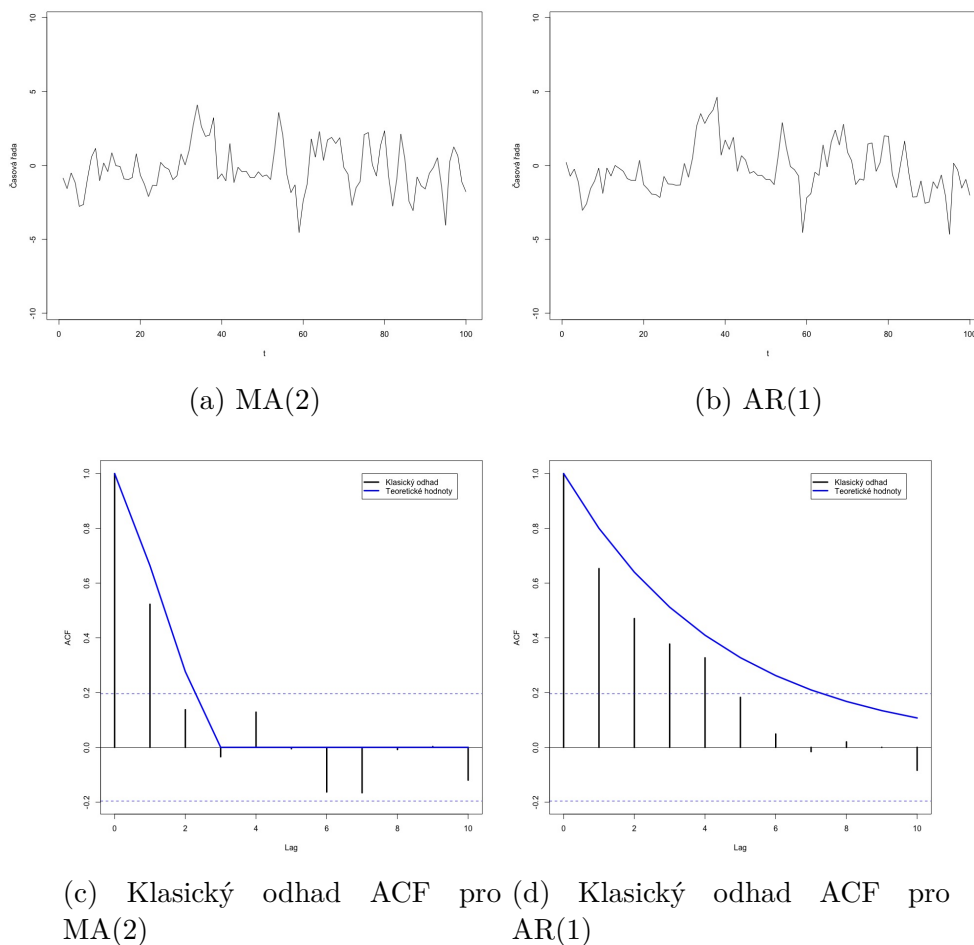


(c) AR(1) s aditivními odlehlými pozorováními (d) Klasický odhad ACF pro aditivní odlehlá pozorování



(e) AR(1) s inovačními odlehlými pozorováními (f) Klasický odhad ACF pro inovační odlehlá pozorování

Obrázek 2.2: Porovnání klasického odhadu s a bez odlehlých pozorování pro AR(1). Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.



Obrázek 2.3: Klasický odhad pro časové řady generované bílým šumem pocházejícím z rozdělení s těžkými chvosty. Časová řada MA(2) je $X_t = Y_t + 0.9Y_{t-1} + 0.6Y_{t-2}$ o rozsahu $n = 100$. Časová řada AR(1) je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$. Y_t pochází ze Studentova t -rozdělení se 3 stupni volnosti t_3 .

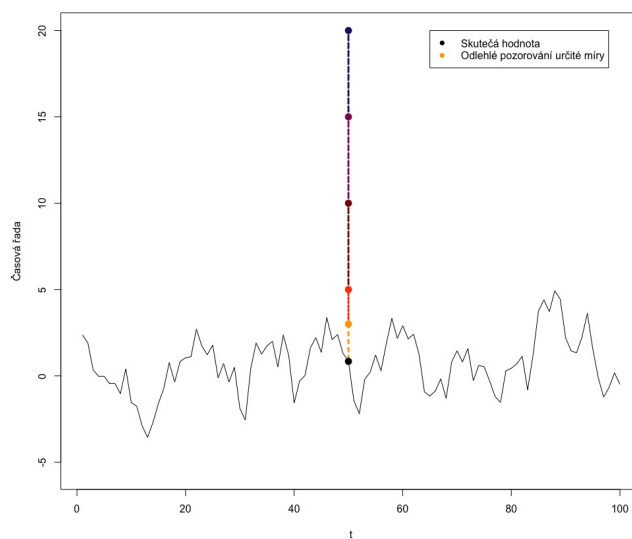
s autokorelační funkcí

$$\rho(t) = 0.8^{|t|}, \quad t \in \mathbb{Z}. \quad (2.5)$$

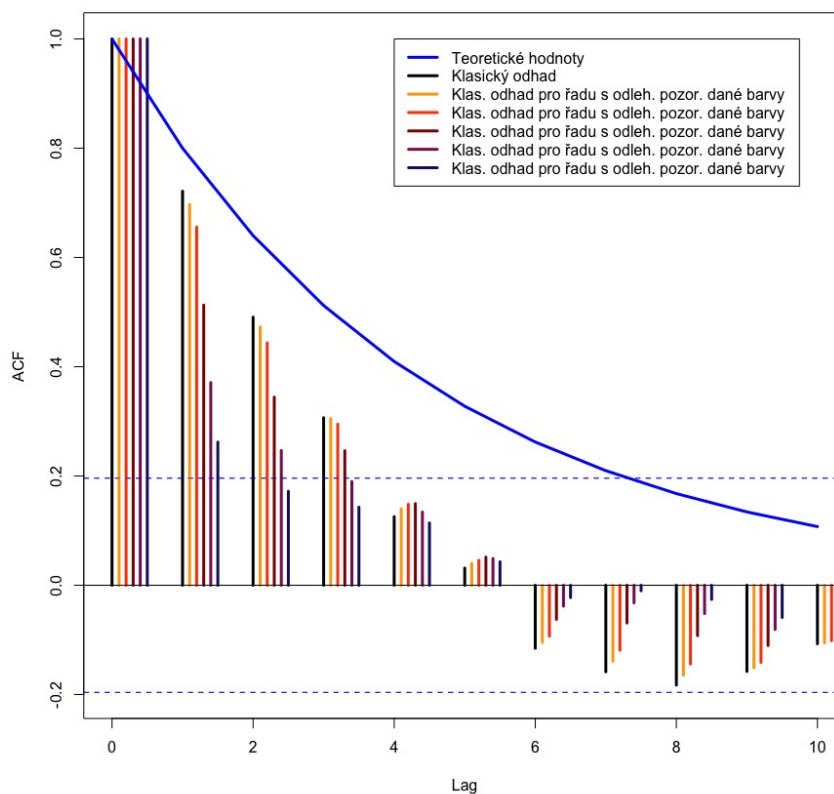
Spočítáme opět klasický odhad autokorelační funkce a porovnáme s teoretickými hodnotami. Vykreslené časové řady a porovnání klasického odhadu s teoretickými hodnotami nalezneme na obrázku 2.3. V tomto konkrétním případě nám klasický odhad pro MA(2) vychází velmi podobně jako teoretické hodnoty. Pro AR(1) klasický odhad dává nižší hodnoty, než jsou teoretické, a klesá k nule mnohem rychleji než teoretické hodnoty.

Příklad. Podíváme se na chování klasického odhadu autokorelační funkce řady s jedním odlehlým pozorováním jehož hodnotu budeme měnit. Tím budeme simulovat vliv zvyšující se odlehlosti na klasický odhad dle vztahu (2.4).

Výstup nalezneme na obrázku 2.4. Daná řada má pouze jedno odlehlé pozorování. S vyšší odlehlostí se stává klasický odhad nevýznamný.



(a) AR(1)



(b) Klasický odhad ACF pro AR(1)

Obrázek 2.4: Porovnání klasického odhadu časové řady s jedním odlehlým pozorováním měnící velikost. Časová řada bez odlehlého pozorování AR(1) je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Odlehlé pozorování se nachází v bodě $t = 5$ a postupně nabývá hodnoty odlehlosti $X_t = 3, 5, 10, 15, 20$.

3. Robustní statistika

V této práci se zabýváme robustním odhadem autokorelační funkce, proto by bylo dobré představit základy robustní statistiky. Zmíněné poznatky nám umožní lépe nahlédnout na danou problematiku a některé obecné předpisy budeme dále využívat ve specifickém tvaru. Hned v úvodu poznamenejme, že zde uvedeme pouze základní přehled robustních metod. Pro podrobnější popis, vlastnosti a příklady je třeba se podívat do odborné literatury. Čerpáme z literatury (Maronna a kol., 2006).

Robustní metody mají zajistit adekvátnost odhadu v situacích, kdy data obsahují odlehlá pozorování. Odlehlým pozorováním myslíme pozorování, které neodpovídá struktuře zkoumaného modelu. Níže přesněji definujeme a popíšeme si typy odlehlých pozorování. Odlehlé pozorování musíme brát vždy vzhledem k určitému modelu. Co je odlehlé pro jeden model, může být typickým projevem jiného modelu. Příkladem mohou být rozdělení s těžkými chvosty jako Cauchyovo nebo Studentovo rozdělení. Klasické metody jsou značně citlivé na odlehlá pozorování a výsledný odhad poté může být zavádějící a neadekvátní. Robustní metody řeší, jakým způsobem zmírnit dopad odlehlých pozorování a získat použitelný odhad. Příkladem může být robustní odhad střední hodnoty, rozptylu či korelace.

Na druhou stranu není vhodné klasické metody permanentně nahradit robustními. V případě, kdy data neobsahují odlehlá pozorování, robustní odhady nemají takovou eficienci jako klasické odhady. Použitím robustních metod vždy ztratíme či zkreslíme informace pocházející z původních dat.

Pracujeme s časovou řadou $\{X_t, t \in T\}$ a nechť všechny náhodné veličiny časové řady pocházejí ze stejného rozdělení. Definujeme si *kontaminované rozdělení*.

Definice 16 (Kontaminované rozdělení). *Necheť $\{X_t, t \in T\}$ je časová řada. Řekneme, že náhodná veličina X_t pro $t \in T$ pochází z kontaminovaného rozdělení G a H s váhami $1 - \varepsilon$ a ε , pokud pro $t \in T$*

$$X_t = U_t G_t + (1 - U_t) H_t,$$

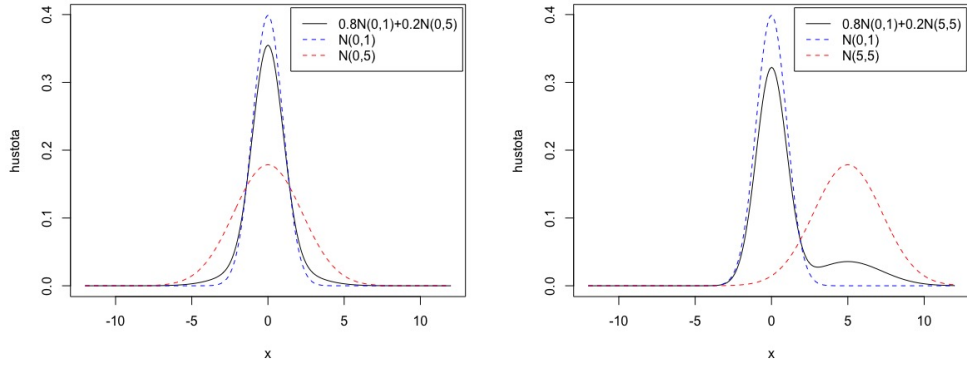
kde U_t má alternativní rozdělení s $P(U_t = 1) = 1 - \varepsilon$ a $P(U_t = 0) = \varepsilon$, $\varepsilon \in (0, 1)$, G_t je náhodná veličina pocházející z rozdělení G , H_t je náhodná veličina pocházející z rozdělení H a náhodné veličiny U_t , G_t a H_t jsou nezávislé.

Daná náhodná veličina X_t je směsí dvou náhodných veličin. Náhodnou veličinu G_t budeme pokládat za základní. H_t je náhodná veličina, pomocí které můžeme modelovat odlehlá pozorování nebo těžké chvosty. Náhodná veličina U_t slouží k modelování přechodu mezi rozděleními G_t a H_t .

Nyní odvodíme distribuční funkci pro náhodnou veličinu pocházející z kontaminovaného rozdělení.

Věta 12 (Distribuční funkce kontaminovaného rozdělení). *Necheť náhodná veličina X_t pro $t \in T$ pochází z kontaminovaného rozdělení G a H s váhami $1 - \varepsilon$ a ε . $G(x)$ je distribuční funkce rozdělení G a $H(x)$ je distribuční funkce rozdělení H . Distribuční funkce kontaminovaného rozdělení G a H s váhami $1 - \varepsilon$ a ε je*

$$P(X_t \leq x) = (1 - \varepsilon)G(x) + \varepsilon H(x).$$



(a) Hustota pro kontaminované rozdělení $G \sim N(0,1)$ a $H \sim N(0,5)$ s váhami 0.8 a 0.2 (b) Hustota pro kontaminované rozdělení $G \sim N(0,1)$ a $H \sim N(5,5)$ s váhami 0.8 a 0.2

Obrázek 3.1: Hustota kontaminovaného rozdělení $G \sim N(0,1)$ a $H \sim N(0,5)$ s váhami 0.8 a 0.2 a rozdělení $G \sim N(0,1)$ a $H \sim N(5,5)$ s váhami 0.8 a 0.2.

Důkaz. Dokážeme z definice 16.

$$\begin{aligned} P(X_t \leq x) &= P(U_t = 1)P(X_t \leq x|U_t = 1) + P(U_t = 0)P(X_t \leq x|U_t = 0) = \\ &= (1 - \varepsilon)G(x) + \varepsilon H(x), \end{aligned}$$

kde $G(x)$ je distribuční funkce rozdělení G a $H(x)$ je distribuční funkce rozdělení H . □

Příklad. Uvedeme dva příklady kontaminovaného rozdělení a graf jejich hustot nalezneme na obrázku 3.1. Zvolíme $\varepsilon = 0.20$. Pak s pravděpodobností 80 % má náhodná veličina základní rozdělení dané náhodnou veličinou G_t a s pravděpodobností 20 % dojde k modelování pomocí H_t , které může reprezentovat odlehlá pozorování.

Prvním příkladem je kontaminované rozdělení s $G_t \sim N(0,1)$ a $H_t \sim N(0,5)$. Jeho hustotu nalezneme na obrázku 3.1a. Zde pozorujeme pouze změnu rozptylu mezi rozděleními. Druhým příkladem je rozdělení s $G_t \sim N(0,1)$ a $H_t \sim N(5,5)$. Jeho hustotu nalezneme na obrázku 3.1b. Oproti předchozímu příkladu, zde pozorujeme i posun střední hodnoty. Výsledné rozdělení je bimodální.

V celé kapitole pracujeme s náhodnými veličinami $\mathbf{X} = (X_1, \dots, X_n)^\top$ o rozsahu n .

3.1 Robustní odhad střední hodnoty

Klasickým odhadem střední hodnoty je výběrový průměr dán předpisem $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$. Pro data pocházející ze symetrického rozdělení lze robustně odhadnout střední hodnotu pomocí výběrového mediánu, který značíme $\widehat{\text{med}}(\mathbf{X})$. Ten není tolik náchylný na odlehlá pozorování. Daný odhad založený na výběrovém

mediánu je $\hat{\mu}^{(med)} = \widehat{\text{med}}(\mathbf{X})$. Ukážeme si to na primitivní posloupnosti pozorování (3,2,1,3,1), kde výběrový průměr i výběrový medián jsou 2. Pro data s jedním odlehlým pozorováním na páté pozici (3,2,1,3,71) je výběrový průměr 16 a výběrový medián 3.

Další možností je *odhad pomocí oříznutí*. Zvolíme koeficient β z intervalu $[0, 0.5]$. Symbol $\lfloor \beta n \rfloor$ reprezentuje dolní celou část βn a $X_{(k)}$ je k -tá pořádková statistika. Robustní odhad získáme vynecháním největších a nejmenších hodnot, je dán předpisem

$$\hat{\mu}^{(\beta)} = \frac{1}{n - 2\lfloor \beta n \rfloor} \sum_{k=\lfloor \beta n \rfloor + 1}^{n - \lfloor \beta n \rfloor} X_{(k)}.$$

Obecnějším odhadem je takzvaný *L-odhad* daný předpisem

$$\hat{\mu}^{(L)} = \sum_{k=1}^n w_k X_{(k)},$$

kde pro $k = 1, \dots, n$ jsou w_k dané váhy. Většinou se berou následující váhy viz (Maronna a kol., 2006), pro každé $k = 1, \dots, n$ platí $w_k \geq 0$ (kladné váhy), $w_k = w_{n-k+1}$ (symetrie) a $\sum_{k=1}^n w_k = 1$ (součet vah je jedna). Pro odhad $\hat{\mu}^{(\beta)}$ se jedná o váhy

$$w_k = \frac{1}{n - 2\lfloor \beta n \rfloor} I[\lfloor \beta n \rfloor + 1 \leq i \leq n - \lfloor \beta n \rfloor],$$

kde

$$I[\lfloor \beta n \rfloor + 1 \leq i \leq n - \lfloor \beta n \rfloor] = \begin{cases} 1, & \lfloor \beta n \rfloor + 1 \leq i \leq n - \lfloor \beta n \rfloor \\ 0, & \text{jinak.} \end{cases}$$

3.2 Robustní odhad polohy

Pracujeme s náhodnými veličinami

$$X_t = \mu + Y_t, \quad t = 1, \dots, n, \quad (3.1)$$

kde $\mu \in \mathbb{R}$ je neznámý parametr *polohy* a pro $t = 1, \dots, n$ jsou Y_t náhodné veličiny reprezentující chybovou složku mající distribuční funkci F_0 a hustotu f_0 . Předpokládejme, že pro $t = 1, \dots, n$ jsou X_t nezávislé a stejně rozdělené náhodné veličiny s distribuční funkcí F . Platí vztah $F(x) = F_0(x - \mu)$. Cílem je odhadnout μ . Ve speciálním případě se jedná i o odhad střední hodnoty.

Věrohodnostní funkce je dána předpisem

$$L(X_1, \dots, X_n, \mu) = \prod_{k=1}^n f_0(X_k - \mu).$$

MLE-odhad $\hat{\mu}^{(MLE)}$ parametru μ je dán vztahem

$$\hat{\mu}^{(MLE)} = \arg \max_{\mu \in \mathbb{R}} L(X_1, \dots, X_n, \mu).$$

Po zlogaritmování dostaneme

$$\log L(X_1, \dots, X_n, \mu) = \sum_{k=1}^n \log f_0(X_k - \mu).$$

Pokud je $\log f_0$ diferencovatelná, zderivujeme ji podle μ a položíme rovnou nule. *MLE-odhad* získáme řešením

$$\sum_{k=1}^n \frac{f_0'(X_k - \hat{\mu}^{(MLE)})}{f_0(X_k - \hat{\mu}^{(MLE)})} = 0,$$

Postup odhadu si ukážeme na příkladu rozdělení $N(0, 1)$, které má hustotu

$$f_0(y) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{y^2}{2}}, \quad y \in \mathbb{R}.$$

Pro data $\{X_t, t = 1, \dots, n\}$ maximalizujeme věrohodnostní funkci

$$L(X_1, \dots, X_n, \mu) = \prod_{k=1}^n f_0(X_k - \mu) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp^{-\sum_{k=1}^n \frac{(X_k - \mu)^2}{2}}.$$

Po zlogaritmování a zderivování dostaneme odhad

$$\hat{\mu}^{(MLE)} = \frac{1}{n} \sum_{k=1}^n X_k.$$

Ve věrohodnostní funkci pracujeme s kvadráty hodnot pozorování, což může ještě umocnit vliv odlehlých pozorování. *MLE-odhad* střední hodnoty je aritmetický průměr pozorování, o kterém víme, že je náchylný na odlehlá pozorování. Z tohoto důvodu byl navržen *M-odhad*, který má být robustnější.

M-odhad parametru polohy μ obecně získáme řešením

$$\hat{\mu}^{(M)} = \arg \min_{\mu \in \mathbb{R}} \Upsilon(X_k - \mu),$$

kde Υ je námi volená funkce. Pro Υ diferencovatelnou dostaneme $\Psi = \Upsilon'$. Odhad získáme řešením rovnice

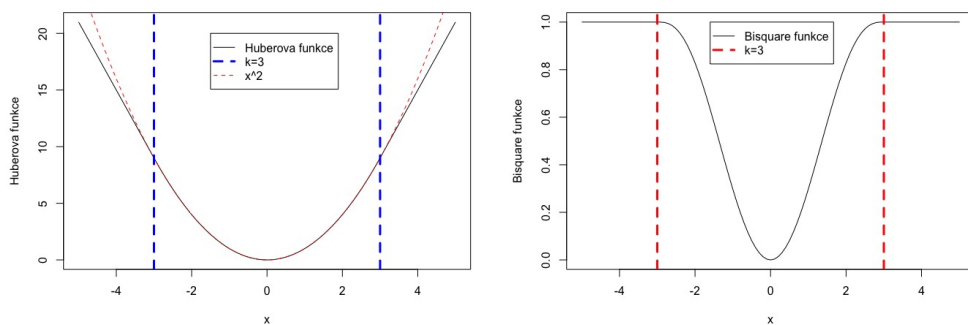
$$\sum_{k=1}^n \Psi(X_k - \hat{\mu}) = 0.$$

Nyní se podíváme na vhodné volby Υ . *Huberova funkce* je dána pro zvolené $k \in \mathbb{R}$ předpisem

$$\Upsilon_k(x) = \begin{cases} x^2, & |x| \leq k, \\ 2k|x| - k^2, & |x| > k, \end{cases}$$

s derivací $\Upsilon_k'(x) = 2\Psi_k(x)$, kde

$$\Psi_k(x) = \begin{cases} x, & |x| \leq k, \\ \text{sgn}(x)k, & |x| > k. \end{cases}$$



(a) Huberova funkce pro $k = 3$.

(b) Bisquare funkce pro $k = 3$.

Obrázek 3.2: Graf Huberovy a bisquare funkcí pro $k = 3$.

Další možností je *bisquare funkce*

$$\Upsilon_k(x) = \begin{cases} 1 - (1 - (x/k)^2)^3, & |x| \leq k, \\ 1, & |x| > k, \end{cases}$$

s derivací $\Upsilon'_k(x) = \frac{6\Psi(x)}{k^2}$, kde

$$\Psi_k(x) = \begin{cases} x \left(1 - \left(\frac{x}{k}\right)^2\right)^2, & |x| \leq k, \\ 0, & |x| > k. \end{cases}$$

Graf Huberovy a bisquare funkcí nalezneme na obrázku 3.2.

3.3 Robustní odhad charakteristiky variability

Základním ukazatelem variability náhodné veličiny je rozptyl a směrodatná odchylka. Klasickým odhadem směrodatné odchylky je odmocnina z výběrového rozptylu. Odhad je dán předpisem $\hat{\sigma} = \left(\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2\right)^{\frac{1}{2}}$.

Prvním robustním odhadem variability je *střední absolutní odchylka* (anglicky *mean absolute deviation*) dána předpisem

$$MD(\mathbf{X}) = \frac{1}{n} \sum_{k=1}^n |X_k - \bar{X}|.$$

Nejedná se přímo o odhad rozptylu, ale o odhad $E|X_k - EX_k|$. Pro získání odhadu rozptylu či směrodatné odchylky je třeba MD přenásobit vhodnou konstantou, která závisí na rozdělení dat.

Dalším robustním odhadem je *odhad mediánové absolutní odchylky od mediánu* (anglicky *median absolute deviation about the median*) dán předpisem

$$MAD(\mathbf{X}) = \widehat{\text{med}}(|\mathbf{X} - \widehat{\text{med}}(\mathbf{X})|).$$

Jedná se o odhad $\widehat{\text{med}}(|X_t - \widehat{\text{med}}(X_t)|)$.

Zmíňme *mezikvartilové rozpětí* (anglicky *interquartile range*) dané předpisem

$$IQR(\mathbf{X}) = Q_3(\mathbf{X}) - Q_1(\mathbf{X}),$$

kde Q_3 je horní (75%) kvartil a Q_1 je dolní (25%) kvartil. Poznamenejme, že existuje více definic výběrových kvartilů a závisí na naší volbě.

Jak bylo výše zmíněno, tyto odhady nejsou přímo odhadem směrodatné odchylky, je třeba je přenásobit vhodnou konstantou závislou na rozdělení. Podíváme se na robustní odhady směrodatné odchylky za pomoci MD a MAD. Pro data pocházející z $N(\mu, \sigma^2)$ dle (Maronna a kol., 2006, kapitola 2.4) a článku (Rousseeuw a Croux, 1993) platí

$$\begin{aligned}\hat{\sigma}^{(MD)} &= \sqrt{\frac{\pi}{2}} \times MD(\mathbf{X}) \doteq 1.2533 \times MD(\mathbf{X}), \\ \hat{\sigma}^{(MAD)} &= \frac{1}{u_{0.75}} \times MAD(\mathbf{X}) \doteq 1.4826 \times MAD(\mathbf{X}),\end{aligned}$$

kde $u_{0.75}$ je 0.75 kvantil normovaného normálního rozdělení a $\hat{\sigma}^{(MD)}$ a $\hat{\sigma}^{(MAD)}$ reprezentuje příslušné robustní odhady směrodatné odchylky. Pokud je rozdělení jiné než normální, tak jsou konstanty samozřejmě jiné.

Odvození *M-odhadu* a dalších odhadů pro směrodatnou chybu či rozptyl nalezneme například v (Maronna a kol., 2006).

Podíváme se na obecnější přístup k variabilitě. Pracujeme s modelem

$$X_t = \sigma Y_t, \quad t = 1, \dots, n,$$

kde Y_t jsou nezávislé stejně rozdělené náhodné veličiny a $\sigma > 0$ je parametrem *měřítko* (anglicky *scale*).

Měřítko je obecnější charakteristikou volatility náhodné veličiny. Můžeme se zabývat měřítkem, i když Y_t bude pocházet z rozdělení bez konečných druhých momentů. Tedy rozptyl poté není konečný pro X_t . Pokud existuje rozptyl Y_t a je konečný, tedy $0 < \text{var} Y_t < \infty$, pak $\text{var} X_t = \sigma^2 \text{var} Y_t$.

Předpokládáme, že Y_t má hustotu $f_0(x)$, pak X_t má hustotu $\frac{1}{\sigma} f_0\left(\frac{x}{\sigma}\right)$. Cílem je odhadnout σ .

Věrohodnostní funkce je

$$L(X_1, \dots, X_n, \sigma) = \prod_{k=1}^n \frac{1}{\sigma} f_0\left(\frac{X_k}{\sigma}\right).$$

MLE-odhad $\hat{\sigma}^{(MLE)}$ měřítko σ je dán vztahem

$$\hat{\sigma}^{(MLE)} = \arg \max_{\sigma \in (0, \infty)} L(X_1, \dots, X_n, \sigma).$$

Po zlogaritmování dostaneme

$$\log L(X_1, \dots, X_n, \sigma) = -n \log(\sigma) + \sum_{k=1}^n \log\left(f_0\left(\frac{X_k}{\sigma}\right)\right).$$

Pokud je $\log(f_0)$ diferencovatelná, zderivujeme ji podle σ a položíme rovnou nule

$$\frac{-n}{\hat{\sigma}^{(MLE)}} + \sum_{k=1}^n \frac{-f_0'\left(\frac{X_k}{\hat{\sigma}^{(MLE)}}\right)}{f_0\left(\frac{X_k}{\hat{\sigma}^{(MLE)}}\right)} \frac{X_k}{(\hat{\sigma}^{(MLE)})^2} = 0.$$

Pak náš odhad $\hat{\sigma}^{(MLE)}$ parametru σ musí splňovat následující rovnost

$$\frac{1}{n} \sum_{k=1}^n \frac{-f'_0\left(\frac{X_k}{\hat{\sigma}^{(MLE)}}\right)}{f_0\left(\frac{X_k}{\hat{\sigma}^{(MLE)}}\right)} \frac{X_k}{\hat{\sigma}^{(MLE)}} = 1.$$

Postup odhadu si ukážeme na příkladu rozdělení $N(0, 1)$, které má hustotu

$$f_0(y) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{y^2}{2}}, \quad y \in \mathbb{R}.$$

Pro data $\{X_t, t = 1, \dots, n\}$ maximalizujeme věrohodnostní funkci

$$L(X_1, \dots, X_n, \sigma) = \prod_{k=1}^n \frac{1}{\sigma} f_0\left(\frac{X_k}{\sigma}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp^{-\sum_{k=1}^n \left(\frac{X_k^2}{2\sigma^2}\right)}.$$

Po zlogaritmování a zderivování dostaneme odhad

$$\hat{\sigma}^{(MLE)} = \sqrt{\frac{1}{n} \sum_{k=1}^n X_k^2}.$$

Opět se nám objevují kvadráty hodnot pozorování. Z tohoto důvodu byl navržen *M-odhad*, který má být robustnější.

Obecný *M-odhad* $\hat{\sigma}^{(M)}$ parametru měřítka σ je řešením rovnice

$$\frac{1}{n\delta} \sum_{k=1}^n \Upsilon\left(\frac{X_k}{\hat{\sigma}^{(M)}}\right) = 1,$$

kde Υ je funkce a $\delta > 0$ je konstanta.

Pro MLE-odhad je

$$\Upsilon(x) = x \frac{-f'_0(x)}{f_0(x)}$$

a $\delta = 1$.

Jako jiný příklad uvedme *bisquare měřítko*, kde je

$$\Upsilon(x) = \min\left(1 - (1 - x^2)^3, 1\right)$$

a $\delta = 0.5$.

Další možností je použít odhad pomocí váhové funkce, kterou nalezneme například v (Maronna a kol., 2006).

3.4 Robustní statistika pro lineární regresi

V této části se budeme zabývat robustními metodami v lineární regresi, jelikož poté robustní odhady regresní funkce použijeme k odhadu autokorelační funkce. Odhad regresní funkce metodou nejmenších čtverců je také velmi náchylný na odlehlá pozorování, která mohou způsobit zcela zcestný tvar a chování odhadu regresní funkce. Například místo klesajícího lineárního trendu nám odhad dá rostoucí lineární trend jako na obrázku 3.3.

Nejdříve si definujeme základní pojmy, budeme čerpat z (Anděl, 2005).

Definice 17 (Lineární regresní model). *Nechť \mathbf{Y} je vektor odpovědí velikosti $n \times 1$, reálná matice regresorů $\mathbf{X} = (x_{i,j})_{k,l=1}^{n,m}$ s rozměry $n \times m$ nenáhodných čísel a $m < n$. Nechť platí $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kde $\boldsymbol{\beta} \in \mathbb{R}^m$ je vektor neznámých parametrů a $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ je vektor nezávislých stejně rozdělených náhodných chyb splňujících $E\varepsilon_k = 0$ a $0 < \text{var} \varepsilon_k = \sigma^2 < \infty$ a σ^2 je neznámý parametr. Tento model je lineární regresní model.*

Lineární regresní model je normální, pokud pro $k = 1, \dots, n$ platí pro chybovou složku $\varepsilon_k \sim N(0, \sigma^2)$. Odhad $\hat{\boldsymbol{\beta}}$ je odhadem vektoru parametrů $\boldsymbol{\beta}$. Vektorem vyrovnaných hodnot je $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ a vektor reziduí je $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$.

Naměřená data jsou uspořádány do dvojic, k -té pozorování (\mathbf{x}_k^\top, y_k) se skládá z reálného nenáhodného vektoru regresorů \mathbf{x}_k a jednorozměrné reálné náhodné odezvy y_k .

LS-odhad vektoru parametrů $\boldsymbol{\beta}$ je získaný metodou nejmenších čtverců řešením

$$\hat{\boldsymbol{\beta}}^{(LS)} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \sum_{k=1}^n (Y_k - \mathbf{X}_k^\top \boldsymbol{\beta})^2,$$

kde \mathbf{X}_k je k -tý řádek matice \mathbf{X} , odpovídající k -tému pozorování.

Tento odhad lze spočítat za splnění podmínky, že \mathbf{X} má lineárně nezávislé sloupce, pomocí následující věty.

Věta 13 (Odhad metodou nejmenších čtverců). *Nechť matice \mathbf{X} má lineárně nezávislé sloupce. Odhad metodou nejmenších čtverců je*

$$\hat{\boldsymbol{\beta}}^{(LS)} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}.$$

Důkaz. Dle Anděl, 2005 (Věta 5.1). □

Vlastnosti a chování LS-odhadu či testování hypotéz a intervaly spolehlivosti pomocí metody nejmenších čtverců jsou náplní každého kurzu regrese, proto zde zmíníme jen úplné základy. Podrobnější informace nalezneme například v (Zvára, 2008) a (Khuri, 2010).

