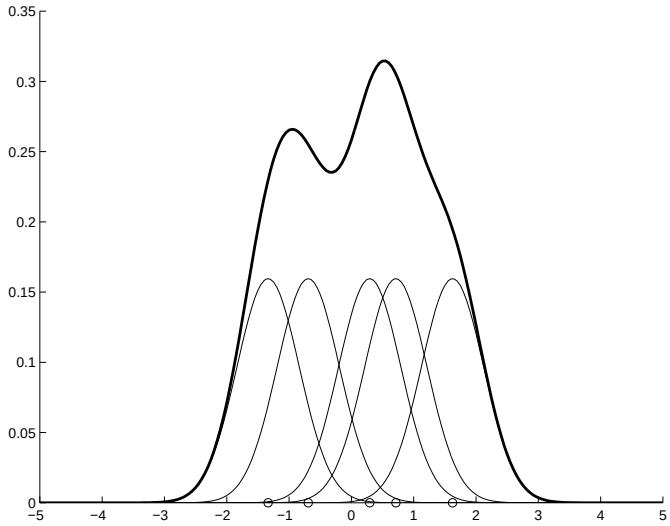
Funktioiden estimointi

Lasse Holmström
Matemaattisten tieteiden laitos
Oulun yliopisto

Kevät 2011



Sisällys

Luku 1

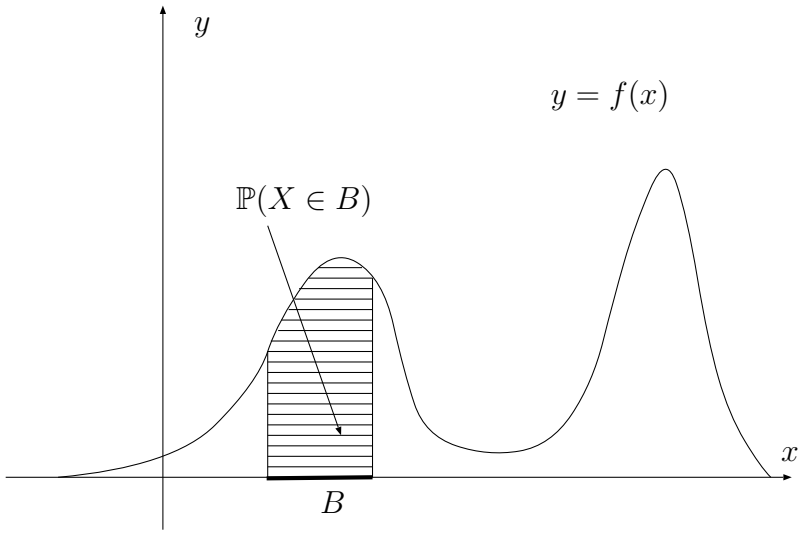
Esimerkkejä funktion estimoinnista

1.1 Tiheysfunktio

Olkoon X (reaaliarvoinen) satunnaismuuttuja. Oletetaan, että X:n jakaumalla on tiheysfunktio $f : \mathbb{R} \to [0, \infty[$,

$$\mathbb{P}(X \in B) = \int_B f(x)dx, \tag{1.1}$$

(vrt. Kuva 1.1).



Kuva 1.1: Satunnaismuuttujan X tiheysfunktio f.

Olkoon $X_1, \ldots, X_n$ satunnaisotos X:n jakaumasta.

Tehtävä: Estimoi f:ää otoksen $X_1, \ldots, X_n$ avulla.

Ratkaisuna saadaan estimaattori $\hat{f}(x; X_1, \ldots, X_n)$, $x \in \mathbb{R}$. Merkitään tätä lyhyesti $\hat{f}(x)$. Perinteisesti yleisin estimointitapa on käyttää *histogrammia*.

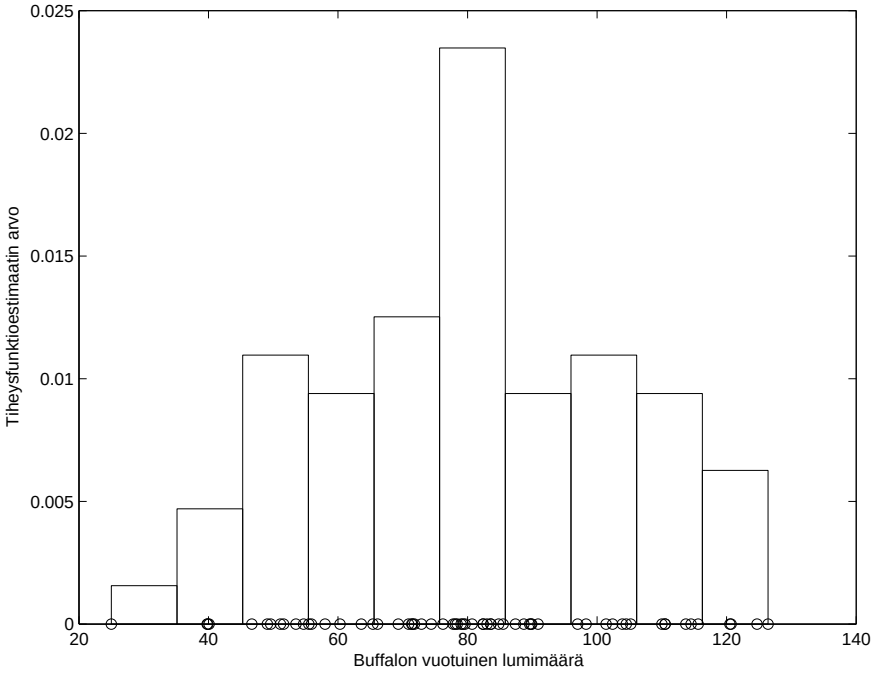
Tiheysfunktion estimointi on avain monen ongelman ratkaisuun:

- aineiston havainnollistaminen, visualisointi
- moodien paikallistaminen (yksi, monta?)
- ryhmittelyanalyysi (engl. clustering)
- simulointi
- hahmontunnistus
- jne.

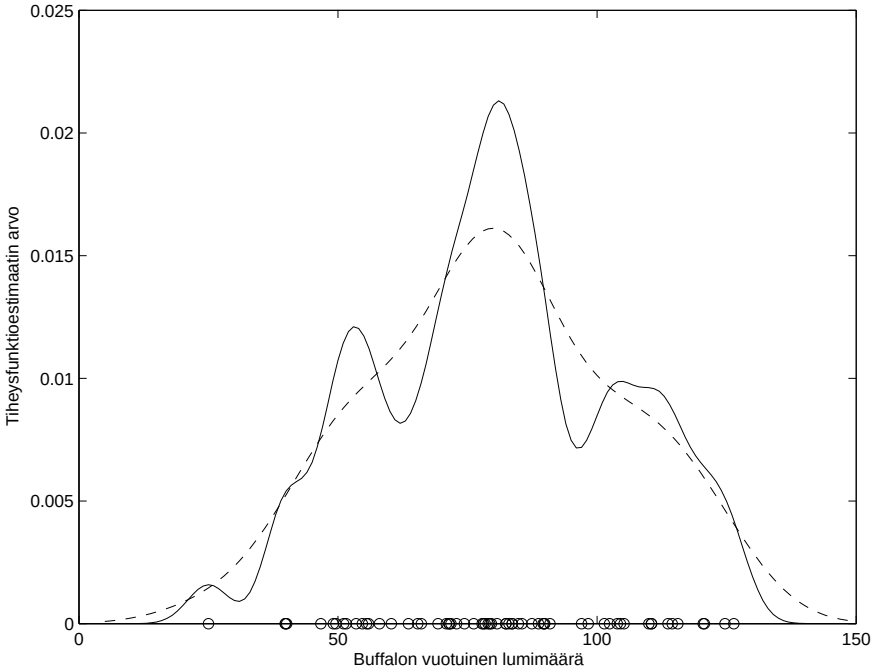
Esimerkki 1.1 Tarkastellaan Buffalon kaupungin (New York, USA) vuotuista lumisademäärää vuosina 1910 - 1972. Satunnaismuuttujana X on vuoden aikana satanut lumi tuumina. Otos ($n = 63$) koostuu seuraavista mitatuista arvoista [8]:

```
126.4  82.4  78.1  51.1  90.9  76.2 104.5  87.4 110.5  25.0
 69.3  53.5  39.8  63.6  46.7  72.9  79.6  83.6  80.7  60.3
 79.0  74.4  49.6  54.7  71.8  49.1 103.9  51.6  82.4  83.6
 77.8  79.3  89.6  85.5  58.0 120.7 110.5  65.4  39.9  40.1
 88.7  71.4  83.0  55.9  89.9  84.8 105.2 113.7 124.7 114.5
115.6 102.4 101.4  89.8  71.5  70.9  98.3  55.5  66.1  78.4
120.5  97.0 110.0
```

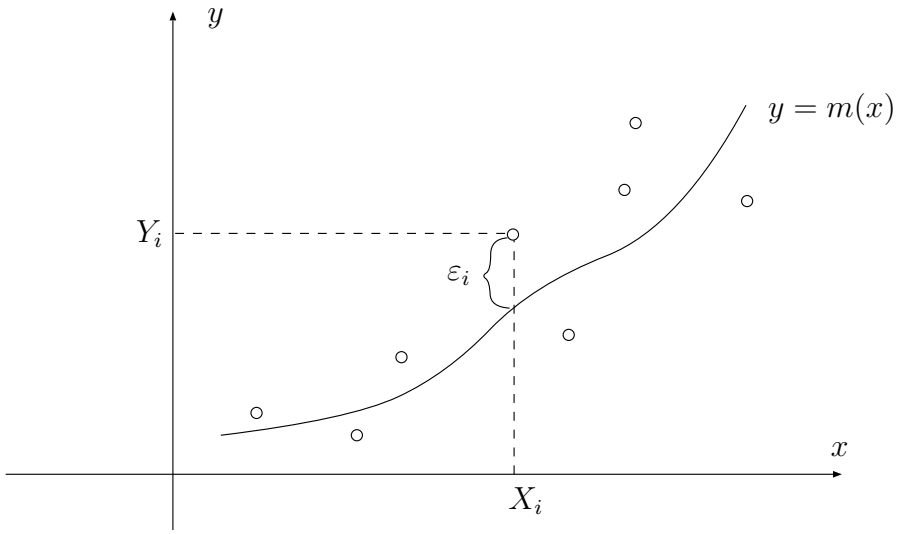
Kuvassa 1.2 on tästä aineistosta muodostettu histogrammi ja kuvassa 1.3 kaksi muuta tiheysfunktion estimaattia, jotka on muodostettu käyttäen ns. ydinmenetelmää (luku 3.2). Estimointitavoista riippuen tiheysfunktion moodien (lokaalien maksimien) lukumäärälle saadaan erilaisia arvioita. ‖



Kuva 1.2: Histogrammi Buffalon kaupungin vuotuisesta lumisademäärästä (tuumissa) Otospisteiden arvot on merkitty vaaka-akselille pienillä ympyröillä.



Kuva 1.3: Buffalon kaupungin vuotuisen lumisademäärän (tuumissa) tiheysfunktion kaksi eri estimaattia (yhtenäinen viiva ja katkoviiva). Otospisteiden arvot on merkitty vaaka-akselille pienillä ympyröillä.



Kuva 1.4: Regressiotehtävä. Kuvassa Y on selitettävä muuttuja ja X on selittävä muuttuja.

1.2 Regressio

Olkoot X ja Y satunnaismuuttujia ja olkoon $(X_1, Y_1), \ldots, (X_n, Y_n)$ satunnaisotos (X, Y):n jakaumasta. Esimerkkinä voisi olla vaikka X_i = henkilön i paino, Y_i = henkilön i pituus. Mallitetaan X:n ja Y:n riippuvuutta toisistaan kaavalla

$$Y_i = m(X_i) + \varepsilon_i.$$

Tässä $m : \mathbb{R} \to \mathbb{R}$ on (regressio)funktio ja ε_i mallittaa satunnaisvirhettä (esim. mittausvirhe tai puuttuva informaatio). Tilannetta on havainnollistettu kuvassa 1.4.

Tehtävä: Estimoi m:ää otoksen $(X_1, Y_1), \ldots, (X_n, Y_n)$ avulla.

Ratkaisuna konstruoidaan estimaattori $\hat{m}(x; (X_1, Y_1) \ldots, (X_n, Y_n))$, $x \in \mathbb{R}$, tai lyhyesti $\hat{m}(x)$.

Regressiota voidaan käyttää monenlaisiin tehtäviin:

- aineiston havainnollistaminen
- muuttujien välisten riipuvuuksien tutkiminen

- ennustaminen
- jne.

Joskus arvot X_i eivät ole satunnaisia vaan ennalta valittuja kiinteitä lukuja.

Esimerkki 1.2 Kahdeksan miehen paino ja pituus on annettu talukossa 1.1.

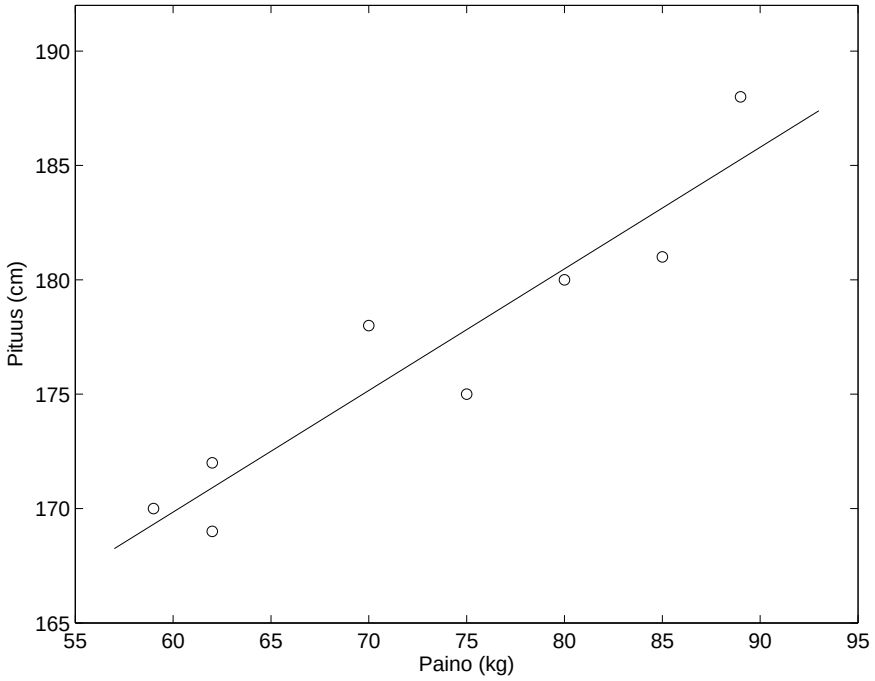
Paino	*Pituus*
62	172
70	178
59	170
85	181
80	180
75	175
62	169
89	188

Taulukko 1.1: Kahdeksan miehen paino (kg) ja pituus (cm).

Kuvassa 1.5 on tähän aineistoon sovitettu suora, joka näyttääkin kuvaavan paino ja pituuden keskimääräistä riippuvuutta melko hyvin. ‖

Esimerkki 1.3 Moottoripyöräilijän kypäriä testattiin simuloiduilla törmäyskokeilla (esim. [5]). Kuvassa 1.6 on esitetty ajajan pään kiihtyvyys törmäyshetkestä kuluneen ajan funktiona $n = 133$ mittauspisteen avulla. Kuvassa 1.7 on tähän aineistoon sovitettu eri asteisia polynomeja regressiofunktion estimaateiksi. Polynomit ovat selvästikin liian jäykkiä funktioita hyvän sovitteen saamiseksi. Kuvassa 1.8 on käytetty ns. lokaalia lineaarista regressiota (luku 4.6) jolloin tulos on paljon parempi. ‖

Esimerkki 1.4 Englannissa on tutkittu perunan kulutuksen riippuvuutta perheen tuloista [7]. Kuvassa 1.9 on esitetty kerätty aineisto 4094 perheen osalta vuodelta 1973. Yksiköinä on käytetty keskiarvon monikertoja (1 = keskiarvo) ja mukaan on otettu vain perheet, joissa perunaa on ylipäätänsä kulutettu ja joiden tulot ja kulutus ovat korkeintaan kolminkertaiset keskiarvoon nähden. Kuvaan on piirretty



Kuva 1.5: Kahdeksan miehen paino ja pituus ja tähän aineistoon sovitettu suora.

sekä lineaarinen että Nadarayan-Watsonin menetelmällä (4.2) saatava regressiofunktion estimaatti. Huomaa kuinka epäuskottavan kuvan lineaarinen estimaatti antaa perunankulutuksen riippuvuudesta tuloista. Nadararyan-Watsonin menetelmä antaa selvästi luontevamman tuloksen: kulutuksen kasvu tasaantuu tietyn tulotason jälkeen ja itseasiassa vähenee korkeimmissa tuloluokissa. ‖

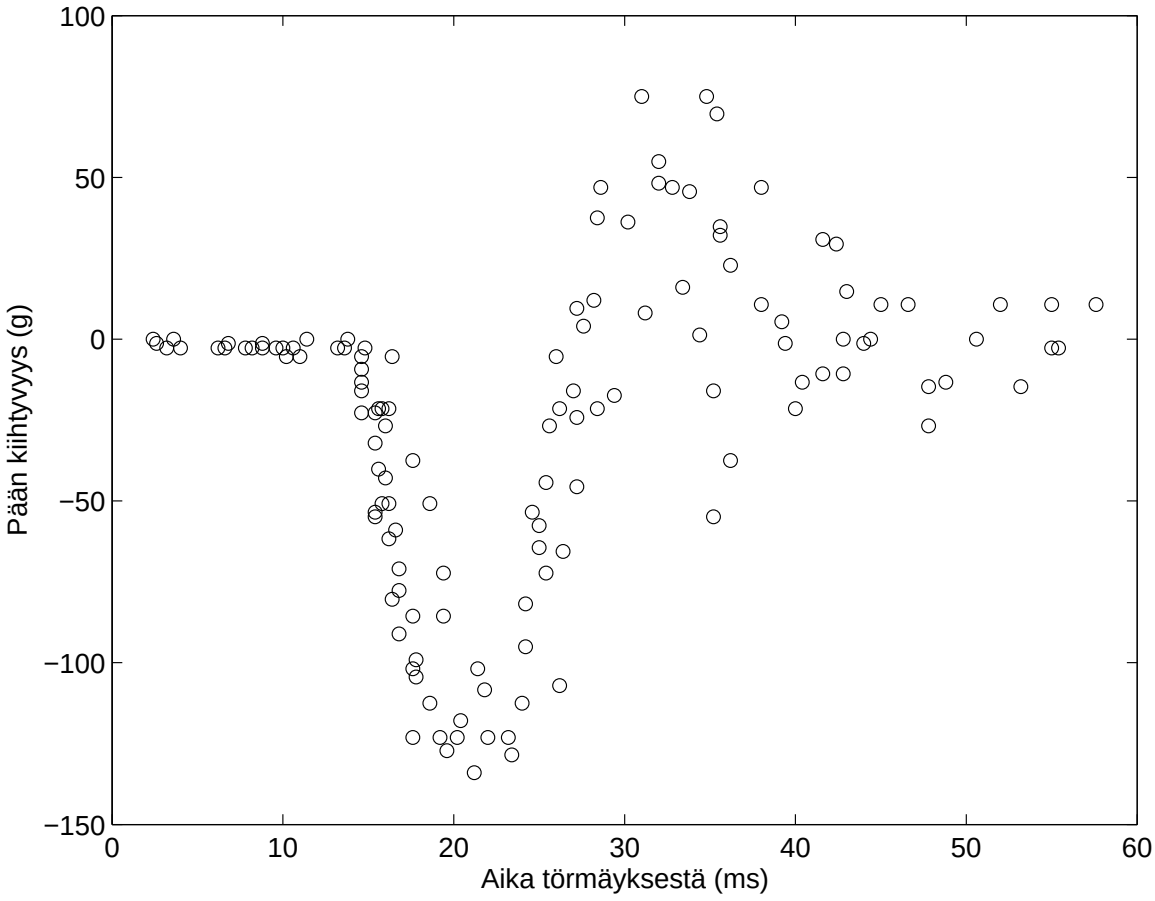
1.3 Tehospektri

Tarkastellaan aikasarjaa $(X_t)_{t\in\mathbb{Z}}$ missä indeksijoukkona ovat kokonaisluvut, $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots\}$ ja meillä on ajassa etenevä jono havaintoja

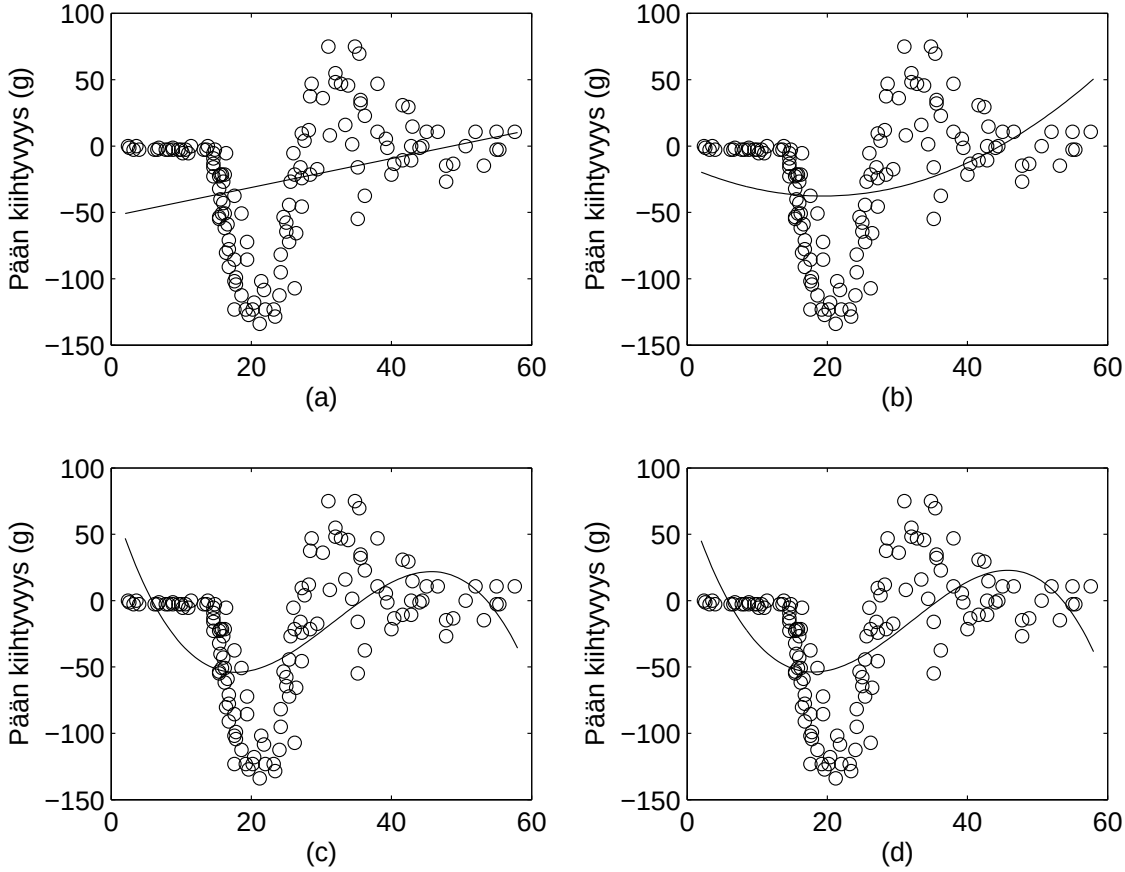
$$\dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots .$$

Tässä kukin X_t on satunnaismuuttuja. Oletetaan, että aikasarja on (heikosti) stationaarinen ja keskiarvoltaan 0:

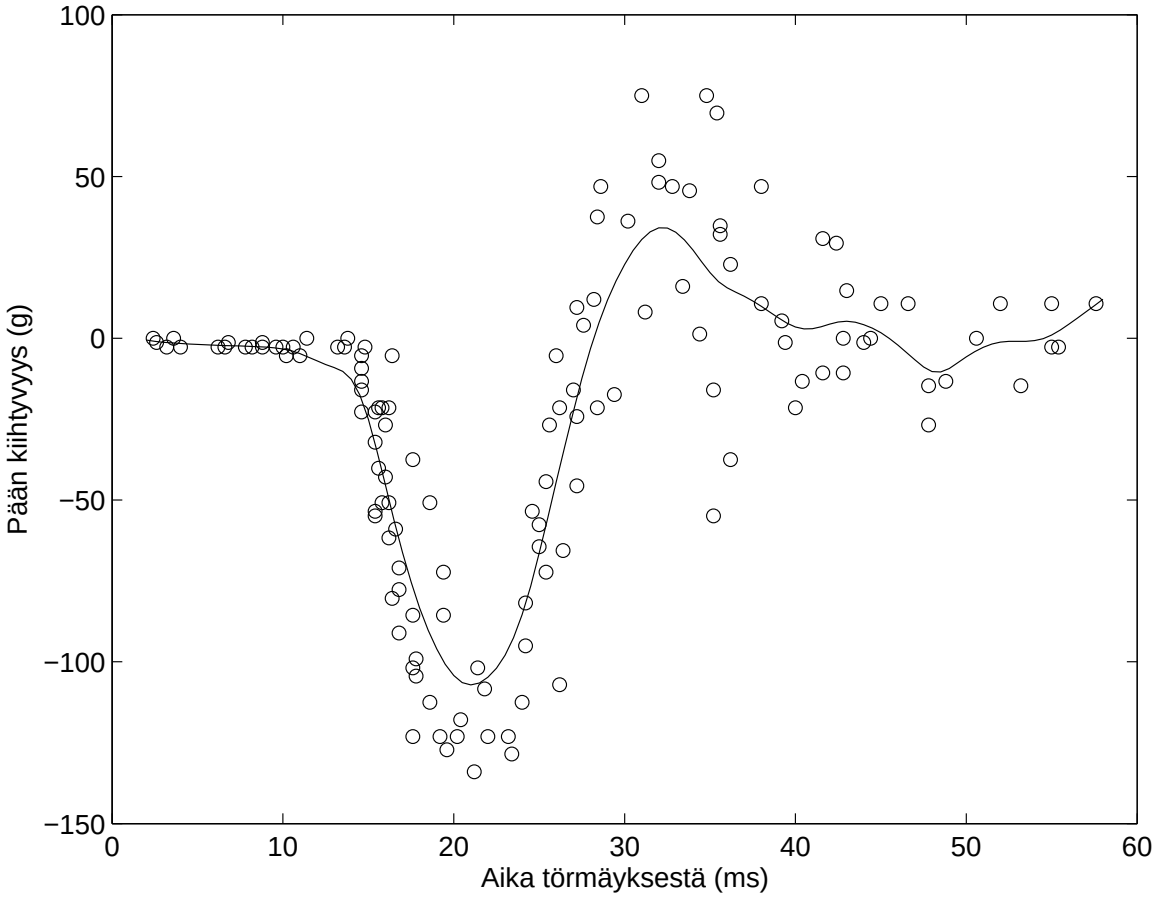
$$\mathbb{E}X_t = 0 \text{ kaikilla } t,$$



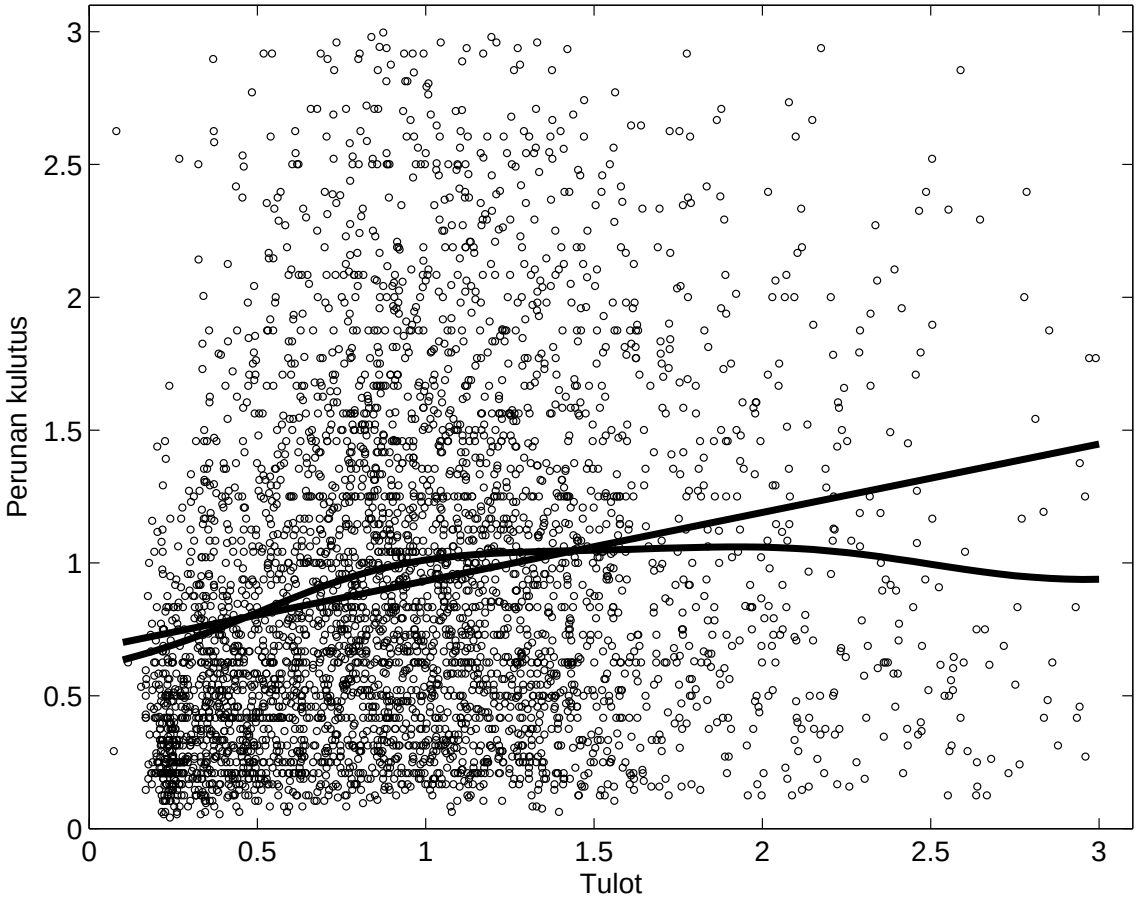
Kuva 1.6: Simuloiduissa törmäyksissä mitattu moottoripyöräilijän pään kiihtyvyys törmäyksestä kuluneen ajan funktiona.



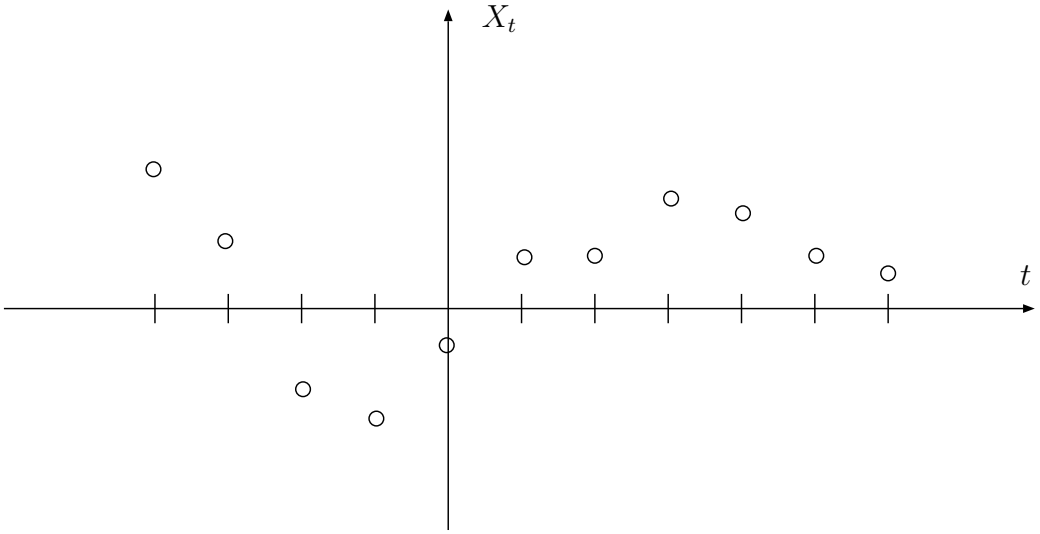
Kuva 1.7: Regressiofunktion estimointi polynomilla kuvan 1.6 aineistosta: ensimmäisen asteen polynomi (a), toisen asteen polynomi (b), kolmannen asteen polynomi (c) ja neljännen asteen polynomi (d).



Kuva 1.8: Regressiofunktion estimointi lokaalilla lineaarisella regressiolla kuvan 1.6 aineistosta.



Kuva 1.9: Perunan kulutuksen riippuvuus perheen tulotasosta (Englanti 1973). Tiedot on kerätty $n = 4094$ perheeltä ja yksikköinä on käytetty keskiarvon moninkertoja. Kuvaan on piirretty sekä lineaarinen että Nadarayan-Watsonin menetelmällä saatava regressiofunktion estimaatti.



Kuva 1.10: Aikasarja $(X_t)_{t\in\mathbb{Z}}$.

$$\gamma(u) = \mathbb{E}X_t X_{t+u},\ u \in \mathbb{Z},\ \text{ei riipu } t\text{:stä.}$$

Yllä $\mathbb{E}$ merkitsee odotusarvoa ja $(\gamma(u))_{u\in\mathbb{Z}}$ on aikasarjan $(X_t)_{t\in\mathbb{Z}}$ *autokovarianssijono*.

Aikasarjoja on monenlaisia:

- taloudelliset aikasarjat
- säähavainnot
- erilaiset mittaussignaalit teknisissä sovelluksissa

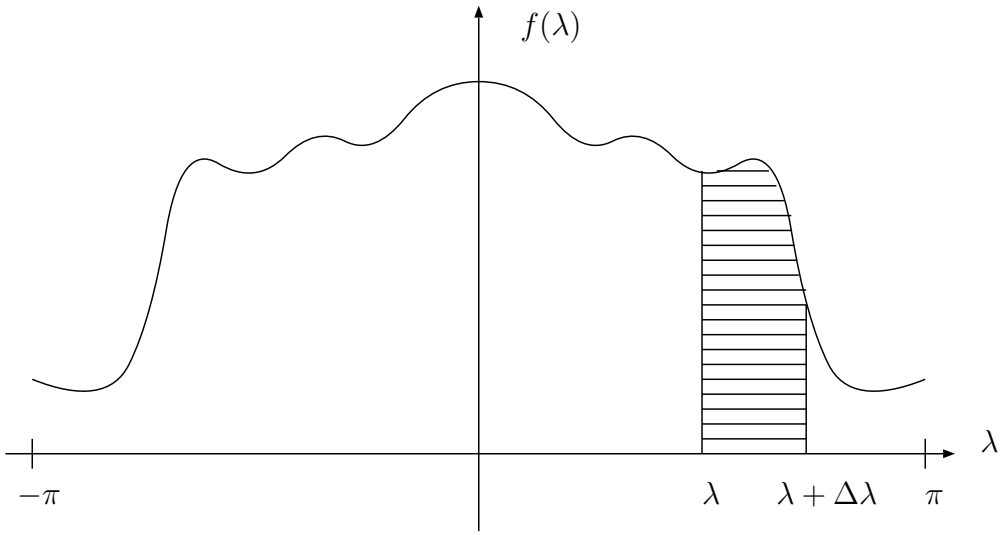
Probleema: Onko aikasarjassa periodisuutta, piileviä syklisiä komponentteja?

Oletetaan, että $\sum_{u=-\infty}^{\infty} |\gamma(u)| < \infty$. Määritellään aikasarjan $(X_t)_{t\in\mathbb{Z}}$ *tehospektri* $f : \mathbb{R} \to \mathbb{R}$ kaavalla

$$f(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \gamma(u) e^{-i\lambda u},\ \lambda \in \mathbb{R}.$$

Kyseessä on siis itseasiassa autokovarianssijonon $(\gamma(u))_{u\in\mathbb{Z}}$ Fourier-muunnos. Voidaan osoittaa, että

(i) f on 2π-jaksollinen.



Kuva 1.11: Tehospektrin tulkinta.

(ii) $f(\lambda) \geq 0$ kaikilla λ.

(iii) f on symmetrinen, eli $f(\lambda) = f(-\lambda)$ kaikilla λ.

(iv) $f(\lambda)\Delta\lambda \propto$ taajuskaistaa $[\lambda, \lambda + \Delta\lambda]$ vastaava teho aikasarjassa $(X_t)_{t\in\mathbb{Z}}$ (vrt. kuva 1.11).

Idea: f:n lokaali maksimi vastaa aikasarjan $(X_t)_{t\in\mathbb{Z}}$ periodista komponenttia.

Käytännössä pystymme kuitenkin havaitsemaan vain äärellisen pituisen aikasarjan.

Tehtävä: Estimoi f:ää äärellisen havaintojonon $X_1, \ldots, X_n$ avulla.

Estimoidaan ensin $\gamma(u)$:ta,

$$\hat{\gamma}(u) = \begin{cases} (1/n)\sum_{t=1}^{n-u} X_t X_{t+u}, & u = 0, \ldots, n-1 \\ 0, & u \geq n \\ \hat{\gamma}(-u), & u < 0. \end{cases}$$

Sitten otetaan f:n estimaattoriksi *periodogrammi*

$$\hat{f}(\lambda) = \frac{1}{2\pi}\sum_{u=-\infty}^{\infty} \hat{\gamma}(u)e^{-i\lambda u} \overset{\text{HT}}{=} \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t e^{-i\lambda t}\right|^2 .$$

(Tässä samoin kuin jatkossa merkintä $\stackrel{\text{HT}}{=}$ tarkoittaa, että yhtäsuuruus todistetaan harjoitustehtävänä).

Nyt voidaan osoittaa, että $\mathbb{E}\hat{f}(\lambda) \to f(\lambda)$, kun $n \to \infty$ ja $\lambda \neq 0 \pmod{2\pi}$. Ikävä kyllä kuitenkin $\mathbb{E}[\hat{f}(\lambda) - f(\lambda)]^2 \not\to 0$ kun $n \to \infty$, joten $\hat{f}(\lambda)$ heilahtelee $f(\lambda)$:n ympäristössä suurillakin n:n arvoilla vaikka sen arvo keskimäärin onkin $f(\lambda)$. Tilanne saadaan korjattua silottamalla periodogrammia lähekkäisiä arvoja keskiarvoistamalla. Historiallisesti tämä periodogrammiin liittyvä estimointiongelma motivoi myös todennäköisyystiheysfunktion estimointimenetelmien kehittämistä.

Esimerkki 1.5 Kuvassa 1.12 on esitetty keskimääräinen vuotuinen auringonpilkkujen määrä vuosina 1770 - 1988 (lähteenä H. Tong: *Non-linear time series: a dynamical systems approach*, 1991, Oxford University Press). Kuvassa 1.13 on logaritmisella asteikolla tästä aineistosta laskettu periodogrammi ja sen silotettu versio. Symmetrisyyden ja jaksollisuuden vuoksi periodogrammit on piirretty vain välillä $[0, \pi]$. Huomaa voimakas maksimi noin taajuudella 0.6, joka vastaa alkuperäisessä aikasarjassa periodia $2\pi/0.6 \approx 10.5$ vuotta. ‖

1.4 Hasardifunktio

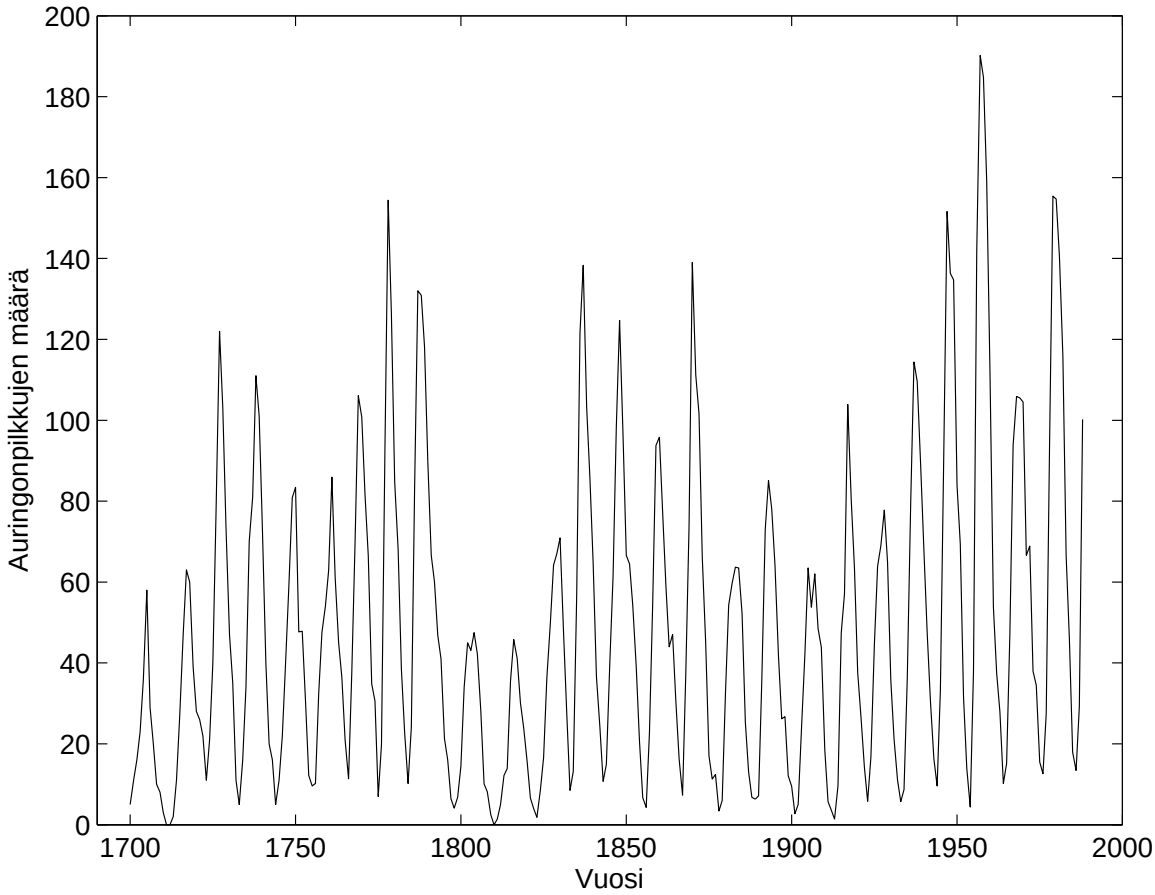
Hasardifunktion käsitettä (engl. hazard function) käytetään mm. luotettavuuden valvonnassa ja elinaika-analyysissä. Laitteiden luotettavuuden yhteydessä käytetään myös nimitystä "vikaantumisintensiteetti".

Olkoon T tarkasteltavan kohteen, esimerkiksi laitteen tai henkilön, toiminta tai elinaika (vrt. kuva 1.14). Olkoon T:llä tiheysfunktio f ja merkitään sen kertymäfunktiota F:llä,

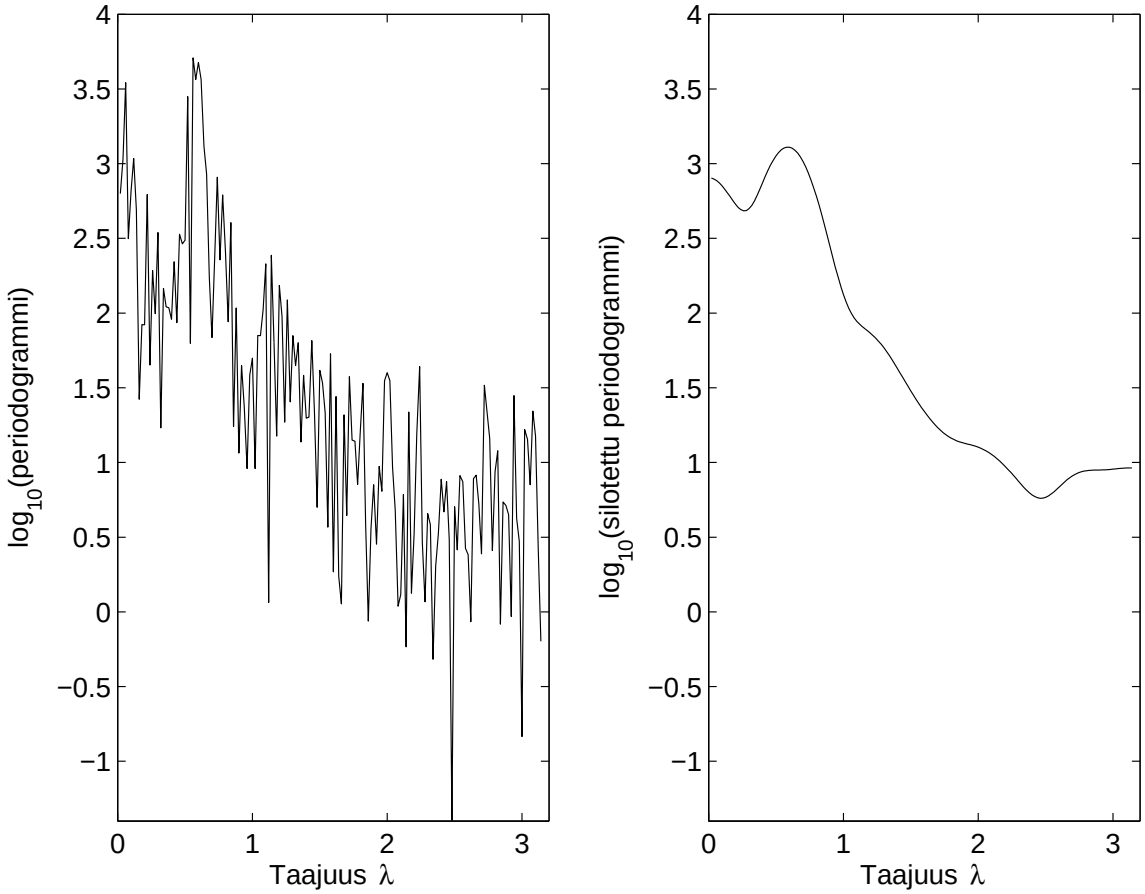
$$F(t) = \int_0^t f(\tau)d\tau.$$

Hasardifunktio määritellään nyt kaavalla

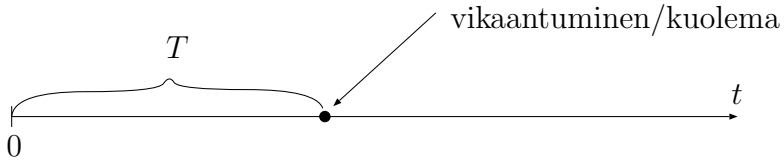
$$H(t) = \frac{f(t)}{1 - F(t)}, \qquad t \geq 0.$$



Kuva 1.12: Auringonpilkkujen keskimääräinen vuotuinen määrä vuosina 1770 - 1988.



Kuva 1.13: Kuvan 1.12 aineistosta laskettu periodogrammi (vasemmalla) ja sen silotettu versio (oikealla).



Kuva 1.14: Laitteen/henkilön toiminta/elinaika T.

Huomataan, että

$$H(t)\Delta t = \frac{f(t)\Delta t}{1 - F(t)}$$
$$\approx \mathbb{P}(\text{vikaantuminen välillä } [t, t+\Delta t] \mid \text{ehjä hetkeen } t \text{ asti}).$$

Kun käytettävissä on satunnaisotos $T_1, \ldots, T_n$, voidaan tiheysfunktiolle f muodostaa estimaattori $\hat{f}$ (vrt. tämän luvun osa 1.1). Siten saadaan estimaattori

$$\hat{F}(t) = \int_0^t \hat{f}(\tau) d\tau$$

mistä edelleen

$$\hat{H}(t) = \frac{\hat{f}(t)}{1 - \hat{F}(t)}, \qquad t \geq 0.$$

1.5 Hahmontunnistus

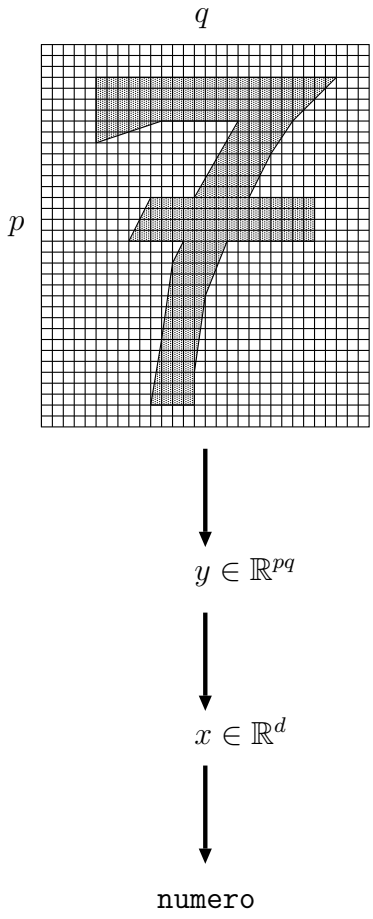
Olkoon X jostain kohteesta tehty havainto; hahmontunnistuksessa tämä on itseasiassa tavallisesti paremminkin tulkittavissa satunnais*vektorina* kuin reaaliarvoisena satunnaismuuttujana. Esimerkkejä ovat

- digitaalinen kuva (esim. konenäkö)
- puhespektri (puheentunnistus)
- EEG käyrä (potilaan tilan seuranta)

Probleema: Mistä "luokasta" havainto on peräisin?

Edellisten esimerkkien kohdalla tämä voisi tarkoittaa esimerkiksi seuraavia ksysymyksiä.

- Mikä esine on kuvassa?
- Mikä foneemi?
- Mikä on potilaan tila?



Kuva 1.15: Käsinkirjoitetun numeron tunnistus.

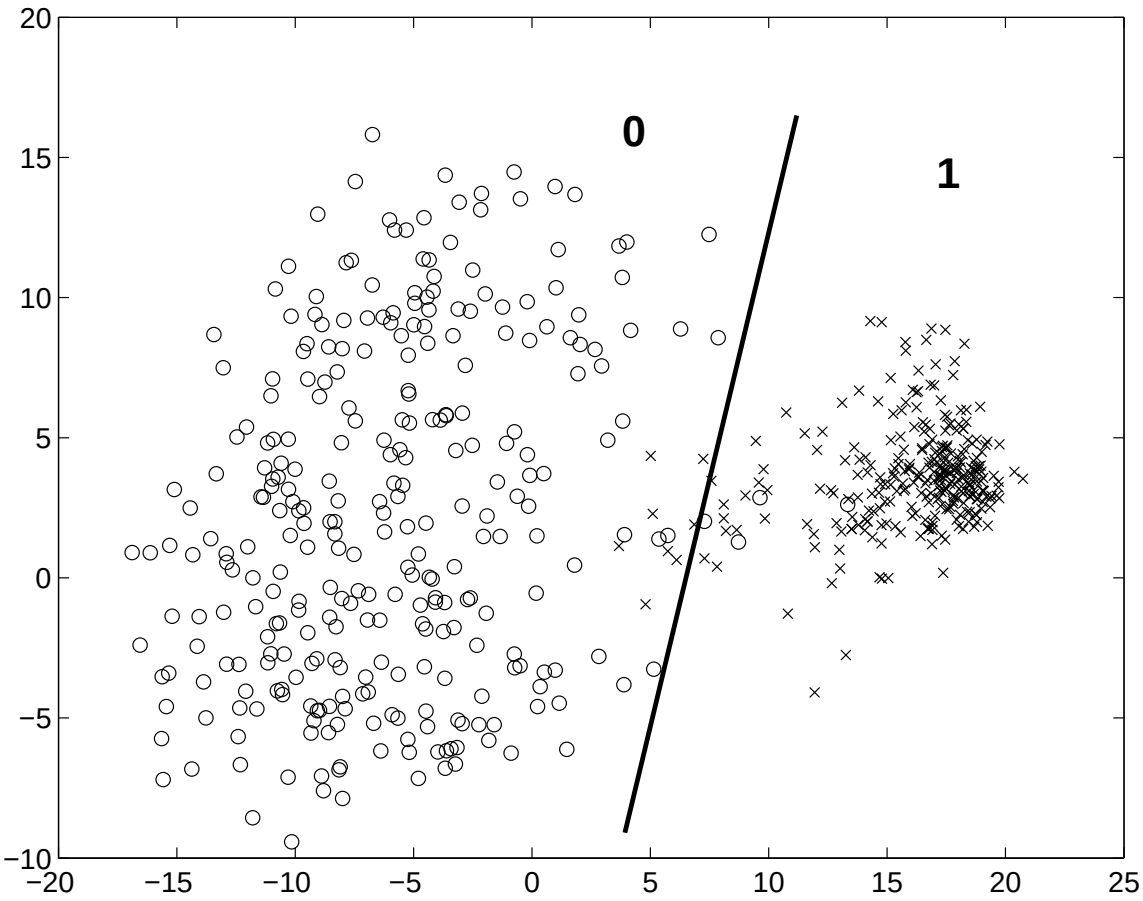
Esimerkki 1.6 Kuva 1.15 esittää yksinkertaistetussa muodossa erään mahdollisen lähestymistavan käsinkirjoitettujen numeroiden automaattiseksi tunnistamiseksi. Numerosta otetaan ensin $p \times q$ digitaalinen kuva y. Sitten kuva esikäsitellään ja sen sisältämä informaatio tiivistetään "piirrevektoriksi" x, jonka dimensio d on tyypillisesti paljon alempi kuin y:n dimensio pq. Piirrevektori x luokitellaan viimein johonkin luokista $0, 1, \ldots, 9$ käyttäen jotain päätössääntöä.

Kuvassa 1.16 on esimerkki eräästä todellisesta aineistosta. Alunperin 1024-ulotteiset (32×32 digitaalinen kuva), käsinkirjoitettuja numeroita 0 ja 1 esittävien vektoreiden sisältämä informaatio tiivistettiin kaksiulotteisiin piirrevektoreihin. Varsin hyvä päätössääntö näiden kahden numeron erottamiseksi toisistaan saadaan jakamalla taso kahteen osaan kuvaan piirretyn suoran avulla. Virheitä toki tehdään mutta näin yleensä onkin, sillä eri luokkia vastaavien piirrevektoreiden paikat eivät useinkaan satu pistevieraisiin joukkoihin. ‖

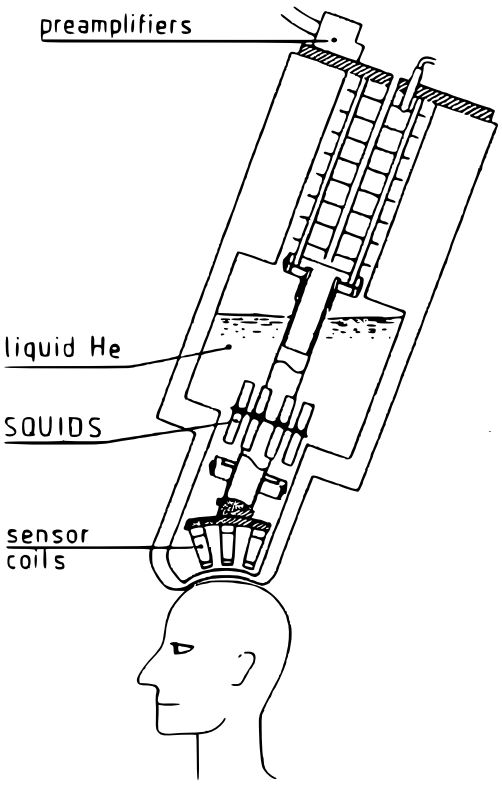
Esimerkki 1.7 Aivojen toimintaa voidaan mitata monella eri tavoilla. Perinteisen EEG:n rinnalle on viime aikoina noussut esimerkiksi MRI-kuvaus. Suomessa on kehitetty myös pään ulkopuolelta aivoista mitattuun magneettikenttään perustuvaa MEG menetelmää (MagnetoEncePhalograhy). Mittauslaitteistossa käytetään äärimmäisen herkkiä SQUID sensoreita (Superconducting QUantum Inteference Device) hyvin heikkojen magneettikenttien muutoksien havaitsemiseksi (ks. kuva 1.17).

Eräässä kokeessa koehenkilön tuli reagoida painamalla oikeassa tai vasemmassa kädessä olevaa painonappia kuullessaan joko korkean (1000 Hz) tai matalan (500 Hz) äänen kuulokkeista. Yhdessä mittaussessiossa korkeita ja matalia ääniä tuotettiin kumpiakin satunnaisessa järjestykssä n. 100 kappaletta. Tehtävänä oli konstruoida hahmontunnistusjärjestelmä, joka MEG-mittausten avulla luokittelee napin painallukset kahteen luokkaan (oikea/vasen). Kokeen tarkoituksena oli testata alustavasti mahdollisuutta käyttää MEG laitteistoa aivotoiminnan diagnosointiin kliinisissä tutkimuksissa.

Kuvassa 1.18a on esitetty kaksi esimerkkiä kummankin luokan esikäsitellyistä mittaussignaaleista, jotka on saatu yhdestä kaikkiaan seitsemästä kanavasta. Käytetty mittausjakso vastaa napin painalluksesta alkanutta 150 ms pituista ajanjaksoa ja yksittäisessä signaalissa on 60 2.5 millisekunnin välein mittattua arvoa. Lopullisena tehtävänä on luokitella näitä 60-ulotteisia piirrevektoreita. Mittaussignaaleissa



Kuva 1.16: Kaksiulotteisia piirrevektoreita, jotka on muodostettu käsinkirjoitetuista numeroista 0 (ympyrät) ja 1 (ristit). Kummastakin numerosta on 300 esimerkkiä. Kuvaan on myös piirretty suora, joka erottaa nämä kaksi luokkaa hyvin.



