Løsningsforslag

TMA 4250: ROMLIG STATISTIKK

16 Mai 2009

1a)

$$(y - X\beta)'(y - X\beta) = SSQ(\beta)$$

$$+ X'X\beta - X'y = \frac{dSSQ}{d\beta}$$

$$\hat{\beta} = [X'X]^{-1} \cdot X' \cdot y$$

$$var(\hat{\beta}) = (X'X)^{-1} \cdot \sigma^2$$

Generelt

$$(y - X\beta)' \Sigma^{-1} (y - X\beta) = \text{Weighted} SSQ(\beta)$$

$$+ X'\Sigma^{-1}X \cdot \beta - X'\Sigma^{-1} y = \frac{d\,SSQ(\beta)}{d\beta}$$

$$\hat{\beta} = [X'\Sigma^{-1}X]^{-1} X' \Sigma^{-1} y$$

$$Var(\hat{\beta}) = [X'\Sigma^{-1}X]^{-1}$$

1b)

$$\pi(y|\beta) = N(X\beta, \Sigma) \quad , \pi(\beta) = N(b, A)$$

$$\pi(\beta|y) \propto \exp\left\{ -\frac{1}{2}(y-X\beta)'\Sigma^{-1}(y-X\beta) - \frac{1}{2}(\beta-b)'A^{-1}(\beta-b) \right\}$$

$$\propto \exp\left\{ -\frac{1}{2}\beta'\left[X'\Sigma^{-1}X + A^{-1}\right]\beta + \beta'\left[X'\Sigma^{-1}y + A^{-1}b\right] \right\}$$

$$E(\beta|y) : \left.\frac{\partial \pi(\beta|y)}{\partial \beta}\right|_{\beta=} = -\left[\cancel{X}'\Sigma^{-1}X + A^{-1}\right]\beta + X'\Sigma^{-1}y + A^{-1}b = 0$$

Dvs

$$\underline{E(\beta|y) = \left[X'\Sigma^{-1}X + A^{-1}\right]^{-1}\left(X'\Sigma^{-1}y + A^{-1}b\right)}$$

$$\underline{Var(\beta|y) = \left[\frac{\partial^2 \pi(\beta|y)}{\partial \beta^2}\right]^{-1} = \left[X'\Sigma^{-1}X + A^{-1}\right]^{-1}}$$

Når $A \to \infty$, ingen a priori informasjon, så vil $A^{-1} \to 0$.

Da blir $E(\beta|y) \approx \left[X'\Sigma^{-1}X\right]^{-1}(X'\Sigma^{-1}y)$

$Var(\beta|y) \approx (X'\Sigma^{-1}X)^{-1}$

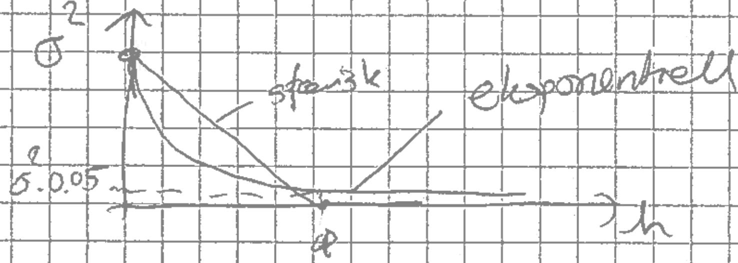
som er likt med den siste i (a)

E(β|y) er en vekting av data y og a priori forventning b. Vektene bestemmes av antall data, presisjon til data og a priori forståelse, dvs X, Σ, A.

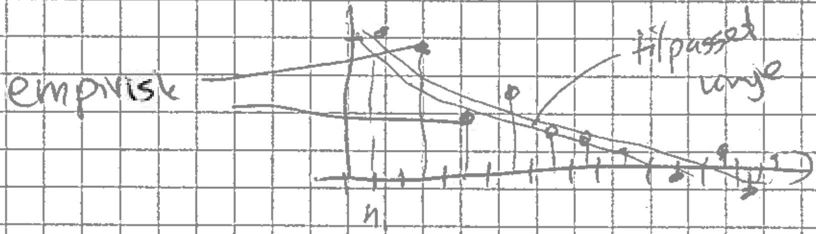
1c)