Pokud má matice \mathbf{X} plnou hodnotu, tedy $\text{rank}(\mathbf{X}) = m$, pak dostaneme pro odhad $\hat{\boldsymbol{\beta}}^{(LS)}$ vlastnosti

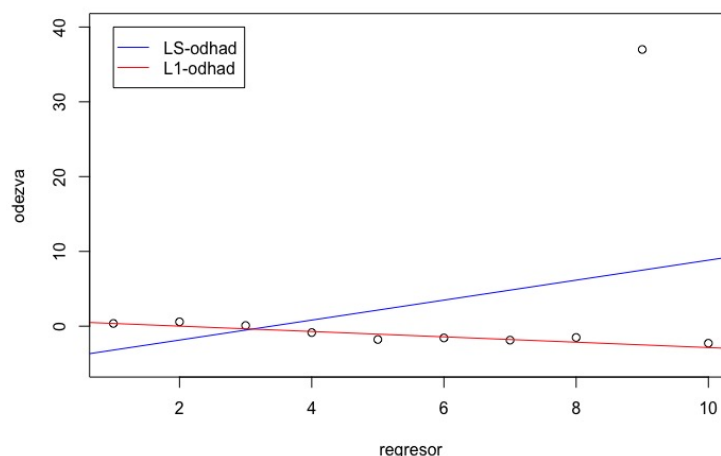
$$\begin{aligned} E\hat{\boldsymbol{\beta}}^{(LS)} &= \boldsymbol{\beta} \\ \text{var}(\hat{\boldsymbol{\beta}}^{(LS)}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Pokud je model normální, pak

$$\hat{\boldsymbol{\beta}}^{(LS)} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

LS-odhad je velmi náchylný na odlehlá pozorování. Proto je třeba zavést robustnější. Tím může být například *L1-odhad* vektoru parametrů $\boldsymbol{\beta}$, který získáme řešením

$$\hat{\boldsymbol{\beta}}^{(L1)} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \sum_{k=1}^n |Y_k - \mathbf{X}_k^\top \boldsymbol{\beta}|.$$



Obrázek 3.3: LS-odhad a L1-odhad regresní přímky pro data s jedním odlehlým pozorováním.

Na obrázku 3.3 můžeme pozorovat porovnání LS-odhadu a L1-odhadu regresní přímky pro data s jedním odlehlým pozorováním. Data vykazují mírný klesající trend, který je L1-odhadem regresní přímky detekován. Zato LS-odhad nám dá znatelný rostoucí trend.

Obecný robustní M -odhad vektoru parametrů β získáme řešením

$$\hat{\beta}^{(M)} = \arg \min_{\beta \in \mathbb{R}^m} \sum_{k=1}^n \Upsilon \left(\frac{Y_k - \mathbf{X}_k \beta}{\hat{\sigma}} \right),$$

kde $\hat{\sigma}$ je robustní odhad měřítka.

V případě diferencovatelnosti Υ pracujeme s $\Psi = \Upsilon'$ a řešíme vztah

$$\sum_{k=1}^n \Psi \left(\frac{Y_i - \mathbf{X}_k^\top \hat{\beta}^{(M)}}{\hat{\sigma}} \right) \mathbf{X}_k = \mathbf{0}.$$

3.5 Struktura odlehlosti pro časové řady

Časové řady se vyznačují závislostí (korelovaností) mezi jednotlivými hodnotami. Výskyt odlehlých pozorování má nejen vliv na odhad parametrů či autokorelační funkce, ale i na naměřené budoucí hodnoty v některých modelech. Budeme pracovat se slabě stacionární časovou řadou.

Odlehlá pozorování (anglicky *outliers*) se liší způsobem své odlehlosti a struktury. Zmíníme některé případy, čím může být odlehlost způsobena. Následující terminologie je naším zobecněním informací čerpaných z (Maronna a kol., 2006). Odlehlé pozorování může být *aditivní*, kdy jeho vliv nepůsobí na jiné pozorování časové řady. *Primární inovační* odlehlé pozorování svou odlehlostí zapříčiní odlehlost budoucích pozorování časové řady a jeho odlehlost není zapříčiněna jiným odlehlým pozorováním v minulosti časové řady. *Sekundární inovační* odlehlé pozorování je takové pozorování, jehož odlehlost je zapříčiněna odlehlostí některého z pozorování v minulosti časové řady. Inovační odlehlost je častým jevem u AR

a ARMA procesů. Ukázku těchto odlehlých pozorování nalezneme na obrázku 3.4. Odlehlé pozorování s *posunem* pochází ze stejného rozdělení jako zbytek řady, jen s jinou střední hodnotou. Odlehlé pozorování s *jiným rozptylem* pochází ze stejného rozdělení jako zbytek řady, jen s jiným rozptylem. Ukázku těchto odlehlých pozorování nalezneme na obrázku 3.5. Odlehlost tedy může být zapříčiněna mnoha okolnostmi a může mít nejrůznější vliv na časovou řadu a ostatní pozorování.

Definujeme si formálně odlehlá pozorování.

Definice 18 (Odehlá pozorování). Časová řada s aditivními odlehlými pozorováními $\{X_t, t \in \mathbb{Z}\}$ je dána předpisem

$$X_t = U_t G_t + (1 - U_t) H_t,$$

kde $\{U_t, t \in \mathbb{Z}\}$ U_t je slabě stacionární časová řada 0 a 1, pro každé $t \in \mathbb{Z}$, U_t má alternativní rozdělení s $P(U_t = 1) = 1 - \varepsilon$ a $P(U_t = 0) = \varepsilon$, $\varepsilon \in (0, 1)$, $\{G_t, t \in \mathbb{Z}\}$ je slabě stacionární časová řada a $\{H_t, t \in \mathbb{Z}\}$ je časová řada.

Časová řada s inovačními odlehlými pozorováními je dána předpisem

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_m X_{t-m} + \tilde{Y}_t + \theta_1 \tilde{Y}_{t-1} + \dots + \theta_n \tilde{Y}_{t-n}, \quad t \in \mathbb{Z}$$

kde $m, n \in \mathbb{N}_0$ a $\{\tilde{Y}_t, t \in \mathbb{Z}\}$ je časová řada, kde pro každé $t \in \mathbb{Z}$, \tilde{Y}_t pochází z kontaminovaného rozdělení dle definice 16, kde G_t pochází z $WN(0, \sigma^2)$.

Pozorujeme data z časové řady s odlehlými pozorováními.

Řada s *aditivním odlehlým pozorováním* je obdobou řady pocházející z kontaminovaného rozdělení dle definice 16, kde specifikujeme $\{G_t, t \in T\}$ jako stacionární časovou řadu a nevyžadujeme nezávislost U_t a H_t . H_t reprezentuje kontaminaci odlehlými pozorováními a U_t reprezentuje, kdy ke kontaminaci dochází.

Inovační odlehlé pozorování vzniká modifikací bílého šumu v ARMA modelu. Poznamenejme, že zde dané odlehlé pozorování ovlivňuje svou odlehlostí budoucí hodnoty časové řady, záleží na struktuře řady.

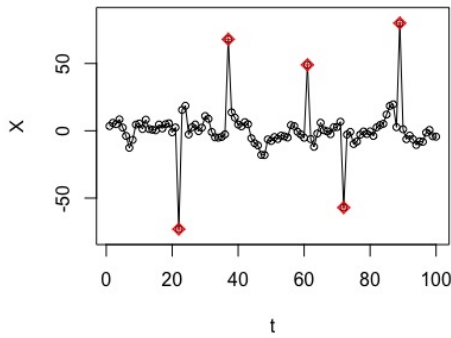
Odehlá pozorování můžeme rozlišovat i podle jejich výskytu vzhledem k časové řadě. Časovou řadu lze nazvat *se symetrickými odlehlými pozorováními*, pokud počet odlehlých pozorování nad a pod střední hodnotou X_t je přibližně stejný. Pojem symetrie, ale nebudeme formálně definovat.

Příklad. Podíváme se na příklad aditivních odlehlých pozorování. Necht $\{G_t, t \in T\}$ je stacionární časová řada, $m \in \mathbb{N}$ je počet odlehlých pozorování, $t_1, \dots, t_m \in T$ jsou časy výskytu odlehlých pozorování, $\omega_1, \dots, \omega_m \in \mathbb{R}$, $\mathbf{I}(t)$ je indikátor výskytu odlehlého pozorování daný předpisem

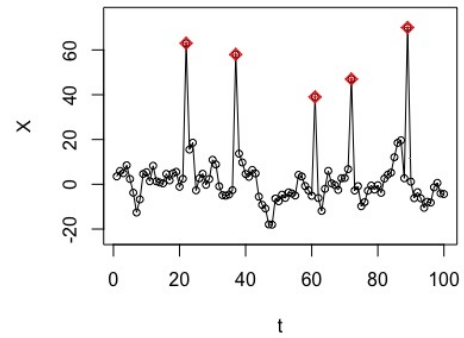
$$\mathbf{I}(t) = \begin{cases} 1, & t = t_1, \dots, t_m, \\ 0, & \text{jinak.} \end{cases}$$

Pak časová řada s aditivními odlehlými pozorováními je

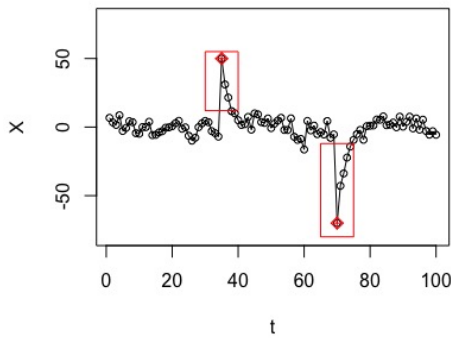
$$X_t = G_t + \sum_{i=1}^m \omega_i \mathbf{I}(t).$$



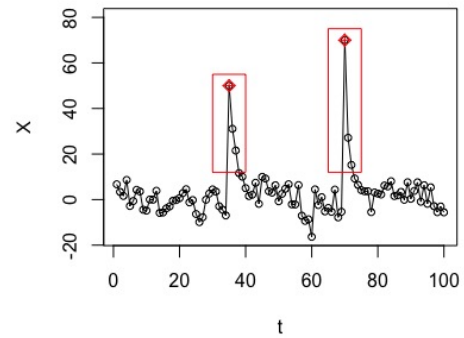
(a) Symetrická aditivní odlehlá pozorování



(b) Nesymetrická aditivní odlehlá pozorování

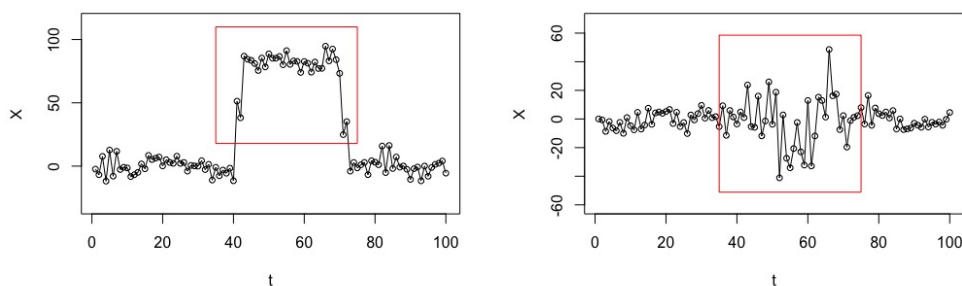


(c) Symetrická inovační odlehlá pozorování



(d) Nesymetrická inovační odlehlá pozorování

Obrázek 3.4: Časová řada se symetrickými a nesymetrickými aditivními a inovačními odlehlými pozorováními. Symetrická odlehlá pozorování jsou vychýlena na obě strany oproti řadě, tedy nabývají nejmenších i největších hodnot. Nesymetrická odlehlá pozorování jsou vychýlena pouze na jednu stranu oproti řadě, tedy buď všechna nabývají největších hodnot nebo naopak nejmenších hodnot. Aditivní odlehlá pozorování neovlivňují další pozorování. Inovační odlehlá pozorování působí a ovlivňují i budoucí pozorování a způsobují i jejich odlehnutí.



(a) Odlehlá pozorování s posunem střední hodnoty (b) Odlehlá pozorování s jiným rozptylem

Obrázek 3.5: Časová řada s odlehlými pozorováními s posunem a jiným rozptylem.

V tomto případě jsou hodnoty U_t a H_t určeny deterministicky, kdy

$$U_t = \begin{cases} 0, & t = t_1, \dots, t_m, \\ 1, & \text{jinak,} \end{cases}$$

a pro $j \in \{1, \dots, m\}$

$$H_t = \begin{cases} G_t + \omega_t, & t = t_1, \dots, t_m, \\ \text{libovolná hodnota,} & \text{jinak,} \end{cases}$$

Příklad. V dalším příkladě se podíváme na časovou řadu $\{X_t, t \in \mathbb{Z}\}$ s inovačními odlehlými pozorováními

$$X_t = 0.9X_{t-1} + 0.3X_{t-2} + \tilde{Y}_t + 0.6\tilde{Y}_{t-1} - 0.1\tilde{Y}_{t-2}, \quad t \in \mathbb{Z}$$

kde $\{\tilde{Y}_t, t \in \mathbb{Z}\}$ je modifikovaný bílý šum a pochází z kontaminovaného rozdělení dle definice 16 a je dán předpisem

$$\tilde{Y}_t = U_t G_t + (1 - U_t) R_t V_t,$$

kde G_t je $WN(0,1)$, R_t je z rovnoměrného rozdělení $U(2,4)$ a

$$V_t = 2Z_t - 1,$$

kde Z_t je z alternativního rozdělení $Alt(0.5)$. U_t pochází z alternativního rozdělení $Alt(0.9)$.

Časová řada G_t reprezentuje slabě stacionární bílý šum bez odlehlých pozorování. U_t určuje zda v daném čase t dojde k odlehlosti. R_t nám dává velikost odlehlosti. V_t zajišťuje symetrické rozmístění odlehlých pozorování. Tím vzniká modifikovaný bílý šum, který obsahuje odlehlost a přenáší ji do časové řady $\{X_t, t \in \mathbb{Z}\}$.

Příklad vlivu odlehlého pozorování na odhad autokorelační funkce je uveden v kapitole 2.2. Robustní statistika se také zabývá robustním odhadem parametrů ARMA modelu, který ale není předmětem této práce. Nebudeme je zde proto uvádět. Popis vhodných metod nalezneme například v (Maronna a kol., 2006, kapitola 8).

3.6 Míra robustnosti

U zvolené metody je třeba určit její míru robustnosti, její schopnost vypořádat se s odlehlými pozorováními. Poté jsme schopni porovnat metody mezi sebou a určit, která bude vhodnější.

K tomuto lze použít *influenční funkci*, která popisuje vliv odlehlého pozorování na odhad parametru.

Dalším významným ukazatelem je *bod selhání* (anglicky *breakdown point*) reprezentující největší možné proporční zastoupení odlehlých pozorování v analyzovaném výběru tak, aby odhad stále poskytoval relevantní informace o základním rozdělení dat.

U odhadu parametru nás zajímá jeho *vychýlení* udávající rozdíl mezi střední hodnotou odhadu a odhadovaným parametrem. U robustních odhadů často řešíme *asymptotické vychýlení*. Také je velmi důležitý rozptyl odhadu, preferujeme odhad s menším rozptylem. S tímto souvisí pojem *eficience*.

Pro detekci odlehlých pozorování ve výběru dat lze použít *pravidlo 6- σ* , které využívá normovaného normálního rozdělení. V lineární regresi se používá *Cookova vzdálenost* udávající vliv daného pozorování na odhad parametrů.

Podrobnější popis měř robustnosti s odkazy na literaturu vkládáme do dodatku A.2.

4. Robustní odhad autokorelační funkce

V této kapitole se budeme zabývat různými metodami odhadu autokorelační funkce, které jsou robustní. Klasický odhad autokorelační funkce je velmi citlivý na odlehlá pozorování a v případě jejich výskytu můžeme získat nesprávný a zavádějící odhad. Cílem robustních odhadů je eliminovat nebo zmírnit vliv odlehlých pozorování na odhad autokorelační a parciální autokorelační funkce.

Příklad. Metody budeme aplikovat na simulovaná data pro časovou řadu $AR(1)$

$$X_t = 0.8X_{t-1} + Y_t, \quad t \in \mathbb{Z}, \quad (4.1)$$

s autokorelační funkcí

$$\rho(t) = 0.8^{|t|}, \quad t \in \mathbb{Z},$$

dle vzorce (2.5).

Pracujeme s aditivními odlehlými pozorováními. Odlehlost je způsobena změnou hodnoty X_t , nikoliv odlehlostí chybové složky Y_t pocházející z $N(0,1)$. Tedy se jedná o izolovanou odlehlost bez vlivu na ostatní hodnoty X_t .

Grafy časové řady nalezneme na obrázku 4.1. Graf 4.1a reprezentuje výše popsanou stacionární časovou řadu. Graf 4.1b ukazuje časovou řadu kontaminovanou symetrickými odlehlými pozorováními. Graf 4.1c vykresluje kontaminaci nesymetrickými odlehlými pozorováními, která jsou odlehlá v kladném měřítku oproti zbytku řady.

4.1 Vynechávací metoda

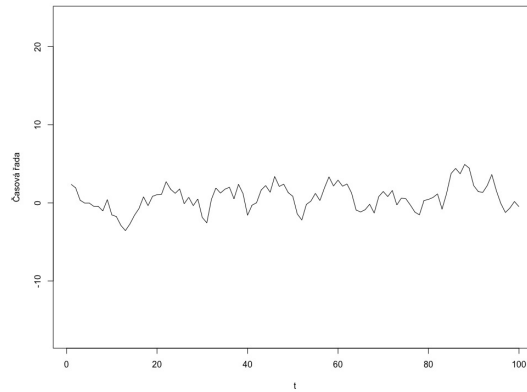
Tato metoda je velmi intuitivní. Negativní vliv odlehlých pozorování řeší jejich vynecháním. Tím nejsou použity pro výpočet odhadu a nemohou způsobit nespolehlivost výsledku. Problémem může být vynechání velkého množství pozorování a případné odstranění pozorování, která však nelze pokládat za odlehlá. Odstraňováním pozorování ztrácíme informace o dané řadě, proto je velmi důležité odstranit pouze zavádějící pozorování, které nám narušují odhad. Čerpáme z literatury (Dürre a kol., 2015) a (Chan a Wei, 1992).

Mějme časovou posloupnost (X_1, \dots, X_n) , z ní sestavíme uspořádanou verzi $X_{(1)} \leq \dots \leq X_{(n)}$. Metoda odstraní předem určený počet nejmenších a největších hodnot a ze zbylých spočítá odhad. Tím, že odstraníme nejnížší a nejvyšší hodnoty, se dá očekávat, že odstraníme odlehlá pozorování. *Robustní odhad autokovarianční funkce vynechávací metodou je*

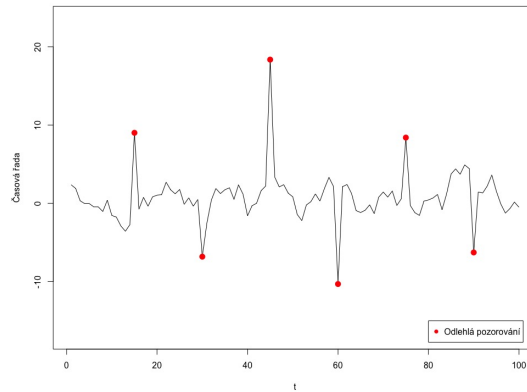
$$\hat{\gamma}^{(\beta)}(t) = \frac{1}{\sum_{k=1}^{n-t} I_k^{(\beta)} I_{t+k}^{(\beta)}} \left\{ \sum_{k=1}^{n-t} (X_k - \bar{X}^{(\beta)})(X_{t+k} - \bar{X}^{(\beta)}) I_k^{(\beta)} I_{t+k}^{(\beta)} \right\},$$

kde

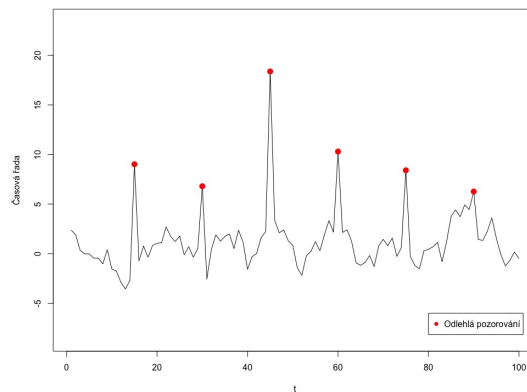
$$\bar{X}^{(\beta)} = \frac{1}{\sum_{k=1}^n I_k^{(\beta)}} \sum_{k=1}^n X_k I_k^{(\beta)},$$



(a) AR(1) bez odlehlých pozorování



(b) AR(1) s aditivními symetrickými odlehlými pozorováními



(c) AR(1) s aditivními nesymetrickými odlehlými pozorováními

Obrázek 4.1: Simulovaná data pro časovou řadu AR(1). Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$ a jsou symetrická a nesymetrická.

$$I_k^{(\beta)} = \begin{cases} 1, & X_{(\lfloor \beta n \rfloor)} < X_k < X_{(n+1-\lfloor \beta n \rfloor)}, \\ 0, & \text{jinak,} \end{cases}$$

kde $X_{(0)} = -\infty$ a $X_{(n+1)} = \infty$. Symbol $\lfloor \beta n \rfloor$ reprezentuje dolní celou část βn a $X_{(k)}$ je k -tá pořádková statistika. Koeficient β volíme z intervalu $[0, 0.5)$. Jinak bychom měli prázdný výběr měření. Pro $\beta = 0$ se jedná o klasický odhad autokovarianční funkce.

Robustní odhad autokorelační funkce vynechávací metodou je

$$\hat{\rho}^{(\beta)}(t) = \frac{\hat{\gamma}^{(\beta)}(t)}{\hat{\gamma}^{(\beta)}(0)}, \quad (4.2)$$

pokud je splněna podmínka $\hat{\gamma}^{(\beta)}(0) > 0$.

S rostoucím β se zbavujeme potenciálních odlehlých pozorování a zlepšujeme robustnost odhadu, ale hrozí snížení efektivity odhadu ve výběru bez odlehlých pozorování. Tím, že získáváme větší robustnost, současně přicházíme o část informací. V případě dat bez odlehlých pozorování dochází k odstraňování relevantních pozorování, čímž přicházíme o užitečné informace, které bychom mohli použít pro odhad. V (Chan a Wei, 1992) je doporučeno volit $\beta \in [0.01, 0.02]$ pro obecné řady, $\beta \in [0.03, 0.05]$ pro řady s objektivním drobným výskytem odlehlých pozorování a $\beta \in [0.06, 0.10]$ pro řady s častým výskytem odlehlých pozorování. V článku však není zmíněno, na základě čeho rozlišíme řady s drobným a častým výskytem. K tomu by bylo třeba se zabývat detekcí a četností odlehlých pozorování.

Budeme se zabývat asymptotickými vlastnostmi dle (Chan a Wei, 1992).

Věta 14 (Asymptotické rozdělení odhadu autokorelační funkce metodou vynechání). *Nechť $\{X_t, t = 1, \dots, n\}$ je časová řada bílého šumu s autokorelační funkcí ρ . Pak $\{\sqrt{n}\hat{\rho}^{(\beta)}(t), t = 1, 2, \dots\}$ jsou asymptoticky nezávislé náhodné veličiny s normálním rozdělením a asymptotickým rozptylem $\sigma^{(\beta)}(t)^{-1}$, kde*

$$\sigma^{(\beta)}(t) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^{n-t} I_k^{(\beta)} I_{t+k}^{(\beta)}.$$

Standardní chybu odhadu lze aproximovat

$$SE[\hat{\rho}^{(\beta)}(t)] \approx \frac{1}{\sqrt{n\hat{\sigma}_n^{(\beta)}(t)}}, \quad (4.3)$$

kde

$$\hat{\sigma}_n^{(\beta)}(t) = \frac{1}{n} \sum_{k=1}^{n-t} I_k^{(\beta)} I_{t+k}^{(\beta)}. \quad (4.4)$$

Důkaz. Viz (Chan a Wei, 1992). □

Poznamenejme, že v modelu s bílým šumem je standardní chyba pro odhad pomocí vynechání větší než pro klasický odhad. Odstraněním pozorování přijdeme

o určitou část informace o originální řadě v situaci bez odlehlých pozorování. Dle (4.4) pro $\beta > 0$ platí $\hat{\sigma}_n^{(\beta)}(t) < \hat{\sigma}_n^{(0)}(t)$. Použijeme aproximaci (4.3) a dostaneme

$$SE[\hat{\rho}^{(\beta)}(t)] \approx \frac{1}{\sqrt{n\hat{\sigma}_n^{(\beta)}(t)}} > \frac{1}{\sqrt{n\hat{\sigma}_n^{(0)}(t)}} \approx SE[\hat{\rho}^{(0)}(t)],$$

kde $\hat{\rho}^{(0)}(t)$ reprezentuje klasický odhad $\hat{\rho}(t)$, kde nedošlo k vynechání žádného pozorování.

Příklad. Podíváme se na příklad odhadu vynechávací metodou s různými volbami β . Budeme pracovat s časovou řadou $AR(1)$ v (4.1) s aditivními symetrickými odlehlými pozorováními zobrazenými na grafu 4.1b.

Pro řadu s odlehlými pozorováními spočítáme klasický odhad dle (2.2) a odhad vynechávací metodou dle (4.2) s volbami $\beta = 0.01, 0.03, 0.10$, které poté porovnáme s teoretickými hodnotami autokorelační funkce. Při volbě $\beta = 0.01$ jsou z výběru odstraněna 2 % pozorování. Simulovaná data obsahují 6 % odlehlých pozorování, tedy volba $\beta = 0.03$ odstraní všechna odlehlá pozorování. V případě $\beta = 0.01$ nám stále některá odlehlá pozorování zůstanou pro výpočet odhadu autokorelační funkce. Při volbě $\beta = 0.10$ zase odstraníme zbytečně moc pozorování a ztratíme i relevantní informace.

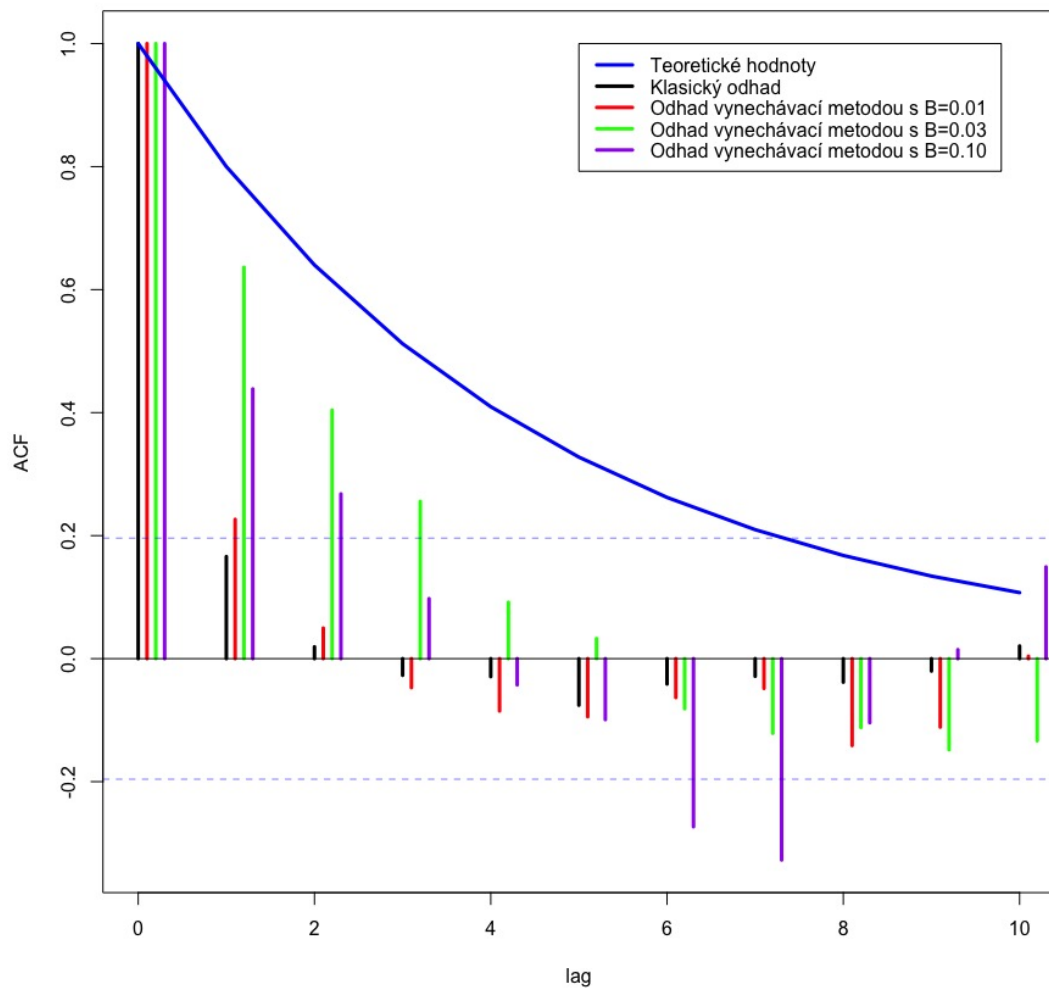
Zkoumanou časovou řadu nalezneme na obrázku 4.1 a příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou na obrázku 4.2. Vidíme, že klasický odhad nám dává extrémně špatný a zavádějící výsledek. Porovnání odhadů vynechávací metodou nám ukazuje, že je rozhodující zvolit správnou hodnotu β .

Problémem může být, že neodstraníme všechna odlehlá pozorování. Odhad získáme lepší, ale stále může být zavádějící. Na druhou stranu mohou být odstraněna i pozorování, která rozhodně nebyla odlehlá. Pak přicházíme o užitečné informace o dané řadě, které ochudí odhad. V případě symetrického ořezávání (stejný počet největších a nejmenších hodnot) můžeme opět přijít o užitečné informace, když jsou odlehlá pozorování nesymetricky rozmístěna (například nabývají pouze extrémně vysokých hodnot). Způsob rozmístění odlehlých pozorování nalezneme na obrázku 3.4.

4.1.1 Kroková vynechávací metoda

Následující postup je vlastní (autorovou) modifikací předchozího postupu. Je vhodný pro řady s nesymetrickým rozložením odlehlých pozorování. Příkladem mohou být data, která jsou ze své podstaty omezena, například nabývají pouze kladných hodnot. Použijeme krokovou metodu pro odstraňování pozorování.

Pracujeme s uspořádaným výběrem $X_{(1)}, \dots, X_{(n)}$. Stanovíme si $\beta \in [0, 0.5)$, pak z původního výběru postupně odstraníme $2\lfloor \beta n \rfloor$ pozorování. V úvodu položíme $I_k^{(\beta)} = 1$ pro každé $k \in \{1, \dots, n\}$. V prvním kroku porovnáme odlehlost $X_{(1)}$ a $X_{(n)}$ a to odlehlejší později nebudeme používat k odhadu autokorelační funkce. Toho docílíme tím, že položíme $I_k^{(\beta)} = 0$ pro X_k , které shledáme odlehlejší. Bez újmy na obecnosti řekněme, že odlehlejší pozorováním je $X_{(1)}$. V druhém kroku porovnáme $X_{(2)}$ a $X_{(n)}$ a postupujeme jako v prvním kroku. Tento algoritmus opakujeme, dokud není vynecháno $2\lfloor \beta n \rfloor$ pozorování z původního výběru.



Obrázek 4.2: Odhad autokorelační funkce vynechávací metodou pro různé volby β . Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní symetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.

Otázkou je, jak porovnáme míru odlehlosti pozorování. Není vhodné použít klasický odhad střední hodnoty (výběrový průměr), který je velmi citlivý na odlehlá pozorování. Proto je třeba spočítat robustní odhad střední hodnoty $\hat{\mu}^{(rob)}$. Robustními odhady střední hodnoty se zabývala kapitola 3.1. Lze použít odhad pomocí výběrového mediánu $\hat{\mu}^{(med)}$, odhad pomocí oříznutí $\hat{\mu}^{(\beta)}$ či L-odhad $\hat{\mu}^{(L)}$. Pro $X_{(t)}$ spočítáme statistiku

$$KO_1(t) = (X_{(t)} - \hat{\mu}^{(rob)})^2. \quad (4.5)$$

Pak z výběru odstraníme pozorování s větší hodnotou $KO_1(t)$ pomocí předefinování $I_k^{(\beta)} = 0$ pro takové k , že odstraňované $X_{(t)} = X_k$.

Takto získaná data použijeme pro *robustní odhad autokorelační funkce jednoduchou krokovou vynechávací metodou* $\hat{\rho}_{KO_1}^{(\beta)}(t)$ dle předpisu (4.2).

Předchozí statistika KO_1 pro odhad střední hodnoty použije všechna pozorování, i ta, která se chystáme odstranit. Proto navrhujeme modifikaci této metody, kdy se odhad střední hodnoty v daném kroku bude počítat bez již odstraněných pozorování a bez pozorování v daném kroku testovaných.

V prvním kroku porovnáme odlehlost $X_{(1)}$ a $X_{(n)}$ a to odlehlejší odstraníme z výběru. Odlehlost porovnáme obdobně jako výše ve statistice KO_1 , ale pro robustní odhad střední hodnoty v prvním kroku $\hat{\mu}_1^{(rob)}$ použijeme pouze hodnoty $X_{(2)}, \dots, X_{(n-1)}$. Postupujeme obdobně jako v předchozím algoritmu. Bez újmy na obecnosti řekněme, že odlehlejší pozorováním je $X_{(1)}$. V druhém kroku porovnáme $X_{(2)}$ a $X_{(n)}$ a pro robustní odhad střední hodnoty v druhém kroku $\hat{\mu}_2^{(rob)}$ použijeme pouze $X_{(3)}, \dots, X_{(n-1)}$. Tento algoritmus opakujeme, dokud není vynecháno $2\lfloor \beta n \rfloor$ pozorování z původního výběru. $\hat{\mu}_m^{(rob)}$ nám udává robustní odhad střední hodnoty v m -tém kroku algoritmu dle výše popsaného pravidla.

