LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

NEPARAMETRISKĀ REGRESIJA LIETOJOT ORTOGONĀLAS FUNKCIJAS UN VEIVLETUS

MAĢISTRA DARBS

Autors: **Haralds Plivčs**

Stud. apl. MaSt020016

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2009

Anotācija

Darbā ir apskatītas pēdējā laikā populārākās neparametriskās regresijas funkcijas novērtēšanas metodes ar ortogonālām funkcijām un veivletiem. Katrai metodei gan teorētiski, gan konkrētām praktiskām datu problēmām apskatīta gludināšanas parametra izvēle un regresijas funkcijas novērtēšana. Ar simulāciju un datu analīzi pētītas metožu priekšrocības un trūkumi.

Atslēgas vārdi: veivletu neparametriskā regresija, ortogonālu funkciju regresija

Abstract

In this work recently popular nonparametric statistical regression estimation methods have been considered using orthogonal functions and wavelets. For both methods theoretically and practically the estimation for smoothing parameters and regression functions are considered. Finally simulation study has been made and data examples analysed.

Keywords: wavelet regression, regression with orthogonal functions

Saturs

Ievads

Mūsdienās gandrīz katrā nozarē ir novērojamas dažāda veida sakarības starp faktoriem, lietām, mainīgiem lielumiem, kā viena faktora izmaiņa ietekmē citu faktoru. Matemātiskajā statistikā svarīga loma ir klāsteru, dispersiju, regresijas, laikrindu u.c. analīzei, kas analizē dažādu faktoru ietekmi uz citiem faktoriem. Šajā darbā tiks aplūkota regresiju analīze, lietojot neparametriskās statisktikas metodes.

Kā regresijas analīzes pirmsākumu var uzskatīt 19. gadsimta sākumu, kad Ležandrs 1805.gadā un Gauss 1809.gadā publicēja "mazāko kvadrātu metodi". Vēlāk 1821.gadā Gauss publicāja arī Gausa-Markova teorēmu, kura ir ļoti svarīga klasiskās statistikas regresijas analīzē.

Regresijas analīze ir statistikas metode, ar kuras palīdzību tiek pētīta un analizēta sakarība starp atkarīgo mainīgo, kuru sauc arī par atbildes mainīgo, un vienu vai vairākiem neatkarīgajiem mainīgajiem, kurus savukārt sauc par regresoriem vai skaidrojošiem mainīgajiem.

Doti n neatkarīgi un vienādi sadalīti novērojumu pāri $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, regresijas modelis ir formā

$$Y_i = r(X_i) + \epsilon_i,\ \mathbb{E}(\epsilon_i) = 0,\ i = 1, 2, \ldots, n,$$

kur funkcija r ir nezināmā regresijas funkcija un ϵ_i ir statistiskā kļūda.
Funkcija r principā ir nosacītā matemātiskā cerība atbildes mainīgajam Y pie fiksētas X vērtības, tas ir,

$$r(x) = \mathbb{E}(Y|X = x).$$

Klasiskajā statistikā parametriskās regresijas modelī funkcija r tiek aplūkota kādā iepriekš noteiktā formā. Visbiežāk šīs funkcijas r forma tiek pieņemta kā lineāra. Piemēram, funkcija r uzdota formā $r(x) = a+bx+\epsilon$. Šajā vienādībā ir divi nezināmi parametri, kurus vajag novērtēt. Tādējādi funkcijas r novērtēšanas problēma tiek reducēta uz parametru a un b novērtēšanas problēmu. Šo parametru novērtējumus iegūst pēc mazāko kvadrātu metodes minimizējot kvadrātu atlikumu (rezidiju) summu. Tādēļ arī 19.gs. termina "regresija" vietā lietoja terminu "mazāko kvadrātu metode". Tālāk, kad parametri ir novērtēti, tos ievietojot izteiksmē $r(x) = a+bx$, iegūst funkcijas r novērtējumu $\hat{r}(x) = \hat{a}+\hat{b}x$. Beigās tiek analizēts, vai izvēlētais modelis ir atbilstošs funkcijai r.

Savukārt neparametriskajā regresijā funkcija r netiek aplūkota kādā iepriekš noteiktā formā, bet funkcija tiek novērtēta un konstruēta saskaņā ar informāciju, kuru iegūst no datiem. Tādējādi neparametriskās regresijas metodes nodrošina efektīvu un vienkāršu metodi datu kopas struktūras atrašanai. Neparametriskās regresijas metodes ietver ortogonālo rindu novērtējumus, kodolu novērtējumus, lokālos polinomus, splainus un veivletus. Visās tikko pieminētajās metodēs ir parametrs, kuru sauc par gludināšanas parametru. Kodolu novērtējumos un lokālajos polinomos tas ir joslas platums, ortogonālajās rindās tas ir saskaitāmo skaits, savukārt veivletiem tas ir slieksnis.

Ja šis gludināšanas parametrs ir izvēlēts par lielu, tad regresijas funkcija var būt pārgludināta. Un pretēji, ja gludināšanas parametrs ir izvēlēts par mazu, tad regresijas līkne var būt nenogludināta.

Tādējādi viena no neparametriskās regresijas problēmām ir optimāla gludināšanas parametra izvēle.

Līdz šim biežāk lietotajām regresijas metodēm trūkums ir "telpiskā adaptivitāte" un problēmas rodas, kad funkcijas ir telpiski nehomogēnas. Telpiski nehomogēnas funkcijas ir tādas funkcijas, kurām gludums mainās atkarībā no skaidrojošā mainīgā vērtības. Citiem vārdiem, šīs funkcijas var raksturot kā funkcijas, kurām ir novērojama pārtrauktība vai citas pēkšņas izmaiņas datu struktūrā, piemēram, lēcieni, pīķi. Tradicionālajām neparametriskās regresijas metodēm bieži radās problēma noteikt precīzi šos lēcienus, pīķus, jo šīs metodes ir bāzētas uz fiksētu telpisku mērogu. Veivletiem piemīt laika-frekvences lokalizācija. Tādējādi veivleti ir telpiski pielāgoti, un tiem nav problēmu aproksimēt negaidītus lēcienus un pīķus.

Veivletu lietojumi neparametriskajā regresijā sākās ar Donoho un Johnstone [1], kuri ieviesa "vieglo" sliekšņošanu. Vēlāk parādījās arī cita veida sliekšņi un sliekšņošanas operatori. Pašlaik šos sliekšus jau dažādi kombinē, lai iegūtu vēl labākus regresijas līknes novērtējumus.

Šī darba mērķis ir:

- aplūkot tradicionālās (līdz šim lietotās) un nesen ieviestās (pēdējos 10-15 gados) neparametriskās regresijas metodes nezināmas regresijas funkcijas novērtēšanai.

- veicot simulācijas un analizējot reālus datus ar programmu R, noteikt metožu priekšrocības un iespējamos trūkumus.

- veikt dažādu neparametrisku regresiju novērtējumu salīdzinājumu.

Darbs sastāv no 6 nodaļām. Pirmajā nodaļā ir izklāstīta parametriskā un neparametriskā regresija un to atšķirības. Otrajā nodaļā ir aplūkotas divas neparametriskās regresijas metodes - kodolu regresija un lokālā polinomu regresija. Trešajā nodaļā ir apskatīta REACT metode, kura parādījās samērā nesen - pirms 10 gadiem. Ceturtajā nodaļā ir neliels ieskats par veivletiem un aplūkota regresijas metode, kura balstīta uz veivletu lietojumiem. Piektajā nodaļā ir šo metožu salīdzināšana un analīze simulētām funkcijām un reāliem datiem. Pēdējā nodaļā ir izklāstīti secinājumi par apskatīto metožu priekšrocībām un trūkumiem. Pielikumā atrodas programmas R kodi.
Darbā ir 21 attēls un 2 tabulas.

1. Parametriskā un neparametriskā regresija

1.1. Parametriskā regresija

Parametriskās regresijas modelī regresijas funkcija r tiek aplūkota iepriekš noteiktā formā un ir definēta kā funkcija $Y = r(X, \beta)$, kur Y ir atkarīgais mainīgais, X ir viens vai vairāki neatkarīgie mainīgie un β ir nezināmi parametri.

Visizplatītākais un praksē biežāk lietotais modelis ir lineārās regresijas modelis.

Lineārās regresijas modelis. Lineāras regresijas modeli pieraksta formā

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \cdots + \beta_k X_i^{(k)} + \epsilon_i, \ i = 1, 2, \ldots, n. \tag{1.1}$$

Modelis vektoru formā

$$Y = X\beta + \epsilon,$$

kur

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_1^{(1)} & X_1^{(2)} & \ldots & X_1^{(k)} \\ 1 & X_2^{(1)} & X_2^{(2)} & \ldots & X_2^{(k)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_n^{(1)} & X_n^{(2)} & \ldots & X_n^{(k)} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{pmatrix}.$$

Matricā $\mathbf{X}$ kolonna ar vieniniekiem atbilst konstantes β_0 iekļaušanai regresijas modelī.

Lineāras regresijas modeļos svarīgi ir šādi pieņēmumi:

1. skaidrojošie mainīgie $X^{(j)}$ nav stohastiska rakstura;

2. matrica $\mathbf{X}$ ir ar rangu k+1;

3. $\mathbb{E}\epsilon_i = 0 \ \ \mathbb{E}\epsilon_i^2 = \sigma^2, \ \ \mathbb{E}\epsilon_i\epsilon_j = 0, \ \ \epsilon_i \sim N(0, \sigma^2 I)$.

Mērķis ir novērtēt modeļa parametrus $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$. Parametru novērtējumus $\hat{\beta}$ iegūst pēc mazāko kvadrātu metodes, minimizējot kvadrātu atlikumu (rezidiju) summu

$$\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{k} X_i^{(j)} \beta_j)^2.$$

Ja matricai $\mathbf{X}^T\mathbf{X}$ eksistē inversā matrica, tad mazāko kvadrātu novērtējums ir formā

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

Funkcijas $r(x)$ novērtējums punktā $X = (X_0, X_1, \ldots, X_k)^T$ ir

$$\hat{r}_n(X) = \sum_{j=0}^{k} \hat{\beta}_j X_j = \mathbf{X}^T\hat{\beta}.$$

Novērtētās vērtības $r = (\hat{r}_n(X_0), \hat{r}_n(X_1), \ldots, \hat{r}_n(X_n))$ var pierakstīt kā

$$r = \mathbf{X}\hat{\beta} = LY,$$

kur $L = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ sauc par projekcijas matricu un vektora $\hat{\epsilon} = Y - r$ elementus sauc par rezidijiem. Matrica L ir simetriska un idempotenta, tas ir, $L = L^T$ un $L^2 = L$.

1.2. Neparametriskā regresija

Neparametriskajā regresijā funkcija r netiek aplūkota kādā iepriekš noteiktā formā, bet tiek konstruēta saskaņā ar informāciju, kuru iegūst no datiem.
1980-jos un 1990-jos gados tradicionālie neparametriskās regresijas novērtējumi pārsvarā bija lineāri gludinātāji (skatīt definīciju 1) - kodolu novērtējumi, gludinošie splaini.
1990-tajos gados Donoho un Johnstone piedāvāja nelineāru aproksimāciju neparametriskajai regresijai - veivletu samazināšanas un veivletu sliekšņošanas novērtējumus.

Definīcija 1. Funkcijas r novērtējums $\hat{r}_n$ ir lineārs gludinātājs, ja katram x eksistē tāds vektors $l(x) = (l_1(x), \ldots, l_n(x))^T$, ka

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i. \tag{1.2}$$

Novērtēto vērtību vektors $r = (\hat{r}_n(X_1), \ldots, \hat{r}_n(X_n))^T$ un $Y = (Y_1, \ldots, Y_n)^T$. Tad

$$r = LY,$$

kur L ir $n \times n$ matrica, kuras i-tā rindiņa ir $l(X_i)^T$. Matricu L sauc par gludināšanas matricu. Tādējādi $L_{ij} = l_j(x_i)$. i-tās rindiņas elementi ir svari, kādi piešķirti Y_i veidojot

$\hat{r}_n(X_i)$.

Visas neparametriskās metodes vieno fakts, ka, lai novērtētu funkciju r, ir nepieciešams noteikt gludināšanas parametru. Tādējādi lielākā problēma neparametriskas regresijas novērtējumos ir pareizi izvēlēties gludināšanas parametru.

1.3. Regresijas novērtējuma precizitātes noteikšana

Gan ar parametrisko, gan ar neparametrisko regresiju tiek novērtēta regresijas funkcija r. Lai noteiktu regresijas funkcijas novērtējuma atbilstību, visbiežāk lieto zaudējuma funkciju un risku.

Definīcija 2. Funkcija $\hat{r}_n(x)$ ir funkcijas $r(x)$ novērtējums punktā x. Tad kvadrātiskās kļūdas zaudējumu funkcija tiek definēta ar izteiksmi

$$L(r(x), \hat{r}_n(x)) = (r(x) - \hat{r}_n(x))^2. \tag{1.3}$$

Definīcija 3. Zaudējuma funkcijas vidējo vērtību sauc par risku jeb vidējo kvadrātisko kļūdu (angliski *mean squared error* - MSE) un to definē ar izteiksmi

$$MSE = R(r(x), \hat{r}_n(x)) = \mathbb{E}(L(r(x), \hat{r}_n(x))). \tag{1.4}$$

Veicot pārveidojumus, iegūst, ka risks

$$R(r(x), \hat{r}_n(x)) = (\mathbb{E}(\hat{r}_n(x)) - r(x))^2 + \mathbb{D}(\hat{r}_n(x)), \tag{1.5}$$

kur

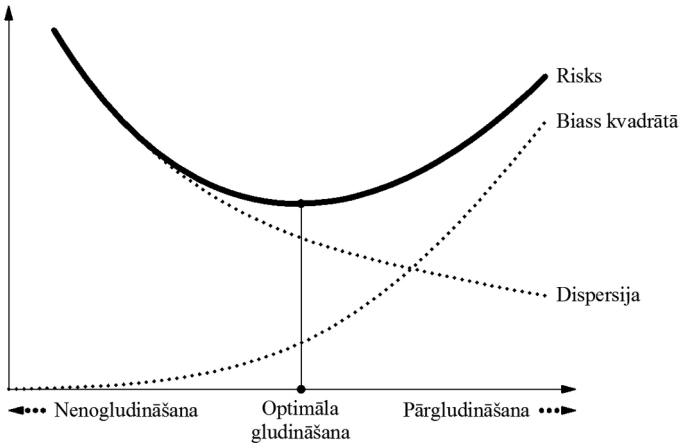
$$\mathbb{E}(\hat{r}_n(x)) - r(x) = biass(\hat{r}_n(x)).$$

Jo mazāka ir $R(r(x), \hat{r}_n(x))$ vērtība, jo labāks ir novērtējums.
Definīcijā 3 risks definēts ir punktā x. Ja vēlas summēt risku visām x vērtībām, tad lieto integrēto MSE

$$R(r, \hat{r}_n) = \int R(r(x), \hat{r}_n(x))dx.$$

Kā jau iepriekš tika pieminēts, tad no gludināšanas parametra izvēles ir atkarīgs regresijas līknes gludums. Ja dati ir pārgludināti, tad vienādībā (1.5) biasa saskaitāmais ir liels

un dispersijas saskaitāmais ir mazs. Un pretēji, ja dati ir nenogludināti, tad biasa saskaitāmais ir mazs un dispersijas saskaitāmais ir liels. Šo problēmu dēvē par biasa-dispersijas kompromisu (angliski *bias-variance tradeoff*) [2].