Eksponentiell: $C(h) = \sigma^2 e^{-3\frac{|h|}{\varphi}}$

Sfærisk: $C(h) = \begin{cases} \sigma^2\left(1-\frac{3}{2}\frac{|h|}{\varphi}+\frac{1}{2}\frac{|h|^3}{\varphi^3}\right) & |h|<\varphi \\ 0 & |h|>\varphi \end{cases}$



Tilpasses ved å se på data innen $(h-\Delta, h+\Delta)$ distanser. Lage empiriske kovariansestimater. For små-middels h har vi ofte mye data. For store h, mindre data



typisk tilpasset plott

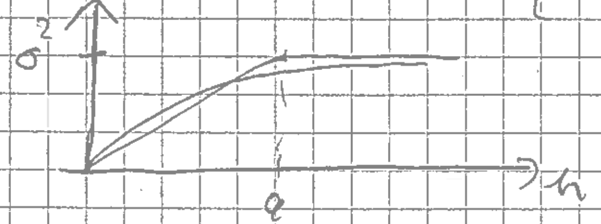
Variogram: $\gamma(h) = \mathrm{Var}(Y(s+h) - Y(s))$

$= 2\,\mathrm{Var}(Y(s)) - 2\,\mathrm{Cov}(Y(s+h), Y(s))$

dersom disse eksisterer

Eksponensiell: $\gamma(h) = \sigma^2\left(1 - e^{-3\frac{|h|}{\varphi}}\right)$

Sfærisk: $\gamma(h) = \begin{cases} \sigma^2\left(\frac{3}{2}\left|\frac{h}{\varphi}\right| - \frac{1}{2}\left|\frac{h}{\varphi}\right|^3\right) & |h|<a \\ 0 & |h|>a \end{cases}$



1d)

$$\Sigma_{0,y} = \begin{pmatrix} \mathrm{Cov}(Y(s_0), Y(s_1)) \\ \vdots \\ \mathrm{Cov}(Y(s_0), Y(s_{20})) \end{pmatrix}^T$$

i) Covariansene er 0, Σ_{0y} nesten 0, $\hat{y}_0 \approx (1, x_0)\beta$.

$V_0 \approx \sigma^2$

ii) $\Sigma_{0y} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, $\hat{y}_0 \approx \sigma^2 \cdot \underbrace{\Sigma^{-1}_{1\cdot}}_{\text{nesten 0 unna element 1, der } \sigma^{-2}} \cdot (y - X\beta) + (1, x_0)\beta$.

$$\approx \sigma^2 (\sigma^{-2} \; 0 \ldots 0) \begin{pmatrix} y_1 - X_1 \beta \\ \vdots \\ y_{20} - X_{20}\beta \end{pmatrix} + (1\, x_0 \beta)$$

$$\approx \underline{y_1}.$$

$$\underline{\hat{V}_0 \approx 0.}$$

$$\pi(y_0 \mid \sigma, y) = N\big(E(E(y_0 \mid \beta, \sigma, y)), E(\mathrm{Var}(y_0 \mid \beta, \sigma, y)) + \mathrm{Var}(E(y_0 \mid \beta, \sigma, y))\big)$$

$$= N\Big(\underset{\beta \mid y, \sigma}{E}(\hat{y}_0(\beta)), \underset{\beta \mid y, \sigma}{E}(V_0) + \underset{\beta \mid y, \sigma}{\mathrm{Var}}(\hat{y}_0(\beta))\Big)$$

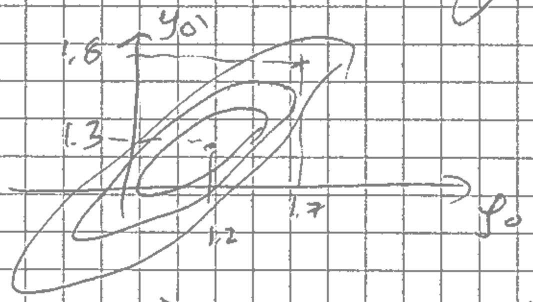
$$= N\Big\{\underline{\Sigma_{0y}\Sigma^{-1}(y - Xm) + (1, x_0)\cdot m}, \; \underline{V_0 + \Sigma_{0y}\Sigma^{-1} X S X' \Sigma^{-1} \Sigma_{0y} + (1, x_0) S \begin{pmatrix} 1 \\ x_0 \end{pmatrix}}\Big\}$$

Istedet for dobbel forventning og varians, kan man sette opp $[y_0, y \mid \beta]$, og betinge på y.

