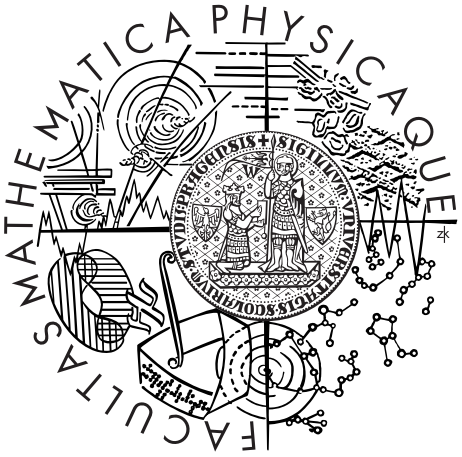
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Diplomová Práce



Jakub Petrásek

Metody bootstrap pro závislá pozorování

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Doc. RNDr. Zuzana Prášková, CSc.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost a matematická statistika
Studijní plán:	Ekonometrie

2008

Poděkování

Na tomto místě bych rád poděkoval vedoucí práce, Doc. RNDr. Zuzaně Práškové, CSc., za zajímavé a aktuální téma, cenné rady a podněty, čas, který strávila při četbě této práce, trpělivost při četných konzultacích a za zapůjčení potřebné literatury. Také děkuji Mgr. Zdeňku Hlávkovi, PhD. za pomoc s programem R.

Prohlášení:

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 18. dubna 2008 Jakub Petrásek

Obsah

Značení

$\mathcal{X} = \{X_1, ..., X_n\}$	výběr náhodných veličin
$\mathcal{X}^* = \{X_1^*, ..., X_n^*\}$	bootstrapový výběr náhodných veličin
θ_0	správná hodnota neznámého parametru θ
$T_n = T_n(X_1, ..., X_n)$	odhad neznámého parametru θ na základě výběru $\mathcal{X}$
$T_n^* = T_n(X_1^*, ..., X_n^*)$	odhad neznámého parametru θ na základě bootstrapového výběru $\mathcal{X}^*$ (hvězdička bude vždy značit bootstrapovou verzi)

P	sdružená pravděpodobnostní míra výběru $\mathcal{X}$
$R_n = R_n(X_1, ..., X_n; \mathrm{P})$	standardizovaný odhad parametru $\theta(\mathrm{P})$
$\widetilde{R}_n = \widetilde{R}_n(X_1, ..., X_n; \mathrm{P})$	studentizovaný odhad parametru $\theta(\mathrm{P})$
G_n	distribuční funkce náhodné veličiny T_n
H_n	distribuční funkce standardizované náhodné veličiny T_n
$\widetilde{H}_n$	distribuční funkce studentizované náhodné veličiny T_n
$H_n^{-1}(\alpha)$	α kvantil náhodné veličiny s rozdělením daným distribuční funkcí H_n

Φ	distribuční funkce standardizovaného normálního rozdělení
ϕ	hustota standardizovaného normálního rozdělení
$u(\alpha)$	kritická hodnota standardizovaného normálního rozdělení na hladině α

$\boldsymbol{A}_{m\times n}$	matice $m \times n$
$\boldsymbol{I}_n$	n rozměrná jednotková matice
$\boldsymbol{A}^T$	transponovaná matice
$\boldsymbol{A}^\dagger$	hermitovsky sdružená (transponovaná a komplexně sdružená) matice

$\xrightarrow{P}$	konvergence v pravděpodobnosti
$\xrightarrow{D}$	konvergence v distribuci
$\xrightarrow{s.j}$	konvergence skoro jistě

Pro reálnou posloupnost $\{x_n, n \in \mathbb{Z}\}$ značíme

$x_n = o(n^k)$	je-li $x_n/n^k \to 0,\ n \to \infty$,
$x_n = O(n^k)$	je-li $\limsup_{n\to\infty} \lvert x_n\rvert/n^k < \infty$.

Pokud je $\{X_n, n \in \mathbb{Z}\}$ posloupnost náhodných veličin pak

$X_n = o_P(n^k)$	je-li $X_n/n^k \xrightarrow{P} 0,\ n \to \infty$,
$X_n = O_P(n^k)$	je-li $(\text{pro}\,\forall \epsilon > 0)\ (\exists M \in (0, \infty))$ takové, že $\sup_{n\geq 1} \mathrm{P}\left(\lvert X_n\rvert/n^k > M\right) < \epsilon$.

Název práce: Metody bootstrap pro závislá pozorování
Autor: Jakub Petrásek
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Doc. RNDr. Zuzana Prášková, CSc.
E-mail vedoucího: praskova@karlin.mff.cuni.cz

Abstrakt: Předložená práce se zabývá principy, asymptotickými vlastnostmi a vzájemným srovnáním metod bootstrap pro závislá pozorování. V první části je čtenář seznámen se základními myšlenkami a výhodami metody bootstrap pro nezávislá data, aby vzápětí tyto znalosti mohl uplatnit při aplikaci na data závislá. V práci jsou představeny metody blokový, frekvenční a sítový bootstrap. Princip každé je popsán v širších souvislostech, jsou vždy uvedeny asymptotické vlastnosti a některé jsou odvozeny. V případě metody blokový bootstrap je ukázána silná závislost na vhodné volbě délky bloku, proto jsou do práce zahrnuty také dva algoritmy sloužící k jejímu empirickému odhadu. Hlavním cílem této práce je porovnat jednotlivé metody z teoretického hlediska a také pomocí simulační studie. V poslední části je prezentováno několik příkladů, jak aplikovat vyložené metody na reálná data. Diskutované postupy jsou implementovány v jazyce R a jazyce Fortran.
Klíčová slova: *metody bootstrap, závislá data, Fourierova transformace, Edgeworthův rozvoj*

Title: Bootstrap Methods for Dependent Observations
Author: Jakub Petrásek
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. RNDr. Zuzana Prášková, CSc.
Supervisor's e-mail address: praskova@karlin.mff.cuni.cz

Abstract: This Diploma thesis deals with principles, asymptotic properties and comparison of bootstrap methods for dependent observations. In the first chapter principal ideas and benefits of bootstrap method for independent data are introduced. Subsequently, these knowledge are applied to data exhibiting dependency. Block, frequency and sieve bootstrap methods are presented. Afterwards, principle of each method is described in broader context, asymptotic properties are presented and some of them are derived. Strong dependency of block bootstrap method on block length is discussed and algorithms for empirical choice of optimal block length are described. The main aim of this work is to compare discussed methods from theoretical point of view and via simulation study. Eventually, a few examples, which are based on real data sets, are presented. Discussed principles are implemented in software R and software Fortran.
Keywords: *bootstrap methods, dependent data, Fourier transform, Edgeworth expansion*

Úvod

Výraz *bootstrap* v doslovném překladu znamená *poutko u bot*. Název pochází z jednoho z příběhů "Dobrodružství barona Münchhausena", kde se baron pomalu utápěl v bahně a zachránil se zatažením za poutka u svých bot, což by žádný tonoucí nikdy neučinil ke své záchraně. Stejně tak metoda bootstrap se může jevit jako zdánlivě nedůvěryhodná.

Základní principy metody bootstrap byly publikovány v článku *Bootstrap Methods: Another look at the jackknife* roku 1979 *Bradem Efronem*. Tento článek vzbudil velký ohlas. Metoda dokázala, že svou přesností předčí klasickou aproximaci normálním rozdělením. Poskytla odpovědi na chování statistik, pro které asymptotické rozdělení nebylo známé. Mnoho statistiků se proto zaměřilo na výzkum vlastností metody bootstrap a na její rozšíření na případ závislých dat.

Metoda bootstrap patří mezi takzvané počítačově intenzivní metody pro statistickou analýzu dat, neboli úspěšná aplikace obnáší nutnost velkého množství výpočtů. S rozvojem výpočetní techniky je její užití stále snažší a s tím roste četnost využívání metody bootstrap v praxi.

Šíře aplikovatelnosti metody bootstrap se stále rozrůstá. V posledních letech se výzkum soustředí na způsoby, jak aplikovat metodu na závislá data. Již bylo navrženo několik možných přístupů. Jednotlivé přístupy jsou srovnávány pomocí metod teorie pravděpodobnosti a na simulovaných datech. Avšak v žádném případě nelze zvolit jeden obecně optimální přístup.

Hlavním cílem práce je seznámit čtenáře s obecnými principy metod bootstrap pro závislá data. Vysvětlit, za jakých okolností lze tyto metody aplikovat, a nakonec jednotlivé metody mezi sebou srovnat z teoretického hlediska a také na základě simulační studie.

Práce je rozdělena do 6 kapitol. V 1. kapitole se čtenář seznámí se základním principem metody bootstrap. Nahlédne do teoretického pozadí metody a naučí se, jak ji používat v praxi.

Ve 2. kapitole jsou popsány metody pro závislá data a nastíněny principy použití.

V dalších kapitolách jsou pak podrobně studovány jednotlivé metody. Text je prokládán relevantními příklady, případně simulacemi.

Nakonec v 5. kapitole jsou metody aplikovány na reálná data a pro srovnání je provedeno několik větších simulačních experimentů. Některé metody doposud nebyly vzájemně srovnány.

Vyložené postupy byly implementovány v jazyce R s podpůrnými procedurami naprogramovanými v jazyce Fortran a spolu s popisy shromážděny ve vytvořené knihovně. Knihovnu a také reálná data použitá v poslední kapitole lze nalézt na přiloženém CD.

1 Princip a základní teoretické výsledky pro nezávislý bootstrap

V této kapitole si vysvětlíme základní principy metody bootstrap. Budeme se zabývat její aplikací na nezávislé stejně rozdělené náhodné veličiny. Tato kapitola není pouze ilustrativní, neboť vše, co si vysvětlíme, využijeme později.

1.1 Základní princip metody bootstrap

Uvažujme náhodný výběr $\mathcal{X} = \{X_1, ..., X_n\}$ daný pravděpodobnostní mírou P. Jelikož jde o nezávislé stejně rozdělené náhodné veličiny, víme, že lze sdruženou míru faktorizovat, tj. lze psát $\mathrm{P} = \mathrm{P}_{X_1} \otimes \cdots \otimes \mathrm{P}_{X_n}$, kde P_{X_i} je pravděpodobnostní míra určující rozdělení veličiny X_i a $\otimes$ značí součin měr. Namísto P_{X_i} pracujme s distribuční funkcí náhodné veličiny X_i, označme ji F. Zajímejme se o parametr θ, který je určen rozdělením F, můžeme proto psát $\theta = \theta(F)$. Pokud bychom znali rozdělení F, mohli bychom správnou hodnotu parametru získat přímým výpočtem, neboť

$$\theta_0 = \int s(x) dF(x),$$

pro nějakou funkci $s(\cdot)$. Rozdělení F však neznáme, proto na základě náhodného výběru $\mathcal{X}$ hledáme veličinu $T_n = T_n(X_1, ..., X_n)$ pro odhad parametru θ, po které požadujeme splnění jistých vlastností. Určením bodového odhadu T_n z pozorovaných hodnot $x_1, ..., x_n$ však statistická práce zřídka kdy končí. Často se dále ptáme:

- Jaké má odhad T_n vychýlení?
- Jaký je rozptyl odhadu T_n parametru θ?
- Jaký je interval spolehlivosti? Jaké jsou testové statistiky a kritické hodnoty pro testování hypotéz?

Právě k zodpovězení těchto otázek lze užít metodu bootstrap. Metoda bootstrap tedy slouží k odhadu statististických vlastností veličiny T_n nikoliv však k získání veličiny T_n samotné.

Základní myšlenka neparametrické metody bootstrap pro nezávislá stejně rozdělená data spočívá v nahrazení původní distribuce F empirickou distribuční funkcí F_n, definovanou na základě náhodného výběru vztahem

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i \leq x]}. \tag{1.1}$$

Poznámka 1. Víme-li, že rozdělení výběru $\mathcal{X}$ je dáno distribuční funkcí $F(\boldsymbol{\beta})$, kde $\boldsymbol{\beta}$ jsou neznámé parametry, lze s výhodou použít metodu parametrický bootstrap, kdy odhadneme parametry $\boldsymbol{\beta}$ a dále pracujeme s distribucí $F(\widehat{\boldsymbol{\beta}})$. Síla metody bootstrap spočívá v tom, že jsme schopni odhadovat charakteristiky náhodné veličiny T_n, i když máme o distribuci F minimum informací. Neparametrická metoda bootstrap totiž odhad charakteristik odhadu řeší bez jakékoliv znalosti rozdělení původního výběru.

Použijeme znalost empirické distribuční funkce. Dosazením vidíme, že

$$\theta(F_n) = \int s(x) dF_n(x).$$

Nahradili jsme populační distribuční funkci distribuční funkcí empirickou. Tento postup se nazývá *plug-in* princip. Hodnotu $\theta(F_n)$ nazveme bootstrapovou verzí parametru $\theta(F)$. Poznamenejme, že jde o náhodnou veličinu, neboť funkce F_n je podmíněná náhodným výběrem $\mathcal{X}$.

Vlastnosti empirické distribuční funkce

Mezi empirickou distribuční funkcí a náhodným výběrem existuje vzájemně jednoznačný vztah, to znamená, že přechodem k distribuční funkci F_n neztrácíme o původním vzorku žádnou informaci.

Víme, že obecným principem metody bootstrap je aproximace populační distribuční funkce F empirickou verzí F_n. O jak kvalitní aproximaci se však jedná?

Zvolíme-li pevné $x \in \mathbb{R}$, pak náhodné veličiny $I_{[X_i \leq x]}$ vyskytující se ve vzorci (1.1) jsou nezávislé stejně rozdělené náhodné veličiny s alternativním rozdělením s parametrem $F(x)$. Proto má distribuční funkce $nF_n(x)$ jako jejich součet binomické rozdělení $Bi(n, F(x))$. Neboli

$$\mathrm{E}\, F_n(x) = F(x), \tag{1.2}$$

$$\mathrm{var}\, F_n(x) = \frac{1}{n} F(x)(1 - F(x)) \to 0, \quad n \to \infty. \tag{1.3}$$

Vztahy (1.2) a (1.3) znamenají, že $F_n(x)$ je konzistentní odhad distribuční funkce F, neboť z Čebyševovy nerovnosti plyne

$$\mathrm{P}\left(|F_n(x) - F(x)| > \epsilon\right) \leq \frac{F(x)(1 - F(x))}{n\epsilon^2} \to 0, \quad n \to \infty.$$

Aplikací silného zákona velkých čísel se dokáže, že dokonce

$$F_n(x) \overset{s.j.}{\to} F(x), \quad n \to \infty. \tag{1.4}$$

Dosavadní úvahy byly provedeny pro pevné x. Dá se však ukázat, že uvedená vlastnost (1.4) platí stejnoměrně vzhledem k x.

Věta 1. *(Glivenkova)*

$$P\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0\right) = 1.$$

Důkaz: Věta s citovaným důkazem je uvedena v knize Anděl (2005), tvrzení 11.10. □

Konstrukce bootstrapových odhadů

Naším cílem je odhadnout charakteristiky rozdělení náhodné veličiny T_n, které jsou dány distribucí F. Jde zejména o vychýlení $B_n = \mathrm{E}\, T_n - \theta_0$ a rozptyl $\mathrm{var}\, T_n$. Pro vyhodnocení statistických závěrů potřebujeme znát rozdělení standardizované statistiky $R_n(X_1, ..., X_n; F)$. Definujme

$$H_n(x) = \mathrm{P}\left(R_n(X_1, ..., X_n; F) \leq x\right).$$

Všechny tyto charakteristiky lze odhadnout aplikací metody bootstrap. Opět nahradíme neznámou distribuční funkci F známou empirickou distribuční funkcí F_n. To znamená odhadovat hledané charakteristiky na bootstrapovém výběru $\mathcal{X}^* = \{X_1^*, ..., X_n^*\}$ se známým rozdělením určeným distribuční funkcí F_n. Protože pracujeme s empirickou distribuční funkcí, platí, že

$$\mathrm{P}^*(X_i^* = X_i) = \frac{1}{n}, \qquad i = 1, ..., n$$

a pro náhodné veličiny $X_1^*, ..., X_n^*$

$$\mathrm{E}^* X_1^* = \sum_{i=1}^{n} X_i \mathrm{P}^*(X_1^* = X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n, \tag{1.5}$$

$$\mathrm{var}^* X_1^* = \sum_{i=1}^{n} (X_i - \mathrm{E}^* X_i^*)^2 \mathrm{P}^*(X_1^* = X_i) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{1.6}$$

$\mathrm{E}^*(\cdot)$ respektive $\mathrm{var}^*(\cdot)$ značí, že jde o podmíněnou střední hodnotu respektive podmíněný rozptyl, tzn. že platí

$$\mathrm{E}^*(\cdot) = \mathrm{E}\,(\cdot|\mathcal{X}), \quad \mathrm{var}^*(\cdot) = \mathrm{var}\,(\cdot|\mathcal{X}).$$

Definujme bootstrapový odhad $T_n^* = T_n(X_1^*, ..., X_n^*; F_n)$ parametru $\theta(F_n)$, bootstrapovou verzi standardizované statistiky označme jako $R_n^* = R_n(X_1^*, ..., X_n^*; F_n)$ a její distribuční funkci H_n^*.

Příklad 1. Na základě pozorování $\mathcal{X} = \{X_1, ..., X_n\}$ pocházejících z rozdělení s distribuční funkcí F s parametry (μ, σ^2) chceme odhadnout parametr μ. Vhodným odhadem pro střední hodnotu je výběrový průměr $\overline{X}_n$. Standardizovanou statistikou je výraz $R_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$, kde $\mu = \mathrm{E}\, X_1$.

Na základě původního výběru $\mathcal{X}$ zkonstruujeme empirickou distribuční funkci F_n. Přechodem do bootstrapového světa se původní výběr $\mathcal{X}$ stává populací pro konstrukci bootstrapových výběrů $\mathcal{X}^*$. Protože již známe vzorce pro první dva bootstrapové momenty ((1.5) a (1.6)), platí

$$\mathrm{E}^* X_1^* = \overline{X}_n, \qquad \mathrm{var}^* X_1^* = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

a standardizovaná bootstrapová statistika má tedy tvar

$$R_n^* = \sqrt{n}(\overline{X}_n^* - \mathrm{E}^* X_1^*)/\mathrm{var}^* X_1^* = \sqrt{n}(\overline{X}_n^* - \overline{X}_n)/s_n,$$

kde jsme označili $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

Je-li například $s(x) = x$ (odhadujeme střední hodnotu), lze rozptyl či jiné charakteristiky vypočítat přímo z definice explicitním vzorcem. V opačném případě by stanovení přesné hodnoty charakteristiky vyžadovalo provedení n^n výběrů a pro každý je třeba určit $T_n^* = T_n(X_1^*, ..., X_n^*|\mathcal{X})$. Jelikož však odhad T_n^* nezávisí na pořadí výběru veličin $X_1^*, ..., X_n^*$, lze bootstrapové výběry, které se shodují až na pořadí, považovat za ekvivalentní. Takto lze vytvořit $\binom{2n-1}{n}$ výběrů (viz věta 4 v kapitole 2). Takový počet operací není většinou proveditelný, proto se postupuje metodou *Monte-Carlo*.

Metoda Monte-Carlo

Principem metody Monte Carlo je mnohokrát (B-krát) vygenerovat bootstrapový výběr $\mathcal{X}^* = \{X_1^*, ..., X_n^*\}$ z původní populace $\mathcal{X} = \{X_1, ..., X_n\}$. Aplikací metody Monte Carlo získáme hodnoty $T_{n,1}^*, ..., T_{n,B}^*$ a $R_{n,1}^*, ..., R_{n,B}^*$. Odhad vychýlení je pak dán vztahem

$$\widehat{B_n^*} = \frac{1}{B}\sum_{i=1}^{B} T_{n,i}^* - \theta(F_n). \tag{1.7}$$

Zajímá-li nás rozptyl, užijeme vztahu

$$\widehat{var^*} T_n^* = \frac{1}{B}\sum_{i=1}^{B}\left(T_{n,i}^* - \frac{1}{B}\sum_{j=1}^{B} T_{n,j}^*\right)^2. \tag{1.8}$$

Chceme-li odhadnout distribuční funkci H_n, můžeme využít verzi centrální limitní věty pro nezávislé stejně rozdělené náhodné veličiny. Avšak i v tomto případě lze využít metodu bootstrap. Bootstrapovým odhadem distribuční funkce H_n nazveme funkci

$$\widehat{H_n^*}(x) = \frac{1}{B}\sum_{i=1}^{B} \mathrm{I}_{[R_{n,i}^* \le x]}. \tag{1.9}$$

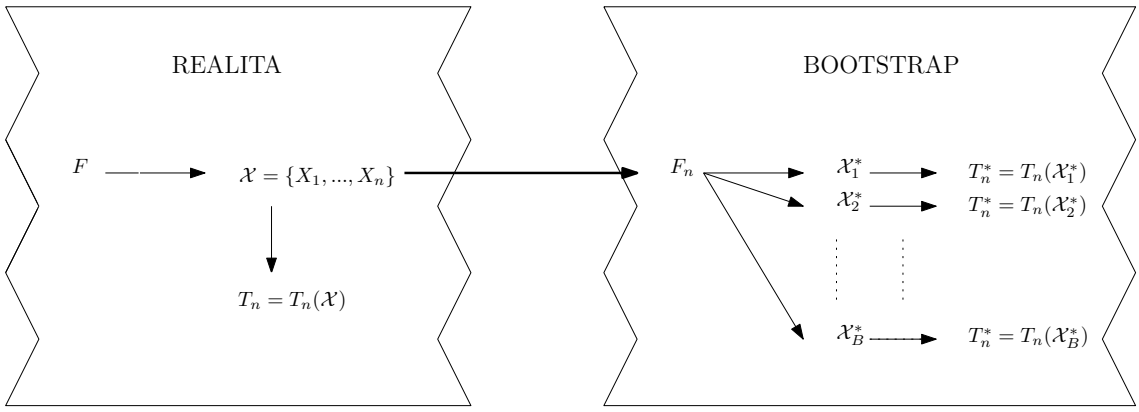
Posledním krokem je použít bootstrapové odhady charakteristik resp. rozdělení jako odhadu těchto na původním výběru $\mathcal{X}$. Z čehož vyplývá zásadní předpoklad při aplikaci metody bootstrap, totiž, že vztah výběru bootstrapového ke skutečnému je analogický vztahu skutečného výběru k populaci.

Ukázali jsme, že metoda bootstrap se používá k odhadu charakteristik a rozdělení původních odhadů. V následující části si odvodíme přesnost aproximace bootstrapovou distribuční funkcí. Ukážeme si, že bootstrapový odhad distribuční funkce ve většině případů předčí svou přesností klasickou normální aproximaci, což je jedna z jeho hlavních předností.

1.2 Asymptotické vlastnosti

V celé sekci budeme používat následující distribuční funkce a jejich bootstrapové protějšky:

$$\begin{aligned}
G_n(x) &= \mathrm{P}(T_n \le x), & G_n^*(x) &= \mathrm{P}^*(T_n^* \le x),\\
H_n(x) &= \mathrm{P}\left(\sqrt{n}\tfrac{T_n-\theta}{\sigma} \le x\right), & H_n^*(x) &= \mathrm{P}^*\left(\sqrt{n}\tfrac{T_n^*-T_n}{\sigma^*} \le x\right),\\
\widetilde{H}_n(x) &= \mathrm{P}\left(\sqrt{n}\tfrac{T_n-\theta}{\widehat{\sigma}} \le x\right), & \widetilde{H}_n^*(x) &= \mathrm{P}^*\left(\sqrt{n}\tfrac{T_n^*-T_n}{\widehat{\sigma}^*} \le x\right).
\end{aligned}$$



Obrázek 1.1: Schéma principu metody bootstrap. V reálném světě máme k dispozici pouze jedinou hodnotu odhadu T_n, ale v bootstrapovém protějšku můžeme na základě znalosti jediného výběru $X_1, ..., X_n$ vygenerovat tolik bootstrapových odhadů T_n^*, kolika jsme schopni. To nám umožňuje odhadovat charakteristiky rozdělení bootstrapového odhadu T_n^*, tyto lze pak dále považovat za odhady charakteristik rozdělení původního odhadu T_n (převzato z Efron a Tibshirani (1993)).

Berry-Essenova nerovnost

Jelikož při aplikaci metody bootstrap aproximujeme rozdělení odhadů bootstrapovými verzemi distribucí, žádáme, aby tato byla alespoň konzistentním odhadem a pak se dále ptáme, o jak přesný odhad se jedná. Nechť $T_n = \overline{X}_n$ a zkoumejme, zda je například distribuční funkce H_n^* konzistentním odhadem funkce H_n. Rychlost konvergence standardizovaných statistik k normálnímu rozdělení popisuje následující věta.

Věta 2. *(Berry-Essenova nerovnost) Buď $n \in \mathbb{N}$ a nechť jsou dány centrované nezávislé náhodné veličiny $X_1, \ldots, X_n$ splňující*

$$0 < S_n^2 = \sum_{i=1}^{n} var\, X_i < \infty.$$

Potom

$$\sup_{x \in \mathbb{R}} \left| P\left(\frac{1}{S_n} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq \frac{C}{S_n^3} \sum_{i=1}^{n} E\, |X_i|^3 .$$

Ukázalo se, že nerovnost platí s $C = 0.7995$.

Důkaz: Věta s citovaným důkazem je uvedena v knize Lachout (2004), věta 18.2. □

Označme $\mu = \mathrm{E}\, X_1$ a $\sigma^2 = \mathrm{var}\, X_1$. Dle věty platí

$$\sup_{x \in \mathbb{R}} |H_n(x) - \Phi(x)| \leq \frac{C}{(\sqrt{n}\sigma)^3} \sum_{i=1}^{n} \mathrm{E}\, |X_i - \mu|^3 = \frac{C}{\sqrt{n}\sigma^3} \mathrm{E}\, |X_1 - \mu|^3 .$$

Je-li $\mathrm{E}\, |X_1 - \mu|^3 < \infty$, pak $H_n(x) - \Phi(x) = O(n^{-1/2})$ pro $\forall x \in \mathbb{R}$. Analogicky lze větu aplikovat na bootstrapový výběr, neboť se opět jedná o nezávislé stejně rozdělené náhodné

veličiny, avšak s odlišným rozdělením a vzhledem k původnímu výběru $\mathcal{X}$. Nakonec lze psát

$$\begin{aligned}\sup_{x\in\mathbb{R}}|H_n(x)-H_n^*(x)| &= \sup_{x\in\mathbb{R}}|(H_n(x)-\Phi(x))-(H_n^*(x)-\Phi(x))| \\ &\leq \frac{C}{\sqrt{n}\sigma^3}\mathrm{E}\,|X_1-\mu|^3+\frac{C}{\sqrt{n}\sigma^{*3}}\mathrm{E}\,\left|X_1^*-\overline{X}_n\right|^3,\end{aligned}$$

což je odhad stejného řádu jako při aproximaci normálním rozdělením. Řád konvergence lze přesněji popsat pomocí takzvaného *Edgeworthova rozvoje*.

Edgeworthův rozvoj

V minulé sekci jsme se zmínili o vyšší přesnosti bootstrapové aproximace rozdělení v porovnání s tradiční aproximací normálním rozdělením. Abychom mohli porozumět tomuto jevu z teoretické roviny, je nutné si vysvětlit základní principy *Edgeworthova rozvoje*. Vyslovíme věty týkající se statistik typu hladké funkce průměru. Do této skupiny patří kromě výběrového průměru také například výběrový rozptyl a zejména mnoho odhadů důležitých při práci se závislými daty, jakým je kupříkladu korelace.

Pracujme s odhadem $T_n = g(\overline{\boldsymbol{Z}}_n)$ parametru $\theta = g(\boldsymbol{\mu})$, kde $g:\mathbb{R}^d\to\mathbb{R}$. Nechť σ značí směrodatnou odchylku odhadu $\sqrt{n}T_n$. Zabývejme se asymptotickými vlastnostmi standardizované statistiky odhadu T_n.

Věta 3. *Nechť $(\boldsymbol{Z}_1,...,\boldsymbol{Z}_n)$ je náhodný výběr z absolutně spojitého d-rozměrného rozdělení s distribuční funkcí F a charakteristickou funkcí χ, který splňuje*

$$E\,\|\boldsymbol{Z}_1\|^4<\infty \quad a \quad \limsup_{\|\boldsymbol{t}\|\to\infty}|\chi(\boldsymbol{t})|<1.$$

Nechť dále funkce $g(\overline{\boldsymbol{Z}}_n)\in\mathcal{C}^4$ v okolí $E\,\boldsymbol{Z}_1$, pak platí

$$P\left(\sqrt{n}\frac{g(\overline{\boldsymbol{Z}}_n)-g(\boldsymbol{\mu})}{\sigma}\leq x\right)=\Phi(x)+n^{-1/2}p_1(x)\phi(x)+n^{-1}p_2(x)\phi(x)+o(n^{-1}), \quad (1.10)$$

kde $p_1(x)$ a $p_2(x)$ jsou polynomy konečného stupně proměnné x, které jsou nezávislé na n.

Důkaz: Pro standardizovaný výběrový průměr odvozeno v (Hall (1992), str. 39-45), pro obecné statistiky tvaru hladkých funkcí výběrového průměru dokázáno v (Hall (1992), str. 52-67). Nerovnost $\limsup_{t\to\infty}|\chi(t)|<1$ se nazývá *Cramerova podmínka*. □

Poznámka 2. V předchozí větě jsme přešli od našeho značení náhodného výběru $\mathcal{X}$ k d-rozměrnému výběru $\{\boldsymbol{Z}_1,...,\boldsymbol{Z}_n\}$. Chceme-li totiž odhadnout například rozdělení rozptylu výběru $\{X_1,...,X_n\}$, definujeme $\boldsymbol{Z}_i=(X_i,X_i^2)$, pro $i=1,...,n$. Funkce g musí být tvaru

$$g(x_1,x_2)=x_2-x_1^2,$$

neboť pak

$$g(\boldsymbol{\mu})=\mathrm{E}\,X_1^2-(\mathrm{E}\,X_1)^2=\mathrm{var}\,X_1,\quad g(\overline{\boldsymbol{Z}}_n)=\overline{X_n^2}-\left(\overline{X}_n\right)^2=\frac{1}{n}\sum_{i=1}^n X_i^2-\left(\overline{X}_n\right)^2=s_n^2.$$

Poznámka 3. Větu jsme vyslovili jen pro případ, kdy známe rozptyl odhadu T_n. Obdobné výsledky platí také pro studentizované statistiky (Hall (1992), kapitola 2).

Výraz na pravé straně rovnosti (1.10) nazveme Edgeworthovým rozvojem. Ze vztahu (1.10) je také patrná rychlost konvergence standardizované statistiky k normálnímu rozdělení, a to ve shodě s Berry-Essenovou nerovností $O(n^{-1/2})$. Obecně platí, že funkce $p_j(x)$ je polynom stupně $3j-1$, je lichá pro sudá j a sudá pro lichá, což má značné důsledky.

Příklad 2. Odhadujme střední hodnotu, přirozeným odhadem je výběrový průměr. Pracujeme-li se standardizovanou statistikou, pak pro funkce $p_1(x)$ resp. $p_2(x)$ platí

$$p_1(x) = -\frac{1}{6}\kappa_3(x^2-1), \tag{1.11}$$

$$p_2(x) = -x\left[\frac{1}{24}\kappa_4(x^2-3)+\frac{1}{72}\kappa_3^2(x^4-10x^2+15)\right], \tag{1.12}$$

kde $\kappa_3 = \mathrm{E}\,(X_1-\mathrm{E}\,X_1)^3$ a $\kappa_4 = \mathrm{E}\,(X_1-\mathrm{E}\,X_1)^4 - 3\left[\mathrm{E}\,(X_1-\mathrm{E}\,X_1)^2\right]^2$. Funkce $p_1(x)$ se proto nazývá redukce šikmosti, funkce $p_2(x)$ pak redukce špičatosti. Hodnoty koeficientů κ_i, $i=3,4$ jsou v tomto případě shodné s hodnotami kumulantů.

