Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Lenka Blažková

Metody odhadování rozptylů statistických odhadů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.

Studijní program: Matematika
Studijní plán: Matematická statistika

2008

Ráda bych poděkovala RNDr. Zbyňku Pawlasovi, Ph.D. za cenné rady a podporu při psaní této diplomové práce. Dík patří i mým rodičům, jejichž obětavost a porozumění též nemalou měrou přispěly k jejímu dokončení.

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

V Praze dne 6. 8. 2008

Lenka Blažková

Obsah

Úvod **5**

1 Odhad rozptylu pro stacionární posloupnosti **7**
1.1 Základní pojmy 7
1.2 Subsampling 8
1.3 Jackknife a bootstrap 10
1.3.1 Jackknife 10
1.3.2 Blokový bootstrap 12
1.3.3 Vlastnosti odhadů 16
1.4 Využití teorie časových řad 18
1.5 Porovnání odhadů 20
1.5.1 Aritmetický průměr 21
1.5.2 Výběrový rozptyl 27
1.5.3 Výběrový α-kvantil 30

2 Odhad rozptylu statistiky pro prostorová data **32**
2.1 Diskrétní náhodné pole 32
2.1.1 Základní značení 32
2.1.2 Subsampling 33
2.1.3 Bootstrap 34
2.1.4 Teoretické vlastnosti odhadů 35
2.1.5 Aritmetický průměr 36
2.2 Kótovaný bodový proces 38
2.2.1 Aritmetický průměr 40

Literatura **42**

Název práce: Metody odhadování rozptylů statistických odhadů
Autor: Lenka Blažková
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.
e-mail vedoucího: zbynek.pawlas@mff.cuni.cz

Abstrakt: Tato diplomová práce popisuje a porovnává některé vybrané metody používané k odhadu rozptylu statistik pro závislá data. Pro stacionární posloupnosti představuje metody *OBS*, jackknife, blokový bootstrap a odhady využívající poznatků z teorie časových řad, které jsou založené na ,,plug-in" přístupu. Pro zvolený konečný rozsah výběru jsou jednotlivé odhady porovnány na základě jejich střední čtvercové chyby. V případě odhadu rozptylu výběrového průměru je střední čtvercová chyba určena přesným výpočtem. Pro výběrový rozptyl a výběrový kvantil jsou použity simulace. Metody užívané pro prostorová data v $\mathbb{Z}^d$ nebo v $\mathbb{R}^d$ zastupuje subsampling, zobecněný blokový bootstrap a odhad rozptylu, který vychází z odhadu autokovarianční funkce. V textu uvedené teoretické asymptotické vlastnosti odhadů jsou podmíněny dalšími předpoklady, například splněním mixing podmínek.

Klíčová slova: odhad rozptylu, závislá pozorování, jackknife, blokový bootstrap, subsampling

Title: Variance Estimation Methods for Statistical Estimates
Author: Lenka Blažková
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Zbyněk Pawlas, Ph.D.
Supervisor's e-mail address: zbynek.pawlas@mff.cuni.cz

Abstract: This thesis describes and compares some of commonly used methods of variance estimation of various statistics for dependent data. In case of stationary sequences, *OBS*, jackknife, moving block bootstrap and plug-in estimates that use information from time series theory are implemeted. The estimators are compared according to their mean squared errors. In case of variance estimation of sample mean for finite sample size is its exact value determined by a theoretical formula. Mean squared errors of variance estimators of sample variance and sample mean are based on simulation. Methods employed in case of spatial data in $\mathbb{Z}^d$ or $\mathbb{R}^d$ are represented by subsampling or generalized moving block bootstrap as well as by the estimate based on autocovariance function estimation. Theoretical asymptotical properties of different variance estimators usually require additional assumptions such as mixing conditions.

Keywords: variance estimation, dependent data, jackknife, moving block bootstrap, subsampling

Úvod

Stanovit bodový odhad neznámého parametru pouze na základě dostupných pozorování je častým statistickým problémem. Důležitým ukazatelem kvality každého odhadu je jeho rozptyl, který v mnoha případech také potřebujeme odhadnout. Právě metodami pro odhad rozptylu konkrétních odhadů se tato diplomová práce zabývá, jejím cílem je některé používané metody popsat, navzájem porovnat a demonstrovat vlastnosti jednotlivých odhadů rozptylů na vygenerovaných datech v rámci simulační studie. Stručně zmíníme způsob volby parametrů jednotlivých metod.

Bootstrap, resampling a subsampling jsou metody opírající se o následující myšlenku: Vztah mezi celou populací a pozorovanými daty se snažíme reprodukovat tak, že pozorovaná data považujeme za model dané populace. Z tohoto modelu vhodně znovu náhodně vybíráme (resample) bootstrapový vzorek, nebo pozorování rozdělíme na menší sektory (subsample) a jeden sektor vybereme, čímž získáme druhou skupinu dat představující ,,nová“ pozorování. Výhoda těchto metod je v tom, že se vyhneme omezujícím předpokladům na rozdělení náhodných veličin.

Osmdesátá léta minulého století byla obdobím zkoumání bootstrapové metody, poprvé uvedené v článku Efron (1979). Šlo o studium vlastností bootstrapu především pro nezávislá pozorování ze stejného rozdělení. Resampling pro prostorová data v $\mathbb{R}^d$ poprvé zavádí Hall (1985).

Naproti tomu v letech devadesátých si pozornost statistiků získaly resamplingové a subsamplingové metody pro závislé náhodné veličiny, např. časové řady a náhodná pole. Zpočátku byly pro resampling, respektive subsampling, používány nepřekrývající se sektory původní množiny pozorování, záhy se však dospělo k překrývajícím se sektorům, které poskytly lepší asymptotické výsledky. Metody uvedené v této práci proto užívají pouze překrývající se sektory.

První kapitola je věnována metodám odhadu rozptylu statistik pro stacionární posloupnosti. Kromě subsamplingové metody *OBS* převzaté z článku Schmeiser a kol. (1990), blokového bootstrapu a metody jackknife je zde uveden i ,,plug-in“ přístup využívající teorii časových řad. V kapitole druhé představujeme některé metody vhodné pro prostorová data. Rozlišujeme dva modely: polohy jednotlivých pozorování buď tvoří body diskrétní pravidelné mříže $\mathbb{Z}^d$, nebo jsou generovány nějakým bodovým procesem. V obou situacích asymptotické výsledky vyžadují

pouze pouze slabou závislost „vzdálených“ pozorování.

Součástí práce jsou kromě popisu jednotlivých metod a přehledu teoretických vlastností jimi získaných odhadů též výsledky praktické simulační studie umožňující porovnání uvedených metod pro daný konečný rozsah výběru. Konkrétně odhadujeme rozptyl aritmetického průměru, výběrového rozptylu, případně 75% výběrového kvantilu a porovnáváme jednotlivé metody na základě středních čtvercových chyb příslušných odhadů. Pokud data pocházejí z normálního rozdělení, je často možné střední čtvercovou chybu spočítat bez simulací.

Numerická a simulační studie byla provedena za použití softwaru R vytvořeného týmem autorů R Development Core Team. Zdrojový kód simulací včetně hodnot použitých pro nastavení náhodných generátorů je na přiloženém CD.

Kapitola 1

Odhad rozptylu statistiky pro stacionární posloupnosti

1.1 Základní pojmy

Nechť X je náhodná veličina s rozdělením P, které patří do rodiny rozdělení $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$, kde θ je reálný parametr. Posloupnost pozorování $X_1, \ldots, X_n$ budeme považovat za součást reálného stacionárního procesu $\{X_k, k \in \mathbb{Z}\}$ definovaného na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbb{P})$. Buď P_n rozdělení náhodného vektoru $(X_1, \ldots, X_n)$. Dále označme $\mathbb{E}X_k = \mu$, $\operatorname{var} X_k = \sigma^2, \sigma^2 > 0$. Autokovarianční funkce

$$R(i,j) = \operatorname{cov}(X_i, X_j) = R(i-j) = \sigma^2 r(i-j)$$

je pouze funkcí rozdílu $i-j, i, j \in \mathbb{Z}$.

Odhad hodnoty θ budeme značit

$$\hat{\theta} = t(X_1, \ldots, X_n),$$

kde t je funkce $t : \mathbb{R}^n \to \mathbb{R}$. Pro přehlednost budeme označení t používat i pro analogické funkce vektorového argumentu o $p < n$ složkách, tj. pro $t : \mathbb{R}^p \to \mathbb{R}$.

Definice 1.1.1 (fisherovsky konsistentní odhad)
Odhad $\hat{\theta}$ založený na pozorováních $X_1, \ldots, X_n$ nechť je dán statistickým funkcionálem empirické pravděpodobnostní míry $\rho_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}$ ($\delta_{(\cdot)}$ označuje Diracovu míru v příslušném bodě) definovaném na $\mathcal{P}$, tedy

$$\hat{\theta} = T(\rho_n).$$

Pokud pro funkcionál T platí $T(P) = \theta$, řekneme, že $\hat{\theta}$ je fisherovsky konsistentním odhadem parametru θ.

Poznámka:
Do třídy těchto statistik patří např. M-, L- a R-odhady parametrů polohy a měřítka.

Jako měřítko přesnosti odhadu $\hat{\theta}$ je velice často používán rozptyl $\operatorname{var}\hat{\theta}$, zaveďme pro něj označení τ. Naším cílem je odhad $\hat{\tau}$, který se využívá při konstrukci konfidenčních či predikčních intervalů, při porovnávání různých metod odhadu atd.

Definice 1.1.2 (konsistence odhadu rozptylu)
Buď $\hat{\tau}$ odhadem $\tau = \operatorname{var}\hat{\theta}$ a předpokládejme, že $\lim_{n\to\infty} n \operatorname{var}\hat{\theta} = \kappa, \kappa \in \mathbb{R}^+$. Řekneme, že odhad $\hat{\tau}$ je konsistentním odhadem τ, pokud $n\hat{\tau} \xrightarrow{P} \kappa$.

Tato kapitola popisuje některé neparametrické metody odhadu τ a jejich vlastnosti. Na neparametrické metody se soustředíme proto, že předem nepředpokládají platnost žádného konkrétního modelu struktury dat (např. modely teorie časových řad, které pro srovnání uvádíme na straně 18) a oproti parametrickým metodám mají tedy výhodu obecnějšího použití. Navíc nevyžadují složité teoretické výpočty ani u závislých pozorování, pro která se teoretické vzorce parametrických metod stávají komplikovanými a často proto v praxi nepoužitelnými.

1.2 Subsampling

Při odhadování τ lze užít metodu *OBS* (overlapping batch statistics), která k odhadu rozptylu využívá hodnoty funkce t na překrývajících se skupinách po sobě jdoucích empirických dat. Metoda *OBS* patří mezi subsamplingové metody, kterým je věnována publikace Politis a kol. (1999).

Empirickou posloupnost $X_1, \ldots, X_n$ rozdělíme do $(n-m+1)$ překrývajících se úseků po sobě jdoucích pozorování délky m, kde $2 \le m \le n-1$. V j-tém úseku jsou tedy obsažena data $X_j, \ldots, X_{j+m-1}$. Spočítáme hodnoty statistiky t v j-tém úseku

$$\hat{\theta}_j = t(X_j, \ldots, X_{j+m-1})$$

pro $j = 1, \ldots, n-m+1$. Odhad rozptylu $\hat{\theta}$ metodou *OBS* je roven

$$\hat{\tau}(m) = \frac{m}{n-m} \, \frac{1}{n-m+1} \sum_{j=1}^{n-m+1} (\hat{\theta}_j - \hat{\theta})^2.$$

Nestrannost odhadu $\hat{\tau}(m)$ zajistíme (viz Schmeiser a kol. (1990)) následujícími třemi předpoklady:
Nechť pro $j = 1, \ldots, n-m+1$ platí

$$\mathbb{E}\,\hat{\theta} = \mathbb{E}\,\hat{\theta}_1 = \cdots = \mathbb{E}\,\hat{\theta}_{n-m+1} = \theta, \tag{1.1}$$

$$\exists\, \kappa > 0 : \quad \operatorname{var}(\hat{\theta}_1) = \cdots = \operatorname{var}(\hat{\theta}_{n-m+1}) = \kappa/m, \quad \tau = \kappa/n, \tag{1.2}$$

$$\operatorname{cov}(\hat{\theta}_j, \hat{\theta}) = \tau. \tag{1.3}$$

Jsou-li tyto předpoklady splněny, lze psát

$$\mathbb{E}\,\hat{\tau}(m) = \frac{m}{n-m}\mathbb{E}\,(\hat{\theta}_j - \hat{\theta})^2 = \frac{m}{n-m}\left[\frac{n-m}{n}\operatorname{var}\hat{\theta}_j\right] = \tau.$$

Předpoklad (1.2) znamená, že τ konverguje k nule s rostoucím n, což spolu s předpokladem (1.1) dává, že $\hat{\theta}$ je konsistentním odhadem θ. Podle (1.2) je $n \operatorname{var} \hat{\theta} = \kappa$ a odhad této konstanty je roven $n\hat{\tau}(m)$. Za dodatečného předpokladu, že posloupnost $(\hat{\theta}_j - \hat{\theta})^2$ je kovariančně stacionární proces s konečným součtem autokovariancí, lze dokázat, že pokud $m/n \to 0$ při $n \to \infty$, pak je $\hat{\tau}$ konsistentní podle definice 1.1.2, detaily viz Schmeiser a kol. (1990). Tento čtvrtý předpoklad znamená, že vzdálená pozorování jsou pouze slabě korelovaná, což zaručuje, že každé další pozorování obsahuje novou informaci.

Uvedené předpoklady jsou však zřídkakdy splněny u výběrů konečného rozsahu, nicméně asymptoticky platí při dostatečné velikosti úseku m pro většinu klasicky používaných statistik. Rozptyl odhadu $\hat{\tau}(m)$ klesá s rostoucím rozsahem výběru n pro libovolné pevné m.

Příklady

- *Výběrový průměr*
  Pro odhad střední hodnoty μ používáme výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Odhad $\operatorname{var} \bar{X}$ metodou *OBS* je roven

$$\widehat{\operatorname{var} \bar{X}}(m) = \frac{m}{n-m} \frac{1}{n-m+1} \sum_{j=1}^{n-m+1} (\bar{X}_j - \bar{X})^2,$$

  kde $\bar{X}_j$ značí aritmetický průměr v j-tém úseku $\bar{X}_j = \frac{1}{m}\sum_{i=j}^{j+m-1} X_i$. Ve speciálním případě odhadu rozptylu výběrového průměru nezávislých stejně rozdělených náhodných veličin jsou předpoklady (1.1) až (1.3) splněny pro konečnou velikost úseku m i pro konečný rozsah výběru n, jedná se proto o nestranný odhad. Všimněme si, že pro $m = 1$ bychom za použití metody *OBS* dostali

$$\widehat{\operatorname{var} \bar{X}}(1) = \frac{1}{n} S^2,$$

  což je přirozený odhad $\operatorname{var} \bar{X}$ v případě, že náhodné veličiny $X_1, \ldots, X_n$ jsou nezávislé a stejně rozdělené. Symbolem S^2 značíme výběrový rozptyl.

- *Výběrový rozptyl*
  Odhadem parametru σ^2 je výběrový rozptyl $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$. Odhad rozptylu S^2 je dán předpisem

$$\widehat{\operatorname{var} S^2}(m) = \frac{m}{n-m} \frac{1}{n-m+1} \sum_{j=1}^{n-m+1} (S_j^2 - S^2)^2,$$

  kde S_j^2 je výběrový rozptyl spočtený pouze na základě dat $X_j, \ldots, X_{j+m-1}$.