1. att. Biasa-dispersijas kompromiss

Kā redzams 1. attēlā, lai iegūtu labu novērtējumu, ir jālīdzsvaro biass un dispersija. To var izdarīt minimizējot risku. Minimizējot risku tiks atrasts optimālais gludināšanas parametrs.

2. Kodolu regresija un lokālā polinomu regresija

2.1. Kodolu regresija

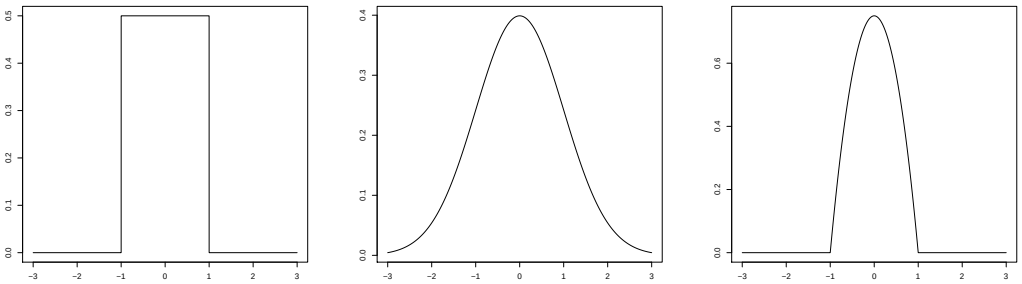
Definīcija 4. Par kodolu sauc tādu gludu funkciju $K(x)$, kurai $\forall x \in \mathbb{R}$ $K(x) \geq 0$ un

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0 \text{ un } \sigma_K^2 = \int x^2 K(x)dx > 0.$$

Biežāk lietotie kodoli funkciju gludināšanā ir:

1. boxcar kodols: $K(x) = \frac{1}{2} I_{\{|x| \leq 1\}}$,
2. Gausa (Gaussian) kodols: $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$,
3. Epanečnikova (Epanechnikov) kodols: $K(x) = \frac{3}{4}(1 - x^2) I_{\{|x| \leq 1\}}$.

Šie kodoli redzami 2.att.



(a) Boxcar kodols (b) Gausa kodols (c) Epanečnikova kodols

2. att. Kodolu piemēri

Nadaraja-Vatsona kodola regresija

Regresijas līkni ir iespējams izteikt ar nosacītajām blīvuma funkcijām. Tātad

$$r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)dy = \int y \frac{f(x,y)}{f_X(x)} dy. \tag{2.1}$$

Definīcija 5. Ja dots kodols K un joslas platums $h > 0$, tad kodola blīvuma novērtējums ir formā

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Apgalvojums 1. *Ja f nepārtraukta punktā x un $h \to 0$, $nh \to \infty$, kad $n \to \infty$, tad $\hat{f}_n(x) \xrightarrow{p} f(x)$.*

Novērtē blīvuma funkcijas $f(x, y)$ un $f_X(x)$ ar kodoliem, t.i.,

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right),$$

un

$$\hat{f}_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right) \cdot K_g\left(y - Y_i\right).$$

Tātad

$$\begin{aligned}\int y\hat{f}_n(x, y)dy &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\int \frac{y}{g}K\left(\frac{y - Y_i}{g}\right) \\ &= \left|\frac{y - Y_i}{g} = s\right| = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\int (sg + Y_i)K(s)ds = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)Y_i. \quad (2.2)\end{aligned}$$

Tādējādi novērtētās blīvuma funkcijas ievietojot (2.1), iegūst, ka regresijas funkcijas novērtējums ir formā

$$\hat{r}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{n^{-1}\sum_{j=1}^{n} K_h(x - X_j)}. \quad (2.3)$$

Definīcija 6 (Nadaraya-Watson(1964))**.** Nadaraja-Vatsona kodola novērtējums regresijas funkcijai r tiek definēts ar izteiksmi

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x)Y_i,$$

kur

$$l_i(x) = \frac{K(\frac{x - x_i}{h})}{\sum_{j=1}^{n} K(\frac{x - x_j}{h})}$$

sauc par svariem, $K(\cdot)$ ir kodols un $h > 0$ ir joslas platums.

Definīcija 7. Risks jeb vidējā kvadrātiskā kļūda tiek definēta ar izteiksmi

$$R(h) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i) - r(X_i))^2\right). \quad (2.4)$$

Kā iepriekšējās nodaļas beigās tika konstatēts, tad, lai iegūtu optimālo gludināšanas parametru h, ir jāminimizē risks $R(h)$. Bet problēma ir tāda, ka $R(h)$ ir atkarīgs no $r(X)$. Viens variants ir minimizēt risku, kurš ir formā

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i) - Y_i)^2.$$

Bet šim variantam trūkums ir tāds, ka šis novērtējums ir nenogludināts, jo dati tiek izmantoti divreiz. Vienreiz, lai novērtētu funkciju r, un otreiz, lai novērtētu risku.
Daudz populārāka un praksē biežāk lietota tiek krosvalidācija (angliski *cross-validation*).

Definīcija 8. Vienu-atstāt-ārā (angliski *leave-one-out*) krosvalidācijas funkcija definēta ar izteiksmi

$$CV = \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_{(-i)}(X_i))^2, \tag{2.5}$$

kur

$$\hat{r}_{(-i)}(x) = \sum_{j=1}^{n} Y_j l_{j,(-i)}(x)$$

ir novērtējums, kas iegūts izlaižot novērojumu pāri (X_i, Y_i), un

$$l_{j,(-i)}(x) = \begin{cases} 0 & , j = i, \\ \frac{l_j(x)}{\sum_{k\neq i} l_k(x)} & , j \neq i. \end{cases}$$

Tādējādi lineāriem gludinātājiem $\hat{r}_n$ risks ir formā

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{r}_n(X_i)}{1 - L_{ii}}\right)^2, \tag{2.6}$$

kur $L_{ii} = l_i(X_i)$.

Teorēma 2. *Risks Nadaraja-Vatsona kodola novērtējumam ir*

$$R(\hat{r}_n, r) = \frac{h^4}{4}\left(\int x^2 K(x)dx\right)^2 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx + \frac{\sigma^2 \int K^2(x)dx}{nh}\int \frac{1}{f(x)}dx + o(\frac{n}{h}) + o(h^4). \tag{2.7}$$

Turklāt, ja $h \to 0,\ nh \to \infty,\ \mathbb{E}Y^2 < \infty,\ f_x(x) > 0,$ *tad* $\hat{r}_n(x) \xrightarrow{P} r(x)$.

Pierādījums atrodams [3].

Piezīme 3. *Lielumu* $2r'(x)\frac{f'(x)}{f(x)}$ *sauc par dizaina novirzi, jo tas ir atkarīgs no* x *sadalījuma* ($x \sim f$). *Tas nozīmē, ka biass ir jūtīgs pret* x *atrašanās vietu. Kodolu novērtējumiem ir liela novirze netālu no robežām, to sauc par robežu novirzi. Lietojot lokālo polinomu regresiju to var uzlabot.*

Piezīme 4. *Ja diferencē risku vienādojumā* (2.7) *un pielīdzina nullei, tad iegūst, ka optimālais joslas platums* $h_* = O(n^{-1/5})$. *Pēc tam šo optimālo joslas platumu* h_* *ievietojot atpakaļ vienādojumā* (2.7), *redzams, ka risks* $R(\hat{r}_n; r) = O(n^{-4/5})$. *Savukārt parametriskos modeļos vislielākās ticamības novērtējumu riski tiecas uz 0 ar ātrumu* n^{-1}.

2.2. Lokālā polinomu regresija

Vispirms apskata novērtējumu formā $\hat{r}_n(x) \equiv a$. Minimizējot $\sum_{i=1}^n (Y_i - a)^2$ iegūst, ka atrisinājums ir konstanta funkcija $\hat{r}_n(x) = \bar{Y}$. Šis novērtējums acīmredzami nav labs novērtējums funkcijai $r(x)$.

Tālāk izvēlas svarus $w_i(x) = K((x_i - x)/h)$. Par funkcijas r novērtējumu izvēlas tādu $\hat{r}_n(x) \equiv a$, kurš minimizētu svērto kvadrātu summu

$$\sum_{i=1}^{n} w_i(x)(Y_i - a)^2.$$

Minimizēšanas rezultātā iegūst novērtējumu

$$\hat{r}_n(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)},$$

kurš ir kodolu regresijas novērtējums, ja svari ir formā $w_i(x) = K((x_i - x)/h)$. No šejienes seko, ka kodolu novērtējums ir lokāli konstants novērtējums. Tātad lokālas konstantes a vietā var lietot lokālu polinomu ar kārtu p.

Pieņemsim, ka x ir kāda fiksēta vērtība, pie kuras vēlas novērtēt $r(x)$. Tad vērtībām u, kuras atrodas x tuvumā, definē polinomu

$$P_x(u, a) = a_0 + a_1(u - x) + \frac{a_2}{2!}(u - x)^2 + \cdots + \frac{a_p}{p!}(u - x)^p.$$

Funkciju $r(u)$ aproksimē ar polinomu: $r(u) \approx P_x(u;a)$. Novērtē $a = (a_0, a_1, \ldots, a_p)^T$, izvēloties tādu $\hat{a} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)$, kurš minimizē lokāli svērto kvadrātu summu

$$\sum_{i=1}^{n} w_i(x)(Y_i - P_x(X_i;a))^2. \tag{2.8}$$

Piezīme 5. *Ja $p = 0$, tad iegūst kodolu novērtējumu. Ja $p = 1$, tad šādu gadījumu sauc par lokālo lineāro regresiju.*

Lai atrastu $\hat{a}(x)$, lietderīgi ir problēmu pārrakstīt vektoru formā.
Tātad izteiksmi (2.8) pārraksta formā

$$(Y - X_x a)^T W_x (Y - X_x a), \tag{2.9}$$

kur W_x ir $n \times n$ diagonālmatrica, kuras (i,i) elements ir $w_i(x)$, un dizaina matrica ir

$$X_x = \begin{pmatrix} 1 & x_1 - x & \ldots & \frac{(x_1-x)^p}{p!} \\ 1 & x_2 - x & \ldots & \frac{(x_2-x)^p}{p!} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n - x & \ldots & \frac{(x_n-x)^p}{p!} \end{pmatrix}.$$

Minimizējot (2.9), iegūst svērto vismazāko kvadrātu novērtējumu

$$\hat{a}(x) = (X_x^T W_x X_x)^{-1} X_x^T W_x Y.$$

Teorēma 6. *Lokālās polinomu regresijas novērtējums tiek definēts ar izteiksmi*

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i,$$

kur $l(x)^T = (l_1(x), \ldots, l_n(x))$,

$$l(x)^T = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x,$$

$e_1^T = (1, 0, \ldots, 0)^T$.

Ir spēkā, ka

$$\mathbb{E}\left(\hat{r}_n\left(x\right)\right) = \sum_{i=1}^{n} l_i\left(x\right) r\left(X_i\right) \;\; un \; \mathbb{D}\left(\hat{r}_n\left(x\right)\right) = \sigma^2 \sum_{i=1}^{n} l_i^2\left(x\right) = \sigma^2 \left\|l\left(x\right)\right\|^2.$$

Teorēmas pierādījums atrodams [4].

Apgalvojums 7. *Ja* $p = 1$, *tad*

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i,$$

kur

$$l_i(x) = \frac{b_j(x)}{\sum_{j=1}^{n} b_j(x)},$$

$$b_i(x) = K\left(\frac{X_i - x}{h}\right)\left(S_{n,2}(x) - (X_i - x)S_{n,1}(x)\right)$$

un

$$S_{n,j}(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(x_i - x)^j, \;\; j = 1, 2.$$

Nākamā teorēma demonstrē, kāpēc lokālā lineārā regresija ir labāka nekā kodolu regresija. Pierādījumu var atrast Fan(1992) un Fan un Gijbels (1996).

Teorēma 8. *Pieņemsim, ka* $Y_i = r(X_i) + \sigma(X_i)\epsilon, \;\; i = 1, \ldots, n$ *un* $a \leq X_i \leq b$. *Pieņemsim, ka* $X_1, \ldots, X_n$ *ir izlase no sadalījuma ar blīvuma funkciju* f *un ka* $x \in (a, b)$

1. $f(x) > 0$
2. f, r'' *un* σ^2 *ir nepārtraukti punkta* x *apkārtnē*
3. $h \to 0$ *un* $nh \to \infty$.

Dotiem $X_1, \ldots, X_n$ *izpildās sekojošais: lokālajam lineārajam novērtējumam un kodolu novērtējumam abiem dispersija ir formā*

$$\frac{\sigma^2(x)}{f(x)nh} \int K^2(u)du + o(\frac{1}{nh}).$$

Nadaraja-Vatsona kodolu novērtējumam biass ir

$$h^2(\frac{1}{2}r''(x) + \frac{r'(x)f'(x)}{f(x)}) \int u^2 K(u)du + o(h^2),$$

turpretīm lokālajam lineārajam novērtējumam ir asimptotiskais biass

$$h^2 \frac{1}{2} r''(x) \int u^2 K(u)du + o(h^2).$$

Tādējādi lokālais lineārais novērtējums ir brīvs no dizaina biasa. Robežu punktos a un b Nadaraja-Vatsona kodolu novērtējumam ir asimptotiskais biass ar kārtu h, turpretī lokālajam lineārajam novērtējumam biass ir ar kārtu h^2. Tātad lokālā lineārā novērtēšana likvidē robežu biasu.

3. Neparametriskā regresija ar REACT metodi

Šajā nodaļa tiks apskatīta metode, kuru detalizēti attīstīja Beran un D*ü*mbgen [5],[6]. Viņi apzīmē šo metodi kā REACT, kura veidojas no angļu valodas vārdu **R**isk **E**stimation and **A**daptation after **C**oordinate **T**ransfarmation pirmajiem burtiem.

Pirms tiek tālāk aplūkota REACT metode regresijas funkcijas novērtēšanai, nepieciešams fiksēt apzīmējumus un pamatjēdzienus no funkciju telpu teorijas.

