Funktiomuotoisen heritabiliteetin estimointi bayesiläisillä malleilla

Tilastotieteen pro gradu -tutkielma
Arttu Arjas
2319498
Matemaattisten tieteiden yksikkö
Oulun yliopisto
Kevät 2019

Tiivistelmä

Heritabiliteetti eli perinnöllisyysaste on perinnöllisyystieteen tunnusluku, joka mittaa sitä, kuinka suuri osa jonkin piirteen vaihtelusta tietyssä biologisessa populaatiossa riippuu perintötekijöistä. Sen avulla voidaan muun muassa selvittää, kuinka nopeasti valinnalla voidaan muuttaa piirteen keskiarvoa tulevissa sukupolvissa. Heritabiliteetti ei välttämättä ole vakio, vaan se voi olla ajan funktio. Kirjallisuudessa ei kuitenkaan ole kovin paljon esimerkkejä heritabiliteettifunktion estimoinnista.

Tässä työssä esitellään kaksi bayesiläistä menetelmää funktiomuotoisen heritabiliteetin estimointiin. Ensimmäisessä, kaksivaiheisessa menetelmässä heritabiliteetti ja sen luottamusväli estimoidaan jokaisessa aikapisteessä erikseen lineaarista sekamallia ja bootstrap-menetelmää apuna käyttäen, jonka jälkeen muodostuneet aikasarjat silotetaan. Toisessa menetelmässä sekamalli on ikään kuin yleistetty pitkittäisaineistotilanteeseen, jolloin heritabiliteetti saadaan estimoitua jokaisessa aikapisteessä samaan aikaan yksivaiheisesti sopivien silottavien priorien avulla.

Tutkimuksessa havaittiin, että yksivaiheisessa menetelmässä estimaatit ovat melko täsmällisiä estimoitaville parametreille asetettavien priorien takia, koska ne muodostavat eri aikapisteiden välille tietynlaisen korrelaatiorakenteen. Toisaalta tämä estimointitapa on kuitenkin laskennallisesti raskas. Kaksivaiheisessa menetelmässä estimaattien luottamusvälit ovat leveämmät, mutta menetelmä ei ole läheskään niin aikaavievä kuin yksivaiheinen estimointi.

Sisältö

1 Johdanto

Heritabiliteetti on tärkeä parametri perinnöllisyystieteessä. Se mittaa sitä, kuinka suuri osa jatkuva-asteikollisen piirteen vaihtelusta selittyy geneettisillä tekijöillä tietyssä populaatiossa. Se vaikuttaa muun muassa siihen, kuinka nopeasti valinnalla voidaan vaikuttaa ominaisuuden keskiarvoon tulevissa sukupolvissa. Heritabiliteetin mittaaminen on vaikeaa klassisilla menetelmillä, koska niitä varten tarvitaan joko tietoa populaation sukupuusta tai risteytyskokeita (Sillanpää, 2011).

Heritabiliteetin määritelmä lähtee siitä, että jonkin piirteen tai fenotyypin havaittu arvo jaetaan perintötekijöistä riippuvaan osaan ja ympäristötekijöistä riippuvaan osaan:

$$\text{Fenotyyppi}(P) = \text{Geenit}(G) + \text{Ympäristö}(E). \tag{1}$$

Fenotyypin varianssi voidaan esittää geneettisen varianssikomponentin ja ympäristövarianssikomponentin summana, jos geeneistä ja ympäristöstä riippuvat satunnaistermit oletetaan toisistaan riippumattomiksi

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2. \tag{2}$$

Yleensä riippumattomuusoletus perustuu ainoastaan siihen, että riippuvutta on vaikeaa tai mahdotonta mitata. (Visscher ym. 2008). Lopulta heritabiliteetti voidaan määritellä geneettisen varianssin osuutena kokonaisvarianssista

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}. \tag{3}$$

Varianssikomponentit ovat populaatioparametreja ja ne täytyy estimoida sopivasta aineistosta. Geneettisen varianssin estimointia varten täytyy

silloin olla informaatiota populaation sukulaisuusrakenteesta. Rakenteen voi ilmaista nk. sukulaisuusmatriisin avulla. Sukulaisuusmatriisin konstruoinnista on kerrottu alaluvussa 2.2. Heritabiliteetti ei välttämättä ole vakio, vaan se voi riippua ajasta. Esimerkiksi piirteeseen vaikuttavan geenin ilmentyminen saattaa vaihdella ajan suhteen (Bryois ym. 2017) ja ympäristö voi muuttua. Uusia automatisoituja menetelmiä eri fenotyyppien mittaamiseen on tullut käyttöön. Ne mahdollistavat esimerkiksi kasvien kasvuun ja kehitykseen liittyvien piirteiden täsmällisen mittaamisen yli ajan. Laadukas pitkittäisaineisto mahdollistaa biologisten ilmiöiden funktionaalisen mallintamisen, jossa voidaan hyödyntää sopivia tilastollisia malleja. Funktionaalisen mallintamisen iso hyöty on se, että siinä yhdistyy useiden ajanhetkien informaatio, jolloin on mahdollista, että estimointi tarkentuu verrattuna estimointiin vain yhdellä ajanhetkellä. (Li ja Sillanpää, 2015)

Automatisoidut aineistonkeräysmenetelmät luovat tarpeen tilastollisille menetelmille, jotka ottavat huomioon aineiston aikaulottuvuuden. Yleensä näiden menetelmien toivotaan myös olevan laskennallisesti nopeita (Kwak ym. 2014). Tässä tutkielmassa esitellään kaksi bayesiläistä menetelmää ajasta riippuvan heritabiliteetin mallintamiseen. Bayesiläisessä analyysissä (alaluku 2.1) voidaan luontevasti ottaa huomioon ennakkokäsitykset ja -tiedot tutkittavasta ilmiöstä nk. priorijakaumien avulla. Menetelmissä hyödynnettävät priorit perustuvat ajatukseen, että heritabiliteettifunktio (kuten oletettavasti moni muukin biologinen prosessi) on sileä (Pletcher ja Geyer, 1999). Eräs tällainen silottava priori on multinormaalijakauma, joka muodostaa aikapisteiden välille korrelaatiorakenteen kovarianssimatriisin avulla, mikä tarkentaa estimointia verrattuna estimointiin vain yhdellä ajanhetkellä.

Ensimmäinen menetelmä on kaksivaiheinen. Siinä heritabiliteetti ja sen

luottamusväli estimoidaan jokaisella ajanhetkellä erikseen, jonka jälkeen saadut aikasarjat silotetaan. Samantyylistä aikapisteittäistä estimointia funktioarvoisille ilmiöille ehdottaa myös Kwak ym. (2014). Menetelmä on nopea ja se skaalautuu hyvin aineiston koon kasvaessa, mutta se ei ota huomioon aikapisteiden välistä riippuvutta, jolloin koko aineiston tilastollinen voima jää hyödyntämättä. Toinen menetelmä estimoi heritabiliteettifunktion yhdellä kertaa. Hyvää tässä menetelmässä on se, että se ottaa aikapisteiden välisen riippuvuuden huomioon ja tuottaa täsmällisempiä estimaatteja kuin kaksivaiheinen menetelmä. Se on kuitenkin huomattavasti hitaampi eikä skaalaudu hyvin aineiston koon kasvaessa.

Tämän tutkielman rakenne on seuraava: Ensin avataan yleisesti bayesiläistä viitekehystä tilastollisessa päättelyssä, jonka jälkeen rakennetaan malli fenotyypille yhdellä ajanhetkellä. Kaksivaiheinen menetelmä perustuu tähän malliin. Tämän jälkeen malli yleistetään tilanteeseen, jossa havaintoja on usealta ajanhetkeltä. Yksivaiheinen menetelmä perustuu tähän yleistettyyn malliin. Seuraavaksi perehdytään siihen, miten biologisiin prosesseihin liittyvät ennakkokäsitykset saadaan liitettyä osaksi malleja, ja miten mallien parametrit voidaan estimoida. Lopuksi esitellään malleilla analysoitavat aineistot, analyysiin käytettävät algoritmit ja tulokset.

2 Mallien ja menetelmien rakentaminen

Tässä luvussa kerrotaan yleisesti bayesiläisestä analyysistä, jonka jälkeen rakennetaan mallit, joiden avulla heritabiliteettifunktiota estimoidaan. Lisäksi esitellään sukulaisuusmatriisi, joka on keskeinen osa malleja.

2.1 Bayesiläinen analyysi

Bayesiläinen päättely perustuu havaintoaineiston ja tutkittavaan aiheeseen liittyvien ennakkokäsitysten eli priorin yhdistämiseen. Gelman ym. (2013, s. 3) jakaa bayesiläisen data-analyysiprosessin kolmeen vaiheeseen.

1. Parametrien yhteis(priori)jakauman asettaminen: yhteistodennäköisyysjakauma kaikille mallin tuntemattomille parametreille.

2. Posteriorijakauman tutkiminen: kun parametrien yhteisjakauma ehdollistetaan havaintoaineistolle, saadaan posteriorijakauma, johon tilastollinen päättely kohdistetaan.

3. Päättely ja mallidiagnostiikka: kuinka hyvin malli sopii dataan ja millaisia päätelmiä posteriorijakaumasta voidaan tehdä sekä kuinka herkkä posteriorijakauma on mallioletusten muutoksille?

Bayesiläisen ja frekventistisen lähestymistavan perustavanlaatuinen ero on se, että bayesiläisessä mallin parametrit ovat ikään kuin ei-havaittavia satunnaismuuttujia, joilla on todennäköisyysjakauma, kun taas frekventistisessä ne ovat kiinteitä arvoja. Tämä muuttaa tapaa, jolla tilastollisen analyysin tuloksia voidaan tulkita. Frekventistinen 95 %:n luottamusväli voidaan tulkita niin, että jos tehtäisiin useita samanlaisia satunnaiskokeita, 95 %:ssa niistä luottamusväli peittäisi parametrin oikean arvon. Bayesiläinen 95 %:n posterioriväli taas voidaan tulkita niin, että sillä on 95 %:n todennäköisyys sisältää tuntematon parametrin arvo. Posterioriväli sisältää

myös priori-informaatiota parametrista, kun taas luottamusväli perustuu vain havaintoaineistoon ja havainnoille oletettuun malliin. Posteriorivälin tulkinta siis tuntuu hieman helpommalta ja intuitiivisemmalta kuin luottamusvälin tulkinta. Bayesiläistä analyysiä rajoittavat lähinnä käytännön ongelmat, jotka liittyvät monimutkaisten mallien rakentamiseen ja tietokoneiden laskentakapasiteetin rajallisuuteen.

Bayes-päättely perustuu Bayesin kaavaan. Olkoon $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ vektori, joka sisältää kaikki tilastollisen mallin parametrit ja $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ havaintoaineisto. Mallinnus aloitetaan havaintoaineiston ja parametrien yhteistodennäköisyystiheydestä $p(\boldsymbol{\theta}, \boldsymbol{y})$. Ehdollisen todennäköisyystiheyden kaavasta saadaan

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{4}$$

jossa $p(\boldsymbol{y}|\boldsymbol{\theta})$ on tilastollinen malli havainnoille tai uskottavuusfunktio $\boldsymbol{\theta}$:n funktiona ja $p(\boldsymbol{\theta})$ on parametrien reunatiheys, jota kutsutaan prioritiheydeksi. Nyt posterioritiheys $p(\boldsymbol{\theta}|\boldsymbol{y})$ saadaan Bayesin kaavasta

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{5}$$

jossa jatkuva-asteikollisten parametrien tapauksessa $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ on nk. normalisointivakio, joka varmistaa, että posterioritiheyden integraali yli parametriavaruuden on yksi. (Gelman ym. 2013, s. 6–7). Monesti (esimerkiksi MCMC-algoritmeissa) normalisointivakio ei ole kiinnostava eikä sitä tarvitse evaluoida. Tällöin on tyypillistä kirjoittaa kaava normalisoimattomalle posterioritiheydelle

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{6}$$

Monissa tilastollisissa malleissa on parametreja, jotka riippuvat toisistaan luonnollisella tavalla ongelman rakenteen kautta. Tällöin puhutaan hierarkkisista malleista. Esimerkiksi bayesiläisessä silotusongelmassa ajatellaan,

että sileä trendi $\boldsymbol{f}$ on tuntematon parametrivektori jota estimoidaan, ja sille määrätään silottava priori. Tämä priorikin saattaa riippua jostakin tuntemattomasta skalaariparametrista λ, joka voidaan myös estimoida. λ voi esimerkiksi kontrolloida sitä, kuinka sileä trendi on. Tällöin yhteisprioritiheys voidaan kirjoittaa muodossa

$$p(\boldsymbol{f}, \lambda) = p(\boldsymbol{f}|\lambda)p(\lambda). \tag{7}$$

Posterioritiheydeksi saadaan

$$p(\boldsymbol{f}, \lambda|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{f}, \lambda)p(\boldsymbol{f}, \lambda) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\lambda)p(\lambda). \tag{8}$$

Nyt uskottavuusfunktio $p(\boldsymbol{y}|\boldsymbol{f}, \lambda)$ riippuu vain parametrivektorista $\boldsymbol{f}$. λ on nimeltään hyperparametri eli parametri, josta $\boldsymbol{f}$:n priori riippuu. Hierarkkinen mallinnus auttaa hahmottamaan ongelmaa ja se helpottaa Bayes-mallien tietokonelaskentaa. (Gelman ym. 2013, s. 107–108)

2.2 Sukulaisuusmatriisi

Sukulaisuusmatriisi $\boldsymbol{A}$ on symmetrinen kovarianssimatriisi, jossa matriisin alkio A_{ij} kuvaa yksilöiden i ja j välistä sukulaisuuden astetta. Sukulaisuusmatriisin avulla havaintoaineistosta saadaan identifioitua geneettinen varianssikomponentti. Klassinen tapa konstruoida sukulaisuusmatriisi perustuu sukupuuhun. Sukupuuta ei ole kuitenkaan aina saatavilla. Uudempi tapa perustuu molekulaarisiin markkereihin, joilla selvitetään alleelin tyyppi eri kohdista kromosomia. Sukulaisuusmatriisi perustuu sukulaisuuskertoimeen (kinship coefficient), ja matriisin alkio A_{ij} saadaan kertomalla yksilöiden i ja j välisen sukulaisuuskertoimen arvo kahdella. Sukulaisuuskertoimen määrittelyssä taas tärkeitä termejä ovat syntyperältään identtiset (identical by descent, IBD) ja tilaltaan identtiset (identical by state, IBS) alleelit.

Jos tietyssä geenipaikassa, joka on määrätty paikka kromosomissa, yksilön

i alleeli on fyysisesti sama kuin yksilöllä j, ja ne ovat molemmat kopioita yksilöiden yhteiseltä esi-isältä peritystä alleelista, sanotaan, että alleelit ovat syntyperältään identtiset. Jos taas tietyssä geenipaikassa riippumatta alleelin alkuperästä yksilön i alleeli on samantyyppinen kuin yksilöllä j, sanotaan, että alleelit ovat tilaltaan identtiset.

