Bayes-päättely, 1. harjoitukset (25.–26.1.2017)

1. Poltetaan loppuun n hehkulamppua, ja merkitään muistiin niiden kestoiät, eli havaitaan aineisto $y_1, \dots, y_n$. Oletetaan, että kestoiät $Y_1, \dots, Y_n$ noudattavat eksponenttijakaumaa riippumattomasti ehdolla parametri $\theta > 0$, eli $Y_1, \dots, Y_n \,|\, \theta \sim Exp(\theta) \perp\!\!\!\perp$, jolloin

$$f_{Y_i|\Theta}(y_i|\theta) = \theta e^{-\theta y_i}$$

kaikille $i \in 1, \dots, n$.

(a) Johda parametrin θ suurimman uskottavuuden estimaatti.

(b) Johda parametrin θ posteriorijakauma $f_{\Theta|\mathbf{Y}}$, kun priorijakauma on gammajakauma $\mathrm{Gam}(\alpha, \beta)$, eli

$$f_\Theta(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}.$$

(c) Valitaan hyperparametrit $\alpha = \beta = 0.01$. Havaitaan kolmen lampun kestoiät $y_1, \dots, y_3$, joiden keskiarvo $\bar{y} = \frac{1}{n} \sum_{i=1}^3 y_i = 1000$ Piirrä posteriorijakauman tiheysfunktion kuvaaja[1] `R`:llä. Havaitaan kaksi muuta aineistoa, joiden keskiarvo on sama, mutta toisessa aineistossa lamppuja on $n = 5$, ja kolmannessa $n = 25$. Piirrä myös näiden aineistojen kuvaajat samaan kuvaan eri väreillä. Piirrä vielä priorijakauman kuvaaja (gammajakauman tiheysfunktio saadaan funktiolla `dgamma`) katkoviivalla. Mitä voit sanoa näiden perusteella sanoa parametrin arvoon liittyvästä epävarmuudesta otoskoon kasvaessa?

2. Jatkoa edelliseen tehtävään.

(a) Johda jakauma seuraavan hehkulampun kestoiälle $\tilde{Y}$ ehdolla aiemmat havainnot, kun $\tilde{Y}, Y_1, \dots, Y_n \,|\, \theta \sim Exp(\theta) \perp\!\!\!\perp$, eli oletetaan, että se on uusi havainto samasta jakaumasta. Toisin sanoen, johda $\tilde{Y}$:n posterioriennustejakauma $f_{\tilde{Y}|\mathbf{Y}}(\tilde{y}|\mathbf{y})$. Osoita, että se on ($\tilde{y}$:n funktiona) muotoa

$$f(x) = \frac{a\lambda^a}{(x+\lambda)^{a+1}},$$

missä

$$a := \alpha + n,$$
$$\lambda := \beta + \sum_{i=1}^n y_i.$$

(b) Piirrä posterioriennustejakauman tiheysfunktion kuvaajat `R`:llä (tämän jakauman tiheysfunktiota ei löydy `R`:stä valmiina, mutta voit kirjoittaa funktion itse!) edellisen tehtävän kolmella eri aineistolla (esimerkiksi välillä (0, 5000)). Piirrä samaan kuvaan myös $\tilde{Y}$:n jakauman tiheysfunktio, jos sen ennustamiseen käytetään piste-estimaattia, eli oletetaan, että $\tilde{Y} \sim \mathrm{Exp}(\hat{\theta})$, missä $\hat{\theta}$ on parametrin θ edellisessä tehtävässä laskettu SU-estimaatti.

Millä näistä jakaumista on pienin, ja millä suurin varianssi (tämän voit päätellä suoraan tiheysfunktion kuvaajista)? Mistä arvelisit tämän johtuvan?

[1] Kuvaajan x-akseliksi kannattaa välitä hyvin lyhyt väli, esim. (0, 0.005).

3. Binomijakauman yleistys useammalle kuin kahdelle eri vaihtoehdolle. Oletetaan, että satunnaisvektori $\mathbf{Y} = (Y_1, \dots, Y_d)$ noudattaa multinomijakaumaa parametrilla $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$, $\sum_{i=1}^d \theta_i = 1$, eli $\mathbf{Y} \,|\, \boldsymbol{\theta} \sim \text{Multin}(\boldsymbol{\theta})$, jolloin

$$f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \,|\, \boldsymbol{\theta}) = \binom{n}{y_1, \dots, y_d} \prod_{i=1}^d \theta_i^{y_i}$$

Johda parametrin $\boldsymbol{\theta}$ posteriorijakauma $f_{\boldsymbol{\Theta}|\mathbf{Y}}$, kun priorijakauma on Dirichlet-jakauma[2] $\text{Dir}(\boldsymbol{\alpha})$, missä $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$, eli

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_d)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \prod_{i=1}^d \theta_i^{\alpha_i - 1}.$$

4. Jatkoa edelliseen tehtävään.

(a) Tarkastellaan tilannetta, jossa $d = 3$. Simuloi[3] 5000 pisteen otokset priorijakaumasta $\text{Dir}(\boldsymbol{\alpha})$ parametreillä $\boldsymbol{\alpha} = (1/3, 1/3, 1/3)$, $\boldsymbol{\alpha} = (1, 1, 1)$ ja $\boldsymbol{\alpha} = (10, 10, 10)$, ja piirrä 3-ulotteiset hajontakuvat näistä otoksista[4]. Minne priorijakauman todennäköisyysmassa on keskittynyt $\boldsymbol{\alpha}$:n eri arvoilla?

(b) Käytetään priorina tasajakaumaa $\text{Dir}(1, 1, 1)$, ja havaitaan aineisto $\mathbf{y} = (30, 15, 2)$. Simuloi 5000:n otos posteriorijakaumasta $f_{\Theta|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y})$, ja piirrä siitä 3-ulotteinen hajontakuva. Vertaa todennäköisyysmassan jakautumista edellisessä kohdassa simuloituun priorijakaumaan.

[2] Dirichlet-jakauma on beta-jakauman yleistys useampaan ulottuvuuteen: kaksiulotteinen Dirichlet-jakauma $\text{Dir}(\alpha_1, \alpha_2)$ on sama asia kuin beta-jakauma $\text{Beta}(\alpha_1, \alpha_2)$.

[3] Kirjastosta `gtools` löytyy funktio `rdirichlet`

[4] `R`:stä löytyy vakiona funktio `scatterplot3d`, mutta hienomman pyöriteltävän 3-ulotteisen kuvan saa kirjaston `car` funktiolla `scatter3d` (vaatii myös kirjaston `rgl` toimiakseen).