Pro $X_{(t)}$ v m -tém kroku spočítáme statistiku

$$KO_2(t) = (X_{(t)} - \hat{\mu}_m^{(rob)})^2. \quad (4.6)$$

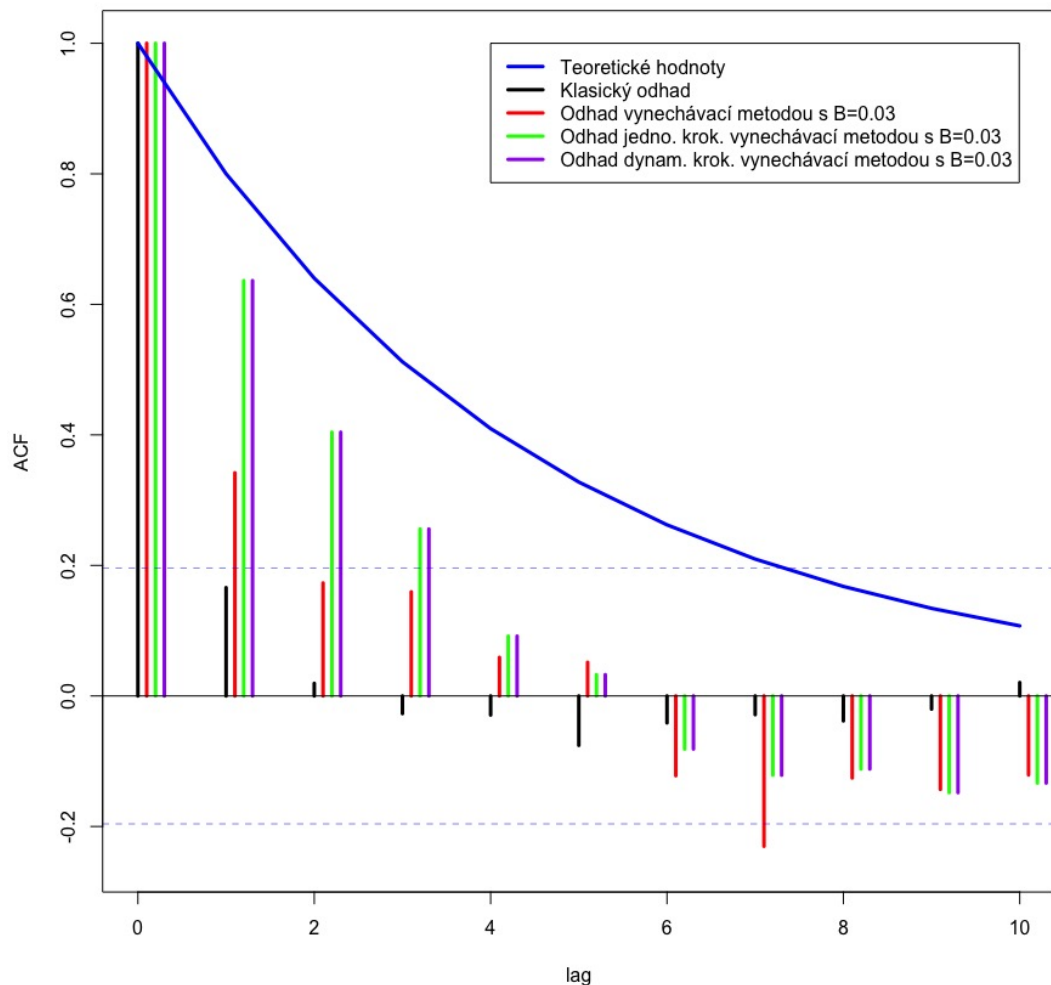
Pravidlo pro rozhodování zůstane stejné. Z výběru odstraníme pozorování s větší hodnotou $KO_2(t)$ pomocí předefinování $I_k^{(\beta)} = 0$ pro takové k , že odstraňované $X_{(t)} = X_k$.

Takto získaná data použijeme pro *robustní odhad autokorelační funkce dynamickou krokovou vynechávací metodou* $\hat{\rho}_{KO_2}^{(\beta)}(t)$ dle předpisu (4.2).

Příklad. Podíváme se na příklad odhadu krokovou vynechávací metodou. Budeme pracovat s časovou řadou $AR(1)$ v (4.1) s aditivními nesymetrickými odlehlými pozorováními zobrazenými na grafu 4.1c.

Pro řadu s aditivními nesymetrickými odlehlými pozorováními spočítáme klasický odhad dle (2.2), odhad vynechávací metodou dle (4.2) s volbou $\beta = 0.03$, robustní odhad autokorelační funkce jednoduchou krokovou vynechávací metodou dle (4.5) s volbou $\beta = 0.03$ a robustní odhad autokorelační funkce dynamickou krokovou vynechávací metodou dle (4.6) s volbou $\beta = 0.03$, které poté porovnáme s teoretickými hodnotami autokorelační funkce. K odhadu střední hodnoty v krokových metodách použijeme výběrový medián.

Simulovaná data obsahují 6 % odlehlých pozorování, tedy volba $\beta = 0.03$ by měla odstranit všechna odlehlá pozorování. Protože, ale data jsou rozložena nesymetricky dle obrázku 4.1c, tak vynechávací metoda nedokáže všechna odlehlá



Obrázek 4.3: Odhad autokorelační funkce krokovou vynechávací metodou. Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní nesymetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.

pozorování odstranit. Oproti tomu kroková vynechávací metoda tuto nesymetričnost dokáže podchytit a eliminovat její vliv.

Zkoumanou časovou řadu nalezneme na obrázku 4.1 a příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou na obrázku 4.3. Vidíme, že klasický odhad nám dává extrémně špatný a zavádějící výsledek. Taktéž vynechávací metoda nám dává velmi špatný výsledek, jelikož nedokáže eliminovat všechna odlehlá pozorování. Výsledek krokových vynechávacích metod je velmi dobrý a podporuje jejich důvod zavedení.

4.1.2 Odhad vynechávacího parametru

Nyní se podíváme na způsob, jak zvolit hodnotu parametru β , který je potřebný ve vynechávací a krokové vynechávací metodě. Jak bylo ukázáno výše, hodnota tohoto parametru je zásadní pro chování odhadu. Je třeba určit četnost

odlehých pozorování v časové řadě $\{X_t, t \in 1, \dots, n\}$. K tomu použijeme pravidlo $6\text{-}\sigma$ uvedené v dodatku A.2.5. Pro $t = 1, \dots, n$ spočteme

$$Z_t = \frac{X_t - \hat{\mu}}{\hat{\sigma}},$$

kde $\hat{\mu}$ je odhad střední hodnoty a $\hat{\sigma}$ je odhad měřítka. Je vhodné použít robustní odhad střední hodnoty dle kapitoly 3.1 a robustní odhad měřítka dle kapitoly 3.3.

Pozorování s hodnotou $|Z_t| > 3$ budeme považovat za odlehlé. Tímto pravidlem získáme počet odlehých pozorování \tilde{n} . Poté zvolíme odhad koeficientu β jako

$$\hat{\beta} = \frac{\tilde{n}}{2n}.$$

4.2 Mediánová metoda

Tato metoda je založena na využití mediánu namísto průměrů. Medián je méně náchylný na odlehlá pozorování a jeho využitím získáme robustnější odhady, jak bylo zmíněno v kapitole 3.1. Postup je navrhnut v (Chakhchoukh, 2010) a (Dürre a kol., 2015). Pracujeme s centrovanou časovou řadou $\{\tilde{X}_t, t = 1, \dots, n\}$. *Robustní odhad autokorelační funkce mediánovou metodou* je dán předpisem

$$\hat{\rho}^{(med)}(t) = \frac{\text{medián}(\tilde{X}_1\tilde{X}_{1+t}, \tilde{X}_2\tilde{X}_{2+t}, \dots, \tilde{X}_{n-t-1}\tilde{X}_{n-1}, \tilde{X}_{n-t}\tilde{X}_n)}{\text{medián}(\tilde{X}_1^2, \tilde{X}_2^2, \dots, \tilde{X}_{n-1}^2, \tilde{X}_n^2)}. \quad (4.7)$$

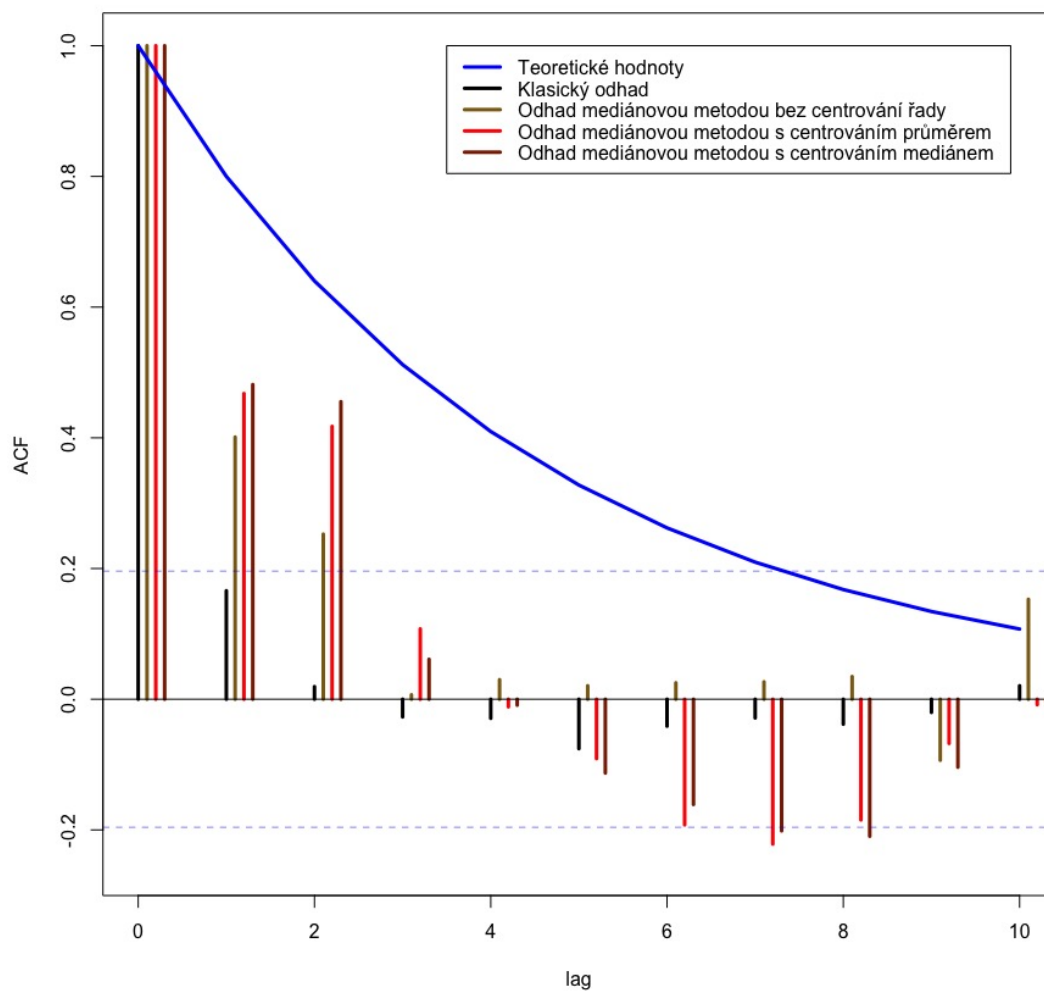
Tento odhad nahradí ve výpočtu střední hodnotu mediánem. Toto nahrazení je vhodné u dat pocházejících ze symetrického rozdělení, pro která navíc i $\tilde{X}_k\tilde{X}_{k+t}$ pochází ze symetrického rozdělení pro každé $k = 1, \dots, n - t$. Stejně tak i odlehlá pozorování by měla být symetrická. V případě porušení těchto předpokladů nám odhad může dávat zavádějící výsledky.

Dle (Dürre a kol., 2015) mediánový odhad lze také získat jako limitu výsledku metody vynechávání zmíněné výše pro $\beta = 0.5$.

Příklad. Podíváme se na příklad odhadu mediánovou metodou. Budeme pracovat s časovou řadou $AR(1)$ v (4.1) s aditivními symetrickými odlehlými pozorováními zobrazenými na grafu 4.1b.

Pro řadu s odlehlými pozorováními spočítáme klasický odhad dle (2.2) a odhad mediánovou metodou dle (4.7), které poté porovnáme s teoretickými hodnotami autokorelační funkce. Tato metoda vyžaduje centrovanou časovou řadu, proto spočítáme odhad mediánovou metodou ještě pro časovou řadu po centrování pomocí výběrového průměru a výběrového mediánu. Simulovaná data obsahují 6 % odlehých pozorování.

Zkoumanou časovou řadu nalezneme na obrázku 4.1 a příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou na obrázku 4.4. Klasický odhad nám opět dává velmi špatný výsledek. Oproti tomu, odhady mediánovou metodou jsou znatelně lepší a jsou ve větším souladu s teoretickými hodnotami. Ukazuje se vliv centrování. Nejpřesnější odhad mediánovou metodou dostaneme po centrování řady výběrovým mediánem. Na druhou stranu, pro centrované časové řady dostaneme zavádějící výsledky odhadu $\hat{\rho}^{(med)}(t)$ pro $t = 6, 7, 8$. Nabývají příliš vysokých záporných hodnot.



Obrázek 4.4: Odhad autokorelační funkce mediánovou metodou. Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní symetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.

4.3 Transformační metoda

Vhodnou transformací časové řady můžeme potlačit vliv odlehlých pozorování, snížit váhu měření. Uvedeme zde dvě transformace zmíněné v (Dürre a kol., 2015) a (Bustos a Yohai, 1986). Časovou řadu $\{X_t, t = 1, \dots, n\}$ transformujeme na novou řadu $\{\tilde{X}_t, t = 1, \dots, n\}$ pomocí vhodné funkce předpisem

$$\tilde{X}_t = \Psi\left(\frac{X_t - \hat{\mu}}{\hat{s}}\right), \quad (4.8)$$

kde $\hat{\mu}$ je odhad střední hodnoty a \hat{s} je odhad rozptylu, tedy $\gamma(0)$. Vzhledem k vlivu odlehlých pozorování je vhodné použít robustní odhady.

Hubertova transformace je

$$\Psi(x) = \text{sign}(x) \min(c_1, |x|).$$

Tukeyho transformace je

$$\Psi(x) = \begin{cases} x(1 - \frac{x^2}{c_2^2})^2, & 0 \leq |x| \leq c_2, \\ 0, & c_2 < |x|. \end{cases}$$

Konstanty c_1 a c_2 jsou ladícími parametry, které je třeba zvolit. Takto získaná data poté použijeme pro výpočet klasického odhadu dle předpisu (2.2). *Robustní odhad transformační metodou pomocí Ψ* je

$$\hat{\rho}^{(\Psi)}(t) = \frac{\sum_{k=1}^{n-t} (\tilde{X}_k - \tilde{\bar{X}}_n)(\tilde{X}_{k+t} - \tilde{\bar{X}}_n)}{\sum_{k=1}^n (\tilde{X}_t - \tilde{\bar{X}}_n)^2}. \quad (4.9)$$

V případě použití transformované časové řady pro odhad autokorelační funkce pomocí klasického vzorce dochází ke zkreslení hodnot.

Cílem této metody uveřejněné v (Bustos a Yohai, 1986) nebyl primárně odhad autokorelační funkce, takže nebylo navrženo jak řešit zkreslení a vychýlení autokorelační funkce pro transformovaná data. Dalším problémem je volba parametrů c_1 a c_2 . Dle (Dürre a kol., 2015) při volbě $c_1 = 1.37$ a $c_2 = 4.68$ dochází při transformaci jen k drobné změně dat pro Gaussovskou časovou řadu bez odlehlých pozorování, což vede pouze na malé vychýlení.

Tento postup není příliš vhodný. Může nám však posloužit v situaci, kdy nás například zajímá především znaménko autokorelační funkce a nejde nám o přesné hodnoty jejího odhadu.

Příklad. Podíváme se na příklad odhadu transformační metodou pomocí různých Ψ . Budeme pracovat s časovou řadou $AR(1)$ v (4.1) s aditivními symetrickými odlehlými pozorováními zobrazenými na grafu 4.1b.

Časovou řadu transformujeme dle předpisu (4.8). Jako odhad střední hodnoty použijeme výběrový průměr a jako odhad rozptylu použijeme výběrový rozptyl. Použijeme Huberovu i Tukeyho transformaci. Simulovaná data obsahují 6 % odlehlých pozorování. Zvolíme $c_1 = c_2$ jako 0.93 výběrový kvantil $\{|\frac{X_t - \hat{\mu}}{\hat{s}}|, t = 1, \dots, n\}$. Hodnotu 0.93 volíme kvůli 6 % kontaminace.

Pro řadu s odlehlými pozorováními a transformované řady spočítáme klasický odhad dle (2.2). Zkoumanou časovou řadu nalezneme na obrázku (4.1). Normovanou časovou řadu $\frac{X_t - \hat{\mu}}{\hat{s}}$ a řady po Huberově i Tukeyho transformaci nalezneme na obrázku 4.5a.

Příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou nalezneme na obrázku 4.5. V tomto příkladu transformační metody dávají znatelně lepší výsledek než klasický odhad.

Jak bylo výše řečeno, problémem je zkreslení a vychýlení odhadu autokorelační funkce po transformaci dat. Dalším problémem je volba parametrů c_1 a c_2 , která silně ovlivňuje transformaci dat. Poznamenejme, že při volbě $c_1 = 1.37$ a $c_2 = 4.68$ dle (Dürre a kol., 2015) dostaneme v tomto konkrétním příkladě extrémně špatné výsledky srovnatelné s klasickým odhadem.

4.4 Koeficientová metoda

Pro odhad autokorelační funkce je možné použít odhad Kendallova a Spearmanova koeficientu. Jejich popisem a odhadem se zabývá například (Dürre a kol., 2015), (Mottonen a kol., 1999) a (Croux a Dehon, 2010). Obecné předpisy, ze kterých budeme vycházet, uvádíme v A.3.

Koeficienty obecně hodnotí korelaci dvou náhodných výběrů Z_1, \dots, Z_{n-t} a Y_1, \dots, Y_{n-t} . Pro použití koeficientů k odhadu autokorelační funkce definujeme $Z_k = X_k$ a $Y_k = X_{k+t}$ pro $k = 1, \dots, n-t$. Pro $t \in \mathbb{N}_0$ pracujeme s časovými řadami $\{X_k, k = 1, \dots, n-t\}$ a $\{X_k, k = t+1, \dots, n\}$. Rozsah dat se změnil na $n-t$.

Odhad Kendallova koeficientu je

$$\hat{\tau}(t) = \frac{2}{(n-t)(n-t-1)} \sum_{i < j} \text{sign}((X_i - X_j)(X_{i+t} - X_{j+t})),$$

a odhad Spearmanova koeficientu je

$$\widehat{Sp}(t) = \frac{\sum_{k=1}^{n-t} \left(R_k - \frac{n+1}{2}\right) \left(R_{k+t} - \frac{n+1}{2}\right)}{\sum_{k=1}^n \left(R_k - \frac{n+1}{2}\right)^2}.$$

Dle (Mottonen a kol., 1999) a (Croux a Dehon, 2010) je třeba odhad Kendallova a Spearmanova koeficientu transformovat pro získání odhadu korelačního koeficientu, který poté ztotožníme s odhadem autokorelační funkce. Pro data pocházející z normálního rozdělení (obecně následující vztahy platí pro eliptické symetrické rozdělení) dostaneme následující *robustní odhad autokorelační funkce za použití Kendallova koeficientu*

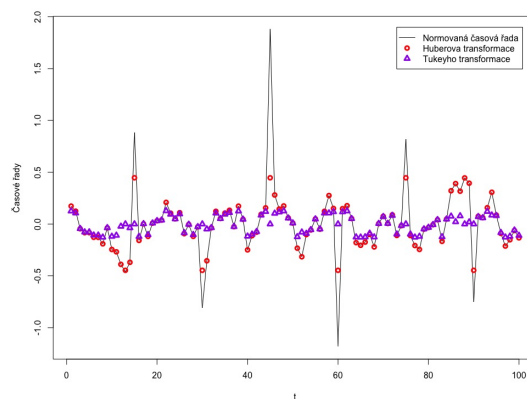
$$\hat{\rho}^{(\tau)}(t) = \sin\left(\frac{\pi \hat{\tau}(t)}{2}\right), \quad (4.10)$$

a za použití *Spearmanova koeficientu*

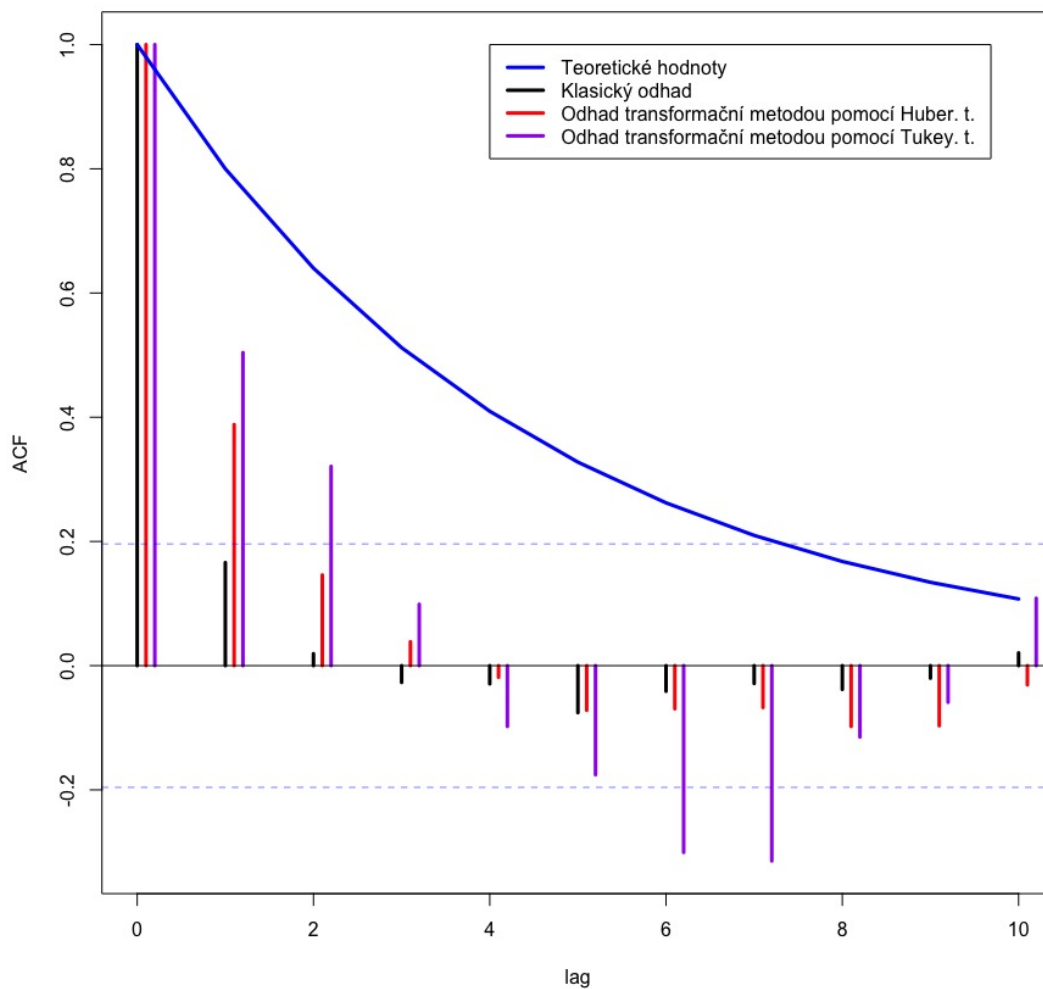
$$\hat{\rho}^{(Sp)}(t) = 2 \sin\left(\frac{\pi \widehat{Sp}(t)}{6}\right). \quad (4.11)$$

Tyto transformace však mohou způsobit porušení předpokladu pozitivní semidefinitnosti.

V (Dürre a kol., 2015) a (Garel a Hallin, 1999) je uveden obecnější předpis koeficientu využívající pořadové statistiky jako například Spearmanův koeficient.



(a) Transformované časové řady



(b) Odhady ACF pro AR(1)

Obrázek 4.5: Odhad autokorelační funkce transformační metodou pro různé volby β . Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní symetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$. Normovaná časová řada je $\{\frac{X_t - \hat{\mu}}{\hat{\sigma}}, t = 1, \dots, n\}$.

Odhad obecného pořadového koeficientu je dán předpisem

$$\widehat{S}p_g(t) = c \sum_{k=1}^{n-t} S\left(\frac{R_k}{n+1}\right) S\left(\frac{R_{k+t}}{n+1}\right),$$

kde

$$c = \frac{1}{\sum_{k=1}^n \left(S\left(\frac{R_k}{n+1}\right)\right)^2},$$

$S(x)$ je skorová funkce a R_k je pořadí X_k v uspořádané verzi.

Normální skorová funkce (nazývaná též Van der Waerden skorová funkce) je dána předpisem

$$S(x) = \Phi^{-1}(x), \quad x \in (0,1),$$

kde $\Phi(x)$ je distribuční funkce $N(0,1)$.

Lze použít Spearmanovu skorovou funkci vedoucí na Spearmanův korelační koeficient, která je dána předpisem

$$S(x) = x - \frac{1}{2}.$$

Dle (Boudt a kol., 2012) odhad získaný použitím normální skorové funkce je méně robustní než při použití Kendallova korelačního koeficientu nebo Spearmanovy skorové funkce.

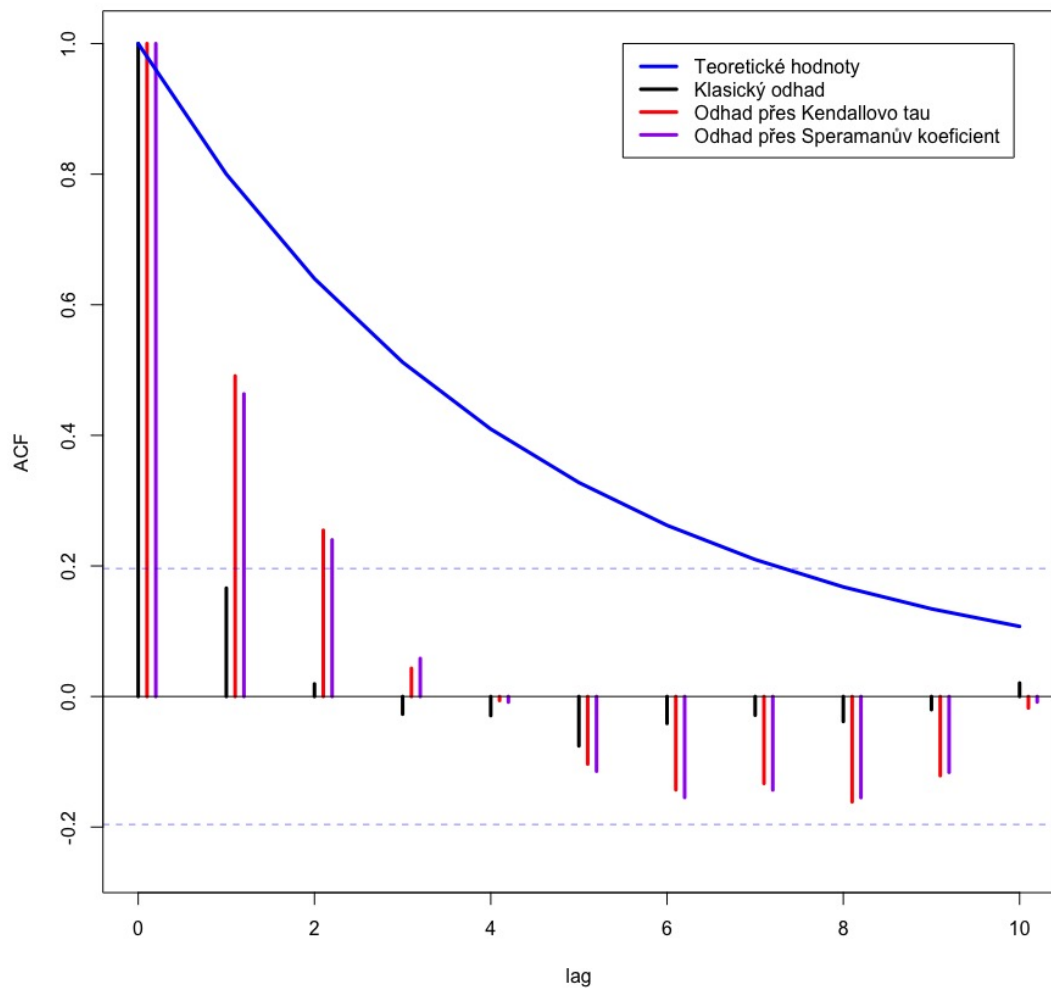
Příklad. Podíváme se na příklad odhadu koeficientovou metodou. Budeme pracovat s časovou řadou $AR(1)$ v (4.1) s aditivními symetrickými odlehlými pozorováními zobrazenými na grafu 4.1b.

Pro řadu s odlehlými pozorováními spočítáme klasický odhad dle (2.2) a odhad koeficientovou metodou, které poté porovnáme s teoretickými hodnotami autokorelační funkce. K odhadu použijeme Kendallův koeficient dle (4.10) a Spearmanův koeficient dle (4.11). Simulovaná data obsahují 6 % odlehlých pozorování.

Zkoumanou časovou řadu nalezneme na obrázku 4.1 a příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou na obrázku 4.6. Klasický odhad nám dává velmi špatný výsledek. Robustní odhady přes Kendallovo tau a Spearmanův koeficient dávají rapidně lepší výsledek oproti klasickému odhadu. V tomto konkrétním případě vychází tyto robustní odhady velmi podobně.

4.5 Metoda odhadu pomocí robustního odhadu parciální autokorelační funkce

Odhad autokorelační funkce lze získat pomocí odhadu parciální autokorelační funkce a využití vztahu mezi těmito funkcemi z věty 5. Dle požadavků věty, pracujeme s centrovanou časovou řadou. Počáteční nastavení volíme jako $\hat{\rho}(1) = \hat{\pi}(1)$. Prvně odhadneme parciální autokorelační funkci pomocí odhadu z (Masarotto, 1987). Tento postup je poté použit i v (Chan a Wei, 1992) a (Dürre a kol., 2015).



Obrázek 4.6: Odhad autokorelační funkce koeficientovou metodou za použití Kendallova tau a Spearmanova koeficientu. Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní symetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.

Vyjdeme ze vztahu pro parciální autokorelační funkci.

$$\pi(t) = \begin{cases} \text{corr}(X_k, X_{k+1}), & t = 1, \\ \text{corr}(U_{t,k}, V_{t,k}), & t > 1. \end{cases}$$

Podíváme se na situaci, kdy $t > 1$. Pak

$$\pi(t) = \text{corr}(U_{t,k}, V_{t,k}) = \frac{EU_{t,k}V_{t,k}}{\sqrt{EU_{t,k}^2EV_{t,k}^2}} = \frac{2EU_{t,k}V_{t,k}}{EU_{t,k}^2 + EV_{t,k}^2},$$

a po úpravách dostaneme

$$\pi(t)(EU_{t,k}^2 + EV_{t,k}^2) - 2EU_{t,k}V_{t,k} = 0.$$

Použijeme odhad parciální autokorelační funkce daný předpisem dle 2.3 obohacený o váhy

$$\hat{\pi}(t) = 2 \frac{\sum_{k=t+1}^n W_{t,k} \hat{U}_{t,k} \hat{V}_{t,k}}{\sum_{k=t+1}^n W_{t,k} (\hat{U}_{t,k}^2 + \hat{V}_{t,k}^2)}, \quad (4.12)$$

kde $W_{t,k}$ jsou vhodné váhy. Po úpravě dostaneme, že hledaný odhad řeší rovnici

$$\hat{\pi}(t) \sum_{k=t+1}^n W_{t,k} (\hat{U}_{t,k}^2 + \hat{V}_{t,k}^2) - 2 \sum_{k=t+1}^n W_{t,k} \hat{U}_{t,k} \hat{V}_{t,k} = 0.$$

Do věty 5 dosadíme výše získaný odhad parciální autokorelační funkce a spočítáme *robustní odhad autokorelační funkce pomocí odhadu parciální autokorelační funkce* za využití rekurze

$$\hat{\rho}^{(pacf)}(t) = \sum_{k=1}^{t-1} \hat{\pi}_{t-1,k} \hat{\rho}^{(pacf)}(t-k) + \hat{\pi}(t) \left(1 - \sum_{k=1}^{t-1} \hat{\pi}_{t-1,k} \hat{\rho}^{(pacf)}(k) \right), \quad t = 1, \dots, n-1. \quad (4.13)$$

Otázkou je, jak vhodně zvolit váhy. Dle (Masarotto, 1987) má být $W_{t,k}$ nezáporná a nerostoucí funkce. V (Dürre a kol., 2015) je navržen *robustní odhad parciální autokorelační funkce* dle předpisu 4.12 s volbami

$$W_{t,k} = w \left(\frac{d_{t,k}(\hat{\pi}(t))}{s_{t,k}^2} \right), \quad (4.14)$$

$$w(x) = \frac{3}{1+x}, \quad (4.15)$$

$$d_{t,k}(x) = \hat{U}_{t,k}^2 - 2x\hat{U}_{t,k}\hat{V}_{t,k} + \hat{V}_{t,k}^2, \quad (4.16)$$

a $s_{t,k}$ je řešením rovnice

$$s_{t,k}^2 = \frac{1}{2(n-t)} \sum_{k=t+1}^n w \left(\frac{d_{t,k}(\hat{\pi}(t))}{s_{t,k}^2} \right) d_{t,k}(\hat{\pi}(t)). \quad (4.17)$$

Takto navržený algoritmus funguje na bázi iterací odhadu $\hat{\pi}(t)$. Zvolíme si dostatečně malou míru chybovosti $\varepsilon > 0$. $\hat{U}_{t,k}$ a $\hat{V}_{t,k}$ volíme dle věty 5. V prvním kroku volíme

$$\begin{aligned}\hat{\pi}(t) &= 0, \\ d_{t,k}(x) &= \hat{U}_{t,k}^2 + \hat{V}_{t,k}^2, \\ s_{t,k}^2 &= \frac{1}{2n} \sum_{k=t+1}^n d_{t,k}(\hat{\pi}(t)).\end{aligned}$$

V dalším kroku spočítáme nové hodnoty $W_{t,k}$, $d_{t,k}(x)$ a $s_{t,k}^2$ za využití hodnot z předchozího kroku (včetně odhadu $\hat{\pi}(t)$). Nové hodnoty použijeme pro získání nového odhadu parciální autokorelační funkce. Spočítáme absolutní hodnotu rozdílu odhadu parciální autokovarianční funkce v tomto a předchozím kroku a tento algoritmus opakujeme dokud je tato absolutní hodnota rozdílu větší než námi zvolené ε .

