Markovin ketju Monte Carlo, Gibbsin otanta-algoritmi ja Metropolis-Hastings algoritmi

Topias Karjalainen

30. huhtikuuta 2020

Sisältö

Luku 1

Johdanto

Tilastotieteissä *frekventistinen* koulukunta oli pitkään vallitseva koulukunta. Viimeaikoina kuitenkin suosiotaan on kasvattanut Bayesilainen koulukunta. Aiemmin Bayesiläinen päättely ei päässyt leviämään, sillä toisin kuin frekventistinen koulukunta, Bayesiläisyys ei tarjonnut suurinpaan osaan kysymyksiä analyyttisiä ratkaisuja. Vasta tietokoneiden aikakautena *Markovin ketju Monte Carlo -menetelmät* (MCMC-menetelmät) ovat antaneet mahdollisuuden ratkaista *posteriori*-jakaumat monimutkaisemmilta malleilta.

Monte Carlo menetelmän kehitteli 50-luvulla *Los Alamosissa* työskennelleet *Nicholas Metropolis*, *Stanislav Ulam* ja yleisnero *John von Neumann*. Yleinen määritelmä Monte Carlo menetelmälle on toistuva satunnainen arvojen arpominen. Yksinkertainen esimerkki Monte Carlo simuloinnista on esimerkiksi π:n arvon estimointi arpomalla sattumanvaraisesti pisteitä tasosta, ja laskemalla kuinka moni niistä on ympyrän säteen sisällä.

Markovin ketjut ovat *stokastisia prosesseja*, jotka on nimetty venäläisen matemaatikon *Andrey Markov*'n mukaan.

Tässä tutkielmassa tulen ensin antamaan lyhyen johdatuksen Markovin ketjuista ja selostan MCMC-menetelmien kannalta relevantin teorian. Tulen esittelemään lyhyesti kaksi algoritmia, joita käytetään MCMC-menetelmissä, *Gibbsin otanta-algoritmin* ja *Metropolis–Hastings algoritmin*. Esitän myös kaksi tärkeää diagnostiikkaa menetelmien tulosten arviointiin ja lopuksi vielä tarkastellaan hieman laajempaa käytännön esimerkkiä MCMC-algoritmistä.

Luku 2

Teoriaa

2.1 Perusmääritelmiä

Määritellään ensiksi todennäköisyys.

Määritelmä 2.1. *σ-algebra.* Olkoot Ω mielivaltainen epätyhjä joukko. Sigma-algebra perusjoukolla Ω on sen osajoukkojen joukkoperhe $\mathcal{F}$, joka toteuttaa ehdot:

1. $\emptyset \in \mathcal{F}$
2. jos $A \in \mathcal{F},$ *niin* $A^c \in \mathcal{F}$
3. jos jos $A_k \in \mathcal{F},$ *kaikilla* $k \in K$, missä K on numeroituva joukko, niin $\bigcup_{k\in K} A_k \in \mathcal{F}$

Määritelmä 2.2. Kuvaus $\mathbf{P}$ liittää kuhunkin tapahtumaan A *todennäköisyyden*, joka on luku suljetulla välillä [0,1] ja sille pätee:

1. $\mathbf{P}(\Omega) = 1$
2. Jos A on tapahtuma, niin sen komplementtitapahtuman A^c todennäköisyys on $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$
3. Jos $(A_k)_{k\in\mathbb{N}}$ ovat erillisiä tapahtumia, niin

$$\mathbf{P}(\bigcup_{k\in\mathbb{N}} A_k) = \sum_{k\in\mathbb{N}} \mathbf{P}(A_k)$$

Määritelmä 2.3. Olkoot $\mathcal{A}$ σ-algebra, ja olkoot X joukko. Pari $(X, \mathcal{A})$ on mitallinen avaruus.

Määritelmä 2.4. Kolmikkoa $(\Omega, \mathcal{F}, \mathbf{P})$ kutsutaan *todennäköisyysavaruudeksi.*

Määritelmä 2.5. *Satunnaismuuttuja* X on (lähes) mielivaltainen kuvaus $X : \Omega \to S$, jossa S on *tilajoukko.*

2.2 Markovin ketjut

Esitellään ensiksi joitain perus asioita Markovin Ketjuista, sillä ne eivät kuulu sellaisenaan opetussuunnitelmaan. Aloitetaan määrittelemällä stokastinen prosessi.

Määritelmä 2.6. Jono $(X_n : n = 1, 2, 3, ...)$ satunnaismuuttujia on diskreettiaikainen *stokastinen prosessi*.

Merkintä 2.7. Merkitään stokastista prosessia merkinnällä (X_n)

Määritellään nyt siirtymäydin, eli jakauma, joka määrittelee tilojen välisten siirtymien todennäköisyydet ja esitetään sitten *Markovin ehto* diskreetille tila-avaruudelle määritelmässä 2.11, ja laajennetaan se jatkuvalle tila-avaruudelle määritelmässä 2.14.

Määritelmä 2.8. Olkoot $(S, \mathcal{S})$ ja $(T, \mathcal{T})$ mitallisia avaruuksia. Siirtymäydin on funktio $T : S \times \mathcal{T} \to [0, \infty[$, jolle pätee

(i) $\forall s \in S : A \to T(s, A)$ on todennäköisyysmitta.

(ii) $\forall A \in \mathcal{T} : s \to T(s, A)$ on mitallinen

Diskreetissä tapauksessa siirtymäydintä kutsutaan *siirtymä matriisiksi*, joka on

$$p_{ij} = \mathbf{P}(X_n = i | X_{n-1} = j),\ \forall i, j \in S \tag{2.9}$$

Jatkuvassa tapauksessaydin kuvaa sitä ehdollista todennäköisyyttä, että siirtymä tapahtuu, eli $P(X \in A|x) = \int_A T(x, x')dx'$. Tälle pätee, että

$$\int_S T(x, x')dx' = 1 \tag{2.10}$$

Määritelmä 2.11. Stokastinen prosessi (X_n) on *Markovin ketju*diskreetissä tila-avaruudessa, jos kaikilla alkuhetkillä m, n ja tiloilla $i, j \in S$ on voimassa

$$\begin{aligned} &\mathbf{P}(X_{n+1} = j | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}, X_n = i) \\ &= \mathbf{P}(X_{n+1} = j | X_n = i) \end{aligned} \tag{2.12}$$

ja *siirtymätodennäköisyyksille* on voimassa

$$p_{ij} = \mathbf{P}(X_{n+1} = j | X_n = i) = \mathbf{P}(X_{m+1} = j | X_m = i) \tag{2.13}$$

Yhtälöä 2.12 kutsutaan *Markovin-ehdoksi* ja yhtälöä 2.13 taas kutsutaan *stationarisuusehdoksi*, mikä tarkoittaa, että siirtymätodennäköisyys tilojen i ja j välillä ei riipu ajasta m ja n, vaan pelkästään tiloista i ja j.

