Sisältö

Luku 1

Johdanto

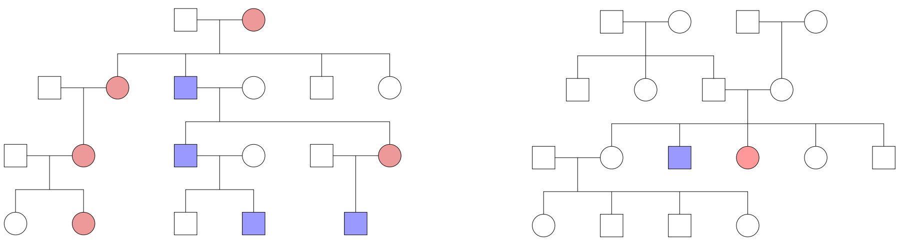
Perinnöllisten tautien tutkimuksessa yksi keskeinen tavoite on taudin *mallintaminen*. Mallintamisella tarkoitetaan tautiin vaikuttavien tekijöiden kuvaamista erityisellä *mallilla*. Hyvä malli sopii tautiin liittyviin havaintoihin ja auttaa jatkotutkimuksissa, joita voivat olla mm. tautiin vaikuttavien tekijöiden fyysinen paikantaminen, taudin biologisten mekanismien selvittäminen ja lääkkeiden kehittäminen.

Usein taudin mallintaminen alkaa sairaiksi havaittujen yksilöiden sukulaisten diagnosoinnilla. Tarkoituksena on tällöin muodostaa *sukupuita*, jotka kertovat taudin esiintymisestä perheissä ja laajemmin suvuissa. Sukupuiden perusteella voidaan arvioida tautiin vaikuttavien perintötekijöiden eli *geenien* lukumäärää ja vaikutusta. Tällaista arviointia kutsutaan *monimutkaiseksi segregaatioanalyysiksi* (engl. complex segregation analysis, jäljempänä lyhyemmin *segregaatioanalyysi* [1]). Kun periytymismalli tunnetaan, siirrytään tyypillisesti *kytkentäanalyysiin* (engl. linkage analysis), joka pyrkii määrittämään geenien sijainnin perimässä.

Yksinkertaisimmillaan taudit ovat yhden geenivirheen aiheuttamia. Tällöin ne usein luokitellaan vallitsevasti (engl. dominant), peittyvästi (engl. resessive) tai sukupuoleen sidotusti periytyviin. Tarkastellaan esimerkkinä kuvan 1.1 sukupuita olettaen, että lukija tuntee perinnöllisyystieteen peruskäsitteet ja -

[1] Tarkasti ottaen segregaatioanalyysillä tarkoitetaan ns. segregaatiosuhteiden selvittämistä, kun vaikuttavia geenejä oletetaan olevan vain yksi. *Monimutkaisesta* segregaatioanalyysistä puhutaan yleensä silloin, kun tarkoitetaan periytymismallin laajempaa selvittämistä. Tällöin voidaan sallia myös useampia vaikuttavia geenejä.

mekanismit (hyviä lähteitä ovat esimerkiksi [Val95, UVV95]). Kuvan 1.1.a sukupuun havaitaan sopivan vallitsevaan periytymiseen. Vallitsevassa periytymisessä tauti puhkeaa, vaikka viallinen geeni peritään ainoastaan toiselta vanhemmalta. Sairaan vanhemman jälkeläisistä on tällöin puolet sairaita. Kuvan 1.1.b sukupuu sopii puolestaan peittyvään periytymiseen. Peittyvästi periytyvissä sairauksissa taudin oireet ilmenevät vain, jos lapsi saa virheellisen geenin sekä isältään että äidiltään. Jos näihin yksinkertaisiin malleihin sopivia havaintoja on riittävästi, voidaan päätellä, että tautiin vaikuttaa vain yksi geeni.



(a) (b)

Kuva 1.1: Vallitsevaa (a) ja peittyvää (b) periytymistä kuvaavat sukupuut (mukailtu lähdettä [UVV95, s. 53]). Miehet on kuvattu neliöinä ja naiset ympyröinä. Sairaat yksilöt on merkitty värjätyin symbolein.

Selkeä luokittelu edellä mainittuihin muotoihin ei aina ole mahdollista. Tautiin voi vaikuttaa useita geenejä tai sairaus voi olla seurausta ympäristön ja perimän monimutkaisesta yhteisvaikutuksesta. Tällaisissa tapauksissa puhutaan *monitekijäisestä* (engl. multifactorial) taudista. Esimerkiksi astma, diabetes ja skitsofrenia ovat yleisiä sairauksia, joiden arvellaan olevan monitekijäisiä.

Moderni segregaatioanalyysi lähestyy monitekijäisyyden ongelmaa todennäköisyysmallien ja tilastollisen päättelyn avulla. Tautiin sopivien mallien etsiminen nojautuu laskennalliseen analyysiin, jonka tärkein ja vaativin osa on yleensä havaintojen uskottavuuden laskeminen annetussa mallissa. Kun havainnot koostuvat suurista sukupuista, tai kun mallit ovat monimutkaisia, uskottavuuden laskeminen tarjoaa tietojenkäsittelytieteellisiä haasteita. Mallintajan on tasapainoteltava mallien kuvausvoiman ja laskennallisen vaativuuden välillä. Mallien valinnassa on lisäksi on otettava huomioon käytössä olevien havaintojen rajallisuus. Esimerkiksi muutamasta pienestä sukupuusta koostuvaa aineistoa ei ole järkevää selittää mallilla, jossa on satoja parametreja.

Tässä tutkielmassa käsitellään segregaatioanalyysin muotoa, jossa havainnot rajoittuvat suurten sukupuiden sijaan sukulaisriskeihin. Sukulaisriskit kuvaavat sairaan yksilön sukulaisten todennäköisyyttä sairastua. Esimerkiksi sisarusriski on todennäköisyys, että yksilö on sairas, kun hänen vanhemmalla sisarellaan tiedetään olevan tauti. Sukulaisriskien käyttö rajaa tarkastelun kaksivasteisiin tauteihin. Toisin sanoen jokaisen yksilön oletetaan olevan kyseisen taudin suhteen joko sairas tai terve – ilman välimuotoja. Oletus sulkee pois tavallisesti segregaatioanalyysin kohteena olevat kvantitatiiviset eli jatkuva- tai monivasteiset taudit.

Havainnot sukulaisriskeistä antavat tietoa taudin periytymismallista. Esimerkiksi identtisen kaksosen korkea sairastumisriski kertoo taudin perinnöllisyydestä. Jos riski sen sijaan on lähellä taudin yleisyyttä väestössä, on luultavaa, että ympäristötekijät vaikuttavat sairauden ilmenemiseen perinnöllisiä tekijöitä enemmän. Sisarusriski voi puolestaan antaa tietoa vallitsevuudesta ja peittyvyydestä kuten kuvan 1.1 esimerkissä todettiin. Tällaisia nyrkkisääntöjä hienostuneempiakin menetelmiä on esitetty [CaE71, Ris90]. Tässä työssä esitetään uusi, bayesiläinen lähestymistapa sukulaisriskihavaintoja hyödyntävään perinnöllisten tautien mallintamiseen.

Todennäköisyysmallintamista varten tautia koskevia havaintoja on syytä tarkastella huolella. Tämä tehdään luvussa 2, jossa sukulaisriskihavainnot määritellään yleisten sukupuuhavaintojen avulla.

Luvussa 3 siirrytään tarkastelemaan perinnöllisten tautien todennäköisyysmalleja eli *geneettisiä malleja*. Lähtökohdaksi otetaan Elstonin ja Stewartin [ElS71] yleinen sukupuuhavaintojen mallintamiseen tarkoitettu kehys. Kun segregaatioanalyysissä usein oletetaan, että useat geenit vaikuttavat eräässä mielessä toisistaan riippumatta, tässä työssä pitäydytään malleissa, jotka kykenevät kuvaamaan geenien mielivaltaisia yhteisvaikutuksia. Keskeisten biologisten olettamusten vaikutusta mallien parametrisointiin käsitellään.

Työn alkuosan päättää luku 4, jossa käsitellään havaintojen ja mallien suhdetta. Tilastollista päättelyä varten sukulaisriskihavaintojen uskottavuus määritellään geneettisten mallien funktiona. Bayesiläistä lähestymistapaa tarjotaan vaihtoehdoksi klassisen suurimman uskottavuuden menetelmälle. Lisäksi pohditaan havaintojen määrän ja laadun merkitystä.

Työn alkuosa selvittää mallintamisen perusasiat: havainnot, mallit ja todennäköisyyksiin nojautuvan päättelyn. Tarkasteltavan menetelmän toteutuksellisten ongelmien ratkaisemiselle luodaan kuitenkin vasta pohja. Alkuosan jälkeen tutkielmassa keskitytään kahteen laskennalliseen kysymykseen. Niitä yhdistää työn aiheena olevat sukulaisriskit ja geneettiset mallit, mutta monilta osin niitä voidaan pitää toisistaan riippumattomina ongelmina.

Ensin, luvussa 5, syvennytään sukulaisriskien laskemiseen, joka on välttämätöntä sukulaisriskihavaintojen uskottavuuden määräämiseksi annetussa geneettisessä mallissa. Näkökulma on tietojenkäsittelytieteellinen: etsitään mahdollisimman nopeita laskentamenetelmiä ja kuvataan ne aikavaativuuden arvioimista varten riittävän yksityiskohtaisesti.

Luvussa 6 tarkastellaan bayesiläisen päättelyn vaatimien moniulotteisten integraalien numeerista laskemista Markovin ketjua hyödyntävällä Monte Carlo -menetelmällä (engl. Markov chain Monte Carlo, lyh. MCMC). Ensin käsitellään lyhyesti MCMC-menetelmän teoriaa, minkä jälkeen esitellään toteutetun simulaatio-ohjelman keskeisiä periaatteita. Ohjelman ja menetelmän ominaisuuksia tarkastellaan luvussa 7 keinotekoisten ja todellisten havaintoaineistojen valossa.

Luvussa 8 tehdään yhteenveto ja hahmotellaan joitakin jatkotutkimuksen aiheita.

Luku 2

Sukulaisriskihavainnot

Tässä luvussa käsitellään perinnölliseen tautiin liittyvien havaintojen rakennetta. Tarkastelu aloitetaan sukupuiden ja sukupuuhavaintojen määrittelemisellä. Tämän jälkeen taudin yleisyyttä ja sukulaisriskejä koskevat havainnot muotoillaan yleisten sukupuuhavaintojen erikoistapauksina.

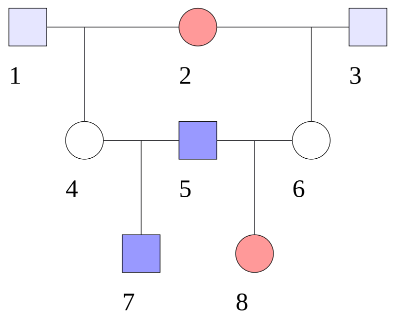
2.1 Populaatio ja sukupuuhavainnot

Perinnöllistä tautia tutkittaessa tarkastellaan taudin esiintymistä jossakin kohteeksi valitussa *populaatiossa* eli *yksilöiden* muodostamassa joukossa. Tautia koskevat havainnot esitetään usein sukupuuna, joka paljastaa sukupuun yksilöiden sukulaisuussuhteet ja ilmiasun taudin suhteen. Joidenkin yksilöiden ilmiasu voi olla myös tuntematon. Niitä yksilöitä, joiden ilmiasu on *havaittu*, kutsutaan jäljempänä *havainnoiduiksi*. Lähes poikkeuksetta sukupuu kattaa vain osan kaikista populaation yksilöistä. Seuraavissa kappaleissa sukupuuta ja sukupuuhavaintoja käsitellään täsmällisemmin.

Määritellään sukupuu parina $\mathcal{S} = (\mathcal{V}, \mathcal{E})$, missä $\mathcal{V}$ on populaation osajoukko ja $\mathcal{E}$ on kokoelma joukon $\mathcal{V}$ yksilöiden ominaisuuksia ja sukulaisuussuhteita kuvaavia relaatioita. Yleensä relaatioilla määritellään ainakin jokaisen yksilön sukupuoli, biologinen äiti ja isä sekä kahden yksilön identtinen sisaruus. Relaatiot noudattavat luonnollisia rajoituksia, joiden tarkempi määritteleminen sivuutetaan tässä. Lisärajoituksen antaa sopimus, jonka mukaan jokaisella sukupuun yksilöllä on sukupuussa joko molemmat vanhemmat tai ei kumpaakaan. Sukupuun *perustajiksi* (engl. founder) kutsutaan niitä yksilöitä, joiden vanhempia ei ole sukupuussa. Sukupuun perustajien joukosta käytetään merkintää $\mathcal{F}$. Yleensä pyritään siihen, että perustajat ovat toisilleen kaukaista sukua, tai että heillä ei ole populaatiossa yhteisiä esivanhempia. Tällöin sukupuu kuvaa yksilöidensä oleelliset (läheisimmät) sukulaisuussuhteet. Teknisen mukavuuden vuoksi yksilöt numeroidaan luonnollisilla luvuilla ja yksilöiden joukko $\mathcal{V}$ samaistetaan joukon $\{1, 2, \dots, |\mathcal{V}|\}$ kanssa.