Sukulaisuuskerroin pohjaa IBD-käsitteeseen. Se on todennäköisyys, jolla yksilöillä i ja j satunnaisesta geenipaikasta otetut alleelit ovat syntyperältään identtiset. Sukulaisuuskertoimen arvo kerrottuna kahdella antaa odotetun IBD-alleelien osuuden kahden yksilön välillä. Sukupuuhun perustuva sukulaisuusmatriisi sisältää kaikkien yksilöiden väliset odotetut IBD-alleelien osuudet. Jos sukupuuta ei ole saatavilla, voidaan odotetun IBD-alleelien osuuden sijasta laskea realisoitunut IBD-alleelien osuus kahden yksilön välillä, joka perustuu molekulaarisiin markkereihin. Näistä luvuista koostuvaa matriisia kutsutaan markkeripohjaiseksi sukulaisuusmatriisiksi. (Sillanpää, 2011)

Eräs tapa muodostaa markkeripohjainen sukulaisuusmatriisi on kuvattu seuraavassa. Tietyssä markkerissa genotyyppi voi olla joko AA, AB tai BB, jotka voidaan koodata numeroiksi järjestyksessä -1, 0 ja 1. Olkoon $\boldsymbol{X}$ $N \times M$ -matriisi, jossa N viittaa yksilöiden määrään ja M markkerien määrään ja $X_{ij} \in \{-1, 0, 1\}$ riippuen yksilön i genotyypistä markkerissa j. Lisäksi olkoon p_j alleelin A alleelifrekvenssi markkerissa j. Nyt voidaan muodostaa sukulaisuusmatriisi

$$\boldsymbol{A} = \frac{1}{c}\boldsymbol{Z}\boldsymbol{Z}^T, \tag{9}$$

jossa $Z_{ij} = X_{ij} + 1 - 2p_j$ ja $c = 2\sum_j p_j(1 - p_j)$. Tämä ei ole ainoa keino muodostaa markkeripohjaista sukulaisuusmatriisia. Tämän ja muita menetelmiä on listannut esimerkiksi VanRaden (2008).

2.3 Lineaarinen sekamalli

Jatkuva-asteikollista piirrettä voidaan mallintaa lineaarisen sekamallin avulla, joka on eräänlainen laajennus tavanomaiseen lineaariseen malliin. Sillä voidaan kätevästi ottaa huomioon vastevektorin alkioiden välinen riippuvuus, joka tässä tapauksessa johtuu populaation yksilöiden välisestä sukulaisuudesta. Oletetaan, että mitataan jotain piirrettä N:ltä yksilöltä jossain populaatiossa. Lineaarinen sekamalli määritellään (Kang ym. 2008)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{10}$$

jossa $\boldsymbol{y}$ on $N \times 1$ -vektori vasteen havaittuja arvoja, $\boldsymbol{X}$ on kiinteät vaikutukset sisältävä $N \times P$ -matriisi, $P \times 1$ -vektori $\boldsymbol{\beta}$ sisältää kiinteiden vaikutusten regressiokertoimet, $N \times Q$ -matriisi $\boldsymbol{Z}$ liittää satunnaisvaikutukset $\boldsymbol{u}$ ($Q \times 1$ -vektori) siihen tasoon, jolla satunnaisvaikutus halutaan laskea, ja $N \times 1$ -vektori $\boldsymbol{\epsilon}$ on yksilökohtainen virhetermien vektori. Heritabiliteettia estimoitaessa kiinteä vaikutus $\boldsymbol{\beta}$ redusoituu pelkkään viitetasoon, joka on vektrorin $\boldsymbol{y}$ odotusarvo, ja $\boldsymbol{X}$ on $N \times 1$ -vektori, jossa jokainen alkio on 1. Lisäksi satunnaisvaikutus lasketaan jokaiselle yksilölle, joten $\boldsymbol{Z} = \boldsymbol{I}$. Oletetaan, että $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \sigma_G^2\boldsymbol{A})$, jossa $\boldsymbol{A}$ on sukulaisuusmatriisi, ja $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_E^2\boldsymbol{I})$. Kun lisäksi $\boldsymbol{y}$ keskistetään, voidaan malli kirjoittaa muodossa

$$\tilde{\boldsymbol{y}} = \boldsymbol{u} + \boldsymbol{\epsilon}, \tag{11}$$

jossa $\tilde{\boldsymbol{y}}$ on keskistetty havaintovektori ja joka muistuttaa kaavaa (1). Yhtälöä (11) ja sen oletuksia vastaava uskottavuusfunktio parametreille σ_E^2 ja σ_G^2 on

$$p(\tilde{\boldsymbol{y}}|\sigma_G^2, \sigma_E^2) = \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{K}|}} \exp\left\{-\frac{1}{2}\tilde{\boldsymbol{y}}^T\boldsymbol{K}^{-1}\tilde{\boldsymbol{y}}\right\}, \tag{12}$$

jossa $\boldsymbol{K} = \sigma_G^2\boldsymbol{A} + \sigma_E^2\boldsymbol{I}$.

2.4 Sekamallin yleistys usealle ajanhetkelle

Kun estimoidaan heritabiliteettifunktiota, aineistossa täytyy olla mittauksia kiinnostavasta piirteestä populaatiossa usealla eri ajanhetkellä. Kaava (2) voidaan tällöin kirjoittaa muodossa

$$\sigma_P^2(t) = \sigma_G^2(t) + \sigma_E^2(t), \quad t = 1, ..., T, \tag{13}$$

jossa $\sigma_x^2(t)$ viittaa varianssikomponentin x arvoon ajanhetkellä t. Tavoitteena on luoda malli, joka määrittelee yhteisjakauman havaintovektorille $\tilde{\boldsymbol{y}}_T$, joka sisältää jokaisen ajanhetken oman keskistetyn havaintovektorin ladottuna peräkkäin. Määrittely onnistuu Kroneckerin matriisitulon avulla (liite A), ja tuloksena saadaan $\tilde{\boldsymbol{y}}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_T)$, jossa $\boldsymbol{K}_T = \text{diag}(\sigma_G^2(1), ..., \sigma_G^2(T)) \otimes \boldsymbol{A} + \text{diag}(\sigma_E^2(1), ..., \sigma_E^2(T)) \otimes \boldsymbol{I}$. Aikapistekohtaiset varianssikomponentit on siis kerätty diagonaalimatriiseihin ja $\boldsymbol{A}$ on jälleen sukulaisuusmatriisi. Symboli $\otimes$ viittaa Kroneckerin tuloon. Uskottavuusfunktio vektoriarvoisille parametreille $\boldsymbol{\sigma}_E^2$ ja $\boldsymbol{\sigma}_G^2$, jotka sisältävät jokaisen aikapisteen varianssikomponentit, on

$$p(\tilde{\boldsymbol{y}}_T | \boldsymbol{\sigma}_G^2, \boldsymbol{\sigma}_E^2) = \frac{1}{\sqrt{(2\pi)^{NT} |\boldsymbol{K}_T|}} \exp\left\{ -\frac{1}{2} \tilde{\boldsymbol{y}}_T^T \boldsymbol{K}_T^{-1} \tilde{\boldsymbol{y}}_T \right\}. \tag{14}$$

Jakauman kovarianssimatriisin rakennetta voi olla vaikea hahmottaa. Se on seuraavanlainen:

$$\boldsymbol{K}_T = \begin{bmatrix} \sigma_G^2(1)\boldsymbol{A} + \sigma_E^2(1)\boldsymbol{I} & 0 & \dots & 0 \\ 0 & \sigma_G^2(2)\boldsymbol{A} + \sigma_E^2(2)\boldsymbol{I} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_G^2(T)\boldsymbol{A} + \sigma_E^2(T)\boldsymbol{I} \end{bmatrix}.$$

Rakenne on siis lohkodiagonaalinen ja diagonaalilla toistuvat aikapistekohtaiset kovarianssimatriisit. Lohkodiagonaalinen rakenne nopeuttaa estimointia, koska matriisin determinantin laskeminen ja matriisin kääntäminen helpottuu. Determinantti on diagonaalilla olevien alimatriisien determinanttien tulo, ja käänteismatriisia varten täytyy kääntää vain alimatriisit erikseen.

2.5 Silotus ja priorit

Varianssifunktioiden priorit valitaan siis sillä perusteella, että funktioiden ajatellaan olevan sileitä. Tästä seuraa se, että heritabiliteettifunktiokin on sileä. Silottava priori on bayesiläinen tapa tehdä silotusta. Silotus tarkoittaa sitä, että havaintoaineistoon sovitetaan funktio, joka kiteyttää aineiston tärkeät piirteet ja poistaa epäsäännöllisen kohinan. Silotusta käytetään esimerkiksi kuvankäsittelyssä ja aikasarjojen avulla ennustamisessa. Silotukseen on kehitetty monia eri menetelmiä. Klassisia tapoja ovat eksponentiaalinen silotus ja liukuva keskiarvo. Kehittyneempiä menetelmiä ovat frekventistisellä puolella rosoisuuden sakottaminen ja bayesiläisellä puolella vastaavat silottavat priorit (Fahrmeir ja Kneib, 2011, s. 18).

Fahrmeir ja Kneib (2011, s. 19) määrittelee klassisen silotusongelman seuraavasti: Olkoon $\boldsymbol{y} = (y_1, ..., y_T)^T$ vektori tasavälisiä havaintoja. Havainnoille oletetaan malli

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{15}$$

jossa $\boldsymbol{f} = (f_1, ..., f_T)^T$ on sileä trendi ja $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_T)^T$ on joukko riippumattomia ja samoin jakautuneita virhetermejä niin, että $\mathbb{E}(\epsilon_t) = 0$. Trendin estimointiin on ehdotettu sakotetun neliösumman (PLS; Penalized Least Squares) minimointia:

$$\underset{\text{[mallin istuvuus]}}{\text{PLS}(\boldsymbol{f}) = \sum_{t=1}^{T}(y_t - f_t)^2} + \underset{\text{[sakkotermi]}}{\lambda \sum_{t=3}^{T}(f_t - 2f_{t-1} + f_{t-2})^2}. \tag{16}$$

Silotusparametri $\lambda \geq 0$ kontrolloi tasapainotusta harhan ja varianssin välillä ja se pitää säätää sopivaksi jokaiselle aineistolle erikseen. (Fahrmeir ja Kneib, 2011, s. 19). Parametrin vaikutusta on havainnollistettu kuvassa 1.

Bayesiläisessä silotuksessa oletetaan, että tuntematon trendi on satunnaisvektori $\boldsymbol{f}$, jolla on prioritiheys $p(\boldsymbol{f})$. Yleisin virhetermeille oletettu jakauma on normaalijakauma, jolloin yhtälöä (15) vastaava malli on

$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I}). \tag{17}$$

Kun σ^2 oletetaan tunnetuksi, voidaan normalisointivakio sivuuttaa ja saadaan

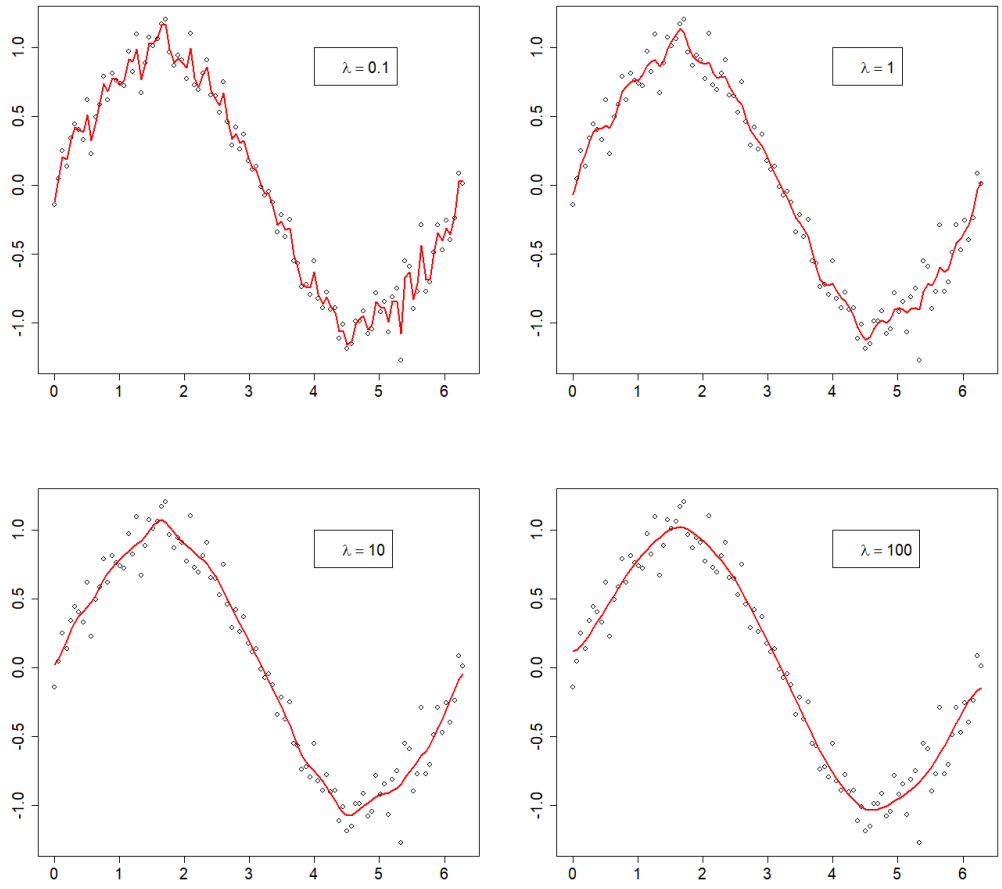
$$p(\boldsymbol{y}|\boldsymbol{f}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{f})^T(\boldsymbol{y}-\boldsymbol{f})\right\}. \tag{18}$$

Valittu priori on yleensä gaussinen, $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{P}^{-1})$, jossa täsmällisyysmatriisina on jokin sakkomatriisi $\boldsymbol{P}$, jonka muoto riippuu siitä, millainen silotus halutaan tehdä. Tällöin prioritiheys τ^2:n ollessa tunnettu on

$$p(\boldsymbol{f}) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{f}^T\boldsymbol{P}\boldsymbol{f}\right\}. \tag{19}$$

Kun sekä havaintomalli että priori ovat gaussiset, voidaan trendin posteriorijakauma johtaa analyyttisesti. Kun merkitään $\lambda = \sigma^2/\tau^2$, se on

$$\boldsymbol{f}|\boldsymbol{y} \sim \mathcal{N}((\boldsymbol{I}+\lambda\boldsymbol{P})^{-1}\boldsymbol{y}, \sigma^2(\boldsymbol{I}+\lambda\boldsymbol{P})^{-1}). \tag{20}$$



Kuva 1: Silotusparametrin vaikutus trendiin.