- *Výběrový α-kvantil*
  Připomeňme, že pro náhodnou veličinu X je α-kvantil q_α definován následujícím způsobem

$$q_\alpha = F^{-1}(\alpha) = \inf_{q \in \mathbb{R}} \{F(q) \geq \alpha\}, \quad \alpha \in (0,1),$$

  kde F^{-1} je kvantilová funkce odpovídající distribuční funkci F náhodné veličiny X. Fisherovsky konsistentním odhadem q_α je výběrový α-kvantil

$$\hat{q}_\alpha = F_n^{-1}(\alpha) = X_{(\lceil n\alpha \rceil)},$$

  kde $X_{(j)}$ značí j-tou pořádkovou statistiku, F_n je empirická distribuční funkce odpovídající míře ρ_n a $\lceil \cdot \rceil$ je horní celá část. Odhad rozptylu $\operatorname{var} \hat{q}_\alpha$ metodou *OBS* je roven

$$\widehat{\operatorname{var}\, \hat{q}_\alpha}(m) = \frac{m}{n-m} \, \frac{1}{n-m+1} \sum_{j=1}^{n-m+1} (\hat{q}_{\alpha,j} - \hat{q}_\alpha)^2,$$

  kde $\hat{q}_{\alpha,j}$ je výběrový α-kvantil $X_{(\lceil m\alpha \rceil)}$ v j-tém úseku dat.

1.3 Jackknife a bootstrap

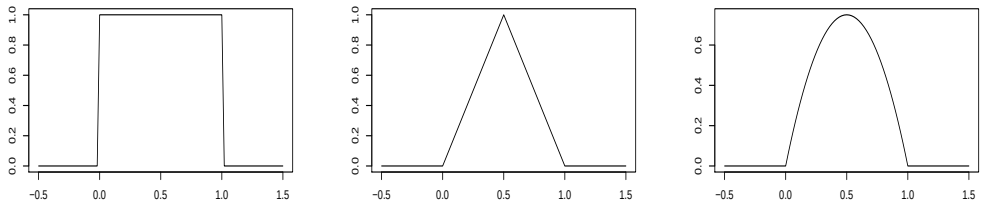
Obě metody k výpočtu odhadu τ používají hodnoty funkce t na vhodně zvolených (pod)skupinách empirických pozorování $X_1, \ldots, X_n$. Z rozdělení takto získaných statistik usuzujeme o rozdělení a vlastnostech odhadu $\hat{\theta}$. V této práci nás především zajímá rozptyl odhadu $\hat{\theta}$.

Abychom neztratili informaci o struktuře závislosti dat, spočítáme – podobně jako v případě metody *OBS* – hodnoty $\hat{\theta}$ na úsecích po sobě jdoucích pozorování. Přesnost odhadu $\hat{\tau}$ pak závisí na zvolené délce úseků a jejich počtu. Úseky dat se mohou (Künsch (1989)) nebo nemusejí překrývat (Carlstein (1986)). V této práci popíšeme pouze odhady vytvořené na základě překrývajících se úseků dat, neboť dávají lepší asymptotické výsledky. Carlsteinovy odhady užívající disjunktní úseky dat se vytvoří analogicky.

Carlstein a Künsch zobecnili metody jackknife a bootstrap pro obecnou stacionární posloupnost $X_1, \ldots, X_n$ – na rozdíl od klasického bootstrapu a jackknifu nepožadují níže popsaná zobecnění nezávislost dat. Nicméně budeme požadovat fisherovskou konsistenci odhadu $\hat{\theta}$ (viz definice 1.1.1) a také, jak později uvidíme, pouze slabě korelovaná vzdálená pozorování (např. splnění mixing podmínek).

1.3.1 Jackknife

Postupujeme tak, že z pozorovaných dat $X_1, \ldots, X_n$ spočítáme jackknifové statistiky – vždy vynecháme (anebo do výpočtu zahrneme pouze s malými vahami)



Obrázek 1.3.1: Příklady volby funkce h v (1.4): indikátor $I_{(0,1)}$, trojúhelníková a Epanečnikovova funkce.

blok po sobě jdoucích pozorování délky l. Tedy empirickou pravděpodobnostní míru a odhad $\hat{\theta}$ pro j-tý blok zapíšeme následujícím způsobem

$$\rho_n^{(j)} = (n - \|w_n\|_1)^{-1} \sum_{k=1}^{n} (1 - w_n(k-j))\delta_{X_k},$$

$$\hat{\theta}^{(j)} = T(\rho_n^{(j)}), \quad j = 0, \ldots, n-l.$$

Váhy $w_n(i)$ splňují nerovnost $0 \leq w_n(i) \leq 1, i \in \mathbb{Z}$ a $w_n(i) > 0 \iff 1 \leq i \leq l$. Norma $\|w_n\|_1$ je dána předpisem $\|w_n\|_1 = \sum_{i=1}^{l} w_n(i)$. V mnoha případech volíme váhy

$$w_n(i) = h((i - \frac{1}{2})/l), \quad 1 \leq i \leq l, \tag{1.4}$$

kde funkce $h : (0,1) \to (0,1)$ je symetrická kolem $x = \frac{1}{2}$ a rostoucí na $(0, \frac{1}{2})$. Pokud je h indikátor intervalu $(0,1)$, dostáváme speciální případ vynechání bloků délky l. Obrázek 1.3.1 znázorňuje funkce h nejčastěji používané pro tvorbu vah.

Jackknifový odhad τ je pak jednoduše výběrovým rozptylem statistik $\hat{\theta}^{(j)}$ s vhodnou standardizací (viz Künsch (1989))

$$\hat{\tau}_{jack} = (n - \|w_n\|_1)^2 n^{-1} (n - l + 1)^{-1} \|w_n\|_2^{-2} \sum_{j=0}^{n-l} (\hat{\theta}^{(j)} - \hat{\theta}^{(\cdot)})^2,$$

kde

$$\hat{\theta}^{(\cdot)} = (n - l + 1)^{-1} \sum_{j=0}^{n-l} \hat{\theta}^{(j)}, \qquad \|w_n\|_2^2 = \sum_{i=1}^{l} w_n(i)^2.$$

Příklad

- *Výběrový průměr*
  Pro výběrový průměr máme $T(P_n) = \int x_t P_n(\mathrm{d}x_1 \ldots \mathrm{d}x_n) = \mathbb{E}\, X_t = \mu$. Označme

$$\alpha_n(s) = (\|w_n\|_1)^{-1} (n - l + 1)^{-1} \sum_{j=0}^{n-l} w_n(s-j),$$

pak $\hat{\mu}_n = \sum_{s=1}^{n} \alpha_n(s) X_s$ je nestranným odhadem μ. Označme ještě v_n následující konvoluci

$$v_n(k) = \sum_{j=1}^{l-|k|} w_n(j) w_n(j+|k|), \quad |k| < l,$$

platí, že $v_n(0) = \|w_n\|_2^2$. Zaveďme funkci $\beta_n(s,k)$,

$$\beta_n(s,k) = v_n(k)^{-1}(n-l+1)^{-1} \sum_{j=0}^{n-l} w_n(s-j) w_n(s+|k|-j), \quad |k| < l.$$

Potom

$$\tilde{R}_n(k) = \sum_{s=1}^{n-|k|} \beta_n(s,k)(X_s - \hat{\mu}_n)(X_{s+|k|} - \hat{\mu}_n), \quad |k| < l,$$

je odhadem autokovarianční funkce $R(k) = \mathbb{E}\,(X_s - \mu)(X_{s+k} - \mu)$. Tento odhad je podobný výběrové autokovarianci, ale má menší vychýlení. Odhad rozptylu výběrového průměru nyní můžeme zapsat jako (viz Künsch (1989))

$$\hat{\tau}_{jack} = \widehat{\operatorname{var}\bar{X}}_{jack} = \frac{1}{n} \sum_{k=-l+1}^{l-1} \frac{v_n(k)}{v_n(0)} \tilde{R}_n(k). \tag{1.5}$$

Abychom zaručili konsistenci odhadu, volíme délku bloku $l = l(n)$ v závislosti na n tak, aby $l \to \infty$ pro $n \to \infty$.

Poznámka:
Za předpokladu $\sum_{k=-\infty}^{\infty} |R(k)| < \infty$ platí (viz Prášková (2004), str. 81), že

$$\kappa = \lim_{n\to\infty} n \operatorname{var}\bar{X}_n = \sum_{k=-\infty}^{\infty} R(k) = 2\pi f(0),$$

kde $f(\lambda),\ \lambda \in [-\pi, \pi]$, je spektrální hustota procesu $\{X_k, k \in \mathbb{Z}\}$. Odhad $\hat{\tau}_{jack}$ je konsistentním odhadem τ a ze zápisu (1.5) je patrné, že $n\hat{\tau}_{jack}$ je váženým odhadem $2\pi f(0)$. Odhad limitního rozptylu κ tak obnáší odhad spektrální hustoty v nule.

1.3.2 Blokový bootstrap

Klasický bootstrap spočívá v tom, že z dat $X_1, \ldots, X_n$ vybereme n pozorování prostým náhodným výběrem s vracením. Získáme tak skupinu bootstrapových dat $X_1^*, \ldots, X_n^*$, z nichž spočítáme hodnotu statistiky

$$\hat{\theta}^* = t(X_1^*, \ldots, X_n^*).$$

Rozdělení $\hat{\theta}^*$ slouží jako aproximace rozdělení $\hat{\theta}$. Taková aproximace má dobré vlastnosti pro nezávislá pozorování $X_1, \ldots, X_n$, pokud jsou ovšem data třeba jen slabě korelovaná, bootstrapové výběry nezachovají informaci o korelaci.

Rozdělme proto empirická data $X_1, \ldots, X_n$ do bloků po sobě jdoucích pozorování délky l, předpokládejme, že $n = kl, k \in \mathbb{N}$ (pokud rovnost neplatí, nahradíme n hodnotou $n' = \lfloor n/l \rfloor$, z hlediska asymptotických vlastností odhadu $\hat{\tau}$ můžeme $n - n'$ koncových pozorování vynechat, anebo do výpočtu odhadu $\hat{\theta}$ zahrnout i hodnotu $t(X_{n-n'+1}, \ldots, X_n)$, oba přístupy vedou ke stejným asymptotickým závěrům, viz Hall a kol. (1995)).

Künschem navržený blokový bootstrap spočívá v tom, že z těchto překrývajících se $n-l+1$ bloků vybereme k bloků – opět k tomu použijeme prostý náhodný výběr s vracením. Takto získaných k bloků dohromady tvoří jeden bootstrapový výběr $X_1^*, \ldots, X_n^*$. Bootstrapová empirická pravděpodobnostní míra je rovna

$$\rho_n^* = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=S_j+1}^{S_j+l} \delta_{X_i},$$

kde náhodné veličiny $S_1, \ldots, S_k$ jsou navzájem nezávislé a rovnoměrně rozdělené na $\{0, 1, \ldots, n-l\}$. Poznamenejme, že v případě asymetrické statistiky $\hat{\theta}$ má vliv pořadí, ve kterém k vybraných bloků pospojujeme, nicméně pro uvažované fisherovsky konsistentní odhady $\hat{\theta}$ se tomuto problému vyhneme.

Určíme hodnotu bootstrapové statistiky $\hat{\theta}^* = T(\rho_n^*)$ a rozdělení $T(\rho_n) - T(P_n)$ aproximujeme rozdělením $\hat{\theta}^* - T(\rho_n)$, hodnoty náhodných veličin $X_1, \ldots, X_n$ jsou pevné, náhodné jsou pouze indexy $S_1, \ldots, S_k$ identifikující bloky získané bootstrapovým výběrem. Rozptyl τ odhadujeme výrazem

$$\hat{\tau}_{boot} = \mathrm{var}^*(\hat{\theta}^*) = \mathbb{E}^*(\hat{\theta}^* - \mathbb{E}^*\,\hat{\theta}^*)^2,$$

v němž $\mathbb{E}^*$ označuje střední hodnotu vzhledem k $S_1, \ldots, S_k$. Pro $l = n$ dostáváme $\mathrm{var}^*(\hat{\theta}^*) = 0$, neboť $\rho_n^* = \rho_n$.

Postup obvykle opakujeme několikrát (simulujeme indexy $S_1, \ldots, S_k$) a získané empirické hodnoty $\hat{\theta}^*$ pak slouží k určení odhadu $\hat{\tau}_{boot}$. Jak uvidíme později, v některých případech je však možné $\hat{\tau}_{boot}$ vyjádřit explicitně.

Příklady

- *Výběrový průměr*
  Bootstrapovou statistiku $\hat{\theta}^*$ výběrového průměru lze zapsat jako

  $$\hat{\theta}^* = k^{-1} \sum_{i=1}^{k} U_{n,i},$$

  kde $\{U_{n,i}, i = 1, \ldots, k\}$ jsou takové nezávislé stejně rozdělené náhodné veličiny, že platí

  $$\mathbb{P}[U_{n,i} = (X_{j+1} + \cdots + X_{j+l})/l] = (n-l+1)^{-1}, \quad j = 0, \ldots, n-l.$$

Pro odhad rozptylu výběrového průměru použijeme

$$\widehat{\operatorname{var} \bar{X}}_{boot} = \operatorname{var}^*(\hat{\theta}^*) = \mathbb{E}^*(\hat{\theta}^* - \mathbb{E}^* \hat{\theta}^*)^2.$$

Podmíněnou střední hodnotu $\mathbb{E}^*\hat{\theta}^* = \mathbb{E}[\hat{\theta}^*|X_1, \ldots, X_n]$ můžeme přepsat jako

$$\begin{aligned} \mathbb{E}\, U_{n,1} &= (n-l+1)^{-1} l^{-1} \sum_{j=0}^{n-l} \sum_{s=1}^{l} X_{j+s} \\ &= (n-l+1)^{-1} l^{-1} \sum_{s=1}^{n} X_s (\min(s-1, n-l) - \max(s-l, 0) + 1) \end{aligned}$$

a odhad rozptylu výběrového průměru bude tedy roven

$$\begin{aligned} \operatorname{var}[\hat{\theta}^*|X_1, \ldots, X_n] &= k^{-1} \operatorname{var}(U_{n,1}) \\ &= k^{-1}(n-l+1)^{-1} l^{-2} \sum_{j=0}^{n-l} \left(\sum_{s=1}^{l} (X_{s+j} - \mathbb{E}\, U_{n,1}) \right)^2 . \end{aligned}$$

Poznámka:
Pokud pro jackknifový odhad $\hat{\tau}_{jack}$ použijeme váhy $w_n(i) = 1, 1 \le i \le l$, pak ve speciálním případě statistiky aritmetického průměru $\hat{\theta} = \bar{X}$ dostáváme stejný odhad jako metodou blokového bootstrapu, totiž $\hat{\tau}_{boot} = \hat{\tau}_{jack}$ (viz Künsch (1989), Theorem 3.4).