Ve větě 3 jsme ukázali platnost a tvar Edgeworthova rozvoje pro odhady typu hladké funkce průměru. Nyní uvedeme vztah původního rozdělení H_n k bootstrapovému H_n^* pro tyto odhady.

Uvažujme splnění předpokladů z věty 3 pro funkci g a také pro výběr $\{\boldsymbol{Z}_1,...,\boldsymbol{Z}_n\}$, který vznikne vhodnou transformací náhodného výběru $\mathcal{X}$. Platí

$$H_n(x) = P(R_n(X_1,...,X_n;F)\le x) = \Phi(x)+n^{-1/2}p_1(x)\phi(x)+n^{-1}p_2(x)\phi(x)+o(n^{-1}), \tag{1.13}$$

a bootstrapovou verzi lze zapsat ve tvaru

$$H_n^*(x) = P(R_n^*(X_1^*,...,X_n^*;F_n)\le x) = \Phi(x)+n^{-1/2}\widehat{p}_1(x)\phi(x)+n^{-1}\widehat{p}_2(x)\phi(x)+o_P(n^{-1}). \tag{1.14}$$

Dále platí

$$H_n(x)-H_n^*(x) = o_P(n^{-1/2}), \quad \forall x.$$

Platnost tvrzení je zdůvodněna v Hall (1992), str. 83 - 84.

Jinými slovy, bootstrapová aproximace je výrazně přesnější nežli klasická aproximace normálním rozdělením. Ovšem toto zvýšení přesnosti platí pouze, pracujeme-li se standardizovanými statistikami, jelikož

$$G_n(x)-G_n^*(x) = \Phi\left(\sqrt{n}\frac{x-g(\boldsymbol{\mu})}{\sigma}\right)-\Phi\left(\sqrt{n}\frac{x-g(\overline{\boldsymbol{Z}}_n)}{s_n}\right)+o_P(n^{-1/2}).$$

a např. $g(\overline{\boldsymbol{Z}}_n)-g(\boldsymbol{\mu}) = O_P(n^{-1/2})$. Dalším zjištěním je velká síla bootstrapové aproximace zejména v případech, kdy rozdělení vykazuje výraznou nesymetrii. Obdobné závěry lze učinit i pro rozdělení studentizovaných verzí statistik (Hall (1992), str. 83-84).

Chceme-li testovat hypotézy či konstruovat intervaly spolehlivosti, musíme vhodně zvolit kvantily. Klasickým způsobem je opět užít centrální limitní větu a tedy použít kvantily rozdělení $N(0,1)$. Jinou možností je použít metodu bootstrap.

1.3 Intervaly spolehlivosti

Ve statistice se v případě konstrukce intervalů spolehlivosti opíráme o tvrzení z teorie pravděpodobnosti, konkrétně o centrální limitní věty. Z předpokladů o asymptotickém chování konstruujeme pro obecné odhady T_n intervaly typu

$$(T_n - u(\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}, T_n + u(\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}), \tag{1.15}$$

kde $\widehat{\sigma}$ je odhad směrodatné odchylky náhodné veličiny $\sqrt{n}T_n$. Interval (1.15) pak nazýváme $100(1-\alpha)$ procentním intervalem spolehlivosti pro odhad T_n. K samotné konstrukci takového intervalu je kromě znalosti CLV potřeba znát odhad směrodatné odchylky $\widehat{\sigma}$. V předchozí sekci jsme vysvětlili, jak lze tento problém elegantně obejít aplikací metody bootstrap. Dále jsme ukázali, že rozdělení *pivotních* (nezávisejících na neznámých parametrech) statistik lze aplikací metody bootstrap aproximovat přesněji než s pomocí normovaného normálního rozdělení. V této sekci obě tato zjištění využijeme ke konstrukci intervalů spolehlivosti.

První možnou ideou, jak vytvářet bootstrapové intervaly spolehlivosti, je použití bootstrapového odhadu funkce $\widetilde{H}_n^*$. Tento postup vede na studentizované intervaly spolehlivosti, intervaly se pak nazývají *bootstrap-t intervaly*.

Algoritmus 1 (Bootstrap-t intervaly)**.**

1. *Generujeme B výběrů $\mathcal{X}_1^*, ..., \mathcal{X}_B^*$ z původního výběru $\mathcal{X}$.*
2. *Pro každý výběr zkonstruujeme studentizovanou statistiku*

$$R_n^*(b) = \sqrt{n}\frac{T_n^*(b) - T_n}{\widehat{\sigma}^*(b)}.$$

3. *Určíme α-kvantily $\widetilde{H}_n^{*-1}(\alpha)$ jako $X_{(\lfloor B\alpha \rfloor)}$, kde $X_{(k)}$ je k-tá pořádková statistika výběru $\{R_n^*(b); 1 \le b \le B\}$.*
4. *$100(1-\alpha)$ procentním intervalem spolehlivosti odhadu T_n je interval*

$$(T_n + \widetilde{H}_n^{*-1}(\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}, T_n + \widetilde{H}_n^{*-1}(1-\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}).$$

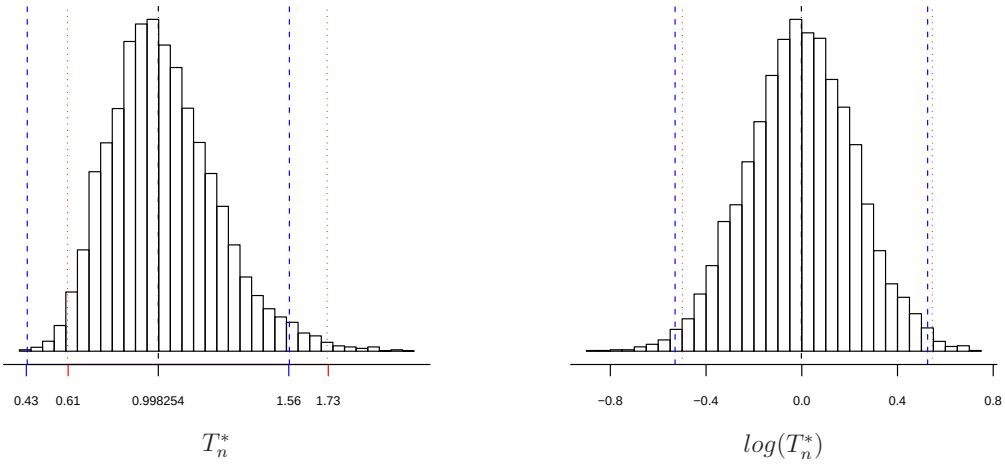
Algoritmus funguje, pokud jsme schopni odhadnout směrodatnou odchylku, v opačném případě je nutné před krokem 2 ještě aplikovat bootstrap na odhad charakteristiky $\widehat{\sigma}^*$. Metoda bootstrap-t interval se pak stává příliš výpočetně náročnou. Proto byly navrženy jiné metody pro konstrukci intervalů spolehlivosti. První z nich je založena na odhadu rozdělení nestandardizované statistiky. Protože interval je odvozen přímo z kvantilů (percentilů) rozdělení G nestandardizované statistiky dostal název *percentilový interval*.

Příklad 3. Nechť $\{X_1, ..., X_{20}\}$ je náhodný výběr, kde $X_1 \sim N(0,1)$. Zabývejme se odhadem statistiky e^{μ}, kde $\mu = \mathrm{E}\, X_1$. Správná hodnota našeho odhadu je e^0. Jako odhad použijme konzistentní $e^{\overline{X}_n}$. Pro odhad založený na CLV odhadněme hodnotu směrodatné odchylky σ pomocí metody bootstrap. Srovnejme nyní vlastnosti následujících intervalů, kde první

je percentilový a druhý pak interval založený na aproximaci normálním rozdělením.

$$\left(G_n^{*-1}(\alpha/2), G_n^{*-1}(1-\alpha/2)\right), \tag{1.16}$$

$$\left(e^{\overline{X}_n} - u(\alpha)\frac{\widehat{\sigma^*}}{\sqrt{n}}, e^{\overline{X}_n} + u(\alpha)\frac{\widehat{\sigma^*}}{\sqrt{n}}\right). \tag{1.17}$$



Obrázek 1.2: Vlevo je histogram rozdělení statistiky $T_n = e^{\overline{X}_n}$ odhadnutý pomocí 10 000 simulací. Vpravo pak histogram rozdělení statistiky $log(T_n) = \overline{X}_n$. Čárkovaně jsou vyznačeny intervaly spolehlivosti dané vztahem (1.17), tečkovaně pak percentilové bootstrapové intervaly tvaru (1.16) získané na základě 1 000 bootstrapových simulací pro jednu realizaci. Čerchovaně je označen průměr.

Ačkoliv mohou být intervaly vychýleny důsledkem odhadu na základě jediné simulace, lze z obrázku 1.2 vyčíst, že v případě, kdy se statistika řídí výrazně šikmým rozdělením, nedává klasický interval založený na CLV dobré výsledky. Velkou slabinu představuje jeho symetričnost. Naopak percentilový interval tímto omezením netrpí a tak odhaduje přesněji. Aplikujeme-li na data normalizující transformaci (v tomto případě logaritmus), pak percentilový i klasický dávají téměř shodné výsledky.

V příkladu 3 jsme ukázali, na jaké statistiky lze aplikovat metodu percentilový interval, která na rozdíl od bootstrap-t intervalu není tak početně náročná. Avšak tato metoda nedává vždy uspokojivé výsledky. Problémem je, že takto vytvořené intervaly postrádají informaci o původním odhadu T_n. Navíc víme, že bootstrapové nestudentizované rozdělení dosahuje přesnosti pouze $O_P(n^{-1/2})$. Je tedy nutné nějakým způsobem do intervalu vnést informaci o původním odhadu a směrodatné odchylce.

Byla proto navržena metoda BC_a, která oba problémy řeší pomocí dvou přidaných veličin *bias-correction* a *acceleration*. Bias correction z_0 odstraňuje neshodu mezi mediánem bootstrapovýcch odhadů a původním odhadem. Veličina acceleration a zachycuje hodnotu směrodatné odchylky. Samotný BC_α interval je konstruován pomocí kvantilů rozdělení nestandardizované statistiky, a to jako interval

$$(G_n^{*-1}(\alpha_1), G_n^{*-1}(\alpha_2)),$$

kde

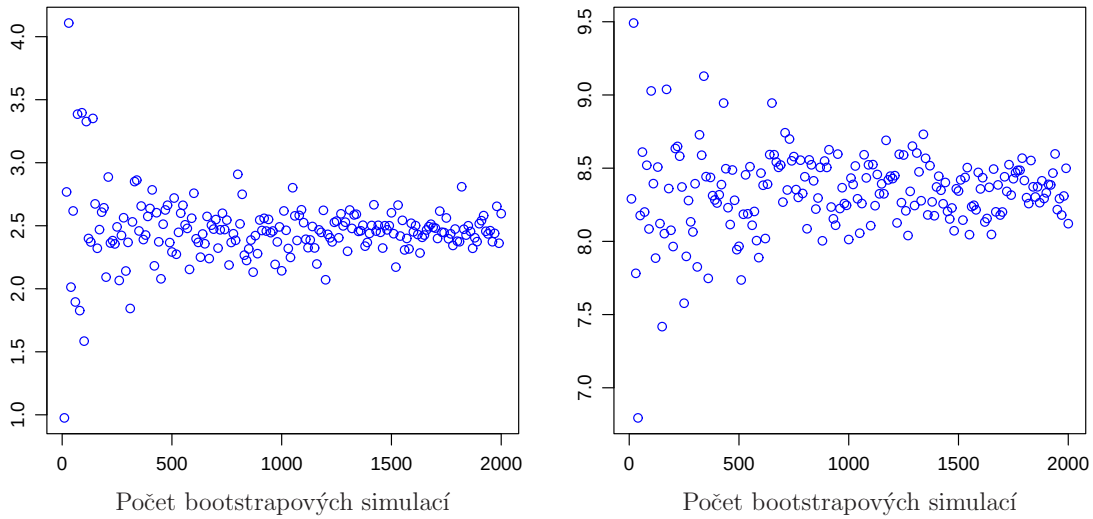
$$\alpha_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + u_{\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + u_{\alpha/2})}\right), \quad \alpha_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + u_{1-\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + u_{1-\alpha/2})}\right).$$

Hlavním pozitivem intervalu BC_a je, že aniž bychom museli statistiku standardizovat (studentizovat), dosahujeme přesnosti $O(n^{-1})$, která platí jen pro standardizované (studentizované) statistiky. Více o BC_a intervalech a dalších metodách konstrukce bootstrapových intervalů lze nalézt v (Efron a Tibshirani (1993), str. 169 - 198, 321 - 334).

1.4 Volba počtu Monte Carlo simulací

Je důležité si uvědomit, že všechny v předchozích sekcích vyslovené teoretické vlastnosti platily pro skutečné bootstrapové rozdělení H_n^*. Toto rozdělení obecně zcela přesně určit nedokážeme, avšak umíme ho libovolně přesně odhadnout pomocí metody Monte Carlo (MC) (viz výrazy (1.7), (1.8) a (1.9)).

Příklad 4. Pokračujme v příkladu 3 a studujme, jak se mění bootstrapový odhad charakteristik $n\mathrm{var}\,(e^{\overline{X}_n})$ a 95% kvantilu rozdělení statistiky $\sqrt{n}e^{\overline{X}_n}$ v závislosti na počtu bootstrapových simulací.



Obrázek 1.3: Vlevo jsou znázorněny bootstrapové odhady statistiky $n\mathrm{var}\,(e^{\overline{X}_n})$, vpravo pak bootstrapové odhady 95% kvantilu rozdělení statistiky $\sqrt{n}e^{\overline{X}_n}$ v závislosti na počtu bootstrapových simulací.

Na obrázku 1.3 je vidět, že vyšší počet než 500 simulací pro rozptyl nevede k výraznému zlepšení. V případě odhadu kvantilu se zdá být 1 000 simulací minimální mez. Pro malý počet simulací jsou bootstrapové odhady velmi nestabilní. Na základě této malé simulační studie bychom mohli soudit, že minimální počet simulací je zhruba 300 resp. 1 000 pro odhady momentů resp. v případě odhadů distribuční funkce pro výběr o rozsahu 20.

Obecně platí, že výsledná variabilita bootstrapových odhadů je složena jak z variability dané původním výběrem tak následnou MC simulací. Je proto nutné zvolit B tak velké, aby

variabilita způsobená MC simulací hrála jen velmi malou roli. Dá se ukázat, že chceme-li variabilitu způsobenou MC simulací snížit na 10 % skutečné variability, pak pro odhad rozptylu statistiky dané hladkou funkcí výběrového průměru je dle van Es a Putter (2006), str. 17 - 20, nutné zhruba

$$B \geq \frac{20\mathrm{var}\,(X_1)}{\mathrm{E}\,(X_1 - \mathrm{E}\,X_1)^4 - \mathrm{var}\,(X_1)^2} n$$

simulací. Počet simulací tedy závisí na rozptylu a špičatosti náhodné veličiny X_1. Doporučuje se zhruba 200 MC simulací. Odhadujeme-li charakteristiky založené na znalosti distribuční funkce, lze vycházet ze vztahu

$$\sup_{x\in\mathbb{R}} \left| \widehat{H_n^*} - H_n^* \right| = \epsilon_n + \sqrt{B^{-1} \log\log B},$$

kde $\epsilon_n = \sup_{x\in\mathbb{R}} |H_n^* - H_n|$ (Prášková (2004a), str. 11). Je-li tedy $\epsilon_n = O_p(n^{-1})$, volíme $B = n^2 \log n$, pak $\sqrt{B^{-1} \log\log B} = o(\epsilon_n)$. Obecně se doporučuje alespoň 1 000 MC simulací.

2 Přehled metod bootstrap pro závislá data

V této kapitole vysvětlíme, jak lze aplikovat metodu bootstrap na data, která vykazují závislostní strukturu. Pokud bychom na takováto data přímo aplikovali metodu bootstrap, nemuseli bychom dostat ani konzistentní výsledky, viz např. Lahiri (2003), str. 21 - 22. Obecně lze říci, že aplikujeme-li přímo metodu bootstrap, porušíme závislost mezi daty. Proto byly navrženy metody, které závislost zohledňují. Tyto metody se liší s ohledem na typ závislosti případně znalost modelu, kterým jsou data generována. Nicméně všechny metody mají jedno společné. Vždy je nutné se nějakým způsobem (alespoň asymptoticky) zbavit závislostní struktury. Pokud se nám podaří získat (asymptoticky) nezávislá data (bloky dat), pak se již lze opřít o poznatky z předchozí kapitoly.

2.1 Reziduální bootstrap

Nejpřímočařejší z hlediska aplikace metody bootstrap je situace, kdy známe model, kterým se data řídí. Předpokládejme, že se naše data řídí stacionárním autoregresním modelem řádu p

$$X_t = \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \epsilon_t, \qquad t \in \mathbb{Z}. \tag{2.1}$$

Standardními předpoklady jsou, že $\{\epsilon_t, t \in \mathbb{Z}\}$ jsou nezávislé stejně rozdělené náhodné veličiny dané distribuční funkcí F s konečným kladným druhým momentem σ^2. Proces chyb $\{\epsilon_t, t \in \mathbb{Z}\}$ nezachycuje systematickou složku, to znamená, že $\mathrm{E}\,\epsilon_1 = 0$. Chyby můžeme odhadnout minimalizací ztrátové funkce, nejčastěji se používá metoda nejmenších čtverců. Pro správnost celé procedury potřebujeme konzistentní odhad.

Uvažujme vektor n po sobě jdoucích náhodných veličin $\boldsymbol{X} = (X_1, ..., X_n)^T$, které se řídí modelem (2.1). Odhadneme parametry modelu metodu nejmenších čtverců a předpokládejme, že odhadnutá rezidua $\{\widehat{\epsilon}_t, t = p+1, ...n\}$ se chovají jako nezávislé stejně rozdělené náhodné veličiny, na které již lze aplikovat metodu bootstrap vysvětlenou v předchozí kapitole. Uvedeme celý algoritmus

Algoritmus 2 (Reziduální bootstrap)**.**

1. *Odhadneme neznámé parametry $\varphi_1, ..., \varphi_p$ a získáme odhad reziduí*

$$\widehat{\epsilon}_t = X_t - \widehat{\varphi}_1 X_{t-1} - \ldots - \widehat{\varphi}_p X_{t-p}, \qquad t = p+1, ..., n.$$

2. *Vypočteme empirickou distribuční funkci F_n centrovaných reziduí* $\widetilde{\epsilon}_t = \widehat{\epsilon}_t - \overline{\widehat{\epsilon}}$, *kde*

$$\overline{\widehat{\epsilon}} = \frac{1}{n-p} \sum_{t=p+1}^{n} \widehat{\epsilon}_t.$$

3. *Generujeme bootstrapové verze reziduí* $\widetilde{\epsilon}_{p+1}, ..., \widetilde{\epsilon}_n$ *pomocí funkce* F_n. *Označme je* $\epsilon^*_{p+1}, ..., \epsilon^*_n$.
4. *Konstruujeme bootstrapový výběr*

$$X^*_t = \widehat{\varphi}_1 X^*_{t-1} + \ldots + \widehat{\varphi}_p X^*_{t-p} + \epsilon^*_t, \qquad t = p+1, ..., n.$$

Počáteční hodnoty lze volit například následovně: $X^*_{1-p} = \ldots = X^*_0 = 0$.

Poznámka 4. Centrování je nutné, neboť odhadnutá rezidua musejí kopírovat původní strukturu. Ani bootstrapová rezidua nesmějí obsahovat systematickou chybu.

Ukázalo se, že pokud jsou chyby v modelu heteroskedastické, není uvedená metoda konzistentní. Pro takový případ byla odvozena metoda *wild bootstrap*. Celý algoritmus výpočtu reziduí je shodný s metodou reziduální bootstrap až na princip, jakým se získávají bootstrapové verze reziduí. Nový proces chyb je generován podle předpisu

$$\epsilon^*_t = \widehat{\epsilon}_t W_t,$$

kde W_t jsou nezávislé náhodné veličiny s rozdělením N(0, 1) a nezávislé na původním vektoru $(X_1, ..., X_n)^T$.

Možnosti aplikace metody reziduální bootstrap se samozřejmě neomezují pouze na kauzální autoregresní procesy. Lze je aplikovat i na obecnější stacionární procesy, viz Lahiri (2003), str. 206-220. Ovšem kritickým předpokladem těchto metod je znalost modelu, kterým se data řídí. Byly proto navrženy další přístupy, které nejsou omezeny tímto předpokladem. Jejich typickým představitelem jsou *blokové metody*.

2.2 Blokový bootstrap

Pokud nevíme, z jakého modelu data pocházejí, je samozřejmě možné se pokusit nalézt vhodný model a na ten aplikovat příslušnou metodu reziduální boostrap. Avšak je zatím neznámou, zda odhadnutá rezidua takového modelu lze považovat za nezávislé stejně rozdělené náhodné veličiny. Pro tyto případy byla odvozena metoda *blokový bootstrap*. Narozdíl od klasické metody nezávislý bootstrap, která bootstrapové výběry tvoří na základě jednotlivých pozorování, metoda blokový bootstrap k tomuto využívá celé bloky dat. Důsledkem toho pak blokové bootstrapové výběry částečně zachycují původní závislostní strukturu. Intuice říká, že délka bloku by měla odrážet sílu závislosti v datech. Správné určení délky bloku l je zásadním momentem při aplikaci blokového bootstrapu. Sekce 3.3 řeší problém určení vhodné délky bloku.

Obecným principem metody je rozdělit původní data na bloky $\mathcal{B}_1, ..., \mathcal{B}_k$. Bootstrapový výběr je pak tvořen náhodným výběrem z těchto bloků a jejich seřazením v pořadí, v jakém

byly vybrány. V zásadě nejpoužívanějšími jsou dvě metody, a to *Moving block bootstrap*, dále jen MBB, a *Nonoverlapping block bootstrap*, NBB. Aplikujeme-li metodu MBB, data rozdělíme na klouzavé bloky

$$\mathcal{B}_1, ..., \mathcal{B}_{n-l+1} = (X_1, ..., X_l), (X_2, ..., X_{l+1}), ..., (X_{n-l+1}, ..., X_n),$$

chceme-li použít metodu NBB, rozdělíme data na nepřekrývající se bloky

$$\mathcal{B}_1, ..., \mathcal{B}_b = (X_1, ..., X_l), (X_{l+1}, ..., X_{2l}), ..., (X_{(b-1)l+1}, ..., X_{bl}),$$

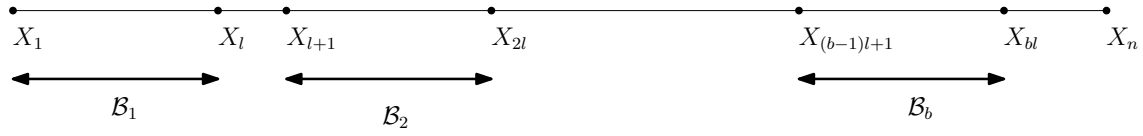
kde $b = \lfloor n/l \rfloor$. Bootstrapové verze původního výběru se pak tvoří náhodným výběrem z bloků $\mathcal{B}_1, ..., \mathcal{B}_k$. To pak například pro metodu MBB znamená, že vektory délky l

$$(X_1^*, ..., X_l^*)^T, ..., (X_{(b-1)l+1}^*, ..., X_{bl}^*)^T$$

v bootstrapovém výběru $\{X_1, ..., X_{bl}^*\}$ jsou podmíněně nezávislé stejně rozdělené náhodné vektory dané rozdělením

$$\mathrm{P}^*((X_1^*, ..., X_l^*)^T = (X_i, ..., X_{i+l-1})^T) = \frac{1}{n-l+1}, \qquad i = 1, ..., n-l+1.$$

Obrázek 2.1: Grafické znázornění volby bloků při aplikaci metody MBB



Obrázek 2.2: Grafické znázornění volby bloků při aplikaci metody NBB. b je největší přírozené číslo vyhovující podmínce $bl \leq n$.

Tedy i v situaci, že data vykazují neznámou závislostní strukturu, jsme schopni generovat jejich bootstrapové replikace. Opět generujeme B bootstrapových výběrů, pro každý odhadneme neznámou hodnotu statistiky $T_n^* = T_n(X_1^*, ..., X_n^*)$ (pokud neplatí $bl = n$, bude poslední blok v bootstrapovém výběru neúplný tak, aby velikost bootstrapového výběru byla stále n). Pak již snadno odhadneme charakteristiky rozdělení, případně samotné rozdělení odhadu T_n.

Ještě než si uvedeme další metodu, poznamenejme, že bootstrapové výběry zkonstruované na základě blokového bootstrapu nezachovávají závislost v bodech, kde se napojují jednotlivé bloky. Takto vzniklý nedostatek lze částečně napravit vhodným standardizováním (viz sekce 3.2).

2.3 Bootstrap založený na tranformaci

Jak jsme již řekli, metody bootstrap pro závislá data spojuje společná idea, a to zbavit se závislostní struktury v datech. Představili jsme již metody, které jsou založeny na znalosti modelu, kterým jsou data generována. Zabývali jsme se také metodami založenými na generování celých bloků pozorování. V této kapitole si ukážeme, že závislostní strukturu lze potlačit i zcela jiným způsobem.

Nechť $\theta = \theta(\mathrm{P})$ je parametr, který odhadujeme na základě náhodné veličiny $T_n = T_n(\mathcal{X})$. Základní myšlenkou je transformovat původní výběr tak, abychom vytvořili alespoň asymptoticky nezávislá data $\mathcal{Y} = f_n(\mathcal{X})$. Máme-li data $\mathcal{Y}$, pak, jelikož jde o (asymptoticky) nezávislá data, není chybou aplikovat klasickou metodu nezávislý bootstrap. Pro úspěšnou aplikaci je samozřejmě důležité, abychom buď dokázali vyjádřit hledanou charakteristiku resp. rozdělení veličiny T_n vzhledem k veličinám $\mathcal{Y}$, nebo byli schopni aplikovat zpětnou inverzi.

My se budeme zabývat metodou založenou na *Fourierově transformaci*, kterou definujeme jako

$$w_x(\lambda) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j e^{-ij\lambda}, \quad \lambda \in [0, 2\pi).$$

Dá se ukázat, že za určitých podmínek lze zvolit frekvence $\{\lambda_k, k = 1, ..., m\}$ pro $m < n$ pevné tak, že náhodné veličiny $\{w(\lambda_k), k = 1, ..., m\}$ jsou asymptoticky nezávislé, byť byla data v původním výběru závislá (viz kapitola 4).

2.4 Sieve bootstrap (Sítový bootstrap)

Tato metoda zatím nemá pojmenování v českém jazyce, nazývejme ji *sítový bootstrap*. V sekci 2.1 jsme se věnovali bootstrapové metodě, kterou lze aplikovat na data řídící se předem známým autoregresním modelem. Nyní jsme však v situaci, že model neznáme. V takovém případě lze neznámý správný model aproximovat jiným modelem. V praxi se používá aproximace pomocí konečně rozměrné autoregresní posloupnosti, jejíž řád volíme minimalizací informačního kritéria, např. Akaikeho informačního kritéria (AIC).

Název *sieve (sítový) bootstrap*, stejně jako u ostatních metod pro závislá data, vznikl z postupu, jakým aproximujeme sdružené pravděpobnostní rozdělení P. Lze psát (dle Lahiri (2003) str. 41-42)

$$G_n(B) = \mathrm{P}(T_n \in B) = \mathrm{P} \circ T_n^{-1}(B) \tag{2.2}$$

pro B borelovsky měřitelnou. Idea metody spočívá v postupné aproximaci míry P autoregresními posloupnostmi s rozděleními $\{\mathrm{P}_k\}_{k\geq 1}$ takovými, že P_{k+1} je přesnější než P_k.

V této souvislosti si uvědomme, že obecným principem metody bootstrap je odhad rozdělení P. V případě blokových metod předem zvolíme pevnou strukturu aproximace sdružené pravděpodobnosti P a následně odhadujeme jako empirické rozdělení. Naopak aplikujeme-li metodu sítový bootstrap, volíme jako aproximaci nejlepší z posloupnosti pravděpodobnostních měr $\{\mathrm{P}_k\}_{k\geq 1}$, které jsou určeny autoregresními posloupnostmi. Odhad je pak určen odhadem parametrů ve zvoleném modelu.

typ dat	metoda	aproximace sdruženého rozdělení P
nezávislá stejně rozdělená	nezávislý bootstrap	$\mathrm{P}_{X_1} \otimes \mathrm{P}_{X_2} \otimes \cdots \otimes \mathrm{P}_{X_n}$
závislá (stacionární)	MBB	$\mathrm{P}_{\mathcal{B}_1} \otimes \mathrm{P}_{\mathcal{B}_2} \otimes \cdots \otimes \mathrm{P}_{\mathcal{B}_{n-l+1}}$
	NBB	$\mathrm{P}_{\mathcal{B}_1} \otimes \mathrm{P}_{\mathcal{B}_2} \otimes \cdots \otimes \mathrm{P}_{\mathcal{B}_b}$
	TB	$\mathrm{P}_{Y_1} \otimes \mathrm{P}_{Y_2} \otimes \cdots \otimes \mathrm{P}_{Y_n}$
	SB	$\{\mathrm{P}_k\}_{k\geq 1}$

Tabulka 2.1: Přehled volby bootstrapových rozdělení v závislosti na typu dat. TB značí metodu bootstrap založenou na transformaci, SB pak sítový bootstrap.

2.5 Metoda Subsampling

Jedná se o alternativu bootstrapových metod. V krátkosti si vysvětleme její základní princip a vztah k blokovým metodám.

Stejně jako v případě metody MBB rozdělíme původní výběr $\mathcal{X}$ na překrývající se bloky $\mathcal{B}_1, ..., \mathcal{B}_{n-l+1}$, každý délky l. Namísto generování b bloků s cílem vytvořit bootstrapový výběr velikosti n, tvoříme podvýběry generováním jediného bloku. Tímto způsobem lze vytvořit pouhých $n - l + 1$ podvýběrů, to znamená, že jsme schopni přesně vypočítat charakteristiku odhadu metodou subsampling, aniž bychom byli nuceni aplikovat výpočetně náročnou metodu Monte Carlo.

Za odhad rozdělení studentizované statistiky $\widetilde{R}_n = (T_n - \theta_0)/s_n$ lze vzít

$$\widetilde{H}_n^*(x) = \frac{1}{n-l+1} \sum_{i=1}^{n-l+1} I_{\left[\frac{T_{l,i}-T_n}{s_l} \leq x\right]}, \quad x \in \mathbb{R},$$

kde $T_{l,i} = T_l(\mathcal{B}_i)$ (Lahiri (2003), str. 37 - 40), je to tedy odhad založený na datech z bloku $\mathcal{B}_i$, a s_l je vhodná normovací veličina. Obdobně lze definovat odhad vychýlení a rozptylu.

Vidíme, že metoda subsampling je speciálním případem blokové metody MBB, zvolíme-li $b = 1$. Hlavními výhodami jsou výpočetně snažší aplikace a absence problémů s normováním ve srovnání s blokovými metodami (více v další kapitole), neboť při aplikaci metody subsampling již nedochází ke spojování nesouvisejících částí původního výběru. S tím také souvisí, že je vhodnější k aplikaci na procesy vykazující velmi silnou závislost nežli klasické blokové metody (Lahiri (2003), str. 244 - 257). Obecně však nelze od metody očekávat příliš kvalitní výsledky, neboť při aplikaci máme k dispozici pouhých $n - l + 1$ podvýběrů.

Počet bootstrapových výběrů

Věta 4. *Nechť $(\boldsymbol{z}_1, ..., \boldsymbol{z}_k)$ je realizace náhodného výběru $(\boldsymbol{Z}_1, ..., \boldsymbol{Z}_k)$, kde $\boldsymbol{z}_i \neq \boldsymbol{z}_j$ pro $i \neq j$, pak existuje*

$$\binom{k+b-1}{b}$$

bootstrapových výběrů o velikosti b.