Sakkomatriisin $\boldsymbol{P}$ rooli on ikään kuin vetää vierekkäisiä havaintoja toisiaan kohti tietyn funktiomuodon perusteella. Silotusparametrin λ arvo kontrolloi silotuksen astetta (l. trendin sileyttä), joka tarkoittaa sitä kuinka vahvasti havaintoja halutaan vetää toisiaan kohti edellämainitun $\boldsymbol{P}$:n sisällä. (Fahrmeir ja Kneib, 2011, s. 23–25)

Tässä tutkielmassa silotus tehdään priorin avulla, mutta sakkomatriisin

$\boldsymbol{P}$ sijaan muodostetaan kovarianssimatriisi $\boldsymbol{C}$ tietyn kovarianssifunktion mukaan. Kovarianssifunktio määrää jokaisen aikapisteparin välille kovarianssin niiden etäisyyden perusteella. Toisiaan lähellä olevilla aikapisteillä on suurempi kovarianssi kuin toisistaan kaukana olevilla aikapisteillä. Kovarianssifunktion muoto ja parametrit määräävät muun muassa sen, kuinka sileä estimoitu funktio on.

Merkitään $d = |t_i - t_j|$ kahden ajanhetken välillä kuluvaa aikaa. Neliöeksponentiaalinen kovarianssifunktio määritellään

$$k_{\mathrm{SE}}(d) = \tau^2 \exp\left\{-\frac{d^2}{2l^2}\right\}, \tag{21}$$

jossa $l > 0$ on pituusskaalaparametri ja $\tau^2 > 0$ on voimakkuusparametri. Pituusskaalaparametri vaikuttaa estimoidun funktion sileyteen ja voimakkuusparametri sen vaihteluun. Priori tällä kovarianssifunktiolla tuottaa todella sileitä funktioestimaatteja. Se on yksi käytetyimmistä kovarianssifunktioista. Eksponentiaalinen kovarianssifunktio on

$$k_{\mathrm{E}}(d) = \tau^2 \exp\left\{-\frac{d}{l}\right\}, \tag{22}$$

jossa parametrit ovat samat kuin neliöeksponentiaalisessa kovarianssifunktiossa. Priorit eksponentiaalisella kovarianssifunktiolla tuottavat huomattavasti rosoisempia funktioestimaatteja kuin neliöeksponentiaalisella kovarianssifunktiolla. Nämä kaksi kovarianssifunktiota ovat itse asiassa erityistapauksia yleisemmästä Matérn-luokasta, joka määritellään

$$k_{\text{Matérn}}(d) = \tau^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}d}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}d}{l}\right), \tag{23}$$

jossa parametri $\nu > 0$ määrää funktion sileysominaisuuksia ja K_ν on toisen lajin Besselin funktio. Eksponentiaalinen kovarianssifunktio saadaan, kun asetetaan $\nu = \frac{1}{2}$ ja neliöeksponentiaalinen kovarianssifunktio raja-arvona, kun $\nu \to \infty$. (Rasmussen ja Williams, 2006, s. 83–86)

2.6 Kaksivaiheinen menetelmä

Kaksivaiheinen menetelmä hyödyntää pelkästään alaluvussa 2.3 kaavassa (11) määriteltyä mallia yhden aikapisteen vastevektorille. Siinä varianssikomponentit estimoidaan jokaisessa aikapisteessä erikseen maksimoimalla rajoitettu uskottavuusfunktio, josta kerrotaan alaluvussa 3.1. Menetelmä ei ole bayesiläinen, mutta se osoittautui yhden ajanhetken tapauksessa nopeammaksi ja täsmällisemmäksi tavaksi estimoida varianssikomponentit. Tämän lisäksi komponenteille estimoidaan luottamusvälit bootstrap-menetelmällä, josta kerrotaan alaluvussa 3.2. Tällä tavalla tuotettuja estimaatteja voidaan pitää aikasarjoina, jotka voivat sisältää kohinaa, ja ne silotetaan priorien avulla. Silotusmallina käytetään kaavassa (15) määriteltyä yhtälöä. Aikasarjat voidaan ilmaista muodossa

$$\log \boldsymbol{y} = \log \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{24}$$

jossa $\boldsymbol{y}$ on varianssikomponentin tai sen luottamusvälirajan aikasarja, $\boldsymbol{f}$ on aikasarjaa kuvaava sileä funktio ja $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ on virhetermi. Aikasarjat logaritmoidaan, jotta virhetermien normaalisuusoletus olisi järkevä. Logaritmoidulle funktiolle $\boldsymbol{f}$ asetetaan priori $\log \boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_l)$, jossa matriisin $\boldsymbol{C}_l$ alkio $C_{lij} = k_{\text{Matérn}}(|t_i - t_j|)$. Alaindeksi l tarkoittaa sitä, että matriisin muoto riippuu pituusskaalasta l, joka myös estimoidaan. Sille ja virhevarianssille σ^2 asetetaan laakeat priorit, jotka sisältävät vain vähän informaatiota siitä, missä parametrien todennäköisimmät arvot ovat. Nämäkin parametrit logaritmoidaan, koska ne voivat saada vain positiivisia arvoja. Kovarianssifunktion muut parametrit ν ja τ^2 täytyy kiinnittää identifioituvuusongelmien takia (Monterrubio-Gómez ym. 2018). Näillä parametreillä, kun $\log \boldsymbol{y}$ keskis-

tetään, mallin voi kirjoittaa hierarkkisessa muodossa

$$\begin{aligned} \log \boldsymbol{y} | \boldsymbol{f}, \sigma^2 &\sim \mathcal{N}(\log \boldsymbol{f}, \sigma^2 \boldsymbol{I}), \\ \log \boldsymbol{f} | l &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_l), \\ \log l &\sim u_{\log l}(\boldsymbol{\theta}_{\log l}), \\ \log \sigma^2 &\sim u_{\log \sigma^2}(\boldsymbol{\theta}_{\log \sigma^2}), \end{aligned} \tag{25}$$

jossa u viittaa johonkin laakeaan priorijakaumaan ja $\boldsymbol{\theta}$ sen parametreihin. Mallin parametrien normalisoimaton posterioritiheys on

$$\begin{aligned} &p(\log \boldsymbol{f}, \log l, \log \sigma^2 | \log \boldsymbol{y}) \propto p_{\mathcal{N}}(\log \boldsymbol{y} | \log \boldsymbol{f}, \sigma^2 \boldsymbol{I}) p_{\mathcal{N}}(\log \boldsymbol{f} | \boldsymbol{0}, \boldsymbol{C}_l) \\ &p_{u;\log l}(\log l | \boldsymbol{\theta}_{\log l}) p_{u;\log \sigma^2}(\log \sigma^2 | \boldsymbol{\theta}_{\log \sigma^2}), \end{aligned} \tag{26}$$

jossa $p_{\mathcal{N}}$ viittaa multinormaalijakauman tiheysfunktioon ja p_u laakean priorijakauman tiheysfunktioon. Koska malli sisältää tuntemattomia hyperparametrejä, funktion $\boldsymbol{f}$ posteriorijakaumaa ei voida johtaa analyyttisesti, kuten kaavassa (20). Sen sijaan parametrien estimointiin täytyy käyttää numeerisia approksimointimenetelmiä, joista kerrotaan lisää luvussa 3.

2.7 Yksivaiheinen menetelmä

Yksivaiheinen menetelmä heritabiliteettifunktion estimointiin perustuu jokaisen aikapisteen havaintovektorin sisältävän vektorin $\tilde{\boldsymbol{y}}_T$ yhteisjakaumaan, joka on johdettu alaluvussa 2.4. Logaritmoiduille varianssikomponenttivektoreille $\boldsymbol{\sigma}_E^2$ ja $\boldsymbol{\sigma}_G^2$ asetetaan jälleen multinormaalipriorit: $\log \boldsymbol{\sigma}_E^2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_E)$ ja $\log \boldsymbol{\sigma}_G^2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_G)$. Malli voidaan esittää seuraavasti:

$$\begin{aligned} \tilde{\boldsymbol{y}}_T &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_T), \\ \log \boldsymbol{\sigma}_G^2 &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_G), \\ \log \boldsymbol{\sigma}_E^2 &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_E). \end{aligned} \tag{27}$$

Käytännössä havaittiin, että kovarianssifunktion pituusskaala ei identifioidu yhtä hyvin kuin kaksivaiheisella menetelmällä. Niinpä sen arvo kiinnitettiin

kaksivaiheisella menetelmällä estimoituun posteriorikeskiarvoon. Niinpä mallin parametreinä ovat ainoastaan varianssikomponenttivektorit. Mallia vastaava normalisoimaton posterioritiheys on

$$p(\boldsymbol{\sigma}_G^2, \boldsymbol{\sigma}_E^2 | \tilde{\boldsymbol{y}}_T) \propto p_{\mathcal{N}}(\tilde{\boldsymbol{y}}_T | \mathbf{0}, \boldsymbol{K}_T) p_{\mathcal{N}}(\log \boldsymbol{\sigma}_G^2 | \mathbf{0}, \boldsymbol{C}_G) p_{\mathcal{N}}(\log \boldsymbol{\sigma}_E^2 | \mathbf{0}, \boldsymbol{C}_E). \quad (28)$$

Tämänkin mallin parametrit estimoidaan numeerisilla menetelmillä, joista kerrotaan seuraavassa luvussa.

3 Parametrien estimointimenetelmät

Bayesiläisessä analyysissä tilastollinen päättely joudutaan usein perustamaan numeerisiin approksimaatioihin, koska posteriorijakauman tiheysfunktiolla ei välttämättä ole tunnettua parametrista muotoa, josta parametrien arvot voisi päätellä. Tässä luvussa on lyhyt katsaus tutkielmassa hyödynnettäviin numeerisiin menetelmiin. Lisäksi esitellään kaksivaiheisessa menetelmässä hyödynnettävät klassiset tilastollisen päättelyn menetelmät rajoitettu uskottavuusfunktio ja bootstrap.

3.1 Rajoitettu uskottavuusfunktio (REML)

Skalaarimuotoiset varianssikomponentit σ_E^2 ja σ_G^2 voidaan estimoida maksimoimalla lineaarista sekamallia (alaluku 2.3) vastaava rajoitettu uskottavuusfunktio. Se on määritelmältään täysi uskottavuusfunktio, josta on analyyttisesti integroitu satunnaistekijät pois ja joka on lineaarisen muunnoksen jälkeen riippumaton ja kohtisuorassa kiinteisiin tekijöihin nähden. REML-estimoinnissa vältetään varianssin aliarviointi, koska siinä otetaan huomioon kiinteiden vaikutusten estimoinnista aiheutuva vapausasteiden menetys.

Yhtälöä (10) ja sen oletuksia vastaava log-uskottavuusfunktio ja rajoitettu log-uskottavuusfunktio ovat (Kang ym. 2008)

$$l_F(\boldsymbol{y};\boldsymbol{\beta},\sigma_G^2,\delta) = \frac{1}{2}\left[-N\log(2\pi\sigma_G^2) - \log|\boldsymbol{H}| - \frac{1}{\sigma_G^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] \tag{29}$$

ja

$$l_R(\boldsymbol{y};\sigma_G^2,\delta) = l_F(\boldsymbol{y};\hat{\boldsymbol{\beta}},\sigma_G^2,\delta) + \frac{1}{2}\left[P\log(2\pi\sigma_G^2) + \log|\boldsymbol{X}^T\boldsymbol{X}| - \log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|\right], \tag{30}$$

joissa $\boldsymbol{H} = \boldsymbol{A} + \delta\boldsymbol{I}$, $\delta = \frac{\sigma_E^2}{\sigma_G^2}$ ja $\boldsymbol{A}$ on sukulaisuusmatriisi. Rajoitetun uskottavuusfunktion maksimoimiseen tarvitaan numeerisia optimointimenetelmiä,

joihin ei tässä työssä syvennytä tarkemmin, mutta niistä kertoo enemmän Kang ym. (2008). Tässä tutkielmassa varianssikomponenttien estimointiin yhdessä aikapisteessä käytettiin R-ohjelmiston paketin `rrBLUP` (Endelman, 2011) funktiota `mixed.solve`.

3.2 Bootstrap-menetelmä

Usein jonkin kiinnostavan tunnusluvun (asymptoottista) otantajakaumaa ei ole mahdollista johtaa analyyttisesti. Tällöin suureelle voi olla vaikea laskea esimerkiksi luottamusvälejä. Tietokoneiden tehostuminen on kuitenkin luonut mahdollisuuden uudelleenkäyttää otosta niin, että kiinnostaville suureille voidaan simuloida empiirinen otantajakauma ja tarkastella sen käyttäytymistä. Tässä työssä bootstrap-menetelmää käytetään aikapistekohtaisten varianssikomponenttien luottamusvälien estimointiin.