- *Výběrový rozptyl*
  Bootstrapovou statistiku $\hat{\theta}^*$ pro výběrový rozptyl získáme analogickým postupem jako u aritmetického průměru. Označme $W_{n,i}$ nezávislé stejně rozdělené náhodné veličiny, pro které platí

$$\mathbb{P}[W_{n,i} = (X_{j+1}^2 + \ldots + X_{j+l}^2)/l] = (n-l+1)^{-1}, \quad j = 0, \ldots, n-l.$$

  Dále $U_{n,i}$ nechť jsou nezávislé stejně rozdělené náhodné veličiny jako v předchozím příkladě, tedy platí

$$\mathbb{P}[U_{n,i} = (X_{j+1} + \cdots + X_{j+l})/l] = (n-l+1)^{-1}, \quad j = 0, \ldots, n-l.$$

  Označme $\bar{X}^* = \sum_{i=1}^{n} X_i^*/n$. Bootstrapový odhad výběrového rozptylu

$$\hat{\theta}^* = \frac{1}{n-1} \left(\sum_{i=1}^{n} (X_i^*)^2 - n(\bar{X}^*)^2 \right),$$

  můžeme zapsat pomocí $W_{n,i}$ a $U_{n,i}$ následujícím způsobem

$$\hat{\theta}^* = \frac{1}{n-1} \left[l \sum_{i=1}^{k} W_{n,i} - n \left(\frac{1}{k} \sum_{i=1}^{k} U_{n,i} \right)^2 \right].$$

Podmíněný rozptyl $\mathrm{var}[\hat{\theta}^*|X_1,\ldots,X_n]$ je tudíž roven výrazu (1.6)

$$\frac{1}{(n-1)^2}\,\mathrm{var}\left[l\sum_{i=1}^{k}W_{n,i}-\frac{n}{k^2}\left(\sum_{i=1}^{k}U_{n,i}\right)^2\right]. \tag{1.6}$$

Připomeňme, že jsou-li náhodné veličiny X, Y nezávislé, platí vztah $\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$. Pokud navíc mají X a Y stejné rozdělení, jsou i náhodné veličiny X^2 a Y^2 nezávislé a stejně rozdělené. S využitím tohoto vztahu lze střední hodnotu výrazu $H := l\sum_{i=1}^{k}W_{n,i} - \frac{n}{k^2}(\sum_{i=1}^{k}U_{n,i})^2$ upravit do tvaru

$$\begin{aligned} lk\,\mathbb{E}W_{n,1}-\frac{n}{k^2}\mathbb{E}(\sum_{i=1}^{k}U_{n,i})^2 &= lk\,\mathbb{E}W_{n,1}-\frac{n}{k^2}\left(\mathbb{E}\sum_{i=1}^{k}U_{n,i}^2+\mathbb{E}\sum_{i\neq j}U_{n,i}U_{n,j}\right)\\ &= lk\,\mathbb{E}\,W_{n,1}-\frac{n}{k}\left(\mathbb{E}\,U_{n,1}^2+(k-1)(\mathbb{E}\,U_{n,1})^2\right), \end{aligned}$$

neboť součet $\sum_{i\neq j}U_{n,i}U_{n,j}$ má $k(k-1)$ sčítanců a vzhledem k nezávislosti náhodných veličin $U_{n,i}$ a $U_{n,j}$ je $\mathbb{E}\,U_{n,i}U_{n,j} = (\mathbb{E}U_{n,1})^2$ pro $i \neq j$. Nyní ještě upravíme vzorec pro střední hodnotu druhé mocniny výrazu H. Umocněním dostáváme

$$l^2\mathbb{E}\left(\sum_{i=1}^{k}W_{n,i}\right)^2-2\frac{nl}{k^2}\mathbb{E}\sum_{j=1}^{k}W_{n,j}\left(\sum_{i=1}^{k}U_{n,i}\right)^2+\frac{n^2}{k^4}\mathbb{E}\left(\sum_{i=1}^{k}U_{n,i}\right)^4. \tag{1.7}$$

Platí

$$\mathbb{E}\left(\sum_{i=1}^{k}W_{n,i}\right)^2 = k\,\mathbb{E}\,W_{n,1}^2+k(k-1)(\mathbb{E}\,W_{n,1})^2, \tag{1.8}$$

$$\begin{aligned} \mathbb{E}\sum_{j=1}^{k}W_{n,j}\left(\sum_{i=1}^{k}U_{n,i}\right)^2 &= \mathbb{E}\sum_{i=1}^{k}U_{n,i}^2W_{n,i}+\mathbb{E}\sum_{i\neq j}U_{n,i}^2W_{n,j}+\\ &\quad+\mathbb{E}\sum_{i\neq j}U_{n,i}U_{n,j}W_{n,j}+\mathbb{E}\sum_{i\neq j\neq p}U_{n,i}U_{n,p}W_{n,j}=\\ &= k\,\mathbb{E}(U_{n,1}^2W_{n,1})+k(k-1)\mathbb{E}\,U_{n,1}^2\mathbb{E}\,W_{n,1}+\\ &\quad+k(k-1)\left[2\mathbb{E}\,U_{n,1}\mathbb{E}(U_{n,1}W_{n,1})+\right.\\ &\quad\left.+(k-2)(\mathbb{E}\,U_{n,1})^2\mathbb{E}\,W_{n,1}\right], \end{aligned} \tag{1.9}$$

$$\begin{aligned} \mathbb{E}\left(\sum_{i=1}^{k}U_{n,i}\right)^4 &= k\,\mathbb{E}(U_{n,1}^4)+k(k-1)\left[4\,\mathbb{E}\,(U_{n,1}^3)\mathbb{E}\,U_{n,1}+3(\mathbb{E}\,U_{n,1}^2)^2\right]+\\ &\quad+k(k-1)(k-2)\left[6\,\mathbb{E}\,U_{n,1}^2(\mathbb{E}\,U_{n,1})^2+(k-3)(\mathbb{E}\,U_{n,1})^4\right]. \end{aligned} \tag{1.10}$$

Označíme-li

$$\begin{aligned}\mathbb{E}\,W_{n,1} &= w_1, \quad \mathbb{E}\,W_{n,1}^2 = w_2,\\ \mathbb{E}\,U_{n,1}^i &= u_i, \quad i = 1,\dots,4,\\ \mathbb{E}\,U_{n,1}W_{n,1} &= uw, \quad \mathbb{E}\,U_{n,1}^2 W_{n,1} = u_0 w_0,\end{aligned}$$

můžeme předpis (1.6) pro bootstrapový odhad rozptylu statistiky S^2 přepsat ve tvaru (1.11)

$$\begin{aligned}\frac{1}{(n-1)^2}\Bigg(&-2\frac{nl}{k}\left[u_0w_0 + (k-1)\left(u_2w_1 + 2u_1uw + (k-2)u_1^2w_1\right)\right]\\ &+l^2k(w_2 - w_1^2) + \frac{n^2}{k^3}\left[u_4 + (k-1)(4u_3u_1 + 3u_2^2)\right.\\ &\left.+(k-1)(k-2)\left(6u_2u_1^2 + (k-3)u_1^4\right)\right]\\ &-\frac{n^2}{k^2}\left[u_2^2 + (k-1)\left(2u_2u_1^2 + (k-1)u_1^4\right)\right] + 2lnw_1\left(u_2 + (k-1)u_1^2\right)\Bigg).\end{aligned} \tag{1.11}$$

1.3.3 Vlastnosti odhadů

Ke zformulování vlastností odhadů rozptylu $\hat{\tau}_{jack}$ a $\hat{\tau}_{boot}$ je třeba zavést několik teoretických pojmů.

Stupeň závislosti dvou σ-algeber $\mathcal{A}_1, \mathcal{A}_2 \subseteq \mathcal{A}$ lze měřit například pomocí následující funkce

$$\alpha(\mathcal{A}_1, \mathcal{A}_2) \equiv \sup\{|\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2)| : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}. \tag{1.12}$$

Definice 1.3.1 (α-mixing podmínka, α-mixing koeficient)
Označme $\mathcal{F}_a^b$ σ-algebru generovanou množinou $\{X_i : a < i < b\}$, kde $-\infty \le a \le b \le \infty$. Definujeme α-mixing koeficienty procesu $\{X_k, k \in \mathbb{Z}\}$ jako

$$\alpha(N) = \sup\{\alpha(\mathcal{F}_{-\infty}^{k+1}, \mathcal{F}_{k+N-1}^{\infty}) : k \in \mathbb{N}\}, \quad N \in \mathbb{N},$$

kde $\alpha(\cdot,\cdot)$ je dáno předpisem (1.12). Proces $\{X_k, k \in \mathbb{Z}\}$ splňuje α-mixing podmínku, jestliže

$$\lim_{N\to\infty} \alpha(N) = 0.$$

Zaměřme se nyní na výběrový průměr. Předpokládejme, že existuje konečná $\lim_{n\to\infty} n \operatorname{var} \bar{X}_n$, a označme její hodnotu κ. Následující věta zaručuje konsistenci bootstrapového odhadu $\hat{\tau}_{boot} = \operatorname{var}^* \bar{X}^*$ parametru $\tau = \operatorname{var} \bar{X}$.

Věta 1.3.1 (konsistence odhadu $\hat{\tau}_{boot}$ pro výběrový průměr)
Nechť existuje $\delta > 0$ takové, že $\mathbb{E}|X_1|^{2+\delta} < \infty$ a $\sum_{i=1}^{\infty} \alpha(i)^{\delta/(2+\delta)} < \infty$. Pokud navíc pro velikost bloku $l = l(n)$ platí $l \to \infty$ a $l/n \to 0$ při $n \to \infty$, pak

$$n \operatorname{var}^* \bar{X}^* \xrightarrow{P} \kappa, \quad n \to \infty.$$

Důkaz: Viz Lahiri (2003), Theorem 3.1, strana 51.

Odhad $\hat{\tau}_{boot} = \operatorname{var}^* \bar{X}^*$ (a tudíž i $\hat{\tau}_{jack}$ s vahami založenými na indikátoru intervalu $[0, 1]$) je tedy konsistentním odhadem rozptylu $\operatorname{var} \bar{X}$ ve smyslu definice 1.1.2. Následující věta ukazuje, že za splnění určitých předpokladů je tento odhad asymptoticky nestranný, a dokonce udává řád vychýlení.

Věta 1.3.2 (vychýlení odhadu $\hat{\tau}_{boot}$ pro výběrový průměr)
Nechť pro nějaké $\delta > 0$ je $\mathbb{E}\,|X_1|^{6+\delta} < \infty$ a pro $l = l(n)$ platí $l \to \infty$ a $l/\sqrt{n} \to 0$ při $n \to \infty$. Nechť α-mixing koeficienty procesu $\{X_k, k \in \mathbb{Z}\}$ splňují podmínku

$$\sum_{n=1}^{\infty} n^5 \alpha(n)^{\delta/(6+\delta)} < \infty.$$

Potom

$$\mathbb{E}\,\hat{\tau}_{boot} - \tau = -\frac{1}{nl} \sum_{k=-\infty}^{\infty} |k| R(k) + o(n^{-1} l^{-1}).$$

Důkaz: Viz Künsch (1989), Theorem 3.2 nebo Lahiri (2003), Theorem 5.1 (b). Předpoklady na mixing koeficienty jsou v Lahiri (2003) o něco silnější, na druhou stranu lze však Lahiriho větu aplikovat na širší třídu odhadů. Kupříkladu je možné ji využít k získání asymptotického chování vychýlení odhadu $\hat{\tau}_{boot}$ pro $\tau = \operatorname{var} S^2$.

Velikost rozptylu $\operatorname{var} \hat{\tau}_{boot}$ je dána vztahem $\operatorname{var} \hat{\tau}_{boot} = Cn^{-3}l + o(n^{-3}l)$, kde C je poměrně komplikovaná konstanta, jejíž přesný tvar lze najít v Lahiri (2003), Theorem 5.2.

Přejděme k vlastnostem odhadu rozptylu τ metodou jackknife. Uvádíme opět asymptotický tvar vychýlení a rozptylu odhadu $\hat{\tau}_{jack}$ pro případ výběrového průměru.

Věta 1.3.3 (vychýlení odhadu $\hat{\tau}_{jack}$ pro výběrový průměr)
Nechť je konvoluce $(h * h)(x) = \int h(z) h(x - z)\,\mathrm{d}z$ dvakrát spojitě diferencovatelná v okolí bodu $x = 1$. Pokud $l = o(n^{1/3})$ a součet $\sum_{k=-\infty}^{\infty} k^2 |R(k)|$ je konečný, pak

$$\mathbb{E}\,\hat{\tau}_{jack} - \tau = \frac{1}{2nl^2} \frac{(h * h)''(1)}{(h * h)(1)} \sum_{k=-\infty}^{\infty} k^2 R(k) + o(n^{-1} l^{-2}).$$

Důkaz: Viz Künsch (1989), Theorem 3.2.

Je tedy vhodné k vytvoření vah (1.4) použít hladkou funkci h, protože pak vychýlení klesá rychleji k nule. Podmínka na součet autokovariancí vylučuje procesy se silnou závislostí vzdálených pozorování.

Věta 1.3.4 (rozptyl odhadu $\hat{\tau}_{jack}$ pro výběrový průměr)
Předpokládejme, že pro nějaké $\delta > 0$ je $\mathbb{E}\,|X_1|^{6+\delta} < \infty$ a α-mixing koeficienty procesu $\{X_k, k \in \mathbb{Z}\}$ vyhovují podmínce $\sum_{k=-\infty}^{\infty} k^2 \alpha(k)^{\delta/(6+\delta)} < \infty$. Jestliže platí $l = o(n)$, pak

$$\operatorname{var} \hat{\tau}_{jack} = \frac{l}{n^3} 2\kappa^2 \int_0^2 \frac{(h * h)(x)^2}{(h * h)(1)^2}\, \mathrm{d}x + o(n^{-3} l).$$

Důkaz: Viz Künsch (1989), Theorem 3.3.

1.4 Využití teorie časových řad

Pro lepší představu popíšeme tzv. plug-in přístup při odhadu rozptylu statistiky, a sice na jednoduchém konkrétním příkladě výběrového průměru. Rozptyl průměru vyjádříme pomocí autokovarianční funkce, kterou můžeme odhadnout neparametricky nebo parametricky. Dosazením odhadů autokovariancí do vzorce pro rozptyl dostaneme pak odhad rozptylu. Snadno nahlédneme, že pro složitější statistiku a typ závislosti dat bude odhad příslušných parametrů a teoretické vyjádření vzorce pro rozptyl mnohem komplikovanější. Navíc je třeba při volbě parametrického přístupu ověřit platnost zvoleného modelu závislosti dat.

Zřejmě

$$\begin{aligned} \operatorname{var} \bar{X} &= \frac{1}{n^2} \operatorname{var} \sum_{i=1}^{n} X_i = \frac{1}{n^2} \sum_{i,j=1}^{n} R(i-j) = \frac{1}{n^2} \sum_{k=-n+1}^{n-1} (n - |k|) R(k) = \\ &= \frac{1}{n} R(0) + \frac{2}{n^2} \sum_{k=1}^{n-1} (n-k) R(k). \end{aligned} \tag{1.13}$$

Poznámka:
Pro posloupnost nezávislých stejně rozdělených pozorování se poslední výraz redukuje na σ^2/n a odhadujeme jej hodnotou $\hat{\sigma}^2/n = \hat{R}(0)/n$.

Ze zápisu (1.13) je patrné, že k odhadu rozptylu výběrového průměru budeme potřebovat odhadnout nejprve hodnoty autokovarianční funkce dané stacionární posloupnosti. Přímá metoda spočívá v tom, že odhadneme autokovarianční funkci a potom její odhad dosadíme do upraveného vzorce (1.13) pro rozptyl výběrového průměru. Pro neparametrický odhad autokovarianční funkce používáme jeden ze dvou následujících vzorců

$$\hat{R}(k) = \frac{1}{n} \sum_{j=1}^{n-k} (X_j - \bar{X})(X_{j+k} - \bar{X}), \tag{1.14}$$

$$\hat{R}(k) \;=\; \frac{1}{n-k}\sum_{j=1}^{n-k}(X_j-\bar{X})(X_{j+k}-\bar{X}), \tag{1.15}$$

kde $k=0,\ldots,l, l<n$. První uvedený odhad autokovariance je vždy vychýlený, ale dává většinou menší střední čtvercovou chybu než druhý odhad. Ten je, pokud známe skutečnou střední hodnotu μ posloupnosti $X_1,\ldots,X_n$, nestranný.