Důkaz. V bootstrapových výběrech nezávisí na pořadí. Jejich tvorbu lze přeformulovat jako počet způsobů, jakými je možné umístit b předmětů do k polí. Graficky můžeme znázornit jeden bootstrapový výběr, kdy $k = 8$, $b = 5$ například následovně

$$|\,|\bullet|\bullet\bullet|\,|\bullet|\bullet|.$$

Toto rozložení znamená, že zastoupení prvků původního výběru ve výběru bootstrapovém je $\{0, 0, 1, 2, 0, 1, 1, 0\}$ (na i-tém místě je vyjádřen počet prvků $\boldsymbol{z}_i$ v bootstrapovém výběru). Platí, že $(k+b-1)!$ je počet všech možných permutací $k+b-1$ prvků (předmětů a hranic polí), ale protože jsou předměty nerozlišitelné (nezávisí na pořadí výběru) a hranice políček samozřejmě také, vidíme, že počet bootstrapových výběrů je roven

$$\frac{(k+b-1)!}{b!(k-1)!}.$$

Interpretaci kombinačního čísla lépe odpovídá alternativní definice problému, totiž výběr b | z počtu $(n+k-1)$ | a jejich nahrazení předměty $\bullet$. □

n \ l	2	4	8	16
32	9.9×10^{11}	3.0×10^{7}	20475	153
	(3.0×10^{8})	(6435)	(35)	(3)
64	1.3×10^{25}	1.1×10^{16}	4.4×10^{9}	2.7×10^{5}
	(9.1×10^{17})	(3.0×10^{8})	(6435)	(35)
128	3.2×10^{51}	1.9×10^{33}	2.6×10^{20}	8.4×10^{11}
	(1.2×10^{37})	(9.2×10^{17})	(3.0×10^{8})	(6435)
256	2.7×10^{104}	7.9×10^{67}	1.2×10^{42}	1.0×10^{25}
	(2.9×10^{75})	(1.2×10^{37})	(9.2×10^{17})	(3.0×10^{8})

Tabulka 2.2: Počet bootstrapových výběrů v závislosti na velikosti výběru a délce bloku. Hodnoty jsou uvedeny pro metodu MBB, v závorce pak pro metodu NBB. Dle věty 4 je počet roven $\binom{n-l+1+n/l-1}{n/l}$ pro metodu MBB, $\binom{n/l+n/l-1}{n/l}$ pak pro NBB.

3 Blokové metody

V předchozí kapitole jsme představili obecně aplikovatelnou bootstrapovou metodu na závislá data, a to *blokový bootstrap*. Vysvětlili jsme její základní princip. Byť se princip metody zdá na první pohled velmi jednoduchým, při aplikaci narážíme na několik skrytých, avšak napravitelných vad.

V následujícím nejprve vysvětlíme, jak aplikovat metodu blokový bootstrap na statistiky závisející na vícerozměrném marginálním rozdělení. Poté se zaměříme na vlastnosti odhadů, které se dají zapsat jako hladké funkce průměru. Pro tyto odvodíme první dva momenty, které jsou nutné pro konstrukci standardizovaných statistik. Ukážeme, jaké následky může mít nesprávná volba délky bloku. Uvedeme některé asymptotické výsledky, na základě kterých odvodíme optimální délku bloku. Nakonec se zmíníme o empirických metodách na volbu vhodné délky bloku.

3.1 Vektorizovaný blokový bootstrap

Obecný postup metody blokový bootstrap uvedený v předchozí kapitole lze aplikovat na mnoho odhadů, jakými jsou např. průměr nebo medián. Ovšem všechny tyto odhady spojuje to, že jsou určeny pouze jednorozměrným marginálním rozdělením. Při analýze časových řad nás však mnohem častěji zajímají vlastnosti odhadů závisejících na vícerozměrném marginálním rozdělení, jako například kovariance. Pro tyto případy je nutné náš algoritmus mírně modifikovat.

Bude-li charakteristika θ záviset na p-dimenzionálním marginálním rozdělení, definujeme nejdříve p-rozměrné vektory $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_{n-p+1}$, kde $\boldsymbol{Y}_i = (X_i, ..., X_{i+p-1})^T$. Nyní chceme zachytit závislostní strukturu těchto nově vytvořených vektorů. Zkonstruujeme bloky $\mathcal{B}_1, ..., \mathcal{B}_{n-p-l+2}$, kde $\mathcal{B}_i = (\boldsymbol{Y}_i, ..., \boldsymbol{Y}_{i+l-1})$. K získání bootstrapového výběru $\mathcal{B}_1^*, ..., \mathcal{B}_k^*$ generujeme náhodně z kolekce bloků $\{\mathcal{B}_i; 1 \leq i \leq n-p-l+2\}$. Představenou metodu nazveme *vektorizovaný blokový bootstrap*.

Příklad 5. Představme si, že studujeme vlastnosti autokovariance řádu k. Výběrovou autokovarianci definujeme vztahem

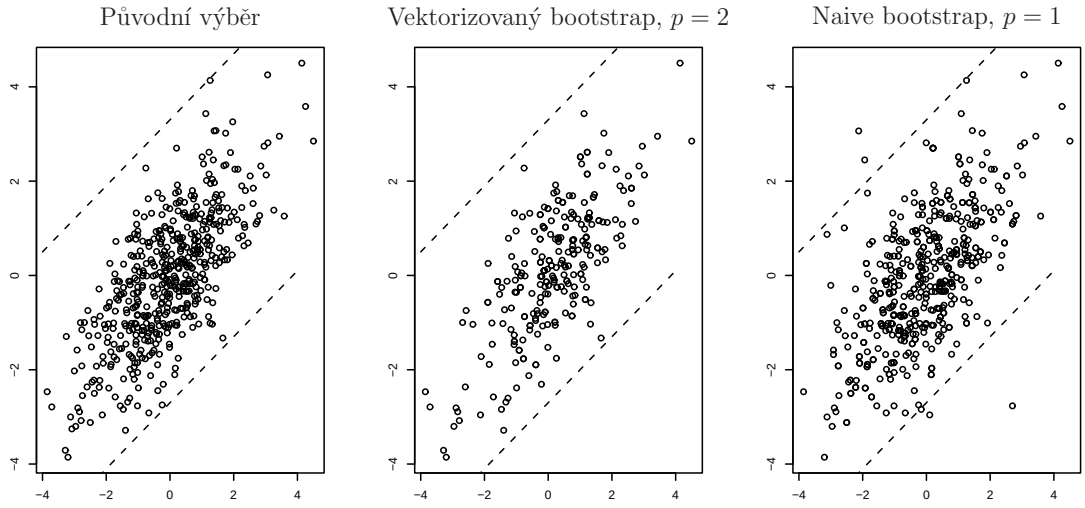
$$R(k) = \frac{1}{n} \sum_{i=1}^{n-k} (X_{i+k} - \overline{X}_n)(X_i - \overline{X}_n).$$

Zřejmě statistika $R(k)$ závisí na k+1-rozměrném marginálním rozdělení vektoru $(X_1, ..., X_n)^T$. Je zřejmé, že pokud bychom přímo tvořili bloky a z nich pak generovali bootstrapový výběr, dostaneme zkreslené výsledky. Ukažme to na jednoduchém příkladu (převzato z Bühlmann

(2002)). Nechť náhodný vektor $(X_1, ..., X_{512})^T$ se řídí procesem

$$X_t = 0.7X_{t-1} + \epsilon_t, \quad t \in \mathbb{Z},$$

kde $\{\epsilon_t, t \in \mathbb{Z}\}$ jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0, 1)$. Odhadujme autokovarianci.



Obrázek 3.1: Levý scatterplot popisuje kovarianční strukturu v původním výběru. Je sestaven z bodů $[X_{i-1}, X_i]$. Uprostřed je stejný scatterplot pro náhodně zvolený bootstrapový výběr vytvořený pro $p = 2$, v třetím grafu pak pro $p = 1$. Čárkovaně jsou vyznačeny subjektivně volené hranice na základě původního výběru.

Na obrázku 3.1 je vidět, že generujeme-li bloky přímo z výběru $\mathcal{X}$, metoda selhává, neboť statistiky vypočtené na základě takto vytvořené bootstrapových výběrů mohou být silně ovlivněny nově vytvořenými odlehlými body (body neležící v pásu).

3.2 Normování

Ač je v praxi nejčastěji využívána metoda MBB, má své nevýhody. Při této metodě nejsou v bootstrapových výběrech zastoupeny pozorování rovnoměrně. Stejný problém s sebou nese také metoda NBB, často se pozorování s indexem $i \geq bl + 1$ nepoužívají. Podívejme se na následky tohoto nesymetrického zastoupení v případě, že odhadujeme střední hodnotu. Poznamenejme, že analyzovat chování metod při odhadu střední hodnoty není příliš restriktivní, neboť se ukazuje, že mnoho statistik lze zapsat jako hladké funkce průměru. Patří sem například kovariance a korelace, které jsou zásadní při analýze časových řad. Pro větší přehlednost definujme bootstrapovou verzi výběrového průměru vztahem $\overline{X}^*_m = m^{-1} \sum_{i=1}^{m} X^*_i$, kde $m = l \lfloor n/l \rfloor$.

Abychom mohli odhadovat rozptyl, případně celé rozdělení, je zaprvé potřeba statistiku centrovat. Je nasnadě zvolit pro centrování výběrový průměr původní populace, avšak výraz $\overline{X}^*_m - \overline{X}_n$ má nenulové vychýlení. Stejně opatrní musíme být i při studentizování.

Věta 5. *Nechť $\mathcal{X} = \{X_1, ..., X_n\}$ je výběr a $\mathcal{X}^* = \{X_1^*, ..., X_n^*\}$ jeho bootstrapový protějšek, označme $m = lb$, kde l je délka bloku a $b = \lfloor n/l \rfloor$. Jestliže jsme výběr $\mathcal{X}^*$ získali aplikací metody MBB, platí*

$$\begin{aligned}
\mu^*_{MBB} &= E^*(\overline{X}^*_m) = (n-l+1)^{-1}\left[n\overline{X}_n - l^{-1}\sum_{i=1}^{l-1}(l-i)(X_i + X_{n-i+1})\right], \\
\sigma^{2*}_{MBB} &= var^*(m^{1/2}\overline{X}^*_m) = l(n-l+1)^{-1}\sum_{i=1}^{n-l+1} S_i^2 - l\,\mu^{*2}_{MBB}
\end{aligned}$$

a pokud jsme bootstrapový výběr vytvořili metodou NBB, platí

$$\begin{aligned}
\mu^*_{NBB} &= E^*(\overline{X}^*_m) = (bl)^{-1}\left[n\overline{X}_n - \sum_{i=bl+1}^{n} X_i\right], \\
\sigma^{2*}_{NBB} &= var^*(m^{1/2}\overline{X}^*_m) = lb^{-1}\sum_{i=1}^{b} S^2_{(i-1)l+1} - l\,\mu^{*2}_{NBB},
\end{aligned}$$

kde $S_i = l^{-1}\sum_{j=i}^{i+l-1} X_j$.

Důkaz: Označme $S_i^* = l^{-1}\sum_{j=i}^{i+l-1} X_j^*$. A uvědomme si, že náhodné veličiny S_i^* jsou nezávislé stejně rozdělené náhodné veličiny podmíněně na výběru $\mathcal{X}$.

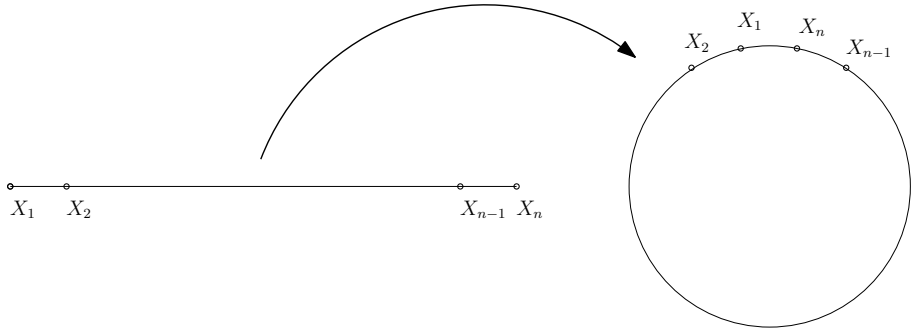
$$\begin{aligned}
\mu^*_{MBB} &= \mathrm{E}^*(\overline{X}^*_m) = \mathrm{E}^*\left(b^{-1}\sum_{i=1}^{b} S_i^*\right) = \mathrm{E}^*(S_1^*) = (n-l+1)^{-1}\sum_{i=1}^{n-l+1} S_i \\
&= (n-l+1)^{-1}\left[n\overline{X}_n - l^{-1}\sum_{i=1}^{l-1}(l-i)(X_i + X_{n-i+1})\right], \\
\sigma^{2*}_{MBB} &= m\mathrm{var}^*\left(\overline{X}^*_m\right) = m\mathrm{var}^*\left(b^{-1}\sum_{i=1}^{b} S_i^*\right) = mb^{-1}\mathrm{var}^*(S_1^*) \\
&= l\left[\mathrm{E}\,S_1^{*2} - \left(\mathrm{E}\,S_1^{*2}\right)^2\right] = l(n-l+1)^{-1}\sum_{i=1}^{n-l+1} S_i^2 - l\,\mu^{*2}_{MBB}.
\end{aligned}$$

Pro metodu NBB, platí

$$\begin{aligned}
\mu^*_{NBB} &= \mathrm{E}^*(\overline{X}^*_m) = \mathrm{E}^*\left(b^{-1}\sum_{i=1}^{b} S_i^*\right) = \mathrm{E}^*(S_1^*) = b^{-1}\sum_{i=1}^{b} S_{(i-1)l+1} \\
&= (bl)^{-1}\left[n\overline{X}_n - \sum_{i=bl+1}^{n} X_i\right], \\
\sigma^{2*}_{NBB} &= m\mathrm{var}^*\left(\overline{X}^*_m\right) = m\mathrm{var}^*\left(b^{-1}\sum_{i=1}^{b} S_i^*\right) = mb^{-1}\mathrm{var}^*(S_1^*) \\
&= l\left[\mathrm{E}\,S_1^{*2} - \left(\mathrm{E}\,S_1^{*2}\right)^2\right] = lb^{-1}\sum_{i=1}^{b} S^2_{(i-1)l+1} - l\,\mu^{*2}_{NBB}. \qquad \square
\end{aligned}$$

Důsledek 1. *Při konstrukci bootstrapových statistik je nutné centrovat výrazem* $E^*(T_m^*)$.

Existuje ale i metoda, jak vytvářet bloky, při které se vyhneme nutnosti takto centrovat a nerovnoměrného využívání dat při generování bootstrapového výběru. S ohledem na způsob, jakým jsou bootstrapové výběry generovány, dostala název *Circular Block Bootstrap*, dále jen CBB. Metoda je založena na "přepisu" dat do kružnice, jak je ukázáno na obrázku 3.2. Přepíšeme-li data do tohoto formátu, pak pro vytvoření bootstrapového výběru stačí zvolit náhodně počátek bloku a do bloku pak kromě tohoto pozorování zahrnout ještě následujících $l-1$ hodnot na kružnici.



Obrázek 3.2: Přepis původního souboru dat pro aplikaci metody CBB.

Metoda CBB zaručuje, že průměr bootstrapového výběru bude roven průměru původního souboru a že data do bootstrapového výběru vstupují rovnoměrně. Ovšem je otázkou, zda-li tvorbou bloků typu $(X_j, ..., X_n, X_1, ..., X_{l-n+j-1})$ nepokazíme původní závislostní strukturu dat.

Byť metoda CBB zaručuje vlastnost $\mathrm{E}^*(\overline{X}_m^*) = \overline{X}_n$, již neplatí $\mathrm{var}^*\left(\sqrt{m}\overline{X}_m^*\right) = var\left(\sqrt{n}\overline{X}_n\right)$. Při studentizovaci statistiky $\sqrt{m}(\overline{X}_m^* - \mathrm{E}^*\overline{X}_m^*)$, která má tedy nulové vychýlení, nelze jednoduše použít výběrový ekvivalent

$$s_n^{2*} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^* - \overline{X}_n^*\right)^2.$$

Pro přehlednost předpokládejme dále, že $n = bl$ a zabývejme se rozptylem výběrového průměru v bootstrapovém výběru získaném metodou NBB. Z věty 5 plyne, že je-li $n = bl$, pak $\mu_{NBB}^* = \mathrm{E}^*\overline{X}_n^* = \overline{X}_n$ a

$$\begin{aligned}\sigma_{NBB}^{2*} &= \frac{l}{b}\sum_{i=1}^{b} S_{(i-1)l+1}^2 - l\,\mu_{NBB}^{*2} = \frac{1}{bl}\sum_{i=1}^{b}\sum_{s=1}^{l}\sum_{t=1}^{l} X_{(i-1)l+s}X_{(i-1)l+t} - l\overline{X}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^{b}\sum_{s=1}^{l}\sum_{t=1}^{l}(X_{(i-1)l+s} - \overline{X}_n)(X_{(i-1)l+t} - \overline{X}_n). \qquad (3.1)\end{aligned}$$

V knize Härdle et al. (2001), str. 10-11, je uvedeno, že $s_n^{2*} - \sigma_{NBB}^{2*} = O_P((l/n)^{1/2})$. Jinými slovy, aproximací studentizované statistiky bootstrapovým protějškem studentizovaným veličinou $\sqrt{s_n^{2*}}$ bychom získali méně přesné výsledky, než jaké poskytuje aproximace normálním rozdělením. Vhodnou volbu studentizace ukážeme v další sekci, na které naváže věta shrnující výsledky asymptotické přesnosti aproximace bootstrapovým rozdělením.

3.3 Přesnost blokového bootstrapu a optimální délka bloku

Pro správnou aplikaci metody blokový bootstrap, ať už v jakémkoliv tvaru, je zásadní otázka volby vhodné délky bloku. Nesprávná volba může odhady výrazně vychýlit, případně zvýšit rozptyl. Pro optimální volbu neexistuje jednoduchý recept. Je ovlivněna v zásadě třemi faktory, a to procesem, kterým byla data generována, statistikou, kterou studujeme, a nakonec cílovou charakteristikou (vychýlení, rozptyl nebo odhad rozdělení). Obvyklým pravidlem pro volbu délky bloku je minimalizace střední čtvercové chyby (MSE). Uveďme nyní hlavní teoretické výsledky pro bootstrapový odhad vychýlení ($\widehat{\text{bias}}\, T_n$), rozptylu ($\widehat{\text{var}}\, T_n$) a distribuční funkce standardizované a studentizované statistiky ($\widehat{H_n}$) a ($\widehat{\widetilde{H}_n}$) pro třídu odhadů spadajících do množiny hladkých funkcí průměru.

Přesnost aproximace vychýlení a rozptylu

Mějme $\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n\}$ *n po sobě jdoucích náhodných vektorů, které se řídí d-rozměrným stacionárním procesem* $\{\boldsymbol{Z}_t, t \in \mathbb{Z}\}$. *Nechť pro délku bloku l platí* $l^{-1}+n^{-1/2}l = o(1)$ *pro* $n \to \infty$. *Za odhad parametru* $\theta = g(\boldsymbol{\mu})$ *volme konzistentní* $T_n = g(\overline{\boldsymbol{Z}}_n)$ *a ptejme se, jaké má náš odhad* T_n *vychýlení a rozptyl. Za platnosti obecných předpokladů, kterými jsou omezení na hladkost funkce g a konečnost absolutních momentů procesu* $\{\boldsymbol{Z}_t\}$, *pro odhady získané metodou MBB platí*

$$MSE(\widehat{bias}\, T_n) = A_1^2 n^{-2} l^{-2} + g_1(0) n^{-3} l + o(n^{-2}l^{-2} + n^{-3}l), \tag{3.2}$$

$$MSE(\widehat{var}\, T_n) = A_2^2 n^{-2} l^{-2} + g_2(0) n^{-3} l + o(n^{-2}l^{-2} + n^{-3}l), \tag{3.3}$$

kde A_i *jsou hodnoty závisející na autokovarianční funkci procesu* $\{\boldsymbol{Z}_t\}$ *a* $g_i(x)$ *jsou nezáporné reálné funkce pro* $i = 1, 2$. *Pokud bychom odhadovali s využitím metody NBB, pak vztahy (3.2) a (3.3) opět platí, pouze s nahrazením* $g_i = 3g_i/2$ *pro* $i = 1, 2$.

Pro všechny nutné předpoklady tvrzení a důkaz odkazujeme na knihu Lahiri (2003), kapitola 5 (tvrzení 5.1).

Poznámka 5. Ve vzorcích (3.2) a (3.3) je první člen odvozen z vychýlení, druhý je pak dán rozptylem odhadované charakteristiky. Povšimněme si, že vychýlení je pro metody MBB a NBB shodné, zatímco rozptyl je 1.5 krát vyšší pro metodu NBB. Výsledek odpovídá intuici, totiž že pokud povolíme blokům se překrývat, bude celková variabilita odhadu menší. Z tohoto důvodu se dále zaměříme jen na metodu MBB, která jak z teoretického hlediska, tak na základě numerických výsledků (Lahiri (2003), str. 115 - 118) dosahuje přesnějších výsledků.

Poznámka 6. Vychýlení není monotónní funkcí délky bloku, přesnějším asymptotickým rozvojem bychom zjistili, že jde ve skutečnosti o konvexní funkci délky bloku (Hall (1992), str. 564 - 568).

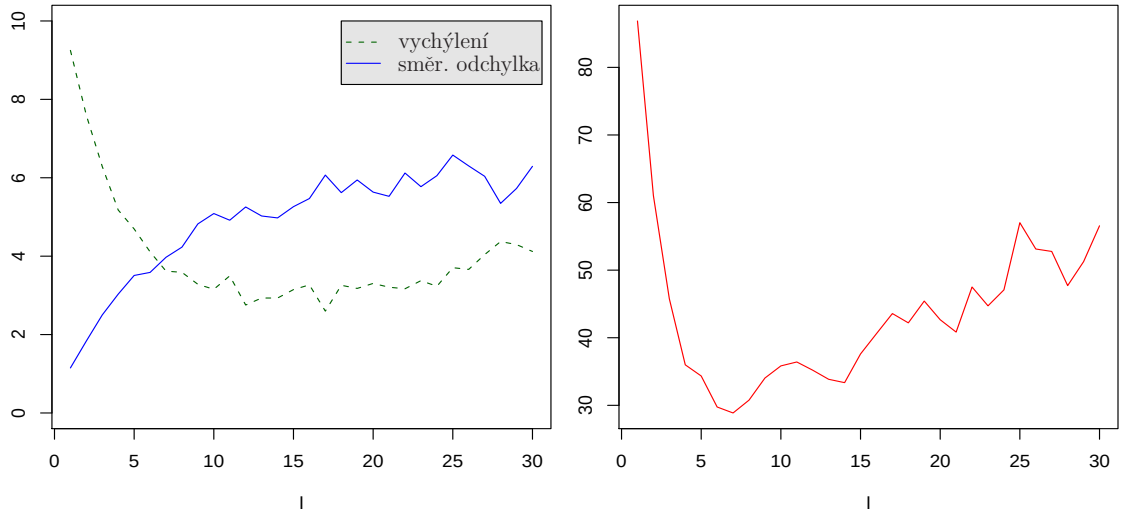
Vztahy (3.2) a (3.3) mohou sloužit k prvotní představě o optimální volbě bloku pro odhad vychýlení a rozptylu. Minimální MSE řádu $O(n^{-2-2/3})$ pro oba případy dosáhneme volbou $l \approx n^{1/3}$. Přesná optimální délka bloku však závisí na hodnotách neznámé autokovarianční funkce. Proto je nutné při volbě optimální délky bloku postupovat například na základě minimalizace MSE. Algoritmům určeným na volbu optimální délky bloku je věnována sekce 3.4.

Příklad 6. Mějme autoregresní posloupnost tvaru

$$X_t = 0,6X_{t-1} + \epsilon_t, \quad t = 1, ..., 125, \tag{3.4}$$

kde ϵ_t jsou nezávislé stejně rozdělené náhodné veličiny s centrovaným χ^2 rozdělením o jednom stupni volnosti. Zabývejme se vlastnostmi výběrového průměru, odhadujme charakteristiku $n\text{var}\,(\overline{X}_n)$.

Na obrázku 3.3 je vidět, že směrodatná odchylka je rostoucí funkcí délky bloku, vychýlení je konvexní funkcí délky bloku. Ve střední čtvercové chybě se odrážejí obě tyto vlastnosti a dosahuje minima pro hodnotu $l = 7$. V tabulce 3.1 jsou pak vypsány odhady pro délky bloku $1, ..., 15$.



Obrázek 3.3: Závislost vychýlení, směrodatné odchylky (vlevo) a střední čtvercové chyby (vpravo) na délce bloku.

Nyní se zaměříme na přesnost aproximace standardizovanými a studentizovanými bootstrapovými rozděleními. Stále pracujme s odhadem $T_n = g(\overline{\boldsymbol{Z}}_n)$. Víme, že bootstrapové odhady centrujeme výrazem $g(\boldsymbol{\mu}^*)$. Ukažme, jak postupovat při studentizaci původních i bootstrapových odhadů a definujme veličiny vyskytující se v tvrzení o přesnosti bootstrapové aproximace rozdělení. Následující úvahy jsou provedeny dle Götze a Künsch (1996). Označme $\mathrm{D}g(\boldsymbol{x})$ gradient funkce g v bodě $\boldsymbol{x} \in \mathbb{R}^d$. Využijme Taylorova rozvoje odhadu T_n

$$T_n = g(\overline{\boldsymbol{Z}}_n) \approx g(\boldsymbol{\mu}) + \mathrm{D}g(\boldsymbol{\mu})^T(\overline{\boldsymbol{Z}}_n - \boldsymbol{\mu}),$$

proto

$$\sqrt{n}(T_n - g(\boldsymbol{\mu})) \approx \sqrt{n}\mathrm{D}g(\boldsymbol{\mu})^T(\overline{\boldsymbol{Z}}_n - \boldsymbol{\mu}) =: \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i.$$

Z posledního výrazu a výrazu pro rozptyl výběrového průměru plyne, že lze psát

$$\text{var}\,\left(\sqrt{n}(T_n - g(\boldsymbol{\mu})\right) \approx \sum_{k=-n}^{n}\left(1 - \frac{|k|}{n}\right)\mathrm{E}\,(U_0U_k) =: \sigma_n^2.$$

l	odhad	vychýlení	směr. odch.	MSE	
1	3.06337	−9.24913	1.14905	86.86664	
2	4.71993	−7.59257	1.83658	61.02013	
3	6.02679	−6.28571	2.50143	45.76728	
4	7.13557	−5.17693	3.03196	35.99338	
5	7.61822	−4.69428	3.50799	34.34230	
6	8.20212	−4.11038	3.58604	29.75495	
7	8.69394	−3.61856	3.97176	28.86887	*
8	8.72491	−3.58759	4.23049	30.76791	
9	9.02988	−3.28262	4.82300	34.03687	
10	9.15792	−3.15458	5.08745	35.83353	
11	8.81463	−3.49787	4.91729	36.41481	
12	9.55884	−2.75366	5.25498	35.19744	
13	9.37976	−2.93274	5.02504	33.85192	
14	9.38047	−2.93203	4.97582	33.35561	
15	9.16551	−3.14699	5.26280	37.60057	

Tabulka 3.1: Kvalita odhadů získaných metodou MBB v závislosti na délce bloku. Pro každou délku bloku provedeno 500 simulací. Správná teoretická hodnota je 12.3125.

Za odhad σ_n^2 zvolme

$$s_n^2 = \sum_{k=0}^{l-1} w_k \widehat{\mathrm{cov}}\,(U_0, U_k) = \sum_{k=0}^{l-1} w_k \frac{1}{n} \sum_{j=1}^{n-l} \widehat{U}_j \widehat{U}_{j+k},$$

kde $\widehat{U}_j = \mathrm{D}g(\overline{\boldsymbol{Z}}_n)^T(\boldsymbol{Z}_j - \overline{\boldsymbol{Z}}_n)$ a w_k váhy. V podstatě to znamená, že namísto odhadu veličiny σ_n^2 odhadujeme

$$\tau_n^2 = \sum_{k=0}^{l-1} w_k \mathrm{E}\,(U_0 U_k).$$

Vhodnou volbou vah w_k získáme konzistentní odhad. Dle Götze a Künsch (1996) je nutné, aby váhy byly tvaru $w_0 = 1$, $w_k = 2w(k/l)$ pro $0 < k < l$, kde w značí funkci s vlastnostmi

$$w : [0, 1) \to [0, 1], \qquad w(0) = 1.$$

Ukazuje se, že asymptotická vzdálenost $\tau_n^2 - \sigma_n^2$ je minimální, a to $O(n^{-1})$, při volbě $w \equiv 1$.

Pro úplnost uveďme výraz, jakým studentizovat bootstrapová rozdělení

$$s_m^{2*} = \frac{1}{m} \sum_{i=1}^{b} \sum_{s=1}^{l} \sum_{t=1}^{l} \widehat{U}^*_{(i-1)l+s} \widehat{U}^*_{(i-1)l+t},$$

kde $\widehat{U}_j^* = \mathrm{D}g(\overline{\boldsymbol{Z}}^*_m)^T(\boldsymbol{Z}_j^* - \overline{\boldsymbol{Z}}^*_m)$.

Přesnost aproximace rozdělení

Nechť R_n resp. $\widetilde{R}_n$ je normovaná resp. studentizovaná statistika odhadu T_n a R_n^* resp. $\widetilde{R}_n^*$ jsou jejich bootstrapové protějšky získané aplikací blokové metody s pevnou délkou bloku l. Pak za splnění obecných předpokladů platí

$$\sup_{x\in\mathbb{R}} |P^*(R_n^* \leq x) - P(R_n \leq x)| = O_P(n^{-1}l + n^{-1/2}l^{-1}), \tag{3.5}$$

$$\sup_{x\in\mathbb{R}} |P^*(\widetilde{R}_n^* \leq x) - P(\widetilde{R}_n \leq x)| = O_P(n^{-1+\delta}l + n^{-1/2}l^{-1} + |\tau_n^2 - \sigma_n^2|) \tag{3.6}$$

pro libovolné $\delta > 0$.

Všechny nutné předpoklady tvrzení a důkaz lze nalézt v knize Lahiri (2003), kapitola 6 (tvrzení 6.7 a 6.8) a článku Götze a Künsch (1996).

Výraz (3.5) je minimalizován při volbě $l = n^{1/4}$, pro níž vychází aproximace řádu $O_P(n^{-3/4})$. Poznamenejme, že jsme dosáhli zlepšení, neboť aproximace normálním rozdělením je řádu $O(n^{-1/2})$.

Ačkoliv jsme dosáhli zpřesnění odhadů rozdělení, mají blokové metody velkou slabinu a tou je nutnost standardizovat odhady. Pro nestandardizované resp. nestudentizované statistiky uvedené vlastnosti aproximace neplatí. Připomeňme, že studium asymptotických vlastností blokových metod se omezuje jen na odhady, které lze zapsat jako hladké funkce výběrového průměru.

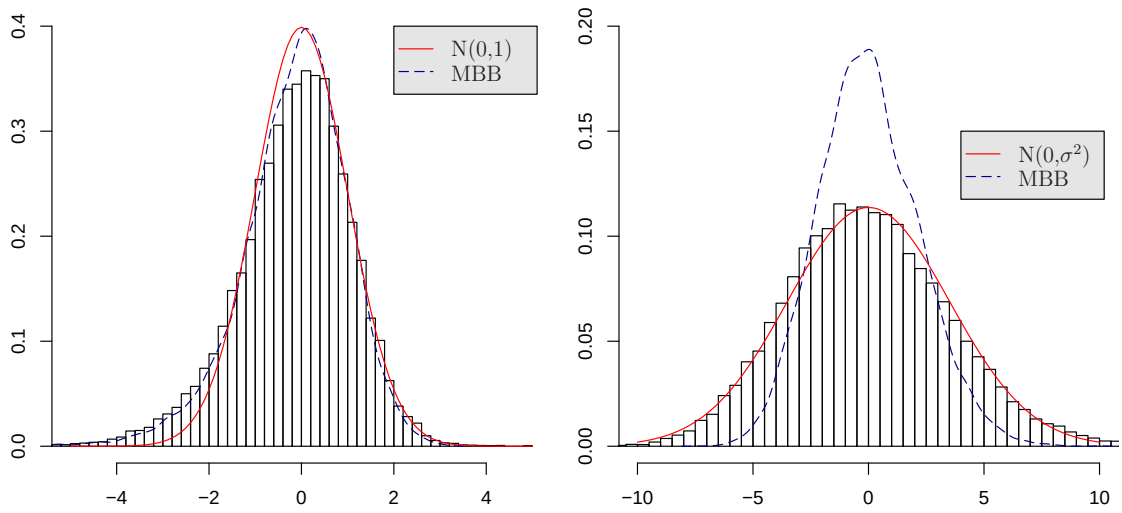
Příklad 7. Uvedli jsme teoretické asymptotické vlastnosti blokových metod. Pokračujme nyní v příkladu 6, studujme studentizované a pouze centrované rozdělení odhadu $T_n = \overline{X}_n$. Pravý histogram na obrázku 3.4 ukazuje, že bootstrapové odhady mají menší rozptyl než odhady získané na základě původního výběru. Tento rozdíl vzniká nezávislostí pozorování v místě navazování bloků. Je proto žádoucí odhady vhodně studentizovat. Na levém histogramu vidíme studentizované verze. Při pohledu na obrázek lze tvrdit, že bootstrapová aproximace oproti symetrickému normálnímu rozdělení zachycuje mírné zešikmení.