Olkoon $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ havaintoaineisto, jossa kukin alkio y_i on realisaatio samasta parametrisesta jakaumasta, jonka kertymäfunktio on $F(y; \theta)$. Oletetaan, että suurimman uskottavuuden estimaatti $\hat{\theta}$ on kohtuullisen helposti laskettavissa, ja että jakaumasta on mahdollista simuloida satunnaislukuja millä tahansa θ:n arvolla. Parametrisessa bootstrapissa (Efron ja Hastie, 2016, s. 169) simuloidaan otos $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, ..., Y_n^*)$ jakaumasta, jonka kertymäfunktio on $F(y; \hat{\theta})$. Olkoon nyt $T = t(Y)$ jokin havainnoista laskettava tunnusluku. Se voi olla vaikkapa suurimman uskottavuuden estimaatti $\hat{\theta}$ tai joku sen muunnos. Tunnusluvun T otantajakaumaa simuloidaan toistamalla seuraavaa B kertaa:

1. Poimitaan otos $\boldsymbol{Y}^{*(b)} = (Y_1^{*(b)}, Y_2^{*(b)}, ..., Y_n^{*(b)})$ jakaumasta, jonka kertymäfunktio on $F(y; \hat{\theta})$
2. Lasketaan $T^{*(b)} = t(\boldsymbol{Y}^{*(b)})$

Parametrittomassa bootstrapissa (Efron ja Hastie, 2016, s. 159-160) otosten simulointi toteutetaan eri tavalla. Siinä otetaan alkuperäisestä otoksesta takaisinpanolla n havaintoa ja lasketaan näistä otoksista tunnusluvun T arvo. Nyt $(T^{*(1)}, T^{*(2)}, ..., T^{*(B)})$ on otos T:n otantajakaumasta. Otoksesta voidaan laskea esimerkiksi otoskeskiarvo ja sen keskivirhe

$$\overline{T}^* = \frac{1}{B}\sum_{b=1}^{B} T^{*(b)}, \tag{31}$$

$$SE(\overline{T}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(T^{*(b)} - \overline{T}^*)}. \tag{32}$$

T:n $1-\gamma$ prosentin luottamusväli saadaan bootstrap-otoksen empiirisistä fraktiileista. Alarajaksi valitaan $100\gamma/2$ %:n fraktiili $T^*_{\gamma/2}$ ja ylärajaksi $100(1-\gamma/2)$ %:n fraktiili $T^*_{1-\gamma/2}$. (Efron ja Hastie, 2016, s. 185-187)

3.3 Monte Carlo -integrointi

Yleinen ongelma bayesiläisessa analyysissä on posteriorijakauman parametrien odotusarvojen laskeminen. Jatkuvilla satunnaismuuttujilla odotusarvon lauseke on integraalimuotoinen. Integraalien analyyttinen laskeminen on hankalaa jo muutamissa ulottuvuuksissa, joten niiden laskemisessa täytyy turvautua numeerisiin menetelmiin. Olkoon X satunnaismuuttuja, jonka tiheysfunktio on p_X. Muunnoksen $g(X)$ odotusarvo määritellään

$$\mathbb{E}(g(X)) = \int_{\mathcal{X}} g(x)p_X(x)dx, \tag{33}$$

jossa $\mathcal{X}$ on tiheysfunktion p_X kantaja. Monte Carlo -integroinnin periaate on tuottaa otos $(X_1, X_2, ..., X_n)$ X:n jakaumasta ja approksimoida odotusarvoa

$$\mathbb{E}(g(X)) \approx \overline{g(X)}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i). \tag{34}$$

Suurten lukujen lain avulla voidaan osoittaa, että otoskeskiarvo suppenee vähintään stokastisesti kohti odotusarvoa, kun otoskoko kasvaa rajatta. Myös estimaattorin $\overline{g(X)_n}$ varianssi voidaan estimoida otoksesta kaavalla

$$\text{var}(\overline{g(X)_n}) \approx \frac{1}{n-1}\sum_{i=1}^{n}(g(X_i) - \overline{g(X)}_n)^2. \tag{35}$$

(Robert ja Casella, 2009, s. 65)

3.4 Markovin ketju Monte Carlo -menetelmät (MCMC)

MCMC-menetelmät perustuvat siihen, että ne tuottavat otoksen parametrien posteriorijakaumasta. Tämä otos on matemaattisessa mielessä Markovin ketju, jossa alkiot riippuvat toisistaan. Vaikka riippuva otos on epäinformatiivisempi kuin riippumaton otos, MCMC-menetelmistä on niin paljon hyötyjä, että ne ovat erittäin suosittuja bayesiläisessä analyysissä. Ensinnäkään posteriorijakauman muodolle ei ole juuri rajoituksia. Lisäksi hankalatkin moniulotteiset ongelmat on helppo pilkkoa pienempiin osiin.

Markovin ketju on sellainen jono satunnaismuuttujia

$$X^{(0)}, X^{(1)}, ..., X^{(t)}, ..., \tag{36}$$

että $X^{(t)}$:n edellisistä satunnaismuuttujista ehdollinen todennäköisyysjakauma riippuu vain edellisestä satunnaismuuttujasta $X^{(t-1)}$. Ehdollista todennäköisyysjakaumaa

$$X^{(t+1)}|X^{(0)}, X^{(1)}, ..., X^{(t)} \sim K(X^{(t)}, X^{(t+1)}) \tag{37}$$

kutsutaan ketjun siirtymäksi. Esimerkiksi satunnaiskulkuprosessi

$$X^{(t+1)} = X^{(t)} + \epsilon_t, \quad \epsilon_t \sim N(0,1), \tag{38}$$

jossa virhetermit ϵ_t ovat toisistaan riippumattomia, on Markovin ketju, jonka siirtymä on $N(X^{(t)}, 1)$. Yleisesti ottaen MCMC-menetelmissä toteutuu stationäärisyysehto, eli jonon peräkkäiset alkiot ovat realisaatioita samasta jakaumasta. Tämä tarkoittaa sitä, että olipa aloituspiste $X^{(0)}$ mikä tahansa, jonon $\{X^{(t)}\}$ jakauma suppenee toivottua kohdejakaumaa, kun t kasvaa rajatta. Siispä satunnaismuuttujan X odotusarvon estimaattorina voidaan käyttää otoskeskiarvoa

$$\mathbb{E}(X) \approx \frac{1}{T} \sum_{t=1}^{T} X^{(t)}. \tag{39}$$

(Robert ja Casella, 2009, s. 168–169)

3.5 Metropolisin–Hastingsin algoritmi

Metropolisin–Hastingsin algoritmi on ehkäpä käytetyin MCMC-menetelmä. Siinä simuloidaan otos posteriorijakaumasta, jolla on tiheysfunktio p, simuloimalla satunnaislukuja jostain ehdotusjakaumasta, jolla on tiheysfunktio q, ja hyväksymällä ehdotusluku tietyllä todennäköisyydellä. Todennäköisyys riippuu tiheysfunktioiden arvoista. Oletetaan, että ollaan simuloitu jo ketju $(X^{(0)}, X^{(1)}, ..., X^{(t)})$, ja ehdotusarvo on X'. Hyväksymistodennäköisyys on

$$\alpha_t = \min \left\{ \frac{p(X')}{p(X^{(t)})} \frac{q(X^{(t)}|X')}{q(X'|X^{(t)})}, 1 \right\}. \tag{40}$$

Todennäköisyydellä α_t asetetaan ketjun seuraavaksi alkioksi $X^{(t+1)} = X'$, ja todennäköisyydellä $1 - \alpha_t$ asetetaan $X^{(t+1)} = X^{(t)}$. Niinpä ketjussa voi olla useita samoja lukuja peräkkäin. On hyvä huomata, että jos ehdotusjakauma on symmetrinen edellisen hyväksytyn alkion $X^{(t)}$ suhteen, $q(X^{(t)}|X') = q(X'|X^{(t)})$, eli hyväksymistodennäköisyys on riippumaton ehdotusjakaumasta. Hyväksymisaste

$$\overline{\alpha} = \frac{1}{T} \sum_{t=1}^{T} \alpha_t \tag{41}$$

kuvaa sitä, kuinka pitkiä jaksoja samaa lukua ketjussa keskimäärin on. Hyväksymisastetta voi kontrolloida esimerkiksi ehdotusjakauman varianssia muuttamalla. (Robert ja Casella, 2009, s. 170–171)

Vaikka teoriassa ehdotusjakauman valinnalla ei ole juurikaan merkitystä, käytännössä se vaikuttaa huomattavasti algoritmin tehokkuuteen ja suppenemisnopeuteen. Eräs luonnollinen ja paljon käytetty valinta ehdotusjakaumaksi on nk. satunnaiskulkuehdotusjakauma, joka on symmetrinen. Siinä ehdotusarvon X' lauseke on muotoa

$$X' = X^{(t)} + \epsilon_t, \tag{42}$$

jossa ϵ_t on nollakeskinen satunnaisluku esimerkiksi tasa- tai normaalijakaumasta. Satunnaiskulkuehdotusjakauma yksinkertaistaa algoritmia siinäkin mielessä, että ehdotusjakaumaan liittyvä osa hyväksymistodennäköisyyden lausekkeessa (40) supistuu pois. (Robert ja Casella, 2009, s. 182)

3.6 Gibbsin otanta

Gibbsin otanta on hieman erityylinen MCMC-menetelmä kuin Metropolisin–Hastingsin algoritmi. Sen vahvuus piilee siinä, että sen avulla voidaan pilkkoa hankala moniulotteinen ongelma pienempiin, helpommin ratkaistaviin palasiin. Oletetaan, että satunnaismuuttuja $\boldsymbol{X}$ koostuu useista komponenteista $(X_1, X_2, ..., X_p)$. Yksittäinen komponentti voi olla yksi- tai moniulotteinen. Gibbsin otanta perustuu siihen, että jokainen $\boldsymbol{X}$:n komponentti ehdollistetaan erikseen kaikille muille komponenteille: $X_i|X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_p$. Oletetaan, että kaikkien näiden ehdollisten satunnaismuuttujien jakaumista osataan simuloida satunnaislukuja. Näitä jakaumia kutsutaan täysehdollisiksi jakaumiksi. Tämän jälkeen Gibbsin otanta -algoritmin rakentaminen on helppoa. Simuloidaan vuorotellen uusi arvo jokaiselle komponentille X_i omasta

täysehdollisesta jakaumastaan niin, että kaikkien muiden komponenttien arvo on kiinnitetty niiden edelliseen simuloituun arvoon. Kun algoritmia on suoritettu tarpeeksi monta kierrosta, simuloidut arvot approksimoivat $\boldsymbol{X}$:n yhteisjakaumaa. (Robert ja Casella, 2009, s. 199–206)

Joskus täysehdollisten jakaumien lausekkeita voi olla vaikeaa tai mahdotonta johtaa. Tällöin voidaan yhdistää Gibbsin otanta- ja Metropolisin–Hastingsin -algoritmit. Jos jostain täysehdollisesta jakaumasta ei osata suoraan simuloida satunnaislukuja, komponentti X_i voidaan simuloida jostakin ehdotusjakaumasta ja laskea sille hyväksymistodennäköisyys kuten Metropolisin–Hastingsin algoritmissa. Jos hyväkysymistodennäköisyyden lausekkeessa esiintyy jotain muita vektorin $\boldsymbol{X}$ komponentteja, niiden arvot kiinnitetään viimeisimpään simuloituun arvoon.

3.7 Elliptinen viipaleotanta (Elliptical slice sampling)

Elliptinen viipaleotanta on melko uusi MCMC-menetelmä, jonka on huomattu toimivan hyvin bayesiläisissä piilomuuttujamalleissa, joissa käytetään multinormaaliprioria. Olkoon $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ vektori piiloparametreja ja $\boldsymbol{y}$ havaintoaineisto. Vektorille $\boldsymbol{X}$ asetetaan gaussinen priori, jonka tiheysfunktio on

$$p(\boldsymbol{X}|\boldsymbol{0}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \right\}. \tag{43}$$

Normalisoimaton posterioritiheys on

$$p(\boldsymbol{X}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X}) p(\boldsymbol{X}|\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{44}$$

jossa on $p(\boldsymbol{y}|\boldsymbol{X})$ on havainnoille oletetun mallin uskottavuusfunktio ja jossa oletetaan, että $\boldsymbol{\Sigma}$ on tunnettu. Menetelmä lähtee ajatuksesta, että Metropo-

lisin–Hastingsin ehdotusarvona käytetään vektoria

$$\boldsymbol{X}' = \sqrt{1-\rho^2}\boldsymbol{X}^{(t)} + \rho\boldsymbol{v}, \quad \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{45}$$

jossa $\rho \in [-1, 1]$ on askelkokoparametri. Kun $\rho = 1$, ehdotus on itse asiassa otos priorijakaumasta, kun taas ρ:n lähestyessä nollaa se on lähellä edellisellä iteraatiolla hyväksyttyä vektoria. Tässä ehdotusarvossa ongelmana on kuitenkin se, että askelkoko täytyy säätää aina uudelleen sopivaksi eri tilanteisiin. Olisi toivottavaa, että askelkoon valinta olisi jotenkin automatisoitu. (Murray ym. 2010)

Kun askelkoon annetaan vaihdella välillä $[-1, 1]$, ehdotusarvon lauseke (45) määrää puoliellipsin. Luonnollisempi parametrointi ehdotusarvolle on

$$\boldsymbol{X}' = \boldsymbol{X}^{(t)} \cos\rho + \boldsymbol{v} \sin\rho, \tag{46}$$

joka määrää kokonaisen ellipsin. Askelkokoa siirretään automaattisesti jokaisella iteraatiolla kohti nollaa, kunnes tarpeeksi hyvä ehdotusarvo löytyy. (Murray ym. 2010)

Elliptisen viipaleotannan algoritmi on seuraavanlainen iteraatiolla t:

1. simuloi $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$
2. laske $\kappa = \log p(\boldsymbol{y}|\boldsymbol{X}^{(t)}) + \log u, \quad u \sim \mathcal{U}(0, 1)$
3. simuloi $\rho \sim \mathcal{U}(0, 2\pi)$
4. aseta $(\rho_{\min}, \rho_{\max}) = (\rho - 2\pi, \rho)$
5. laske $\boldsymbol{X}' = \boldsymbol{X}^{(t)} \cos\rho + \boldsymbol{v} \sin\rho$
6. jos $\log p(\boldsymbol{y}|\boldsymbol{X}') > \kappa$, aseta $\boldsymbol{X}^{(t+1)} = \boldsymbol{X}'$

7. muutoin:

8. jos $\rho < 0$, aseta $\rho_{\min} = \rho$, muutoin aseta $\rho_{\max} = \rho$

9. simuloi $\rho \sim \mathcal{U}(\rho_{\min}, \rho_{\max})$

10. palaa kohtaan 5

Algoritmin yksi iteraatio tuottaa yhden realisaation parametrivektorin $\boldsymbol{X}$ posteriorijakaumasta. (Murray ym. 2010). Vektorin $\boldsymbol{v}$ tuottaminen onnistuu helposti. Jos $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}}$, jossa $\boldsymbol{L}$ on Choleskyn hajotelman yläkolmiomatriisi, niin $\boldsymbol{v} = \boldsymbol{L}\boldsymbol{d}$, jossa $\boldsymbol{d} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

