

Oulun yliopisto
 Matemaattisten tieteiden laitos
 Funktioiden estimointi
 4. harjoitus, viikko 8, 2011

1. Olkoon $f(\cdot; \theta)$ jakauman $N(\theta, 1)$ tiheysfunktio. Tarkastellaan tiheysfunktio-perhettä

$$\mathcal{F} = \{f(\cdot; \theta) \mid \theta \in \mathbb{R}\}$$

sekä i.i.d. otokseen X_1, \dots, X_n perustuvaa estimaattoria

$$\hat{f}_n(x) = t_n(x, X_1, \dots, X_n) = \frac{1}{\sqrt{2\pi(n-1)/n}} \exp\left(-\frac{n}{2(n-1)}\left(\frac{1}{n} \sum_{i=1}^n X_i - x\right)^2\right),$$

missä $n \geq 2$.

Osoita, että $\hat{f}_n(x)$ on harhaton perheessä \mathcal{F} eli että

$$E_f t_n(x, X_1, \dots, X_n) = f(x), \quad \forall f \in \mathcal{F}.$$

(Vihjeitä: $t_n(x, X_1, \dots, X_n) = u_n(x, Y)$, missä $Y = \frac{1}{n} \sum_{i=1}^n X_i$. Tunnet Y :n jakauman. Edellisissä harjoituksissa johdettiin hyödyllinen kaava.)

2. Olkoot x_1, \dots, x_n reaalityköt, ja tarkastellaan ydineestimaatin kaavaa

$$\hat{f}_n(x; h) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R}, \quad h > 0,$$

jossa K on Gaussin ydin (eli $N(0, 1)$ -jakauman tiheysfunktio). Osoita, että

$$\lim_{h \rightarrow 0^+} \hat{f}_n(x; h) = 0, \quad \text{kun } x \notin \{x_1, \dots, x_n\}.$$

Mitä saadaan raja-arvoksi, jos $x = x_i$?

3. Olkoot x_1, \dots, x_n kiinteitä reaalityköt, ja olkoon $h > 0$. Tarkastelemme kahden riippumattoman satunnaismuuttujaa I ja Z , joilla on seuraavat jakaumat. I on diskreetti satunnaismuuttuja, jolla on tasainen jakauma joukossa $\{1, \dots, n\}$ (ts. $\mathbb{P}(I = i) = 1/n$, kun $1 < i < n$). Satunnaismuuttujalla Z on jatkuva jakauma tiheysfunktioilla K . Määrittelemme vielä satunnaismuuttujan Y kaavalla

$$Y = x_I + hZ.$$

(a) Osoita, että Y :llä on tiheysfunktio

$$f(y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{y - x_i}{h}\right), \quad y \in \mathbb{R}.$$

- (b) Hahmottele tietokoneohjelma, joka generoi kokoa N olevan satunnaisotoksen ylläolevasta tiheysfunktioista, kun käytössäsi on generaattori `rand` välin $[0, 1]$ tasaiselle jakaumalle sekä tästä (todennäköisyyslaskennan mielessä) riippumaton generaattori `randK` tiheysfunktioita K vastaavalle jakaumalle.

4. Tässä tehtävässä tarkastelemme kahden tiheysfunktion, f ja g , välisen L^1 -etäisyyden

$$\|f - g\|_1 = \int_{-\infty}^{\infty} |f(x) - g(x)| dx = \int |f - g|$$

ominaisuuksia.

- (a) Osoita, että $\|f - g\|_1 < \infty$.
- (b) Olkoon $a : \mathbb{R} \rightarrow \mathbb{R}$ monotoninen funktio, joka on niin säännöllinen, että muuttujanvaihtokaavan käyttö integraalissa on sallittua. Olkoon satunnaismuuttujalla X tiheysfunktio f ja satunnaismuuttujalla Y tiheysfunktio g . Olkoot vastaavasti \tilde{f} ja \tilde{g} satunnaismuuttujien $a(X)$ ja $a(Y)$ tiheysfunktioita. Osoita, että

$$\|f - g\|_1 = \|\tilde{f} - \tilde{g}\|_1.$$

- (c) Osoita seuraava identiteetti (joka tunnetaan nimellä Scheffén lause)

$$2 \sup_B \left| \int_B f - \int_B g \right| = \int |f - g|.$$

Tässä supremum otetaan kaikkien (Borelin) joukkojen yli. Vihe: seuraavien tulosten todistamisesta on hyötyä

$$2 \int_B (f - g) \leq 2 \int_{\{f > g\}} (f - g) = \int |f - g|.$$

Tässä $\{f > g\} = \{x : f(x) > g(x)\}$, ja se on Borelin joukko.

5. (Tietokonetehtävä.) Kokeilemme tiheysfunktion ydinestimointia `snowfall`-aineistolla, jonka löydät kurssin kotisivulta. Aineisto sisältää tuumissa mitattuna vuosittaisen lumentulon v. 1910–1972 Buffalon kaupungissa New Yorkin osavaltiossa (vrt. esimerkki 1.1). Voit joko käyttää valmiita ohjelmia tai toteuttaa ydinestimaattorin haluamassasi ohjelmointiympäristössä.

- Etsi sellainen silotusparametrin h arvo, jolla estimaatissa näkyy selkeästi kolme moodia (eli lokaalia maksimia), kun ytimenä käytetään Gaussin ydintä. Seuraavaksi etsi sellainen h , jolla estimaatissa näkyy ainoastaan yksi moodi.
- Koeta toistaa edelliset kohdat käyttämällä ns. tasaista ydintä $K = \frac{1}{2}1_{[-1,1]}$.