3.4 Empirická volba optimální délky bloku

V předešlé části jsme odvodili optimální délku bloku. Ta však závisí na neznámé autokovarianční funkci. Byly proto navrženy algoritmy, jak optimální délku bloku odhadnout.

Optimální délka tvorbou podvýběrů

Dle výrazů (3.2) a (3.3) platí, že optimální délka bloku pro odhad vychýlení či rozptylu je řádu $n^{-1/3}$, odhadujeme-li rozdělení pak řádu $n^{-1/4}$. S ohledem na tyto výsledky byla navržena metoda, která hledá optimální délku bloku tvorbou podvýběrů, ve kterých určí optimální délku bloku a následně délku upraví na celý výběr, publikováno v Hall et al. (1995).



Obrázek 3.4: Histogram studentizované statistiky $\widetilde{R}_n = \sqrt{n}(T_n - \mathrm{E}\,T_n)/s_n$ (pro studentizaci zvoleny obdélníkové váhy) na levém obrázku a centrované $\sqrt{n}(T_n - \mathrm{E}\,T_n)$ na pravém a jejich aproximace normálním rozdělením resp. metodou blokový bootstrap. Skutečné rozdělení odhatnuto na základě 20 000 simulací, počet bootstrapových simulací 5 000, $l = 5$, $\sigma^2 = 12.3125$.

Algoritmus 3 (Volba délky bloku tvorbou podvýběrů)**.**

1. *Výběr $\mathcal{X}$ o n pozorováních rozdělíme na překrývající se podvýběry o velikosti m, označme $\mathcal{S} = \{\mathcal{X}_i, i = 1, ..., n - m + 1\}$, kde $\mathcal{X}_i = \{X_i, ..., X_{i+m-1}\}$ (např. $m = \lfloor n/2 \rfloor$).*
2. *Odhadneme charakteristiku studované statistiky, označme ji $\varphi(T_n)$, na základě blokové metody s délkou bloku l_n z výběru $\mathcal{X}$.*
3. *Provedeme stejné odhady $\varphi(T^{l_m}_{mi})$ pro $i = 1, ..., n - m + 1$ jako v bodě 2., tzn. pro každý podvýběr z množiny $\mathcal{S}$ a s délkou bloku $l_m \in L$, kde L značí vhodnou množinu délek bloků.*
4. *Definujme délku $l^{opt}_{\mathcal{S}}$ jako*

$$l^{opt}_{\mathcal{S}} = \underset{l \in L}{argmin} \sum_{i=1}^{n-m+1} \left(\varphi(T^{l}_{mi}) - \varphi(T_n)\right)^2.$$

5. *V závislosti na charakteristice, kterou studujeme, je optimální délka bloku v celém výběru*

$$l^{opt} = \left(\frac{n}{m}\right)^k l^{opt}_{\mathcal{S}},$$

kde $k = 1/3$ pro vychýlení a ropztyl a $k = 1/4$ pro odhady rozdělení.

Celý algoritmus je možno iterovat, to znamená použít délku l^{opt} jako počáteční volbu v kroku 2.

Příklad 8. Připomeňme příklad 6. Uvažujeme autoregresní posloupnost tvaru

$$X_t = 0,6X_{t-1} + \epsilon_t, \quad t = 1, ..., 125, \tag{3.7}$$

kde ϵ_t jsou nezávislé stejně rozdělené náhodné veličiny s centrovaným χ^2 rozdělením o jednom stupni volnosti. Zabýváme se vlastnostmi výběrového průměru, odhadujeme charakteristiku $n\mathrm{var}\,(\overline{X}_n)$.

V příkladu 6 jsme odhadli optimální délku bloku pro proces tvaru (3.4). Aplikujme na tento proces uvedený algoritmus na hledání vhodné délky bloku. Tabulka 3.2 ilustruje kvalitu představené metody. Přibližně dvě třetiny odhadnutých délek bloku leží v intervalu $[5, 9]$. Varovná je zhruba třetina výsledků v intervalu $[2, 3]$ a nevýznamná změna zvýšením počtu iterací.

	l														
# iterací	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	4	27	0	34	0	21	0	11	3	0	0	0	0	0
2	0	14	24	0	24	0	18	0	17	3	0	0	0	0	0
3	0	17	17	0	34	0	10	0	18	4	0	0	0	0	0

Tabulka 3.2: Frekvence optimální délky bloku odhadnuté na základě metody založené na tvorbě podvýběrů. Celkem provedeno 100 simulací, pro odhady charakteristik provedeno vždy 500 bootstrapových výběrů.

Optimální délka odhadem parametrů

O následující metodě se lze dočíst v knize Lahiri (2003), str. 185 - 197. Metoda přímo vychází z asymptotických rozvojů (3.2), (3.3) a (3.5) a je založena na odhadu parametrů v těchto rozvojích.

V případě odhadu vychýlení a rozptylu lze vedoucí člen rozvoje střední čtvercové chyby zapsat ve tvaru

$$A_i^2 n^{-2} l^{-2} + g_i n^{-3} l, \quad i = 1, 2, \tag{3.8}$$

kde A_i a g_i jsou neznámé konstanty. Jedná se o ryze konvexní funkci proměnné l. Výraz (3.8) je proto minimalizován pro délku bloku l, splňující rovnici

$$-2A_i^2 n^{-2} l^{-3} + g_i n^{-3} = 0, \quad i = 1, 2.$$

Střední čtvercová chyba je minimálizována pro délku bloku

$$l_i^{opt} = \left(\frac{2nA_i^2}{g_i} \right)^{\frac{1}{3}}, \quad i = 1, 2.$$

Dalšími úvahami lze dospět k výrazu pro optimální délku bloku v případě odhadu standardizované distribuční funkce (Lahiri (2003), str. 180 - 182, 187), platí

$$l_3^{opt} = \left(\frac{2nA_3^2}{2g_3} \right)^{\frac{1}{4}}.$$

Vidíme, že optimální délka bloku závisí na neznámých parametrech A_i a g_i. Odhady těchto parametrů lze nalézt v knize Lahiri (2003), str. 187 - 192.

Příklad 9. Pokračujme dále v příkladu 8. Srovnejme výkonnost parametrické metody hledání délky bloku s metodou založenou na tvorbě podvýběrů. Z tabulky 3.3 plyne, že téměř 80 % výsledných vhodných délek dle algoritmu leží v intervalu $[4, 7]$, ve kterém leží optimální délka (7) a MSE se od optimální neliší významně. Pouze pětina výsledných délek leží v intervalu $[2, 3]$, kde již MSE dosahuje významně vyšších hodnot.

Na základě výsledků tohoto příkladu lze tvrdit, že parametrická metoda dosahuje vyrovnanějších výsledků (většina výsledných délek leží v intervalu $[4, 6]$), pouze pětina dat leží v kritickém intervalu $[1, 3]$ a navíc jde o metodu výpočetně méně náročnou.

l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
frekvence	1	14	27	54	50	40	10	3	1	0	0	0	0	0	0

Tabulka 3.3: Frekvence optimální délky bloku odhadnuté na základě metody odhadu parametrů. Celkem provedeno 200 simulací a pro každou provedeno 5 000 bootstrapových výběrů k získání konzistenstních odhadů vychýlení a rozptylu.

3.5 Stacionární bootstrap

V některých případech požadujeme, aby nově vzniklý bootstrapový výběr zachovával stacionaritu. Poznamenejme, že ani jedna ze tří dosud probraných blokových metod stacionaritu nezachovává. Z tohoto důvodu byla navržena nová metoda pro tvorbu bloků, *stacionární bootstrap*.

Délka bloku i volba počátku bloku jsou nezávislé náhodné veličiny (samozřejmě nezávislé také na původním výběru $\mathcal{X}$). Pro stacionární bootstrap je důležité si data zapsat do kružnice. Při konstrukci bloku v první fázi se zvolí počáteční index, generujeme náhodnou veličinu P z rovnoměrného rozdělení $R(1, ..., n)$. Ve druhé fázi určujeme délku bloku, ta je dána geometrickým rozdělením s parametrem p, to znamená

$$\mathrm{P}(D = l) = p(1 - p)^{l-1}, \quad j = 1, 2, ...$$

(Data máme uspořádána na kružnici). Navíc jsou náhodné veličiny P a D vzájemně nezávislé.

Konstrukce bootstrapového výběru se může na první pohled zdát zbytečně složitá. Avšak má dvě výhody. Zbavujeme se nutnosti volit délku bloku. A za druhé a především, takto zkonstruovaný bootstrapový výběr je stacionární. Stacionární bootstrap se neužívá příliš často, neboť se ukázalo, že metody s pevnou délkou bloku dosahují lepších výsledků (Lahiri (2003), kap. 5).

4 Bootstrap založený na Fourierově transformaci

4.1 Vlastnosti Fourierovy transformace

V následující části zavedeme potřebné definice a vyslovíme lemmata nutná k důkazu věty o asymptotickém chování diskrétních Fourierových transformací.

Definice 1. *Fourierovu transformaci* náhodného vektoru $\boldsymbol{X} = (X_1, ..., X_n)^T$ definujeme vztahem

$$w_x(\lambda) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j e^{-ij\lambda}, \qquad \lambda \in [0, 2\pi). \tag{4.1}$$

Pracujme s frekvencemi $\{\lambda_j, j = 0, ..., n-1\}$, kde $\lambda_j = 2\pi j/n$. Příslušné $w(\lambda_j)$ budeme nazývat *diskrétní Fourierovy transformace (DFT)*. Zavedeme následující značení:

$$w_{x,j} = w_x(\lambda_j),\ w_{x,-j} = w_x(-\lambda_j) = \overline{w_x(\lambda_j)},$$

kde index x znamená, že se jedná o DFT vzhledem k vektoru $(X_1, ..., X_n)^T$.

Definice 2. Skalární součin na prostoru $\mathbb{C}^n$ definujeme jako $\langle u, v\rangle = \sum_{j=1}^{n} u_j \overline{v}_j$.

Lemma 6. *Nechť*

$$\boldsymbol{e}_j = \frac{1}{\sqrt{n}} \left(e^{i\lambda_j}, e^{i2\lambda_j}, ..., e^{in\lambda_j}\right)^T,$$

kde $\lambda_j = 2\pi j/n$. Pak soubor vektorů $\{\boldsymbol{e}_j, j = 0, ..., n-1\}$ tvoří ortonormální bázi v $\mathbb{C}^n$.

Důkaz:

$$\langle \boldsymbol{e}_j, \boldsymbol{e}_k\rangle = \frac{1}{n} \sum_{l=1}^{n} e^{il(\lambda_j - \lambda_k)} = \begin{cases} 1, & j = k \\ \frac{1}{n} e^{i(\lambda_j - \lambda_k)} \frac{1 - e^{in(\lambda_j - \lambda_k)}}{1 - e^{i(\lambda_j - \lambda_k)}} = 0, & j \neq k. \end{cases}$$

Soubor $\{\boldsymbol{e}_j, j = 0, ..., n-1\}$ tvoří n lineárně nezávislých vektorů, je to tedy báze prostoru $\mathbb{C}^n$. □

Poznámka 7. Zřejmě lze psát $w_{x,j} = \langle \boldsymbol{X}, \boldsymbol{e}_j\rangle$.

Důsledek 2. *Platí*

$$X_t = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} w_{x,j} e^{it\lambda_j}, \qquad t = 1, ..., n, \tag{4.2}$$

vektorově pak

$$\boldsymbol{X} = \sum_{j=0}^{n-1} w_{x,j} \boldsymbol{e}_j.$$

Důkaz:

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} w_{x,j} e^{it\lambda_j} &= \frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} X_k e^{-ik\lambda_j} e^{it\lambda_j} = \frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} X_k e^{-i\lambda_j (k-t)} \\ &= \frac{1}{n} \sum_{j=0}^{n-1} X_t = X_t \end{aligned}$$ □

Vztah (4.2) nazveme inverzní Fourierovou transformací. Vektor $\boldsymbol{X}$ jsme vyjádřili vzhledem k bázi $\{\boldsymbol{e}_j\}_{j=0}^{n-1}$ s koeficienty $\{w_{x,j}\}_{j=0}^{n-1}$.

Definice 3. Posloupnost $\{Y_t, t \in \mathbb{Z}\}$ nekorelovaných náhodných veličin s nulovou střední hodnotou a konečným kladným rozptylem σ^2 nazveme *bílý šum* a budeme značit $\{Y_t\} \sim WN(0, \sigma^2)$.

Ještě než vyslovíme větu, která odhalí asymptotické chování diskrétních Fourierových transformací, uvedeme několik tvrzení, která jsou k jejímu důkazu nezbytná.

Věta 7. *Budiž* $\{X_t, t \in \mathbb{Z}\}$ *lineární proces*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \xi_{t-j}, \quad t \in \mathbb{Z},$$

kde $\{\xi_t\} \sim WN(0, \sigma^2)$. *Nechť dále* $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. *Pak pro spektrální hustotu procesu* $\{X_t\}$ *platí*

$$f(\lambda) = \frac{\sigma^2}{2\pi} \psi(e^{-i\lambda}) = \frac{\sigma^2}{2\pi} \left| \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda} \right|^2.$$

Důkaz: viz Prášková (2004b), věta 5.6. □

Definice 4. *Komplexní náhodnou veličinu* definujeme jako $X = \operatorname{Re} X + i \operatorname{Im} X$, kde $\operatorname{Re} X$ a $\operatorname{Im} X$ jsou reálné náhodné veličiny. Řekneme, že X je *komplexní náhodná veličina s normálním rozdělením s parametry* μ *a* σ^2, označme $X \sim N^C(\mu, \sigma^2)$, je-li vektor $(\operatorname{Re} X, \operatorname{Im} X)^T$ rozdělen jako

$$\begin{pmatrix} \operatorname{Re} X \\ \operatorname{Im} X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \operatorname{Re} \mu \\ \operatorname{Im} \mu \end{pmatrix}, \frac{1}{2} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right).$$

Lemma 8 (Cramérova-Slutzkého věta)**.** *Nechť* $\{X_t\}$, $\{Y_t\}$ *jsou posloupnosti náhodných veličin a* X *náhodná veličina. Pak platí implikace*

$$X_t \xrightarrow{D} X,\ Y_t \xrightarrow{P} 0 \Rightarrow X_t + Y_t \xrightarrow{D} X, \quad \textit{pro } t \to \infty.$$

Důkaz: viz Brockwell a Davis (1996), tvrzení 6.3.3. □

Věta 9 (Varianční matice Fourierových transformací)**.** *Nechť $\xi_1, \ldots, \xi_n$ jsou nekorelované centrované náhodné veličiny s konečným kladným rozptylem σ^2, definujme*

$$\begin{aligned}\alpha(\lambda_j) &= \sqrt{\frac{2}{n}} \sum_{t=1}^{n} \xi_t \cos(\lambda_j t),\\ \beta(\lambda_j) &= \sqrt{\frac{2}{n}} \sum_{t=1}^{n} \xi_t \sin(\lambda_j t).\end{aligned}$$

Pak vektor $\boldsymbol{g} = (\alpha(\lambda_1), ..., \alpha(\lambda_{\lfloor (n-1)/2 \rfloor}), \beta(\lambda_1), ..., \beta(\lambda_{\lfloor (n-1)/2 \rfloor}))^T$ je vektor nekorelovaných náhodných veličin se shodným rozptylem σ^2.

Důkaz. Definujme matici $\boldsymbol{G}$ typu $(n \times (n-2))$ pro sudé n resp. $(n \times (n-1))$ pro n liché, pro jejíž t-tý řádek platí

$$\boldsymbol{G}_{t\bullet} = \left(\sqrt{\frac{2}{n}} \cos(\lambda_1 t), ..., \sqrt{\frac{2}{n}} \cos(\lambda_{\lfloor (n-1)/2 \rfloor} t), \sqrt{\frac{2}{n}} \sin(\lambda_1 t), ..., \sqrt{\frac{2}{n}} \sin(\lambda_{\lfloor (n-1)/2 \rfloor} t) \right),$$

pak zřejmě $\boldsymbol{g} = \boldsymbol{G}^T \boldsymbol{\xi}$, kde $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$. Počítejme přímo

$$\operatorname{var} \boldsymbol{g} = \mathrm{E}\, \boldsymbol{G}^T \boldsymbol{\xi} \boldsymbol{\xi}^T \boldsymbol{G} = \boldsymbol{G}^T \mathrm{E}\, \boldsymbol{\xi} \boldsymbol{\xi}^T \boldsymbol{G} = \sigma^2 \boldsymbol{G}^T \boldsymbol{G}.$$

Zbývá ukázat, že sloupcové vektory matice $\boldsymbol{G}$ jsou ortonormální. Užijme vzorce

$$\cos xt = \frac{e^{ixt} + e^{-ixt}}{2}, \ \sin xt = \frac{e^{ixt} - e^{-ixt}}{2i}$$

a uvědomme si, že

$$\sqrt{\frac{2}{n}} \sum_{t=1}^{n} e^{it2\pi j/n} = \begin{cases} \sqrt{2}, & j = 0 \\ \sqrt{\frac{2}{n}} e^{i2\pi j/n} \frac{1 - e^{i2\pi j}}{1 - e^{i2\pi j/n}} = 0, & j \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{4.3}$$

Prvky g_{jk} matice $\boldsymbol{G}^T \boldsymbol{G}$ lze zapsat jako

$$\begin{aligned}\frac{2}{n} \sum_t \cos(\lambda_j t) \cos(\lambda_k t) &= \frac{2}{n} \sum_t \frac{e^{i(\lambda_j + \lambda_k)t} + e^{i(\lambda_j - \lambda_k)t} + e^{-i(\lambda_j - \lambda_k)t} + e^{-i(\lambda_j + \lambda_k)t}}{4},\\ \frac{2}{n} \sum_t \cos(\lambda_j t) \sin(\lambda_k t) &= \frac{2}{n} \sum_t \frac{e^{i(\lambda_j + \lambda_k)t} - e^{i(\lambda_j - \lambda_k)t} + e^{-i(\lambda_j - \lambda_k)t} - e^{-i(\lambda_j + \lambda_k)t}}{4i},\\ \frac{2}{n} \sum_t \sin(\lambda_j t) \sin(\lambda_k t) &= \frac{2}{n} \sum_t \frac{e^{i(\lambda_j + \lambda_k)t} - e^{i(\lambda_j - \lambda_k)t} - e^{-i(\lambda_j - \lambda_k)t} + e^{-i(\lambda_j + \lambda_k)t}}{-4}.\end{aligned}$$

S ohledem na vlastnost (4.3) je vektor $\boldsymbol{g}$ skutečně vektorem nekorelovaných náhodných veličin. □

Poznámka 8. Přidáme-li k matici $\boldsymbol{G}$ sloupce ($\boldsymbol{e}_{n/2}$, je-li n sudé)

$$\boldsymbol{e}_0 = \frac{1}{\sqrt{n}}\mathbf{1}, \qquad \boldsymbol{e}_{n/2} = \frac{1}{\sqrt{n}}\left(\cos(\lambda_{n/2}), ..., \cos(n\lambda_{n/2})\right)^T$$

získáme ortonormální bázi $\mathbb{R}^n$.

Věta 10. *Budiž $\{X_t, t \in \mathbb{Z}\}$ lineární proces jako ve větě 7. Předpokládejme navíc, že posloupnost $\{\xi_t\}$ je posloupnost nezávislých stejně rozdělených náhodných veličin. Nechť opět $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Uvažujme vektor n po sobě jdoucích náhodných veličin $(X_1, ..., X_n)^T$, které se řídí procesem $\{X_t\}$. Pak:*

$$cov\,(w_{x,j}, w_{x,k}) = \begin{cases} 2\pi f(\lambda_j) + o(1) & pro\ \lambda_j = \lambda_k, \\ o(1) & pro\ \lambda_j \neq \lambda_k, \end{cases} \qquad 1 \leq j, k \leq n-1.$$

Je-li navíc

$$\sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2} < \infty, \tag{4.4}$$

pak

$$cov\,(w_{x,j}, w_{x,k}) = \begin{cases} 2\pi f(\lambda_j) + O(n^{-1}) & pro\ \lambda_j = \lambda_k, \\ O(n^{-1}) & pro\ \lambda_j \neq \lambda_k, \end{cases} \qquad 1 \leq j, k \leq n-1.$$

kde $f(\lambda)$ značí spektrální hustotu.

Důkaz: Nejprve je nutné Fourierovy transformace vyjádřit vzhledem k náhodným veličinám $\{\xi_t\}$, které jsou nezávislé a stejně rozdělené, a proto na ně lze aplikovat CLV.

$$\begin{aligned} w_x(\lambda) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t e^{-it\lambda} = \frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}\left(\sum_{t=1}^{n}\xi_{t-j}e^{-i(t-j)\lambda}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}\left(\sum_{t=1-j}^{n-j}\xi_t e^{-it\lambda}\right) = \sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}\frac{1}{\sqrt{n}}\left(\sum_{t=1}^{n}\xi_t e^{-it\lambda}\right) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}\left[\sum_{t=1-j}^{0}\xi_t e^{-it\lambda} - \sum_{t=n-j+1}^{n}\xi_t e^{-it\lambda}\right] \\ &=: \boldsymbol{\psi}(e^{-i\lambda})w_\xi(\lambda) + \frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}U_{nj} =: \boldsymbol{\psi}(e^{-i\lambda})w_\xi(\lambda) + R_n, \end{aligned} \tag{4.5}$$

kde $\boldsymbol{\psi}(e^{-i\lambda})$ značí řadu $\sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}$ a $w_\xi(\lambda)$ jsou diskrétní Fourierovy transformace náhodné posloupnosti $\{\xi_t\}$. Náhodné veličiny U_{nj} jsou součty nekorelovaných náhodných veličin s nulovou střední hodnotou a konečným kladným rozptylem, proto

$$\begin{aligned} \mathrm{E}\,U_{nj} &= 0, \\ \mathrm{E}\,|U_{nj}|^2 &\leq 2\sigma^2 min(|j|, n), \end{aligned}$$

neboť se jedná o součet $2|j|$ pro $j < n$ respektive $2n$ pro $j \geq n$ nezávislých stejně rozdělených náhodných veličin. Člen R_n má zřejmě nulovou střední hodnotu a rozptyl lze odhadnout výrazem

$$\begin{aligned} \mathrm{E}\,|R_n|^2 &= \mathrm{E}\left|\frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}\psi_j e^{-ij\lambda}U_{nj}\right|^2 \leq \left(\frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}|\psi_j|\left(\mathrm{E}\,|U_{nj}|^2\right)^{1/2}\right)^2 \\ &\leq 2\sigma^2\left(\frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}|\psi_j|min(|j|,n)^{1/2}\right)^2, \end{aligned} \tag{4.6}$$

vzhledem k Minkowského nerovnosti, ze které plyne že

$$\mathrm{E}\left|\sum_i X_i\right|^2 \leq \left[\sum_i\left(\mathrm{E}\,|X_i|^2\right)^{1/2}\right]^2.$$

Zvolme nyní posloupnost $m(n)$ závisející na n s vlastnostmi $m \to \infty$ pro $n \to \infty$ a zároveň $\frac{m(n)}{n} \to 0$. Potom

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{j=-\infty}^{\infty}|\psi_j|min(|j|,n)^{1/2} &= \frac{1}{\sqrt{n}}\sum_{|j|\leq m(n)}|\psi_j||j|^{1/2} + \sum_{|j|>m(n)}|\psi_j| \\ &\leq \left(\frac{m(n)}{n}\right)^{1/2}\sum_{|j|\leq m(n)}|\psi_j| + \sum_{|j|>m(n)}|\psi_j| \end{aligned}$$

a vzhledem k předpokládaným vlastnostem posloupnosti $m(n)$ konverguje poslední výraz k nule pro $n \to \infty$. S použitím *Čebyševovy nerovnosti* také $R_n \xrightarrow{P} 0$ pro $n \to \infty$. Připomeňme, že naše dosavadní úvahy nezávisely na volbě frekvence λ. Z definice DFT plyne, že

$$\begin{aligned} \mathrm{cov}\,(w_{\xi,j}, w_{\xi,k}) &= \mathrm{E}\,w_{\xi,j}w_{\xi,-k} = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}\,\xi_t\xi_s e^{-it\lambda_j}e^{is\lambda_k} \\ &= \frac{\sigma^2}{n}\sum_{t=1}^{n}e^{-it(\lambda_j-\lambda_k)}, \end{aligned}$$

což znamená, že

$$\mathrm{cov}\,(w_{\xi,j}, w_{\xi,k}) = \begin{cases} \sigma^2 & \text{pro } \lambda_j = \lambda_k, \\ 0 & \text{pro } \lambda_j \neq \lambda_k, \end{cases} \qquad 1 \leq j,k \leq n-1.$$

Z rozpisu (4.5) plyne, že

$$\begin{aligned} \mathrm{cov}\,(w_{x,j}, w_{x,k}) &= \boldsymbol{\psi}(e^{-i\lambda_j})\boldsymbol{\psi}(e^{i\lambda_k})\mathrm{cov}\,(w_{\xi,j}, w_{\xi,k}) + Z_n \\ &= 2\pi f(\lambda_j)I_{[j=k]} + Z_n. \end{aligned}$$

Pro zbytek Z_n užitím Jensenovy a Schwarzovy nerovnosti platí

$$\begin{aligned} |Z_n| &\leq \mathrm{E}\left|\boldsymbol{\psi}(e^{-i\lambda_j})w_{\xi,j}\overline{R_n}\right| + \mathrm{E}\left|\boldsymbol{\psi}(e^{i\lambda_j})w_{\xi,-j}R_n\right| + \mathrm{E}\,|R_n|^2 \\ &\leq 2|\boldsymbol{\psi}(e^{-i\lambda_j})|\left[\mathrm{var}\,(w_{\xi,j})\right]^{1/2}\left[\mathrm{E}\,|R_n|^2\right]^{1/2} + \mathrm{E}\,|R_n|^2 \approx o(1), \end{aligned}$$

jelikož jsme ukázali, že $\mathrm{E}\,|R_n|^2 \to 0$ pro $n \to \infty$. Platí-li dále předpoklad (4.4), pak

$$\mathrm{E}\,|R_n|^2 \leq \frac{K}{n}, \quad n \to \infty$$

pro nějaké $K \in (0, \infty)$ s ohledem na nerovnost (4.6). A tedy $Z_n \approx O(n^{-1})$. □

Věta 11. *Nechť jsou splněny předpoklady věty 10 a je dáno $m < n$ pevné. Pak pro frekvence $0 < \lambda_{i_1} < \ldots < \lambda_{i_m} < \pi$ platí*

$$(w_x(\lambda_{i_1}), ..., w_x(\lambda_{i_m}))^T \xrightarrow{D} (Z_{i_1}, ..., Z_{i_m})^T, \quad n \to \infty,$$

kde $Z_j \sim N^C(0, 2\pi f(\lambda_j))$. Navíc vektor $(Z_{i_1}, ..., Z_{i_m})^T$ je vektor nezávislých náhodných veličin.

Důkaz: Využijme rozpisu z předchozího důkazu. Definujme

$$\boldsymbol{Y}_t = \left(\sqrt{2}\xi_t \cos(\lambda_{i_1} t), \sqrt{2}\xi_t \sin(\lambda_{i_1} t), ..., \sqrt{2}\xi_t \cos(\lambda_{i_m} t), \sqrt{2}\xi_t \sin(\lambda_{i_m} t)\right)^T, \quad t = 1, ..., n$$

a ukažme pomocí Feller-Lindebergovy verze centrální limitní věty pro náhodné vektory (Lachout (2004), str. 106 - 107), že

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{Y}_t \xrightarrow{D} \boldsymbol{U}, \quad n \to \infty,$$

kde $\boldsymbol{U} \sim N_{2m}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{2m})$. Vzhledem k vlastnostem náhodných veličin $\{\xi_t\}$ a větě 9 platí

$$\begin{aligned} \mathrm{E}\,\boldsymbol{Y}_t &= \boldsymbol{0}, \qquad t = 1, ..., n, \\ \mathrm{var}\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{Y}_t\right) &= \sigma^2 \boldsymbol{I}_{2m}. \end{aligned}$$

Nakonec ověřme Feller-Lindebergovu podmínku

$$\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\,\|\boldsymbol{Y}_t\|^2 I_{\left[\|\boldsymbol{Y}_t\| \geq \sqrt{n}\epsilon\right]} \leq \mathrm{E}\left[m\xi_1^2 I_{\left[|m\xi_1| \geq \sqrt{n}\epsilon\right]}\right] \to 0, \quad n \to \infty, \tag{4.7}$$

neboť náhodná veličina ξ_1 má konečný druhý moment. Z definice diskrétních Fourierových transformací platí

$$w_{\xi,j} = \frac{1}{\sqrt{2n}} \sum_{t=1}^{n} (\boldsymbol{Y}_{t,j} - i\boldsymbol{Y}_{t,j+1}).$$

Proto

$$\begin{pmatrix} \mathrm{Re}\, w_{\xi,j} \\ \mathrm{Im}\, w_{\xi,j} \end{pmatrix} \xrightarrow{D} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right).$$

S ohledem na definici 4 zřejmě $w_{\xi,j} \xrightarrow{D} N^C(0, \sigma^2)$ a navíc $(w_{\xi,1}, ..., w_{\xi,m})^T$ je vzhledem k diagonalitě varianční matice vektor nekorelovaných komplexních náhodných veličin. V případě

normálního rozdělení to znamená, že jde také o *nezávislé* náhodné veličiny. V předchozím důkazu jsme došli ke vztahu

$$w_x(\lambda) = \boldsymbol{\psi}(e^{-i\lambda})w_\xi(\lambda) + R_n,$$

kde $\mathrm{E}\,|R_n|^2 \to 0$ pro $n \to \infty$. Dále platí

$$\mathrm{var}\left[\boldsymbol{\psi}(e^{-i\lambda_j})w_{\xi,j}\right] = |\boldsymbol{\psi}(e^{-i\lambda_j})|^2\sigma^2.$$

Aplikací Cramér-Slutzkého věty a věty 7 se dokáže, že

$$(w_{x,i_1}, ..., w_{x,i_m})^T \xrightarrow{D} (Z_{i_1}, ..., Z_{i_m})^T,$$

kde $Z_j \sim N^C(0, 2\pi f(\lambda_j))$ a $(Z_{i_1}, ..., Z_{i_m})^T$ je vektor nezávislých komplexních náhodných veličin. □

Poznámka 9. Nezávislost jsme dokázali pro frekvence $0 < \lambda_j < \pi$, platnost věty lze rozšířit na frekvence $0 \le \lambda_j \le \pi$, pro které

$$\begin{aligned} \omega_{x,0} &\xrightarrow{D} Z_{(0)} \sim N(\sqrt{n}\mathrm{E}\,X_1, 2\pi f(\lambda_0)), \quad n \to \infty, \\ \omega_{x,n/2} &\xrightarrow{D} Z_{(n/2)} \sim N(0, 2\pi f(\lambda_{n/2})), \quad n \text{ sudá}, n \to \infty. \end{aligned}$$

O dalších vlastnostech diskrétních Fourierových transformací se lze dočist v knize Brillinger (1975).