Tato metoda je rekurzivní, proto pro odhad $\hat{\pi}(t)$ pro $t \geq 1$ je třeba nejdříve spočítat odhady všech hodnot parciální autokorelační funkce $\hat{\pi}(r)$ pro $1 \leq r < t$.

Příklad. Podíváme se na příklad odhadu pomocí robustního odhadu parciální autokorelační funkce. Budeme pracovat s časovou řadou $AR(1)$ v 4.1 s aditivními symetrickými odlehlými pozorováními zobrazenými na grafu 4.1b.

Pro řadu s odlehlými pozorováními spočítáme klasický odhad dle 2.2 a odhad pomocí robustního odhadu autokorelační funkce 4.13, kde využijeme robustní odhad parciální autokorelační funkce dle 4.12 s volbami 4.15 až 4.17. Námi používaná věta 5 vyžaduje centrovanost časové řady. Proto spočítáme odhad ještě pro řady centrované výběrovým průměrem a výběrovým mediánem. Odhady porovnáme s teoretickými hodnotami autokorelační funkce. Simulovaná data obsahují 6 % odlehlých pozorování. Míru chybovosti volíme $\varepsilon = 10^{-10}$.

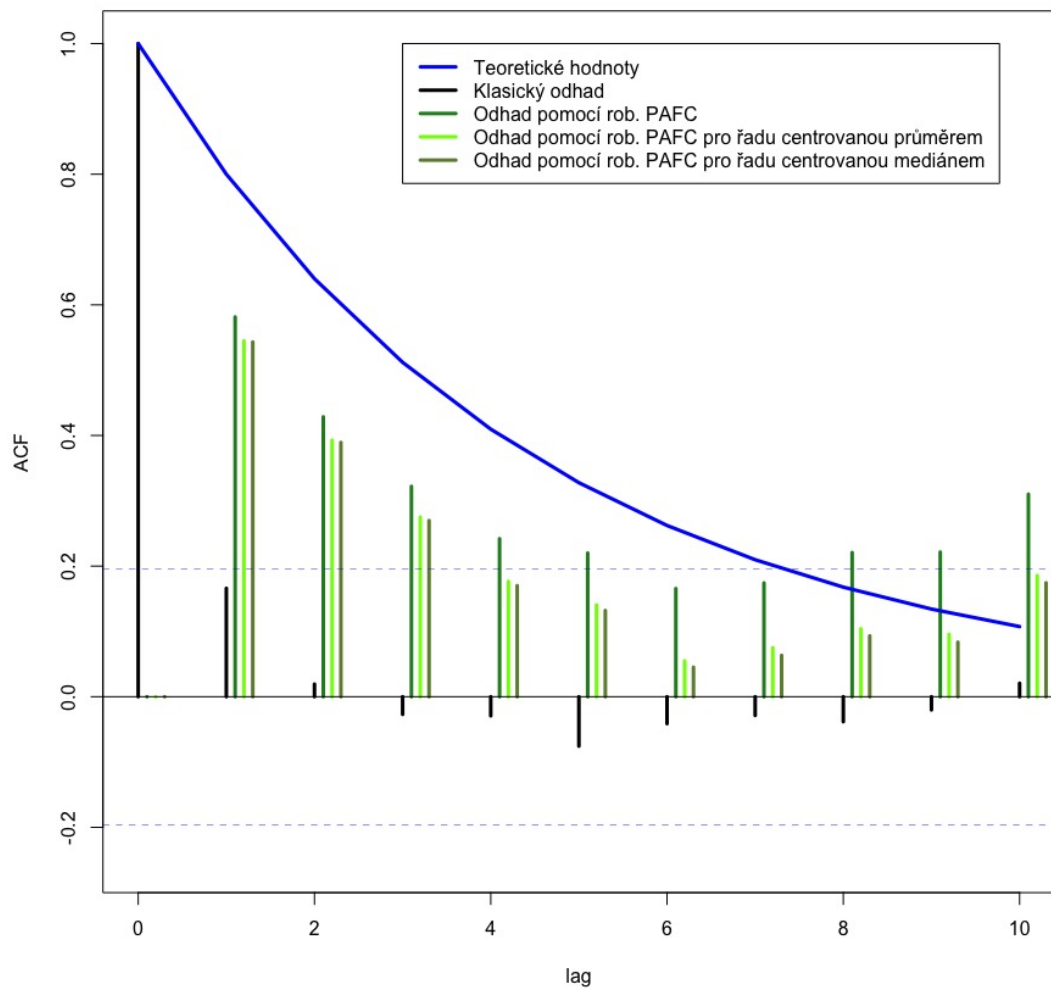
Zkoumanou časovou řadu nalezneme na obrázku 4.1 a příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou na obrázku 4.7. Klasický odhad nám dává velmi špatný výsledek. Robustní odhad dává rapidně lepší výsledek oproti klasickému odhadu. Nejlepší výsledek dostáváme pro řadu, která nebyla centrovaná. Pro centrované řady jsou robustní odhady velmi podobné. Pro necentrovanou časovou řadu dostáváme přehnaně vysoké hodnoty $\hat{\rho}^{(pacf)}(t)$ pro $t = 8, 9, 10$, což ale může zapříčinit špatné rozpoznání modelu.

Poznamenejme, že tento odhad je definovaný až pro $t \geq 1$.

4.6 Regresní metoda

Tento odhad je založen na autoregresním scatterplotu AR , na který je použita robustní lineární regrese. Čerpáme z literatury (Chang a Politis, 2014). Označíme si

$$\begin{aligned}\bar{X}_{i,\dots,j} &= \frac{1}{j-i+1} \sum_{l=i}^j X_l, \\ \bar{X}_n &= \frac{1}{n} \sum_{l=1}^n X_l.\end{aligned}$$



Obrázek 4.7: Odhad autokorelační funkce pomocí robustního odhadu parciální autokorelační funkce. Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní symetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.

Vytvoříme množinu uspořádaných dvojic (x_k, y_k) ve tvaru

$$\{(X_k - \bar{X}_{1,\dots,n-t}, X_{k+t} - \bar{X}_{t+1,\dots,n}), k \in \{1, \dots, n-t\}\}.$$

Tyto dvojice použijeme pro sestavení lineárního regresního modelu. Vysvětlující proměnnou (regresory) tvoří x_k , tedy $X_k - \bar{X}_{1,\dots,n-t}$. Vysvětlovanou proměnnou (odezvu) tvoří y_k , tedy $X_{k+t} - \bar{X}_{t+1,\dots,n}$. Regresní vztah je dán předpisem

$$y_k = c_1(t)x_k + \varepsilon_k, \quad k = 1, \dots, n-t,$$

kde ε_k je chybovou složkou, náhodnou veličinou stejně rozdělenou pro každé k s nulovou střední hodnotou a kladným, konstantním a konečným rozptylem. ε_k jsou nekorelované s x_k . Odhad $\hat{c}_1(t)$ parametru $c_1(t)$ metodou nejmenších čtverců je dle (Dupač a Hušková, 2013) dán předpisem

$$\hat{c}_1^{(OLS)}(t) = \frac{\sum_{k=1}^{n-t} x_k y_k}{\sum_{k=1}^{n-t} x_k^2}.$$

Z tohoto odhadu pozorujeme, že $c_1(t)$ reprezentuje autokorelaci dle

$$\text{corr}(y_k, x_k) = \text{corr}(c_1(t)x_k + \varepsilon_k, x_k) = c_1(t).$$

K odvození odhadu budeme potřebovat ergodickou větu. Připomeňme, že v celé kapitole pracujeme se slabě stacionární časovou řadou se střední hodnotou $\mu \in \mathbb{R}$ a s autokovarianční funkcí, pro kterou budeme vyžadovat

$$\gamma(n) \xrightarrow{n \rightarrow \infty} 0.$$

Odhad poté bude

$$\begin{aligned} \hat{c}_1^{(OLS)}(t) &= \frac{\sum_{k=1}^{n-t} x_k y_k}{\sum_{k=1}^{n-t} x_k^2} \\ &= \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_{1,\dots,n-t})(X_{k+t} - \bar{X}_{t+1,\dots,n})}{\sum_{k=1}^{n-t} (X_k - \bar{X}_{1,\dots,n-t})^2} \\ &= \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_{1,\dots,n-t})(X_{k+t} - \bar{X}_{t+1,\dots,n}) \sum_{k=1}^n (X_k - \bar{X}_n)^2}{\sum_{k=1}^{n-t} (X_k - \bar{X}_{1,\dots,n-t})^2 \sum_{k=1}^n (X_k - \bar{X}_n)^2} \\ &= \frac{\sum_{k=1}^n (X_k - \bar{X}_n)^2}{\sum_{k=1}^{n-t} (X_k - \bar{X}_{1,\dots,n-t})^2} \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_{1,\dots,n-t})(X_{k+t} - \bar{X}_{t+1,\dots,n})}{\sum_{k=1}^n (X_k - \bar{X}_n)^2}. \end{aligned}$$

Označíme si

$$A_{n,n-t} = \frac{\frac{1}{n} \sum_{l=1}^n (X_l - \bar{X}_n)^2}{\frac{1}{n-t} \sum_{l=1}^{n-t} (X_l - \bar{X}_{1,\dots,n-t})^2}.$$

Dosadíme $A_{n,n-t}$ a výraz v čitateli rozšíříme přičtením a odečtením výběrového průměru následovně

$$\begin{aligned} &\frac{n}{n-t} A_{n,n-t} \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_n + \bar{X}_n - \bar{X}_{1,\dots,n-t})(X_{k+t} - \bar{X}_n + \bar{X}_n - \bar{X}_{t+1,\dots,n})}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} \\ &= \frac{n}{n-t} \frac{A_{n,n-t}}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} \times \\ &\times \sum_{k=1}^{n-t} [(X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n) + (X_k - \bar{X}_n)(\bar{X}_n - \bar{X}_{t+1,\dots,n}) + \\ &+ (\bar{X}_n - \bar{X}_{1,\dots,n-t})(X_{k+t} - \bar{X}_n) + (\bar{X}_n - \bar{X}_{1,\dots,n-t})(\bar{X}_n - \bar{X}_{t+1,\dots,n})]. \end{aligned}$$

Poslední výraz si rozložíme do jednotlivých částí a dostaneme

$$\begin{aligned}
& \frac{n}{n-t} A_{n,n-t} \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} + \\
& + A_{n,n-t} \frac{\frac{1}{n-t} (\bar{X}_n - \bar{X}_{t+1,\dots,n}) \sum_{k=1}^{n-t} (X_k - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} + \\
& + A_{n,n-t} \frac{\frac{1}{n-t} (\bar{X}_n - \bar{X}_{1,\dots,n-t}) \sum_{k=1}^{n-t} (X_{k+t} - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} + \\
& + A_{n,n-t} \frac{\frac{n-t}{n-t} (\bar{X}_n - \bar{X}_{1,\dots,n-t})(\bar{X}_n - \bar{X}_{t+1,\dots,n})}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2}.
\end{aligned}$$

Což upravíme na

$$\begin{aligned}
& \frac{n}{n-t} A_{n,n-t} \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} + \\
& + A_{n,n-t} \frac{(\bar{X}_n - \bar{X}_{t+1,\dots,n})(\bar{X}_{1,\dots,n-t} - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} + \\
& + A_{n,n-t} \frac{(\bar{X}_n - \bar{X}_{1,\dots,n-t})(\bar{X}_{t+1,\dots,n} - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} + \\
& + A_{n,n-t} \frac{(\bar{X}_n - \bar{X}_{1,\dots,n-t})(\bar{X}_n - \bar{X}_{t+1,\dots,n})}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2}.
\end{aligned}$$

Podíváme se na chování posledních tří zlomků z výrazu výše. K tomu využijeme následující vztahy pro pevné t

$$\begin{aligned}
\bar{X}_n & \xrightarrow{P}_{n \rightarrow \infty} \mu, \\
\bar{X}_{1,\dots,n-t} & \xrightarrow{P}_{n \rightarrow \infty} \mu, \\
\bar{X}_{t+1,\dots,n} & \xrightarrow{P}_{n \rightarrow \infty} \mu, \\
\frac{1}{n} \sum_{l=1}^n (X_l - \bar{X}_n)^2 & \xrightarrow{P}_{n \rightarrow \infty} \sigma^2 = \gamma(0),
\end{aligned}$$

které nám plynou z ergodické věty (Brockwell a kol., 1991, věta 7.1.1), pokud $\gamma(n) \xrightarrow{n \rightarrow \infty} 0$, pak $\text{var} \bar{X}_t \xrightarrow{n \rightarrow \infty} 0$ a dostaneme $\bar{X}_n \xrightarrow{P}_{n \rightarrow \infty} \mu$. Obdobně pro $\bar{X}_{1,\dots,n-t}$ a $\bar{X}_{t+1,\dots,n}$.

Poté dostaneme

$$(\bar{X}_n - \bar{X}_{t+1,\dots,n})(\bar{X}_{1,\dots,n-t} - \bar{X}_n) \xrightarrow{P}_{n \rightarrow \infty} 0,$$

a

$$\frac{(\bar{X}_n - \bar{X}_{t+1,\dots,n})(\bar{X}_{1,\dots,n-t} - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} \xrightarrow{P}_{n \rightarrow \infty} \frac{0}{\sigma^2} = 0.$$

Poslední vztah platí i pro další výrazy se součinem rozdílu aritmetických průměrů.

Pak pro jednotlivé zlomky dostaneme

$$\begin{aligned} \frac{(\bar{X}_n - \bar{X}_{t+1, \dots, n})(\bar{X}_{1, \dots, n-t} - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} &\xrightarrow{P}_{n \rightarrow \infty} 0, \\ \frac{(\bar{X}_n - \bar{X}_{1, \dots, n-t})(\bar{X}_{t+1, \dots, n} - \bar{X}_n)}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} &\xrightarrow{P}_{n \rightarrow \infty} 0, \\ \frac{(\bar{X}_n - \bar{X}_{1, \dots, n-t})(\bar{X}_n - \bar{X}_{t+1, \dots, n})}{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2} &\xrightarrow{P}_{n \rightarrow \infty} 0. \end{aligned}$$

Nyní se podíváme na chování $A_{n, n-t}$, k čemuž využijeme vztahy

$$\begin{aligned} \frac{1}{n} \sum_{l=1}^n (X_l - \bar{X}_n)^2 &\xrightarrow{P}_{n \rightarrow \infty} \sigma^2 = \gamma(0), \\ \frac{1}{n-t} \sum_{l=1}^{n-t} (X_l - \bar{X}_{1, \dots, n-t})^2 &\xrightarrow{P}_{n \rightarrow \infty} \sigma^2 = \gamma(0), \end{aligned}$$

z toho plyne

$$A_{n, n-t} \xrightarrow{P}_{n \rightarrow \infty} 1.$$

Dle výše odvozeného dostaneme

$$\begin{aligned} \hat{c}_1^{(OLS)}(t) &\approx \frac{n}{n-t} \cdot 1 \cdot \frac{\sum_{k=1}^{n-t} (X_k - \bar{X}_n)(X_{k+t} - \bar{X}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} + 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0 \\ &\approx \frac{n}{n-t} \hat{\rho}(t). \end{aligned}$$

Výraz $\frac{n}{n-t}$ je zapříčiněn tím, že odhad směrnice metodou nejmenších čtverců je nevychýlený, zatímco odhad autokorelační funkce je dle své konstrukce vychýlený.

Pro odhad autokorelační funkce získáme vztah

$$\hat{\rho}(t) = \frac{n-t}{n} \hat{c}_1^{(OLS)}(t).$$

V předchozí části jsme pracovali s dvojicemi

$$\{(X_k - \bar{X}_{1, \dots, n-t}, X_{k+t} - \bar{X}_{t+1, \dots, n}), k \in \{1, \dots, n-t\}\},$$

kde pracujeme s výběrovým průměrem. Nyní budeme pracovat s dvojicemi

$$\{(X_k, X_{k+t}), k \in \{1, \dots, n-t\}\}.$$

Poté je třeba regresní přímku obohatit o konstantní člen nahrazující výběrový průměr. Regresní vztah je poté dán předpisem

$$y_k = c_1(t)x_k + c_0(t) + \varepsilon_k, \quad k = 1, \dots, n-t,$$

kde $c_0(t)$ a $c_1(t)$ jsou neznámé parametry, které je třeba odhadnout. Po dosazení dostaneme

$$X_{k+t} = c_1(t)X_k + c_0(t) + \varepsilon_k, \quad k = 1, \dots, n-t.$$

Nyní spočítáme odhad $\hat{c}_1^{(OLS)}(t)$ metodou nejmenších čtverců pro parametr $c_1(t)$ daný dle (Dupač a Hušková, 2013) předpisem

$$\hat{c}_1^{(OLS)}(t) = \frac{\sum_{k=1}^{n-t} (X_k - \frac{1}{n-t} \sum_{l=1}^{n-t} X_l)(X_{k+t} - \frac{1}{n-t} \sum_{l=t+1}^n X_l)}{\sum_{k=1}^{n-t} (X_k - \frac{1}{n-t} \sum_{l=1}^{n-t} X_l)^2}.$$

Tento odhad opět použijeme a získáme

$$\hat{\rho}(t) = \frac{n-t}{n} \hat{c}_1^{(OLS)}(t).$$

Ze vztahu mezi odhadem směrnice regresní přímky a autokorelační funkce plyne, že získáním robustního odhadu $\hat{c}_1^{(rob)}(t)$ směrnice dostaneme i robustní odhad autokorelační funkce. *Robustní odhad autokorelační funkce regresní metodou* je

$$\hat{\rho}^{(reg)}(t) = \frac{n-t}{n} \hat{c}_1^{(rob)}(t). \quad (4.18)$$

Poznamenejme, že po využití odhadu $\hat{c}_1^{(rob)}(t)$ pro $\hat{\rho}^{(reg)}(t)$ může nastat situace $\hat{\rho}^{(reg)}(t) > 1$ nebo $\hat{\rho}^{(reg)}(t) < -1$. Odhad pomocí robustní regrese nám nezaručí $\hat{\rho}^{(reg)}(t) \in [-1, 1]$. Toto je třeba řešit a budeme se tomu věnovat dále.

Jako robustní odhad pro lineární regresi lze použít L1-odhad nebo M-odhad zmíněné v kapitole 3.4. Podrobnější popis a další metody odhadu nalezneme například v (Maronna a kol., 2006) a (Chang a Politis, 2014).

Tato metoda může dle (Chang a Politis, 2014) způsobit porušení předpokladu pozitivní semidefinitnosti.

Příklad. Podíváme se na příklad odhadu regresní metodou. Budeme pracovat s časovou řadou $AR(1)$ v (4.1) s aditivními symetrickými odlehlými pozorováními zobrazenými na grafu 4.1b.

Pro řadu s odlehlými pozorováními spočítáme klasický odhad dle (2.2) a odhad regresní metodou dle (4.18), které poté porovnáme s teoretickými hodnotami autokorelační funkce. Simulovaná data obsahují 6 % odlehlých pozorování.

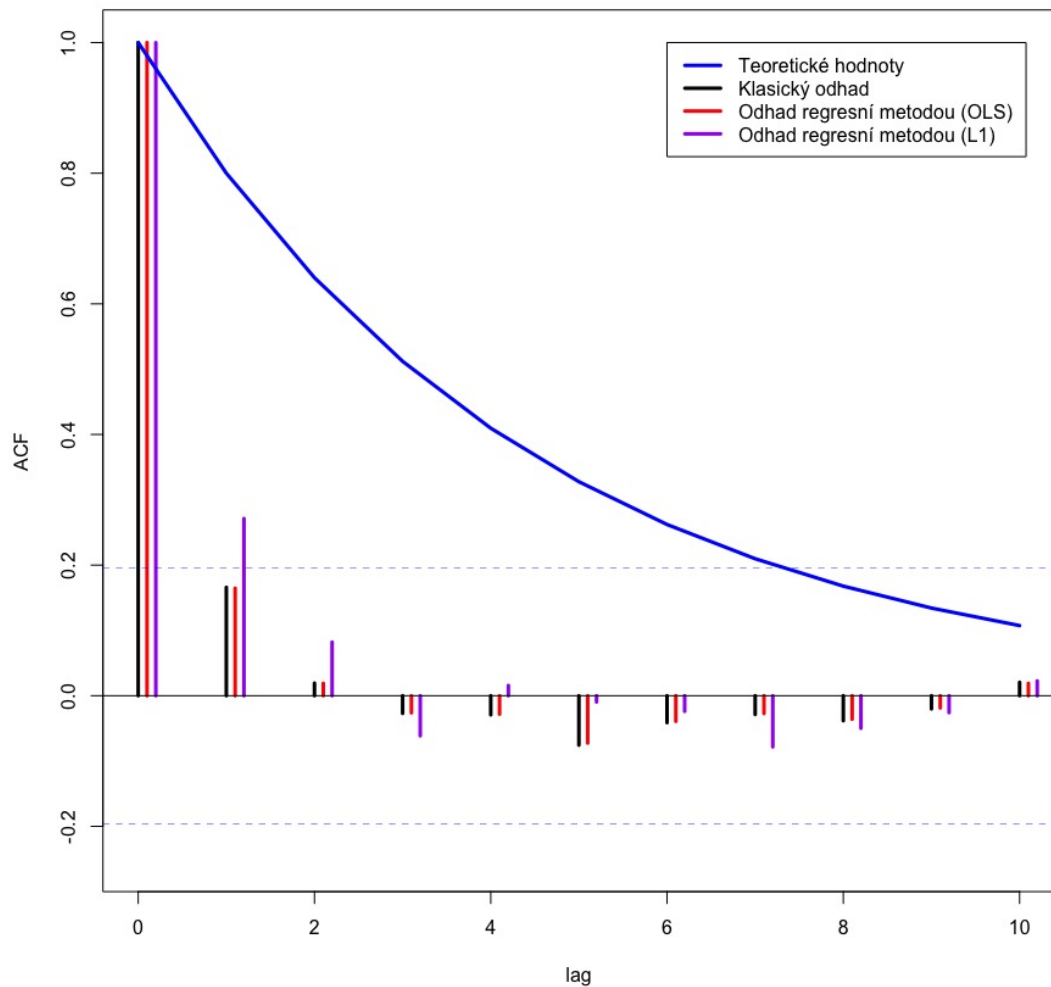
Zkoumanou časovou řadu nalezneme na obrázku 4.1 a příslušné odhady autokorelační funkce porovnané s teoretickou hodnotou na obrázku 4.8. Klasický odhad nám opět dává velmi špatný výsledek, obdobně jako regresní metoda při použití metody nejmenší čtverců pro odhad směrnice regresní přímky. V tomto konkrétním případě i robustnějšího L1-odhad nedává příliš uspokojivý výsledek, přesto lepší než klasický odhad.

4.7 Porušení vlastností

Některé metody mohou zapříčinit porušení některých vlastností odhadu autokorelační funkce. Podíváme se na řešení.

4.7.1 Omezenost odhadu

Autokorelační funkce ρ nabývá pouze hodnot z intervalu $[-1, 1]$. Některé robustní metody mohou tento předpoklad porušovat. Nechť $\hat{\rho}^{(rob)}(t)$ je robustní odhad $\rho(t)$. Pokud nastane situace, že $\hat{\rho}^{(rob)}(t) > 1$, pak je třeba položit $\hat{\rho}^{(rob)}(t) = 1$. Obdobně pro vypočtené $\hat{\rho}^{(rob)}(t) < -1$ položíme $\hat{\rho}^{(rob)}(t) = -1$.



Obrázek 4.8: Odhad autokorelační funkce regresní metodou. Časová řada bez odlehlých pozorování je $X_t = 0.8X_{t-1} + Y_t$ o rozsahu $n = 100$, kde Y_t je $N(0,1)$. Aditivní symetrická odlehlá pozorování se nachází v bodech $t = 15, 30, 45, 60, 75, 90$.

Situace, kdy $\hat{\rho}^{(rob)}(t) = \pm 1$, je degenerovaná. Dalším řešením je pro vypočtené $\hat{\rho}^{(rob)}(t) > 1$ položit $\hat{\rho}^{(rob)}(t) = 1 - \epsilon$ a pro $\hat{\rho}^{(rob)}(t) < -1$ položit $\hat{\rho}^{(rob)}(t) = -1 + \epsilon$, kde $\epsilon \in (0, 1)$ si stanovíme předem jako malé číslo.

Tato situace může velmi snadno nastat u regresní metody dle (4.18). Také jsme se s tím setkali při porušení předpokladů mediánové metody (4.7).

4.7.2 Pozitivní semidefinitnost

O matici $\mathbf{\Gamma}^{(t)}$ velikosti $t \times t$ z (1.1), mající na pozici (i, j) prvek $\gamma(i - j)$, víme, že je pozitivně semidefinitní. Některé odhady mohou tento předpoklad porušovat. Místo autokovarianční matice budeme pracovat s autokorelační maticí $\mathbf{\Sigma}^{(t)}$ velikosti $t \times t$ mající na pozici (i, j) prvek $\rho(i - j)$, která je taktéž pozitivně semidefinitní. Spočítáme robustní odhad autokorelační funkce $\hat{\rho}^{(rob)}$ a dostaneme matici $\hat{\mathbf{\Sigma}}^{(t)}$, mající na pozici (i, j) prvek $\hat{\rho}^{(rob)}(i - j)$.

Kromě pozitivní semidefinitnosti matice $\mathbf{\Sigma}^{(t)}$ splňuje i specifickou Toeplitzovu strukturu, na hlavní diagonále jsou stejné prvky $\rho(0)$, vedle hlavní diagonály z obou stran prvky $\rho(1)$ a tak dále dle zavedení matice.

Potřebné znalosti z oblasti lineární algebry budeme čerpat z (Anděl, 2005). Zde jen zmíníme, jaké metody řešení se objevují v literatuře.

Matrice $\hat{\mathbf{\Sigma}}^{(t)}$ je dle své definice reálná a symetrická. Dle spektrální věty pro symetrické matice existuje rozklad

$$\hat{\mathbf{\Sigma}}^{(t)} = \hat{\mathbf{U}}_t \hat{\mathbf{D}}_t \hat{\mathbf{U}}_t^\top, \quad (4.19)$$

kde $\hat{\mathbf{D}}_t = \text{diag}(\hat{d}_1, \dots, \hat{d}_t)$ je diagonální matice a $\hat{\mathbf{U}}_t$ je ortogonální matice. Diagonální prvky matice $\hat{\mathbf{D}}_t$ jsou vlastními čísly matice $\hat{\mathbf{\Sigma}}^{(t)}$. Matici $\hat{\mathbf{U}}_t$ tvoří ortonormální vlastní vektory.

Symetrická matice je pozitivně semidefinitní právě tehdy, když jsou všechna její vlastní čísla nezáporná. Pozitivní definitnost vyžaduje, aby všechna vlastní čísla byla kladná. Tedy rozložením $\hat{\mathbf{\Sigma}}^{(t)}$ zjistíme dle tvaru $\hat{\mathbf{D}}_t$, zda je matice $\hat{\mathbf{\Sigma}}^{(t)}$ pozitivně semidefinitní.

V (Dürre a kol., 2015) je použit rozklad matice $\hat{\mathbf{\Sigma}}^{(t)}$ dle (4.19). Dle matice $\hat{\mathbf{D}}_t$ vytvoříme matici $\hat{\hat{\mathbf{D}}}_t = \text{diag}(\hat{\hat{d}}_1, \dots, \hat{\hat{d}}_t)$ předpisem

$$\hat{\hat{d}}_i = \begin{cases} \hat{d}_i, & \hat{d}_i \geq 0, \\ 0, & \hat{d}_i < 0. \end{cases}$$

Poté dostaneme matici

$$\hat{\hat{\mathbf{\Sigma}}}^{(t)} = \hat{\mathbf{U}}_t \hat{\hat{\mathbf{D}}}_t \hat{\mathbf{U}}_t^\top, \quad (4.20)$$

která bude pozitivně semidefinitní. Poznamenejme, že touto úpravou můžeme přijít o Toeplitzovu strukturu matice. V (Dürre a kol., 2015) je matice $\hat{\hat{\mathbf{\Sigma}}}^{(t)}$ dále upravována iteračně, aby výsledná matice měla požadovanou Toeplitzovu strukturu.

V článku (Chang a Politis, 2014), který vychází z (McMurry a Politis, 2010), se pracuje s transformovanou symetrickou maticí $\hat{\hat{\mathbf{\Sigma}}}^{(t)}$ se spektrálním rozkladem

$$\hat{\hat{\mathbf{\Sigma}}}^{(t)} = \hat{\hat{\mathbf{U}}}_t \hat{\hat{\mathbf{D}}}_t \hat{\hat{\mathbf{U}}}_t^\top,$$

kde $\hat{\mathbf{D}}_t = \text{diag}(\hat{d}_1, \dots, \hat{d}_t)$ obsahuje vlastní čísla matice $\hat{\Sigma}^{(t)}$. Definujeme si matici $\hat{\mathbf{D}}_t^{(\epsilon, \zeta)} = \text{diag}(\hat{d}_1^{(\epsilon, \zeta)}, \dots, \hat{d}_t^{(\epsilon, \zeta)})$ s prvky

$$\hat{d}_i^{(\epsilon, \zeta)} = \max\left(\hat{d}_i, \frac{\epsilon}{n^\zeta}\right), \quad i = 1, \dots, t,$$

kde parametry volíme $\epsilon \geq 0$, $\zeta > 0.5$ a n je délka časové řady. V (Chang a Politis, 2014) je doporučeno $\epsilon = 1$ a $\zeta = 1$. Poté dostaneme výslednou matici

$$\hat{\Sigma}^{(t)(\epsilon, \zeta)} = \hat{\mathbf{U}}_t \hat{\mathbf{D}}_t^{(\epsilon, \zeta)} \hat{\mathbf{U}}_t^\top,$$

která bude reprezentovat odhad autokorelační matice, který bude pozitivně semidefinitní. Při volbě $\epsilon > 0$ dostaneme navíc pozitivně definitní matici. Tento postup je velmi podobný jako předchozí řešení ve (4.20), jen nenahrazujeme prvky diagonální matice přesně nulou, ale malým kladným číslem, což zaručí pozitivní definitnost. Dále rozkládáme odhad autokorelační matice s transformovanými prvky místo původní matice. Opět je potřeba řešit porušení Toeplitzovy struktury. Pro podrobnější popis odkážeme do zmíněné odborné literatury.

Porušení pozitivní semidefinitnosti může nastat u koeficientové metody dle (4.10) a (4.11) a u regresní metody dle (4.18).

5. Aplikace

V této části porovnáme mezi sebou metody z kapitoly 4 na simulovaných datech a poté je použijeme na reálná data. Simulace jsou provedeny v programu R (R Core Team, 2018).

Odhady vynechávací metodou dle (4.2), jednoduchou i dynamickou krokovou vynechávací metodou dle (4.5) a (4.6) a odhad pomocí robustního odhadu parciální autokorelační funkce dle (4.13) jsou počítány pomocí vlastních funkcí.

Mediánová metoda dle (4.7) a koeficientová metoda za použití Kendallova τ dle (4.10) či Spearmanova koeficientu dle (4.11) je spočtena pomocí vlastní funkce využívající základní funkce programu.

Odhad regresní metodou dle (4.18) je spočten pomocí funkce *lfit* z balíku *Llpack* (Osorio, 14.11.2017).