Odhad $\operatorname{var}\bar{X}$ za použití přímé metody je tedy roven

$$\widehat{\operatorname{var}\bar{X}} = \frac{1}{n}\hat{R}(0)+\frac{2}{n^2}\sum_{k=1}^{l}(n-k)\hat{R}(k). \tag{1.16}$$

Poznámka:
Pro $l=n-1$ dává odhad rozptylu (1.16) výběrového průměru (použijeme-li druhý z odhadů autokovarianční funkce) vždy hodnotu nula. Potíž je v tom, že pro velká k nemáme dost dvojic vzdálených pozorování, abychom na jejich základě vytvořili dostatečně přesný odhad $\hat{R}(k)$. V praxi proto volíme l menší, doporučená je volba přibližně řádu $\sqrt{n}$. Na rozdíl od ostatních uvedených odhadů rozptylu τ může odhad (1.16) nabývat záporných hodnot.

Nyní budeme předpokládat už konkrétní typ závislosti dat $X_1,\ldots,X_n$. Nechť se posloupnost $X_1,\ldots,X_n$ alespoň přibližně řídí modelem AR(1) nebo MA(1), potom můžeme použít výsledky teorie časových řad týkající se těchto dvou procesů. Připomeňme, jak odhadujeme parametry každého z nich.

AR(1)
Nechť $\{X_k, k\in\mathbb{Z}\}$ tvoří autoregresní posloupnost AR(1), tj.

$$X_k = \varphi X_{k-1}+Y_k, \quad k\in\mathbb{Z},$$

kde $|\varphi|<1$ a $\{Y_k, k\in\mathbb{Z}\}$ je bílý šum $\mathrm{WN}(0,\varsigma^2)$. Pro takovou posloupnost je autokovarianční funkce rovna (viz Prášková (2004))

$$R(k) = \varsigma^2\frac{\varphi^{|k|}}{1-\varphi^2}, \quad k\in\mathbb{Z}.$$

Momentový odhad parametru φ je shodný s výběrovým autokorelačním koeficientem a je roven $\widehat{\varphi}=\hat{r}(1)=\frac{\hat{R}(1)}{\hat{R}(0)}$. První dvě hodnoty autokovarianční funkce odhadneme pomocí výběrové autokovariance (1.14)

$$\hat{R}(k) = \frac{1}{n}\sum_{j=1}^{n-k}(X_j-\bar{X})(X_{j+k}-\bar{X}), \quad k=0,1.$$

Odhad $\hat{\varsigma}^2$ vyjádříme pomocí $\hat{\varphi}$ jako

$$\hat{\varsigma}^2 = \hat{R}(0)(1-\hat{\varphi}^2).$$

Odhad $\operatorname{var}\bar{X}$ je potom roven

$$\widehat{\operatorname{var}\bar{X}} = \frac{1}{n^2}\sum_{k=-n+1}^{n-1}(n-|k|)\hat{\varsigma}^2\frac{\hat{\varphi}^{|k|}}{1-\hat{\varphi}^2} = \frac{\hat{R}(0)}{n^2}\sum_{k=-n+1}^{n-1}(n-|k|)\hat{\varphi}^{|k|}. \tag{1.17}$$

MA(1)

Nechť $\{X_k, k\in\mathbb{Z}\}$ tvoří MA(1) proces, tj.

$$X_k = Y_k + \vartheta Y_{k-1}, \quad k\in\mathbb{Z},$$

kde $|\vartheta|<1$ a $\{Y_k, k\in\mathbb{Z}\}$ je bílý šum $\mathrm{WN}(0,\varsigma^2)$. Autokovarianční funkce je v tomto případě rovna

$$\begin{aligned} R(k) &= \varsigma^2(1+\vartheta^2), \quad k=0,\\ &= \varsigma^2\vartheta, \qquad\quad |k|=1,\\ &= 0, \qquad\quad |k|>2. \end{aligned} \tag{1.18}$$

Pro účely odhadu parametrů ji nahradíme výběrovou autokovarianční funkcí (1.14). Odhad parametru ϑ dostaneme jako reálný kořen rovnice

$$\vartheta^2\hat{r}(1) - \vartheta + \hat{r}(1) = 0,$$

kde $\hat{r}(1) = \frac{\hat{R}(1)}{\hat{R}(0)}$.

Pro $|\hat{r}(1)|\le\frac{1}{2}$ dostáváme reálné $\vartheta_{1,2} = \frac{1\pm\sqrt{1-4\hat{r}^2(1)}}{2\hat{r}(1)}$, přičemž vezmeme tu hodnotu ϑ, která je v absolutní hodnotě menší než 1. Momentové odhady parametrů ϑ a ς^2 jsou

$$\hat{\vartheta} = \frac{1-\sqrt{1-4\hat{r}^2(1)}}{2\hat{r}(1)},$$

$$\hat{\varsigma}^2 = \frac{\hat{R}(0)}{1+\hat{\vartheta}^2}.$$

Odhad $\operatorname{var}\bar{X}$ je tedy roven

$$\widehat{\operatorname{var}\bar{X}} = \frac{1}{n}\hat{\varsigma}^2(1+\hat{\vartheta}^2) + 2\frac{(n-1)}{n^2}\hat{\varsigma}^2\hat{\vartheta}. \tag{1.19}$$

Pro $|\hat{r}(1)|>\frac{1}{2}$ za odhady považujeme hodnoty získané při $|\hat{r}(1)|=\frac{1}{2}$.

1.5 Porovnání odhadů

Přesnost odhadu rozptylu statistiky $\hat{\theta}$ měříme pomocí střední kvadratické chyby MSE (mean squared error), která je funkcí druhé mocniny rozdílu odhadnutého a skutečného rozptylu:

$$\mathrm{MSE}\,\hat{\tau} = \mathbb{E}(\hat{\tau}-\tau)^2.$$

V některých případech budeme schopni $\mathrm{MSE}\,\hat{\tau}$ spočítat přesně, jinde se spokojíme s aproximací

$$\frac{1}{N}\sum(\hat{\tau}-\tau)^2,$$

kde sčítáme přes N odhadů rozptylu získaných pomocí simulací. Generujeme vždy $N = 10\,000$ posloupností daného typu (AR(1), MA(1), aj.) délky $n = 100$ pro vybrané pevné hodnoty parametrů příslušného procesu.

U metod popsaných v této kapitole je třeba zvolit hodnoty příslušných parametrů každé metody. Pro *OBS* jsme volili délku úseku pozorování m rovnu 2, 5, 10 a 20. V případě výběrového průměru je optimální doporučená hodnota m řádově $n^{1/3}$ (viz Schmeiser a kol. (1990)).

Stejné hodnoty jsme přiřadili i parametru velikosti bloku l v metodách jackknife a bootstrap. Doporučená optimální délka bloku pro bootstrap v Hall a kol. (1995) je opět řádu $n^{1/3}$.

Pokud byly k stanovení MSE použity simulace, pro výpočet chyby bootstrapového odhadu $\hat{\tau}_{boot}$ bylo vždy použito 1000 výběrů s vracením.

Tučně jsou v tabulkách zvýrazněny nejmenší dosažené $\mathrm{MSE}\,\hat{\tau}$ pro konkrétní hodnoty parametrů procesu, hvězdička označuje nejlepší výsledek příslušné metody.

1.5.1 Aritmetický průměr

Začněme statistikou výběrového průměru $\hat{\theta} = \bar{X}$. Odhad $\hat{\tau}$ (kromě odhadu získaného užitím metod popsaných v části 1.4) můžeme vyjádřit pomocí vhodné matice $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ ve tvaru kvadratické formy

$$(\mathbf{X}-\mu\mathbf{1})'\mathbf{A}(\mathbf{X}-\mu\mathbf{1}), \tag{1.20}$$

kde $\mathbf{X}-\mu\mathbf{1}$ je vektor centrovaných hodnot $(X_1-\mu,\ldots,X_n-\mu)'$. Explicitní vyjádření prvků matice $\mathbf{A}$ je u jednotlivých metod poměrně komplikované, nicméně s použitím softwaru můžeme získat hodnoty $\mathbb{E}\,\hat{\tau}$ a $\mathrm{var}\,\hat{\tau}$ pro konkrétní rozsah výběru n, zvolenou autokovarianční funkci R a dané nastavení příslušných parametrů metody. Odtud je potom $\mathrm{MSE}\,\hat{\tau} = \mathrm{var}\,\hat{\tau} + (\mathbb{E}\,\hat{\tau}-\tau)^2$. Pro $\mathbb{E}\,\hat{\tau}$ dostáváme

$$\mathbb{E}(\mathbf{X}-\mu\mathbf{1})'\mathbf{A}(\mathbf{X}-\mu\mathbf{1}) = \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}(X_i-\mu)(X_j-\mu) = \sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}R(i-j).$$

Rozptyl $\mathrm{var}\,\hat{\tau} = \mathbb{E}\,\hat{\tau}^2 - (\mathbb{E}\,\hat{\tau})^2$ lze tedy rozepsat jako

$$\begin{aligned}\mathrm{var}\,\hat{\tau} = \quad & \mathbb{E}\quad \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\sum_{q=1}^{n}a_{ij}(X_i-\mu)(X_j-\mu)a_{pq}(X_p-\mu)(X_q-\mu)\\ & -\quad \left(\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}R(i-j)\right)^2.\end{aligned} \tag{1.21}$$

Pokud mají náhodné veličiny $X_1, \ldots, X_n$ mnohorozměrné normální rozdělení, lze použitím Isserlisovy formule (1.22) (viz Isserlis (1916))

$$\mu_4 = R(i-j)R(p-q) + R(i-p)R(j-q) + R(i-q)R(j-p), \tag{1.22}$$

kde μ_4 označuje střední hodnotu $\mathbb{E}\,(X_i-\mu)(X_j-\mu)(X_p-\mu)(X_q-\mu)$, vzorec (1.21) zjednodušit na

$$\operatorname{var}\hat{\tau} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\sum_{q=1}^{n} a_{ij}a_{pq}(R(i-p)R(j-q) + R(i-q)R(j-p)). \tag{1.23}$$

Tento přesný výpočet nahradíme simulacemi pouze v případě metod značených AR a MA, neboť odhady (1.17) a (1.19) nelze převést na kvadratickou formu.

Střední čtvercové chyby $\operatorname{MSE}\hat{\tau}$ jednotlivých metod získané výpočtem či simulací zaznamenáváme v tabulkách následujícím způsobem: písmena A a B označují přímou metodu využívající pro odhad $\hat{\tau}$ předpis (1.16), do nějž dosazujeme odhad autokovarianční funkce $\hat{R}(k)$ získaný na základě vzorce (1.14), respektive (1.15), s parametrem l rovným $2, 5$ a 10.

Metody AR a MA předpokládají, že se pozorování řídí modelem AR(1), respektive MA(1), odhadují příslušné parametry modelů a na jejich základě teprve určují odhad autokovarianční funkce a odhad $\hat{\tau}$, který je dán předpisem (1.17) nebo (1.19).

Metoda s označením *iid* předpokládá, že simulovaná data jsou realizací posloupnosti nezávislých stejně rozdělených pozorování a τ odhaduje jako

$$\hat{\tau} = \frac{\hat{R}(0)}{n} = \frac{1}{n^2}\sum_{j=1}^{n}(X_j - \bar{X})^2.$$

Při použití vah pro jackknife založených na indikátoru intervalu $I_{[0,1]}$ dostaneme v případě aritmetického průměru výsledky totožné s teoretickým bootstrapem. Abychom ukázali vliv volby vah na velikost $\operatorname{MSE}\hat{\tau}$, vytvářeli jsme váhy (1.4) na základě Epanečnikovovy funkce

$$h(x) = \begin{cases} \frac{3}{4}\,(1-(2x-1)^2), & x \in [0,1], \\ 0, & x \notin [0,1]. \end{cases} \tag{1.24}$$

AR(1)

Předpokládejme, že se náhodný proces $\{X_k, k \in \mathbb{Z}\}$ řídí modelem $X_k = \varphi X_{k-1} + Y_k$ s nezávislými chybami $Y_k \sim N(0, \varsigma^2)$. Pozorování $X_1, \ldots, X_n$ generujeme jako náhodné veličiny z n-rozměrného normálního rozdělení $N_n(0, \Sigma)$, kde kovarianční

$\tau = \operatorname{var} \bar{X}$	$\varphi = 0{,}2$	$\varphi = 0{,}6$		$\varphi = 0{,}2$	$\varphi = 0{,}6$
Metoda	MSE$\times 10^4$	MSE$\times 10^4$	Metoda	MSE$\times 10^4$	MSE$\times 10^4$
A 2	$*$ 0,20027	$*$ 5,52413	MA	**0,11913**	**4,27876**
A 5	0,43735	6,11821			
A 10	0,75710	10,89623	AR	1,59728	7,03971
B 2	$*$ 0,20360	$*$ 5,42056	jackknife 2	$*$ 0,16680	14,45537
B 5	0,45832	6,22716	jackknife 5	0,16701	7,39126
B 10	0,83317	11,67931	jackknife 10	0,30111	$*$ 5,92221
iid	0,30378	21,54741	jackknife 20	0,59444	9,54317
OBS 2	$*$ 0,15277	14,10639	bootstrap 2	$*$ 0,16680	14,45537
OBS 5	0,18049	6,39541	bootstrap 5	0,18881	7,14114
OBS 10	0,36126	$*$ 6,20937	bootstrap 10	0,33797	$*$ 6,78614
OBS 20	0,78908	12,19246	bootstrap 20	0,63482	10,87244

Tabulka 1.5.1: Hodnoty MSE $\hat{\tau}$ pro proces AR(1)

matice Σ je rovna

$$\Sigma = \frac{1}{1-\varphi^2} \begin{pmatrix} 1 & \varphi & \varphi^2 & \cdots & \varphi^{n-1} \\ \varphi & 1 & \varphi & \cdots & \varphi^{n-2} \\ \varphi^2 & \varphi & 1 & & \varphi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \cdots & \cdots & 1 \end{pmatrix}.$$

Pro simulaci volíme hodnoty $\varphi = 0{,}2$ a $0{,}6$, $\varsigma^2 = 1$. Hodnoty MSE $\hat{\tau}$ pro proces AR(1) jsou uvedeny v tabulce 1.5.1.

MA(1)

Dále simulujeme náhodný proces $\{X_k, k \in \mathbb{Z}\}$, kde předpokládáme platnost modelu $X_k = Y_k + \vartheta Y_{k-1}$, $Y_k \sim N(0,1)$, $k \in \mathbb{Z}$. Hodnotu parametru ϑ volíme 0,3 a 0,8. Hodnoty MSE $\hat{\tau}$ pro proces MA(1) jsou uvedeny v tabulce 1.5.2.

Náhodné veličiny, iid $N(\mu, \sigma^2)$

Předpokládáme, že proces $\{X_k, k \in \mathbb{Z}\}$ je tvořen nezávislými náhodnými veličinami s rozdělením $X_k \sim N(0, \sigma^2)$. Hodnotu parametru σ^2 volíme 1 a 1,5. Výsledné hodnoty MSE $\hat{\tau}$ zachycuje tabulka 1.5.3.