Havainnot sukupuun yksilöiden ilmiasuista voitaisiin esittää joukon $\mathcal{V}$ relaatioina. Otetaan kuitenkin käyttöön erilainen esitys. Sukupuun havainnot ilmaistaan funktiona $x : \mathcal{U} \to \{0, 1\}$, missä $\mathcal{U}$ on sukupuun havainnoitujen yksilöiden joukko. Jos yksilö u on sairas, asetetetaan $x_u = 1$, muuten asetetaan $x_u = 0$. Tarpeen mukaan funktiota x käsitellään jonona $(x_{u_1}, x_{u_2}, \dots, x_{u_n})$, kun $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$. Jos havainnoitujen yksilöiden joukko halutaan merkitä näkyviin, käytetään merkintää $x_{\mathcal{U}}$. Vastaava joukolla indeksoiminen otetaan käyttöön yleisemminkin.

Esimerkki 2.1. Kuva 2.1 esittää kahdeksan yksilön sukupuuta. Havainnoidut yksilöt muodostavat joukon $\mathcal{U} = \{2, 4, 5, 6, 7, 8\}$. Havainnot voidaan kirjoittaa jonona $x = (x_2, x_4, x_5, x_6, x_7, x_8) = (1, 0, 1, 0, 1, 1)$. □



Kuva 2.1: Kahdeksan yksilön sukupuu, jossa on kuusi havainnoitua yksilöä. Yksilöt 4 ja 6 on havaittu terveiksi ja yksilöt 2, 5, 7 ja 8 sairaiksi. Yksilöiden 1 ja 3 ilmiasut ovat tuntemattomia.

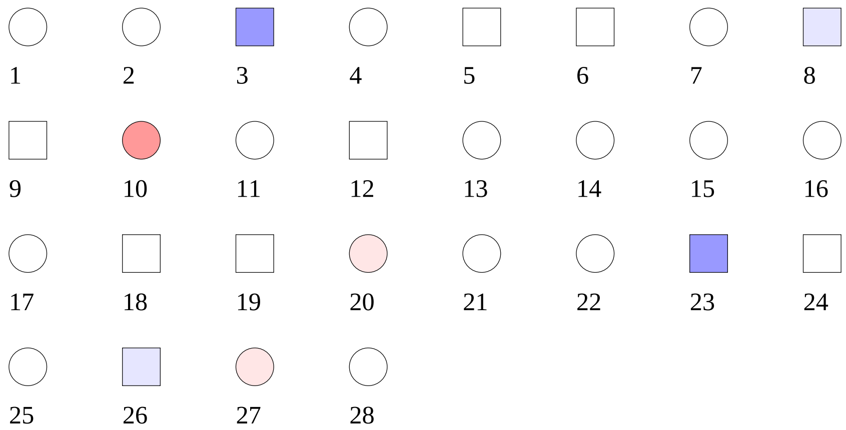
Sukupuuhavaintojen käyttöön perinnöllisten tautien mallintamisessa liittyy ongelmia. Sukupuiden muodostaminen on työlästä, minkä vuoksi havainnot voivat jäädä niukoiksi. Suuren ja monimutkaisen sukupuun analysointi voi puolestaan olla laskennallisesti vaativaa Seuraavissa osioissa tarkastellaan yksinkertaisempia suureita, jotka mittaavat taudin esiintyvyyttä ja periytyvyyttä populaatiossa.

2.2 Taudin yleisyys

Keskeinen tautia kuvaava suure on taudin *yleisyys* populaatiossa (engl, population prevalence). Se määritellään sairaiden yksilöiden osuutena koko populaatiosta. Tässä yhteydessä populaatiolla tarkoitetaan yleensä väestön elossa olevia tai nuorimpien sukupolvien jäseniä.

Taudin yleisyyden arvioiminen kuvan 2.1 kaltaisista sukupuista voi helposti tuottaa virhearvion. Syynä on paitsi pieni yksilömäärä, myös havainnon valikoituminen niin, että sukupuussa on sairaita yksilöitä suhteessa enemmän kuin populaatiossa keskimäärin. Vaikka sukupuu valittaisiin populaatiosta satunnaisesti (riippumatta yksilöiden ilmiasuista), pienessä sukupuussa joko sairaiden tai terveiden osuus korostuu.

Ongelmat vältetään valitsemalla yksilöitä satunnaisesti ja toisistaan riippumatta. Jos populaatio on suuri, valitut yksilöt voidaan ajatella ilmiasultaan riippumattomiksi. He muodostavat tällöin kuvan 2.2 kaltaisen sukupuun, jossa kaikki yksilöt ovat sukupuun perustajia. Valittujen yksilöiden oletetaan olevan kaukaista sukua toisilleen. Kuvan 2.2 havaintoihin perustuva arvio taudin yleisyydelle on $3/(21+3) = 0.125$. Yleisesti havainto taudin yleisyydestä esitetään lukuparina (N, N_1), missä N on kaikkien havainnoitujen lukumäärä ja N_1 on sairaiksi havaittujen lukumäärä.



Kuva 2.2: Sukupuu, jossa kaikki 28 yksilöä ovat perustajia. 3 yksilöä on havaittu sairaiksi ja 21 yksilöä terveiksi.

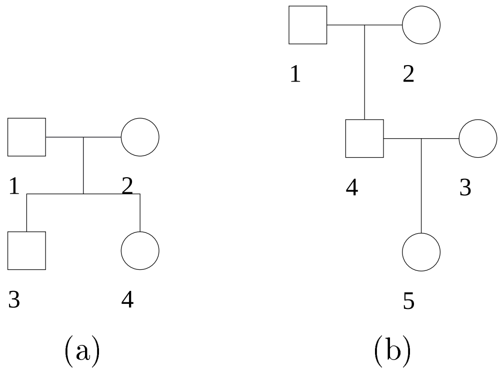
2.3 Sukulaisuus ja sukulaisriskit

Kahden yksilön välinen sukulaisuussuhde määritellään luontevimmin populaatiota kuvaavan sukupuun avulla. Koska koko populaatiota kuvaavaa sukupuuta ei yleensä tunneta, on sukulaisuussuhteet järkevää määritellä ehdollisena annetulle sukupuulle. Seuraavassa tarkastellaan mielivaltaista sukupuuta ja sen kahta yksilöä u ja v.

Sukulaisuuden määritelmä on yksinkertainen: yksilöt u ja v ovat sukulaisia, jos heillä on sukupuussa vähintään yksi yhteinen esivanhempi (yksilön esivanhemmaksi hyväksytään myös yksilö itse). Yksilöiden sukulaisuussuhteet määräytyvät sukupuun relaatioiden perusteella. Yleisesti sanotaan, että yksilö v on tyypin R sukulainen yksilölle u, jos sukupuun yksilöitä u ja v yhdistävä osa on tyypin R mukainen (tällaista osaa voitaisiin kutsua esimerkiksi alisukupuuksi). Usein on järkevää tarkastella sukulaisuussuhteita unohtaen yksilöiden sukupuoli.

Sukulaisuussuhteen muodollisen määrittelyn sijaan valotetaan ajatusta esimerkein. Yksilö v on yksilölle u sukulainen tyyppiä *sisar*, jos heillä on yhteiset vanhemmat, jotka eivät ole keskenään sukua. Kuvassa 2.3.a on esitetty sisaruuden määrittelevä sukupuu. Sukulaisuustyyppiä *isänäiti* vastaava sukupuu on kuvassa 2.3.b. Yksilö v on yksilön u isänäiti, jos yksilö v on yksilön u isän äiti, ja jos lisäksi yksilön u äiti ja yksilön u isän vanhemmat eivät ole keskenään sukua. Huomautettakoon, että luonnollinen kieli antaa nimet vain tavallisimmin tarkasteltaville sukulaisuustyypeille (esim. sisar, lapsi, lapsenlapsi, sisarpuoli, serkku, pikkuserkku). Luonnollisen kielen käsitteet eivät ole tarkkoja. Esimerkiksi sisareksi voidaan sanoa yhtä hyvin sellaisia yksilöitä, joiden vanhemmat ovat toisilleen kaukaista sukua kuin myös yksilöitä, joiden vanhemmat ovat toisilleen pikkuserkkuja. Kuvassa 2.3.a annettu määritelmä sopii ensin mainittuun; jälkimmäisen sukulaisuustyypin määrittelevässä sukupuussa olisi viisi sukupolvea.

Sukulaisriskit ovat taudin periytymistä kuvaavia tunnuslukuja, joiden määrittelemiseksi riittää tarkastella populaation yksilöpareja. *Sukulaisriski populaatiossa tyypin R sukulaiselle* määritellään seuraavasti: Tarkastellaan koko populaation kattavan sukupuun sellaisia yksilöpareja (u, v), joille v on tyypin R sukulainen yksilölle u. Olkoon n_0^R niiden yksilöparien (u, v) lukumäärä, joille u on sairas ja v terve. Olkoon vastaavasti n_1^R niiden yksilöparien (u, v) lukumäärä, joille sekä u että v ovat sairaita. Sukulaisriski tyypin R sukulaiselle



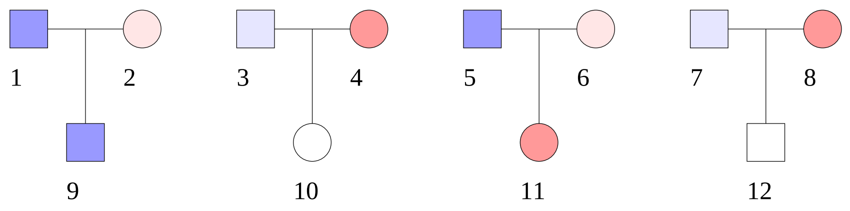
Kuva 2.3: Sisaruuden (a) ja isovanhemmuuden (b) määrittelevät sukupuut. Sukupuussa (a) yksilö 4 on yksilölle 3 sukulainen tyyppiä *sisar*. Sukupuussa (b) yksilö 2 on yksilölle 5 sukulainen tyyppiä *isänäiti*, kun yksilöiden sukupuoli huomioidaan.

määritellään osamääränä $n_1^R/(n_0^R + n_1^R)$. Määritelmä on siten yhdenmukainen taudin yleisyyden määritelmän kanssa.

Sukulaisriskejä voidaan arvioida poimimalla populaatiosta yksilöpareja, joiden sukulaisuussuhde tunnetaan. Yksi tapa on käyttää kuvien 1.1.a ja 1.1.b kaltaisia sukupuuhavaintoja. Virhearvionnin vaara on kuitenkin ilmeinen, jos sukupuussa on vähän yksilöitä. Ongelmat vältetään valitsemalla populaatiosta satunnaisesti sukulaispareja, joiden ensimmäinen jäsen tiedetään sairaaksi. Parit poimitaan siis ehdollisesti (esimerkiksi taudin esiintymistä kartoittavia rekistereitä käyttäen). Havaintoja kuvaava sukupuu koostuu tällöin useista erillisistä alisukupuista. Olkoon N_0^R niiden sukupuiden lukumäärä, joissa sukulaisparin toinen jäsen on havaittu terveeksi, ja olkoon N_1^R niiden sukupuiden lukumäärä, joissa sukulaisparin toinen jäsen on havaittu sairaaksi. Havaintojen mukainen arvio sukulaisriskille on tällöin osamäärä $N_1^R/(N_0^R + N_1^R)$. Havainnoitujen parien yhteismäärästä $N_0^R + N_1^R$ käytetään merkintää N^R.

Esimerkki 2.2. Kuva 2.4 esittää lapsiriskihavaintoja. Havaintojen mukainen arvio lapsiriskille on $2/(2 + 2) = 1/2$. Merkille pantavaa on, että lapsiriskihavainnot eivät erottele vanhempien eivätkä lasten sukupuolta. $\square$

Sukulaisriskihavainnot voidaan kullekin tarkasteltavalle sukulaisuustyypille pelkistää lukupariksi tai jopa yhdeksi luvuksi. Verrattuna suuriin sukupuihin sukulaisriskejä kuvaavien lukujen ja lukuparien raportointi on helppoa, vaikka havaittuja yksilöitä olisi tuhansia. Erityisesti lapsen, sisaren ja identtisen kaksosen havaittuja riskejä on raportoitu ja analysoitu useille taudeille (astma [Lai98], skitsofrenia [Ris90], diabetes [Ric90, Onk99a]). Tässä tutkielmassa



Kuva 2.4: Neljä riippumatonta havaintoa taudin esiintymisestä vanhemmalla ja lapsella.

esitettävän menetelmän kannalta nämä *ydinperheriskeiksi* kutsuttavat suureet ovat erityisasemassa. Sukulaisuustyypistä *sisar* käytetään jatkossa lyhennettä S, tyypistä *lapsi* (tai *jälkeläinen*) lyhennettä O ja tyypistä *identtinen kaksonen* lyhennettä MZ.

Esimerkki 2.3. Taulukon 2.1 esittämät diabeteksen havaitut sukulaisriskit perustuvat lähteeseen [Onk99a]. Koska alkuperäiset havainnot sisältävät enemmän mm. yksilöiden sukupuoleen sidottua tietoa, taulukon luvut ovat pyöristettyjä keskiarvoja todellisista havainnoista. Taulukon luvuista nähdään, että havainnot ovat tarkempia taudin yleisyydelle kuin identtisten kaksosten riskille ja että lapsi- ja sisarusriskihavaintojen tarkkuudet sijoittuvat edellisten välille. Diabetesaineistoa käsitellään lisää luvussa 7. □

N	N_1	N^O	N_1^O	N^S	N_1^S	N^{MZ}	N_1^{MZ}
50 000	250	5 000	265	5 000	500	32	10

Taulukko 2.1: Diabeteksen havaittuja ydinperheriskejä ja taudin yleisyyttä kuvaavat luvut. N tutkitusta yksilöstä N_1 yksilöä on havaittu sairaaksi ja loput $N - N_1$ terveeksi. Merkinnät N^O, N_1^O, N^S, N_1^S ja N^{MZ}, N_1^{MZ} viittaavat vastaavasti jälkeläisriskiin, sisarusriskiin ja identtisen kaksosen riskiin.