5.1 Simulace

Budeme simulovat data pro AR a MA časovou řadu délky $n = 100$ s aditivními a inovačními odlehlými pozorováními. Také rozlišíme zda odlehlá pozorování jsou rozmístěna symetricky či nesymetricky. V každém z těchto případů provedeme $m = 100$ iterací. V i -té iteraci spočítáme $\hat{\rho}_i(t)$ jako odhad $\rho(t)$ pro dané $t \in \mathbb{N}_0$. Poté spočteme průměrné vychýlení

$$BIAS = \frac{1}{m} \sum_{i=1}^m (\hat{\rho}_i(t) - \rho(t)),$$

a průměrnou střední kvadratickou chybu

$$\begin{aligned} MSE &= \frac{1}{m} \sum_{i=1}^m (\hat{\rho}_i(t) - \rho(t))^2 \\ &= \frac{1}{m-1} \sum_{i=1}^m (\hat{\rho}_i(t) - \tilde{\rho}_m(t))^2 + \left[\frac{1}{m} \sum_{i=1}^m (\hat{\rho}_i(t) - \rho(t)) \right]^2, \end{aligned}$$

kde $\tilde{\rho}_m(t) = \frac{1}{m} \sum_{i=1}^m \hat{\rho}_i(t)$. Výstup uvedeme v tabulkách a grafech pro odhady $\hat{\rho}(t)$ s $t = 1, \dots, 5$.

Do simulací nebudeme zahrnovat odhad transformační metodou dle (4.9), jelikož tato metoda má své vážné nedostatky, jak bylo zmíněno při jejím popisu.

5.1.1 Aditivní symetrická odlehlá pozorování

Časová řada obsahuje aditivní symetrická odlehlá pozorování. $\{G_t, t \in \mathbb{Z}\}$ je stacionární časová řada, v naší situaci AR(1) nebo MA(4). Časová řada s odlehlými pozorováními je

$$X_t = U_t G_t + (1 - U_t) R_t V_t, \quad (5.1)$$

kde R_t je z rovnoměrného rozdělení $R(a, b)$ a

$$V_t = 2Z_t - 1,$$

kde Z_t je z alternativního rozdělení $Alt(0.5)$. Tím získáme přibližně symetrickou strukturu odlehlých pozorování. U_t pochází z alternativního rozdělení $Alt(0.9)$. Tím pádem v řadě bude přibližně 10 % odlehlých pozorování.

V těchto konkrétních simulacích volíme

$$a = 2\sigma,$$

$$b = 4\sigma,$$

kde σ^2 je rozptyl stacionární řady G_t .

Mediánová metoda dle (4.7) a *odhad pomocí robustního odhadu parciální autokorelační funkce* dle (4.13) vyžadují práci s centrovanou časovou řadou. Řada, se kterou pracujeme dle 5.1, je centrovaná. To platí i proto, že odlehlá pozorování jsou symetrická. V případě nesymetrických odlehlých pozorování již bude předpoklad centrovanosti porušen. Proto budeme výše zmíněné metody aplikovat i na řadu, kterou budeme centrovat výběrovým průměrem a nebo výběrovým mediánem. Pak lze mluvit o necentrované metodě, metodě centrované výběrovým průměrem a metodě centrované výběrovým mediánem.

AR(1)

Pracujeme s časovou řadou s odlehlými pozorováními $\{X_t, t \in \mathbb{Z}\}$ dle (5.1), jejíž stacionární složka je AR(1)

$$G_t = 0.8G_{t-1} + Y_t, \quad t \in \mathbb{Z},$$

s autokorelační funkcí

$$\rho(t) = 0.8^{|t|}, \quad t \in \mathbb{Z}.$$

$\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $N(0,1)$ a

$$\begin{aligned} \text{var}(G_t) &= \text{var}(0.8G_{t-1} + Y_t) = \text{var}(0.8^2G_{t-2} + 0.8Y_{t-1} + Y_t) = \\ &= \text{var}\left(\sum_{k=0}^{\infty} 0.8^k Y_t\right) = \sum_{k=0}^{\infty} 0.8^k \text{var}(Y_t) = \sum_{k=0}^{\infty} 0.8^k = \\ &= \frac{1}{1-0.8} = \frac{25}{9}. \end{aligned}$$

Výsledky simulace nalezneme v tabulce 5.1 a na obrázku 5.1. Vidíme, že se vychýlení i střední čtvercová chyba chovají obdobně pro různé $t = 1, \dots, 5$, proto budeme metody hodnotit celkově dle obou kritérií současně.

U vynechávací metody dle (4.2) se ukazuje důležitost správného rozpoznání koeficientu β určující míru oříznutí. Při správném určení $\beta = 0.050$ dostáváme velmi dobré výsledky. Při srovnání vynechávací metody s jednoduchou krokovou vynechávací metodou dle (4.5) a dynamickou krokovou vynechávací metodou dle (4.6) dostáváme pro krokové metody lepší výsledek, což je očekávané. Krokové metody jsou dle svého navržení flexibilnější na eliminaci odlehlých pozorování. Jejich lepší vlastnosti se projevily i v tomto případě, kdy jsou odlehlá pozorování přibližně symetrická, přestože byly navrženy primárně pro nesymetrická odlehlá pozorování. Obě krokové metody vycházejí extrémně podobně, dynamická je o něco málo lepší, což opět plyne z její konstrukce. Poznamenejme, že zde

uvádíme srovnání vynechávací metody a krokové vynechávací metody pouze pro $\beta = 0.050$, které odpovídá přibližnému počtu simulovaných odlehlých pozorování. Správné určení β je důležité i pro krokovou vynechávací metodu, při nesprávné hodnotě bude i odhad špatný.

Ke správné volbě β je třeba určit míru odlehlých pozorování. K tomu lze použít například $6\text{-}\sigma$ pravidlo zmíněné v A.2.5, jehož použití jsme nastínili v kapitole 4.1.2. Pro výběr velikosti n určíme, že počet odlehlých pozorování je $\tilde{n} < n$. Pak zvolíme $\beta = \frac{\tilde{n}}{2}$. Tento postup volby budeme dále podrobněji zkoumat v samostatných simulacích.

Mediánová metoda dle (4.7) dává dobré hodnoty. Vidíme, že nejlepší odhad dostáváme pro necentrovanou metodu. BIAS pro necentrovanou metodu je srovnatelný s vynechávací metodou se správně určeným β . MSE je oproti vynechávací metodě větší. Výhodou této metody je, že není třeba určit hodnotu dalšího parametru. Dle simulace centrované metody dávají horší chování odhadu. Připomeňme, že v našem případě řada je dle své definice centrovaná a centrování průměrem nebo mediánem již není nutné.

Koeficientová metoda za použití Kendallova τ dle (4.10) či Spearmanova koeficientu dle (4.11) dává velmi dobré výsledky. Oba odhady se chovají podobně jako odhad založený na vynechávací metodě pro $\beta = 0.050$ jak z pohledu BIAS tak MSE. Odhad spočtený pomocí τ se chová o něco lépe než odhad založený na Spearmanově koeficientu.

Velmi dobré výsledky pozorujeme pro odhad pomocí robustního odhadu partiální autokorelační funkce dle (4.13). Nejlepší výsledek dostáváme pro necentrovanou metodu. Poznamenejme, že v simulacích bylo zjištěno, že s rostoucím t roste i BIAS tohoto odhadu pro necentrovanou metodu a MSE pro všechny varianty této metody. Začalo se to projevovat ztlačněji od $t \geq 8$. Toto chování shrnuje obrázek 5.2. Obdobný problém jsme viděli již v příkladu pro daný odhad v kapitole 4.

Poslední regresní metoda dle (4.18) dává velmi nepřesné odhady jak z pohledu BIAS, tak MSE. U OLS odhadu jsme to očekávali. L1 odhad dává o něco lepší výsledky než OLS, ale stále velmi špatné oproti ostatním metodám.

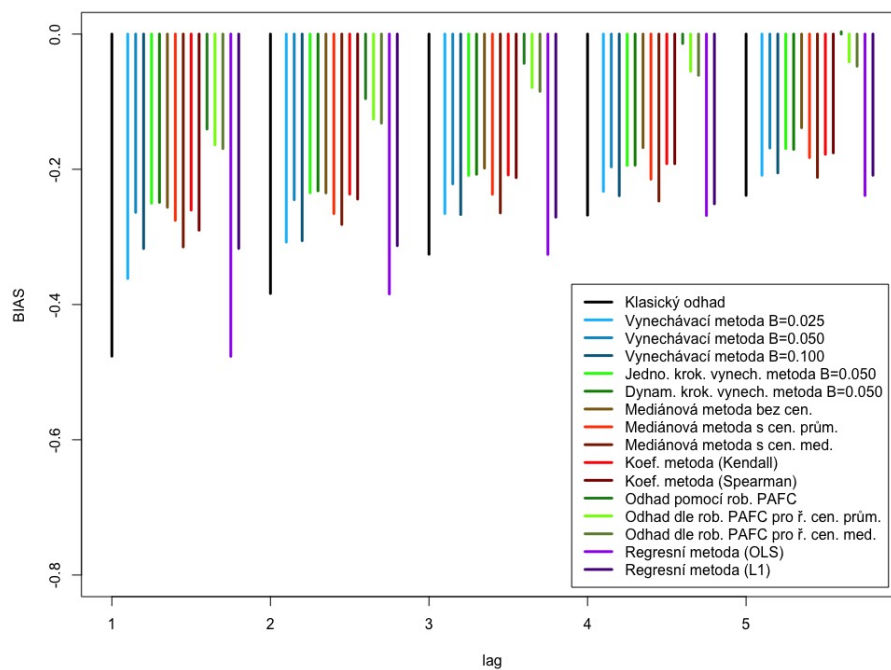
Pro toto konkrétní nastavení se podíváme na simulaci, kde budeme odhadovat koeficient β potřebný ve vynechávací metodě dle (4.2) a krokové vynechávací metodě dle (4.5) a (4.6). Způsob odhadu je popsán v kapitole 4.1.2. Pro robustní odhad polohy použijeme výběrový medián a pro robustní odhad měřítka použijeme několik odhadů z kapitoly 3.3. Prvním odhadem měřítka bude klasický odhad směrodatné odchyly, druhý odhad bude založen na

$$MAD(\mathbf{X}) = \widehat{\text{med}}(|\mathbf{X} - \widehat{\text{med}}(\mathbf{X})|),$$

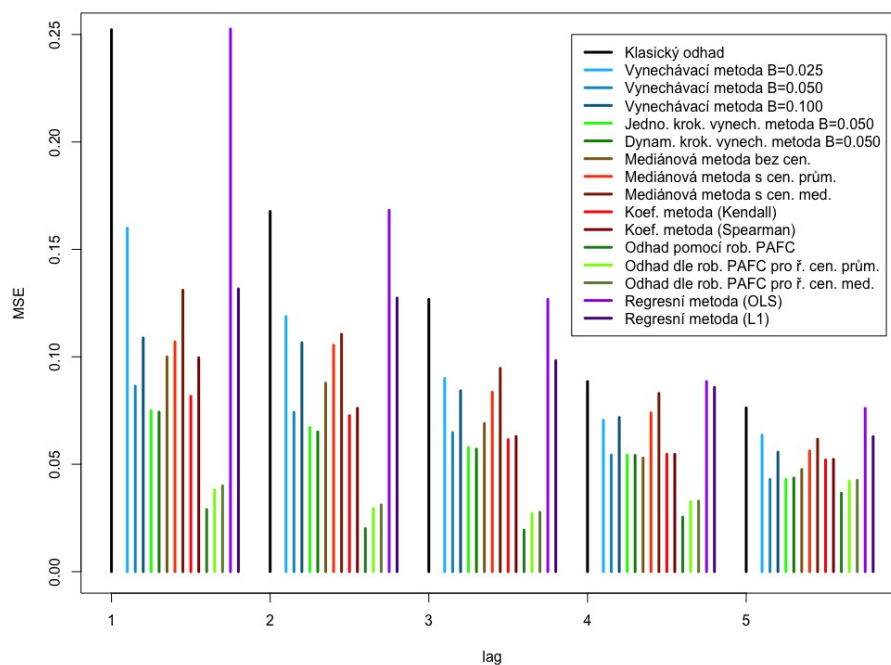
a poslední na

$$MD(\mathbf{X}) = \frac{1}{n} \sum_{k=1}^n |X_k - \bar{X}|.$$

Odhady MAD a MD nejsou přímo odhady směrodatné odchyly, je třeba je transformovat, čemuž se věnujeme v kapitole 3.3. Připomeňme následující pravidlo pro robustní odhad směrodatné odchyly, které platí za předpokladu normálního

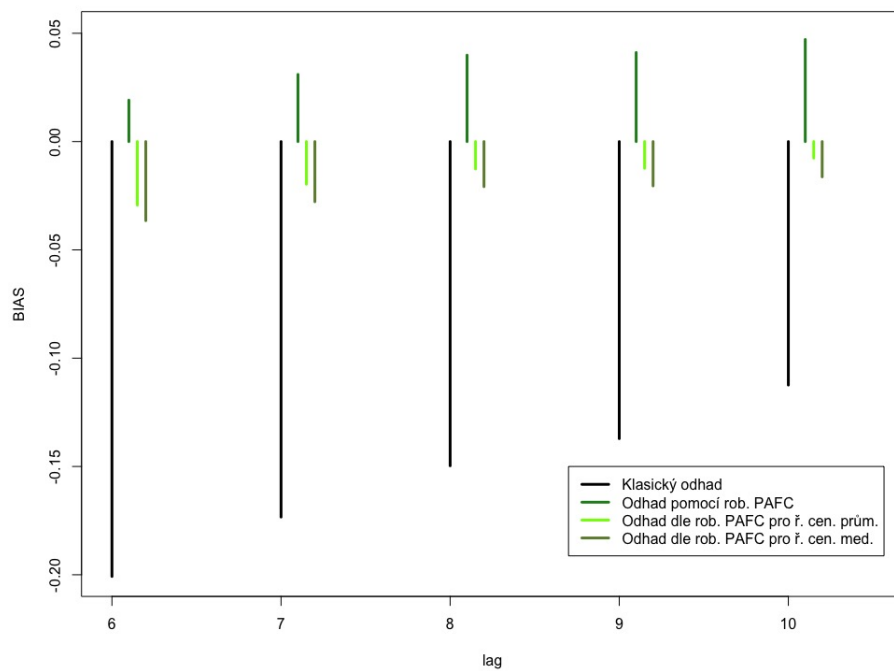


(a) BIAS

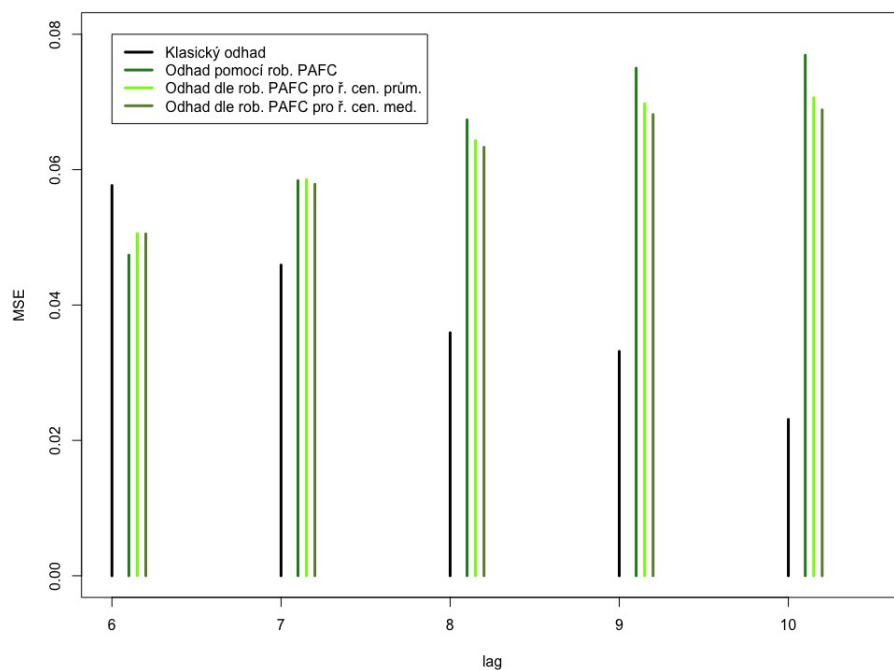


(b) MSE

Obrázek 5.1: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování. Grafický výstup.



(a) BIAS



(b) MSE

Obrázek 5.2: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování, porovnání klasického odhadu a odhadu pomocí robustního odhadu parciální autokorelační funkce pro $t = 6, \dots, 10$. Grafický výstup.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.47637	-0.38362	-0.32567	-0.26785	-0.23857
Vynechávací metoda B=0.025	-0.36161	-0.30767	-0.26541	-0.23254	-0.20881
Vynechávací metoda B=0.050	-0.26378	-0.24497	-0.22152	-0.19655	-0.16851
Vynechávací metoda B=0.100	-0.31721	-0.30562	-0.26699	-0.23910	-0.20531
Jedno. krok. vynech. metoda B=0.050	-0.25019	-0.23460	-0.20903	-0.19425	-0.16948
Dynam. krok. vynech. metoda B=0.050	-0.24889	-0.23187	-0.20737	-0.19385	-0.17049
Mediánová metoda bez cen.	-0.25609	-0.23493	-0.19826	-0.16799	-0.13851
Mediánová metoda s cen. prům.	-0.27544	-0.26549	-0.23690	-0.21464	-0.18264
Mediánová metoda s cen. med.	-0.31483	-0.28159	-0.26449	-0.24697	-0.21196
Koef. metoda (Kendall)	-0.26034	-0.23672	-0.20841	-0.19173	-0.17783
Koef. metoda (Spearman)	-0.29010	-0.24403	-0.21231	-0.19173	-0.17567
Odhad pomocí rob. PAFC	-0.14048	-0.09552	-0.04328	-0.01397	0.00332
Odhad dle rob. PAFC pro ř. cen. prům.	-0.16384	-0.12570	-0.07901	-0.05491	-0.04101
Odhad dle rob. PAFC pro ř. cen. med.	-0.16934	-0.13175	-0.08474	-0.06101	-0.04756
Regresní metoda (OLS)	-0.47658	-0.38442	-0.32587	-0.26829	-0.23874
Regresní metoda (L1)	-0.31703	-0.31292	-0.27082	-0.25106	-0.20879
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.25226	0.16769	0.12674	0.08853	0.07624
Vynechávací metoda B=0.025	0.15988	0.11873	0.08996	0.07054	0.06355
Vynechávací metoda B=0.050	0.08637	0.07419	0.06476	0.05430	0.04298
Vynechávací metoda B=0.100	0.10879	0.10659	0.08419	0.07179	0.05565
Jedno. krok. vynech. metoda B=0.050	0.07495	0.06707	0.05782	0.05427	0.04295
Dynam. krok. vynech. metoda B=0.050	0.07425	0.06507	0.05709	0.05414	0.04361
Mediánová metoda bez cen.	0.10001	0.08775	0.06905	0.05292	0.04757
Mediánová metoda s cen. prům.	0.10697	0.10545	0.08356	0.07391	0.05618
Mediánová metoda s cen. med.	0.13099	0.11049	0.09463	0.08298	0.06167
Koef. metoda (Kendall)	0.08165	0.07265	0.06147	0.05471	0.05204
Koef. metoda (Spearman)	0.09956	0.07604	0.06292	0.05464	0.05229
Odhad pomocí rob. PAFC	0.02894	0.02007	0.01936	0.02542	0.03652
Odhad dle rob. PAFC pro ř. cen. prům.	0.03807	0.02940	0.02707	0.03254	0.04220
Odhad dle rob. PAFC pro ř. cen. med.	0.03997	0.03110	0.02767	0.03285	0.04252
Regresní metoda (OLS)	0.25255	0.16824	0.12678	0.08852	0.07602
Regresní metoda (L1)	0.13164	0.12737	0.09826	0.08581	0.06290

Tabulka 5.1: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování. Tabulkový výstup.

rozdělení

$$\hat{\sigma}^{(MD)} = \sqrt{\frac{\pi}{2}} \times MD(\mathbf{X}) \doteq 1.2533 \times MD(\mathbf{X}),$$

$$\hat{\sigma}^{(MAD)} = \frac{1}{u_{0.75}} \times MAD(\mathbf{X}) \doteq 1.4826 \times MAD(\mathbf{X}),$$

kde $u_{0.75}$ je 0.75 kvantil normovaného normálního rozdělení a $\hat{\sigma}^{(MD)}$ a $\hat{\sigma}^{(MAD)}$ reprezentuje příslušné robustní odhady směrodatné odchylky, které použijeme.

V kapitole 4.1.2 se jako pravidlo pro určení odlehlosti používá $|Z_t| > 3$. Ukážeme si odhad i pro $|Z_t| > 2$. Připomeňme, že v našem případě

$$Z_t = \frac{X_t - \widehat{\text{med}}(\mathbf{X})}{\hat{\sigma}},$$

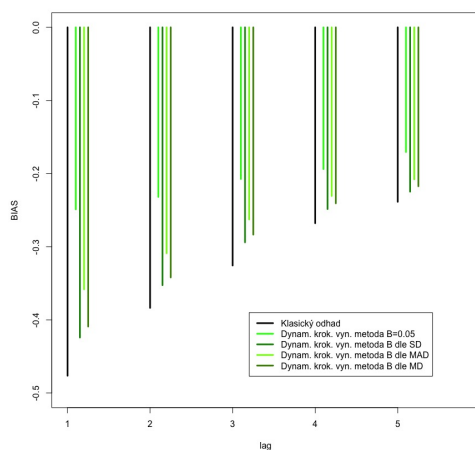
a obdobně použijeme i $\hat{\sigma}^{(MD)}$ a $\hat{\sigma}^{(MAD)}$ jako robustní odhad měřítka.

V simulacích provedeme odhad β pro každou řadu zvlášť. K odhadu autokorelační funkce použijeme pouze dynamickou krokovou vynechávací metodu (4.6). Výsledky simulace nalezneme v tabulce 5.2 a na obrázku 5.3. Vidíme, že nejlépe odpovídá odhad pomocí MAD. Nejhorší výsledek dostáváme za použití SD, což je očekávané, jelikož tento odhad měřítka není příliš robustní.

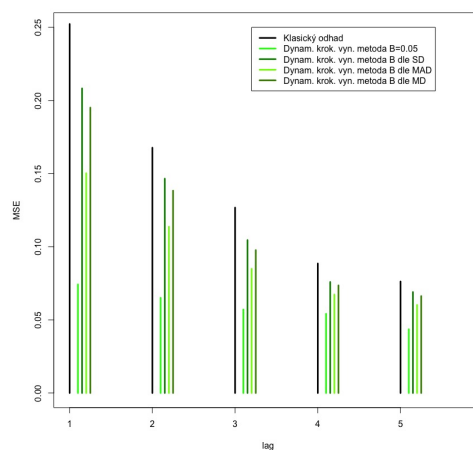
Při použití pravidla $|Z_t| > 3$ dostáváme lepší výsledky než pro klasickou metodu, ale při porovnání s pevně voleným $\beta = 0.05$ se ukazuje, že detekce odlehlých pozorování je nedostatečná. Oproti tomu při použití pravidla $|Z_t| > 2$ v našem případě dostáváme výsledky již srovnatelné s pevným $\beta = 0.05$, při použití MAD dokonce i lepší.

Pravidlo pro rozhodnutí o odlehlosti: $ Z_t > 3$					
BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.47637	-0.38362	-0.32567	-0.26785	-0.23857
Dynam. krok. vyn. metoda B=0.05	-0.24889	-0.23187	-0.20737	-0.19385	-0.17049
Dynam. krok. vyn. metoda B dle SD	-0.42416	-0.35256	-0.29398	-0.24849	-0.22468
Dynam. krok. vyn. metoda B dle MAD	-0.35818	-0.30893	-0.26269	-0.23064	-0.20787
Dynam. krok. vyn. metoda B dle MD	-0.40916	-0.34200	-0.28340	-0.24063	-0.21732
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.25226	0.16769	0.12674	0.08853	0.07624
Dynam. krok. vyn. metoda B=0.05	0.07425	0.06507	0.05709	0.05414	0.04361
Dynam. krok. vyn. metoda B dle SD	0.20825	0.14653	0.10456	0.07590	0.06901
Dynam. krok. vyn. metoda B dle MAD	0.15022	0.11371	0.08491	0.06732	0.06022
Dynam. krok. vyn. metoda B dle MD	0.19510	0.13834	0.09769	0.07357	0.06628
Pravidlo pro rozhodnutí o odlehlosti: $ Z_t > 2$					
BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.47637	-0.38362	-0.32567	-0.26785	-0.23857
Dynam. krok. vyn. metoda B=0.05	-0.24889	-0.23187	-0.20737	-0.19385	-0.17049
Dynam. krok. vyn. metoda B dle SD	-0.28464	-0.25351	-0.22407	-0.20620	-0.18331
Dynam. krok. vyn. metoda B dle MAD	-0.24433	-0.22376	-0.19792	-0.18855	-0.16094
Dynam. krok. vyn. metoda B dle MD	-0.25980	-0.23525	-0.20703	-0.19477	-0.17291
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.25226	0.16769	0.12674	0.08853	0.07624
Dynam. krok. vyn. metoda B=0.05	0.07425	0.06507	0.05709	0.05414	0.04361
Dynam. krok. vyn. metoda B dle SD	0.10220	0.08084	0.06605	0.05934	0.05013
Dynam. krok. vyn. metoda B dle MAD	0.07251	0.06242	0.05280	0.05258	0.04186
Dynam. krok. vyn. metoda B dle MD	0.08531	0.07035	0.05793	0.05520	0.04696

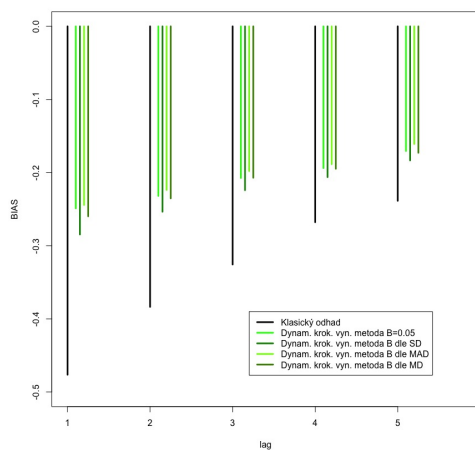
Tabulka 5.2: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování, porovnání klasického odhadu a odhadu pomocí krokové vynechávací metody pro různé způsoby odhadu koeficientu β . K posouzení odlehlosti je použito σ pravidlo pomocí, kterého odhadneme koeficient β . Pro odhad polohy v pravidle jsme použili medián a porovnáváme odhady měřítka za použití SD a robustních MAD, MD. Tabulkový výstup.



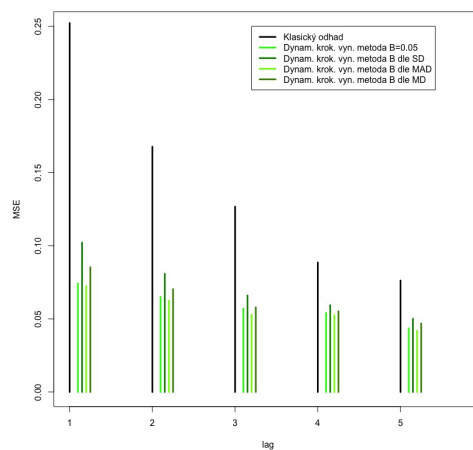
(a) BIAS, $|Z_t| > 3$



(b) MSE, $|Z_t| > 3$



(c) BIAS, $|Z_t| > 2$



(d) MSE, $|Z_t| > 2$

Obrázek 5.3: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování, porovnání klasického odhadu a odhadu pomocí krokové vynechávací metody pro různé způsoby odhadu koeficientu β . K posouzení odlehlosti je použito σ pravidlo pomocí, kterého odhadneme koeficient β . Pro odhad polohy v pravidle jsme použili medián a porovnáváme odhady měřítka za použití SD a robustních MAD, MD. Grafický výstup.

MA(4)

Pracujeme s časovou řadou s odlehlými pozorováními $\{X_t, t \in \mathbb{Z}\}$ dle (5.1), jejíž stacionární složka je MA(4)

$$G_t = Y_t + 0.9Y_{t-1} + 0.6Y_{t-2} + 0.5Y_{t-3} + 0.3Y_{t-4}, \quad t \in \mathbb{Z},$$

s autokorelační funkcí

$$\rho(t) = \begin{cases} = 1, & t = 0 \\ \doteq 0.753, & t = 1, -1 \\ \doteq 0.490, & t = 2, -2 \\ \doteq 0.307, & t = 3, -3 \\ \doteq 0.120, & t = 4, -4 \\ = 0, & |t| > 4. \end{cases}$$

$\{Y_t, t \in \mathbb{Z}\}$ je bílý šum $N(0,1)$ a

$$\begin{aligned} \text{var}(G_t) &= \text{var}(Y_t + 0.9Y_{t-1} + 0.6Y_{t-2} + 0.5Y_{t-3} + 0.3Y_{t-4}) = \\ &= \text{var}(Y_t) + 0.9^2 \text{var}(Y_{t-1}) + 0.6^2 \text{var}(Y_{t-2}) + \\ &+ 0.5^2 \text{var}(Y_{t-3}) + 0.3^2 \text{var}(Y_{t-4}) = \\ &= 1 + 0.9^2 + 0.6^2 + 0.5^2 + 0.3^2 = 2.51. \end{aligned}$$

Výsledky simulace nalezneme v tabulce 5.3 a na obrázku 5.4. Chování metod je obdobné jako v předchozím případě pro AR(1) v kapitole 5.1.1. Nejznatelnější rozdíl mezi klasickým odhadem a robustními metodami pozorujeme pro $t = 1, 2$. Pro $t = 4$ jsou výsledky všech odhadů velmi podobné, kromě odhadu pomocí robustního odhadu parciální autokorelační funkce dle (4.13), pro který s rostoucím t roste i BIAS tohoto odhadu pro necentrovanou metodu a MSE pro všechny varianty této metody. Poznamenejme, že $\rho(5) = 0$.

5.1.2 Aditivní nesymetrická odlehlá pozorování

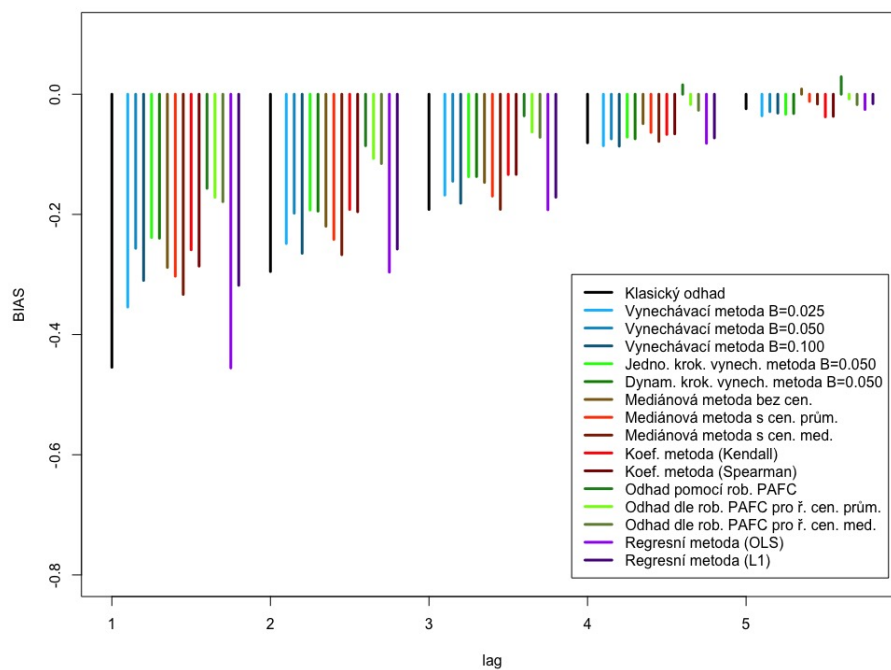
Pracujeme s řadou dle (5.1), jen pro každé $t \in T$ je $V_t = 1$. Z nesymetrické struktury odlehlých pozorování vyplývá, že nyní pracujeme s necentrovanou časovou řadou.

AR(1)

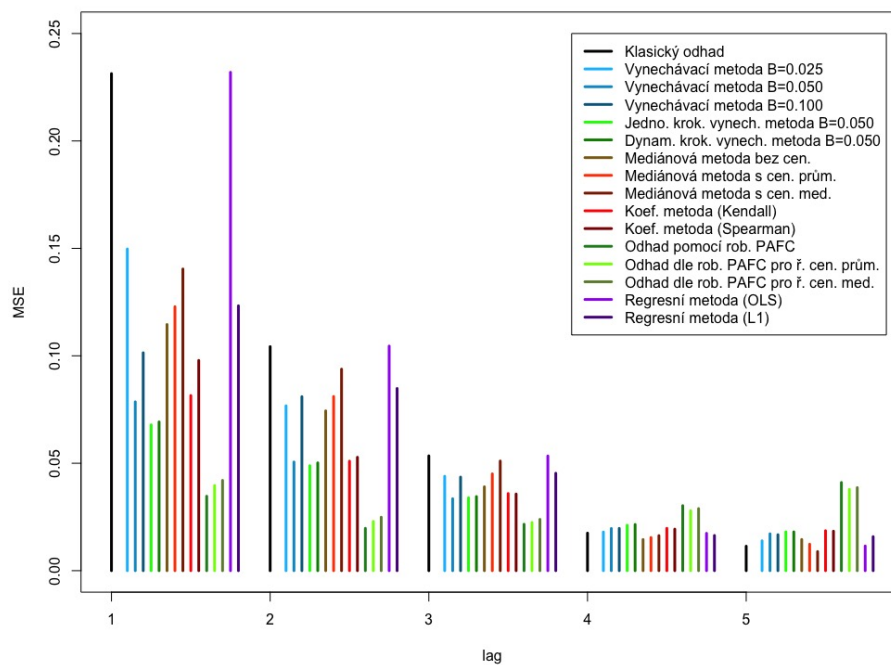
Stacionární složka je AR(1) dle (5.1.1). Výsledky simulace nalezneme v tabulce 5.4 a na obrázku 5.5.