- $L^p(\mathbb{R})$ $1 \leq p < \infty$ ir mērojamu funkciju telpa un to definē

$$L^p(\mathbb{R}) = \left\{ f : \int_{-\infty}^{\infty} |f(x)|^p \, dx < +\infty \right\}.$$

- $L^2(\mathbb{R})$ ir Hilberta telpa. Divu funkciju $f \in L^2(\mathbb{R})$ un $g \in L^2(\mathbb{R})$ skalārais reizinājums tiek definēts

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x)g(x)dx.$$

- Funkcijas $f(x) \in L^2(\mathbb{R})$ normu definē ar izteiksmi

$$\|f\|_{L^2}^2 = \langle f, f \rangle = \int_{-\infty}^{\infty} |f(x)|^2 \, dx.$$

- Ar $L^2[0,1]$ apzīmē visu intervālā $[0,1]$ kvadrātā integrējamu funkciju telpu.

$$L^2[0,1] = \left\{ f : \; f(x) = 0 \; ja \; x \notin [0,1] \; un \; \int_0^1 |f(x)|^2 \, dx < +\infty \right\}.$$

Definīcija 9. Funkciju virkne $\{\phi_j\}_{j=1}^{\infty}$ veido $L^2[0,1]$ ortonormālu bāzi, ja izpildās:

1. $\langle \phi_i, \phi_j \rangle = \delta_{i,j} = \begin{cases} 0 & ,\text{ja } i \neq j \\ 1 & ,\text{ja } i = j, \end{cases}$
2. $\forall j \;\; \|\phi_j\|_{L^2[0,1]} = 1,$
3. ja $r \in L^2[0,1]$ un $\forall j \;\; r \perp \phi_j$, tad $r = 0$.

Šajā darbā kā ortonormālā bāze tiks lietota kosinusu bāze, kura definēta ar izteiksmēm

$$\phi_0(x) = 1, \;\; \phi_j(x) = \sqrt{2} cos(2\pi j x), \;\; j = 1, 2 \ldots . \tag{3.1}$$

REACT. Tiek apskatīts regresijas modelis

$$Y_i = r(x_i) + \sigma\epsilon_i,$$

kur $\epsilon_i \sim N(0,1)$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi. Tiek pieņemts, ka ir regulārs dizains, t.i., $x_i = i/n, \;\; i = 1, \ldots, n$. Segmentā $[0,1]$ ir definēta ortonormāla bāze $\phi_1, \phi_2, \ldots$. Tad funkciju r pārraksta kā lineāro kombināciju no ortonormālas bāzes un parametriem

$$r(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x),$$

kur $\theta_j = \int_0^1 \phi_j(x) r(x) dx$.

Funkciju r aproksimē ar galīgu summu

$$r_n(x) \approx \sum_{j=1}^{n} \theta_j \phi_j(x),$$

kas ir r projekcija uz $\{\phi_1, \ldots, \phi_n\}$.

Lai novērtētu parametrus $\theta = (\theta_1, \ldots, \theta_n)$, tiek ieviests jauns mainīgais, kurš ir formā

$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(x_i), \;\; j = 1, 2, \ldots. \tag{3.2}$$

Mainīgais Z_j ir sadalīts pēc normālā sadalījuma, jo Z_j ir lineāra kombinācija no Normālajiem sadalījumiem. Turklāt

$$\mathbb{E}(Z_j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i) \phi_j(x_i) = \frac{1}{n} \sum_{i=1}^{n} r(x_i) \phi_j(x_i) \approx \int r(x) \phi_j(x) dx = \theta_j. \tag{3.3}$$

Savukārt dispersija ir

$$\mathbb{D}(Z_j) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{D}(Y_i) \phi_j^2(x_i) = \frac{\sigma^2}{n} \frac{1}{n} \sum_{i=1}^{n} \phi_j^2(x_i) \approx \frac{\sigma^2}{n} \int \phi_j^2(x) dx = \frac{\sigma^2}{n} \equiv \sigma_n^2. \tag{3.4}$$

Tātad Z_j ir neatkarīgi un $Z_j \sim N(\theta_j, \sigma_n^2), \;\; \sigma_n^2 = \frac{\sigma^2}{n}$.

Tālāk tiek ieviests modulators, kas ir vektors $b = (b_1, \ldots, b_n)$, kuram $0 \leq b_j \leq 1$, $j = 1, \ldots, n$. Konstants modulators ir formā $(b, \ldots, b)$. Sakārtotu apakškopu selekcijas (angliski *nested subset selection*) modulators ir formā $b = (1, \ldots, 1, 0, \ldots, 0)$. Monotons

modulators ir formā $1 \geq b_1 \geq \ldots \geq b_n \geq 0$.
Konstantu modulatoru kopu apzīmē ar M_{CONS}. Sakārtotu apakškopu selekcijas modulatoru kopu apzīmē ar M_{NSS}. Monotonu modulatoru kopu apzīmē ar M_{MON}.

θ novērtējumu pārraksta lietojot modulatoru

$$\hat{\theta} = bZ = (b_1 Z_1, \ldots, b_n Z_n).$$

Dotam modulatoram $b = (b_1, \ldots, b_n)$ funkcijas r novērtējums ir

$$\hat{r}_n(x) = \sum_{j=1}^{n} \hat{\theta}_j \phi_j(x) = \sum_{j=1}^{n} b_j Z_j \phi_j(x). \tag{3.5}$$

Ievērojot, ka $\hat{r}_n(x) = \sum_{i=1}^{n} Y_i l_i(x)$, kur $l_i(x) = \frac{1}{n} \sum_{j=1}^{n} b_j \phi_j(x) \phi_j(x_i)$, seko, ka novērtējums $\hat{r}_n$ arī ir lineārs gludinātājs.

Modulatori samazina Z_j vērtību tuvāk nullei, un šī samazināšana nodrošina funkcijas gludumu. Tādējādi samazināšanas lieluma izvēle atbilst joslas platuma izvēles problēmai kodolu novērtējumos. Atšķirība no kodolu novērtējumiem ir riska minimizēšanā. Ja iepriekšējās nodaļās, lai novērtētu risku, tika lietota krosvalidācija, tad tagad krosvalidācijas vietā lieto SURE - Steina nenovirzītu riska novērtējumu [7].
$\hat{\theta} = (b_1 Z_1, \ldots, b_n Z_n)$ risku definē ar izteiksmi

$$R(b) = \mathbb{E}_\theta \left[\sum_{j=1}^{n} (b_j Z_j - \theta_j)^2 \right]. \tag{3.6}$$

Teorēma 9 (Stein). *$Z \sim N^n(\theta, V)$. Ar $\hat{\theta} = \hat{\theta}(Z)$ apzīmē θ novērtējumu. $g(Z_1, \ldots, Z_n) = \hat{\theta} - Z$. Definē*

$$\hat{R}(z) = tr(V) + 2tr(VD) + \sum_i g_i^2(z), \tag{3.7}$$

kur $g_i = \hat{\theta}_i - Z_i$. Matricas D (i, j) komponente ir funkcijas $g(z_1, \ldots, z_n)$ i-tās komponentes parciālais atvasinājums pēc z_j.
Ja g ir diferencējama, tad

$$\mathbb{E}_\theta(\hat{R}(Z)) = R(\theta, \hat{\theta}). \tag{3.8}$$

Pierādījums. Tiks pierādīts gadījumā, kad $V = \sigma I$.

Ja $X \sim N(\mu, \sigma^2)$, tad $\mathbb{E}(g(X)(X-\mu)) = \sigma^2 \mathbb{E} g'(X)$.
Līdz ar to $\sigma^2 \mathbb{E}_\theta D_i = \mathbb{E}_\theta g_i(Z_i - \theta)$ un

$$\begin{aligned}
\mathbb{E}_\theta(\hat{R}(Z)) &= n\sigma^2 + 2\sigma^2 \sum_{i=1}^{n} \mathbb{E}_\theta D_i + \sum_{i=1}^{n} \mathbb{E}_\theta(\hat{\theta}_i - Z_i)^2 \\
&= n\sigma^2 + 2\sum_{i=1}^{n} \mathbb{E}_\theta(g_i(Z_i - \theta_i)) + \sum_{i=1}^{n} \mathbb{E}_\theta(\hat{\theta}_i - Z_i)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}_\theta(Z_i - \theta_i)^2 + 2\sum_{i=1}^{n} \mathbb{E}_\theta\left((\hat{\theta}_i - Z_i)(Z_i - \theta_i)\right) + \sum_{i=1}^{n} \mathbb{E}_\theta(\hat{\theta}_i - Z_i)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}_\theta(\hat{\theta}_i - Z_i + Z_i - \theta_i)^2 = \sum_{i=1}^{n} \mathbb{E}_\theta(\hat{\theta}_i - \theta_i)^2 = R(\hat{\theta}, \theta). \quad (3.9)
\end{aligned}$$

□

Tātad REACT ideja ir novērtēt risku $R(b)$ un izvēlēties tādu $\hat{b}$, lai minimizētu novērtēto risku modulatoru kopā M. Minimizējot M_{CONS} kopā, tiek iegūts James-Stein novērtējums. Tādējādi REACT ir James-Stein novērtējuma vispārinājums.

Teorēma 10. *Modulatora b risks ir*

$$R(b) = \sum_{j=1}^{n} \theta_j^2 (1 - b_j)^2 + \frac{\sigma^2}{n} \sum_{j=1}^{n} b_j^2.$$

$R(b)$ modificēts SURE novērtējums ir

$$\hat{R}(b) = \sum_{j=1}^{n} (Z_j^2 - \frac{\hat{\sigma}^2}{n})_+ (1 - b_j)^2 + \frac{\hat{\sigma}^2}{n} \sum_{j=1}^{n} b_j^2,$$

kur $\hat{\sigma}^2$ ir σ^2 novērtējums.

Definīcija 10. M ir modulatoru kopa. θ novērtējums ir $\hat{\theta} = (\hat{b}_1 Z_1, \ldots, \hat{b}_n Z_n)$, kur $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_n)$ minimizē $\hat{R}(b)$ kopā M. Tad REACT funkcijas novērtējums ir

$$\hat{r}_n(x) = \sum_{j=1}^{n} \hat{\theta}_j \phi_j(x) = \sum_{j=1}^{n} \hat{b}_j Z_j \phi_j(x). \quad (3.10)$$

Lai ieviestu šo metodi, ir jāatrod tāds $\hat{b}$, kurš minimizētu $\hat{R}(b)$. $\hat{R}(b)$ minimums kopā M_{CONS} ir James-Stein novērtējums. Lai minimizētu $\hat{R}(b)$ kopā M_{NSS}, aprēķina $\hat{R}(b)$

katram modulatoram, kurš ir formā $(1, \ldots, 1, 0, \ldots, 0)$. Citiem vārdiem sakot, jāatrod vesels pozitīvs skaitlis $\hat{J}$, kurš minimizētu

$$\hat{R}(J) = \frac{J\hat{\sigma}^2}{n} + \sum_{j=J+1}^{n} (Z_j^2 - \frac{\hat{\sigma}^2}{n})_+. \tag{3.11}$$

Funkcijas r novērtējums tad ir formā

$$\hat{r}_n(x) = \sum_{j=1}^{\hat{J}} Z_j \phi_j(x).$$

REACT metodes īss kopsavilkums:

1. definē $Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(x_i), \;\; j = 1, \ldots, n.$
2. atrod $\hat{J}$, kurš minimizē risku $\hat{R}(J)$, kurš ir uzdots ar vienādojumu (3.11).
3. funkcijas r novērtējums ir formā $\hat{r}_n(x) = \sum_{j=1}^{\hat{J}} Z_j \phi_j(x)$.

4. Veivletu regresija

Šajā nodaļā tiks apskatīta telpiski nehomogēnu funkciju novērtēšana. Telpiski nehomogēnas funkcijas ir tādas funkcijas, kurām gludums mainās atkarībā no neatkarīgā mainīgā X vērtības. Šādas funkcijas ir sarežģīti novērtēt.

Piemēram, apskatam bloku funkciju (3.(a) attēls), kuru ieviesa Donoho [8] un kura ir formā

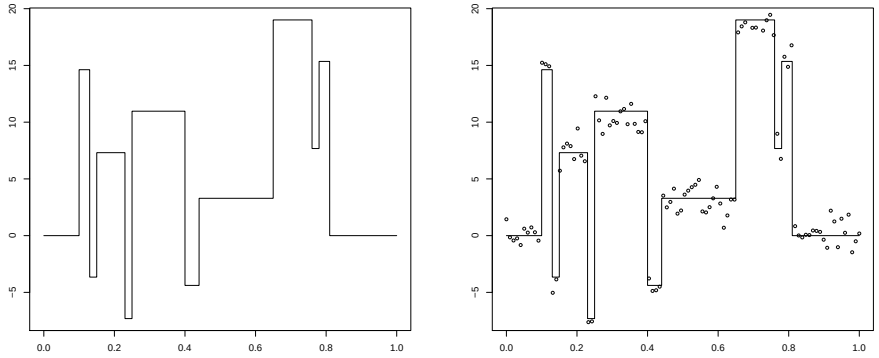
$$r(x) = \sum_{j=1}^{11} h[j] \cdot B(x - t[j]),$$

kur $B(x) = \frac{1+sgn(x)}{2}$, $sgn(x) = I_{\{x>0\}} - I_{\{x<0\}}$ un

$$t = [0.10, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81]$$

$$h = [4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2].$$

Ar Monte Karlo simulāciju palīdzību ģenerē izlasi apjomā n=100 no modeļa $Y_i = r(x_i) + \epsilon_i$, kur $\epsilon_i \sim N(0, 1)$ un $x_i = i/n$ (3.(b) attēls).



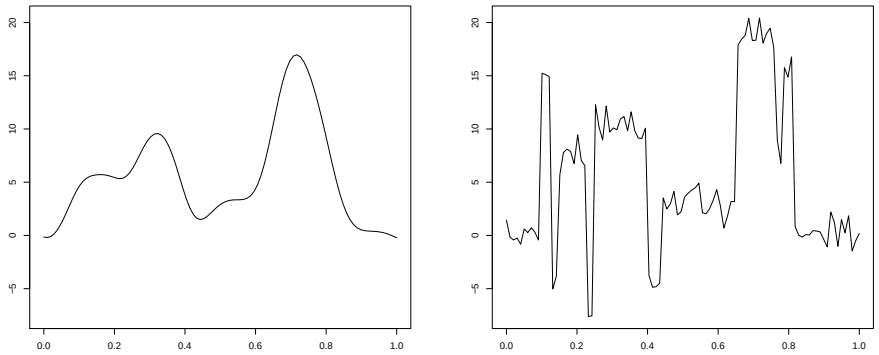
(a) Bloku funkcija

(b) Bloku funkcija un ģenerētā izlase ar apjomu 100

3. att. Bloku funkcija

Ja regresijas līknes novērtēšanai lieto lokālo lineāro regresiju ar lielu joslas platumu, tad regresijas līkne ir gluda, bet tiek nogludināti lēcieni. Savukārt, ja lieto mazu joslas platumu, tad lēcieni tiek atrasti, bet lielākā daļa regresijas līknes ir oscilējoša (4.attēls). Līdzīgi ir arī, ja lieto ortogonālas funkcijas. Ja lieto ortogonālās funkcijas ar zemākas kārtas saskaitāmiem, tad tiek izlaisti lēcieni. Bet, ja tiek lietoti augstākas kārtas saskaitāmie,

tad atrod lēcienus, bet regresijas līkne ir oscilējoša.