(d)

$$\pi\left(\begin{pmatrix} y_0 \\ y_0' \end{pmatrix} \middle| \sigma, y\right) = N\left(\begin{pmatrix}1.2\\1.3\end{pmatrix}, \begin{bmatrix}0.21^2 & \varrho \\ \varrho & 0.22^2\end{bmatrix}\right)$$

Dersom ϱ er veldig stor, som er naturlig dersom romlig korrelasjon er stor, så er at y_0, y_0' er store samtidig ikke veldig sannsynlig.



1e)

$$\pi(\xi|y) \propto \pi(y|\xi)\cdot\pi(\xi) \propto e^{-\frac{\xi}{2}(y-X\beta)'R(y-X\beta) - d\xi}\, \frac{\xi^{20/2}}{|R|} \cdot \xi^{c-1}$$

$$\propto \xi^{c+10-1}\, e^{-\xi\left[d+\frac{1}{2}(y-X\beta)'R(y-X\beta)\right]}$$

Gjenkjenner $c \to c+10$, $d \to d+\frac{1}{2}(y-X\beta)'R(y-X\beta)$

$$\pi(y_0|\beta,y) = \int \pi(y_0|\beta,\xi,y)\cdot\pi(\xi|y)\,d\xi$$

$$= \int N(\hat{y}_0, V_0)\cdot \text{Gamma}\left(c+10, \tfrac{1}{2}(y-X\beta)'R(y-X\beta)+d\right)d\xi$$

$$= \int \frac{\xi^{1/2}}{|1-R_{0y}R^{-1}R_{0,y}^T|^{1/2}\sqrt{2\pi}}\, e^{-\frac{\xi}{2}(y_0-\hat{y}_0)^2/R_0}\, \frac{1}{k}\,\xi^{\overbrace{c+10-1}^{c'}}\, e^{-\xi\overbrace{\left(d+\frac{1}{2}(y-X\beta)'R(y-X\beta)\right)}^{d'}}\,d\xi$$

$$= \frac{1}{|1-R_{0y}R^{-1}R_{0y}^T|\sqrt{2\pi}}\,\frac{1}{k}\; k' \;\frac{1}{k'}\int \xi^{1/2+c'-1}\, e^{-\xi\left(\frac{1}{2}\frac{(y_0-\hat{y}_0)^2}{R_0}+d'\right)}\,d\xi$$

$$k' = \left(\frac{\left(\frac{1}{2}\frac{(y_0-\hat{y}_0)^2}{R_0}+d'\right)^{1/2+c'}}{\Gamma(1/2+c')}\right)^{-1}, \quad \text{norm. konst i Gamma}$$

Merk: Her avhenger k, k' av y, y_0.

k er normaliseringskonstanten i

$$\text{Gamma}\Big(\underbrace{c+\theta}_{c'},\ \underbrace{\tfrac{1}{2}(y-X\beta)' R (y-X\beta)+d}_{d'}\Big)$$

k' er normaliseringskonstanten i

$$\text{Gamma}\Big(\tfrac{1}{2}+c',\ \tfrac{1}{2}\frac{(y_0-\hat{y}_0)^2}{R_0}+d'\Big)$$

Integralet $\displaystyle\int \frac{1}{k'}\,\xi^{1/2+c'-1}\, e^{-\xi\left(\frac{1}{2}\frac{(y_0-\hat{y}_0)^2}{R_0}+d'\right)}\,d\xi$

er over hele tettheten og lik 1.