Na základě diskrétní Fourierovy transformace se nám podařilo eliminovat závislostní strukturu v datech a získat asymptoticky nezávislé náhodné veličiny. Na takto transformované náhodné veličiny můžeme aplikovat klasický nezávislý bootstrap. Tuto metodu s ohledem na transformaci nazveme *frekvenční bootstrap*.

Asymptotických vlatností diskrétních Fourierových transformací a periodogramu se hojně využívá, často je však nutné provést zpětnou transformaci z frekvenční domény zpět do časové. Nyní si ukážeme způsob, jak aplikovat metodu frekvenční bootstrap na regresní modely a to bez nutnosti zpětné transformace.

4.2 Regresní model ve frekvenční doméně

V této části se budeme zabývat *modely nekorelované stochastické regrese* tvaru

$$y_t = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_t + e_t, \quad t = 1, ..., n, \tag{4.8}$$

kde $\{e_t\}$ je slabě stacionární proces s vlastnostmi $\mathrm{E}\,e_t = 0$, $\mathrm{var}\,e_t = \sigma^2 > 0$ pro $t = 1, ..., n$, $\{\boldsymbol{x}_t\}$ je slabě stacionární p-rozměrný proces s regulární varianční maticí $\mathrm{E}\left(\boldsymbol{x}_t\boldsymbol{x}_t^T\right) = \boldsymbol{\Sigma} > 0$ pro $t = 1, ..., n$ a procesy $\{\boldsymbol{x}_t\}$ a $\{\epsilon_s\}$ jsou vzájemně nekorelované pro $t, s = 1, ..., n$.

Příkladem může být model, kde se regresory i chyby řídí kauzálním autoregresním procesem prvního řádu. Vedle těchto modelů se pracuje i s procesy s tzv. dlouhou pamětí, které jsou definovány níže.

Procesy s dlouhou pamětí

Definice 5. Nechť $\{X_t, t \in \mathbb{Z}\}$ je stacionární proces. Řekneme, že $\{X_t, t \in \mathbb{Z}\}$ je proces s krátkou pamětí, platí-li

$$\sum_{k=0}^{\infty} |R_x(k)| < \infty,$$

kde R_x značí autokovarianční funkci procesu $\{X_t, t \in \mathbb{Z}\}$, a pro jeho spektrální hustotu platí

$$f_x(\lambda) > 0, \quad \lambda \in [0, 2\pi].$$

Příklad 10. Příkladem procesu s krátkou pamětí může být autoregresní proces prvního řádu $\{X_t\}$ daný vztahem

$$X_t = \varphi X_{t-1} + \epsilon_t, \quad t \in \mathbb{Z},$$

kde $|\varphi| < 1$ a $\{\epsilon_t\} \sim WN(0, \sigma_\epsilon^2)$. Potom je

$$R_x(k) = \frac{\sigma_\epsilon^2}{1-\varphi^2}\varphi^{|k|}, \quad k \in \mathbb{Z},$$

$$f_x(\lambda) = \frac{\sigma_\epsilon^2}{2\pi}\frac{1}{1-2\varphi\cos\lambda+\varphi^2}, \quad \lambda \in [0, 2\pi].$$

Odtud je vidět, že skutečně $\sum_{k=0}^{\infty} |R_x(k)| < \infty$ a $f_x(\lambda) > 0, \quad \lambda \in [0, 2\pi]$. Podobně lze ukázat, že obecné kauzální procesy ARMA jsou modely s krátkou pamětí.

Definice 6. Budiž $\{X_t, t \in \mathbb{Z}\}$ stacionární proces. Řekneme, že $\{X_t, t \in \mathbb{Z}\}$ je s dlouhou pamětí, existuje-li $d \in (0, 0.5)$ takové, že proces $\{Y_t\}$ definovaný vztahem

$$Y_t = (1-B)^d X_t$$

je proces s krátkou pamětí. B značí operátor posunutí definovaný vztahem $BX_t = X_{t-1}$.

Proces $\{X_t\}$ lze získat aplikací nekonečného filtru $(1-B)^{-d}$ na proces $\{Y_t\}$. Pro spektrální hustoty proto platí vztah

$$f_x(\lambda) = |1 - e^{-i\lambda}|^{-2d} f_y(\lambda).$$

A jelikož

$$|1 - e^{-i\lambda}| = \sqrt{(1-\cos(\lambda))^2 + \sin(\lambda)^2} = \sqrt{2\,(1-\cos(\lambda))}$$

a

$$\lim_{\lambda \to 0} \frac{\sqrt{2\,(1-\cos(\lambda))}}{|\lambda|} = 1,$$

pro chování spektrální hustoty u nuly platí

$$f_x(\lambda) = O(\lambda^{-2d}).$$

Odhad parametrů regresního modelu ve frekvenční doméně

V následujícím studujme vlastnosti odhadu metodou nejmenších čtverců (OLS) v modelu nekorelované stochastické regrese ve frekvenční doméně. Zabýváme se vlastnostmi odhadu směrového koeficientu $\boldsymbol{\beta}$, proto předpokládejme, že matice regresorů i vektor závisle proměnných jsou centrované.

Definice 7. Regresní model

$$y_t = \boldsymbol{\beta}^T \boldsymbol{x}_t + e_t, \quad t = 1, ..., n, \tag{4.9}$$

nazveme modelem v časové doméně, model tvaru

$$w_{y,j} = \boldsymbol{\beta}^T \boldsymbol{w}_{x,j} + w_{e,j}, \quad j = 1, ..., n-1, \tag{4.10}$$

pak jeho protějškem v doméně frekvenční. Odhady metodou nejmenších čtverců parametru $\boldsymbol{\beta}$ označme jako $\widehat{\boldsymbol{\beta}}_{OLS}$ v modelu (4.9) a v modelu (4.10) jako $\widehat{\boldsymbol{\beta}}_{FOLS}$. Je tedy

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{OLS} &= \left[\boldsymbol{X}^T \boldsymbol{X}\right]^{-1} \left[\boldsymbol{X}^T \boldsymbol{y}\right], \\ \widehat{\boldsymbol{\beta}}_{FOLS} &= \left(\boldsymbol{W}_x^\dagger \boldsymbol{W}_x\right)^{-1} \left(\boldsymbol{W}_x^\dagger \boldsymbol{w}_y\right) = \left(\sum_{j=1}^{n-1} \boldsymbol{w}_{x,j} \boldsymbol{w}_{x,j}^\dagger\right)^{-1} \left(\sum_{j=1}^{n-1} \boldsymbol{w}_{x,j} w_{y,-j}\right), \end{aligned}$$

kde $\boldsymbol{W}_x = (\boldsymbol{w}_{x,1}, ..., \boldsymbol{w}_{x,n-1})^T$.

Lemma 12. *Je-li $\boldsymbol{x} \in \mathbb{R}^n$ pak*

$$w_{x,n-j} = \overline{w_{x,j}}, \qquad j = 1, ..., \lfloor n/2 \rfloor .$$

Důkaz: Z periodicity komplexní funkce e^{iz} a vlastností skalárního součinu plyne

$$w_{x,n-j} = \langle \boldsymbol{x}, \boldsymbol{e}_{n-j} \rangle = \langle \boldsymbol{x}, \overline{\boldsymbol{e}_j} \rangle = \overline{\langle \boldsymbol{x}, \boldsymbol{e}_j \rangle} = \overline{w_{x,j}}.$$

□

Pro úspěšnou aplikaci metody bootstrap je nutné zachovat tuto symetrii. Dále předchozí lemma poskytuje prostor ke snížení výpočetní náročnosti. Můžeme totiž odhady počítat pouze na základě prvních $\lfloor n/2 \rfloor$ diskrétních Fourierových transformací.

Důsledek 3. *Uvažujme regresní model tvaru (4.10). Pak pro odhad směrového koeficientu platí*

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{FOLS} &= \left[\sum_{j=1}^{(n-1)/2} \boldsymbol{w}_{x,j} \boldsymbol{w}_{x,j}^\dagger\right]^{-1} \left[Re \sum_{j=1}^{(n-1)/2} \boldsymbol{w}_{x,j} w_{y,j}^\dagger\right], \text{ pro } n \text{ liché} \\ &= \left[\left(\sum_{j=1}^{(n-2)/2} 2\boldsymbol{w}_{x,j} \boldsymbol{w}_{x,j}^\dagger + \boldsymbol{w}_{x,n/2} \boldsymbol{w}_{x,n/2}^\dagger\right)\right]^{-1} \left[Re \left(\sum_{j=1}^{(n-2)/2} 2\boldsymbol{w}_{x,j} w_{u,j}^\dagger + \boldsymbol{w}_{x,n/2} w_{y,n/2}^\dagger\right)\right], \\ &\text{pro } n \text{ sudé.} \end{aligned}$$

V následující větě vyslovíme tvrzení o vztahu odhadů $\widehat{\boldsymbol{\beta}}_{FOLS}$ a $\widehat{\boldsymbol{\beta}}_{OLS}$.

Věta 13. *Uvažujme model nekorelované stochastické regrese tvaru (4.9) resp. (4.10). Platí, že*

$$\widehat{\boldsymbol{\beta}}_{FOLS} = \widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}.$$

Důkaz: Připomeňme, že předpokládáme, že matice nezávislých proměnných $\boldsymbol{X}$ resp. vektor závisle proměnných $\boldsymbol{y}$ jsou centrované. Centrování je nezbytné při odhadu (směrového) parametru $\boldsymbol{\beta}$. Definujme matici $\boldsymbol{W} = (\boldsymbol{e}_0, ..., \boldsymbol{e}_{n-1})^\dagger$. Dle lemmatu 6 řádky matice $\boldsymbol{W}$ tvoří ortonormální bázi v $\mathbb{C}^n$, proto $\boldsymbol{W}\boldsymbol{W}^\dagger = \boldsymbol{I}$ a také $\boldsymbol{W}^\dagger\boldsymbol{W} = \boldsymbol{I}$. Model (4.10) je transformací modelu (4.9) tvaru ($\boldsymbol{y}$ a $\boldsymbol{X}$ jsou centrované, proto lze přidat řádek s frekvencí $2\pi 0/n$)

$$\boldsymbol{W}\boldsymbol{y} = \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}e. \tag{4.11}$$

Pro OLS odhad transformavaného modelu (4.11) platí

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{FOLS} &= \left[(\boldsymbol{W}\boldsymbol{X})^\dagger(\boldsymbol{W}\boldsymbol{X})\right]^{-1}\left[(\boldsymbol{W}\boldsymbol{X})^\dagger(\boldsymbol{W}\boldsymbol{y})\right] = \left[\boldsymbol{X}^T\boldsymbol{W}^\dagger\boldsymbol{W}\boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}^T\boldsymbol{W}^\dagger\boldsymbol{W}\boldsymbol{y}\right] \\ &= \left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}^T\boldsymbol{y}\right] = \widehat{\boldsymbol{\beta}}_{OLS}.\end{aligned}$$

□

V článku Hidalgo (2003) (věta 2.1) je dokázána asymptotická vlastnost

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\right), \quad n \to \infty, \tag{4.12}$$

kde $\boldsymbol{\Omega} = 2\pi\int_{-\pi}^{\pi} f_x(\lambda)f_e(\lambda)d\lambda$ *a* f_x, f_e *jsou spektrální hustoty procesů* $\{\boldsymbol{x}_t\}$, $\{e_t\}$. *Ta platí za složitých obecných předpokladů, které však neomezují množinu modelů nekorelované stochastické regrese pouze na modely, jejichž chybové složky, případně regresory, se řídí procesy s krátkou pamětí. Předpoklady lze nalézt v práci Hidalgo (2003), str. 5 - 6.*

Pohledem na asymptotickou varianční matici lze soudit, že jednou z podmínek bude zřejmě také konečnost integrálu $\int_{-\pi}^{\pi} f_x(\lambda)f_e(\lambda)d\lambda$. Integrál je konečný, je-li integrand omezen výrazem $O(\lambda^{-1+\epsilon})$, $\lambda \to 0$ pro nějaké $\epsilon > 0$. Proto budou-li regresory i chyby procesy s dlouhou pamětí, je nutné, aby $d_e + d_{x_j} < 1/2$ pro $j = 1, ..., p$, kde d_e, resp. d_{x_j} je parametr d z definice 6 procesu chyb, resp. j-té složky procesu regresorů.

4.3 Frekvenční bootstrap

V předchozí části jsme dokázali větu o asymptotické nezávislosti vektoru diskrétních Fourierových transformací. Následně jsme definovali model ve frekvenční doméně a dokázali shodu OLS odhadu s OLS odhadem v modelu v časové doméně.

Aplikujme nyní dokázané vlastnosti na následující standardní problém regresní analýzy. Uvažujme model nezávislé lineární regrese, chtěli bychom učinit statistické závěry o vektoru parametrů $\boldsymbol{\beta}$ odhadnutého OLS metodou. Popišme nyní algoritmus frekvenčního bootstrapu, publikováno v Hidalgo (2003). Poznamenejme, že $\boldsymbol{y}$ a sloupce matice $\boldsymbol{X}$ a tedy i odhadnutá rezidua jsou stále centrované náhodné vektory.

Algoritmus 4 (Frekvenční bootstrap (FDB))**.**

1. *Odhadneme neznámé parametry* $\beta_1, ..., \beta_p$ *a získáme odhad reziduí*

$$\widehat{e}_t = y_t - \widehat{\beta}_1 x_{t1} - ... - \widehat{\beta}_p x_{tp}, \qquad t = 1, ..., n.$$

2. *Vypočteme Fourierovu transformaci* $\{w_{\widehat{e},j}\}$ *odhadnutých reziduí* $\{\widehat{e}_t\}$.
3. *Definujeme*

$$v_{\widehat{e},j} = w_{\widehat{e},j}/|w_{\widehat{e},j}|, \qquad j = 1, ..., \lfloor n/2 \rfloor$$

a z těchto vypočteme výběrový průměr a výběrový rozptyl

$$\overline{v}_{\widehat{e},j} = \frac{1}{\lfloor n/2 \rfloor} \sum_{j=1}^{\lfloor n/2 \rfloor} v_{\widehat{e},j}, \quad \widehat{\sigma}_v^2 = \frac{1}{\lfloor n/2 \rfloor} \sum_{j=1}^{\lfloor n/2 \rfloor} |v_{\widehat{e},j} - \overline{v}_{\widehat{e},j}|^2.$$

Nakonec definujeme standardizovaná rezidua

$$\widetilde{v}_{\widehat{e},j} = \frac{v_{\widehat{e},j} - \overline{v}_{\widehat{e},j}}{\widehat{\sigma}_v}, \qquad j = 1, ..., \lfloor n/2 \rfloor .$$

4. *Standardizovaná rezidua* $\{\widetilde{v}_{\widehat{e},j}\}$ *pro* $j = 1, ..., \lfloor n/2 \rfloor$ *použijeme jako základní populaci pro generování bootstrapových replik:*

$$P\left[\widetilde{v}^*_{\widehat{e},j} = \widetilde{v}_{\widehat{e},j}\right] = \frac{1}{\lfloor n/2 \rfloor}.$$

5. *Pomocí generovaných bootstrapových reziduí replikujeme regresní model ve spektrální doméně*

$$w^*_{y,j} = \widehat{\boldsymbol{\beta}}^T \boldsymbol{w}_{x,j} + \widetilde{v}^*_{\widehat{e},j} |w_{\widehat{e},j}|, \qquad j = 1, ..., \lfloor n/2 \rfloor .$$

6. *Odhadneme parametr* $\widehat{\boldsymbol{\beta}}^*$ *v novém bootstrapovém regresním modelu.*

Abychom porozuměli smyslu standardizace prováděné v algoritmu frekvenčního bootstrapu, vypočtěme první dva momenty bootstrapového odhadu parametru $\boldsymbol{\beta}$.

Věta 14.

$$\begin{aligned} E^*\widehat{\boldsymbol{\beta}}^* &= \widehat{\boldsymbol{\beta}} \\ var^*\widehat{\boldsymbol{\beta}}^* &= \widehat{var}\,\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\Sigma}}^{-1} \frac{1}{\lfloor n/2 \rfloor} \sum_{j=1}^{\lfloor n/2 \rfloor} \left(|w_{\widehat{e},j}|^2 \boldsymbol{w}_{x,j} \boldsymbol{w}^\dagger_{x,j} \right) \widehat{\boldsymbol{\Sigma}}^{-1} \end{aligned}$$

Důkaz: Platí: $\widehat{\boldsymbol{\Sigma}} = \frac{1}{\lfloor n/2 \rfloor} \boldsymbol{W}^\dagger_{x,r} \boldsymbol{W}_{x,r}$.

$$\mathrm{E}^*\widehat{\boldsymbol{\beta}}^* = \mathrm{E}^* \left[\widehat{\boldsymbol{\Sigma}}^{-1} \mathrm{Re} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{j=1}^{\lfloor n/2 \rfloor} \boldsymbol{w}_{x,j} w^*_{y,-j} \right) \right]$$

$$= \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Sigma}}^{-1}\mathrm{Re}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}\boldsymbol{w}_{x,j}\mathrm{E}^*\left[\widetilde{v}^*_{\widehat{e},-j}\right]|w_{\widehat{e},j}|\right) = \widehat{\boldsymbol{\beta}},$$

$$\begin{aligned}
\mathrm{var}^*\widehat{\boldsymbol{\beta}}^* &= \mathrm{E}^*\widehat{\boldsymbol{\Sigma}}^{-1}\frac{1}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}\boldsymbol{w}_{x,j}\widetilde{v}_{\widehat{e},-j}(\boldsymbol{w}_{x,j}\widetilde{v}_{\widehat{e},-j})^*|w_{\widehat{e},j}|^2\widehat{\boldsymbol{\Sigma}}^{-1} \\
&= \widehat{\boldsymbol{\Sigma}}^{-1}\frac{1}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}\left(|w_{\widehat{e},j}|^2\mathrm{E}^*|\widetilde{v}^*_{\widehat{e},-j}|^2\boldsymbol{w}_{x,j}\boldsymbol{w}^\dagger_{x,j}\right)\widehat{\boldsymbol{\Sigma}}^{-1} \\
&= \widehat{\boldsymbol{\Sigma}}^{-1}\frac{1}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}\left(|w_{\widehat{e},j}|^2\boldsymbol{w}_{x,j}\boldsymbol{w}^\dagger_{x,j}\right)\widehat{\boldsymbol{\Sigma}}^{-1}\frac{1}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}|\widetilde{v}_{\widehat{e},j}|^2 \\
&= \widehat{\boldsymbol{\Sigma}}^{-1}\frac{1}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}\left(|w_{\widehat{e},j}|^2\boldsymbol{w}_{x,j}\boldsymbol{w}^\dagger_{x,j}\right)\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{\mathrm{var}}\,\widehat{\boldsymbol{\beta}},
\end{aligned}$$

kde $\widehat{\mathrm{var}}\,\widehat{\boldsymbol{\beta}}$ je konzistentní odhad $\mathrm{var}\,\widehat{\boldsymbol{\beta}}$. □

Dokázali jsme, že bootstrapový odhad je nestranný a jeho rozptyl konzistentně odhaduje $\mathrm{var}\,\widehat{\boldsymbol{\beta}}$.

V práci Hidalgo (2003) (věta 2.2) je ukázáno, že i pro bootstrapový odhad $\widehat{\boldsymbol{\beta}}^*$ platí za složitých obecných předpokladů asymptotická normalita

$$P^*\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}\right) \le \boldsymbol{x}\right) \xrightarrow{P} \Gamma(\boldsymbol{x}), \quad n \to \infty \tag{4.13}$$

pro libovolné $\boldsymbol{x} \in \mathbb{R}^p$, kde Γ je distribuční funkce náhodného vektoru s rozdělením $N_p\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\right)$.

Algoritmus 4 předpokládá, že diskrétní Fourierovy transformace chyb $w_{e,1}, ..., w_{e,\lfloor n/2\rfloor}$ jsou asymptoticky nezávislé. Asymptotická vlastnost však platí pouze pro konečný počet DFT pevné délky m. Navíc jsme ukázali platnost této vlastnosti (věta 11) pouze pro procesy s krátkou pamětí. Byť obě tyto poznámky mohou naznačovat, že algoritmus selhává, bylo dokázáno, že rozdělení bootstrapového a původního odhadu parametru $\boldsymbol{\beta}$ mají za uvažovaných podmínek shodná asymptotická rozdělení.

Dosud jsme k získání statistických vlastností odhadů mohli užít buďto centrální limitní větu a silný zákon velkých čísel nebo aplikovat metodu blokový bootstrap. Ovšem metoda blokový bootstrap s sebou nese nutnost vhodné volby délky bloku. Právě představená metoda nevyžaduje znalost dalšího parametru.

Poznámka 10. Byla navržena alternativní metoda frekvenčního bootstrapu, která se od představené se liší v popsaném algoritmu v bodě *2*. Při této metodě se jako populace pro bootstrapové výběry používají samotná standardizovaná odhadnutá rezidua, nikoliv standardizované DFT odhadnutých reziduí. Fourierova transformace je provedena až na bootstrapových výběrech (Hidalgo (2003)). Bylo dokázáno, že asymptotická rozdělení takto získané bootstrapové repliky parametru $\boldsymbol{\beta}$ a původního odhadu parametru $\boldsymbol{\beta}$ jsou opět shodná. Simulační experimenty v práci Hidalgo (2003) naznačují téměř shodnou výkonnost obou FDB metod. My se soustředíme na metodu popsanou algoritmem 4, neboť možnost generování bootstrapových replik lze v tomto případě lépe zdůvodnit.

4.4 Aplikace metody frekvenční bootstrap

Pro ilustraci výkonnosti frekvenčního bootstrapu proveďme malou simulaci. Srovnejme krycí schopnosti intervalů spolehlivosti vytvořených na základě metody frekvenční bootstrap a asymptotického rozdělení. Na základě metody frekvenční bootstrap jsme odvodili dva typy intervalů, a to studentizované dané vztahem

$$\left(\widehat{\beta} + \sqrt{\widehat{\operatorname{var}}\,\widehat{\boldsymbol{\beta}}}\,\widetilde{H}_n^{-1}(\alpha/2), \widehat{\beta} + \sqrt{\widehat{\operatorname{var}}\,\widehat{\boldsymbol{\beta}}}\,\widetilde{H}_n^{-1}(1-\alpha/2)\right)$$

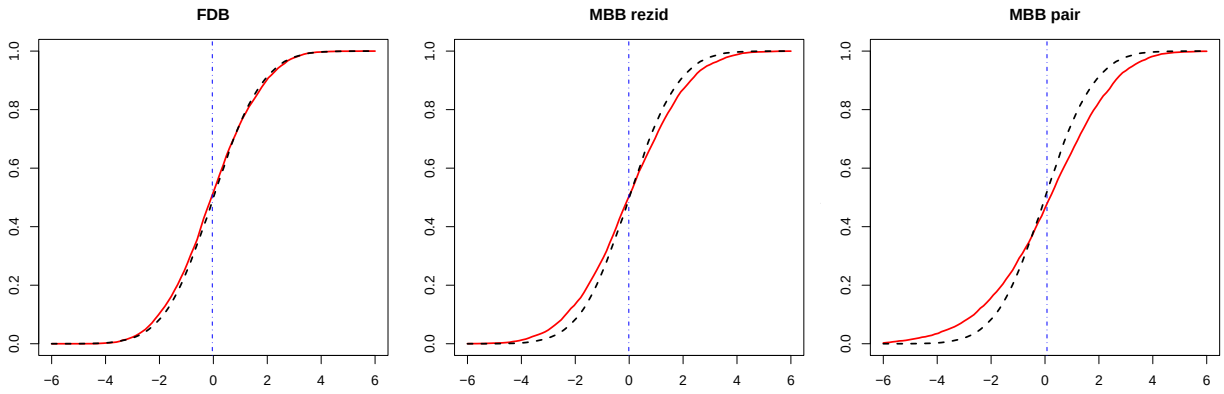
a percentilové

$$\left(G_n^{-1}(\alpha/2), G_n^{-1}(1-\alpha/2)\right).$$

Asymptotické intervaly jsou odvozeny z limitního rozdělení, kde je rozptyl odhadován výrazem $\widehat{\operatorname{var}}\,\widehat{\boldsymbol{\beta}}$ (viz výraz (4.12)). Studie je provedena na modelu nekorelované stochastické regrese, kde jsou chyby i regresory modelovány jako autoregresní proces prvního řádu s nezávislými stejně rozdělenými inovacemi.

Metoda *MBB rezid* bude značit aplikaci metody blokový bootstrap na centrovaná odhadnutá rezidua a následnou konstrukci bootstrapové repliky. Naopak v případě *MBB pair* bude bloková metoda aplikována přímo na množinu párů $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.

Uvědomme si, že aplikujeme-li blokovou metodu přímo na odhadnutá rezidua, obecně nezískáme ani nestranné odhady. Obrázek by mohl naznačovat, že aplikujeme-li metodu blokový bootstrap na odhadnutá rezidua, mají bootstrapové odhady menší rozptyl.



Obrázek 4.1: Bootstrapový odhad rozdělení statitistiky $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ jedné realizace jednoparametrického regresního modelu, kde regresory i chyby jsou generovány jako AR(1) s autoregresním koeficientem 0.3. Skutečné rozdělení odhadnuto na základě 20 000 simulací, $l = 10$.

Z tabulky 4.1 je jasně průkazná lepší krycí schopnost studentizovaných intervalů. Naopak percentilové intervaly vycházejí téměř ve všech případech hůře než asymptotické protějšky. Dále lze soudit, že kvalita intervalů spolehlivosti je choulostivější na silnou závislost mezi regresory než na závislost mezi chybami.

Simulací můžeme také srovnat vlastnosti odhadu momentů parametru $\boldsymbol{\beta}$. Pracujme se shodnými modely jako v předešlém a odhadujme rozptyl $n\operatorname{var}\widehat{\boldsymbol{\beta}}$. K tomuto odhadu lze přímo použít blokovou metodu.

			studentizované			percentilové			asymptotické		
n	β_x	β_e	99%	95%	90%	99%	95%	90%	99%	95%	90%
65	0.50	0.50	98.44	94.02	88.56	96.62	91.26	85.86	97.32	91.32	85.50
65	0.50	0.90	98.58	94.74	89.46	95.88	90.24	84.74	97.80	92.10	85.28
65	0.90	0.50	95.06	91.04	85.98	91.62	85.62	80.08	93.56	86.32	80.08
65	0.90	0.90	95.20	90.90	85.58	85.12	79.32	74.48	90.38	82.10	75.10
125	0.50	0.50	98.78	94.84	89.76	97.94	93.32	87.80	98.30	93.26	87.48
125	0.50	0.90	98.90	94.62	90.18	97.36	92.30	87.02	98.54	93.32	87.32
125	0.90	0.50	97.18	92.86	87.78	94.90	88.84	83.82	96.24	89.66	83.68
125	0.90	0.90	97.36	92.76	87.52	91.44	84.62	79.48	94.84	86.94	79.80

Tabulka 4.1: Průměrná krycí schopnost intervalů spolehlivosti v závislosti na velikosti a typu procesu. Srovnání dvou bootstrapových a asymptotické metody. Pro každý model provedeno 5 000 simulací a pro každou simulaci dále 2 000 bootstrapových výběrů.

				vychýlení			směr. odchylka			MSE		
n	β_x	β_e	var $\widehat{\beta}$	FDB	MBBr	MBBp	FDB	MBBr	MBBp	FDB	MBBr	MBBp
65	0.50	0.50	1.73	−0.18	−0.29	−0.40	0.74	0.58	0.85	0.57	0.42	0.89
65	0.50	0.90	7.57	−0.27	−1.67	−2.78	6.21	4.15	3.60	38.70	20.06	20.68
65	0.90	0.50	1.03	−0.19	−0.29	−0.20	0.66	0.46	0.58	0.48	0.30	0.37
65	0.90	0.90	8.21	−2.35	−4.20	−4.24	5.05	3.13	3.31	31.07	27.46	28.91
125	0.50	0.50	1.67	−0.08	−0.17	−0.23	0.58	0.40	0.69	0.34	0.19	0.53
125	0.50	0.90	8.85	−0.65	−1.71	−2.44	5.11	3.21	3.65	26.54	13.22	19.33
125	0.90	0.50	0.83	−0.10	−0.17	−0.13	0.45	0.29	0.40	0.22	0.11	0.18
125	0.90	0.90	8.58	−1.21	−3.78	−3.99	5.87	2.78	3.07	35.93	22.07	25.31

Tabulka 4.2: Vlastnosti odhadu $n\widehat{\text{var}}\,\widehat{\boldsymbol{\beta}}$ v závislosti na velikosti a typu procesu. Srovnání frekvenčního bootstrapu a blokových metod. Správné hodnoty odhadnuty na základě 2 000 simulací. K odhadům vlastností bootstrapových metod pak použito 500 simulačních experimentů a pro každý 500 bootstrapových výběrů, $l = 10$.

Z tabulky 4.2 lze vyčíst, že metoda frekvenční bootstrap jasně dominuje v minimalizaci vychýlení. Avšak odhady vykazují příliš vysokou variabilitu. Vyšší variabilita může vznikat také tím, že populace pro generování bootstrapových replik má rozsah pouze $\lfloor n/2 \rfloor$ pro frekvenční bootstrap, zatímco při MBB metodě je populace tvořena celkem $n - l + 1$ bloky. Celkově se metoda MBB rezid zdá být optimální z hlediska minimalizace střední čtvercové chyby a to pro celý průřez procesů uvažovaných v simulaci. Provedená simulace ukazuje vysokou variabilitu odhadu regresního parametru, jsou-li chyby mezi sebou silně korelované. Na korelaci regresorů variabilita příliš nezávisí.

5 Sítový bootstrap

Již víme, že metoda *sítový bootstrap* je založena na vhodné aproximaci výběru autoregresní posloupností. V následující kapitole popíšeme podrobněji celý algoritmus metody sítový bootstrap, uvedeme některé asymptotické výsledky a provedeme srovnání s blokovými metodami.

5.1 Princip a základní předpoklady

Nechť $\boldsymbol{X} = (X_1, ..., X_n)^T$ značí vektor n po sobě jdoucích náhodných veličin, které se řídí kauzálním lineárním procesem $\{X_t, t \in \mathbb{Z}\}$. Tento lze zapsat jako

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j Y_{t-j}, \quad t \in \mathbb{Z}, \tag{5.1}$$

kde $Y_t \sim WN(0, \sigma^2)$. Označme $\mu = \mathrm{E}\, X_t$. Nyní je nutné vektor $\boldsymbol{X}$ popsat (konečně rozměrným) modelem a odhadnout jeho parametry.

Pokud bychom vycházeli přímo z reprezentace (5.1), museli bychom kromě řádu modelu a parametrů $\{\psi_j\}$ odhadovat také neznámý bílý šum $\{Y_t\}$. Proto se přechází k procesům, jejichž parametry lze odhadnout snáze, a to k reprezentaci autoregresními posloupnostmi.

