Bayes-päättely, 5. harjoitukset (22.–23.2.2017)

1. Osoita, että Jeffreyn priori on säännöllisille malleille invariantti mallin parametrisoinnin suhteen, eli että parametrin θ muunnoksen $\phi = g(\theta)$ Jeffreyn priori J_Φ saadaan sijoittamalla käänteismuunnos $h(\phi) = g^{-1}(\phi)$ alkuperäisen θ:n avulla parametrisoidun mallin Jeffreyn prioriin J_Θ.

Tämän voit tehdä suoralla laskulla osoittamalla tiheysfunktion muuntokaavan avulla, että

$$J_\Phi(\phi) \propto J_\Theta(h(\phi)).$$

Voit käyttää hyväksi tietoa, että logaritmisen uskottavuusfunktion ensimmäisen derivaatan odotusarvo on 0 säännöllisille malleille:

$$E\left[\frac{\partial \log f_{Y|\theta}(Y|\theta)}{\partial \theta} \,\middle|\, \theta\right] = 0.$$

2. Tarkastellaan klassista aineistoa kauden 1970 amerikkalaisen pesäpallon lyöjätilastoista [1]. Tiedosto `baseball75.csv` sisältää 18 pelaajan kauden 45 ensimmäisen lyöntivuoron osumat `Hits` (merkitään $y_1, \ldots, y_{18}$), loppukauden lyöntivuorojen osumat `RemainingHits` (merkitään $\tilde{y}_1, \ldots, \tilde{y}_{18}$) ja loppukauden lyöntivuorojen määrät `RemainingAB` (merk. $m_1, \ldots, m_{18}$).

Tarkoituksena on ennustaa loppukauden ns. batting averagea, eli osumien määrää jaettuna lyöntivuorojen määrällä, ensimmäisten 45 lyöntivuoron perusteella. Toteutetaan ensin kaksi yksinkertaisinta mallia. Nämä ovat kumpikin konjugaattimalleja, joten ratkaise ne käyttämättä stania (vihje: mediaanit ja kvantiilit voit laskea posteriorijakaumasta `R`:n funktiolla `qbeta`).

(a) Mallinnetaan jokaista pelaaja erikseen, eli oletetaan, että pelaajien todelliset osumatarkkuudet $\theta_1, \ldots, \theta_{18}$ ovat riippumattomia, ja että

$$Y_j \sim \text{Binom}(45, \theta_j) \quad \text{kaikille } j = 1, \ldots, 18.$$

Voit käyttää priorina esimerkiksi tasajakaumia $\theta_j \sim \text{Beta}(1, 1)$ kaikille $i = 1, \ldots, 18$. Voit siis yksinkertaisesti laskea posteriorijakauman parametrille θ_j jokaiselle pelaajalle erikseen.

Piirrä mallille kuva (vrt. kirjan sivun 113 kuva 5.4), jossa x-akselilla ovat havaitut batting averaget $y_i/45$, ja y-akselilla posteriorijakaumien mediaanit osumatarkkudelle θ_j. Piirrä kuvaan myös parametrien θ_j posteriorijakaumien keskimmäiset kvantiilit $(0.25, 0.75)$ ja viiva 45 asteen kulmassa origosta (Kuvassa 1 esimerkki tehtävän 2 (b) mallille).

(b) Oletetaan, että kaikkien pelaajien todellinen osumatarkkuus on sama, eli että kaikkien 18 pelaajien tulokset ovat riippumaton otos samasta binomijakaumasta:

$$Y_j \sim \text{Binom}(45, \theta) \quad \text{kaikille } j = 1, \ldots, 18.$$

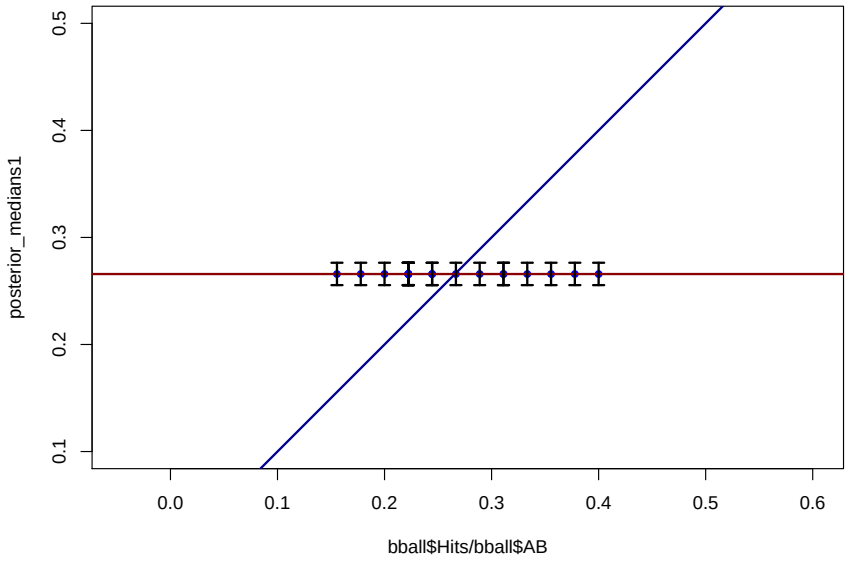
Voit käyttää priorina esimerkiksi tasajakaumaa $\theta \sim \text{Beta}(1, 1)$.

Piirrä vastaava kuva kuin a-kohdassa, ja lisää vielä vaakasuora viiva parametrin tämän mallin θ:n mediaanin kohdalle kumpaankin kuvaan.

3. Jatkoa edelliseen tehtävään.

(a) Toteutetaan oikea hierarkkinen malli, eli oletetaan, että pelaajien osumatarkkuudet θ_j ovat riippumaton otos samasta betajakaumasta tuntemattomilla parametreilla α ja β. Parametrisoidaan tämä betajakauma uudelleen sen odotusarvon $\phi = \frac{\alpha}{\alpha+\beta}$ ja "otoskoon" $\lambda = \alpha + \beta$

Kuva 1:



avulla, jolloin malli on

$$\begin{aligned} Y_j \,|\, \theta_j &\sim \text{Binom}(45, \theta_j), \\ \theta_j \,|\, \lambda, \phi &\sim \text{Beta}(\lambda\phi, \lambda(1-\phi)), \\ \phi \sim \text{Beta}(1,1),\ \lambda &\sim \text{Pareto}(0.1, 1.5),\ \lambda > 0.1 \end{aligned}$$

kaikille $j = 1, \ldots, 18$. Priorijakauma parametrille $p(\phi, \lambda)$ siis voidaan esittää tasajakauman ja Pareto-jakauman tulona. Estimoi tämä malli stanilla (vihje: λ:n jakauman voit määritellä komennolla `pareto(0.1,1.5)`. Pareto-jakauma on määritelty vain sen ensimmäistä argumenttia suuremmilla arvoilla, joten anna λ:lle määre `<lower=0.1>` kun määrittelet sen `parameters`-blokissa. Samoin muista rajata odotusarvo ϕ välille $(0, 1)$.

Piirrä vastaava kuva kuin edellisessä tehtävässä (vihje: tällä kertaa voit arvioida θ_j:n posteriorijakauman mediaanit ja kvantiilit `quantile`-funktiolla posteriorijakauman simuloiduista arvoista). Tulkitse kuvaa: miksi pisteet sijaitsevat samalla suoralla?

(b) Sovita aineistoon vielä hierarkkinen malli logit-muunnoksella $\alpha_j = \text{logit}(\theta_j) = \log \frac{\theta_j}{1-\theta_j}$ ja normaalilla populaatiojakaumalla:

$$\begin{aligned} Y_j \,|\, \alpha_j &\sim \text{Binom}(45, \text{logit}^{-1}(\alpha_j)), \\ \alpha_j \,|\, \mu, \sigma &\sim N(\mu, \sigma), \\ \mu \sim N(-1, 1),\ p(\sigma) &\propto N(0, 1),\ \sigma > 0 \end{aligned}$$

kaikille $j = 1, \ldots, 18$. Populaatiojakauman parametrin μ priorina käytetään siis normaalijakauma, ja varianssille (puolikasta) normaalijakaumaa positiivisella reaaliakselilla (vihje: kun annat parametrille σ määreen `<lower=0>`, niin voit normaalisti määritellä sen noudattamaan

normaalijakaumaa `model`-blokissa: stan tajuaa automaattisesti, että kyseessä on puolinormaalijakauma).

Piirrä vastaava kuva kuin edellisissä kohdissa. Kumpi hierarkkisista malleista ”kutisti” yksittäisten pelaajien mediaaneja enemmän kohti yhteistä mediaania?

