Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen:
Jo Eidsvik
Tlf. 90127472

EKSAMEN I FAG TMA4250
ROMLIG STATISTIKK

Lørdag 16. mai 2009

Tid: 09:00–13:00

Tillatte hjelpemidler:
Statistiske tabeller og formler, Tapir. Egetprodusert gult ark. Godkjent kalkulator.

Sensur: 31. mai 2009

Oppgave 1

Data $y = (y_1, \ldots, y_{20})'$ er manuelle målinger av snødybde (i meter) samlet inn ved $s_1, \ldots, s_{20}$ ulike nord, øst koordinater i et fjellområde i Trøndelag. Data brukes blant annet til å forutsi tilsig til vannmagasin. Korteste distanser til nabosteder varierer omlag $300m - 2km$. Høyde over havet $x_1, \ldots, x_{20}$ vurderes som en viktig forklaringsvariabel. En foreslått modell er

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, 20. \tag{1}$$

På kortform: $y = X\beta + \epsilon$, der $\beta = (\beta_0, \beta_1)'$ og linje $i = 1, \ldots, 20$ i matrisa X er $X_i = (1, x_i)$. Støyleddene $\epsilon = (\epsilon_1, \ldots, \epsilon_{20})'$ er normalfordelte med forventning 0 og kovariansmatrise Σ.

a) Anta først at $\Sigma = \sigma^2 I_{20}$, dvs at støyleddene er uavhengige med kjent, konstant standardavvik σ. Finn uttrykket for minste kvadraters estimator for β. Hva er kovariansmatrisa til estimatoren $\hat{\beta}$?

Anta nå at $\Sigma = \sigma^2 R$, der R er en kjent matrise med romlig korrelasjon mellom de 20 målingene. Hva blir estimatoren for β i dette tilfellet? Hva er nå kovariansmatrisa til estimatoren $\hat{\beta}$?

b) Anta at vi fra tidligere år har a priori informasjon om β i form av tetthet $\pi(\beta) = N(b, A)$, der b er 2×1 vektor med forventningsverdier og A er 2×2 kovariansmatrise.

Vis at a posteriori fordelingen $\pi(\beta|y) = N(m, S)$, der $S = [A^{-1} + X'\Sigma^{-1}X]^{-1}$ og $m = S[A^{-1}b + X'\Sigma^{-1}y]$. Sammenlign løsningen med de i punkt a) ?

c) To romlige korrelasjonsfunksjoner er eksponensiell og sfærisk. Hvilken matematisk form har disse? Skisser dem i et plott, sammenlign og diskuter. Beskriv kort hvordan du vil tilpasse en korrelasjonsmodell fra data y.

Hva er definisjonen på et variogram? Hva er matematisk form for eksponensiell og sfærisk variogram? Skisser dem i et plott og sammenlign.

Vi ønsker å finne kriging prediksjonen i en ny koordinat s_0, med kjent høyde over havet x_0. For kjent $\beta = (\beta_0, \beta_1)'$ og σ kan prediksjonen skrives på formen $\hat{y}_0 = \Sigma_{0,y}\Sigma^{-1}(y - X\beta) + (1, x_0)\beta$, med prediksjonsvarians $V_0 = \sigma^2 - \Sigma_{0,y}\Sigma^{-1}\Sigma'_{0,y}$. Dvs at $\pi(y_0|\beta, \sigma, y) = N(\hat{y}_0, V_0)$. Vi antar her korrelasjonsmatrisa R som kjent.

d) Hva er $\Sigma_{0,y}$ i uttrykkene for $\hat{y}_0$ og V_0? Hva skjer med uttrykkene dersom i) s_0 er langt unna alle koordinater $s_1, \ldots, s_{20}$, ii) s_0 er svært nær koordinat s_1, men langt unna alle andre?

Anta at σ er kjent, mens β er ukjent. Bruk resultatet fra punkt b) til å utlede prediktiv fordeling $\pi(y_0|\sigma, y)$.