Määritelmä 2.14. Stokastinen prosessi (X_n) on *Markovin ketju* jatkuvassa tila-avaruudessa, jos kaikilla alkuhetkillä t, X_t:n ehdolliselle jakaumalle pätee

$$P(X_t \in A|x_{t-1}, x_{t-2}, ..., x_0) = P(X_t \in A|x_{t-1}) \tag{2.15}$$

Määritelmä 2.16. Satunnaismuuttujan X_0 jakaumaa kutsutaan *alkujakaumaksi.*

MCMC-menetelmien kannalta keskeinen ominaisuus Markovin ketjulle on sen tasapaino jakauma (eng. *invariant distribution*). Siihen perustuu koko idea menetelmän takana. Esitetään seuraavaksi tämä ominaisuus, sekä määritellään toinen ominaisuus eli *kääntyvä Markovin ketju* esittämällä ns. *detailed balance*-yhtälö, jota tarvitaan Metropolis-hastings algoritmin tasapainojakauman olemassa olon osoittamiseen.

Määritelmä 2.17. Todennäköisyysjakauma $\pi = (\pi)_{i \in S}$ on diskreetin tila-avaruuden Markovin ketjun (X_n) tasapainojakauma, jos

$$\sum_{i \in S} \pi_i p_{ij} = \pi_j, \forall j \in S \tag{2.18}$$

Yhtälö 2.18 voidaan kirjoittaa myös muotoon

$$\pi^T \mathbf{P} = \pi^T \tag{2.19}$$

Jakauma π on jatkuvan tila-avaruuden Markovin ketjun (X_n) tasapainojakauma jos

$$\pi(y) = \int_S \pi(x) T(x, y) dx \tag{2.20}$$

Määritelmä 2.21. Markovin ketju on *kääntyvä*, jos löytyy sellainen TN-jakauma $\lambda = (\lambda_i)_{i \in S}$, että

$$\lambda_i p_{ij} = \lambda_j p_{ji}, \forall i, j \in S \tag{2.22}$$

Määritelmä 2.23. Markovin ketju jatkuvassa S:ssä on kääntyvä, jos on olemassa

$$\pi(x) T(x, y) = \pi(y) T(y, x), \forall x, y \in S \tag{2.24}$$

Kääntyvällä Markovin ketjulla on sellainen mukava ominaisuus, että ketjun kääntyvyys on riittävä ehto tasapainojakauman olemassa ololle. Osoitamme tämän seuraavaksi.

Lause 2.25. *Jos Markovin ketju on kääntyvä, niin $\lambda = \pi$ on sen tasapainojakauma.*

Todistus.

$$\sum_{i \in S} \lambda_i p_{ij} = \sum_{i \in S} \lambda_j p_{ji} = \lambda_j \sum_{i \in S} p_{ij} = \lambda_j$$

□

Lause 2.26. *Jos Markovin ketju (X_n) on kääntyvä ja tilajoukko S on jatkuva, niin π on sen tasapainojakauma.*

Todistus. Yhtälön 2.10 mukaan $\int_S T(y,x)dx = 1$, joten

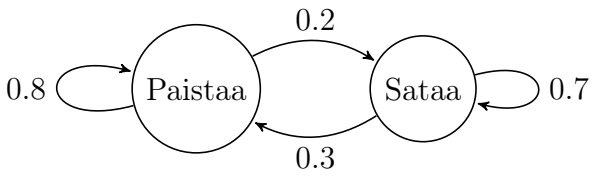
$$\int_S \pi(x)T(x,y)dx = \int_S \pi(y)T(y,x)dx = \pi(y)\int_S T(y,x)dx = \pi(y) \tag{2.27}$$

□

Esimerkki 2.28. Pohditaan lyhyttä esimerkkiä, jossa tilajoukko on $S = \{"sataa", "paistaa"\}$. Määritellään siirtymätodennäköisyydet siirtymämatriisilla

$$\mathbf{P}^{(1)} = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}$$

Tämä voidaan visualisoida kuvan 2.28 mukaisesti. Ketju on äärellinen, joten sillä on ta-



Kuva 2.1: Esimerkki 2.28

sapainojakauma. Yhtälö 2.19 implikoi, että jakauma π on siirtymämatriisin $\mathbf{P}$ vasen ominaisvektori ($\pi^T\mathbf{P} = \lambda\pi^T$, jossa $\lambda = 1$). Tämä voidaan ratkaista numeerisesti, ja ratkaisu on $\pi^T = (0.4, 0.6)$. Helposti nyt nähdään, että 2.19 pätee.

Seuraavaksi esitellään lyhyeksi *syklinen siirtymäydin*, ja osoitetaan tulos, jota tarvitsemme Gibbsin otanta-algoritmin tasapainojakauman todistuksen yhteydessä.

Määritelmä 2.29. Markovin ketjuissa voimme myös tietyin ehdoin yhdistää siirtymätiheyksiä. Tällöin puhutaan syklisistä siirtymätiheyksistä. Voidaan esimerksiksi määritellä siirtymätiheys

$$T_1 ... T_d \tag{2.30}$$

Lause 2.31. *Jos π on tasapainojakauma kaikille siirtymätiheyksille $T_1, ..., T_d$, niin se on tasapainojakauma siirtymätiheydelle $T = T_1...T_d$.*

Todistus.

$$\pi T = \pi T_1...T_d = \pi T_2...T_d = ... = \pi T_d = \pi$$

□

Käydään vielä nopeasti läpi muutama ominaisuus, joita tarvitaan kun haluamme osoittaa kappaleessa 3 algoritmiemme toimivan.

Määritelmä 2.32. Markovin ketju (X_n), siirtymäytimellä $T(x, y)$ on *pelkistymätön* jos kaikilla $A \in \mathcal{T}$, joilla $\mathbf{P}(A) > 0$, on olemassa sellainen n, että $T^n(x, A) > 0$ kaikilla $x \in S$

Käytännössä pelkistymättömyys tarkoittaa siis sitä, että jokaisesta tila-avaruuden kolkasta on mahdollista päästä jokaiseen muuhun pisteeseen avaruutta, eli ketju ei voi jäädä jumiin johonkin alueelle.

Määritelmä 2.33. Markovin ketju (X_n) on *palautuva* jos

(i) ketju (X_n) on pelkistymätön ja

(ii) kaikilla $A \in \mathcal{T}$, joilla $\mathbf{P}(A) > 0$, $\mathbb{E}_x[\eta_A] = \infty$ kaikilla $x \in A$

Missä η_A on käyntien määrä joukossa A.

Palautuvalla ketju tarkoittaa sitä, että ketju palaa alueelle, jossa se on jo käynyt. Palautumista vahvempi ominaisuus on *Harris*-palautuvuus.

