UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I:	*ST 391 – Ikkeparametrisk tetthetsestimering*
VED:	*Nils Lid Hjort*
FOR:	*Dr. Scient. og Cand. Scient.-studenter*
TID FOR EKSAMEN:	*Utlevering mandag 11. desember 1989 kl.* 9^{00}, *innlevering av rapport mandag 18. desember 1989 kl.* $\leq 16^{00}$
HJELPEMIDLER:	*Alle*

Oppgavesettet inneholder fire oppgaver, og er på fire sider. Cand. Scient.-studentene skal besvare oppgavene 1, 2, 3, mens Dr. Scient.-studentene også skal besvare oppgave 4.

Oppgave 1

La $X_1, \ldots, X_n$ være uavhengige observasjoner med samme tetthetsfunksjon $f(x)$ i $\mathcal{R} = (-\infty, \infty)$. Blant de vanligste ikkeparametriske estimatorene for f er kjernemetode-estimatoren

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)\frac{1}{h}\,,$$

der funksjonen $K(z)$ i denne oppgaven antas å være symmetrisk rundt 0, ikkenegativ, med integral 1, skalert slik at $\int z^2 K(z)\,\mathrm{d}z = 1$, og med eksisterende $\beta_K = \int K(z)^2\,\mathrm{d}z$. h er en positiv glattingsparameter. Denne oppgaven er hovedsaklig en liten ekskursjon inn i de L_1-baserte vurderinger av egenskapene til $\widehat{f}$.

(a) Ved hjelp av Taylor-utviklinger skal du gjøre rede for at

$$\mathrm{E}\,\widehat{f}(x) \doteq f(x) + \tfrac{1}{2}h^2 f''(x), \quad \mathrm{Var}\,\widehat{f}(x) \doteq \frac{\beta_K}{nh} f(x) - \frac{1}{n} f(x)^2.$$

Forklar hvilke antagelser som gjøres.

(b) Det vanligste matematiske kriterium for vurdering av egenskaper ved $\widehat{f}$ er den punktvise forventede kvadratfeil $\mathrm{MSE}(x) = \mathrm{E}\,\{\widehat{f}(x) - f(x)\}^2$, og dennes integrerte $\mathrm{IMSE} = \mathrm{E}\int\{\widehat{f}(x) - f(x)\}^2\,\mathrm{d}x$. Finn uttrykk for den punktvis optimale h, si $h_o(x)$, og den globalt optimale h, si h_o, som minimerer henholdsvis (den naturlige approksimasjon til) $\mathrm{MSE}(x)$ og (den naturlige approksimasjon til) IMSE. [Disse uttrykkene må delvis avhenge av den ukjente f.] Utled også tilnærmede uttrykk for de minimale $\mathrm{MSE}(x)$ og IMSE.

(c) Illustrer dette ved å beregne den punktvis optimale og den globalt optimale glattingsparameter i to situasjoner: (i) når den sanne f er en normalfordeling; (ii) når den sanne f er gitt ved $f(x) = f_0(x/\tau)/\tau$, for passende skalaparameter τ, hvor $f_0(y) = ye^{-y}$ for $y \geq 0$. (Dette er en skalert χ^2-fordeling med 4 frihetsgrader.) Anvend i begge situasjoner $K(z) = \phi(z)$, den standardnormale tetthetsfunksjonen. Sammenlign & kommenter.

(d) Alternative kriterier til de L_2-baserte over er de L_1-baserte

$$J_0(x) = \mathrm{E}\,\big|\widehat{f}(x) - f(x)\big|, \quad J_0 = \mathrm{E}\int \big|\widehat{f}(x) - f(x)\big|\,\mathrm{d}x.$$

Gi noen for- & mot-argumenter for at man skal studere egenskaper ved $\widehat{f}(x)$ ved hjelp av disse kriteriene, istedet for ved L_2-kriteriene.

(e) Generelt viser det seg å være nesten umulig å minimere $J_0(x)$ og J_0 med hensyn på h. Det viser seg imidlertid fruktbart å arbeide med visse naturlige øvre skranker: Vis at

$$J_0(x) \le J(x) = \mathrm{Scylla}(x) + \mathrm{Charybdis}(x), \quad J_0 \le J = \mathrm{Scylla} + \mathrm{Charybdis},$$

der $\mathrm{Scylla}(x) = |\mathrm{E}\,\widehat{f}(x) - f(x)|$ er den absolutte skjevhet [absolute bias], der $\mathrm{Charybdis}(x) = \mathrm{E}\,|\widehat{f}(x) - \mathrm{E}\,\widehat{f}(x)|$ er den forventede absolutte avstand til forventningen [mean absolute deviation], og der Scylla og Charybdis er disses integrerte.

(f) Vis at

$$(nh)^{1/2}\mathrm{Charybdis}(x) \to (2/\pi)^{1/2}\beta_K^{1/2} f(x)^{1/2},$$

under visse forutsetninger, ved hjelp av følgende nyttige generelle setning, som du ikke skal vise [under eksamen]:

Dersom $Y_1, \ldots, Y_n$ *er uavhengige med samme fordeling, og* $\mathrm{E}\,Y_i = 0$, $\mathrm{E}\,Y_i^2 = \sigma_n^2$, $\mathrm{E}\,|Y_i|^3 = \rho_n$, *så vil*

$$\mathrm{E}\,\big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\big| = \sqrt{\frac{2}{\pi}}\sigma_n + \varepsilon_n,$$

der $|\varepsilon_n| \le 1.999\rho_n/(\sqrt{n}\sigma_n^2)$. ♣

(g) Vis også at

$$J_R = \mathrm{Scylla}_R + \mathrm{Charybdis}_R \doteq \tfrac{1}{2}h^2\int_R |f''(x)|\,\mathrm{d}x + (2/\pi)^{1/2}\beta_K^{1/2}\int_R \sqrt{f(x)}\,\mathrm{d}x\,(nh)^{-1/2},$$

der Scylla og Charybdis er integrert over et *begrenset* intervall R. Hvordan må Odyssevs velge h, for voksende n, for at $\mathrm{Scylla}_R + \mathrm{Charybdis}_R$ skal forsvinne?