4. Jatkoa kahteen edelliseen tehtävään. Verrataan malleja ennustamalla ”validation set”:in, eli kaikkien pelaajien havaituilla loppukauden osuneiden lyöntien $\tilde{y}_1, \ldots, \tilde{y}_{18}$ (suhteessa lyöntivuoroihin), arvoja.

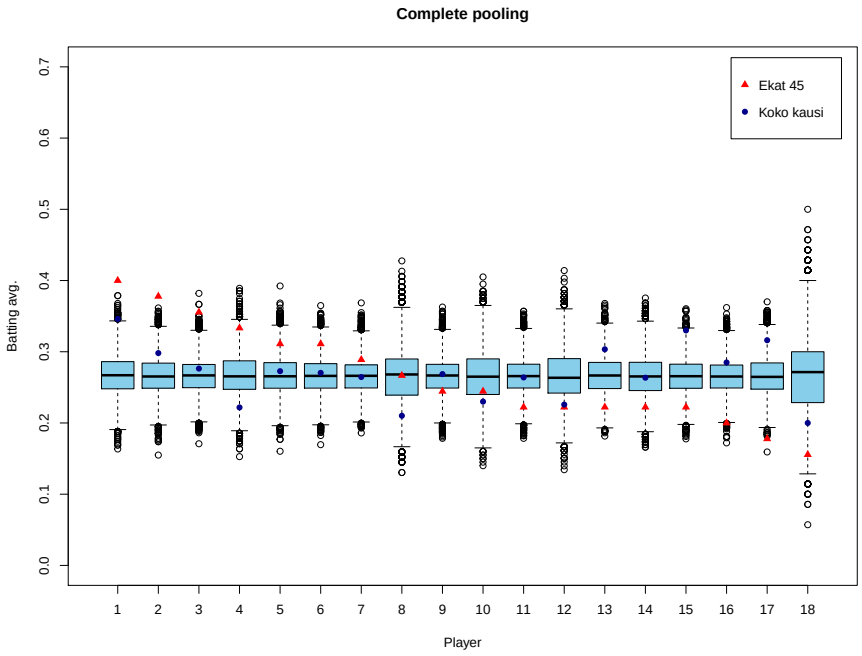
(a) Minkä malleista arvelisit etukäteen antavan parhaimman, ja minkä huonoimman ennusteen? Miksi?

(b) Oletetaan ensin riippumaton otos binomijakaumasta $\tilde{Y}_1, \ldots, \tilde{Y}_{18}$, joille

$$\tilde{Y}_j \,|\, \theta_j \sim \text{Binom}(m_j, \theta_j)$$

kaikille $j = 1, \ldots, 18$ ($m_1, \ldots, m_{18}$ ovat siis muuttujan `RemainingAB` arvot). Simuloi otokset posterioriennustejakaumasta $p(\tilde{\mathbf{y}}|\mathbf{y})$ kaikille malleille, ja piirrä niiden jakaumista viiksilaatikkokuvat, joihin merkitset lisäksi alkukauden ja loppukauden batting averaget kaikille pelaajille (Kuvassa 2 esimerkki tehtävän 2 (b) mallille). Miten vahvasti alku- ja loppukauden batting averaget riippuvat kuvan perusteella toisistaan?

Kuva 2:



(c) Laske sen jälkeen posterioriennustejakauman logaritmi

$$\log p(\tilde{\mathbf{y}}|\mathbf{y}) = \sum_{j=1}^{18} \log p(\tilde{y}_j|\mathbf{y})$$

(jota kutsutaan monesti log likelihoodiksi) havaituille loppukauden osumamäärille kaikille malleille.

Kahdelle konjugaattimallille saat posterioriennustejakauman arvot suoraan `VGAM`-kirjaston `dbetabinom.ab`-funktion avulla. Hierarkkisille malleille voit käyttää posteriorijakaumasta simuloituja parametrin θ_j:n arvoja (merkitään i:ttä simuloitua arvoa $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{18}^{(i)})$):

$$\log p(\tilde{\mathbf{y}}|\mathbf{y}) \approx \log \left(\frac{1}{L} \sum_{i=1}^{L} p(\mathbf{y}|\boldsymbol{\theta}^{(i)}) \right),$$

missä L on simuloidun otoksen koko.

(d) Mikä malli ennusti huonoiten ja mikä parhaiten (eli mikä malli antoi pienimmän, ja mikä suurimman todennäköisyyden oikeasti toteutuneelle aineistolle, eli loppukauden batting averageille)? Oliko järjestys sama kuin etukäteen veikkasit? Miten selittäisit tulosta?

Viitteet

[1] Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.