Määritelmä 2.34. Joukko A on Harris palautuva, jos $\mathbf{P}_x(\eta_A = \infty) = 1$ kaikilla $x \in A$. Markovin ketju (X_n) on Harris palautuva jos (X_n) on pelkistymätön, ja jokainen joukko A, jolla $\mathbf{P}(A) > 0$, on Harris palautuva.

Määritelmä 2.35. Markovin ketju (X_n) on perioidinen, jos on olemassa erilliset osajoukot $A_1, A_2, ..., A_d \subset S, d > 1$, että

$$T(x, A_{i+1}) = 1, \ \forall \in A_i, \ i = 1, ..., d-1 \tag{2.36}$$

ja

$$T(x, A_1) = 1, \ i = d \tag{2.37}$$

Jos ketju ei ole perioidinen, se on *aperiodinen.*

Määritellään sitten viekä ergodisuus, joka on tärkeä ominaisuus Markovin ketjun tasapainojakauman olemassaolon kannalta.

Määritelmä 2.38. Markovin ketju (X_n) on ergodinen, jos se on pelkistymätön, aperiodinen ja Harris palautuva.

Ergodisuus on MCMC-menetelmien kannalta tärkeä ominaisuus, sillä se takaa Markovin ketjun (X_n) konvergoitumisen uniikkiin tasapainojakaumaansa mistä tahansa tila-avaruuden S pisteestä. Tämän osoittaminen on melko hankalaa ja ylittää reilusti tämän tutkielman laajuuden, joten jätämme sen tekemättä.

Luku 3

Markovin ketju Monte Carlo menetelmät

Tässä luvussa aiomme esitellä MCMC-metodeja. Esittelemme alaluvussa 3.4 Gibbsin otanta-algoritmina tunnetun MCMC-menetelmän, ja sitten alaluvussa 3.2 esittelemme Metropolis–Hastings algoritmin. Osoitamme myös, että todellisuudessa alaluvun 3.4 algoritmi onkin todella vain erikoistapaus luvun 3.2 algoritmista. Pohditaan kuitenkin ensin menetelmän motivaatiota.

Määritelmä 3.1. MCMC-menetelmiksi kutsutaan jakaumaa p simuloiviin menetelmiin, jotka perustuvat siihen, että luodaan ergodinen Markovin ketju (X_n), jolla on tasapainojakaumana jakauma p.

Ketjun ergodisuus takaa siis sen, että ketju konvergoituu jokaisesta tila-avaruuden pisteestä. Se takaa, että ketjun empiirinen keskiarvo konvergoituu odotusarvoon

$$\mathfrak{J}_N = \frac{1}{N}\sum_{t=1}^{N} h(X_t) \tag{3.2}$$

konvergoituu odotusarvoon, eli

$$\lim_{N\to\infty} \mathfrak{J}_N \to \mathbb{E}_p[h(X_t)] = \int h(x)p(x)dx \tag{3.3}$$

Tällöin ketjun tiloja voidaan siis käsitellä i.i.d. otoksena tasapainojakaumasta.

Menetelmän ydin on siis rakentaa systemaattisella tavalla Markovin ketju, jonka tasapainojakaumana on haluamamme simuloitava jakauma. Tämän voisi kuvitella olevan kovin vaikeaa, mutta yllättävästi se onkin melko triviaalia.

3.1 Gibbsin otanta-algoritmi

Gibbsin otanta-algoritmi on tapa simuloida Bayesiläistä moniulotteista posteriorijakaumaa (eli ulottuvuuksia vähintään 2), kun suora otanta on hankalaa. Algoritmi on nimetty

amerikkalaisen fyysikon, *Josiah Willard Gibbs*'n (1839-1903) mukaan, mutta sen todellinen kehittäjä on veljekset *Donald Geman* (1943-) ja *Stuart Geman* (1949-) vuonna 1984 artikkelissa *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*.

Määritelmä 3.4. Olkoot θ parametrivektori, joka jaetaan d:hen osaan tai osavektoriin, eli $\theta = (\theta_1, \theta_2, ..., \theta_d)$. Gibbsin otanta-algoritmi määritellään seuraavanlaisesti:

1. Valitaan satunnainen indeksi j jolle $1 \leq j \leq d$
2. Arvotaan uusi tila jokaiselle osavektorille θ_j ehdollistamalla se jokaiselle muulle parametrille, eli vedetään arvot $\theta = (\theta_1, \theta_2, ..., \theta_d)$ jakaumista

   $$p(\theta_j | \theta_{-j}^{(n-1)}, y) \tag{3.5}$$

   missä $\theta_{-j}^{(n-1)}$ on kaikki muut θ:n komponentit paitsi js komponentti, näiden tämänhetkisillä arvoilla eli

   $$\theta_{-j}^{(n-1)} = (\theta_1^{(n)}, ..., \theta_{j-1}^{(n)}, \theta_{j+1}^{(n-1)}, ..., \theta_d^{(n-1)})$$

Yleensä ylläolevassa määritelmässä kohdassa 1. ei arvota uutta järjestystä, vaan järjestys päätetään alussa, ja sitä pidetään koko algortimin ajan samana. Tällöin kyseessä on ns. *systemic scan Gibbs sampler*.

Lause 3.6. *Määritelmän 3.4 mukaisen algoritmin tuottamalla Markovin ketjulla on tasapainojakauma $p(\theta)$. [5]*

Todistus. Olkoot $\theta^{(n-1)}$ alkuperäienen tila, ja $\theta_j^{(n)}$ uusi j:nennen parametrin tila. Nyt $\theta^{(n-1)}$:n ja $\theta_j^{(n)}$:n yhteis tiheys on

$$p(\theta^{(n-1)}) p_j(\theta_j^{(n)} | \theta_{-j}^{(n-1)}) \tag{3.7}$$

Nyt voidaan integroida

(3.8)

$$\begin{aligned}
\int p(\theta^{(n-1)}) p_j(\theta_j^{(n)} | \theta_{-j}^{(n-1)}) d\theta_j^{(n-1)} &= \int p(\theta_j^{(n-1)} | \theta_{-j}^{(n-1)}) p(\theta_{-j}^{(n-1)}) p_j(\theta_j^{(n)} | \theta_{-j}^{(n-1)}) d\theta_j^{(n-1)} \\
&= p(\theta_{-j}^{(n-1)}) p_j(\theta_j^{(n)} | \theta_{-j}^{(n-1)}) \int p(\theta_j^{(n-1)} | \theta_{-j}^{(n-1)}) d\theta_j^{(n-1)} \\
&= p(\theta_{-j}^{(n-1)}) p_j(\theta_j^{(n)} | \theta_{-j}^{(n-1)}) \\
&= p(\theta_j^{(n)}, \theta_{-j}^{(n-1)})
\end{aligned}$$

Eli Gibbs otanta-algoritmin päivitys ei muuta jakaumaa. Nyt voidaan soveltaa lausetta 2.31, jolloin voidaan todeta, että koska p tasapainojakauma jokaiselle $p_j(\theta_j^{(n)}|\theta_{-j}^{(n-1)})$, niin se on tasapainojakauma niiden yhteis siirtymätiheydelle

$$T = \prod_{j=1}^{d} p_j(\theta_j^{(n)}|\theta_{-j}^{(n-1)}) \tag{3.9}$$

□

Ohitetaan Gibbsin otanta-algoritmin kohdalla toistaiseksi esimerkit, ja palataan siihen kappaleessa 4, jossa tarkastelemme laajempaa esimerkkiä lineaarisen regression parissa. Toteutamme tämän Gibbsin otanta-algoritmina.