Kuva 1.17: Kaaviokuva aivotoiminnan mittauksesta MEG-laitteistolla.

on paljon "kohinaa", joka ei tosin johdu niinkään itse mittalaitteesta kuin itse mitattavan ilmiön (napin painallus peukalolla) kannalta epäoleellisten aivoprosessien aiheuttamista magneettikentän vaihteluista. Keskiarvoistamalla suuri joukko mittaussignaaleja nähdään, että kahden luokan välillä kuitenkin on selvä ero (kuva 1.18b). Eri menetelmiä testattaessa päästiin parhaimmillaan 13% virhetodennäköisyyteen yksittäisiä signaaleja luokiteltaessa. ‖

Oletetaan yksinkertaisuuden vuoksi, että meillä on kaksi mahdollista luokkaa, "1" ja "2". Useamman luokan tapaus voidaan käsitellä vastaavasti. Havaintoa ja siihen liittyvää luokkainformaatiota mallitetaan parilla (X, J), missä

$$\begin{aligned} X &= d\text{-ulotteinen satunnaisvektori, "hahmo",} \\ J &= X\text{:n luokka (1 tai 2).} \end{aligned}$$

Probleema: Kun X havaitaan, on arvattava vastaava J!

Tehtävälle konstruoidaan ratkaisu muodostamalla *luokitin* $g : \mathbb{R}^d \to \{1, 2\}$ ja asettamalla (ks. kuva 1.19)

$$\begin{aligned} g(X) = 1 &\Rightarrow X \text{ luokitellaan luokkaan 1,} \\ g(X) = 2 &\Rightarrow X \text{ luokitellaan luokkaan 2.} \end{aligned}$$

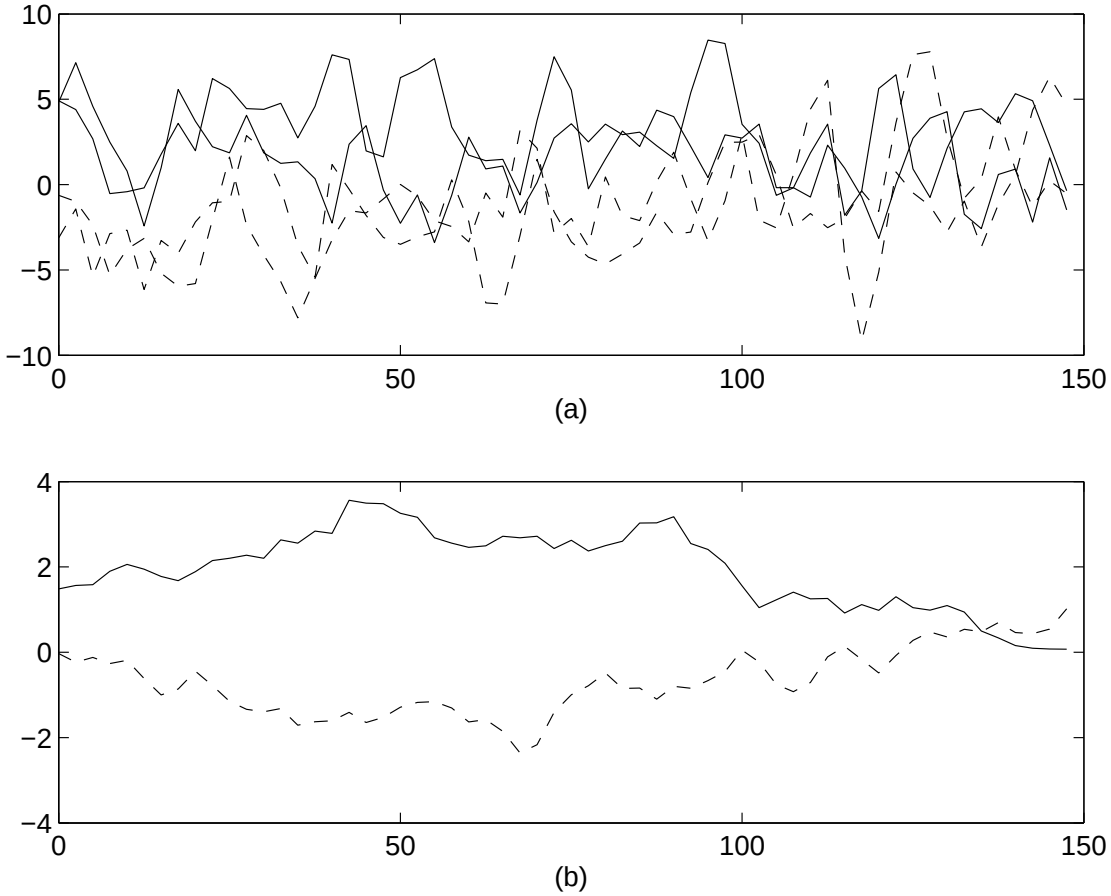
Oletetaan, että on annettu *opetusdata*, eli joukko luokiteltuja hahmoja $(X_1, J_1), \ldots, (X_n, J_n)$.

Tehtävä: Estimoi luokitin $\hat{g}$, joka on hyvä siinä mielessä, että se luokittelee myöhemmin tehtävät havainnot mahdollisimman oikein.

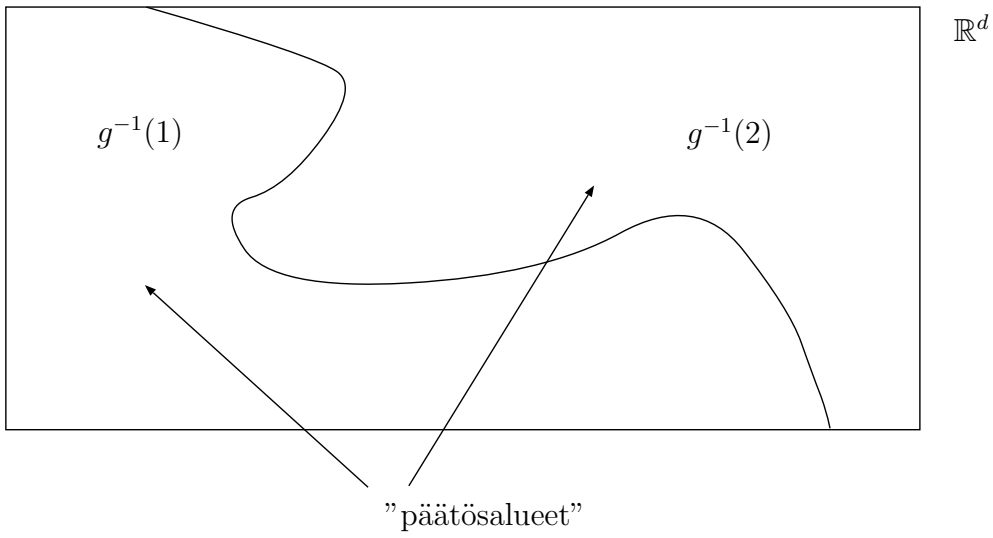
Oletetaan, että

(i) $X|J = 1 \sim f_1$, $X|J = 2 \sim f_2$, eli X:n jakauman luokassa i määrää tiheysfunktio f_i.

(ii) $\mathbb{P}(J = 1) = \mathbb{P}(J = 2)$, eli luokilla 1 ja 2 on yhtäsuuret "prioritodennäköisyydet".



Kuva 1.18: Kuvassa a on kaksi signaalia oikean (yhtenäinen viiva) ja kaksi signaalia vasemman (katkoviiva) käden napinpainalluksista mitatuista MEG signaaleista. Kuvassa b on keskiarvoistettu 94 oikean käden ja 92 vasemman käden signaalia. Vaakaakselilla on aika millisekunteina napin painalluksesta.



Kuva 1.19: Luokittelija g euklidisessa avaruudessa.

Voidaan osoittaa, että tällöin paras mahdollinen luokitin (tekee vähiten virheitä) saadaan määrittelemällä

$$g(X) = \underset{i=1,2}{\operatorname{argmax}} f_i(X) = \begin{cases} 1, & \text{kun} f_1(X) > f_2(X) \\ 2, & \text{kun} f_1(X) \leq f_2(X). \end{cases}$$

Nyt, jos $\hat{f}_i$ on opetusdataan perustuva f_i:n estimaattori, $i = 1, 2$, niin tätä optimaalista luokitinta voidaan estimoida määrittelemällä

$$\hat{g}(X) = \underset{i=1,2}{\operatorname{argmax}} \hat{f}_i(X).$$

Luku 2

Parametrinen ja parametriton funktion estimointi

2.1 Perusasioita

Olkoon $(\Omega, \mathcal{A}, \mathbb{P})$ todennäköisyyskenttä. Siis $\mathcal{A}$ on Ω:n sigma-algebra (tapahtumien joukko) ja kun $A \in \mathcal{A}$, on $\mathbb{P}(A)$ tapahtuman A todennäköisyys.

Olkoon $X : \Omega \to \mathbb{R}$ satunnaismuuttuja. Siis, X on mitallinen eli

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{A}$$

kaikilla Borelin joukoilla $B \subset \mathbb{R}$.

Määritellään X:n odotusarvo $\mathbb{E}X$ ja varianssi $\mathrm{Var}(X)$ kaavoilla

$$\mathbb{E}X = \int_\Omega X d\mathbb{P}, \qquad \mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$

Jotta nämä suureet olisivat hyvin määriteltyjä, oletetaan odotusarvon tapauksessa, että X on integroituva ja varianssin tapauksessa, että X^2 on integroituva .

Olkoon sitten X:n jakaumalla tiheysfunktio f (kaava (1.1)); merkitsemme tällöin $X \sim f$. Silloin

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x) dx, \qquad \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}X)^2 f(x) dx.$$

Olkoon edelleen $a \in \mathbb{R}$. Silloin pätee

$$\mathbb{E}(X - a)^2 = \mathbb{E}[(X - \mathbb{E}X) + (\mathbb{E}X - a)]^2 = \mathrm{Var}(X) + (\mathbb{E}X - a)^2. \qquad (2.1)$$

Olkoon sitten $\Theta \subset \mathbb{R}^d$ ja olkoon jokaisella $\theta \in \Theta$ annettu tiheysfunktio $f(\cdot;\theta) : \mathbb{R} \to [0,\infty[$. Merkitään näin saatavaa θ:lla parametroitua tiheysfunktioperhettä

$$\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\}. \tag{2.2}$$

Esimerkki 2.1 Otetaan parametrijoukoksi $\Theta = \mathbb{R}\times]0,\infty[$ ja asetetaan

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}\left(\frac{x-\theta_1}{\sqrt{\theta_2}}\right)^2},\ x \in \mathbb{R},\ \theta = (\theta_1,\theta_2).$$

Tavallisesti merkitään $\theta_1 = \mu$, $\theta_2 = \sigma^2$. Kyseessä on tietenkin normaalijakauman $N(\mu,\sigma^2)$ tiheysfunktio. Merkitsemme tiheysfunktiolle joskus $f(\cdot;(\mu,\sigma^2)) \sim N(\mu,\sigma^2)$. Jos X:llä on jakauma $N(\mu,\sigma^2)$, merkitsemme myös $X \sim N(\mu,\sigma^2)$. Tunnetusti μ ja σ^2 ovat X:n odotusarvo ja varianssi. ‖

Tarkastellaan jotain parametrista perhettä (2.2) ja olkoon $X \sim f(\cdot;\theta)$ jollain $\theta \in \Theta$, jota emme tunne. Oletetaan, että meillä on kuitenkin käytössä i.i.d. otos $X_1,\ldots,X_n \sim f(\cdot;\theta)$, eli satunnaismuuttujat X_i ovat riippumattomia ja samoin jakautuneita tiheysfunktiolla $f(\cdot;\theta)$ (i.i.d. = independent and identically distributed). Parametrivektoria θ estimoidaan sopivan *estimaattorin* $\hat{\theta}_n = t_n(X_1,\ldots,X_n)$ avulla. Tässä $t_n : \mathbb{R}^n \to \Theta$ on (Borel) mitallinen funktio ja tavallisten todennäköisyyslaskennan merkintäsopimusten mukaisesti estimaattori on siis kuvaus $\hat{\theta}_n : \Omega \to \Theta$, $\hat{\theta}_n(\omega) = t_n(X_1(\omega),\ldots,X_n(\omega))$, $\omega \in \Omega$, itseasiassa satunnaismuuttuja.

Esimerkki 2.2 Olkoon $X \sim N(\theta,1)$ eli X:n tiheysfunktio on

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2},\ x \in \mathbb{R}.$$

Koska θ on X:n odotusarvo, on satunnaisotokseen $X_1,\ldots,X_n$ perustuva θ:n luonteva estimaattori $\hat{\theta}_n = (1/n)\sum_{i=1}^n X_i$. Tässä siis $t_n(x_1,\ldots,x_n) = (1/n)\sum_{i=1}^n x_i$, $(x_1,\ldots,x_n) \in \mathbb{R}^n$. ‖

Huomautus 2.3 Käytännössä havaitaan vain lukujoukko $x_1,\ldots,x_n$ eli $X_1(\omega),\ldots,X_n(\omega)$ jollain kiinteällä $\omega \in \Omega$. Sanomme, että $t_n(x_1,\ldots,x_n)$ on θ:n *estimaatti*. Usein merkitään kuitenkin (epätäsmällisesti) $\hat{\theta}_n = t_n(x_1,\ldots,x_n)$. ‖

2.2 Cramerin-Raon alaraja

Tarkastellaan 1-ulotteista tilannetta, $\Theta \subset \mathbb{R}$, $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\}$. Olkoon $X_1, \ldots, X_n \sim f(\cdot;\theta)$ i.i.d. otos ja $\hat{\theta}_n = t_n(X_1, \ldots, X_n)$ θ:n estimaattori. Olemme kiinnostuneita estimaattorin $\hat{\theta}_n$ neliöllisestä virheestä $\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2$, missä odotusarvo lasketaan satunnaisvektorin $(X_1, \ldots, X_n)$ jakauman suhteen ja merkinnällä $\mathbb{E}_\theta$ halutaan korostaa sitä, että muuttujilla X_i on tiheysfunktiona $f(\cdot;\theta)$:

$$\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [t_n(x_1, \ldots, x_n) - \theta]^2 f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n.$$

Huomaa, että $(X_1, \ldots, X_n)$:n tiheysfunktio on tulo satunnaismuuttujien X_i tiheysfunktioista, koska oletamme riippumattomuuden.

Kaavan (2.1) perusteella

$$\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = (\mathbb{E}_\theta\hat{\theta}_n - \theta)^2 + \mathrm{Var}_\theta(\hat{\theta}_n).$$

Tässä $\mathbb{E}_\theta\hat{\theta}_n - \theta$ on estimaattorin *harha* ja käytämme sille merkintää $\mathrm{Bias}_\theta(\hat{\theta}_n)$. Voimassa on siis seuraava *harha-varianssi hajoitelma*:

$$\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = \mathrm{Bias}_\theta^2(\hat{\theta}_n) + \mathrm{Var}_\theta(\hat{\theta}_n). \tag{2.3}$$

Olkoon S joukko. Merkitsemme S:n karakteristista funktiota

$$1_S(x) = \begin{cases} 1 & \text{kun } x \in S, \\ 0 & \text{kun } x \notin S. \end{cases}$$

Olkoon $\Theta =]a, b[$ avoin väli ja asetetaan seuraavat säännöllisyysehdot kun $\theta \in \Theta$.

(i) Joukko $S = \{x \in \mathbb{R} \mid f(x;\theta) > 0\}$ ei riipu θ:sta.

(ii) $\partial f(\cdot;\theta)/\partial\theta$ on integroituva ja

$$\frac{d}{d\theta}\int_S f(x;\theta)dx = \int_S \frac{\partial f(x;\theta)}{\partial\theta}dx.$$

(iii)

$$\begin{aligned}\frac{d}{d\theta}\int_{S^n} t_n(x_1, \ldots, x_n)\prod_{i=1}^n f(x_i;\theta)dx_1 \cdots dx_n \\ = \int_{S^n} t_n(x_1, \ldots, x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^n f(x_i;\theta)dx_1 \cdots dx_n.\end{aligned}$$

(Erityisesti merkittyjen integraalien tulee olla olemassa.)

(iv)

$$0 < \mathbb{E}_\theta \left[\left(\frac{\partial \log f(X_i;\theta)}{\partial \theta} \right)^2 \right] < \infty,$$

missä log tarkoittaa luonnollista logaritmia.

Huomautus 2.4 Tarkkaan ottaen ehdossa (iv) oleva satunnaismuuttuja $Y_1 = \partial \log f(X_1;\theta)/\partial\theta$ ei ole määritelty kun $X_1 \notin S$, koska tällöin $f(X_1;\theta) = 0$ kaikilla θ. Sovimme, että $Y_1 = 0$, kun $X_1 \notin S$. Odotusarvon laskemisen kannalta tällä uudelleen määrittelyllä ei ole merkitystä, koska johtuen joukon S määritelmästä, tapahtuman $X_1 \notin S$ todennäköisyys on tietysti nolla. ‖

Merkitään $b(\theta) = \mathrm{Bias}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta\hat{\theta}_n - \theta$, $\theta \in \Theta$.

Lause 2.5 (Cramerin-Raon alaraja) *Jos em. ehdot (i) - (iv) ovat voimassa, pätee kaikilla $\theta \in \Theta$ epäyhtälö*

$$\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 \geq \frac{[1 + b'(\theta)]^2}{\mathbb{E}_\theta \left[\left(\frac{\partial \log f(X_1;\theta)}{\partial \theta} \right)^2 \right]} \cdot \frac{1}{n} \,. \tag{2.4}$$

Todistus: Satunnaisvektorilla $(X_1, \ldots, X_n)$ on tiheysfunktio $(x_1, \ldots, x_n) \mapsto \prod_{i=1}^n f(x_i;\theta)$. Siten

$$\theta + b(\theta) = \mathbb{E}_\theta\hat{\theta}_n = \int_{S^n} t_n(x_1, \ldots, x_n) \prod_{i=1}^n f(x_i;\theta) dx_1 \cdots x_n.$$

Derivoimalla θ:n suhteen ja käyttäen ehtoa (iii) saadaan tästä

$$\begin{aligned} 1 + b'(\theta) &= \int_{S^n} t_n(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \prod_{i=1}^n f(x_i;\theta) dx_1 \cdots dx_n \\ &= \int_{S^n} t_n(x_1, \ldots, x_n) \left[\sum_{i=1}^n \frac{1}{f(x_i;\theta)} \frac{\partial f(x_i;\theta)}{\partial\theta} \right] \prod_{i=1}^n f(x_i;\theta) dx_1 \cdots x_n. \end{aligned} \tag{2.5}$$

Olkoon $S^c = \mathbb{R} \setminus S$ joukon S komplementti. Huomautuksen 2.4 mukaisesti määrittelemme satunnaismuuttujan Y_i, $i = 1, \ldots, n$, kaavalla

$$Y_i = \frac{\partial \log[f(X_i;\theta) + 1_{S^c}(X_i)]}{\partial\theta} = \frac{1}{f(X_i;\theta) + 1_{S^c}(X_i)} \cdot \frac{\partial f(X_i;\theta)}{\partial\theta}.$$

Silloin $Y_i = 0$, kun $X_i \notin S$ ja

$$\mathbb{E}_\theta Y_i = \int_S \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta} f(x;\theta) dx = \frac{d}{d\theta} \int_S f(x;\theta) dx = \frac{d1}{d\theta} = 0,$$

missä toinen yhtäsuuruus seuraa ehdosta (ii). Olkoon sitten $Z = \sum_{i=1}^n Y_i$, jolloin $\mathbb{E}_\theta Z = \sum_{i=1}^n \mathbb{E}_\theta Y_i = 0$ ja

$$\begin{aligned} \mathbb{E}_\theta[(\hat{\theta}_n - \theta)Z] &= \mathbb{E}_\theta(\hat{\theta}_n Z) \\ &= \int_{S^n} t_n(x_1, \ldots, x_n) \left[\sum_{i=1}^n \frac{1}{f(x_i;\theta)} \frac{\partial f(x_i;\theta)}{\partial \theta} \right] \prod_{i=1}^n f(x_i;\theta) dx_1 \cdots x_n \\ &= 1 + b'(\theta), \end{aligned}$$

missä ensimmäinen yhtäsuuruus seuraa siitä, että $\mathbb{E}_\theta Z = 0$ ja viimeinen yhtäsuuruus seuraa kaavasta (2.5). Schwarzin epäyhtälöstä saadaan nyt

$$1 + b'(\theta) = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)Z] \leq \sqrt{\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2} \sqrt{\mathbb{E}_\theta Z^2}. \tag{2.6}$$

Tässä

$$\mathbb{E}_\theta Z^2 = \mathrm{Var}_\theta(Z) = n\mathrm{Var}_\theta(Y_1) = n\mathbb{E}_\theta Y_1^2 = n\mathbb{E}_\theta \left[\left(\frac{\partial \log f(X_1;\theta)}{\partial \theta} \right)^2 \right],$$

missä ensimmäinen yhtäsuuruus seuraa siitä, että $\mathbb{E}_\theta Z = 0$, toinen siitä, että Y_i't ovat riippumattomia ja samoin jakautuneita ja kolmas siitä, että $\mathbb{E}_\theta Y_1 = 0$. Korottamalla nyt (2.6) puolittain toiseen ja jakamalla $\mathbb{E}_\theta Z^2$:lla saadaan väite. □

Huomautus 2.6 Yleensä $1 + b'(\theta) \neq 0$. Erityisesti näin on kun $\hat{\theta}_n$ on *harhaton*, eli $\mathbb{E}_\theta \hat{\theta}_n = \theta$, kun $\theta \in \Theta$, jolloin $b(\theta) = 0, \theta \in \Theta$. ‖

Lause 2.5 sanoo, että tehdyillä oletuksilla paras mahdollinen neliöllisen virheen suppenemisnopeus on $1/n$. Kuten seuraavasta kohdasta ilmenee, paljon käytetty suurimman uskottavuuden estimaattori itse asiassa *saavuttaa* tämän nopeuden.

2.3 Suurimman uskottavuuden estimointi

Tarkastellaan edelleen 1-ulotteisen parametriavaruuden $\Theta \subset \mathbb{R}$ tilannetta ja tiheysfunktioperhettä $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\}$. Olkoon $\theta_0 \in \Theta$ kiinteä ja $X_1, \ldots, X_n \sim f(\cdot;\theta_0)$ i.i.d. otos. Funktio $L(\cdot; X_1, \ldots, X_n) : \Theta \to [0, \infty[$,

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i; \theta), \qquad \theta \in \Theta,$$

on ns. *uskottavuusfunktio*. Joskus merkitään lyhyesti $L(\theta; X_1, \ldots, X_n) = L(\theta)$.

Suurimman uskottavuuden ("su") estimoinnissa otetaan θ_0:n estimaatiksi L:n (jokin) maksimoija $\hat{\theta}_n$ joukossa Θ. Ajatuksena on, että $\hat{\theta}_n$ "maksimoi otoksen $X_1, \ldots, X_n$ todennäköisyyden". Usein itseasissa maksimoidaan $\log L$ ratkaisemalla $\hat{\theta}_n$ *uskottavuusyhtälöstä*

$$\left.\frac{\partial \log L(\theta; X_1, \ldots, X_n)}{\partial \theta}\right|_{\theta=\hat{\theta}_n} = 0. \tag{2.7}$$

Esimerkki 2.7 Tarkastellaan normaalijakaumia $N(\theta, 1), \theta \in \mathbb{R}$, eli $\mathcal{F}$ koostuu tiheysfunktioista

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}, \qquad x \in \mathbb{R}.$$

Olkoon $\theta_0 \in \mathbb{R}$ ja $X_1, \ldots, X_n \sim f(\cdot; \theta_0)$ i.i.d. otos. Tällöin

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i-\theta)^2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2}$$

ja

$$\log L(\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2.$$

Saamme

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n}(X_i - \theta) = \sum_{i=1}^{n} X_i - n\theta,$$

joka häviää kun $\theta = \hat{\theta}_n = (1/n)\sum_{i=1}^{n} X_i$. Selvästi $\hat{\theta}_n$ on uskottavuusfunktion maksimoija. Siten odotusarvon θ_0 su-estimaattori on otoskeskiarvo (= otoksen keskiarvo). ‖

Jatketaan vielä edellisen esimerkin tarkastelua. Estimaattorilla $\hat{\theta}_n$ on seuraavat ominaisuudet.

(A) Vahvan suurten lukujen lain nojalla $\hat{\theta}_n \to \theta_0$ m.v. (= "melkein varmasti", todennäköisyydellä 1).

(B) $\hat{\theta}_n \sim N(\theta_0, 1/n)$, joten $\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N(0,1)$.

Tutkitaan Cramerin-Raon alarajan tiukkuutta estimaattorille $\hat{\theta}_n$. Ehdon (i) joukko $S = \{x \mid f(x;\theta) > 0\} = \mathbb{R}$ ei riipu θ:sta ja

$$\begin{aligned}
&\mathbb{E}_\theta\left[\left(\frac{\partial \log f(X_1;\theta)}{\partial\theta}\right)^2\right] \\
&\quad = \mathbb{E}_\theta\left\{\left[\frac{\partial}{\partial\theta}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}(X_1-\theta)^2\right)\right]^2\right\} \\
&\quad = \mathbb{E}_\theta[(X_1-\theta)^2] = \mathrm{Var}_\theta(X_1) = 1,
\end{aligned}$$

joten ehto (iv) on voimassa. Myös (ii) pätee ja $\hat{\theta}_n$ toteuttaa (iii):n. Lisäksi kohdan (B) nojalla $\hat{\theta}_n$ on harhaton, $\mathbb{E}_\theta\hat{\theta}_n = \theta$ (vrt. (B) yllä). Olkoon $\tilde{\theta}_n$ mikä tahansa toinen harhaton θ:n estimaattori, joka toteuttaa lauseen 2.5 ehdot. Silloin erityisesti (2.4):n osoittaja saa arvon 1, joten Cramerin-Raon epäyhtälö saa muodon

$$\mathbb{E}_\theta(\tilde{\theta}_n - \theta)^2 \geq \frac{1}{n}.$$

Mutta estimaattorille $\hat{\theta}_n$ itseasiassa pätee (vrt. (B)),

$$\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = \mathrm{Var}_\theta(\hat{\theta}_n) = \frac{1}{n},$$

joten suurimman uskottavuuden estimaattori $\hat{\theta}_n$ itseasiassa *saavuttaa Cramerin-Raon alarajan* ja on siis siinä mielessä optimaalinen harhaton estimaattori, että sillä on pienin varianssi.

Tekemällä tiheysfunktioperheestä $\mathcal{F}$ tietyt säännöllisyysoletukset saadaan suestimaattorin optimaalisuutta koskeva yleisempi tulos. Olkoon $\Theta =]a,b[$ jälleen väli ja oletetaan, että seuraavat ehdot ovat voimassa.

(i) Kaikilla $\theta \in \Theta$ on olemassa derivaatat $\partial^k \log f(x;\theta)/\partial\theta^k$, $k = 1,2,3,\ x \in \mathbb{R}$.

(ii) Kaikilla $\theta_0 \in \Theta$ on olemassa $\delta > 0$ ja funktiot $g_k : \mathbb{R} \to [0,\infty[,\ k = 1,2,3$, s.e. kaikilla $\theta \in]\theta_0 - \delta, \theta_0 + \delta[$ pätee

$$\left|\frac{\partial^k f(x;\theta)}{\partial\theta^k}\right| \leq g_k(x),\ k = 1,2,3,\ x \in \mathbb{R}$$

ja

$$\int_{-\infty}^{\infty} g_k(x)dx < \infty,\ k = 1, 2, \qquad \mathbb{E}_\theta[g_3(X)] < \infty.$$

(iii) Kaikilla $\theta \in \Theta$,

$$0 < \mathbb{E}_\theta \left[\left(\frac{\partial \log f(X_1; \theta)}{\partial \theta} \right)^2 \right] < \infty.$$

Lause 2.8 *Olkoot (i), (ii) ja (iii) voimassa. Olkoon $\theta_0 \in \Theta$ ja $X_1, X_2, \ldots \sim f(\cdot; \theta_0)$ jono riippumattomia ja samoin jakautuneita satunnaismuuttujia. Silloin, todennäköisyydellä 1, uskottavuusyhtälöllä (2.7) on otokseen $X_1, \ldots, X_n$ perustuva ratkaisu $\hat{\theta}_n$, $n \in \mathbb{N}$, s.e.*

(A) $\hat{\theta}_n \to \theta_0$ m.v. kun $n \to \infty$,

(B) $c\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, 1)$ (jakaumakonvergenssi) kun $n \to \infty$, missä

$$c = \sqrt{\mathbb{E}_{\theta_0} \left[\left(\frac{\partial \log f(X_1; \theta_0)}{\partial \theta} \right)^2 \right]}.$$

Todistus: Ks. Serfling: "Approximation theorems of mathematical statistics", Wiley 1980, ss. 145 - 148. □

Palautetaan mieleen, mitä kohdan (B) jakaumakonvergrenssilla tarkoitetaan. Olkoon Φ jakauman $N(0, 1)$ kertymäfunktio,

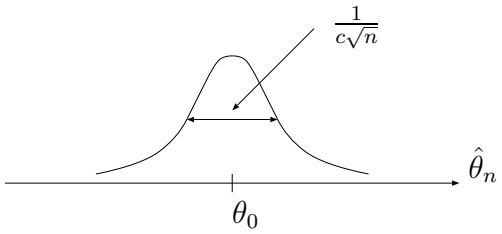
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt.$$

Jos silloin F_n on satunnaismuuttujan $c\sqrt{n}(\hat{\theta}_n - \theta_0)$ kertymäfunktio, tarkoittaa (B) sitä, että $\lim_{n\to\infty} F_n(x) = \Phi(x)$ kaikilla x. Siten suurella otoskoolla (eli suurella n) estimaattorin $\hat{\theta}_n$ jakauma on likimain $N(\theta_0, 1/(c^2 n))$ (ks. kuva 2.1).

Siten $\mathbb{E}_{\theta_0}(\hat{\theta}_n - \theta_0)^2 \approx 1/(c^2 n)$ eli Cramerin-Raon epäyhtälön alaraja harhattomalle estimaattorille likimain saavutetaan. Suurimman uskottavuuden estimaatori on tässä mielessä *asymptoottisesti optimaalinen.*

Huomautus 2.9 Usein pätee itseasiassa

$$\lim_{n\to\infty} \frac{\mathbb{E}_{\theta_0}(\hat{\theta}_n - \theta_0)^2}{1/(c^2 n)} = 1,$$



Kuva 2.1: Estimaattorin $\hat{\theta}_n$ jakauma kun otoskoko n on suuri.

jolloin Cramerin-Raon alaraja todella saavutetaan asymptoottisesti. Eräs riittävä ehto tälle on, että

$$\sup_n \mathbb{E}_{\theta_0} |\sqrt{n}(\hat{\theta}_n - \theta_0)|^{2+\varepsilon} < \infty$$

jollain $\varepsilon > 0$ (ks. Serfling, ss. 13-14). Vastaavat ehdot voidaan antaa korkeammillekin momenteille jolloin saadaan

$$\lim_{n\to\infty} \mathbb{E}_{\theta_0} \left| c\sqrt{n}(\hat{\theta}_n - \theta_0) \right|^k = m_k,$$

missä m_k on jakauman $N(0,1)$ k:s absoluuttinen momentti. ‖

2.4 Parametrinen funktion estimointi

Olkoot (a_n) ja (b_n) reaalilukujonoja. Käytämme jatkossa seuraavia merkintöjä:

$$\begin{aligned} a_n = o(b_n) &\quad\Leftrightarrow\quad \lim\nolimits_{n\to\infty}(a_n/b_n) = 0, \\ a_n = O(b_n) &\quad\Leftrightarrow\quad \limsup\nolimits_{n\to\infty} |a_n/b_n| < \infty, \\ a_n \sim b_n &\quad\Leftrightarrow\quad \lim\nolimits_{n\to\infty}(a_n/b_n) = 1. \end{aligned}$$

Joissain kaavoissa esiintyvä termi $o(1/n)$ esimerkiksi tarkoittaa jonoa (a_n), jolla $a_n = o(1/n)$ jne.

2.4.1 Tiheysfunktio

Olkoon f tuntematon tiheysfunktio, jota haluamme estimoida i.i.d. otoksen $X_1, \ldots, X_n \sim f$ avulla. Parametrinen estimointi tapahtuu seuraavasti.

(i) Valitaan parametrinen tiheysfunktioiden perhe $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\}$.

(ii) Oletetaan, että $f = f(\cdot;\theta_0)$ jollain $\theta_0 \in \Theta$.

(iii) Estimoidaan θ_0 jollain sopivalla estimaattorilla $\hat{\theta}_n = t_n(X_1, \ldots, X_n)$.

(iv) Otetaan f:n estimaattoriksi $\hat{f}_n = f(\cdot;\hat{\theta}_n)$.

Esimerkki 2.10 Esimerkissä 2.7 oli $X_1, \ldots, X_n \sim N(\theta_0, 1)$ ja su-estimaattoriksi saatiin otoskeskiarvo $\hat{\theta}_n = (1/n)\sum_{i=1}^n X_i$. Tällöin siis asetamme

$$\hat{f}_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\frac{1}{n}\sum_{i=1}^n X_i)^2}, \qquad x \in \mathbb{R}.$$

‖

Säännöllisissä tapauksissa on neliöllinen estimointivirhe yleensä luokkaa $1/n$ mikä nähdään seuraavasti. Taylorin kehitelmä antaa

$$f(x;\hat{\theta}_n) - f(x;\theta_0) = \frac{\partial f(x;\theta_0)}{\partial\theta}(\hat{\theta}_n - \theta_0) + \frac{1}{2}\frac{\partial^2 f(x;\tilde{\theta}_n)}{\partial\theta^2}(\hat{\theta}_n - \theta_0)^2,$$

missä (satunnaismuuttuja) $\tilde{\theta}_n$ on $\hat{\theta}_n$:n ja θ_0:n välissä. Saamme tästä

$$\begin{aligned}[f(x;\hat{\theta}_n) - f(x;\theta_0)]^2 &= \left(\frac{\partial f(x;\theta_0)}{\partial\theta}\right)^2 (\hat{\theta}_n - \theta_0)^2 \\ &+ \frac{\partial f(x;\theta_0)}{\partial\theta}\frac{\partial^2 f(x;\tilde{\theta}_n)}{\partial\theta^2}(\hat{\theta}_n - \theta_0)^3 + \frac{1}{4}\left(\frac{\partial^2 f(x;\tilde{\theta}_n)}{\partial\theta^2}\right)^2 (\hat{\theta}_n - \theta_0)^4.\end{aligned} \tag{2.8}$$

Nyt esimerkiksi su-estimaattorille $\hat{\theta}_n$ pätee, että $c\sqrt{n}(\hat{\theta}_n - \theta_0)$:n jakauma konvergoi otoskoon kasvaessa kohti jakaumaa $N(0,1)$ ja sopivilla oletuksilla (ks. Huomautus 2.9) saadaan

$$\lim_{n\to\infty} \mathbb{E}_{\theta_0} \left| c\sqrt{n}(\hat{\theta}_n - \theta_0) \right|^k = m_k > 0, \qquad k = 2, 3, 4,$$

jolloin

$$\mathbb{E}_{\theta_0} \left| \hat{\theta}_n - \theta_0 \right|^k \sim \frac{m_k}{c^k n^{k/2}}, \qquad k = 2, 3, 4.$$

Tästä saadaan edelleen pisteittäiselle keskimääräiselle neliövirheelle $\mathrm{MSE}[\hat{f}_n(x)]$ (MSE = Mean Squared Error) kaava

$$\begin{aligned}\mathrm{MSE}[\hat{f}_n(x)] &= \mathrm{MSE}[f(x;\hat{\theta}_n)] \equiv \mathbb{E}_{\theta_0}[f(x;\hat{\theta}_n) - f(x;\theta_0)]^2 \\ &= \left(\frac{\partial f(x;\theta_0)}{\partial\theta}\right)^2 \cdot \frac{C_1}{n} + O(\frac{1}{n^{3/2}}) \sim \left(\frac{\partial f(x;\theta_0)}{\partial\theta}\right)^2 \cdot \frac{C_1}{n},\end{aligned}$$

missä C_1 on vakio. Sopivilla säännöllisyysoletuksilla voidaan $\mathrm{MSE}[\hat{f}_n(x)]$ edelleen integroida, jolloin saadaan integroitu keskimääräinen neliövirhe $\mathrm{MISE}[\hat{f}_n]$,

$$\mathrm{MISE}[\hat{f}_n] = \int_{-\infty}^{\infty} \mathrm{MSE}[\hat{f}_n(x)]dx \sim \frac{C_2}{n},$$

missä

$$C_2 = C_1 \int_{-\infty}^{\infty} \left(\frac{\partial f(x;\theta_0)}{\partial \theta} \right)^2 dx > 0.$$

Siten integroitu neliövirhe on samaa luokkaa $1/n$ kuin yksittäisen parametrin neliöllinen estimointivirhe.

Esimerkki 2.11 Tarkastellaan jälleen normaalijakaumien tiheysfunktioiden perhettä

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}, \qquad x \in \mathbb{R},$$

ja olkoon $X_1, \ldots, X_n \sim f(\cdot;\theta_0)$, $\hat{\theta}_n = (1/n)\sum_{i=1}^n X_i$. Tällöin

$$\frac{\partial f(x;\theta_0)}{\partial \theta} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta_0)^2}(x-\theta_0) = (x-\theta_0)f(x;\theta_0),$$

ja

$$\frac{\partial^2 f(x;\theta_0)}{\partial \theta^2} = -f(x;\theta_0) + (x-\theta_0)^2 f(x;\theta_0) = [(x-\theta_0)^2 - 1]f(x;\theta_0).$$

Sijoittamalla kehitelmään (2.8) saadaan

$$\begin{aligned}
[f(x;\hat{\theta}_n) - f(x;\theta_0)]^2 &= (x-\theta_0)^2 f(x;\theta_0)^2 (\hat{\theta}_n - \theta_0)^2 \\
&\quad + (x-\theta_0) f(x;\theta_0)[(x-\tilde{\theta}_n)^2 - 1] f(x;\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^3 \\
&\quad + \frac{1}{4}[(x-\tilde{\theta}_n)^2 - 1]^2 f(x;\tilde{\theta}_n)^2 (\hat{\theta}_n - \theta_0)^4 \\
&\equiv (x-\theta_0)^2 f(x;\theta_0)^2 (\hat{\theta}_n - \theta_0)^2 + R_1(x;\theta_0,\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^3 + R_2(x;\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^4,
\end{aligned}$$

missä $(\hat{\theta}_n - \theta_0)^3$:n ja $(\hat{\theta}_n - \theta_0)^4$:n kertoimia on merkitty R_1:llä ja R_2:lla. Sijoittamalla $y = x - \theta_0$ saadaan

$$\int_{-\infty}^{\infty} (x-\theta_0)^2 f(x;\theta_0)^2 dx = \int_{-\infty}^{\infty} y^2 f(y;0)^2 dy = \frac{1}{2\pi} \int_{-\infty}^{\infty} y^2 e^{-y^2} dy = \frac{1}{4\sqrt{\pi}}.$$

Edelleen, Schwarzin epäyhtälöstä ja sijoituksilla $y = x - \theta_0$, $z = x - \tilde{\theta}_n$ saadaan

$$\begin{aligned}\left|\int_{-\infty}^{\infty} R_1(x;\theta_0,\tilde{\theta}_n)dx\right| &\leq \int_{-\infty}^{\infty} |R_1(x;\theta_0,\tilde{\theta}_n)|dx \\ &\leq \sqrt{\int_{-\infty}^{\infty}(x-\theta_0)^2 f(x;\theta_0)^2 dx}\sqrt{\int_{-\infty}^{\infty}[(x-\tilde{\theta}_n)^2-1]^2 f(x;\tilde{\theta}_n)^2 dx} \\ &= \sqrt{\int_{-\infty}^{\infty} y^2 f(y;0)^2 dy}\sqrt{\int_{-\infty}^{\infty}[z^2-1]^2 f(z;0)^2 dz} \equiv C_1 < \infty.\end{aligned}$$

Lisäksi sijoituksella $y = x - \tilde{\theta}_n$ saadaan

$$\int_{-\infty}^{\infty} R_2(x;\tilde{\theta}_n)dx = \frac{1}{4}\int_{-\infty}^{\infty}[y^2-1]^2 f(y;0)^2 dy \equiv C_2 < \infty. \tag{2.9}$$

Siten

$$\begin{aligned}\mathrm{MISE}[\hat{f}_n] &= \int_{-\infty}^{\infty}\mathbb{E}_{\theta_0}[f(x;\hat{\theta}_n)-f(x;\theta_0)]^2 dx \\ &= \mathbb{E}_{\theta_0}\int_{-\infty}^{\infty}[f(x;\hat{\theta}_n)-f(x;\theta_0)]^2 dx \\ &= \frac{1}{4\sqrt{\pi}}\mathbb{E}_{\theta_0}(\hat{\theta}_n-\theta_0)^2 + \mathbb{E}_{\theta_0}\left[\int_{-\infty}^{\infty} R_1(x;\theta_0,\tilde{\theta}_n)dx(\hat{\theta}_n-\theta_0)^3\right] \\ &\quad +\mathbb{E}_{\theta_0}\left[\int_{-\infty}^{\infty} R_2(x;\tilde{\theta}_n)dx(\hat{\theta}_n-\theta_0)^4\right],\end{aligned}$$

missä

$$\begin{aligned}&\left|\mathbb{E}_{\theta_0}\left[\int_{-\infty}^{\infty} R_1(x;\theta_0,\tilde{\theta}_n)dx(\hat{\theta}_n-\theta_0)^3\right]\right| \\ &\quad\leq \mathbb{E}_{\theta_0}\left[\int_{-\infty}^{\infty}|R_1(x;\theta_0,\tilde{\theta}_n)|dx|\hat{\theta}_n-\theta_0|^3\right] \leq C_1\mathbb{E}_{\theta_0}|\hat{\theta}_n-\theta_0|^3\end{aligned}$$

ja (2.9):n perusteella

$$\mathbb{E}_{\theta_0}\left[\int_{-\infty}^{\infty} R_2(x;\tilde{\theta}_n)dx(\hat{\theta}_n-\theta_0)^4\right] = C_2\mathbb{E}_{\theta_0}(\hat{\theta}_n-\theta_0)^4.$$

Nyt $\hat{\theta}_n = (1/n)\sum_{i=1}^n X_i \sim N(\theta_0, 1/n)$, joten

$$\begin{aligned}\mathbb{E}_{\theta_0}(\hat{\theta}_n-\theta_0)^2 &= \frac{1}{n}, \\ \mathbb{E}_{\theta_0}|\hat{\theta}_n-\theta_0|^3 &= \frac{4}{\sqrt{2\pi}n^{3/2}}, \\ \mathbb{E}_{\theta_0}(\hat{\theta}_n-\theta_0)^4 &= \frac{3}{n^2},\end{aligned}$$

missä kaksi jälkimmäistä yhtälöä todetaan helpoilla laskuilla. Siten saamme lopulta, että

$$\mathrm{MISE}[\hat{f}_n] \sim \frac{1}{4\sqrt{\pi}n}.$$

‖

2.4.2 Regressiofunktio

Olkoot X ja Y satunnaismuuttujia. Haluamme estimoida Y:n regressiofunktiota m X:n suhteen eli funktiota

$$m(x) = \mathbb{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

(vrt. HT). Voimme mallittaa X:n ja Y:n välistä riippuvuutta regressiofunktion avulla esimerkiksi olettamalla että

$$Y = m(X) + \varepsilon, \tag{2.10}$$

missä $\varepsilon \sim N(0, \sigma^2)$ on riippumaton X:stä.

Oletetaan sitten, että meillä on käytössä i.i.d. otos $(X_1, Y_1), \ldots, (X_n, Y_n)$ parin (X, Y) yhteisjakaumasta. Parametrinen regressiofunktion estimointi tapahtuu tällöin seuraavasti.

(i) Valitaan parametrinen funktioperhe $\mathcal{F} = \{m(\cdot;\theta) \mid \theta \in \Theta\}$, missä $\Theta \subset \mathbb{R}^d$.

(ii) Oletetaan, että $m = m(\cdot;\theta_0)$ eräällä $\theta_0 \in \Theta$.

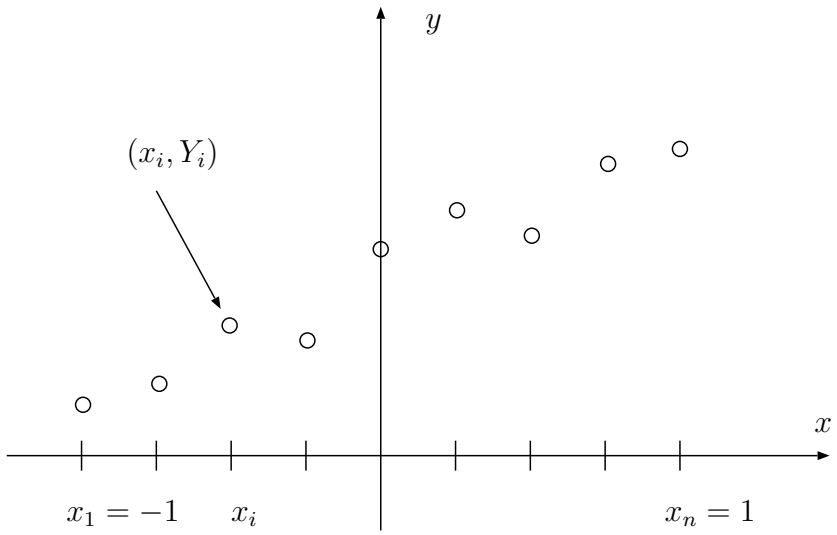
(iii) Estimoidaan θ_0 jollain sopivalla estimaattorilla $\hat{\theta}_n = t_n((X_1, Y_1), \ldots, (X_n, Y_n))$.

(iv) Otetaan m:n estimaattoriksi $\hat{m}_n = m(\cdot;\hat{\theta}_n)$.

Esimerkki 2.12 Eräitä mahdollisia funktioperheitä ovat esimerkiksi

$$\begin{aligned}
m(x;\theta) &= ax + b,\ x \in \mathbb{R},\ \text{missä } \theta = (a,b) \in \mathbb{R}^2,\\
m(x;\theta) &= ax^b,\ x > 0,\ \text{missä } \theta = (a,b) \in \mathbb{R}^2,\\
m(x;\theta) &= a\cos bx + c\sin dx,\ x \in \mathbb{R},\ \text{missä } \theta = (a,b,c,d) \in \mathbb{R}^4.
\end{aligned}$$

‖



Kuva 2.2: Lineaarinen regressiotehtävä välillä $[-1, 1]$.

Huomautus 2.13 Joskus X ei ole stokastinen, jolloin estimointi perustuu otokseen $(x_1, Y_1), \ldots, (x_n, Y_n)$, missä $x_1, \ldots, x_n \in \mathbb{R}$ ovat kiinteitä. Seuraava esimerkki on tällaisesta tilanteesta. ‖

Esimerkki 2.14 (Lineaarinen regressio) Tarkastellaan funktioperhettä

$$m(x; \theta) = ax + b, \quad x \in [-1, 1], \ \theta = (a, b) \in \mathbb{R}^2.$$

Olkoot $x_i = -1 + 2(i-1)/(n-1)$, $i = 1, \ldots, n$, välin $[-1, 1]$ tasavälinen jako ja $(x_1, Y_1), \ldots, (x_n, Y_n)$ i.i.d. otos. Oletamme, että regressiofunktio on suora $m(\cdot; \theta_0)$, $\theta_0 = (a_0, b_0)$, ja pyrimme estimoimaan parametrit a_0 ja b_0 mallissa

$$Y_i = a_0 x_i + b_0 + \varepsilon_i, \qquad i = 1, \ldots, n,$$

missä satunnaismuuttujat ε_i ovat riippumattomat, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, ja $\sigma > 0$ oletetaan tunnetuksi (ks. kuva 2.2).

Nyt $Y_1, \ldots, Y_n$ ovat riippumattomia ja

$$Y_i \sim N(a_0 x_i + b_0, \sigma^2), \qquad i = 1, \ldots, n.$$

Käytetään suurimman uskottavuuden estimointia ja maksimoidaan uskottavuus

$$\begin{aligned} L(a,b) &= L(a,b;Y_1,\ldots,Y_n) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{Y_i-ax_i-b}{\sigma})^2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-ax_i-b)^2}. \end{aligned}$$

Pitää siis *minimoida*

$$\lambda(a,b) = \sum_{i=1}^{n}(Y_i - ax_i - b)^2, \qquad (a,b) \in \mathbb{R}^2.$$

Toisin sanoen, on minimoitava "neliöllisten virheiden summa", eli ratkaistava pienimmän neliösumman tehtävä.

Nyt $\sum_{i=1}^{n} x_i = 0$, joten

$$\frac{\partial\lambda}{\partial a} = 2\sum_{i=1}^{n}(Y_i - ax_i - b)(-x_i) = 2\{-\sum_{i=1}^{n} x_iY_i + a\sum_{i=1}^{n} x_i^2\},$$

$$\frac{\partial\lambda}{\partial b} = 2\sum_{i=1}^{n}(Y_i - ax_i - b)(-1) = 2\{-\sum_{i=1}^{n} Y_i + nb\}.$$

Asettamalla $\partial\lambda/\partial a = 0$ ja $\partial\lambda/\partial b = 0$ saadaan ratkaisu $(\hat{a}_n, \hat{b}_n)$,

$$\hat{a}_n = \frac{\sum_{i=1}^{n} x_iY_i}{\sum_{i=1}^{n} x_i^2}, \qquad \hat{b}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

On helppo nähdä, että $\hat{\theta}_n = (\hat{a}_n, \hat{b}_n)$ on todella funktion λ minimikohta ja $\hat{\theta}_n$ on siten suurimman uskottavuuden estimaattori. Käyttäen hyväksi sitä, että $\mathbb{E}_{\theta_0} Y_i = a_0x_i + b_0$ ja $\sum_{i=1}^{n} x_i = 0$, nähdään helposti, että

$$\begin{cases} \mathbb{E}_{\theta_0}\hat{a}_n = a_0 \\ \mathbb{E}_{\theta_0}\hat{b}_n = b_0 \end{cases}, \qquad \begin{cases} \mathrm{Var}_{\theta_0}(\hat{a}_n) = \sigma^2/\sum_{i=1}^{n} x_i^2 \\ \mathrm{Var}_{\theta_0}(\hat{b}_n) = \sigma^2/n. \end{cases}$$

Itseasiassa

$$\hat{a}_n \sim N(a_0, \sigma^2/\sum_{i=1}^{n} x_i^2), \qquad \hat{b}_n \sim N(b_0, \sigma^2/n),$$

kuten helposti nähdään. Edelleen, pienellä laskulla havaitaan, että

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}\left(-1 + \frac{2(i-1)}{n-1}\right)^2 = \frac{4}{3}(n-1) + 2 + \frac{2}{3(n-1)},$$

joten

$$\frac{\sigma^2}{\sum_{i=1}^n x_i^2} = \frac{\sigma^2}{\frac{4}{3}(n-1)+2+\frac{2}{3(n-1)}} = \frac{3\sigma^2}{4n} + o(\frac{1}{n}).$$

Neliöllinen virhe pisteessä $x \in [-1,1]$ on nyt

$$\begin{aligned}
\mathrm{MSE}[\hat{m}_n(x)] &= \mathbb{E}_{\theta_0}[m(x;\hat{\theta}_n) - m(x;\theta_0)]^2 \\
&= \mathbb{E}_{\theta_0}[\hat{a}_n x + \hat{b}_n - (a_0 x + b_0)]^2 \\
&= \mathbb{E}_{\theta_0}[(\hat{a}_n - a_0)x + (\hat{b}_n - b_0)]^2 \\
&= x^2\mathbb{E}_{\theta_0}(\hat{a}_n - a_0)^2 + 2x\mathbb{E}_{\theta_0}[(\hat{a}_n - a_0)(\hat{b}_n - b_0)] + \mathbb{E}_{\theta_0}(\hat{b}_n - b_0)^2 \\
&= x^2\mathrm{Var}_{\theta_0}(\hat{a}_n) + 2x\mathrm{Cov}_{\theta_0}(\hat{a}_n, \hat{b}_n) + \mathrm{Var}_{\theta_0}(\hat{b}_n).
\end{aligned}$$

Helposti nähdään (HT), että $\mathrm{Cov}_{\theta_0}(\hat{a}_n, \hat{b}_n) = 0$. Siten

$$\mathrm{MSE}[\hat{m}_n] = x^2\left(\frac{3\sigma^2}{4n} + o(\frac{1}{n})\right) + \frac{\sigma^2}{n} \sim \frac{\sigma^2}{n}(1 + \frac{3}{4}x^2).$$

Estimointivirhe on siis tässä parametrisessa menetelmässä jälleen luokkaa $1/n$. ‖

2.5 Kohti parametritonta funktion estimointia

Tarkastellaan joukossa $D \subset \mathbb{R}$ määriteltyjä funktioita $f : D \to \mathbb{R}$. Tyypillisesti esimerkiksi $D = \mathbb{R}$ tai $D = [0,1]$. Olkoon

$$L^2(D) = \{f \mid \int_D [f(x)]^2 dx < \infty\},$$

missä $\int_D$ tarkoittaa Lebesguen integraalia. Olkoon $(\varphi_k)_{k\in\mathbb{N}}$ ortonormaali kanta funktioavaruudessa $L^2(D)$. Siis $\varphi_k \in L^2(D)$, $k \in \mathbb{N}$,

$$\int_D \varphi_k(x)\varphi_\ell(x)dx = \begin{cases} 1, & k = \ell, \\ 0, & k \neq \ell, \end{cases}$$

ja jokaisella $f \in L^2(D)$ on yksikäsitteinen esitys muodossa $f = \sum_{k=1}^\infty a_k\varphi_k$, $a_k \in \mathbb{R}$, $k \in \mathbb{N}$.

Esimerkki 2.15 Olkoon $D = [0,1]$. Tällöin voidaan osoittaa, että funktiot

$$\varphi_k(x) = \begin{cases} 1, & k = 1, \\ \sqrt{2}\cos(2\pi\ell x), & k = 2\ell, \\ \sqrt{2}\sin(2\pi\ell x), & k = 2\ell + 1 \end{cases}$$

$x \in \mathbb{R}$, $\ell \in \mathbb{N}$, muodostavat ortonormaalin kannan avaruudessa $L^2([0,1])$. ‖

Huomautus 2.16 Jos melkein kaikkialla (Lebesguen mitan mielessä) yhtyvät funktiot samaistetaan, on avaruus $L^2(D)$ itseasiassa ns. Hilbertin avaruus. Tämä tarkoittaa sitä, että $L^2(D)$ on vektoriavaruus, jossa on määritelty sisätulo ja tämän sisätulon määräämä metrinen avaruus on täydellinen. Sisätulona on $< f, g >= \int_D fg$ ja jos $(\varphi_k)_{k\in\mathbb{N}}$ on ortonormaali kanta ja $f \in L^2(D)$ on esitetty muodossa $f = \sum_{k=1}^{\infty} a_k\varphi_k$, keroimet a_k saadaan laskettua kaavasta

$$a_k =< \varphi_k, f >= \int_D \varphi_k(x)f(x)dx, \qquad k \in \mathbb{N}.$$

Edelleen, sarjan suppeneminen kehitelmässä $f = \sum_{k=1}^{\infty} a_k\varphi_k$ siis itse asiassa tarkoittaa, että

$$\lim_{n\to\infty} \int_D \left(\sum_{k=1}^{n} a_k\varphi_k(x) - f(x) \right)^2 dx = 0.$$

‖

Olkoon sitten $m \in \mathbb{N}$ kiinteä, $\Theta = \mathbb{R}^m$, ja $\theta = (a_1, \ldots, a_m) \in \Theta$. Merkitään

$$f(x;\theta) = \sum_{k=1}^{m} a_k\varphi_k(x), \qquad x \in D,$$

ja määritellään parametrinen funktioperhe $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\} \subset L^2(D)$. Olkoon $f \in L^2(D)$ jonkun satunnaismuuttujan tiheysfunktio ja oletetaan, että $f \in \mathcal{F}$, eli $f = f(\cdot;\theta_0)$ eräällä $\theta_0 = (a_{10}, \ldots, a_{m0}) \in \Theta$,

$$f = \sum_{k=1}^{m} a_{k0}\varphi_k.$$

Olkoon $X_1, \ldots, X_n \sim f$ i.i.d. otos. Koska

$$a_{k0} = \int_D \varphi_k(x)f(x)dx = \mathbb{E}_{\theta_0}\varphi_k(X_1),$$

saadaan a_{k0}:lle luonteva estimaattori

$$\hat{a}_{kn} = \frac{1}{n}\sum_{i=1}^{n} \varphi_k(X_i)$$

ja tästä f:lle estimaattori $\hat{f}_n = f(\cdot;\hat{\theta}_n)$, $\hat{\theta}_n = (\hat{a}_{1n}, \ldots, \hat{a}_{mn})$,

$$\hat{f}_n = \sum_{k=1}^{m} \hat{a}_{kn}\varphi_k. \tag{2.11}$$

Helposti nähdään, että $\hat{a}_{kn}$ on harhaton:

$$\mathbb{E}_{\theta_0}\hat{a}_{kn} = \frac{1}{n}\cdot n\mathbb{E}_{\theta_0}\varphi_k(X_1) = \int_D \varphi_k(x)f(x)dx = a_{k0}.$$

Siten myös $\hat{f}_n$ on harhaton,

$$\mathbb{E}_{\theta_0}\hat{f}_n = \sum_{k=1}^{m}(\mathbb{E}_{\theta_0}\hat{a}_{kn})\varphi_k = \sum_{k=1}^{m} a_{k0}\varphi_k = f.$$

Mitataan estimaattorin $\hat{f}_n$ virhettä integroidulla neliöllisellä virheellä,

$$\mathrm{MISE}[\hat{f}_n] = \mathbb{E}_{\theta_0}\int_D (\hat{f}_n - f)^2.$$

Saamme,

$$\begin{aligned}
\int_D (\hat{f}_n - f)^2 &= \int_D \left(\sum_{k=1}^{m}(\hat{a}_{kn} - a_{k0})\varphi_k\right)^2 \\
&= \int_D \sum_{k,\ell=1}^{m}(\hat{a}_{kn} - a_{k0})(\hat{a}_{\ell n} - a_{\ell 0})\varphi_k\varphi_\ell \\
&= \sum_{k,\ell=1}^{m}(\hat{a}_{kn} - a_{k0})(\hat{a}_{\ell n} - a_{\ell 0})\int_D \varphi_k\varphi_\ell \\
&= \sum_{k=1}^{m}(\hat{a}_{kn} - a_{k0})^2,
\end{aligned}$$

missä viimeisessä yhtäsuuruudessa käytettiin kannan $(\varphi_k)_{k\in\mathbb{N}}$ ortonormaalisuutta. Siten

$$\mathrm{MISE}[\hat{f}_n] = \mathbb{E}_{\theta_0}\sum_{k=1}^{m}(\hat{a}_{kn} - a_{k0})^2 = \sum_{k=1}^{m}\mathbb{E}_{\theta_0}(\hat{a}_{kn} - a_{k0})^2 = \sum_{k=1}^{m}\mathrm{Var}_{\theta_0}(\hat{a}_{kn}).$$

Edelleen,

$$\mathrm{Var}_{\theta_0}(\hat{a}_{kn}) = \mathrm{Var}_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_k(X_i)\right) = \frac{1}{n}\mathrm{Var}_{\theta_0}[\varphi_k(X_1)].$$

Tästä seuraa, että

$$\mathrm{MISE}[\hat{f}_n] = \frac{c_m}{n}, \qquad c_m = \sum_{k=1}^{m} \mathrm{Var}_{\theta_0}[\varphi_k(X_1)],$$

missä $c_m < \infty$, kunhan vaan kaikilla k pätee

$$\mathbb{E}_{\theta_0}[\varphi_k(X_1)^2] = \int_D [\varphi_k(x)]^2 f(x)dx < \infty.$$

Estimaattorin $\hat{f}_n$ virhe on siis tavanomaisen parametrisen estimaattorin luokkaa $1/n$.