Sukulaisriskihavaintojen käyttö on ongelmallista, jos havainnot eivät ole riippumattomia, kuten edellä muotoiltu riippumattomien alisukupuiden käyttö olettaa. Tähän ongelmaan ei syvennytä tässä työssä. Seuraavissa luvuissa tarkastellaan ihanteellista tapausta, jossa otantavirhettä ei ole.

Luku 3

Geneettiset mallit

Edellisessä luvussa 2 taudin yleisyys ja sukulaisriskit määriteltiin sairauden esiintymistä ja periytyvyyttä kuvaavina tunnuslukuina, joiden arviointi perustuu sukupuuhavaintoihin. Havaintoja voidaan selittää sairauden perinnöllisten ja ympäristöstä riippuvien tekijöiden avulla. Tässä luvussa tarkastellaan geneettisiä malleja, jotka pyrkivät kuvaamaan sairauden taustalla olevat perintötekijät, niiden periytymismekanismin ja yhteyden taudin ilmenemiseen. Samalla rajataan joukko seuraavien lukujen kannalta kiinnostavia malliperheitä.

3.1 Mallintamisen lähtökohta

Taudin geneettinen malli rajoittuu perintötekijöiden ja ilmiasun yhteyden stokastiseen eli todennäköisyyksiin perustuvaan mallintamiseen. Mallilla ei pyritä kuvaamaan koko perimää vaan ainoastaan joitakin taudin kannalta oleellisia geenejä. Usein geenejä ja näiden sijaintia kromosomistossa ei tiedetä, ja siksi perintötekijöinä tarkastellaan muutamia hypoteettisia geenejä, jotka yhdessä muodostavat yksilön genotyypin.

Kun tarkastellaan populaatiosta satunnaisesti valittua sukupuuta $\mathcal{S} = (\mathcal{V}, \mathcal{E})$, on sukupuun yksilöiden ilmiasuja ja genotyyppejä luontevaa tarkastella satunnaismuuttujina. Jokaiseen sukupuun yksilöön $u \in \mathcal{V}$ liitetään 0/1 -arvoinen satunnaismuuttuja X_u. Muuttuja saa arvon 1, jos u on sairas, ja arvon 0, jos hän on terve. Edellisessä luvussa 4 määritelty havainto x_u on siten satunnaismuuttujan X_u realisaatio. Yksilön u genotyyppi on myös satunnaismuuttuja, merkitään H^u. Genotyyppien arvojoukko on yleensä sukupuun yksilöille yhteinen koostuen hypoteettisten geenien kaikista mahdollisista yhdistelmistä.

Geneettisen mallin tehtävä on selittää havainnot. Ensimmäinen vaatimus on, että malli määrittelee perintötekijöiden rakenteen eli satunnaismuuttujien H^u arvojoukon. Lisäksi mallin on kuvattava genotyyppien yhteys taudin ilmenemiseen ja mahdollistettava mallin hyvyyden tilastollinen koeteltavuus. Tätä varten riittää, jos malli määrittelee ilmiasujen ja genotyyppien yhteistodennäköisyysjakauman $P(X_\mathcal{V}, H^\mathcal{V})$. Tässä on syytä huomauttaa, että geneettisen mallin tehtävä ei ole mallintaa sukupuun $\mathcal{S}$ valintaa populaatiosta.

Elston ja Stewart [ElS71] esittivät sukupuiden analysointiin soveltuvan yleisen mallin, jossa jakauma $P(X_\mathcal{V}, H^\mathcal{V})$ muodostetaan kolmeen komponenttiin perustuvana tulona. Oleellinen tulomuoto johdetaan seuraavaksi ketjusääntöä ja riippumattomuusoletuksia käyttäen.

Ketjusäännön nojalla

$$P(X_\mathcal{V}, H^\mathcal{V}) = P(X_\mathcal{V} \mid H^\mathcal{V})\, P(H^\mathcal{V}),$$

missä tulon ensimmäinen tekijä kuvaa genotyyppien vaikutuksen taudin ilmenemiseen. Toinen tekijä ilmaisee genotyyppien yhteisjakauman sukupuun yksilöillä. Olettamalla, että yksilön ilmiasu on riippumaton muiden yksilöiden genotyypeistä ja ilmiasuista, kun yksilön genotyyppi tunnetaan, voidaan kirjoittaa

$$P(X_\mathcal{V} \mid H^\mathcal{V}) = \prod_{u \in \mathcal{V}} P(X_u \mid H^u). \tag{3.1}$$

Soveltamalla ketjusääntöä genotyyppien yhteisjakaumaan saadaan

$$P(H^\mathcal{V}) = P(H^\mathcal{F})\, P(H^{\mathcal{V} \setminus \mathcal{F}} \mid H^\mathcal{F}).$$

Tulon ensimmäinen termi kertoo genotyyppien jakauman perustajilla ja jälkimmäinen genotyyppien siroamisen (engl. segregation) perustajilta heidän jälkeläisilleen. Jos perustajien genotyypit oletetaan kaukaisen sukulaisuuden vuoksi toisistaan riippumattomiksi, voidaan kirjoittaa

$$P(H^\mathcal{F}) = \prod_{u \in \mathcal{F}} P(H^u). \tag{3.2}$$

Kun genotyyppien periytymistapahtumat vanhemmilta lapsille oletetaan ehdollisesti riippumattomiksi annettuna vanhempien genotyypit, saadaan

$$P(H^{\mathcal{V} \setminus \mathcal{F}} \mid H^\mathcal{F}) = \prod_{u \in \mathcal{V} \setminus \mathcal{F}} P(H^u \mid H^{F(u)}, H^{M(u)}), \tag{3.3}$$

missä $F(u)$ tarkoittaa yksilön u isää ja $M(u)$ yksilön u äitiä. Jos sukupuussa on identtisiä kaksosia (tai kolmosia jne.), niin periytymistapahtumat eivät ole edellisessä mielessä riippumattomia. Tällöin tuloon (3.3) otetaan mukaan vain yksi kustakin keskenään identtisten sisarusten joukosta.

Elstonin ja Stewartin mallin komponentit ovat tulojen (3.1), (3.2) ja (3.3) jakaumat. Kun jakaumat oletetaan yksilöistä riippumattomiksi, geneettisen mallin tehtäväksi jää määritellä kolme jakaumaa: genotyyppien vaikutus taudin ilmenemiseen ($P(X_u \mid H^u)$), genotyyppien frekvenssit sukupuun perustajilla ($P(H^u)$) ja genotyyppien periytyminen vanhemmilta lapsille ($P(H^u \mid H^{F(u)}, H^{M(u)})$).

Osioissa 3.3–3.5 tarkastellaan kutakin mallin komponenttia erikseen. Samalla esitellään joukko riippumattomuusoletuksia, jotka yksinkertaistavat yleistä mallia. Lisäksi esitellään muutamia myöhemmin tarvittavia merkintöjä.

3.2 Genotyypit

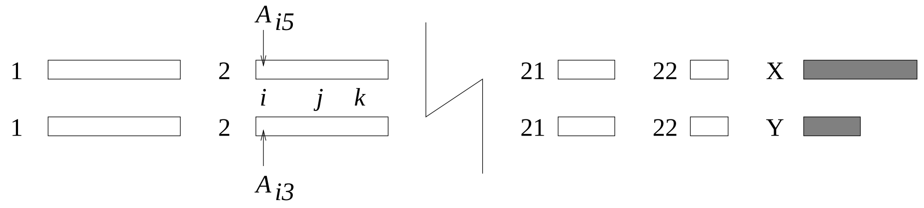
Geneettinen malli määrittelee perintötekijöille hypoteettisen rakenteen kiinnittämällä geenipaikkojen eli *lokusten* ja lokusten geenimuunnelmien eli *alleelien* lukumäärän. Merkitään lokusten lukumäärää kirjaimella ℓ, ja numeroidaan lokukset luvuilla $1, 2, \ldots, \ell$. Lokusten numeroinnin sovitaan vastaavan niiden fyysistä sijaintia kromosomistossa seuraavassa mielessä: jos i, j ja k ovat saman kromosomin lokuksia, joille $i < j < k$, niin lokus j on lokusten i ja k välissä (ks. kuva 3.1). Lokusten oletetaan olevan autosomeja (eli eivät sijaitse sukupuolikromosomeissa X ja Y).

Lokuksen i alleelien joukko olkoon

$$\mathcal{A}_i = \{A_{ij} : 1 \leq j \leq m_i\},$$

missä m_i on alleelien lukumäärä lokuksessa i. Joukon $\mathcal{A}_i$ alkiot ovat niitä alleeleja, joita mallin mukaan populaation yksilöillä voi esiintyä lokuksessa i. Kuva 3.1 havainnollistaa tilannetta. Termin 'geeni' monikäsitteisyyden vuoksi jäljempänä käytetään vain termejä 'alleeli' ja 'lokus' (vrt. [Sha98, s. 11]).

Tavallisesti oletetaan ([Smi71, DBC93]), että tautiin vaikuttavia lokuksia on vain muutamia ja että lokuksissa on vain muutama alleeli. Usein lokusten oletetaan olevan jopa kaksialleelisia. Tällöin ajatuksena on, että taudin aiheuttaa



Kuva 3.1: Kaaviokuva miehen kromosomistosta. 23 parin ylemmät vastinkromosomit yksilö on perinyt isältään ja alemmat äidiltään. Lokukseen i kromosomien kantaja on perinyt isältään alleelin A_{i5} ja äidiltään alleelin A_{i3}.

yksi mutatoitunut alleeli muiden alleelien ollessa "terveitä". Mutatoitunutta alleelia merkitään yleensä isolla kirjaimella $(A, B, C, \dots)$ ja terveitä alleeleja, tai paremminkin näiden luokkaa, merkitään pienellä kirjaimella $(a, b, c, \dots)$.

Genotyypit ovat alleelien järjestämättömiä tai järjestettyjä yhdistelmiä. Genotyyppien täsmällistä käsittelyä varten tukeudutaan joukko-opillisiin määritelmiin. Lokuksen i järjestämättömien genotyyppien joukko on

$$\mathcal{G}_i = \big\{\{A_{ij}, A_{ik}\} : A_{ij}, A_{ik} \in \mathcal{A}_i\big\}$$

Järjestettyjen genotyyppien joukko on

$$\mathcal{H}_i = \big\{(A_{ij}, A_{ik}) : A_{ij}, A_{ik} \in \mathcal{A}_i\big\},$$

missä alleeliparin ensimmäinen jäsen tulkitaan isältä perityksi ja toinen jäsen äidiltä perityksi. Järjestämättömien ja järjestettyjen genotyyppien joukot yli kaikkien lokusten määritellään lokuksittaisten joukkojen karteesisina tuloina

$$\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_\ell$$

ja

$$\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_\ell.$$

Järjestetyt genotyypit sisältävät perimästä täydellisemmän tiedon kuin järjestämättömät genotyypit. Aina ei ole kuitenkaan tarpeellista erottaa isältä ja äidiltä perittyjä alleeleja. Järjestämättömien genotyyppien etu on niiden pienempi määrä. Seuraava esimerkki havainnollistaa joukkojen $\mathcal{H}$ ja $\mathcal{G}$ kokojen suhdetta.

Esimerkki 3.1. Olkoon lokuksia ℓ, ja olkoon jokaisessa lokuksessa m alleelia. Järjestettyjen genotyyppien lukumäärä on tällöin $|\mathcal{H}| = (m^2)^\ell = m^{2\,\ell}$ ja järjestämättömien genotyyppien lukumäärä on $|\mathcal{G}| = (\frac{1}{2}\, m\,(m+1))^\ell$. Nähdään, että lukumäärät kasvavat ekponentiaalisesti suhteessa lokusten lukumäärään ja polynomisesti suhteessa alleelien lukumäärään. Merkille pantavaa on, että niin alleelien kuin lokustenkin lukumäärän kasvaessa ero joukkojen $\mathcal{H}$ ja $\mathcal{G}$ suuruuksissa kasvaa: Jos $m = \ell = 2$, niin $|\mathcal{H}| = 4^2 = 16$ ja $|\mathcal{G}| = 3^2 = 9$. Jos $m = \ell = 3$, niin $|\mathcal{H}| = 9^3 = 729$ ja $|\mathcal{G}| = 6^3 = 216$. □

Otetaan sitten käyttöön joukko merkintöjä, jotka myöhemmin havaitaan käytännöllisiksi. Olkoon

$$h = \big((A_{1j_1}, A_{1k_1}), \ldots, (A_{\ell j_\ell}, A_{\ell k_\ell})\big) \in \mathcal{H}. \tag{3.4}$$

Tällöin merkinnällä h_i tarkoitetaan alleeliparia (A_{ij_i}, A_{ik_i}). Alleelijonosta $(A_{1j_1}, \ldots, A_{\ell j_\ell})$ käytetään merkintää $h_{\cdot 1}$ ja alleelijonosta $(A_{1k_1}, \ldots, A_{\ell k_\ell})$ merkintää $h_{\cdot 2}$. Merkinnällä h_{is}, tarkoitetaan alleelia A_{ij_i}, jos $s = 1$, ja alleelia A_{i,k_i}, jos $s = 2$. Järjestämättömille genotyypeille käytetään vastaavia merkintöjä kuin järjestetyille genotyypeille. Olkoon

$$g = \big(\{A_{1j_1}, A_{1k_1}\}, \ldots, \{A_{\ell j_\ell}, A_{\ell k_\ell}\}\big) \in \mathcal{G}, \tag{3.5}$$

missä $j_i \leq k_i$ jokaiselle lokukselle i. Tällöin merkinnällä g_i tarkoitetaan alleelijoukkoa $\{A_{ij_i}, A_{ik_i}\}$. Merkintä g_{i1} tarkoittaa nyt alleelia A_{ij_i} ja g_{i2} alleelia A_{ik_i}. Tässä on hyvä huomata erikoistapaus $A_{ij_i} = A_{ik_i}$, jolloin $g_{i1} = g_{i2}$. Edelleen jonoista $(g_{1s}, \ldots, g_{\ell s})$ käytetään merkintää $g_{\cdot s}$.