3.2 Metropolis–Hastings algoritmi

Metropolis–Hastings algoritmi on kehittelijöidenssä *Nicholas Metropolisksen* (1915-1999) ja *Wilfred Keith Hastings*:n (1930-2016) mukaan nimetty MCMC-menetelmä, jolla voidaan simuloida Bayesiläisessä analyysissa käytettäviä posteriori jakaumia myös silloin kun tiheys on mahdotonta määrittää analyyttisesti.

Algoritmin pohjan kehitti *Stanislav Ulam* ja *Metropolis* työskennellessään *Los Alamosissa* ja myöhemmin *Metropolis* kehitteli nykyään *Metropolis-algoritmina* tunnettua algoritmiä ja esittelivät sen artikkelissa *Equation of state calculations by fast computing machines*[6]. Tämä versio algoritmista vaati, että pian esiteltävä *ehdotusjakauma* on symmetrinen. Myöhemmin *Hastings* laajenti algoritmin koskemaan myös epäsymmetrisiä ehdotusjakaumia artikkelissa *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*

Merkintä 3.10. TN-jakauma $J_n(\cdot|\cdot)$ on niin sanottu *ehdotusjakauma* (*proposal distribution, jumping distribution*), josta *MH-algoritmissa* arvotaan ehdotus tila.

Määritelmä 3.11. Metropolis–Hastings algoritmi on seuraavanlainen

1. Valitaan aloitus tila θ_0 ja asetetaan $n = 0$
2. Generoidaan kandidaatti tila θ' satunnaisesti jakaumasta $J_n(\theta'|\theta_{n-1})$
3. Lasketaan suhde
$$r = \frac{p(\theta'|y)/J_n(\theta'|\theta_{n-1})}{p(\theta_{n-1}|y)/J_n(\theta_{n-1}|\theta')}$$

4. Asetetaan

$$\theta_t = \begin{cases} \theta', \text{todennäköisyydellä} \quad \min(r, 1) \\ \theta_{t-1}, \text{muuten} \end{cases}$$

Jossa $J_t(\theta'|\theta^{t-1})$ on ns. ehdotusjakauma (eng. proposal distribution).

Lause 3.12. *Määritelmän 3.11 algoritmi tuottaa Markovin ketjun jolla on uniikki tasapainojakauma $p(\theta)$*

Todistus. Todistus nojautuu Markovin ketjun kääntyvyysominaisuuteen (2.21 ja 2.23), eli

$$T(\theta^{(n)}|\theta^{(n-1)})p(\theta^{(n-1)}) = T(\theta^{(n-1)}|\theta^{(n)})p(\theta^{(n)}) \tag{3.13}$$

joka on siis riittävä ehto tasapainojakauman olemassaololle. Mietitään kahta tapausta: (1) $\theta^{(n)} \neq \theta^{(n-1)}$ ja (2) $\theta^{(n)} = \theta^{(n-1)}$. Tapauksen (2) siirtymä voi tapahtua kahdella tavalla. Joko kohdassa 4. ehdotus θ' hylätään, tai se hyväksytään, mutta osutaan sattumanvaraisesti takaisin samaan kohtaan. Kuitenkin selvästi nähdään, että ehto 3.13 pätee tilanteessa (2).

Tilanteessa (1) siirtymätodennäköisyys pisteestä $\theta^{(n-1)}$ pisteeseen $\theta^{(n)}$ on

$$T(\theta^{(n)}|\theta^{(n-1)}) = J_n(\theta^{(n)}|\theta^{(n-1)}) \min\left(\frac{p(\theta^{(n)})J_n(\theta^{(n-1)}|\theta^{(n)})}{p(\theta^{(n-1)})J_n(\theta^{(n)}|\theta^{(n-1)})}, 1\right) \tag{3.14}$$

Jota voidaan muokata helposti

$$\begin{aligned} T(\theta^{(n)}|\theta^{(n-1)}) &= J_n(\theta^{(n)}|\theta^{(n-1)}) \min\left(\frac{p(\theta^{(n)})J_n(\theta^{(n-1)}|\theta^{(n)})}{p(\theta^{(n-1)})J_n(\theta^{(n)}|\theta^{(n-1)})}, 1\right) \\ &= \frac{1}{p(\theta^{(n-1)})} \min\left(p(\theta^{(n)})J_n(\theta^{(n-1)}|\theta^{(n)}), p(\theta^{(n-1)})J_n(\theta^{(n)}|\theta^{(n-1)})\right) \end{aligned} \tag{3.15}$$

Nähdään kuitenkin, että yhtälön 3.15 alempi yhtäläisyys on symmetrinen eli

$$T(\theta^{(n-1)}|\theta^{(n)}) = \frac{1}{p(\theta^{(n)})} \min\left(p(\theta^{(n-1)})J_n(\theta^{(n)}|\theta^{(n-1)}), p(\theta^{(n)})J_n(\theta^{(n-1)}|\theta^{(n)})\right) \tag{3.16}$$

joten kerrotaan 3.15 termillä $p(\theta^{(n-1)})$ ja hyödynnetään 3.16 ominaisuutta

$$\begin{aligned} T(\theta^{(n)}|\theta^{(n-1)})p(\theta^{(n-1)}) &= \frac{1}{p(\theta^{(n-1)})} \min\left(p(\theta^{(n)})J_n(\theta^{(n-1)}|\theta^{(n)}), p(\theta^{(n-1)})J_n(\theta^{(n)}|\theta^{(n-1)})\right)p(\theta^{(n-1)}) \\ &= \frac{1}{p(\theta^{(n)})} \min\left(p(\theta^{(n-1)})J_n(\theta^{(n)}|\theta^{(n-1)}), p(\theta^{(n)})J_n(\theta^{(n-1)}|\theta^{(n)})\right)p(\theta^{(n)}) \\ &= T(\theta^{(n-1)}|\theta^{(n)})p(\theta^{(n)}) \end{aligned}$$