Olkoon sitten

$$\Theta = \{(a_k)_{k\in\mathbb{N}} \mid \sum_{k=1}^{\infty} a_k^2 < \infty\}$$

ääretönulotteinen parametriavaruus. Voidaan osoittaa, että kaikilla $\theta = (a_k)_{k\in\mathbb{N}} \in \Theta$ sarja $f(\cdot;\theta) = \sum_{k=1}^{\infty} a_k\varphi_k$ suppenee avaruuden $L^2(D)$ mielessä (vrt. Huomautus 2.16). Siten $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\} \subset L^2(D)$. Toisaalta, kaikilla $f \in L^2(D)$ on olemassa $\theta = (a_k)_{k\in\mathbb{N}} \in \Theta$ s.e. $f = \sum_{k=1}^{\infty} a_k\varphi_k$. Siten ($L^2$-mielessä), $\mathcal{F} = L^2(D)$.

Olkoon nyt $f \in L^2(D)$ tiheysfunktio, $f = f(\cdot;\theta_0)$, $\theta_0 = (a_{k0})_{k\in\mathbb{N}} \in \Theta$. Funktion f estimointi otoksen $X_1, \ldots, X_n \sim f$ perusteella on nyt *parametritonta*, koska se ei perustu *äärellisulotteisen* parametrivektorin estimointiin. Oleellisesti ottaen nyt $\theta_0 = f$ ja avaruus Θ, josta θ_0:aa haetaan on ääretönulotteinen. Käytämme aikaisempaa estimaattoria (2.11) ja kehitelmän $f = \sum_{k=1}^{\infty} a_{k0}\varphi_k$ perusteella saamme

$$\begin{aligned}
\mathrm{MISE}[\hat{f}_n] &= \mathbb{E}_{\theta_0} \int_D (\hat{f}_n - f)^2 \\
&= \mathbb{E}_{\theta_0} \int_D \left[\sum_{k=1}^{m} (\hat{a}_{kn} - a_{k0})\varphi_k - \sum_{k=m+1}^{\infty} a_{k0}\varphi_k \right]^2 \\
&= \mathbb{E}_{\theta_0} \left[\sum_{k=1}^{m} (\hat{a}_{kn} - a_{k0})^2 + \sum_{k=m+1}^{\infty} a_{k0}^2 \right] \\
&= \frac{c_m}{n} + \sum_{k=m+1}^{\infty} a_{k0}^2,
\end{aligned}$$

missä $c_m = \sum_{k=1}^{m} \mathrm{Var}_{\theta_0}[\varphi_k(X_1)]$. Jos siis halutaan, että $\lim_{n\to\infty} \mathrm{MISE}[\hat{f}_n] = 0$, tulisi ilmeisesti

(i) $m \to \infty$, kun $n \to \infty$ (jotta $\sum_{k=m+1}^{\infty} a_k^2 \to 0$),

(ii) $c_m/n \to 0$.

Vaikka kohdan (i) mukaan m:n täytyy kasvaa rajatta otoskoon mukana, kohta (ii) kuitenkin edellyttää, että kasvu ei saa olla *liian* nopeaa. Ongelmana on näin ollen valita jono $(m_n)_{n\in\mathbb{N}}$ siten, että

$$m_n \to \infty, \qquad \frac{c_{m_n}}{n} \to 0, \qquad \text{kun } n \to \infty.$$

Lopullinen estimattori on sitten

$$\hat{f}_n = \sum_{k=1}^{m_n} \hat{a}_{kn}\varphi_k.$$

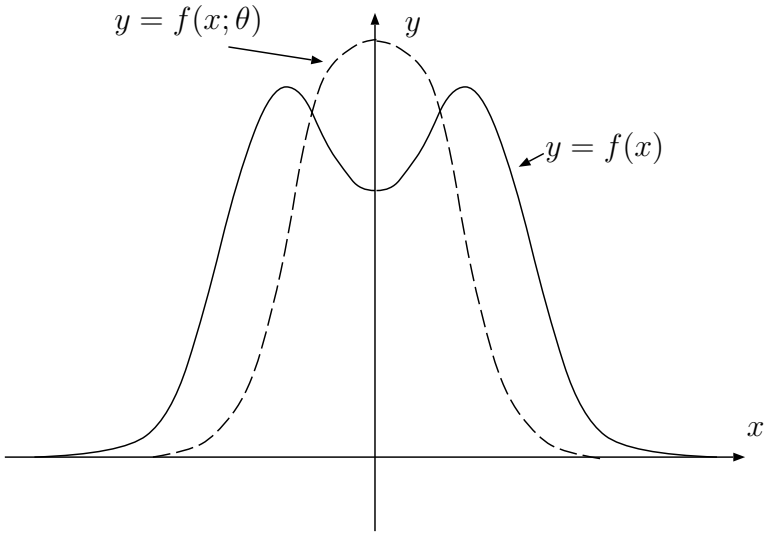
Osoittautuu, että menetellen yllä esitetyllä tavalla päästään tyypillisesti virheeseen

$$\mathrm{MISE}[\hat{f}_n] \sim \frac{c}{n^\delta},$$

missä $0 < \delta < 1$ ja $c > 0$ on jokin vakio. Virhe on siis konvergenssinopeudeltaan huonompi kuin parametrisen estimoinnin tapauksessa. On kuitenkin huomattava, että parametrisen menetelmän tehokkuus (vauhti $1/n$) perustuu siihen oletukseen, että $f \in \mathcal{F}$, eli että $f = f(\cdot;\theta_0)$ jollain $\theta_0 \in \Theta$. Vain pienen määrän eri funktiomuotoja sisältävän perheen $\mathcal{F}$ tapauksessa (matalaulotteinen Θ) tämä voi olla erittäin rajoittava oletus. Kahden tilastotieteen suuren nimen, R. Fisherin ja K. Pearsonin oppiriita 1900-luvun alkupuolella liittyi juuri tähän seikkaan (ks. esimerkiksi [10]). Fisher kiersi ongelman esittämällä, että

- sopivan parametrisen perheen $\mathcal{F}$ valitseminen on soveltajan asia ("specification"),
- parametrivektorin θ_0 estimoiminen on tilastotieteilijän asia ("estimation").

Resepti ei kuitenkaan välttämättä toimi käytännössä. Se toimii erityisen huonosti monissa nykyajan data-analyysitehtävissä, jossa tarkasteltavat jakaumat voivat olla rakenteeltaan hyvin monimutkaisia. Myös erittäin yksinkertaisia varoittavia esimerkkejä on helppo konstruoida.



Kuva 2.3: Estimoitava kaksihuippuinen tiheysfunktio (yhtenäinen viiva) ja käytettävissä olevan funktioperheen tyypillinen jäsen (katkoviiva).

Esimerkki 2.17 Olkoon $\Theta = \mathbb{R}\times]0,\infty[$, $\theta = (\mu, \sigma^2)$,

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad x \in \mathbb{R}.$$

Olkoon estimoitava tiheys $f = (1/2)f(\cdot;(-2,1)) + (1/2)f(\cdot;(2,1))$ (ks. kuva 2.3). Nyt selvästi $\mathbb{E}\int_{\mathbb{R}}[f(\cdot;\hat{\theta}_n) - f]^2 \not\to 0$ riippumatta estimaattorista $\hat{\theta}_n : \Omega \to \Theta$! ‖

Luku 3

Parametriton tiheysfunktion estimointi

3.1 Pakollinen harha

Olkoon $\mathcal{F} \subset \{f \mid f : \mathbb{R} \to [0, \infty[\text{ tiheysfunktio}\}$ jokin tiheysfunktioiden joukko. Oletetaan, että satunnaismuuttujan jakaumalla on tiheysfunktio $f \in \mathcal{F}$ ja että $X_1, \ldots, X_n \sim f$ on i.i.d. otos.

Saamme f:lle estimaattorin valitsemalla sopivan (Borelin) funktion $t_n : \mathbb{R}^{n+1} \to \mathbb{R}$ ja asettamalla $\hat{f}_n(x) = t_n(x, X_1, \ldots, X_n),\ x \in \mathbb{R}$. Analogisesti parametrin estimoinnin kanssa sanomme, että $\hat{f}_n$ on harhaton, jos

$$\mathbb{E}_f \hat{f}_n(x) = f(x), \quad x \in \mathbb{R}, \quad \text{kaikilla } f \in \mathcal{F}.$$

Tässä merkintä $\mathbb{E}_f$ tarkoittaa, että odotusarvo lasketaan satunnaismuuttujien $X_1, \ldots, X_n$ yhteisjakauman suhteen, kun $X_i \sim f$ kaikilla i. Harhattoman estimaattorin saattaa löytää, kun perhe $\mathcal{F}$ on riittävän "pieni".

Esimerkki 3.1 Jos $f(\cdot; (\mu, \sigma^2)) \sim N(\mu, \sigma^2)$ ja $\mathcal{F}$ on normaalijakauman tiheysfunktioiden perhe $\{f(\cdot; (\mu, \sigma^2)) \mid (\mu, \sigma^2) \in \mathbb{R} \times]0, \infty[\}$, niin harhaton estimaattori löytyy (ks. esim. [1], Exercise 7.14). ‖

Funktion estimaattorin harhattomuus on luonteva kysymys itse asiassa vain *jatkuville* funktiolle. Nimittäin, jos $f = g$ lukuun ottamatta yhtä pistettä (tai itseasiassa

mitä tahansa nollamittaista joukkoa), on mille tahansa estimaattorille $t_n(\cdot, X_1, \ldots, X_n)$ voimassa

$$\mathbb{E}_f t_n(x, X_1, \ldots, X_n) = \mathbb{E}_g t_n(x, X_1, \ldots, X_n)$$

kaikilla $x \in \mathbb{R}$. Jos siis $t_n(\cdot, X_1, \ldots, X_n)$ on harhaton, tulee olla $f(x) = g(x)$ kaikilla $x \in \mathbb{R}$, mikä on ristiriita. Jatkuvat funktiot f ja g eivät voi poiketa toisistaan vain yhdessä pisteessä (tai nollamittaisessa joukossa).

Nyt kuitenkin M. Rosenblatt osoitti v. 1956, että perheen

$$\mathcal{F} = \{f \mid f : \mathbb{R} \to [0, \infty[\text{ on jatkuva tiheysfunktio}\} \tag{3.1}$$

tapauksessa ei löydykään yhtään harhatonta estimaattoria! Todistamme tämän seuraavassa lauseessa. Rosenblattin tulos oli aikoinaan paha pettymys. Parametrisessa estimoinnissa oltiin totuttu hakemaan optimaalisia harhattomia estimaattoreita ja nyt ilmeni, että parametriton estimointi onkin harhaista. Toisaalta nykyään jopa parametrisessa estimoinnissa toisinaan käytetään harhaisia estimaattoreita (esim. ns. harjanneregressio) ja harhan määrää käytetään optimoitavana "säätöparametrina". Parametrittomassa estimoinnissa harha on hinta, joka maksetaan joustavuudesta, eli isosta $\mathcal{F}$:stä.

Lause 3.2 *Olkoon $\mathcal{F}$ kaikkien jatkuvien tiheysfunktioden perhe (3.1). Jos $x \in \mathbb{R}$ ja $n \in \mathbb{N}$, niin kaikilla (Borelin) funktioilla $t_n : \mathbb{R}^{n+1} \to \mathbb{R}$, on olemassa $f \in \mathcal{F}$ s.e., jos $X_1, \ldots, X_n \sim f$ on i.i.d. otos, niin $\mathbb{E}_f t_n(x, X_1, \ldots, X_n) \neq f(x)$.*

Todistus: Tehdään vastaoletus: on olemassa $x \in \mathbb{R}$, $n \in \mathbb{N}$ ja funktio t_n s.e.

$$\mathbb{E}_f t_n(x, X_1, \ldots, X_n) = f(x) \text{ kaikilla } f \in \mathcal{F}. \tag{3.2}$$

Oletetaan ensin, että $n \geq 2$. Valitaan kiinteä $f \in \mathcal{F}$. Silloin kaikilla $g \in \mathcal{F}$, $\lambda \in [0, 1]$, on

$$\int_{\mathbb{R}} [\lambda f + (1-\lambda)g] = \lambda \int_{\mathbb{R}} f + (1-\lambda) \int_{\mathbb{R}} g = \lambda + (1-\lambda) = 1,$$

joten $\lambda f + (1-\lambda)g \in \mathcal{F}$. Siten (3.2):n nojalla

$$\begin{aligned}
\lambda f(x) + (1-\lambda)g(x) &= \mathbb{E}_{\lambda f + (1-\lambda)g} t_n(x, X_1, \ldots, X_n) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t_n(x, x_1, \ldots, x_n) \prod_{i=1}^{n} [\lambda f(x_i) + (1-\lambda)g(x_i)] dx_1 \cdots dx_n \qquad (3.3) \\
&= \sum_{r=0}^{n} \lambda^r (1-\lambda)^{n-r} b_r(f, g),
\end{aligned}$$

missä

$$b_r(f,g) = \sum_{i_1<\cdots<i_r} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t_n(x, x_1, \ldots, x_n) \prod_{k=1}^{r} f(x_{i_k}) \prod_{\ell=1}^{n-r} g(x_{j_\ell}) dx_1 \cdots dx_n,$$

$1 \le i_1 < \cdots < i_r \le n$, $j_1 < \cdots < j_{n-r}$, ja

$$\{i_1, \ldots, i_r\} \cup \{j_1, \ldots, j_{n-r}\} = \{1, \ldots, n\}.$$

Merkitään

$$t_{n,\{i_1,\ldots,i_r\}}(x, x_{j_1}, \ldots, x_{j_{n-r}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t_n(x, x_1, \ldots, x_n) \prod_{k=1}^{r} f(x_{i_k}) dx_{i_1} \cdots dx_{i_r}.$$

(Moniulotteisia integraaleja koskevan Fubinin lauseen mukaan funktio $t_{n,\{i_1,\ldots,i_r\}}$ on määritelty ainakin m.k. (so. "melkein kaikkialla") ja se voidaan tarvittaessa laajentaa koko $\mathbb{R}^{n-r+1}$:n (Borelin) funktioksi.) Kun $r = 0$, ylläoleva kaava tulkitaan siten, että summa $\sum_{i_1<\cdots<i_r}$ ja tulo $\prod_{k=1}^{r} f(x_{i_k})$ puuttuvat. Vaihtamalla integroimisjärjestystä (sallittu samaisen Fubinin lauseen nojalla) saadaan

$$\begin{aligned} b_r(f,g) &= \sum_{i_1<\cdots<i_r} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t_{n,\{i_1,\ldots,i_r\}}(x, x_{j_1}, \ldots, x_{j_{n-r}}) \prod_{\ell=1}^{n-r} g(x_{j_\ell}) dx_{j_1} \cdots dx_{j_{n-r}} \\ &= \sum_{i_1<\cdots<i_r} \mathbb{E}_g[t_{n,\{i_1,\ldots,i_r\}}(x, X_1, \ldots, X_{n-r})]. \end{aligned}$$

Kaavan (3.3) oikea puoli on muotoa $\sum_{i=0}^{n} a_i \lambda^i$, missä λ^n:n kerroin on

$$a_n = \sum_{r=0}^{n} (-1)^{n-r} b_r(f,g).$$

Toisaalta (3.3):n vasen puoli on ensimmäistä astetta λ:n suhteen. Koska oletimme, että $n \ge 2$, tulee siis olla $a_n = 0$, mikä voidaan kirjoittaa muotoon

$$(-1)^n b_0(f,g) + \sum_{r=1}^{n} (-1)^{n-r} b_r(f,g) = 0$$

eli

$$b_0(f,g) = \sum_{r=1}^{n} (-1)^{-r+1} b_r(f,g).$$

Mutta käyttäen $b_r(f,g)$:n ja $t_{n,\{i_1,\dots,i_r\}}(x,x_{j_1},\dots,x_{j_{n-r}})$:n määritelmiä voidaan tämä yhtälö kirjoittaa myös muodossa

$$\begin{aligned}\mathbb{E}_g[t_n(x,X_1,\dots,X_n)] &= \sum_{r=1}^{n}(-1)^{-r+1}\sum_{i_1<\cdots<i_r}\mathbb{E}_g[t_{n,\{i_1,\dots,i_r\}}(x,X_1,\dots,X_{n-r})] \\ &= \mathbb{E}_g\left[\sum_{r=1}^{n}(-1)^{-r+1}\sum_{i_1<\cdots<i_r}t_{n,\{i_1,\dots,i_r\}}(x,X_1,\dots,X_{n-r})\right] \\ &\equiv \mathbb{E}_g[t_{n-1}(x,X_1,\dots,X_{n-1})],\end{aligned}$$

missä viimeisessä vaiheessa huomattiin, että satunnaismuuttuja, josta arvoarvo $\mathbb{E}_g$ lasketaan ei riipu X_n:stä ja on siksi eräs otokseen $X,\dots,X_{n-1}$ perustuva estimaattori. Vastaoletuksen nojalla yhtälön vasen puoli on $g(x)$, joten olemme löytäneet $n-1$ kokoiseen otokseen perustuvan estimaattorin, jolle pätee

$$\mathbb{E}_g[t_{n-1}(x,X_1,\dots,X_{n-1}) = g(x) \text{ kaikilla } g\in\mathcal{F}.$$

Suorittamalla induktion alaspäin löydämme lopulta funktion t_1 s.e.

$$\mathbb{E}_g t_1(x,X_1) = g(x) \text{ kaikilla } g\in\mathcal{F},$$

mikä voidaan kirjoittaa myös muodossa

$$\int_{-\infty}^{\infty} t_1(x,u)g(u)du = g(x) \text{ kaikilla } g\in\mathcal{F}. \tag{3.4}$$

Tarkastelemalla sopivia tiheysfunktioita g osoitamme, että tämä johtaa ristiriitaan.

Olkoon $g_k \sim N(x,1/k^2)$, $k\in\mathbb{N}$, eli

$$g_k(u) = \frac{k}{\sqrt{2\pi}}e^{-\frac{k^2}{2}(u-x)^2}, \quad u\in\mathbb{R}.$$

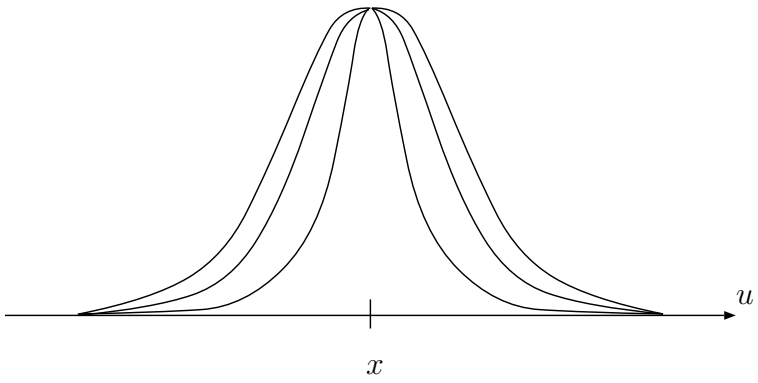
Selvästi $g_k\in\mathcal{F}$, joten yhtälön (3.4) nojalla

$$\int_{-\infty}^{\infty} t_1(x,u)\frac{k}{\sqrt{2\pi}}e^{-\frac{k^2}{2}(u-x)^2}du = \frac{k}{\sqrt{2\pi}}, \qquad \text{kaikilla } k\in\mathbb{N},$$

eli

$$\int_{-\infty}^{\infty} t_1(x,u)e^{-\frac{k^2}{2}(u-x)^2}du = 1, \qquad \text{kaikilla } k\in\mathbb{N}. \tag{3.5}$$

Olkoon h_k yhtälön (3.5) vasemman puolen integroitava funktio ,



Kuva 3.1: Funktion $u \mapsto e^{-\frac{k^2}{2}(u-x)^2}$ kuvaajia eri k:n arvoilla. Kun k on suuri, on funktio likimain 0 kun $u \neq x$.

$$h_k(u) = t_1(x,u)e^{-\frac{k^2}{2}(u-x)^2},\ u \in \mathbb{R},\ k \in \mathbb{N}.$$

Silloin pätee (vrt. kuva 3.1),

$$\lim_{k\to\infty} h_k(u) = \begin{cases} 0, & u \neq x \\ t_1(x,x), & u = x. \end{cases}$$

Lisäksi $|h_k(u)| \leq |h_1(u)|$ kaikilla $u \in \mathbb{R}$ ja $\int_{-\infty}^{\infty} |h_1(u)|du < \infty$ (koska odotusarvo $\mathbb{E}_{g_1}[t_1(x, X_1)]$ on olemassa). Lebesguen dominoidun konvergenssin lauseen nojalla yhtälöstä (3.5) seuraa nyt

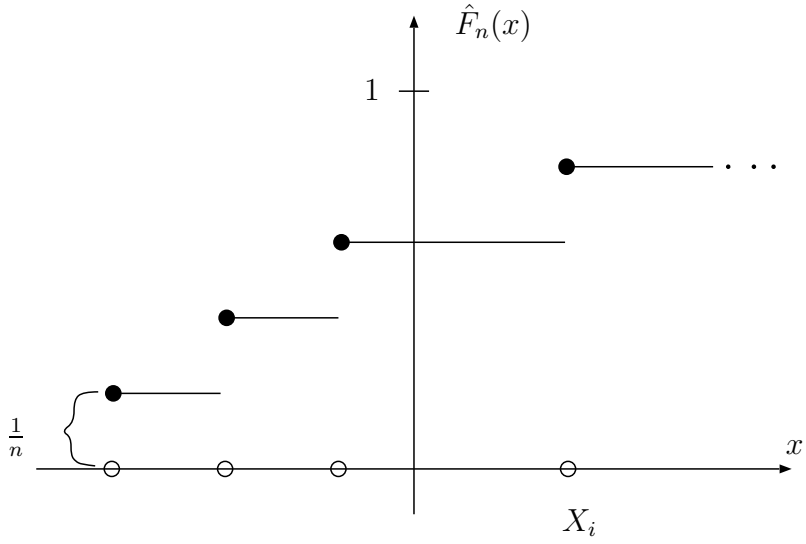
$$1 = \lim_{k\to\infty} \int_{-\infty}^{\infty} h_k(u)du = \int_{-\infty}^{\infty} \lim_{k\to\infty} h_k(u)du = 0,$$

joten vastaoletuksesta on päädytty ristiriitaan. □

3.2 Ydinestimointi

Esitämme aluksi heuristisen johdon ydinestimaattorille. Olkoon X satunnaismuuttuja, jonka jakaumalla on tiheysfunktio f. Olkoon F vastaava kertymäfunktio,

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$



Kuva 3.2: Empiirinen kertymäfunktio

Siis $F'(x) = f(x)$ (kaikilla x, jos f on jatkuva). Jos $X_1, \ldots, X_n \sim f$ on i.i.d. otos, saadaan F:lle luonteva estimaattori määrittelemällä

$$\hat{F}_n(x) = \frac{1}{n}\#\{i \mid X_i \leq x,\ i = 1, \ldots, n\} = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,x]}(X_i),$$

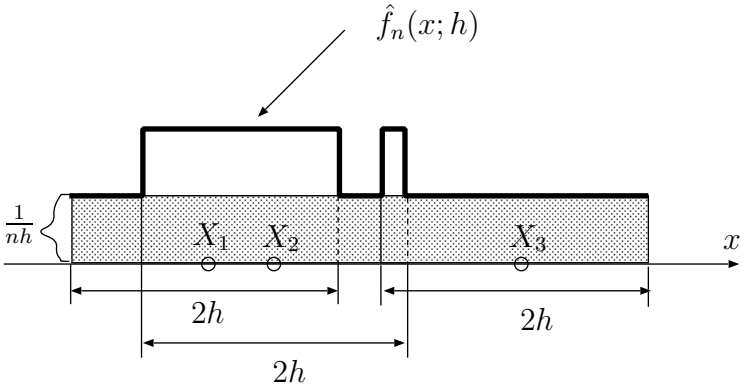
missä käytämme äärellisen joukon S alkioiden lukumäärälle merkintää $\#S$. Estimaattori $\hat{F}_n$ on ns. *empiirinen kertymäfunktio* (kuva 3.2).

Olkoon $h > 0$ pieni positiivinen luku. Silloin

$$\begin{aligned}
f(x) = F'(x) &\approx \frac{1}{2}\left[\frac{F(x+h) - F(x)}{h} + \frac{F(x) - F(x-h)}{h}\right] \\
&= \frac{1}{2h}[F(x+h) - F(x-h)] \\
&\approx \frac{1}{2h}[\hat{F}_n(x+h) - \hat{F}_n(x-h)] \\
&= \frac{1}{2h} \cdot \frac{1}{n}\left[\#\{i \mid X_i \leq x+h\} - \#\{i \mid X_i \leq x-h\}\right] \\
&= \frac{1}{2hn}\#\{i \mid x-h < X_i \leq x+h\} \equiv \hat{f}_n(x;h).
\end{aligned}$$

Näin määritelty estimaattori $\hat{f}_n(\cdot;h)$ on tiheysfunktion f *naiivi estimaattori*.

Määritellään sitten $K = (1/2)1_{[-1,1[}$, eli



Kuva 3.3: Tiheysfunktion naiivi estimaattori.

$$K(x) = \begin{cases} 1/2, & -1 \leq x < 1, \\ 0, & \text{muulloin.} \end{cases}$$

Silloin

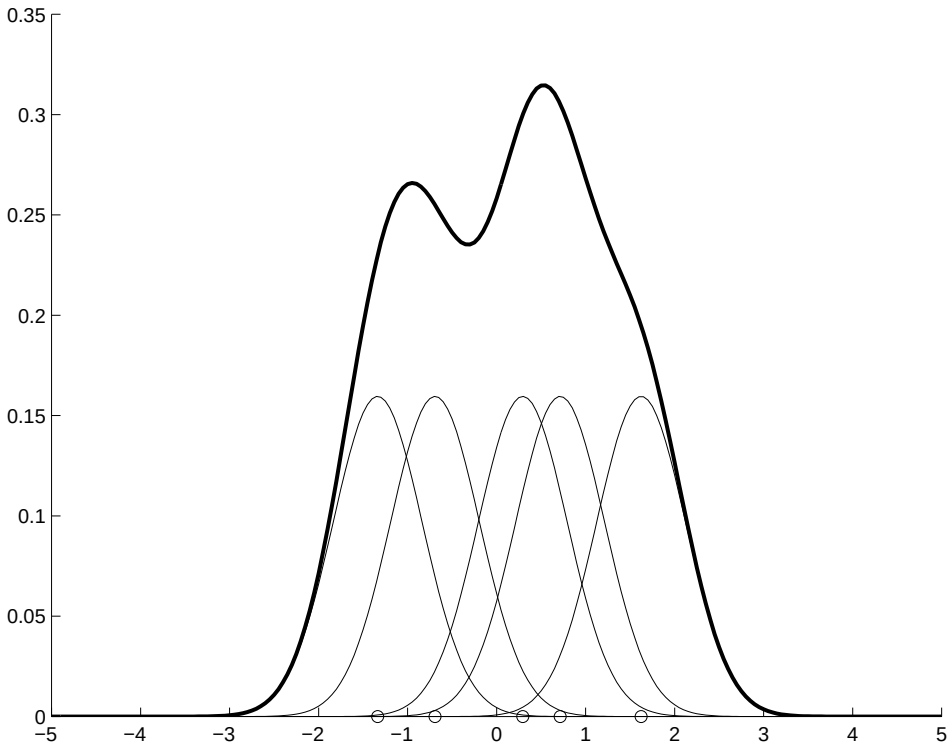
$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{3.6}$$

ja

$$\begin{aligned} \hat{f}_n(x;h) &= \frac{1}{2hn} \# \left\{ i \mid -1 < \frac{X_i - x}{h} \leq 1 \right\} \\ &= \frac{1}{2hn} \# \left\{ i \mid -1 \leq \frac{x - X_i}{h} < 1 \right\} \\ &= \frac{1}{2hn} \sum_{i=1}^{n} 1_{[-1,1[} \left(\frac{x - X_i}{h} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left(\frac{x - X_i}{h} \right). \end{aligned} \tag{3.7}$$

Kaavaa (3.7) voi lukea siten, että $\hat{f}_n(x;h)$ saadaan asettamalla jokaiseen otospisteeseen X_i positiivisella luvulla h ja otoskoolla n skaalattu K, eli $(1/(nh))K((x-X_i)/h)$, ja summaamalla sitten yli ostospisteiden (kuva 3.3). Estimaattorin arvo $\hat{f}_n(x;h)$ on sitä suurempi mitä tiheämmässä ostospisteitä X_i on x:n ympäristössä.

Ydinestimaattori saadaan nyt yksinkertaisesti korvaamalla $K = (1/2)1_{[-1,1[}$ kaavassa (3.7) yleisemmällä funktiolla K, joka kuitenkin toteuttaa ehdon (3.6).



Kuva 3.4: Tiheysfunktion ydinestimointi Gaussin ydintä käyttäen. Otospisteiden (5 kappaletta) paikkoja vaaka-akselilla on merkitty pienillä ympyröillä. Estimaatti saadaan ytimien summana ja on piirretty paksummalla viivalla.

Määritelmä 3.3 *Olkoon* $f : \mathbb{R} \to [0, \infty[$ *tiheysfunktio ja* $X_1, \ldots, X_n \sim f$ *i.i.d. otos. Olkoon* $K : \mathbb{R} \to \mathbb{R}$ *funktio, jolle* $\int_{-\infty}^{\infty} K(x)dx = 1$ *ja* $h > 0$. *Silloin funktion* f ydinestimaattori *on*

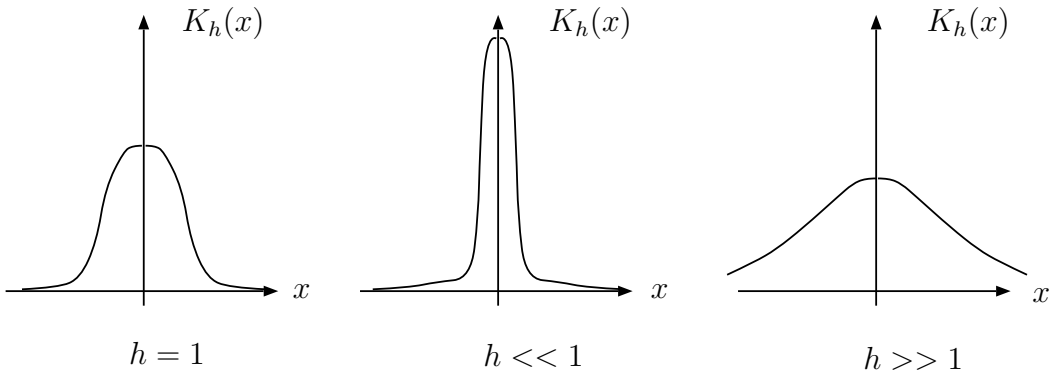
$$\hat{f}_n(x; h) = \hat{f}_n(x, X_1, \ldots, X_n; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Sanomme, että K on *ydin* ja h *silotusparametri*. Kuvassa 3.4 on esimerkki ydinestimaatista, jossa ytimenä on standardi normaalijakauman $N(0, 1)$ tiheysfunktio, jota tässä yhteydessä tavallisesti kutsutaan Gaussin ytimeksi.

Merkitään

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \quad x \in \mathbb{R}.$$

Silloin



Kuva 3.5: Ytimen skaalaaminen silotusparametrilla h.

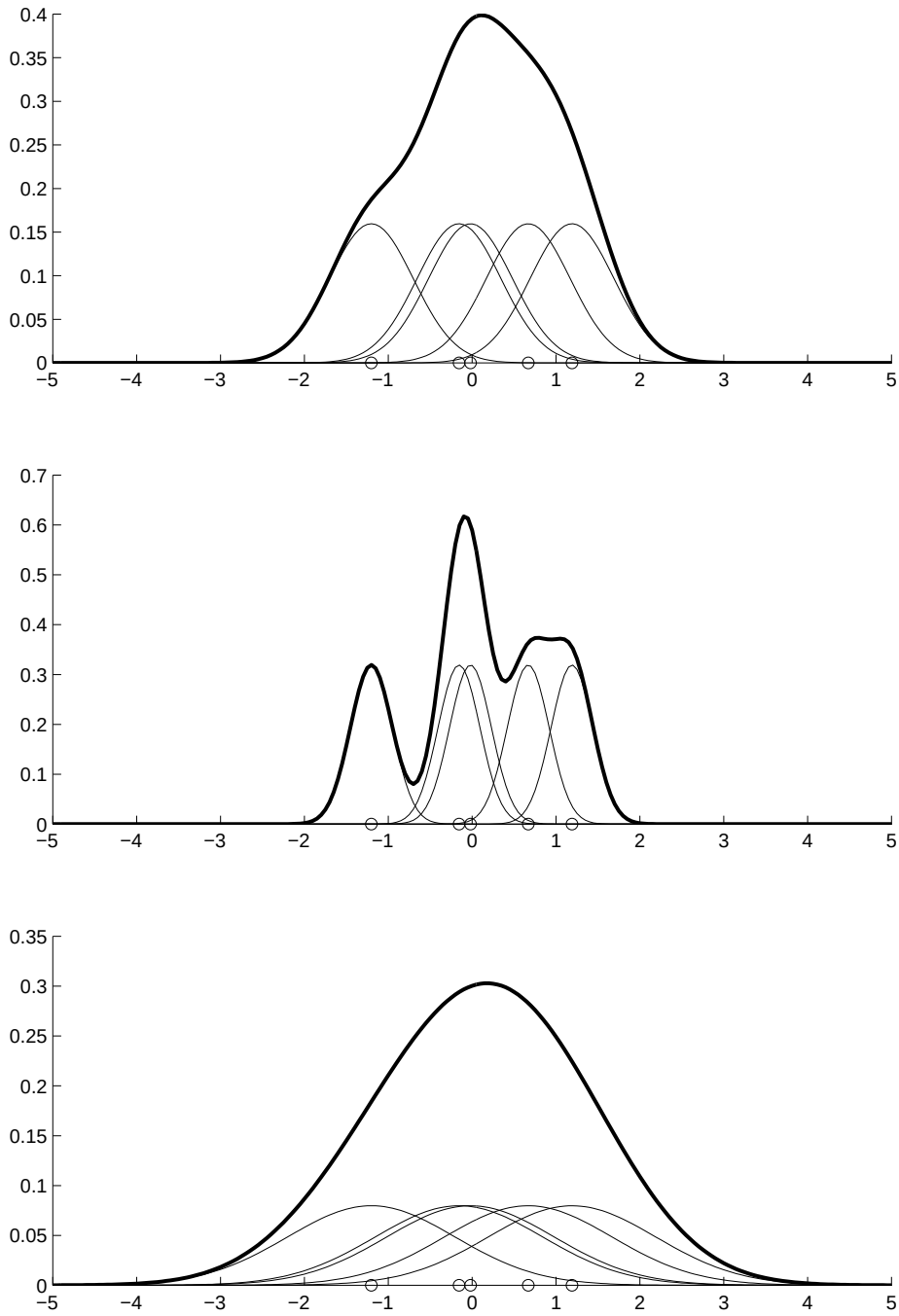
$$\hat{f}_n(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i).$$

Suorittamalla integroinnissa muuttujan vaihto $y = (x - X_i)/h$, $dx = hdy$, havaitaan, että

$$\int_{-\infty}^{\infty} K_h(x - X_i)dx = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) dx = \int_{-\infty}^{\infty} K(y)dy = 1.$$

Tästä seuraa, että $\int_{-\infty}^{\infty} \hat{f}_n(x;h)dx = 1$. Jos lisäksi $K(x) \geq 0$ kaikilla $x \in \mathbb{R}$, havaitaan, että $\hat{f}_n(\cdot;h)$ on itseasiassa tiheysfunktio, kun otos $X_1, \ldots, X_n$ on kiinnitetty.

Silotusparametrin h suuruuden vaikutus skaalattuun ytimeen K_h on esitetty kuvassa 3.5. Kuvassa 3.6 puolestaan näytetään silotusparametrin suuruuden vaikutus koko estimaatin $\hat{f}_n(\cdot;h)$ muotoon. Nähdään, että h:n pienentäminen tuo esiin yksittäisten otospisteiden vaikutuksen, jolloin estimaatista tulee hyvin rosoinen ja se sisältää useita paikallisia ääriarvoja, joita todellisessa estimoitavassa tiheydessä ei välttämättä lainkaan ole. Toisaalta, kun h valitaan suureksi, tulee estimaatista hyvin sileä ja siitä häviävät kaikki mahdollisesti kiinnostavatkin yksityiskohdat. Perusongelma ydinestimaatin käytössä onkin "oikean" suuruisen silotusparametrin valinta.



Kuva 3.6: Silotusparametrin h suuruuden vaikutus tiheysfunktion ydinestimaattiin (paksu viiva). Liian pieni h (keskimmäinen kuva) johtaa rosoiseen estimaattiin ja liian suuri h (alin kuva) puolestaan liian sileään estimaattiin. Otospisteet (jakaumasta $N(0,1)$) ovat kaikissa kuvissa samat ja ne on merkitty pienillä ympyröillä vaaka-akselille.

3.3 Virhekriteerejä

Olkoon $\hat{f}_n$ tiheysfunktion f jokin estimaattori (ei siis välttämättä ydinestimaattori). Estimaattorin tekemää virhettä voidaan mitata joko yksittäisessä pisteessä tai globaalisti koko $\mathbb{R}$:ssä.

3.3.1 Pisteittäinen virhe

Olkoon $x \in \mathbb{R}$ kiinteä. Tällöin voidaan määritellä esimerkiksi seuraavat luonetevat virhekriteerit:

absoluuttinen virhe	$\lvert\hat{f}_n(x) - f(x)\rvert$
keskimääräinen absoluuttinen virhe	$\mathbb{E}\lvert\hat{f}_n(x) - f(x)\rvert$.

Absoluuttinen virhe on satunnaismuuttuja kun taas keskimääräinen absoluuttinen virhe vain ei-negatiivinen luku. Odotusarvo lasketaan tietysti otoksen $X_1, \ldots, X_n \sim f$ jakauman suhteen ja merkitsemme jatkossakin yksinkertaisuuden vuoksi odotusarvoa usein vain symbolilla $\mathbb{E}$ tarkemman merkinnän $\mathbb{E}_f$ sijaan ja samoin varianssille Var tarkemman Var_f sijaan.

Keksimääräinen absoluuttinen virhe voidaan yleistää valitsemalla $p > 0$ ja ottamalla kriteeriksi

$$\mathbb{E}|\hat{f}_n(x) - f(x)|^p. \tag{3.8}$$

Erityisen suosittu erikoistapaus on $p = 2$, jolla saadaan keskimääräinen neliöllinen virhe

$$\mathrm{MSE}[\hat{f}_n(x)] = \mathbb{E}[\hat{f}_n(x) - f(x)]^2.$$

Tähän liittyen määritellään edelleen pisteessä x laskettu harha ja varianssi

$$\mathrm{Bias}[\hat{f}_n(x)] = \mathbb{E}\hat{f}_n(x) - f(x),$$

$$\mathrm{Var}[\hat{f}_n(x)] = \mathbb{E}[\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)]^2.$$

Silloin (vrt. harjoitustehtävä 1.2)

$$\mathrm{MSE}[\hat{f}_n(x)] = \mathrm{Bias}^2[\hat{f}_n(x)] + \mathrm{Var}[\hat{f}_n(x)].$$

3.3.2 Globaali virhe

Yksi mahdollinen globaali virhekriteeri on tietenkin $\sup_x |\hat{f}_n(x) - f(x)|$. Todellisuudessa suositumpi ja helpommin analysoitava kriteeri saadaan integroimalla. Olkoon $1 \leq p < \infty$ ja määritellään (mitallisten) funktioiden $g : \mathbb{R} \to \mathbb{R}$ avaruus

$$L^p = L^p(\mathbb{R}) = \{g \mid \int_{-\infty}^{\infty} |g(x)|^p dx < \infty\}.$$

Huomautus 3.4 Voidaan osoittaa, että jos avaruudessa L^p samaistetaan melkein kaikkialla yhtyvät funktiot g, niin kaava $\|g\|_p = (\int_{-\infty}^{\infty} |g(x)|^p dx)^{1/p}$ määrittelee normin ja syntyvä normiavaruus $(L^p, ||\cdot||)$ on täydellinen, eli ns. Banachin avaruus. ‖

Luonnollinen virhekriteeri saadaan nyt integroimalla (3.8),

$$\int_{-\infty}^{\infty} \mathbb{E}|\hat{f}_n(x) - f(x)|^p dx = \mathbb{E} \int_{-\infty}^{\infty} |\hat{f}_n(x) - f(x)|^p = \mathbb{E}\|\hat{f}_n - f\|_p^p.$$

Ensimmäisessä yhtälössä odotusarvon ja integroinnin järjestyksen saa vaihtaa Fubinin lauseen nojalla (odotusarvo voidaan kirjoittaa integraalina käyttäen otoksen $X_1, \ldots, X_n$ tiheysfunktiota). On huomattava, että jos esimerkiksi $\hat{f}_n$ on ydinestimaattori ja $f, K \in L^p$, niin silloin aina $\mathbb{E}\|\hat{f}_n - f\|_p^p < \infty$.

Tavallisimmin valitaan $p = 1$ tai $p = 2$. Tapauksessa $p = 2$ saadaan kriteeriksi keskimääräinen integroitu neliöllinen virhe,

$$\mathrm{MISE}[\hat{f}_n] = \mathbb{E} \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx = \int_{-\infty}^{\infty} \mathrm{Bias}^2[\hat{f}_n(x)] dx + \int_{-\infty}^{\infty} \mathrm{Var}[\hat{f}_n(x)] dx. \tag{3.9}$$

Tapauksen $p = 1$ käsittely on matemaattisesti melko hankalaa eikä siksi ole ollut kovin suosittua alan kirjallisuudessa (ks. kuitenkin [1] ja [2]).

Myös muita funktioiden välisen etäisyyden mittoja voitaisiin käyttää. Yksi mahdollisuus on *Kullbackin-Leiblerin* luku tai etäisyys (silloin kun se on määritelty),

$$K(f, \hat{f}_n) = \int_{-\infty}^{\infty} f(x) \log \left[\frac{f(x)}{\hat{f}_n(x)} \right] dx.$$

Voidaan osoittaa, että $K(f, \hat{f}_n) \geq 0$ aina. Tämä virhekriteeri itse asiassa liittyy läheisesti suurimman uskottavuuden estimointiin mikä nähdään seuraavasti. Kullbackin-Leiblerin etäisyys $K(f, g)$ on pieni kun g on sellainen, että $-K(f, g)$ on suuri, eli

$$\int_{-\infty}^{\infty} f(x) \log \frac{g(x)}{f(x)} dx = \int_{-\infty}^{\infty} f(x) \log g(x) dx - \int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

on suuri. Siten itseasiassa integraalin $\int_{-\infty}^{\infty} f(x) \log g(x) dx$, eli odotusarvon $\mathbb{E}_f \log g(X_1)$, missä $X_1, \ldots, X_n \sim f$ on i.i.d. otos, tulee olla suuri. Korvaamalla odotusarvo otoskeskiarvolla nähdään, että g tekee suureksi summan $(1/n) \sum_{i=1}^{n} \log g(X_i)$ ja siis myös tulon $\prod_{i=1}^{n} g(X_i)$. Siten se, että $K(f, \hat{f}_n)$ on pieni tarkoittaa sitä, että $\hat{f}_n$ on lähellä f:ää suurimman uskottavuuden estimoinnin mielessä.

Vielä eräs mahdollisuus on käyttää Hellingerin etäisyyttä

$$H_p(\hat{f}_n, f) = \left[\int_{-\infty}^{\infty} [\hat{f}_n(x)^{1/p} - f(x)^{1/p}]^p \right]^{1/p},$$

missä $p > 0$ (ks. [1]).

3.4 L^2-virhe

Käymme aluksi läpi joitain L^p-avaruuksiin liittyviä tuloksia. Oletamme, että $p = 1$ tai $p = 2$. Lemmojen 3.5, 3.6 ja 3.9 tulokset kuitenkin pätevät kaikille $1 \leq p < \infty$.

Lemma 3.5 *Olkoot $f, g \in L^p$. Silloin*

(i) $\|f + g\|_p \leq \|f\|_p + \|g\|_p$ (Minkowskin epäyhtälö).

(ii) $\|\lambda f\|_p = |\lambda| \|f\|_p$ kaikilla $\lambda \in \mathbb{R}$.

Todistus: HT. □

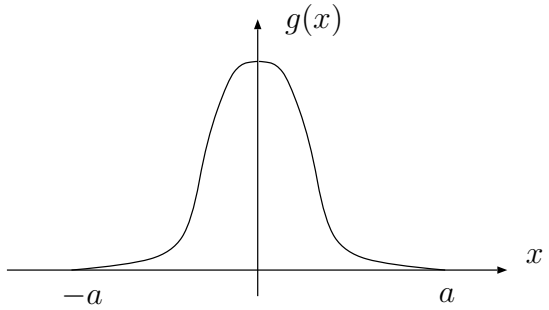
Lemma 3.6 *Olkoon $f : \mathbb{R} \times \mathbb{R} \to [0, \infty[$ mitallinen. Silloin*

$$\| \int_{-\infty}^{\infty} f(\cdot, y) dy \|_p \leq \int_{-\infty}^{\infty} \|f(\cdot, y)\|_p dy, \quad \text{(yleistetty Minkowskin epäyhtälö)}$$

eli toisin sanoen

$$\left[\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y) dy \right)^p dx \right]^{1/p} \leq \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} (f(x, y))^p dx \right]^{1/p} dy.$$

Todistus: HT. □



Kuva 3.7: Konvoluutiossa käytettävä silottava funktio.

Määritelmä 3.7 *Olkoon $f \in L^p$ ja $g \in L^1$. Silloin f:n ja g:n* konvoluutio *on*

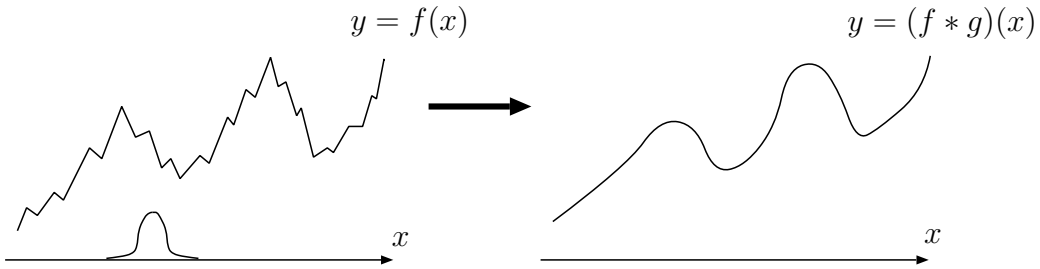
$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy, \quad x \in \mathbb{R}. \tag{3.10}$$

Huomautus 3.8 Konvoluutio $(f * g)(x)$ on määritelty ainakin m.k. (melkein kaikkialla). Tämä seuraa siitä, että

$$\begin{aligned}
&\left[\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |f(x-y)g(y)|dy\right)^p dx\right]^{1/p} \\
&\quad = \| \int_{-\infty}^{\infty} |f(\cdot - y)g(y)|dy\|_p \\
&\quad \leq \int_{-\infty}^{\infty} \|f(\cdot - y)g(y)\|_p dy \\
&\quad = \int_{-\infty}^{\infty} \|f(\cdot - y)\|_p |g(y)|dy \\
&\quad = \|f\|_p \|g\|_1 < \infty,
\end{aligned}$$

missä ensimmäinen epäyhtälö seuraa lemmasta 3.6 ja sen jälkeen hyödynnettiin lemman 3.5 kohtaa (ii) ja viimeisessä vaiheessa käytettiin yhtälöä $\|f(\cdot - y)\|_p = \|f\|_p$. Siten integraalin $\int_{-\infty}^{\infty} |f(x-y)g(y)|dy$ täytyy olla äärellinen melkein kaikilla $x \in \mathbb{R}$, joten $(f * g)(x)$ on määritelty m.k. x. Voidaan osoittaa, että (3.10) voidaan laajentaa mitalliseksi funktioksi $f * g : \mathbb{R} \to \mathbb{R}$. Edellinen lasku puolestaan osoittaa, että $f * g \in L^p$. Huomaa myös, että konvoluutiolle pätee $f * g = g * f$, mikä nähdään suorittamalla muuttujan vaihdos $z = x - y$.

Konvoluution avulla on mahdollista silottaa funktioita. Jos nimittäin g on sileä ei-negatiivinen funktio, jolle $\int_{-\infty}^{\infty} g(x)dx = 1$ ja joka häviää pienen välin $[-a, a]$ ulkopuolella (kuva 3.7), voidaan $(f * g)(x)$ tulkita funktion f painofunktiolla g lasketuksi keskiarvoksi pisteen x ympäristössä,



Kuva 3.8: Rosoinen funktio voidaan silottaa konvoloimalla se kuvan 3.7 kaltaisen funktion kanssa.

$$(f * g)(x) = \int_{-a}^{a} f(x-y)g(y)dy.$$

Siten konvoluutiolla voidaan haluttaessa silottaa mahdolliseti alunperin hyvinkin "rosoista" funktiota (kuva 3.8).

Lemma 3.9 *Olkoon $f \in L^p$ jatkuva, $g \in L^1$, $\int_{-\infty}^{\infty} g(x)dx = 1$, ja $g_h(x) = (1/h)g(x/h)$, $x \in \mathbb{R}$, $h > 0$. Silloin*

(i) $f * g_h \to f$, kun $h \to 0+$. Tässä konvergenssi on L^p-normin mielessä, eli $\|f * g_h - f\|_p \to 0$, kun $h \to 0+$.

(ii) $\int_{-\infty}^{\infty}[(f * g_h)(x)]^p dx \to \int_{-\infty}^{\infty}[f(x)]^p dx$, kun $h \to 0+$.

(iii) Jos f on rajoitettu, niin $(f * g_h)(x) \to f(x)$ kaikilla $x \in \mathbb{R}$, kun $h \to 0+$.

Todistus: Todistetaan ensin (i). Saamme

$$\begin{aligned}
\|f * g_h - f\|_p &= \|\int_{-\infty}^{\infty} f(\cdot - y)g_h(y)dy - f(\cdot)\|_p \\
&= \|\int_{-\infty}^{\infty}[f(\cdot - y) - f(\cdot)]g_h(y)dy\|_p \\
&\leq \int_{-\infty}^{\infty}\|[f(\cdot - y) - f(\cdot)]g_h(y)\|_p dy \\
&= \int_{-\infty}^{\infty}\|f(\cdot - y) - f(\cdot)\|_p |g_h(y)|dy \\
&= \int_{-\infty}^{\infty}\|f(\cdot - hz) - f(\cdot)\|_p |g(z)|dz. \qquad (3.11)
\end{aligned}$$

Toisessa yhtäsuuruudessa käytettiin ehtoa $\int_{-\infty}^{\infty} g(x)dx = 1$, sitten sovellettiin lemmaa 3.6, lemman 3.5 kohtaa (ii) ja viimeisessä yhtälössä tehtiin integroinnissa muuttujan vaihdos $z = y/h$. Määritellään (3.11):ssa

$$\Delta(hz) = \|f(\cdot - hz) - f(\cdot)\|_p$$

ja osoitetaan, että $\lim_{h\to 0+} \Delta(hz) = 0$ kaikilla $z \in \mathbb{R}$. Olkoon $\varepsilon > 0$ ja valitaan $a > 0$ s.e.

$$\int_{|x|\geq a} |f(x)|^p dx < \varepsilon. \tag{3.12}$$

Olkoon $z \in \mathbb{R}$ kiinteä ja h niin pieni, että $|hz| < 1$. Jaetaan $[\Delta(hz)]^p$ kahden termin summaksi kirjoittamalla

$$\begin{aligned}[\Delta(hz)]^p &= \int_{-\infty}^{\infty} |f(x - hz) - f(x)|^p dx \\ &= \int_{|x|\leq a+1} |f(x - hz) - f(x)|^p dx + \int_{|x|>a+1} |f(x - hz) - f(x)|^p dx\end{aligned}$$

ja osoitetaan, että kumpikin integraali kaavan oikealla puolella lähestyy nollaa, kun $h \to 0+$. Ensimmäisessä integraalissa havaitaan, että $x, x - hz \in [-a - 2, a + 2]$ ja koska f on tasaisesti jatkuva välillä $[-a - 2, a + 2]$ ja $|x - (x - hz)| = |hz| \to 0$, kun $h \to 0+$, lähenee integraali nollaa, kun $h \to 0+$. Toisessa integraalissa

$$|f(x - hz) - f(x)|^p \leq 2^{p-1}[|f(x - hz)|^p + |f(x)|^p].$$

Edelleen, $|x - hz| > |x| - |hz| > a + 1 - 1 = a$. Siten ehdon (3.12) nojalla toinen integraali on pienenmpi kuin $2^{p-1}(\varepsilon + \varepsilon) = 2^p\varepsilon$. On siis näytetty, että $\Delta(hz) \to 0$, kun $h \to 0+$. Toisaalta lemman 3.5 kohdan (i) nojalla

$$\Delta(hz)|g(z)| \leq [\|f(\cdot - hz)\|_p + \|f\|_p]|g(z)| = 2\|f\|_p|g(z)|$$

ja

$$\int_{-\infty}^{\infty} 2\|f\|_p|g(z)|dz < \infty,$$

koska $g \in L^1$. Lebesguen dominoidun konvergenssin lauseesta seuraa (3.11):n nojalla nyt, että

$$\lim_{h\to 0+} \|f * g_h - f\|_p \leq \lim_{h\to 0+} \int_{-\infty}^{\infty} \Delta(hz)|g(z)|dz = \int_{-\infty}^{\infty} \lim_{h\to 0+} \Delta(hz)|g(z)|dz = 0.$$

Kohta (ii) todistetaan erikseen tapauksissa $p = 1$ ja $p = 2$. Kun $p = 1$, saamme

$$\left|\int_{-\infty}^{\infty}(f * g_h)(x)dx - \int_{-\infty}^{\infty} f(x)dx\right| \leq \int_{-\infty}^{\infty}|(f * g_h)(x) - f(x)|dx = \|f * g_h - f\|_1$$

ja kohdan (i) nojalla oikea puoli lähenee nollaa, kun $h \to 0+$. Kun $p = 2$, saamme

$$\begin{aligned}
&\left|\int_{-\infty}^{\infty}[(f * g_h)(x)]^2dx - \int_{-\infty}^{\infty} f(x)^2dx\right| = \left|\int_{-\infty}^{\infty}\{[(f * g_h)(x)]^2 - f(x)^2\}dx\right| \\
&= \left|\int_{-\infty}^{\infty}[(f * g_h)(x) - f(x)][(f * g_h)(x) + f(x)]dx\right| \\
&\leq \int_{-\infty}^{\infty}|(f * g_h)(x) - f(x)||(f * g_h)(x) + f(x)|dx \\
&\leq \sqrt{\int_{-\infty}^{\infty}[(f * g_h)(x) - f(x)]^2dx}\sqrt{\int_{-\infty}^{\infty}[(f * g_h)(x) + f(x)]^2dx},
\end{aligned}$$

missä viimeisessä vaiheessa käytettiin Schwarzin epäyhtälöä. Tässä ensimmäinen neliöjuurilauseke on $\|f * g_h - f\|_2$, joka (i) kohdan nojalla lähestyy nollaa, kun $h \to 0+$. Toinen neliöjuurilauseke puolestaan pysyy rajoitettuna, koska

$$\sqrt{\int_{-\infty}^{\infty}[(f * g_h)(x) + f(x)]^2dx} = \|f * g_h + f\|_2 \leq \|f * g_h - f\|_2 + 2\|f\|_2,$$

missä viimeinen epäyhtälö seuraa lemman 3.5 kohdasta (i). Nyt todistettavan lemman kohdan (i) mukaan $\|f * g_h - f\|_2 \to 0$, kun $h \to 0+$, joten yhtälön oikea puoli on rajoitettu, kun $h \to 0+$. Näin on kohta (ii) todistettu.