3.8 MCMC-algoritmien suppenemisen tarkastelu

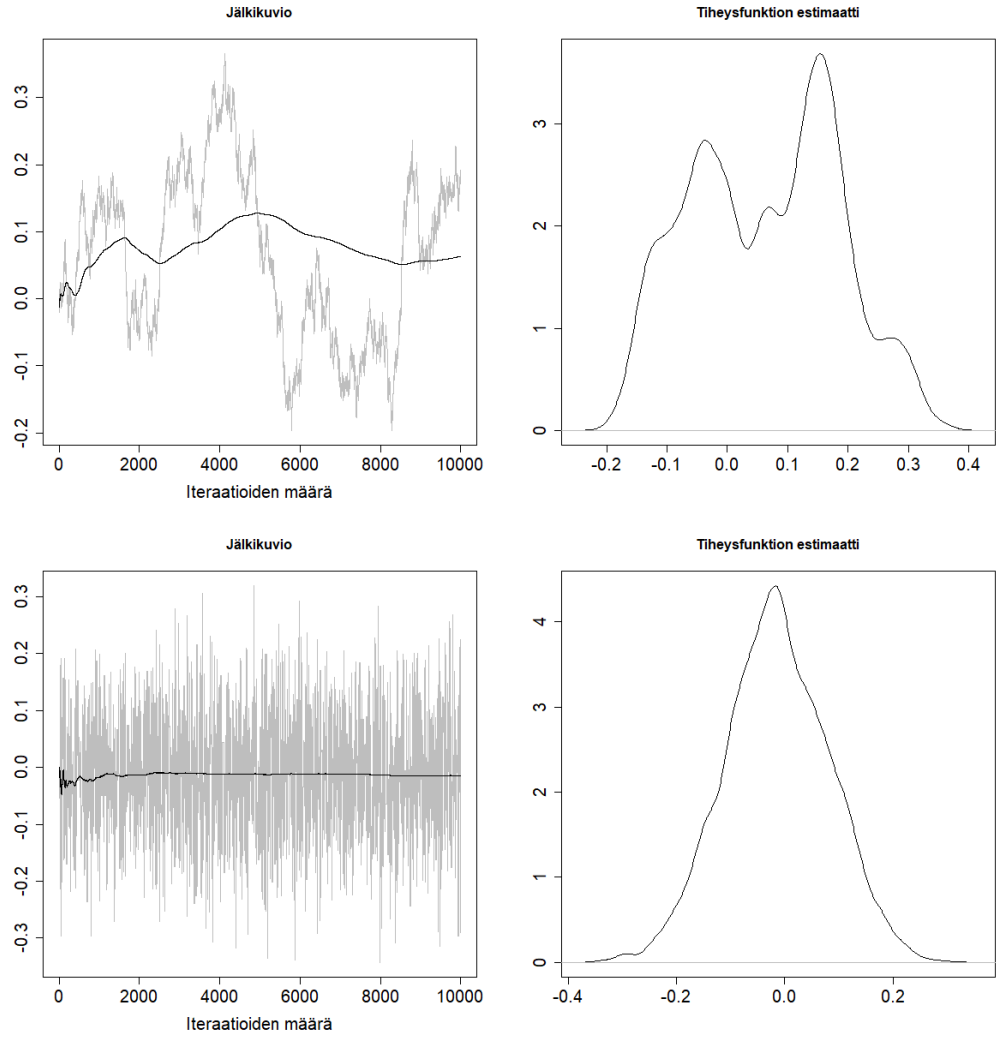
Koska MCMC-menetelmät perustuvat satunnaislukujen simulointiin, on perusteltua kysyä kuinka paljon satunnaislukuja on tarpeeksi. Liian pieni määrä ei edusta posteriorijakaumaa tarpeeksi hyvin ja tietyn pisteen jälkeen simuloinnin jatkaminen ei enää merkittävästi muuta estimaatteja.

Tärkein kriteeri MCMC-algoritmin suppenemiselle on se, että tuotetun ketjun jakauman tulisi olla sama kuin posteriorijakauma. On kuitenkin vaikeaa todentaa että näin on, koska posteriorijakauman muotoa ei välttämättä edes tunneta. On kuitenkin mahdollista tutkia, kuinka riippumaton tai riippuva ketju on aloituspisteestä eri ketjujen realisaatioita vertaamalla. Eräs yksinkertainen tapa on verrata eri ketjujen keskiarvoja. Niiden tulisi luonnollisesti olla lähellä toisiaan. (Robert ja Casella, 2009, s. 238–240)

Suosittu tapa tutkia suppenemista on piirtää kuvia parametrien Markovin ketjuista. Eräs tällainen kuva on nk. jälkikuvio, jossa on erotettu viivalla parametrin arvo kullakin iteraatiolla. Jälkikuvion pitäisi näyttää sta-

tionääriseltä, eli sen pitäisi vaihdella tasaisesti saman keskiarvon ympärillä ja vaihteluvälin tulisi olla suurin piirtein vakio. (Robert ja Casella, 2009, s. 242–244). Kuvassa 2 on esitetty jälkikuvio ja tiheysfunktion ydinestimaatti Gaussin ytimellä kahdelle eri Markovin ketjulle. Ylärivillä näkyy huonolaatuinen jälkikuvio, jossa kumulatiivinen keskiarvo ei vakiinnu. Lisäksi tiheysfunktio on lievästi kaksihuippuinen. Alarivillä on hyvälaatuinen jälkikuvio. Erot kuvissa saatiin aikaan Metropolisin–Hastingsin algoritmin ehdotusjakauman varianssia muuttamalla.

MCMC-algoritmeissa on myös tyypillistä jättää osa Markovin ketjun alusta käyttämättä analyysissä, koska voidaan ajatella, että ketju ei ole vielä ehtinyt päästä kohdejakauman todennäköisimmälle alueelle. Tätä alkuosaa kutsutaan burn-in jaksoksi.



Kuva 2: Kahden eri Markovin ketjun jälkikuvio harmaalla, kumulatiivinen keskiarvo mustalla viivalla ja tiheysfunktion estimaatti. Molemmissa on estimoitu normaalijakauman odotusarvoa samasta aineistosta, joka on simuloitu standardinormaalijakaumasta.

4 Aineistot ja niiden analyysi

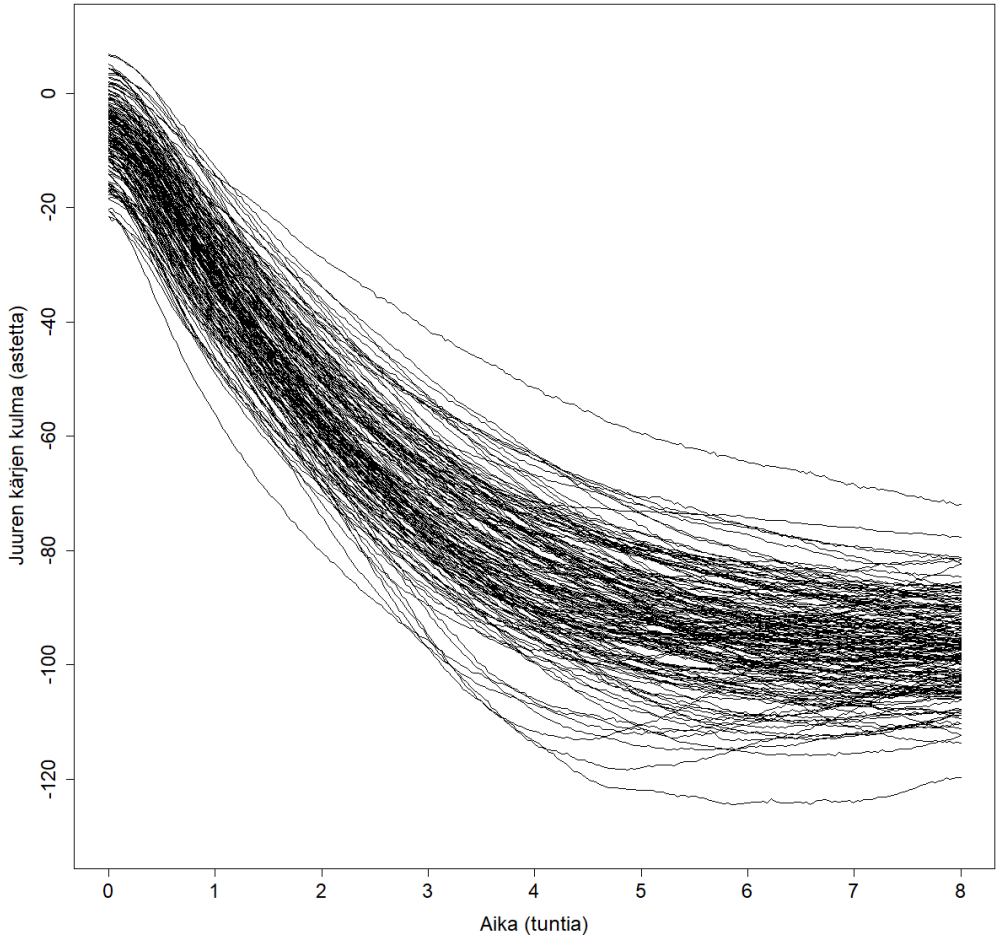
Tässä luvussa esitellään aineistot, joihin sovelletaan alaluvuissa 2.6 ja 2.7 esiteltyjä menetelmiä. Lisäksi esitellään MCMC-algoritmit, joiden avulla tilastollinen päättely tehdään.

4.1 Simuloitu aineisto

Simuloitu aineisto luotiin sen takia, että estimointimenetelmiä voitaisiin testata ja tuloksia verrata aineiston generoineen mallin parametrien arvoihin. Simuloidussa aineistossa oli $N = 100$ yksilöä, joilla kullakin oli $T = 50$ havaintoa tasavälein välillä $[0, 24]$ tuntia. Aineistoon simuloitiin ajasta riippuva geneettinen varianssi $\sigma_G^2(t) = \cos\frac{2\pi}{24}t + 2$ ja ympäristövarianssi $\sigma_E^2(t) = \sin\frac{2\pi}{24}t + 2$. Aineisto voisi vertautua oikeassa elämässä mittauksiin jostakin vuorokauden- tai vuodenajasta syklisesti riippuvasta fenotyypistä. Aikapistekohtainen kovarianssimatriisi oli $\boldsymbol{K}(t) = \sigma_E^2(t)\boldsymbol{I} + \sigma_G^2(t)\boldsymbol{A}$, jossa sukulaisuusmatriisin $\boldsymbol{A}$ alkio $A_{ij} = \frac{1}{2^{|i-j|}},\ i, j = 1, ..., N$. Matriisille $\boldsymbol{K}(t)$ tehtiin Choleskyn hajotelma $\boldsymbol{K}(t) = \boldsymbol{L}(t)\boldsymbol{L}^T(t)$. Vektorit $\boldsymbol{y}(t)$ simuloitiin jakaumasta $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K}(t))$ niin, että $\boldsymbol{y}(t) = \boldsymbol{L}(t)\boldsymbol{d}$, jossa $\boldsymbol{d} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

4.2 Todellinen aineisto

Toinen analysoitava aineisto koostuu lituruohon (*Arabidopsis thaliana*) siemenmittauksista. Siinä on mitattu 162 lituruohon siemenistä kasvavan juuren päähän syntyvän kulman suuruutta. Siemenet olivat petrimaljoissa, jotka asetettiin niin, että niiden kulma maahan nähden oli 90 astetta. Näin painovoima vaikutti kasvavaan juureen kohtisuorasti. Juuria kuvattiin kahdeksan tunnin ajan kahden minuutin välein, joten aineistossa on yhteensä 162 aikasarjaa, joissa kussakin on 241 aikapistettä. Koska aikapisteitä oli näin paljon, valittiin analyysiin joka kuudes eli 40 nopeuttamaan estimointia. Kulmat mi-



Kuva 3: Kaikkien yksittäisten siementen havaintoaikasarjat visualisoituna.

tattiin automaattisesti tietokoneen avulla. Mittaustapaa kuvailee tarkemmin Moore ym. (2013). Aineisto sisältää myös genotyyppitiedot 234 eri markkerista, joten yksilöiden välinen sukulaisuusmatriisi $\boldsymbol{A}$ on myös konstruoitavissa.

Ilmiötä, jossa kasvin juuret tai versot muuttavat kasvusuuntaansa paino-

voimavektoriin nähden, kutsutaan gravitropismiksi. Kasvin kyky aistia painovoima mahdollistaa versojen kasvun valoa kohti, jolloin yhteyttämistä tapahtuu enemmän, ja juurten kasvun syvemmälle maahan, jolloin kasvi on tukevammin pystyssä ja saa paremmin vettä ja mineraaleja. (Masson ym. 2002). Muun muassa tähän ja moneen muuhun tutkimukseen hyödynnetään lituruohoa, joka on erittäin suosittu kasvi biologien keskuudessa. On huomattu, että sen avulla voidaan tutkia monien muidenkin kasvien biologista rakennetta ja toimintaa. Lituruohon hyviä puolia ovat muun muassa sen nopea kasvu ja pieni genomi, joka mahdollistaa yksityiskohtaisen molekulaarisen analyysin. (Meinke ym. 1998)

4.3 Algoritmit

Aineistojen analyysiin kaksivaiheisella menetelmällä käytettiin algoritmia 1 (liite B), joka on muunneltu versio Monterrubio-Gómez ym. (2018) algoritmista 1. Algoritmissa hyödynnetään Gibbsin otantaa, eli siinä simuloidaan yksi parametri tai parametrien osajoukko kerrallaan niin, että muiden parametrien arvot on kiinnitetty. Simulaatioita ei kuitenkaan tehdä parametrien täysehdollisista jakaumista, vaan sileän funktion $\boldsymbol{f}$ arvot simuloidaan käyttämällä elliptistä viipaleotantaa ja skalaariparametrit σ^2 ja l simuloidaan Metropolisin–Hastingsin algoritmilla. Ehdotusjakaumana on satunnaiskulkunormaalijakauma, jonka keskipiste on parametrin edellinen hyväksytty arvo. Jakauman varianssia mukautetaan niin, että ehdotusparametrien hyväksymisaste on noin 0.44. Algoritmia ajettiin 150 000 MCMC-iteraatiota, jossa burn-in -jakso oli 50 000 iteraatiota. Posteriorijakaumista laskettiin parametrien posteriorikeskiarvot ja algoritmin suppenemista tarkasteltiin Markovin ketjujen jälkikuvioista.

Algoritmi 2 (liite B) on yksivaiheisen menetelmän analyysialgoritmi. Se

on hyvin samankaltainen kuin kaksivaiheisessa menetelmässä. Se on Gibbs-henkinen siinä mielessä, että se simuloi toisen varianssikomponenttivektorin kun toinen on kiinnitetty edelliseen hyväksyttyyn arvoonsa. Vektoreiden simulointiin käytetään elliptistä viipaleotantaa. Algoritmia ajettiin 150 000 MCMC-iteraatiota, ja burn-in -jaksona oli ensimmäiset 50 000 iteraatiota. Posteriorijakaumista laskettiin parametrien posteriorikeskiarvot ja niiden 95 %:n posteriorivälit. Algoritmin suppenemista tutkittiin Markovin ketjujen jälkikuvioista. Molemmat esitellyt algoritmit on implementoitu C++-ohjelmointikielellä.

5 Tulokset

Tässä luvussa esitellään analyysien tulokset, verrataan menetelmien estimaatteja ja tutkitaan algoritmien suppenemista Markovin ketjujen jälkikuvioista. Lisäksi siemenaineistosta saatavia tuloksia verrataan kirjallisuudessa aikaisemmin esitettyihin tuloksiin.