Eli myös tapauksessa (1) yhtälö 3.13 pätee. □

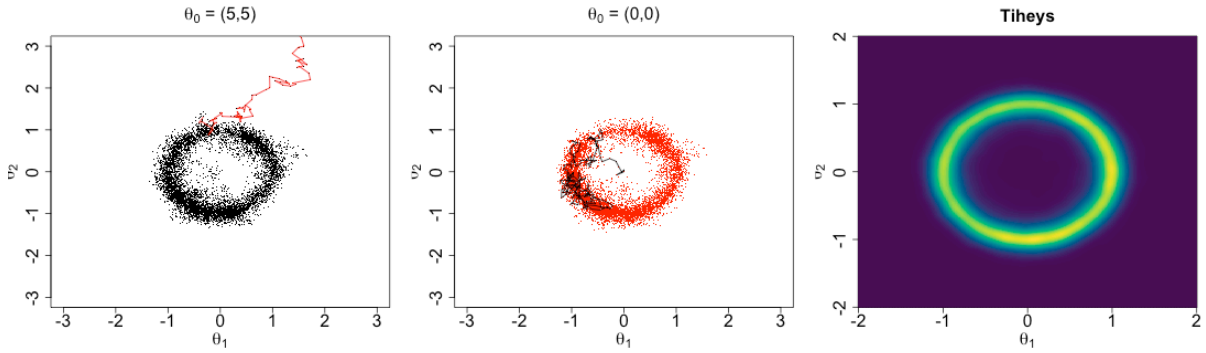
Esimerkki 3.17. Ajatellaan kuvitteellista tapausta, jossa meillä jatkuva kaksiulotteinen todennäköisyysjakauma, jonka tiheysfunktio on

$$p(\theta) \propto \exp\left(-5|\theta_1^2 + \theta_2^2 - 1|\right) \tag{3.18}$$

joka muodostaa regasmaisen 2-ulotteisen jakauman. Valitaan ehdotusjakaumaksi $J_n(\theta_n|\theta_{n-1})$ 2d-multinormaalijakauma

$$J_n(\theta_n|\theta_{n-1}) \sim N(\theta_{n-1}, \sigma^2 I_2) \tag{3.19}$$

jossa I_2 on 2x2 yksikkömatriisi ja olkoot $\sigma^2 = 0.01$. Nyt Metropolis Hastings algoritmin avulla voidaan simuloida jakaumaa $p(\theta)$ algoritmilla 3.11. Simuloidaan kaksi Markovin ketjua asettemalla aloitustiloiksi $(0, 0)$ ja $(5, 5)$, kummastakin 10 000 tilaa. Simuloimme



Kuva 3.1: *Vasemmalla: (5,5). Keskellä: (0,0) Oikealla: tiheysestimaatti (huomaa eri skaala).*

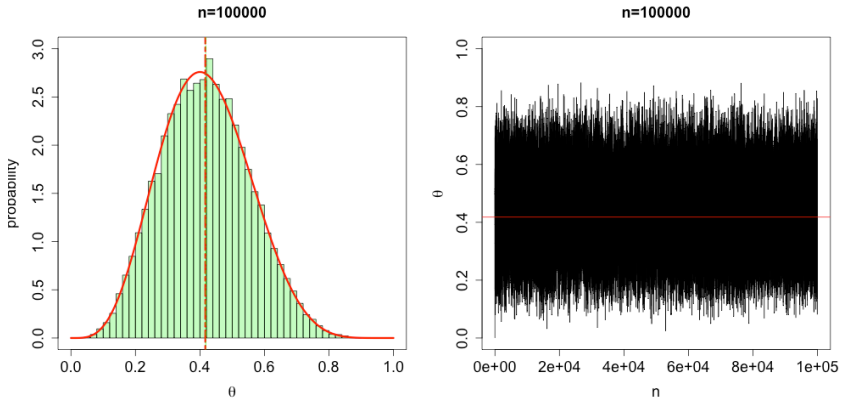
myös 200 000 pistettä aloitusarvolla (0,0), joista luodaan tiheysestimaatti. Tulokset löytyy kuvasta 4.2. Kahdessa ekassa kuvassa viiva on ensimmäisen 250 pisteen polku. Selvästi nähdään, että aloituspisteellä ei ole väliä. Markovin ketjun tasapainojakauma on sama huolimatta aloituspisteestä.

Esimerkki 3.20. Otetaan toisena esimerkkinä klassinen tapaus, jossa oletetaan, että havainnot ovat jauakautuneet *Bernoulli-jakauman* mukaan, $y_i \sim Bernoulli(\theta)$, ja että priori on tasajakauma. Tällöin tiedetään, että analyyttinen posteriori on

$$p(\theta|y_i) = Beta(\sum_{i=1}^{n} y_i + 1, n - \sum_{i=1}^{n} y_i + 1)$$

Valitaan hieman eksoottinen ehdotusjakauma esimerkin vuoksi:

$$J_n(\theta_n|\theta_{n-1}) \sim \begin{cases} Unif(\theta_{n-1}, 1) & \text{kun } \theta_{n-1} < 0.5 \\ Unif(0, \theta_{n-1}) & \text{kun } \theta_{n-1} \geq 0.5 \end{cases} \tag{3.21}$$



Kuva 3.2: Esimerkin 3.20 tulokset

Toisin kuin esimerkissä 3.17, nyt ehdotusjakauma ei olekkaan symmetrinen.

Oletetaan että, meillä on havainnot $(1, 1, 1, 0, 0, 1, 0, 0, 0, 0)$ ja tarkastellaan sekä analyyttistä että MH-algoritmin tuottamaa jakaumaa ja niiden eroja.

Kuvasta 3.2 nähdään esimerkin tulokset. Huomataan, että vaikka ehdotusjakauma on melko kummallinen, niin kuitenkin tarpeaksi monella iteraatiolla saavutetaan tasapainojakauma. Huomaa, että vasemmassa kuvaajassa on vihreällä simulaatio keskiarvo, ja punaisella analyyttinen keskiarvo, mutta nämä arvot ovat niin lähellä toisiaan, että viivat ovat päällekkäin.

3.2.1 Ehdotusjakauman valinnasta

Kummassakin kappaleen 3.2 esimerkissä valitsimme ehdotusjakauman melko satunnaisesti. Varsinkin esimerkissä 3.20 se on erittäin epätavallinen, mistä syystä hyvän approksimaation saavuttaminen vie todella monta iteraatiota. Yleensä jos haluamme oikeasti tehokkaasti ja ekonomisesti simuloida jakaumia esitetyllä algoritmilla, haluamme valita ehdotusjakauman jollakin järkevällä, systemaattisella tavalla, joka minimoisi tarvittavien iteraatioiden määrän.