Kohdan (iii) todistamiseksi oletetaan, että f on rajoitettu,

$$\|f\|_\infty \equiv \sup_{x\in\mathbb{R}}|f(x)| < \infty.$$

Silloin konvoluutio $(f * g_h)(x)$ on olemassa kaikilla $x \in \mathbb{R}$, sillä

$$\int_{-\infty}^{\infty}|f(x - y)g_h(y)|dy \leq \|f\|_\infty\int_{-\infty}^{\infty}|g_h(y)|dy < \infty$$

kaikilla $x \in \mathbb{R}$. Edelleen,

$$\begin{aligned}
|(f * g_h)(x) - f(x)| &= \left|\int_{-\infty}^{\infty} f(x - y)g_h(y)dy - f(x)\right| \\
&= \left|\int_{-\infty}^{\infty}[f(x - y) - f(x)]g_h(y)dy\right| \\
&\leq \int_{-\infty}^{\infty}|f(x - y) - f(x)||g_h(y)|dy \\
&= \int_{-\infty}^{\infty}|f(x - hz) - f(x)||g(z)|dz,
\end{aligned}$$

missä viimeisessä yhtäsuuruudessa tehtiin muuttujan vaihdos $z = y/h$. Koska f on jatkuva, on $\lim_{h\to 0+} |f(x-hz) - f(x)| = 0$ kaikilla $z \in \mathbb{R}$. Edelleen,

$$|f(x-hz) - f(x)||g(z)| \leq 2\|f\|_\infty |g(z)|$$

ja $\int_{-\infty}^{\infty} 2\|f\|_\infty |g(z)|dz < \infty$. Siten Lebesguen dominoidun konvergenssin lauseesta seuraa, että

$$\begin{aligned}\lim_{h\to 0+} |(f * g_h)(x) - f(x)| &\leq \lim_{h\to 0+} \int_{-\infty}^{\infty} |f(x-hz) - f(x)||g(z)|dz \\ &= \int_{-\infty}^{\infty} \lim_{h\to 0+} |f(x-hz) - f(x)||g(z)|dz = 0.\end{aligned}$$

Näin on koko lemma todistettu. □

Huomautus 3.10 Oletus f:n jatkuvuudesta on tarpeeton lemman 3.9 kohdissa (i) ja (ii), koska voidaan osoittaa, että L^p-funktioita voidaan approksimoida mielivaltaisen tarkasti jatkuvilla L^p-funktioilla.

Huomautus 3.11 Lemman 3.9 kohta (ii) on itseasiassa selvä jo (i):n perusteella, jos tuntee normiavaruuksien perusominaisuuksia: $\|f * g_h\|_p^p \to \|f\|_p^p$, koska $f * g_h \to f$ normin $\|\cdot\|_p$-mielessä.

Merkitään jatkossa

$$C^s = C^s(\mathbb{R}) = \{f \mid f : \mathbb{R} \to \mathbb{R} \text{ on } s \text{ kertaa jatkuvasti derivoituva}\}.$$

Edelleen, merkitsemme funktiolle $g \in L^2$, ja ytimelle K,

$$R(g) = \|g\|_2^2 = \int_{-\infty}^{\infty} g(x)^2 dx, \quad \mu_\ell(K) = \int_{-\infty}^{\infty} x^\ell K(x)dx.$$

Suureen $R(g)$ ajatellaan kuvaavan funktion g "rosoisuutta" (engl. roughness) ja $\mu_\ell(K)$ on ytimen K ℓ:s momentti (silloin, kun se on olemassa).

Lause 3.12 *Olkoon f tiheysfunktio, $f \in C^2$ ja $f, f', f'' \in L^2$. Olkoon $K \in L^1 \cap L^2$, $\int_{-\infty}^{\infty} K(x)dx = 1$, $K(x) \geq 0$, $K(x) = K(-x)$ kaikilla $x \in \mathbb{R}$ ja $\mu_2(K) < \infty$. Olkoon vielä $X_1, \ldots, X_n \sim f$ i.i.d. otos, $h_n > 0$, $n \in \mathbb{N}$, ja*

$$\hat{f}_n(x; h_n) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - X_i) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \quad x \in \mathbb{R}$$

f:n ydinestimaattori. Silloin, jos $\lim_{n\to\infty} h_n = 0$*, niin*

$$\begin{aligned}\mathrm{MISE}[\hat{f}_n(\cdot;h_n)] &= \mathbb{E}\int_{-\infty}^{\infty}[\hat{f}_n(x;h_n) - f(x)]^2 dx \\ &= \frac{1}{4}h_n^4\mu_2(K)^2 R(f'') + \frac{1}{nh_n}R(K) + o\left(h_n^4 + \frac{1}{nh_n}\right).\end{aligned} \tag{3.13}$$

Jos vielä $\lim_{n\to\infty} nh_n = \infty$*, saamme erityisesti, että* $\lim_{n\to\infty}\mathrm{MISE}[\hat{f}_n(\cdot;h_n)] = 0$*, eli estimaattori* $\hat{f}_n(\cdot;h_n)$ *on tarkentuva.*

Todistus: Olkoon ensin $h > 0$ mielivaltainen silotusparametrin arvo. Silloin

$$\begin{aligned}\mathbb{E}\int_{-\infty}^{\infty}[\hat{f}_n(x;h) - f(x)]^2 dx &= \int_{-\infty}^{\infty}\mathbb{E}[\hat{f}_n(x;h) - f(x)]^2 dx \\ &= \int_{-\infty}^{\infty}\left\{[\mathbb{E}\hat{f}_n(x;h) - f(x)]^2 + \mathrm{Var}[\hat{f}_n(x;h)]\right\}dx \\ &= \int_{-\infty}^{\infty}\mathrm{Bias}^2[\hat{f}_n(x;h)]dx + \int_{-\infty}^{\infty}\mathrm{Var}[\hat{f}_n(x;h)]dx.\end{aligned} \tag{3.14}$$

Toinen yhtälö seuraa siitä, että $\mathbb{E}[\hat{f}_n(x;h_n)^2] < \infty$ m.k. x, sillä

$$\mathbb{E}[K_h(x-X_i)K_h(x-X_j)] = \begin{cases} \int_{-\infty}^{\infty} K_h(x-y)^2 f(y)dy, & i = j, \\ \left[\int_{-\infty}^{\infty} K_h(x-y)f(y)dy\right]^2, & i \neq j, \end{cases}$$

ja $(K_h^2 * f)(x)$, $(K_h * f)(x) < \infty$, m.k. x, koska $K_h, K_h^2, f \in L^1$ (ks. huomautus 3.8).

Nyt saamme

$$\begin{aligned}\mathbb{E}[\hat{f}_n(x;h)] &= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}K_h(x-X_i) \\ &= \frac{1}{n}\cdot n\mathbb{E}K_h(x-X_1) \\ &= \int_{-\infty}^{\infty}K_h(x-y)f(y)dy \\ &= \int_{-\infty}^{\infty}f(x-z)K_h(z)dz \\ &= (f*K_h)(x),\end{aligned} \tag{3.15}$$

kaikilla x, missä neljännessä yhtälössä tehtiin muuttujan vaihdos $z = x - y$, jotta konvoluutioissa saadaan haluttu järjestys. Konvoluutio on olemassa Schwarzin

epäyhtälön nojalla, koska $f(x-\cdot), K_h \in L^2$. Edelleen, m.k. x saamme

$$\begin{aligned}
\mathrm{Var}[\hat{f}_n(x;h)] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\right] \\
&= \frac{1}{n^2}\cdot n\mathrm{Var}[K_h(x-X_1)] \\
&= \frac{1}{n}\left\{\mathbb{E}[K_h(x-X_1)^2] - [\mathbb{E}K_h(x-X_1)]^2\right\} \qquad (3.16)\\
&= \frac{1}{n}\left\{\int_{-\infty}^{\infty}\frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left[\int_{-\infty}^{\infty} K_h(x-y)f(y)dy\right]^2\right\} \\
&= \frac{1}{n}\left\{\frac{1}{h}(f*(K^2)_h)(x) - [(f*K_h)(x)]^2\right\} \\
&= \frac{1}{nh}(f*(K^2)_h)(x) - \frac{1}{n}[(f*K_h)(x)]^2.
\end{aligned}$$

Kaavoista (3.14),(3.15) ja (3.16) seuraa nyt, että

$$\begin{aligned}
\mathrm{MISE}[\hat{f}_n(\cdot;h)] &= \int_{-\infty}^{\infty}[(f*K_h)(x)-f(x)]^2dx \\
&+\frac{1}{nh}\int_{-\infty}^{\infty}(f*(K^2)_h)(x)dx - \frac{1}{n}\int_{-\infty}^{\infty}[(f*K_h)(x)]^2dx, \qquad (3.17)
\end{aligned}$$

Johdetussa kaavassa ensimmäinen termi on integroitu harha ja kaksi jälkimmäistä termiä muodostavat integroidun varianssin.

Funktio $K^2/R(K)$ toteuttaa ehdon $\int_{-\infty}^{\infty} K(x)^2/R(K)dx = 1$, joten lemman 3.9 kohdan (ii) perusteella saamme

$$\begin{aligned}
\int_{-\infty}^{\infty}(f*(K^2)_h)(x)dx &= R(K)\int_{-\infty}^{\infty}\left(f*\left(\frac{K^2}{R(K)}\right)_h\right)(x)dx \\
&= R(K)\left\{\int_{-\infty}^{\infty} f(x)dx + \varepsilon_1(h)\right\} \qquad (3.18)\\
&= R(K)(1+\varepsilon_1(h)) \\
&= R(K)+\varepsilon_2(h), \qquad (3.19)
\end{aligned}$$

missä $\varepsilon_1(h), \varepsilon_2(h) \to 0$, kun $h \to 0+$. Lemman 3.9 kohdan (ii) perusteella saadaan myös, että

$$\int_{-\infty}^{\infty}[(f*K_h)(x)]^2dx = \int_{-\infty}^{\infty} f(x)^2dx + \varepsilon_3(h) = R(f)+\varepsilon_3(h), \qquad (3.20)$$

missä $\varepsilon_3(h) \to 0$, kun $h \to 0+$.

Lopuksi käsitellään kaavan (3.17) harhatermi. Siinä

$$(f * K_h)(x) - f(x) = \int_{-\infty}^{\infty} [f(x-y) - f(x)] K_h(y) dy.$$

Sovelletaan Taylorin kaavaa integraali-jäännöstermillä,

$$f(x-y) - f(x) = -f'(x)y + \int_0^1 (1-u)y^2 f''(x-uy) du.$$

Funktion K_h symmetrisyydestä seuraa, että $\int_{-\infty}^{\infty} y K_h(y) dy = 0$ joten, edellyttäen että merkityt kaksinkertaisten integraalien ja iteroitujen integraalien yhtäsuuruudet ovat voimassa,

$$\begin{aligned}
(f * K_h)(x) - f(x) &= \int_{-\infty}^{\infty} \left\{ \int_0^1 (1-u)y^2 f''(x-uy) du \right\} K_h(y) dy \\
&= \int_{-\infty}^{\infty} \int_0^1 (1-u) y^2 K_h(y) f''(x-uy) du dy \\
&= \int_{-\infty}^{\infty} \int_1^{\infty} \left(1 - \frac{1}{\tau}\right) \tau \eta^2 K_h(\tau\eta) f''(x-\eta) d\tau d\eta \qquad (3.21) \\
&= \int_{-\infty}^{\infty} \left[\int_1^{\infty} (\tau - 1) \eta^2 \frac{1}{h} K\left(\frac{\tau\eta}{h}\right) d\tau \right] f''(x-\eta) d\eta,
\end{aligned}$$

missä tehtiin muuttujien vaihdos $y = \tau\eta$, $u = 1/\tau$, $\eta \in \mathbb{R}$, $\tau > 0$, jonka Jacobin determinantti on $-1/\tau$. Perustelemme laskun hetken kuluttua. Merkitään kuitenkin ensin

$$L(\eta) = \int_1^{\infty} (\tau - 1) \eta^2 K(\tau\eta) d\tau,$$

jolloin edellisen kaavan viimeisessä integraalissa hakasuluissa oleva lauseke voidaan kirjoittaa muotoon

$$\int_1^{\infty} (\tau - 1) \eta^2 \frac{1}{h} K\left(\frac{\tau\eta}{h}\right) du = h^2 \cdot \frac{1}{h} \int_1^{\infty} (\tau - 1) \left(\frac{\eta}{h}\right)^2 K\left(\tau \frac{\eta}{h}\right) du = h^2 L_h(\eta).$$

Ei ole vaikeaa nähdä, että $L(\eta) < \infty$ kaikilla $\eta \in \mathbb{R}$. Silloin saamme (3.21):sta hyödyllisen kaavan

$$(f * K_h)(x) - f(x) = h^2 (f'' * L_h)(x). \qquad (3.22)$$

Perustellaan nyt (3.21):ssa tehdyt integroinnit. Ensiksi saadaan

$$\begin{aligned}\int_{-\infty}^{\infty} L(\eta)d\eta &= \int_1^{\infty}(\tau-1)\left[\int_{-\infty}^{\infty}\eta^2 K(\tau\eta)d\eta\right]d\tau \\ &= \int_1^{\infty}(\tau-1)\left[\int_{-\infty}^{\infty}\frac{y^2}{\tau^2}K(y)\frac{1}{\tau}dy\right] \\ &= \int_1^{\infty}\left(\frac{1}{\tau^2}-\frac{1}{\tau^3}\right)d\tau\int_{-\infty}^{\infty}y^2K(y)dy \\ &= \frac{1}{2}\mu_2(K)<\infty,\end{aligned} \tag{3.23}$$

missä toisessa yhtälössä tehtiin sijoitus $\tau\eta = y$. Siten $L \in L^1$. Sijoittamalla $y = \tau\eta$, $u = 1/\tau$ saadaan nyt

$$\int_{-\infty}^{\infty}\int_0^1 |(1-u)y^2K_h(y)f''(x-uy)|dudy = h^2(|f''| * L_h)(x) < \infty$$

m.k. x, koska $f'' \in L^2$ ja $L_h \in L^1$ (ks. huomautus 3.8). Fubinin lauseesta seuraa nyt, että (3.21):ssa tehdyt integroinnit ovat sallittuja.

Koska siis $\int_{-\infty}^{\infty} L(\eta)d\eta = (1/2)\mu_2(K)$, saadaan lemman 3.9 kohdan (ii) perusteella

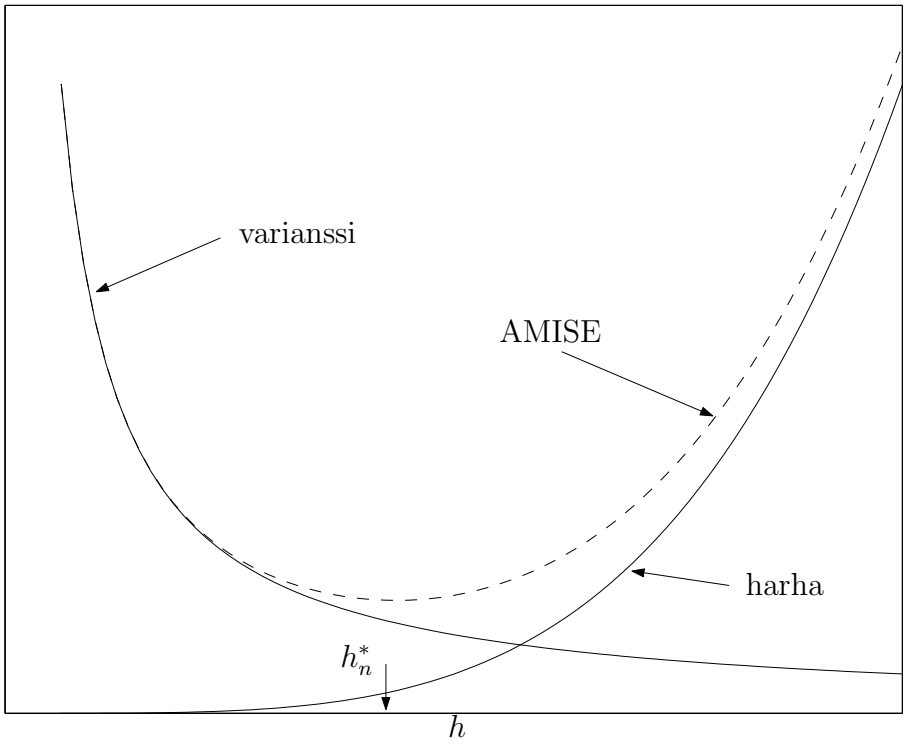
$$\begin{aligned}\int_{-\infty}^{\infty}[(f*K_h)(x)-f(x)]^2dx &= \int_{-\infty}^{\infty}[h^2(f''*L_h)(x)]^2dx \\ &= h^4\left(\frac{\mu_2(K)}{2}\right)^2\int_{-\infty}^{\infty}\left[\left(f''*\left(\frac{L}{\mu_2(K)/2}\right)_h\right)\right]^2dx \\ &= \frac{1}{4}h^4\mu_2(K)^2\left\{\int_{-\infty}^{\infty}[f''(x)]^2dx+\varepsilon_4(h)\right\} \\ &= \frac{1}{4}h^4\mu_2(K)^2R(f'')+h^4\varepsilon_5(h),\end{aligned} \tag{3.24}$$

missä $\varepsilon_4(h), \varepsilon_5(h) \to 0$, kun $h \to 0+$. Kaavoista (3.17),(3.18),(3.20) ja (3.24) seuraa nyt, että

$$\begin{aligned}\mathrm{MISE}[\hat{f}_n(\cdot;h)] = \frac{1}{4}h^4\mu_2(K)^2R(f'')+\frac{1}{nh}R(K) \\ -\frac{1}{n}R(f)+h^4\varepsilon_5(h)+\frac{1}{nh}\varepsilon_2(h)-\frac{1}{n}\varepsilon_3(h).\end{aligned}$$

Jos nyt h:n paikalle sijoitetaan h_n ja $h_n \to 0+$, seuraa tästä (3.13). □

Määritellään ydinestimaattorin asymptoottinen keskimääräinen integroitu neliöllinen virhe kaavalla



Kuva 3.9: Asymptoottinen integroitu neliöllinen virhe AMISE (katkoviiva) saadaan laskemalla yhteen harhan ja varianssin osuudet (yhtenäiset viivat). Optimaalinen silotusparametrin arvo h_n^* minimoi asymptoottisen virheen.

$$\text{AMISE}[\hat{f}_n(\cdot;h)] = \frac{1}{4}h^4\mu_2(K)^2R(f'') + \frac{1}{nh}R(K). \tag{3.25}$$

Tässä ensimmäinen termi edustaa asymptoottista integroitua harhaa ja toinen termi vastaavasti asymptoottista integroitua varianssia. Olkoon otoskoko n kiinteä. Silloin havaitsemme, että kun $h \to 0+$, niin harhan osuus lähestyy nollaa ja varianssin puolestaan kasvaa rajatta. Kun taas $h \to \infty$, tapahtuu täsmälleen päinvastoin, harhan osuus kasvaa rajatta ja varianssin puolestaan lähestyy nollaa. Tilanne on esitetty kuvassa 3.9. Jotta koko asymptoottinen virhe AMISE$[\hat{f}_n(\cdot;h)]$ saataisiin pieneksi, on siis löydettävä tasapaino harhan ja varianssin välillä. Tätä ongelmaa kutsutaan joskus nimellä "bias-variance trade-off".

Asymptoottisesti optimaalisen silotusparamterin löytämiseksi tulee AMISE$[\hat{f}_n(\cdot;h)]$ minimoida h:n suhteen. Asettamalla (3.25):n oikean puolen h:n

suhteen laskettu derivaatta nollaksi ja ratkaisemalla h saadaan optimaaliseksi silotusparametrin arvoksi

$$h_n^* = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')}\right]^{1/5} \cdot n^{-1/5}. \tag{3.26}$$

Sijoittamalla h_n^* h:n paikalle kaavaan (3.25) saadaan optimaalinen asymptoottinen integroitu neliöllinen virhe

$$\mathrm{AMISE}[\hat{f}_n(\cdot; h_n^*)] = \frac{5}{4}[\mu_2(K)^2 R(f'')]^{1/5}[R(K)]^{4/5} \cdot n^{-4/5}. \tag{3.27}$$

Siten optimaalinen ydinestimaattori konvergoi kohti estimoitavaa funktiota vauhdilla $n^{-4/5}$. Osoittautuu, että tämä vauhti on tietyssä mielessä *paras* mikä parametrittomalla menetelmällä ylipäätänsä voidaan saavuttaa. Tässä mielessä ydinestimaattori on optimaalinen parametriton tiheysfunktion estimaattori.

3.5 Minimax-virhe

Lause 3.12 koski yhtä kiinteätä tiheysfunktiota f. Tämän lisäksi on kiinnostavaa tutkia tiheysfunktioestimaattorin suorituskykyä kokonaisessa funktioluokassa.

Olkoon

$$\mathcal{F} \subset \{f \mid f : \mathbb{R} \to [0, \infty[\text{ tiheysfunktio}\},$$

$f \in \mathcal{F}$, $X_1, \ldots, X_n \sim f$ i.i.d. otos ja $\hat{f}_n(x) = t_n(x, X_1, \ldots, X_n), x \in \mathbb{R}$, f:n estimaattori. Tarkastellaan suurinta mahdollista virhettä luokassa $f \in \mathcal{F}$, eli suuretta

$$\begin{aligned}\sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx &= \sup_{f\in\mathcal{F}} \int_{-\infty}^{\infty} \mathbb{E}_f [\hat{f}_n(x) - f(x)]^2 dx \\ &= \sup_{f\in\mathcal{F}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} [t(x, x_1, \ldots, x_n) - f(x)]^2 \prod_{i=1}^{n} f(x_i) dx_1 \cdots dx_n dx.\end{aligned}$$

Kun etsitään parasta estimaattoria koko luokassa $\mathcal{F}$, ollaan kiinnostuneita niin sanotusta *minimax-virheestä*

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx,$$

missä infimum otetaan yli kaikkien f:n estimaattoreiden (eli siis yli kaikkien Borelin funktioiden t_n). Termi "minimax" tulee siitä, että ensin otetaan infimum ja sitten supremum.

Oletetaan, että on olemassa jono $(a_n)_{n\in\mathbb{N}}$ positiivisia lukuja ja $A > 0$ siten että

$$\liminf_{n\to\infty} a_n \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx \geq A. \tag{3.28}$$

Silloin parhaankin estimaattorijonon $(\hat{f}_n)_{n\in\mathbb{N}}$ konvergenssivauhti on korkeintaan $1/a_n$ yli luokan $\mathcal{F}$: löytyy $\varepsilon > 0$ s.e., jos $A' = A - \varepsilon$, niin

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx \geq \frac{A'}{a_n}, \tag{3.29}$$

kaikilla $\hat{f}_n$, kun n on riittävän suuri. Kaavassa (3.28) voitaisiin vaihtoehtoisesti vaatia, että

$$a_n \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx \geq A''$$

kaikilla n mutta tällöin luultavasti A'' olisi paljon pienempi kuin A ja saatu alaraja A''/a_n vastaavasti huonompi kuin A'/a_n ja siksi käytämmekin mieluummin ehtoa (3.28).

Jos sitten löytyy estimaattorijono $(\hat{f}_n^*)_{n\in\mathbb{N}}$ s.e. jollain $B < \infty$ pätee

$$\limsup_{n\to\infty} a_n \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n^*(x) - f(x)]^2 dx \leq B, \tag{3.30}$$

sanomme, että $(\hat{f}_n^*)_{n\in\mathbb{N}}$ on *minimax-optimaalinen*. (Tässä lim sup:n käyttöön pätee sama huomio kuin lim inf:n käyttöön kaavassa (3.28).) Silloin siis löytyy $B' < \infty$ s.e. suurilla n,

$$\frac{A'}{a_n} \leq \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n^*(x) - f(x)]^2 dx \leq \frac{B'}{a_n}. \tag{3.31}$$

Niinpä epäyhtälöistä (3.29) ja (3.31) seuraa, että löytyy $C < \infty$ s.e. suurilla n,

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n^*(x) - f(x)]^2 dx \leq C \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f \int_{-\infty}^{\infty} [\hat{f}_n(x) - f(x)]^2 dx.$$

Sanomme, että $1/a_n$ on tiheysfunktioestimaattorin (minimax-)*optimaalinen konvergenssivauhti* luokassa $\mathcal{F}$.

Osoittautuu, että tiheysfunktion ydinestimaattorilla on optimaalinen konvergenssivauhti monilla luokilla $\mathcal{F}$ ja erilaisilla virhekriteereillä mitattuna (ks. [1, 2]). Yksinkertaisuuden vuoksi todistamme tämän optimaalisuuden seuraavaksi eräällä melko yksinkertaisella $\mathcal{F}$ käyttäen pisteittäistä virhettä (integroidun virheen käsittely vaatisi vielä jonkun verran lisätyötä).

Olkoon siis $C > 0$ ja merkitään

$$\mathcal{F}_{2,C} = \{f \mid f \in C^2(\mathbb{R}) \text{ tiheysfunktio }, \|f\|_\infty \leq C, \|f''\|_\infty \leq C\}.$$

Siis $|f(x)| \leq C$, $|f''(x)| \leq C$ kaikilla $x \in \mathbb{R}$. Olkoon $x_0 \in \mathbb{R}$ kiinteä ja $\hat{\varphi}_n = t_n(X_1, \ldots, X_n)$ arvon $f(x_0)$ jokin estimaattori, $X_1, \ldots, X_n \sim f$.

Lause 3.13 *Olkoon $\hat{\varphi}_n^* = (1/n)\sum_{i=1}^n K_{h_n}(x_0 - X_i)$, missä $h_n = \lambda n^{-1/5}$, $\lambda > 0$ ja K kuten lauseessa 3.12. Silloin on olemassa $B > 0$ s.e.*

$$n^{4/5} \sup_{f \in \mathcal{F}_{2,C}} \mathbb{E}_f[\hat{\varphi}_n^* - f(x_0)]^2 \leq B, \quad n \in \mathbb{N}.$$

Todistus: Merkitään $h = h_n$. Olkoon $f \in \mathcal{F}_{2,C}$. Koska f on rajoitettu, on olemassa odotusarvo

$$\mathbb{E}_f \hat{\varphi}_n^* = \int_{-\infty}^{\infty} f(x_0 - y) K_h(y) dy = (f * K_h)(x_0).$$

Samaten $\mathbb{E}_f[(\hat{\varphi}_n^*)^2] < \infty$ ja

$$\begin{aligned}
\mathbb{E}_f[\hat{\varphi}_n^* - f(x_0)]^2 &= [\mathbb{E}_f \hat{\varphi}_n^* - f(x_0)]^2 + \mathrm{Var}[\hat{\varphi}_n^*] \\
&= [(f * K_h)(x_0) - f(x_0)]^2 + \frac{1}{nh}(f * (K^2)_h)(x_0) - \frac{1}{n}[(f * K_h)(x_0)]^2 \\
&\leq [(f * K_h)(x_0) - f(x_0)]^2 + \frac{1}{nh}(f * (K^2)_h)(x_0) \qquad (3.32) \\
&= [h^2 (f'' * L_h)(x_0)]^2 + \frac{1}{nh}(f * (K^2)_h)(x_0) \\
&= h^4 [(f'' * L_h)(x_0)]^2 + \frac{1}{nh}(f * (K^2)_h)(x_0),
\end{aligned}$$

missä toisessa yhtälössä käytettiin kaavaa (3.16) ja toiseksi viimeisessä yhtälössä

kaavaa (3.22). Tässä

$$\begin{aligned} |(f'' * L_h)(x_0)| &= \left| \int_{-\infty}^{\infty} f''(x_0 - y) L_h(y) dy \right| \\ &\leq \int_{-\infty}^{\infty} |f''(x_0 - y)| L_h(y) dy \\ &\leq C \int_{-\infty}^{\infty} L_h(y) dy \\ &= \frac{1}{2} \mu_2(K) C, \end{aligned}$$

missä toisessa epäyhtälössä käytettiin sitä, että $f \in \mathcal{F}_{2,C}$ ja viimeisessä yhtälössä hyödynnettiin kaavaa (3.23). Edelleen,

$$\begin{aligned} |(f * (K^2)_h)(x_0)| &= \left| \int_{-\infty}^{\infty} f(x_0 - y) (K^2)_h(y) dy \right| \\ &\leq \int_{-\infty}^{\infty} |f(x_0 - y)| (K^2)_h(y) dy \\ &\leq C \int_{-\infty}^{\infty} (K^2)_h(y) dy \\ &= C \int_{-\infty}^{\infty} K(y)^2 dy. \\ &= CR(K). \end{aligned}$$

Sijoittamalla nämä arviot kaavaan (3.32) saadaan

$$\mathbb{E}_f[\hat{\varphi}_n^* - f(x_0)]^2 \leq \frac{1}{4} \mu_2(K)^2 C^2 h^4 + \frac{1}{nh} CR(K).$$

Valitsemalla $h = \lambda n^{-1/5}$ tämän kaavan oikea puoli saadaan muotoon $Bn^{-4/5}$, missä

$$B = \frac{1}{4} \mu_2(K)^2 C^2 \lambda^4 + \frac{CR(K)}{\lambda}. \qquad \square$$

Huomautus 3.14 Lauseen 3.13 todistuksessa saatua vakiota B voidaan optimoida valitsemalla λ sopivasti. Pienin arvo B:lle saavutetaan, kun

$$\lambda = \left(\frac{R(K)}{\mu_2(K)^2 C} \right)^{1/5} \tag{3.33}$$

(vrt. (3.26)). Tällöin

$$B = \frac{5}{4} C^{6/5} \left(\sqrt{\mu_2(K)} R(K) \right)^{4/5} \tag{3.34}$$

(vrt. (3.27)).

Nyt siis olemme osoittaneet, että (3.30) pätee ydinestimaattorille (pisteittäin), kun $a_n = n^{4/5}$. Seuraavaksi osoitamme vastaavan alarajatuloksen, eli että (3.28) pätee (pisteittäin), kun $a_n = n^{4/5}$ ja $\mathcal{F} = \mathcal{F}_{2,C}$. Todistuksen idea on seuraava. Kullakin otoskoolla n funktioluokka $\mathcal{F}_{2,C}$ korvataan kahden funktion osajoukollaan $\{f_n^+, f_n^-\} \subset \mathcal{F}_{2,C}$, jossa f_n^+ ja f_n^- poikkeavat toisistaan juuri *sopivan paljon*:

$$\sup_{f \in \mathcal{F}_{2,C}} \mathbb{E}_f[\hat{\varphi}_n - f(x_0)]^2 \geq \max_{f \in \{f_n^+, f_n^-\}} \mathbb{E}_f[\hat{\varphi}_n - f(x_0)]^2 \geq An^{-4/5}$$

riippumatta estimaattorista $\hat{\varphi}_n$ ja missä A on otoskoosta n riippumaton vakio. Funktioiden f_n^+ ja f_n^- valinnassa on oltava tarkka. Jos nimittäin $f_n^+ = f_n^-$, saadaan virheen alarajaksi nolla valitsemalla $\hat{\varphi}_n = f_n^+(x_0)$. Jos taas f_n^+ ja f_n^- ovat täysin erilaiset, voidaan nolla-virheeseen päästä yksinkertaisella testillä $X_1 \geq \alpha$ eli ottamalla otokseen $X_1, \ldots, X_n$ perustuvaksi estimaattoriksi

$$\hat{\varphi}_n = \begin{cases} f_n^+(x_0), & \text{jos } X_1 \geq \alpha, \\ f_n^-(x_0), & \text{jos } X_1 < \alpha \end{cases}$$

(ks. kuva 3.10). Alarajatuloksen todistuksessa tarvitsemme seuraavaa lemmaa.

Lemma 3.15 *Olkoot f ja g tiheysfunktioita. Silloin*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \min\{\prod_{i=1}^{n} f(x_i), \prod_{i=1}^{n} g(x_i)\} dx_1 \cdots dx_n \geq 1 - \sqrt{2n\left(1 - \int_{-\infty}^{\infty} \sqrt{f(x)g(x)}dx\right)}.$$

Todistus: HT. □

Lause 3.16 *On olemassa $A > 0$ s.e. kaikilla n pätee*

$$n^{4/5} \inf_{\hat{\varphi}_n} \sup_{f \in \mathcal{F}_{2,C}} \mathbb{E}_f[\hat{\varphi}_n - f(x_0)]^2 \ \geq A.$$

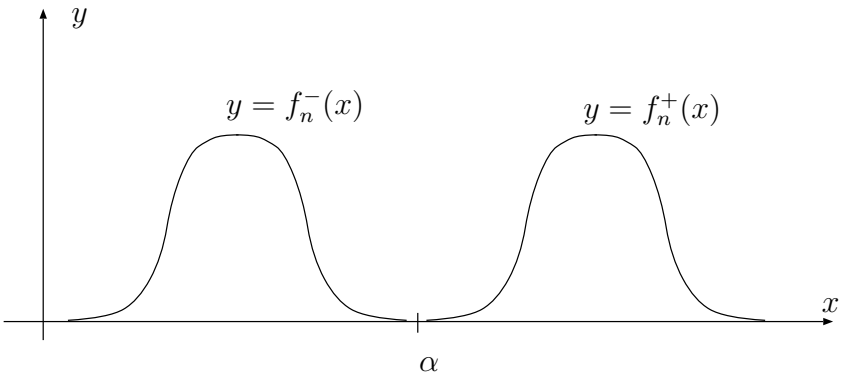
Todistus: Valitaan ensin kiinteä $f_0 \in \mathcal{F}_{2,C/2}$ s.e. eräällä $D > 0$ pätee

(i) $f_0(x) \geq D$, kun $|x - x_0| \leq 1$.

Valitaan sitten $g \in C^2$ s.e.

(ii) $\|g\|_\infty \leq C/2$, $\|g''\|_\infty \leq C/2$ ja $g(x) = 0$, kun $|x| > 1$,

(iii) $\int_{-\infty}^{\infty} g(y)dy = 0$,



Kuva 3.10: Kaksi tiheysfunktiota, jotka voidaan erottaa toisistaan yhteen otospisteeseen perustuvan testin $X_1 \geq \alpha$ avulla.

(iv) $\int_{-\infty}^{\infty} g(y)^2 dy < D/2$,

(v) $g(0) \neq 0$.

Määritellään sitten $g_n(x) = n^{-2/5} g(n^{1/5}(x - x_0))$, $x \in \mathbb{R}$ ja vaaditaan vielä g:stä, että

(vi) $|g_n(x)| \leq f_0(x)$ kaikilla $x \in \mathbb{R}$, $n \in \mathbb{N}$.

Sopivan parin f_0, g konstruointi on harjoitustehtävänä.

Määritellään nyt

$$\begin{cases} f_n^+ &= f_0 + g_n, \\ f_n^- &= f_0 - g_n. \end{cases} \qquad n \in \mathbb{N}.$$

Silloin $f_n^+, f_n^- \in C^2$, $\int_{-\infty}^{\infty} f_n^+(x)dx = \int_{-\infty}^{\infty} f_n^-(x)dx = 1$ kohdan (iii) nojalla ja $f_n^+(x), f_n^-(x) \geq 0$ kaikilla $x \in \mathbb{R}$ kohdan (vi) nojalla, joten f_n^+ ja f_n^- ovat C^2-tiheysfunktioita. Edelleen,

$$g_n''(x) = n^{-2/5} n^{2/5} g''(n^{1/5}(x - x_0)) = g''(n^{1/5}(x - x_0)), \quad x \in \mathbb{R}, n \in \mathbb{N}.$$

Ehdosta (ii) ja siitä, että $f_0 \in \mathcal{F}_{2,C/2}$ seuraa nyt, että $f_n^+, f_n^- \in \mathcal{F}_{2,C}$ kaikilla n.

Nyt

$$\begin{aligned}
\inf_{\hat{\varphi}_n} \sup_{f\in\mathcal{F}_{2,C}} \mathbb{E}_f[\hat{\varphi}_n - f(x_0)]^2 &\geq \inf_{\hat{\varphi}_n} \max_{f\in\{f_n^+,f_n^-\}} \mathbb{E}_f[\hat{\varphi}_n - f(x_0)]^2 \\
&\geq \inf_{\hat{\varphi}_n} \frac{1}{2}\left\{\mathbb{E}_{f_n^+}[\hat{\varphi}_n - f_n^+(x_0)]^2 + \mathbb{E}_{f_n^-}[\hat{\varphi}_n - f_n^-(x_0)]^2\right\} \\
&= n^{-4/5} \inf_{\hat{\varphi}_n} \frac{1}{2}\left\{\mathbb{E}_{f_n^+}[g(0) + n^{2/5}(f_0(x_0) - \hat{\varphi}_n)]^2 + \mathbb{E}_{f_n^-}[g(0) - n^{2/5}(f_0(x_0) - \hat{\varphi}_n)]^2\right\} \\
&= n^{-4/5} \inf_{\hat{\varphi}_n} \frac{1}{2}\left\{\mathbb{E}_{f_n^+}[g(0) + \hat{\varphi}_n]^2 + \mathbb{E}_{f_n^-}[g(0) - \hat{\varphi}_n]^2\right\},
\end{aligned}$$

missä toiseksi viimeisessä yhtälössä käytettiin f_n^+:n ja f_n^-:n määritelmiä ja viimeinen yhtäsuuruus seuraa siitä, että $n^{2/5}(f_0(x_0) - \hat{\varphi}_n)$ käy läpi kaikki estimaattorit, kun $\hat{\varphi}_n$ käy läpi kaikki estimaattorit.

Edelleen, jos $\hat{\varphi}_n = t_n(X_1, \ldots, X_n)$, niin

$$\begin{aligned}
\mathbb{E}_{f_n^+}[g(0) + \hat{\varphi}_n]^2 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [g(0) + t_n(x_1, \ldots, x_n)]^2 \prod_{i=1}^n f_n^+(x_i) dx_1 \cdots dx_n \\
&\geq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [g(0) + t_n(x_1, \ldots, x_n)]^2 \min\left\{\prod_{i=1}^n f_n^+(x_i), \prod_{i=1}^n f_n^-(x_i)\right\} dx_1 \cdots dx_n
\end{aligned}$$

ja samoin funktiolle f_n^-. Siten

$$\begin{aligned}
&\mathbb{E}_{f_n^+}[g(0) + \hat{\varphi}_n]^2 + \mathbb{E}_{f_n^-}[g(0) - \hat{\varphi}_n]^2 \\
&\quad\geq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{[g(0) + t_n(x_1, \ldots, x_n)]^2 + [g(0) - t_n(x_1, \ldots, x_n)]^2\right\} \times \\
&\quad\min\left\{\prod_{i=1}^n f_n^+(x_i), \prod_{i=1}^n f_n^-(x_i)\right\} dx_1 \cdots dx_n.
\end{aligned}$$

Toisaalta, kaikilla $a, b \in \mathbb{R}$ pätee

$$(a+b)^2 + (a-b)^2 = 2a^2 + 2b^2 \geq 2a^2.$$

Valitsemalla yllä $a = g(0)$, $b = t_n(x_1, \ldots, x_n)$, saadaan yo. integraalille ala-arvio

$$2g(0)^2 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \min\left\{\prod_{i=1}^n f_n^+(x_i), \prod_{i=1}^n f_n^-(x_i)\right\} dx_1 \cdots dx_n.$$

Valitsemalla lemmassa 3.15 $f = f_n^+$ ja $g = f_n^-$ saadaan edelleen alaraja

$$2g(0)^2 \left[1 - \sqrt{2n\left(1 - \int_{-\infty}^{\infty} \sqrt{f_n^+(x) f_n^-(x)} dx\right)}\right].$$

Yhdistämällä tähänastiset tulokset saadaan

$$\inf_{\hat{\varphi}_n} \sup_{f \in \mathcal{F}_{2,C}} \mathbb{E}_f[\hat{\varphi}_n - f(x_0)]^2 \tag{3.35}$$
$$\geq n^{-4/5} \frac{1}{2} \cdot 2g(0)^2 \left[1 - \sqrt{2n \left(1 - \int_{-\infty}^{\infty} \sqrt{f_n^+(x) f_n^-(x)} dx \right)} \right].$$

Huomaa, että ehdon (v) nojalla $g(0)^2 > 0$. Lopuksi osoitetaan, että kaavassa (3.35) hakasuluissa oleva lauseke on rajoitettu alhaalta positiivisella luvulla λ, joka ei riipu n:stä. Olkoon $S = \{x \in \mathbb{R} \mid f_0(x) > 0\}$. Silloin saadaan

$$\begin{aligned}
\int_{-\infty}^{\infty} \sqrt{f_n^+(x) f_n^-(x)} dx &= \int_{-\infty}^{\infty} \sqrt{[f_0(x) + g_n(x)][f_0(x) - g_n(x)]} dx \\
&= \int_{-\infty}^{\infty} \sqrt{f_0(x)^2 - g_n(x)^2} dx \\
&= \int_S \sqrt{f_0(x)^2 - g_n(x)^2} dx \\
&= \int_S f_0(x) \sqrt{1 - [g_n(x)/f_0(x)]^2} dx \\
&\geq \int_S f_0(x) \{1 - [g_n(x)/f_0(x)]^2\} dx \\
&= \int_S f_0(x) dx - \int_S [g_n(x)^2/f_0(x)] dx \\
&= 1 - \int_S [g_n(x)^2/f_0(x)] dx.
\end{aligned}$$

missä kolmas yhtälö seuraa siitä, että (vi):n nojalla $g_n(x) = 0$ aina kun $f_0(x) = 0$ ja viidennen vaiheen epäyhtälö seuraa siitä, että samoin (vi):n nojalla $1 - [g_n(x)/f_0(x)]^2 \leq 1$, kun $x \in S$. Edelleen, ehdon (ii) nojalla $g_n(x) = 0$, kun $n^{1/5}|x - x_0| > 1$ eli ainakin silloin, kun $|x - x_0| > 1$. Siten

$$\int_S [g_n(x)^2/f_0(x)] dx = \int_{|x-x_0| \leq 1} [g_n(x)^2/f_0(x)] dx \leq \frac{1}{D} \int_{-\infty}^{\infty} g_n(x)^2 dx,$$

missä epäyhtälössä käytettiin ehtoa (i). Edelleen,

$$\begin{aligned}
\int_{-\infty}^{\infty} g_n(x)^2 dx &= n^{-4/5} \int_{-\infty}^{\infty} g(n^{1/5}(x - x_0))^2 dx \\
&= n^{-4/5} n^{-1/5} \int_{-\infty}^{\infty} g(y)^2 dy \\
&= \frac{1}{n} \int_{-\infty}^{\infty} g(y)^2 dy,
\end{aligned}$$

missä toisessa yhtälössä tehtiin muuttujan vaihdos $y = n^{1/5}(x - x_0)$. Siten

$$\int_{-\infty}^{\infty} \sqrt{f_n^+(x) f_n^-(x)} dx \geq 1 - \frac{1}{Dn} \int_{-\infty}^{\infty} g(y)^2 dy$$

ja saamme, että

$$\begin{aligned}
&1 - \sqrt{2n \left(1 - \int_{-\infty}^{\infty} \sqrt{f_n^+(x) f_n^-(x)} dx\right)} \\
&\quad \geq 1 - \sqrt{2n \left(1 - \left(1 - \frac{1}{Dn} \int_{-\infty}^{\infty} g(x)^2 dx\right)\right)} \\
&\quad = 1 - \sqrt{\frac{2n}{Dn} \int_{-\infty}^{\infty} g(x)^2 dx} \\
&\quad = 1 - \sqrt{\frac{2}{D} \int_{-\infty}^{\infty} g(x)^2 dx} \;\equiv\; \lambda,
\end{aligned}$$

missä $\lambda > 0$ ehdon (iv) nojalla. □

3.6 Optimaalinen ydin

Kaavoissa (3.25), (3.26) ja (3.27) on annettu lauseen 3.12 ydinestimaattorin asymptoottinen integroitu neliöllinen virhe, tämän minimoiva optimaalinen silotusparametri h_n^* ja sitä vastaava minimaalinen asymptoottinen integroitu neliöllinen virhe, joka voidaan kirjoittaa muotoon

$$\mathrm{AMISE}[\hat{f}_n(\cdot; h_n^*)] = \frac{5}{4} C(K) R(f'')^{1/5} n^{-4/5},$$

missä

$$C(K) = [\mu_2(K) R(K)^2]^{2/5}. \tag{3.36}$$

Optimaalinen ydin on sellainen K, jolla $C(K)$ saavuttaa pienimmän arvonsa.

Merkitään $D(K) = \mu_2(K) R(K)^2$, jolloin $C(K) = D(K)^{2/5}$. Suure $D(K)$ on

invariantti skaalauksen suhteen:

$$\begin{aligned} D(K_h) &= \int_{-\infty}^{\infty} x^2 K_h(x)dx \left[\int_{-\infty}^{\infty} [K_h(x)]^2 dx\right]^2 \\ &= \int_{-\infty}^{\infty} x^2 \frac{1}{h} K\left(\frac{x}{h}\right) dx \left[\int_{-\infty}^{\infty} \frac{1}{h^2} K\left(\frac{x}{h}\right)^2 dx\right]^2 \\ &= \int_{-\infty}^{\infty} h^2 y^2 K(y)dy \left[\int_{-\infty}^{\infty} \frac{1}{h} K(y)^2 dy\right]^2 \\ &= \int_{-\infty}^{\infty} y^2 K(y)dy \left[\int_{-\infty}^{\infty} K(y)^2 dy\right]^2 = D(K), \end{aligned}$$

missä kolmas yhtäsuuruus seuraa muuttujan vaihdosta $y = x/h$. Ydin K minimoi suureen $C(K)$ täsmälleen silloin, kun se minimoi suureen $D(K)$ ja seuraavassa lauseessa osoitamme, että minimoiva ydin on

$$K^*(x) = \frac{3}{4}(1-x^2)_+ = \begin{cases} \frac{3}{4}(1-x^2), & \text{kun } |x| \le 1, \\ 0, & \text{kun } |x| > 1. \end{cases}$$

Lause 3.17 *Olkoon K tiheysfunktio ja $\mu_2(K) < \infty$. Silloin $D(K) \ge D(K^*)$.*

Todistus: Ytimen K^* toinen momentti on

$$\mu_2(K^*) = \int_{-1}^{1} x^2 \cdot \frac{3}{4}(1-x^2)dx = \frac{1}{5}.$$

Koska $D(K)$ on invarinatti skaalauksessa, voimme olettaa, että $\mu_2(K) = \mu_2(K^*) = 1/5$. Näin siksi, että $D(K_h) = D(K)$ ja

$$\mu_2(K_h) = \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{h} K\left(\frac{x}{h}\right) dx = h^2 \mu_2(K) = \frac{1}{5},$$

kun $h = \sqrt{1/[5\mu_2(K)]}$. Saamme

$$\begin{aligned} \int_{-\infty}^{\infty} K(x)^2 dx &= \int_{-\infty}^{\infty} \{[(K(x) - K^*(x)] + K^*(x)\}^2 dx \\ &= \int_{-\infty}^{\infty} [K(x) - K^*(x)]^2 dx + \int_{-\infty}^{\infty} [K^*(x)]^2 dx \\ &+ 2\int_{-\infty}^{\infty} K^*(x)[K(x) - K^*(x)]dx. \end{aligned}$$

Tässä $\int_{-\infty}^{\infty}[K(x)-K^*(x)]^2dx \geq 0$ ja

$$\begin{aligned}\int_{-\infty}^{\infty} K^*(x)[K(x)-K^*(x)]dx &= \int_{-1}^{1}\frac{3}{4}(1-x^2)[K(x)-K^*(x)]dx \\ &= -\int_{\mathbb{R}\setminus[-1,1]}\frac{3}{4}(1-x^2)[K(x)-K^*(x)]dx \\ &= \int_{\mathbb{R}\setminus[-1,1]}\frac{3}{4}(x^2-1)K(x)dx \quad \geq \quad 0,\end{aligned}$$

missä viimeinen epäyhtälö seuraa siitä, että K on ei-negatiivinen ja toisessa yhtälössä käytettiin sitä, että

$$\begin{aligned}&\int_{-\infty}^{\infty}\frac{3}{4}(1-x^2)[K(x)-K^*(x)]dx \\ &\quad = \frac{3}{4}\left\{\int_{-\infty}^{\infty}K(x)dx - \int_{-\infty}^{\infty}K^*(x)dx - \mu_2(K)+\mu_2(K^*)\right\} = 0,\end{aligned}$$

koska $\int_{-\infty}^{\infty}K(x)dx = \int_{-\infty}^{\infty}K^*(x)dx = 1$ ja $\mu_2(K) = \mu_2(K^*) = 1/5$. Siten

$$\int_{-\infty}^{\infty}K(x)^2dx \geq \int_{-\infty}^{\infty}[K^*(x)]^2dx,$$

eli $R(K) \geq R(K^*)$ ja saamme viimein, että

$$D(K) = \mu_2(K)R(K)^2 = \mu_2(K^*)R(K)^2 \geq \mu_2(K^*)R(K^*)^2 = D(K^*). \qquad \square$$

Funktio K^* on ns. *Epanechnikovin ydin*. Kyseessä on erikoistapaus beta-jakauman tiheysfunktiosta. Yleisesti $f \sim \text{Beta}(\alpha,\beta,\gamma,\delta)$, kun

$$f(x) = \begin{cases} C(x-\gamma)^{\alpha-1}(\delta-x)^{\beta-1}, & \text{kun } x \in [\gamma,\delta] \\ 0, & \text{muulloin.}\end{cases}$$

Tässä $\alpha,\beta > 0$, $\gamma < \delta$ ja $C = C(\alpha,\beta,\gamma,\delta)$ normalisointivakio, joka takaa ehdon $\int_{-\infty}^{\infty}f(x)dx = 1$. Olkoon $p \in \{0,1,2,\dots\}$ ja $\alpha = \beta = p+1$, $\gamma = -1$, $\delta = 1$. Määritellään

$$K(x;p) = \begin{cases} C_p(x+1)^p(1-x)^p, & \text{kun } |x| \leq 1, \\ 0, & \text{muulloin,}\end{cases}$$

eli $K(x;p) = C_p(1-x^2)_+^p$. Taulukossa 3.1 on listattu eri p:n arvoilla saatavia ytimiä.

p	nimi	sileys
0	Tasainen	
1	Epanechnikov	$C(\mathbb{R})$
2	Biweight	$C^1(\mathbb{R})$
3	Triweight	$C^2(\mathbb{R})$

Taulukko 3.1: Beta-jakauman tiheysfunktiosta saatavia ytimiä.

Jos estimoinnissa ei käytetä optimaalista Epanechnikovin ydintä, voidaan kysyä miten paljon suurempi otoskoko tarvitaan jotta saavutetaan sama asymptoottinen virhe kuin jos käytettäisiin optimaalista ydintä. Merkitään ytimeen K perustuvaa ydinestimaattoria $\hat{f}_n(\cdot; h, K)$. Vaadimme siis, että

$$1 = \frac{\mathrm{AMISE}[\hat{f}_{n_1}(\cdot; h^*_{K^*,n_1}, K^*)]}{\mathrm{AMISE}[\hat{f}_{n_2}(\cdot; h^*_{K,n_2}, K)]} = \frac{C(K^*)}{C(K)} \left(\frac{n_1}{n_2}\right)^{-4/5},$$

missä $h^*_{K,n}$ on ytimeen K perustuvan ydinestimaattorin asymptoottisesti optimaalinen silotusparamtetri ja $C(K)$ saadaan kaavasta (3.36). Saamme tästä ehdon

$$\frac{n_1}{n_2} = \left[\frac{C(K^*)}{C(K)}\right]^{5/4}.$$

Suhde n_1/n_2 mittaa ytimen K tehokkuutta optimaalisen ytimen K^* suhteen. Taulukossa 3.2 on laskettu eri ytimien tehokkuuksia. Mukaan otettu kolmioydin on $K(x) = (1 - |x|)_+$. Johtopäätöksenä voidaan todeta, että ytimen muodolla on melko vähän vaikutusta estimoinnin tehokkuuteen. Oleellisempia ytimen valintakriteereitä ovatkin haluttava tiheysfunktioestimaatin sileys tai vaikka ytimen häviäminen rajoitetun välin ulkopuolella, mistä voi olla etua laskentatyön vähentämisessä, kun annetulla x summassa $(1/n)\sum_{i=1}^n K_h(x - X_i)$ osa termeistä häviää.

3.7 Korkeamman kertaluvun ytimet

Lauseen 3.12 todistuksessa saatiin ydinestimaattorin harhan neliön integraalille kaava

ydin	$[C(K^*)/C(K)]^{5/4}$
Epanechnikov	1.000
Biweight	0.994
Triweight	0.987
Gauss	0.951
Kolmio	0.986
Tasainen	0.930

Taulukko 3.2: Eri ytimien tehokkuudet optimaalisen Epanechnikovin ytimen suhteen.

$$\int_{-\infty}^{\infty} \text{Bias}^2[\hat{f}_n(x;h)]dx = \int_{-\infty}^{\infty} [\mathbb{E}\hat{f}_n(x;h) - f(x)]^2 dx = \int_{-\infty}^{\infty} [(f * K_h)(x) - f(x)]^2 dx$$

ja tässä

$$(f * K_h)(x) - f(x) = \int_{-\infty}^{\infty} [f(x-y) - f(x)]K_h(y)dy.$$

Todistuksessa hyödynnettiin sitten Taylorin kehitelmää

$$f(x-y) - f(x) = f'(x)(-y) + R_2(x,y)$$

ja ytimen K symmetrisyydestä seuraavaa ehtoa $\int_{-\infty}^{\infty} yK_h(y)dy = 0$, jolloin saatiin

$$(f * K_h)(x) - f(x) = \int_{-\infty}^{\infty} R_2(x,y)K_h(y)dy = h^2(f'' * L_h)(x),$$

missä L on eräs ytimestä K riippuva funktio. Tästä sitten seurasi kertaluokkaa h^4 oleva harhatermi f:n keskimääräiseen integroituun neliölliseen virheeseen (vrt. (3.13).

Jos kuitenkin itseasiassa olisi

$$\int_{-\infty}^{\infty} y^k K_h(y)dy = 0, \qquad k = 1, 2, \ldots, s-1$$

ja $f \in C^s$, saataisiin samoin menetellen, että

$$f(x-y) - f(x) = \sum_{k=1}^{s-1} \frac{1}{k!} f^{(k)}(x)(-y)^k + R_s(x,y)$$

ja

$$(f * K_h)(x) - f(x) = \int_{-\infty}^{\infty} R_s(x,y) K_h(y) dy = h^s (f^{(s)} * L_h^s)(x),$$

missä L^s on eräs funktio. Kun $h \to 0+$, on h^s pienempää kertalukua kuin h^2 jos $s > 2$. Tämä motivoi seuraavan määritelmän.

Määritelmä 3.18 *Olkoon $K \in L^1$ ja $\int_{-\infty}^{\infty} K(y)dy = 1$, $K(x) = K(-x)$, $x \in \mathbb{R}$. Jos $s \geq 2$ on parillinen, sanomme että K on* kertalukua s *oleva ydin mikäli*

(i) $\mu_k(K) = \int_{-\infty}^{\infty} x^k K(x)dx = 0, \qquad k = 1, \ldots, s-1,$

(ii) $\mu_s(K) = \int_{-\infty}^{\infty} x^s K(x)dx \neq 0.$

Huomautus 3.19 Ei-negatiivisen symmetrisen ytimen kertaluku on aina 2. Edelleen, symmetrisellä K on $\mu_k(K) = 0$ aina kun k on pariton.

Yllä esitetyn epämuodollisen päättelyn perusteella on helppo uskoa todeksi (tai itse asiassa myös todistaa) seuraava lauseen 3.12 yleistys.

Lause 3.20 *Olkoon $s \geq 2$ parillinen, $f \in C^s$ tiheysfunktio, $f^{(k)} \in L^2$ kaikilla $k = 0, 1, \ldots, s$ ja olkoon $K \in L^1 \cap L^2$ kertalukua s oleva ydin. Silloin , jos $\lim_{n\to\infty} h_n = 0$, niin*

$$\begin{aligned}\mathrm{MISE}[\hat{f}_n(\cdot; h_n)] &= \mathbb{E} \int_{-\infty}^{\infty} [\hat{f}_n(x; h_n) - f(x)]^2 dx \\ &= \frac{1}{(s!)^2} h_n^{2s} \mu_s(K)^2 R(f^{(s)}) + \frac{1}{nh_n} R(K) + o\left(h_n^{2s} + \frac{1}{nh_n}\right).\end{aligned}$$

Jos vielä $\lim_{n\to\infty} nh_n = \infty$, saamme erityisesti, että $\lim_{n\to\infty} \mathrm{MISE}[\hat{f}_n(\cdot; h_n)] = 0$, eli estimaattori $\hat{f}_n(\cdot; h_n)$ on tarkentuva.

Ydinestimaattorin asymptoottiselle keskimääräiselle integroidulle neliövirheelle saamme nyt kaavan

$$\mathrm{AMISE}[\hat{f}_n(\cdot;h_n)] = \frac{1}{(s!)^2}h_n^{2s}\mu_s(K)^2R(f^{(s)}) + \frac{1}{nh_n}R(K),$$

asymptoottinen optimaalinen silotusparametri on

$$h_n^* = \left[\frac{(s!)^2R(K)}{2s\mu_s(K)^2R(f^{(s)})}\right]^{\frac{1}{2s+1}} \cdot n^{-\frac{1}{2s+1}}$$

ja vastaava asymptoottinen virhe on

$$\begin{aligned}\mathrm{AMISE}[\hat{f}_n(\cdot;h_n^*)] &= \frac{2s+1}{2s}\left[\frac{2s}{(s!)^2}\right]^{\frac{1}{2s+1}}\left[\mu_s(K)^2R(f^{(s)})\right]^{\frac{1}{2s+1}}[R(K)]^{\frac{2s}{2s+1}} \cdot n^{-\frac{2s}{2s+1}} \\ &= \frac{2s+1}{2s}\left[\frac{2s}{(s!)^2}\right]^{\frac{1}{2s+1}}\left[\mu_s(K)^2R(K)^{2s}\right]^{\frac{1}{2s+1}}\left[R(f^{(s)})\right]^{\frac{1}{2s+1}} \cdot n^{-\frac{2s}{2s+1}}\end{aligned}$$

(vrt. (3.26) ja (3.27)). Erityisesti nähdään, että harhan pienenemisen johdosta virheen konvergenssinopeus on parantunut: tapauksessa $s = 2$ (lause 3.12) se oli $n^{-4/5}$ mutta nyt $n^{-\frac{2s}{2s+1}}$. Jos $s \to \infty$ (eli f on hyvin sileä ja K korkeata kertalukua), lähestyy $n^{-\frac{2s}{2s+1}}$ parametrista vauhtia n^{-1}.