(a) Regresijas līkne ar lielu joslas platumu

(b) Regresijas līkne ar mazu joslas platumu

4. att. Regresijas līknes dažādiem joslas platumiem

Tā kā veivleti ir veidoti telpiski pielāgoti, tad tas ļauj efektīvi novērtēt nehomogēnas funkcijas, kurām ir pārtraukumi, pīķi vai citas datu struktūras izmaiņas.

4.1. Veivleti

Kaut arī pirmo veivletu Alfred Haar uzkonstruēja jau 1910.gadā, nopietnāka veivletu pētniecība un analīze statistikā notikusi tikai pēdējos 15-20 gados. Vārdu veivlets ieviesa Morlet un Grossmann 1980-tajos. Viņi lietoja franču valodas vārdu ondelette, ar to apzīmējot"mazus vilnīšus." Drīz pēc tam tas tika ievists arī angļu valodā, pārtulkojot "onde" uz "wave', iegūstot "wavelet". Teoriju par veivletiem galvenokārt attīstīja Y.Meyer ar saviem kolēģiem. Savukārt S.Mallat 1989.gadā izveidoja *fast wavelet transform* (FWT) algoritmu, ar kura palīdzību realizācija notiek ātrāk. Statistikā visvairāk veivletus pētījuši un lietojuši ir Vidakovic, Ogden, Donoho un Johnstone.

Veivleti ir diezgan jauna bāzes funkciju saime, kuru lieto, lai izteiktu un aproksimētu citas funkcijas. Veivletu koeficienti spēj noteikt datos dažādas funkcijas izmaiņas, ko citas metodes varētu izlaist, piemēram, pārtrauktību un pēkšņus pīķus. Veivletu galvenās īpašības ir ortogonalitāte, lokalizācija laikā un mērogā, ierobežotība intervālā. Ortogonalitāte nodrošina, ka datu reprezentācijā nenotiks lieka dublēšanās, jo informācija, kas bāzēta vienā saskaitāmā ir neatkarīga no informācijas, kas bāzēta citos saskaitāmajos.

Veivleti tiek veidoti no "tēva" veivleta ϕ (angliski *father wavelet*) un "mātes" veivleta

ψ (angliski *mother wavelet*). Tēva veivletu lieto, lai iegūtu gludus, zemas frekvences datus, turpretī mātes veivleti tiek lietoti, lai iegūtu detalizētus, augstākas frekvences datus.

Definē funkcijas

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k),\ j,k \in \mathbb{Z}, \tag{4.1}$$

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ j,k \in \mathbb{Z} \tag{4.2}$$

Skaitli j sauc par mērogu, skaitlis k ir atrašanās vieta.

Tādējādi funkcijai ϕ_{jk} ir tāda pati forma kā ϕ, bet tai ir cits mērogs, jo tā tiek pareizināta ar $2^{J/2}$, un tā ir pārbīdīta par skaitli k.

Kā jau nodaļas sākumā pieminēts, tad pirmais un vienkāršākais veivlets ir Haar veivlets. Haar tēva veivlets definēts ar izteiksmi

$$\phi(x) = \begin{cases} 1 & 0 \le x < 1 \\ 0 & \text{citur.} \end{cases}$$

Savukārt Haar mātes veivlets definēts ar izteiksmi

$$\psi(x) = \begin{cases} -1 & 0 \le x \le \frac{1}{2} \\ 1 & \frac{1}{2} < x \le 1 \\ 0 & \text{citur.} \end{cases}$$

Haar veivletiem piemīt īpašība, ka ārpus kāda noteikta intervāla tie ir nulle. Bet Haar veivletu trūkums ir tas, ka tie nav gludi. Kā uzkonstruēt lokalizētus, gludus veivletus? Detalizētāk to ir aplūkojuši *Härdle* [9] un Daubechies [10].

Vispārīgi veivletiem (izņemot Haar veivletus) nav analītiskas uzdošanas formas. Tos ģenerē lietojot MRA un skaitliskas metodes (detalizētāk [10]).

MRA definēšanas procesā tiek izmantota Furjē transformācija un inversā Furjē transformācija:

- Funkcijas f **Furjē transformācija** f^* ir

$$f^*(t) = \int_{-\infty}^{\infty} e^{-ixt} f(x)dx,\ \text{kur } i = \sqrt{-1}.$$

- Ja f^* ir absolūti integrējama, tad f iegūst ar **inverso Furjē transformāciju**

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ixt} f^*(t) dt.$$

Definīcija 11. Dotai funkcijai ϕ definē kopas

$$\begin{gathered} V_0 = \left\{ f : f(x) = \sum_{k\in\mathbb{Z}} c_k \phi(x-k), \sum_{k\in\mathbb{Z}} c_k^2 < \infty \right\}, \\ V_1 = \{ f(x) = g(2x) : g \in V_0 \}, \\ V_2 = \{ f(x) = g(2x) : g \in V_1 \} \\ \vdots \qquad \vdots \end{gathered}$$

Teiksim, ka funkcija ϕ ģenerē multilīmeņu analīzi (*multiresolution analysis* (MRA)) telpā $\mathbb{R}$, ja

$$V_j \subset V_{j+1}, \ j \geq 0 \tag{4.3}$$

un

$$\bigcup_{j\geq 0} \text{ ir blīva telpā } L_2(\in \mathbb{R}). \tag{4.4}$$

Funkciju ϕ sauks par tēva veivletu vai mēroga funkciju.

Nosacījums (4.4) nozīmē, ka jebkurai funkcijai $f \in L_2(\mathbb{R})$ eksistē tāda funkciju virkne $f_1, f_2, \ldots$, ka katra $f_r \in \bigcup_{j\geq 0} V_j$ un $\|f_r - f\| \to 0$ kad $r \to \infty$.

Teorēma 11. *Ja $V_0, V_1, \ldots$ ir MRA, kura ģenerēta ar funkciju ϕ, tad $\forall j \in \mathbb{Z}$ $\{\phi_{jk}, k \in \mathbb{Z}\}$ ir ortonormāla bāze kopā V_j.*

Teorēmas pierādījums atrodams [11].

Pieņemsim, ka ir dots MRA. Tad $\phi \in V_1$, jo $\phi \in V_0$ un $V_0 \subset V_1$. Tā kā $\{\phi_{1k}, k \in \mathbb{Z}\}$ ir ortonormāla bāze kopai V_1, funkciju ϕ var izteikt kā lineāru kombināciju no kopas V_1 funkcijām

$$\phi(x) = \sum_k l_k \phi_{1k}(x), \tag{4.5}$$

kur $l_k = \int \phi(x)\phi_{1k}(x)dx$ un $\sum_k l_k^2 < \infty$. Koeficientus $\{l_k\}$ sauc par mēroga koeficientiem.

Piezīme 12. *Haar veivletiem $l_0 = l_1 = 2^{-1/2}$ un $l_k = 0$, ja $k \neq 0, 1$.*

Vienādību (4.5) sauc par divu-mērogu vienādojumu. Šo vienādību var pārrakstī formā

$$\phi^*(t) = m_0(t/2)\phi^*(t/2),$$

kur $m_0(t) = \sum_k l_k e^{itk}/\sqrt{2}$.

Pielietojot šo formulu rekursīvi, iegūst

$$\phi^*(t) = m_0(t/2) \prod_{k=1}^{\infty} m_0(t/2^k)\phi^*.$$

Tādējādi, pie dotiem mēroga koeficientiem ir iespējams aprēķināt $\phi^*(t)$, un tad pēc inversās Furjē transformācijas atrast $\phi(t)$ [11].

Nākamajā teorēmā parādīts, kā no mēroga koeficientiem konstruēt tēva veivletu.

Teorēma 13. *Dotiem koeficientiem* $\{l_k, k \in \mathbb{Z}\}$ *definē funkciju*

$$m_0(t) = \frac{1}{\sqrt{2}} \sum_k l_k e^{-ikt}. \tag{4.6}$$

$$\phi^* = \prod_{j=1}^{\infty} m_0(t/2^j)$$

un funkcija ϕ *ir inversā Furjē transformācija no* ϕ^*. *Pieņemsim, ka*

$$\frac{1}{\sqrt{2}} \sum_{k=N_0}^{N_1} l_k = 1$$

kādiem $N_0 < N_1$ *un ka*

$$|m_0(t)|^2 + |m_0(t+\pi)|^2 = 1,$$

un ka $m_0(t) \neq 0$, $|t| \leq \pi/2$ *un ka eksistē ierobežota funkcija* Φ *tāda, ka* $\int \Phi(|u|)du < \infty$ *un* $|\phi(x)| \leq \Phi(|x|)$ *gandrīz visiem* x.

Tad funkcija ϕ *ir tēva veivlets, kurš ir nulle ārpus intervāla* N_0, N_1.

Pierādījums atrodams [10].

Tālāk definē kopu W_k, lai tā būtu kopas V_k ortogonāls papildinājums kopā V_{k+1}. Citiem vārdiem, $\forall f \in V_{k+1}$ var uzrakstīt kā summu $f = v_k + w_k$, kur $v_k \in V_k$ un $w \in W_k$, un v_k un w_k ir ortogonāli. Raksta

$$V_{k+1} = V_k \oplus W_k.$$

Tādējādi

$$L_2(\mathbb{R}) = \bigcup_k V_k = V_0 \oplus W_0 \oplus W_1 \oplus \cdots .$$

Tad mātes veivletu definē ar izteiksmi

$$\psi = \sqrt{2}\sum_k (-1)^{k+1} l_{1-k}\phi(2x-k).$$

Tādējādi, lai atrastu atbilstošu pāri (ϕ, ψ), pietiek atrast tēva veivletu ϕ. Pēc tam mātes veivletu ψ iegūst no tēva veivleta.

Teorēma 14. *Funkcijas* $\{\phi_{jk}, k \in \mathbb{Z}\}$ *veido* W_j *bāzi. Funkcijas* $\{\phi_k, \psi_{jk}, k \in \mathbb{Z}, j \in \mathbb{Z}_+\}$ *ir ortonormāla bāze telpā* $L_2(\mathbb{R})$.

Šīs teorēmas pierādījums atrodams [10].

Tādējādi jebkuru funkciju $f \in L_2$ var uzrakstīt kā

$$f(x) = \sum_k \alpha_{0k}\phi_{0k}(x) + \sum_{j=0}^{\infty}\sum_{j=0}^{\infty} \beta_{jk}\psi_{jk}, \tag{4.7}$$

kur

$$\alpha = \int f(x)\psi_{0k}(x)dx$$

tiek saukts par mēroga koeficientu un

$$\beta_{jk} = \int f(x)\psi_{jk}(x)dx$$

sauc par datalizācijas koficientiem.

Galīga summa

$$f_J(x) = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x). \tag{4.8}$$

ir funkcijas f līmeņa J aproksimācija.