Definujme operátor zpětného posunutí $B : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$, pro který platí

$$BX_t = X_{t-1}, \qquad B^k X_t = B^{k-1}\left(BX_t\right) = X_{t-k}, \quad k \in \mathbb{Z}.$$

Dále označme mocninnou řadu

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j, \quad \psi_0 = 1,\ z \in \mathbb{C}. \tag{5.2}$$

Je-li řada (5.2) absolutně konvergentní v uzavřeném jednotkovém kruhu, má smysl operátor $\psi(B)$ a řadu (5.1) lze zřejmě zapsat jako

$$X_t = \mu + \psi(B) Y_t, \quad t \in \mathbb{Z}.$$

Autoregresní reprezentace procesu $\{X_t, t \in \mathbb{Z}\}$ musí být tvaru

$$Y_t = \varphi(B)\left(X_t - \mu\right), \quad t \in \mathbb{Z} \tag{5.3}$$

pro nějaký operátor $\varphi(B)$. Porovnáním ale vidíme, že musí platit $\varphi(z) = (\psi(z))^{-1}$ pro $|z| \leq 1$. Řadu $\varphi(z)$ rozepišme jako

$$\varphi(z) = 1 - \sum_{j=1}^{\infty} \varphi_j z^j, \quad z \in \mathbb{C},$$

což znamená, že jsme získali reprezentaci

$$X_t = \mu + \sum_{j=1}^{\infty} \varphi_j \left(X_{t-j} - \mu\right) + Y_t, \quad t \in \mathbb{Z}. \tag{5.4}$$

Definice 8. Stacionární proces $\{X_t, t \in \mathbb{Z}\}$ definovaný vztahem

$$X_t = \mu + \psi(B)Y_t, \quad t \in \mathbb{Z},$$

kde $Y_t \sim WN(0,1)$ nazveme *invertibilním*, pokud existuje řada $\varphi(z)$ taková, že $\varphi(z) < \infty$ pro $|z| \leq 1$ a

$$Y_t = \varphi(B)\left(X_t - \mu\right), \quad t \in \mathbb{Z}.$$

Z předchozích úvah vyplývá, že k úspěšné aplikaci metody sítový bootstrap nutně $\sum_{j=0}^{\infty} |\psi_j| < \infty$ a navíc požadujeme, aby řada $\psi(z)$ splňovala podmínku $|\psi(z)| > 0$ uvnitř jednotkového kruhu. Jelikož nelze pracovat s autoregresním procesem nekonečného řádu, ale tento aproximujeme pouze autoregresním procesem řádu konečného, je třeba některé předpoklady dále zesílit.

Definujme filtraci $\mathcal{F}_t = \sigma\{Y_s, s \leq t\}$. V článku Bühlmann (1997) jsou jako obecné nutné předpoklady úspěšné aplikace metody uvedeny následující.

P1) $\{Y_t, t \in \mathbb{Z}\}$ jsou nezávislé stejně rozdělené náhodné veličiny, $\mathrm{E}\, Y_1 = 0$ a $\mathrm{E}\,|Y_1|^s < \infty$ pro nějaké $s \geq 4$.

P2) Řada $|\psi(z)| > 0$ pro $|z| \leq 1$, $\sum_{j=0}^{\infty} j^r |\psi_j| < \infty$ pro nějaké $r \in \mathbb{N}$.

První část předpokladu P2 zaručuje invertibilitu procesu, jeho druhá část pak omezuje třídu procesů na procesy se slabou závislostí. Předpoklad splňují například modely typu ARMA(p, q). Z teoretického hlediska (stejně jako v případě délky bloku) je nutné, aby $p = p(n) \to \infty$ pro $n \to \infty$ a zároveň $p(n) = o(n)$.

Stěžejní pro přesnost metody je vhodná volba řádu modelu. Ukazuje se, že lze s výhodou použít *Akaikeho informační kritérium (AIC)*. Kritérium lze obecně zapsat jako

$$AIC(\widehat{\boldsymbol{\theta}}) = -2\ell(\widehat{\boldsymbol{\theta}}) + 2(p+1), \tag{5.5}$$

kde $\ell(\widehat{\boldsymbol{\theta}})$ je logaritmická věrohodnostní funkce, do které jsou dosazeny maximálně věrohodné odhady, a $p+1$ je počet parametrů modelu. Tyto odhady neznáme, proto se v praxi postupuje tak, že se do věrohodnostní funkce dosazují odhady získané např. metodou momentů (více viz dodatek). Obecně informační kritéria volí optimální model jako kompromis mezi počtem parametrů a přesností aproximace. Z výrazu (5.5) je vidět, jak kritérium AIC penalizuje počet parametrů.

Algoritmus 5 (Sítový bootstrap)**.**

1. *Aproximujeme výběr $\mathcal{X}$ autoregresním modelem řádu p. Řád p volíme na základě informačního (např. AIC) kritéria a odhadneme parametry μ a $\varphi_1, ..., \varphi_p$, které lze odhadnout metodou Yule-Walkerových rovnic, případně OLS metodou.*
2. *Získáme odhad reziduí*

$$\widehat{Y}_t = X_t - \overline{X}_n - \sum_{j=1}^{p} \widehat{\varphi}_j \left(X_{t-j} - \overline{X}_n \right), \qquad t = p+1, ..., n.$$

3. *Centrujeme rezidua $\widetilde{Y}_t = \widehat{Y}_t - \overline{\widehat{Y}}$, $t = 1, ..., n-p$, kde $\overline{\widehat{Y}} = \frac{1}{n-p} \sum_{t=p+1}^{n} \widehat{Y}_t$.*
4. *Centrovaná rezidua použijeme jako základní populaci pro generování bootstrapových reziduí $\{Y_t^*, t \geq 1\}$ tak, že*

$$P\left(Y_t^* = \widetilde{Y}_t\right) = \frac{1}{n-p}, \qquad t \geq 1.$$

5. *Zvolíme počáteční hodnoty $X_{1-p}^*, ..., X_0^*$ jako $\overline{X}_n$ a konstruujeme bootstrapovou řadu*

$$X_t^* - \overline{X}_n = \sum_{j=1}^{p} \widehat{\varphi}_j \left(X_{t-j}^* - \overline{X}_n \right) + Y_t^*, \qquad t \geq 1,$$

než dosáhneme stacionarity. Za bootstrapový výběr volíme stacionární úsek.

5.2 Asymptotické vlastnosti

Následující dvě věty shrnují základní vlastnosti asymptotického chování bootstrapového odhadu rozptylu a distribuční funkce výběrového průměru.

Věta 15. *Nechť platí předpoklady P1 a P2 s $r = 1$, nechť dále $p(n) = o\left([n/\log(n)]^{1/(2r+2)}\right)$. Potom*

$$var^*(\sqrt{n}\overline{X}_n^*) - var(\sqrt{n}\overline{X}_n) = o_P(1), \quad n \to \infty.$$

Jestliže navíc $\sum_{t_1,t_2,t_3} |\kappa_4(X_0, X_{t_1}, X_{t_2}, X_{t_3})| < \infty$, kde $\kappa_4(X_0, X_{t_1}, X_{t_2}, X_{t_3})$ značí čtvrtý kumulant veličin X_0, X_{t_1}, X_{t_2} a X_{t_3}, a platí-li předpoklad P2 s $r \geq 1$, pak

$$var^*(\sqrt{n}\overline{X}_n^*) - var(\sqrt{n}\overline{X}_n) = O_P((p/n)^{1/2}) + O_P(p^{-r}), \quad n \to \infty. \tag{5.6}$$

Důkaz: viz Bühlmann (1997), str. 136 - 141. □

Z předchozí věty plyne, že přesnost odhadu závisí na síle závislosti v původním procesu. Teoreticky to znamená následující. Nechť existuje $\delta > 0$ takové, že pro libovolné $\kappa \in (\delta, 1/2)$ existuje r

$$r \in \left(\frac{1}{2\kappa} - 1, \infty \right), \quad r \in \mathbb{N}$$

splňující předpoklad P2. Pak volbou $p(n) = Cn^{1/(2r+2)} \log(n)^{-1/(2r+2)-1}$, kde $C \in (0, \infty)$, a dosazením r do výrazu pro $p(n)$

$$(p/n)^{1/2} \approx n^{-1/2+\kappa}, \quad p^{-r} \approx n^{-1/2+\kappa}, \qquad \kappa \in (\delta, 1/2)$$

Proto vzhledem k výrazu (5.6) platí pro $\kappa \in (\delta, 1/2)$

$$\text{var}^*(\sqrt{n}\overline{X}_n^*) - \text{var}\,(\sqrt{n}\overline{X}_n) = O_P(n^{-1/2+\kappa}), \quad n \to \infty.$$

Nyní se zaměříme na aproximaci rozdělení centrované statistiky.

Věta 16. *Nechť platí předpoklady P1 a P2 s* $r = 1$*, nechť dále* $p(n) = o\left([n/\log(n)]^{1/4}\right)$. *Jestliže*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N\left(0, \sum_{k=-\infty}^{\infty} R(k)\right), \quad n \to \infty,$$

pak

$$\sup_{x \in \mathbb{R}} \left| P^*\left(\sqrt{n}(\overline{X}_n^* - \overline{X}_n) \le x\right) - P\left(\sqrt{n}(\overline{X}_n - \mu) \le x\right) \right| = o_P(1), \quad n \to \infty.$$

Důkaz: viz Bühlmann (1997), str. 136 - 143. □

Vidíme, že metoda konzistentně odhaduje distribuční funkci centrovaného výběrového průměru. Stejně jako tomu bylo v případě blokových metod, lze platnost věty za dodatečných předpokladů rozšířit na centrované statistiky typu hladkých funkcí výběrového průměru.

Srovnání s metodou blokový bootstrap

Při aplikaci blokové metody je nejprve nutné vhodně určit délku bloku. Přesnost bootstrapových odhadů pak na této volbě velmi silně závisí. Uvedli jsme algoritmy pomocí nichž lze vhodnou délku odhadnout, avšak, jak se ukázalo v příkladu, metody ne vždy fungují zcela podle očekávání. V případě metody sítový bootstrap je nejprve nutné odhadnout řád autoregresního procesu, ale přesnost metody nezávisí na řádu modelu tak silně.

Pokud se nám při aplikaci blokové metody podaří vhodně určit délku bloku, dosáhneme při odhadu rozptylu nejvýše přesnosti řádu $O_P(n^{-1/3})$, nezávisle na síle závislosti v původním výběru. Naproti tomu rychlost aproximace metody sítový bootstrap souvisí s touto závislostí, jak je ukázáno v diskusi za větou 15. Navíc jde pro velkou množinu procesů o přesnější aproximaci. Bylo provedeno několik simulačních experimentů s cílem srovnat metodu blokový, sítový a frekvenční bootstrap, případně také s asymptotickými metodami, viz kapitola 6.

6 Aplikace

6.1 Simulace

Numerické výsledky simulací jsou uvedeny v příloze.

Regresní modely

Simulace byla provedena pro regresní modely R1 - R6, kde se vektor regresorů $\boldsymbol{x}$ a vektor chyb $\boldsymbol{e}$ řídí následujícími stacionárními procesy.

Ve všech modelech jsou regresory generovány jako autoregresní proces prvního řádu

$$X_t = \varphi_1 X_{t-1} + Y_t,$$

kde $\varphi_1 = 0.95$ pro model R3, $\varphi_1 = 0.2$ pro zbylé. $\{Y_t\}$ jsou nezávislé stejně rozdělené náhodné veličiny, $Y_1 \sim N(0,1)$ pro R1 - R5. V modelu R6 jsou náhodné veličiny Y_t modelovány procesem s dlouhou pamětí FARIMA(0, d, 0), modelem

$$Y_t = (1-B)^{-d}\epsilon_t, \qquad t \in \mathbb{Z},$$

kde $\{\epsilon_t\}$ jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0,1)$ a $d = 0.4$.

Chyby jsou pro modely R1 - R3 generovány jako autoregresní posloupnosti prvního řádu s parametrem $\varphi_1 = 0.95$ pro model R2 a $\varphi_1 = 0.2$ pro zbylé. V případě modelů R4 - R6 pak jako ARMA(1, 1) posloupnosti

$$X_t = 0.2X_{t-1} + \psi_1 Y_{t-1} + Y_t, \qquad t \in \mathbb{Z},$$

kde $\psi_1 = 0.9$ a Y_t jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0,1)$ pro model R4. Pro modely R5 a R6 $\psi_1 = 0.5$ a Y_t se řídí FARIMA(0, d, 0) procesem s parametrem $d = 0.45$ resp. $d = 0.05$ pro model R6.

Tabulka (B.1) shrnuje krycí schopnosti studentizovaných bootstrapových intervalů a intervalů asymptotických pro modely R1 - R6. V modelu R1, kde regresory i chyby vykazují slabou závislost, výkonnost metod sítový i frekvenční bootstrap mírně převyšuje asymptotický přístup. V ostatních případech je lepší krycí schopnost již znatelnější. Zvýšíme-li sílu závislosti v procesu chyb, intervaly spolehlivosti konstruované metodou sítový bootstrap pokrývají správnou hodnotu s pravděpodobností vyšší než jakou je nominální hodnota. Naopak frekvenční bootstrap poskytuje kratší intervaly.

Tabulka dále naznačuje náchylnost metody frekvenční bootstrap na závislost mezi regresory. Avšak stále udržuje krycí schopnost intervalů lépe než klasická aproximace normálním rozdělením. Toto zjištění je ve shodě se simulací provedenou v kapitole 4 a také s výsledky

publikovanými v článku Hidalgo (2003). V modelu R4, kde jsou chyby generovány jako ARMA proces, metoda sítový bootstrap mírně ztrácí na metodu frekvenční bootstrap. Je překvapující, jak dobrých výsledků dosahuje metoda sítový bootstrap v modelech, kde jsou regresory a chyby generovány procesem s dlouhou pamětí. V těchto mírně převyšuje metodu frekvenční bootstrap a výrazně asymptotický přístup. Zvýšení počtu pozorování nevede k výrazným změnám ve výkonosti intervalů.

Prvním pohledem na tabulku (B.2) vidíme, že metoda frekvenční bootstrap zaostává v měřítku střední čtvercové chyby za metodami blokový i sítový bootstrap. Metoda frekvenční bootstrap minimalizuje vychýlení, jehož výrazné zlepšení oproti ostatním metodám můžeme vidět u modelů, kde jsou chyby silně závislé (zejména v modelech R2 a R5). Avšak odhady mají velmi vysokou variabilitu, což odpovídá výsledkům získaným v kapitole 4. V modelech R1 a R2 dosahuje metoda sítový bootstrap nejlepších výsledků. Je však nutné brát tento výsledek s rezervou, neboť v těchto modelech jsou chyby generovány jako autoregresní posloupnosti prvního řádu. Zajímavější je srovnání pro komplexnější modely. Ukazuje se, že metody blokový a sítový bootstrap dávají srovnatelné výsledky z hlediska minimalizace střední čtvercové chyby. Uvědomme si, že simulaci provádíme pro optimální délku bloku, která je určena minimalizací střední čtvercové chyby.

Srovnání algoritmů pro optimální délku bloku

Již v kapitole 3 jsme srovnali výkonnost obou představených algoritmů. Studujme vlastnosti algoritmů pro nelineární statistiku, pro jiné typy procesů a nakonec pro odhad celé distribuční funkce. Definujme modely

$$\begin{aligned}
\mathrm{D1}: X_t &= 0.6X_{t-1} + Y_t, \quad t = 1, ..., 125, \\
\mathrm{D2}: X_t &= \frac{\sqrt{2}}{2}\left(Y_t + Y_{t-1}\right), \quad t = 1, ..., 100, \\
\mathrm{D3}: X_t &= 0.3X_{t-1} + 0.2X_{t-2} + 0.7Y_{t-1} + Y_t, \quad t = 1, ..., 125,
\end{aligned}$$

kde $\{Y_t\}$ jsou nezávislé stejně rozdělené náhodné veličiny s centrovaným χ^2 rozdělením o jednom stupni volnosti pro modely D1 a D2 a s rozdělením $N(0, 1)$ pro model D3. Model D1 je uveden pro srovnání s příkladem v textu, kde jsme studovali rozptyl výběrového průměru. Model D2 ilustruje chování algoritmů pro procesy s odlišnou autokovarianční funkcí, zvolen také pro srovnání s výsledky v pracích Hall (1992) a Lahiri (2003). Nakonec model D3 je představitelem komplexnějšího modelu. Výsledky jsou shrnuty v tabulkách (B.3), (B.4), (B.5) a (B.6). Hvězdičkami jsou označeny délky bloků, pro které není střední čtvercová chyba výrazně odlišná od minimální (v tabulce (3.1) bychom mohli volit délky 6, 7 a 8).

Tabulka (B.3) ukazuje, že i pro nelineární statistiky, dávají metody výsledky blízké optimu. Zdá se, že parametrická metoda mírně podhodnocuje optimální délku, naopak metoda založená na minimalizaci MSE je nestabilní. Generujeme-li data modelem D2, pak median navržených délek pro odhad rozptylu algoritmem MSE nabývá hodnoty 3, což se ukázalo být jako optimální délka. Tento model byl použit v práci Hall (1992) a také v Lahiri (2003) s velikostí vzorku 125 namísto 100 pozorování. Výsledky odpovídají uvedeným v Lahiri (2003). Nakonec pozorujeme výrazně podhodnocené délky bloků pro model D3 a to při aplikaci obou metod.

Srovnání metod blokový a sítový bootstrap

Provedeme 4 sady simulací a to postupně pro modely klouzavých součtů, ARMA(1, 1) modely a nakonec pro modely vymykající se množině, pro které je dokázána platnost z teoretického hlediska. Ve čtvrté části srovnáme kvalitu aproximace celé distribuční funkce.

Modely klouzavých součtů

V první části srovnejme kvality bootstrapových odhadů v následujících invertibilních modelech klouzavých součtů.

$$\begin{aligned}
\mathrm{C(MA)1} : X_t &= 0.95Y_{t-1} + Y_t, \\
\mathrm{C(MA)2, C(MA)4} : X_t &= \sum_{j=1}^{30} \varphi_j Y_{t-j} + Y_t, \quad \varphi_j = (-1)^{j+1}\frac{1}{j+1}, \\
\mathrm{C(MA)3, C(MA)5} : X_t &= \sum_{j=1}^{30} \varphi_j Y_{t-j} + Y_t, \quad \varphi_j = (-1)^{j+1}\frac{1}{(j+1)^2},
\end{aligned}$$

kde $\{Y_t\}$ jsou nezávislé stejně rozdělené náhodné veličiny, $Y_1 \sim N(0,1)$ pro modely C(MA)1, C(MA)2 a C(MA)4 a Y_1 má centrované χ^2 rozdělení o jednom stupni volnosti pro modely C(MA)3 a C(MA)5.

Odhadujeme-li rozptyl výběrového průměru, metoda sítový bootstrap dominuje v minimálizaci vychýlení, které s rostoucí velikostí vzorku klesá rychleji než v případě blokové metody. Blokové metody lze pokládat v tomto případě za kvalitnější z hlediska minimalizace střední čtvercové chyby. Přesně opačně je tomu v případě odhadu rozptylu mediánu. Je překvapivé, že metoda sítový bootstrap dosahuje lepšího výsledku pro model C(MA)1.

ARMA(1,1) modely

Ve druhé části pak studujme chování ve stacionárních a invertibilních ARMA(1, 1) modelech.

$$\begin{aligned}
\mathrm{C(ARMA)1} : X_t &= 0.3X_{t-1} + 0.4Y_{t-1} + Y_t, \\
\mathrm{C(ARMA)2} : X_t &= 0.7X_{t-1} + 0.5Y_{t-1} + Y_t, \\
\mathrm{C(ARMA)3} : X_t &= -0.7X_{t-1} + 0.5Y_{t-1} + Y_t, \\
\mathrm{C(ARMA)4} : X_t &= 0.2X_{t-1} + 0.5Y_{t-1} + Y_t, \\
\mathrm{C(ARMA)5} : X_t &= -0.2X_{t-1} + 0.5Y_{t-1} + Y_t,
\end{aligned}$$

kde $\{Y_t\}$ jsou nezávislé stejně rozdělené náhodné veličiny, $Y_1 \sim N(0,1)$.

Opět pozorujeme stejnou tendenci minimalizace MSE ve srovnání kvality odhadu rozptylu robustní statistiky a výběrového průměru. Pro modely C(ARMA)3 a C(ARMA)5 je autokovarianční funkce periodická, viz Prášková (2004b), str. 69. Pro tyto dle Bühlmann (2002), str. 133 - 134, jsou autoregresní aproximace velmi spolehlivé. Proto pro tyto procesy dosahuje odhad metodou sítový bootstrap výrazně lepších výsledků nežli odhad blokovou metodou. Výsledky simulace odpovídají těm z článku Bühlmann (2002).

Neinvertibilní a heteroskedastické modely

Nakonec se zabývejme modely, pro něž věty o asymptotickém chování neplatí, případně z teoretického hlediska nedávají konzistentní odhady. Prvními budou neinvertibilní modely, dále pak nestacionární model.

$$\begin{aligned} \mathrm{C(OUT)1, C(OUT)2} : X_t &= Y_{t-1} + Y_t, \\ \mathrm{C(OUT)3} : X_t &= 0.5X_{t-1} + 0.4Y_{t-1} + Y_t, \end{aligned}$$

kde $\{Y_t\}$ jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0,1)$ pro model C(OUT)1 respektive s χ^2 rozdělením o jednom stupni volnosti pro model C(OUT)2. V modelu C(OUT)3 jsou chyby heteroskedastické s rozdělením $N(0, \sigma_t^2)$, kde $\sigma_t^2 = 1 + \frac{1}{2}(-1)^t$.

Celkově pozorujeme větší odolnost metody blokový bootstrap proti nesplnění předpokladů.

Odhad celého rozdělení

Poslední simulace se zaměřuje na odhad celé distribuční funkce. Jednou možností je odhadnout skutečné rozdělení jednotlivých statistik a na základě jedné realizace pak zkoumat aproximaci bootstrapovými rozděleními. Avšak tato rozdělení mohou být značně vychýlená důsledkem jediné realizace. Jinou možností je srovnávat zachování nominální krycí schopností intervalů spolehlivosti. Protože jsme však již detailně studovali kvalitu odhadu rozptylu, srovnejme kvalitu odhadu dalších charakteristik rozdělení, a to šikmosti a špičatosti. Poznamenejme, že právě v odhadu šikmosti obecně dochází ke zlomu mezi kvalitou aproximace normálním rozdělením a rozdělením bootstrapovým. Abychom mohli tyto charakteristiky lépe porovnat, pracujme s výrazně šikmým rozdělením statistiky $T_n = e^{\overline{X}_n}$ respektive $T_n = e^{\mathrm{med}(\boldsymbol{x})}$.

V případě šikmosti metoda sítový bootstrap dominuje pro modely klouzavých součtů, to je pro modely s velmi slabou závislostí. Pro ostatní jsou pak výsledky srovnatelné. Podobné výsledky vidíme také v tabulce B.14. V modelu C(ARMA)2 však metoda sítový bootstrap dosahuje horších výsledků, které jsou způsobeny velkým rozptylem.

6.2 Reálná data

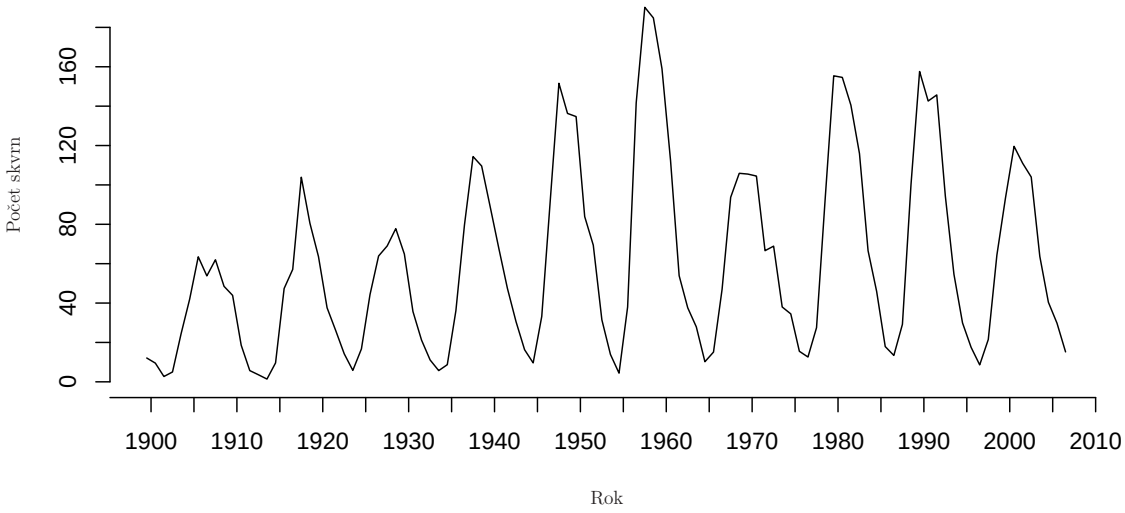
V této části ukážeme, jak lze některé metody bootstrap pro závislá data uplatnit v praktických příkladech. Prvním bude rozbor časové řady udávající počet skvrn na slunečním povrchu.

Sluneční skvrny

Již od 17. století se sleduje počet skvrn na povrchu Slunce. Tento počet vykazuje periodicitu s periodou okolo 11 let. Pokud bychom chtěli učinit závěry o střední hodnotě, případně charakteristice, kterou lze zapsat jako hladkou funkci střední hodnoty, potřebujeme zejména odhad rozptylu.

Aplikujme postupně metody blokový a sítový bootstrap k odhadu rozptylu výběrového průměru. V tabulce 6.1 je vidět, jak se mění odhad rozptylu s rostoucí délkou bloku. Ukazuje se, že v případě periodické řady, může nesprávná volba délky bloku výrazně zkreslit odhady. Délka bloku by měla odpovídat délce periody případně jejím násobkům.

Zdroj dat: webová stránka http://sidc.oma.be/.



Obrázek 6.1: Počet slunčních skvrn v letech 1900 - 2006.

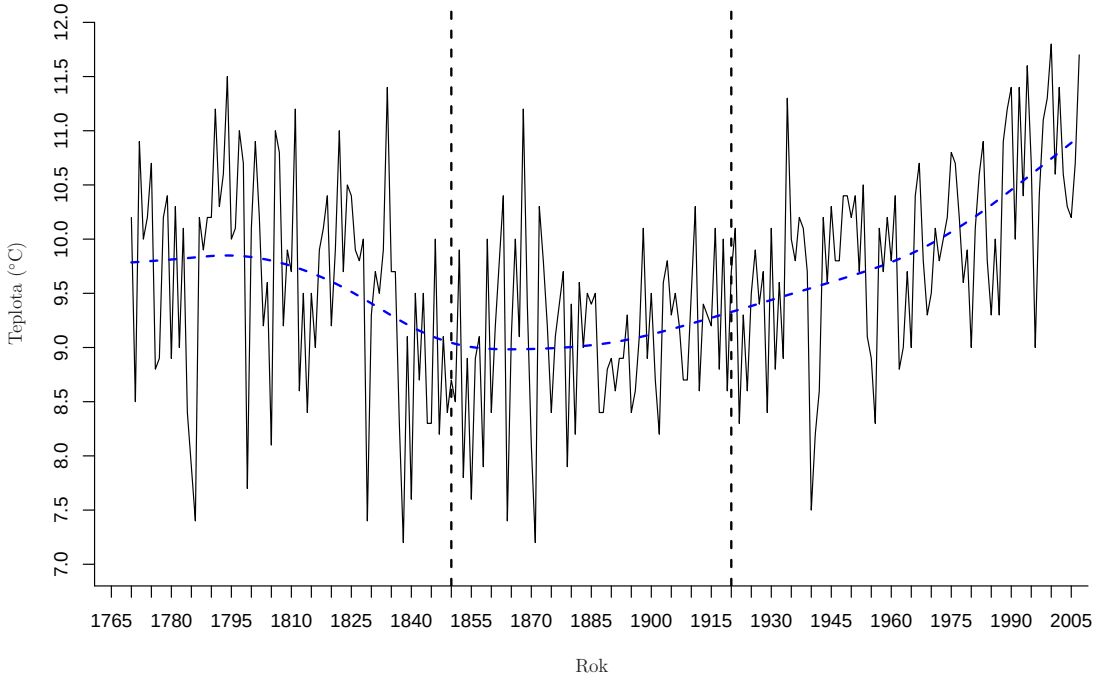
	Odhady blokovou metodou v závislosti na délce bloku					
l	4	5	6	7	8	9
$n\widehat{\mathrm{var}}(\overline{X}_n)$	6480.053	5562.953	4577.344	4479.054	4192.805	3216.095
l	10	11	12	13	14	15
$n\widehat{\mathrm{var}}(\overline{X}_n)$	3433.967	3635.143	3940.507	4827.763	5034.791	5973.273

Tabulka 6.1: Odhady rozptylu výběrového průměru počtu Slunečních skvrn v letech 1900 - 2006. Při metodě sítový bootstrap zvolen autoregresní model třetího řádu s parametry $\widehat{\varphi}_1 = 1.1423991$, $\widehat{\varphi}_2 = -0.2545654$ a $\widehat{\varphi}_3 = -0.3180392$. Odhad pak $n\widehat{\mathrm{var}}(\overline{X}_n) = 2074.807$.

Průměrné roční teploty

Druhým datovým souborem jsou průměrné roční teploty naměřené v pražském Klementinu od roku 1750 do současnosti.

Máme podezřezní na růst průměrné roční teploty. Řada znárorněná na obrázku 6.2 naznačuje, že v minulém století skutečně docházelo k růstu průměrné roční teploty. Zvyšování teploty není lineární, zdá se dokonce, že strmější růst nastává zhruba okolo roku 1920. Podobný problém je řešen v knize Davison a Hinkley (1997), příklad 8.6.



Obrázek 6.2: Průměrné roční teploty naměřené v pražském Klementinu od roku 1770 do roku 2007 (průměrná teplota v roce 2007 doplněna na základě dat na stránkách českého hydrometeorologického ústavu). Pro znázornění trendu byl řadou proložen kubický spline. V grafu jsou dále znázorněny subjektivně zvolené intervaly (1850 - 1920) a (1920 - 2007), pro které budeme testovat existenci monotónního trendu.

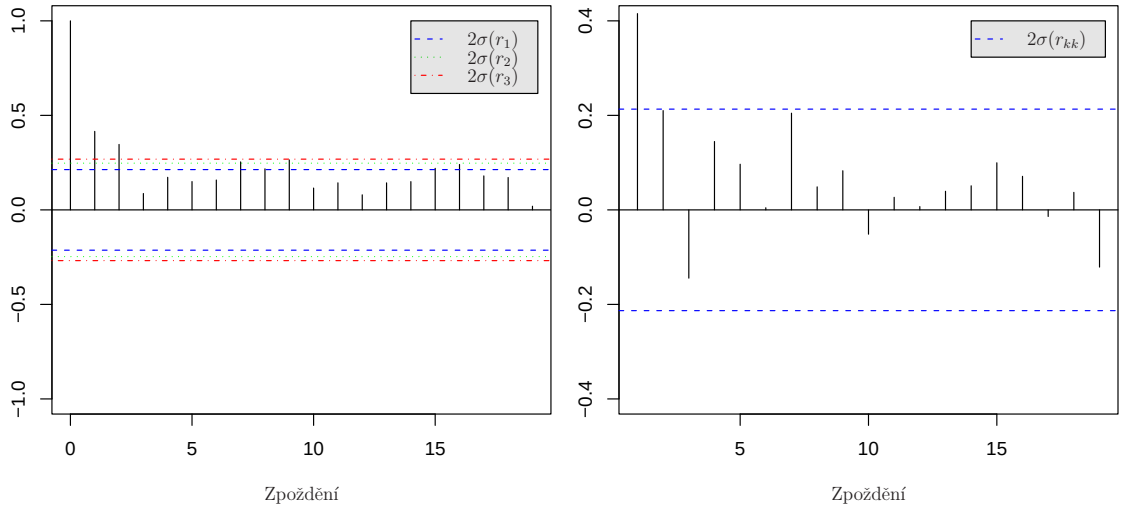
Na obrázku 6.3 je zobrazena struktura autokorelací a parciálních autokorelací pro časový úsek řady 1920 - 2007, ze které vyplývá, že řada by mohla být popsána modelem klouzavých součtů druhého řádu. To znamená, že lze užít metodu blokový i sítový bootstrap.

Náš model pro časový úsek 1920 - 2007 můžeme zapsat jako

$$X_t = \mathrm{Tr}_t + Y_t, \qquad t = 1, ..., n,$$

kde $\{\mathrm{Tr}_t\}$ je trendová složka a $\{Y_t\}$ stacionární posloupnost. Často se předpokládá například lineární trend tvaru $\mathrm{Tr}_t = \alpha + \beta t$, odhadnou se koeficienty a na základě odhadnutých koeficientů se učiní statistické závěry. Při konstrukci statistik je nutné mít na paměti, že chyby $\{Y_t\}$ jsou korelované. My bychom však rádi testovali přítomnost obecného monotónního trendu, neboť pohledem na obrázek 6.2 nelze tvrdit, že se jedná o ryze lineární trend. V práci Brillinger (1989) je odvozen test na ověřování přítomnosti monotónního trendu. Testujme hypotézu H0 proti alternativě H1, kde

$$\begin{aligned} \mathrm{H0}: &\quad \mathrm{Tr}_t \text{ je konstantní,} \\ \mathrm{H1}: &\quad \mathrm{Tr}_t \text{ je rostoucí.} \end{aligned}$$



Obrázek 6.3: Autokorelogram (vlevo) a parciální autokorelogram. $\sigma(r_k)$ respektive $\sigma(r_{kk})$ značí odhad směrodatné odchylky výběrové autokorelace (Bartlettova aproximace) respektive parciální výběrové autokorelace (Quenoueillova aproximace).