Eräs helppo tapa generoida korkeamman kertaluvun ytimiä on lähteä kertalukua k olevasta ytimestä $K_{[k]}$ ja muodostaa kertalukua $k+2$ oleva ydin $K_{[k+2]}$ kaavalla

$$K_{[k+2]}(x) = \alpha_k K_{[k]}(x) + \beta x K'_{[k]}(x),$$

missä α_k ja β_k ovat sopivasti valittuja kertoimia (todistus harjoitustehtävänä).

Todellisuudessa korkean kertaluvun ytimestä saatava hyöty voi kuitenkin olla vaatimaton ja tulla merkittäväksi vasta todella suurella otoskoolla n. Lisäksi on huomattava, että kertalukua $s > 2$ olevat ytimet saavat välttämättä negatiivisia arvoja, joten estimaatti $\hat{f}_n(\cdot;h)$ ei ole enää tiheysfunktio.

Lopuksi todettakoon, että kertalukua s olevalla ytimellä saatava ydinestimaattori on minimax-mielessä optimaalinen: vauhti $n^{-\frac{2s}{2s+1}}$ on paras mahdollinen sopivassa funktioluokassa.

3.8 Silotusparametrin valinta

Seuraava esitys perustuu pitkälti kirjaan [12], josta haluttaessa voi löytää lisää yksityiskohtia.

Tarkasteltavana ongelmana on nyt siis silotusparametrin h valinta käytännössä. Ainakin seuraavia tapoja voidaan ajatella:

- Subjektiivinen valinta: valitaan h, joka "näyttää hyvältä". Tämä lähestymistapa voi olla perusteltu esimerkiksi aineiston alustavassa tarkastelussa ja visualisoinnissa.
- Automaattinen valinta: valitaan h käytettävissä olevan otoksen $X_1, \ldots, X_n$ perusteella pyrkien minimoimaan esmerkiksi MISE. Tämä vaihtoehto voi olla järkevä, kun aineiston jakaumasta ei ole ennakkoon käsitystä, kun ydinestimointi on työkaluna tilastollisessa ohjelmistossa tai myös korkeaulotteisissa tapauksissa, missä subjektiivinen valinta on hankalaa.
- Sovelluskohtaisen kriteerin käyttäminen: esimerkiksi hahmontunnistuksessa h voidaan valita siten, että luokitteluvirhe minimoituu.

Toisaalta, yhden h:n sijasta on toisinaan parempi tarkastella ydinestimaatteja monilla eri silotusparametrin arvoilla. Tällöin estimoitavasta tiheysfunktiosta voi olla mahdollista saada tietoa monissa eri mittakaavoissa, eri "resoluutiotasoilla" (ns. family approach).

3.8.1 Nopeita ja yksinkertaisia menetelmiä

Normaaliskaala (normal scale) Tämä on yksinkertainen peukalosääntö, joka perustuu normaalijakauman optimaaliseen silotukseen. Kun $f \sim N(\mu, \sigma^2)$, voidaan helposti laskea, että $R(f'') = 3/(8\sqrt{\pi}\sigma^5)$. Kaavasta (3.26) saadaan silloin

$$h_n^* = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2(K)^2}\right]^{1/5} \sigma n^{-1/5}.$$

Menetelmän idea on laskea σ:lle estimaatti $\hat{\sigma}$ otoksesta $X_1, \ldots, X_n$ ja ottaa silotusparametriksi

$$\hat{h}_n^{\mathrm{NS}} = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2(K)^2}\right]^{1/5} \hat{\sigma} n^{-1/5}.$$

Voidaan esimerkiksi valita

$$\hat{\sigma} = s \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

missä käytetään tavallista merkintää s otoskeskihajonnalle.

Ylisilotus (oversmoothing) Voidaan osoittaa, että

$$\min\{R(f'') \mid f \text{ tiheys}, \mu_2(f) = \sigma^2\} = \frac{35}{243\sigma^5}\ .$$

Minimi itseasiassa saavutetaan valinnalla $f \sim \text{Beta}(4, 4, -3, 3)$, eli kun

$$f(x) = C(9 - x^2)_+^3, \quad x \in \mathbb{R}$$

(ks. [8, ss. 165-166]). Siten, kun $\mu_2(f) = \int_{-\infty}^{\infty} x^2 f(x)dx = \sigma^2$, saadaan asymptoottisesti optimaaliselle silotusparametrille (3.26) yläraja

$$h_n^* \leq \left[\frac{243R(K)}{35\mu_2(K)^2}\right]^{1/5} \sigma n^{-1/5} = 3\left[\frac{R(K)}{35\mu_2(K)^2}\right]^{1/5} \sigma n^{-1/5}.$$

Idea ylisilotuksessa on estimoida σ otoksesta käyttäen otoskeskihajontaa s ja asettaa silotusparametriksi

$$\hat{h}_n^{\text{OS}} = 3\left[\frac{R(K)}{35\mu_2(K)^2}\right]^{1/5} sn^{-1/5}.$$

Laskemalla nähdään, että kun normaaliskaalan silotusparametrissa $\hat{h}_n^{\text{NS}}$ valitaan $\hat{\sigma} = s$, saadaan $\hat{h}_n^{\text{NS}}/\hat{h}_n^{\text{OS}} \approx 0.93$. Käytännössä $\hat{h}_n^{\text{OS}}$ on hyvä "lähtöarvo" h:lle, jota voi sitten asteittain pienentää halutun silotustason saavuttamiseksi.

3.8.2 Kehittyneempiä menetelmiä

Ristiinvalidointi (cross-validation) Olkoon $\hat{f}_n(x; h) = (1/n)\sum_{i=1}^{n} K_h(x - X_i)$ tiheysfunktion f ydinestimaattori. Kirjoitetaan $\text{MISE}[\hat{f}_n(\cdot; h)]$ muotoon

$$\begin{aligned}
\text{MISE}[\hat{f}_n(\cdot; h)] &= \mathbb{E}\int_{-\infty}^{\infty}[\hat{f}_n(x; h) - f(x)]^2 dx \\
&= \mathbb{E}\left\{\int_{-\infty}^{\infty}[\hat{f}_n(x; h)]^2 dx - 2\int_{-\infty}^{\infty}\hat{f}_n(x; h)f(x)dx + \int_{-\infty}^{\infty} f(x)^2 dx\right\} \\
&= \mathbb{E}\int_{-\infty}^{\infty}[\hat{f}_n(x; h)]^2 dx - 2\mathbb{E}\int_{-\infty}^{\infty}\hat{f}_n(x; h)f(x)dx + \int_{-\infty}^{\infty} f(x)^2 dx. \qquad (3.37)
\end{aligned}$$

Tarkoituksena on pyrkiä minimoimaan MISE$[\hat{f}_n(\cdot;h)]$ h:n suhteen. Ongelmana on se, että kehitelmässä (3.37) termi $2\mathbb{E}\int_{-\infty}^{\infty}\hat{f}_n(x;h)f(x)dx$ riippuu tuntemattomasta tiheysfunktiosta f. Huomaa, että viimeinen termi $\int_{-\infty}^{\infty}f(x)^2dx$ ei kuitenkaan ole ongelma, koska se ei riipu lainkaan h:sta. Nyt huomataan, että

$$\int_{-\infty}^{\infty}\hat{f}_n(x;h)f(x)dx = \mathbb{E}(\hat{f}_n(X;h) \mid X_1,\ldots,X_n),$$

missä $\mathbb{E}(\ \cdot\ \mid X_1,\ldots,X_n)$ tarkoittaa ehdollista odotusarvoa annetulla otoksella $X_1,\ldots,X_n$. Korvaamme tämän ehdollisen odotusarvon *ristiinvalidoidulla* keskiarvolla,

$$\mathbb{E}(\hat{f}_n(X;h) \mid X_1,\ldots,X_n) \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i,n}(X_i;h),$$

missä estimaattori $\hat{f}_{-i,n}(\cdot;h)$ perustuu otokseen, josta on jätetty pois X_i,

$$\hat{f}_{-i,n}(x;h) = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}K_h(x-X_j),\quad x\in\mathbb{R}.$$

Idea ristiinvalidoinnissa on, että ottamalla $\hat{f}_{-i,n}(X_i;h)$ suureen $\hat{f}_n(X_i;h)$ sijaan pyritään poistamaan se optimoinnin kannalta ikävä ongelma, että tavallisesti $\hat{f}_n(X_i;h)\to\infty$, kun $h\to 0+$.

On helppo nähdä, että

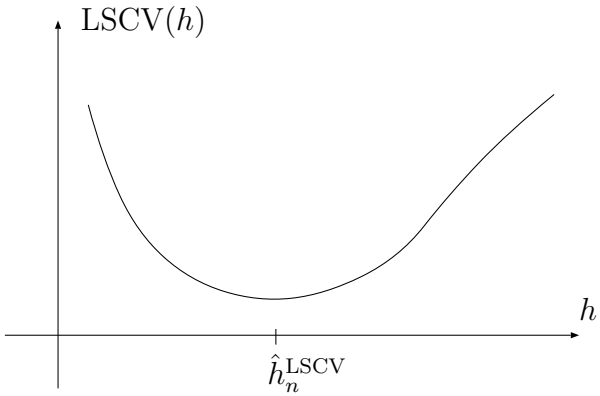
$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i,n}(X_i;h)\right] = \mathbb{E}\int_{-\infty}^{\infty}\hat{f}_n(x;h)f(x)dx.$$

Kun pudotetaan kaavasta (3.37) vielä pois h:sta riippumaton termi $\int_{-\infty}^{\infty}f(x)^2dx$, päädytään minimoimaan suureen MISE$[\hat{f}_n(\cdot;h)]-\int_{-\infty}^{\infty}f(x)^2dx$ harhatonta estimaattoria

$$\mathrm{LSCV}(h) = \int_{-\infty}^{\infty}[\hat{f}_n(x;h)]^2dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i,n}(X_i;h).$$

(LSCV = Least Squares Cross-Validation). Ristiinvalidointimenetelmä valitsee silotusparametriksi tämän minimoijan $\hat{h}_n^{\mathrm{LSCV}}$ (vrt. kuva 3.11).

Tämän menetelmän yksi käytännön ongelma on se, että tilanne ei aina ole kuvan 3.11 kaltainen vaan funktiossa LSCV voi olla useita lokaaleja minimejä. Ratkaisu tällöin usein on valita minimikohdista suurin. Toinen ongelma on se, että silotusparametrilla $\hat{h}_n^{\mathrm{LSCV}}$, joka nyt on satunnaismuuttuja, on varsin suuri varianssi.



Kuva 3.11: Ristiinvalidointimenetelmässä minimoitava funktio LSCV.

Harhainen ristiinvalidointi (biased cross-validation) Tässä menetelmässä ristiinvalidointimenetelmän varianssia pienennetään lisäämällä harhaa. Voidaan osoittaa, että

$$\mathbb{E}R(\hat{f}_n''(\cdot;h)) = R(f'') + \frac{R(K'')}{nh^5} + O(h^2).$$

Ideana on estimoida $R(f'')$:ää suureella

$$\widetilde{R(f'')} = R(\hat{f}_n''(\cdot;h)) - \frac{R(K'')}{nh^5},$$

jolloin

$$\mathbb{E}\widetilde{R(f'')} = R(f'') + O(h^2)$$

ja minimoida sitten

$$\mathrm{BCV}(h) = \frac{1}{4}h^4\mu_2(K)^2\widetilde{R(f'')} + \frac{R(K)}{nh}$$

(vrt. (3.25)). Funktion BCV minimoijalle käytämme merkintää $\hat{h}_n^{\mathrm{BCV}}$.

Esimerkki 3.21 Tämän esimerkin simulointi on myös harjoitustehtävänä. Olkoon $f(\cdot;\mu,\sigma^2) \sim N(\mu,\sigma^2)$ ja tarkastellaan tiheysfunktion

$$f = \frac{3}{4}f(\cdot;0,1) + \frac{1}{4}f(\cdot;3/2,1/9)$$

ydinestimointia otoksen $X_1,\ldots,X_{100} \sim f$ perusteella käyttäen Gaussin ydintä, kun silotusparametriksi valitaan $\hat{h}_n^{\mathrm{LSCV}}$ tai $\hat{h}_n^{\mathrm{BCV}}$. Koe toistettiin 500 kertaa ja saatujen silotusparametrien jakaumien tiheysfunktiot (niiden ydinestimaatit) on esitetty kuvassa 3.12. Kuvaan on pisteviivalla piirretty myös se silotusparametrin arvo

$h^{\mathrm{MISE}} = 0.318$, joka minimoi ydinestimaattorin keskimääräisen integroidun neliöllisen virheen $\mathrm{MISE}[\hat{f}_n(\cdot; h)]$ (3.9). Tämä optimaalinen silotusparametrin arvo voidaan itse asiassa laskea tarkasti esimerkkimme tilanteessa (ks. [12], kappale 2.6). Havaitaan, että ristiinvalidoinnilla saatava silotusparametri $\hat{h}_n^{\mathrm{LSCV}}$ estimoi arvoa h^{MISE} varsin harhattomasti mutta sillä on jonkun verran suurempi varianssi kuin harhaisella ristiinvalidoinnilla saatavalla silotusparametrilla $\hat{h}_n^{\mathrm{BCV}}$. Funktioiden LSCV ja BCV minimoijat valittiin väliltä $[0.05, 2]$. Tämä on tärkeää erityisesti BCV:n kohdalla, koska $\mathrm{BCV}(h) \to 0$, kun $h \to \infty$, mikä nähdään tarkastelemalla harjoitustehtävänä funktiolle BCV johdettavaa eksplisiittistä lauseketta (Gaussin ytimen tapauksessa). ‖

Suora sijoitus (direct plug-in) Edellä arvioitiin suuretta $R(f'') = \int_{-\infty}^{\infty} f''(x)^2 dx$. Jatkossa joudumme yleisemmin estimoimaan suuretta

$$R(f^{(s)}) = \int_{-\infty}^{\infty} f^{(s)}(x)^2 dx, \quad s \geq 2.$$

Kun f on riittävän säännöllinen, saamme osittaisintegroinnilla

$$\int_{-\infty}^{\infty} f^{(s)}(x)^2 dx = \int_{-\infty}^{\infty} f^{(s)}(x) f^{(s)}(x) dx = (-1)^s \int_{-\infty}^{\infty} f^{(2s)}(x) f(x) dx.$$

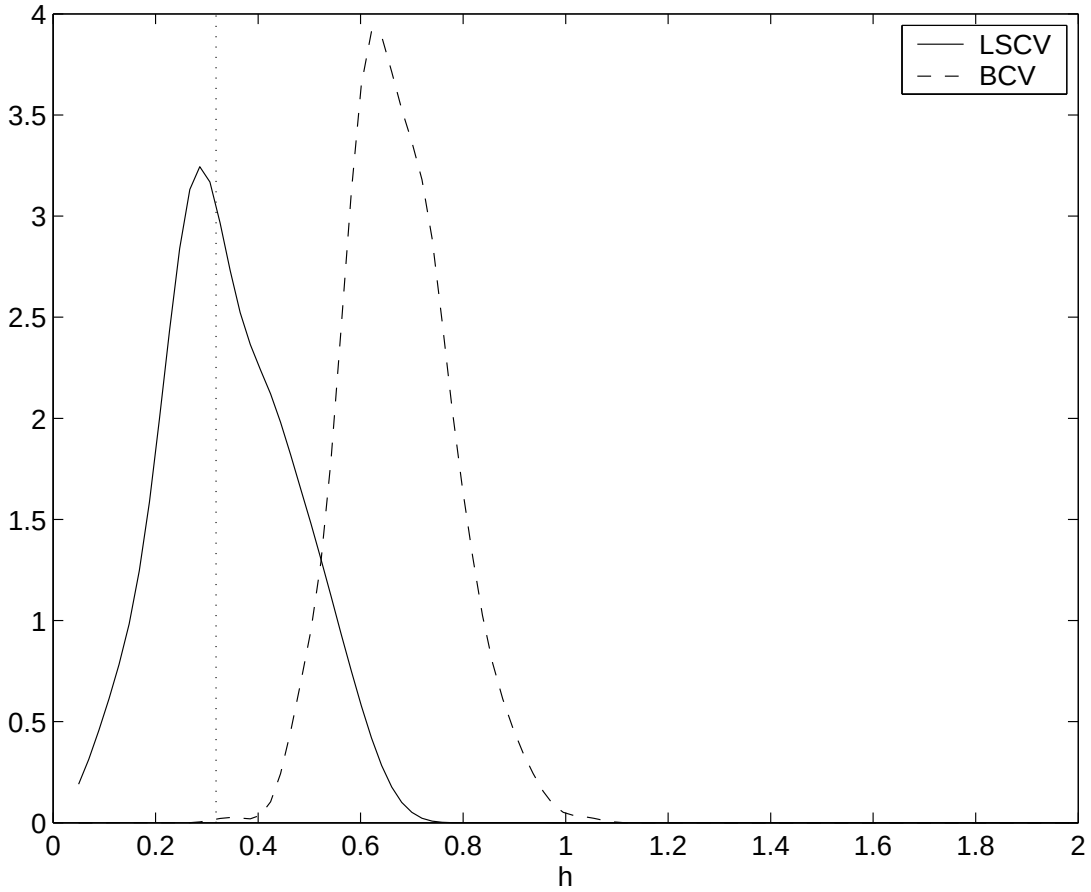
Kun r on parillinen, merkitään

$$\psi_r = \int_{-\infty}^{\infty} f^{(r)}(x) f(x) dx.$$

Suureen ψ_r estimointi voidaan perustaa siihen, että $\psi_r = \mathbb{E} f^{(r)}(X)$. Olkoon L riittävän säännöllinen ydin ja $\lambda > 0$ silotusparametri. Muodostetaan L:n ja λ:n avulla f:n ydinestimaattori ja derivoidaan se,

$$\begin{aligned} \hat{f}_n(x; \lambda) &= \frac{1}{n} \sum_{j=1}^{n} L_\lambda(x - X_j) \\ \hat{f}_n^{(r)}(x; \lambda) &= \frac{1}{n} \sum_{j=1}^{n} (L_\lambda)^{(r)}(x - X_j). \end{aligned}$$

Tästä saamme ψ_r:lle silotusparametrista λ riippuvan estimaattorin



Kuva 3.12: Silotusparametrien $\hat{h}_n^{\mathrm{LSCV}}$ (yhtenäinen viiva) ja $\hat{h}_n^{\mathrm{BCV}}$ (katkoviiva) jakaumat toistokokeessa, jossa kahden normaalijakauman sekoitteesta $f = \frac{3}{4}f(\cdot; 0, 1) + \frac{1}{4}f(\cdot; 3/2, 1/9)$ otettiin $n = 100$ pisteen satunnaisotos 500 kertaa. Pystyssä kulkeva pisteviiva osoittaa ydinestimaattorin keskimääräisen integroidun neliöllisen virheen $\mathrm{MISE}[\hat{f}_n(\cdot; h)]$ minimoivan silotusparametrin arvon.

$$\hat{\psi}_r(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_n^{(r)}(X_i;\lambda) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(L_\lambda)^{(r)}(X_i - X_j).$$

Voidaan osoittaa (ks. [12, luku 3.5]), että AMISE-optimaalinen λ on nyt

$$\lambda_n^* = \left[\frac{k!L^{(r)}(0)}{-\mu_k(L)\psi_{r+k}}\right]^{\frac{1}{r+k+1}} \cdot n^{-\frac{1}{r+k+1}}, \tag{3.38}$$

missä k on L:n kertaluku ja missä oletetaan, että eräät säännöllisyysoletukset ovat voimassa.

Ydinestimaattorille (kun $K \geq 0$) saatiin aikaisemmin

$$h_n^* = \left[\frac{R(K)}{\mu_2(K)^2\psi_4}\right]^{\frac{1}{5}} \cdot n^{-\frac{1}{5}},$$

missä

$$\psi_4 = \int_{-\infty}^{\infty} f^{(4)}(x)f(x)dx = \int_{-\infty}^{\infty} f''(x)^2 dx = R(f'')$$

(vrt. (3.26)). Ajatuksena suoran sijoituksen menetelmässä on korvata ψ_4 estimaattorilla $\hat{\psi}_4(\lambda)$ ja määritellä silotusparametri kaavalla

$$h_n^{\mathrm{DPI}} = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(\lambda)}\right]^{\frac{1}{5}} \cdot n^{-\frac{1}{5}}.$$

Ongelmana on se, että ensin tulisi valita sopiva arvo λ:lle. Tällaiselle apusilotusparametrille käytetään joskus nimeä "pilottisilotusparametri" (engl. pilot smoothing parameter). Voitaisiin esimerkiksi valita $L = K$ ja käyttää (3.38):n arvoa $\lambda = \lambda_n^*$. Mutta silloin joudutaan estimoimaan suuretta ψ_6 ($r = 4$, $k = 2$). Tämän estimointi vaatisi puolestaan ψ_8:n estimoinnin ja niin edelleen!

Ratkaisu tähän "muna ja kana" ongelmaan on estimoida jokin ψ_r nopealla ja helpolla menetelmällä ja muodostaa sitten sen avulla rekursiivisesti ψ_{r-2}, ψ_{r-4}, ja niin edelleen. Voidaan vaikka korvata f tiheysfunktiolla $g \sim N(0, \hat{\sigma}^2)$, missä $\hat{\sigma}^2$ on estimoitu otoksesta $X_1, \ldots, X_n \sim f$ ja ottaa

$$\psi_r = \int_{-\infty}^{\infty} g^{(r)}(x)g(x)dx = \frac{(-1)^{r/2}r!}{(2\hat{\sigma})^{r+1}(r/2)!\sqrt{\pi}}, \tag{3.39}$$

missä jälkimmäinen yhtäsuuruus saadaan suoralla laskulla. Tavallinen lähestymistapa on soveltaa 2-vaiheista rekursiota ($L = K \geq 0$):

1.

$$\hat{\psi}_8^{\mathrm{NS}} = \frac{105}{32\sqrt{\pi}\hat{\sigma}^9}$$

(kaavasta (3.39)) ja missä esimerkiksi $\hat{\sigma} = s$.

2.

$$\lambda_1 = \left[\frac{-2K^{(6)}(0)}{\mu_2(K)\hat{\psi}_8^{\mathrm{NS}}}\right]^{1/9} \cdot n^{-1/9}$$

(vrt. (3.38); $r = 6$, $k = 2$) on silotusparametri $\hat{\psi}_6(\lambda)$:lle.

3.

$$\lambda_2 = \left[\frac{-2K^{(4)}(0)}{\mu_2(K)\hat{\psi}_6(\lambda_1)}\right]^{1/7} \cdot n^{-1/7}$$

(vrt. (3.38); $r = 4$, $k = 2$) on silotusparametri $\hat{\psi}_4(\lambda)$:lle.

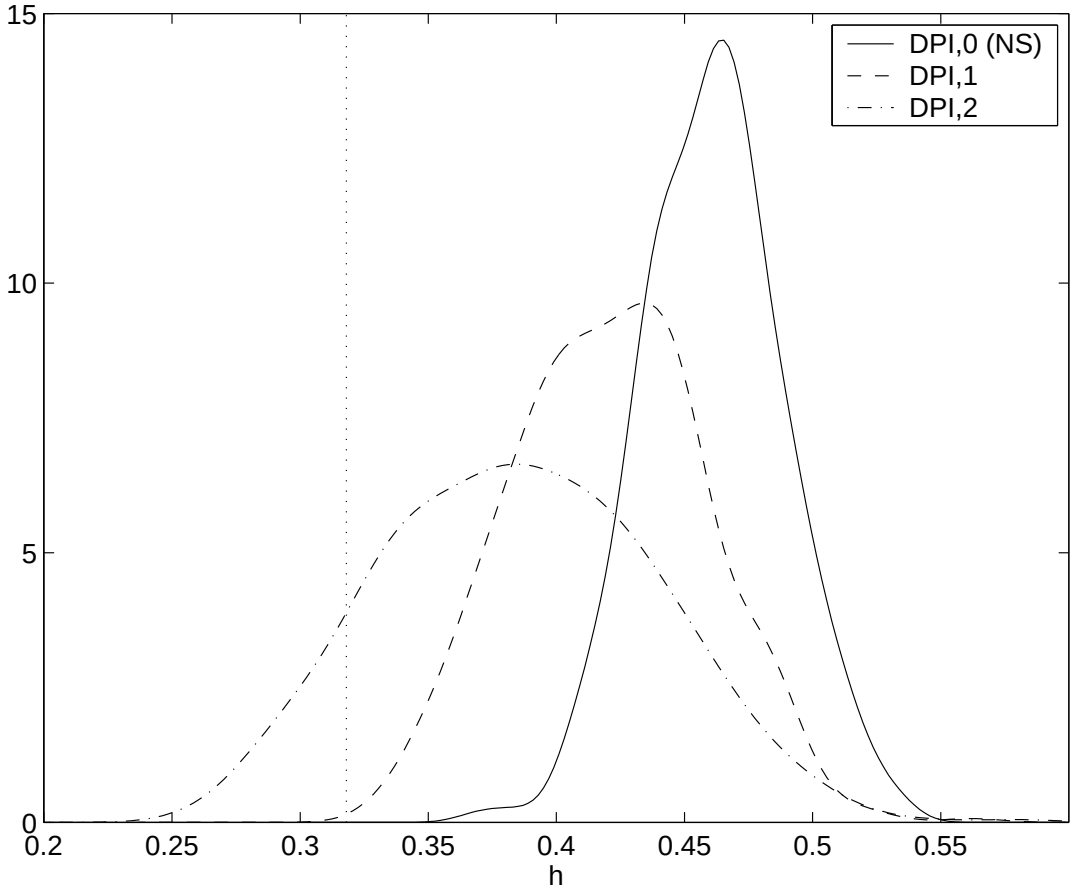
4.

$$\hat{h}_n^{\mathrm{DPI},2} = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(\lambda_2)}\right]^{1/5} \cdot n^{-1/5}$$

on silotusparametri f:n ydinestimaattorille.

Yleisesti, käyttämällä ℓ-vaiheista algoritmia saadaan tällä suoralla sijoitusmenetelmällä silotusparametri $\hat{h}_n^{\mathrm{DPI},\ell}$.

Esimerkki 3.22 Tämän esimerkin käytännön implementointiin palataan harjoitustehtävässä. Tarkastellaan esimerkin 3.21 tiheysfunktiota f ja silotusparametrien $\hat{h}_n^{\mathrm{DPI},\ell}$, $\ell = 0, 1, 2$ jakaumia käytettäessä otosta $X_1, \ldots, X_{100} \sim f$ ja Gaussin ydintä. Koe toistettiin taas 500 kertaa ja saatujen silotusparametrien jakaumien tiheysfunktiot (niiden ydinestimaatit) on esitetty kuvassa 3.13. Teoreettisesti optimaalisen silotusparamterin arvo on osoitettu pisteviivalla ja olemme ottaneet $\hat{h}_n^{\mathrm{DPI},0} = \hat{h}_n^{\mathrm{NS}}$ missä normaalijakauman varianssina on käytetty otoksen varianssia. Havaitaan, että vaiheiden lukumäärän ℓ kasvattaminen pienentää harhaa mutta suurentaa varianssia. Sopivan lukumäärän hakemisessa on siis taas kyse harhan ja varianssin tasapainoittamisesta. ‖



Kuva 3.13: Silotusparametrien $\hat{h}_n^{\mathrm{DPI},\ell}$, $\ell = 0, 1, 2$ jakaumat toistokokeessa, jossa kahden normaalijakauman sekoitteesta $f = \frac{3}{4} f(\cdot; 0, 1) + \frac{1}{4} f(\cdot; 3/2, 1/9)$ otettiin $n = 100$ pisteen satunnaisotos 500 kertaa. Pystyssä kulkeva pisteviiva osoittaa ydinestimaattorin keskimääräisen integroidun neliöllisen virheen $\mathrm{MISE}[\hat{f}_n(\cdot; h)]$ minimoivan silotusparametrin arvon. Olemme ottaneet $\hat{h}_n^{\mathrm{DPI},0} = \hat{h}_n^{\mathrm{NS}}$, missä normaalijakauman varianssina on käytetty otoksen varianssia.

Yhtälön ratkaisu Tässä menetelmässä otetaan $\lambda = \gamma(h)$ sopivalla funktiolla γ ja ratkaistaan numeerisesti yhtälö

$$h = \left[\frac{R(K)}{\mu_2(K)^2 \hat{\psi}_4(\gamma(h))}\right]^{1/5} \cdot n^{-1/5}. \tag{3.40}$$

Voidaan vaikka yrittää valita γ siten, että AMISE-optimaaliselle h_n^* ja λ_n^* (kaavat (3.26) ja (3.38)) pätee $\lambda_n^* \approx \gamma(h_n^*)$. Silloin, jos $\hat{\psi}_4(\lambda_n^*) \approx \psi_4$, on h_n^* likimain (3.40):n ratkaisu. Kaavojen (3.26) ja (3.38) perusteella voidaan päätellä γ:lle sopiva muoto:

$$\lambda_n^* = \left[\frac{2L^{(4)}(0)\mu_2(K)^2}{R(K)\mu_2(L)}\right]^{1/7} (-\psi_4/\psi_6)^{1/7} (h_n^*)^{5/7}.$$

Käytännössä ψ_4 ja ψ_6 joudutaan tietysti estimoimaan ja näin päädytään useampivaiheiseen algoritmiin aivan kuten edellä. Saatua silotuparametria merkitään $\hat{h}_n^{\mathrm{STE}}$ (STE = Solve The Equation).

Silotettu ristiinvalidointi Tarkastellaan approksimaatiota

$$\mathrm{MISE}[\hat{f}_n(\cdot; h)] \approx \int_{-\infty}^{\infty} [(f * K_h)(x) - f(x)]^2 dx + \frac{R(K)}{nh}.$$

Valitaan ydin L ja silotusparametri λ ja minimoidaan h:n suhteen suure

$$\mathrm{SCV}(h) = \int_{-\infty}^{\infty} [(\hat{f}_n(\cdot; \lambda, L) * K_h)(x) - \hat{f}_n(x; \lambda, L)]^2 dx + \frac{R(K)}{nh},$$

(SCV = Smoothed Cross-Validation) missä

$$\hat{f}_n(x; \lambda, L) = \frac{1}{n} \sum_{i=1}^{n} L_\lambda(x - X_i)$$

ja $\lambda = \gamma(h)$ jollain γ. Sopiva funktio γ voidaan taas löytää teoriaan perustuen ja tuloksena on jälleen monivaiheinen algoritmi.

Nimitys "silotettu ristiinvalidointi" johtuu tietystä analogiasta (jota ei perustella tässä) LSCV:n kanssa.

Suorituskyvyn vertailu Oletetaan, että h_n^{MISE} minimoi todellisen (ei siis asymptoottisen) keskimääräisen integroidun neliövirheen $\mathrm{MISE}[\hat{f}_n(\cdot; h)]$. Sopivilla säännöllisyysoletuksilla pätee silloin, että

$$n^{1/10}\left(\hat{h}_n^{\mathrm{LSCV}}/h_n^{\mathrm{MISE}} - 1\right) \to N(0, \sigma^2_{\mathrm{LSCV}}), \text{ kun } n \to \infty,$$

missä konvergennsi tapahtuu jakaumamielessä. Siten $\hat{h}_n^{\mathrm{LSCV}}$:n suhteellinen virhe konvergoi kohti nollaa nopeudella $n^{-1/10}$. Sama pätee BCV:lle.

Edelleen,

$$n^{5/14}\left(\hat{h}_n^{\mathrm{DPI},2}/h_n^{\mathrm{MISE}} - 1\right) \to N(0, \sigma^2_{\mathrm{DPI}}), \text{ kun } n \to \infty.$$

Siten pilottisilotus auttaa huomattavan paljon, ainakin asymptoottisessa mielessä.

Samaa vauhtia konvergoivat silotusparametrin valintamenetelmät saadaan myös STE:stä ja SCV:stä. DPI- ja SCV-menetelmistä on myös versiot, joissa konvergenssinopeus on $n^{-1/2}$, eli parametrista luokkaa.

Käytännön estimointiin voi suositella DPI-, STE- ja SCV-pohjaisia algoritmeja. Sen sijaan LSCV ja BCV eivät ole yhtä hyviä.

3.9 Adaptiivinen ydinestimointi

Esimerkki 3.23 Olkoon $f(\cdot; \mu, \sigma^2) \sim N(\mu, \sigma^2)$ ja tarkastellaan tiheysfunktion

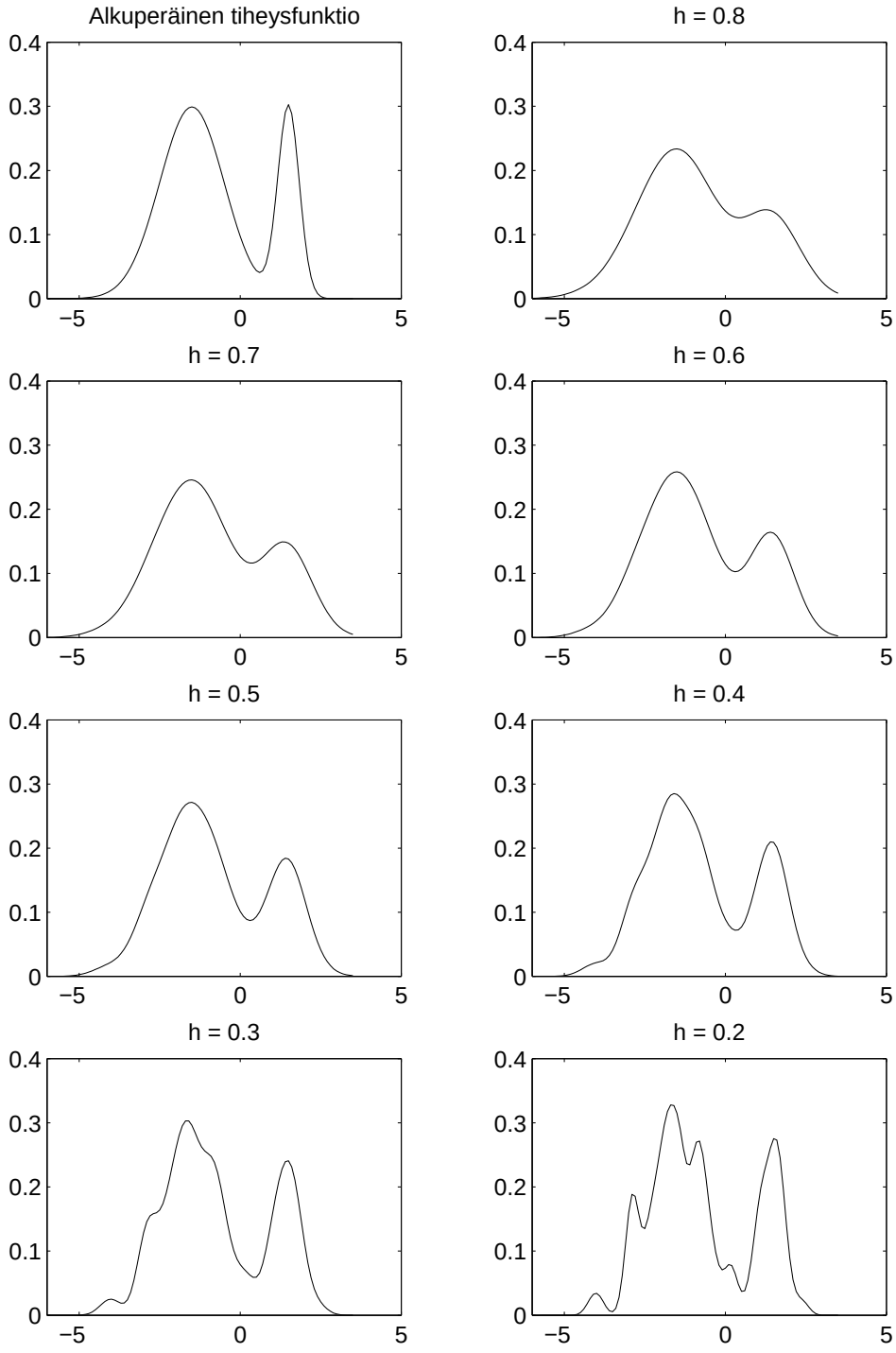
$$f = \frac{3}{4} f(\cdot; -3/2, 1) + \frac{1}{4} f(\cdot; 3/2, 1/9)$$

ydinestimointia otoksen $X_1, \ldots, X_{100} \sim f$ perusteella käyttäen Gaussin ydintä. Kuvassa 3.14 on esitetty eri silotusparametrin arvoilla saatavia estimaatteja. Selvästikään sama h ei ole hyvä kaikilla x: oikean puoleisen piikin hyvä estimointi vaatii pienen silotusparamterin arvon, joka puolestaan johtaa vasemman puoleisen piikin alisilotukseen. ‖

Johtopäätös edellisestä esimerkistä on, että h:n voisi olla syytä muuttua estimointialueen mukana. Voisimme esimerkiksi korvata kiinteän luvun h funktiolla $h : \mathbb{R} \to]0, \infty[$, ja määritellä ydinestimaattorin kaavalla

$$\hat{f}_n(x; h(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x)} K\left(\frac{x - X_i}{h(x)}\right), \quad x \in \mathbb{R}. \tag{3.41}$$

Tällöin siis itseasiassa eri x:n arvoilla saadaan eri estimaattorit.



Kuva 3.14: Eri silotusparametrin arvoilla saatavia ydinestimaatteja kun kahden normaalijakauman sekoitteesta $f = \frac{3}{4}f(\cdot; -3/2, 1) + \frac{1}{4}f(\cdot; 3/2, 1/9)$ on otettu $n = 100$ pisteen satunnaisotos.

Vaihtoehtoinen idea on ottaa oma silotusparametri h_i kullekin otospisteelle. Silloin saamme

$$\hat{f}_n(x;(h_i)) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-X_i}{h_i}\right), \quad x \in \mathbb{R}.$$

Tässä tapauksessa meillä siis on vain yksi ydinestimaattori, joka käyttää silotusparametrijonoa $(h_1, \ldots, h_n)$. Summassa olevan ytimen leveys riippuu otospisteestä ja voimme valita h_i:n siten, että se heijastaa ostospisteiden tiheyttä X_i:n lähellä: iso h_i, kun otospisteitä on harvassa ja pieni h_i, kun niitä on tiheässä.

Kummallakin *adaptiivisella* estimointitavalla on mahdollista parantaa estimointitulosta MISE-mielessä. Eräs jälkimmäisen menetelmän versio on Breimanin-Meiselin-Purcelin estimaattori, joka perustuu ns. k-lähinaapuri etäisyyteen.

Olkoon siis $k \in \{1, \ldots, n-1\}$ ja merkitään $d_k(X_i)$:llä etäisyyttä X_i:stä sen k. lähinaapuriin. Toisin sanoen, joukko $\{|X_i - X_j| \mid j = 1, \ldots, n,\ j \neq i\}$ permutoidaan,

$$|X_i - X_{(1)}| \leq |X_i - X_{(2)}| \leq \cdots \leq |X_i - X_{(n-1)}|$$

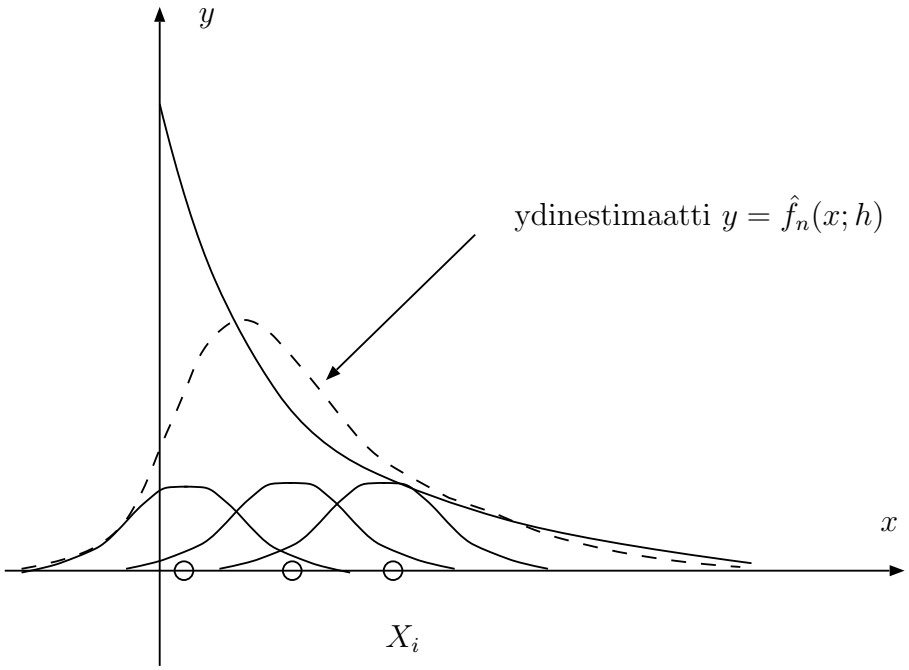
ja merkitään $d_k(X_i) = |X_i - X_{(k)}|$. Sitten otetaan $h_i = hd_k(X_i)$, $i = 1, \ldots, n$, missä $h > 0$ on kaikille otospisteille yhteinen globaali silotusparametri.

3.10 Reunat

Esimerkki 3.24 Olkoon $f \sim \mathrm{Exp}(1)$ eksponenttijakautuneen (odotusarvolla 1) satunnaismuuttujan tiheysfunktio. Siis,

$$f(x) = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Jos nyt käytetään jotain tavanomaista ydintä, käy helposti niin, että $\hat{f}_n(x;h) > 0$ vaikka $x < 0$. Tämä voi olla epätoivottavaa, koska itse tiheysfuktio tällöin kuitenkin häviää (ks. kuva 3.15). Eräänä ratkaisuna onkin ottaa käyttöön erityiset reunaytimet. ‖



Kuva 3.15: Jos exponenttijakauman tiheyttä (yhtenäinen viiva) estimoidaan tavanomaiseen ytimeen perustuvalla ydinestimaattorilla (katkoviiva), voi tulos olla epätoivottava, koska myös arvoille $x < 0$ tulee estimaatissa positiivinen todennäköisyys.

Reunaytimet (boundary kernels) Reunaytimistä löytyy tarvittaessa lisätietoa kirjan [8] luvusta 6.2.3.5. Oletetaan, että f on tiheys ja $f(x) = 0$, kun $x < 0$. Ideana reunaydinmenetelmässä on annetulla silotusparametrilla h käyttää erikoisydintä välillä $[0, h]$ oleville otospisteille ja tavallista ydintä välillä $]h, \infty[$ oleville otospisteille. Yksi vaihtoehto erikoisytimeksi on ns. leijuva (engl. floating) reunaydin:

$$K^c(x) = \frac{3}{4}[(c+1) - \frac{5}{4}(1+2c)(x-c)^2][x-(c+2)]^2 1_{[c,c+2]}(x).$$

Tässä parametri $c \in [-1, 0]$ määrää ytimen muodon. Erityisesti K^c häviää välin $[c, c+2]$ ulkopuolella (ks. kuva 3.16). Tämä ydin on luvussa 3.6 määritellyn biweight-ytimen $K(x) = (15/16)(1-x^2)_+^2$ modifikaatio. Reunalla modifioitu ydinestimaattori saadaan nyt määrittelemällä $c_i = -X_i/h$, $i = 1, \ldots, n$ ja

$$\hat{f}_n(x; h) = \frac{1}{n}\left\{ \sum_{\substack{i=1 \\ X_i \in [0,h]}}^{n} (K^{c_i})_h(x - X_i) + \sum_{\substack{i=1 \\ X_i \notin [0,h]}}^{n} K_h(x - X_i) \right\}.$$

Ydin K^c on määritelty siten, että momenttiehdot

$$\int_{-\infty}^{\infty} K^c(x)dx = 1 \text{ ja } \int_{-\infty}^{\infty} xK^c(x)dx = 0$$

ovat voimassa; jälkimmäinen ehto takaa harhan pienenemisen kuten Lauseessa 3.12.

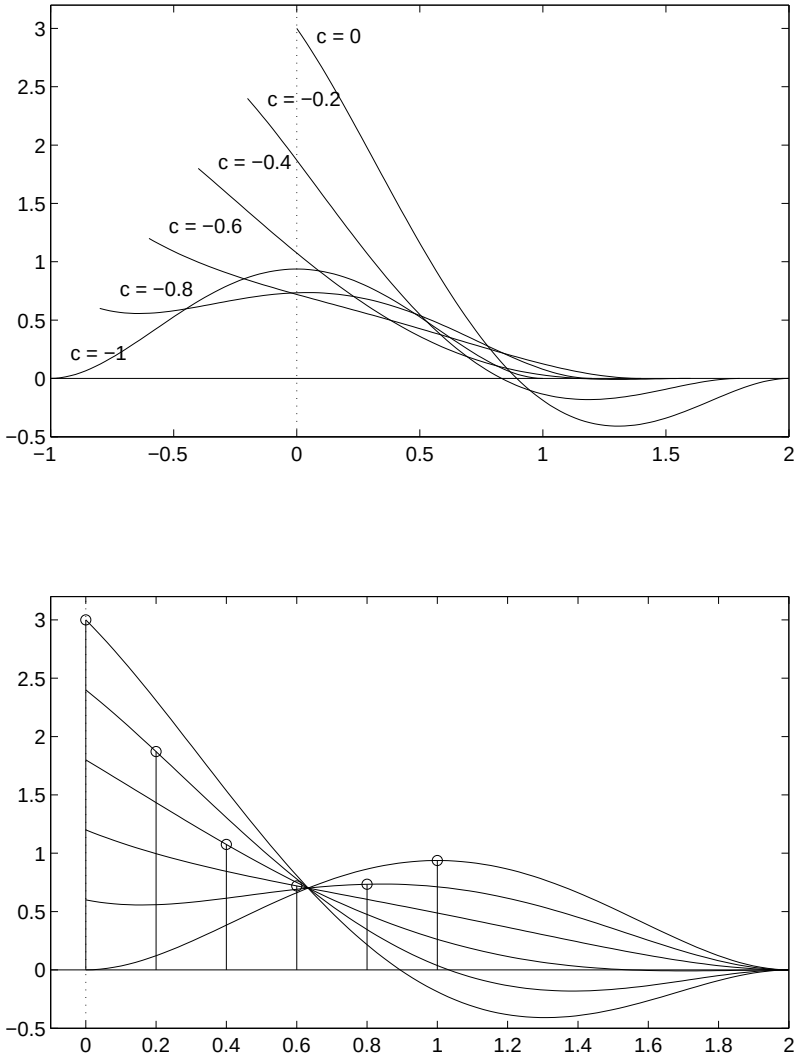
Näin määritelty ydin toimii hyvin esimerkiksi juuri tiheysfunktiolle $f \sim \text{Exp}(1)$ mutta ei esimerkiksi jos $f \sim \text{Beta}(3, 9, 0, 1)$ (ks. kuva 3.18). Parempi reunaydin onkin nyt ns. "nollaydin"

$$K^{0,c}(x) = \frac{15}{6}(x-c)^2(2+c-x)^2[(7c^2+14c+8) - 7x(c+1)]1_{[c,c+2]}(x),$$

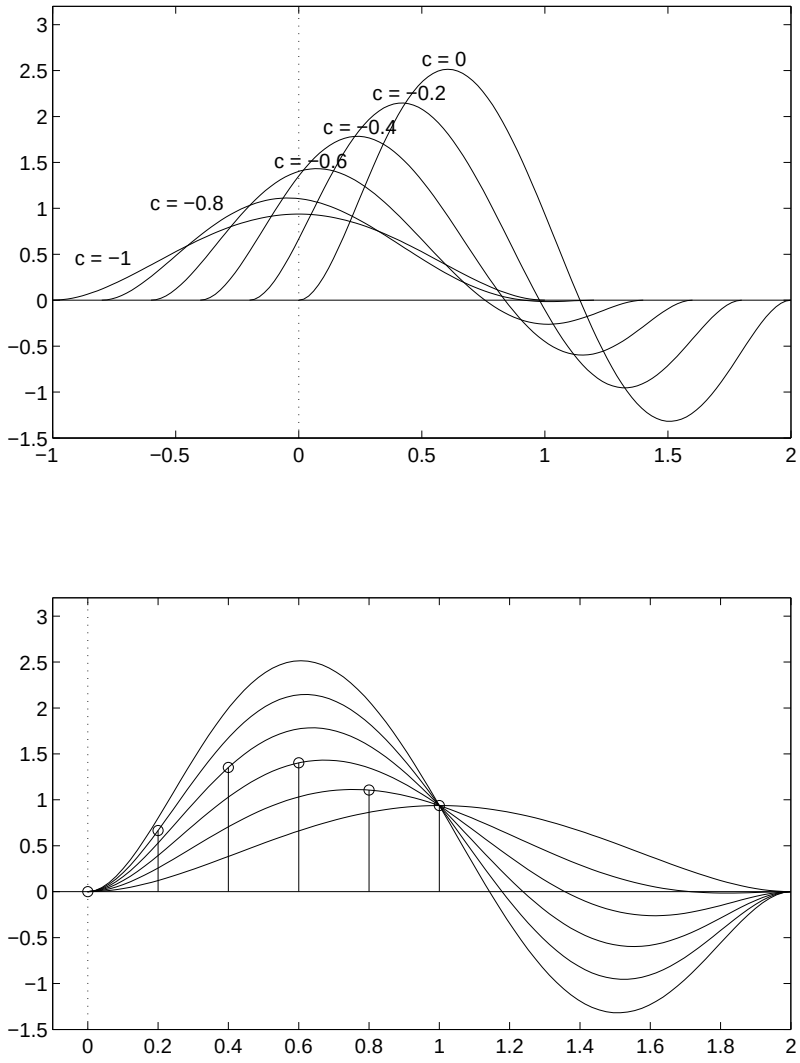
koska sille pätee $(K^{0,c})(0) = (K^{0,c})'(0) = 0$, eli ydin toteuttaa samat reunaehdot kuin estimoitava tiheysfunktio (kuva 3.17). Tätä ydintä käytettäessä $\text{Beta}(3, 9, 0, 1)$ tiheyden estimointi onnistuu origon lähellä jo varsin hyvin (kuva 3.18). Reunaytimiä valittaessa on siis oltava tarkkana erilaisten haitallisten efektien varalta.

Reunaheijastus Oletetaan taas, että $f(x) = 0$, kun $x < 0$. Lisätään otokseen $X_1, \ldots, X_n \sim f$ pisteet $-X_1, \ldots, -X_n$ ja määritellään

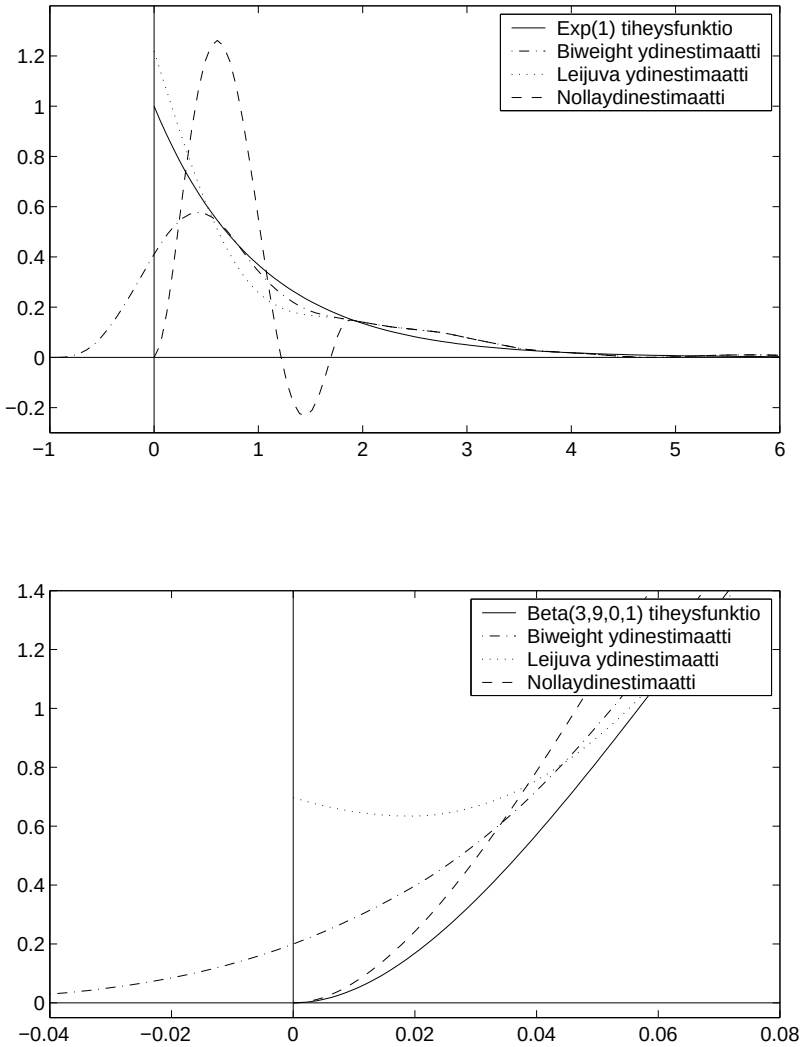
$$\tilde{f}_n(x; h) = \frac{1}{2n}\left\{ \sum_{i=1}^{n} K_h(x - X_i) + \sum_{i=1}^{n} K_h(x + X_i) \right\}.$$



Kuva 3.16: Biweight-ytimestä modifioimalla saatu leijuva ydin. Ylemmässä kuvassa leijuvan ytimen K^c kuvaajat kun $c = 0, -0.2, -0.4, -0.6, -0.8$ ja -1. Alemmassa kuvassa vastaavasti datapisteisiin $X_i = 0, 0.2, 0.4, 0.6, 0.8$ ja 1 liittyvät ytimet oikeille paikoilleen asetettuna kun niitä käytetään ydinestimaatissa ja $h = 1$. Datapisteet ja niitä vastaavat ytimet on merkitty pystysuorilla viivoilla ja niiden päissä olevilla pienillä ympyröillä.



Kuva 3.17: Biweight-ytimestä modifioimalla saatu nollaydin. Esitys kuten kuvassa 3.16.



Kuva 3.18: Esimerkkejä ytimen valinnan vaikutuksesta estimointialueen reunalla. Ylemmässä kuvassa on Exp(1)-jakauman tiheysfunktiota estimoitu tavallisella biweight-ytimellä, leijuvalla ytimellä ja nollaytimellä kun $n = 100$ ja $h = 0.93$. Alemmassa kuvassa on vastaavasti estimoitu Beta$(3, 9, 0, 1)$-jakauman tiheysfunktiota, kun $n = 100$ ja $h = 0.11$. Kuva näyttää tuloksen tästä jälkimmäisestä tilanteesta origon läheisyydessä.

Lopulliseksi estimaattoriksi otetaan sitten

$$\hat{f}_n(x;h) = \begin{cases} 2\tilde{f}_n(x;h), & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Silotusparametria h valittaessa on otoskokona oltava n eikä $2n$.

Lopuksi todettakoon, että laskuharjoituksissa tarkasteltu transformaatiomenetelmä on joskus myös käyttökelpoinen reunaongelmissa.

3.11 Ydinestimointi avaruudessa $\mathbb{R}^d$

3.11.1 Dimensiokirous

Termi "dimensiokirous" (curse of dimensionality) on peräisin R.E. Bellmanilta vuodelta 1961. Hän käytti sitä kuvaamaan kombinatorisen optimoinnin vaikeuden eksponentiaalista kasvua avaruuden dimension funktiona.

Olkoon $d \geq 1$ ja ajatellaan avaruuden $\mathbb{R}^d$ pistettä x pystyvektorina

$$x = [x_1, \ldots, x_n]^T = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d.$$

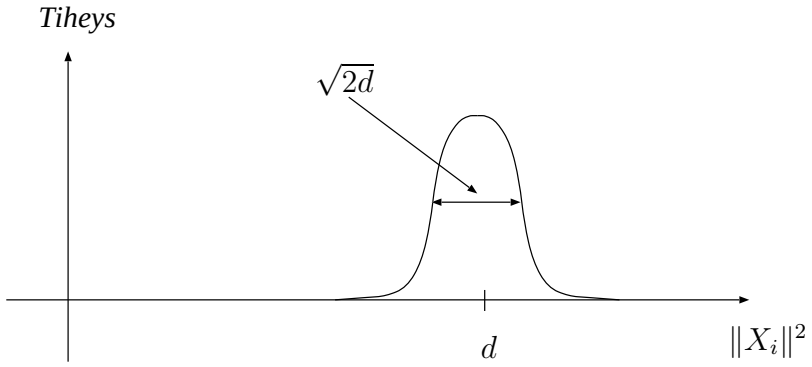
Tässä T siis merkitsee transpoosia. Käytämme x:n normille merkintää $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$.

Olkoon nyt $X : \Omega \to \mathbb{R}^d$ satunnaisvektori. Olkoon $\mu \in \mathbb{R}^d$ ja $\Sigma \in \mathbb{R}^{d \times d}$ symmetrinen positiivisesti definiitti $d \times d$ matriisi (eli $\Sigma^T = \Sigma$ ja $x^T \Sigma x > 0$ kaikilla $x \in \mathbb{R}^d \setminus \{0\}$). Sanomme, että X:llä on *multinormaalijakauma odotusarvolla μ ja kovarianssimatriisilla* Σ, jos X:n jakaumalla on tiheysfunktio

$$f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^d,$$

missä $|\Sigma|$ on matriisin Σ determinantti. Merkitsemme tällöin $X \sim N(\mu, \Sigma)$. Voidaan osoittaa, että $\mathbb{E}X = \mu$ ja $\mathbb{E}[(X-\mu)(X-\mu)^T] = \Sigma$.