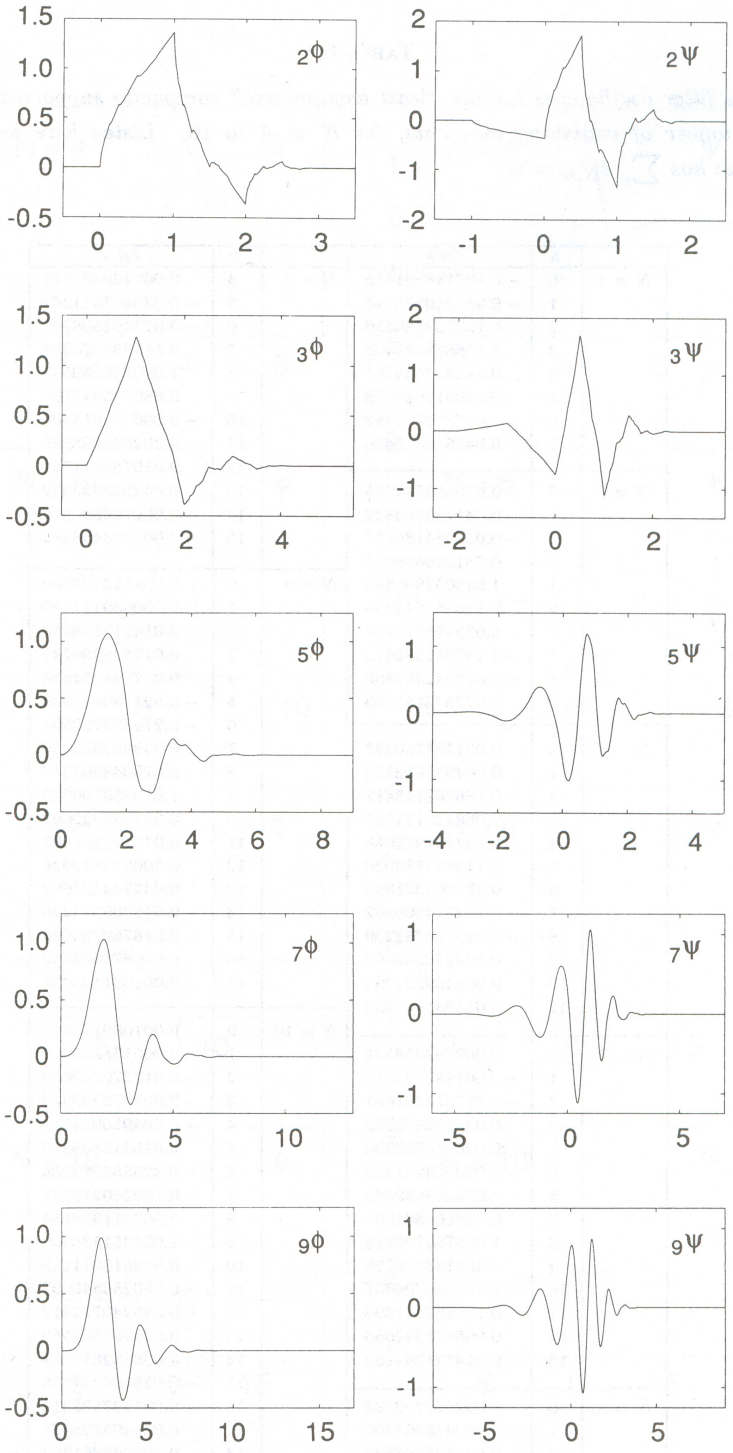
Momentu izzušana. Ja izpildās

$$\int_{\mathbb{R}} \psi(x)dx = 0, \tag{4.9}$$

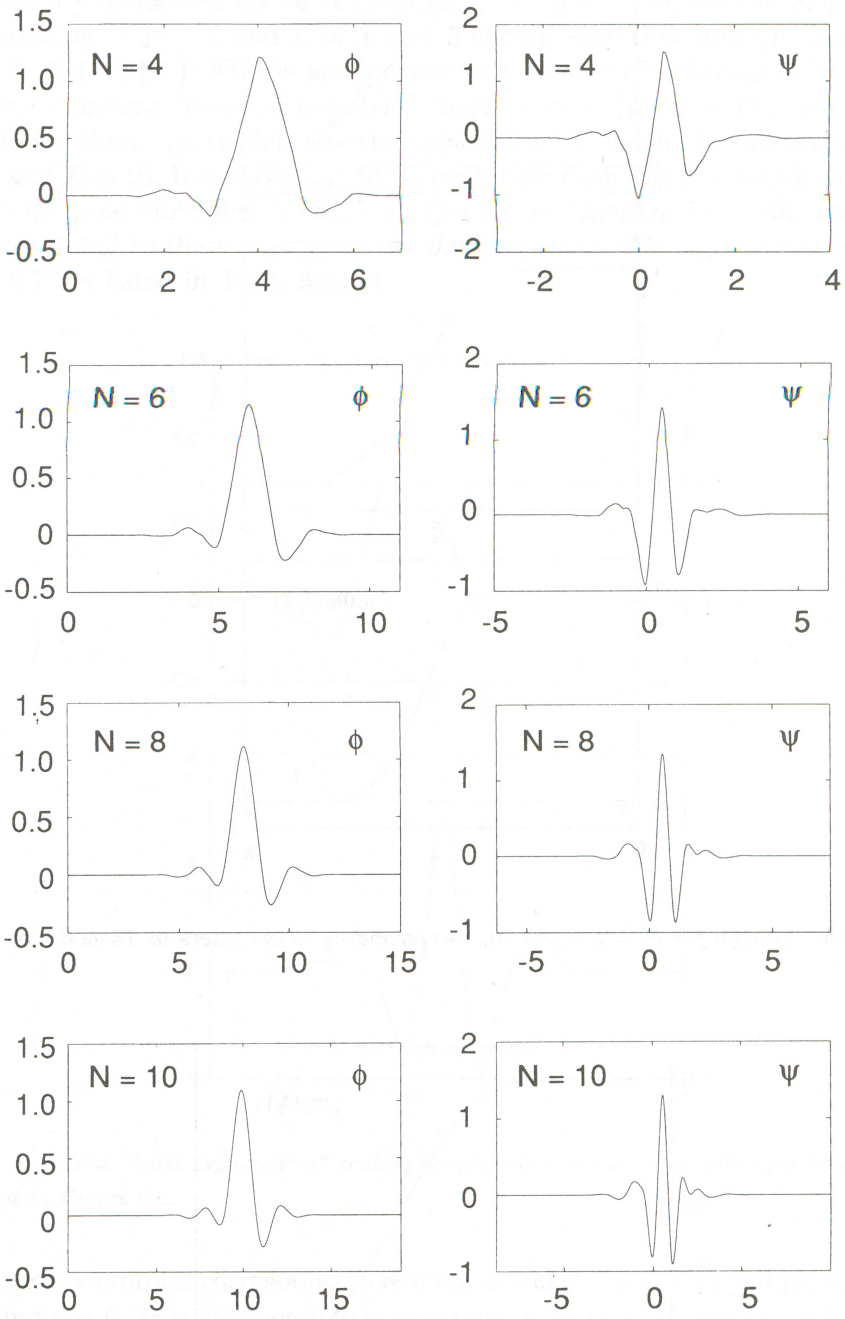
tad saka, ka funkcijai $\psi(x)$ piemīt nulltā momenta izzušana. Savukārt, ja

$$\int_{\mathbb{R}} x^k\psi(x)dx = 0,$$

tad funkcijai $\psi(x)$ piemīt k-tā momenta izzušana.



5. att.: Tēva N^{ϕ} un mātes N^{ψ} veivletu attēli ar maksimālo skaitu izzūdošo momentu *"extremal phase"* veivletu klasei, N=2,3,5,7,9



6. att.: Vismazākās asimetrijas tēva ϕ un mātes ψ veivletu attēli ar maksimālo skaitu izzūdošo momentu *"least asymmetric"* veivletu klasei, N=4,6,8,10

Tādējādi, ja nepieciešams uzkonstruēt gludu ortonormālu veivletu, jāmeklē veivlets $\psi(x)$ ar daudz izzūdošiem momentiem. Šo approksimāciju izmantoja Daubechies, kura uzkonstruēja gludu veivletu klasi. Daubechies veivletam ar kārtu N ir N izzūdošie momenti un tas ir ierobežots intervālā $[0, 2N-1]$. Ja palielina N, tad Daubechies veivleti kļūst gludāki. Daubechies uzkonstruēja divas veivletu klases (skatīt 5. un 6. attēlu).

Tomēr biežāk statistikā tiek lietoti vismazākās asimetrijas Daubechies veivleti.

4.2. Neparametriskā regresija ar veivletiem

Aplūkots tiek regresijas modelis

$$Y_i = r(x_i) + \sigma\epsilon_i,\ i = 1, 2, \ldots, n,$$

kur $\epsilon_i \sim N(0,1)$. Tāpat kā REACT metodes gadījumā tiek pieņemts, ka ir fiksēts dizains, proti, $x_i = i/n$. Vēl tiek pieņemts, ka $n = 2^J$, kur J ir pozitīvs vesels skaitlis. Šie pieņēmumi ļauj pielietot Mallata FWT algoritmu.

Gadījumos, kad nav fiksēts dizains vai datu skaits nav divnieka pakāpe, ir nepieciešams datus modificēt. Iespējami ir vairāki veidi, kā datus modificēt, lai apjoms būtu divnieka pakāpe. Var ņemt tikai esošos pēdējos novērojumus, kuri veido divnieka pakāpi, un sākuma novērojumus ignorēt. Vai otrādāk, ignorēt beigu novērojumus un ņemt tikai sākuma novērojumus, kuru apjoms ir divnieka pakāpe. Iespējama ir arī šo pieminēto veidu kombinēšana. Bet šādi praksē rīkojas ļoti reti. Daudz biežāk lieto datu replicēšanu, kopēšanu vai interpolāciju. Ideja datu replicēšanai ir paplašināt novērojumu skaitu līdz nākamajai divnieka pakāpei ar esošo novērojumu piekopēšanu. Datu skaita samazināšanu līdz tuvākajai mazākai divnieka pakāpei veic ar interpolāciju.

Lai novērtētu funkciju r ar veivletiem, vispirms funkciju r aproksimē ar izvirzījumu

$$r(x) \approx r_n(x) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x), \tag{4.10}$$

kur $\alpha_{j_0,k} = \int r(x)\phi_{j_0,k}(x)dx$ un $\beta_{j,k} = \int r(x)\psi_{jk}(x)dx$.

Tā kā reāli funkcija r nav zināma, tad šos koeficientus pēc formulām aprēķināt nevar. Tāpēc šie koficienti ir jānovērtē.

Koeficientu empīriskai novērtēšanai lieto izteiksmes

$$S_k = \frac{1}{n}\sum_i \phi_{j_0,k}(x_i)Y_i \ \text{ un } \ D_{jk} = \frac{1}{n}\sum_i \psi_{jk}(x_i)Y_i, \tag{4.11}$$

kur S_k ir empīriskie mērogošanas koeficienti un D_{jk} ir empīriskie detalizācijas koeficienti.

Par mērogošanas koeficientu novērtējumu tiek ņemts nepārveidots empīriskais mērogošanas koeficients, proti, $\hat{\alpha}_{j_0,k} = S_k$. Savukārt, lai novērtētu koeficientus β_{jk}, tiek lietots

speciāls empīrisko koeficientu D_{jk} samazināšanas veids, kuru sauc par sliekšņošanu. Šie novērtētie koeficienti pēc tam tiek ievietoti (4.10) un tiek iegūts funkcijas r novērtējums

$$\hat{r}_n(x) = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} \hat{\beta}_{jk}\psi_{jk}(x). \tag{4.12}$$

Veivletu regresija ir līdzīga ortogonālo rindu regresijas metodei. Ir tikai divas atšķirības:

- ir cita ortonormāla bāze;
- tiek lietota sliekšņošana - samazināšanas veids, kurā, ja empīriskais detalizācijas koeficients D_{jk} ir mazs, tad $\hat{\beta}_{jk}$ tiek pietuvināts nullei vai pārvērsts par nulli.

Veivletu sliekšņošana. Tēva veivletu koeficientu $\alpha_{j_0,k}$ novērtējumi ir vienādi ar empīriskajiem mērogošanas koeficientiem S_k. Tiem netiek pielietota nekāda veida sliekšošana vai samazināšana. Savukārt mātes veivletu koeficientu novērtējumi ir bāzēti uz empīrisko detalizācijas koeficientu D_{jk} samazināšanu. Veivletiem tiek pielietota nelineāra samazināšana- sliekšņošana, kuru savukārt iedala "stiprā" sliekšņošana un "vājā" sliekšņošana.

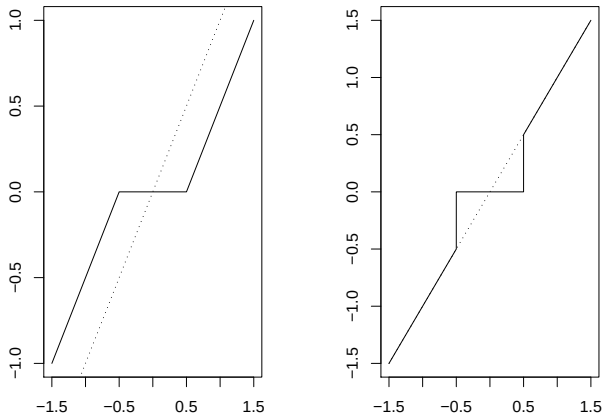
Koeficientu β_{jk} novērtējumu pēc stiprās sliekšņošanas definē ar izteiksmi

$$\hat{\beta}_{jk} = \begin{cases} 0 & \text{, ja } |D_{jk}| < \lambda \\ D_{jk} & \text{, ja } |D_{jk}| \geq \lambda. \end{cases} \tag{4.13}$$

Savukārt pēc vājās sliekšņošanas koeficientu β_{jk} novērtējumi tiek definēti ar izteiksmi

$$\hat{\beta}_{jk} = \begin{cases} D_{jk} + \lambda & \text{, ja } D_{jk} < -\lambda \\ 0 & \text{, ja } -\lambda \leq D_{jk} < \lambda \\ D_{jk} - \lambda & \text{, ja } D_{jk} > \lambda. \end{cases} \tag{4.14}$$

Piemēram, 7.attēla ir redzama sliekšošanas veidu atšķirība, ja $\lambda = 0.5$.



7. att. Kreisā pusē ir vājā sliekšņošana, labā pusē ir stirpā sliekšņošana

Sliekšņošanas algoritms ļauj datiem pašiem noteikt, kuri veivletu koeficienti ir nozīmīgi. Stiprā sliekšņošana balstās uz principu "paturēt" vai "likvidēt", turpretī vājās sliekšņošanas pamatā ir "samazināt" vai "likvidēt".

Kā redzams izteiksmēs (4.13) un (4.14), koeficientu β_{jk} novērtēšana ir atkarīga no parametra λ. Parametru λ sauc par slieksni. Ja izvēlas ļoti lielu slieksni, tad koeficientus saīsina par daudz, un veidojas pārgludināšana. Un pretēji, ja slieksnis ir par mazu, tad rekonstruēšanā tiek pieļauti daudz vairāk koeficienti, veidojot oscilējošu, nenogludinātu novērtējumu.

Eksistē dažādas metodes kā izvēlēties slieksni. Sliekšņus var iedalīt divās kategorijās. Ir globālie sliekšņi, kuros izvēlas vienu λ vērtību, kuru pielieto visiem veivletu koeficientiem visos līmeņos j. Un ir arī lokālie sliekņi, kuros katram līmenim j izvēlas atšķirīgus sliekņus λ_j.

Vienkāršākais un visbiežāk lietotais slieksnis ir Donoho un Johnstone [12] ieviestais universālais slieksnis, kurā λ tiek definēta ar izteiksmi

$$\lambda = \hat{\sigma}\sqrt{\frac{2\log n}{n}}, \tag{4.15}$$

kur σ^2 ir sekojošs σ novērtējums

$$\hat{\sigma} = \sqrt{n} \times \frac{median\left(|D_{J-1,k} - median(D_{J-1,k})|\right)}{0.6745}. \tag{4.16}$$

Piezīme 15. *Šis nav vienīgais σ novērtējums. Drīkst izmantot arī jebkuru citu σ^2 novērtējumu $\hat{\sigma}^2$.*

Savukārt no lokāliem sliekšņiem visizplatītākais ir slieksnis λ_j, kuru izvēlas tā, lai tas minimizētu Steina nenovirzītu riska novērtējumu, kurš šajā gadījumā definēts ar izteiksmi

$$S(\lambda_j) = \sum_{k=1}^{n_j} \left[\frac{\hat{\sigma}^2}{n} - 2\frac{\hat{\sigma}^2}{n} I\left(\left|\tilde{\beta}_{jk}\right| \leq \lambda_j\right) + \min\left(\tilde{\beta}_{jk}^2, \lambda_j^2\right) \right], \tag{4.17}$$

kur $n_j = 2^{j-1}$ ir parametru skaits j-tajā līmenī.

Pēdējos gados veivletu sliekšņošanas tehnikas ir ļoti attīstījušās. Ir izveidoti jauni sliekšņu veidi, kuri ir balstīti uz esošo sliekšņu dažādu kombinēšanu.

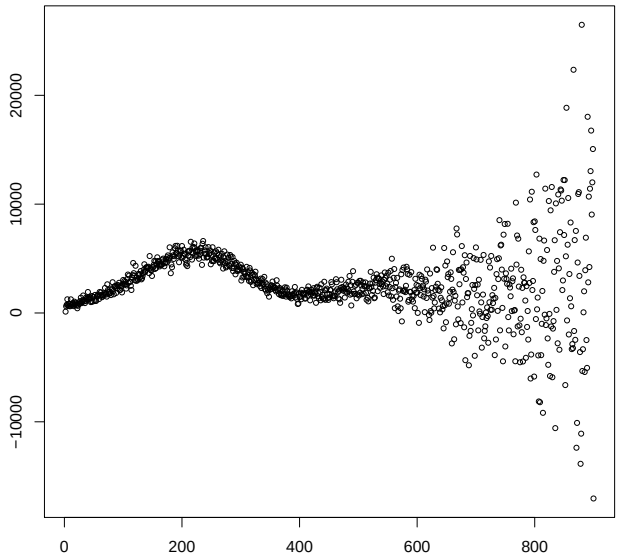
5. Simulācijas un datu analīze

Simulācijās un datu analīzē tiek lietota statistikas programma R. Programmā R uzrakstītie kodi (skatīt pielikumā) tiek lietoti, lai konstruētu un novērtētu regresijas funkciju r un secinātu, kura no apskatītajām neparametriskās regresijas metodēm ir labāka, bet kuru metodi iespējams nevajadzētu izvēlēties regresijas funkcijas novērtēšanai. Šajā analīzē pētāmais objekts ir reāli dati un testa funkciju ģenerēti dati.