Formen på k', k viser at $\pi(y_0|y,\beta)$ er en t-fordeling:

$$\pi(y_0|y,\beta) \propto \frac{k'}{k} = \frac{\Gamma(\frac{1}{2}+c')}{\Gamma(c')}\cdot\frac{d'^{\,c'}}{\left(\frac{1}{2}\frac{(y_0-\hat{y}_0)^2}{R_0}+d'\right)^{c'+1/2}}$$

$$\propto \frac{1}{\left(\frac{1}{2}\frac{(y_0-\hat{y}_0)^2}{R_0}+d+\frac{1}{2}(y-X\beta)'R^{-1}(y-X\beta)\right)^{c+\theta+1/2}}$$

kvadratisk form i y_0

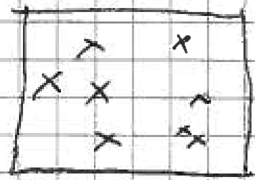
relatert til frihetsgrader

t-fordeling $\dfrac{1}{(1+y_0^2)^b}$

2a)

- Antall punkter er Poisson($\lambda \cdot |S|$)
- Antall punkter i disjunkte områder uavhengige
-

* Trekk fra Poissonfordeling $\lambda \cdot |S| \rightsquigarrow n$.
* Spre n punkter uniformt på S.



8 punkter

10 punkter

Forventet antall punkter er 10. Punktene er ikke regulære, heller ikke for klustrete.

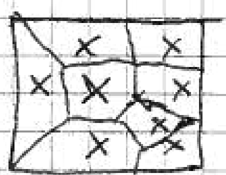
Alt 1 ikke-homogen

* Ikke-homogen. Trekk fra Poisson $\lambda = 20$, $\lambda \cdot |S|$.
* Spre punktene som følger: 1. Uniformt på S
2. Thinning av de som havner med $s_1 < 1/2$.

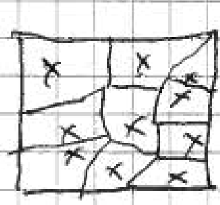
Alt 2 ikke-homogen

~~* Del områdene i 2~~ : • Trekk Venstre for seg $\lambda = 10 \cdot 1/2$
• Trekk Høyre for seg. $\lambda = 20 \cdot 1/2$

2b)



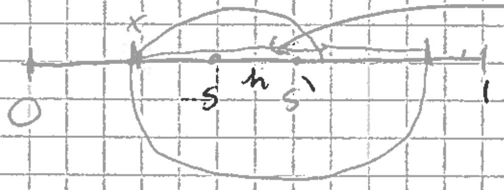
8 regioner



10 regioner

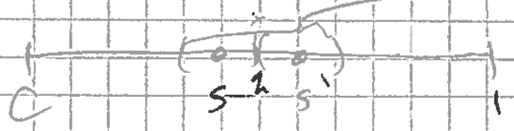
2b)

$x < s$
$s < s'$



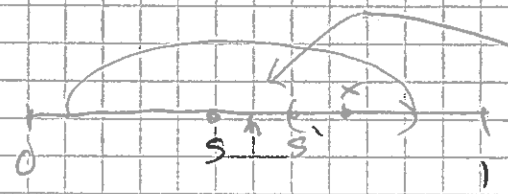
Null punkter i region.
$\rightarrow x$ nærmeste punkt til s og s'.
Regionlengde $= (s+h-x)\cdot 2$

$s < x \leq s'$



Null punkter i region
$\rightarrow x$ nærmeste punkt til s og s'.
Regionslengde $= 2h$

$s < s' < x$



Null punkter i region
$\rightarrow x$ nærmeste punkt til s og s'.
Regionlengde $= (x-s)\cdot 2$

Vi må integrere over alle $x \in (0,1)$

$$P(s, s' \text{ i samme celle} \mid x) = e^{-\lambda \cdot \text{Lengde region}(x,s,s')}$$

$$P(s, s' \text{ i samme celle}) = \int_0^1 \lambda \cdot e^{-\lambda \cdot \text{Lengde region}(x,s,s')} \, dx$$

$$= \lambda \left[\int_0^s e^{-2\lambda(s+h-x)} \, dx + \int_s^{s+h} e^{-2\lambda h} \, dx + \int_{s+h}^1 e^{-2\lambda(x-s)} \, dx \right]$$

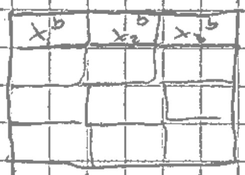
$$= \lambda \cdot \frac{1}{2\lambda} \left\{ e^{-2\lambda h}(1+2\lambda h) - e^{-2\lambda(s+h)} - e^{-2\lambda(1-s)} \right\}$$

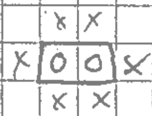
Alternativ: simulering.
Trekk B Poissonprosesser på (0,1), intensitet λ.
Tell andelen av B, der s, $s+h$ havner i samme intervall.