Lokuksen i järjestettyä genotyyppiä h_i sanotaan *homotsygootiksi*, jos $h_{i1} = h_{i2}$, ja *heterotsygootiksi*, jos $h_{i1} \neq h_{i2}$. Termejä käytetään vastaavasti myös järjestämättömistä genotyypeistä.

Lähes vakiintuneen tavan mukaan (esim. [Lan97, Sha98]) järjestettyjä ja järjestämättömiä genotyyppejä merkitään jättämällä sulut ja pilkut pois. Tällöin järjestetty genotyyppi (3.4) kirjoitetaan muodossa

$$A_{1j_1} A_{2j_2} \ldots A_{\ell j_\ell} / A_{1k_1} A_{2k_2} \ldots A_{\ell k_\ell}$$

ja järjestämätön genotyyppi (3.5) muodossa

$$A_{1j_1} A_{1k_1} A_{2j_2} A_{2k_2} \ldots A_{\ell j_\ell} A_{\ell k_\ell}.$$

Seuraava esimerkki valottaa edellä esitettyjä käsitteitä ja merkintöjä.

Esimerkki 3.2. Kahden lokuksen kaksialleelisen mallin 16 järjestettyä genotyyppiä ovat AB/AB, AB/Ab, Ab/AB, Ab/Ab, ..., aB/ab, ab/aB ja ab/ab. Järjestämättömiä genotyyppejä on yhdeksän: $AABB$, $AABb$, $AAbb$, $AaBB$, $AaBb$, $Aabb$, $aaBB$, $aaBb$ ja $aabb$. Jos $h = Ab/aB$, niin $h_{\cdot 1}$ tarkoittaa alleelijonoa Ab, ja $h_{\cdot 2}$ tarkoittaa vastaavasti alleelijonoa aB. Genotyyppi h voidaan hajottaa myös lokuksittaisiin genotyyppeihin $h_1 = A/a$ ja $h_2 = b/B$. □

Populaatiosta satunnaisesti valitun yksilön u alleelit ja genotyypit ovat satunnaismuuttujia. Edellisessä osiossa esitelty muuttuja H^u asetetaan nyt vastaamaan yksilön u järjestettyä genotyyppiä. Järjestämättömästä genotyypistä käytetään merkintää G^u. Alaindeksejä käytetään samoin kuin satunnaismuuttujien realisaatioille. Siis esimerkiksi $H^u_{\cdot 2}$ tarkoittaa yksilön u äidiltään perimää alleelijonoa.

3.3 Genotyyppien ilmenemisyleisyydet

Perintötekijät liitetään taudin ilmiasuun todennäköisyysjakauman avulla. Edellä tehtyjen rajoitusten mukaisesti tarkastellaan geneettisiä malleja, joissa perintötekijät vastaavat genotyyppejä kiinnitetyssä joukossa ($\mathcal{H}$ tai $\mathcal{G}$) ja taudin ilmiasu on joko 'sairas' (arvo 1) tai 'terve' (arvo 0).

Genotyyppien vaikutus taudin ilmenemiseen, eli kaavan (3.1) komponentti $P(X_u \mid H^u)$, määritellään geneettisessä mallissa genotyyppien *ilmenemisyleisyyksillä* (engl. penetrance). Genotyypin $h \in \mathcal{H}$ ilmenemisyleisyyttä merkitään $f(h)$. Yhteys taudin ilmiasuun määritellään asettamalla

$$P(X_u = 1 \mid H^u = h) \quad = \quad f(h),$$

missä u on populaatiosta satunnaisesti valittu yksilö. Yksilön ilmiasu oletetaan riippumattomaksi sukupuun muiden yksilöiden ilmiasuista, jos hänen genotyyppinsä tunnetaan. Tämä oletus sulkee pois esimerkiksi lähisukulaisten yhtäläisten ympäristötekijöiden mallintamisen.

Usein oletetaan, että $f(h) = f(h')$, jos järjestetyt genotyypit h ja h' sisältävät samat alleelit [Sha98, s. 81]. Tällöin riittää määritellä ilmenemisyleisyys järjestämättömille genotyypeille $g \in \mathcal{G}$:

$$P(X_u = 1 \mid G^u = g) \quad = \quad f(g).$$

Jäljempänä funktion f argumenteiksi voidaan kirjoittaa sekä järjestettyjä että järjestämättömiä genotyyppejä.

Kun genotyypeille sallitaan toisistaan riippumattomat ilmenemisyleisyydet, voidaan lokusten mielivaltaiset yhteisvaikutukset kuvata. Ongelmia aiheuttaa se, että mallin tarvitsemien parametrien määrä kasvaa nopeasti lokusten määrän kasvaessa. Biologisen tulkinnan löytäminen erikoisille yhteisvaikutuksille voi myös olla vaikeaa.

Ongelmat ratkaistaan usein esittämällä genotyyppien ilmenemisyleisyydet lokuskohtaisten vaikutusten avulla. Ajatuksena on tällöin, että malli määrittelee jokaiselle lokukselle i funktion $f_i : \mathcal{G} \to [0, 1]$. Genotyypin $g \in \mathcal{G}$ ilmenemisyleisyys muodostetaan lokuskohtaisten vaikutusten $f_i(g_i)$ sopivana yhdistelmänä. Risch [Ris90] käsitteli kolmea ilmenemisyleisyyksien mallia: *multiplikatiivisessa* mallissa $f(g)$ on tulo $\prod_i f_i(g_i)$, *additiivisessa* mallissa summa $\sum_i f_i(g_i)$ ja *geneettisen heterogeenisuuden* mallissa komplementtivaikutusten tuloon perustuen $1 - \prod_i(1 - f_i(g_i))$.

Tällaisten esitysten etu on, että malli tarvitsee ilmenemisyleisyyksien määrittelemiseen ainoastaan $\sum_i |\mathcal{G}_i|$ parametria. Mallien laskennallinen käsittely helpottuu, ja yhteisvaikutusten biologisessa tulkinnassa voidaan noudattaa esityksen mukaista logiikkaa [Ris90, Gor92]. Heikkoutena on mallien kuvausvoiman voimakas väheneminen. Ei ole syytä olettaa, että yhteisvaikutukset noudattavat jotakin edellä esitettyä, tai niiden kaltaista, yksinkertaista kaavaa.

3.4 Genotyyppien frekvenssit

Genotyyppien esiintymisyleisyyksiä populaatiofrekvenssejä kuvaa Elstonin ja Stewartin mallin komponentti $P(H^u)$. Yleisimmillään geneettinen malli tarvitsee $|\mathcal{H}| - 1$ parametria tämän jakauman määrittelemiseksi. Seuraavaksi esitettävät riippumattomuusoletukset mahdollistavat genotyyppien frekvenssien kuvaamisen ainoastaan $\sum_i(|\mathcal{A}_i| - 1)$ parametrilla.

Hardyn ja Weinbergin tasapaino: Olkoon alleelin A_{ij} frekvenssi populaatiossa p_{ij}. Tällöin satunnaisesti valitun yksilön genotyyppi on A_{ij}/A_{ik} todennäköisyydellä $p_{ij}\, p_{ik}$.

Kytkentätasapaino: Olkoon alleelin A_{ij} frekvenssi populaatiossa p_{ij}. Tällöin satunnaisesti valitun yksilön isältään – samoin äidiltään – perimä alleelijono on $A_{1j_1} \ldots A_{\ell j_\ell}$ todennäköisyydellä $\prod_{i=1}^{\ell} p_{ij_i}$.

Yllä esitetyt oletukset eivät yleensä ole tarkalleen ottaen voimassa. Voidaan kuitenkin osoittaa, että tietyillä populaatiota koskevilla oletuksilla Hardyn ja Weinbergin tasapaino saavutetaan jo yhdessä sukupolvessa ja kytkentätasapaino asymptoottisesti sukupolvien määrän kasvaessa [Lan97, s. 4, 8], [Bul80, s. 34–38]. Tässä on hyvä panna merkille, että esimerkiksi Hardyn ja Weinbergin tasapaino vaatii oletuksen äärettömästä populaatiosta. Äärellisessä populaatiossa homotsygootin genotyypin frekvenssin pitäisi olla vähemmän kuin alleelifrekvenssien tulo, koska otanta tehdään takaisinpanotta. Periaate oli käytössä jo aiemmin esitettyssä Elstonin ja Stewartin komponenttijaossa: sukupuun perustajien genotyyppien oletettiin olevan riippumattomia.

Merkitään nyt todennäköisyyttä, että sukupuun perustajalla on genotyyppi $h \in \mathcal{H}$, lyhyesti $p(h)$. Merkintätapaa käytetään myöhemmin myös muille tarkasteltaville satunnaismuuttujille ja niiden arvoille, mikäli ei ole sekaannuksen vaaraa.

Tasapaino-oletuksista seuraa, että järjestettyjen genotyyppien frekvenssit voidaan esittää alleelifrekvenssien tulona:

$$p(h) = \prod_{i=1}^{\ell} p(h_i) = \prod_{i=1}^{\ell} p(h_{i1})\, p(h_{i2}).$$

Myös järjestämättömien genotyyppien frekvenssit voidaan kirjoittaa tulomuodossa:

$$p(g) = \prod_{i=1}^{\ell} p(g_i) = \prod_{i=1}^{\ell} |g_i|\, p(g_{i1})\, p(g_{i2}). \tag{3.6}$$

Tässä lienee syytä muistuttaa, että g_i on joukko $\{g_{i1}, g_{i2}\}$, jonka koko $|g_i|$ on 1, jos g_i on homotsygootti. Jos g_i on heterotsygootti, sen koko on 2.

3.5 Genotyyppien periytyminen

Yleisimmillään geneettinen malli sallii periytymistodennäköisyyksille $p(h^u \mid h^{F(u)}, h^{M(u)})$ lähes mielivaltaisia arvoja (kunhan ne toteuttavat ehdollisten todennäköisyysjakaumien vaatimukset). Näin mahdollistetaan alleelimutaatioiden, monimutkaisten periytymismekanismien ja joidenkin valinnan (engl. selection) muotojen mallintaminen. Tässä työssä edellä mainittujen piirteiden kuvaamiseen ei perehdytä. Yleisimmillään mallin voidaan ajatella tarvitsevan lähes $|\mathcal{H}|^3$ periytymisparametria.

Seuraavaksi esitellään muutama keskeinen oletus, jotka vähentävät tarvittavien parametrien määrää huomattavasti ja mahdollistavat periytymistodennäköisyyksien yksinkertaisen laskemisen. Havainnollisuuden vuoksi kahdessa jälkimmäisessä oletuksessa tarkastellaan vain perimistä isältä, vaikka oletukset koskevatkin yhtälailla perimistä äidiltä.

Riippumaton periytyminen: Lapsi perii alleelijonon $h^u_{\cdot 1}$ isältään ja alleelijonon $h^u_{\cdot 2}$ äidiltään riippumattomasti. Tällöin on voimassa

$$p(h^u \mid h^{M(u)}, h^{F(u)}) \quad = \quad p(h^u_{\cdot 1} \mid h^{F(u)})\, p(h^u_{\cdot 2} \mid h^{M(u)}).$$

Tasainen periytyminen: Yksilö siirtää jälkeläiselleen yhtä suurella todennäköisyydellä isältään perimän alleelin kuin äidiltään perimän alleelin. Tästä seuraa, että jokaiselle lokukselle i periytymistodennäköisyys $p(h^u_{i1} \mid h^{F(u)}_i)$ on joko 1, $\frac{1}{2}$ tai 0. Todennäköisyys on 1, jos isän kumpikin alleeli on sama kuin lapsen alleeli h^u_{i1}. Jos vain toinen alleeli on sama, todennäköisyys on $\frac{1}{2}$. Jos kumpikaan alleeleista ei ole sama kuin lapsen, periytymistodennäköisyys on 0. Esimerkiksi $p(A \mid A/a) = \frac{1}{2}$. Huomataan, että tämä oletus on vahvempi kuin oletus riippumattomasta periytymisestä.

Kytkemätön periytyminen: Lokukset sijaitsevat joko eri kromosomeissa tai niin kaukana toisistaan, että niiden alleelit periytyvät riippumattomasti. Periytymistodennäköisyydet voidaan kirjoittaa tulona:

$$p(h^u_{\cdot 1} \mid h^{F(u)}) \quad = \quad \prod_{i=1}^{\ell} p(h^u_{i1} \mid h^{F(u)}_i).$$

Kaksi ensimmäistä oletusta voidaan tulkita *Mendelin ensimmäisen säännön* muotoiluksi ja kolmas oletus *Mendelin toisen säännön* muotoiluksi (vrt. [Tir93, s. 311]). Todettakoon, että mendelistisestä periytymisestä puhutaan joskus väljemmin (vrt. [Sha98, s. 21] ja [Lan97, Ch. 7]). Merkille pantavaa on, että tekemällä kaikki yllä esitetyt oletukset malli ei tarvitse yhtään parametria periytymistodennäköisyyksien määrittelemiseen. Jokainen periytymistodennäköisyys $p(h^u_{\cdot 1} \mid h^{F(u)})$ on joko 0 tai muotoa $(\frac{1}{2})^k$, missä $0 \leq k \leq \ell$.