5.1 Simuloitu aineisto

Kuvassa 4 on rinnakkain kaksivaiheisen ja yksivaiheisen menetelmän estimaatit simuloidun aineiston varianssikomponentti- ja heritabiliteettifunktioille. Kuvista huomaa selkeän eron kaksivaiheisen menetelmän luottamusvälien ja yksivaiheisen menetelmän posteriorivälien välillä: luottamusvälit ovat huomattavasti leveämpiä kuin posteriorivälit. Ero johtuu siitä, että yksivaiheinen menetelmä ottaa kaikkien aikapisteiden informaation huomioon priorin kautta, jolloin estimoinnin epävarmuus pienenee. Välit ovat kuitenkin määritelmällisesti eri asia, joten niitä ei voi täysin verrata toisiinsa. Näyttää siltä, että näin pienen aineiston perusteella heritabiliteettifunktiosta ei voi sanoa kaksivaiheisen menetelmän perusteella juuri mitään varmaa: luottamusväli täyttää melkein koko parametrin arvoalueen. Simuloidusta aineistosta estimoituja funktioita voidaan verrata oikeisiin funktioihin, koska ne tiedetään. Kaksivaiheisen menetelmän estimaatit näyttävät seuraavan parametrien oikeita arvoja huonommin kuin yksivaiheisen menetelmän estimaatit.

Kaksivaiheisen menetelmän Markovin ketjujen jälkikuviot (liite C) näyttävät kauttaaltaan hyvin käyttäytyviltä: kumulatiivinen keskiarvo vakiintuu nopeasti ja vaihteluväli pysyy samana. Yksivaiheisen menetelmän jälkikuviotkin näyttävät hyväksyttäviltä, mutta niistä huomaa että algoritmia olisi

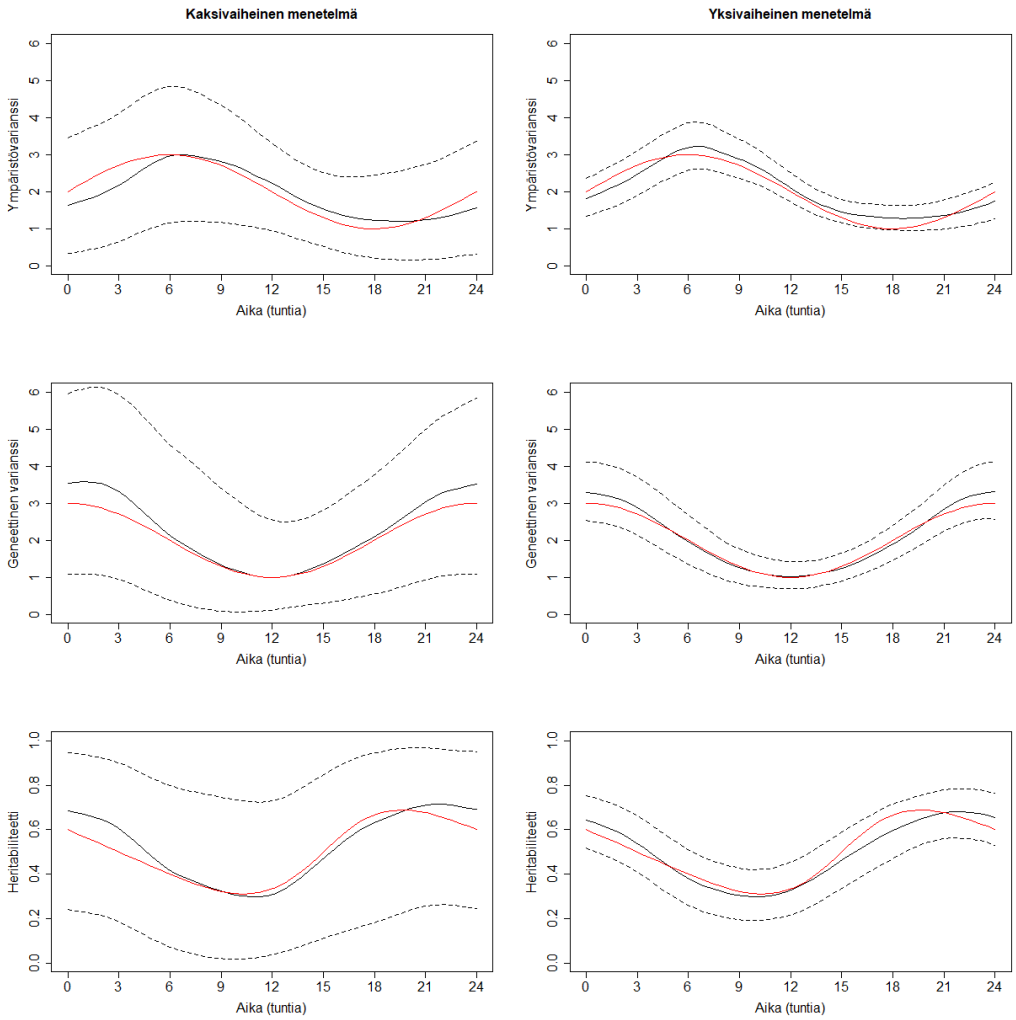
voinut ajaa kauemminkin.

5.2 Todellinen aineisto

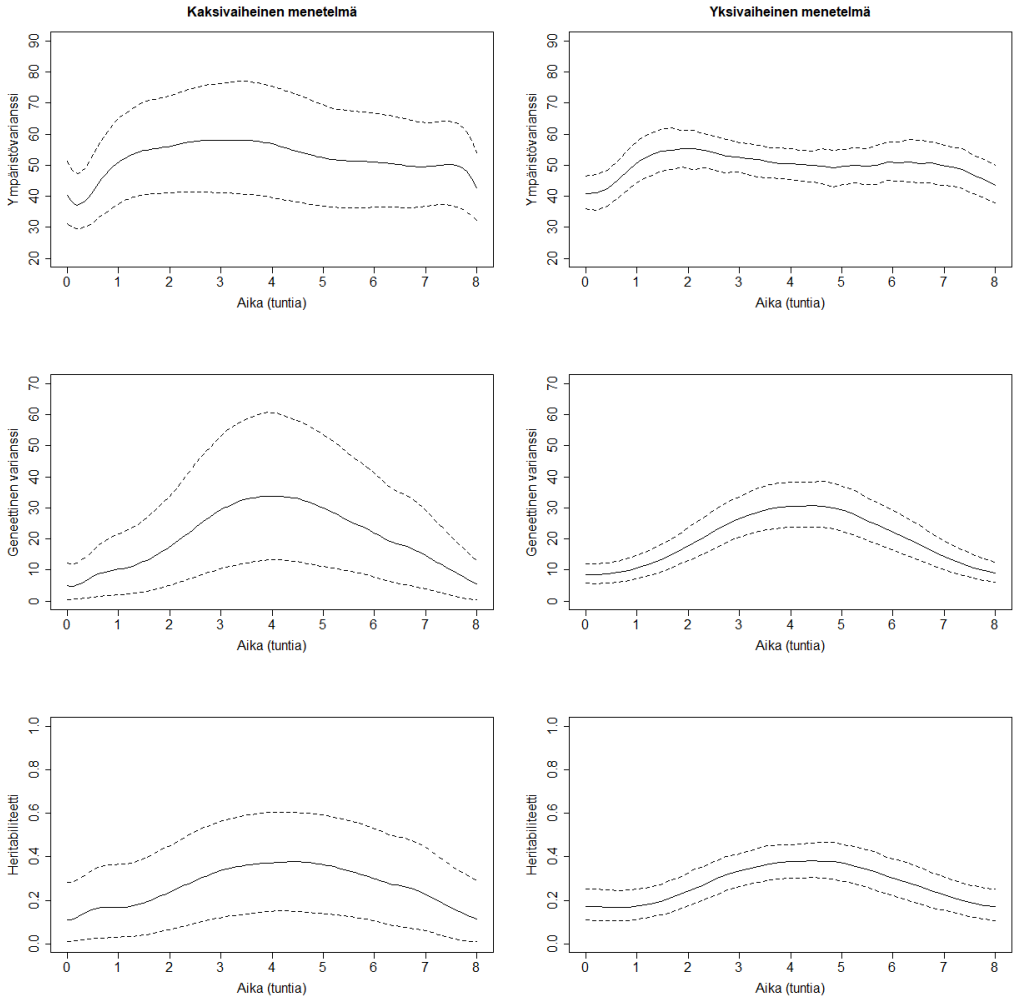
Kuvassa 5 on rinnakkain kaksivaiheisen ja yksivaiheisen menetelmän estimaatit siemenaineiston varianssikomponentti- ja heritabiliteettifunktioille. Luottamusvälien ja posteriorivälien ero on jälleen huomattava, mutta simuloitua aineistoa suuremman yksilömäärän ansiosta kaksivaiheisen menetelmän heritabiliteettiestimaatti on täsmällisempi. Keskiarvofunktioissa ei menetelmien välillä ole suurta eroa. Ympäristövarianssifunktio näyttää melko tasaiselta, kun taas geneettinen varianssi on koholla kokeen keskivaiheilla. Tämä aiheuttaa kummun myös heritabiliteettifunktioon.

Kaksivaiheisen menetelmän jälkikuviot (liite C) näyttävät hyvin käyttäytyviltä. Yksivaiheisen menetelmän jälkikuvioista huomaa jälleen, että MCMC-iteraatioiden määrä olisi voinut olla suurempi.

Samaa siemenaineistoa on analysoitu aikaisemminkin kirjallisuudessa, mutta pääpainona ei ole ollut heritabiliteettifunktion estimointi. Funktiosta on kuitenkin piirretty kuvia, joihin tämän tutkielman tuloksia voi verrata. Muun muassa Vanhatalo ym. (2019) on analysoinut samaa aineistoa. Siinä heritabiliteettifunktion estimaatti on hyvin lähellä tämän työn estimaatteja. Tosin siinä funktiolle ei ole estimoitu posteriori- tai luottamusvälejä. Myös Moore ym. (2013) on analysoinut siemenaineistoa. Siinä varianssikomponentti- ja heritabiliteettifunktiot ovat samanmuotoisia kuin tässä tutkielmassa, mutta niiden saamat arvot ovat erikoisella tavalla aivan eri luokkaa. Tässäkään artikkelissa ei ole posteriori- tai luottamusvälejä funktioille.



Kuva 4: Simuloidusta aineistosta eri menetelmillä estimoidut ympäristövarianssi, geneettinen varianssi ja heritabiliteetti mustalla viivalla ja niiden 95 %:n luottamusvälit tai posteriorivälit katkoviivalla. Parametrien oikeat arvot on piirretty punaisella viivalla.



Kuva 5: Todellisesta aineistosta eri menetelmillä estimoidut ympäristövarianssi, geneettinen varianssi ja heritabiliteetti mustalla viivalla ja niiden 95 %:n luottamusvälit tai posteriorivälit katkoviivalla.

6 Pohdinta

Tässä työssä esiteltiin kaksi bayesiläistä menetelmää funktiomuotoisen heritabiliteetin estimoimiseen pitkittäisaineistosta: kaksivaiheinen menetelmä ja yksivaiheinen menetelmä. Menetelmät ovat siinä mielessä joustavia, että ne eivät aseta juurikaan rajoituksia heritabiliteettifunktion muodolle. Sen tulee olla ainoastaan riittävän sileä. Lisäksi menetelmät yleistyvät kaikkiin sellaisiin tilanteisiin, joissa biologisen populaation sukulaisuusmatriisi tiedetään, oli se sitten konstruoitu molekulaarisen markkeriaineiston tai sukupuun perusteella. Yksivaiheisen menetelmän vahvuuksiin vaikuttaa kuuluvan myös estimoinnin täsmällisyys. Algoritmeja testattaessa huomattiin, että heritabiliteettiestimaatti vain yhdelle aikapisteelle, riippumatta muista pisteistä, on melko epätäsmällinen. Kun taas heritabiliteetti estimoidaan useille aikapisteille yhtä aikaa, niin priorin kautta tuleva lisäinformaatio helpottaa tehtävää huomattavasti. Ilmiön voi huomata vertaamalla menetelmien luottamus- ja posteriorivälejä tulokset-luvussa. Kaksivaiheisen menetelmän vahvuus taas on sen hyvä skaalautuvuus. Sillä saa isoillekin aineistoille estimaatteja kohtuullisessa ajassa. Tässä työssä analysoiduilla aineistoilla kaksivaiheisella menetelmällä kesti noin tunnin saada tuotettua kaikki estimaatit. Yksivaiheisella menetelmällä siinä kesti simuloidulle aineistolle noin vuorokausi ja todelliselle aineistolle noin kolme vuorokautta.

Hyvä jatkotutkimuksen aihe voisi olla yksivaiheisen menetelmän kehittäminen nopeammaksi. Suurin pullonkaula algoritmissa on $N \times N$ -kokoisen matriisin determinantin laskeminen ja kääntäminen, joka tehdään joka MCMC-iteraatiolla T kertaa. Matriisi on yleisemmässä muodossa $a\boldsymbol{A} + b\boldsymbol{I}$, jossa pienet kirjaimet ovat skalaareja ja isot matriiseja. Kaavassa päivittyy aina toinen skalaarikomponentti, kun toinen pysyy samana. Voisi yrittää selvittää, onko ao. matriisin determinantille ja käänteismatriisille olemassa käyttökelpoista

päivityskaavaa, joka nopeuttaisi laskentaa.

Toinen kehityskohde on kiinnitettävien hyperparametrien arvojen valinta. Tällä hetkellä ne täytyy valita kokeilemalla, johon voi kulua paljon aikaa. Yksi vaihtoehto on niiden estimointi muiden parametrien tapaan, mutta käytännössä niiden identifioituvuus havaintoaineistosta voi osoittautua ongelmalliseksi.

Viitteet

Abadir, Karim M. ja Magnus, Jan R. (2005). *Matrix Algebra*. Cambridge University Press, Cambridge.

Bryois, Julien, Buil, Alfonso, Ferreira, Pedro G., Panousis, Nikolaos I., Brown, Andrew A., Viñuela, Ana, Planchon, Alexandra, Bielser, Deborah, Small, Kerrin, Spector, Tim ja Dermitzakis, Emmanouil T. (2017). Time-dependent genetic effects on gene expression implicate aging processes. *Genome Research* 27: 542–552.

Efron, Bradley ja Hastie, Trevor (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4: 250–255.

Fahrmeir, Ludwig ja Kneib, Thomas (2011). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press, Oxford.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki ja Rubin, Donald B. (2013). *Bayesian Data Analysis*. CRC Press LLC, Boca Raton.

Kwak, Il-Youp, Moore, Candace R., Spalding, Edgar P. ja Broman, Karl W. (2014). A Simple Regression-Based Method to Map Quantitative Trait Loci Underlying Function-Valued Phenotypes. *Genetics* 197 (4): 1409–1416.

Kang, Hyun Min, Zaitlen, Noah A., Wade, Claire M., Kirby, Andrew, Heckerman, David, Daly, Mark J. ja Eskin, Eleazar (2008). Efficient control

of population structure in model organism association mapping. *Genetics* 178 (3): 1709–1723.

Li, Zitong ja Sillanpää, Mikko J. (2015). Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data. *Trends in Plant Science* 20 (12): 822–833.