Yleisesti ottaen hyvällä ehdotusjakaumalla on muutama ominaisuus[1]

1. Kaikilla θ:n arvoilla on helppo arpoa arvo $J(\theta'|\theta)$
2. Suhde r on helppo laskea
3. Siirtymät ovat tarpeaksi pitkiä. Muuten Markovin Ketju etenee liian hitaasti ja hyvän estimaatin saaminen kestää liian pitkään.

4. Siirtymiä ei hylätä liian usein. Muuten Markovin Ketju ei etene vaan seisoo paikallaan.

Lisäksi simulointia voidaan nopeuttaa mm. käyttämällä adaptiivista ehdotusjakaumaa, eli toisinsanoen ehdotusjakaumaa muunnellaan riippuen ketjun liikkeistä.

3.3 Yleisiä käytäntöjä MCMC-menetelmissä

Usein on tapana simuloida useampi kuin yksi ketju kustakin tapauksesta. Tällöin asetetaan näiden ketjujen aloitustilat eriäviksi, jotta saadaan vähennettyä aloitustilan vaikutusta lopputukokseen.

Toinen yleinen käytäntö on jättää ketjun alkupäästä jonkin verran tiloja huomiotta, sillä ketjun alkupäässä ei se ei välttämättä ole vielä saavuttanut tasapainojakaumaa. Tätä kutsutaan *Burn-in periodiksi*.

Joskus taas on hyvä tapa pudottaa joka n:s tila ketjusta. Tilojen tiputtaminen ei vaikuta tasapainojakaumaan, kunhan ketju vain on saavuttanut sen. Tilojen tiputtamista on hyötyä jos mallissa on paljon parametreja, jolloin tietokoneessa voi tulla ongelmia tilan kanssa.

3.4 Konvergenssi ja Diagnostiikka

Kappaleissa 3.4 ja 3.2 esiteltyjen algoritmien kohdalla voi herätä kysymys, että mikä on riittävä määrä iteraatioita, jotta Markovin ketju on saavuttanut tasapainojakaumansa ja otanta on riittävän hyvä aproksimaatio posteriorijaukaumasta. Esimerkiksi kun katsotaan kuvan 4.2 vasemman puolen kuvan punaista polkua, niin voimme sanoa, että se ei ole vielä saavuttanut tasapainojakaumaa sillä tunnemme melko hyvin halutun jakauman, mutta yleensä emme välttämättä osaa sanoa tätä suoraan.

Toinen yleinen ongelma, joka kaipaa pohdintaa on se, että ketjujen sisällä on korrelaatiota, mikä vaikeuttaa päättelyä simuloinnin tuloksista. Otannat eivät siis ole välttämättä täysin riippumattomia.

3.4.1 Gelmanin $\hat{R}$

Yksi yleisimmistä estimaattoreista, joita käytetään MCMC-metodeissa Markovin ketjujen konvergenssin arvioimiseen on *Andrew Gelmanin* $\hat{R}$ [1]. Se mittaa ketjujen sisäistä sekoittumista (eng. *mixing*) ja erillisten ketjujen välistä sekoittumista.

Määritelmä 3.22. *Gelmanin* $\hat{R}$ [2][3] lasketaan jakamalla ensin jokin määrä simuloituja ketjua erillisillä aloituspisteillä keskeltä kahtia. Olkoon nyt m ketjujen määrä jaon jälkeen ja n ketjujen pituus. Olkoot $\psi_{ij}(i = 1, ..., n; j = 1, ..., m)$ tila i ketjussa j. Nyt merkataan

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\psi}_{.j} - \overline{\psi}_{..})^2, \quad \text{jossa} \tag{3.23a}$$

(3.23b) $$\overline{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}$$

(3.23c) $$\overline{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\overline{\psi}_{.j}$$

Ja merkataan

(3.24) $$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \;\; jossa \;\; s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \overline{\psi}_{.j})^2$$

Jossa siis B on ketjujen välinen varianssi (*between sequence*) ja W on ketjujen sisäinen varianssi (*within sequence*). Näiden painotettuna keskiarvona saadaan estimaattori ψ:n marginaaliselle posteriori varianssille

(3.25) $$\hat{\sigma}^+(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

Nyt $\hat{R}$ voidaan määritellä

(3.26) $$\hat{R} = \sqrt{\frac{\hat{\sigma}^+(\psi|y)}{W}}$$

Gelmanin $\hat{R}$ mittaa sitä miten paljon kullakin hetkellä ψ:n jakauma voisi supistua, jos simulaatioiden annettaisiin jatkua loputtomasti. Eli siis kun $\hat{R} \approx 1$, niin voidaan sanoa, että Markovin ketju on todennäköisesti saavuttanut tasapainojakaumansa, sillä tällöin eri pisteitä aloitettujen ketjujen välinen varianssi on vakaantunut.

Tällä diagnostiikalla on kuitenkin onglemansa ja rajoitteensa. Se saattaa virheellisesti diagnosoida konvergoituneen ketjun esimerkiksi jos ketjulla on pitkät hännät. Sen takia käytetäänkin mielummin muita diagnostiikka keinoja. Esimerkiksi *Stan* käyttää korjattua $\hat{R}$ diagnostiikkaa [7].

3.4.2 ESS

Toinen hyödyllinen diagnostiikka MCMC-menetelmissä on niin sanottu *effective sample size (ESS)*, joka mittaa sitä kuinka monta effektiivisesti riippumatonta otosta ketjussa on. Tämä tarkoittaa siis sitä, että se mittaa kuinka paljon ketjun autokorrelaatio vaikuttaa keskivirheeseen verrattuna täysin riippumattomiin otoksiin.

Määritelmä 3.27. ESS lasketaan kaavalla

$$\hat{n}_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t} \tag{3.28}$$

jossa

$$\hat{\rho}_t = 1 - \frac{V_t}{2\hat{\sigma}^+} \tag{3.29a}$$

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{ij} - \psi_{i-t,j})^2 \tag{3.29b}$$

$\hat{\sigma}^+$ saadaan kaavasta 3.25. Termissä $\sum_{t=1}^{T} \hat{\rho}_t$ autokorrelaatioita summataan, kunnes kahden peräkkäisen kovariaatin summa on negatiivinen [4].

Luku 4

Laajempi esimerkki

Tarkastellaan vielä lopuksi laajempaa esimerkkiä. Tarkastellaan normaalia lineaarista regressiomallia. Käytetään `R`:stä löytyvää `airquality` dataa, jossa on dataa New Yorkin ilmanlaadusta. Regressoidaan datasetin muuttuja `Ozone` muuttujille `Solar.R` ja `Wind`. Sovitetaan tämä malli Gibbsin otanta-algoritmin avulla. Sitä varten meidän on ensiksi määriteltävä marginaaliset ehdolliset jakaumat parametreille.