For en koordinat s_0 blir $\pi(y_0|\sigma, y) = N(1.2, 0.21^2)$. For en annen koordinat $s_{0'}$, kun 50 m unna s_0, er $\pi(y_{0'}|\sigma, y) = N(1.3, 0.22^2)$. Ved begge koordinatene måles snødybden, og blir henholdsvis 1.7 og 1.8 meter. Begge målingene y_0 og $y_{0'}$ er relativt langt ut i halen. Forklar hvorfor to så store målinger likevel ikke trenger å være veldig usannsynlig i denne situasjonen?

e) Anta nå at β er kjent, mens σ er ukjent. Vi parameteriserer støynivået med $\zeta = \sigma^{-2}$. Som a priori fordeling bruker vi en gammafordeling *Gamma*(c,d), dvs

$$\pi(\zeta) = \frac{d^c \zeta^{c-1}}{\Gamma(c)} \exp(-d\zeta). \tag{2}$$

Vis at a posteriori fordelingen $\pi(\zeta|y) = Gamma(c + 10, d + 1/2(y - X\beta)'R^{-1}(y - X\beta))$.

Regn ut prediktiv fordeling $\pi(y_0|\beta, y)$.

Oppgave 2

a) Gi definisjonen til en homogen Poissonprosess over et to dimensjonalt område S.

Beskriv hvordan du vil simulere en realisasjon av en homogen Poissonprosess på enhetskvadratet $S = (0, 1)^2$, med intensitet $\lambda = 10$. Tegn to skissemessige 'realisasjoner' av en slik prosess. Kommenter.

Vurder nå en ikke-homogen Poissonprosess med intenstitet $\lambda = 10$ for koordinater med $s_1 < 1/2$, og $\lambda = 20$ for koordinater med $s_1 \geq 1/2$. Her er s_1 første koordinat i enhetskvadratet $S = (0, 1)^2$. Beskriv hvordan du kan simulere en realisasjon av en denne.

b) Din realisasjon av en homogen Poissonprosess fra punkt a) gir grunnlag for en tesselering (oppdeling) av enhetskvadratet. Denne kalles en Dirichlet-Voronoi tesselering. Et område A_i, $i = 1, \ldots, n$, er definert ved alle koordinater som har x_i som nærmeste punktrealisasjon. Tegn cirka tesseleringen som oppstår basert på dine to realisasjoner over.

Anta en homogen Poissonprosess (i tid) på enhetsintervallet $(0, 1)$, med intensitet $\lambda = 10$. Denne gir grunnlag for en Dirichlet-Voronoi celleoppdeling av enhetsintervallet. Regn ut sannsynligheten for at to koordinater s og s', der $|s - s'| = h = 0.1$, er innen samme celle.

Oppgave 3

Anta at x representerer metninger (vann/olje) langs en geologisk formasjon i Nordsjøen. Metningene er definert på et to dimensjonalt grid av størrelse 50×50. Vi modellerer $x = (x_1, \ldots, x_{2500})'$ som et binært Markov felt med celleverdier $x_i \in \{0, 1\}$, der 0 er vann og 1 er olje. Anta første ordens naboskap, med potensial β for like cliquer, potensial 0 for ulike cliquer, dvs en vanlig Ising modell.

a) Anta først at en geolog skisserer et tiltenkt felt x, og at vi utfra dette estimerer β.

Definer blokk-pseudolikelihood som produktet av de fulle betingete fordelinger over alle disjunkte blokker av 1×2 celler. Skriv opp denne blokk-pseudolikelihooden og beskriv hvordan du fra den kan estimere β. Hvordan vil denne estimeringsalgoritmen fungere i forhold til en vanlig en-celle pseudolikelihood?

b) Anta nå at seismiske data $y = (y_1, \ldots, y_{2500})'$ er tilgjengelig. Seismiske data antas i dette tilfellet betinget uavhengig i alle gridceller, med $y_i \in \{0, 1\}$, der $y_i = x_i$ med sannsynlighet $p = 0.75$.

Skriv opp a posteriori fordelingen til x gitt y.

Finn den fulle betingete fordelingen for en 1×2 blokk av celler, gitt data y og alle andre celle-verdier. Beskriv hvordan du vil bruke disse fulle betingete fordelingene i en Gibbs sampler. Hvordan vil denne Gibbs sampling algoritmen fungere i forhold til den vanlige en-celle oppdateringen?

c) Et petroleumselskap vurderer å plassere en brønn i dette området. De bestemmer seg for å borre dersom minst 200 av cellene er oljemettet. Bruk resultatet fra Gibbs sampling i b) til å bistå selskapet i denne beslutningen.