Masson, Patrick H., Tasaka, Masao, Morita, Miyo T., Guan, Changhui, Chen, Rujin ja Boonsirichai, Kanokporn (2002). Arabidopsis thaliana: A Model for the Study of Root and Shoot Gravitropism. *The Arabidopsis book / American Society of Plant Biologists* 1: e0043.

Meinke, David W., Cherry, J. Michael, Dean, Caroline, Rounsley, Steven D. ja Koornneef, Maarten (1998). Arabidopsis thaliana: A Model Plant for Genome Analysis. *Science* 282: 662, 679–82.

Monterrubio-Gómez, Karla, Roininen, Lassi, Wade, Sara, Damoulas, Theo ja Girolami, Mark (2018). Posterior Inference for Sparse Hierarchical Non-stationary Models. *arXiv e-prints*, arXiv:1804.01431: arXiv:1804.01431. arXiv: `1804.01431 [stat.CO]`.

Moore, Candace R., Johnson, Logan S., Kwak, Il-Youp, Livny, Miron, Broman, Karl W. ja Spalding, Edgar P. (2013). High-Throughput Computer Vision Introduces the Time Axis to a Quantitative Trait Map of a Plant Growth Response. *Genetics* 195 (3): 1077–1086.

Murray, Iain, Adams, Ryan Prescott ja MacKay, David J.C. (2010). Elliptical Slice Sampling. *Proceedings of Machine Learning Research* 9: 541–548.

Pletcher, Scott D. ja Geyer, Charler J. (1999). The Genetic Analysis of Age-Dependent Traits: Modeling the Character Process. *Genetics* 153: 825–835.

Rasmussen, Carl Edward ja Williams, Christopher K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge.

Robert, Christian P. ja Casella, George (2009). *Introducing Monte Carlo Methods with R*. Springer, New York.

Sillanpää, Mikko J. (2011). On statistical methods for estimating heritability in wild populations. *Molecular Ecology* 20: 1324–1332.

Vanhatalo, Jarno, Li, Zitong ja Sillanpää, Mikko J. (2019). A Gaussian process model and Bayesian variable selection for mapping function-valued quantitative traits with incomplete phenotypic data. *Bioinformatics*. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz164. eprint: http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz164/28355391/btz164.pdf. URL: https://doi.org/10.1093/bioinformatics/btz164.

VanRaden, Paul M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91 (11): 4414–4423.

Visscher, Peter M., Hill, William G. ja Wray, Naomi R. (2008). Heritability in the genomics era – concepts and misconceptions. *Nature Reviews Genetics* 9: 255–266.

A Kroneckerin matriisitulo

Tässä työssä tilastollisen mallin määrittelyssä tarvitaan Kroneckerin tuloa. Hieman harvinaisempana operaationa sitä on syytä avata. Kahden reaaliarvoisen matriisin $\boldsymbol{A}_{n\times m}$ ja $\boldsymbol{B}_{p\times q}$ Kroneckerin tulo määritellään

$$\boldsymbol{A}\otimes\boldsymbol{B}=\begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1m}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2m}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & \dots & a_{nm}\boldsymbol{B} \end{bmatrix}. \tag{47}$$

Tulomatriisi on kooltaan $np \times mq$. Kroneckerin tulolle on johdettu monia ominaisuuksia, joista todettakoon tässä työssä tärkeät

$$\begin{aligned} &(\boldsymbol{A}\otimes\boldsymbol{B})^{-1}=\boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}, \\ &\quad|\boldsymbol{A}\otimes\boldsymbol{B}|=|\boldsymbol{A}|^{n}|\boldsymbol{B}|^{m}, \ \text{ kun } \boldsymbol{A}\in\mathbb{R}^{n\times n} \text{ ja } \boldsymbol{B}\in\mathbb{R}^{m\times m}. \end{aligned} \tag{48}$$

(Abadir ja Magnus, 2005, s. 273-279)

B MCMC-algoritmit

Algoritmi 1

Alkuarvot: $\boldsymbol{f}^{(1)}, \sigma^{2(1)}, l^{(1)}$

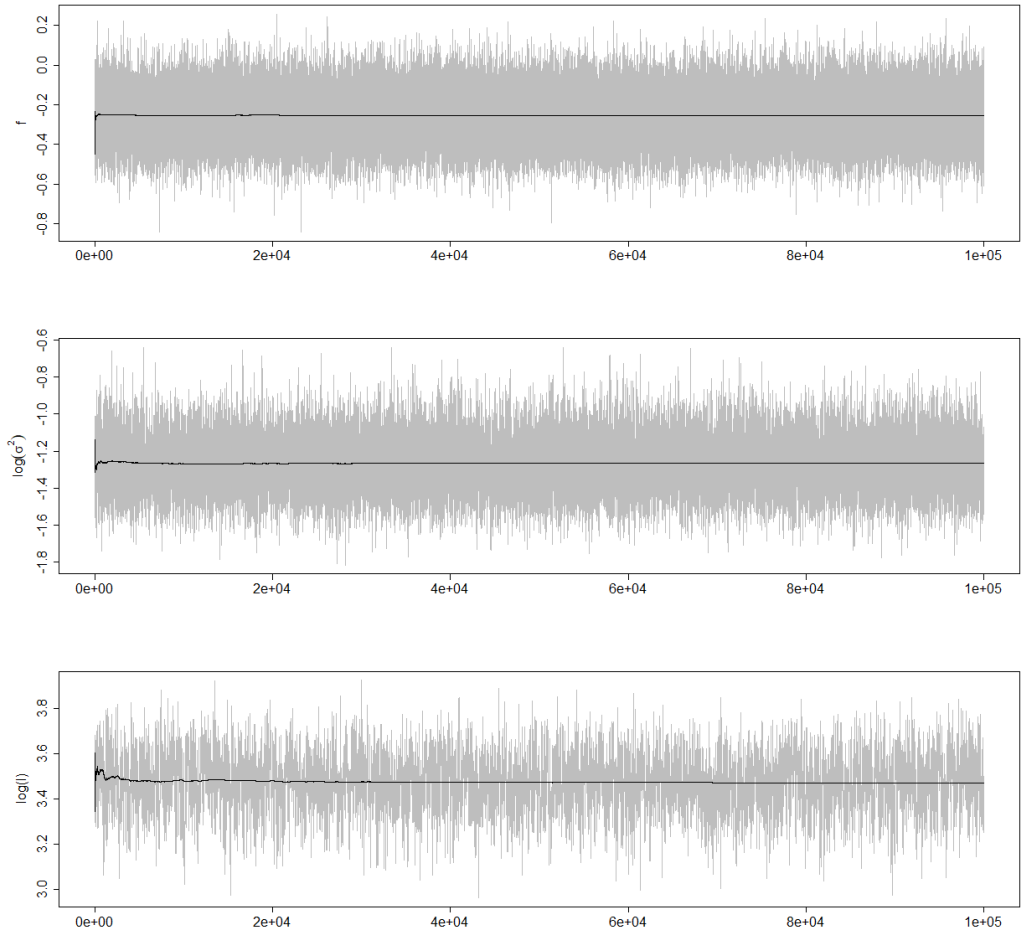
1: t = 1:(T - 1)
2: simuloi $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$
3: simuloi $u \sim \mathcal{U}(0,1)$
4: laske $\kappa = \log p_{\mathcal{N}}(\boldsymbol{y}|\boldsymbol{f}^{(t)}, \sigma^{2(t)}) + \log u$
5: simuloi $\rho \sim \mathcal{U}(0, 2\pi)$
6: aseta $(\rho_{\min}, \rho_{\max}) = (\rho - 2\pi, \rho)$
7: laske $\boldsymbol{f}' = \boldsymbol{f}^{(t)} \cos\rho + \boldsymbol{v} \sin\rho$
8: **jos** $p_{\mathcal{N}}(\boldsymbol{y}|\boldsymbol{f}',\ \sigma^{2(t)}) > \kappa$ **niin**
9: aseta $\boldsymbol{f}^{(t+1)} = \boldsymbol{f}'$
10: **muutoin**
11: **jos** $\rho < 0$ **niin**
12: aseta $\rho_{\min} = \rho$
13: **muutoin**
14: aseta $\rho_{\max} = \rho$
15: **lopeta jos**
16: simuloi $\rho \sim \mathcal{U}(\rho_{\min}, \rho_{\max})$
17: palaa kohtaan 7.
18: **lopeta jos**
19: simuloi $\log \sigma^{2'} \sim N(\log \sigma^{2(t)}, s_1)$
20: laske $\alpha_{\sigma^2} = \min\left\{1, \frac{p_{\mathcal{N}}(\boldsymbol{y}|\boldsymbol{f}^{(t+1)},\ \sigma^{2'}) p_{u;\log\sigma^2}(\log\sigma^{2'})}{p_{\mathcal{N}}(\boldsymbol{y}|\boldsymbol{f}^{(t+1)},\ \sigma^{2(t)}) p_{u;\log\sigma^2}(\log\sigma^{2(t)})}\right\}$
21: todennäköisyydellä α_{σ^2} aseta $\log\sigma^{2(t+1)} = \log\sigma^{2'}$, muutoin aseta $\log\sigma^{2(t+1)} = \log\sigma^{2(t)}$
22: mukauta s_1
23: simuloi $\log l' \sim N(\log l^{(t)}, s_2)$
24: laske $\alpha_l = \min\left\{1, \frac{p_{\mathcal{N}}(\boldsymbol{f}^{(t+1)}|\boldsymbol{0},\ \boldsymbol{C}_{l'}) p_{u;\log l}(\log l')}{p_{\mathcal{N}}(\boldsymbol{f}^{(t+1)}|\boldsymbol{0},\ \boldsymbol{C}_{l^{(t)}}) p_{u;\log l}(\log l^{(t)})}\right\}$
25: todennäköisyydellä α_l aseta $\log l^{(t+1)} = \log l'$, muutoin aseta $\log l^{(t+1)} = \log l^{(t)}$
26: mukauta s_2
27: **lopeta**

Algoritmi 2

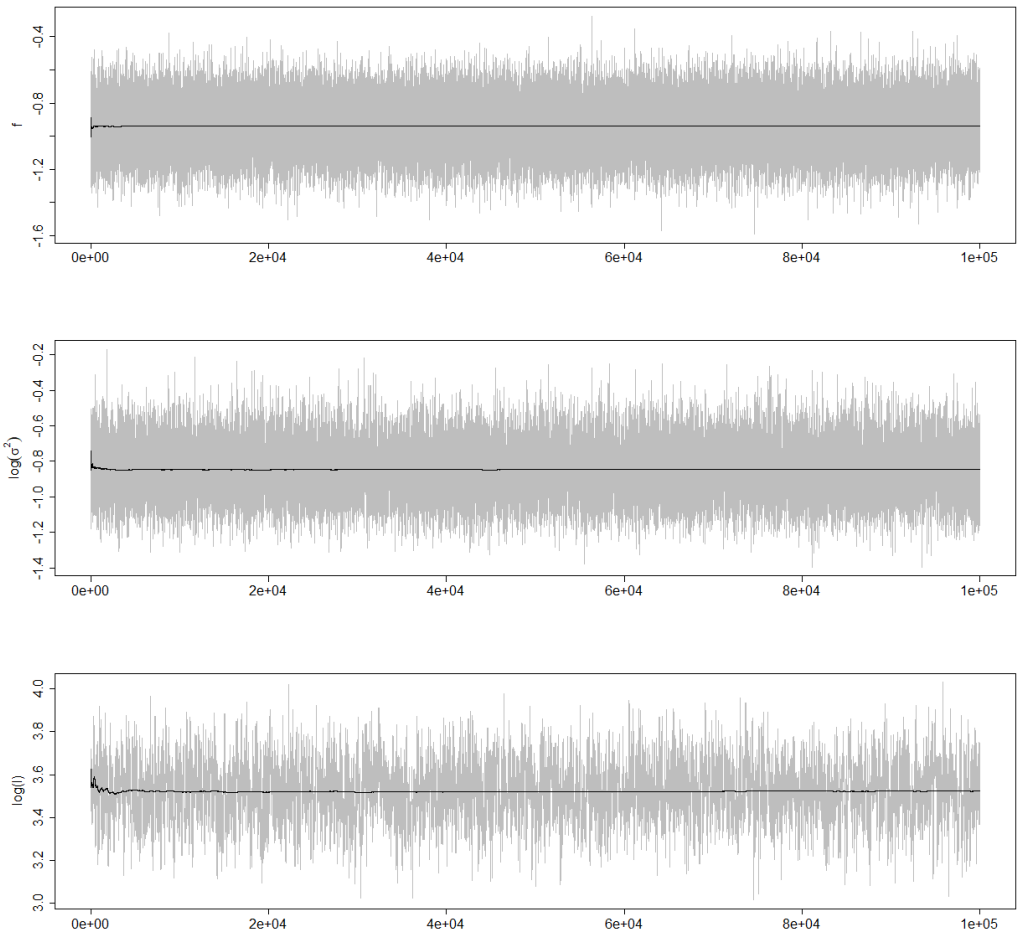
Alkuarvot: $\boldsymbol{\sigma}_G^{2(1)}, \boldsymbol{\sigma}_E^{2(1)}$

t = 1:(T - 1)
  simuloi $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_G)$
  simuloi $u \sim \mathcal{U}(0,1)$
  laske $\kappa = \log p_{\mathcal{N}}(\tilde{\boldsymbol{y}}_T|\mathbf{0}, \boldsymbol{K}_T) + \log u$
  simuloi $\rho \sim \mathcal{U}(0, 2\pi)$
  aseta $(\rho_{\min}, \rho_{\max}) = (\rho - 2\pi, \rho)$
  laske $\boldsymbol{\sigma}_G^{2'} = \boldsymbol{\sigma}_G^{2(t)} \cos\rho + \boldsymbol{v} \sin\rho$
  päivitä $\boldsymbol{K}_T$
  jos $\log p_{\mathcal{N}}(\tilde{\boldsymbol{y}}_T|\mathbf{0}, \boldsymbol{K}_T) > \kappa$ **niin**
    aseta $\boldsymbol{\sigma}_G^{2(t+1)} = \boldsymbol{\sigma}_G^{2'}$
  muutoin
    jos $\rho < 0$ **niin**
      aseta $\rho_{\min} = \rho$
    muutoin
      aseta $\rho_{\max} = \rho$
    lopeta jos
    simuloi $\rho \sim \mathcal{U}(\rho_{\min}, \rho_{\max})$
    palaa kohtaan 7.
  lopeta jos
  simuloi $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_E)$
  simuloi $u \sim \mathcal{U}(0,1)$
  laske $\kappa = \log p_{\mathcal{N}}(\tilde{\boldsymbol{y}}_T|\mathbf{0}, \boldsymbol{K}_T) + \log u$
  simuloi $\rho \sim \mathcal{U}(0, 2\pi)$
  aseta $(\rho_{\min}, \rho_{\max}) = (\rho - 2\pi, \rho)$
  laske $\boldsymbol{\sigma}_E^{2'} = \boldsymbol{\sigma}_E^{2(t)} \cos\rho + \boldsymbol{v} \sin\rho$
  päivitä $\boldsymbol{K}_T$
  jos $\log p_{\mathcal{N}}(\tilde{\boldsymbol{y}}_T|\mathbf{0}, \boldsymbol{K}_T) > \kappa$ **niin**
    aseta $\boldsymbol{\sigma}_E^{2(t+1)} = \boldsymbol{\sigma}_E^{2'}$
  muutoin
    jos $\rho < 0$ **niin**
      aseta $\rho_{\min} = \rho$
    muutoin
      aseta $\rho_{\max} = \rho$
    lopeta jos
    simuloi $\rho \sim \mathcal{U}(\rho_{\min}, \rho_{\max})$
    palaa kohtaan 25.
  lopeta jos
lopeta