3a)

$$\text{Blokk-pseudolikelihood} = \prod_{i=1}^{n/2} P(x_i^b \mid x_{-i}^b)$$

$$x_i^b = (x_{2i-1}, x_{2i}) \in \left(\{0,0\}, \{0,1\}, \{1,0\}, \{1,1\} \right) = S_i^b$$





$$P(x_{2i-1}, x_{2i} \mid x_1, \dots x_{2i-2}, x_{2i+1}, \dots x_n)$$
$$= P(x_{2i-1}, x_{2i} \mid \text{naboer}) \propto e^{\beta \sum_{3\,\text{celler}} I(x_{2i-1} = y_k) + \beta \sum_{3\,\text{celler}} I(x_{2i} = z_k) + \beta I(x_{2i-1} = x_{2i})}$$

y_k er venstre celler:  3 cliquer

z_k er høyre celler:  3 cliquer

$I(x_{2i-1} = x_{2i})$ er clique mellom de to sentrale.

$$P(x_{2i-1} = 0, x_{2i} = 0 \mid \text{naboer}) = \frac{e^{\beta \sum_{3\,\text{celler}} I(0 = y_k) + \beta \sum_{3\,\text{celler}} I(0 = z_k) + \beta}}{\sum_{S_i^b} P(x_{2i-1}, x_{2i} \mid \text{naboer})}$$

tilsvarende for $\{0,1\}$, $\{1,0\}$ og $\{1,1\}$.
Den siste β i eksponent i teller er ikke med for $\{0,1\}$, $\{1,0\}$.

* Blokk-pseudolikelihood maksimeres numerisk, over et grid av β, eller for eks med Newton's Metode.
* Blokk-pseudolikelihood er nærmere sann likelihood $p(x)$, og er sånn sett bedre enn kun bruk av en-celle.

3b)

$$p(x|y) \propto p(x) \prod_{i=1}^{n} p(y_i|x_i)$$

$$p(y_i|x_i) = \begin{cases} p & y_i = x_i \\ 1-p & y_i \neq x_i \end{cases}$$

$$p(x_i^b | x_{-i}^b, y) = p(x_i^b | \text{naboer}, y)$$

$$\propto e^{\beta \sum_{3\,\text{celler}} I(x_{2i-1} = y_x) + \beta \sum_{3\,\text{celler}} I(x_{2i} = y) + \beta I(x_{2i-1} = x_{2i}) + \ln p(y_{2i-1}|x_{2i-1}) + \ln p(y_{2i}|x_{2i})}$$

Som i 3a) må sannsynlighet summeres til 1 over utfallsrommet $\{0,0\}, \{0,1\}, \{1,0\}, \{1,1\}$

Iterer:

* Gibbs sampler: - Trekk tilfeldig blokk $i = 1, \ldots, n/2$
  - Trekk ny $x_i^b \sim p(x_i^b | \text{naboer}, y)$

* Pga at større blokker oppdateres vil trolig denne Gibbs-sampler algoritmen fungere bedre enn en en-celle algoritme. Miksingen i MCMC blir bedre.

3c) I hver iterasjon av Gibbs-samplingen, kan vi telle antall 1-ere.

$$S = \sum_{i=1}^{n} x_i$$

Sett antall = 0;

Iterer Nantall ganger:

Gibbsampling:

- Trekk tilfeldig blokk $i = 1, \ldots, n/2$
- Trekk ny $x_i^b \sim p(x_i^b \mid \text{naboer}, y)$
- $S = \sum_{i=1}^{n} x_i$
- Dersom $S > 200$, sett antall++;

$$\text{Estimat} = \frac{\text{antall}}{\text{Nantall}}$$

Dersom Estimat $> 1/2$ vil man borre.