5.1. Reālu datu analīze

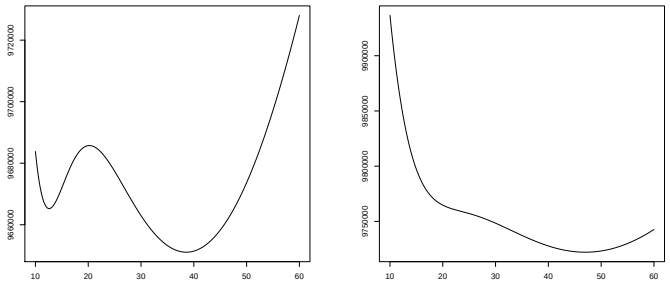
Šajā nodaļā analizēti tiks "CMB" un "SP500" dati. "CMB" datus var apakšielādēt Larry Wassermann mājas lapā ($http : //www.stat.cmu.edu/ \sim larry/all - of - nonpar/data.html$).

CMB (cosmic microwave background radiation) datos (8.attēls) X_i dati attēlo temperatūras fluktuācijas frekvenci un Y_i dati reprezentē fluktāciju spēku katrā frekvencē.



8. att. CMB datu izkliedes attēls

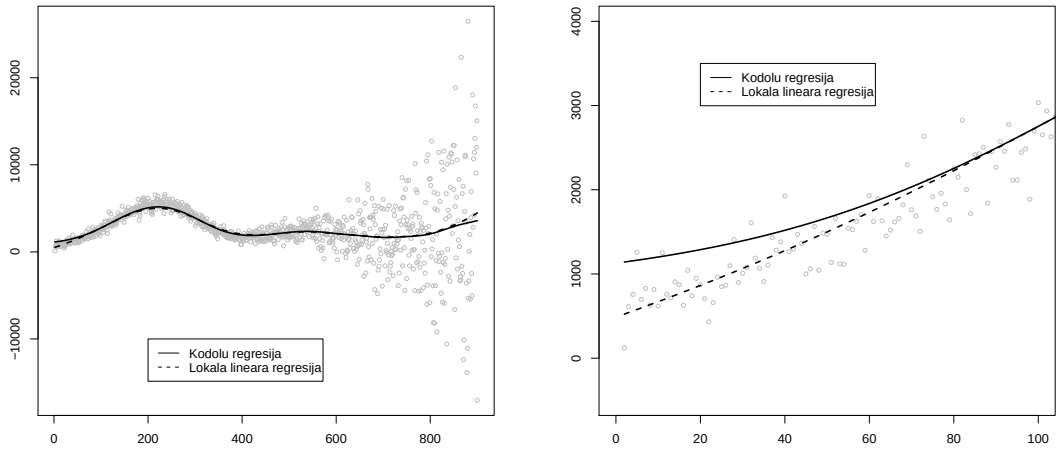
Apskatām, kādus rezultātus iegūst lietojot tradicionālās neparametriskās regresijas metodes - kodolu regresiju un lokālo lineāro regresiju. Lai novērtētu regresijas funkciju, nepieciešams atrast optimālo joslas platumu. To atrod minimizējot risku pēc krosvalidācijas. 9. attēlā ir redzami riski atkarībā no joslas platuma.



(a) Risks kodolu regresijas gadījumā (b) Risks lokālās lineārās regresijas gadījumā

9. att. Riski atkarībā no joslas platuma

Tātad kodolu regresijai optimālais joslas platums h=38.6, savukārt lokālai lineārai regresijai optimālais joslas platums h=47. Lietojot šos joslas platumus, regresijas funkcijas novērtējumi ir redzami 10. attēlā.

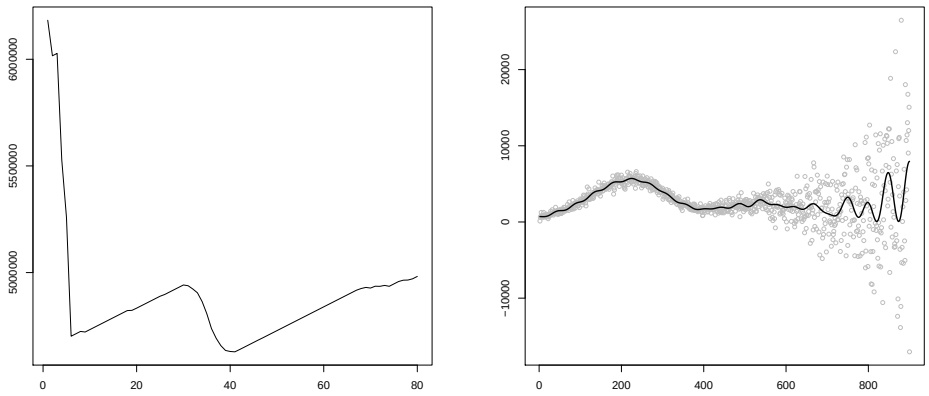


(a) Regresija visiem novērojumiem (b) Regresija sākuma novērojumiem

10. att. Regresijas funkcijas novērtējumi CMB datiem

Kodolu un lokālās lineārās regresijas novērtējumos vispārīgi būtiskas atšķirības nav saskatāmas (10.(a)attēls). Bet, ja apskatās situāciju robežu galapunktos, tad tur ir redzamas atšķirības. Un, kā teorijas apskatā jau tika parādīts, tad labāks novērtējums ir lokāli lineārās regresijas metodei, jo regresijas funkcija labāk tiek aproksimēta robežu galapunktos, uzlabojot robežu biasu. Tas arī redzams 10.(b) attēlā, kurā ir attēloti novērojumi intervālā X=[0,100].

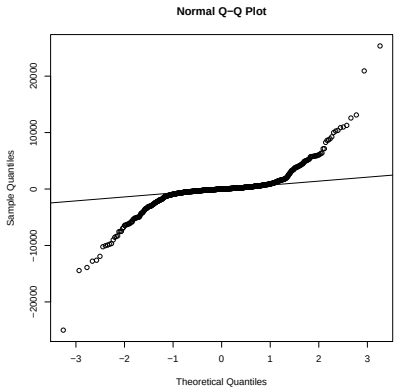
Nākamā metode, ar kuru tiek analizēti CMB dati, ir REACT metode. Šajā gadījumā gludināšanas parametrs ir saskaitāmo skaits rindā. Optimālais saskaitāmo skaits tiek noteikts minimizējot risku nevis pēc krosvalidācijas, bet gan minimizējot izteiksmi (3.11), kas ir Steina nenovirzīts riska novērtējums. CMB datiem REACT metodes gadījumā risks atkarībā no saskaitāmo skaita ir redzams 11.(a)attēlā. Tātad optimālais saskaitāmo skaits $\hat{J}$ ir 41. Regresijas līkne kopā ar CMB datiem redzama 11.(b) attēlā.



(a) Risks atkarībā no saskaitāmo skaita (b) REACT metodes regresija

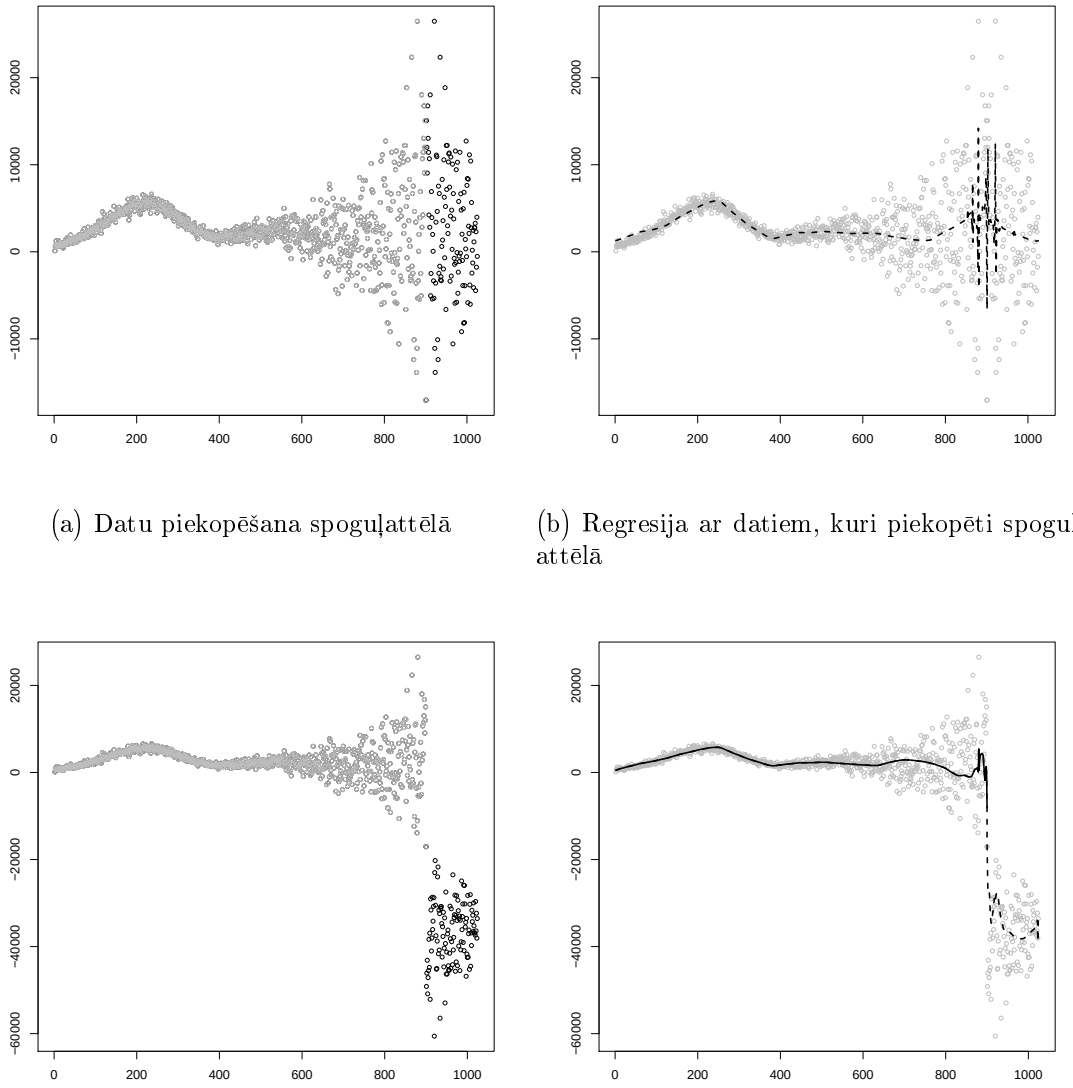
11. att. REACT metode CMB datiem

Kā redzams 11.(b) attēlā, tad REACT metodei regresijas novērtējums beigās nav gluds. Iespējams, ka neizpildījās kāds no nosacījumiem. Tāpēc tiek pārbaudīts, vai atlikumi ir sadalīti pēc Normālā sadalījuma. Uzzīmējot Q-Q plot (12. attēls), redzams, ka CMB datiem atlikumi nav sadalīti pēc Normālā sadalījuma. Tādēļ REACT metode CMB datiem nav īsti piemērota.



12. att. Q-Q plot

CMB datiem n=899 un tas nav divnieka pakāpe. Lai varētu pielietot diskrēto veivletu tranformāciju, datu skaitu ir nepieciešams palielināt līdz 1024. No sākuma tiek pielietotas programmā R automātiski iebūvētās metodes datu paplašināšanai [13]. Tās ir inversā un parastā datu pārkopēšana. Un tad iegūtās veivletu regresijas līknes redzamas 13.attēlā.



(a) Datu piekopēšana spoguļattēlā

(b) Regresija ar datiem, kuri piekopēti spoguļattēlā

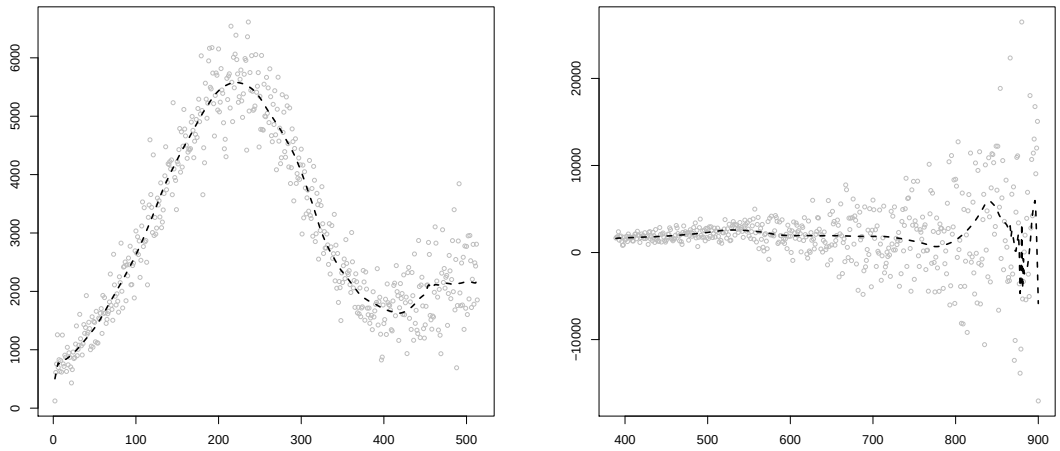
(c) Datu piekopēšana inversā spoguļattēlā

(d) Regresija ar datiem, kuri piekopēti inversā spoguļattēla

13. att. Veivletu regresija CMB datiem

Redzams, ka līkne beigās nav gluda, bet ir oscilējoša. Tātad šīs datu papildināšanas metodes nav labas. Cita alternatīva ir lietot tikai jau esošus datus un sadalīt tos divās daļās. Viena daļa ir 512 novērojumi no sākuma, bet otra daļa ir 512 novēojumi no beigām. Ideja varētu būt apskatīt katru daļu atsevišķi un tad apvienot. Bet arī šeit beigu

novērtējumos līkne nav gluda (14. attēls).