Jos lokukset ovat kytkeytyneitä, lokusten alleelit eivät periydy riippumattomasti vanhemmalta lapselle. Kahdessa kytkeytyneessä lokuksessa vanhempi siirtää lapselleen todennäköisemmin samalta isovanhemmalta peräisin olevan alleeliparin kuin parin, jossa toinen alleeli on peräisin isoisältä ja toinen isoäidiltä. Jälkimmäinen on mahdollista vain, jos tapahtuu *tekijänvaihto* (engl. recombination). Kytkentä mallinnetaan yleensä Haldanen mallin [Hal19] mukaisesti. Tekijänvaihtojen oletetaan silloin olevan riippumattomia erillissä lokusväleillä.

Käytetään todennäköisyydestä, että lokusten i ja j välillä tapahtuu tekijänvaihto, merkitää θ_{ij}. Vierekkäisten lokusten $i-1$ ja i tekijänvaihtotodennäköisyydelle käytetään lyhennysmerkintää θ_i, kun $1 < i \leq \ell$. Teknisen mukavuuden vuoksi sovitaan lisäksi, että $\theta_{0i} = \theta_1 = \frac{1}{2}$ kaikilla i.

Tekijänvaihtotodennäköisyydet voidaan kirjoittaa vierekkäisten lokusten tekijänvaihtotodennäköisyyksien avulla. Helpolla induktiolla voidaan osoittaa, että Haldanen mallissa pätee Trow'n kaava [Gor92]

$$1 - 2\,\theta_{ij} \;=\; \prod_{k=i+1}^{j} (1 - 2\,\theta_k). \tag{3.7}$$

Malli tarvitsee siten $\ell - 1$ parametria tekijänvaihtotodennäköisyyksien ilmaisemiseen. Seuraava esimerkki havainnollistaa periytymistodennäköisyyksien muodostamista, kun lokukset ovat kytkeytyneitä.

Esimerkki 3.3. Olkoon lokusten 1, 2 ja 3 alleelit vastavasti A ja a, B ja b sekä C ja c. Olkoon isän järjestetty genotyyppi $h^{F(u)} = ABc/aBC$. Lasketaan todennäköisyys sille, että lapsen isältään perimä alleelijono on $h^u_{\cdot 1} = ABC$, kun periytymisen oletetaan olevan tasaista, mutta sallitaan Haldanen mallin mukainen kytkentä lokusten välillä.

Huomataan, että todennäköisyydellä $\frac{1}{2}$ lapsi perii alleelin A. Jotta lapsi perisi myös alleelin C, täytyy lokusten 1 ja 3 välillä tapahtua tekijänvaihto. Ei sen sijaan ole väliä, tapahtuuko tekijänvahto lokusten 1 ja 2 vai lokusten 2 ja 3 välillä. Kirjoitettuna

$$\begin{aligned} p(h^u_{\cdot 1} \mid h^{F(u)}) &= \frac{1}{2}\,(\theta_{12}\,(1-\theta_{23}) + (1-\theta_{12})\,\theta_{23}) \\ &= \theta_1\,(\theta_2\,(1-\theta_3) + (1-\theta_2)\,\theta_3). \end{aligned} \tag{3.8}$$

Toisaalta soveltaen Trow'n kaavaa saadaan

$$\begin{aligned} p(h^u_{\cdot 1} \mid h^{F(u)}) &= \frac{1}{2}\,\theta_{13} \\ &= \frac{1}{2}\cdot\frac{1}{2}\,(1-(1-2\,\theta_{12})(1-2\,\theta_{23})) \\ &= \frac{1}{2}\,\theta_1\,(1-(1-2\,\theta_2)(1-2\,\theta_3)). \end{aligned} \tag{3.9}$$

Lähempi tarkastelu osoittaa lausekkeiden (3.8) ja (3.9) johtavan samaan tulokseen. □

3.6 Malliperheet ja mallien parametrit

Edellä huomattiin, että erivahvuiset biologiset olettamukset vaikuttavat mallin parametrisointiin ja kuvausvoimaan. Seuraavassa käsitellään mallien ryhmittelemistä perheiksi ja luokiksi. Näissä käsitteissä noudatetaan lähteen [MyT98] käytäntöä.

Numeroidaan edellisissä osioissa esitellyt geneettisten mallien komponentteja koskevat keskeiset oletukset seuraavasti:

1. Riippumaton tasainen periytyminen
2. Hardyn ja Weinbergin tasapaino
3. Kytkentätasapaino
4. Kytkemätön periytyminen
5. Kaksialleelisuus
6. Genotyyppien lokuskohtaiset vaikutukset

Jokaisen oletusten yhdistelmän voidaan ajatella määräävän geneettisten mallien perheen. Malliperhe on tällöin joukko malleja, joilla on väljästi sanoen yhdenmukainen rakenne.

Esimerkki 3.4. Tarkastellaan oletusten 1–5 määräämän malliperheen jäsentä M. Olkoon mallin M lokusten lukumäärä ℓ. Koska jokaisessa lokuksessa on kaksi alleelia, mallissa on 3^ℓ järjestämätöntä genotyyppiä ja siksi myös näin monta parametria genotyyppien ilmenemisyleisyyksille. Genotyyppien frekvenssit malli M kuvaa ℓ alleelifrekvenssillä, koska kaksialleelisuuden lisäksi oletetaan Hardyn ja Weinbergin tasapainon sekä kytkentätasapainon olevan voimassa. Oletusten 1 ja 4 vuoksi genotyyppien periytymistodennäköisyydet ovat täysin määrätyt, eikä lisäparametreja tarvita. $\square$

Oletusten erilaisia yhdistelmiä, siten myös erilaisia malliperheitä, on $2^6 = 64$. Jäljempänä rajoitutaan tarkastelemaan vain muutamaa malliperhettä. Suurimman huomion saa oletusten 1–4 määräämä perhe, jolle otetaan käyttöön merkintä $\mathcal{M}_{[4]}$. Malliperheen jäseniä kutsutaan *kytkemättömän periytymisen* malleiksi tai lyhyemmin *kytkemättömiksi* malleiksi. Tärkeä erikoistapaus saadaan, kun oletuksiin lisätään kaksialleelisuuden rajoite. Rajoitettua malliperhettä merkitään $\mathcal{M}_{[5]}$. Luvussa 5 tarkastellaan edellisten lisäksi oletusten 1–3 määräämää malliperhettä, jonka mallit huomioivat lokusten välisen kytkennän. Tätä perhettä merkitään $\mathcal{M}_{[3]}$. Sen jäseniä kutsutaan *kytkennällisen periytymisen* malleiksi, lyhyemmin *kytkennällisiksi* malleiksi. Kun malleja käsitellään yleisesti tekemättä mitään oletuksista 1–6, puhutaan *yleisistä* malleista. Yleisten mallien perhettä merkitään $\mathcal{M}$. Oletuksesta 6 luovutaan kokonaan. Pannaan vielä merkille malliperheiden sisältyvyyssuhteet:

$$\mathcal{M} \supset \mathcal{M}_{[3]} \supset \mathcal{M}_{[4]} \supset \mathcal{M}_{[5]}.$$

Malliperheet voidaan edelleen osittaa *malliluokiksi*. Osittaminen edellyttää mallin hajottamista kahteen osaan: parametrien rakenteeksi ja parametrien arvoiksi. Tässä työssä tarkastellaan lähinnä perheen $\mathcal{M}_{[5]}$ malliluokkia. Havainnollistetaan siksi osittamista tässä perheessä. Kuten esimerkissä 3.4 todettiin, perheen $\mathcal{M}_{[5]}$ jäsenillä on $\ell + 3^\ell$ parametria. Malliluokat määräytyvät nyt parametrien lukumäärän ekvivalenssiluokkina: kaksi mallia kuuluvat samaan luokkaan, jos ja vain jos niissä on yhtä monta parametria. Perheen $\mathcal{M}_{[5]}$ malliluokat ovat siten yhden lokuksen mallien luokka, kahden lokuksen mallien luokka, kolmen lokuksen mallien luokka jne. Otetaan ℓ lokuksen malliluokalle käyttöön merkintä $\mathcal{M}_{[5]}^{(\ell)}$, lyhyemmin $\mathcal{M}^{(\ell)}$.

Mallien parametrien arvojoukkoihin ei edellä otettu selvää kantaa. Matemaattisen käsittelyn kannalta on usein mukavaa, jos todennäköisyyksiä kuvaaven parametrien arvojen sallitaan olla mielivaltaisia välin $[0, 1]$ reaalilukuja. Kun malleja käsitellään tietokoneella, joudutaan kuitenkin rajoittumaan lukuihin, joiden binääriesitys on äärellinen. Esimerkiksi genotyyppien frekvenssien arvojoukkojen äärellisyyttä voidaan puolustella myös tarkasteltavan populaation rajoitetulla koolla. Lokusten lukumäärä on myös rajoitettu, vaikka määrä potentiaalisesti onkin hyvin suuri. Jos ei toisin mainita, oletetaan parametrien arvojoukkojen jäljempänä olevan äärellisiä. Tällöin malliperheet ja -luokat ovat mallien äärellisiä joukkoja.

Mallien parametreihin viitataan jäljempänä samoin kuin edellisissä osioissa. Genotyyppien ilmenemisyleisyyksiin viitataan tunnuksella f ja tekijänvaihtotodennäköisyyksiin symbolilla θ. Alleelien ja genotyyppien frekvensseihin sekä periytymistodennäköisyyksiin viitataan yhteisesti kirjaimella p, jota käytetään mallin mukaisten todennäköisyyksien ilmaisemiseen yleisemminkin. Esimerkiksi kytkennällisen periytymisen mallin M parametrit voidaan kirjoittaa parina $\langle p_M; f_M \rangle$. Tässä yhteydessä p_M viittaa ainoastaan alleelifrekvensseihin.

Tarkastellaan vielä parametrien arvojen rajoittamista kytkemättömissä ja kytkennällisissä malleissa. Esitetään kaksi periaatetta, jotka eivät vähennä mallin parametrien määrää mutta rajoittavat niille annettavia arvoja.

Symmetrian eliminointi: Kaksialleelisessa mallissa voidaan olettaa, että $0 < p(A_{i1}) \leq \frac{1}{2}$ kaikilla i, sillä muuten jokaisella mallilla on 2^ℓ sairaan ja terveen alleelin nimentää vaille ekvivalenttia mallia. Usean alleelin mallissa vastaava rajoite jokaiselle lokukselle i on

$$0 < p(A_{i1}) \leq p(A_{i2}) \leq \cdots \leq p(A_{im_i}) < 1.$$

Ilmenemisyleisyyksien monotonisuus: Kaksialleelisessa mallissa voidaan olettaa, että genotyypin ilmenemisyleisyys on sitä suurempi mitä enemmän se sisältää mutaatioalleeleja. Sanotaan, että ilmenemisyleisyydet ovat *monotoniset*, jos kaikille genotyypeille g pätee ehto: jos genotyyppi g' saadaan genotyypistä g vaihtamalla jonkin lokuksen terve alleeli mutaatioalleeliksi, niin $f(g') \geq f(g)$.

Ehto voidaan yleistää usean alleelin malleille: mallin ilmenemisyleisyydet ovat monotoniset, jos kaikille genotyypeille $g = A_{1j_1}A_{1k_1}\ldots A_{\ell j_\ell}A_{\ell k_\ell}$ ja $g' = A_{1j'_1}A_{1k'_1}\ldots A_{\ell j'_\ell}A_{\ell k'_\ell}$ on voimassa

$$(\forall i : (j'_i \leq j_i \text{ ja } k'_i \leq k_i)) \Rightarrow f(g') \geq f(g).$$

Monotoniset ilmenemisyleisyydet määräävät siten genotyypeille osittaisen järjestyksen.

Symmetrian eliminointi ei rajoita malliperheen kuvausvoimaa, ja ilmenemisyleisyyksien monotonisuusehto on biologisesti järkevä. Ehtojen yhdistelmä on rajoittavampi: vaaditaan, että harvinaisempi alleeli on aina haitallisempi. Myöhemmissä tarkasteluissa mallien ei edellytetä noudattavan näitä ehtoja, mikäli tästä ei erikseen mainita.

Seuraava esimerkki havainnollistaa numeerisesti geneettisen mallin komponentteja ja edellä kuvattuja symmetrian eliminointiehtoa ja ilmenemisyleisyyksien monotonisuusehtoa.

Esimerkki 3.5. Tarkastellaan oletusten 1–5 mukaista kahden lokuksen mallia. Olkoot ensimmäisen lokuksen sairaan alleelin A frekvenssi populaatiossa 0.10 ja toisen lokuksen sairaan alleelin B frekvenssi 0.20. Olkoon genotyyppien ilmenemisyleisyydet taulukon 3.1 mukaiset. Todetaan, että alleelifrekvenssit toteuttavat symmetrian eliminointiehdon ja että ilmenemisyleisyydet ovat monotoniset.