(h) Finn brukbare uttrykk for den $h_o(x)$ som (tilnærmet) minimerer den punktvise $J(x)$, og for den h_o som (tilnærmet) minimerer J_R. Utled også uttrykk for de tilhørende minimale $J(x)$ og J_R. Kommenter!

(i) Illustrer resultatene ved å beregne $h_o(x)$ og h_o i de to situasjonene beskrevet i punkt (c). Anvend igjen $K = \phi$, og ved beregning av h_o skal du bruke det tilnærmede uttrykk for J_R fra punkt (g), men med R hemningsløst satt lik hele $\mathcal{R}$ (selv om en detalj av utledningen under (g) krever begrenset R). Kommenter svarene.

Oppgave 2

Utgangspunktet er som i Oppgave 1, og vi skal fortsatt arbeide med kjernemetoden. Det som skal studeres nå er visse egenskaper ved den deriverte $\widehat{f}'(x)$, betraktet som en estimator

for den sanne deriverte $f'(x)$. I tillegg til antagelsene om kjernefunksjonen K fra Oppgave 1 skal det forutsettes at $\gamma_K = \int K'(z)^2\,\mathrm{d}z$ er endelig.

(a) Gi noen grunner til å være interessert i å estimere den deriverte av f.

(b) Vis ved Taylor-utvikling at

$$\mathrm{E}\,\widehat{f}'(x) \doteq f'(x) + \frac{3}{6}h^2 f'''(x) \quad \text{og} \quad \mathrm{Var}\,\widehat{f}'(x) = \frac{\gamma_K}{nh^3}f(x) + O\Big(\frac{f''(x)}{nh} + \frac{f'(x)^2}{n}\Big),$$

under visse forutsetninger om glatthet av f.

(c) Hvordan må h oppføre seg, for et voksende antall observasjoner, for at $\mathrm{MSE}(x) = \mathrm{E}\,\{\widehat{f}'(x) - f'(x)\}^2$ skal konvergere mot null?

(d) Utled uttrykk for den punktvise optimale $h_o(x)$, som tilnærmet minimerer $\mathrm{MSE}(x)$, og den globalt optimale h_o, som tilnærmet minimerer $\mathrm{IMSE}(x) = \mathrm{E}\int(\widehat{f}' - f')^2\,\mathrm{d}x$. Utled dessuten uttrykk for de assosierte minimale $\mathrm{MSE}(x)$ og IMSE, og kommenter!

(e) Illustrer disse resultatene ved å beregne $h_o(x)$ og h_o i de to situasjonene beskrevet i Oppgave 1s punkt (c). Kommenter.

Oppgave 3

I denne oppgaven skal vi se på en brøkdel av et interessant datamateriale hentet fra Scott, Gotto, Cole, Gory (1978): "Plasma lipids as collateral risk factors in coronary heart disease — a study of 371 males with chest pain", *Journal of Chronic Diseases* **31**, 337–345. Materialet består av plasma-lipid-konsentrasjonsmålinger, mer presist av henholdsvis plasma cholesterol concentration og plasma triglyceride concentration (begge målt i mg/100 ml). Noen figurer i Silvermans bok er faktisk relatert til dette materialet, men de baserer seg på de siste 320 data-parene (som svarte til personer som var døde pr. 1978), mens vi skal se på data fra de første 51 personene (som levde pr. 1978). Endelig skal vi [ved denne anledning] kun interessere oss for den ene størrelsen, kolesterol-konsentrasjonen. Her er datapunktene:

116.0	158.0	177.8	192.8	206.9	222.2	265.0
129.9	160.1	178.0	193.9	207.1	227.9	266.2
147.1	162.1	178.3	194.4	207.6	233.8	289.1
149.1	167.0	180.0	199.9	208.8	234.3	
149.7	168.0	187.6	200.7	210.0	237.2	
155.0	168.1	190.0	201.2	217.1	238.2	
156.0	168.6	190.1	204.8	218.7	243.2	
156.9	170.1	190.3	206.2	221.6	251.0	

MINITABs upretensiøse automatiske histogram over dem ser slik ut:

```
Midpoint   Count
     120       2  **
     140       3  ***
     160      10  **********
     180       6  ******
     200      15  ***************
     220       6  ******
     240       5  *****
     260       3  ***
     280       1  *
```

Vel: Anta at disse 51 målingene kommer fra en felles homogen populasjon. Estimer den underliggende sannsynlighetstettheten for kolesterol-konsentrasjon i denne populasjonen, basert på de 51 datapunktene. Forklar dine valg av metoder.

Oppgave 4 — kun for Dr. Scient.-studentene

Finn et bibliotek. Let gjennom matematisk/statistisk journal-litteratur fra 1987, 1988, 1989, og finn herfra to artikler som handler om tetthetsestimering og som interesserer deg. Forklar ganske kort i din eksamensrapport hva disse to artiklene handler om, med vekt på hovedidéer og hovedresultater. Gi også din egen (korte) vurdering av dem.

↪ *Signe Arbeidet* ↩