Porovnání chyb odhadů

Pro zvolená l u přímých metod chyba odhadu roste s rostoucím l. Odhad založený na (1.14) je vždy přesnější než odhad využívající (1.15). Nejnižší dosažená MSE

$\tau = \text{var}\,\bar{X}$	$\vartheta = 0{,}3$	$\vartheta = 0{,}8$		$\vartheta = 0{,}3$	$\vartheta = 0{,}8$
Metoda	MSE$\times 10^4$	MSE$\times 10^4$	Metoda	MSE$\times 10^4$	MSE$\times 10^4$
A 2	$*$ 0,23734	$*$ 0,84090	MA	**0,13392**	**0,40183**
A 5	0,51828	1,87629			
A 10	0,89143	3,25209	AR	1,70652	3,41932
B 2	$*$ 0,24161	$*$ 0,85401	jackknife 2	$*$ 0,17370	0,94383
B 5	0,54366	1,96347	jackknife 5	0,18772	$*$ 0,71574
B 10	0,98220	3,57416	jackknife 10	0,35101	1,29179
iid	0,39988	2,68928	jackknife 20	0,69538	2,55203
OBS 2	$*$ 0,15898	0,87283	bootstrap 2	$*$ 0,17370	0,94383
OBS 5	0,21048	$*$ 0,81173	bootstrap 5	0,21691	$*$ 0,85958
OBS 10	0,42469	1,56638	bootstrap 10	0,39375	1,47900
OBS 20	0,92539	3,39918	bootstrap 20	0,74095	2,74830

Tabulka 1.5.2: Hodnoty MSE $\hat{\tau}$ pro proces MA(1)

$\tau = \text{var}\,\bar{X}$	$\varsigma^2 = 1$	$\varsigma^2 = 1{,}5$		$\varsigma = 1$	$\varsigma = 1{,}5$
Metoda	MSE$\times 10^4$	MSE$\times 10^4$	Metoda	MSE$\times 10^4$	MSE$\times 10^4$
A 2	$*$ 0,09440	$*$ 0,21159	MA	0,05932	0,13348
A 5	0,19144	0,43073			
A 10	0,32059	0,72134	AR	1,01206	2,27715
B 2	$*$ 0,09632	$*$ 0,21672	jackknife 2	$*$ 0,02979	$*$ 0,06704
B 5	0,20215	0,45483	jackknife 5	0,06280	0,14130
B 10	0,35580	0,80057	jackknife 10	0,12271	0,27610
iid	**0,01990**	**0,04478**	jackknife 20	0,24441	0,54991
OBS 2	$*$ 0,03060	$*$ 0,06886	bootstrap 2	$*$ 0,02979	$*$ 0,06704
OBS 5	0,07171	0,16136	bootstrap 5	0,06729	0,15141
OBS 10	0,14927	0,33585	bootstrap 10	0,13159	0,29609
OBS 20	0,32507	0,73141	bootstrap 20	0,25364	0,57069

Tabulka 1.5.3: Hodnoty MSE $\hat{\tau}$ pro $X_t \sim N(\mu, \sigma^2)$

při použití přímých metod je pro pozorování z AR(1) s parametrem $\varphi = 0{,}6$ a pro pozorování z MA(1) s parametrem $\vartheta = 0{,}8$ srovnatelná s nejlepšími výsledky ostatních použitých metod. Pro zbylé uvažované procesy jsou odhady rozptylu odvozené přímými metodami nepatrně horší než nejpřesnější získané subsamplingové a resamplingové odhady.

Postup, kdy z dat odhadneme parametr φ autoregresní posloupnosti AR(1) a tento odhad dosadíme do vzorce (1.17), nedává ve srovnání s ostatními postupy dobré výsledky. Takto získaný odhad má asi desetkrát větší MSE než ostatní postupy. Pouze v případě, že pozorování skutečně pocházejí z AR(1) a jsou

silně závislá ($\varphi = 0{,}6$) dává chybu srovnatelnou s ostatními metodami. Problém je nejspíš v nepřesnosti odhadu $\hat{\varphi}$, možným vylepšením by bylo do součtu (1.17) zahrnout méně členů.

Pro data splňující modely AR(1) a MA(1) má nejmenší střední kvadratickou chybu odhad metodou MA. Dosažená MSE je řádově 10^{-5}, kromě dat z AR(1) s parametrem $\varphi = 0{,}6$, kdy je řádu 10^{-4}.

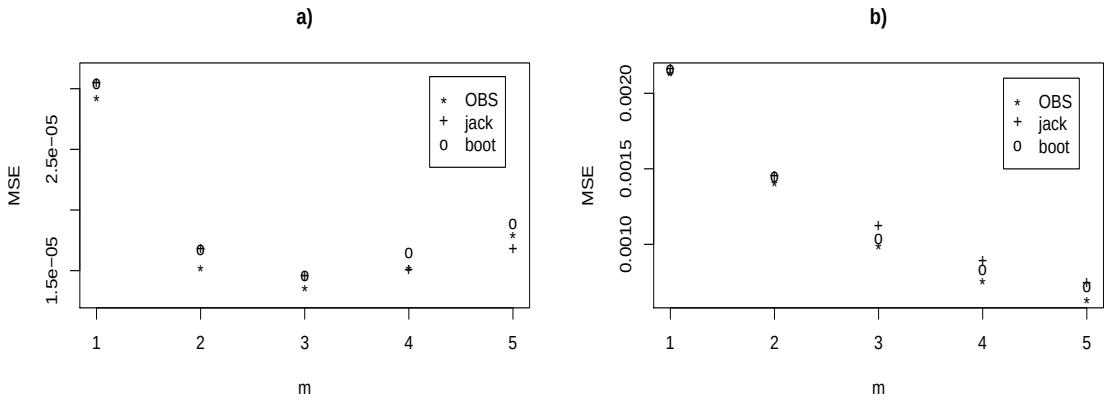
Data z MA(1) se vyznačují korelacemi pouze malého dosahu a podobně data z AR(1) při nastavení parametru $\varphi = 0{,}2$ mají korelace vyšších řádů nevýrazné, proto z nich odvozené resamplingové a subsamplingové odhady dávají menší chybu při nízké hodnotě parametru $m = 2$ nebo $m = 5$. Naproti tomu pro data z AR(1) s $\varphi = 0{,}6$ nejlépe vychází volba $m = 10$, v této situaci má samozřejmě největší MSE metoda *iid*, která nepočítá se závislostí dat.

Pro nezávislá data z normálního rozdělení je nejlepší metoda *iid*, dosažená MSE je řádu 10^{-6}. Ostatní metody mají větší MSE, protože se snaží do odhadu $\hat{\tau}$ promítnout závislost dat, což je v tomto případě nežádoucí. Proto lépe vycházejí subsamplingové a resamplingové metody s kratší délkou bloku $m = 2$.

Většinou je při pevné hodnotě m metoda jackknife s Epanečnikovovými vahami (1.24) nepatrně lepší než bootstrap (a tedy i než jackknife s vahami založenými na indikátoru intervalu $[0, 1]$).

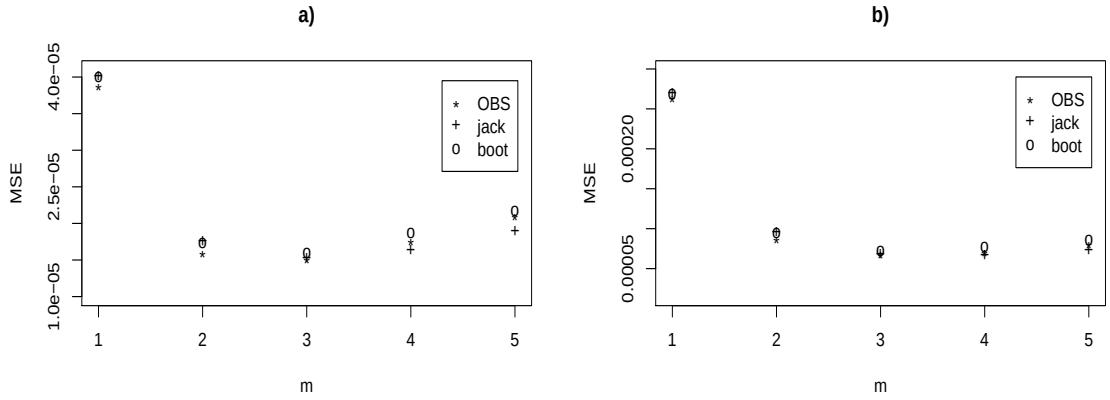
Při volbě metody odhadu pro konkrétní situaci je třeba si uvědomit, že bootstrap má podstatně vyšší časovou složitost než ostatní uvedené metody, proto doporučujeme dát přednost metodám *OBS* nebo jackknife.

Jaký vliv má zvolená délka bloku m na velikost chyby $\mathrm{MSE}\,\hat{\tau}$ pro metody *OBS*, jackknife a blokový bootstrap? Závislost $\mathrm{MSE}\,\hat{\tau}$ na m znázorníme graficky alespoň pro $m \in \{1, \ldots, 5\}$. Pro $m = 1, 2$ jsou velikosti chyb všech tří jmenovaných metod většinou téměř shodné (pokud tomu tak není, nejmenší chybu dává metoda *OBS*).

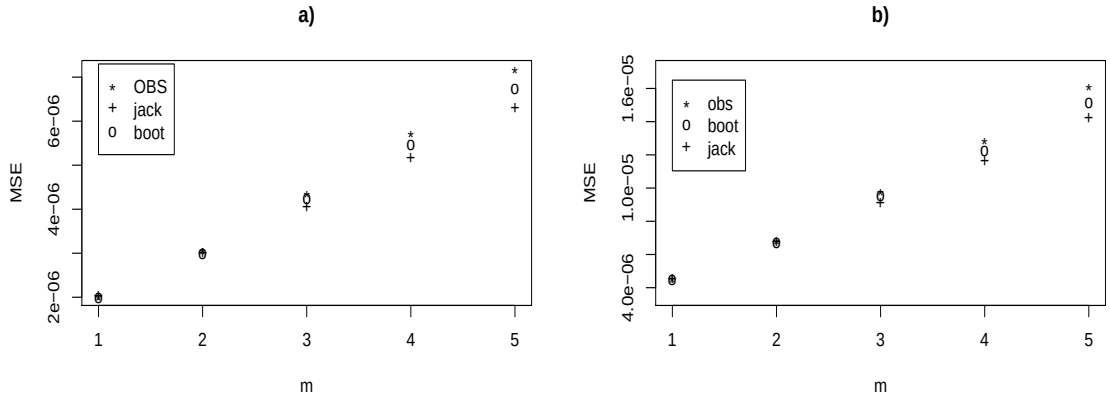


Obrázek 1.5.2: $\mathrm{MSE}\,\hat{\tau}$ v závislosti na m pro AR(1): a) $\varphi = 0{,}2$, b) $\varphi = 0{,}6$.

Pro závislá data (viz obrázky 1.5.2 a 1.5.3) dosahuje $\mathrm{MSE}\,\hat{\tau}$ v bodě $m = 1$ lokálního maxima. S rostoucím m nejprve střední čtvercová chyba rychle klesá,



Obrázek 1.5.3: MSE $\hat{\tau}$ v závislosti na m pro MA(1): a) $\vartheta = 0{,}3$, b) $\vartheta = 0{,}8$.



Obrázek 1.5.4: MSE $\hat{\tau}$ v závislosti na m pro nezávislé náhodné veličiny s rozdělením $N(0, \sigma^2)$: a) $\sigma^2 = 1$, b) $\sigma^2 = 1{,}5$.

od $m = 3$ má mírně klesající (obrázek 1.5.2 b), téměř konstantní (obrázek 1.5.3 b), anebo dokonce mírně rostoucí průběh (např. obrázek 1.5.2 a). Nejmenší střední čtvercovou chybu dávají v konkrétních případech metody jackknife (pozorování z MA(1)) nebo *OBS* (pozorování z AR(1)).

Co se týká nezávislých veličin z normálního rozdělení, s rostoucím m roste MSE $\hat{\tau}$ každé ze tří uvažovaných metod přibližně lineárně, přitom nejmenší chybu dává metoda jackknife a největší *OBS* – viz obrázek 1.5.4. Přirozeně tedy nejlépe vycházejí metody, které předpokládají co nejslabší závislost dat (ideální nastavení parametru m je tedy $m = 1$).

1.5.2 Výběrový rozptyl

Při odhadu rozptylu $\tau = \operatorname{var} S^2$ budeme k určení $\operatorname{MSE} \hat{\tau}$ používat výhradně simulace. Budeme potřebovat teoretické vyjádření $\operatorname{var} S^2$. Statistiku S^2 můžeme obecně vyjádřit ve tvaru kvadratické formy (1.20) s maticí $\mathbf{A}$ danou následujícím předpisem

$$\mathbf{A} = \begin{pmatrix} \frac{1}{n} & -\frac{1}{n(n-1)} & \cdots & -\frac{1}{n(n-1)} \\ -\frac{1}{n(n-1)} & \frac{1}{n} & \cdots & -\frac{1}{n(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & \cdots & \frac{1}{n} \end{pmatrix}.$$

Pro rozptyl $\operatorname{var} S^2$ platí vzorec (1.23), v němž symbol $\hat{\tau}$ nahradíme statistikou S^2. Dosadíme-li do (1.23) za $a_{ij}, i, j = 1, \ldots, n$, dostáváme

$$\operatorname{var} S^2 = \frac{2}{(n-1)^2} \quad \Big[\sum_{i,j=1}^{n} \left(R(i-j)\right)^2 - \tfrac{2}{n} \sum_{i,j,k=1}^{n} R(i-k)R(j-k) + \tfrac{1}{n^2} \left(\sum_{i,j=1}^{n} R(i-j)\right)^2 \Big]. \tag{1.25}$$

Tento vztah pro $\operatorname{var} S^2$ nám stačí, neboť znovu generujeme data, pro něž platí modely AR(1) či MA(1) s gaussovským bílým šumem, popřípadě jde přímo o data z normálního rozdělení. Obecné vyjádření $\operatorname{var} S^2$ jako funkce parametrů modelu a momentů vektoru chyb (nejvýše čtvrtého řádu) lze nalézt v Sharma (1986).

Ve výsledných tabulkách značí A a B opět odhady získané na základě odhadů autokovarianční funkce (1.14) a (1.15) s parametrem $l = 2, 5, 10$, které dosazujeme do vzorce (1.25).

Metoda *iid* předpokládá nezávislé stejně rozdělené veličiny z normálního rozdělení a odhaduje $\operatorname{var} S^2 = \frac{2\varsigma^4}{n-1}$ výrazem

$$\hat{\tau} = \frac{2}{n-1} \hat{R}(0)^2 = \frac{2}{n^2(n-1)} \left(\sum_{j=1}^{n} (X_j - \bar{X})^2 \right)^2.$$

Pro jackknife používáme tentokrát váhy (1.4) založené na indikátoru intervalu $[0, 1]$. Střední čtvercová chyba $\operatorname{MSE} \hat{\tau}$ pro teoretický bootstrap s označením t.boot je v případě výběrového rozptylu spočtena na základě vzorce odvozeného na straně 16.