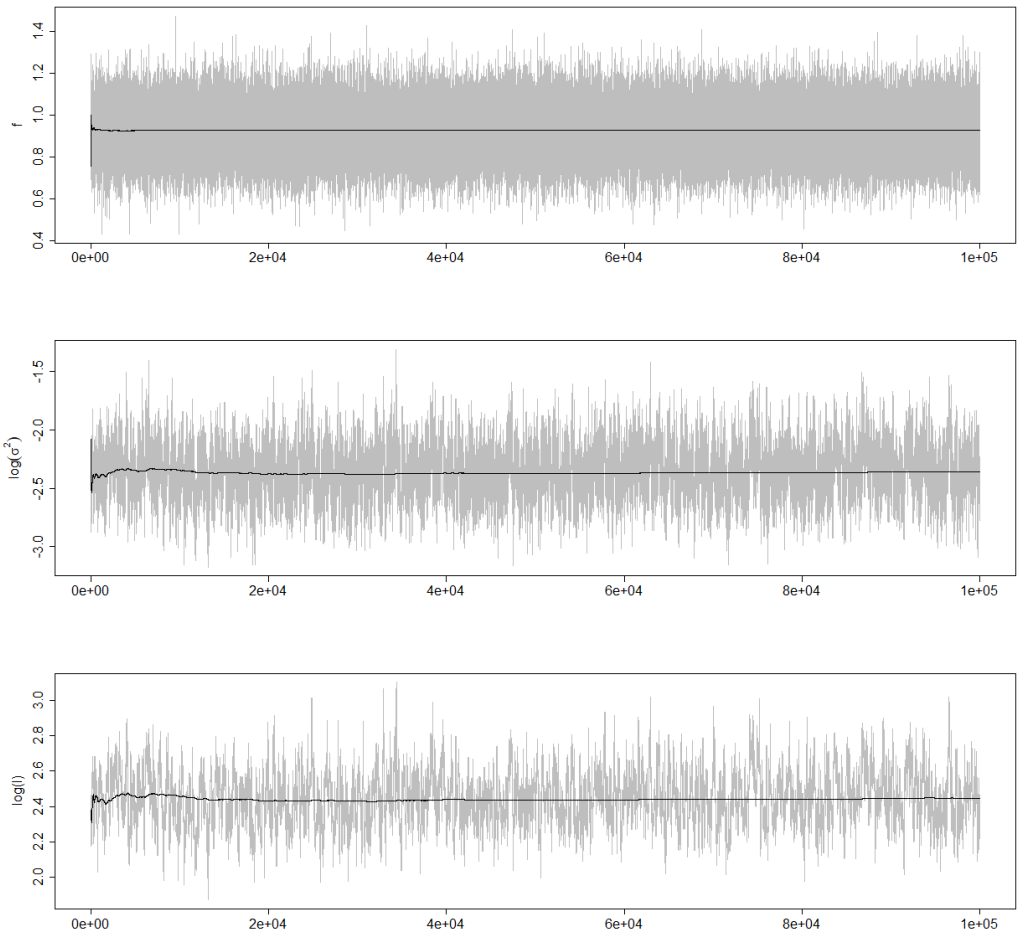
C Jälkikuviot



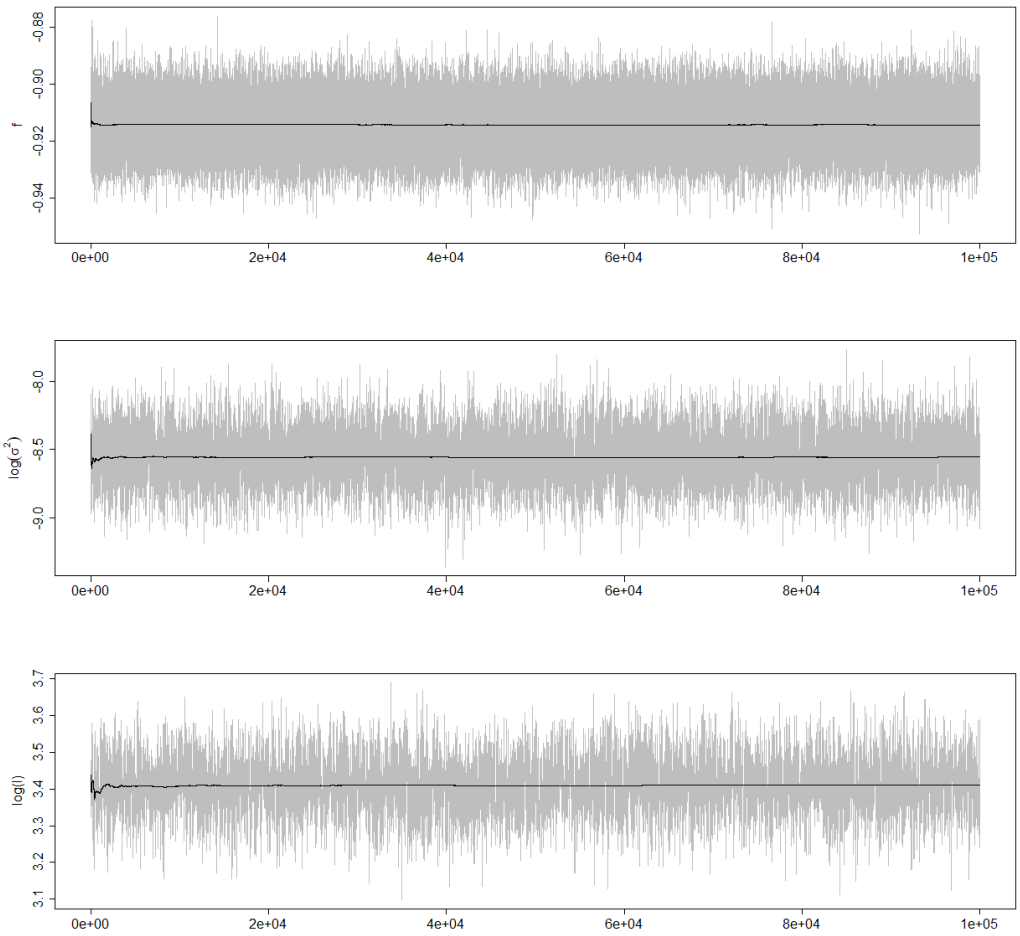
Kuva 6: Kaksivaiheisen menetelmän simuloidun aineiston ympäristövarianssin silotusmallin parametrien Markovin ketjujen jälkikuviot ja kumulatiiviset keskiarvot. Funktion $\boldsymbol{f}$ komponenttiparametri on satunnaisesti valittu.



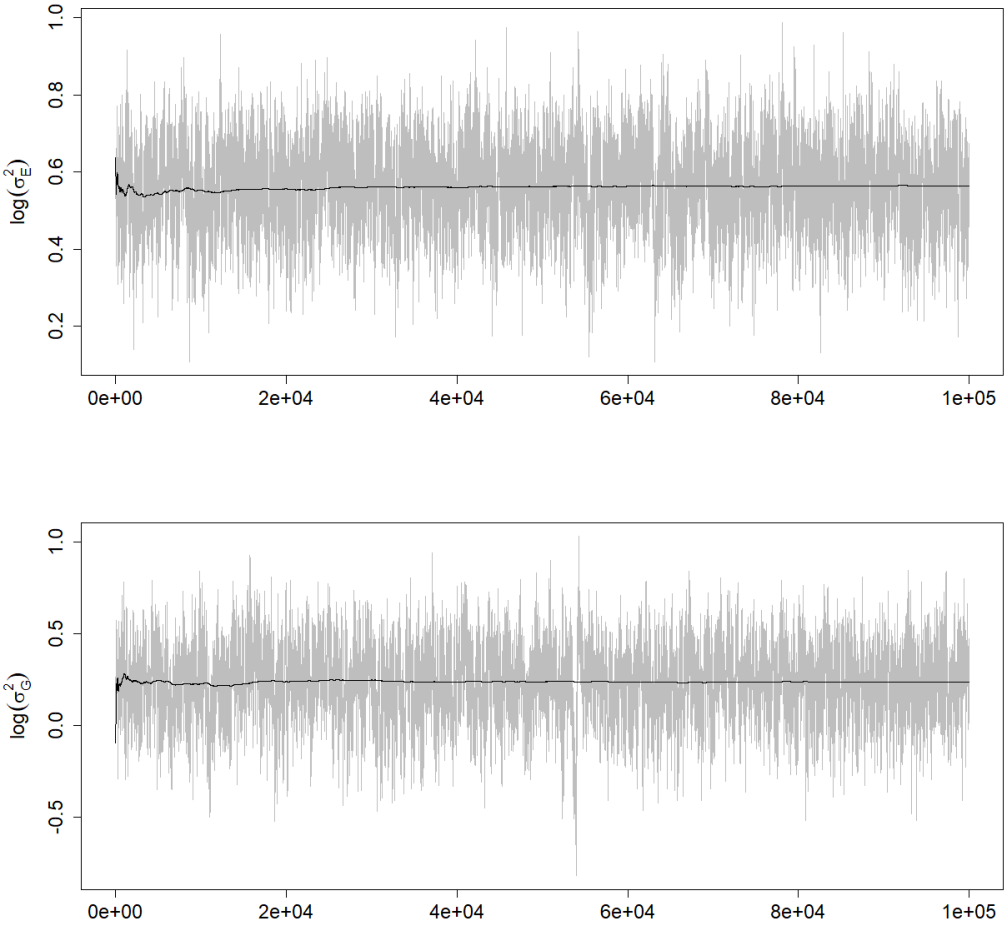
Kuva 7: Kaksivaiheisen menetelmän simuloidun aineiston geneettisen varianssin silotusmallin parametrien Markovin ketjujen jälkikuviot ja kumulatiiviset keskiarvot. Funktion $\boldsymbol{f}$ komponenttiparametri on satunnaisesti valittu.



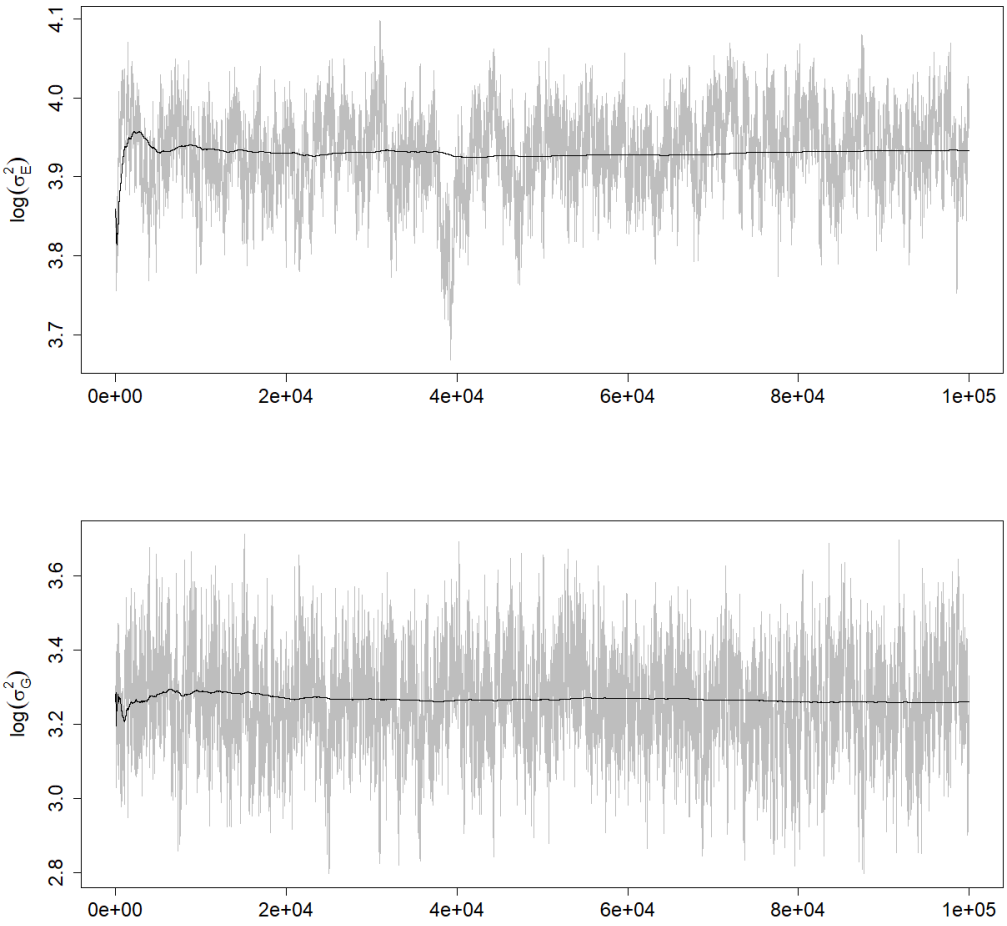
Kuva 8: Kaksivaiheisen menetelmän siemenaineiston ympäristövarianssin silotusmallin parametrien Markovin ketjujen jälkikuviot ja kumulatiiviset keskiarvot. Funktion $\boldsymbol{f}$ komponenttiparametri on satunnaisesti valittu.



Kuva 9: Kaksivaiheisen menetelmän siemenaineiston geneettisen varianssin silotusmallin parametrien Markovin ketjujen jälkikuviot ja kumulatiiviset keskiarvot. Funktion $\boldsymbol{f}$ komponenttiparametri on satunnaisesti valittu.



Kuva 10: Yksivaiheisen menetelmän simuloidun aineiston ympäristö- ja geneettisen varianssin satunnaisen komponenttiparametrin Markovin ketjujen jälkikuviot ja kumulatiiviset keskiarvot.



Kuva 11: Yksivaiheisen menetelmän siemenaineiston ympäristö- ja geneettisen varianssin satunnaisen komponenttiparametrin Markovin ketjujen jälkikuviot ja kumulatiiviset keskiarvot.