Vidíme, že se vychýlení i střední čtvercová chyba chovají obdobně pro různé $t = 1, \dots, 5$, proto budeme metody hodnotit celkově dle obou kritérií současně.

Vynechávací metoda dle (4.2) dává velmi neuspokojivý výsledek. Metoda ořezává pozorování symetricky, jenže v této situaci jsou odlehlá pozorování nesymetrická. Takže, i když zvolíme dostatečně velké β , které odstraní všechna odlehlá pozorování, tak na druhou stranu přijdeme o velké množství objektivních pozorování, která jsou pro odhad důležitá. V situaci s nesymetrickými odlehlými pozorováními je tato metoda velmi nevhodná, což simulace potvrzují.



(a) BIAS



(b) MSE

Obrázek 5.4: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování. Grafický výstup.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.45447	-0.29505	-0.19178	-0.08076	-0.02413
Vynechávací metoda B=0.025	-0.35414	-0.24845	-0.16789	-0.08570	-0.03588
Vynechávací metoda B=0.050	-0.25641	-0.19798	-0.14481	-0.07431	-0.02901
Vynechávací metoda B=0.100	-0.30985	-0.26471	-0.18142	-0.08638	-0.03144
Jedno. krok. vynech. metoda B=0.050	-0.23856	-0.19271	-0.13717	-0.07144	-0.03327
Dynam. krok. vynech. metoda B=0.050	-0.23959	-0.19459	-0.13696	-0.07413	-0.03212
Mediánová metoda bez cen.	-0.28831	-0.21967	-0.14676	-0.04892	0.00869
Mediánová metoda s cen. prům.	-0.30284	-0.24161	-0.16957	-0.06356	-0.01188
Mediánová metoda s cen. med.	-0.33318	-0.26707	-0.19159	-0.07850	-0.01631
Koef. metoda (Kendall)	-0.25889	-0.19181	-0.13360	-0.06687	-0.03764
Koef. metoda (Spearman)	-0.28603	-0.19555	-0.13344	-0.06594	-0.03679
Odhad pomocí rob. PAFC	-0.15664	-0.08563	-0.03591	0.01568	0.02913
Odhad dle rob. PAFC pro ř. cen. prům.	-0.17147	-0.10693	-0.06292	-0.01714	-0.00789
Odhad dle rob. PAFC pro ř. cen. med.	-0.17881	-0.11534	-0.07165	-0.02665	-0.01726
Regresní metoda (OLS)	-0.45551	-0.29606	-0.19228	-0.08164	-0.02535
Regresní metoda (L1)	-0.31804	-0.25767	-0.17140	-0.07271	-0.01571
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.23134	0.10428	0.05343	0.01748	0.01139
Vynechávací metoda B=0.025	0.14973	0.07673	0.04394	0.01800	0.01391
Vynechávací metoda B=0.050	0.07858	0.05058	0.03351	0.01964	0.01716
Vynechávací metoda B=0.100	0.10141	0.08099	0.04355	0.01968	0.01674
Jedno. krok. vynech. metoda B=0.050	0.06788	0.04887	0.03394	0.02114	0.01806
Dynam. krok. vynech. metoda B=0.050	0.06924	0.05019	0.03449	0.02149	0.01806
Mediánová metoda bez cen.	0.11458	0.07438	0.03903	0.01450	0.01451
Mediánová metoda s cen. prům.	0.12286	0.08107	0.04506	0.01543	0.01232
Mediánová metoda s cen. med.	0.14040	0.09380	0.05103	0.01632	0.00891
Koef. metoda (Kendall)	0.08150	0.05097	0.03589	0.01972	0.01860
Koef. metoda (Spearman)	0.09785	0.05278	0.03569	0.01933	0.01839
Odhad pomocí rob. PAFC	0.03464	0.01970	0.02160	0.03026	0.04104
Odhad dle rob. PAFC pro ř. cen. prům.	0.03962	0.02292	0.02244	0.02791	0.03785
Odhad dle rob. PAFC pro ř. cen. med.	0.04202	0.02489	0.02387	0.02889	0.03866
Regresní metoda (OLS)	0.23194	0.10453	0.05338	0.01738	0.01149
Regresní metoda (L1)	0.12328	0.08481	0.04534	0.01638	0.01583

Tabulka 5.3: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování. Tabulkový výstup.

Oproti tomu jednoduchá kroková vynechávací metoda dle (4.5) a dynamická kroková vynechávací metoda (4.6) jsou flexibilní a přizpůsobí se odlehlosti pozorování. Proto i odhady jsou velmi dobré. Přesto je stále důležité zvolit správné β .

Mediánová metoda dle (4.7) pro necentrovanou metodu dává velmi uspokojivý výsledek srovnatelný s krokovou vynechávací metodou. Pro řady, které jsme centrovali pomocí průměru či mediánu je odhad horší.

Koeficientová metoda za použití Kendallova τ dle (4.10) či Spearmanova koeficientu dle (4.11) dává uspokojivý výsledek srovnatelný s krokovou vynechávací metodou a mediánovou metodou s řadou bez centrování. Kendallovo τ dává o něco lepší výsledek než Spearmanův koeficient.

Velmi dobré výsledky pozorujeme pro odhad pomocí robustního odhadu parciální autokorelační funkce dle (4.13). Nejlepší výsledek dostáváme opět pro necentrovanou metodu. BIAS a MSE opět roste s rostoucím t , znatelněji se projevuje

od $t \geq 8$, což není v tabulkách ani na grafech.

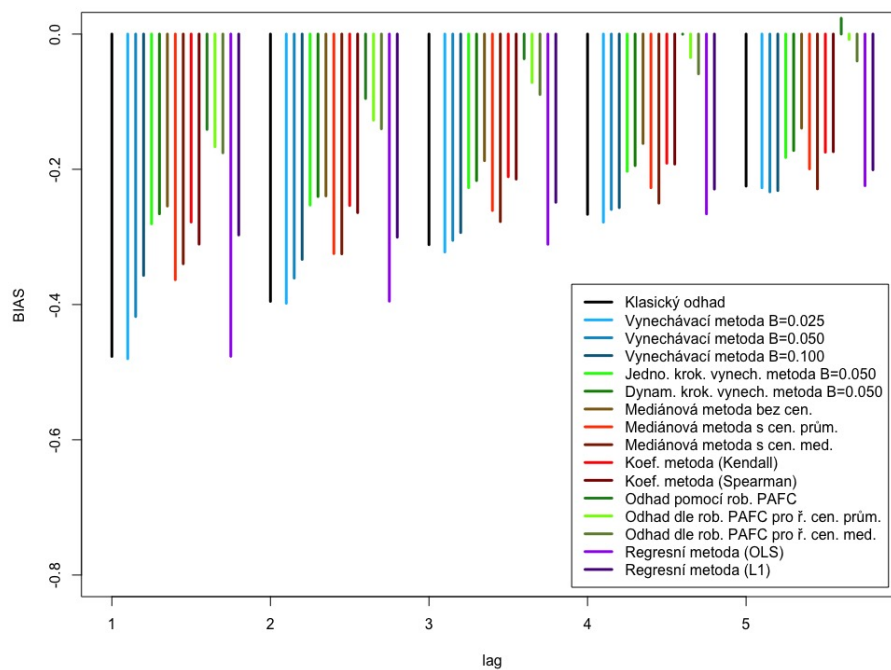
Poslední regresní metoda dle (4.18) opět vychází velmi špatně. L1 odhad o něco lépe než OLS, přesto oproti ostatním metodám výsledek je stále neuspokojivý.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.47677	-0.39530	-0.31144	-0.26657	-0.22483
Vynechávací metoda B=0.025	-0.48012	-0.39809	-0.32232	-0.27820	-0.22733
Vynechávací metoda B=0.050	-0.41778	-0.36093	-0.30508	-0.25941	-0.23336
Vynechávací metoda B=0.100	-0.35689	-0.33326	-0.29355	-0.25667	-0.23143
Jedno. krok. vynech. metoda B=0.050	-0.28061	-0.25293	-0.22714	-0.20303	-0.18250
Dynam. krok. vynech. metoda B=0.050	-0.26583	-0.24025	-0.21668	-0.19451	-0.17219
Mediánová metoda bez cen.	-0.25434	-0.23938	-0.18723	-0.16183	-0.13916
Mediánová metoda s cen. prům.	-0.36327	-0.32466	-0.26093	-0.22728	-0.19939
Mediánová metoda s cen. med.	-0.33969	-0.32501	-0.27734	-0.25004	-0.22891
Koef. metoda (Kendall)	-0.27803	-0.25353	-0.21099	-0.19099	-0.17467
Koef. metoda (Spearman)	-0.31064	-0.26405	-0.21449	-0.19231	-0.17393
Odhad pomocí rob. PAFC	-0.14102	-0.09531	-0.03652	-0.00022	0.02334
Odhad dle rob. PAFC pro ř. cen. prům.	-0.16661	-0.12736	-0.07168	-0.03478	-0.00800
Odhad dle rob. PAFC pro ř. cen. med.	-0.17564	-0.14012	-0.08938	-0.05882	-0.03982
Regresní metoda (OLS)	-0.47651	-0.39502	-0.31084	-0.26589	-0.22421
Regresní metoda (L1)	-0.29712	-0.30042	-0.24877	-0.22922	-0.20076
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.25493	0.17929	0.11825	0.08635	0.06460
Vynechávací metoda B=0.025	0.25635	0.17810	0.12111	0.09132	0.06447
Vynechávací metoda B=0.050	0.19912	0.14984	0.10855	0.08066	0.06999
Vynechávací metoda B=0.100	0.14350	0.12521	0.09862	0.07914	0.06714
Jedno. krok. vynech. metoda B=0.050	0.09334	0.07849	0.06510	0.05403	0.04758
Dynam. krok. vynech. metoda B=0.050	0.08589	0.07209	0.05992	0.05127	0.04493
Mediánová metoda bez cen.	0.10502	0.09534	0.07005	0.05805	0.05205
Mediánová metoda s cen. prům.	0.16183	0.13311	0.08906	0.07036	0.05637
Mediánová metoda s cen. med.	0.14369	0.13388	0.10311	0.08089	0.06709
Koef. metoda (Kendall)	0.09467	0.08192	0.06416	0.05446	0.04936
Koef. metoda (Spearman)	0.11427	0.08689	0.06490	0.05465	0.04909
Odhad pomocí rob. PAFC	0.03550	0.02538	0.02024	0.02159	0.02453
Odhad dle rob. PAFC pro ř. cen. prům.	0.04067	0.02912	0.02146	0.02010	0.02053
Odhad dle rob. PAFC pro ř. cen. med.	0.04448	0.03386	0.02615	0.02486	0.02549
Regresní metoda (OLS)	0.25515	0.17959	0.11841	0.08665	0.06494
Regresní metoda (L1)	0.12209	0.12079	0.09022	0.07015	0.05805

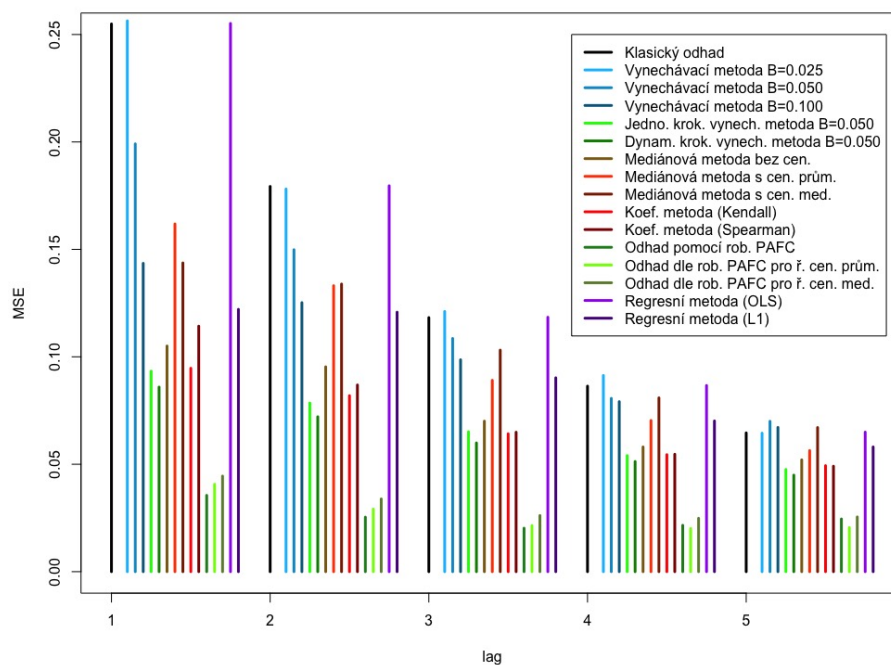
Tabulka 5.4: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování. Tabulkový výstup.

MA(4)

Stacionární složka je MA(4) dle 5.1.1. Výsledky simulace nalezneme v tabulce 5.5 a na obrázku 5.6. Chování metod je obdobné jako v předchozím případě pro AR(1) v kapitole 5.1.2. Odhad pomocí robustního odhadu parciální autokorelační funkce dle (4.13) se opět vyvíjí jinak oproti ostatním odhadům pro $t = 4, 5$ a lze jej popsat stejně jako v kapitole 5.1.1. Znatelně nám roste MSE pro všechny varianty této metody.

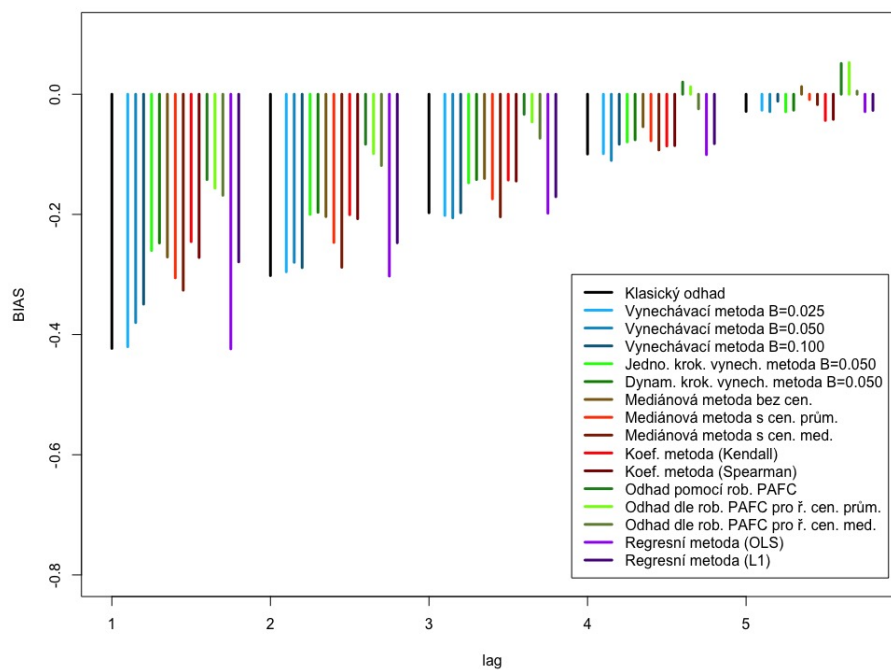


(a) BIAS

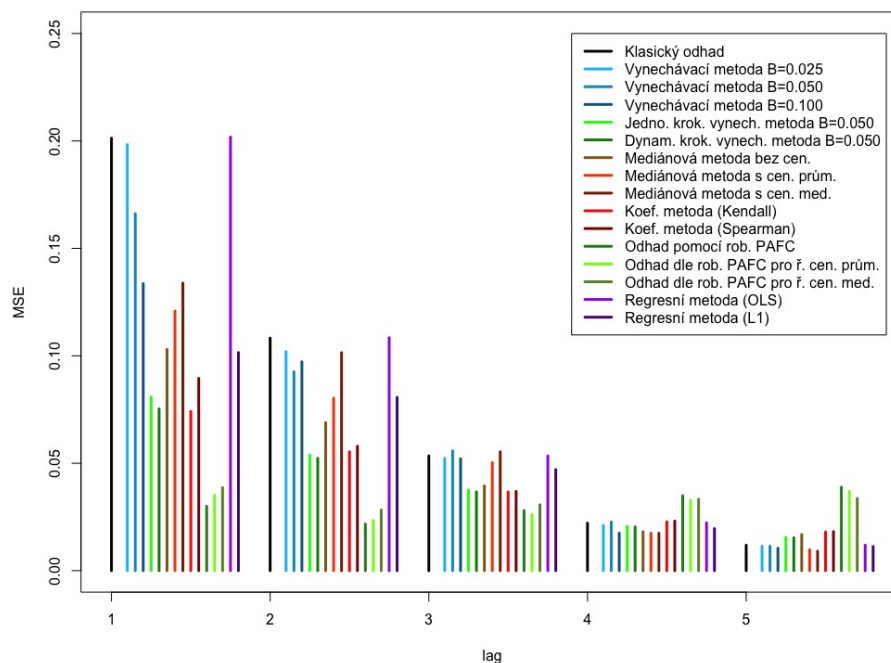


(b) MSE

Obrázek 5.5: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování.



(a) BIAS



(b) MSE

Obrázek 5.6: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování. Grafický výstup.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.42306	-0.30184	-0.19724	-0.09952	-0.02861
Vynechávací metoda B=0.025	-0.42020	-0.29548	-0.20176	-0.09886	-0.02653
Vynechávací metoda B=0.050	-0.38007	-0.27975	-0.20570	-0.11022	-0.02896
Vynechávací metoda B=0.100	-0.34931	-0.28860	-0.19722	-0.08322	-0.01161
Jedno. krok. vynech. metoda B=0.050	-0.26017	-0.20008	-0.14749	-0.07948	-0.02917
Dynam. krok. vynech. metoda B=0.050	-0.24755	-0.19654	-0.14204	-0.07577	-0.02651
Mediánová metoda bez cen.	-0.27096	-0.20363	-0.14017	-0.05406	0.01272
Mediánová metoda s cen. prům.	-0.30563	-0.24670	-0.17416	-0.07738	-0.00882
Mediánová metoda s cen. med.	-0.32614	-0.28803	-0.20401	-0.09273	-0.01724
Koef. metoda (Kendall)	-0.24533	-0.20052	-0.14282	-0.08609	-0.04358
Koef. metoda (Spearman)	-0.27165	-0.20718	-0.14452	-0.08550	-0.04185
Odhad pomocí rob. PAFC	-0.14204	-0.08319	-0.03343	0.02026	0.05119
Odhad dle rob. PAFC pro ř. cen. prům.	-0.15633	-0.09886	-0.04606	0.01229	0.05260
Odhad dle rob. PAFC pro ř. cen. med.	-0.16820	-0.11860	-0.07317	-0.02428	0.00511
Regresní metoda (OLS)	-0.42378	-0.30264	-0.19798	-0.10047	-0.02896
Regresní metoda (L1)	-0.27902	-0.24721	-0.17056	-0.08225	-0.02688
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.20132	0.10828	0.05340	0.02213	0.01188
Vynechávací metoda B=0.025	0.19830	0.10192	0.05229	0.02106	0.01140
Vynechávací metoda B=0.050	0.16616	0.09259	0.05583	0.02263	0.01141
Vynechávací metoda B=0.100	0.13361	0.09723	0.05209	0.01751	0.01048
Jedno. krok. vynech. metoda B=0.050	0.08085	0.05381	0.03753	0.02066	0.01551
Dynam. krok. vynech. metoda B=0.050	0.07530	0.05230	0.03668	0.02041	0.01536
Mediánová metoda bez cen.	0.10297	0.06890	0.03945	0.01816	0.01685
Mediánová metoda s cen. prům.	0.12081	0.08035	0.05028	0.01747	0.00975
Mediánová metoda s cen. med.	0.13386	0.10149	0.05534	0.01745	0.00911
Koef. metoda (Kendall)	0.07414	0.05537	0.03665	0.02274	0.01805
Koef. metoda (Spearman)	0.08953	0.05791	0.03694	0.02312	0.01820
Odhad pomocí rob. PAFC	0.02997	0.02174	0.02799	0.03491	0.03893
Odhad dle rob. PAFC pro ř. cen. prům.	0.03514	0.02336	0.02617	0.03266	0.03695
Odhad dle rob. PAFC pro ř. cen. med.	0.03867	0.02825	0.03066	0.03327	0.03368
Regresní metoda (OLS)	0.20178	0.10846	0.05341	0.02227	0.01185
Regresní metoda (L1)	0.10154	0.08067	0.04704	0.01960	0.01127

Tabulka 5.5: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování. Tabulkový výstup.

5.1.3 Inovační symetrická odlehlá pozorování

Časová řada obsahuje inovační symetrická odlehlá pozorování

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_m X_{t-m} + \tilde{Y}_t + \theta_1 \tilde{Y}_{t-1} + \dots + \theta_n \tilde{Y}_{t-n}, \quad t \in \mathbb{Z} \quad (5.2)$$

kde $m, n \in \mathbb{N}_0$ a $\{\tilde{Y}_t, t \in \mathbb{Z}\}$ pochází z kontaminovaného rozdělení dle definice 16, kde navíc $\{G_t, t \in \mathbb{Z}\}$ je bílý šum pocházející z $N(0,1)$. Časová řada $\{\tilde{Y}_t, t \in \mathbb{Z}\}$ je

$$\tilde{Y}_t = U_t G_t + (1 - U_t) R_t V_t,$$

kde R_t je z rovnoměrného rozdělení $R(2, 4)$ a

$$V_t = 2Z_t - 1,$$

kde Z_t je z alternativního rozdělení $Alt(0.5)$. Tím získáme přibližně symetrickou strukturu odlehlých pozorování. U_t pochází z alternativního rozdělení $Alt(0.9)$. $\{\tilde{Y}_t, t \in \mathbb{Z}\}$ zastupuje roli bílého šumu v ARMA struktuře.

AR(1)

Pracujeme s časovou řadou s odlehlými pozorováními $\{X_t, t \in \mathbb{Z}\}$ dle (5.2) danou předpisem

$$X_t = 0.8X_{t-1} + \tilde{Y}_t, \quad t \in \mathbb{Z}.$$

Výsledky simulace nalezneme v tabulce 5.6 a na obrázku 5.7. Vidíme, že oproti aditivním odlehlým pozorováním nemají inovační odlehlá pozorování takový vliv na odhad autokorelační funkce. Toto jsme již pozorovali v kapitole 2. Klasický odhad dle (2.2) dává velmi dobré a přesné hodnoty. Poznamenejme, že toto chování závisí i na naší definici struktury inovačních odlehlých pozorování dle definice 18.

Vynechávací metoda dle (4.2) a kroková vynechávací metoda dle (4.5) a (4.6) dávají velmi zavádějící a špatné výsledky. Vidíme, že s rostoucím β roste BIAS i MSE.

Mediánová metoda dle (4.7) pro necentrovanou metodu dává přesnější výsledek. Pro centrované varianty je odhad opět špatný.

Koeficientová metoda za použití Kendallova τ dle (4.10) či Spearmanova koeficientu dle (4.11) dává srovnatelně dobrý výsledek jako klasický odhad, stejně tak regresní metoda dle (4.18).

Velmi dobré počáteční výsledky pozorujeme opět pro odhad pomocí robustního odhadu parciální autokorelační funkce dle 4.13. S rostoucím t opět roste BIAS a MSE. U odhadu za použití necentrované metody pozorujeme opačné znaménko vychýlení než u zbylých metod.

MA(4)

Pracujeme s časovou řadou s odlehlými pozorováními $\{X_t, t \in \mathbb{Z}\}$ dle (5.2) danou předpisem

$$X_t = \tilde{Y}_t + 0.9\tilde{Y}_{t-1} + 0.6\tilde{Y}_{t-2} + 0.5\tilde{Y}_{t-3} + 0.3\tilde{Y}_{t-4}, \quad t \in \mathbb{Z}.$$

Výsledky simulace nalezneme v tabulce 5.7 a na obrázku 5.8. Chování metod je obdobné jako v předchozím případě pro AR(1) v kapitole 5.1.3. Povšimněme si, že MSE je velmi podobné pro všechny metody pro $t = 4, 5$

5.1.4 Inovační nesymetrická odlehlá pozorování

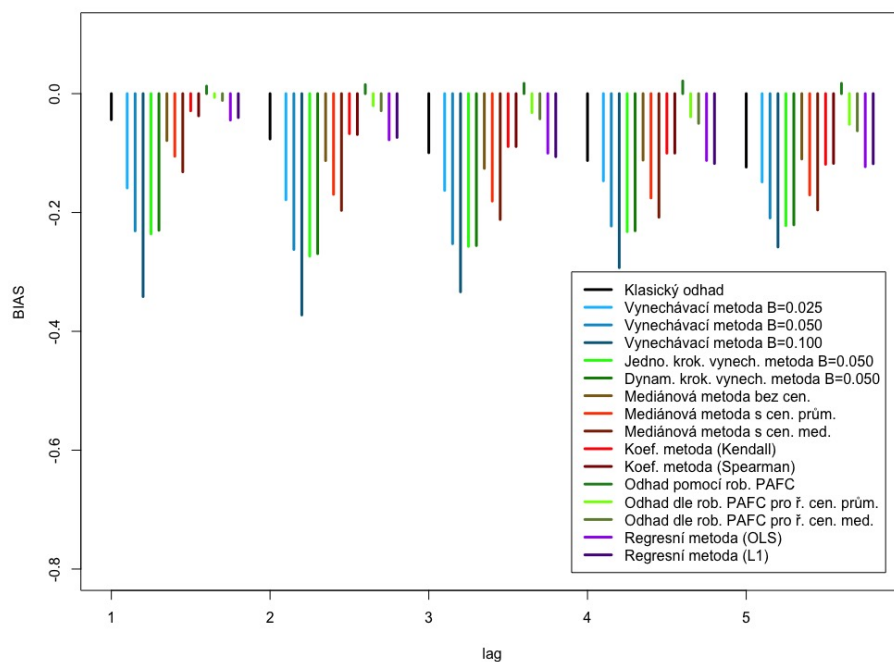
Pracujeme s řadou dle (5.2), jen pro každé $t \in T$ je $V_t = 1$. Z nesymetrické struktury odlehlých pozorování pracujeme s necentrovanou časovou řadou.

AR(1)

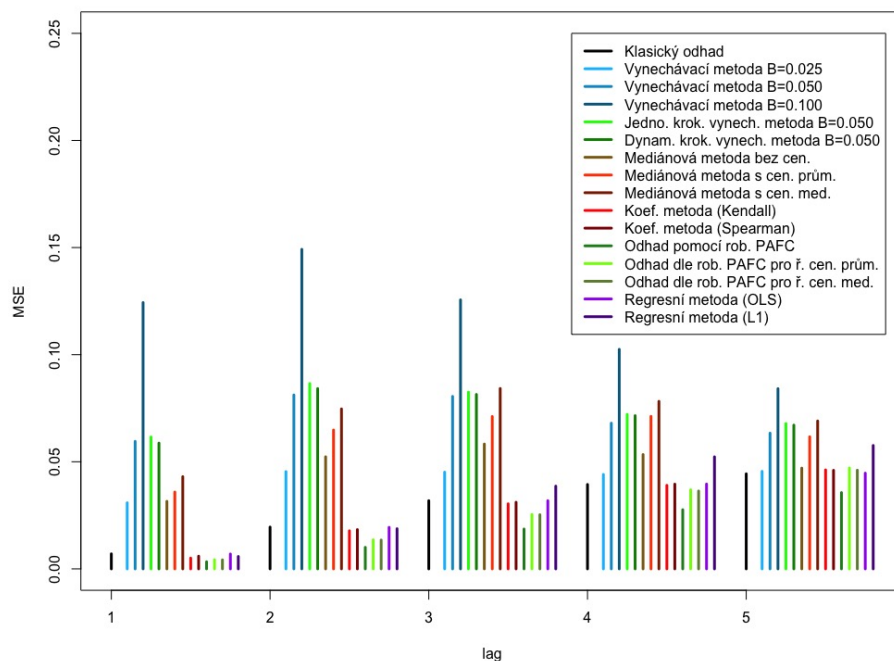
Časová řada $\{X_t, t \in \mathbb{Z}\}$ je dána předpisem (5.1.3). Výsledky simulace nalezneme v tabulce 5.8 a na obrázku 5.9.

Klasický odhad dle (2.2) opět vychází velmi dobře a vynechávací metody dle (4.2), (4.5), (4.6) oproti tomu velmi špatně.

Mediánová metoda dle (4.7) dává neuspokojivý výsledek oproti předchozím výsledkům. Pro necentrovanou metodu nabývá BIAS podobných hodnot jako pro klasický odhad, ale v opačném znaménku. MSE vychází znatelně vyšší. Dále pro centrované metody je BIAS i MSE velmi špatné.

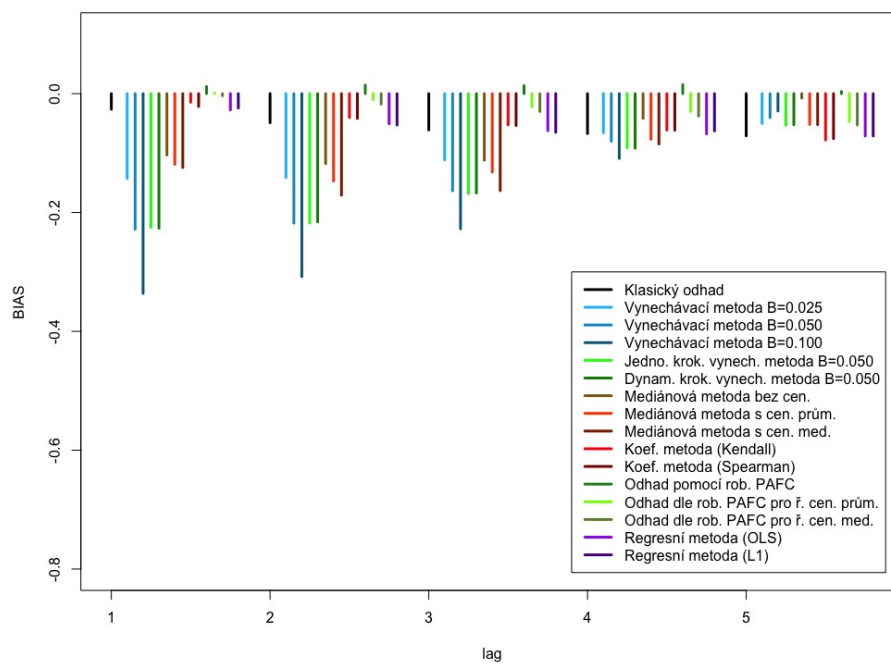


(a) BIAS

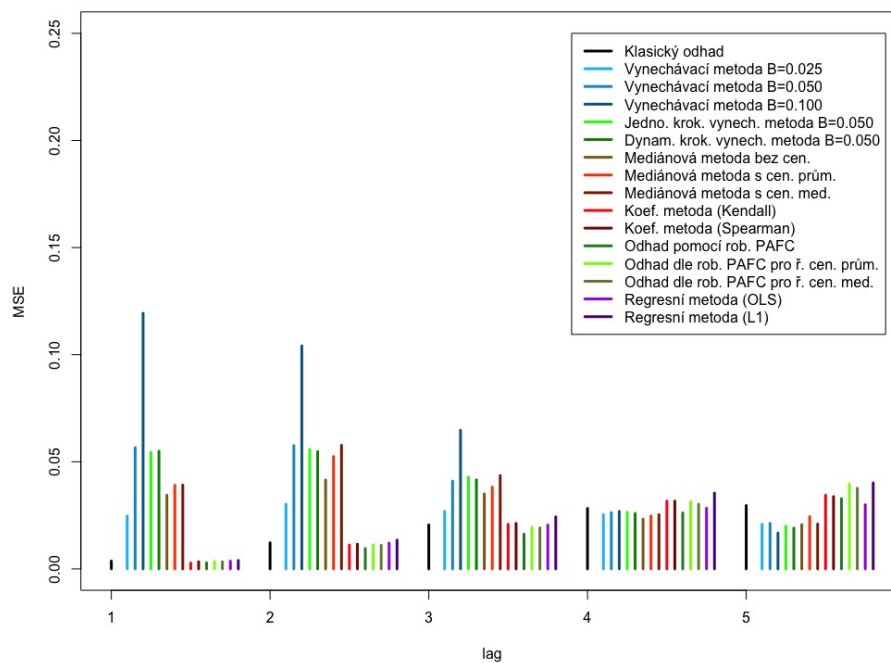


(b) MSE

Obrázek 5.7: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování. Grafický výstup.