AR(1)

V případě procesu AR(1) platí pro autokovarianční funkci vztah

$$R(i-j) = \varsigma^2 \frac{\varphi^{|i-j|}}{1-\varphi^2}, \quad i, j \in \mathbb{Z},$$

$\tau = \operatorname{var} S^2$	$\varphi = 0{,}2$	$\varphi = 0{,}6$		$\varphi = 0{,}2$	$\varphi = 0{,}6$
Metoda	MSE$\times 10^4$	MSE$\times 10^4$	Metoda	MSE$\times 10^4$	MSE$\times 10^4$
A 2	$*$ 0,72402	$*$ 32,151	iid	**0,45871**	28,741
A 5	0,86728	50,916	AR	0,61506	34,167
A 10	1,15434	64,725	MA	0,59833	**17,001**
B 2	$*$ 0,73334	$*$ 33,314			
B 5	0,89444	54,406			
B 10	1,25939	71,992			
OBS 2	1,61408	42,278	jackknife 2	$*$ 0,84838	$*$ 26,047
OBS 5	$*$ 1,10827	28,480	jackknife 5	1,11184	35,895
OBS 10	1,58125	$*$ 27,041	jackknife 10	1,42217	49,814
OBS 20	2,53999	52,741	jackknife 20	1,93496	58,367
t.boot 2	$*$ 0,81219	$*$ 26,503	bootstrap 2	$*$ 0,80796	$*$ 27,008
t.boot 5	1,02573	31,782	bootstrap 5	1,01707	31,731
t.boot 10	1,31390	38,310	bootstrap 10	1,30413	37,759
t.boot 20	1,82121	45,965	bootstrap 20	1,81109	45,318

Tabulka 1.5.4: Hodnoty MSE $\hat{\tau}$ pro proces AR(1)

proto vzorec (1.25) pro $\operatorname{var} S^2$ můžeme přepsat jako

$$\frac{2\varsigma^4}{(n-1)^2(1-\varphi^2)^2}\left[\sum_{i,j=1}^{n}(\varphi^{|i-j|})^2 - \frac{2}{n}\sum_{i,j,k=1}^{n}\varphi^{|i-j|}\varphi^{|i-k|} + \frac{1}{n^2}\left(\sum_{i,j=1}^{n}\varphi^{|i-j|}\right)^2\right].$$

Toto vyjádření použijeme například pro výpočet odhadu metodou AR.

Proces AR(1) generujeme způsobem popsaným na straně 22, hodnotu parametru φ volíme opět 0,2 a 0,6. V tabulce 1.5.3 jsou uvedeny hodnoty MSE $\hat{\tau}$ získané simulační studií.

MA(1)

Pro proces MA(1) upravíme vzorec (1.25) do následujícího tvaru

$$\frac{2\varsigma^4}{(n-1)^2}\left[n(1+\vartheta^2)^2 + 2(n-1)\vartheta^2 - \frac{2}{n}\left(n(1+\vartheta)^2 - 2\vartheta\right) + \left((1+\vartheta)^2 - \frac{2}{n}\vartheta\right)^2\right],$$

který kromě výpočtu MSE $\hat{\tau}$ využívá i metoda MA. Parametr ϑ volíme stejně jako pro případ aritmetického průměru, viz strana 23. Hodnoty MSE $\hat{\tau}$ jsou zapsány v tabulce 1.5.5.

Náhodné veličiny, iid $N(\mu, \sigma^2)$

Generujeme nezávislé náhodné veličiny z rozdělení $N(0, \sigma^2)$, hodnoty parametru σ^2 volíme opět $\sigma^2 = 1$ a $\sigma^2 = 1{,}5$. Hodnoty MSE $\hat{\tau}$ uvádíme v tabulce 1.5.6.

$\tau = \text{var}\, S^2$	$\vartheta = 0{,}3$	$\vartheta = 0{,}8$		$\vartheta = 0{,}3$	$\vartheta = 0{,}8$
Metoda	MSE$\times 10^4$	MSE$\times 10^4$	Metoda	MSE$\times 10^4$	MSE$\times 10^4$
A 2	$*$ 1,03647	$*$ 11,298	iid	**0,65757**	9,708
A 5	1,23194	13,246	AR	0,93255	13,070
A 10	1,63706	17,145	MA	0,87782	**8,119**
B 2	$*$ 1,05209	$*$ 11,559			
B 5	1,27306	13,786			
B 10	1,78956	18,797			
OBS 2	$*$ 1,03541	15,789	jackknife 2	$*$ 1,21105	$*$ 11,814
OBS 5	1,48851	$*$ 11,820	jackknife 5	1,59680	15,664
OBS 10	2,21577	19,204	jackknife 10	2,02522	19,627
OBS 20	3,68776	32,493	jackknife 20	2,73222	25,176
t.boot 2	$*$ 1,16299	$*$ 11,656	bootstrap 2	$*$ 1,16453	$*$ 11,790
t.boot 5	1,47345	14,326	bootstrap 5	1,46185	14,240
t.boot 10	1,87618	17,834	bootstrap 10	1,86943	17,773
t.boot 20	2,60002	23,562	bootstrap 20	2,57223	23,240

Tabulka 1.5.5: Hodnoty MSE $\hat{\tau}$ pro proces MA(1)

$\tau = \text{var}\, S^2$	$\sigma^2 = 1$	$\sigma^2 = 1{,}5$		$\sigma^2 = 1$	$\sigma^2 = 1{,}5$
Metoda	MSE$\times 10^4$	MSE$\times 10^4$	Metoda	MSE$\times 10^4$	MSE$\times 10^4$
A 2	$*$ 0,40303	$*$ 2,04033	iid	**0,33700**	**1,70607**
A 5	0,48701	2,46552	AR	0,35987	1,82185
A 10	0,67363	3,41023	MA	0,35916	1,81824
B 2	$*$ 0,40500	$*$ 2,05032			
B 5	0,49888	2,52557			
B 10	0,73499	3,72088			
OBS 2	7,38690	37,39617	jackknife 2	$*$ 0,59218	$*$ 2,99790
OBS 5	1,53249	7,75825	jackknife 5	0,71071	3,59799
OBS 10	$*$ 1,39016	$*$ 7,03769	jackknife 10	0,92113	4,66324
OBS 20	1,96733	9,95963	jackknife 20	1,31314	6,66478
t.boot 2	$*$ 0,57989	$*$ 2,93571	bootstrap 2	$*$ 0,57134	$*$ 2,89239
t.boot 5	0,69834	3,53533	bootstrap 5	0,69324	3,50953
t.boot 10	0,90754	4,59441	bootstrap 10	0,90064	4,55950
t.boot 20	1,30153	6,58899	bootstrap 20	1,29180	6,53975

Tabulka 1.5.6: Hodnoty MSE $\hat{\tau}$ pro $X_t \sim N(\mu, \sigma^2)$

Porovnání chyb odhadů

Střední čtvercová chyba odhadů A a B opět roste se zvyšující se hodnotou parametru l, odhad metodou A je vždy přesnější než odhad metodou B. Kromě

situace, kdy pozorování splňující model $AR(1)$ s parametrem $\varphi = 0{,}6$, mají odhady rozptylu statistiky S^2 přímými metodami menší MSE než resamplingové a subsamplingové odhady.

Metoda AR dává vždy větší střední čtvercovou chybu než metody MA a *iid*, nicméně vyjma procesů AR(1) s $\varphi = 0{,}6$ a MA(1) s $\vartheta = 0{,}8$ je tato chyba menší než chyby přímých, resamplingových a subsamplingových metod.

Nejlepší výsledky jsme získali při použití metod MA (proces AR(1) s parametrem $\varphi = 0{,}6$ a proces MA(1) s parametrem $\vartheta = 0{,}8$) a *iid* (ostatní simulované procesy).

Porovnáme-li teoretickou MSE blokového bootstrapu s MSE spočtenou na základě simulací náhodných výběrů s vracením při stejné volbě parametru l, zjistíme, že se téměř neliší a že hodnota teoretické MSE je vždy vyšší. Vzhledem k vyšší časové složitosti simulace bootstrapových výběrů doporučujeme k výpočtu MSE použít spíše vzorec (1.11).

Střední čtvercová chyba jackknifových odhadů je v naprosté většině případů větší než chyba bootstrapových odhadů se stejnou délkou úseku l. Jak u metody jackknife, tak u blokového bootstrapu roste velikost MSE s rostoucí délkou úseku l, nejlépe tedy vychází odhady s parametrem $l = 2$.

Jinak je tomu u metody *OBS* – zde není volba délky úseku m jednoznačná. Pro nezávislá pozorování z normálního rozdělení a pro proces AR(1) s parametrem $\varphi = 0{,}6$ je MSE nejmenší při $m = 10$. Ve zbylých třech simulovaných případech vychází nejlépe volby $m = 2$ a $m = 5$. Nejnižší dosažená MSE za použití metody *OBS* je (až na proces MA(1) s parametrem $\vartheta = 0{,}3$) větší než chyby nejlepších bootstrapových a jackknifových odhadů.

1.5.3 Výběrový α-kvantil

Uvažujme proces $\{X_k, k \in \mathbb{Z}\}$, který se řídí modelem MA(3)

$$X_k = Y_k + \vartheta_1 Y_{k-1} + \vartheta_2 Y_{k-2} + \vartheta_3 Y_{k-3}$$

s exponenciálním rozdělením chyb $Y_k \sim \mathrm{Exp}(1)$. Hodnoty parametrů stanovíme takto: $\vartheta_1 = 0{,}8, \vartheta_2 = 0{,}5$ a $\vartheta_3 = 0{,}4$.

Pro odhad α-kvantilu použijeme odhad

$$\hat{q}_\alpha = t(X_1, \ldots, X_n) = F_n^{-1}(\alpha) = X_{(\lceil n\alpha \rceil)}.$$

Na základě generovaných dat odhadujeme rozptyl α-kvantilu pro $\alpha = 0{,}75$. Abychom určili $\mathrm{MSE}\,\hat{\tau}$, aproximovali jsme teoretickou hodnotu

$$\tau = \mathrm{var}\,\hat{q}_\alpha$$

na základě $p = 10\,000$ simulací procesu MA(3) s výše uvedenými parametry výrazem

$$\frac{1}{p-1} \sum_{i=1}^{p} (X^i_{(\lceil n\alpha \rceil)} - \bar{X}_{(\lceil n\alpha \rceil)})^2, \tag{1.26}$$

kde $X^i_{(\lceil n\alpha \rceil)}$ je hodnota statistiky $t_i := t(X_{1,i}, \ldots, X_{n,i})$ v i-té generované posloupnosti a $\bar{X}_{(\lceil n\alpha \rceil)}$ je aritmetický průměr $\sum_{i=1}^{p} t_i/p$.

Ke konstrukci odhadu $\operatorname{var} \hat{\tau}$ používáme metody *OBS*, jackknife a bootstrap. Hodnoty parametrů m, respektive l, jsme znovu volili $2, 5, 10$ a 20.

Abychom mohli použít pro jackknifový odhad kvantilu váhy (1.4), vytvořili jsme vlastní funkci pro výpočet α-kvantilu. Tato funkce vychází z empirické distribuční funkce F_n odpovídající míře ρ_n. Empirická distribuční funkce F_n je po částech konstantní se skoky v bodech $X_1, \ldots, X_n$. Velikosti skoků upravíme právě vahami (1.4) na

$$\frac{1 - w_n(i)}{n - \|w_n\|_1}, \quad i = 1, \ldots, n.$$

Odhadem α-kvantilu je pak pořádková statistika $X_{(k)}$, kde $k \leq n$ je nejmenší index splňující nerovnost

$$\sum_{i=1}^{k} \frac{1 - w_n(i)}{n - \|w_n\|_1} \geq \alpha.$$

U metody jackknife používáme tři různé váhové funkce. Očíslování v tabulce odpovídá pořadí příslušných funkcí h na obrázku 1.3.1 (strana 11). Výsledky simulace najdeme v tabulce 1.5.7.

Porovnání chyb odhadů

Při odhadu 75% kvantilu dává nejmenší střední čtvercovou chybu metoda *OBS* a sice při délce úseku $m = 5$. Shodně s metodou *OBS* dává chybu řádu 10^{-3} i bootstrap. Jackknife je za použití libovolných uvažovaných vah podstatně horší než *OBS* a bootstrap, přičemž při dané pevné délce úseku l je vždy nejlepší odhad využívající váhy založené na indikátoru intervalu $[0, 1]$. Pro jackknife je lepší volit větší délku úseku $l = 20$, bootstrap má naopak nejmenší MSE pro $l = 2$. Možné zlepšení jackknifového odhadu by nejspíš přinesla vhodná standardizace (původní standardizace byla navržena pro odhad $\operatorname{var} \bar{X}$).

$\tau = \operatorname{var} \hat{q}_\alpha$	$l = m = 2$	$l = m = 5$	$l = m = 10$	$l = m = 20$
OBS	0,00882	∗**0,00364**	0,00567	0,00754
jackknife 1	0,03717	0,03009	0,01784	∗**0,01300**
jackknife 2	0,27798	0,05263	0,06027	∗**0,02708**
jackknife 3	0,16861	0,07780	0,06538	∗**0,02822**
bootstrap	∗**0,00508**	0,00541	0,00707	0,00857

Tabulka 1.5.7: Hodnoty MSE $\hat{\tau}$ pro proces MA(3)

Kapitola 2

Odhad rozptylu statistiky pro prostorová data

2.1 Diskrétní náhodné pole

2.1.1 Základní značení

Uvažujme množinu mřížových bodů $\mathbb{Z}^d$, $d \in \mathbb{N}$. Nechť $\{X_j, j \in \mathbb{Z}^d\}$ je striktně stacionární náhodné pole a θ je reálný parametr, jehož hodnotu odhadujeme na základě pozorování $\mathbb{X}_{D_n} = \{X_j, j \in D_n\}$, $D_n \subseteq \mathbb{Z}^d$ jako

$$\hat{\theta} = t(\mathbb{X}_{D_n}) = t(X_j, j \in D_n).$$

Nechť $A \subseteq (-\frac{1}{2}, \frac{1}{2}]^d$ je otevřená souvislá množina obsahující počátek souřadnic a A_0 je taková podmnožina $\mathbb{R}^d$, že $A \subseteq A_0 \subseteq \bar{A}$, kde $\bar{A}$ je označení pro uzávěr množiny A. Dále buď $\{\lambda_n\}_{n\in\mathbb{N}}$, $\lambda_n \in [1, \infty)$ rostoucí posloupnost reálných čísel taková, že $\lim_{n\to\infty} \lambda_n = \infty$. Nechť je množina A_n dána tvarem množiny A_0 a konstantou λ_n udávající poměr zvětšení, totiž

$$A_n = A_0\lambda_n. \tag{2.1}$$

Nechť množina A_0 splňuje následující podmínku:

Podmínka 2.1.1 :
Nechť je pro každou kladnou reálnou posloupnost $\{a_n\}_{n\in\mathbb{N}}$ s $\lim_{n\to\infty} a_n = 0$ počet d-rozměrných krychlí, které protínají množinu A_0 i její doplněk A_0^C, daných předpisem $a_n(i + [0, 1)^d), i \in \mathbb{Z}^d$, řádu $O([a_n]^{-(d-1)})$ při $n \to \infty$.

Podmínka 2.1.1 je například pro $d = 2$ splněna, když A_0 tvoří vnitřek jednoduché uzavřené křivky konečné délky.

Předpokládejme, že pro množinu D_n platí $D_n = A_n \cap \mathbb{Z}^d$. Důležitým důsledkem podmínky 2.1.1 pro pozorování ležící na mřížce $\mathbb{Z}^d$ je, že počet pozorování ležících

blízko hranice A_n je zanedbatelný v porovnání s celkovým počtem všech pozorování $\mathbb{X}_{D_n}$.

Zaměříme se opět na odhad rozptylu

$$\tau = \operatorname{var} t(\mathbb{X}_{D_n}) = \mathbb{E}(t(\mathbb{X}_{D_n}) - \mathbb{E}[t(\mathbb{X}_{D_n})])^2. \tag{2.2}$$

Symbolem $|D|$ budeme značit počet bodů množiny D, je-li D konečnou podmnožinou $\mathbb{Z}^d$. Pro $A \subseteq \mathbb{R}^d$ kladné Lebesgueovy míry bude $|A|$ představovat objem množiny A.

2.1.2 Subsampling

Metodu subsampling uvedenou v Sherman (1996) lze použít při konstrukci neparametrických odhadů teoretických momentů statistiky $\hat{\theta}$, my popíšeme její verzi pro rozptyl.