Definujme koeficienty

$$c_j = \left[(j-1)\left(1-\frac{j-1}{n}\right)\right]^{1/2} - \left[j\left(1-\frac{j}{n}\right)\right]^{1/2}, \quad j = 1, ..., n$$

a testovou statistiku výrazem $T_n = \sum_{t=1}^{n} c_t X_t$. Rozdělení testové statistiky je za nulové hypotézy (konstantní trend) blízké normálnímu a vzhledem k symetrii koeficientů c_j je $\mathrm{E}\, T_n = 0$ za platnosti H0. Odhadněme proto rozptyl statistiky T_n a p-hodnoty. Za hypotézy H0 platí, že T_n má asymptoticky rozdělení $N(0, \operatorname{var} T_n)$, proto lze p-hodnoty odhadnout jako $\mathrm{P}(X > T_n)$, kde $X \sim N(0, \widehat{\operatorname{var}} T_n)$. Výsledky jsou uvedeny v tabulce 6.2.

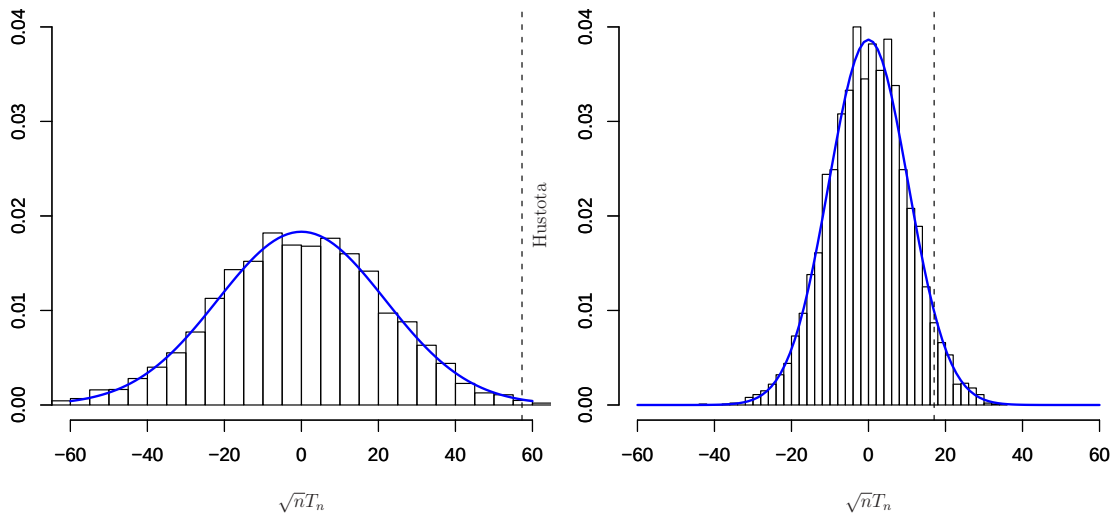
	1920-2007		1850-1920	
	Blokový	Sítový	Blokový	Sítový
$\widehat{\operatorname{var}} T_n$	2.078793	2.403175	1.127121	1.269317
$\mathrm{P}(X > T_n)$	0.001654528	0.005527345	0.036345764	0.055501208

Tabulka 6.2: Odhady rozptylu statistiky T_n a p-hodnoty testu. Pro blokový bootstrap zvolena délka bloku metodou založenou na tvorbě podvýběrů ($l = 6$). V případě metody sítový bootstrap navržen model druhého řádu s parametry $\widehat{\varphi}_1 = 0.3280688$ a $\widehat{\varphi}_2 = 0.2100961$ pro data z let 1920-2007, pro průměrné teploty v letech 1850-1920, pak model prvního řádu s parametrem $\widehat{\varphi}_1 = -0.1183577$.

Metodou sítový bootstrap můžeme odhadnout celé rozdělení statistiky T_n. Z obrázku 6.4 vyplývá, že statistika má skutečně symetrické rozdělení blízké normálnímu. Pro srovnání uveďme, že percentilové p-hodnoty definované vztahem $1 - G^*(T_n)$ nabývají hodnot 0.0040 pro teploty z let 1920 - 2007 respektive 0.0512 pro pozorování z období 1850 - 1920.

Na základě všech výsledků lze shodně prohlásit, že na hladině 1 % zamítáme konstantní trend průměrných teplot v letech 1920 - 2007 ve prospěch trendu rostoucího. Pro období 1850 - 1920 již nelze ve shodě všech metod ani na 5 % hladině tvrdit, že zamítáme hypotézu konstatního trendu průměrných teplot proti alternativě rostoucího trendu.

Data byla stažena z webové stránky http://zmeny-klima.ic.cz/, kde je autor stránky převzal z knihy Václav Cílek, Jiří Svoboda, Zdeněk Vašků: Velká kniha o klimatu Zemí koruny české.



Obrázek 6.4: Rozdělení statistiky na základě řady z let 1920 - 2007 vlevo a z let 1850 - 1920 vpravo. Plnou čarou hustota normálního rozdělení s odhadnutým rozptylem, čárkovaně hodnoty statistik.

Finanční časové řady

Zvažujeme-li výhodnost investice, potřebujeme spolehlivé odhady budoucích hodnot. Mějme akciový index a ptejme se, jak modelovat vývoj ceny akcie a jak tento model použít k predikci.

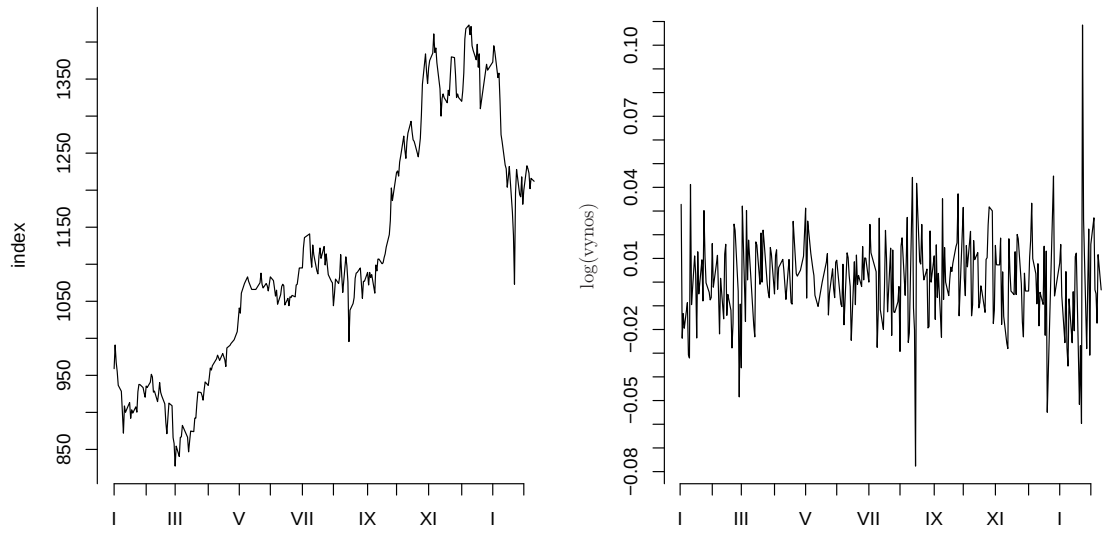
Pracujme s akciemi společnosti ČEZ z období 1.1.2007 - 11.2.2008. Denní uzavírací ceny akcií byly staženy z internetových stránek společnosti ČEZ (http://www.cez.cz/edee/cs/akcie/akcie.jsf/).

Označme P_t cenu akcie v čase t, předpokládejme, že dynamiku vývoje ceny P_t lze popsat rovnicí

$$P_{t+1} - P_t = \alpha(t, P_t)P_t + \sigma(t, P_t)P_t Z_t,$$

kde členy $\alpha(t, P_t)$ a $\sigma(t, P_t)$ jsou neznámé funkce a Z_t nezávislé náhodné veličiny s nulovou střední hodnotou a jednotkovým rozptylem. Předpokládejme, že drift $\alpha(t, P_t)$ je konstantní a volatilita $\sigma(t, P_t)$ je funkcí ceny akcie.

Pro další analýzu se uvažují buďto relativní změny ceny $\frac{P_{t+1}-P_t}{P_t}$ nebo logaritmy výnosů $\log \frac{P_{t+1}}{P_t}$. Používejme dále logaritmickou transformaci a označme $X_t = \log \frac{P_{t+1}}{P_t}$. Takto transformavané ceny akcií se chovají jako bílý šum, jak je vidět i na obrázku 6.6, avšak s tzv. podmíněnou heteroskedasticitou. Tu lze modelovat např. pomocí GARCH modelu.



Obrázek 6.5: Vývoj ceny akcie společnosti ČEZ od počátku roku 2007 do 11.2.2008. Na pravém grafu pak logaritmy výnosů.

Definice 9. Řekneme, že proces $\{\epsilon_t, t \in \mathbb{Z}\}$ je GARCH(p, q) proces, platí-li

$$\begin{aligned} \mathrm{E}\,[\epsilon_t|\mathcal{F}_{t-1}] &= 0, \\ \mathrm{var}\,[\epsilon_t|\mathcal{F}_{t-1}] &= \sigma_t^2, \\ \sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \end{aligned}$$

kde $\mathcal{F}_t = \sigma\{\epsilon_s, s \leq t\}$.

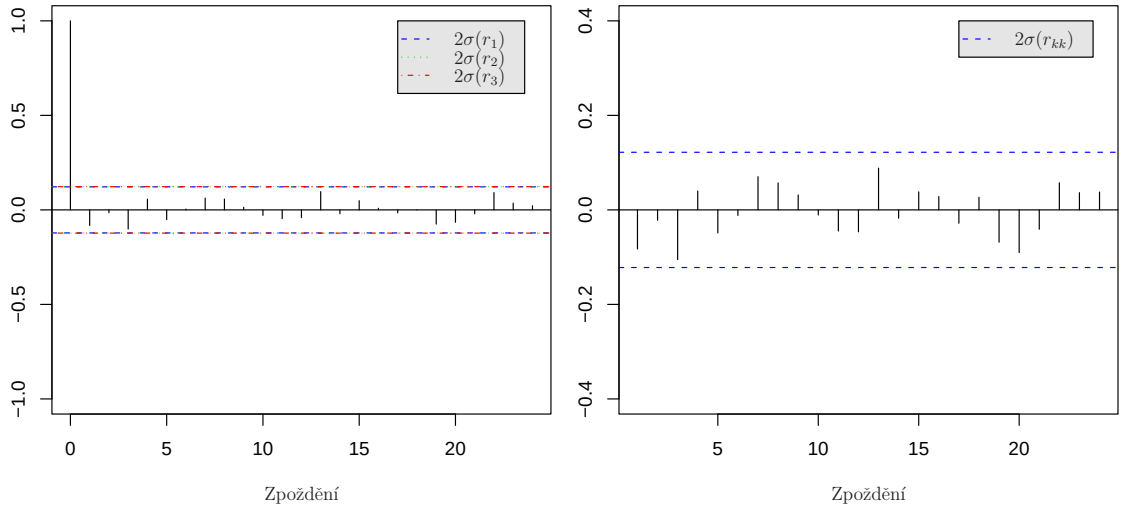
Proces $\{\epsilon_t^2, t \in \mathbb{Z}\}$ lze za jistých podmínek reprezentovat jako ARMA proces, proto se k identifikaci GARCH modelu využívají autokorelogramy a parciální autokorelogramy procesu $\{\epsilon_t^2, t \in \mathbb{Z}\}$. Na obrázku 6.7 vidíme, že první člen odhadnuté autokorelační a parciální autokorelační funkce v tomto novém procesu již vychází významný, což svědčí o podmíněné heteroskedasticitě.

S ohledem na obrázek (6.7) volme k popisu procesu logaritmů výnosů $\{X_t\}$ model GARCH(1,1). Proces $\{X_t\}$ popíšeme modelem

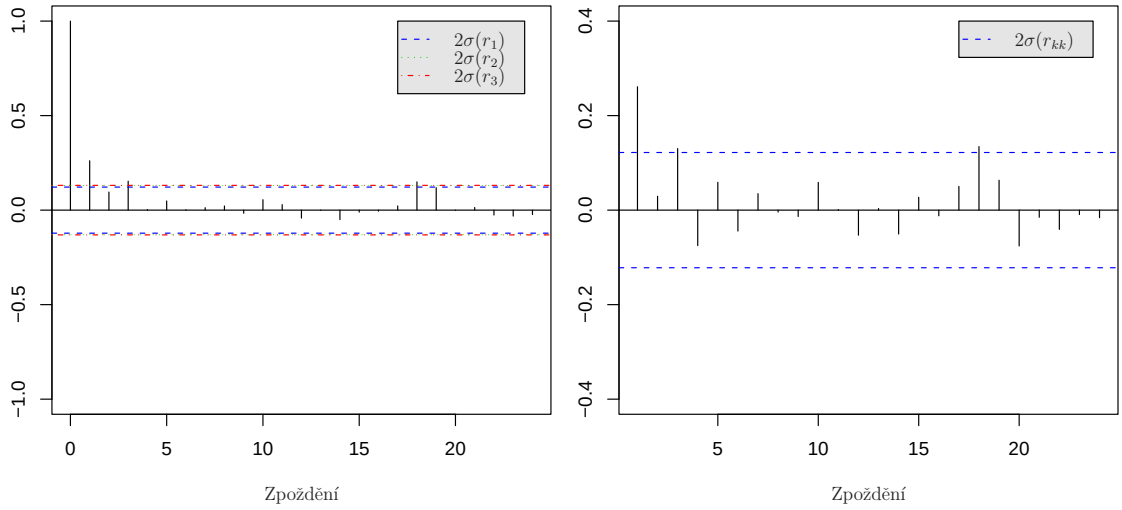
$$\begin{aligned} X_t &= \mu + \epsilon_t, \\ \epsilon_t &= \sigma_t Z_t, \\ \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \qquad t = 1, ..., 269, \end{aligned}$$

kde Z_t jsou nezávislé stejně rozdělené náhodné veličiny s nulovou střední hodnotou a jednotkovým rozptylem. Zapsaný model je mírně zobecněný GARCH model, kde chyby $\{\epsilon_t\}$ se ve shodě s definicí 9 řídí GARCH(1,1) procesem, neboť

$$\begin{aligned} \mathrm{E}\,[\epsilon_t|\mathcal{F}_{t-1}] &= \sigma_t \mathrm{E}\, Z_t = 0, \\ \mathrm{var}\,[\epsilon_t|\mathcal{F}_{t-1}] &= \sigma_t^2 \mathrm{E}\, Z_t^2 = \sigma_t^2 \end{aligned}$$



Obrázek 6.6: Autokorelogram a parciální autokorelogram logaritmů výnosů akcií společnosti ČEZ.



Obrázek 6.7: Autokorelogram a parciální autokorelogram řady $\{X_t^2, t = 1, ..., 269\}$, kde X_t jsou logaritmy výnosů akcií společnosti ČEZ.

a poslední rovnost také platí. Pomocí funkce garchFit v progamu R odhadneme neznámé parametry metodou maximální věrohodnoti. Dosadíme odhadnuté parametry

$$\begin{aligned} X_t &= 0.001379 + \epsilon_t, \\ \epsilon_t &= \sigma_t Z_t, \\ \sigma_t^2 &= 0.00005701 + 0.2957\epsilon_{t-1}^2 + 0.5615\sigma_{t-1}^2, \qquad t = 1, ..., 269. \end{aligned}$$

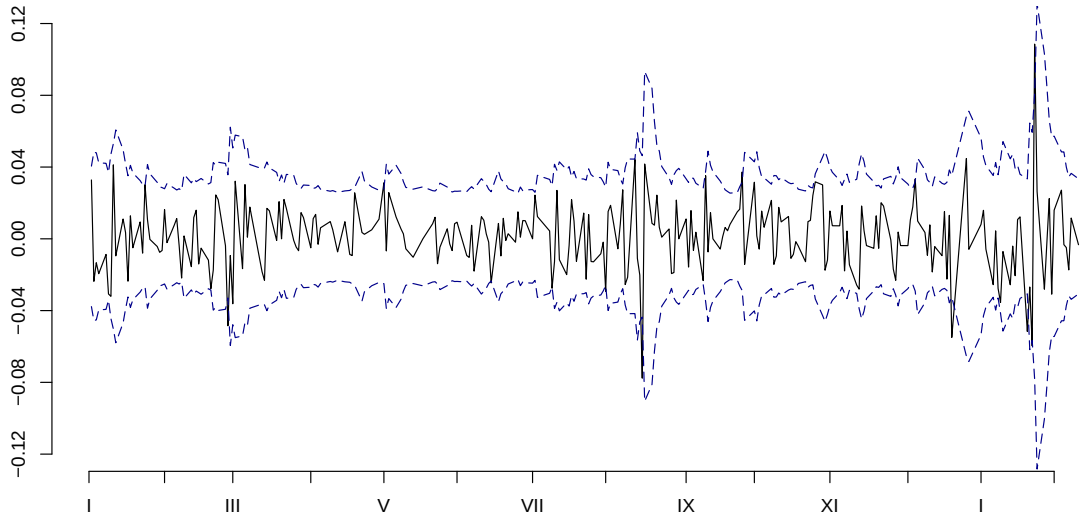
Platí, že GARCH(1,1) proces je slabě stacionární, je-li $\alpha + \beta < 1$. V našem případě je $0.2957 + 0.5615 < 1$, tedy jedná se o stacionární proces. Popsané definice a vlastnosti byly převzaty z knihy Franke et al. (2004).

Predikční intervaly pro budoucí pozorování

Označme $n = 269$ (délka řady). Konfidenční intervaly o jeden krok dopředu se konstruují dle vztahu

$$(\mu - u(\alpha/2)\sigma_{n+1}, \mu + u(\alpha/2)\sigma_{n+1}),$$

kde $\sigma_{n+1} = \sqrt{\omega + \alpha(X_n - \mu)^2 + \beta\sigma_n^2}$. V praxi se postupuje tak, že se nejprve dosadí odhady a následně se rekurentně vypočte rozptyl. Na obrázku 6.8 je řada logaritmů výnosů akcií společnosti ČEZ s odhadnutými 95% intervaly spolehlivosti pro pozorování o jeden krok dopředu.



Obrázek 6.8: Logaritmy výnosů a jejich 95% intervaly spolehlivosti o jeden krok dopředu.

Naznačme, jak lze konstruovat asymptotické intervaly spolehlivosti pro vzdálenější budoucí pozorování. Jako odhad budoucích volatilit volme podmíněnou střední hodnotu

$$\begin{aligned}
\mathrm{E}\left[\sigma_{n+k}^2|\mathcal{F}_n\right] &= \mathrm{E}\left[\omega + \alpha\epsilon_{n+k-1}^2 Z_{n+k-1}^2 + \beta\sigma_{n+k-1}^2|\sigma_1^2, ..., \sigma_{n+1}^2\right] \\
&= \ldots = \sum_{j=0}^{k-2} \omega\left(\alpha+\beta\right)^j + \left(\alpha+\beta\right)^{k-1}\sigma_{n+1}^2 \\
&= \frac{\omega}{1-\alpha-\beta} + \left(\alpha+\beta\right)^{k-1}\left(\sigma_{n+1}^2 + \frac{\omega}{1-\alpha-\beta}\right), \quad k \geq 1.
\end{aligned}$$

Rekurentně vypočtené odhady rozptylu označme $\widehat{\sigma_t}^2$. Odhad intervalu spolehlivosti pro budoucí hodnotu X_{n+k} je pak tvaru

$$\left(\widehat{\mu} - u(\alpha/2)\widehat{\sigma_{n+k}}, \widehat{\mu} + u(\alpha/2)\widehat{\sigma_{n+k}}\right). \tag{6.1}$$

V řadě případů pozorujeme u rozdělení logaritmů výnosů výrazné zešikmení a těžké chvosty. Normalitu testujme pomocí *Shapiro Wilk* testu, p-hodnota vychází 0.0000007886.

S ohledem na výraznou nenormalitu dat nelze předpokládat výraznou kvalitu takto konstruovaných intervalů. Uvědomme si, že intervaly typu (6.1) jsou symetrické. Bootstrapové intervaly, které nepředpokládají normalitu dat, proto slibují spolehlivější intervaly pro budoucí hodnoty.

Postup konstrukce bootstrapových konfidenčních intervalů je popsán v článku Pascual et al. (2000). Základním principem použitého bootstrapového odhadu intervalů spolehlivosti je tvorba bootstrapových výběrů dle předpisu odhadnutého GARCH modelu, a to generováním z centrovaných odhadnutých reziduí. Předpokládáme, že každý takto získaný bootstrapový výběr lze opět modelovat jako GARCH proces, odhadneme jeho parametry a simulujeme budoucí hodnoty opět pomocí původních centrovaných odhadnutých reziduí. Zapišme celý algoritmus pro GARCH model z definice 9.

Algoritmus 6 (Bootstrapové intervaly pro GARCH proces)**.**

1. *Odhadneme neznámé parametry modelu* $(\widehat{\omega}, \widehat{\alpha}, \widehat{\beta})$ *a rekurentně vypočteme odhad volatilit* $\widehat{\sigma}_t$.
2. *Vypočteme rezidua* $\widehat{Z}_t = \frac{\epsilon_t}{\widehat{\sigma}_t}$ *a definujeme centrovaná rezidua* $\widetilde{Z}_t = \widehat{Z}_t - \overline{\widehat{Z}}_t$. *Centrovaná rezidua použijeme jako základní populaci pro generování bootstrapových výběrů* $\{Z_t^*\}$.
3. *Bootstrapový GARCH proces generujeme rekurentně dle vztahů*

$$\begin{aligned} \epsilon_t^* &= \sigma_t^* Z_t^*, \quad t = 1, ..., n, \\ \sigma_t^{2*} &= \widehat{\omega} + \widehat{\alpha}\epsilon_{t-1}^{*2} + \widehat{\beta}\sigma_{t-1}^{2*}, \quad t = 2, ..., n, \end{aligned}$$

   kde $\sigma_1^{2*} = \widehat{\sigma}_1^2 = \frac{\widehat{\omega}}{1-\widehat{\alpha}-\widehat{\beta}}$.
4. *Předpokládáme, že řadu* $\{\epsilon_t^*\}$ *lze modelovat jako GARCH proces stejného řádu. Odhadneme jeho parametry* $(\widehat{\omega^*}, \widehat{\alpha^*}, \widehat{\beta^*})$.
5. *Generujeme budoucí hodnoty pomocí vztahů*

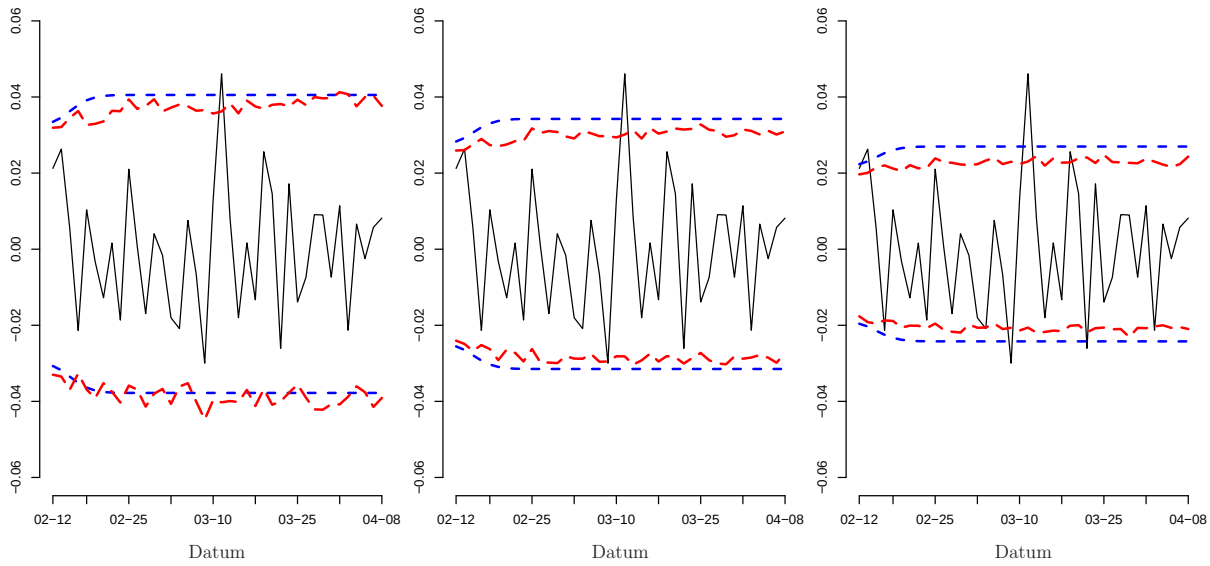
$$\begin{aligned} \epsilon_{n+s}^* &= \sigma_{n+s}^* Z_{n+s}^*, \quad s = 1, ..., k, \\ \sigma_{n+s}^{2*} &= \widehat{\omega^*} + \widehat{\alpha^*}\epsilon_{n+s-1}^{*2} + \widehat{\beta^*}\sigma_{n+s-1}^{2*}, \quad s = 1, ..., k, \end{aligned}$$

   kde $\epsilon_n^* = \epsilon_n$ *a* $\sigma_n^{2*} = \frac{\widehat{\omega^*}}{1-\widehat{\alpha^*}-\widehat{\beta^*}} + \widehat{\alpha^*}\sum_{j=0}^{n-2}\widehat{\beta^{*j}}\left(\epsilon_{n-j-1} - \frac{\widehat{\omega^*}}{1-\widehat{\alpha^*}-\widehat{\beta^*}}\right)$.
6. *Kroky 3. - 5. opakujeme B-krát. Takto získáme B bodových předpovědí hodnoty* ϵ_{n+k}. $(1-\alpha)100\%$ *interval spolehlivosti konstruujeme jako*

$$\left(G_n^*\left(\frac{\alpha}{2}\right), G_n^*\left(1-\frac{\alpha}{2}\right)\right),$$

   kde G_n *značí empirickou distribuční funkci bootstrapových předpovědí* ϵ_{n+k}^*. *Poznamenejme, že pomocí popsaného postupu lze konstruovat také intervaly spolehlovosti pro budoucí hodnoty volatilit.*

Na obrázku 6.9 vidíme srovnání bootstrapových a asymptotických intervalů spolehlovosti. U 95% intervalu je vidět, že bootstrapové intervaly se rozšiřují se vzdálenějšími



Obrázek 6.9: Na grafech jsou postupně srovnány 95%, 90% a 80% intervaly spolehlivosti pro budoucí hodnoty. Krátce čárkovaně (modře) jsou vyznačeny asymptotické intervaly dlouze (červeně) pak bootstapové. Bootstrapové intervaly získány provedením 2000 bootstrapových výběrů.

pozorování, jak bychom očekávali, naopak asymptotické jsou téměř konstantní. V případě intervalů s nominální krycí schopností 90 % pozorujeme, že asymptotické intervaly jsou zbytečně široké. Vzhledem k nominální hodnotě by teoreticky neměly v intervalu ležet 4 pozorování (celkem predikujeme interval pro 40 pozorování). V bootstrapových intervalech leží 37 pozorování, v asymptotických pak 39, v případě 80 % intervalů je to 33 pro bootstrapové a 36 pro asymptotické. Také se zdá, že bootstrapové intervaly jsou mírně posunuté směrem k nižším hodnotám, což je patrně způsobeno výrazně šikmým rozdělením. Toto by se dalo interpretovat tak, že na finančních trzích se častěji objevují výraznější výkyvy cen směrem dolů nežli nahoru.

Charakteristiky průměrného výnosu

Pokud se zajímáme o průměrný výnos, např. zvažujeme-li dlouhodobou výhodnost investice, lze také použít bootstrapovou metodu k odhadu rozptylu a intervalů spolehlivosti. Protože jsme ukázali, že logaritmy výnosů v našich příkladech jsou stacionární, můžeme k odhadu charakteristik např. výnosnosti využít metodu blokový bootstrap. $100(1-\alpha)\%$ interval spolehlivosti je konstruován dle předpisu

$$\left(\overline{X}_n + F^{-1}(\alpha/2)\sqrt{\frac{s_n^2}{n}}, \overline{X}_n + F^{-1}(1-\alpha/2)\sqrt{\frac{s_n^2}{n}}\right),$$

kde

$$s_n^2 = \sum_{k=0}^{l-1} \frac{1}{n} \sum_{j=1}^{n-l} \left(X_j - \overline{X}_n\right)\left(X_{j+k} - \overline{X}_n\right)$$

a F^{-1}je kvantilová funkce studentizovaného bootstrapového rozdělení.

	logaritmy výnosů akcií spol. ČEZ
$\overline{X}_n$	0.0008703944
$n\widehat{\text{var}}\,\overline{X}_n$	0.0002965264
95% interval pro $\overline{X}_n$	(-0.001099910, 0.003028955)
99% interval pro $\overline{X}_n$	(-0.001673492, 0.003713116)

Tabulka 6.3: Odhady rozptylu a intervaly spolehlivosti pro výběrový průměr logaritmů výnosů akcií společnosti ČEZ. Konfidenční intervaly odhadnuty metodou blokový bootstrap s délkou bloku $l = 5$ a obdélníkovými váhami pro studentizaci. Počet bootstrapových výběrů zvolen $N = 5000$.

6.3 Implementace

Všechny výpočty byly provedeny ve statistickém software R. Naprogramované procedury metody nezávislý bootstrap lze nalézt zejména v knihovně bootstrap, metody bootstrap pro závislá data lze řešit funkcemi knihoven boot a tseries.

Pro účely simulací provedených v této práci bylo nutné naprogramovat některé nové procedury, případně již existující přepsat s cílem urychlení doby výpočtu. Zvýšení rychlosti výpočtu bylo docíleno naprogramováním částí algoritmů v jazyce Fortran. Jedná se o procedury na diagnostiku modelu stochastické regrese, výpočet bootstrapových charakteristik na základě metod sítový a blokový bootstrap a hledání optimální délky bloku.

Z těchto byla vytvořena knihovna depboot, ve které jsou všechny procedury opatřeny nápovědou a obohaceny několika příklady použití. Knihovna byla dále doplněna datovými soubory použitými v aplikační části.

Schémata uvedená v práci byla vytvořena v programu Ipe 6.0.

6.4 Obsah přiloženého CD a instalace knihovny

Kromě textu v elektronické podobě lze na přiloženém CD dále nalézt skripty k tvorbě všech grafů a tabulek. Tyto jsou setříděny podle kapitol, ke které náležejí. Všechny skripty jsou spustitelné ve statistickém programu R (http://www.r-project.org/). V adresáři Knihovna jsou uloženy zdrojové soubory potřebné k instalaci knihovny.

Instalace knihovny je provedena příkazem *R CMD INSTALL depboot_1.0.tar.gz*. Je však nutné mít příkaz *R CMD* zprovozněn (problém nastává pod Windows, pod Linuxem funguje), návod lze nalézt kupříkladu na webové stránce
http://www.murdoch-sutherland.com/Rtools/. Alternativou pod operačním systémem Windows je využít soubor depboot_1.0.zip. Instalace pak probíhá přímo z prostředí R kliknutím na `Packages\Install package from local zip files`. Nakonec je potřeba zkopírovat soubory fdbootstrap.dll a blockboot.dll do adresáře libs nainstalované knihovny (obvykle je to cesta `C:\Program Files\R\R-2.6.0\library\depboot\libs`).

