Oulun yliopisto
Matemaattisten tieteiden laitos
Funktioiden estimointi
11. harjoitus, viikko 15, 2011

1. Tarkastelemme otokseen $(X_1, Y_1), \ldots, (X_n, Y_n)$ perustuvan Nadarayan-Watsonin estimaattorin $\hat{m}_n(\cdot; h)$ käyttäytymistä silotusparametrin h vaihdellessa, kun ytimenä on Gaussin ydin. Olkoon x mielivaltainen piste. Määrää estimaatin $\hat{m}_n(x; h)$ raja-arvo, kun (a) $h \to \infty$, (b) $h \to 0$.

2. Eulerin–MacLaurinin summauskaava annetaan oppikirjoissa usein muodossa, jonka toisen kertaluvun versio on

$$\sum_{i=1}^{n} f(i) = \int_1^n f(x)\,dx + \frac{1}{2}(f(n) + f(1)) + \frac{1}{12}(f'(n) - f'(1)) - \frac{1}{2}\int_1^n B_2(x - \lfloor x \rfloor) f''(x)\,dx.$$

Tässä B_2 on toisen asteen Bernoullin polynomi, $B_2(x) = x^2 - x + 1/6$, ja $\lfloor x \rfloor$ tarkoittaa suurinta kokonaislukua, joka on pienempi tai yhtä kuin x (alaspäin pyöristäminen). Kaava on voimassa kaikille $f \in C^2([1, n])$.

Johda yllä annettu Eulerin–MacLaurinin kaavan versio matkimalla luentojen lemman 4.2 todistusta (ts. pilko integraalissa $\int_1^n B_2(x - \lfloor x \rfloor) f''(x)\,dx$ integrointiväli n osaväliin ja osittaisintegroi kahdesti).

3. (Pakollinen harha regressiofunktion estimoinnissa.) Vaikka tehtävänasettelu saattaakin näyttää monimutkaiselta, saa tämän tehtävän ratkaistua yksinkertaisella laskulla!

Olkoon $x_0 \in \mathbb{R}$ kiinteä piste, ja tarkastellaan tiheysfunktioperhettä

$$\mathcal{F} = \{f \mid f \text{ jatkuva kahden muuttujan tiheysfunktio, jolle } \int_{-\infty}^{\infty} |y| f(x, y)\,dy < \infty \text{ kaikilla } x \text{ ja jolle } \int_{-\infty}^{\infty} f(x_0, y)\,dy > 0.\}$$

Olkoon $n \geq 1$ mielivaltainen otoskoko, $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. otos tiheydestä f ja $m_f(x) = \mathbb{E}_f(Y \mid X = x)$ tiheysfunktion f määräämä regressiofunktio. ($\mathbb{E}$:in alaindeksi f viittaa siihen, että ehdollinen odotusarvo lasketaan olettamalla, että parilla (X, Y) on tiheysfunktiona f.) Osoita, että tässä tilanteessa kaikki regressiofunktioestimaattorit ovat harhaisia, eli että mielivaltaiselle (Borelin) funktiolle $t : \mathbb{R}^{2n+1} \to \mathbb{R}$ löytyy $f \in \mathcal{F}$ siten, että

$$\mathbb{E}_f\, t(x_0, X_1, Y_1, \ldots, X_n, Y_n) \neq m_f(x_0).$$

(Alaindeksi tarkoittaa sitä, että vasemmalla puolella oleva odotusarvo lasketaan olettamalla, että $(X_1, Y_1), \ldots, (X_n, Y_n)$ ovat i.i.d. ja jakaumasta f.)

Vihje: Tee vastaoletus, ota kaksi tiheysfunktiota $f, g \in \mathcal{F}$ sekä tarkastele niiden konveksia kombinaatiota $\lambda f + (1-\lambda)g \in \mathcal{F}$, jossa $0 \leq \lambda \leq 1$. Tarkastele λ:n funktioina toisaalta lauseketta

$$\mathbb{E}_{\lambda f+(1-\lambda)g}\, t(x_0, X_1, Y_1, \ldots, X_n, Y_n)$$

ja toisaalta tiheyden $\lambda f + (1-\lambda)g$ määräämän regressiofunktion arvoa pisteessä x_0. Valitsemalla f ja g sopivasti päädyt ristiriitaan.

4. Generoi kokoa $n = 100$ oleva otos homoskedastisesta mallista

$$Y = m(X) + \sigma\varepsilon,$$

jossa X ja ε ovat riippumattomia satunnaismuuttujia, X:llä on tasainen jakauma välillä $[0, 1]$, $\epsilon \sim N(0, 1)$, $\sigma = 0.2$, ja

$$m(x) = 4.26(\exp(-3.5x) - 4\exp(-7x) + 3\exp(-10.5x)).$$

Estimoi regressiofunktio m otoksen perusteella käyttämällä (a) Pristleyn–Chaon estimattoria (b) Gasserin–Müllerin estimaattoria.

Valitse Gasserin–Müllerin estimaattoria varten sellainen ydin, jonka integraalifunktion osaat laskea numeerisesti; esim. Gaussin ytimelle K saadaan integraali $\int_{-\infty}^{a} K(u)\, du$ laskettua MATLABissa komennolla

```
0.5 * (1 + erf(a / sqrt(2)))
```