Postupujeme tak, že množinu mřížových bodů D_n rozdělíme na překrývající se sektory D_n^i, $i = 1, \ldots, k$ stejného tvaru jako D_n, kde $\beta_n < \lambda_n$ určuje velikost jednotlivých sektorů a $k = k(n)$ je jejich počet. Předpokládáme, že $\beta_n \to \infty$ a $\beta_n/\lambda_n \to 0$ pro $n \to \infty$.

Přesněji $D_n = A_0\lambda_n \cap \mathbb{Z}^d$ je podmnožinou d-rozměrné krychle

$$\tilde{A}_n = \left(-\frac{\lambda_n}{2}, \frac{\lambda_n}{2}\right]^d$$

se středem v počátku souřadné soustavy a hranami délky λ_n. Krychli $\tilde{A}_n$ rozdělíme na $(2\left\lfloor \frac{\lambda_n}{2} \right\rfloor - \beta_n + 1)^d$ překrývajících se d-rozměrných krychlí

$$\tilde{A}_i = i + \beta_n \mathcal{U}, \quad i \in \mathbb{Z}^d : \tilde{A}_i \subseteq \tilde{A}_n,$$

kde $\mathcal{U} = (0, 1]^d$ je jednotková krychle v $\mathbb{R}^d$. Označme $m_n = \lambda_n/\beta_n$. Uvnitř každé d-rozměrné krychle $\tilde{A}_i$ určíme množinu D_n^i, která je m_n-krát zmenšeným a vhodně posunutým obrazem množiny D_n. Protože data byla pozorována pouze na množině D_n, použijeme v dalším pouze ty sektory D_n^i, pro něž platí $D_n^i \subseteq A_n$. Bez újmy na obecnosti předpokládejme, že tuto vlastnost mají právě sektory D_n^i, $i = 1, \ldots, k$.

Na základě pozorování $\{X_j, j \in D_n^i\}$ v i-tém sektoru stanovíme hodnotu statistiky $\hat{\theta}_i = t(\mathbb{X}_{D_n^i})$ a parametr τ subsamplingovou metodou odhadujeme jako

$$\hat{\tau}_n = \frac{1}{k|D_n|} \sum_{i=1}^{k} |D_n^i|(\hat{\theta}_i - \bar{\theta})^2, \quad \bar{\theta} = k^{-1} \sum_{i=1}^{k} \hat{\theta}_i. \tag{2.3}$$

2.1.3 Bootstrap

Metodu z knihy Lahiri (2003), kterou nyní představíme, lze považovat za přímé rozšíření blokového bootstrapu z předchozí kapitoly.

Buď $\{\beta_n\}_{n\geq 1}$ posloupnost přirozených čísel taková, že

$$\beta_n/\lambda_n \to 0, \quad \beta_n \to \infty, \quad n \to \infty.$$

Prvním krokem bude rozdělení množiny mřížových bodů D_n pomocí d-rozměrných krychlí o objemu β_n^d. Označme

$$\mathcal{K}_n = \{k \in \mathbb{Z}^d : \beta_n(k + \mathcal{U}) \cap D_n \neq \emptyset\}$$

množinu všech bodů k takových, že d-rozměrná krychle $\beta_n(k + \mathcal{U})$ má s D_n neprázdný průnik.

Bootstrapovou verzi procesu $\mathbb{X}_{D_n} = \{X_j, j \in D_n\}$ vytvoříme na základě bootstrapových verzí jednotlivých sektorů

$$D_{n,k} = D_n \cap [\beta_n(k + \mathcal{U})], \quad k \in \mathcal{K}_n.$$

Označíme-li ještě $\mathcal{I}_n = \{i \in \mathbb{Z}^d : i + \beta_n \mathcal{U} \subseteq D_n\}$ množinu všech mřížových bodů, které tvoří levý dolní roh některé z d-rozměrných krychlí o objemu β_n^d uvnitř D_n, lze potom množinu překrývajících se d-rozměrných krychlových sektorů uvnitř D_n zapsat jako $\{i + \beta_n \mathcal{U} : i \in \mathcal{I}_n\}$.

Každou d-rozměrnou krychli $\beta_n(k+\mathcal{U}), k \in \mathcal{K}_n$ původního dělení při bootstrapu nahradíme nezávisle náhodně vybranou d-rozměrnou krychlí $i + \beta_n \mathcal{U}, i \in \mathcal{I}_n$. Bootstrapovou verzi sektoru $D_{n,k}$ zavádíme následujícím předpisem

$$D^*_{n,k} = [I_k + \beta_n \mathcal{U}] \cap [D_{n,k} - k\beta_n + I_k], \quad k \in \mathcal{K}_n,$$

kde I_k jsou nezávislé stejně rozdělené náhodné vektory s pravděpodobnostím rozdělením

$$P(I_1 = i) = \frac{1}{|\mathcal{I}_n|}, \quad i \in \mathcal{I}_n.$$

Výchozí sektor $D_{n,k}$ a jeho bootstrapová verze $D^*_{n,k}$ mají tedy stejný tvar i velikost a pozorování $\mathbb{X}_{D^*_{n,k}} = \{X_j, j \in D^*_{n,k}\}$ zachovávají původní strukturu závislosti dat $\mathbb{X}_{D_{n,k}}$. Pro každé $k \in \mathcal{K}_n$ je $|\mathbb{X}_{D^*_{n,k}}| = |\mathbb{X}_{D_{n,k}}|$.

Konečně bootstrapovou verzi $\mathbb{X}^*_{D_n}$ původních pozorování $\mathbb{X}_{D_n}$, dostaneme zřetězením všech $D^*_{n,k}$. Bootstrapový odhad parametru θ je pak

$$\hat{\theta}^* = t(\mathbb{X}^*_{D_n}).$$

Rozptyl $\tau = \operatorname{var} \hat{\theta}$ odhadujeme jako

$$\hat{\tau}_{boot}(\beta_n) = \mathbb{E}^*(\hat{\theta}^* - \mathbb{E}\,\hat{\theta}^*)^2, \tag{2.4}$$

kde $\mathbb{E}^*$ je střední hodnota vzhledem k $I_k, k \in \mathcal{K}_n$.

Označíme-li množinu indexů d-rozměrných krychlí o objemu β_n^d uvnitř D_n symbolem

$$\mathcal{K}_{1n} = \{k \in \mathcal{K}_n : \beta_n(k + \mathcal{U}) \subseteq D_n\}$$

a množinu indexů přesahujících krychlí

$$\mathcal{K}_{2n} = \{k \in \mathcal{K}_n : \beta_n(k + \mathcal{U}) \cap D_n^C \neq \emptyset\},$$

kde D_n^C je doplněk množiny D_n, pak pokud je $\hat{\theta} = \bar{X}$, lze odhad rozptylu $\tau = \operatorname{var} \bar{X}$ vyjádřit v následujícím tvaru (viz Lahiri (2003), strana 291)

$$\begin{aligned}
\hat{\tau}_{boot} &= |\mathbb{X}_{D_n}|^{-2}|\mathcal{I}_n|^{-1}\left(|\mathcal{K}_{1n}|\sum_{i\in\mathcal{I}_n} S_n(i;0)^2 + \sum_{k\in\mathcal{K}_{2n}}\sum_{i\in\mathcal{I}_n} S_n(i;k)^2 - |\mathbb{X}_{D_n}|^2 \hat{\mu_n}^2\right), \\
\hat{\mu}_n &= |\mathbb{X}_{D_n}|^{-1}|\mathcal{I}_n|^{-1}\left(|\mathcal{K}_{1n}|\sum_{i\in\mathcal{I}_n} S_n(i;0) + \sum_{k\in\mathcal{K}_{2n}}\sum_{i\in\mathcal{I}_n} S_n(i;k)\right),
\end{aligned}$$

kde $S_n(i;k)$ je součet všech pozorování z množiny $[D_{n,k} - k\beta_n + i] \cap [i + \beta_n\mathcal{U}]$ pro $i \in \mathcal{I}_n$, $k \in \mathcal{K}_n$.

2.1.4 Teoretické vlastnosti odhadů

Střední hodnotu náhodného pole $\{X_j, j \in \mathbb{Z}^d\}$ označme $\mu = \mathbb{E}\, X_j$.

Definice 2.1.1 (α-mixing koeficient)
Nechť $S \subseteq \mathbb{R}^d$ a označme $\mathcal{F}_S$ σ-algebru generovanou náhodnými veličinami $\{X_s : s \in S \cap \mathbb{Z}^d\}$. Dále nechť

$$\alpha(S_1, S_2) = \{|\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2)| : A_1 \in \mathcal{F}_{S_1}, A_2 \in \mathcal{F}_{S_2}\}, \quad S_1, S_2 \subseteq \mathbb{R}^d.$$

Definujeme α-mixing koeficient náhodného pole jako

$$\alpha(a;b) = \sup\{\alpha(S_1, S_2) : \max(|S_1|, |S_2|) \leq b,\ d(S_1, S_2) \geq a\}, \quad a > 0,\ b \geq 1,$$

kde $d(S_1, S_2)$ je vzdálenost množin definovaná jako $\inf\{\|x - y\|, x \in S_1, y \in S_2\}$.

Věta 2.1.1 (konsistence odhadu rozptylu pro subsampling v $\mathbb{Z}^2$)
Buď $D_n \subseteq \mathbb{Z}^2$. Nechť

$$\sup_{b\geq 1}\left\{\frac{\alpha(a;b)}{b}\right\} \leq K \cdot a^{-\varepsilon},$$

kde $\varepsilon > 2$, a $K > 0$. Dále nechť je posloupnost $\{|D_n|^2 t(\mathbb{X}_{D_n})^4, n \in \mathbb{N}\}$ stejnoměrně integrovatelná. Potom

$$\hat{\tau}_n \xrightarrow{L_2} \tau, \quad n \to \infty,$$

kde $\hat{\tau}_n$ je dáno předpisem (2.3).
Důkaz: Viz Sherman (1996), Theorem 2, strana 513.

Věta 2.1.2 (konsistence odhadu $\hat{\tau}_{boot}$ pro výběrový průměr)
Označme z_0 počátek souřadnic $(0, \ldots, 0) \in \mathbb{Z}^d$. Nechť $\mathbb{E}|X_{z_0}|^{6+\delta} < \infty$ a nechť pro α-mixing koeficienty $\alpha(a; b)$ náhodného pole $\{X_j, j \in \mathbb{Z}^d\}$ platí

$$\alpha(a; b) \leq C \cdot a^{-c_1} b^{-c_2}, \quad a, b \geq 1,$$

pro nějaké $\delta > 0$, $c_1 > 5d(6 + \delta)/\delta$ a $0 \leq c_2 \leq c_1/d$. Pak

$$|A_n|\, \hat{\tau}_{boot}(\beta_n) \xrightarrow{P} \kappa, \quad n \to \infty,$$

kde je $\hat{\tau}_{boot}(\beta_n)$ odhad definovaný (2.4) a $\kappa = \sum_{i \in \mathbb{Z}^d} \mathrm{cov}(X_i, X_{z_0})$.

Důkaz: Nalezneme na straně 296 v Lahiri (2003).

Poznámka:
Lahiri formuluje větu obecněji. Jeho verzi lze použít k odvození konsistence odhadu $\hat{\tau}_{boot}$ pro výběrový rozptyl.

2.1.5 Aritmetický průměr

Simulace jsme omezili pouze na případ stacionárního gaussovského náhodného pole realizovaného v $\mathbb{Z}^2$. Při simulaci pozorování $\mathbb{X}_{D_n}$ vycházíme z předpokladu, že autokovarianční funkce R má exponenciální tvar a je dána parametrem $\varphi \in (0, 1)$ a vzdáleností dvojice pozorování, totiž

$$\mathrm{cov}(X_i, X_j) = R(\|i - j\|) = \varphi^{\|i-j\|}, \quad i, j \in \mathbb{Z}^d,$$

kde $\|\cdot\|$ je euklidovská vzdálenost.

Pro snadnou implementaci volíme $D_n = (-\frac{n}{2}, \frac{n}{2}]^2 \cap \mathbb{Z}^d$ vzniklou sjednocením mřížových bodů ve čtverci[1] o straně délky $n = 8$. Máme tedy k dispozici 64 mřížových bodů, v nichž pozorujeme bodový proces. Pro mřížové body $i, j \in D_n$ je hodnota $\|i - j\| \in \Delta \equiv \{0, 1, \sqrt{2}, 2, \sqrt{5}, \ldots, (n-1)\sqrt{2}\}$.

Počet simulovaných náhodných polí je 4000, hodnotu parametru φ jsme volili 0,4 a 0,8.

Nechť

$$\hat{\theta} = \bar{X} = \frac{1}{|D_n|} \sum_{j \in D_n} X_j.$$

[1] O vlivu zvoleného tvaru množiny D_n v $\mathbb{Z}^d$ na vlastnosti odhadu a o optimální volbě příslušného parametru β_n pojednávají například Nordman a Lahiri (2004).

K určení MSE $\hat{\tau}$ jednotlivých odhadů $\hat{\tau}$ využijeme teoretickou hodnotu rozptylu $\tau = \operatorname{var}\hat{\theta}$, která je rovna

$$\frac{1}{|D_n|^2}\operatorname{var}\sum_{j\in D_n} X_j = \frac{1}{|D_n|^2}\sum_{i,j\in D_n} R(\|i-j\|) = \frac{1}{|D_n|^2}\sum_{i,j\in D_n} \varphi^{\|i-j\|} = \frac{1}{|D_n|^2}\sum_{\delta\in\Delta} p(\delta)\varphi^{\delta}, \tag{2.5}$$

kde $p(\delta)$ je počet uspořádaných dvojic $(i,j) \in D_n$ takových, že $\|i-j\| = \delta$.

Přímou metodou, v tabulce značenou písmenem A, stanovíme odhad $\hat{\tau}$ rozptylu $\tau = \operatorname{var}\hat{\theta}$ tak, že do vzorce pro rozptyl výběrového průměru (2.5) dosadíme hodnoty empirické autokovarianční funkce $\hat{R}$, pro niž platí následující vztah

$$\hat{R}(\delta) = \frac{1}{p(\delta)}\sum_{i,j\in D_n:\,\delta(i,j)=\delta} (X_i - \bar{X})(X_j - \bar{X}), \quad \delta\in\Delta,\ \delta < l, \tag{2.6}$$

kde $l \le (n-1)\sqrt{2}$. Použijeme-li $l = (n-1)\sqrt{2}$, je takto získaný odhad rozptylu roven

$$\frac{1}{|D_n|^2}\sum_{i,j}\hat{R}(i-j) = 0,$$

proto volíme l menší, konkrétně $l = \sqrt{2(n-1)}$.

Metoda *iid* předpokládá nezávislá data. Využívá vzorec (2.5) a odhaduje tedy rozptyl τ jako

$$\hat{\tau} = \frac{1}{|D_n|^2}\sum_{i\in D_n}\hat{R}(0) = \frac{1}{|D_n|}\hat{R}(0).$$

Bootstrapové odhady jsou konstruovány na základě 650 výběrů s vracením. Velikost strany β_n čtverců D_n^i pro subsampling a $D_{n,k}$ pro bootstrap jsme volili $2, 3$ a 4. Výsledky simulací jsou uvedeny v tabulce 2.1.1.