Määritellään ensiksi malli. Merkitään $\tau = 1/\sigma^2$, $\theta = (\beta_0, \beta_1, \beta_2, \tau)$

$$
\begin{aligned}
y_i|\beta_0, \beta_1, \beta_2, \tau &\sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i, 1/\tau) \\
\beta_0|\mu_0, \tau_0 &\sim N(\mu_0, 1/\tau_0) \\
\beta_1|\mu_1, \tau_1 &\sim N(\mu_1, 1/\tau_1) \\
\beta_2|\mu_2, \tau_2 &\sim N(\mu_2, 1/\tau_2) \\
\tau|\alpha, \gamma &\sim Gamma(\alpha, \gamma)
\end{aligned} \tag{4.1}
$$

Muodostetaan uskottavuusfunktio

$$
L(y|\theta) = \prod_{i=1}^{N} N(\beta_0 + \beta_1 x_i + \beta_2 x_i, 1/\tau) \tag{4.2}
$$

Ja posterioiri on siten

$$
p(\theta|y) \propto p(\beta_0)p(\beta_1)p(\beta_2)p(\tau) \prod_{i=1}^{N} N(\beta_0 + \beta_1 x_i + \beta_2 x_i, 1/\tau) \tag{4.3}
$$

Tästä saadaan sitten johdettua marginaaliset jakaumat otanta-algoritmia varten

$$
\begin{aligned}
\beta_0|\beta_1,\beta_2,\tau_0,\tau,\mu_0,x,y &\sim N\left(\frac{\tau_0\mu_0+\tau\sum_{i=1}^{N}(y_i-\beta_1x_{i1}-\beta_2x_{i2})}{\tau_0+\tau N},\frac{1}{\tau_0+\tau N}\right)\\
\beta_1|\beta_0,\beta_2,\tau_1,\tau,\mu_1,x,y &\sim N\left(\frac{\tau_1\mu_1+\tau\sum_{i=1}^{N}(y_i-\beta_0-\beta_2x_{i2})x_{i1}}{\tau_1+\tau\sum_{i=1}^{N}x_{i1}^2},\frac{1}{\tau_1+\tau\sum_{i=1}^{N}x_{i1}^2}\right)\\
\beta_2|\beta_0,\beta_1,\tau_2,\tau,\mu_2,x,y &\sim N\left(\frac{\tau_2\mu_2+\tau\sum_{i=1}^{N}(y_i-\beta_0-\beta_1x_{i1})x_{i2}}{\tau_2+\tau\sum_{i=1}^{N}x_{i2}^2},\frac{1}{\tau_2+\tau\sum_{i=1}^{N}x_{i2}^2}\right)\\
\tau|\beta_0,\beta_1,\beta_2,\alpha,\gamma,x,y &\sim Gamma\left(\alpha+\frac{N}{2},\gamma+\frac{N}{2}\sum_{i=1}^{N}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2})^2\right)
\end{aligned}
\tag{4.4}
$$

Asetetaan hyperparametrit $\mu_0 = 80, \mu_1 = 0, \mu_2 = -5, \tau_0 = 1/50, \tau_1 = 1/50, \tau 2 = 1/50, \alpha = 5, \gamma = 0.01$. Parametrit on valittu sen mukaan, että ne asettavat suurimman tiheyden lähelle arvioitua sijaintia ja toisaalta eivät ole kovin informatiivisia vaan leveitä. Simuloidaan nyt tästä Gibbsin otanta-algoritmilla yhtälöitä 4.4 käyttäen 8 ketjua, kunkin pituus 2 000. Burnin-periodi olkoot 2000. Yhteensä siis meillä on 16 000 otosta.

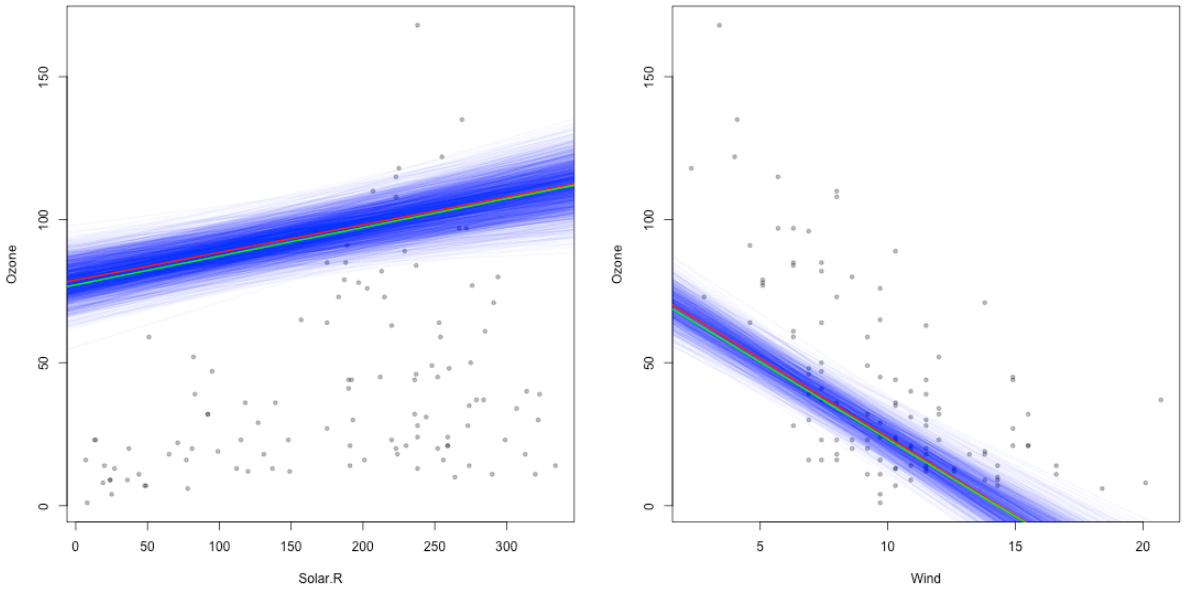
θ	Mean	SD	2.5%	25%	50%	75%	97.5%
β_0	78.89544	5.61842	67.97969	75.14744	78.89385	82.64874	89.86483
β_1	0.09675	0.02263	0.05196	0.08130	0.09688	0.11218	0.14106
β_2	-5.48880	0.51291	-6.48407	-5.83627	-5.49158	-5.13095	-4.48973
τ	0.00177	0.00023	0.00136	0.00161	0.00176	0.00192	0.00226