Hardyn ja Weinbergin tasapainosta ja kytkentätasapainosta seuraa, että genotyypin $AaBB$ frekvenssi on $(2 \cdot 0.10 \cdot 0.90)(0.20 \cdot 0.20) = 0.0072$. Koska genotyypin $AaBB$ ilmenemisyleisyys on 0.80, on genotyyppiä $AaBB$ kantavien sairaiden osuus populaatiossa $0.80 \cdot 0.72\% \approx 0.6\%$. Taulukosta 3.1 nähdään, että mallin mukaan tautiin vaikuttavat selvästi mutatoituneet alleelit: taudin todennäköisyys yksilöllä on selvästi suurempi, jos hänellä on molemmat mutatoituneet alleelit A ja B kuin, jos hänellä on vain toinen näistä tai ei kumpaakaan. Sen sijaan alleliparit AA tai BB eivät juuri lisää yksilön riskiä sairastua verrattuna alleelipareihin Aa ja Bb. □

	AA	Aa	aa
BB	0.90	0.80	0.20
Bb	0.70	0.60	0.10
bb	0.05	0.02	0.01

Taulukko 3.1: Esimerkki ilmenemisyleisyyksistä kahden lokuksen kaksialleelisessa mallissa.

Luku 4

Havaintojen ja mallien suhde

Kun taudin esiintymisestä on tehty havaintoja ja tarkasteltavien geneettisten mallien perhe on valittu, voidaan siirtyä mallintamiseen oleellisimpaan vaiheeseen, tilastolliseen päättelyyn. Tässä päättelyllä tarkoitetaan niin tautiin vaikuttavien lokusten lukumäärän arvioimista kuin myös genotyyppien yhteisvaikutusten selvittämistä tai sairastumisriskin arvioimista lapselle, jonka vanhemmat tiedetään sairaiksi. Jos mitään ei päätellä, mallintaminen on turhaa.

Päättely edellyttää havaintojen ja mallien suhteen analysointia. Todennäköisyysmatematiikka tarjoaa tähän sopivat välineet. Luku aloitetaan muotoilemalla havaintojen todennäköisyys mallien funktiona. Sitten esitetään bayesiläinen lähestymistapa, joka mahdollistaa puhumisen mallien todennäköisyyksistä. Lopuksi käsitellään havaintoaineiston määrän ja laadun merkitystä mallintamisessa.

4.1 Havaintojen uskottavuus

Edellisessä luvussa todettiin, että taudin geneettinen malli mallintaa mitä tahansa sukupuuta ja siten koko populaatiota. Seuraavaksi tarkastellaan populaatiota luonnehtivien sukulaisriskihavaintojen (ks. luku 2) uskottavuutta annetussa mallissa. Tätä varten määritellään taudin yleisyys ja sukulaisriskit kolmannen kerran – nyt suhteessa geneettisiin malleihin.

Taudin yleisyys geneettisessä mallissa määritellään todennäköisyytenä, että populaatiosta satunnaisesti valittu yksilö on sairas. Suureesta käytetään lähes vakiintuneen käytännön mukaan merkintää K [Ris90, Sha98, Lan97] (joskus myös K_p [Jam71]). Täsmällisemmin ilmaistuna

$$K = P(X_u = 1),$$

missä u on populaatiosta satunnaisesti valittu yksilö. Korostetaan, että P viittaa geneettisen mallin määrittelemiin todennäköisyyksiin. *Sukulaisriski tyypin R sukulaiselle geneettisessä mallissa* määritellään todennäköisyytenä, että sairaan yksilön tyyppiä R oleva sukulainen on myös sairas. Sukulaisriskille käytetään merkintää K_R. Täsmällisemmin kirjoitettuna

$$K_R = P(X_v = 1 \mid X_u = 1),$$

missä v on yksilölle u tyypin R sukulainen populaatiosta satunnaisesti valitussa sukupuussa. Täydellisessä mallissa suureiden arvot yhtyvät taudin yleisyyteen ja sukulaisriskeihin populaatiossa.

Geneettisen mallin hyvyyttä voidaan koetella vertaamalla mallin mukaisia riskejä riskihavaintoihin. Lähempää tarkastelua varten olkoon (N^R, N_1^R) sukulaisuustyyppiä R koskeva havainto. Sairaiksi havaittujen lukumäärää N_1^R voidaan ajatella tulokseksi binomikokeesta, jossa onnistumistodennäköisyys on riski K_R ja otoksen koko on N^R. Havainnon todennäköisyys on silloin

$$\binom{N^R}{N_1^R} K_R^{N_1^R} (1 - K_R)^{N^R - N_1^R}.$$

(Yleensä havainnot tehdään kuitenkin takaisinpanotta, jolloin koe ei tarkasti ottaen ole binominen. Myös monet muut syyt voivat johtaa kuvatusta ihannetapauksesta poikkeamiseen. Näin on esimerkiksi silloin, jos havainnoidut sukulaisparit ovat sukua toisilleen.)

Olkoon sitten laajemmin D havaintoaineisto, joka koostuu kokoelman $\mathcal{R}$ sukulaisuustyypeistä (esimerkiksi tyypeistä O, S ja MZ). Havaintoaineiston D uskottavuus mallissa M on nyt

$$L(M; D) = \prod_{R \in \mathcal{R}} \binom{N^R}{N_1^R} K_R^{N_1^R} (1 - K_R)^{N^R - N_1^R}, \qquad (4.1)$$

kun riskit K_R ovat mallin M mukaisia. Tässäkin on oletettava, että havainnot eri sukulaisuustyypeille ovat riippumattomia. Tässä yhteydessä taudin yleisyys voidaan rinnastaa sukulaisriskeihin, ja siten D voi sisältää myös havaintoja taudin yleisyydestä.

Uskottavuusfunktio sitoo mallit ja havainnot. Se mittaa eräässä mielessä mallin kelvollisuutta. Jos kahdesta mallista toinen antaa havainnolle pienen todennäköisyyden ja toinen malli selvästi suuremman todennäköisyyden, voidaan jälkimmäistä mallia pitää aineiston valossa parempana. Tähän lähestymistapaan perustuu tilastollisessa mallintamisessa laajasti käytetty yleinen *suurimman uskottavuuden menetelmä*. Menetelmässä etsitään malli, jossa havaintojen uskottavuus on mahdollisimman suuri. Tämän jälkeen kaikki päätelmät tehdään tähän löydettyyn malliin nojautuen. Löydöksen luotettavuutta arvioidaan käyttämällä sopivia tilastollisia testisuureita. Segregaatioanalyysissä suurimman uskottavuuden menetelmä on ollut suosittu (esim. [Smi71, DBC93]).

Sukulaisriskihavaintoihin perustuvaan mallintamiseen suurimman uskottavuuden menetelmää esittivät mm. Campbell ja Elston [CaE71]. He muokkasivat lauseketta (4.1) jättämällä binomikertoimet pois ja ottamalla logaritmin. Näin he saivat *logaritminen pseudouskottavuus* (engl. pseudo log likelihood)

$$l(M; D) = \sum_{R \in \mathcal{R}} \left(N_1^R \log K_R + (N^R - N_1^R) \log(1 - K_R) \right). \qquad (4.2)$$

Todettakoon, että funktioiden (4.1) ja (4.2) maksimikohdat (eli maksimoivat mallit) ovat samat, koska binomikertoimet eivät riipu mallista M ja logaritmikuvaus on monotoninen. Tulokaavaan (4.1) verrattuna tämän summakaavan käsittely on laskennallisesti helpompaa, kun otoskoot ovat suuria ja riskit ovat lähellä arvoa 0 tai 1. Suurimman uskottavuuden mallin löytäminen ja löydöksen merkittävyyden arviointi numeerisesti onnistuu vaikeuksitta [CaE71].

4.2 Bayesiläinen lähestymistapa

Edellisessä osiossa käsiteltiin havaintojen uskottavuutta annetussa geneettisessä mallissa. Päättelyn kannalta tämä ei ole suoraviivaista, koska itseasiassa havainnot ovat annettuina ja kiinnostavaa on hypoteettisen mallien uskottavuudet. Klassisessa frekventistisessä tilastotieteessä malleilla ja havainnoilla on selvä ero: havaintoja tarkastellaan satunnaismuuttujina, malleja ei. Bayesiläisessä lähestymistavassa myös mallien ajatellaan olevan satunnaismuuttujia, ja siten voidaan puhua mallien todennäköisyyksistä tai todennäköisyystiheyksistä, sen mukaan, ovatko mallit diskreettejä vai jatkuvia. Seuraavassa sovelletaan yleistä bayesiläistä lähestymistapaa nojautuen pääasiallisesti lähteisiin [BoT73, Ch. 1], [GRS96, Ch. 1] ja [MyT98].

4.2.1 Bayesiläinen päättely

Esitetään bayesiläisen lähestymistavan perusperiaatteet sovellettuna geneettisiin malleihin. Olkoot $\mathcal{M}$ tarkasteltava malliperhe, M malliperheen jäsen ja D joukko sukulaisriskihavaintoja. Jotta voidaan puhua mallien todennäköisyyksistä, mallien oletetaan tässä olevan diskreettejä. Pyrkimyksenä on ensin löytää laskettavissa oleva muoto mallin *posterioritodennäköisyydelle* $p(M \mid D)$. (Huomautettakoon, että nyt p viittaa bayesiläiseen todennäköisyyteen eli subjektiivisen uskomuksen asteeseen.)

Ehdollisen todennäköisyyden määritelmää ja ketjusääntöä käyttäen saadaan

$$p(M \mid D) \;=\; \frac{p(M)\,p(D \mid M)}{p(D)}.$$

Tämä kaava tunnetaan *Bayesin teoreemana*. Termi $p(D \mid M)$ on havainnon D uskottavuus mallissa M eli $L(M; D)$, jolle annettiin lauseke osiossa 4.1. Termiä $p(M)$ kutsutaan mallin M prioritodennäköisyydeksi, lyhyemmin *prioriksi*. Priori kertoo, kuinka paljon malliin uskotaan, ennen kuin havaintoja on tehty. Geneettisten mallien prioreja käsitellään tarkemmin seuraavassa osiossa 4.2.2. Kolmas termi on havainnon todennäköisyys $p(D)$, jolle voidaan kirjoittaa kaava

$$p(D) \;=\; \sum_{M' \in \mathcal{M}} p(M')p(D \mid M').$$

Käytännössä summan laskeminen voi olla vaikeaa tai työlästä. Huomataan kuitenkin, että havainnon todennäköisyys ei riipu mallista M. Usein havainnon todennäköisyyden tunteminen ei päättelyn kannalta olekaan tarpeellista.

Bayesiläinen päättely on – lyhyesti sanoen – päättelyä mallien posterioritodennäköisyyksiin $p(M \mid D)$ nojautuen. On tärkeää huomata, että päättely on aina ehdollista valitulle malliperheelle. Seuraavat esimerkeissä valotetaan asiaa.

Esimerkki 4.1. Olkoot $\mathcal{M}$ jokin geneettisten mallien perhe ja D tautiin liittyvä havaintoaineisto. Oletetaan, että ollaan kiinnostuneita tautiin vaikuttavien lokusten lukumäärästä. Bayesiläisen lähestymistavan mukaan lokusten lukumäärälle muodostetaan todennäköisyysjakauma. Posterioritodennäköisyys ℓ

lokukselle on

$$p(\ell \mid D) \;=\; \sum_{M \in \mathcal{M}} p(M \mid D)\, 1_{\{\ell_M = \ell\}},$$

missä ℓ_M on lokusten lukumäärä mallissa M. Bayesin teoreemaa käyttäen saadaan

$$p(\ell \mid D) \;=\; c \sum_{M \in \mathcal{M}} p(M)\, p(D \mid M)\, 1_{\{\ell_M = \ell\}},$$

missä $c = 1/p(D)$ on normeeraava malleista riippumaton vakio. Esimerkiksi lokusten todennäköisin lukumäärä voidaan nyt päätellä laskemalla summaus ilman vakiota c. □

Esimerkki 4.2. Tarkastellaan kaksialleelisten kytkemättömien mallien luokkaa $\mathcal{M}^{(\ell)}_{[5]}$. Järjestetään mallin M parametrit (alleelifrekvenssit ja ilmenemisyleisyydet) ja merkitään mallin M järjestyksessä j:nnen parametrin arvoa M_j. Esimerkin 4.1 tapaan posterioritodennäköisyys sille, että j:nnen parametrin arvo on välillä $[y_1, y_2]$, on

$$c \sum_{M \in \mathcal{M}^{(\ell)}_{[5]}} p(M)\, p(D \mid M)\, 1_{\{y_1 \le M_j \le y_2\}}.$$

Yhden parametrin sijaan voidaan yhtä hyvin tarkastella usean parametrin yhteisjakaumaa. □

Edellä annettujen esimerkkien päättely voidaan yleistää seuraavasti (vrt. [GRS96, s. 3]). Olkoot $\mathcal{M}$ malliperhe tai -luokka ja $\mathbf{M}$ satunnaismuuttuja arvojoukkonaan $\mathcal{M}$. Olkoon muuttujan $\mathbf{M}$ jakauma mallien posteriorijakauma π, jolle siis $\pi(M) = p(M \mid D)$. Päättelyn kannalta kiinnostava suure voidaan nyt kirjoittaa odotusarvona

$$E_\pi(z(\mathbf{M})) \;=\; \sum_{M \in \mathcal{M}} \pi(M)\, z(M), \tag{4.3}$$

missä funktio z määrää kiinnostuksen kohteen (vrt. edellä annettuihin esimerkkeihin). Odotusarvon laskeminen analyyttisesti on usein vaikeaa, jopa mahdotonta. Luvussa 6 esitetään numeerinen simulointimenetelmä, jolla summa (4.3) voidaan laskea likimääräisesti.