Porovnání odhadů rozptylu

V obou simulovaných případech dává metoda subsampling nejmeší MSE při nastavení parametru $\beta_n = 4$, tj. tehdy, když jednotlivé $\hat{\theta}_i$ ve vzorci (2.3) počítáme na základě šestnácti pozorování. Bootstrap naopak dává pro libovolnou velikost β_n chybu přibližně stejnou. Upřednostnili bychom opět $\beta_n = 4$, neboť výpočet

MSE $\hat{\tau}$ pro $\varphi = 0,4$							
A	iid	subs 2	subs 3	subs 4	boot 2	boot 3	boot 4
0,00350	0,00497	0,00356	0,00187	∗**0,00134**	∗0,00357	0,00365	0,00365
MSE $\hat{\tau}$ pro $\varphi = 0,8$							
A	iid	subs 2	subs 3	subs 4	boot 2	boot 3	boot 4
0,14164	0,18270	0,17136	0,15521	∗**0,13803**	0,17139	0,16988	∗0,16247

Tabulka 2.1.1: Hodnoty MSE $\hat{\tau}$ pro gaussovské náhodné pole v $\mathbb{Z}^2$.

odpovídající chyby je nejméně časově náročný. Vhodnější z obou metod je potom subsampling, neboť jeho MSE je menší než střední čtvercová chyba bootstrapových odhadů.

Přímá metoda odhadu rozptylu dává menší nebo stejnou chybu jako bootstrapový odhad, ale je horší než subsampling. Má ovšem navíc ještě tu nevýhodu, že jí stanovený odhad $\hat{\tau}$ nemusí být vždy nezáporný. Největší MSE má metoda *iid*.

2.2 Kótovaný bodový proces

Nechť $\{X_s, s \in \mathbb{R}^d\}$ je striktně stacionární náhodné pole dimenze $d \in \mathbb{N}$. Reálné náhodné veličiny X_s jsou indexované pomocí spojitého parametru s. Předpokládejme, že pozorujeme realizaci náhodného pole v konečně mnoha nepravidelně rozmístěných bodech, tj. souřadnice bodů tvoří homogenní bodový proces N na množině $A_n \subseteq \mathbb{R}^d$. Pozorovaná data lze chápat jako realizaci kótovaného bodového procesu

$$\mathbb{X}_{N,A_n} = \{(s_j, X_{s_j}), j = 1, \ldots, N(A_n)\},$$

kde $N(A_n)$ je (náhodný) rozsah výběru. Nechť $A \subseteq [0,1]^d$ je kompaktní množina a

$$A_n = nA = \{y : y = ns, s \in A\} \subseteq [0,n]^d$$

označuje oblast, kde pozorujeme kótovaný bodový proces.

Hodnotu parametru θ odhadujeme pomocí statistiky $\hat{\theta} = t(\mathbb{X}_{N,A_n})$. Opět se zaměříme na odhad rozptylu

$$\tau = \operatorname{var} \hat{\theta} = \operatorname{var} t(\mathbb{X}_{N,A_n}).$$

Definice 2.2.1 (konsistence odhadu rozptylu)
Nechť $\lim_{n\to\infty} |A_n| \operatorname{var} \hat{\theta} = \kappa$, $\kappa \in \mathbb{R}^+$. Řekneme, že odhad $\hat{\tau}$ je konsistentním odhadem rozptylu $\tau = \operatorname{var} \hat{\theta}$, jestliže

$$|A_n|\, \hat{\tau} \xrightarrow{P} \kappa.$$

Pro $c \in (0,1)$ označme $B_n := A_{\lfloor cn \rfloor}$, množina B_n má tedy stejný tvar jako A_n, ale je menší. Uvažujme všechna posunutí

$$B_n + y = \{b + y : b \in B_n\},$$

která jsou podmnožinou A_n, platí tedy

$$y \in Y_n^c := \{y \in A_n : B_n + y \subseteq A_n\}.$$

Za předpokladu, že množina A_n je dostatečně velká a zároveň máme i dostatek podoblastí vzniklých posunutím množiny B_n, tj. $c \to 0$, ale $cn \to \infty$, můžeme z variability $t(\mathbb{X}_{N,B_n+y}) = t(X_s, s \in B_n + y)$ získat odhad rozptylu $\tau = \operatorname{var} t(\mathbb{X}_{N,A_n})$ jako

$$\hat{\tau} := \frac{1}{|A_n|} \int_{Y_n^c} |B_n| \{t(\mathbb{X}_{N,B_n+y}) - \bar{t}(\mathbb{X}_{N,B_n})\}^2 \mathrm{d}y / |Y_n^c|, \tag{2.7}$$

kde

$$\bar{t}(\mathbb{X}_{N,B_n}) = \int_{Y_n^c} t(\mathbb{X}_{N,B_n+y}) \mathrm{d}y / |Y_n^c|. \tag{2.8}$$

Pro odvození teoretických vlastností odhadu rozptylu statistiky $\hat{\theta}$ budeme stejně jako v případě stacionárních posloupností požadovat, aby náhodné pole $\{X_s, s \in \mathbb{R}^d\}$ splňovalo určité podmínky slabé závislosti – a sice, že závislost klesá se vzdáleností indexů náhodných veličin.

Definice 2.2.2 (α-mixing podmínka)
Označme

$$\begin{aligned} \alpha(k;l) = \quad & \sup\{|\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2)| : A_i \in \mathcal{F}(E_i), \quad i = 1,2, \\ & E_2 = E_1 + s, \quad |E_1| = |E_2| \le l, \quad d(E_1, E_2) \ge k\}, \end{aligned}$$

kde d je vzdálenost dvou množin v $\mathbb{R}^d$, E_1 a E_2 jsou kompaktní a konvexní podmnožiny $\mathbb{R}^d$, $\mathcal{F}(E)$ značí σ-algebru generovanou náhodnými veličinami $\{X_s, s \in E\}$ a $|E|$ je Lebesgueova míra množiny E. Řekneme, že $\{X_s, s \in \mathbb{R}^d\}$ splňuje α-mixing podmínku, jestliže platí $\alpha(l; l^d) \to 0$ při $l \to \infty$.

Definice 2.2.3 (ρ-mixing podmínka)
Označme

$$\begin{aligned} \rho(k;l) = \quad & \sup\{\operatorname{corr}(\varepsilon_1, \varepsilon_2) : \mathbb{E}|\varepsilon_i|^2 < \infty, \quad i = 1,2, \quad E_2 = E_1 + s, \\ & |E_1| = |E_2| \le l, \quad d(E_1, E_2) \ge k\}, \end{aligned}$$

kde ε je $\mathcal{F}_N(E)$-měřitelná funkce a $\mathcal{F}_N(E)$ je σ-algebra generovaná náhodnými body procesu N, které padnou do množiny E.
Bodový proces N splňuje ρ-mixing podmínku, jestliže $\rho_N(l; l^d) \to 0$ při $l \to \infty$.

Věta 2.2.3 Odhad $\bar{t}(\mathbb{X}_{B_n})$ z (2.8) je L_2-konsistentním odhadem hodnoty $\gamma = \lim_{n\to\infty} \mathbb{E}\, t(\mathbb{X}_{N,A_n})$, $\gamma \in \mathbb{R}$, jestliže jsou splněny α- i ρ-mixing podmínka a pro všechna n platí

$$\mathbb{E}|t(\mathbb{X}_{N,A_n})|^{2+\delta} \le C_\delta < \infty,$$

pro nějaké $\delta > 0$ a nějakou konstantu C_δ.

Z konvergence $\bar{t}(\mathbb{X}_{B_n}) \xrightarrow{L_2} \gamma$ lze za určitých podmínek získat konsistenci $\hat{\tau}$ (viz Politis a Sherman (2001)).

Věta 2.2.4 Nechť $\lim_{n\to\infty} |A_n| \operatorname{var} \hat{\theta} = \kappa$. Odhad $\hat{\tau}$ je L_2-konsistentním odhadem, tj.

$$|A_n|\, \hat{\tau} \xrightarrow{L_2} \kappa,$$

jestliže jsou splněny α- i ρ-mixing podmínka a pro všechna n platí

$$\mathbb{E}|t(\mathbb{X}_{N,A_n})|^{4+\delta} \leq C'_\delta < \infty,$$

pro nějaké $\delta > 0$ a nějakou konstantu C'_δ.

2.2.1 Aritmetický průměr

Pro praktický výpočet musíme ovšem integrál v definici $\hat{\tau}$ (2.7) aproximovat konečným součtem. Nabízejí se dva přístupy, jak postupovat.

(i) **Deterministická aproximace** spočívající v rozdělení oblasti Y_n^c např. pravidelnou pravoúhlou mřížkou obsahující k menších útvarů. Integrál $\bar{t}(\mathbb{X}_{N,B_n})$ pak aproximujeme příslušným riemannovským součtem $\bar{t}_{k,R}(\mathbb{X}_{N,B_n})$ přes tyto oblasti.

(ii) **Metoda Monte Carlo** nebo stochastická aproximace, kdy z množiny Y_n^c náhodně vybereme k bodů $y_1, \ldots, y_k$ (náhodné veličiny $y_1, \ldots, y_k$ jsou nezávislé a mají stejné rovnoměrné rozdělení na Y_n^c) a integrál $\bar{t}(\mathbb{X}_{N,B_n})$ aproximujeme průměrem $\bar{t}_{k,MC}(\mathbb{X}_{N,B_n}) = \frac{1}{k}\sum_{i=1}^{k} t(\mathbb{X}_{N,B_n+y_i})$. Při $k \to \infty$ je tato aproximace konsistentní.

Obě aproximace lze použít, pokud máme k dispozici dostatečně velký počet bodů, které zahrnujeme do výpočtu přibližné hodnoty. V našich simulacích jsem využili první z uvedených postupů, kdy jsme aproximaci integrálu spočetli za použití pravidelné čtvercové mřížky s $k = (\frac{1}{8}(1-c)n)^2$ oblastmi.

K určení MSE $\hat{\tau}$ potřebujeme znát teoretickou hodnotu $\tau = \operatorname{var} \hat{\theta}$. Pokud jsou kóty nezávislé se střední hodnotou μ a rozptylem σ^2 a nezávislé na bodovém procesu N, pak pro aritmetický průměr je $\mathbb{E}\,\hat{\theta} = \mu$ a $\operatorname{var} \hat{\theta} = \sigma^2\, \mathbb{E}\, \frac{1}{N(A_n)} \cdot I_{\{N(A_n)>0\}}$. Pro závislé kóty, podobně jako v případě kvantilu (viz 1.5.3), aproximujeme teoretickou hodnotu $\operatorname{var} \hat{\theta}$ pomocí $p = 10\,000$ realizací náhodných polí v A_n generovaných výše popsaným způsobem, totiž

$$\tau \approx \frac{1}{p-1} \sum_{j=1}^{p} (\bar{X}_{N_i(A_n)} - \bar{X})^2,$$

kde $\bar{X}_{N_i(A_n)}$ označuje výběrový průměr v i-tém generovaném souboru dat, jenž má náhodný rozsah $N(A_n, i)$, a $\bar{X} = \frac{1}{p}\sum_{i=1}^{p} \bar{X}_{N_i(A_n)}$ je průměrná hodnota výběrových průměrů přes všechny generované soubory. Tuto teoretickou hodnotu τ uvádíme ve výsledné tabulce spolu s průměrnými hodnotami odhadů $\hat{\tau}$ obou použitých metod.

$\tau = \operatorname{var} \bar{X}$	$c = 0{,}2$	$c = 0{,}3$	$c = 0{,}4$	$c = 0{,}6$	*iid*
$R \equiv 0$					
$\operatorname{MSE} \hat{\tau} \times 10^{-4}$	0,32021	$*$0,31700	0,33082	0,61327	**0,06390**
$\frac{1}{J} \sum \hat{\tau}$	0,01575	0,01347	0,01072	0,00571	0,01265
Teoretická hodnota τ je 0,01266.					
$R(i-j) = \varphi^{\|i-j\|}$					
$\operatorname{MSE} \hat{\tau} \times 10^{-4}$	98,492	83,066	$*$**81,133**	104,076	125,211
$\frac{1}{J} \sum \hat{\tau}$	0,024246	0,03352	**0,03588**	0,02298	0,01123
Teoretická hodnota τ je 0,12308.					

Tabulka 2.2.2: Hodnoty $\operatorname{MSE} \hat{\tau}$ pro kótovaný bodový proces v $\mathbb{R}^2$

.

Pro účely simulace se opět omezíme na stacionární gaussovské náhodné pole v rovině $\mathbb{R}^2$ a odhad aritmetického průměru. Pozorovaná data ohraničíme čtvercem $A_n = [0, n] \times [0, n]$ o straně délky $n = 10$. Pozorování v A_n jsou realizací Poissonova procesu s intenzitou $\lambda = 0{,}8$, tedy počet pozorování má Poissonovo rozdělení s intenzitou $\lambda |A_n|$ a souřadnice jsou vytvořeny jako nezávislé náhodné veličiny z rovnoměrného rozdělení na intervalu $[0, n]$. Simulujeme dvě situace: kóty jsou buď navzájem nezávislé a mají stejné normální rozdělení $N(0, \sigma^2)$, anebo je autokovarianční funkce dána předpisem $R(i-j) = \varphi^{\|i-j\|}$. Nastavení parametrů pro simulaci volíme $\varphi = 0{,}55$ a $\sigma^2 = 1$.

Metoda *iid* předpokládá nezávislost pozorování a odhaduje rozptyl τ jako $\hat{\sigma}_j^2 / N_j$, kde N_j je počet pozorování j-té simulace a σ_j^2 je výběrový rozptyl pozorování z této simulace.

Střední čtvercové chyby uvedené v tabulce 2.2.2 jsou získány na základě $J = 10\,000$ simulací.

Porovnání odhadů roztylu

Odhad (2.7) pro nezávislé náhodné veličiny s normálním rozdělením má nejmenší MSE pro $c = 0{,}2$, přesnější je v tomto případě samozřejmě odhad rozptylu metodou *iid*. Nejlepší odhady jsou provázeny chybou řádu 10^{-5}, metoda *iid* dává dokonce chybu řádu 10^{-6}.

Pro závislá pozorování je odhad (2.7) nejpřesnější, když je $c = 0{,}4$. Ve srovnání s (2.7) metoda *iid* nedává dobrý odhad. Chyba odhadu rozptylu pro závislá data je podstatně větší než pro nezávislá a dosahuje řádu 10^{-3}. Průměrné hodnoty $\hat{\tau}$ v tabulce 2.2.2 výrazně podhonocují teoretickou hodnotu τ.

Literatura

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics 14*(3), 1171–1179.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Hall, P. (1985). Resampling a coverage pattern. *Stochastic Processes and their Applications 20*(2), 231–246.

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

Isserlis, L. (1916). On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression. *Biometrika 11*(3), 185–190.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics 17*(3), 1217–1241.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer-Verlag, New York.

Nordman, D. J. a Lahiri, S. N. (2004). On optimal subsample size for variance estimation. *The Annals of Statistics 32*(5).

Politis, D. N., Romano, J. P. a Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.

Politis, D. N. a Sherman, M. (2001). Moment estimation for statistics from marked point processes. *Journal of the Royal Statistical Society 63*, 261–275.

Prášková, Z. (2004). *Základy náhodných procesů II.* Univerzita Karlova v Praze – Nakladatelství Karolinum.

R Development Core Team (2008). *R Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: `http://www.R-project.org`. ISBN 3-900051-07-0.

Sharma, S. C. (1986). The effects of correlation among observations on the consistency property of sample variance. *Communications in Statistics - Theory and Methods 15*(4), 1125–1152.

Sherman, M. (1996). Variance estimation for statistics computed from spatial lattice data. *Journal of the Royal Statistical Society 58*(3), 509–523.

Schmeiser, B. W., Avramidis, T. N. a Hashem, S. (1990). Overlapping batch statistics. *In: Proceedings of the Winter Simulation Conference, O.Balci, R. P. Sadowski, R. E. Nance (eds)*, 395–398.