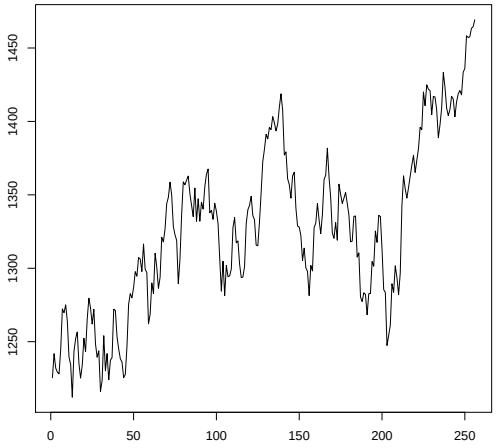
(a) Veivletu regresija 512 novērojumiem no sākuma

(b) Veivletu regresija 512 novērojumiem no beigām

14. att. Veivletu regresija CMB datiem

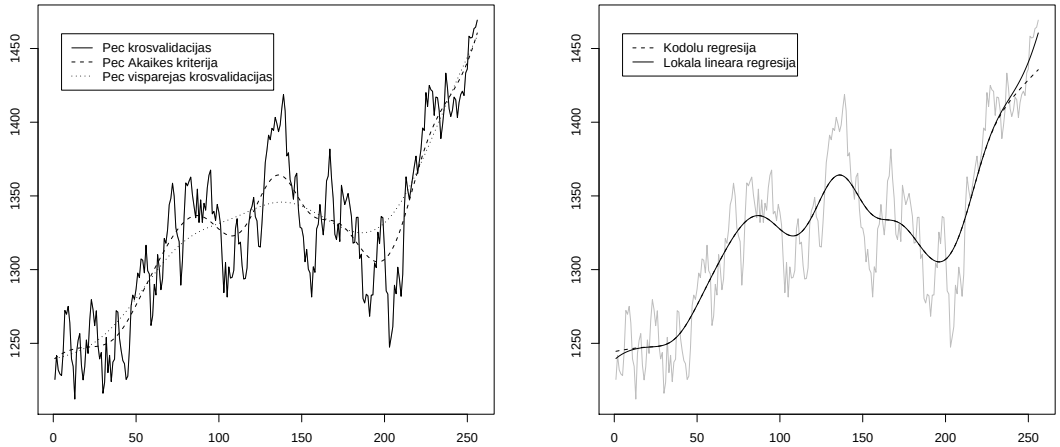
Tātad jāsecina, ka patiesais iemesls nav datu "skaita palielināšanas izvēle", bet gan tas, ka neizpildās nosacījums par atlikumu normalitāti. Vēl iemesls ir arī heteroskedastitāte. Ja sākumā līdz kādam 400 novērojumam ir homoskedastitāte, tad pēc tam parādās heteroskedasticitāte. Kā šo problēmu risināt? Iespēja varētu būt izvēlēties slieksni nevis globālo (universālo), bet gan lokālo, un $\hat{\sigma}$ ņemt katram līmenim savu minimizējot pēc SURE. 2007.gadā ir parādījusies arī publikācija par veivletu lietošanu neparametriskās regresijas gadījumos, kad ir heteroskedastitāte [14]. Bet pagaidām vairāk ar teorētiskiem aprēķiniem un mazāk ar praktiskiem piemēriem.

Otrs reālu datu piemērs ir "SP500" dati. SP500 datus var apakšielādēt Jianqing Fan mājas lapā ($http://www.orfe.princeton.edu/\sim jqfan/fan/nls/datasets.html$). SP500 dati ir *Standard and Poor's* indeksa dati, kas balstās uz 500 lielāko Amerikas uzņēmumu akciju svērtajām vidējām cenām (aptuveni 70% no Amerikas tirgus). Apskatīti tiks indeksa dati laika posmam no 1998.gada 28.decembra līdz 1999.gada 31.decembrim (15. attēls).



15. att. *Standard and Poor's* indeksa dati

Kodolu regresijas un lokālās lineārās regresijas konstruēšanā šajā gadījumā ir problēmas ar joslas platuma noteikšanu, jo krosvalidācija nestrādā. Kā redzams 16.(a) attēlā, ja izvēlēsies joslas platumu ar krosvalidāciju, tad notiks nevis gludināšana, bet gan interpolācija. Tas ir tādēļ, ka dati ir korelēti. Sīkāk šādas problēmas pētījis Jeffrey D.Hart [15]. Savukārt izvēloties joslas platumu pēc Akaikes kritērija, iegūst jau gludāku novērtējumu, bet ar vispārējo krosvalidāciju regresijas līkne tiek pat pārgludināta. Optimālās regresijas līknes pie joslas platuma h=11 ir redzamas 16.(b) attēlā.

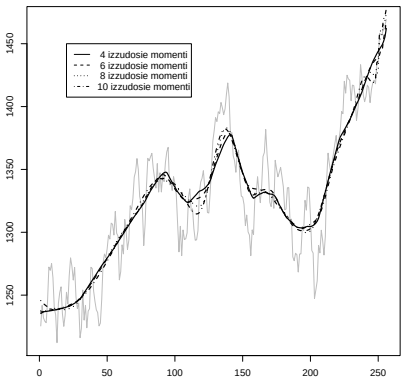


(a) Krosvalidācija pēc dažādām metodēm (b) Regresijas līknes pie h=11

16. att. Kodolu un lokālās lineārās regresijas funkcijas novērtējumi SP500 datiem

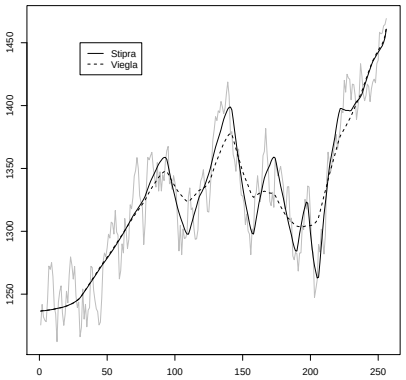
Arī šeit ir redzams, ka robežās labāka ir lokālā lineārā regresija.

Veivletu regresijas gadījumā datus var lietot uzreiz bez modificēšanas, jo ir 256 novērojumi. Regresijas līknes novērtēšanai lietots tiek vismazākās asimetrijas Daubechies veivlets. Kā redzams 17. attēlā, tad regresijas līkne, palielinoties izzūdošajiem momentiem, mainās mazliet galapunktos un vidū. Pārējā daļā ir ļoti līdzīgas. Tā kā pie 10 izzūdošajiem momentiem sākumā parādās novirze, tad jāsecina, ka optimālais regresijas funkcijas novērtējums būs gadījumā, kad ir 4-6 izzūdošie momenti.



17. att. Veivletu novērtējumi SP500 datiem

Kā redzams 18. attēlā, tad novērtējums ar 4 izzūdošajiem momentiem ir mazliet līdzīgs lokālās lineārās regresijas novērtējumam ar joslas platumu h=11. Ja salīdzina stipro sliekšņošanu ar vājo, tad redzams, ka stiprā sliekšņošana vairāk interpolē, turpretī vājā vairāk gludina.



18. att. Veivletu novērtējumi SP500 datiem

Vēl redzams arī, ka, tāpat kā lokālās lineārās regresijas novērtējumam, arī veivletu

novērtējumam galapunktos nav novērojama robežu novirze.

Pēc reālu datu analīzes jāsecina, ka ir grūti izvēlēties, kura gludināšanas parametra vērtība ir optimālākā, kurš novērtējums ir labākais.

5.2. Simulēto funkciju analīze

Tā kā pēc regresijas funkciju novērtējumu attēliem nevar viennozīmīgi secināt, kura metode ir labākā, tiek veiktas simulācijas testa funkcijām pie dažādiem izlases apjomiem.

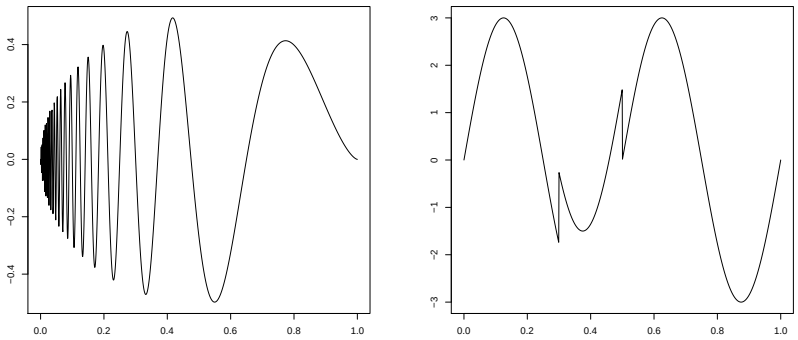
Simulētas funkcijas (19.attēls) ir Doppler funkcija un Jumpsine funkcija.

Doppler funkcija ir formā

$$r(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right), \quad 0 \leq x \leq 1.$$

Jumpsine funkcija ir formā

$$r(x) = 3(\sin(4\pi x) + 0.5 \cdot I_{\{0.3<x\leq 0.5\}})$$



(a) Doppler funkcija (b) Jumpsine funkcija

19. att. Testa funkciju īstie grafiki

Šīs funkcijas ir labas neparametriskās regresijas metožu testēšanai, jo tās ir sareģģīti novērtēt. Šīs funkcijas ir samērā gludas, izņemot dažus lēcienus vai krasu līknes formas izmaiņu.

Tādēļ ar šīm pieminētajām testa funkcijām tiek veiktas simulācijas. Veikto simulāciju procesu var rakstorot sekojoši:

1. Testa funkcijai f tiek simulēti n novērojumi;

2. Aprēķina vidējo kvadrātisko kļūdu šai vienai simulācijai

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(r(x_j) - \hat{r}(x_j))^2.$$

3. Atkārto 1. un 2. punktu m reizes, tādējādi iegūstot m MSE vērtības.

4. Aprēķina vidējo vērtību no visām m iegūtajām vidējo kvadrātisko kļūdu vērtībām (angliski *average mean squared error*)

$$AMSE = \frac{1}{m}\sum_{i=1}^{m} MSE_i.$$

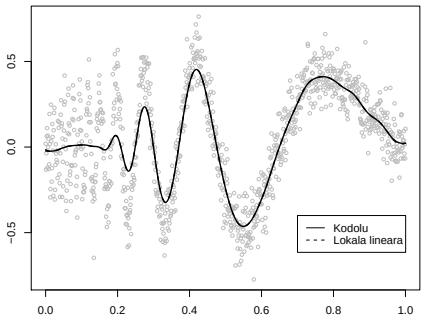
Ģenerēti tiek dati, kuri iegūti testa funkcijai pievienojot kļūdu $\sigma\epsilon$, kur $\epsilon \sim N(0,1)$ un σ ir 0.1. Simulācijās novērojumu skaits n ir 64, 128, 256, 512 un 1024, un simulācijas atkārtošanas reižu skaits $m = 50$.

Šis simulāciju process tiek pielietots Nadaraja-Vatsona kodolu regresijas metodei, lokālai lineārajai regresijai, REACT metodei un veivletiem. Simulācijās ar veivletiem tiek izmantots Daubechies veivlets no vismazākās asimetrijas klases ar 4 izzūdošiem momentiem. Kā slieksnis tiek lietots universālais slieksnis. Kodolu novērtējumiem joslas platums tiek noteikts, minimizējot risku pēc krosvalidācijas. REACT metodē ir viens veids kā nosaka saskaitāmo skaitu, proti, minimizē risku pēc SURE.

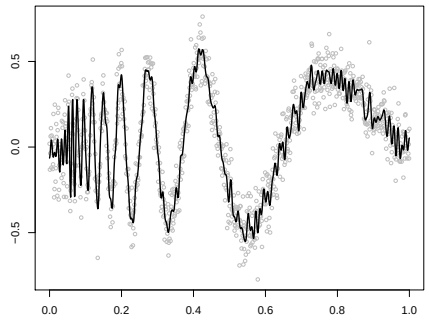
Pirmā testa funkcija ir Doppler funkcija. Doppler funkcija ir sinusoīda ar mainīgu amplitūdu un biežumu. Šīs funkcijas novērtēšanā vislielākās problēmas jebkurai metodei ir sākums, jo tur ir grūti atpazīt, kāda ir patiesā funkcija un datu struktūra. Šīs testa funkcijas regresijas līknes redzamas 20.attēlā.

1. tabula AMSE Doppler funkcijai

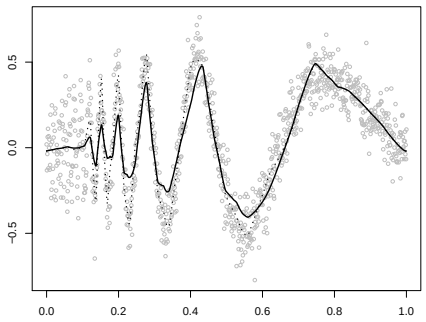
Metode	n=64	n=128	n=256	n=512	n=1024
Kodolu regresija	0.02053	0.01595	0.01156	0.00787	0.00567
Lokālā lineārā	0.02047	0.01599	0.01160	0.00788	0.00568
REACT	0.01251	0.01116	0.00646	0.00409	0.00267
Veivleti (vieglā)	0.02635	0.02905	0.02294	0.01651	0.01172
Veivleti (stiprā)	0.01926	0.01846	0.01157	0.00769	0.00479



(a) Kodolu un lokālās lineārās regresijas novērtējums



(b) REACT metodes regresijas novērtējums



(c) Veivletu regresijas novērtējums

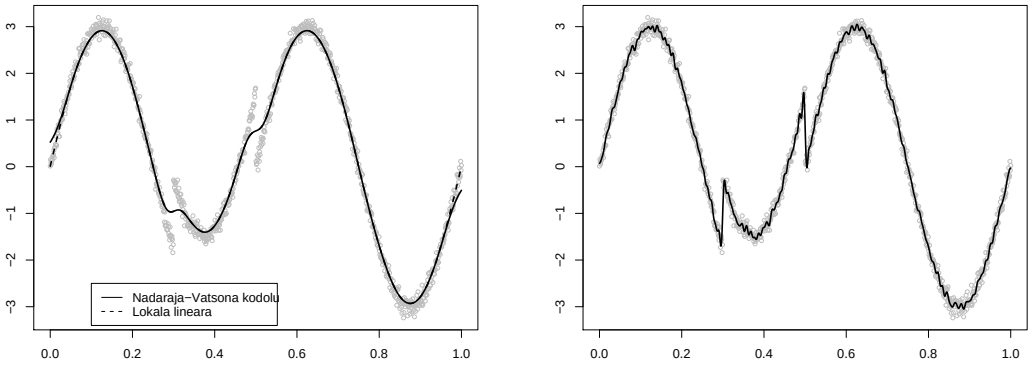
20. att. Regresijas funkcijas novērtējumi Doppler funkcijai

Salīdzinot regresijas novērtējumus, pamanāms, ka nevienai no metodēm pilnībā nav izdevies atpazīt sākumā īsto līkni, jo tur Doppler funkcijai ļoti mainās frekvence. Vistuvāko novērējumu īstajai funkcijai sākumā deva REACT metode, jo tajā ir fiksēta precīzāk īstās līknes oscilācija. Bet trūkums atkal REACT metodei ir tālākajā novērtējumā. Tur dažās vietās un jo īpaši beigās līkne ir ļoti oscilējoša. Bet kodolu un veivletu gadījumā līknes ir ļoti līdzīgas, turklāt abas metodes sākumā nefiksē frekveņču izmaiņu.

