Oulun yliopisto
Matemaattisten tieteiden yksikkö
Tilastollinen hahmontunnistus
Harjoitus 10

1. Osoita, että tiheysfunktion ydinestimaattorille $\hat{f}_n(\cdot;h)$ aina pätee

$$\int_{\mathbb{R}^d} \hat{f}_n(x;h)dx = 1.$$

Milloin $\hat{f}_n(\cdot;h)$ on itseasiassa tiheysfunktio?

2. Olkoot $x_1,\ldots,x_n \in \mathbb{R}^d$ ja tarkastellaan ydinestimaatin kaavaa

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

jossa K on Gaussin ydin (eli $\mathrm{N}(0, I_d)$-jakauman tiheysfunktio). Osoita, että

$$\lim_{h\to 0+} \hat{f}_n(x;h) = 0, \qquad \text{kun } x \notin \{x_1,\ldots,x_n\}.$$

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

3. Olkoon X satunnaisvektori, jonka kovarianssimatriisi on Σ_X, ja jolla on tiheysfunktio f_X, ja olkoon $X_1,\ldots X_n$ otos X:n jakaumasta. Olkoon A valkaisumatriisi, ts. satunnaisvektorin $Y = AX$ kovarianssimatriisi Σ_Y on yksikkömatriisi I_d. Määritellään $Y_i = AX_i, i = 1,\ldots,n$ ja olkoon $\hat{f}_{Y,n}(\cdot;h)$ otokseen $Y_1,\ldots Y_n$ ja ytimeen K perustuva ydinestimaatori Y:n tiheysfunktiolle f_Y. Muodostetaan nyt $\hat{f}_{Y,n}(\cdot;h)$:n avulla estimaattori

$$\hat{f}_{X,n}(x;h) = |\det A|\ \hat{f}_{Y,n}(Ax;h),\ x \in \mathbb{R}^d,$$

X:n tiheysfunktiolle f_X. Osoita, että $\hat{f}_{X,n}(\cdot;h)$ on itse asiassa otokseen $X_1,\ldots X_n$ perustuva ydinestimaattori ytimellä

$$K_A(x) = |\det A|\ K(Ax).$$

Tarkista myös, että K_A todella on ydin eli että $\int_{\mathbb{R}^d} K_A(x)dx = 1$.

4. Tässä tehtävässä tarkastellaan yksiulotteista tasavälistä histogrammia. Reaaliakselilla on määriteltynä tasavälinen pisteistö u_k, $k \in \mathbb{Z}$, jossa $u_k < u_{k+1}, \forall k$ ja jossa perättäisten pisteiden etäisyys toisistaan on vakio h, $u_{k+1} - u_k = h, \forall k$. Näin tulee määritellyksi reaaliakselin ositus väleihin $B_k = [u_k, u_{k+1}[$. Olkoon $X_1,\ldots,X_n$ satunnaisotos satunnaismuuttujan X jakaumasta, jonka tiheysfunktio on f. Kuten luennoilla, määrittelemme, että satunnaismuuttuja N_k saa arvokseen niiden otospisteiden X_i lukumäärän, jotka sattuvat välille B_k. Luennoilla määritelty tiheysfunktion histogrammiestimaattori pisteessä $x \in \mathbb{R}$ saa silloin muodon

$$\hat{f}_n(x) = \frac{N_k}{nh}, \quad \text{kun } x \in B_k.$$

(a) Näytä, että $\int_{\mathbb{R}} \hat{f}_n(x)dx = 1$.

(b) Näytä, että kun $x \in B_k$, niin

$$\mathbb{E}\hat{f}_n(x) = \frac{p_k}{h}, \quad \text{missä} \quad p_k = \int_{B_k} f(x)\,dx.$$

5. Olkoot satunnaisvektorien X ja Y tiheysfunktiot f ja g, olkoon $T : \mathbb{R}^d \to \mathbb{R}^d$ bijektio sekä olkoot satunnaisvektorien $T(X)$ ja $T(Y)$ tiheysfunktiot $\tilde{f}$ ja $\tilde{g}$. Osoita, että tiheysfunktioiden ns. L^1-etäisyys säilyy muunnoksessa T, eli että

$$\int_{\mathbb{R}^d} |f(x) - g(x)|dx = \int_{\mathbb{R}^d} |\tilde{f}(x) - \tilde{g}(x)|dx.$$

Oletamme T:n olevan niin säännöllisen, että muuttujanvaihtokaavaa voidaan soveltaa yo. integraalissa. Päteekö tämä tulos L^2-etäisyydelle, eli onko myös

$$\sqrt{\int_{\mathbb{R}^d} (f(x) - g(x))^2 dx} = \sqrt{\int_{\mathbb{R}^d} (\tilde{f}(x) - \tilde{g}(x))^2 dx} \ ?$$