Taulukko 4.1: Tulokset regressiosta

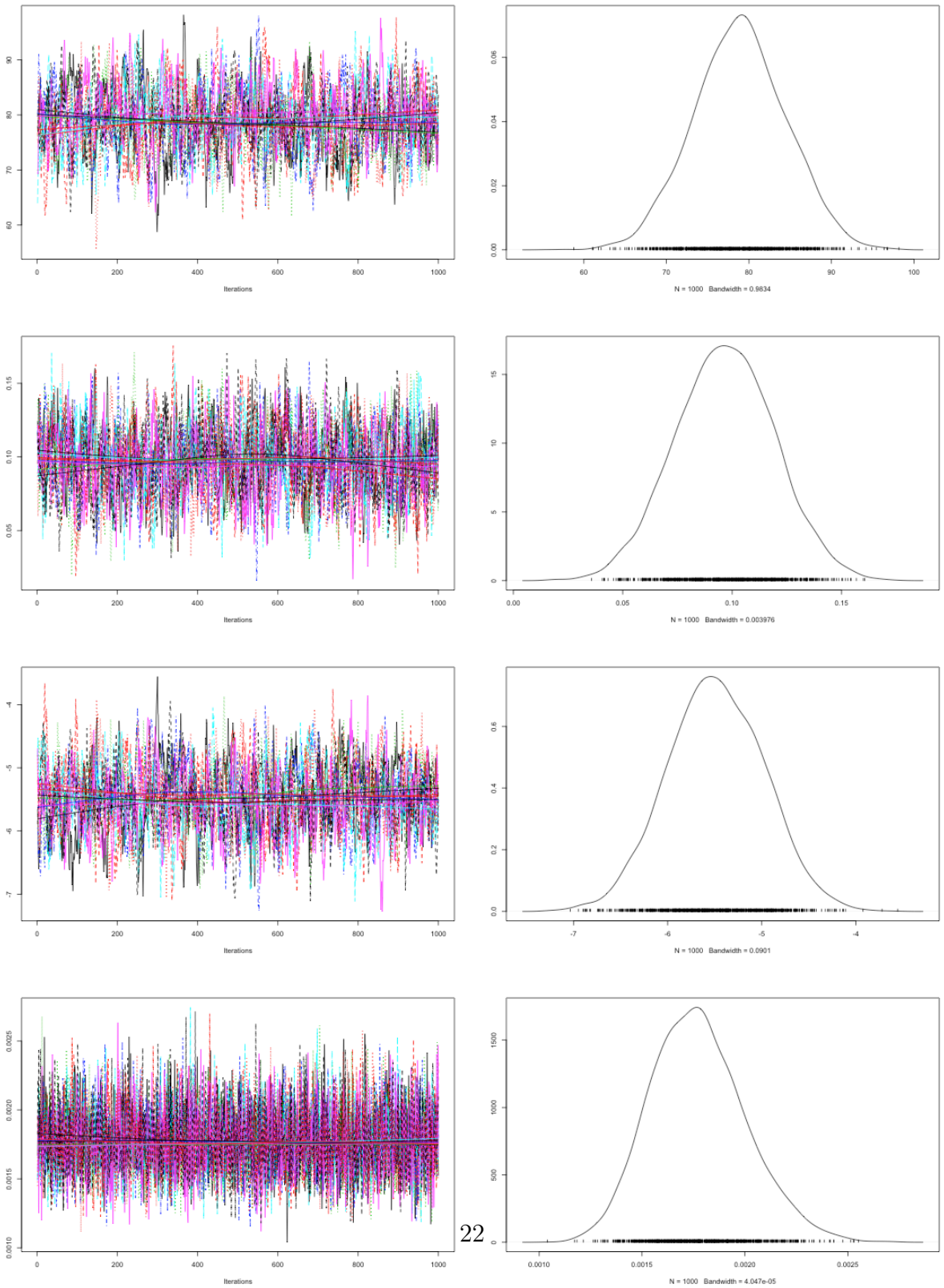
θ	$\hat{R}$	$\hat{n}_{eff}$
β_0	1.00	778
β_1	1.00	1177
β_2	1.01	902
τ	1.00	7656

Taulukko 4.2: Diagnostiikka

Taulukoissa 4.1 ja 4.2 nähdään tulokset ja MCMC-diagnostiikat, joista puhuttiin aiemmin. Nähdään, että ketjut näyttäisivät konvergoituneen, mikä ei sinänsä ihmetytä lainkaan, sillä simulaatio määrämme on valtava. Yhteensä burn-in periodin jälkeen otoksia on 8000.



Kuva 4.1: *Esimerkin regression tulos, vasemmalla Solar.R -akselilla ja oikealla Wind. Punainen viiva on posteriori keskiarvolla piiretty regressio viiva ja vihreä on vertailun vuoksi PNS mentelmällä sovitettu suora. Siniset ovat sovitteita parametrien eri ehdotus arvoilla.*



Kuva 4.2: *Esimerkin ketjut ja parametrien posteriori tiheydet*

Luku 5

Loppusanat

Olemme nyt tarkastelleet perus ajatuksia MCMC-menetelmistä. Kävimme läpi hieman teoriaa, esittelimme algoritmit ja kokeilimme sellaista käytännön tilanteessa. Kuitenkin olemme vain raapaisseet pintaa. Käyttämämme algoritmit ovat monessa mielessä erittäin rajoittuneita. Gibbsin otanta-algortimissa meidän tulee tuntea marginaalijakaumat analyyttisesti, ja Metropolis–Hastings voi olla välillä hidas konvergoitumaan.

Tärkeitä laajennuksia, jotka korjaavat edellisessäkappaleessa mainitsemiani onglemia, on myöhemmin kehitetyt tehokkaammat MCMC-metodit. Tärkeimpänä *Hamiltonian monte carlo* eli HMC. Se pyrkii vähentämään tarvittavien otosten määrää, joka tarvitaan tarpeeksi tarkan posteriori estimaatin saamiseksi, sillä menetelmä vähentää autokorrelaatiota tilojen välillä.

Olen lisäksi tarkastellut tässä työssä menetelmiä enemmänkin vain Bayesiläisen tilastotieteen kannalta. Todellisuudessa näille menetelmille on paljon muitakin potentiaalisia käyttökohteita, esimerkiksi optimointiongelmien parissa.

Kirjallisuus

[1] Andrew Gelman. *Bayesian Data Analysis*. 3. painos. ISBN: 978-1-4398-4095-5.

[2] Donald Rubin Gelman Andrew. "Inference from Iterative Simulation using Multiple Sequences". *Statistical Science* 7 (1992), s. 457–511.

[3] Stephen Brooks Gelman Andrew. "General Methods for Monitoring Convergence of Iterative Simulations". *Journal of Computational and Graphical Statistics* (joulukuu 1998). URL: `http://www.stat.columbia.edu/~gelman/research/published/brooksgelman2.pdf`.

[4] Charles Geyer. "Practical Markov Chain Monte Carlo". *Statistical Science* 7.4 (1992), s. 473–483.

[5] Petri Koistinen. *Computational Statistics*. 2009.

[6] Nicholas Metropolis. "Equation of state calculations by fast computing machines" (1953).

[7] Aki Vehtari et al. "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC". *arXiv e-prints*, arXiv:1903.08008 (maaliskuu 2019), arXiv:1903.08008. arXiv: `1903.08008 [stat.CO]`.