Simulāciju apkopojums redzams 1. tabulā. Skatoties pēc rezultātiem, REACT metodes

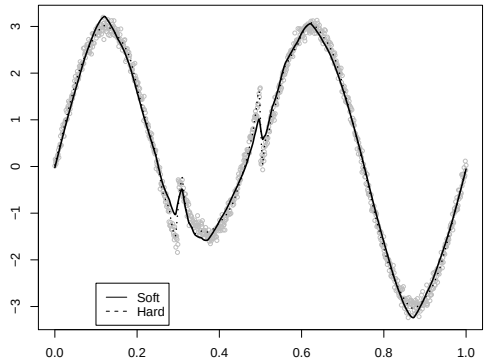
rezultāti ir vislabākie. Tātad regresijas līknes novērtējums ir precīzāks, bet kā jau iepriekš pieminēts un redzams 20.(b) attēlā, tad regresijas līkne nav pietiekoši gluda. Tādēļ nevar teikt, ka REACT novērtējums ir vislabākais. Kodolu regresijai, lokālai lineārajai regresijai un veivletiem pēc stiprās sliekšņošanas AMSE rezultāti ir līdzīgi.

Otrā testa funkcija ir Jumpsine funkcija. Šīs testa funkcijas regresijas līknes redzamas 21.attēlā.



(a) Kodolu un lokālās lineārās regresijas novērtējums

(b) REACT metodes regresijas novērtējums



(c) Veivletu regresijas novērtējums

21. att. Regresijas funkcijas novērtējumi Jumpsine funkcijai

Jumpsine funkcijai kodolu regresijas un lokālās lineārās regresijas gadījumā tiek izlaisti lēcieni pie 0.3 un 0.5. Bet pārējā daļa ir gluda. Savukārt REACT metode un veivleti fiksēja lēcienus punktos 0.3 un 0.5. REACT metodei trūkums tāpat kā Doppler funkcijas gadījumā ir tas, ka citās regresijas līknes vietās parādījās lieka, nevajadzīga oscilācija. Veivletu gadījumā līkne ir gluda.

2. tabula AMSE Jumpsine funkcijai

Metode	n=64	n=128	n=256	n=512	n=1024
Kodolu regresija	42.9802	20.6482	9.7575	8.2246	6.2114
Lokālā lineārā	42.9239	20.5396	9.7058	8.1991	6.1923
REACT	0.0334	0.0168	0.0094	0.0061	0.0043
Veivleti (vieglā)	0.1214	0.0867	0.0652	0.0429	0.0267
Veivleti (stiprā)	0.0905	0.0405	0.0267	0.0119	0.0063

Ja skatās AMSE rezultātus 2. tabulā, tad šeit neapšaubāmi vissliktākie rezultāti ir kodolu regresijai un lokālai lineārajai regresijai. It īpaši, kad izlases apjoms ir mazs. Kad apjoms ir lielāks, tad kļūda jau samazinās, bet tik un tā kļūda ir ļoti liela salīdzinot ar pārējām metodēm. REACT metode līdzīgi kā Doppler funkcijas gadījumā pēc AMSE rezultātiem uzrāda vismazāko kļūdu, bet tā kā regresijas līkne nav gluda, tad es neuzskatu, ka tas ir labākais novērtējums. Veivletu gadījumā līkne ir gluda, kļūda ir diezgan maza un arī lēcieni punktos 0.3 un 0.5 tiek fiksēti. Tātad tik tiešām telpiski nehomogēnām funkcijām veivleti izrādās labāki par citām metodēm. Tas ir pateicoties telpas-laika lokalizācijai.

Bet, kā redzējām reālos datu piemēros, tad tur lokālās lineārās regresijas metode bija pat labāka nekā veivleti un REACT, jo tur datos nebija pēkšņu lēcienu vai būtiskas datu struktūras izmaiņas.

Kuru metodi izvēlēties ir ļoti atkarīgs no datiem konkrētajā situācijā. Nevar vienno-zīmīgi pateikt, kura metode ir labākā. Visām metodēm ir savi trūkumi, priekšrocības.

6. Secinājumi

Darbā tika apskatītas neparametriskās regresijas metodes regresijas funkcijas novērtēšanai. Apskatītas tradicionālās kodolu regresijas un lokālo polinomu regresijas metodes, kā arī jaunākas metodes - veivleti un ortogonālo funkciju metode.

Ortogonālu funkciju un veivletu regresijas novērtējumu izteiksmē saskaitāmie rindā ir neatkarīgi viens no otra, un tādējādi viens saskaitāmais neietekmē to, kāda būs līkne kaut kur tālāk. REACT metodes trūkums varētu būt ortonormālās bāzes izvēle. Ja izvēlas kosinusu bāzi (kā tas ir šajā darbā), tad metode strādā labi, bet, lietojot ortogonālus polinomus, problēma ir ar polinomu kārtu, jo bāzē ir nepieciešami ortogonāli polinomi, kuriem kārta ir vienāda ar izlases apjomu. Pie lieliem izlases apjomiem varētu būt pagrūti uzkonstruēt ortonormālu bāzi.

Kodolu regresijas novērtējumam trūkums ir robežu novirze galapunktos. Bet šo problēmu samazina vai likvidē, ja lieto lokālos polinomus. Tādējādi lokālie polinomi ir labāki nekā kodolu regresija, jo tie jau automātiski likvidē robežu biasu. Arī veivletiem, kur metode strādāja korekti, robežu novirzes nebija.

Veivletiem zināms trūkums ir prasība, lai datu skaits ir divnieka pakāpe. Jo modificējot, pārveidojot datus, lai izlases apjoms būtu divnieka pakāpē, palielinās iespējamība, ka novērtējums varētu būt kļūdains. Izvēršot darba rezultātus varētu pārdomāt, vai ir iespējama kāda vispārīgāka metode datu palielināšanai (samazināšanai) līdz divnieka pakāpei. Darba laikā radās cits variants, kā varētu rīkoties, ja nav divnieka pakāpes. Iespējams, ka tāds variants jau ir izstrādāts, bet ideja ir ņemt m blokus (intervālus), kur $m = 2^J$ un katram blokam aprēķināt vienu vērtību, kas ir vidējā vērtība no visiem punktiem, kuri atrodas konkrētajā blokā.

Tā kā veivletus statistikā sāka lietot samērā nesen, tad vēl ir daudz teorētisku problēmu, kuras var tikt analizētas. Piemēram, varētu apskatīt heteroskedastitātes problēmu, kas nesen 2007.gadā tikusi apskatīta teorētiski. Tomēr praktiskām datu problēmām sīkāka analīze nav veikta. Veivletu metodes lieto arī laikrindu analīzē, it sevišķi prognozes veikšanā (skatīt [16],[17]).

Visām neparametriskās regresijas metodēm nav iespējams viennozīmi noteikt gludinošo parametru neatkarīgi no pielietotā riska minimizācijas veida. Katra riska funkcija un tā minimizācija dod savu rezultātu un līdz ar to piedāvā savu optimālo gludināšanas parametru. Faktiski būtu nepieciešams veikt dziļāku analīzi ar simulāciju palīdzību, lietojot

testa funkcijas. Tomēr praktiskos piemēros vienmēr papildus ir jāveic arī vizuāla datu analīze.

Izmantotā literatūra un avoti

[1] Johnstone I.M. Donoho D.L. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[2] Mark G. Low. Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23:824–835, 1995.

[3] Härdle W. *Applied Nonparametric Regression.* Cambridge University Press. Cambridge., 1990.

[4] Larry Wasserman. *All of Nonparametric Statistics.* Springer, 2006.

[5] Rudolf Beran. React scatterplot smoothers: Superefficiency through basis economy. *Journal of the American Statistical Association*, 95:155–171, 2000.

[6] Dümbgen L. Beran, R. Modulation of estimators and confidence sets. *The Annals of Statistics*, 26:1826–1856, 1998.

[7] Charles M.Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9:1135–1151, 1981.

[8] Johnstone I.M. Donoho D.L. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

[9] Pikard D. un Tsybakov A. Härdle W., Kerkyacharian G. *Wavelets, Approximation, and Statistical Applications.* Springer-Verlag, New York, NY, 1998.

[10] Ingrid Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[11] David F.Walnut. *An Introduction to Wavelet Analysis.* Birkhäuser, Boston, 2002.

[12] Johnstone I.M. Donoho D.L. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26:879–921, 1998.

[13] Nason G.P. *Wavelet Methods in Statistics with R.* Springer, 2008.

[14] Yan-Yan Qi Han-Ying Liang. Asymptotic normality of wavelet estimator of regression function under na assumptions. *Bull.Korean Math. Soc.*, 2007.

[15] Jeffrey D. Hart. Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:173–187, 1991.

[16] Walden A.T. Percival D.B. *Wavelet Methods for Time Series Analysis.* Cambridge University Press, Cambridge, 2000.

[17] F.Murtagh O.Renaud, J.-L.Strack. Wavelet - based forecasting of short and long memory series. 2002.

1. Programmas R kodi

```
#####     REACT  regresija
orto<-function(xdati,ydati)
{
   vec<-c()
   fii<-function(j,x)
   {
   f<-1
   f[j>1]<-sqrt(2)*cos((j-1)*pi*x)
   f
   }
   n<-length(xdati)
   zz<-c()
   zz2<-c()
   for(j in 1:n)
   {
   zz[j]<-1/n*sum(sapply(1:n,function(i){ydati[i]*fii(j,xdati[i])}))
   zz2[j]<-(zz[j])^2
   }
   sigma<-1/2/(n-1)*sum((ydati[2:n]-ydati[1:(n-1)])^2)
   Risks<-function(J)
   {
   n<-length(xdati)
   J*sigma/n+sum(sapply((J+1):n,function(j){max((zz2[j]-sigma/n),0)}))
   }
   hhh<-seq(1,length(xdati)-1,by=1)
   index<-order(sapply(hhh,Risks))[1]
   novJ<-hhh[index]
   ortoNov<-function(J,x){sum(sapply(1:J,function(j){zz[j]*fii(j,x)}))}
   for (k in 1:length(xdati)){
   vec[k]<-ortoNov(novJ,xdati[k])
   }
   vec
}

#####     Nadaraya-Watson kodolu regresija
nadwatson<-function(xdati,ydati)
{
   h<-h.select(xdati,ydati,method="cv")
   locpoly(xdati,ydati,bandwidth=h,degree=0,gridsize=length(xdati))
}

#####      Lokala lineara regresija
linloc<-function(xdati,ydati)
{
   h<-h.select(xdati,ydati,method="cv")
   locpoly(xdati,ydati,bandwidth=h,degree=1,gridsize=length(xdati))
}
```

```
#####      Veivletu regresija
# Soft thresholding veivletu regresija
softwavelet<-function(xdati,ydati)
{
   vec<-c()
   wds<-wd(ydati,filter.number=4,family="DaubLeAsymm",bc="symmetric")
   thS.wds<-threshold(wds,type="soft",boundary=TRUE)
   trecS<-wr(thS.wds)
   vec<-trecS[1:length(xdati)]
   vec
}
# Hard thresholding veivletu regresija
hardwavelet<-function(xdati,ydati)
{
   vec<-c()
   wds<-wd(ydati,filter.number= ,family="DaubLeAsymm",bc="symmetric")
   thH.wds<-threshold(wds,type="hard",boundary=TRUE)
   trecH<-wr(thH.wds)
   vec<-trecH[1:length(xdati)]
   vec
}

#####       Testa funkcijas
# Dopller funkcija
doppler<-function(x)
{
  sqrt(x*(1-x))*sin(2.1*pi/(x+0.05))
}
#  Bloku funkcija
blocks<-function(x)
{
  t<-c(0.10, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)
  h<-c(4,-5, 3,-4, 5,-4.2, 2.1, 4.3,-3.1, 2.1,-4.2)
  sgn<-function(y)
  {
    ss<-0
    ss[y>0]<-1
    ss[y<0]<--1
    ss
  }
  ssgn<-function(z){(1+sgn(z))/2}
  summa<-0
  for (j in 1:11){summa<-summa+(h[j]*ssgn(x-t[j]))}
  summa
}
# JumpSine funkcija
jumpsine<-function(x)
```

```
{
  y<-3*(sin(4*pi*x))
  y[x>0.3 & x<=0.5]<-3*(sin(4*pi*x)+0.5*1)
  y
}

#####        Simulacijas
library(KernSmooth)
library(sm)
library(wavelets)
library(wavethresh)
testafunkcija<-function(x){doppler(x)}
sigma<-0.1
reizes<-50
for (n in c(32,64,128,256,512,1024))
{
  summa.orto<-0
  summa.kodols<-0
  summa.linloc<-0
  summa.soft<-0
  summa.hard<-0
  xdati<-seq(0,1,len=n)
  for (m in 1:reizes)
  {
    ydati<-sapply(xdati,function(x){testafunkcija(x)+sigma*rnorm(1,0,1)})
    ista<-sapply(xdati,function(x){testafunkcija(x)})
    mse.orto<-(sum((ista-orto(xdati,ydati))^2))/n
    summa.orto<-summa.orto+mse.orto
    mse.linloc<-(sum((ista-linloc(xdati,ydati)$y)^2))/n
    summa.linloc<-summa.linloc+mse.linloc
    mse.kodols<-(sum((ista-nadwatson(xdati,ydati)$y)^2))/n
    summa.kodols<-summa.kodols+mse.kodols
    mse.soft<-(sum((ista-softwavelet(xdati,ydati))^2))/n
    summa.soft<-summa.soft+mse.soft
    mse.hard<-(sum((ista-hardwavelet(xdati,ydati))^2))/n
    summa.hard<-summa.hard+mse.hard
  }
  amse.orto<-summa.orto/reizes
  amse.linloc<-summa.linloc/reizes
  amse.kodols<-summa.kodols/reizes
  amse.soft<-summa.soft/reizes
  amse.hard<-summa.hard/reizes
  cat("n=",n,"AMSE.orto=",amse.orto,"AMSE.linloc=",
  amse.linloc,"AMSE.kodols=",amse.kodols,"AMSE.soft=",
  amse.soft,"AMSE.hard=",amse.hard,"\n")
}
```

Maģistra darbs "Neparametriskā regresija lietojot ortogonālas funkcijas un veivletus" izstrādāts LU Fizikas un Matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Haralds Plivčs

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: doc. Dr.math. Nadežda Siņenko

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts maģistra gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)