Olkoon nyt $f \sim N(0, I_d)$, missä I_d on $d \times d$ identtinen matriisi, eli f on standardi d-ulotteisen normaalijakauman tiheysfunktio. Miten sijaitsevat tällöin avaruudessa $\mathbb{R}^d$ i.i.d. otoksen $X_1, \ldots, X_n \sim f$ pisteet X_i? Tilastotieteestä tiedetään, että jos



Kuva 3.19: Kun $X_i \sim N(0, I_d)$, on satunnaismuuttujalla $\|X_i\|^2$ jakauma χ^2_d.

$Z_1, \ldots, Z_d \sim N(0,1)$, on satunnaismuuttujan $\sum_{i=1}^d Z_i^2$ jakauma χ^2_d ("khi toiseen d:llä vapausasteella"). Koska nyt

$$X_i = [X_{i1}, \ldots, X_{id}]^T, \qquad X_{ij} \sim N(0,1),$$

ja $X_{i1}, \ldots, X_{id}$ ovat riippumattomia, on siis

$$\|X_i\|^2 = \sum_{j=1}^{d} X_{ij}^2 \sim \chi^2_d.$$

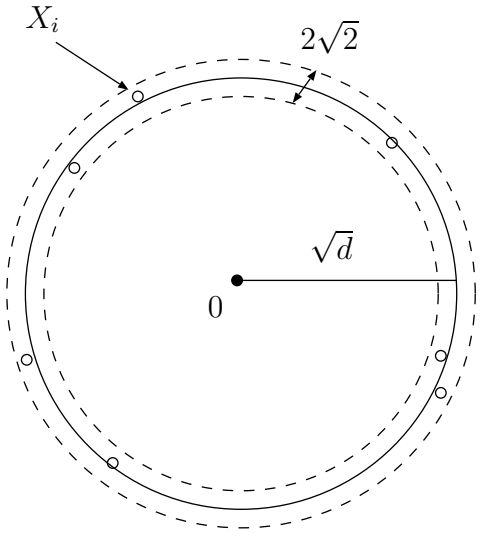
Jakauman χ^2_d odotusarvo on d ja varianssi $2d$, joten

$$\mathbb{E}\|X_i\|^2 = d, \qquad \mathrm{Var}\|X_i\|^2 = 2d$$

(ks. kuva 3.19).

Voidaan siis ajatella, että suurella todennäköisyydellä $|\|X_i\|^2 - d| \leq 2\sqrt{2d}$, mistä pienellä laskulla voidaan päätellä, että kun d on suuri, sijaitsevat otospisteet X_i origo-keskisessä pallonkuoressa, jonka paksuus on $2\sqrt{2}$ ja etäisyys origosta $\sqrt{d}$! Lisäksi $2\sqrt{2}/\sqrt{d} \to 0$, kun $d \to \infty$, joten otospisteiden keskittyminen tämän kuoren sisään on sitä selvempää mitä korkeammasta dimensiosta on kysymys. Tilannetta havainnolistaa kuva 3.20.

Korkeissa dimensiossa kaikki data näyttää siis olevan jakauman *hännillä*. Koska hännillä tiheys on alhainen, ei ole yllätys, että korkeadimensioisissa avaruuksissa ostospisteet ovat hyvin *harvassa*. Tämä voidaan todeta myös suoralla laskulla seuraavasti. Kun $i \neq j$, on $X_i - X_j \sim N(0, 2I_d)$, joten $(1/\sqrt{2})(X_i - X_j) \sim N(0, I_d)$. Tästä seuraa, että



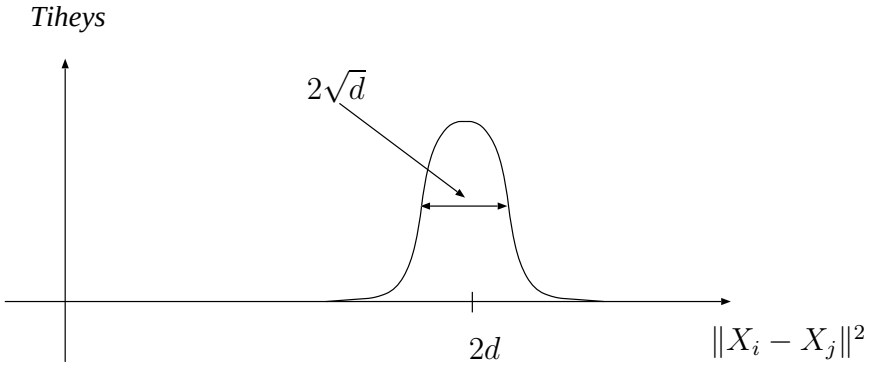
Kuva 3.20: Kun dimensio d on korkea, otospisteet $X_i \sim N(0, I_d)$ sijaitsevat origokeskisessä $\sqrt{d}$-säteisessä pallon kuoressa, jonka paksuus on $\sqrt{d}$.

$$\|(1/\sqrt{2})(X_i - X_j)\|^2 \sim \chi^2_d,$$

joten

$$\mathbb{E}\|X_i - X_j\|^2 = 2d, \qquad \mathrm{Var}\|X_i - X_j\|^2 = 4d$$

(ks. kuva 3.21). Erityisesti näemme, että pisteiden X_i ja X_j välinen keskimääränen etäisyys $\mathbb{E}\|X_i - X_j\|$ kasvaa rajatta, kun $d \to \infty$.
Tämä otoksen harvuus merkitsee sitä, että käytettäessä ydinestimointia korkeissa dimensioissa (ks. seuraava luku) täytyy silotusparamterin h olla suuri varianssin pitämiseksi kurissa. Suuri silotusparametri kuitenkin kasvattaa harhaa (vrt. kuva 3.9), jolloin estimoitavan tiheysfunktion yksityiskohtia ei saada näkyviin. Paras ratkaisu olisikin, jos otoskoko olisi hyvin suuri mutta se ei valitettavsti ole käytännössä useinkaan mahdollista. Tämä "tyhjän avaruuden ongelma" onkin tilastotieteen versio dimensiokirouksesta.



Kuva 3.21: Kun satunnaismuuttujilla X_i on standardi multinormaalijakauma ja $i \neq j$, on satunnaismuuttujalla $(1/\sqrt{2})\|X_i - X_j\|^2$ jakauma χ^2_d, joten $\mathbb{E}\|X_i - X_j\|^2 = 2d$ ja $\mathrm{Var}\|X_i - X_j\|^2 = 4d$.

3.11.2 Ydinestimaattori

Olkoon $f : \mathbb{R}^d \to [0, \infty[$ tiheys, $X_1, \ldots, X_n \sim f$, $K : \mathbb{R}^d \to \mathbb{R}$ ydin, $\int_{\mathbb{R}^d} K(x)dx = 1$ ja H symmetrinen positiivisesti definiitti $d \times d$ matriisi, ns. silotusmatriisi. Määrittelemme tällöin f:n ydinestimaattorin kaavalla

$$\hat{f}_n(x; H) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i), \quad x \in \mathbb{R}^d,$$

missä

$$K_H(x) = \frac{1}{|H^{1/2}|}K(H^{-1/2}x), \quad x \in \mathbb{R}^d.$$

Tässä matriisit $H^{\pm 1/2}$ määritellään kaavalla $H^{\pm 1/2} = U\Lambda^{\pm 1/2}U^T$, kun $H = U\Lambda U^T$ on H:n ominaisarvohajotelma. Siis, U on ortogonaalinen matriisi, jonka sarakkeet muodostavat H:n ortonormaalit ominaisvektorit ja

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$$

on lävistäjämatriisi, jonka lävistäjäalkioina ovat H:n ominaisarvot ja merkitsemme

$$\Lambda^{\pm 1/2} = \mathrm{diag}(\lambda_1^{\pm 1/2}, \ldots, \lambda_d^{\pm 1/2}).$$

Esimerkiksi d-ulotteinen, matriisilla H skaalattu Gaussin ydin on

$$K_H(x) = \frac{1}{(2\pi)^{d/2}|H^{1/2}|} e^{-\frac{1}{2}x^T H^{-1} x}, \quad x \in \mathbb{R}^d$$

eli $K_H \sim N(0, H)$.

Ydinestimointi toimii d-ulotteisessa avaruudessa samaan tapaan kuin 1-ulotteisessa tilanteessa. Otospisteiden X_i kohdalle asetetaan sopivasti skaalatut ydinfunktiot, ytimet summataan yhteen ja tulos normeerataan jakamalla n:llä. Kuvassa 3.22 tätä on havainnollistettu tapauksessa $d = 2$.

Onko käytetyn ytimen muodolla vaikutusta estimointiin? Olkoon $U = [u_1, \ldots, u_d]$, missä u_j on H:n ominaisarvoon λ_j liittyvä ortonormaali ominaisvektori, $Hu_j = \lambda_j u_j$, $j = 1, \ldots, d$. Gaussin ytimen K_H tasa-arvopinnat saadaan ehdosta $x^T H^{-1} x = c$, missä $c > 0$ on vakio. Jos $x = \sum_{j=1}^d \xi_j u_j$, saadaan tästä ehto $\sum_{j=1}^d \xi_j^2/\lambda_j = c$, joka määrää ellipsoidin avaruudessa $\mathbb{R}^d$ (ks. kuva 3.23). Voidaan osoittaa, että joskus estimointitulos paranee, jos f:n ja K_H:n kovarianssit (muodot) ovat lähellä toisiaan (vrt. kuva 3.24).

Olkoon sitten k 1-ulotteinen ydin. Kaksi suoraviivaista tapaa konstruoida k:sta lähtien d-ulotteinen ydin on muodostaa tuloydin tai pallosymmetrinen ydin:

$$\begin{aligned} &K(x) = \textstyle\prod_{i=1}^d k(x_i), \quad x = [x_1, \ldots, x_d]^T \in \mathbb{R}^d \qquad \text{"tuloydin"} \\ &K(x) = C_{k,d} k(\|x\|), \quad x \in \mathbb{R}^d \qquad \text{"pallosymmetrinen ydin"}. \end{aligned}$$

Yksinkertaisimmillaan silotusmatriisin H voi valita diagonaaliseksi:

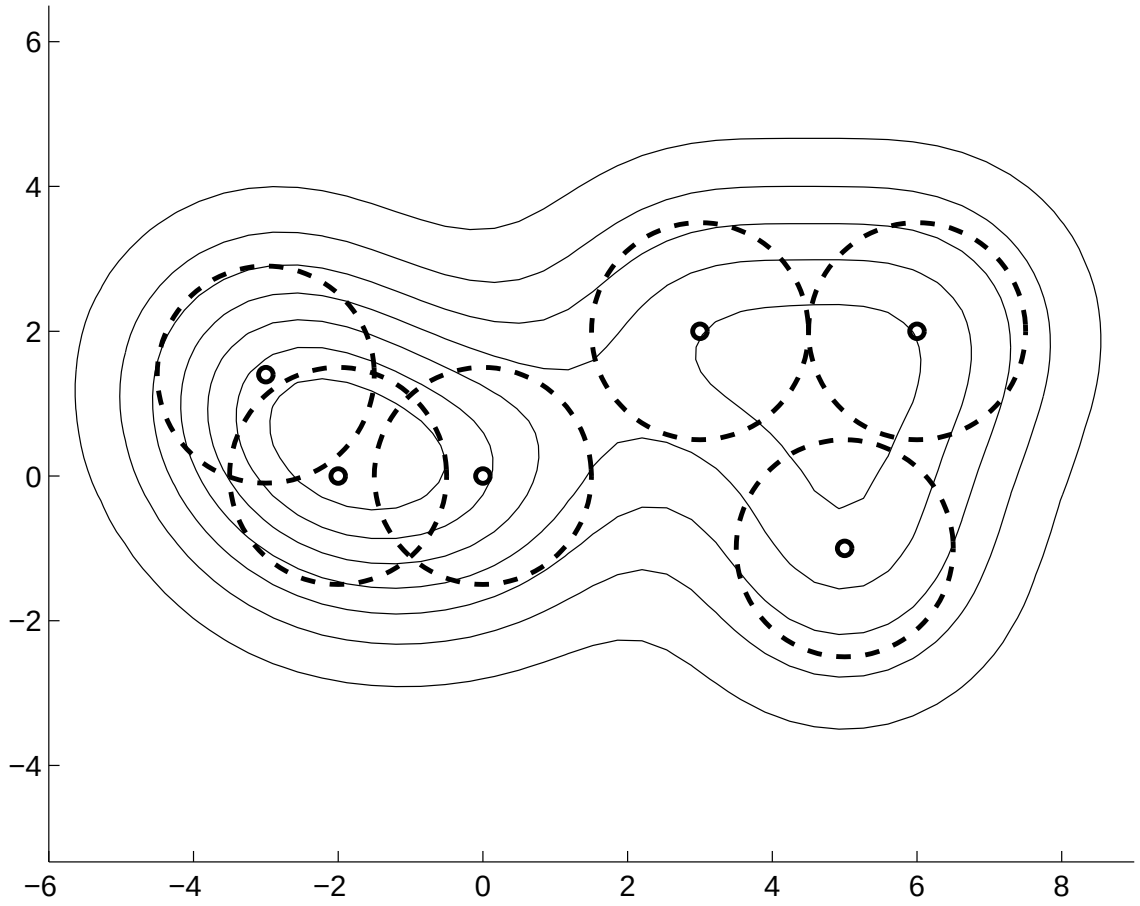
$$H = \operatorname{diag}(h_1^2, \ldots, h_d^2), \quad \hat{f}_n(x; H) = \frac{1}{n h_1 \cdots h_d} \sum_{i=1}^n K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right),$$

tai vielä yksinkertaisemmin

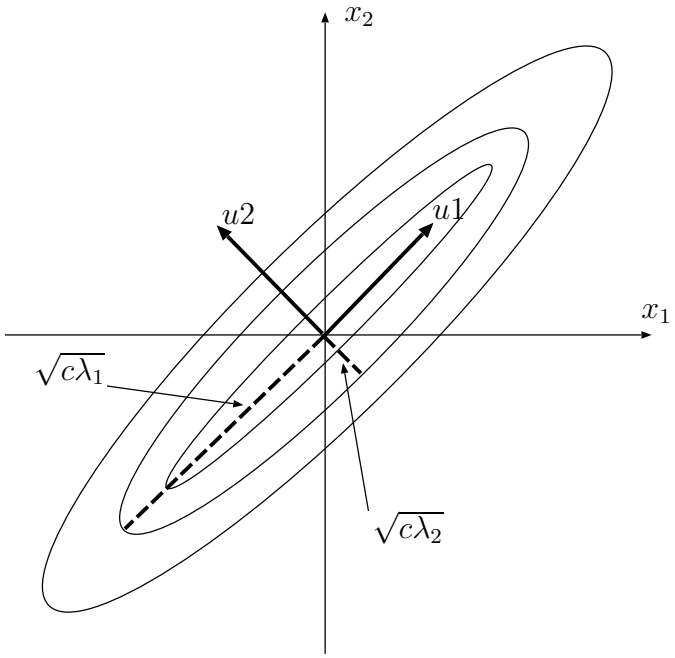
$$H = \operatorname{diag}(h^2, \ldots, h^2) = h^2 I_d, \quad \hat{f}_n(x; H) = \frac{1}{n h^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

Esimerkki 3.25 Tarkastellaan ydinestimointia avaruudessa $\mathbb{R}^2$. Olkoon $f(\cdot, \mu, \Sigma)$ jakauman $N(\mu, \Sigma)$ tiheysfunktio ja tarkastellaan sekoitetta

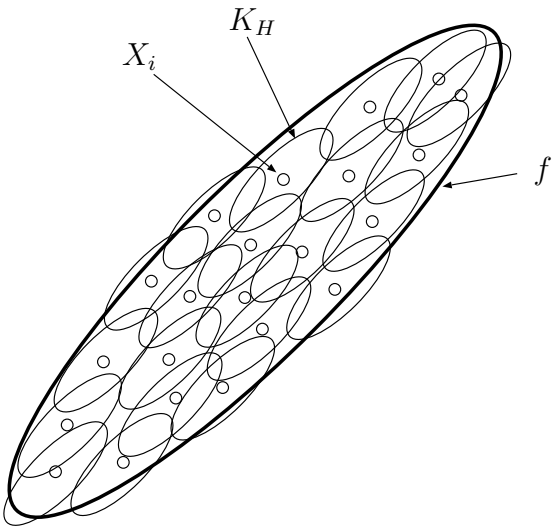
$$f = \frac{1}{3} f(\cdot, \mu_1, I_2) + \frac{1}{3} f(\cdot, \mu_2, I_2) + \frac{1}{3} f(\cdot, \mu_3, I_2),$$



Kuva 3.22: Ydinestimointi avaruudessa $\mathbb{R}^2$. Kuvaan on piirretty 6 otospistettä ja ydinestimaatin tasa-arvokäyrät kun käytetään standardia Gaussin ydintä ja silotusmatriisia $1.5^2 I_2$. Ytimiä on havainnollistettu katkoviivoilla piirretyillä ympyröillä, joiden säde on 1.5.



Kuva 3.23: Gaussin ytimen tasa-arvopinnat ovat ellipsoideja. Ellipsoidin pääakselien suunnat määrätyvät kovarianssimatriisin ominaisvektoreista u_i ja puoliakselien pituuksien suhteet riippuvat vastaavista ominaisarvoista λ_i.



Kuva 3.24: Joskus estimointi on tehokkaampaa, jos estimoitavan tiheyden f ja ydinestimoinnissa käytetyn ytimen K_H muodot ovat lähellä toisiaan.

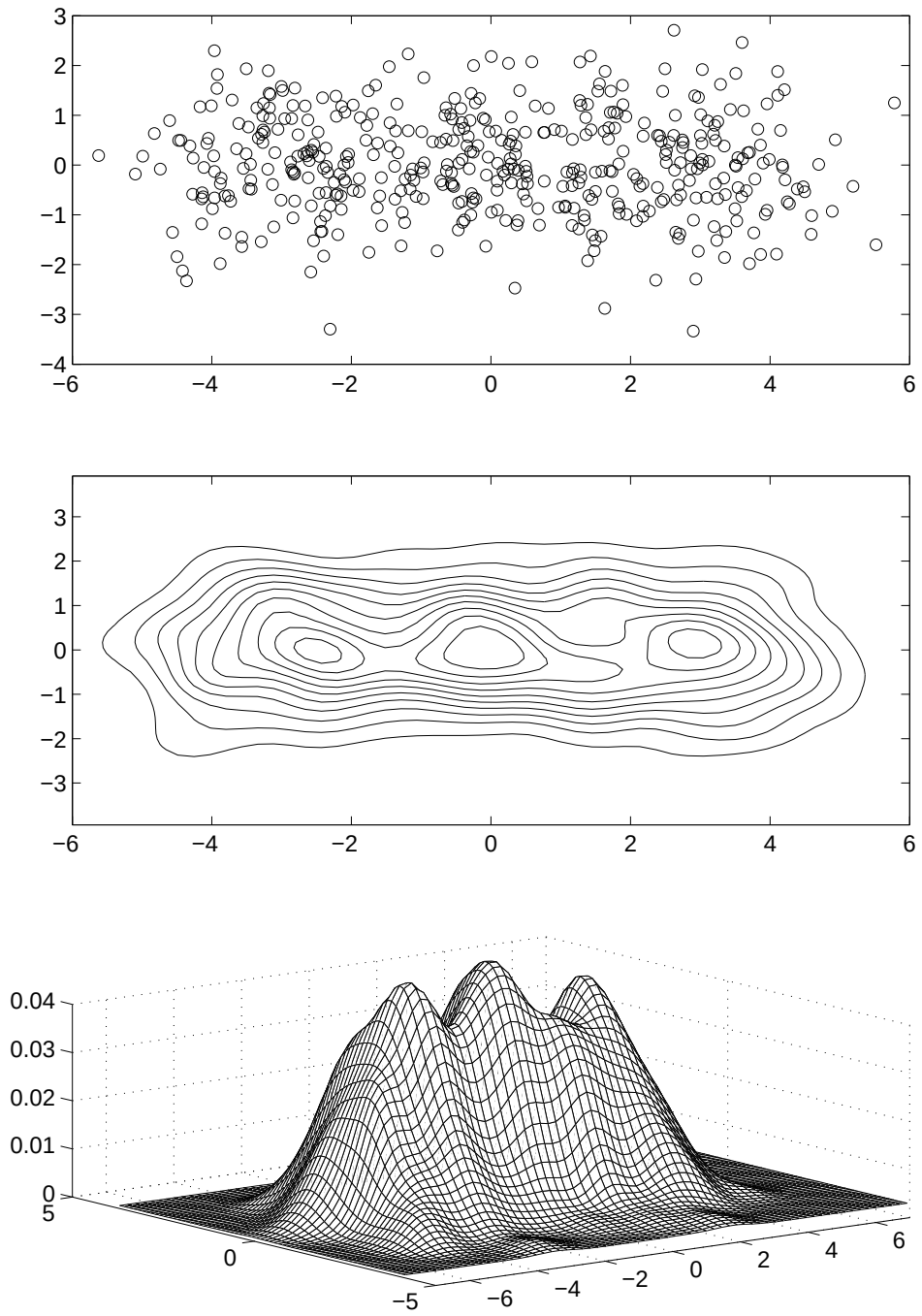
missä $\mu_1 = [-3\ \ 0]^T$, $\mu_2 = [0\ \ 0]^T$ ja $\mu_3 = [3\ \ 0]$. Kuvassa 3.25 on esitetty Gaussin ytimellä saatava ydinestimaatti kokoa $n = 450$ olevalle otokselle tiheysfunktiosta f kun silotusmatriisina on

$$H = \begin{bmatrix} 0.5^2 & 0 \\ 0 & 0.4^2 \end{bmatrix}.$$

Ylimmän kuvan hajontakuviosta ei selvästikään pysty päättelemään f:n kolmihuippuista muotoa. Sensijaan keskimmäisen kuvan tasa-arvokäyrästön ja alimman kuvan pinnan muodon perusteella kolmihuippuisuus on ilmeistä. Esimerkki osoittaa ydinestimaatin potentiaalisen hyödyn aineiston havainnollistamisessa. ‖

Lauseen 3.12 todistus voidaan helposti yleistää tapaukseen $H = h^2 I_d$, jolloin saadaan seuraava tulos.

Lause 3.26 *Olkoon $f \in C^2(\mathbb{R}^d)$ tiheysfunktio ja oletetaan, että $f, \partial f/\partial x_i$, $\partial^2 f/\partial x_i \partial x_j \in L^2(\mathbb{R}^d)$, $i, j = 1, \ldots, d$. Olkoon $K \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} K(x)dx = 1$, $K(x) \geq 0$, $K(-x) = K(x)$ kaikilla $x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} x_i x_j K(x)dx = 0$ kun $i \neq j$ ja*



Kuva 3.25: Esimerkki tiheysfunktion ydinestimoinnista dimensiossa $d = 2$. Ylimmässä kuvassa on käytetty aineisto, $n = 450$ pisteen otos kolmen normaalijakauman sekoitteesta. Keskimmäisessä kuvassa ydinestimaatti on esitetty tasa-arvokäyrinä ja alimmassa kuvassa avaruuden $\mathbb{R}^3$ pintana.

$\int_{\mathbb{R}^d} x_i x_j K(x)dx = \mu_2(K) < \infty$, *kun* $i = j$. *Silloin, jos* $h_n \to 0$ *ja* $H_n = h_n^2 I_d$, *on*

$$\begin{aligned}\mathrm{MISE}[\hat{f}_n(\cdot; H_n)] &= \mathbb{E}\int_{\mathbb{R}^d}[\hat{f}_n(x; H_n) - f(x)]^2 dx \\ &= \frac{1}{4}h_n^4 \mu_2(K)^2 R(\nabla^2 f) + \frac{R(K)}{nh_n^d} + o\left(h_n^4 + \frac{1}{nh_n^d}\right),\end{aligned}$$

missä $R(\nabla^2 f) = \int_{\mathbb{R}^d}[\sum_{j=1}^d \partial^2 f/\partial x_j^2]^2 dx$. *Jos siis vielä* $\lim_{n\to\infty} nh_n^d = \infty$, *saamme erityisesti, että* $\lim_{n\to\infty} \mathrm{MISE}[\hat{f}_n(\cdot; H_n)] = 0$, *eli estimaattori* $\hat{f}_n(\cdot; h_n)$ *on tarkentuva.*

Tavalliseen tapaan saamme lauseen 3.26 perusteella, että asymptoottisen virheen

$$\mathrm{AMISE}[\hat{f}_n(\cdot; H_n)] = \frac{1}{4}h_n^4 \mu_2(K)^2 R(\nabla^2 f) + \frac{R(K)}{nh_n^d}$$

minimoiva silotusmatriisi on $H_n^* = (h_n^*)^2 I_d$, missä

$$h_n^* = \left[\frac{dR(K)}{\mu_2(K)^2 R(\nabla^2 f)}\right]^{\frac{1}{d+4}} \cdot n^{-\frac{1}{d+4}}$$

ja vastaava optimaalinen asymptoottinen virhe on

$$\mathrm{AMISE}[\hat{f}_n(\cdot; H_n^*)] = \left[\mu_2(K)^2 [dR(K)]^{4/d}\right]^{\frac{d}{d+4}} \left[R(\nabla^2 f)\right]^{\frac{d}{d+4}} \cdot n^{-\frac{4}{d+4}}.$$

Jos erityisesti $f \sim N(\mu, \sigma^2 I_d)$ ja $K \sim N(0, I_d)$, saadaan (HT)

$$h_n^* = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \sigma \cdot n^{-\frac{1}{d+4}}$$

ja

$$\mathrm{AMISE}[\hat{f}_n(\cdot; H_n^*)] = (4\pi)^{-d/2}\left(\frac{d+4}{2}\right)\left(\frac{d+2}{4}\right)^{\frac{d}{d+4}} \sigma^{-d} \cdot n^{-\frac{4}{d+4}}.$$

Dimensiokirous näkyy ylläolevissa kaavoissa selvästi: eksponentti $-4/(d+4)$ lähenee nollaa, kun $d \to \infty$, joten estimointivirheen konvergenssivauhti hiipuu suurilla d. Tämä on selvä ero parametriseen estimointiin, jossa konvergenssinopeus suurimman uskottavuuden estimoinnissa on $1/n$ dimensiosta riippumatta.

Esimerkki 3.27 Seuraava dimensiokirousta valaiseva numeerinen esimerkki on peräisin kirjasta [9] ja siinä tutkitaan tiheysfunktion arvon $f(0)$ estimointia, kun d kasvaa. Olkoon $f, K \sim N(0, I_d)$, ja $\hat{f}_n(\cdot; H_n)$ ydinestimaattori, jossa $H_n = h_n^2 I_d$

d	1	2	3	4	5	6	7	8	9	10
n	4	19	67	223	768	2 790	10 700	43 700	187 000	842 000

Taulukko 3.3: Eri dimensioissa d tarvittava otoskoko tarkkuuden (3.42) saavuttamiseksi kun estimoidaan standardi normaalijakauman tiheysfunktiota origossa.

ja h_n on valittu siten, että virhe $\mathbb{E}[\hat{f}_n(0; H_n) - f(0)]^2$ minimoituu. Haluamme niin tarkkaa estimointia, että

$$\frac{\mathbb{E}[\hat{f}_n(0; H_n) - f(0)]^2}{f(0)^2} < 0.1. \tag{3.42}$$

Numeerisella laskulla saadaan selville kussakin dimensiossa tarvittava otoskoko, jolla tämä suhteellisen virheen tarkkuusvaatimus toteutuu. Tulokset on annettu talukossa 3.3 ja niistä nähdään, että estimointitehtävä vaikeutuu erittäin nopeasti dimension kasvaessa.

‖

Tekemällä sopivat oletukset f:stä, K:sta ja jonosta (H_n) voidaan todistaa lauseen 3.26 yleisempi versio (ks.[12, kaava (4.9)]),

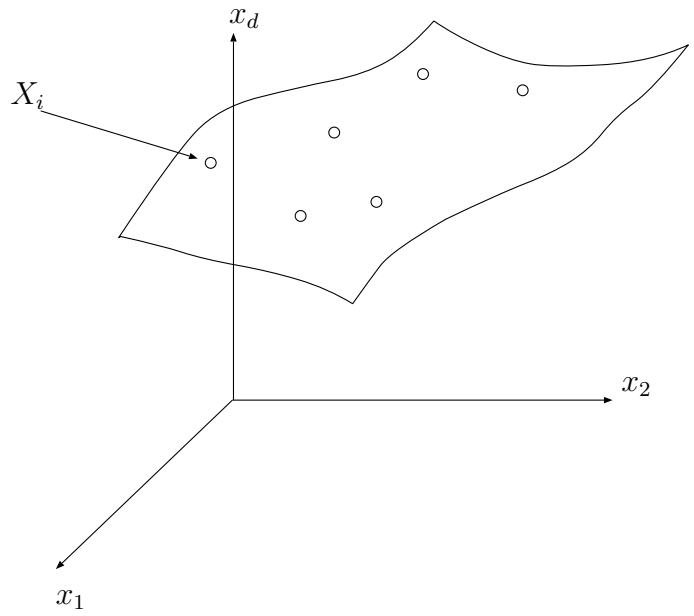
$$\mathrm{AMISE}[\hat{f}_n(\cdot; H_n)] = \frac{1}{4}\mu_2(K)^2 R(\mathrm{tr}[H_n G_f]) + \frac{R(K)}{n|H_n^{1/2}|},$$

missä $G_f = [\partial^2 f/\partial x_i x_j]$ on f:n Hessin matriisi ja tr tarkoittaa matriisin jälkeä, $\mathrm{tr}B = \sum_{i=1}^d b_{ii}$, kun $B = [b_{ij}]$.

Jos $f \sim N(\mu, \Sigma)$ ja $K \sim N(0, I_d)$, voidaan osoittaa ([12, harj. 4.7]), että ydinestimaattorin AMISE-optimaalinen silotusmatriisi ja vastaava asymptoottinen virhe ovat

$$H_n^* = \left(\frac{4}{d+2}\right)^{\frac{2}{d+4}} \Sigma n^{-\frac{1}{d+4}},$$

$$\mathrm{AMISE}[\hat{f}_n(\cdot; H_n^*)] = (4\pi)^{-d/2} \left(\frac{d+4}{4}\right) \left(\frac{d+2}{4}\right)^{\frac{d}{d+4}} |\Sigma|^{-1/2} \cdot n^{-\frac{4}{d+4}}.$$



Kuva 3.26: Todellinen tilastollinen aineisto usein likimääräisesti sijaitsee alkuperäisen avaruuden $\mathbb{R}^d$ melko matalaulotteisella alimonistolla.

Tämän johdosta voidaan pitää jossain määrin perusteltuna myös mielivaltaisen tuntemattoman f:n tapauksessa ottaa $H = h^2\widehat{\Sigma}$, missä $\widehat{\Sigma}$ on otoksesta $X_1, \ldots, X_n \sim f$ estimoitu kovarianssimatriisi (vrt. myös kuva 3.24). Ongelmana on sitten enää löytää sopiva $h > 0$.

D. Scottin [8] mielestä edellä kuvattua suurempi dimensiokirousonglema on se, että otospisteistö $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ usein sijaitsee avaruuden $\mathbb{R}^d$ melko matalaulotteisella monistolla (so. "pinnalla", vrt. kuva 3.26). Scottin mukaan todellisen tilastollisen aineiston efektiivinen dimensio on harvoin suurempi kuin 5. Usein estimoinnissa edetäänkin kahdessa vaiheessa:

1. Datan dimensiota pienennetään sopivalla kuvauksella $\varphi : \mathbb{R}^d \to \mathbb{R}^{d'}$, $d' << d$. Yksi esimerkki on tilastollisessa analyysissä usein käytetty ns. pääkomponenttimuunnos.

2. Suoritetaan tiheysfunktion estimointi aineistolle $\varphi(X_1), \ldots, \varphi(X_n)$ avaruudessa $\mathbb{R}^{d'}$.

3.12 Eräitä muita menetelmiä

3.12.1 Lähinaapuriestimaattori

Olkoon $f : \mathbb{R} \to [0, \infty[$ tiheysfunktio, $X_1, \ldots, X_n \sim f$ i.i.d. otos ja $x \in \mathbb{R}$. Olkoon $(X_{(i)})$ otospisteiden permutaatio siten että

$$|x - X_{(1)}| \leq |x - X_{(2)}| \leq \cdots \leq |x - X_{(n)}|$$

ja $k \in \{1, \ldots, n\}$. Olkoon edelleen $d_k(x) = |x - X_{(k)}|$ x:n ja sen k:nnen lähinaapurin välinen etäisyys ja määritellään

$$\hat{f}_n(x; k) = \frac{k/n}{2d_k(x)}, \quad x \in \mathbb{R}.$$

Sanomme, että $\hat{f}_n(\cdot; k)$ on tiheysfunktion f *k-lähinaapuriestimaattori*. Tämä estimaattori on varsin luonnollinen ainakin kun $1 << k << n$, koska tällöin

$$\frac{k}{n} \approx \mathbb{P}(X \in [x - d_k(x), x + d_k(x)]) = \int_{x-d_k(x)}^{x+d_k(x)} f(t)dt \approx 2d_k(x)f(x),$$

joten

$$f(x) \approx \frac{k/n}{2d_k(x)}.$$

Silotusparametrin rooli on nyt k:lla. Voidaan osoittaa, että jos $k_n \to \infty$ ja $k_n/n \to 0$, niin

$$\hat{f}_n(x; k) \to f(x), \quad \text{kun } n \to \infty,$$

stokastisen konvergenssin mielessä.

Jos $K = (1/2)1_{[-1,1]}$, voidaan myös kirjoittaa

$$\hat{f}_n(x; k) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{d_k(x)} K\left(\frac{x - X_i}{d_k(x)}\right), \tag{3.43}$$

kunhan vaan $|x - X_{(i)}| < |x - X_{(i+1)}|$, $i = 1, \ldots, n$. Nähdään siis, että $\hat{f}_n(\cdot; k)$ on itseasiassa muotoa (3.41) oleva adaptiivinen ydinestimaattori. *Yleistetty k-lähinaapuriestimaattori* saadaan korvaamalla $K = (1/2)1_{[-1,1]}$ mielivaltaisella ytimellä.

Huonoja puolia k-lähinaapuriestimoinnissa ovat:

- $\hat{f}_n(\cdot; k)$ ei ole tiheysfunktio. Itseasiassa $\int_{-\infty}^{\infty} \hat{f}_n(x; k)dx = \infty$ (HT). Siten $\hat{f}_n(x; k)$ on mielekäs estimaattori vain yksittäisissä pisteissä.

- $\hat{f}_n(\cdot; k)$ ei ole sileä (HT).

Lopuksi todettakoon, että yo. tarkastelut yleistyvät suoraviivaisesti avaruuteen $\mathbb{R}^d$.

3.12.2 Otogonaalisarjaestimaattori

Tätä estimaattoria käsiteltiin jo aikaisemmin luvussa 2.5. Tarkastellaan funktioavaruutta $L^2(D)$, missä esimerkiksi $D = [0, 1]$ tai $D = \mathbb{R}$. Olkoon edelleen $(\varphi_k)_{k\in\mathbb{N}}$ ortonormaali kanta $L^2(D)$:ssä. Jos $f \in L^2(D)$ on tiheysfunktio ja $X_1, \ldots, X_n \sim f$ on i.i.d. otos, voidaan f esittää sarjana

$$f = \sum_{k=1}^{\infty} a_k \varphi_k, \quad a_k = \mathbb{E}\varphi_k(X_1)$$

ja luonteva estimaattoori on

$$\hat{f}_n(\cdot; m) = \sum_{k=1}^{m} \hat{a}_{kn} \varphi_k, \quad \hat{a}_{kn} = \frac{1}{n} \sum_{i=1}^{n} \varphi_k(X_i).$$

Integroitu neliövirhe voidaan tavalliseen tapaan hajottaa harha- ja varianssiosiinsa,

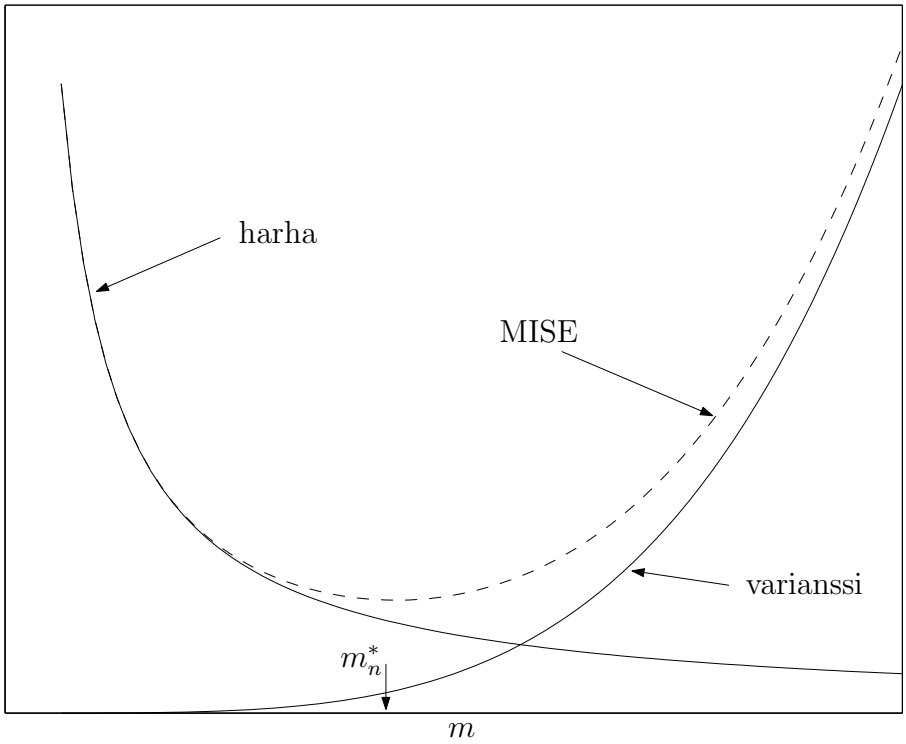
$$\mathrm{MISE}[\hat{f}_n(\cdot; m)] = \int_{-\infty}^{\infty} [\mathbb{E}\hat{f}_n(x; m) - f(x)]^2 dx + \int_{-\infty}^{\infty} \mathrm{Var}[\hat{f}_n(x; m)]dx.$$

Koska $\mathbb{E}\hat{a}_{kn} = a_k$, on $\mathbb{E}\hat{f}_n(x; m) = \sum_{k=1}^{m} a_k \varphi_k(x)$, joten harhalle saamme

$$\int_{-\infty}^{\infty} [\mathbb{E}\hat{f}_n(x; m) - f(x)]^2 dx = \int_{-\infty}^{\infty} \left[\sum_{k=m+1}^{\infty} a_k \varphi_k(x) \right]^2 dx = \sum_{k=m+1}^{\infty} a_k^2.$$

Varianssille puolestaan pätee

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{Var}[\hat{f}_n(x; m)]dx &= \int_{-\infty}^{\infty} \mathbb{E}[\hat{f}_n(x; m) - \mathbb{E}\hat{f}_n(x; m)]^2 dx \\ &= \mathbb{E} \int_{-\infty}^{\infty} [\sum_{k=1}^{m} (\hat{a}_{kn} - a_k)\varphi_k(x)]^2 dx = \mathbb{E} \sum_{k=1}^{m} (\hat{a}_{kn} - a_k)^2 = \sum_{k=1}^{m} \mathbb{E}(\hat{a}_{kn} - a_k)^2. \end{aligned}$$



Kuva 3.27: Integroitu neliöllinen virhe MISE (katkoviiva) saadaan laskemalla yhteen harhan ja varianssin osuudet (yhtenäiset viivat). Optimaalinen silotusparametrin arvo m_n^* minimoi kokonaisvirheen.

Näin saamme harha-varianssihajotelman

$$\mathrm{MISE}[\hat{f}_n(\cdot;m)] = \sum_{k=m+1}^{\infty} a_k^2 + \sum_{k=1}^{m} \mathbb{E}(\hat{a}_{kn} - a_k)^2. \tag{3.44}$$

Silotusparametrin rooli on nyt m:llä ja optimaalisen tuloksen saavuttamiseksi m:n tulee riippua sopivasti otoskoosta n, eli $m = m_n$ (vrt. kuva 3.27).

Lause 3.28 *Olkoon* $|\varphi_k(x)| \leq C$ *kaikilla* $x \in D, k \in \mathbb{N}$. *Oletetaan, että* $m_n \to \infty$, $m_n/n \to 0$, *kun* $n \to \infty$ *ja olkoon*

$$\hat{f}_n(\cdot;m_n) = \sum_{k=1}^{m_n} \hat{a}_{kn}\varphi_k.$$

Silloin $\mathrm{MISE}[\hat{f}_n(\cdot;m)] \to 0$, *kun* $n \to \infty$.

Todistus: Kaavassa (3.44) harhatermi $\sum_{k=m_n+1}^{\infty} a_k^2$ lähestyy nollaa, kun $n \to \infty$, koska $m_n \to \infty$ ja $\sum_{k=1}^{\infty} a_k^2 = \int_{-\infty}^{\infty} f(x)^2 dx < \infty$.

Varianssitermille saadaan

$$\begin{aligned}
\mathbb{E}(\hat{a}_{kn} - a_k)^2 &= \mathrm{Var}[\hat{a}_{kn}] \\
&= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \varphi_k(X_i)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[\varphi_k(X_i)] \\
&= \frac{1}{n}\mathrm{Var}[\varphi_k(X_1)] \\
&= \frac{1}{n}\left\{\mathbb{E}[\varphi_k(X_1)^2] - [\mathbb{E}\varphi_k(X_1)]^2\right\}.
\end{aligned}$$

Tässä $\mathbb{E}\varphi_k(X_1) = a_k$ ja saamme

$$\sum_{k=1}^{m_n} \mathbb{E}(\hat{a}_{kn} - a_k)^2 = \frac{1}{n}\sum_{k=1}^{m_n} \mathbb{E}\varphi_k(X_1)^2 - \frac{1}{n}\sum_{k=1}^{m_n} a_k^2 \leq \frac{1}{n}\sum_{k=1}^{m_n} \mathbb{E}\varphi_k(X_1)^2.$$

Toisaalta

$$\frac{1}{n}\sum_{k=1}^{m_n} \mathbb{E}\varphi_k(X_1)^2 = \frac{1}{n}\sum_{k=1}^{m_n} \int_{-\infty}^{\infty} \varphi_k(x)^2 f(x) dx \leq C^2 \frac{m_n}{n} \to 0,$$

kun $n \to \infty$. □

Yleisempi ortogonaalisarjaestimaattori on

$$\hat{f}_n(\cdot;(\lambda_{kn})) = \sum_{k=1}^{\infty} \lambda_{kn}\hat{a}_{kn}\varphi_k,$$

missä $(\lambda_{kn}) = (\lambda_{kn})_{k\in\mathbb{N}}$ on otoskoosta n riippuva painojono. Jos

$$\lambda_{kn} = \begin{cases} 1, & k \le m_n, \\ 0, & k > m_n, \end{cases}$$

saadaan aikaisempi estimaattori. Nyt voidaan kuitenkin päästä parempaan tulokseen valitsemalla $(\lambda_{kn})_{k\in\mathbb{N}}$ s.e. MISE minimoituu. Saadaan

$$\begin{aligned} \mathrm{MISE}[\hat{f}_n(\cdot;(\lambda_{kn}))] &= \mathbb{E}\int_{-\infty}^{\infty}\left\{\sum_{k=1}^{\infty}[\lambda_{kn}\hat{a}_{kn} - a_k]\varphi_k(x)\right\}^2 dx \\ &= \sum_{k=1}^{\infty}\mathbb{E}\left[\lambda_{kn}\hat{a}_{kn} - a_k\right]^2 \\ &= \sum_{k=1}^{\infty}\left\{[\lambda_{kn}a_k - a_k]^2 + \mathrm{Var}[\lambda_{kn}\hat{a}_{kn}]\right\}. \end{aligned}$$

Tässä

$$\mathrm{Var}[\lambda_{kn}\hat{a}_{kn}] = \lambda_{kn}^2\mathrm{Var}[\hat{a}_{kn}] = \frac{\lambda_{kn}^2}{n}\mathrm{Var}[\varphi_k(X_1)],$$

joten

$$\mathrm{MISE}[\hat{f}_n(\cdot;(\lambda_{kn}))] = \sum_{k=1}^{\infty}\left\{a_k^2(1-\lambda_{kn})^2 + \frac{\lambda_{kn}^2}{n}\mathrm{Var}[\varphi_k(X_1)]\right\}.$$

Funktion $g(\lambda) = a(1-\lambda)^2 + b\lambda^2$, $a, b > 0$, minimi on kohdassa $\lambda = a/(a+b)$ ja $g(a/(a+b)) = ab/(a+b)$. Siten optimaalinen painojono on

$$\lambda_{kn}^* = \frac{a_k^2}{a_k^2 + \frac{1}{n}\mathrm{Var}[\varphi_k(X_1)]},$$

ja

$$\mathrm{MISE}[\hat{f}_n(\cdot;(\lambda_{kn}^*))] = \sum_{k=1}^{\infty}\left\{\frac{a_k^2\mathrm{Var}[\varphi_k(X_1)]}{na_k^2 + \mathrm{Var}[\varphi_k(X_1)]}\right\}.$$

Kertoimet λ_{kn}^* kuitenkin riippuvat (a_k:n ja $\mathrm{Var}[\varphi_k(X_1)]$:n kautta) tuntemattomasta funktiosta f, jota juuri yritetään estimoida, joten käytännössä λ_{kn}^* joudutaan estimoimaan jollain tavoin.

Vielä yksi ehdotus on ottaa

$$\lambda_{kn} = \begin{cases} (1 + \lambda k^{2m})^{-1}, & k \leq n, \\ 0, & k > n \end{cases}$$

ja siis

$$\hat{f}_n(\cdot; (\lambda_{kn})) = \sum_{k=1}^{n} \left(\frac{\hat{a}_{kn}}{1 + \lambda k^{2m}} \right) \varphi_k.$$

Parametrit λ ja m voidaan määrätä hakemalla MISE$[\hat{f}_n(\cdot; (\lambda_{kn}))]$:lle estimaattori ja minimoimalla se λ:n ja m:n suhteen. Jos otetaan $m = 2$ ja $\lambda = Cn^{-4/5}$, $C > 0$, voidaan osoittaa, että

$$\text{MISE}[\hat{f}_n(\cdot; (\lambda_{kn}))] = O(n^{-4/5}),$$

kun $f \in C^2([0,1])$ ja lisäksi $f(0) = f(1)$, $f'(0) = f'(1)$. Tämä tulos seuraa B.L.S. Prakasa Raon kirjan "Nonparametric Functional Estimation" (Academic Press 1983) sivun 162 harjoitustehtävästä 13. Myös muuta ortogonaalisarjaestimaattoriin liittyvää materiaalia löytyy tästä kirjasta.

Haittapuolena ortogonaalisarjamenetelmässä on se, että se ei välttämättä tuota tiheysfunktiota, koska estimaattori voi saada myös negatiivisia arvoja.

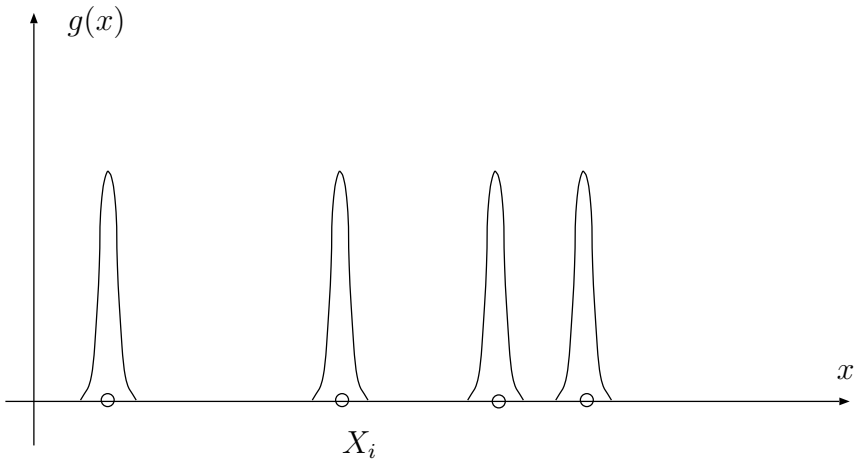
3.12.3 Sakotettu uskottavuus

Olkoon f tiheysfunktio ja $X_1, \ldots, X_n \sim f$ i.i.d. otos. Uskottavuusmenetelmässä estimoidaan f:ää etsimällä funktio $g = \hat{f}_n$, joka maksimoi uskottavuuden

$$L(g) = \prod_{i=1}^{n} g(X_i),$$

kun g kuuluu johonkin tiheysfunktioiden avaruuteen $\mathcal{F}$. Osoittautuu, että todellisuudessa vain tapaus dim $\mathcal{F} < \infty$ on mielekäs, koska $L(g)$ saadaan mielivaltaisen isoksi valitsemalla g siten, että se oleellisesti häviää joukon $\{X_1, \ldots, X_n\}$ ulkopuolella ja suuresta funktioavaruudesta tällaisia funktioita helposti löytyy (kuva 3.28). Sen vuoksi L:ään lisätään sakkotekijä,

$$L_c(g) = \prod_{i=1}^{n} g(X_i) e^{-\lambda c(g)}, \quad \lambda > 0,$$



Kuva 3.28: Suuresta funktioavaruudesta voi löytyä tiheysfunktioita g, jotka oleellisesti häviävät otoksen ulkopuolella.

joka aiheuttaa sileysvaatimuksen g:lle. Eräs mahdollisuus on

$$c(g) = R(g'') = \int_{-\infty}^{\infty} [g''(x)]^2 dx.$$

Nyt λ:lla on silotusparametrin rooli. Tällä *sakotetun uskottavuuden menetelmällä* saatavat estimaattorit ovat tyypillisesti määriteltävissä splinikäyrien avulla.

Luku 4

Parametriton regressio

4.1 Malli

Tarkastellaan paria (X, Y), missä X ja Y ovat satunnaismuuttujia. Kuvaamme Y:n riippuvuutta X:stä mallilla

$$Y = m(X) + \sigma(X)\varepsilon. \tag{4.1}$$

Tässä m ja σ ovat (Borel-mitallisia) reaaliarvoisia funktioita ja σ on ei-negatiivinen. Lisäksi ε on X:stä riippumaton satunnaismuuttuja ja

$$\mathbb{E}\varepsilon = 0, \qquad \mathrm{Var}(\varepsilon) = 1.$$

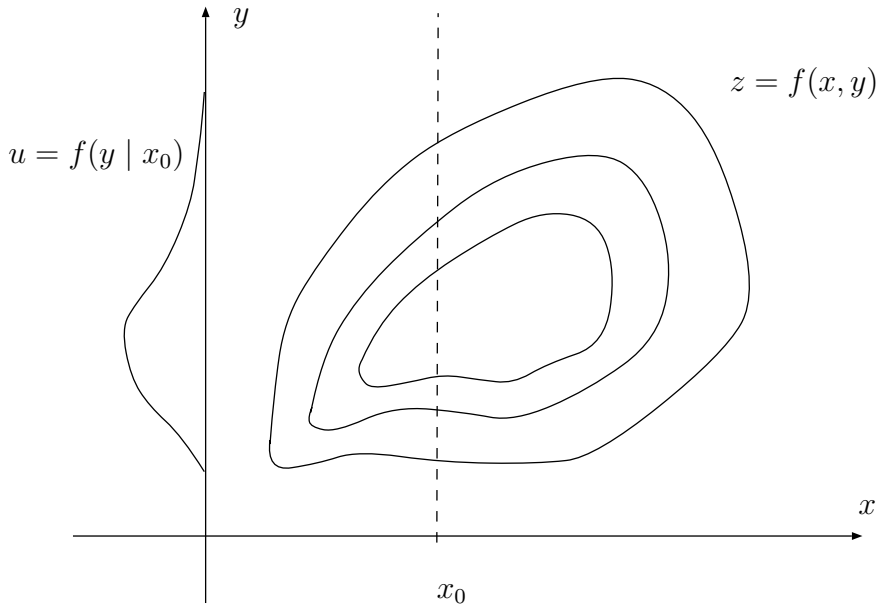
Nyt Y:n ehdollinen odotusarvo ehdolla X on

$$\begin{aligned} \mathbb{E}(Y \mid X) &= \mathbb{E}(m(X) \mid X) + \mathbb{E}(\sigma(X)\varepsilon \mid X) \\ &= m(X) + \sigma(X)\mathbb{E}\varepsilon \\ &= m(X), \end{aligned}$$

koska $\mathbb{E}\varepsilon = 0$. Siten $m(X) = \mathbb{E}(Y \mid X)$, joten m on Y:n regressiofunktio X:n suhteen. Edelleen,

$$\mathrm{Var}(Y \mid X) = \mathbb{E}[(Y - E(Y \mid X))^2 \mid X] = \sigma(X)^2.$$

Tarkkaan ottaen ylläolevat kaavat ovat voimassa melkein varmasti (s.o. todennäköisyydellä 1), kun merkityt odotusarvot ovat olemassa.



Kuva 4.1: Parin (X, Y) tiheysfunktio $z = f(x, y)$ ja Y:n ehdollinen tiheysfunktio $u = f(y \mid x_0)$ ehdolla $X = x_0$.

Jos σ on vakiofunktio, on kyseessä *homoskedastinen* malli. Muussa tapauksessa puhutaan *heteroskedastisesta* mallista.

Olkoon (X, Y):n jakaumalla tiheysfunktio f. Voidaan osoittaa, että kun $\int_{-\infty}^{\infty} f(x, y)dy > 0$, on

$$\begin{aligned} m(x) &= \mathbb{E}(Y \mid X = x) \\ &= \int_{-\infty}^{\infty} y \left[\frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, z)dz} \right] dy = \frac{\int_{-\infty}^{\infty} yf(x, y)dy}{\int_{-\infty}^{\infty} f(x, y)dy} = \int_{-\infty}^{\infty} yf(y \mid x)dy, \end{aligned} \tag{4.2}$$

missä

$$f(y \mid x) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)dy}$$

on Y:n ehdollinen tiheysfunktio ehdolla $X = x$ (kuva 4.1).

4.2 Ydinregressio – Nadarayan-Watsonin menetelmä

Oletetaan malli (4.1), olkoon $f : \mathbb{R}^2 \to [0, \infty[$ parin (X, Y) tiheysfunktio ja $(X_1, Y_1), \ldots, (X_n, Y_n) \sim f$ i.i.d. otos. Siis meillä on i.i.d. otos $\varepsilon_1, \ldots, \varepsilon_n$ satunnaismuuttujan ε jakaumasta (merkitään jatkossa $\varepsilon_1, \ldots, \varepsilon_n \sim \varepsilon$) ja

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \ldots, n.$$

Analogisesti tiheysfunktion parametrittoman estimoinnin kanssa, regressiofunktiolle m saadaan estimaattori valitsemalla sopiva (Borelin) funktio $t_n : \mathbb{R}^{2n+1} \to \mathbb{R}$ ja asettamalla $\hat{m}_n(x) = t_n(x, X_1, Y_1, \ldots, X_n, Y_n)$, $x \in \mathbb{R}$. Tämä voi tapahtua esimerkiksi siten, että estimoidaan tiheysfunktio f otoksen $(X_1, Y_1), \ldots, (X_n, Y_n)$ avulla käyttäen jotain estimaattoria $\hat{f}_n$. Silloin (vrt. (4.2)) on luonnollista ottaa m:n estimaattoriksi

$$\hat{m}_n(x) = \frac{\int_{-\infty}^{\infty} y\hat{f}_n(x, y)dy}{\int_{-\infty}^{\infty} \hat{f}_n(x, y)dy}, \quad x \in \mathbb{R}.$$

Olkoon erityisesti $K : \mathbb{R} \to [0, \infty[$ symmetrinen ydin, $h > 0$ ja muodostetaan tuloydin $L(x, y) = K(x)K(y)$ ja sitä vastaava ydinestimaattori $\mathbb{R}^2$:ssa,

$$\begin{aligned} \hat{f}_n(x, y; h^2 I_2) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^2} L\left(\frac{x - X_i}{h}, \frac{y - Y_i}{h}\right) \\ &= \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right). \end{aligned}$$

Silloin on helppo nähdä (HT), että

$$\hat{m}_n(x) = \hat{m}_n(x; h) = \frac{\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

Tämä regressiofunktion estimaattori tunnetaan nimellä *Nadarayan-Watsonin estimaattori*. Se voidaan kirjoittaa myös muodossa

$$\hat{m}_n(x; h) = \sum_{i=1}^{n} W_h(x - X_i)Y_i,$$

missä

$$W_h(x - X_i) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)},$$

ja pätee

$$\sum_{i=1}^{n} W_h(x - X_i) = 1.$$

Estimaattorin $\hat{m}_n(\cdot; h)$ arvo $\hat{m}_n(x; h)$ pisteessä x on siis arvojen Y_i painotettu keskiarvo x:n ympäristössä. Tavanomaisia ytimiä K (esimerkiksi Gaussin ydin) käytettäessä suurin paino on niillä Y_i, joilla X_i on lähinnä x:ää, koska $K(x - X_i)$ on tavallisesti sitä pienempi mitä suurempi $|x - X_i|$ on (ks. kuva 4.2). Nadarayan-Watsonin estimaattorin idea on siis estimoida ehdollista odotusarvoa $m(x) = \mathbb{E}(Y \mid X = x)$ niiden Y_i:den (painotettuna) keskiarvona, joilla $X_i \approx x$.