Závěr

Metodu bootstrap lze použít nejen jako alternativu klasických asymptotických metod, ale také k odhadu charakteristik a konfidenčních intervalů testových statistik, pro které asymptotické rozdělení není známé. Navíc lze tvrdit, že výsledky získané metodou bootstrap jsou obecně přesnější nežli klasické asymptotické. Fakt, že výše uvedená tvrzení platí i pro případ závislých pozorování, je jedním z důvodů, proč se metoda bootstrap těší velkému zájmu statistiků a stále větší oblibě při aplikacích v praxi.

Dalším přínosem metody bootstrap je snadná a přímočará aplikace. Ovšem jak jsme ukázali v případě blokových metod, může bezhlavý postup vést ke špatným výsledkům.

Statistické závěry v regresních modelech lze řešit metodou frekvenční bootstrap. Na základě simulačního experimentu lze tuto metodu hodnotit jako spolehlivější nežli metody založené na asymptotickém rozdělení.

V oblasti časových řad proti sobě stojí dva bootstrapové přístupy, blokový a sítový. Sítový bootstrap těží ze snažší aplikace, blokové metody naopak z větší univerzálnosti. Kvality odhadů pomocí těchto dvou metod se pro obecné stacionární invertibilní procesy výrazně neodlišují. K odchýlení výkonnosti dochází při přechodu k neinvertibilním případně až nestacionárním procesům, kde blokové metody jasně dominují. Na druhou stranu blokové metody kriticky závisejí na vhodné volbě délky bloku. Existují postupy, jak vhodnou délku bloku odhadnout. Tyto postupy jsou výpočetně velmi náročné a pro složitější modely podhodnocují optimální délku.

V prvním příkladu, kde byla použita reálná data, jsme demonstrovali následky neuvážené aplikace. Nevhodnou volbou délky bloku jsme dospěli k téměř dvojnásobnému nadhodnocení roztpylu výběrového průměru. Při testování hypotézy rostoucích průměrných teplot jsme ukázali, že aplikace metody bootstrap je velmi snadná. Je však vždy nutné ověřit platnost obecných předpokladů. Při řešení problému optimálního investování jsme ukázali, že metodu lze užít také k odhadu konfidenčních intervalů pro budoucí pozorování. Poznamenejme, že výsledky získané metodou bootstrap splňovaly očekávané vlastnosti více nežli výsledky asymptotické.

Nakonec upozorněme, že metoda bootstrap není univerzální metodou k řešení všech statistických problémů. Existuje mnoho případů, kdy metoda bootstrap selhává.

V práci byly samostatně odvozeny některé vlastnosti diskrétních Fourierových transformací. Práce poskytla ucelený přehled metod bootstrap pro závislá data a porovnala je z teoretického hlediska a také pomocí simulační studie. Diskutované postupy jsou implementovány v jazyce R a jazyce Fortran.

Bylo by zajímavé více prozkoumat vlastnosti metody frekvenční bootstrap. Neméně zajímavé by bylo studium možností aplikace metody bootstrap na heteroskedastické modely.

Literatura

Anděl, J. (2005). *Základy matematické statistiky.* Univerzita Karlova v Praze, MATFYZPRESS.

Brillinger, D. R. (1975). *Time Series, Data Analysis and Theory.* Holt, Rinehart and Winston, Inc.

Brillinger, D. R. (1989). Consistent detection of a monotonic trend superposed on a stationary time series. *Biometrika*, 76(1):23–30.

Brockwell, P. J. a Davis, R. A. (1996). *Time Series: Theory and Methods.* Springer-Verlag GmbH.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148.

Bühlmann, P. (2002). Bootstraps for Time Series. *Statist. Sci.*, 17(1):52–72.

Davison, A. C. a Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

Efron, B. a Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

van Es, B. a Putter, H. (2006). *The Bootstrap.* Universiteit van Amsterdam, Korteweg-de Vries Institute for Mathematics.

Franke, J., Härdle, W., a Hafner, C. (2004). *Einführung in die Statistik der Finanzmärkte.* Springer.

Götze, F. a Künsch, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Ann. Stat.*, 24(5):1914–1933.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag.

Hall, P., Horowitz, J. L., a Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Hidalgo, J. (2003). An alternative bootstrap to moving blocks for time series regression models. *J. Econom.*, 117(2):369–399.

Härdle, W., Horowitz, J. L., a Kreiss, J.-P. (2001). *Bootstrap Methods for Time Series.* Diskusní materiál, Humboldt-Universität zu Berlin.

Lachout, P. (2004). *Teorie pravděpodobnosti.* Univerzita Karlova v Praze, nakladatelství Karolinum.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data.* Springer.

Pascual, L., Romo, J., a Ruiz, E. (2000). Forecasting Returns and Volatilities in GARCH Processes Using the Bootstrap. Universidad Carlos III de Madrid.

Prášková, Z. (2004a). Metoda bootstrap. *Robust´2004.*

Prášková, Z. (2004b). *Náhodné procesy II.* Univerzita Karlova v Praze, nakladatelství Karolinum.

A Dodatek

A.1 Odhady parametrů v autoregresních procesech

Odhad parametrů metodou momentů v autoregresním procesu řádu p

Uvažujme autoregresní proces řádu p tvaru

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

kde $\{Y_t\} \sim WN(0, \sigma^2)$. Uvedeme postup, jakým lze odhadnout neznámé parametry $\varphi_1, ..., \varphi_p$, σ^2 a definujeme *Akaikeho Informační kritérium* (postupujeme dle Prášková (2004b)).

Dále se uvažuje, že proces $\{X_t\}$ je centrovaný. Sestaví se rovnice

$$\begin{aligned} \mathrm{E}\, X_t X_t - \varphi_1 \mathrm{E}\, X_{t-1} X_t - \ldots - \varphi_p \mathrm{E}\, X_{t-p} X_t &= \mathrm{E}\, Y_t X_t, \\ \mathrm{E}\, X_t X_{t-k} - \varphi_1 \mathrm{E}\, X_{t-1} X_{t-k} - \ldots - \varphi_p \mathrm{E}\, X_{t-p} X_{t-k} &= \mathrm{E}\, Y_t X_{t-k}, \quad k = 1, ..., p, \end{aligned}$$

ze kterých se úpravou odvodí tzv. *Yule-Walkerovy rovnice*

$$\varphi_1 R(1) + \ldots + \varphi_p R(p) + \sigma^2 = R(0), \tag{A.1}$$

$$\varphi_1 R(k-1) + \ldots + \varphi_p R(k-p) = R(k), \quad k = 1, ..., p, \tag{A.2}$$

kde jsme označili $R(k) = \mathrm{E}\, X_t X_{t-k}$. Autokovarianční funkci $R(k)$ odhadneme výběrovým protějškem $\widehat{R}(k) = \frac{1}{n} \sum_{t=1}^{n-k} X_t X_{t+k}$ pro $k = 0, ..., p$. Odhady autoregresních parametrů $\widehat{\varphi}_1, ..., \widehat{\varphi}_p$ se získají jako řešení soustavy rovnic (A.2). Nakonec se použije rovnice (A.1) k odhadu rozptylu bílého šumu

$$\widehat{\sigma}^2 = \widehat{R}(0) - \widehat{\varphi}_1 \widehat{R}(1) - \ldots - \widehat{\varphi}_p \widehat{R}(p).$$

Odvození Akaikeho informačního kritéria

Známe-li řád procesu, odhadneme pomocí momentové metody jeho parametry, jak jsme ukázali. Řád procesu se obecně odhaduje minimalizací informačního kritéria. Odvodíme tvar Akaikeho informačního kritéria pro autoregresní modely.

Předpokládejme nyní navíc, že Y_t jsou nezávislé náhodné veličiny dané rozdělením $N(0, \sigma^2)$. Definujme dále *podmíněnou věrohodnostní funkci* pro odhad parametrů $\boldsymbol{\varphi}(p) = (\varphi_1, ..., \varphi_p)^T$ a $\sigma^2(p)$ výrazem

$$L(\boldsymbol{\varphi}(p), \sigma^2(p); \boldsymbol{X}) = \frac{1}{(2\pi\sigma^2(p))^{(n-p)/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p})^2 \right\}.$$

Metodou momentů jsme odhadli parametry $\boldsymbol{\varphi}(p)$ a $\sigma^2(p)$ pro autoregresní model řádu p. V praxi se tyto odhady dosadí do *logaritmické podmíněné věrohodnostní funkce* a dále se funkce aproximuje na tvar

$$l(\widehat{\boldsymbol{\varphi}}(p), \widehat{\sigma}^2(p); \boldsymbol{X}) \approx -\frac{n}{2} \log\left(2\pi\widehat{\sigma}^2(p)\right) - \frac{n}{2}.$$

Akaikeho informační kritérium definujeme jako funkci vzhledem k řádu procesu p

$$AIC(p; \widehat{\boldsymbol{\varphi}}(p), \widehat{\sigma}^2(p)) = n\left(\log\left(2\pi\widehat{\sigma}^2\right) + 1\right) + 2\,(p+1)$$

a její minimalizací (vzhledem k p) získáme optimální model.

A.2 Kumulanty

Mějme dánu náhodnou veličinu X, její charakteristickou funkci označme χ. Platí, že $\chi(t) = \mathrm{E}\, e^{itX}$ pro $t \in \mathbb{R}$.

Definice 10. J-tý kumulant κ_j náhodné veličiny X je definován jako koeficient členu $\frac{1}{j!}(it)^j$ v Taylorově rozvoji funkce $\log \chi(t)$ v bodě 0.

Kumulanty lze vyjádřit vzhledem k momentům. Z definice

$$\log \chi(t) = \kappa_1 it + \frac{1}{2}\kappa_2 (it)^2 + \ldots + \frac{1}{j!}\kappa_j (it)^j + \ldots, \tag{A.3}$$

ale zároveň lze formálně psát

$$\chi(t) = 1 + \mathrm{E}\, X it + \frac{1}{2}\mathrm{E}\, X^2 (it)^2 + \ldots + \frac{1}{j!}\mathrm{E}\, X^j (it)^j + \ldots. \tag{A.4}$$

A protože pro Taylorův rozvoj logaritmu platí $\log(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$, dosazením do (A.4) a srovnáním s (A.3) vidíme, že

$$\sum_{j=1}^{\infty} \frac{1}{j!}\kappa_j (it)^j = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\left(\sum_{j=1}^{\infty} \frac{1}{j!}\mathrm{E}\, X^j (it)^j\right)^k}{k}. \tag{A.5}$$

Z rovnosti (A.5) lze porovnáním koeficientů u členů $(it)^j$ vyjádřit kumulanty vzhledem k momentům, totiž že

$$\begin{aligned}
\kappa_1 &= \mathrm{E}\, X, \\
\kappa_2 &= \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2 = \mathrm{var}\, X, \\
\kappa_3 &= \mathrm{E}\, X^3 - 3\mathrm{E}\, X^2 \mathrm{E}\, X + 2(\mathrm{E}\, X)^3 = \mathrm{E}\,(X - \mathrm{E}\, X)^3, \\
\kappa_4 &= \mathrm{E}\, X^4 - 4\mathrm{E}\, X^3 \mathrm{E}\, X - 3(\mathrm{E}\, X^2)^2 + 12\mathrm{E}\, X^2 (\mathrm{E}\, X)^2 - 6(\mathrm{E}\, X)^4 \\
&= \mathrm{E}\,(X - \mathrm{E}\, X)^4 - 3(\mathrm{var}\, X)^2.
\end{aligned}$$

V případě, že pracujeme s d-rozměrným náhodným vektorem $\boldsymbol{X} = (X_1, ..., X_d)^T$, je charakteristická funkce definována výrazem $\chi(\boldsymbol{t}) = \mathrm{E}\, e^{i\langle \boldsymbol{t}, \boldsymbol{X}\rangle}$, $\boldsymbol{t} \in \mathbb{R}^d$. Zavedeme vektorové

indexování $\boldsymbol{j} = (j_1, ..., j_d)^T$ (postupujeme dle Hall (1992), str. 242 - 244). Kumulanty závisejí na celém vektoru $\boldsymbol{X}$, proto J-tý kumulant, kde i-tá složka vektoru $\boldsymbol{X}$ má zastoupení $X_i^{j_i}$ označíme $\kappa_J(X_1^{j_1}...X_d^{j_d})$, $J = j_1 + ... + j_d$. Rozvoj (A.3) přechází ve vícerozměrném případě v rozvoj

$$\log \chi(\boldsymbol{t}) = \sum_{J=1}^{\infty} \frac{1}{j_1!...j_d!} \kappa_J(X_1^{j_1}...X_d^{j_d})(it_1)^{j_1}...(it_d)^{j_d},$$

který lze s využitím vektorového indexování zapsat jako

$$\log \chi(\boldsymbol{t}) = \sum_{\boldsymbol{j}>0} \frac{1}{\boldsymbol{j}!} \kappa_{\boldsymbol{j}} (i\boldsymbol{t})^{\boldsymbol{j}},$$

kde sčítáme přes nezáporné celočíselné vektory s alespoň jednou složkou kladnou,

$$\kappa_{\boldsymbol{j}} = \kappa_J(X_1^{j_1}...X_d^{j_d}), \quad \boldsymbol{j}! = \prod_{s=1}^{d} j_s!, \quad (i\boldsymbol{t})^{\boldsymbol{j}} = \prod_{s=1}^{d} (it_s)^{j_s}.$$

S využitím označení $\mu_{\boldsymbol{j}} = \mathrm{E}\,(X_1^{j_1}...X_d^{j_d})$ přepíšeme rovnost (A.5) ve vícerozměrném případě jako

$$\sum_{\boldsymbol{j}>0} \frac{1}{\boldsymbol{j}!} \kappa_{\boldsymbol{j}} (i\boldsymbol{t})^{\boldsymbol{j}} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\left(\sum_{\boldsymbol{j}>0} \frac{1}{\boldsymbol{j}!} \mu_{\boldsymbol{j}} (i\boldsymbol{t})^{\boldsymbol{j}} \right)^k}{k}.$$

Pro první čtyři kumulanty platí vztahy

$$\begin{aligned}
\kappa_1(X_s) &= \mathrm{E}\, X_s, \\
\kappa_2(X_s X_t) &= \mathrm{E}\,(X_s X_t) - \mathrm{E}\, X_s \mathrm{E}\, X_t = \mathrm{cov}\,(X_s, X_t), \\
\kappa_3(X_s X_t X_u) &= \mathrm{E}\,(X_s X_t X_u) - \mathrm{E}\,(X_s X_t)\,\mathrm{E}\, X_u - \mathrm{E}\,(X_t X_u)\,\mathrm{E}\, X_s \\
&\quad -\mathrm{E}\,(X_s X_u)\,\mathrm{E}\, X_t + 2\mathrm{E}\, X_s \mathrm{E}\, X_t \mathrm{E}\, X_u, \\
\kappa_4(X_s X_t X_u X_v) &= \mathrm{E}\,(X_s X_t X_u X_v) - \mathrm{E}\,(X_s X_t X_u)\,\mathrm{E}\, X_v - \mathrm{E}\,(X_s X_t X_v)\,\mathrm{E}\, X_u \\
&\quad -\mathrm{E}\,(X_s X_u X_v)\,\mathrm{E}\, X_t - \mathrm{E}\,(X_t X_u X_v)\,\mathrm{E}\, X_s - \mathrm{E}\,(X_s X_t)\,\mathrm{E}\,(X_u X_v) \\
&\quad -\mathrm{E}\,(X_s X_u)\,\mathrm{E}\,(X_t X_v) - \mathrm{E}\,(X_s X_v)\,\mathrm{E}\,(X_t X_u) \\
&\quad +2\left[\mathrm{E}\, X_s \mathrm{E}\, X_t \mathrm{E}\,(X_u X_v) + \mathrm{E}\, X_s \mathrm{E}\, X_u \mathrm{E}\,(X_t X_v)\right] \\
&\quad +2\left[\mathrm{E}\, X_s \mathrm{E}\, X_v \mathrm{E}\,(X_t X_u) + \mathrm{E}\, X_t \mathrm{E}\, X_u \mathrm{E}\,(X_s X_v)\right] \\
&\quad +2\left[\mathrm{E}\, X_t \mathrm{E}\, X_v \mathrm{E}\,(X_s X_u) + \mathrm{E}\, X_u \mathrm{E}\, X_v \mathrm{E}\,(X_s X_t)\right] \\
&\quad -6\mathrm{E}\, X_s \mathrm{E}\, X_t \mathrm{E}\, X_u \mathrm{E}\, X_v,
\end{aligned}$$

pro libovolné $s, t, u, v = 1, ..., n$.

B Příloha

Regresní modely

		FDB			SIEVE			ASY		
n	model	99%	95%	90%	99%	95%	90%	99%	95%	90%
65	R1	98.80	94.90*	90.50	98.95*	95.40	90.40*	98.35	93.55	88.00
	R2	98.30	93.80	87.80	99.25	95.55*	90.80*	98.90*	94.10	87.20
	R3	93.85	90.25	85.40	98.45*	94.25*	89.30 *	93.70	85.90	79.20
	R4	99.05*	95.80	90.80*	99.20	95.70	91.15	98.75	94.65*	88.50
	R5	98.75*	94.55	89.45	98.75*	95.25*	89.95*	98.30	93.20	87.50
	R6	97.60	93.45	88.25	98.50*	94.60*	89.35*	96.55	90.15	84.05
125	R1	99.15	94.85*	89.50*	99.00*	94.55	89.30	98.70	93.60	88.15
	R2	98.70	94.25	89.95*	99.25*	95.30	91.10	99.3	95.00*	90.15
	R3	95.30	90.65	85.65	98.20*	94.50*	88.60*	95.35	87.10	80.75
	R4	99.00*	95.40	90.10*	99.00*	95.05*	90.20	98.80	94.05	88.90
	R5	98.55	95.15*	90.60*	99.00*	95.35	90.85	98.85	94.60	89.30
	R6	98.55	93.15	87.60	98.75*	94.45*	88.55*	97.25	90.65	84.90

Tabulka B.1: Průměrné krytí intervalů spolehlivosti v závislosti na velikosti a typu procesu. Srovnání dvou bootstrapových a asymptotické metody. Provedeno 2 000 simulací a pro každou generováno 2 000 bootstrapových výběrů.

				vychýlení			směr. odchylka			MSE		
n	model	l	$n\mathrm{var}\,\widehat{\beta}$	FDB	MBB	SIEVE	FDB	MBB	SIEVE	FDB	MBB	SIEVE
65	R1	3	1.11	−0.03	−0.06	−0.04	0.38	0.31	0.32	0.14	0.10	0.10
	R2	9	8.39	−0.15	−1.72	−3.26	8.25	5.27	3.75	68.13	30.70	24.71
	R3	3	0.35	−0.07	−0.07	−0.05	0.25	0.18	0.21	0.07	0.04	0.05
	R4	4	2.75	−0.10	−0.27	−0.30	1.10	0.85	0.86	1.21	0.80	0.83
	R5	5	4.66	−0.30	−0.76	−1.10	2.92	2.02	1.83	8.59	4.65	4.54
	R6	3	1.85	−0.21	−0.50	−0.29	0.93	0.59	0.74	0.91	0.59	0.64
125	R1	6	1.12	−0.05	−0.06	−0.05	0.29	0.24	0.23	0.08	0.06	0.06
	R2	18	10.69	−0.38	−1.80	−3.25	8.63	6.08	4.65	74.54	40.16	32.20
	R3	6	0.26	−0.04	−0.04	−0.03	0.16	0.12	0.14	0.03	0.02	0.02
	R4	8	2.77	−0.13	−0.21	−0.24	0.80	0.65	0.63	0.66	0.46	0.45
	R5	10	5.14	−0.02	−0.48	−0.75	2.95	2.05	1.88	8.69	4.45	4.07
	R6	6	1.74	−0.18	−0.34	−0.22	0.70	0.46	0.55	0.52	0.33	0.35

Tabulka B.2: Kvalita aproximace rozptylu parametru β v modelu nekorelované stochastické regrese. Srovnání tří bootstrapových metod. Správné hodnoty odhadnuty metodou MC pomocí 10 000 simulací. Délka bloku volena metodou tvorby podvýběrů. Výsledky získány provedením 2 000 simulací a pro každou generováno 500 bootstrapových výběrů.

Srovnání algoritmů pro optimální délku bloku

l	1	2	3	4*	5*	6*	7*	8	9	10	11	12	13	14	15
podvýběry	9	0	18	22	16	13	0	6	5	3	3	0	0	1	4
parametrická	9	25	30	27	8	1	0	0	0	0	0	0	0	0	0

Tabulka B.3: Frekvence optimální délky bloku odhadnuté na základě dvou odlišných metod. Celkem provedeno 100 simulací. Volba délky bloku pro rozptyl statistiky $median(\mathcal{X})$. Data se řídí modelem D1. Optimální délka získaná metodou MC má hodnotu $l^{opt} = 6$.

l	1	2*	3*	4*	5*	6*	7	8	9	10	11	12	13	14	15
podvýběry	6	0	29	20	12	14	0	8	1	3	3	0	2	1	1
parametrická	25	23	33	12	6	1	0	0	0	0	0	0	0	0	0

Tabulka B.4: Frekvence optimální délky bloku odhadnuté na základě dvou odlišných metod. Celkem provedeno 100 simulací. Volba délky bloku pro rozptyl statistiky $\overline{X}_n$. Data se řídí modelem D2. Optimální délka získaná metodou MC má hodnotu $l^{opt} = 3$.

l	1	2*	3*	4*	5	6	7	8	9	10	11	12	13	14	15
podvýběry	19	22	0	16	7	6	4	5	0	5	5	3	4	2	2
parametrická	44	47	9	0	0	0	0	0	0	0	0	0	0	0	0

Tabulka B.5: Frekvence optimální délky bloku odhadnuté na základě dvou odlišných metod. Celkem provedeno 100 simulací. Volba délky bloku pro odhad distribuční funkce statistiky $\overline{X}_n$. Data se řídí modelem D2. Optimální délka získaná metodou MC má hodnotu $l^{opt} = 3$.

l	1	2	3	4	5	6	7*	8*	9*	10*	11*	12*	13*	14	15
podvýběry	0	0	1	14	33	20	0	12	7	5	0	0	1	1	6
parametrická	2	4	12	24	22	22	8	6	0	0	0	0	0	0	0

Tabulka B.6: Frekvence optimální délky bloku odhadnuté na základě dvou odlišných metod. Celkem provedeno 100 simulací. Volba délky bloku pro rozptyl statistiky $\overline{X}_n$. Data se řídí modelem D3. Optimální délka získaná metodou MC má hodnotu $l^{opt} = 9$.

Srovnání metod blokový a sítový bootstrap - MA procesy

				vychýlení		směr. odchylka		MSE	
n	model	l	$n\mathrm{var}\,\overline{X}_n$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(MA)1	4	5.76	−1.56	−1.04	1.52	3.12	4.72	10.81
	C(MA)2	3	1.38	−0.14	−0.07	0.35	0.64	0.14	0.42
	C(MA)3	4	1.63	−0.06	−0.02	0.48	0.73	0.24	0.53
	C(MA)4	3	1.40	−0.16	−0.08	0.35	0.65	0.14	0.43
	C(MA)5	5	1.75	−0.17	−0.13	0.54	0.74	0.32	0.56
125	C(MA)1	8	5.66	−0.87	−0.59	1.61	2.66	3.35	7.40
	C(MA)2	6	1.41	−0.13	−0.09	0.35	0.54	0.14	0.31
	C(MA)3	8	1.66	−0.08	−0.02	0.50	0.58	0.25	0.34
	C(MA)4	6	1.41	−0.14	−0.09	0.36	0.57	0.15	0.33
	C(MA)5	10	1.63	−0.07	0.01	0.52	0.56	0.28	0.32

Tabulka B.7: Kvalita odhadu rozptylu výběrového průměru v ARMA modelech. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 2 000, bootstrapových výběrů 500.

				vychýlení		směr. odchylka		MSE	
n	model	l	$n\mathrm{var}\,T_n$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(MA)1	5	7.03	−0.74	−1.12	3.51	3.28	12.86	11.97
	C(MA)2	6	2.04	−0.02	−0.04	1.18	0.92	1.39	0.85
	C(MA)3	10	2.68	−0.16	0.01	1.63	1.34	2.70	1.80
	C(MA)4	6	2.00	0.02	−0.01	1.13	0.89	1.28	0.80
	C(MA)5	6	2.53	0.16	0.10	1.54	1.31	2.41	1.71
125	C(MA)1	10	7.35	−0.85	−0.92	3.29	2.97	11.55	9.66
	C(MA)2	12	1.93	0.03	0.04	1.05	0.72	1.10	0.51
	C(MA)3	20	2.60	−0.16	0.07	1.55	1.08	2.42	1.16
	C(MA)4	12	1.93	0.05	0.04	1.07	0.72	1.14	0.53
	C(MA)5	12	2.60	−0.07	0.04	1.37	1.08	1.87	1.16

Tabulka B.8: Kvalita odhadu rozptylu mediánu. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 2 000, bootstrapových výběrů 500.

Srovnání metod blokový a sítový bootstrap - ARMA procesy

				vychýlení		směr. odchylka		MSE	
n	model	l	$n\mathrm{var}\,\overline{X}_n$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(ARMA)1	4	4.09	−1.25	−0.66	0.98	1.96	2.52	4.28
	C(ARMA)2	6	23.75	−11.56	−7.21	5.78	12.39	166.94	205.39
	C(ARMA)3	6	0.78	−0.05	−0.07	0.27	0.32	0.07	0.11
	C(ARMA)4	4	3.53	−0.88	−0.48	0.90	1.82	1.59	3.55
	C(ARMA)5	4	1.51	−0.14	−0.01	0.43	0.75	0.21	0.56
125	C(ARMA)1	8	4.20	−0.96	−0.58	1.07	1.66	2.06	3.10
	C(ARMA)2	12	24.36	−7.96	−4.11	7.58	13.05	120.77	187.28
	C(ARMA)3	12	0.78	−0.06	−0.04	0.26	0.27	0.07	0.07
	C(ARMA)4	8	3.63	−0.66	−0.41	1.00	1.53	1.44	2.51
	C(ARMA)5	8	1.51	−0.09	−0.01	0.46	0.60	0.22	0.36

Tabulka B.9: Kvalita odhadu rozptylu výběrového průměru v ARMA modelech. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 2 000, bootstrapových výběrů 500.

				vychýlení		směr. odchylka		MSE	
n	model	l	$n\mathrm{var}\,T_n$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(ARMA)1	4	4.73	−0.42	−0.25	2.35	2.27	5.71	5.23
	C(ARMA)2	6	27.81	−10.07	−8.65	11.56	13.52	234.91	257.70
	C(ARMA)3	6	1.35	0.13	0.03	0.87	0.60	0.77	0.36
	C(ARMA)4	4	4.50	−0.52	−0.60	2.10	2.02	4.69	4.43
	C(ARMA)5	4	2.11	0.09	0.05	1.18	0.97	1.40	0.94
125	C(ARMA)1	8	4.99	−0.39	−0.34	2.25	1.93	5.19	3.84
	C(ARMA)2	12	28.42	−6.35	−5.10	13.02	12.68	209.80	186.88
	C(ARMA)3	12	1.38	0.03	−0.00	0.80	0.46	0.64	0.21
	C(ARMA)4	8	4.24	−0.09	−0.07	2.07	1.69	4.28	2.86
	C(ARMA)5	8	2.23	−0.04	−0.07	1.05	0.75	1.10	0.56

Tabulka B.10: Kvalita odhadu rozptylu mediánu. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 2 000, bootstrapových výběrů 500.

Srovnání metod blokový a sítový bootstrap - neinvertibilní a heteroskedastické procesy

				vychýlení		směr. odchylka		MSE	
n	model	l	$n\mathrm{var}\,\overline{X}_n$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(OUT)1	4	3.97	−0.71	−0.72	1.12	2.36	1.76	6.07
	C(OUT)2	4	7.47	−1.08	−0.99	3.61	5.58	14.19	32.09
	C(OUT)3	5	7.53	−2.55	−1.12	2.08	4.20	10.83	18.88
125	C(OUT)1	8	4.05	−0.57	−0.56	1.11	2.15	1.56	4.94
	C(OUT)2	8	8.22	−1.17	−0.86	3.29	5.00	12.22	25.70
	C(OUT)3	10	7.69	−1.74	−0.82	2.35	3.49	8.55	12.87

Tabulka B.11: Kvalita odhadu rozptylu výběrového průměru v modelech nesplňujících teoretické předpoklady pro úspěšné aplikace metod. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 2 000, bootstrapových výběrů 500.

				vychýlení		směr. odchylka		MSE	
n	model	l	$n\mathrm{var}\,\overline{X}_n$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(OUT)1	4	5.14	−0.23	−0.87	2.63	2.30	6.97	6.06
	C(OUT)2	4	7.06	−0.07	−0.23	4.78	4.91	22.87	24.20
	C(OUT)3	5	8.97	−1.64	−1.52	4.21	4.55	20.46	23.01
125	C(OUT)1	8	5.14	−0.14	−0.52	2.35	2.17	5.55	4.97
	C(OUT)2	8	7.09	−0.21	−0.27	3.79	4.05	14.44	16.46
	C(OUT)3	10	9.57	−1.55	−1.76	4.17	3.62	19.73	16.18

Tabulka B.12: Kvalita odhadu rozptylu mediánu v modelech nesplňujících teoretické předpoklady pro úspěšné aplikace metod. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 2 000, bootstrapových výběrů 500.

Srovnání metod blokový a sítový bootstrap - odhad celého rozdělení

				vychýlení		směr. odchylka		MSE	
n	model	l	$\kappa_3(T_n)$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(MA)2	4	0.447	−0.047	−0.032	0.133	0.128	0.020	0.017
	C(MA)3	5	0.390	0.066	0.076	0.160	0.125	0.030	0.021
	C(ARMA)1	4	0.775	−0.155	−0.087	0.180	0.227	0.056	0.059
	C(ARMA)2	3	2.220	−1.094	−0.439	0.324	1.098	1.301	1.398
	C(ARMA)3	4	0.338	−0.017	−0.032	0.113	0.092	0.013	0.010
125	C(MA)2	8	0.336	−0.040	−0.032	0.131	0.089	0.019	0.009
	C(MA)3	10	0.379	−0.053	−0.041	0.146	0.086	0.024	0.009
	C(ARMA)1	8	0.557	−0.089	−0.063	0.148	0.128	0.030	0.020
	C(ARMA)2	6	1.509	−0.489	−0.203	0.278	0.573	0.316	0.370
	C(ARMA)3	8	0.209	0.025	0.018	0.110	0.072	0.013	0.006

Tabulka B.13: Kvalita odhadu šikmosti odhadu $T_n = e^{\overline{X}_n}$ v ARMA modelech. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 500, bootstrapových výběrů 2000.

				vychýlení		směr. odchylka		MSE	
n	model	l	$\kappa_4(T_n)$	MBB	SIEVE	MBB	SIEVE	MBB	SIEVE
65	C(MA)2	4	0.343	−0.082	−0.016	0.270	0.275	0.079	0.076
	C(MA)3	5	0.174	0.177	0.226	0.383	0.318	0.178	0.152
	C(ARMA)1	4	1.316	−0.623	−0.410	0.533	0.703	0.672	0.662
	C(ARMA)2	3	9.497	−7.092	−1.397	1.761	17.618	53.393	312.331
	C(ARMA)3	4	0.108	0.056	0.056	0.226	0.192	0.054	0.040
125	C(MA)2	8	0.249	−0.107	−0.087	0.215	0.184	0.057	0.042
	C(MA)3	10	0.374	−0.196	−0.171	0.263	0.205	0.107	0.071
	C(ARMA)1	8	0.555	−0.183	−0.102	0.328	0.346	0.141	0.130
	C(ARMA)2	6	3.916	−1.942	−0.043	1.972	9.839	7.659	96.804
	C(ARMA)3	8	0.025	0.050	0.062	0.163	0.157	0.029	0.028

Tabulka B.14: Kvalita odhadu špičatosti odhadu $T_n = e^{\overline{X}_n}$ v ARMA modelech. Srovnání metody blokový a sítový bootstrap. Správné hodnoty odhadnuty metodou MC pomocí 5 000 simulací. Délka bloku volena metodou tvorby podvýběrů, simulací 500, bootstrapových výběrů 2000.