(a) BIAS



(b) MSE

Obrázek 5.8: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování. Grafický výstup.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.04386	-0.07627	-0.09940	-0.11259	-0.12360
Vynechávací metoda B=0.025	-0.15881	-0.17860	-0.16287	-0.14676	-0.14857
Vynechávací metoda B=0.050	-0.23094	-0.26227	-0.25255	-0.22277	-0.20927
Vynechávací metoda B=0.100	-0.34153	-0.37251	-0.33355	-0.29287	-0.25808
Jedno. krok. vynech. metoda B=0.050	-0.23600	-0.27347	-0.25710	-0.23225	-0.22215
Dynam. krok. vynech. metoda B=0.050	-0.22996	-0.26929	-0.25568	-0.23073	-0.22079
Mediánová metoda bez cen.	-0.07896	-0.11272	-0.12582	-0.11187	-0.11005
Mediánová metoda s cen. prům.	-0.10542	-0.16953	-0.18084	-0.17563	-0.17043
Mediánová metoda s cen. med.	-0.13161	-0.19647	-0.21182	-0.20801	-0.19575
Koef. metoda (Kendall)	-0.02890	-0.06735	-0.08906	-0.10015	-0.11891
Koef. metoda (Spearman)	-0.03733	-0.06879	-0.08904	-0.10009	-0.11744
Odhad pomocí rob. PAFC	0.01265	0.01523	0.01729	0.02124	0.01729
Odhad dle rob. PAFC pro ř. cen. prům.	-0.00627	-0.02039	-0.03199	-0.03905	-0.05164
Odhad dle rob. PAFC pro ř. cen. med.	-0.01144	-0.02876	-0.04241	-0.05000	-0.06257
Regresní metoda (OLS)	-0.04450	-0.07769	-0.10017	-0.11246	-0.12294
Regresní metoda (L1)	-0.04027	-0.07375	-0.10611	-0.11760	-0.11800
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.00702	0.01951	0.03183	0.03940	0.04438
Vynechávací metoda B=0.025	0.03086	0.04541	0.04515	0.04406	0.04552
Vynechávací metoda B=0.050	0.05952	0.08119	0.08052	0.06802	0.06337
Vynechávací metoda B=0.100	0.12432	0.14917	0.12558	0.10248	0.08413
Jedno. krok. vynech. metoda B=0.050	0.06155	0.08651	0.08251	0.07213	0.06786
Dynam. krok. vynech. metoda B=0.050	0.05872	0.08415	0.08143	0.07152	0.06711
Mediánová metoda bez cen.	0.03153	0.05232	0.05827	0.05337	0.04704
Mediánová metoda s cen. prům.	0.03587	0.06481	0.07114	0.07120	0.06166
Mediánová metoda s cen. med.	0.04305	0.07468	0.08421	0.07821	0.06905
Koef. metoda (Kendall)	0.00503	0.01778	0.03037	0.03900	0.04620
Koef. metoda (Spearman)	0.00592	0.01833	0.03110	0.03955	0.04602
Odhad pomocí rob. PAFC	0.00333	0.01005	0.01861	0.02762	0.03561
Odhad dle rob. PAFC pro ř. cen. prům.	0.00425	0.01358	0.02539	0.03693	0.04712
Odhad dle rob. PAFC pro ř. cen. med.	0.00421	0.01346	0.02528	0.03635	0.04607
Regresní metoda (OLS)	0.00694	0.01939	0.03182	0.03960	0.04469
Regresní metoda (L1)	0.00578	0.01874	0.03862	0.05230	0.05761

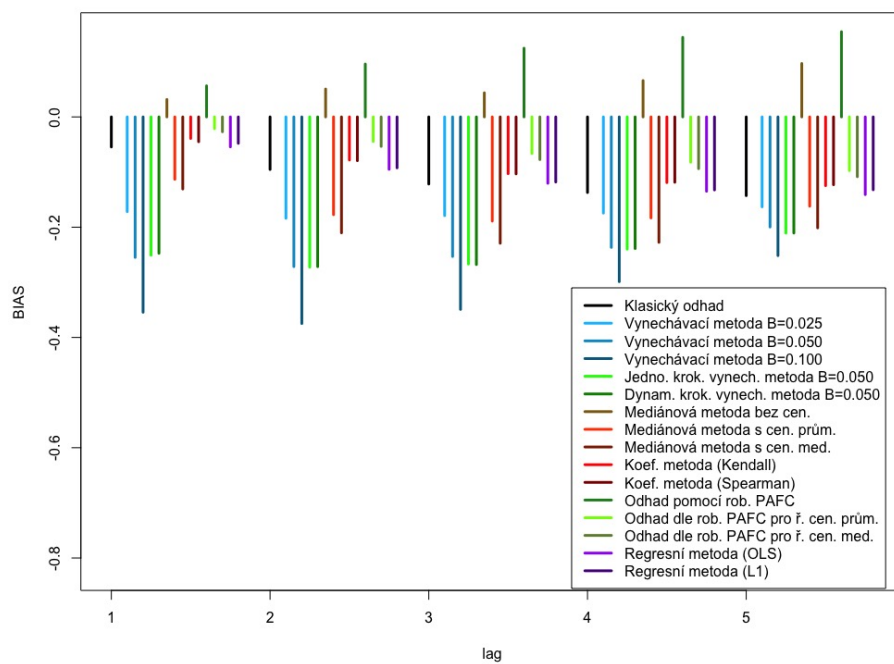
Tabulka 5.6: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování. Tabulkový výstup.

Koeficientová metoda za použití Kendallova τ dle (4.10) či Spearmanova koeficientu dle (4.11) dává srovnatelně dobrý výsledek jako klasický odhad, stejně tak regresní metoda dle (4.18)

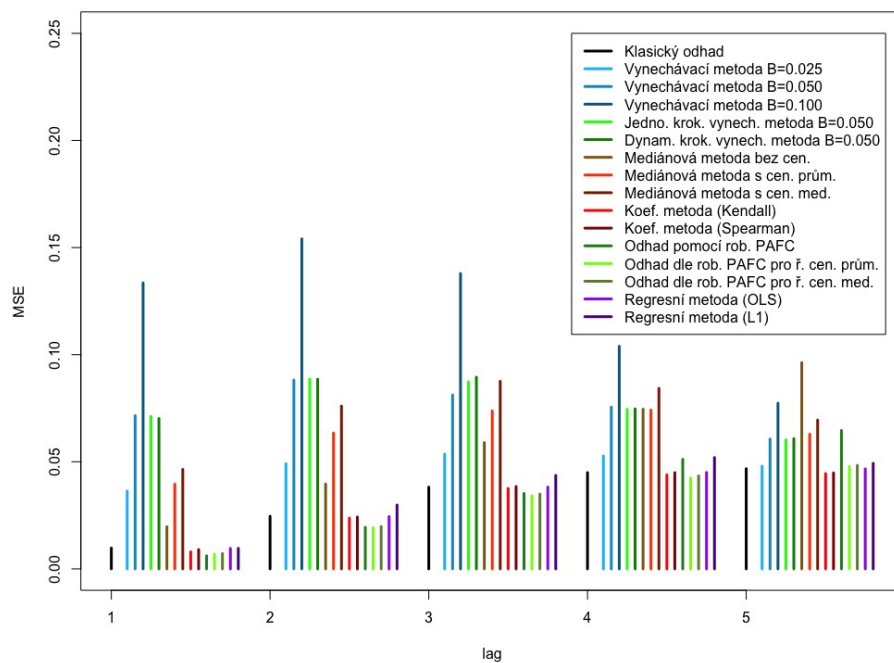
Odhad pomocí robustního odhadu parciální autokorelační funkce dle (4.13) dává dobré výsledky pro centrovanou metodu. U necentrované metody s rostoucím t opět roste BIAS a MSE. U odhadu za použití necentrované metody pozorujeme opačné znaménko vychýlení než u zbylých metod (kromě mediánové, jak bylo výše zmíněno).

MA(4)

Časová řada $\{X_t, t \in \mathbb{Z}\}$ je dána předpisem (5.1.3). Výsledky simulace nalezneme v tabulce 5.9 a na obrázku 5.10. Chování metod je obdobné jako v předchozím případě pro AR(1) v kapitole 5.1.4. Povšimněme si, že MSE je velmi podobné



(a) BIAS



(b) MSE

Obrázek 5.9: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování. Grafický výstup.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.02639	-0.04899	-0.06104	-0.06716	-0.07107
Vynechávací metoda B=0.025	-0.14285	-0.14107	-0.11132	-0.06579	-0.05016
Vynechávací metoda B=0.050	-0.22820	-0.21800	-0.16354	-0.08029	-0.04027
Vynechávací metoda B=0.100	-0.33617	-0.30792	-0.22745	-0.10906	-0.02923
Jedno. krok. vynech. metoda B=0.050	-0.22528	-0.21783	-0.16828	-0.09103	-0.05341
Dynam. krok. vynech. metoda B=0.050	-0.22634	-0.21585	-0.16717	-0.09180	-0.05238
Mediánová metoda bez cen.	-0.10332	-0.11778	-0.11212	-0.04168	-0.00748
Mediánová metoda s cen. prům.	-0.11889	-0.14707	-0.13199	-0.07670	-0.05185
Mediánová metoda s cen. med.	-0.12431	-0.17079	-0.16323	-0.08481	-0.05226
Koef. metoda (Kendall)	-0.01441	-0.04035	-0.05265	-0.06140	-0.07783
Koef. metoda (Spearman)	-0.02162	-0.04171	-0.05367	-0.06167	-0.07571
Odhad pomocí rob. PAFC	0.01209	0.01459	0.01335	0.01512	0.00375
Odhad dle rob. PAFC pro ř. cen. prům.	0.00080	-0.01005	-0.02202	-0.03004	-0.04716
Odhad dle rob. PAFC pro ř. cen. med.	-0.00354	-0.01766	-0.03012	-0.03774	-0.05273
Regresní metoda (OLS)	-0.02741	-0.05085	-0.06238	-0.06768	-0.07129
Regresní metoda (L1)	-0.02418	-0.05288	-0.06518	-0.06291	-0.07121
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.00365	0.01220	0.02054	0.02820	0.02958
Vynechávací metoda B=0.025	0.02468	0.03019	0.02685	0.02530	0.02086
Vynechávací metoda B=0.050	0.05659	0.05758	0.04099	0.02622	0.02123
Vynechávací metoda B=0.100	0.11932	0.10412	0.06473	0.02686	0.01674
Jedno. krok. vynech. metoda B=0.050	0.05439	0.05574	0.04281	0.02647	0.01991
Dynam. krok. vynech. metoda B=0.050	0.05492	0.05474	0.04159	0.02589	0.01904
Mediánová metoda bez cen.	0.03436	0.04150	0.03501	0.02324	0.02063
Mediánová metoda s cen. prům.	0.03901	0.05244	0.03815	0.02472	0.02436
Mediánová metoda s cen. med.	0.03911	0.05770	0.04355	0.02531	0.02094
Koef. metoda (Kendall)	0.00275	0.01113	0.02082	0.03163	0.03438
Koef. metoda (Spearman)	0.00332	0.01153	0.02123	0.03162	0.03375
Odhad pomocí rob. PAFC	0.00288	0.00953	0.01621	0.02611	0.03279
Odhad dle rob. PAFC pro ř. cen. prům.	0.00335	0.01120	0.01940	0.03130	0.03954
Odhad dle rob. PAFC pro ř. cen. med.	0.00330	0.01102	0.01916	0.03024	0.03762
Regresní metoda (OLS)	0.00361	0.01203	0.02047	0.02831	0.02993
Regresní metoda (L1)	0.00394	0.01345	0.02430	0.03537	0.04014

Tabulka 5.7: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování. Tabulkový výstup.

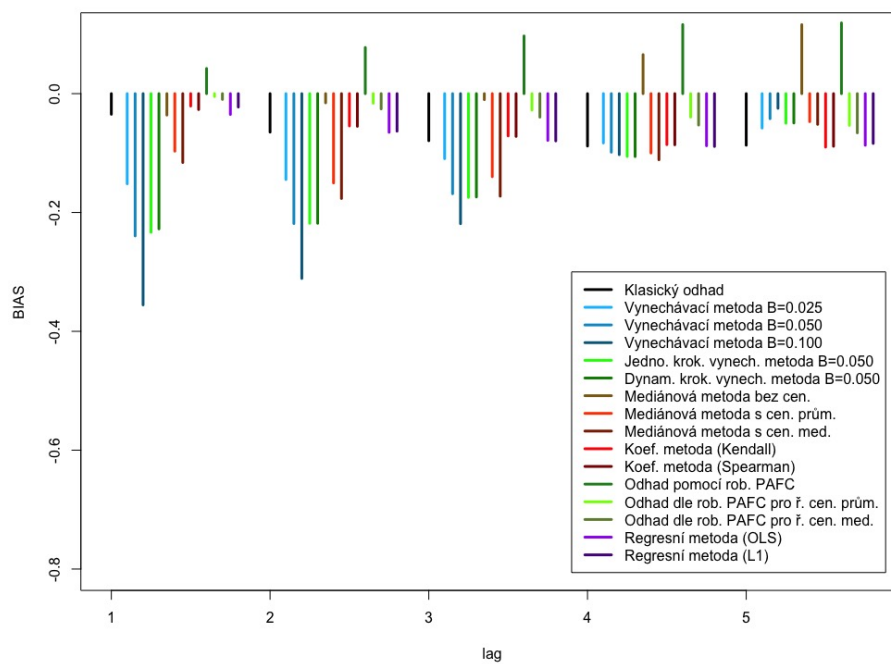
pro všechny metody pro $t = 4, 5$.

5.1.5 Souhrn simulací pro AR(1)

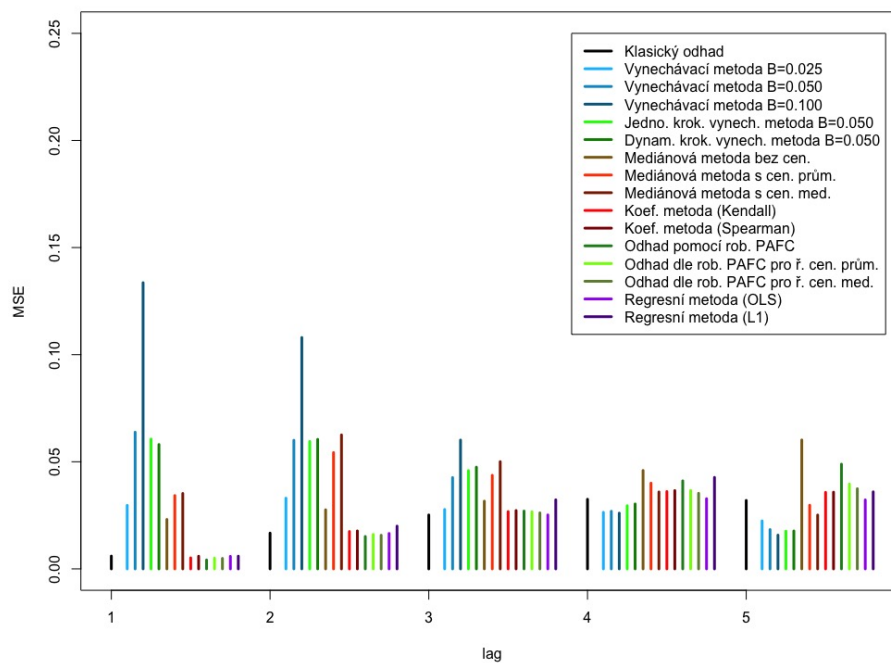
Simulace ukázaly, že aditivní odlehlá pozorování mají značný vliv na chování klasického odhadu a jeho hodnoty mohou být nepoužitelné.

U vynechávací metody je velkou nevýhodou nutnost volby parametru β , který silně ovlivňuje chování odhadu. Dále je tato metoda neúspěšná pro nesymetrická odlehlá pozorování. Kroková metoda dokáže podchytit nesymetričnost a velmi dobře se s ní vypořádá, ale stále je závislá na určení správného parametru β .

Mediánová metoda nám dala uspokojujivý výsledek. V simulaci jsme pracovali s časovou řadou, která byla dle své struktury centrována. Lepší výsledky jsme dostali pro necentrovanou metodu. Poznamenejme, že v situaci, kdy časová řada je necentrována, tak je nutné ji centrovat. Vyžadují to předpoklady metody.



(a) BIAS



(b) MSE

Obrázek 5.10: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování. Grafický výstup.

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.05447	-0.09531	-0.12169	-0.13710	-0.14275
Vynechávací metoda B=0.025	-0.17181	-0.18385	-0.17902	-0.17444	-0.16309
Vynechávací metoda B=0.050	-0.25479	-0.27144	-0.25310	-0.23664	-0.19967
Vynechávací metoda B=0.100	-0.35427	-0.37477	-0.34912	-0.29879	-0.25152
Jedno. krok. vynech. metoda B=0.050	-0.25087	-0.27240	-0.26680	-0.23972	-0.21075
Dynam. krok. vynech. metoda B=0.050	-0.24736	-0.27155	-0.26777	-0.23874	-0.21063
Mediánová metoda bez cen.	0.03165	0.05057	0.04371	0.06591	0.09686
Mediánová metoda s cen. prům.	-0.11315	-0.17732	-0.18887	-0.18323	-0.16193
Mediánová metoda s cen. med.	-0.13071	-0.21023	-0.22913	-0.22749	-0.20124
Koef. metoda (Kendall)	-0.03909	-0.07809	-0.10294	-0.11906	-0.12467
Koef. metoda (Spearman)	-0.04498	-0.07935	-0.10332	-0.11855	-0.12297
Odhad pomocí rob. PAFC	0.05639	0.09588	0.12461	0.14434	0.15456
Odhad dle rob. PAFC pro ř. cen. prům.	-0.02134	-0.04454	-0.06641	-0.08215	-0.09736
Odhad dle rob. PAFC pro ř. cen. med.	-0.02692	-0.05346	-0.07739	-0.09384	-0.10842
Regresní metoda (OLS)	-0.05432	-0.09503	-0.12037	-0.13485	-0.14082
Regresní metoda (L1)	-0.04793	-0.09238	-0.11809	-0.13244	-0.13212
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.00970	0.02456	0.03816	0.04495	0.04678
Vynechávací metoda B=0.025	0.03632	0.04913	0.05362	0.05271	0.04798
Vynechávací metoda B=0.050	0.07151	0.08813	0.08121	0.07551	0.06063
Vynechávací metoda B=0.100	0.13348	0.15405	0.13789	0.10390	0.07736
Jedno. krok. vynech. metoda B=0.050	0.07113	0.08858	0.08731	0.07451	0.06024
Dynam. krok. vynech. metoda B=0.050	0.07018	0.08855	0.08946	0.07470	0.06083
Mediánová metoda bez cen.	0.01966	0.03960	0.05896	0.07454	0.09629
Mediánová metoda s cen. prům.	0.03954	0.06339	0.07375	0.07415	0.06287
Mediánová metoda s cen. med.	0.04646	0.07601	0.08758	0.08426	0.06947
Koef. metoda (Kendall)	0.00796	0.02373	0.03752	0.04387	0.04449
Koef. metoda (Spearman)	0.00901	0.02423	0.03844	0.04493	0.04478
Odhad pomocí rob. PAFC	0.00609	0.01940	0.03522	0.05114	0.06463
Odhad dle rob. PAFC pro ř. cen. prům.	0.00681	0.01918	0.03406	0.04239	0.04777
Odhad dle rob. PAFC pro ř. cen. med.	0.00715	0.01976	0.03492	0.04340	0.04835
Regresní metoda (OLS)	0.00954	0.02439	0.03819	0.04502	0.04667
Regresní metoda (L1)	0.00959	0.02982	0.04364	0.05197	0.04928

Tabulka 5.8: Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování. Tabulkový výstup.

Koeficientová metoda nám dávala konstantně dobré výsledky. Odhady dle Kendallova τ a Spearmanova koeficientu byly velmi podobné, dle Kendallova τ vycházel odhad o něco málo lépe.

Velmi dobře vycházely odhady pomocí robustního odhadu parciální autokorelační funkce. Lépe pro řadu, kterou jsme necentrovali průměrem nebo mediánem. Metoda opět vyžaduje centrovanou časovou řadu. Připomeňme, že časová řada je dle své struktury centrovaná, proto není potřeba ji centrovat. Tento odhad byl často nepřesný pro vyšší hodnoty t , v našem případě $t \geq 8$.

Regresní metoda nám dala neuspokojivé výsledky, L1 odhad byl lepší než klasický, ale nevyrovnal se jiným robustním metodám.

Oproti tomu inovační odlehlá pozorování klasický odhad příliš neovlivnily. Vynechávací a kroková vynechávací metoda vycházela velmi špatně a nebyla vhodná pro odhad autokorelační funkce. Nepříliš dobře vycházela i mediánová metoda. Koeficientová i regresní metoda vyšla velmi dobře a podobně jako klasický od-

BIAS	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	-0.03474	-0.06478	-0.07921	-0.08821	-0.08680
Vynechávací metoda B=0.025	-0.15172	-0.14456	-0.10942	-0.08303	-0.05788
Vynechávací metoda B=0.050	-0.23946	-0.21865	-0.16845	-0.09854	-0.04216
Vynechávací metoda B=0.100	-0.35553	-0.31099	-0.21890	-0.10270	-0.02467
Jedno. krok. vynech. metoda B=0.050	-0.23336	-0.21811	-0.17457	-0.10569	-0.04975
Dynam. krok. vynech. metoda B=0.050	-0.22760	-0.21825	-0.17390	-0.10585	-0.04938
Mediánová metoda bez cen.	-0.03608	-0.01547	-0.01013	0.06566	0.11607
Mediánová metoda s cen. prům.	-0.09682	-0.15029	-0.13975	-0.09991	-0.04709
Mediánová metoda s cen. med.	-0.11592	-0.17644	-0.17257	-0.11121	-0.05171
Koef. metoda (Kendall)	-0.02090	-0.05439	-0.07114	-0.08580	-0.08996
Koef. metoda (Spearman)	-0.02674	-0.05500	-0.07197	-0.08608	-0.08858
Odhad pomocí rob. PAFC	0.04232	0.07766	0.09707	0.11620	0.11903
Odhad dle rob. PAFC pro ř. cen. prům.	-0.00487	-0.01615	-0.02761	-0.03949	-0.05340
Odhad dle rob. PAFC pro ř. cen. med.	-0.00973	-0.02570	-0.03959	-0.05289	-0.06600
Regresní metoda (OLS)	-0.03506	-0.06508	-0.07870	-0.08772	-0.08688
Regresní metoda (L1)	-0.02261	-0.06320	-0.07957	-0.08888	-0.08379
MSE	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.00597	0.01670	0.02516	0.03249	0.03194
Vynechávací metoda B=0.025	0.02966	0.03300	0.02774	0.02639	0.02239
Vynechávací metoda B=0.050	0.06380	0.06003	0.04264	0.02689	0.01835
Vynechávací metoda B=0.100	0.13359	0.10801	0.06013	0.02602	0.01576
Jedno. krok. vynech. metoda B=0.050	0.06058	0.05948	0.04588	0.02948	0.01761
Dynam. krok. vynech. metoda B=0.050	0.05805	0.06044	0.04748	0.03030	0.01770
Mediánová metoda bez cen.	0.02307	0.02757	0.03160	0.04594	0.06017
Mediánová metoda s cen. prům.	0.03422	0.05430	0.04368	0.04001	0.02971
Mediánová metoda s cen. med.	0.03520	0.06253	0.05005	0.03591	0.02514
Koef. metoda (Kendall)	0.00514	0.01744	0.02675	0.03610	0.03572
Koef. metoda (Spearman)	0.00590	0.01772	0.02723	0.03655	0.03581
Odhad pomocí rob. PAFC	0.00417	0.01511	0.02699	0.04109	0.04888
Odhad dle rob. PAFC pro ř. cen. prům.	0.00499	0.01609	0.02664	0.03660	0.03958
Odhad dle rob. PAFC pro ř. cen. med.	0.00488	0.01571	0.02613	0.03528	0.03742
Regresní metoda (OLS)	0.00585	0.01653	0.02518	0.03276	0.03219
Regresní metoda (L1)	0.00594	0.01996	0.03224	0.04269	0.03597

Tabulka 5.9: Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování. Tabulkový výstup.

had. Odhad pomocí robustního odhadu parciální autokorelační funkce dává velmi přesné výsledky. U časových řad centrovaných pomocí průměru či mediánu pozorujeme dobré chování. U řady, kterou jsme necentrovali, nastává problém s rostoucím vychýlením a střední čtvercovou chybou pro rostoucí t . Připomeňme, že toto chování závisí i na naší definici struktury inovačních odlehlých pozorování dle definice 18.

5.1.6 Souhrn simulací pro MA(4)

Pro MA(4) jsme pozorovali obdobné chování odhadů jako pro AR(1). U aditivních odlehlých pozorování se robustní metody uplatnily a odhady jimi získané vykazovaly znatelně přesnější výsledky než pomocí klasického odhadu. U inovačních odlehlých pozorování jsme dostávali obecně horší nebo srovnatelné výsledky pomocí robustních metod jako pomocí klasického odhadu. Totéž jsme pozorovali

pro simulace s AR(1).

Nejvýraznější rozdíly byly pro odhady $t = 1, 2$. Pro $t = 4, 5$ dávaly všechny metody více méně obdobné výsledky. Toto chování lze očekávat jelikož $\rho(5) = 0$ a $\rho(4) = 0.12$, což je velmi blízko nuly a špatně rozpoznatelné.

5.2 Reálná data

Robustní odhady použijeme na reálná data. Budeme se zabývat cenou akcií¹ společnosti Johnson&Johnson, což je globální farmaceutická firma. Pracovat budeme s časovou řadou $\{X_t, t = 1, \dots, n\}$ udávající rozdíl mezi nejvyšší a nejnižší denní cenou akcie od listopadu 2018 do konce února 2019 s denní frekvencí v pracovní dny, rozsah dat je $n = 80$ a cena je udána v USD. Tedy

$$X_t = V_t - W_t, \quad t = 1, \dots, 80,$$

kde $\{V_t, t = 1, \dots, n\}$ je časovou řadou nejvyšších denních cen akcie a $\{W_t, t = 1, \dots, n\}$ je časovou řadou nejnižších denních cen akcie. Poznamenejme, že zkoumaná řada $\{X_t, t = 1, \dots, n\}$ ze své definice nabývá pouze nezáporných hodnot a odlehlá pozorování jsou nesymetrická.

Graf nejvyšších a nejnižších denních cen akcie je uveden na obrázku 5.11a a jejich rozdíl, tedy námi zkoumaná časová řada, na obrázku 5.11b. Pozorujeme značný výkyv oproti zbytku řady ve 30–tém pozorování, odpovídající datu 14. 12. 2018. Tento propad byl zapříčiněn zprávou agentury Reuters (Girion, 14.12.2018), že společnost věděla o příležitostné přítomnosti azbestu ve svém produktu, dětském pudru.

Ve vynechávací metodě jsme volili rozsah $\beta = 0.02, 0.05, 0.10$ na základě odhadu dle kapitoly 4.1.2. V krokové vynechávací metodě jsme volili $\beta = 0.05$. Zkoumaná řada $\{X_t, t = 1, \dots, n\}$ na obrázku 5.11b není zjevně centrovaná, proto nebudeme aplikovat necentrovane verze mediánové metody dle (4.7) a odhadu pomocí robustního odhadu parciální autokorelační funkce dle (4.13), které mají centrovanost řady v předpokladech. Tabulka 5.10 a graf 5.11c shrnují robustní odhady autokorelační funkce.

Klasický odhad nám dává statisticky nevýznamné nebo velmi hraniční výsledky pro $t = 1, 5$. U vynechávací metody dle (4.2) a krokové vynechávací metody dle (4.5) a (4.6) se ukazuje důležitost správného rozpoznání koeficientu β určující míru oříznutí. Pro $t = 1, 2$ dostáváme mírně významnější výsledky než pro klasický odhad. Pro mediánovou metodu dle (4.7) využívající centrování výběrovým průměrem dostáváme statisticky významné odhady. V případě centrování mediánem jsou výsledky znatelně horší. To může být zapříčiněno nesplněním symetrie dat, kterou tato metoda vyžaduje. Koeficientová metoda dle (4.10) a (4.11) dává všechny hodnoty odhadu statisticky významné a v předchozích simulacích se velmi osvědčila.

Odhad pomocí robustního odhadu parciální autokorelační funkce dle (4.13) pro centrovanou metodu za použití výběrového průměru dostává velmi vysoké hodnoty odhadu autokorelační funkce. To může být způsobeno vychýlením výběrového průměru, který je náchylný na odlehlá pozorování. Tím je do výpočtu

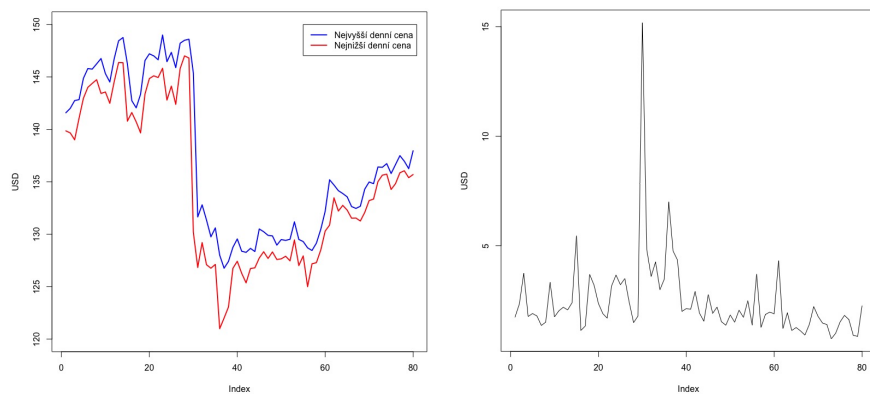
¹<https://www.macrotrends.net/stocks/charts/JNJ/johnson-johnson/stock-price-history>

již od začátku vnesena chyba. Při využití výběrového mediánu dostáváme hodnoty podobné mediánové metodě při centrování výběrovým průměrem. Regresní metoda za využití L1 odhadu vychází statisticky významně pouze pro $t = 1$.

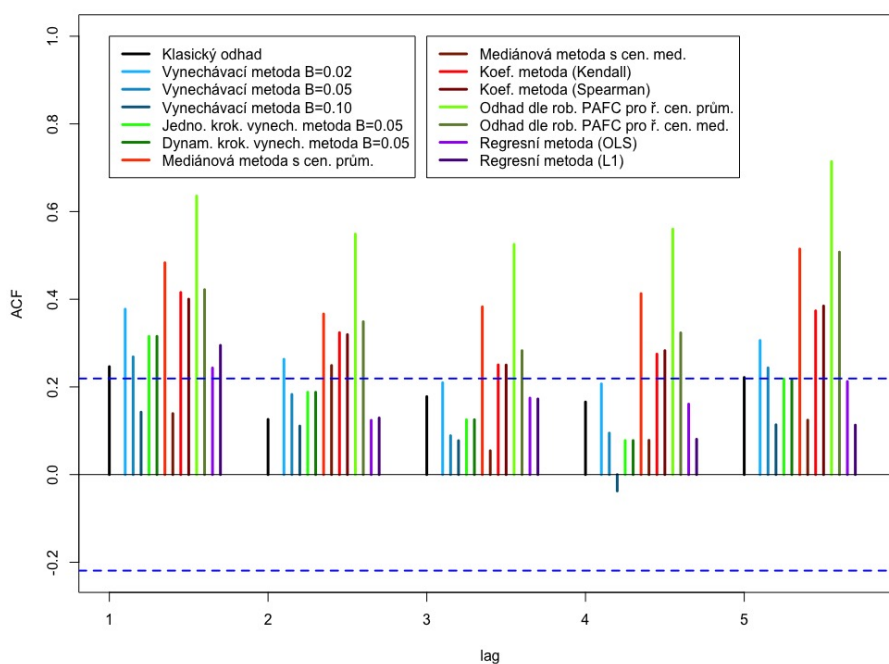
Na reálných datech vidíme užitečnost robustních odhadů. Pro tato data lze doporučit například koeficientovou metodu a odhad pomocí robustního odhadu parciální autokorelační funkce při centrování mediánem, které se ukázaly jako uspokojivé i v simulacích.

Metoda	$\hat{\rho}(1)$	$\hat{\rho}(2)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$
Klasický odhad	0.24629	0.12589	0.17792	0.16566	0.22159
Vynechávací metoda B=0.02	0.37737	0.26312	0.20953	0.20696	0.30600
Vynechávací metoda B=0.05	0.26860	0.18290	0.08874	0.09470	0.24351
Vynechávací metoda B=0.10	0.14246	0.11078	0.07740	-0.03738	0.11368
Jedno. krok. vynech. metoda B=0.05	0.31535	0.18781	0.12515	0.07771	0.21784
Dynam. krok. vynech. metoda B=0.05	0.31535	0.18781	0.12515	0.07771	0.21784
Mediánová metoda s cen. prům.	0.48330	0.36637	0.38266	0.41283	0.51461
Mediánová metoda s cen. med.	0.13910	0.24890	0.05441	0.07823	0.12420
Koef. metoda (Kendall)	0.41548	0.32376	0.25029	0.27490	0.37357
Koef. metoda (Spearman)	0.40025	0.31933	0.24999	0.28276	0.38457
Odhad dle rob. PAFC pro ř. cen. prům.	0.63539	0.54871	0.52512	0.56008	0.71412
Odhad dle rob. PAFC pro ř. cen. med.	0.42186	0.34877	0.28273	0.32338	0.50750
Regresní metoda (OLS)	0.24323	0.12386	0.17458	0.16078	0.21249
Regresní metoda (L1)	0.29482	0.12918	0.17259	0.08066	0.11292

Tabulka 5.10: Robustní odhady ACF pro časovou řadu rozdílů nejvyšší a nejnižší denní ceny akcie společnosti Johnson&Johnson od listopadu 2018 do konce února 2019.