Esimerkki 4.1 Tarkastellaan Nadarayan-Watsonin estimaattorin käyttäytymistä eri silotusparametrin arvoilla. Olkoon $n = 30$, X tasaisesti jakautunut välillä $[0, 1]$, $m(x) = x$, $\varepsilon \sim N(0, 1)$, $\sigma = 0.2$ ja $Y_i = X_i + \sigma\varepsilon_i$, $i = 1, \ldots, n$. Kuvassa 4.3 on neljällä eri silotusparametrin arvolla saatua regressiofunktion m estimaattia. Tässä tapauksessa arvot $h = 0.01$ ja $h = 0.05$ ovat selvästi liian pieniä ja $h = 1$ puolestaan liian iso. Arvo $h = 0.1$ toimii parhaiten vaikka regressiofunktion lineaarisuus ei aivan paljastukaan. Silotusparametrin kasvaessa estimaatti konvergoi kohti vakiofunktiota, jonka arvo on pisteiden Y_i keskiarvo (HT). ‖

4.3 Kiinteä asetelma

Yksinkertaisuuden vuoksi suoritamme Nadarayan-Watsonin estimaattorin tarkemman analyysin mallille

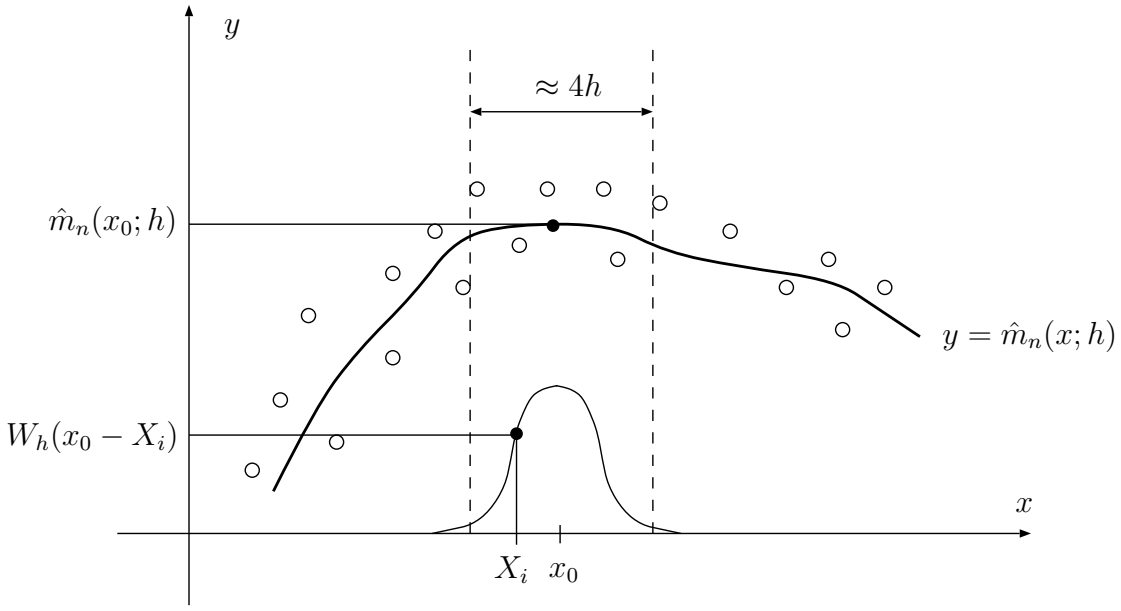
$$Y_i = m(x_i) + \sigma(x_i)\varepsilon_i, \quad i = 1, \ldots, n, \tag{4.3}$$

missä $x_i = i/n$, $m : [0, 1] \to \mathbb{R}$, $\sigma : [0, 1] \to [0, \infty[$, $\varepsilon_1, \ldots, \varepsilon_n \sim \varepsilon$ on i.i.d. otos, $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}[\varepsilon] = 1$. Tämä on *kiinteä asetelma* siinä missä (4.1) on *satunnainen asetelma*.

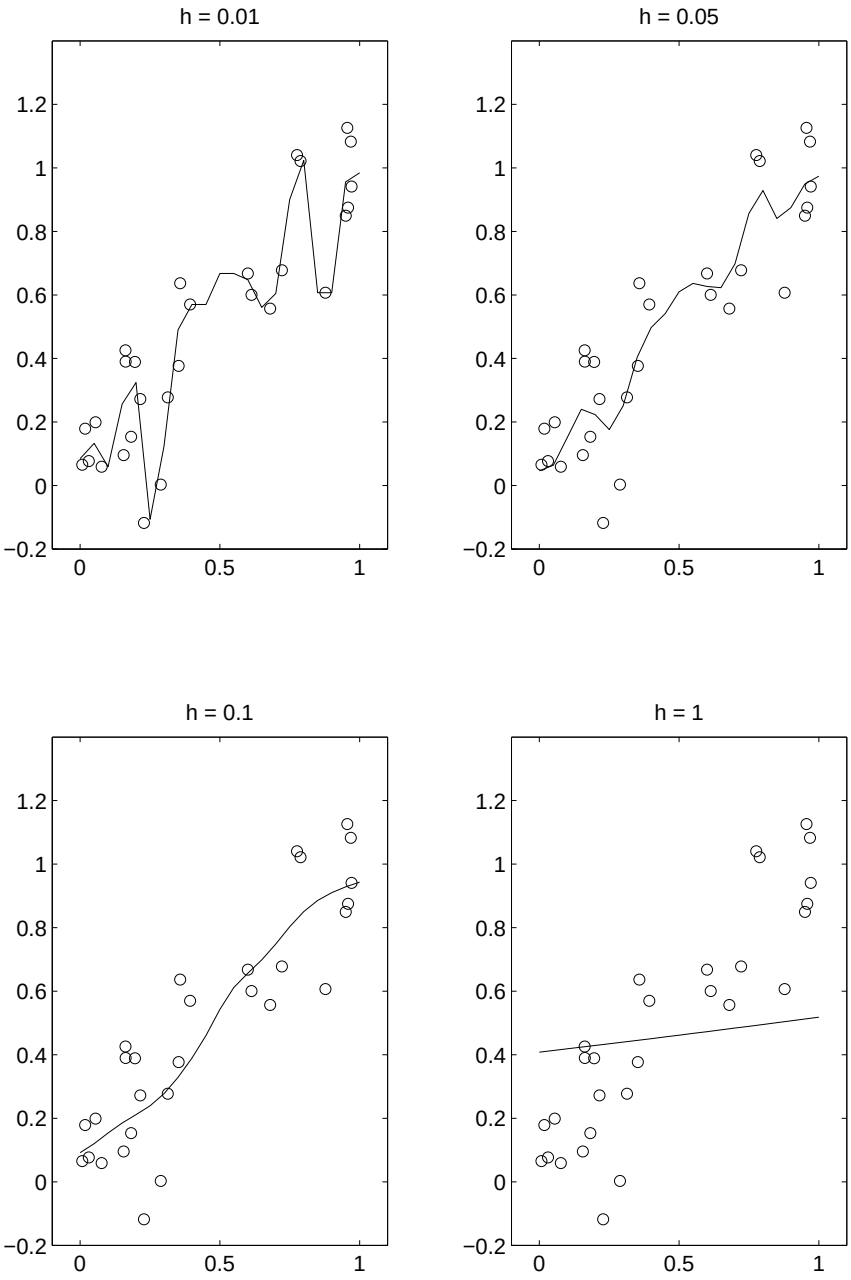
Olkoon

$$\hat{m}_n(x; h_n) = \frac{\sum_{i=1}^{n} K_{h_n}(x - x_i) Y_i}{\sum_{i=1}^{n} K_{h_n}(x - x_i)} = \sum_{i=1}^{n} W_n(x - x_i) Y_i, \tag{4.4}$$

$$W_n(x - x_i) = \frac{K_{h_n}(x - x_i)}{\sum_{j=1}^{n} K_{h_n}(x - x_j)}.$$



Kuva 4.2: Nadarayan-Watsonin estimaattori estimoi regressiofunktion m arvoa pisteessä x_0 tätä pistettä lähellä olevia otospisteitä X_i vastaavien Y_i:den painotettuna keskiarvona. Gaussin ytimen tapauksessa h vastaa keskihajontaa, joten ytimen vaikutusalue on likimain $4h$:n levyinen.



Kuva 4.3: Eri silotusparametrin arvoilla saatavia Nadarayan-Watsonin regressiofunktion estimaatteja kun $n = 30$, X on tasaisesti jakautunut välillä $[0, 1]$, $m(x) = x$, $\varepsilon \sim N(0, 1)$, $\sigma = 0.2$ ja $Y_i = X_i + \sigma\varepsilon_i$, $i = 1, \ldots, n$. Käyetty otos on piirretty pienillä ympyröillä ja se on kaikissa kuvissa sama.

Johdetaan asymptoottinen kehitelmä neliölliselle virheelle

$$\mathrm{MSE}[\hat{m}_n(x;h_n)] = \mathbb{E}[\hat{m}_n(x;h_n) - m(x)]^2.$$

Todistetaan aluksi *Eulerin-Maclaurinin summauskaavan* eräs erikoistapaus.

Lemma 4.2 *Olkoon $g \in C^2([0,1])$ ja $g(0) = g(1) = 0$. Silloin*

$$\frac{1}{n}\sum_{i=1}^{n-1} g(i/n) = \int_0^1 g(x)dx + \frac{1}{2n^2}\int_0^1 p(nx)g''(x)dx,$$

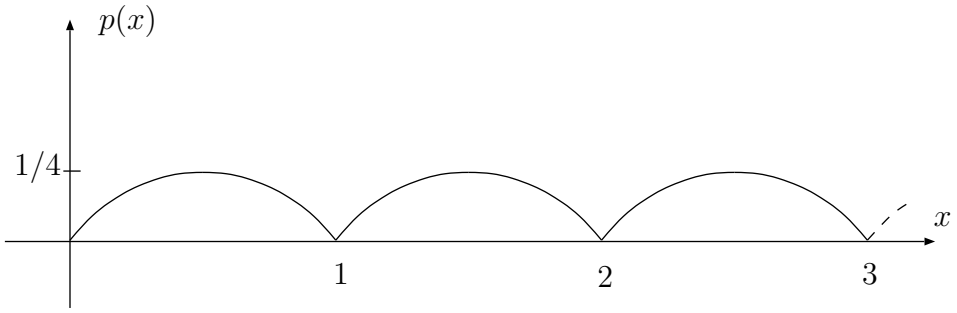
missä $p(x) = (x-i)(i+1-x)$, $x \in [i, i+1[$, $i = 0, 1, 2, \ldots$ ..

Todistus: Funktio p on jaksollinen ja se on saatu kopioimalla funktiota $x \mapsto x(1-x)$, $x \in [0,1[$ väleille $[i, i+1[$, $i = 1, 2, \ldots$ (ks. kuva 4.4). Osittaisintegroidaan kullakin osavälillä $[i/n, (i+1)/n]$ kahdesti:

$$\begin{aligned}
\frac{1}{2n^2}\int_0^1 p(nx)g''(x)dx &= \frac{1}{2n^2}\sum_{i=0}^{n-1}\int_{i/n}^{(i+1)/n} p(nx)g''(x)dx \\
&= \frac{1}{2n^2}\sum_{i=0}^{n-1}\int_{i/n}^{(i+1)/n} (nx-i)(i+1-nx)g''(x)dx \\
&= \frac{1}{2n^2}\sum_{i=0}^{n-1}\left\{ \Big|_{i/n}^{(i+1)/n}(nx-i)(i+1-nx)g'(x) - \int_{i/n}^{(i+1)/n}[-2n^2x + n(2i+1)]g'(x)dx\right\} \\
&= \frac{1}{2n}\sum_{i=0}^{n-1}\int_{i/n}^{(i+1)/n}[2nx-(2i+1)]g'(x)dx \\
&= \frac{1}{2n}\sum_{i=0}^{n-1}\left\{ \Big|_{i/n}^{(i+1)/n}[2nx-(2i+1)]g(x) - 2n\int_{i/n}^{(i+1)/n} g(x)dx\right\} \\
&= \frac{1}{2n}\left\{g(0)+g(1)+2\sum_{i=1}^{n-1} g(i/n)\right\} - \sum_{i=0}^{n-1}\int_{i/n}^{(i+1)/n} g(x)dx \\
&= \frac{1}{n}\sum_{i=1}^{n-1} g(i/n) - \int_0^1 g(x)dx,
\end{aligned}$$

missä kolmannen yhtäsuuruuden sijoitustermi häviää ja toiseksi viimeisessä yhtälössä käytettiin ehtoa $g(0) = g(1) = 0$. □

Lemma 4.3 *Olkoon $K \in C^2(\mathbb{R})$, $K(x) = 0$, kun $|x| > 1$, $x_i = i/n$, $i = 1 \ldots n$, $h_n \to 0$, kun $n \to \infty$ ja $x \in]0,1[$ kiinteä. Silloin kaikilla $r = 0, 1, 2, \ldots$ pätee*



Kuva 4.4: Eulerin-Maclaurinin summauskaavassa esiintyvä jaksollinen funktio.

$$\frac{1}{n}\sum_{i=1}^{n}(x-x_i)^r K_{h_n}(x-x_i) = h_n^r \mu_r(K) + O\left(\frac{1}{(nh_n)^2}\right),$$

missä $\mu_r(K) = \int_{-\infty}^{\infty} z^r K(z)dz = \int_{-1}^{1} z^r K(z)dz$.

Todistus: Olkoon n niin suuri, että $[x-h_n, x+h_n] \subset]0,1[$. Merkitän $g(y) = (x-y)^r K_{h_n}(x-y)$, $y \in [0,1]$. Silloin $g(y) = 0$, kun $|x-y| > h_n$, joten $\{y \mid g(y) \neq 0\} \subset [x-h_n, x+h_n] \subset]0,1[$. Erityisesti $g(0) = g(1) = 0$. Soveltamalla lemmaa 4.2 saamme siten

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x-x_i)^r K_{h_n}(x-x_i) &= \frac{1}{n}\sum_{i=1}^{n-1}(x-i/n)^r K_{h_n}(x-i/n) \\ &= \int_0^1 (x-y)^r K_{h_n}(x-y)dy + \frac{1}{2n^2}\int_0^1 p(ny)\frac{d^2}{dy^2}[(x-y)^r K_{h_n}(x-y)]dy.\end{aligned}$$

Tässä

$$\begin{aligned}\int_0^1 (x-y)^r K_{h_n}(x-y)dy &= h_n^r \int_0^1 \left(\frac{x-y}{h_n}\right)^r \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) dy \\ &= h_n^r \int_{(x-1)/h_n}^{x/h_n} z^r K(z)dz \\ &= h_n^r \int_{-1}^{1} z^r K(z)dz = h_n^r \mu_r(K),\end{aligned}$$

missä toinen yhtäsuuruus seuraa muuttujan vaihdoksesta $z = (x-y)/h_n$. Edelleen

$$\frac{d^2}{dy^2}[(x-y)^r K_{h_n}(x-y)]$$
$$= \begin{cases} \frac{1}{h_n^2}(K'')_{h_n}(x-y), & r=0, \\ \frac{2}{h_n^2}(K')_{h_n}(x-y) + (y-x)\frac{1}{h_n^2}(K'')_{h_n}(x-y), & r=1, \\ r(r-1)K_{h_n}(x-y) + 2r(x-y)^{r-1}\frac{1}{h_n}(K')_{h_n}(x-y) & \\ +(y-x)^r (K'')_{h_n}\frac{1}{h_n^2}(K'')_{h_n}(x-y), & r \geq 2. \end{cases}$$

Nyt kaikilla $y \in [0,1]$ pätee $|p(ny)| \leq 1/4$, $|x-y| \leq 1$ ja lisäksi $\int_{-1}^{1} |(K^{(j)})_{h_n}(z)|dz < \infty$, $j = 0,1,2$. Siten

$$\left| \int_0^1 p(ny) \frac{d^2}{dy^2}[(x-y)^r K_{h_n}(x-y)]dy \right| \leq \frac{C}{h_n^2}, \quad C > 0. \qquad \square$$

Lause 4.4 *Olkoon $K \in C^2(\mathbb{R})$ symmetrinen ydin, $\int_{-\infty}^{\infty} K(x)dx = 1$, $K(x) \geq 0$ kaikilla x ja $K(x) = 0$, kun $|x| > 1$. Olkoon edelleen $m \in C^2([0,1])$, $\sigma \in C([0,1])$ ei-negatiivinen, $x_i = i/n$, ja*

$$Y_i = m(x_i) + \sigma(x_i)\varepsilon_i, \quad i = 1, \ldots, n,$$

missä $\varepsilon_1, \ldots, \varepsilon_n \sim \varepsilon$ on i.i.d. otos, $\mathbb{E}\varepsilon = 0$ ja $\mathrm{Var}[\varepsilon] = 1$. Olkoon $\hat{m}_n(\cdot; h_n)$ kaavan (4.4) määrittelemä estimaattori. Jos $x \in]0,1[$ on kiinteä ja $h_n \to 0$, $nh_n \to \infty$, kun $n \to \infty$, on tällöin

$$\begin{aligned} \mathrm{MSE}[\hat{m}_n(x;h_n)] &= \mathbb{E}[\hat{m}_n(x;h_n) - m(x)]^2 \\ &= \frac{1}{4}h_n^4 \mu_2(K)^2 m''(x)^2 + \frac{R(K)\sigma(x)^2}{nh_n} + o\left(h_n^4 + \frac{1}{nh_n}\right). \end{aligned}$$

Todistus: Tavalliseen tapaan

$$\mathrm{MSE}[\hat{m}_n(x;h_n)] = [\mathbb{E}\hat{m}_n(x;h_n) - m(x)]^2 + \mathrm{Var}[\hat{m}_n(x;h_n)]$$

ja käsittelemme harhan neliön (ensimmäinen termi) ja varianssin (toinen termi) erikseen. Tarkastellaan harhaa ensin. Saamme

$$\begin{aligned} \mathbb{E}\hat{m}_n(x;h_n) &= \mathbb{E}\left(\sum_{i=1}^n W_n(x-x_i)Y_i\right) \\ &= \sum_{i=1}^n W_n(x-x_i)\mathbb{E}Y_i = \sum_{i=1}^n W_n(x-x_i)m(x_i). \end{aligned}$$

Taylorin kehitelmän perusteella

$$m(x_i) = m(x) - m'(x)(x - x_i) + \frac{1}{2}m''(\xi_i)(x - x_i)^2,$$

missä ξ_i on x_i:n ja x:n välissä. Siten

$$\begin{aligned}\mathbb{E}\hat{m}_n(x; h_n) &= m(x)\sum_{i=1}^{n} W_n(x - x_i) \\ &\quad -m'(x)\sum_{i=1}^{n}(x - x_i)W_n(x - x_i) + \frac{1}{2}\sum_{i=1}^{n} m''(\xi_i)(x - x_i)^2 W_n(x - x_i).\end{aligned}$$

Koska $\sum_{i=1}^{n} W_n(x - x_i) = 1$, on

$$m(x)\sum_{i=1}^{n} W_n(x - x_i) = m(x).$$

Edelleen,

$$\begin{aligned}\sum_{i=1}^{n}(x - x_i)W_n(x - x_i) &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x - x_i)K_{h_n}(x - x_i)}{\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - x_i)} \\ &= \frac{h_n\mu_1(K) + O\left(\frac{1}{(nh_n)^2}\right)}{\mu_0(K) + O\left(\frac{1}{(nh_n)^2}\right)} \\ &= O\left(\frac{1}{(nh_n)^2}\right),\end{aligned}$$

missä toisessa yhtäsuuruudessa käytettiin lemmaa 4.3 ja lopuksi sitä, että $\mu_1(K) = 0$ ja $\mu_0(K) = 1$. Sitten kirjoitamme

$$\begin{aligned}\sum_{i=1}^{n} m''(\xi_i)(x - x_i)^2 W_n(x - x_i) &= m''(x)\sum_{i=1}^{n}(x - x_i)^2 W_n(x - x_i) \\ &\quad + \sum_{i=1}^{n}[m''(\xi_i) - m''(x)](x - x_i)^2 W_n(x - x_i).\end{aligned}$$

Tässä

$$\begin{aligned}\sum_{i=1}^{n}(x - x_i)^2 W_n(x - x_i) &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x - x_i)^2 K_{h_n}(x - x_i)}{\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - x_i)} \\ &= \frac{h_n^2\mu_2(K) + O\left(\frac{1}{(nh_n)^2}\right)}{\mu_0(K) + O\left(\frac{1}{(nh_n)^2}\right)} \\ &= h_n^2\mu_2(K) + O\left(\frac{1}{(nh_n)^2}\right),\end{aligned} \tag{4.5}$$

missä käytettiin lemmaa 4.3 ja sitä, että $\mu_0(K) = 1$. Koska $W_n(x - x_i) = 0$, kun $|x - x_i| > h_n$, voimme kirjoittaa

$$\sum_{i=1}^{n}[m''(\xi_i) - m''(x)](x - x_i)^2 W_n(x - x_i) = \sum_{i=1}^{n} \varepsilon_{in}(x - x_i)^2 W_n(x - x_i),$$

missä

$$\varepsilon_{in} = \begin{cases} m''(\xi_i) - m''(x), & |x - x_i| \leq h_n, \\ 0, & |x - x_i| > h_n. \end{cases}$$

Koska m'' on jatkuva pisteessä x ja ξ_i on x:n ja x_i:n välissä, on $\varepsilon_n \equiv \max_{i=1,\ldots,n} |\varepsilon_{in}| \to 0$, kun $n \to \infty$. Siten

$$\left|\sum_{i=1}^{n}[m''(\xi_i) - m''(x)](x - x_i)^2 W_n(x - x_i)\right| \leq \varepsilon_n \sum_{i=1}^{n}(x - x_i)^2 W_n(x - x_i),$$

ja (4.5):n nojalla saadaan

$$\sum_{i=1}^{n}[m''(\xi_i) - m''(x)](x - x_i)^2 W_n(x - x_i) = o(h_n^2) + o\left(\frac{1}{(nh_n)^2}\right).$$

Siten

$$\mathbb{E}\hat{m}_n(x; h_n) = m(x) + \frac{1}{2}h_n^2 \mu_2(K) m''(x) + o(h_n^2) + O\left(\frac{1}{(nh_n)^2}\right).$$

Tästä saadaan harhan neliölle

$$\begin{aligned} [\mathbb{E}\hat{m}_n(x; h_n) - m(x)]^2 &= \frac{1}{4}h_n^4 \mu_2(K)^2 m''(x)^2 \\ &\quad + o(h_n^4) + O\left(\frac{1}{n^2}\right) + O\left(\frac{1}{(nh_n)^4}\right). \end{aligned} \tag{4.6}$$

Sitten käsitellään varianssin osuus. Saamme

$$\begin{aligned} \mathrm{Var}[\hat{m}_n(x; h_n)] &= \mathrm{Var}\left[\sum_{i=1}^{n} W_n(x - x_i) Y_i\right] \\ &= \sum_{i=1}^{n}[W_n(x - x_i)]^2 \mathrm{Var}[Y_i] \\ &= \sum_{i=1}^{n}[W_n(x - x_i)]^2 \sigma(x_i)^2 \\ &= \frac{\frac{1}{n^2}\sum_{i=1}^{n} \sigma(x_i)^2 [K_{h_n}(x - x_i)]^2}{\left[\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - x_i)\right]^2}. \end{aligned}$$

Tässä

$$\frac{1}{n^2}\sum_{i=1}^{n}\sigma(x_i)^2[K_{h_n}(x-x_i)]^2 = \frac{1}{nh_n}\cdot\frac{1}{n}\sum_{i=1}^{n}\sigma(x_i)^2(K^2)_{h_n}(x-x_i)$$
$$= \frac{\sigma(x)^2}{nh_n}\cdot\frac{1}{n}\sum_{i=1}^{n}(K^2)_{h_n}(x-x_i) + \frac{1}{nh_n}\cdot\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{in}(K^2)_{h_n}(x-x_i),$$

missä

$$\varepsilon_{in} = \begin{cases} \sigma(x_i)^2 - \sigma(x)^2, & |x-x_i| \le h_n, \\ 0, & |x-x_i| > h_n. \end{cases}$$

Koska σ on jatkuva pisteessä x, on $\varepsilon_n \equiv \max_{i=1,\ldots,n}|\varepsilon_{in}| \to 0$, kun $n \to \infty$. Lemman 4.3 nojalla

$$\begin{aligned}\frac{\sigma(x)^2}{nh_n}\cdot\frac{1}{n}\sum_{i=1}^{n}(K^2)_{h_n}(x-x_i) &= \frac{\sigma(x)^2}{nh_n}\left[\int_{-\infty}^{\infty}K(x)^2dx + O\left(\frac{1}{(nh_n)^2}\right)\right] \\ &= \frac{R(K)\sigma(x)^2}{nh_n} + O\left(\frac{1}{(nh_n)^3}\right).\end{aligned}$$

Edelleen, koska

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{in}(K^2)_{h_n}(x-x_i)\right| \le \varepsilon_n\left[R(K) + O\left(\frac{1}{(nh_n)^2}\right)\right] = o(1),$$

on

$$\frac{1}{nh_n}\cdot\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{in}(K^2)_{h_n}(x-x_i) = o\left(\frac{1}{nh_n}\right).$$

Siten saamme varianssille

$$\mathrm{Var}[\hat{m}_n(x;h_n)] = \frac{\frac{R(K)\sigma(x)^2}{nh_n} + o\left(\frac{1}{nh_n}\right)}{\left[1+O\left(\frac{1}{(nh_n)^2}\right)\right]^2} = \frac{R(K)\sigma(x)^2}{nh_n} + o\left(\frac{1}{nh_n}\right). \qquad (4.7)$$

Kaavassa (4.6) termit $O(1/n^2)$ ja $O(1/(nh_n)^4)$ ovat tyyppiä $O(1/(nh_n))$. Siten yhtälöistä (4.6) ja (4.7) seuraa väite. □

Edellisen lauseen perusteella asymptoottiselle pisteittäiselle neliölliselle virheelle saadaan kaava

$$\mathrm{AMSE}[\hat{m}_n(x;h_n)] = \frac{1}{4}h_n^4\mu_2(K)^2m''(x)^2 + \frac{R(K)\sigma(x)^2}{nh_n}.$$

Tästä saadaan edelleen optimaalinen silotusparametri

$$h_n^* = \left[\frac{R(K)\sigma(x)^2}{\mu_2(K)^2 m''(x)^2}\right]^{1/5} \cdot n^{-1/5}$$

ja optimaalinen asymptoottinen virhe

$$\mathrm{AMSE}[\hat{m}_n(x; h_n^*)] = \frac{5}{4}[\mu_2(K)^2 m''(x)^2]^{1/5}[R(K)\sigma(x)^2]^{4/5} \cdot n^{-4/5}.$$

Virheen konvergenssivauhti kohti nollaa on siis sama kuin tiheysfunktion ydinestimaattorilla.

4.4 Satunnainen asetelma

Nadarayan-Watsonin estimaattorin matemaattinen analyysi on hankalampaa, jos sekä X että Y ovat satunnaismuuttujia, koska tällöin $\hat{m}_n(x; h)$ on kahden toisistaan riippuvan satunnaismuuttujan osamäärä,

$$\hat{m}_n(x; h) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i}{\sum_{i=1}^n K_h(x - X_i)}.$$

Tilanne muuttu kuitenkin kiinteän asetelman kaltaiseksi, jos tarkastellaan *ehdollista* virhettä

$$\mathbb{E}[[\hat{m}_n(x; h) - m(x)]^2 \mid X_1, \ldots, X_n],$$

jossa otos $X_1, \ldots, X_n$ on kiinnitetty. Sopivilla oletuksilla (ks. [12, luku 5.4]) saadaan

$$\mathbb{E}[\hat{m}_n(x; h_n) - m(x) \mid X_1, \ldots, X_n] = \frac{1}{2}h_n^2 \left\{\frac{2m'(x)f'(x)}{f(x)} + m''(x)\right\} \mu_2(K) + h_n^2 U_n,$$

$$\mathrm{Var}[\hat{m}_n(x; h_n) \mid X_1, \ldots, X_n] = \frac{R(K)\sigma(x)^2}{nh_n f(x)} + \frac{1}{nh_n}V_n,$$

missä f on X:n tiheysfunktio ja satunnaismuuttujat U_n ja V_n konvergoivat kohti nollaa stokastisesti. Tästä saadaan

$$\begin{aligned}\mathrm{MSE}[\hat{m}_n(x; h_n) \mid X_1, \ldots, X_n] &= \frac{1}{4}h_n^4 \left\{\frac{2m'(x)f'(x)}{f(x)} + m''(x)\right\}^2 \mu_2(K)^2 \\ &\quad + \frac{R(K)\sigma(x)^2}{nh_n f(x)} + \left(h_n^4 + \frac{1}{nh_n}\right) W_n,\end{aligned}$$

missä $W_n \to 0$ stokastisesti. Aymptoottinen ehdollinen virhe on

$$\mathrm{AMSE}[\hat{m}_n(x;h_n) \mid X_1,\ldots,X_n] = \frac{1}{4}h_n^4\left\{\frac{2m'(x)f'(x)}{f(x)} + m''(x)\right\}^2 \mu_2(K)^2 + \frac{R(K)\sigma(x)^2}{nh_nf(x)}.$$

Havaitsemme, että jos f on välillä $[0,1]$ tasaisen jakauman tiheysfunktio, päädytään itseasiassa kiinteälle asetelmalle johdettuun tulokseen. Sopivilla oletuksilla (ks. [8, luku 8.1.3]) saadaan sama asymptoottinen virhe vaikka vielä keskiarvoistetaan otoksen $X_1,\ldots,X_n$ suhteen, eli kun tarkastellaan virhettä

$$\mathrm{MSE}[\hat{m}_n(x;h)] = \mathbb{E}[\hat{m}_n(x;h) - m(x)]^2 = \mathbb{E}\{\mathbb{E}[[\hat{m}_n(x;h) - m(x)]^2 \mid X_1,\ldots,X_n]\}.$$

4.5 Eräitä muita ydinregressiomenetelmiä

Olkoon f tiheysfunktio ja $X \sim f$. Nadarayan-Watsonin estimaattori voidaan kirjoittaa muotoon

$$\begin{aligned}\hat{m}_n(x;h) &= \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)Y_i}{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)} \\ &= \sum_{i=1}^n \frac{1}{n\hat{f}_n(x;h)}K_h(x-X_i)Y_i, \qquad (4.8)\end{aligned}$$

missä $\hat{f}_n(x;h) = (1/n)\sum_{i=1}^n K_h(x-X_i)$ on f:n ydinestimaattori. Tätä kaavaa voi käyttää motivoimaan eräitä muita ydinregressiomenetelmiä.

Olkoon $f(x) = 0$, kun $x \notin [0,1]$. *Priestleyn-Chaon* estimaattori on

$$\hat{m}_n(x;h) = \sum_{i=1}^n [X_{(i)} - X_{(i-1)}]K_h(x-X_{(i)})Y_{[i]}, \qquad (4.9)$$

missä $0 \le X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} \le 1$ on otoksen $X_1,\ldots,X_n$ kasvava permutaatio, $X_{(0)} = 0$, ja $Y_{[i]}$ on $X_{(i)}$:n pari alkuperäisessä otoksessa $(X_1,Y_1),\ldots,(X_n,Y_n)$. Jos h on pieni (n suuri), niin vain x:ää lähellä olevia otospisteitä X_i vastaavat termit ovat tärkeitä (4.9):ssa. Jos F ja $\hat{F}_n$ ovat vastaavasti X:n kertymäfunktio ja empiirinen kertymäfunktio (vrt. kuva 3.2) ja approksimoidaan

$$\int_{X_{(i-1)}}^{X_{(i)}} f(t)dt = F(X_{(i)}) - F(X_{i-1}) \approx \hat{F}_n(X_{(i)}) - \hat{F}_n(X_{i-1}) = \frac{i}{n} - \frac{i-1}{n} = \frac{1}{n},$$

on relevanteilla i:n arvoilla voimassa

$$f(x)[X_{(i)} - X_{(i-1)}] \approx \int_{X_{(i-1)}}^{X_{(i)}} f(t)dt \approx \frac{1}{n}$$

ja näemme, että (4.9) ja (4.8) muistuttavat toisiaan.

Gasserin-Müllerin estimaattori on

$$\hat{m}_n(x;h) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(x-t)dtY_{[i]},$$

missä $s_i = (1/2)(X_{(i)} + X_{(i+1)})$ on välin $[X_{(i)}, X_{(i+1)}]$ keskipiste ja $s_0 = 0, s_n = 1$. Jos $X_{(i)}$:t ovat likimain tasavälisiä, on tämä suurin piirten sama estimaattori kuin (4.9), koska

$$\int_{s_{i-1}}^{s_i} K_h(x-t)dt \approx K_h(x - X_{(i)})[X_{(i)} - X_{(i-1)}].$$

4.6 Lokaali regressio

Olkoon $D \subset \mathbb{R}$ väli. Parametrisessa regressiossa tarkastellaan jotain sopivaa funktioperhettä $\mathcal{F} = \{m(\cdot;\theta) \mid \theta \in \Theta\}$ ja pyritään konstruoimaan otokseen $(X_1, Y_1), \ldots, (X_n, Y_n)$ perustuva estimaattori $\hat{\theta}_n$, jolla

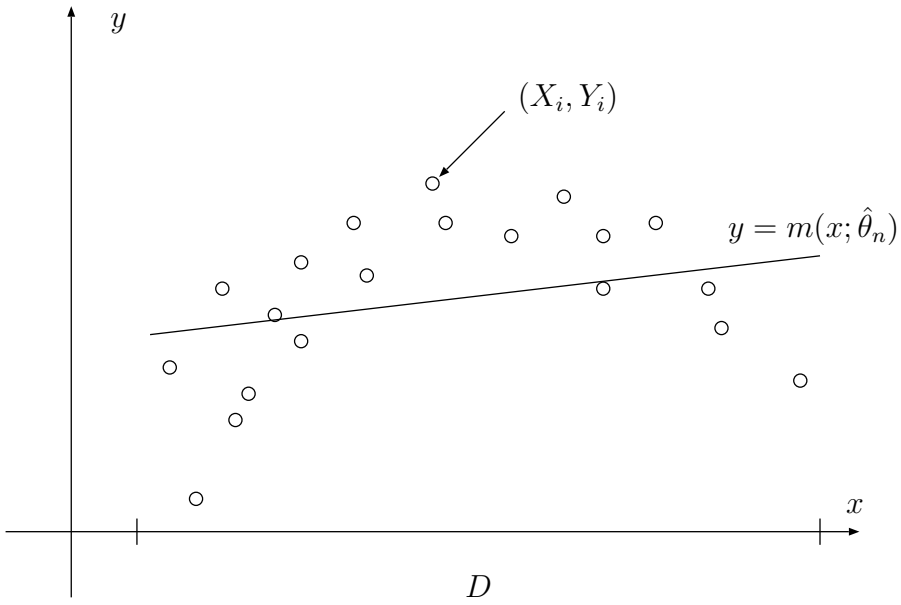
$$m(x) = \mathbb{E}(Y \mid X = x) \approx m(x;\hat{\theta}_n), \quad x \in D.$$

Jos $\mathcal{F}$ on "pieni", eli koostuu hyvin yksinkertaisista funktioista, voi tämä lähestymistapa johtaa huonoon globaaliin, koko aluetta D koskevaan tulokseen (kuva 4.5). Yksinkertaiset funktiot, kuten matala-asteiset polynomit, voivat kuitenkin toimia hyvin *lokaalisti*, kunkin estimointipisteen x pienessä ympäristössä.

Otetaan erityisesti $\mathcal{F}$:ksi vakiofunktioden joukko, $\Theta = \mathbb{R}$, $m(t;a) = a$, $t \in D$, $a \in \Theta$. Olkoon $x \in D$ kiinteä ja etsitään sellainen $m(\cdot;\hat{a}_n) \in \mathcal{F}$, joka estimoi m:ää hyvin pisteessä x. Tähän päästään minimoimalla painotettujen neliövirheiden summa

$$\lambda(a) = \sum_{i=1}^{n} [Y_i - m(X_i;a)]^2 K_h(x - X_i) = \sum_{i=1}^{n} [Y_i - a]^2 K_h(x - X_i),$$

missä K on sopiva ei-negatiivinen symmetrinen ydin, esimerkiksi Gaussin ydin. Eniten painoa $\lambda(a)$:ssa silloin saavat ne parit (X_i, Y_i), joissa X_i on lähellä x:ää. Ehdosta



Kuva 4.5: Esimerkki huonosta globaalista lopputuloksesta, kun kuvan aineistoon on haettu regressiofunktioestimaatti suorien muodostamasta funktioavaruudesta.

$$\lambda'(a) = -2\sum_{i=1}^{n}[Y_i - a]K_h(x - X_i) = 0$$

saadaan

$$a = \frac{\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\sum_{i=1}^{n} K_h(x - X_i)} \equiv \hat{a}_n.$$

Siten $m(x; \hat{a}_n) = \hat{a}_n$ on Nadarayan-Watsonin estimaattori pisteessä x!

Tämä lokaalin regression idea yleistyy välittömästi (korkeintaan) astetta ℓ oleville polynomeille,

$$m(t; a_0, a_1, \ldots, a_\ell) = \sum_{k=0}^{\ell} a_k t^k,$$

jolloin puhutaan *lokaalista polynomisesta regressiosta*. Käytännössä tärkein on tapaus $\ell = 1$, eli *lokaali lineaarinen regressio*. Nadarayan-Watsonin estimaattoriin verrattuna sen huomattavana etuna on parempi käyttäytyminen estimointivälin päätepisteissä. Tavallisesti lokaali polynomi keskistetään x:ään, jolloin lineaarisessa tapauksessa

$$m(t; a_0, a_1) = a_0 + a_1(x - t), \quad t \in D$$

ja minimoidaan lauseke

$$\lambda(a_0, a_1) = \sum_{i=1}^{n} [Y_i - (a_0 + a_1(x - X_i))]^2 K_h(x - X_i).$$

Jos minimoija on $(\hat{a}_{0n}, \hat{a}_{1n})$, otetaan regressiofunktion estimaattoriksi pisteessä x arvo $\hat{a}_{0n}$,

$$\hat{m}_n(x; h) = m(x; \hat{a}_{0n}, \hat{a}_{1n}) = \hat{a}_{0n}.$$

Pienellä laskulla (HT) saadaan tälle eksplisiittinen kaava

$$\hat{m}_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{\hat{s}_2(x) - \hat{s}_1(x)(x - X_i)\} K_h(x - X_i) Y_i}{\hat{s}_2(x)\hat{s}_0(x) - \hat{s}_1(x)^2}, \tag{4.10}$$

missä

$$\hat{s}_r(x) = \frac{1}{n} \sum_{i=1}^{n} (x - X_i)^r K_h(x - X_i)$$

(vrt. [12, luku 5.2]).

On huomattava, että lokaali lineaarinen estimaatti $\hat{m}_n(x; h)$ ei x:n funktiona suinkaan ole paloittain lineaarinen. Sovitettu suora riippu pisteestä x ja niitä on siis äärettömän paljon.

4.7 Silotusparametrin valinta

Olkoon $\hat{m}_n(\cdot; h)$ silotusparametrista $h > 0$ riippuva ydinregressioestimaattori. Kun halutaan h, joka toimii hyvin koko alueella D, on tarkasteltava sopivaa globaalia virhekriteeriä. Eräs mahdollisuus on ehdollinen virhe

$$\begin{aligned} \mathrm{MISE}[\hat{m}_n(\cdot; h) \mid X_1, \ldots, X_n] &= \mathbb{E}\left[\int_D [\hat{m}_n(x; h) - m(x)]^2 f(x) dx \mid X_1, \ldots, X_n \right] \\ &= \mathbb{E}\left[[\hat{m}_n(X; h) - m(X)]^2 \mid X_1, \ldots, X_n \right]. \end{aligned} \tag{4.11}$$

Ottamalla odotusarvo vielä otoksen $X_1, \ldots, X_n$ yli saadaan kriteeri

$$\mathrm{MISE}[\hat{m}_n(\cdot; h)] = \mathbb{E} \int_D [\hat{m}_n(x; h) - m(x)]^2 f(x) dx = \mathbb{E}[\hat{m}_n(X; h) - m(X)]^2.$$

Joskus käytetään vielä ylimääräistä painofunktiota w, jolloin yllä $f(x)$ korvataan $f(x)w(x)$:llä.

Kuten tiheysfunktion estimoinnissa, globaalin virheen asymptoottista arvoa voi käyttää sijoitusestimaattoreiden (plug-in) konstruoimiseen optimaaliselle h:lle. Toinen lähestymistapa on käyttää ristiinvalidointia. Kehitelmästä

$$Y - \hat{m}_n(X; h) = [Y - m(X)] + [m(X) - \hat{m}_n(X; h)]$$

saadaan helposti

$$\begin{aligned} \mathbb{E}[[Y - \hat{m}_n(X; h)]^2 \mid X_1, \ldots, X_n] &= \mathbb{E}[[Y - m(X)]^2 \mid X_1, \ldots, X_n] \\ &+ \mathbb{E}[[m(X) - \hat{m}_n(X; h)]^2 \mid X_1, \ldots, X_n]. \end{aligned}$$

Tässä oikean puolen ensimmäinen termi ei riipu h:sta ja toinen on sama kuin (4.11). Siten h voidaan määrätä minimoimalla lauseke

$$\mathbb{E}[[Y - \hat{m}_n(X; h)]^2 \mid X_1, \ldots, X_n].$$

Tätä voi estimoida lausekkeella

$$\mathrm{LSCV}(h) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{m}_{n,-i}(X_i; h)]^2,$$

missä $\hat{m}_{n,-i}(\cdot; h)$ on laskettu otoksesta $\{(X_j, Y_j) \mid j = 1, \ldots, n,\ j \neq i\}$.

Edelleen, kuten tiheysfunktion estimoinnissa, h voidaan joskus valita myös subjektiivisesti tai jotain sovelluskohtaista kriteeriä optimoimalla.

Ilman todistuksia todetaan vielä, että regressiofunktion ydinestimaattorit ovat sopivissa funktioluokissa minimax-optimaalisia.

4.8 Luottamusvälit

Joskus on tarpeen tietää miten tarkka regressiofunktion estimaatti on esitmointipisteessä x tai koko tarkastelualueessa D. Luottamusväli $\hat{m}_n(x; h)$:lle voidaan konstruoida käyttäen hyväksi sopivasti normeeratun erotuksen $\hat{m}_n(x; h) - m(x)$ asymptoottista

normaalisuutta. Jos esimerkiksi $h_n = Cn^{-1/5}$, niin sopivilla oletuksilla pätee (ks. [7, kappale 4]) Nadarayan-Watsonin estimaattorille jakaumakonvergenssi

$$\frac{\sqrt{nh_n}}{\sqrt{\sigma(x)^2 R(K)/f(x)}}[\hat{m}_n(x;h_n) - m(x)] \to N(b,1),$$

kun $n \to \infty$, missä

$$b = \left(\frac{2m'(x)f'(x)}{f(x)} + m''(x)\right)\mu_2(K),$$

ja f on X:n tiheys.

Jos itseasiassa $h_n = o(n^{-1/5})$, harha b häviää asymptoottisesti. Jos b tällöin jätetään huomiotta, saadaan likimääräinen $(1-\alpha)\cdot 100\%$:n luottamusväli

$$[\hat{m}_n(x;h_n) - \Delta, \hat{m}_n(x;h_n) + \Delta],$$

missä

$$\Delta = \frac{R(K)^{1/2}\hat{\sigma}_n(x)}{\sqrt{nh_n\hat{f}_n(x)}} \cdot c_\alpha,$$

kun $\hat{\sigma}_n$ ja $\hat{f}_n$ ovat σ:n ja f:n estimaattoreita ja

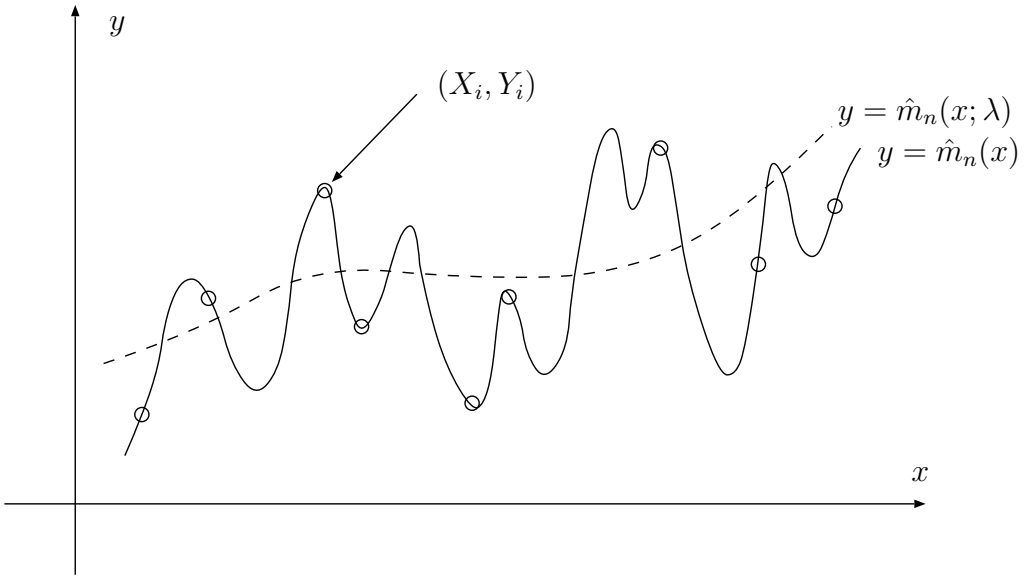
$$\frac{1}{\sqrt{2\pi}}\int_{-c_\alpha}^{c_\alpha} e^{-\frac{1}{2}t^2}dt = 1-\alpha. \tag{4.12}$$

Voidaan esimerkiksi ottaa $\hat{f}_n$:ksi f:n ydinestimaattori ja

$$\hat{\sigma}_n(x)^2 = \sum_{i=1}^n W_h(x-X_i)[Y_i - \hat{m}_n(x;h_n)]^2,$$

$$W_h(x-X_i) = \frac{K_h(x-X_i)}{\sum_{j=1}^n K_h(x-X_j)}.$$

Edelleen voidaan osoittaa, että yo. luottamusvälikaavaa voidaan käyttää *samanaikaisesti* kiinteissä pisteissä $x_1, \ldots, x_\ell$. Normalisoidut erotukset $\hat{m}_n(x_k;h_n) - m(x_k)$ ovat nimittäin asymptoottisesti riippumattomia, jolloin riittää määrätä c_α s.e. (4.12) pätee, kun $(1-\alpha)$:n tilalla on $(1-\alpha)^{1/\ell}$.



Kuva 4.6: Jos regressiofunktion estimaatiksi otetaan virheiden neliösumman minimoiva funktio, päädytään otosta interpoloivaan ratkaisuun (yhtenäinen käyrä). Jos optimiraktaisulta vaaditaaan myös sileyttä, päädytään parempaan estimaattiin (katkoviiva).

4.9 Silottava splini

Tarkastellaan mallia

$$Y_i = m(X_i) + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

missä $\sigma > 0$ on vakio, $\varepsilon_1, \ldots, \varepsilon_n \sim \varepsilon$ on i.i.d. otos, ε on riipumaton X:stä, $\mathbb{E}\varepsilon = 0$ ja $\mathrm{Var}[\varepsilon] = 1$. Naiivissa pienimmän neliösumman menetelmässä etsitään funktio $g = g^*$, joka minimoi virhesumman

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]^2 \tag{4.13}$$

ja otetaan $\hat{m}_n = g^*$. Jos funktioille g ei aseteta mitään rajoituksia (paitsi ehkä jokin sileysvaatimus $g \in C^s$), on $\hat{m}_n$ mikä tahansa (riittävän sileä) funktio, joka kulkee otospisteiden $(X_1, Y_1), \ldots, (X_n, Y_n)$ kautta. Tällainen $\hat{m}_n$ on harvoin järkevä tilastollinen malli (kuva 4.6).

Optimiratkaisun rosoisuutta voi yrittää hillitä lisäämällä (4.13):een sakkotermi, esimerkiksi

$$\lambda \int_a^b [g''(x)]^2 dx,$$

missä $\lambda > 0$ on vakio ja $a = \min\{X_1, \ldots, X_n\}$, $b = \max\{X_1, \ldots, X_n\}$. Sitten haetaan funktio $g*$, joka minimoi lausekkeen

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]^2 + \lambda \int_a^b [g''(x)]^2 dx. \tag{4.14}$$

Minimiratkaisu $g^* = \hat{m}_n(\cdot;\lambda)$ on kompromissi interpoloinnin (pieni interpolointivirhe $(1/n)\sum_{i=1}^{n}[Y_i - g(X_i)]^2$) ja sileyden (pieni $\lambda \int_a^b [g''(x)]^2 dx$) välillä (kuva 4.6). Vakiolla λ on silotusparametrin rooli. Kun λ on pieni, pyrkii $\hat{m}_n(\cdot;\lambda)$ interpoloimaan otospisteitä. Kun $\lambda \to \infty$, voidaan osoittaa, että menetelmä lähestyy tavallista lineaarista regressiota (pienimmän neliösumman suoran sovitusta). Tämä lienee intuitiivisesti uskottavaa, koska suoralle $y = g(x)$ pätee $g'' = 0$, jolloin kaavassa (4.14) jäljelle jää vain virheiden neliösumman minimointi.

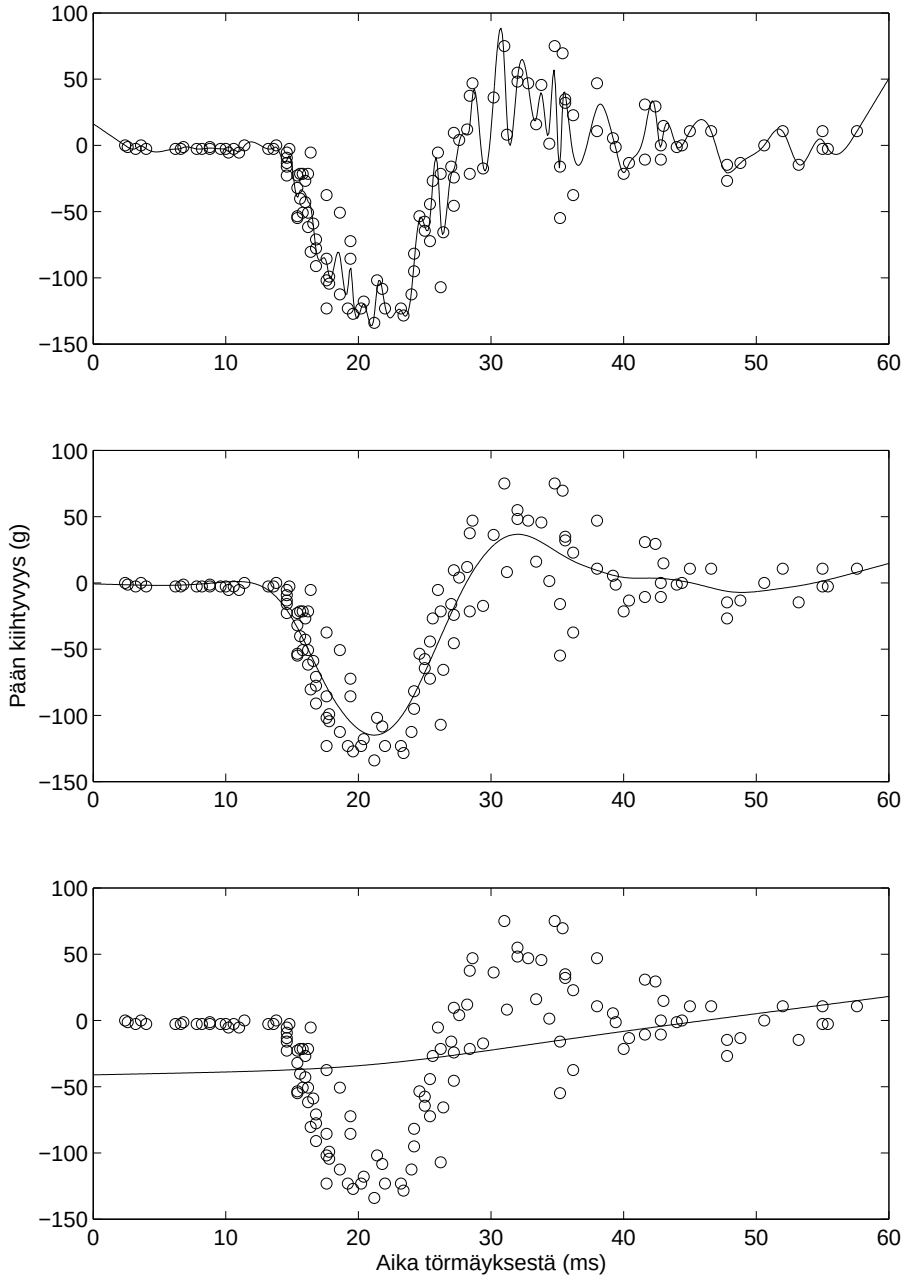
Esimerkki 4.5 Tarkastellaan jäälleen esimerkin 1.3 aineistoa ja sovelletaan siihen estimaattoria $\hat{m}_n(\cdot;\lambda)$ hakemalla minimoivaa funktiota avaruudesta $C^2([a,b])$. Tässä luvussa johdamme tälle optimointiongelmalle yleisen ratkaisun. Tälle esimerkille saatavat tulokset arvoilla $\lambda = 10^{-3}, 20$ ja 10^5 on esitety kuvassa 4.7. Arvo $\lambda = 10^{-3}$ tuottaa liki interpoloivan ratkaisun ja arvo $\lambda = 10^5$ on melkein suora. Lopputulos näyttää varsin hyvältä arvolla $\lambda = 20$. ‖

Tulemme todistamaan, että kun (4.14) minimoidaan funktioavaruudessa $C^2([a,b])$, missä $a > b$ ovat kiinteitä, on $\hat{m}_n(\cdot;\lambda)$ kuutiollinen, eli kertalukua 4 oleva splini.

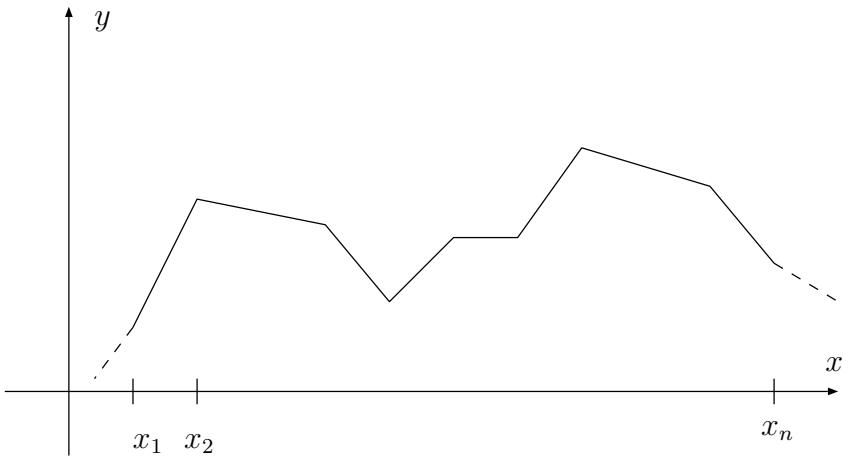
Määritelmä 4.6 *Olkoon $r \geq 2$ kokonaisluku ja $x_1 < \cdots < x_n$. Funktio $g : \mathbb{R} \to \mathbb{R}$ on* kertalukua r oleva splini solmupistein $x_1 < \cdots < x_n$, *jos*

(i) *g rajoitettuna kullekin osavälille $]-\infty, x_1[$, $[x_n, \infty[$, $[x_i, x_{i+1}[$, $i = 1, \ldots, n$, on korkeintaan astetta $r-1$ oleva polynomi,*

(ii) $g \in C^{r-2}(\mathbb{R})$.



Kuva 4.7: Esimerkin 1.3 aineiston silottaminen minimoimalla lauseke (4.14) avaruudessa $C^2([a,b])$. Käytetyt silotusparametrin arvot ovat $\lambda = 10^{-3}$ (ylin kuva), $\lambda = 20$ (keskimmäinen kuva) ja $\lambda = 10^5$ (alin kuva).



Kuva 4.8: Kertalukua 2 oleva splini, eli jatkuva murtoviiva.

Ehto (i) sanoo, että g on kertalukua r (= korkeintaan astetta $r-1$) oleva *palapolynomi*. Ehto (ii) sanoo, että g on maksimaalisen sileä palapolynomin "paloittaisen luonteen" omaava palapolynomi, sillä ehto $g \in C^{r-1}(\mathbb{R})$ pakottaisi g:n olemaan tavallinen polynomi (HT). Kuvassa 4.8 on kertalukua 2 oleva splini.

Nimitys "splini" tulee englanninkielen sanasta "spline", joka tarkoittaa joustavasta materiaalista (esim. puu, teräs) valmistettua ohutta "viivainta", jolla ennen suunniteltiin sileitä käyriä mm. laivanrakennuksessa (kuva 4.9). Spliniviivaimen määräävä käyrä $y = g(x)$ toteuttaa ehdot

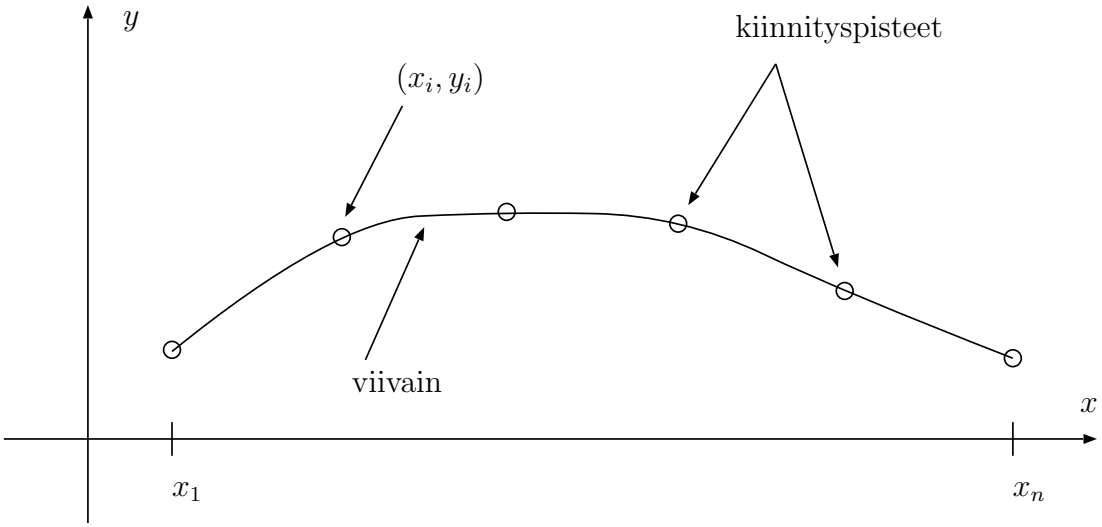
(i) $g \in C^2$,

(ii) g:llä on (ainakin likimain) pienin taivutusenergia kaikista annettujen pisteiden (x_i, y_i) kautta kulkevista käyristä.