4.2.2 Geneettisten mallien prioritodennäköisyyksistä

Edellä mainittiin priorien tarpeellisuus bayesiläisessä päättelyssä. Todettiin, että prioritodennäköisyydet ovat mallintajan valittavissa – prioreja ei voi eikä saa määrätä havaintoaineistoon nojautuen. Bayesiläisen lähestymistavan puolustajat näkevät priorien määrittelemisen yleensä mahdollisuutena ja etuna [MyT98]. Koska priorien tulkinta on lopulta subjektiivista, niiden valinta voi tuottaa ongelmia. Edut ja ongelmat ovat ilmeiset myös geneettisten mallien prioreita valittaessa.

Priorien valinta antaa mahdollisuuden tautiin liittyvän taustatietämyksen kuvaamiseen. Jos taustatietämys perustuu aikaisempaan, mahdollisesti erilaisen havaintoaineiston, bayesiläiseen analyysiin, on priorijakaumaksi luontevaa asettaa aikaisemman havaintoaineiston mukainen posteriorijakauma. Tavallisemmin on kuitenkin nojauduttava asiantuntijatietämykseen. Asiantuntija voisi esimerkiksi sanoa, että sairauksien vallitseva periytyminen on yleisempää kuin peittyvä periytyminen. Toisena uskomuksena voisi olla, että tarkasteltavaa sairautta ei voida selittää yhdellä vaikuttavalla lokuksella. Tällaisen tietämyksen muuntaminen numeeriseen muotoon voi olla vaikeaa.

On tavallista, ettei minkäänlaista taustatietämystä ei ole käytettävissä. Yleisen nyrkkisäännön mukaan tällöin on käytettävä *epäinformatiivisia* tai *informaatioteoreettisia* prioreja [MyT98, s. 34]. Keskeisen ongelman tuottaa se, että malliperheen jäsenillä parametrien lukumäärä vaihtelee. Seuraavassa tarkastellaan havainnollisuuden vuoksi kytkemättömän periytymisen kaksialleelisten mallien perhettä $\mathcal{M}_{[5]}$.

Oletetaan tässä, että kyseinen malliperhe on mallien äärellinen joukko (ks. osio 3.5). Silloin jokaisen mallin prioritodennäköisyys voidaan asettaa yhtäsuureksi. Priorijakaumaa voidaan pitää epäinformatiivisena, koska se ei suosi mitään mallia. Menettely suosii kuitenkin malliluokkia, joiden parametrien määrä on suuri. Tarkasteltavan malliperheen osalta seuraa erikoisesti, että kahden lokuksen malliluokan priori on suurempi kuin yhden lokuksen malliluokan priori. Lisäksi parametrien arvojoukko vaikuttaa malliluokkien prioritodennäköisyyksiin: diskretoinnin tarkentuessa malliluokkien prioritodennäköisyyksien erot kasvavat. Kuvattua menettelyä ei voida soveltaa, jos malliperhe on ääretön.

Informaatioteoreettinen kriteeri ehdottaa prioria, joka on suhteessa mallin pakkautuvuuteen [MyT98, s. 38–40], [OlB94]. Naiivi tarkastelu antaa tällöin suhteen

$$-\log p(M) \quad \approx \quad \ell_M + 3^{\ell_M} + c,$$

missä ℓ_M on lokusten lukumäärä mallissa M ja c on jokin vakio. Tässä oletetaan, että jokaisen parametrin arvo on kuvattavissa rajoitetun pituisena binäärijonona. Menettely johtaa lokusten lukumäärän suhteen ylieksponentiaaliseen prioriin, joka suosii voimakkaasti yhden ja kahden lokuksen malleja. Lähestymistapa on kelvollinen, vaikka malliperhe olisi numeroituvasti ääretön. Informaatioteoreettinen priori sakottaakin paitsi parametrien lukumäärästä, myös niiden tarkkuudesta. Informaatioteoreettisen kriteerin täydellinen noudattaminen vaatisi Kolmogorovin monimutkaisuusmitan (engl. Kolmogorov complexity) mukailua. Tässä työssä ei aihetta voida käsitellä syvällisemmin; varsin uusia teoreettisia pohdintoja aiheeseen liittyen löytyy lähteestä [ViL00].

Kolmas menettely priorien määräämisessä yhdistää edellä selostettujen lähestymistapojen piirteitä. Priori hajotetaan parametrien rakenteen prioriksi ja parametrien arvojen prioriksi, joista jälkimmäinen on ehdollinen parametrien rakenteelle eli lokusten lukumäärälle [MyT98, s. 34]. Tämä voidaan ilmaista ketjusäännön sovelluksena:

$$p(M) \quad = \quad p(\ell_M)\, p(p_M, f_M \mid \ell_M), \tag{4.4}$$

missä ℓ_M, p_M ja f_M kuvaavat lokusten lukumäärän, alleelifrekvenssit ja ilmenemisyleisyydet mallissa M. Jos parametrien arvojoukot ovat äärellisiä, jälkimmäistä komponenttia vastaava jakauma voidaan asettaa tasaiseksi ja siten eräässä mielessä epäinformatiiviseksi. Jos parametreja haluttaisiin käsitellä jatkuva-arvoisina, todennäköisyysjakauma voitaisiin korvata tasaisella tiheysfunktiolla. Tasaisen jakauman periaatteellisena ongelmana on se, että se ei ole invariantti suhteessa mallin parametrien muunnoksiin [BoT73, Ch. 1].

Malliluokkien priorit $p(\ell_M)$ voidaan valita yhtäsuuriksi, jos lokusten lukumäärä oletetaan rajoitetuksi. Koska lukumäärä voi periaatteessa olla hyvin suuri, voi kuitenkin olla järkevää valita priori, joka pienenee lokusten lukumäärän kasvaessa. Jos malliperhe sallisi mielivaltaisen suuria määriä lokuksia, näin olisi pakkokin menetellä.

Jos bayesiläisessä päättelyssä rajoitutaan malliluokkaan malliperheen sijaan, mallin rakenteen prioria ei tarvitse määrätä. Asiaan palataan luvussa 6, jossa toteutuksellisten syiden vuoksi joudutaan tarkastelemaan malliluokkaa malliperheen sijaan. Samassa yhteydessä esitetään myös tulon (4.4) jälkimmäistä komponenttia vastaava "informatiivinen" priorijakauma.

4.3 Havaintojen riittävyys

Edellä tarkasteltiin havaintojen ja mallien suhdetta olettaen, että havaintoaineisto on annettu. Kysymys siitä, minkälaiset havainnot ovat mallintamisen kannalta edullisimpia, sivuutettiin. Seuraavaksi pohditaan, mikä olisi määrällisesti suuremman tai pienemmän aineiston tai laadultaan erilaisen aineiston vaikutus taudin mallintamiseen.

4.3.1 Havaintojen tarkkuus

Sukulaisriskihavainnot ovat sitä tarkempia, mitä suurempi on havaintoaineiston koko. Tarkimmillaan havaintoaineisto kattaa koko populaation. On selvää, että tarkka aineisto erottelee geneettiset mallit paremmin kuin epätarkka aineisto. Tämä nähdään myös havaintojen uskottavuuden lausekkeesta (4.1): geneettisen mallin mukaisten sukulaisriskien poikkeama havaituista riskeistä korostuu voimakkaasti aineiston koon kasvaessa.

Havaintojen tarkkuus huomioidaan eri lailla suurimman uskottavuuden menetelmässä ja bayesiläisessä päättelyssä. Suurimman uskottavuuden menetelmässä päätelmät nojautuvat malliin, joka sopii havaintoihin parhaiten. Jos aineisto on pieni, sen uskottavuus eroaa sopivimmalla ja toiseksi sopivimmalla mallilla vain vähän. Nämä mallit voivat olla päättelyn kannalta hyvin erilaisia, minkä vuoksi ainoastaan sopivimpaan malliin nojautuminen ei ole järkevää. Bayesiläinen menetelmä ottaa huomioon aineiston tarkkuuden käyttämällä päättelyssä useiden mallien posterioritodennäköisyyksillä painotettua keskiarvoa. Jos havaintoja on hyvin vähän, mallien painotus noudattelee priorijakaumaa. Tarkkuuden kasvaessa uskottavuustermistä tulee hallitseva ja priorin merkitys vähenee. Tarkkuuden kasvaessa rajatta menetelmien ero häviää, mikäli suurimman uskottavuuden malli on yksikäsitteinen.

4.3.2 Havaintojen laatu

Esittäessään sukulaisriskihavaintohin nojautuvaa suurimman uskottavuuden menetelmää Campbell ja Elston [CaE71] myönsivät, että sukulaispareihin rajoittuvat havainnot eivät riitä määräämään suurimman uskottavuuden mallia yksikäsitteisesti. Sen sijaan useat erilaiset geneettiset hypoteesit voivat tulla kyseeseen. He totesivat, että hypoteesien erottelu edellyttää monimutkaisempia menetelmiä, kuten segregaatioanalyysiä tai kytkentäanalyysiä. Campbell ja Elston eivät esittäneet perusteluja väitteelleen. Erityisen mielenkiintoinen on avoimeksi jäänyt kysymys sukulaisriskihavaintojen monipuolisuuden merkityksestä mallintamisessa. Sukulaisriskihavainnot ovat – väljästi sanoen – sitä monipuolisempia, mitä useampia sukulaisuustyyppejä ne kattavat. Monipuolisuutta voidaan siten pitää laadullisena ominaisuutena, verrattuna havaintojen määrään

Mallin yksikäsitteistä määräytymistä voidaan ajatella mallin parametrien identifioitumisena havaintojen tarkentuessa. Suurimman uskottavuuden menetelmän ohella identifioituvuutta voidaan tarkastella myös Bayesiläisessä mallintamisessa. Tällöin on kysymys suurimman posterioritodennäköisyyden antavan mallin yksikäsitteisyydestä. Molemmissa tapauksissa kysymys identifioituvuudesta palautuu havaintojen ja mallien suhdetta kuvaavan uskottavuusfunktion ominaisuuksiin.

Uskottavuusfunktion ominaisuuksia on helpointa tarkastella malliluokissa, joiden jäsenillä on samat parametrit. Matemaattinen tarkastelu helpottuu myös, jos parametreja käsitellään jatkuva-arvoisina. Tällöin havaintojen uskottavuus on siisti jatkuva funktio parametriavaruudessa. Epäidentifioituvuuden yhtenä syynä voivat olla funktion symmetriat. Esimerkiksi kytkemättömät mallit ovat symmetrisia lokusten suhteen: parametrien permutointi lokusten suhteen ei vaikuta mallin mukaisiin sukulaisriskeihin. Tärkeämpi syy epäidentifioituvuuteen on mallien *yliparametrisointi* eli parametrien suuri määrä suhteessa havaintoihin. Yliparametrisoinnin vuoksi parametriavaruudessa voi olla kokonaisia mallien jatkumoita, joissa uskottavuusfunktion arvo on sama. Jos symmetrioiden sanotaan johtavan funktion monihuippuisuuteen, yliparametrisoinnin voidaan sanoa aiheuttavan eräänlaisia harjanteita.

Pitäytyen luvussa 2 esitetyn kaltaisissa havainnoissa seuraavaksi esitetään joitakin identifioituvuuteen liittyviä tuloksia kytkemättömille malleille. Aluksi tarkastellaan asetelmaa, jossa perinnölliseksi arvellulle taudille etsitään yhden lokuksen kaksialleelista mallia. Mallissa on tällöin neljä parametria: ensimmäisen alleelin frekvenssi ja ilmenemisyleisyydet kolmelle genotyypille.

James [Jam71] havaitsi, että kaikkien neljän parametrin määrääminen sukulaisriskeihin perustuen on mahdotonta, koska kaikkien sukulaisuustyyppien sukulaisriskit voidaan esittää kolmen suureen avulla. Nämä suureet ovat taudin yleisyys K, additiivinen varianssi V_A ja dominantti varianssi V_D, joita käsitellään sukulaisriskien laskennan yhteydessä tarkemmin (ks. osio 5.3.2). Kun lokuksia on vain yksi, tulos pysyy voimassa, vaikka alleeleja olisi enemmänkin. Jamesin tuloksista seuraa, että parametrien identifioituvuuden kannalta ei ole syytä käyttää havaintoja useille sukulaisuustyypeille, mikäli sukulaisriskihavainnot ovat selitettävissä yhden lokuksen mallilla. Esimerkiksi taudin yleisyyden, lapsiriskin ja sisarusriskin lisänä ei identtisen kaksosen havaittu riski anna lisätietoa alleelifrekvenssien ja ilmenemisyleisyyksien arvoista.

Fishman et al. [FRS78] ovat tutkivat ydinperheisiin pohjautuvien havaintojen riittävyyttä ja ovat esittävät kaksi mielenkiintoista tulosta. Lähempää tarkastelua varten numeroidaan kaksilapsisen ydinperheen yksilöt järjestyksessä: äiti, isä, vanhempi lapsi ja nuorempi lapsi. Merkitään

$$K_{O|i,j} = P(X_3 = 1 \mid X_1 = i, X_2 = j),\ \ i, j \in \{0, 1\}$$

ja

$$K_{O^2|i,j} = P(X_3 = 1, X_4 = 1 \mid X_1 = i, X_2 = j),\ \ i, j \in \{0, 1\}.$$

Fishman et al. laajensivat Jamesin tulosta ja totesivat, että suureet $K_{O|i,j}$ eivät riitä määräämään kaikkia neljää parametria. Toisaalta he osoittivat, että taudin yleisyys K, lapsiriski K_O ja edellä määritellyt $K_{O|i,j}$ ja $K_{O^2|i,j}$ yhdessä määräävät yksikäsitteisesti kaikki mallin neljä parametria. Poikkeuksena on tapaus, jossa $V_A = 0$. Korostettakoon, että suureita K_O ja $K_{O|i,j}$ vastaavat havainnot ovat sukupuuhavaintoja, mutta ne eivät ole sukulaisriskihavaintoja luvun 2 määritelmän mielessä.

