Bertil Wegmann
Avdelning statistik och maskininlärning (STIMA)
IDA
Linköpings universitet
bertil.wegmann@liu.se
2019-11-28

**Inlämningsuppgift - Bayesiansk statistik, 7.5 hp**

**Moment 3 - Regularisering, informationskriterium, modelljämförelse, MCMC**

**Inlämningsuppgiften skall lösas individuellt**. Valfritt program får användas för att lösa uppgifterna nedan, men mina instruktioner och ledtrådar gäller för programmeringsspråket R. **Bifoga all kod tillsammans med lösningen för varje deluppgift**. **Labbrapport** med lösningar på nedanstående uppgifter samt relevanta datorutskrifter **lämnas in på Lisam senast tisdagen den 10:e december kl. 23:59**.

Använd **samma datamaterial från inlämningsuppgiften i moment 2**. **Obs! Redovisa den dag då du föddes i labbrapporten.** Antag följande Bayesianska linjära regressionsmodell för responsvariabeln $y$ med **högst** 6 standardiserade (förutom dummyvariablerna) förklaringsvariabler (lägg till en vektor med 1:or till $X$ för att modellera interceptet $\beta_0$):

$$y_i|\mu_i, \sigma, X \overset{iid}{\sim} N(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i}$$

$$\beta_0 \sim N(3, 10),$$

$$\beta_j \sim N(0, 10),\ j = 1, 2, 3, 4, 5, 6.$$

$$\ln \sigma \sim N(0, 1)$$

Nu till uppgifterna:

1. (8 poäng) I denna uppgift ska du med hjälp av kvadratisk approximation anpassa Bayesianska linjära regressionsmodeller med din beroende variabel $y$.
   (a) Anpassa var och en av följande tre Bayesianska linjära regressionsmodeller som ska användas till uppgifterna **(b) till (d)**:
      i. Bayesiansk linjär regressionsanalys **utan** förklaringsvariabler.
      ii. Bayesiansk linjär regressionsanalys **med** den kontinuerliga förklaringsvariabel $x$ som har högst korrelation med den beroende variabeln $y$.
      iii. Bayesiansk linjär regressionsanalys **med alla** dina förklaringsvariabler.
   (b) Beräkna och redovisa Deviance $D(q)$ för respektive modell på hela datamaterialet och jämför mellan modellerna. Vilken av modellerna verkar anpassa data bäst? Motivera.
   (c) Dela upp data i ett **training** och ett **test sample**. Första halvan av data (första 29 observationerna) används till **training sample** och andra halvan till **test sample**. Beräkna och redovisa Deviance för respektive **sample** och jämför mellan modellerna. Vilken av modellerna verkar ha bäst prediktionsförmåga? Motivera.

(d) Beräkna och redovisa AIC, DIC och WAIC för respektive modell och jämför mellan modellerna. **Funktionerna AIC och DIC ska ej användas, men funktionen WAIC får användas**. Vilken av modellerna verkar vara bäst och varför? Är det något eller några av måtten som lämpar sig bättre att använda än andra? Motivera.

(e) Anpassa var och en av följande tre Bayesianska linjära regressionsmodeller (**notera att modell ii. och iii. redan anpassats i uppgift (a)**) som ska användas till uppgifterna **(f) till (h)**:

i. Bayesiansk linjär regressionsanalys **med** den kontinuerliga förklaringsvariabel $x$ som har nästhögst korrelation med den beroende variabeln $y$.

ii. Bayesiansk linjär regressionsanalys **med** den kontinuerliga förklaringsvariabel $x$ som har högst korrelation med den beroende variabeln $y$.

iii. Bayesiansk linjär regressionsanalys **med alla** dina förklaringsvariabler.

(f) Beräkna själv Akaike-vikter för var och en av de tre modellerna och tolka dom i ord. **Funktionen compare ska ej användas, men funktionerna AIC, DIC och WAIC får användas**.

(g) Använd de genomsnittliga värdena på förklaringsvariablerna för modellerna i uppgift (e). Beräkna ett 95.2 % kredibilitetsintervall för $y$ genom att använda modellernas Akaike-vikter i uppgift (f). Kredibilitetsintervallet för $y$ blir således ett viktat genomsnittligt kredibilitetsintervall över modellerna.

(h) Anpassa **modell iii.** i uppgift (e) med samma normalfördelade regulariserande prior för varje lutningsparameter. **Avgör med hjälp av DIC** vilken standardavvikelse för den regulariserande priorn som ger den bästa modellen. Avgör detta med hjälp av en grid med 20 olika värden för standardavvikelsen på ett lämpligt intervall.

2. (4 poäng) Anpassa en Bayesiansk linjär regressionsanalys **med alla** dina förklaringsvariabler med hjälp av **funktionen map2stan** i R. Välj totala antalet posteriordragningar till **3000 per MCMC kedja efter uppvärmningsfas** och antalet dragningar i **uppvärmningsfasen till 1000 per MCMC kedja** och använd **totalt 4 MCMC kedjor**. Denna anpassning med MCMC ska bland annat jämföras med den kvadratiska approximationen av samma modell från uppgift 1(a) iii.

(a) **Jämför posteriorresultaten från MCMC med den kvadratiska approximationen** genom att använda **funktionen precis()**. Kommentera kring likheter/skillnader i resultat mellan metoderna. Vad kan likheter/skillnader tänkas bero på?

(b) **Redovisa en figur för respektive metod av MCMC och kvadratisk approximation** med alla separata posteriorfördelningar, alla parvisa posteriorfördelningar och alla parvisa korrelationskoefficienter för parametrarna (**använd funktionen pairs()**). Verkar kvadratisk approximation fungera bra utifrån dom separata posteriorfördelningarna?

(c) **Utvärdera och dra slutsats om MCMC dragningarna har lett till att MCMC algoritmen konvergerat till posteriorn** med hjälp av all den MCMC diagnostik som vi tagit upp på momentet. Välj sedan en av lutningsparametrarna och utvärdera om kvartilerna i posteriorfördelningen för denna lutningsparameter verkar ha konvergerat till ett specifikt värde. Hur många MCMC dragningar verkar behövas för att uppnå konvergens för posteriorkvartilerna?