Toisaalta voidaan osoittaa, että käyrän $y = h(x)$ kuvaaman viivaimen likimääräinen taivutusenergia on $R(h'') = \int_{x_1}^{x_n} h''(x)^2 dx$ ja ehdoilla $h(x_i) = g(x_i)$, $i = 1, \ldots, n$, $h'(x_1) = g'(x_1)$, $h'(x_n) = g'(x_n)$, suureen $R(h'')$ minimoi kuutiollinen splini. Siten käyrä $y = g(x)$ on "splini" myös matemaattisessa mielessä!

4.10 Ratkaisun olemassaolo ja yksikäsitteisyys

Olkoon $n \geq 2$ ja $a = x_1 < \cdots < x_n = b$. Määritellään



Kuva 4.9: Spliniviivain kiinnitettynä piirustuslaudan pisteisiin (x_i, y_i).

$$S = \{g \mid g \text{ on kertalukua 4 oleva (kuutiollinen) splini solmupistein } x_1, \ldots, x_n\}.$$

S on selvästi vektoriavaruus ja voidaan osoittaa (HT), että $\dim S = n+4$. Tämä on itseasiassa helppo uskoa, koska kaikkien kuutiollisten palapolynomien avaruudella on dimensio $4(n+1)$, S on tämän avaruuden $3n$:stä lineaarisesta rajoitusehdosta $(g^{(k)}(x_i - 0) = g^{(k)}(x_i + 0)$, $k = 0, 1, 2$, $i = 1, \ldots, n)$ saatava aliavaruus ja $4(n+1) - 3n = n + 4$.

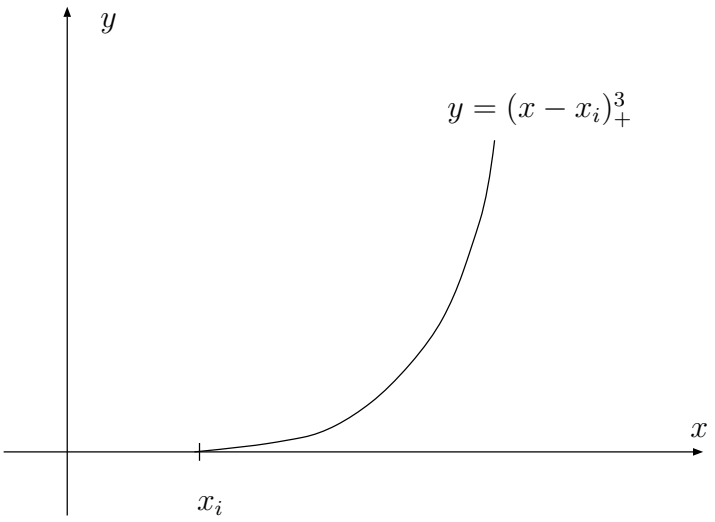
Funktiot

$$\begin{cases} x \mapsto x^k, & k = 0, 1, 2, 3, \\ x \mapsto (x - x_i)_+^3, & i = 1, \ldots, n \end{cases}$$

ovat kaksi kertaa jatkuvasti derivoituvia ja lineaarisesti riippumattomia, joten ne muodostavat S:n erään kannan (vrt. kuva 4.10). Siten kaikilla $g \in S$ on olemassa $a_1, \ldots, a_{n+4} \in \mathbb{R}$ s.e.

$$g(x) = \sum_{k=1}^{4} a_k x^{k-1} + \sum_{k=5}^{n+4} a_k (x - x_{k-4})_+^3, \quad x \in \mathbb{R}. \tag{4.15}$$

Olkoon edelleen $NS \subset S$ niiden kuutiollisten splinien joukko, jotka ovat korkeintaan astetta 1 olevia polynomeja väleillä $]-\infty, a[$ ja $]b, \infty[$. NS on solmupisteitä



Kuva 4.10: Yksi avaruuden S kantafunktioista.

$a = x_1 < \cdots < x_n = b$ vastaava *luonnollisten splinien avaruus*. Kuten S, myös NS on vektoriavaruus ja voidaan osoittaa (HT), että dim $NS = n$. Tämä lienee taas helppo uskoa, koska vaatimus lineaarisuudesta välin $[a, b]$ ulkopuolella tarkoittaa $2 + 2 = 4$:ää lineaarista rajoitusehtoa splinille $g \in S$ ja dim $S - 4 = (n + 4) - 4 = n$.

Lineaarisuusvaatimus välin $[a, b]$ ulkopuolella erityisesti merkitsee, että kehitelmässä (4.15) on $a_3 = a_4 = 0$, kun $g \in NS$.

Lemma 4.7 *Olkoon* $\{\varphi_1, \ldots \varphi_n\}$ *avaruuden NS kanta ja*

$$\varphi_j(x) = b_{0j} + b_{1j}x + \sum_{i=1}^{n} c_{ij}(x - x_i)_+^3, \quad x \in \mathbb{R}. \tag{4.16}$$

Tällöin, jos $g \in C^2([a, b])$ *ja* $s = \sum_{j=1}^{n} a_j\varphi_j \in NS$, *niin*

$$\int_a^b g''(x)s''(x)dx = 6\sum_{i=1}^{n} g(x_i)\sum_{j=1}^{n} a_j c_{ij}.$$

Todistus: Integroidaan osittain osaväleillä $[x_i, x_{i+1}]$:

$$\begin{aligned}\int_a^b g''(x)s''(x)dx &= \sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}} g''(x)s''(x)dx \\ &= \sum_{i=1}^{n-1}\left\{\int_{x_i}^{x_{i+1}} g'(x)s''(x) - \int_{x_i}^{x_{i+1}} g'(x)s'''(x)dx\right\}.\end{aligned}$$

Tässä

$$\sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} g'(x)s''(x) = g'(x_n)s''(x_n) - g'(x_1)s''(x_1),$$

koska funktio $x \mapsto g'(x)s''(x)$ on jatkuva. Edelleen, koska s'' häviää välin $[a,b] = [x_1, x_n]$ ulkopuolella ja s'' on jatkuva, tulee olla $s''(x_1) = s''(x_n) = 0$. Siten sijoitustermien summa häviää. Edelleen,

$$\begin{aligned}
-\sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} g'(x)s'''(x)dx &= -\sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} g'(x) \left[\sum_{j=1}^{n} a_j \sum_{k=1}^{i} 6c_{kj} \right] dx \\
&= -\sum_{i=1}^{n-1} [g(x_{i+1}) - g(x_i)] \left[\sum_{j=1}^{n} a_j \sum_{k=1}^{i} 6c_{kj} \right] \\
&= -6 \sum_{j=1}^{n} a_j \sum_{i=1}^{n-1} [g(x_{i+1}) - g(x_i)] \sum_{k=1}^{i} c_{kj}.
\end{aligned}$$

Tässä

$$\begin{aligned}
\sum_{i=1}^{n-1} [g(x_{i+1}) - g(x_i)] \sum_{k=1}^{i} c_{kj} &= \sum_{i=1}^{n-1} g(x_{i+1}) \sum_{k=1}^{i} c_{kj} - \sum_{i=1}^{n-1} g(x_i) \sum_{k=1}^{i} c_{kj} \\
&= \sum_{i=2}^{n} g(x_i) \sum_{k=1}^{i-1} c_{kj} - \sum_{i=1}^{n-1} g(x_i) \sum_{k=1}^{i} c_{kj} \\
&= g(x_n) \sum_{k=1}^{n} c_{kj} - g(x_1)c_{1j} - \sum_{i=2}^{n} g(x_i)c_{ij} \\
&= g(x_n) \sum_{k=1}^{n} c_{kj} - \sum_{i=1}^{n} g(x_i)c_{ij},
\end{aligned}$$

missä kolmannessa yhtälössä kirjoitettiin $\sum_{k=1}^{i-1} c_{kj} = \sum_{k=1}^{i} c_{kj} - c_{ij}$. Mutta $\sum_{k=1}^{n} c_{kj} = (1/6)\varphi_j'''(x)$, kun $x > x_n$ ja $\varphi_j'''(x) = 0$, kun $x > x_n$, koska $\varphi_j \in NS$. Siten $\sum_{k=1}^{n} c_{kj} = 0$ ja saamme lopulta

$$\int_a^b g''(x)s''(x)dx = -6 \sum_{i=1}^{n} a_j \left[-\sum_{j=1}^{n} g(x_i)c_{ij} \right] = 6 \sum_{i=1}^{n} g(x_i) \sum_{j=1}^{n} a_j c_{ij}. \qquad \square$$

Tulemme todistamaan, että (4.14):lla on yksikäsitteinen minimi $g^* = \hat{m}_n(\cdot;\lambda)$ avaruudessa $C^2([a,b])$ ja että itseasiassa $\hat{m}_n(\cdot;\lambda) \in NS \mid [a,b]$ eli $\hat{m}_n(\cdot;\lambda) = s \mid [a,b]$ eräällä $s \in NS$. Tätä minimiratkaisua sanotaan silotusparametria λ vastaavaksi *silottavaksi spliniksi*. Tietäen tämän olemassaolotuloksen voimme jo ennakkoon johtaa kaavan estimaattorille $\hat{m}_n(\cdot;\lambda)$.

Tulemme suorittamaan tarkastelut kiinteillä $X_1, \ldots, X_n$ eli käsittelemme itseasiassa kiinteätä asetelmaa. Olkoon siis $a = x_1 < \cdots < x_n = b$ kuten edellä ja olkoon $(x_1, Y_1), \ldots, (x_n, Y_n)$ satunnaisotos. Olkoon

$$J(g) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - g(x_i)]^2 + \lambda \int_a^b g''(x)^2 dx \tag{4.17}$$

ja $g_1, g_2 \in C^2([a,b])$, $\delta \in \mathbb{R}$. Silloin

$$J(g_1 + \delta g_2) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - g_1(x_i) - \delta g_2(x_i)]^2 + \lambda \int_a^b [g_1''(x) + \delta g_2''(x)]^2 dx \tag{4.18}$$

ja $\delta \mapsto J(g_1 + \delta g_2)$ on δ:n toisen asteen polynomi. Funktionaalin $J : C^2([a,b]) \to \mathbb{R}$ *Gâteaux derivaatta* pisteessä g_1 *suuntaan* g_2 on silloin

$$J'(g_1)[g_2] = \lim_{\delta \to 0} \frac{J(g_1 + \delta g_2) - J(g_1)}{\delta} = \frac{d}{d\delta} J(g_1 + \delta g_2)\bigg|_{\delta=0}.$$

Olkoon nyt g^* $J(g)$:n minimoiva funktio avaruudessa $C^2([a,b])$. Silloin kaikilla $g \in C^2([a,b])$, $\delta \in \mathbb{R}$, pätee

$$J(g^* + \delta g) \geq J(g^*)$$

joten $J'(g^*)[g] = 0$. Käyttäen kaavaa (4.18) saadaan

$$\frac{d}{d\delta} J(g^* + \delta g) = -\frac{2}{n}\sum_{i=1}^{n}[Y_i - g^*(x_i) - \delta g(x_i)]g(x_i) + 2\lambda \int_a^b [(g^*)''(x) + \delta g''(x)]g''(x)dx$$

josta asettamalla $\delta = 0$ saadaan

$$J'(g^*)[g] = -\frac{2}{n}\sum_{i=1}^{n}[Y_i - g^*(x_i)]g(x_i) + 2\lambda \int_a^b (g^*)''(x)g''(x)dx.$$

Koska toisaalta $J'(g^*)[g] = 0$, saamme ehdon

$$\frac{1}{n\lambda}\sum_{i=1}^{n}[Y_i - g^*(x_i)]g(x_i) = \int_a^b (g^*)''(x)g''(x)dx. \tag{4.19}$$

Olkoon sitten $\{\varphi_1, \ldots, \varphi_n\} \subset NS$ kanta ja b_{0j}, b_{1j}, c_{ij} kuten kaavassa (4.16). Koska tiedetään, että $g^* \in NS \mid [a,b]$, on olemassa $a_1^*, \ldots, a_n^* \in \mathbb{R}$ s.e.

$$g^*(x) = \sum_{j=1}^{n} a_j^* \varphi_j(x), \quad x \in [a,b].$$

Sovelletaan lemmaa 4.7, kun $s = \sum_{j=1}^{n} a_j^* \varphi_j$, jolloin saadaan

$$\int_a^b (g^*)''(x) g''(x) dx = 6 \sum_{i=1}^{n} g(x_i) \sum_{j=1}^{n} a_j^* c_{ij}.$$

Siten (4.19):n kanssa yhtäpitävä ehto on

$$\frac{1}{n\lambda} \sum_{i=1}^{n} [Y_i - \sum_{j=1}^{n} a_j^* \varphi_j(x)] g(x_i) = 6 \sum_{i=1}^{n} g(x_i) \sum_{j=1}^{n} a_j^* c_{ij}.$$

Koska tämä pätee kaikilla $g \in C^2([a,b])$, saadaan

$$\frac{1}{n\lambda} [Y_i - \sum_{j=1}^{n} a_j^* \varphi_j(x)] = 6 \sum_{j=1}^{n} a_j^* c_{ij}, \quad i = 1, \ldots, n$$

eli

$$\sum_{j=1}^{n} [\varphi_j(x_i) + 6n\lambda c_{ij}] a_j^* = Y_i, \quad i = 1, \ldots, n$$

tai

$$(\Phi + n\lambda G) a^* = y, \tag{4.20}$$

missä

$$\begin{cases} \Phi = [\varphi_j(x_i)] \in \mathbb{R}^{n \times n}, \\ G = [6c_{ij}] \in \mathbb{R}^{n \times n}, \\ a^* = [a_1^*, \ldots, a_n^*]^T \in \mathbb{R}^n, \\ y = [Y_1, \ldots, Y_n]^T \in \mathbb{R}^n \end{cases}$$

Kerroinvektori a^* ratkeaa yhtälöstä (4.20) yksikäsitteisesti kunhan osoitamme, että matriisi $\Phi + n\lambda G$ on kääntyvä. Näytämme, että ehdosta $(\Phi + n\lambda G)a = 0$ seuraa, että $a = 0$. Jos $(\Phi + n\lambda G)a = 0$, pätee (4.19) kaikilla $g \in C^2([a,b])$, kun $g^* = \sum_{j=1}^{n} a_j \varphi_j$ ja $Y_1 = \cdots = Y_n = 0$. Otetaan erityisesti $g = g^*$ jolloin saamme

$$\frac{1}{n\lambda} \sum_{i=1}^{n} [g^*(x_i)]^2 + \int_a^b [(g^*)''(x)]^2 dx = 0.$$

Ehdosta $\int_a^b [(g^*)''(x)]^2 dx = 0$ seuraa, että g^* on ensimmäisen asteen polynomi ja ehdon $\sum_{i=1}^{n} [g^*(x_i)]^2 = 0$ perusteella $g^* = 0$, koska $n \geq 2$ ja g^* siis häviää ainakin kahdessa pisteessä. Siten $a_1 = \cdots = a_n = 0$.

Lause 4.8 *Lausekkeen (4.17) minimoija g^* avaruudessa $C^2([a,b])$ on yksikäsitteinen. Itseasiassa $g^* \in NS \mid [a,b]$ ja jos $\{\varphi_1, \ldots, \varphi_n\} \subset NS$ on kanta, on $g^*(x) = \sum_{j=1}^n a_j^* \varphi_j(x)$, $x \in [a,b]$, missä $a^* = [a_1^*, \ldots, a_n^*]^T$ on yhtälön (4.20) yksikäsitteinen ratkaisu.*

Todistus: Näytetään ensin, että funktio $g^* = \sum_{j=1}^n a_j^* \varphi_j$ minimoi (4.17):n. Olkoon $g \in C^2([a,b])$ mielivaltainen. Silloin

$$\begin{aligned}
J(g) &= \frac{1}{n}\sum_{i=1}^n [Y_i - g(x_i)]^2 + \lambda \int_a^b g''(x)^2 dx \\
&= \frac{1}{n}\sum_{i=1}^n [(Y_i - g^*(x_i)) + (g^*(x_i) - g(x_i))]^2 + \lambda \int_a^b [(g^*)''(x) + (g''(x) - (g^*)''(x))]^2 dx \\
&= J(g^*) + \frac{2}{n}\sum_{i=1}^n [Y_i - g^*(x_i)][g^*(x_i) - g(x_i)] \\
&+ 2\lambda \int_a^b (g^*)''(x)[g''(x) - (g^*)''(x)]dx + \frac{1}{n}\sum_{i=1}^n [g^*(x_i) - g(x_i)]^2 \\
&+ \lambda \int_a^b [g''(x) - (g^*)''(x)]^2 dx.
\end{aligned}$$

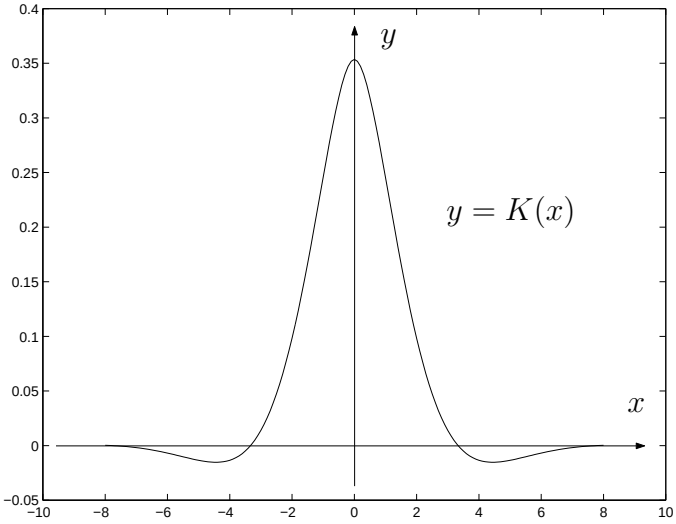
Tässä yhtälön oikean puolen toinen ja kolmas termi ovat yhteensä $J'(g^*)[g - g^*]$ ja tämä Gâteaux derivaatta häviää (4.19):n nojalla (g:n paikalla $g - g^*$). Edelleen, neljäs ja viides termi ovat ei-negatiivisia, joten $J(g) \geq J(g^*)$. Siten g^* on minimoiva funktio.

Sitten osoitetaan ratkaisun yksikäsitteisyys. Olkoon $\tilde{g}$ toinen (4.17):n minimoiva funktio. Edellisen laskun neljäs ja viides termi häviävät silloin kun $g = \tilde{g}$, eli

$$\int_a^b [\tilde{g}''(x) - (g^*)''(x)]^2 dx = 0,$$

$$\sum_{i=1}^n [g^*(x_i) - \tilde{g}(x_i)]^2 = 0.$$

Siten $(\tilde{g} - g^*)''(x) = 0$ kaikilla $x \in [a,b]$ eli $\tilde{g} - g^*$ on ensimmäisen asteen polynomi. Koska toisaalta $\tilde{g} - g^*$ häviää n:ssä pisteessä x_i ja $n \geq 2$, täytyy sen hävitä koko välillä $[a,b]$ eli $\tilde{g} = g^*$. □



Kuva 4.11: Silottavaan spliniin liittyvä kertalukua 4 oleva ydin K.

4.11 Yhteys ydinregressioon

Yhtälöstä (4.20) nähdään, että (4.17):n minimoivan silottavan splinin $\hat{m}_n(\cdot;\lambda)$ arvo pisteessä x riippuu lineaarisesti otospisteistä $Y_1,\ldots,Y_n$:

$$a^* = (\Phi + n\lambda G)^{-1} y, \quad y = [Y_1,\ldots,Y_n]^T,$$

$$\hat{m}_n(x;\lambda) = \sum_{j=1}^{n} a_j^* \varphi_j(x) = \frac{1}{n}\sum_{i=1}^{n} W_i(x,\lambda,x_1,\ldots,x_n) Y_i,$$

eräillä painofunktioilla $W_i(\cdot,\lambda,x_1,\ldots,x_n)$. Voidaan osoittaa, että kun n on suuri, $\lambda > 0$ pieni ja x_i ei ole lähellä päätepisteitä a ja b, on

$$W_i(x,\lambda,x_1,\ldots,x_n) \approx \frac{1}{f(x_i)}\frac{1}{h(x_i)} K\left(\frac{x-x_i}{h(x_i)}\right),$$

missä f on tiheysfunktio, josta pisteet x_i on saatu i.i.d. otoksena ja

$$\begin{cases} h(x_i) = \{\lambda/[nf(x_i)]\}^{1/4}, \\ K(x) = (1/2)e^{-|x|/\sqrt{2}} \sin(|x|/\sqrt{2} + \pi/4). \end{cases}$$

K on itseasiassa kertalukua 4 oleva ydin ($\int_{-\infty}^{\infty} x^2 K(x)dx = 0$) (kuva 4.11). Siten $\hat{m}_n(\cdot;\lambda)$ on likimain adaptiivinen ydinregressioestimaattori!

4.12 Silotusparametrin määrääminen

Kuten muidenkin menetelmien kohdalla, voidaan nytkin tietysti tyytyä valitsemaan λ subjektiivisesti tai joissain tapauksissa sovelluskohtaisen optimointikriteerin perusteella.

Automaattisista menetelmistä mainittakoon ristiinvalidointi ja yleistetty ristiinvalidointi. Kummassakin menetelmässä minimoidaan

$$\mathbb{E}[[\hat{m}_n(X;\lambda) - m(X)]^2]$$

missä siis estimaattorissa $\hat{m}_n(\cdot;\lambda)$ olevat $x_1, \ldots, x_n$ ovat kiinteät otospisteet tiheysfunktiosta f ja $X \sim f$ (vrt. luku 4.7). Odotusarvo lasketaan X:n ja otoksen $Y_1, \ldots, Y_n$ suhteen.

Ristiinvalidoinnissa minimioidaan suure

$$\mathrm{LSCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}_{n,-i}(x_i;\lambda)]^2,$$

missä $\hat{m}_{n,-i}(\cdot;\lambda)$ on muodostettu jättämällä pari (x_i, Y_i) pois otoksesta (vrt. luku 4.7).

Yleistetyn ristiinvalidoinnin määrittelemiseksi palataan vielä kaavaan (4.20), joka kerrotaan puolittain Φ^T:llä:

$$(\Phi^T\Phi + n\lambda\Omega)a^* = \Phi^T y,$$

missä $\Omega = \Phi^T G$. Matriisin Ω alkio $[\Omega]_{ij}$ voidaan lemman 4.7 nojalla kirjoittaa muotoon

$$[\Omega]_{ij} = 6\sum_{k=1}^{n}\varphi_i(x_k)c_{kj} = \int_a^b \varphi_i''(x)\varphi_j''(x)dx.$$

Tästä seuraa helposti (HT), että matriisi $\Phi^T\Phi + n\lambda\Omega$ on positiivisesti definiitti. Erityisesti on siis olemassa $(\Phi^T\Phi + n\lambda\Omega)^{-1}$ ja $\hat{m}_n(\cdot;\lambda)$:n arvot pisteissä $x_1, \ldots, x_n$ saadaan kaavasta

$$\begin{aligned}[\hat{m}_n(x_1;\lambda), \ldots, \hat{m}_n(x_n;\lambda)]^T &= \Phi a^* \\ &= \Phi(\Phi^T\Phi + n\lambda\Omega)^{-1}\Phi^T y \\ &= H(\lambda)y,\end{aligned}$$

missä $H(\lambda) = \Phi(\Phi^T\Phi + n\lambda\Omega)^{-1}\Phi^T$. Tämä kaava osoittaa, että silottava splini on eräänlainen yleistetty harjanneregressiomenetelmä (engl. ridge regression), joka on lineaarisen regression eräs muunnelma.

Yleistetyssä ristiinvalidoinnissa minimoidaan

$$\mathrm{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}_n(x_i;\lambda)]^2}{\left[\frac{1}{n}\mathrm{tr}\{I_n - H(\lambda)\}\right]^2},$$

missä $I_n \in \mathbb{R}^{n\times n}$ on yksikkömatriisi ja tr tarkoittaa matriisin jälkeä. Voidaan osoittaa (HT), että

$$\mathrm{LSCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{m}_n(x_i;\lambda)}{1 - [H(\lambda)]_{ii}}\right]^2.$$

Toisaalta GCV(λ):ssa

$$\frac{1}{n}\mathrm{tr}\{I_n - H(\lambda)\} = \frac{1}{n}\sum_{i=1}^{n}[1 - [H(\lambda)]_{ii}] = 1 - \frac{1}{n}\sum_{i=1}^{n}[H(\lambda)]_{ii},$$

joten GCV(λ) saadaan LSCV(λ):sta korvaamalla $[H(\lambda)]_{ii}$ keskiarvollaan. Näin LSCV(λ) ja GCV(λ) ovat sukua toisilleen.

Lisää materiaalia silottavista splineistä löytyy lähteistä [4] ja [11]. Hyvä lähde implementoinnin käytännön toteutukseen on [6].

4.13 Ortogonaalisarjakehitelmät

Olkoon $D \subset \mathbb{R}$ ja tarkastellaan kiinteää asetelmaa, $\{x_1, \ldots, x_n\} \subset D$,

$$Y_i = m(x_i) + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

missä $\sigma > 0$ on vakio, $\varepsilon_1, \ldots, \varepsilon_n \sim \varepsilon$ on i.i.d. otos ja $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}[\varepsilon] = 1$. Oletetaan edelleen, että $m \in L^2(D)$ ja että $(\varphi_k)_{k\in\mathbb{N}}$ on avaruuden $L^2(D)$ ortonormaali kanta. Silloin

$$m = \sum_{k=1}^{\infty} a_k\varphi_k \tag{4.21}$$

eräillä $a_k \in \mathbb{R}$, $k \in \mathbb{N}$. Itseasiassa

$$a_k = \int_D \varphi_k(x) m(x) dx.$$

Miten kertoimia a_k voidaan estimoida?

Olkoon erityisesti $D = [0,1]$ ja $x_i = i/n$, $i = 1, \ldots, n$. Silloin

$$a_k = \int_0^1 \varphi_k(x) m(x) dx \approx \sum_{i=1}^n \varphi_k(x_i) m(x_i) \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n \varphi_k(x_i) \mathbb{E} Y_i = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \varphi_k(x_i) Y_i \right],$$

missä kolmannessa yhtäsuuruudessa käytettiin ehtoa $\mathbb{E}\varepsilon_i = 0$. Siten on luontevaa ottaa

$$\hat{a}_{kn} = \frac{1}{n} \sum_{i=1}^n \varphi_k(x_i) Y_i, \tag{4.22}$$

jolloin $\mathbb{E}\hat{a}_{kn} \approx a_k$. Tästä saadaan *ortogonaalisarjaestimaattori*

$$\hat{m}_n(x; N) = \sum_{k=1}^N \hat{a}_{kn} \varphi_k(x), \quad x \in [0,1].$$

Tässä N:llä on silotusparametrin rooli ja, kuten tiheysfunktion estimoinnissa, tulee se nytkin ottaa otoskoosta riippuvaksi: $N = N_n$ s.e. N_n kasvaa rajatta, mutta ei liian nopeasti, kun $n \to \infty$.

Tämä menetelmä toimii hyvin, jos m:n kehitelmän (4.21) tärkeimmät termit ovat alkupäässä s.o. vastaavat pieniä k:n arvoja. Jos näin ei ehkä ole, voidaan yrittää *kynnystämistä*: valitaan $\delta > 0$ ja otetaan

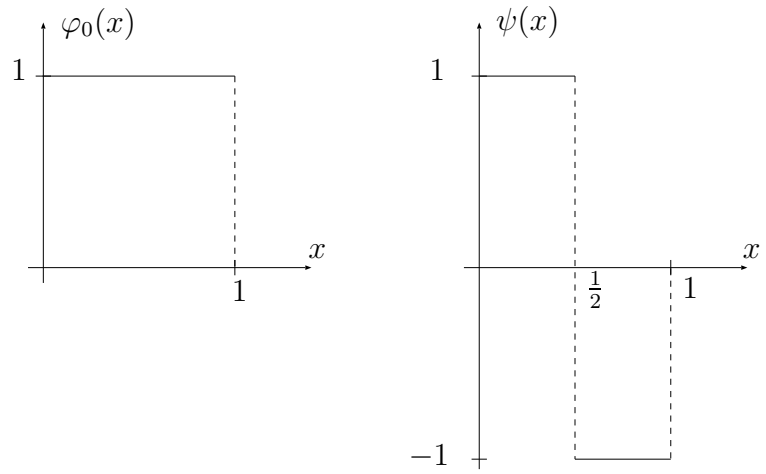
$$\hat{m}_n(x; \delta) = \sum_{k=1}^\infty \hat{a}_{kn} 1_{[\delta,\infty[}(|\hat{a}_{kn}|) \varphi_k(x) \qquad \text{ns. "kova" kynnystys,} \tag{4.23}$$

tai

$$\hat{m}_n(x; \delta) = \sum_{k=1}^\infty \operatorname{sgn}(\hat{a}_{kn})(|\hat{a}_{kn}| - \delta)_+ \varphi_k(x) \qquad \text{ns. "pehmeä" kynnystys,} \tag{4.24}$$

missä $1_{[\delta,\infty[}$ on välin $[\delta, \infty[$ karakteristinen funktio ja $\operatorname{sgn}(\hat{a}_{kn})$ on kertoimen $\hat{a}_{kn}$ merkki.

Fourier-kannan (vrt. esimerkki 2.15) käyttö sopii parhaiten sileille funktiolle, joissa ei ole teräviä paikallisia heilahteluja. Epäsäännöllisimmille funktioille voidaan



Kuva 4.12: Haarin aallokkeita. Vasemmalla isä-aalloke ja oikealla äiti-aalloke.

parempaan tulokseen päästä *aallokkeilla* (engl. wavelets). Aallokkeiden käytön suosio tilastollisessa estimoinnissa on ollut voimakkaassa kasvussa viimeisen kymmenen vuoden aikana.

Tarkastellaan avaruutta $L^2([0,1])$. Yksinkertaisin esimerkki aallokkeista ovat ns. *Haarin aallokkeet*. Ensin määritellään ns. isä- ja äiti-aallokkeet (kuva 4.12),

$$\varphi_0(x) = 1, \quad x \in [0,1] \qquad \text{"isä-aalloke"},$$

$$\psi(x) = \begin{cases} 1, & x \in [0, 1/2[, \\ -1, & x \in [1/2, 1] \\ 0 & \text{muulloin.} \end{cases} \qquad \text{"äiti-aalloke"}.$$

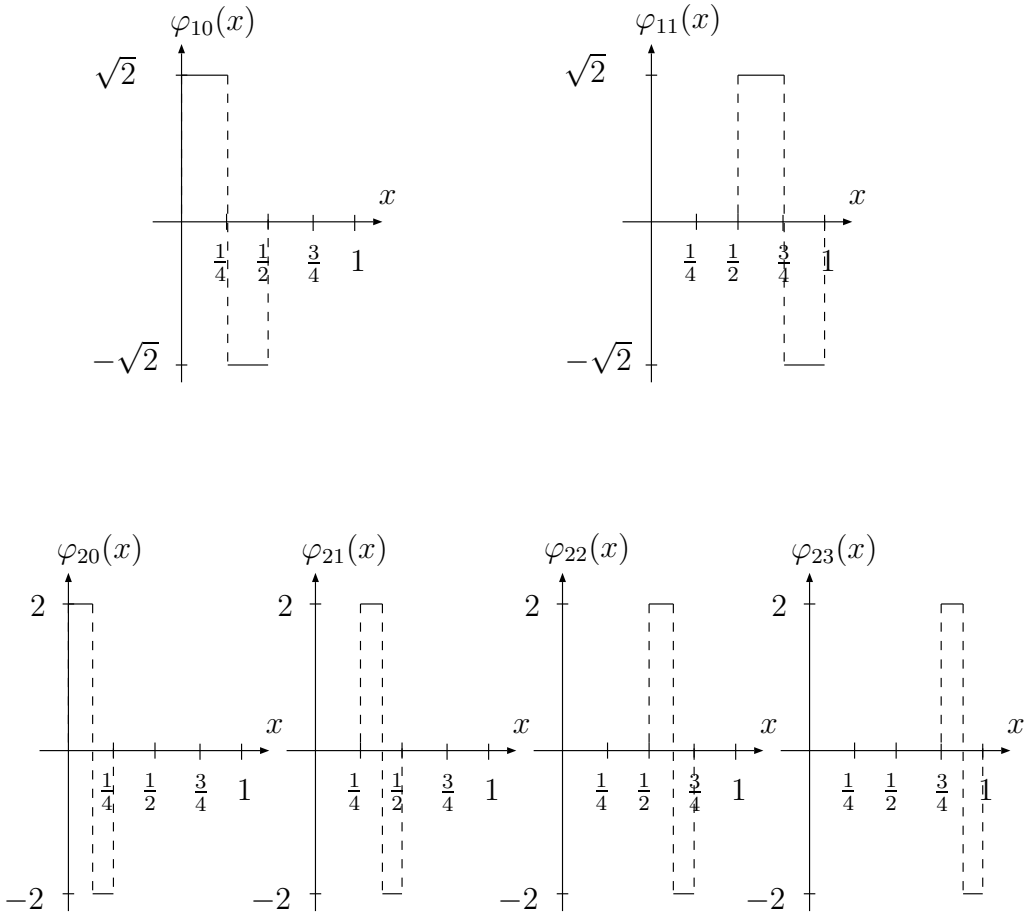
Äiti-aallokkeesta sitten generoidaan loput kantafunktiot dilataatioilla (skaalauksella) ja translaatioilla,

$$\varphi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \quad j = 0, 1, 2, \ldots, k = 0, 1, \ldots, 2^j - 1, \quad x \in [0,1].$$

Kuvassa 4.13 on esitetty arvoilla $j = 1$ ja $j = 2$ saatavat aallokkeet. Indeksi j numeroi *resoluutiotasoja*. Valitsemalla j suureksi päästään esittäämään nopeita vaihteluita estimoitavassa funktiossa.

Voidaan osoittaa, että

$$\{\varphi_0\} \cup \{\varphi_{jk} \mid j = 0, 1, 2, \ldots, k = 0, 1, \ldots, 2^j - 1\}$$



Kuva 4.13: Haarin aallokkeita. Ylärivissä dilataatiota $j = 1$ vastaavat kantafunktiot φ_{10} ja φ_{11}. Alarivissä dilataatiota $j = 2$ vastaavat kantafunktiot $\varphi_{20}, \varphi_{21}, \varphi_{22}$ ja φ_{23}.

on avaruuden $L^2([0,1])$ ortonormaali kanta. Regressiofunktion $m \in L^2([0,1])$ aallokekehtitelmässä

$$m = a_0\varphi_0 + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} a_{jk}\varphi_{jk}$$

voidaan kertoimet a_0 ja a_{jk} estimoida kuten kaavassa (4.22). Suuri j:n arvo vastaa kantafunktiota, joka häviää hyvin pienen välin ulkopuolella. Silloin estimaattorin $\hat{a}_{jkn}$ lausekkeessa (vrt. (4.22)) on vain vähän termejä, joten saatava estimaatti voi olla huono. Siksi valitaan jokin maksimiresoluutiotaso J ja asetetaan $\hat{a}_{jkn} = 0$, kun $j > J$. Sitten voidaan menetellä kuten (4.23):ssa tai (4.24):ssä ja ottaa

$$\hat{m}_n(x;\delta) = \hat{a}_{0n}\varphi_0(x) + \sum_{j=0}^{J}\sum_{k=0}^{2^j-1} \hat{a}_{jkn}1_{[\delta,\infty[}(|\hat{a}_{jkn}|)\varphi_{jk}(x)$$

tai

$$\hat{m}_n(x;\delta) = \hat{a}_{0n}\varphi_0(x) + \sum_{j=0}^{J}\sum_{k=0}^{2^j-1} \mathrm{sgn}(\hat{a}_{jkn})(|\hat{a}_{jkn}| - \delta)_+\varphi_{jk}(x).$$

Parametrille δ voidaan teorian perusteella ehdottaa hyviä arvoja.

Esimerkki 4.9 Kuvissa 4.14 ja 4.15 on kaksi esimerkkiä funktion esittämisestä Haarin kannan avulla. Ideaalista aallokehitelmää on approksimoitu ns. diskreetillä aallokemuunnoksella. Tämän ja muiden aalloke esimerkkien laskut on suoritettu käyttäen Matlabille ilmaiseksi saatavalla WaveLab ohjelmistolla.

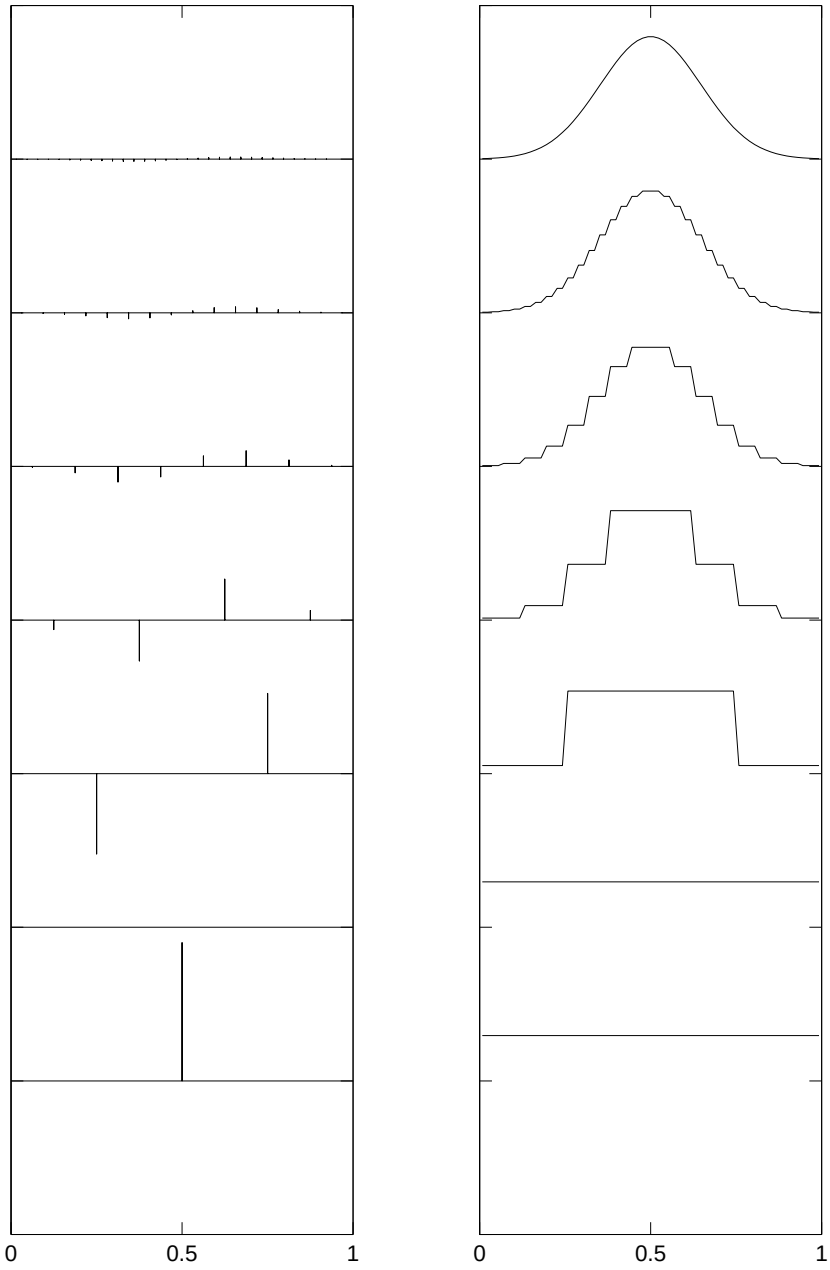
Kuvassa 4.14 on tarkasteltu normaalijakauman $N(0.5, 0.15^2)$ tiheysfunktion rajoittumaa välille $[0,1]$ diskretoituna 64 tasaväliseen pisteeseen. Kuvassa 4.15 on tarkasteltu vastaavasti jakauman $N(0.5, 0.02^2)$ tiheysfunktiota. Maksimiresoluutiotaso on $J = 5$ ja mitään kynnystystä ei ole käytetty ($\delta = 0$). Aallokkeiden kertoimet on esitetty pystyviivoilla, jotka on sijoitettu kohtiin, joissa vastaavat kantafunktiot likimain sijaitsevat. Alimpana on isä-aallokkeen kerroin ja sen jälkeen resoluutiotasoja $j = 0, \ldots, 5$ vastaavat kertoimet. Pystysuorien viivojen korkeudet on piirretty kertoimien koon mukaisesti. Huomaa, että resoluutiotasoa $j = 0$ vastaava kerroin häviää symmetrisyyden johdosta. Oikella on esitetty estimoitavan funktion sarjakehitelmän tarkentuminen summattavia resoluutiotasoja lisättäessä. Ylimpänä alkuperäinen virheetön funktio, jonka diskreetti aallokemuunnos rekonstruoi täydellisesti, kun otetaan mukaan kaikki resoluutiotasot $j = 0, \ldots, 5$.

Vertailun vuoksi on kuvassa 4.16 esitetty jakauman $N(0.5, 0.02^2)$ tiheysfunktion kehittäminen esimerkin 2.15 ortonormaalissa Fourier kannassa käyttäen jälleen välin $[0, 1]$ tasavälistä jakoa 64 pisteeseen. Kuvan 4.15 mukaan otettaessa aallokekehitelmään mukaan 13 kantafunktiota saadaan alkuperäinen funktio esitettyä käytännössä virheettömästi. Sama määrä esimerkin 2.15 kantafunktioita antaa kuitenkin kuvan 4.16 mukaisen paljon huonomman tuloksen. Aallokkeiden lokaalisti rajattu vaikutus mahdollistaa nopeasti muuttuvien funktioiden tehokkaan esittämisen, kun taas sin- ja cos-funktioista muodostetuilla Fourier kantafunktioilla ei tällaista lokaalisuutta ole. ‖

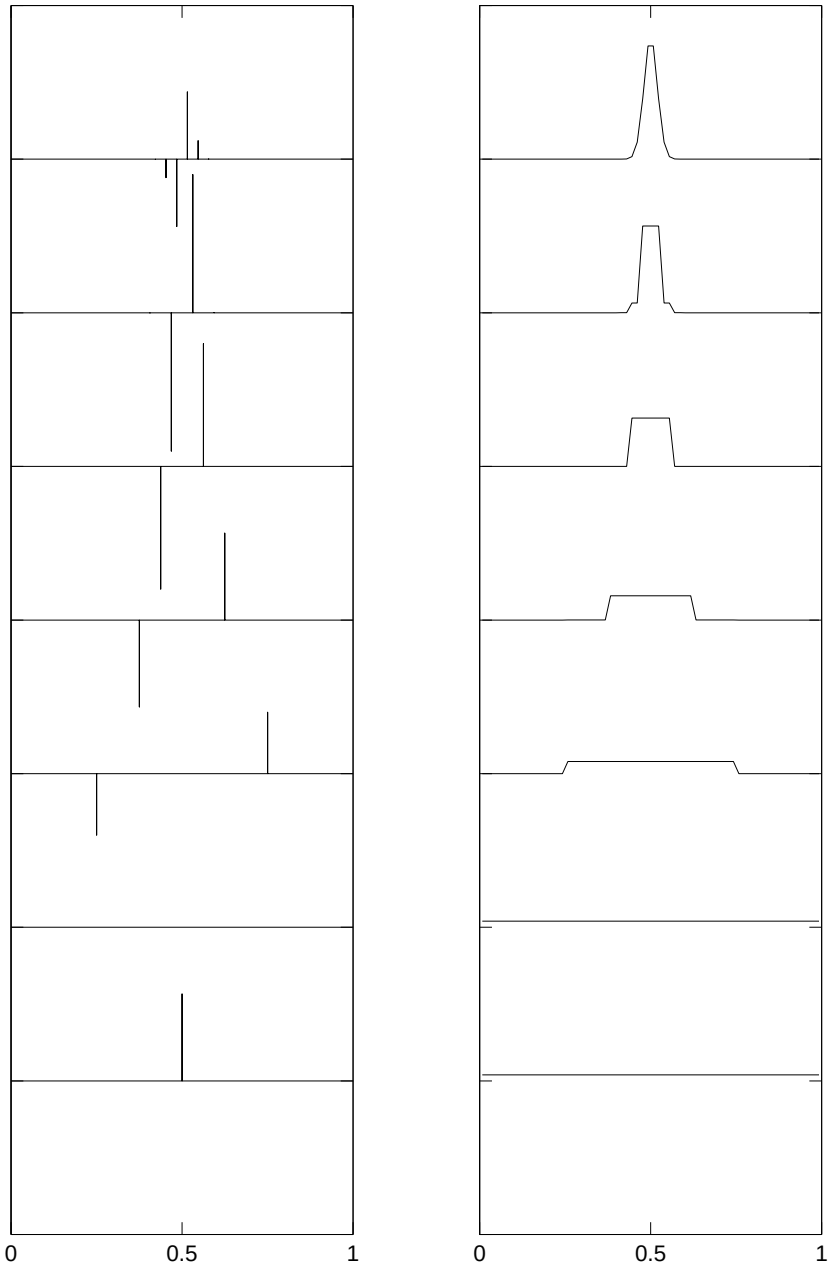
Haarin kanta tuottaa paloittain vakioita, epäsileitä estimaatteja. Sileisiin estimaatteihin päästään käyttämällä sileitä isä- ja äiti-aallokkeita. Sileille äiti-aallokkeille ei kuitenkaaan ole voitu esittää yksinkertaisia analyyttisiä kaavoja.

Esimerkki 4.10 Kuvassa 4.17 on esitetty eräitä jatkuvia aallokkeita, joista voidaan konstruoida ortonormaali kanta avaruuteen $L^2(\mathbb{R})$. Ylärivissä on Daubechien kertalukua 4 vastaavat isä- ja äiti-aallokkeet. Tälle äiti-aallokkeelle pätee $\int \psi(x)dx = \int x\psi(x)dx = 0$. Alarivissä on kertalukua 3 olevat isä ja äiti coiflet-aallokkeet. Isälle pätee $\int x^k\varphi(x)dx = 0$, kun $k = 1, \ldots, 5$ ja äidille pätee $\int x^k\psi(x)dx = 0$, kun $k = 0, \ldots, 5$. Osoittautuu, että aivan kuten korkeamman kertaluvun ydinestimaattoreillakin, momenttien häviäminen johtaa tehokkaampiin estimaattoreihin. ‖

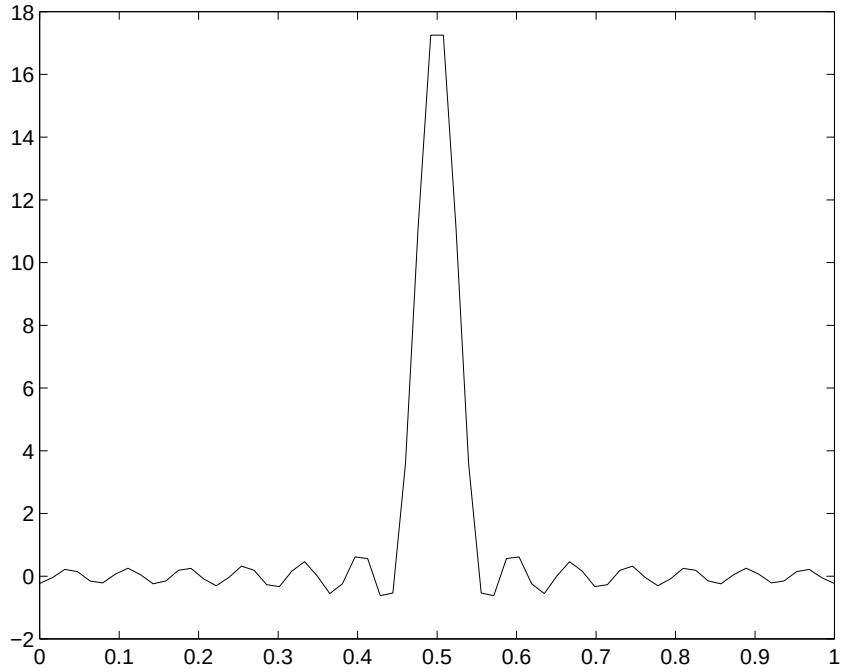
Esimerkki 4.11 Kuvassa 4.18 on analysoitu seisminen signaali aallokkeiden avulla. Käytetty aineisto on yksi WaveLab ohjelmiston esimerkeistä. Yläkuvassa alkuperäinen signaali ja toiseksi alimmaisessa kuvassa kertalukua 3 olevilla coiflet-aallokkeilla tehty analyysi. Aallokkeiden kertoimia on yhteensä 1024 kappaletta. Alimmaisessa kuvassa kovan kynnystämisen ($\delta = 0.0443$) jälkeen jäljelle jäävät kertoimet ja toiseksi ylimmäisessä niiden avulla saatava alkuperäisen signaalin estimaatti. Jäljelle jääneitä kertoimia on 100 kappaletta ja lopputulosta on vaikea erottaa alkuperäisestä. Keskimmäisessä kuvassa on alkuperäisen ja estimoidun signaalin erotus. ‖



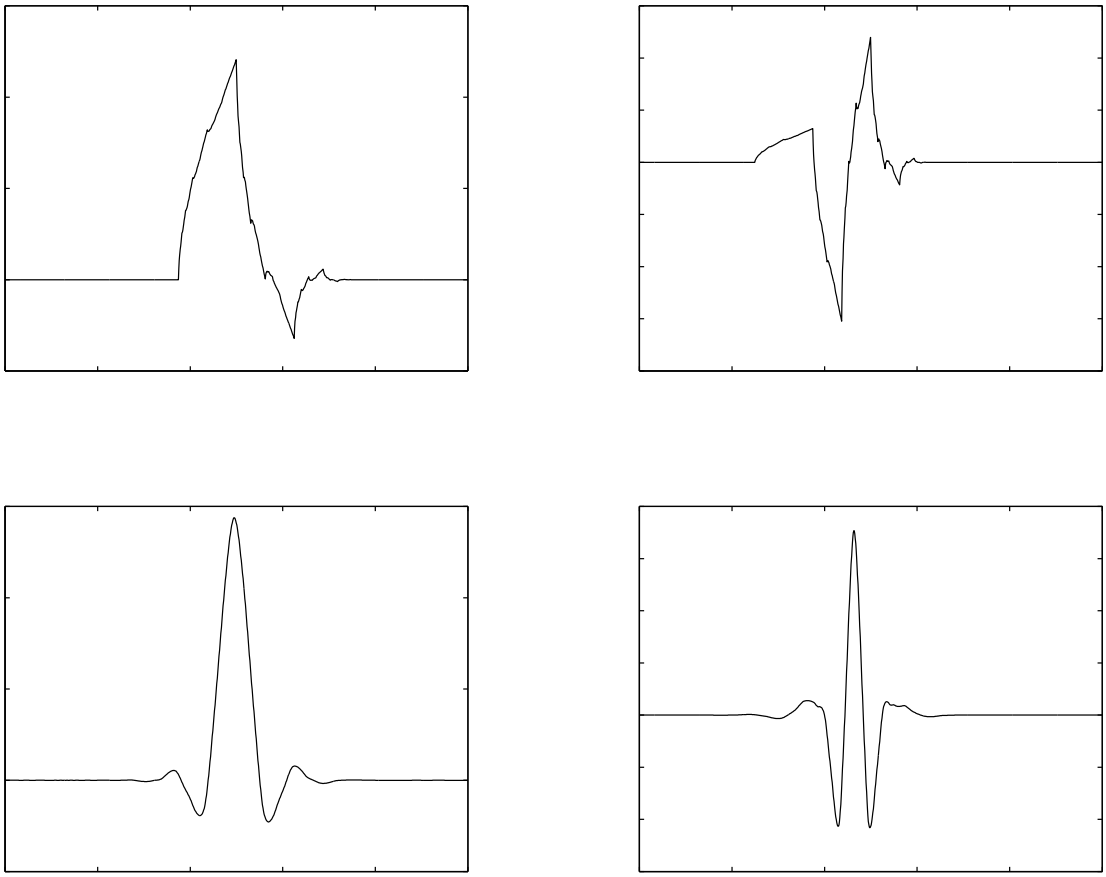
Kuva 4.14: Normaalijakauman $N(0.5, 0.15^2)$ tiheysfunktion analyysi diskreetin aallokemuunnoksen avulla.



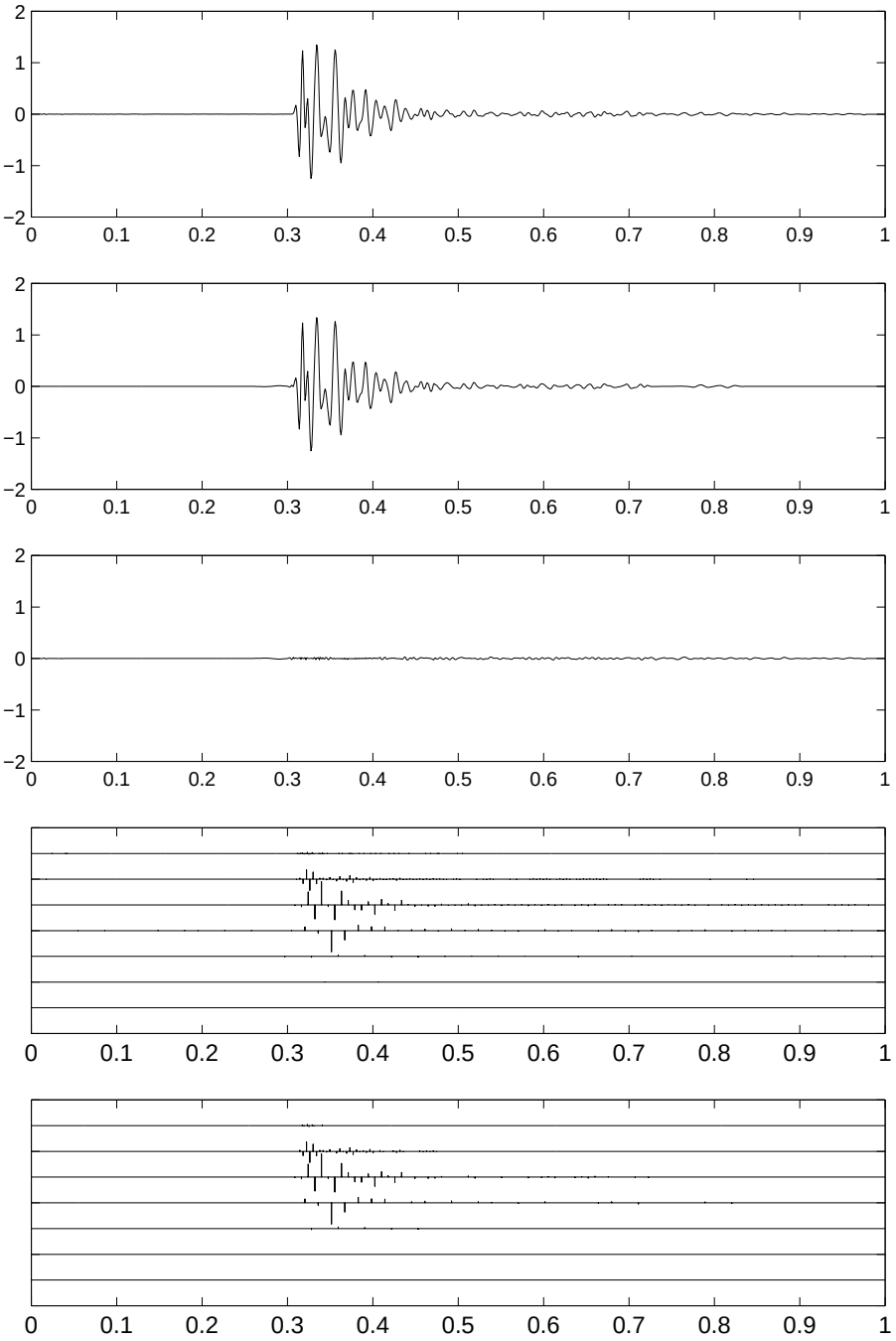
Kuva 4.15: Normaalijakauman $N(0.5, 0.02^2)$ tiheysfunktion analyysi diskreetin aallokemuunnoksen avulla.



Kuva 4.16: Normaalijakauman $N(0.5, 0.02^2)$ tiheysfunktion analyysi käyttäen Fourier kehitelmää, jossa on 13 termiä.



Kuva 4.17: Esimerkkejä aallokkeista. Ylärivissä Daubechien kertalukua 4 olevat isä- ja äiti-aallokkeet ja alarivissä kertalukua 3 olevat isä ja äiti coiflet-aallokkeet.



Kuva 4.18: Seismisen signaalin (yläkuva) analyysi kertalukua 3 olevilla coiflet-aallokkeilla. Kovalla kynnystämisellä on alkuperäisistä kertoimista (toiseksi alimmainen kuva, kertoimia 1024 kappaletta) saatu 100 kerrointa (alin kuva), joiden avulla on suoritettu alkuperäisen signaalin rekonstruktio (toiseksi ylin kuva). Alkupaeräisen signaalin ja rekonstruktion erotus on keskimmäisessä kuvassa.

Kirjallisuusviitteet

[1] L. Devroye. *A Course in Density Estimation.* Birkhäuser, Boston, 1987.

[2] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L^1 View.* John Wiley, New York, 1985.

[3] S. Efromovich. *Nonparametric Curve Estimation. Methods, Theory, and Applications.* Springer, New York, 1999.

[4] R. L. Eubank. *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York, second edition, 1999.

[5] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London, 1996.

[6] P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach.* Chapman & Hall, London, 1994.

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

[8] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, Inc., New York, 1992.

[9] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[10] J. R. Thompson and R. A. Tapia. *Nonparametric Function Estimation, Modeling, and Simulation.* SIAM, Philadelphia, 1990.

[11] G. Wahba. *Spline Models for Observational Data.* SIAM, Philadelphia, 1990.

[12] M. P. Wand and M. C. Jones. *Kernel smoothing*. Chapman & Hall, London, 1995.