(a) Časová řada nejvyšších a nej- (b) Časová řada rozdílů nejvyšších
nižších denních cen a nejnižších denních cen



(c) Odhady ACF

Obrázek 5.11: Časová řada rozdílů nejvyšší a nejnižší denní ceny akcie společnosti Johnson&Johnson od listopadu 2018 do konce února 2019 a robustní odhady ACF.

Závěr

V této práci jsme se zabývali robustním odhadem autokorelační funkce stationární časové řady. Klasický odhad je značně náchylný na odlehlá pozorování v časové řadě, což může zapříčinit nepřesné a zavádějící výsledky. Řešili jsme různé robustní odhady zmírňující vliv odlehlých pozorování.

V první kapitole jsme zavedli potřebný aparát, definovali autokorelační funkci a uvedli její základní vlastnosti.

Druhá kapitola se zabývala klasickým odhadem autokorelační funkce, uvedli jsme formální vzorec pro výpočet a vlastnosti klasického odhadu jako vychýlení, konzistenci a asymptotické chování. Odvodili jsme limitní vliv odlehlých pozorování a na příkladech ukázali závažnost nedostatků klasického odhadu.

Ve třetí kapitole jsme řešili základy robustní statistiky, které jsme potřebovali v dalších kapitolách, a popsali různé typy odlehlých pozorování. Zohlednili jsme aditivní odlehlá pozorování, která mají svojí odlehlostí vliv pouze sama na sebe a odlehlost se nepřenáší na další pozorování, a inovační odlehlá pozorování, která svou odlehlostí ovlivní další pozorování.

Čtvrtá kapitola byla zasvěcena robustním metodám odhadu autokorelační funkce. U každé metody jsme uvedli formální vzorec, nutné předpoklady na časovou řadu, výhody a nevýhody metody a spočetli robustní odhad na konkrétní časové řadě. Metody byly porovnány v simulační studii v páté kapitole.

První byla *vynechávací metoda*, ta odstraní část nejvyšších a nejnižších hodnot časové řady a zbytek použije pro výpočet odhadu. Tato metoda se projevila jako velmi efektivní při odstranění vlivu odlehlých pozorování, ale ukázala se značná slabina ve volbě parametru určujícího počet odstraněných pozorování. V případě špatné volby tohoto parametru buď nebyla odstraněna všechna odlehlá pozorování nebo bylo odstraněno příliš pozorování, i ta adekvátní. Také tato metoda odstraňuje stejný počet nejvyšších a nejnižších hodnot, což v případě nesymetrických odlehlých pozorování vede opět na problém nedostatečného nebo přílišného odstranění pozorování.

Na základě vynechávací metody jsme navrhli vlastní *krokovou vynechávací metodu*. Ta k odstraňování pozorování přistupuje flexibilně a na základě míry odlehlosti vybírá pozorování, která je třeba odstranit. Tato flexibilita je extrémně dobrá pro nesymetrická odlehlá pozorování. Uvedli jsme ještě modifikaci této metody, která je robustnější při posouzení míry odlehlosti daného pozorování. Výsledky byly velmi podobné. Nevýhodou této metody je opět potřeba určit procento pozorování, která nebudou použita pro výpočet odhadu. Nejsme si vědomi, že by daná metoda existovala v literatuře. Nakonec jsme popsali, jak řešit odhad počtu odlehlých pozorování.

Další *mediánová metoda* nahrazuje průměr mediánem pro výpočet odhadu. Chování odhadu bylo velmi dobré a přesné, ale metoda vyžaduje předpoklady na symetrii časové řady a odlehlých pozorování.

Vliv odlehlých pozorování můžeme zmírnit transformací dat, ze které vychází *transformační metoda*. Pomocí různých vah a transformací lze zmírnit odlehlost v časové řadě. Transformovaná data jsme poté dosadili do klasického odhadu. Tato metoda se nejevila obecně příliš vhodná, nebylo navrženo jak řešit zkreslení a vychýlení odhadu pro transformovaná data. Také navržené transformace

závisely na volbě hodnot parametrů, které určují míru transformace.

Koeficientová metoda vychází z odhadu Kendallova τ a Spearmanova koeficientu. Oba tyto koeficienty jsou značně robustní a lze je transformací převést na korelační koeficient, který reprezentuje autokorelační funkci. Pomocí obou koeficientů jsme dostali velmi uspokojivé a podobné výsledky.

Předposlední metoda vycházela z *robustního odhadu parciální autokorelační funkce*. Pomocí iterací jsme spočetli robustní odhad parciální autokorelační funkce založený na specifických vahách. Tento odhad jsme poté použili pro rekurentní přepočítání na robustní odhad autokorelační funkce pomocí Durbinova-Levinsonova algoritmu. Tato metoda vyžaduje centrovanou časovou řadu. Výsledky byly velmi přesné pro nižší hodnoty autokorelační funkce, v našem případě pro $t = 1, \dots, 5$. Pro pozdější hodnoty, přibližně $t \geq 8$, docházelo ke značnému vychýlení a velké střední čtvercové chybě.

Poslední metoda pracovala s *regresní přímkou*, kdy jsme pomocí robustního L1 odhadu odhadli směrnici regresní přímky pro vhodně posunutá data. Tento odhad dával lepší výsledky než klasický, ale oproti jiným robustním metodám byly výsledky horší.

Některé metody mohou zapříčinit porušení oboru hodnot autokorelační funkce či pozitivní semidefinitnosti, na konci čtvrté kapitoly jsme zmínili způsoby řešení těchto problémů.

Pátá kapitola obsahuje simulační studii. Porovnali jsme metody mezi sebou vzhledem k vychýlení a střední čtvercové chybě. Pracovali jsme s klouzavými průměry a autoregresním modelem. Zohlednili jsme aditivní a inovační odlehlá pozorování. Také jsme zohlednili, zda jsou odlehlá pozorování rozmístěna symetricky nebo nesymetricky oproti zbytku časové řady. Výsledky jsme zahrnuli do hodnocení jednotlivých metod výše. Potvrdil se devastující vliv odlehlých pozorování na klasický odhad autokorelační funkce a výhody robustních metod. Také se ukázala významnost symetrie a nesymetrie odlehlých pozorování, kvůli které jsme navrhli krokovou vynechávací metodu. U inovačních odlehlých pozorování se ukázalo, že jejich vliv na klasický odhad není znatelný. Dokonce některé robustní metody v tomto případě dávaly horší výsledky než klasický odhad, jelikož jejich použitím dochází ke zkreslení informace o datech. To může souviset i s naší definicí struktury inovačních odlehlých pozorování. Vše jsme doplnili názornými grafickými výstupy a podrobnými tabulkami. Nakonec jsme metody použili na reálná data.

Seznam použité literatury

- ANDERSON, T. W. (1994). *The statistical analysis of time series*. John Wiley & Sons. ISBN 0-471-04745-7.
- ANDĚL, J. (2005). *Základy matematické statistiky*. Matfyzpress, Praha. ISBN 80-86732-40-1. 1. vydání.
- BOUDT, K., CORNELISSEN, J. a CROUX, C. (2012). The Gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, **22**(2), 471–483.
- BROCKWELL, P. J. a DAVIS, R. A. (2002). *Introduction to time series and forecasting*. Springer. ISBN 0-387-95351-5. 2. vydání.
- BROCKWELL, P. J., DAVIS, R. A. a FIENBERG, S. E. (1991). *Time Series: Theory and Methods*. Springer Science & Business Media. ISBN 1-4419-0319-8. 2. vydání.
- BUSTOS, O. H. a YOHAI, V. J. (1986). Robust estimates for ARMA models. *Journal of the American Statistical Association*, **81**(393), 155–168.
- CHAKHCHOUKH, Y. (2010). A new robust estimation method for ARMA models. *IEEE Transactions on Signal Processing*, **58**(7), 3512–3522.
- CHAN, W.-S. a WEI, W. W. (1992). A comparison of some estimators of time series autocorrelations. *Computational statistics & data analysis*, **14**(2), 149–163.
- CHANG, C. C. a POLITIS, D. N. (2014). Robust autocorrelation estimation. *Journal of Computational and Graphical Statistics*, **25**(1), 144–166.
- CIPRA, T. (2008). *Finanční ekonometrie*. Ekopress. ISBN 978-80-86929-43-9. 1. vydání.
- CROUX, C. a DEHON, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical methods & applications*, **19**(4), 497–515.
- DUPAČ, V. a HUŠKOVÁ, M. (2013). *Pravděpodobnost a matematická statistika*. Karolinum. ISBN 978-80-246-2208-8. 2. vydání.
- DÜRRE, A., FRIED, R. a LIBOSCHIK, T. (2015). Robust estimation of (partial) autocorrelation. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**(3), 205–222.
- GAREL, B. a HALLIN, M. (1999). Rank-based autoregressive order identification. *Journal of the American Statistical Association*, **94**(448), 1357–1371.
- GIRION, L. (14.12.2018). *Johnson&Johnson knew for decades that asbestos lurked in its Baby Powder*. URL <https://www.reuters.com/investigates/special-report/johnsonandjohnson-cancer/>.

- KHURI, A. I. (2010). *Linear model methodology*. Chapman and Hall/CRC. ISBN 978-1-58488-481-1.
- MAKHOUL, J. (1981). Lattice methods in spectral estimation. *Applied Time Series Analysis II*. 301–325.
- MARONNA, R., MARTIN, R. D. a YOHAI, V. (2006). *Robust statistics: theory and practice*. Wiley, New York. ISBN 978-0-470-01092-1.
- MARTIN, R. D. a YOHAI, V. J. (1986). Influence functionals for time series. *The annals of Statistics*. 781–818.
- MASAROTTO, G. (1987). Robust identification of autoregressive moving average models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **36**(2), 214–220.
- McMURRY, T. L. a POLITIS, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, **31**(6), 471–482.
- MOTTONEN, J., KOIVUNEN, V. a OJA, H. (1999). Robust autocovariance estimation based on sign and rank correlation coefficients. In *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics. SPW-HOS'99*. IEEE. 187–190.
- OSORIO, F. (14.11.2017). *Routines for L1 Estimation*. URL <https://cran.r-project.org/web/packages/L1pack/L1pack.pdf>. Package L1pack.
- PRÁGEROVÁ, A. (1971). *Diferenční rovnice*. SNTL. 1. vydání.
- PRÁŠKOVÁ, Z. (2004). *Základy náhodných procesů II*. Univerzita Karlova, Karolinum Press. ISBN 80-246-0971-1. 1. vydání.
- R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- ROUSSEEUW, P. J. a CROUX, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, **88**(424), 1273–1283.
- SHEVLYAKOV, G. L. a OJA, H. (2016). *Robust correlation: Theory and applications*. John Wiley & Sons. ISBN 9781119264538. 3. vydání.
- ZVÁRA, K. (2008). *Regrese*. Matfyzpress, vydavatelství matematickofyzikální fakulty univerzity Karlovy. ISBN 978-80-7378-041-8. 1. vydání.

Seznam obrázků

2.1	Porovnání klasického odhadu s a bez odlehlých pozorování pro MA(2)	23
2.2	Porovnání klasického odhadu s a bez odlehlých pozorování pro AR(1)	24
2.3	Klasický odhad pro časové řady generované bílým šumem pocházejícím z rozdělení s těžkými chvosty	25
2.4	Porovnání klasického odhadu časové řady s jedním odlehlým pozorováním měnící velikost	26
3.1	Hustota kontaminovaného rozdělení	28
3.2	Graf Huberovy a bisquare funkce pro $k = 3$	31
3.3	LS-odhad a L1-odhad regresní přímky	35
3.4	Symetrická a nesymetrická aditivní a inovační odlehlá pozorování	37
3.5	Odlehlá pozorování s posunem a jiným rozptylem	38
4.1	Simulovaná data pro časovou řadu AR(1)	41
4.2	Odhad autokorelační funkce vynechávací metodou pro různé volby β pro časovou řadu AR(1)	44
4.3	Odhad autokorelační funkce krokovou vynechávací metodou pro časovou řadu AR(1)	46
4.4	Odhad autokorelační funkce mediánovou metodou pro časovou řadu AR(1)	48
4.5	Odhad autokorelační funkce transformační metodou pro časovou řadu AR(1)	51
4.6	Odhad autokorelační funkce koeficientovou metodou pro časovou řadu AR(1)	53
4.7	Odhad autokorelační funkce pomocí robustního odhadu parciální autokorelační funkce pro časovou řadu AR(1)	56
4.8	Odhad autokorelační funkce regresní metodou pro časovou řadu AR(1)	61
5.1	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování.	67
5.2	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování, porovnání klasického odhadu a odhadu pomocí robustního odhadu parciální autokorelační funkce pro $t = 6, \dots, 10$.	68
5.3	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování, porovnání klasického odhadu a odhadu pomocí krokové vynechávací metody pro různé způsoby odhadu koeficientu β .	71
5.4	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování.	73
5.6	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování.	77
5.7	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování.	80

5.8	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování.	81
5.9	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování.	83
5.10	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování.	85
5.11	Časová řada rozdílů nejvyšší a nejnižší denní ceny akcie společnosti Johnson&Johnson od listopadu 2018 do konce února 2019 a robustní odhady ACF.	90

Seznam tabulek

1.1	Chování autokorelační a autokovarianční funkce k identifikaci modelu	14
5.1	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování.	69
5.2	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování, porovnání klasického odhadu a odhadu pomocí krokové vynečávací metody pro různé způsoby odhadu koeficientu β .	70
5.3	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická aditivní odlehlá pozorování.	74
5.4	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování.	75
5.5	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická aditivní odlehlá pozorování.	78
5.6	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování.	82
5.7	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, symetrická inovační odlehlá pozorování.	84
5.8	Simulovaná data pro časovou řadu AR(1). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování.	86
5.9	Simulovaná data pro časovou řadu MA(4). BIAS, MSE, $n = 100$, nesymetrická inovační odlehlá pozorování.	87
5.10	Robustní odhady ACF pro časovou řadu rozdílů nejvyšší a nejnižší denní ceny akcie společnosti Johnson&Johnson od listopadu 2018 do konce února 2019.	89

Seznam použitých zkratek

\mathbb{N}	množina přirozených čísel
\mathbb{N}_0	množina přirozených čísel s 0
\mathbb{Z}	množina celých čísel
\mathbb{R}	množina reálných čísel
\mathbb{C}	množina komplexních čísel
a^\top	transpozice vektoru a
$EX_t = \mu_t$	střední hodnota náhodné veličiny (vektoru) X
$var X$	rozptyl náhodné veličiny X
\mathbf{I}	indikátor jevu
P	pravděpodobnostní míra
\mathcal{A}	σ -algebra
Ω	množina elementárních jevů
γ	autokovarianční funkce
ρ	autokorelační funkce
π	parciální autokorelační funkce
$X_{(k)}$	k -tá pořádková statistika
R_i	pořadí X_i v uspořádané verzi
$\Gamma^{(k)}$	matice velikosti k s $\gamma(i-j)$ na pozici (i, j)
t_0	bod useknutí autokorelační funkce
$WN(0, \sigma^2)$	bílý šum
$MA(q)$	Model klouzavých průměrů řádu q
$AR(p)$	Autoregresní model řádu p
$ARMA(p, q)$	Smíšený proces řádu p a q
$\lfloor x \rfloor$	Dolní celá část x
$\lceil x \rceil$	Horní celá část x
f'	Derivace funkce f
$\xrightarrow{P}_{n \rightarrow \infty}$	konvergence v pravděpodobnosti

A. Dodatek

A.1 Yuleovy-Walkerovy rovnice

Yuleovy-Walkerovy rovnice slouží k výpočtu autokovarianční a autokorelační funkce slabě stacionárních ARMA a AR řad. Čerpáme z (Prášková, 2004). Rovnice popíšeme pro ARMA(p, q) řadu splňující předpoklady věty 9 danou předpisem

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + Y_t + \theta_1 Y_{t-1} + \dots + \theta_q Y_{t-q}, \quad t \in \mathbb{Z}.$$

Ten lze přepsat do tvaru

$$X_t - \varphi_1 X_{t-1} - \dots - \varphi_p X_{t-p} = Y_t + \theta_1 Y_{t-1} + \dots + \theta_q Y_{t-q}, \quad t \in \mathbb{Z}.$$

Vynásobíme obě strany rovnice náhodnými veličinami X_t, \dots, X_{t-p} a aplikujeme střední hodnotu. Dostaneme první systém rovnic

$$\begin{aligned} EX_t^2 - \sum_{k=1}^p \varphi_k EX_{t-k} X_t &= EY_t X_t + \sum_{k=1}^q \varphi_k EY_{t-k} X_t, \\ EX_t X_{t-1} - \sum_{k=1}^p \varphi_k EX_{t-k} X_{t-1} &= EY_t X_{t-1} + \sum_{k=1}^q \varphi_k EY_{t-k} X_{t-1}, \\ &\vdots \\ EX_t X_{t-l} - \sum_{k=1}^p \varphi_k EX_{t-k} X_{t-l} &= 0, \quad l \geq \max(p, q + 1). \end{aligned}$$

Pravé strany systému se postupně vynulovávají, protože

$$\begin{aligned} EX_{t-l} Y_t &= 0, \quad l \geq 1, \\ EX_{t-l} Y_{t-1} &= 0, \quad l \geq 2, \\ &\vdots \\ EX_{t-l} Y_{t-q} &= 0, \quad l \geq q + 1. \end{aligned}$$

Přepíšeme do tvaru autokovariančních funkcí a získáme druhý systém rovnic

$$\begin{aligned} \gamma(0) - \sum_{k=1}^p \varphi_k \gamma(k) &= EY_t X_t + \sum_{k=1}^q \varphi_k EY_{t-k} X_t, \\ \gamma(1) - \sum_{k=1}^p \varphi_k \gamma(k-1) &= EY_t X_{t-1} + \sum_{k=1}^q \varphi_k EY_{t-k} X_{t-1}, \\ &\vdots \\ \gamma(l) - \sum_{k=1}^p \varphi_k \gamma(l-k) &= 0, \quad l \geq \max(p, q + 1). \end{aligned}$$

Pro získání hodnot na pravých stranách rovnic z druhého systému, je potřeba vynásobit obě strany rovnic z prvního systému náhodnými veličinami Y_t, \dots, Y_{t-q}

a aplikovat střední hodnotu. Dostaneme třetí systém rovnic

$$\begin{aligned}
EX_t Y_t - \sum_{k=1}^p \varphi_k EX_{t-k} Y_t &= EY_t^2 + \sum_{k=1}^q \varphi_k EY_{t-k} Y_t, \\
EX_t Y_{t-1} - \sum_{k=1}^p \varphi_k EX_{t-k} Y_{t-1} &= EY_t Y_{t-1} + \sum_{k=1}^q \varphi_k EY_{t-k} Y_{t-1}, \\
&\vdots \\
EX_t Y_{t-q} - \sum_{k=1}^p \varphi_k EX_{t-k} Y_{t-q} &= EY_t Y_{t-q} + \sum_{k=1}^q \varphi_k EY_{t-k} Y_{t-q}.
\end{aligned}$$

Ze třetího systému rovnic vyjádříme potřebné hodnoty pro $EX_t Y_{t-l}$, kde $l \in \mathbb{N}_0$, ty poté dosadíme do druhého systému rovnic, čímž získáme počáteční podmínky řešení soustavy lineárních rovnic. Tyto počáteční podmínky využijeme pro diferenční rovnici

$$\gamma(l) - \sum_{k=1}^p \varphi_k \gamma(l-k) = 0, \quad l \geq \max(p, q+1).$$

Způsob řešení lineárních diferenčních rovnic nalezneme například v literatuře (Prágerová, 1971).

A.2 Příklady měr robustnosti

Míra robustnosti metody je schopnost vypořádat se s odlehlými pozorováními. Podle toho jsme schopni porovnat různé metody mezi sebou a určit, která bude vhodnější.

A.2.1 Influenční funkce

Pracujeme se stejně rozdělenými a nezávislými náhodnými veličinami X_t , které pochází z rozdělení F s parametrem λ . Odhad $\hat{\lambda}_n(\mathbf{X})$ parametru λ je získaný z $\mathbf{X} = (X_1, \dots, X_n)$. Nyní není potřeba specifikovat o odhad čeho se přesně jedná, budeme se zabývat robustními vlastnostmi odhadu. $\hat{\lambda}_\infty(F)$ je asymptotická hodnota odhadu parametru pro rozdělení F , tedy

$$\hat{\lambda}_n(\mathbf{X}) \xrightarrow{P}_{n \rightarrow \infty} \hat{\lambda}_\infty(F).$$

Pokud je odhad konzistentní, pak $\hat{\lambda}_\infty(F) = \lambda$.

Pro lepší představu uveďme příklad pro odhad střední hodnoty, kdy $\hat{\lambda}_n(\mathbf{X}) = \frac{1}{n} \sum_{k=1}^n X_k$ a $\hat{\lambda}_\infty(F) = EX_t$. Může se však jednat o odhad směrodatné odchylky, měřítka či určitý M-odhad z předchozích částí.

Dále budeme pracovat s kontaminovaným rozdělením s distribuční funkcí

$$(1 - \varepsilon)F + \varepsilon\Delta_\delta,$$

kde Δ_δ je distribuční funkce Diracovy míry v bodě δ splňující $P(X_t = \delta) = 1$.

Definice 19 (Influenční funkce). *Nechť náhodná veličina X_t pochází z rozdělení F s parametrem λ , odhad $\hat{\lambda}_n(\mathbf{X})$ získaný z $\mathbf{X} = (X_1, \dots, X_n)$ je odhadem parametru λ a $\hat{\lambda}_\infty(F)$ je asymptotická hodnota odhadu parametru pro rozdělení F , tedy*

$$\hat{\lambda}_n(\mathbf{X}) \xrightarrow{P}_{n \rightarrow \infty} \hat{\lambda}_\infty(F).$$

Pak influenční funkci definujeme jako

$$IF_{\hat{\lambda}}(\delta, F) = \lim_{\varepsilon \rightarrow 0^+} \frac{\hat{\lambda}_\infty((1 - \varepsilon)F + \varepsilon\Delta_\delta) - \hat{\lambda}_\infty(F)}{\varepsilon},$$

kde Δ_δ je distribuční funkce Diracovy míry pozorování δ splňující $P(X_t = \delta) = 1$.

Pro malé ε je možné použít aproximaci

$$\hat{\lambda}_\infty((1 - \varepsilon)F + \varepsilon\Delta_\delta) \approx \hat{\lambda}_\infty(F) + \varepsilon IF_{\hat{\lambda}}(\delta, F),$$

a vychýlení

$$\hat{\lambda}_\infty((1 - \varepsilon)F + \varepsilon\Delta_\delta) - \hat{\lambda}_\infty(F),$$

lze aproximovat pomocí $\varepsilon IF_{\hat{\lambda}}(\delta, F)$.

Uvedeme příklad influenční funkce pro střední hodnotu μ a její odhad, výběrový průměr $\hat{\lambda}_n(\mathbf{X}) = \frac{1}{n} \sum_{k=1}^n X_k$

$$\begin{aligned} IF_{\hat{\mu}}(\delta, F) &= \lim_{\varepsilon \rightarrow 0^+} \frac{\hat{\lambda}_\infty((1 - \varepsilon)F + \varepsilon\Delta_\delta) - \hat{\lambda}_\infty(F)}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0^+} \frac{(1 - \varepsilon)\mu + \varepsilon\delta - \mu}{\varepsilon} \\ &= \delta - \mu. \end{aligned}$$

Zavedení influenční funkce pro časové řady zde nebudeme uvádět a nalezneme jej například v (Martin a Yohai, 1986).

A.2.2 Bod selhání

Bodem selhání (anglicky *breakdown point*) ε^* odhadu $\hat{\lambda}$ parametru λ je největší možné proporční zastoupení odlehlých pozorování v analyzovaném výběru tak, aby odhad $\hat{\lambda}$ stále poskytoval informace o λ , tedy o základním rozdělení dat.

U smysluplných dat očekáváme, že odlehlá pozorování budou tvořit méně než polovinu celkového počtu pozorování, tedy $\varepsilon^* \leq 0.5$. Příkladem může být aritmetický průměr jako odhad parametru střední hodnoty s bodem selhání 0 nebo výběrový medián jako odhad mediánu s bodem selhání 0.5.

A.2.3 Asymptotické vychýlení

Hodnota ε^* se zabývá tím, jak extrémně kontaminovaná data mohou být. Nás ale zajímá chování odhadu i pro méně extrémní situaci $\varepsilon \leq \varepsilon^*$. Definujeme si rodinu rozdělení

$$\mathcal{F}_{\varepsilon, \lambda} = \{(1 - \varepsilon)F + \varepsilon G : G \in \mathcal{G}\},$$

kde \mathcal{G} udává rodinu všech rozdělení a F je rozdělení s parametrem λ .

Asymptotické vychýlení odhadu $\hat{\lambda}$ pro parametr λ rozdělení $F \in \mathcal{F}_{\varepsilon, \lambda}$ je

$$\text{bias}(\hat{\lambda}, \lambda, F) = \hat{\lambda}_{\infty}(F) - \lambda$$

a *maximální vychýlení* při ε

$$\text{maxbias}(\hat{\lambda}, \lambda, \varepsilon) = \sup\{|\text{bias}(\hat{\lambda}, \lambda, F)| : F \in \mathcal{F}_{\varepsilon, \lambda}\}.$$

V případě, kdy $\Theta = \mathbb{R}$, platí vztah

$$\varepsilon^*(\hat{\lambda}, F) = \sup\{\varepsilon \geq 0 : \text{maxbias}(\hat{\lambda}, \lambda, \varepsilon) < \infty\}.$$

Poznamenejme, že s vychýlením souvisí i pojmy jako *sensitivita kontaminace* (anglicky *contamination sensitivity*) či *sensitivita hrubých chyb* (anglicky *gross-error sensitivity*) odhadu $\hat{\lambda}$ parametru λ , které zde nebudeme uvádět. Jejich definici a vlastnosti nalezneme například v (Maronna a kol., 2006) (kapitola 3.3).

A.2.4 Eficience

Budeme předpokládat, že pracujeme s rozdělením F_{λ} , konzistentním odhadem $\hat{\lambda}_n$ parametru λ a platí, že $\sqrt{n}(\hat{\lambda}_n - \lambda)$ má asymptotické rozdělení $N(0, \sigma^2)$. V takovém případě je odhad asymptoticky nevychýlený. Pak nás zajímá pouze asymptotický rozptyl $\text{asvar}(\hat{\lambda})$ odhadu $\hat{\lambda}_n$. Preferujeme odhad s menším rozptylem. Necht $\hat{\lambda}_n^{\min}$ je odhad s minimálním možným asymptotickým rozptylem $\text{asvar}(\hat{\lambda}^{\min})$, který má výše zmíněné asymptotické rozdělení. Při splnění podmínek regularity dle (Anděl, 2005, definice 7.8) je tímto odhadem odhad získaný pomocí metody maximální věrohodnosti.

Asymptotická eficience odhadu $\hat{\lambda}_n$ parametru λ je dána předpisem

$$ae(\hat{\lambda}_n, \lambda) = \frac{\text{asvar}(\hat{\lambda}^{\min})}{\text{asvar}(\hat{\lambda})}.$$

Definujeme si *asymptoticky eficientní odhad*.

Definice 20 (Asymptoticky eficientní odhad). *Necht $\hat{\lambda}_n$ je konzistentní odhad parametru λ rozdělení F , $\sqrt{n}(\hat{\lambda}_n - \lambda)$ má asymptotické rozdělení $N(0, \sigma^2)$. Odhad $\hat{\lambda}_n$ nazveme asymptoticky eficientní pokud*

$$ae(\hat{\lambda}_n, \lambda) = 1.$$

Poznamenejme, že pojem eficience souvisí s *Raovou-Cramérovou mezí* a *Fisherovou mírou informace*. Tyto pojmy zde ale nebudeme rozvádět a definovat. Podrobnosti nalezneme například v (Anděl, 2005).

A.2.5 Posouzení odlehlosti

Podíváme na způsob, jak rozpoznat, zda dané pozorování je potenciálním odlehlým pozorováním.

Pravidlo 6-sigma

Jako základní míra pro určování odlehlosti slouží takzvané pravidlo 6-*sigma*. Pro každé $t = 1, \dots, n$ definujeme

$$Z_t = \frac{X_t - \hat{\mu}}{\hat{\sigma}},$$

kde $\hat{\mu}$ je odhad střední hodnoty a $\hat{\sigma}$ je odhad měřítka. Zde se jedná o klasické odhady. Pozorování s hodnotou $|Z_t| > 3$ lze považovat za potenciální odlehlá pozorování. Toto pravidlo vychází z předpokladu, že Z_t pochází z $N(0,1)$. Pak

$$P(|Z_t| > 3) = 1 - \Phi(3) + \Phi(-3) = 2(1 - \Phi(3)) \doteq 0,$$

kde $\Phi(z)$ je distribuční funkce $N(0,1)$.

Tato statistika nefunguje pro malé výběry a robustní verze nalezneme například v (Maronna a kol., 2006).

Cookova vzdálenost

Podíváme se na vliv k -tého pozorování (\mathbf{x}_k, y_k) na odhad regresní přímky. Porovnáme LS-odhad vektoru parametrů spočítaného z celého výběru a z výběru bez k -tého pozorování. LS-odhad $\hat{\beta}^{(LS)}$ je počítán z celého výběru. LS-odhad $\hat{\beta}_{(k)}^{(LS)}$ je počítán z dat bez k -tého pozorování a vektor vyrovnaných hodnot je $\hat{\mathbf{Y}}_{(k)} = \mathbf{X}\hat{\beta}_{(k)}$. Cookovu vzdálenost pro k -té pozorování spočítáme předpisem

$$D_k = \frac{1}{\text{rank}(\mathbf{X})\hat{\sigma}^2} \|\hat{\mathbf{Y}}_{(k)} - \hat{\mathbf{Y}}\|^2,$$

kde pracujeme s klasickou Euklidovskou normou vektoru a $\hat{\sigma}^2$ je odhad rozptylu pomocí reziduí

$$\hat{\sigma}^2 = \frac{1}{n - \text{rank}(\mathbf{X})} \sum_{k=1}^n (Y_k - \hat{Y}_k)^2.$$

Již nepožadujeme plnou hodnost matice \mathbf{X} .

Cookovu vzdálenost lze počítat i pomocí matice ortogonální projekce a poté pomocí Studentova rozdělení či $Q-Q$ grafu lze sestavit test pro určení odlehlosti. Tento postup nalezneme například v (Maronna a kol., 2006).

A.3 Koeficienty korelace

Pracujeme se dvěma náhodnými výběry X_1, \dots, X_n a Y_1, \dots, Y_n . Jejich vzájemný vztah často popisujeme pomocí korelačního koeficientu jehož odhad je

$$\hat{\rho} = \frac{\sum_{k=1}^n (X_k - \bar{X}_n)(Y_k - \bar{Y}_n)}{\left(\sum_{k=1}^n (X_k - \bar{X}_n)^2 \sum_{k=1}^n (Y_k - \bar{Y}_n)^2\right)^{\frac{1}{2}}}. \quad (\text{A.1})$$

Vztah dvou veličin lze popsat nejen pomocí korelačního koeficientu, ale existují i jiné přístupy: Kendallův a Spearmanův koeficient. Pro určitá rozdělení existuje

vztah mezi těmito koeficienty a korelačním koeficientem. Odhady těchto koeficientů poté použijeme pro odhad autokorelační funkce. Čerpáme z (Shevlyakov a Oja, 2016).

Kendallův koeficient τ má odhad

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}((X_i - X_j)(Y_i - Y_j)).$$

Spearmanův koeficient Sp má odhad

$$\widehat{Sp} = \frac{\sum_{k=1}^n (R_k^X - \bar{R}_k^X)(R_k^Y - \bar{R}_k^Y)}{\left(\sum_{k=1}^n (R_k^X - \bar{R}_k^X)^2 \sum_{k=1}^n (R_k^Y - \bar{R}_k^Y)^2 \right)^{\frac{1}{2}}},$$

kde R_k^X je pořadí X_k v uspořádané verzi

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n-1)} \leq X_{(n)},$$

obdobně pro R_k^Y a

$$\begin{aligned} \bar{R}_k^X &= \frac{1}{n} \sum_{k=1}^n R_k^X = \frac{n+1}{2}, \\ \bar{R}_k^Y &= \frac{1}{n} \sum_{k=1}^n R_k^Y = \frac{n+1}{2}. \end{aligned}$$

Dle (Mottonen a kol., 1999) a (Croux a Dehon, 2010) je třeba odhad Kendallova a Spearmanova koeficientu transformovat pro získání odhadu korelačního koeficientu. Pro data pocházející z normálního rozdělení (obecně následující vztahy platí pro eliptické symetrické rozdělení) dostaneme následující odhad autokorelační funkce za použití Kendallova korelačního koeficientu

$$\hat{\rho}^{(\tau)} = \sin\left(\frac{\pi \hat{\tau}}{2}\right),$$

a za použití Spearmanova koeficientu

$$\hat{\rho}^{(Sp)} = 2 \sin\left(\frac{\pi \widehat{Sp}}{6}\right).$$