Whittemore et al. [WKW91] esittivät toisen tavan parametrien määräämiseen. Käytetään todennäköisyydestä, että k-lapsisessa ydinperheessä kaikki k sisarusta ovat sairaita, merkintää p_{S^k}. Kirjoittajat osoittivat numeerisella esimerkillä, että suureet p_{S^1}, p_{S^2}, p_{S^3} ja p_{S^4} voivat määrätä kaikki mallin neljä parametria yksikäsitteisesti. Samalla he kuitenkin totesivat, että joillakin suureiden arvoilla sopivia malleja voi olla useita. He eivät ottaneet kantaa siihen, voiko sopivia malleja olla ääretön määrä, vai onko niitä aina korkeintaan muutama niin kuin heidän esimerkissään. Kirjoittajat erehtyivät ”vastaesimerkkinsä” nojalla väittämään, että Jamesin huomio on väärä. Tämä osoittaa, että he eivät ymmärtäneet Jamesin huomion koskevan ainoastaan yksilöiden välisiä pareittaisia riippuvuuksia: sukulaisriskejä ja korrelaatioita. Heiltä on myös selvästi jäänyt huomioimatta Fishmanin et al. tulokset.

Neljän parametrin identifioituvuusongelma paljastaa sukulaisriskihavaintojen heikkouden verrattuna yleisempiin sukupuuhavaintoihin. Fishmanin et al. sekä Whittemoren et al. tulokset antavat kuitenkin toivoa, että havaintojen rikastaminen muilla kuin yksilöiden pareittaista riippuvuutta kuvaavilla suureilla voi mahdollistaa parametrien arvojen yksikäsitteisen määräämisen.

Jamesin tulos yleistyy myös useammalle lokukselle. Kun mallissa on ℓ lokusta, sukulaisriskihavainnoilla identifioituvien muuttujien lukumäärä on luokkaa $\Theta(\ell^2)$ (ks. kaava (5.14), luku 5). Kiinnostavien parametrien lukumäärä on kuitenkin eksponentiaalinen suhteessa lokusten lukumäärään. Edellä mainittiin, että neljän parametrin malli voidaan täysin määrätä ydinperheisiin perustuviin havaintoihin pohjautuen. Kun malli käsittää useampia lokuksia ja alleeleja, on luultavaa, että mallin parametrien yksikäsitteinen määrääminen vaatii monimutkaisempiin sukupuihin perustuvia havaintoja. Käytännössä yksikäsitteisyyden tavoittelusta onkin luovuttava.

Koska bayesiläisessä päättelyssä huomioidaan kaikki mahdolliset mallit, lähestymistapa kärsii epäidentifioituvuudesta vähemmän kuin suurimman uskottavuuden menetelmä. Ratkaisevaa on kuitenkin se, mitä on tarkoitus päätellä. Jos esimerkiksi yritetään ennustaa havaintojen mukaisia riskejä, ei epäidentifioituvuus ole suurimman uskottavuuden menetelmässäkään ongelma. Jos sen sijaan on tarkoitus päätellä esimerkiksi tautialleelien frekvenssejä, voi suurimman uskottavuuden menetelmä olla heikko jättäessään huomiotta suurimman uskottavuuden mallin vaihtoehtoiset hypoteesit.

Luku 5

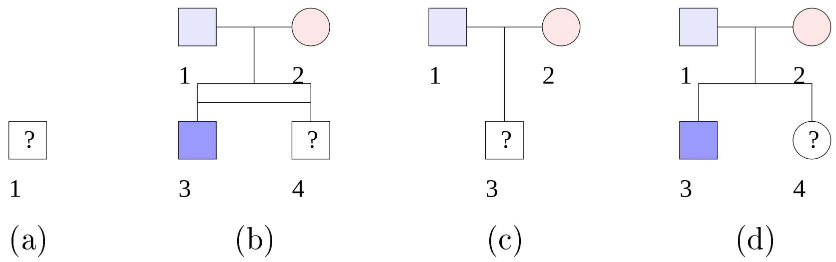
Sukulaisriskien laskenta

Edellisessä luvussa 4 esitettiin, kuinka hypoteettisen geneettisen mallin mukaiset sukulaisriskit määräävät sukulaisriskihavaintojen uskottavuuden kyseisessä mallissa. Avoimeksi jätettiin, kuinka sukulaisriskejä lasketaan annetussa mallissa. Tässä luvussa laskentaongelmaa tarkastellaan kahdessa ulottuvuudessa. Ensimmäisen ulottuvuuden muodostavat laskettavat suureet, joita ovat taudin yleisyys, ydinperheriskit ja yleiset sukulaisriskit. Toisen ulottuvuuden muodostavat geneettisten mallien perheet. Tässä rajoitutaan yleisiin, kytkennällisen periytymisen ja kytkemättömän periytymisen malleihin.

Tavoitteena on löytää kullekin suure-malliperhe -parille paras laskentamenetelmä, kun menetelmän hyvyyttä mitataan yksinkertaisesti laskennan aikavaativuutena. Tullaan huomaamaan, että laskentaongelman ja -menetelmien luonteen vuoksi pahimman ja keskimääräisen tapauksen vaativuudet ovat samoja ja että menetelmien tilantarve ei ole kiinnostavaa. Aikavaativuuksia tarkastellaan lähtökohtaisesti lokusten lukumäärän ja alleelien lukumäärien funktiona. Osoittautuu kuitenkin, että vaativuudet voidaan usein kirjoittaa kohtuullisen yksinkertaisena järjestämättömien tai järjestettyjen genotyyppien lukumäärän funktiona. Saavutettavat aikavaativuudet kootaan taulukoksi osiossa 5.4.

5.1 Taudin yleisyys ja ydinperheriskit

Taudin yleisyys ja ydinperheriskit ovat käytännön mallintamisessa oleellisimpia riskityyppejä. Ne ovat myös helpoimpia laskea, koska niitä vastaavat sukulaisuustyypit voidaan määritellä yksinkertaisilla sukupuilla, jotka koostuvat vain muutamista yksilöistä. Kuvassa 5.1 on esitetty näitä riskejä vastaavat sukupuut. Seuraavissa osioissa 5.1.1–5.1.4 käsitellään kunkin riskin laskemista perustuen kuvan sukupuihin ja yksilöiden numerointiin. Sukupuiden oletetaan olevan satunnaisesti populaatiosta valittuja.



Kuva 5.1: Taudin yleisyyttä (a), identtisen kaksosen riskiä (b), lapsiriskiä (c) ja sisarusriskiä (d) vastaavat sukupuut. Kysymysmerkillä on osoitettu yksilö, jonka sairastumisriskistä on kyse.

Kunkin riskin osalta laskentaa käsitellään ensin yleisessä malliperheessä hyödyntäen Elstonin ja Stewartin komponenttiesitystä. Tämän jälkeen esitetään, kuinka laskentaa voidaan tehostaa kytkennällisen tai kytkemättömän periytymisen malleissa. Myöhemmin osiossa 5.3.3 palataan Elstonin ja Stewartin komponenttiesitykseen. Silloin nähdään, että tässä osiossa esitettävät yleisille malleille soveltuvat laskentamenetelmät ovat yleisemmän sukupuun läpikäyntialgoritmin erikoistapauksia.

5.1.1 Taudin yleisyys

Taudin yleisyydelle pätee määritelmän nojalla

$$K = P(X_1 = 1).$$

Summaamalla yli yksilön tuntemattomien genotyyppien saadaan

$$\begin{aligned} P(X_1 = 1) &= \sum_{h^1 \in \mathcal{H}} P(H^1 = h^1)\, P(X_1 = 1 \mid H^1 = h^1) \\ &= \sum_{h^1 \in \mathcal{H}} p(h^1) f(h^1). \end{aligned}$$

Yleisissä malleissa frekvenssit $p(h^1)$ ovat suoraan mallien parametreja. Ilmenemisyleisyydet $f(h^1)$ voidaan määrätä helposti järjestämättömien genotyyppien ilmenemisyleisyyksien mukaan. Summan suoraviivainen laskenta onnistuu siten ajassa $\Theta(|\mathcal{H}|)$.

Kytkennällisen periytymisen malleissa voidaan hyödyntää tasapaino-oletuksia (ks. kaava (3.6)) ja kirjoittaa summaus lokuksittaisten järjestämättömien genotyyppien avulla:

$$P(X_1 = 1) \quad = \quad \sum_{g_1^1 \in \mathcal{G}_1} |g_1^1|\, p(g_{11}^1)\, p(g_{12}^1) \ldots \sum_{g_\ell^1 \in \mathcal{G}_\ell} |g_\ell^1|\, p(g_{\ell 1}^1)\, p(g_{\ell 2}^1)\, f(g^1). \quad (5.1)$$

Summauksen laskenta voidaan kirjoittaa yksinkertaiseksi rekursiiviseksi algoritmiksi kuten kuvassa 5.2. Kun aliohjelma Yleisyys suoritetaan arvolla $i = 1$, se laskee taudin yleisyyden kaavan (5.1) mukaisesti. Algoritmi ei ota kantaa genotyyppien esitykseen eikä tietorakenteeseen, jossa genotyyppien ilmenemisyleisyydet kuvataan. Yksi mahdollisuus on genotyyppien iteratiivinen tiivis koodaaminen indekseiksi ilmenemisyleisyydet sisältävään taulukkoon. "Keskeneräiden indeksi" annetaan tällöin parametrina rekursiivisessa kutsussa, kunnes se viimeisellä rekursiotasolla on "valmis". Tällainen voidaan selvästikin toteuttaa tehokkaasti. Algoritmin aikavaativuudeksi saadaan $\Theta(|\mathcal{G}|)$.

Yleisyys(i)
1. **jos** $i > \ell$ **niin palauta** $f(g)$
2. $s := 0$
3. **jokaiselle** $g_i \in \mathcal{G}_i$ **:**
4. $\quad s := s + |g_i|\, p(g_{i1})\, p(g_{i2})$ Yleisyys($i + 1$)
5. **palauta** s

Kuva 5.2: Algoritmi taudin yleisyyden K laskemiseen kytkennällisen periytymisen malleissa.

Taudin yleisyyden laskeminen näyttää, kuinka vaativaa riskien laskenta helpoimmillaan on. Koska periytymistodennäköisyyksiä ei tarvinnut huomioida, päästiin yleisissä malleissa alilineaariseen vaativuuteen suhteessa parametrien määrään. Kytkennällisen periytymisen malleissa vaativuus on samaa luokkaa kuin parametrien määrä.

5.1.2 Identtisen kaksosen riski

Määritelmän nojalla identtisen kaksosen riskille on voimassa

$$K_{MZ} \quad = \quad P(X_4 = 1 \mid X_3 = 1).$$

Ehdollisen todennäköisyyden määritelmä antaa

$$P(X_4 = 1 \mid X_3 = 1) \quad = \quad \frac{P(X_4 = 1, X_3 = 1)}{P(X_3 = 1)}.$$

Kun summataan yli vanhempien 1 ja 2 sekä kaksosen 3 mahdollisten genotyyppien, nimittäjä saadaan kirjoitettua muodossa

$$P(X_3 = 1) \;=\; \sum_{h^1,h^2,h^3 \in \mathcal{H}} p(h^1)\, p(h^2)\, p(h^3 \mid h^1, h^2)\, f(h^3).$$

Osoittajan laskemiseksi summataan yli vanhempien 1 ja 2 sekä kaksosten 3 ja 4 mahdollisten genotyyppien:

$$\begin{aligned} P(X_3 = 1, X_4 = 1) &= \sum_{h^1,h^2,h^3,h^4 \in \mathcal{H}} p(h^1)\, p(h^2)\, p(h^3, h^4 \mid h^1, h^2)\, f(h^3)\, f(h^4) \\ &= \sum_{h^1,h^2 \in \mathcal{H}} p(h^1)\, p(h^2) \sum_{h^3 \in \mathcal{H}} p(h^3 \mid h^1, h^2)\, f(h^3)^2. \end{aligned}$$

Jälkimmäisen lausekkeen johdossa käytettiin hyväksi tietoa, että identtisillä kaksosilla on aina yhteinen genotyyppi. Molempien summauksien suoraviivainen laskeminen vie ajan $\Theta(|\mathcal{H}|^3)$.

Tehokkaampaan laskentaan päästään, jos voidaan olettaa, että genotyyppien jakauma kaksosilla on sama kuin heidän vanhemmillaan, toisin sanoen

$$P(H^1 = h) = P(H^2 = h) = P(H^3 = h) = P(H^4 = h)$$

kaikilla $h \in \mathcal{H}$. Ensimmäinen ja viimeinen yhtäsuuruus ovat aina voimassa. Sen sijaan keskimmäinen yhtäsuuruus ei ole voimassa ilman sopivia oletuksia genotyyppien periytymisestä ja frekvensseistä vanhemmilla 1 ja 2. Voidaan osoittaa, että kytkennällisissä malleissa tasapainoehdot ja riippumaton tasainen periytyminen takaavat, että tasapainoehdot säilyvät voimassa myös perustajien lapsilla (tuloksen tarkempi käsittely sivuutetaan). Tästä seuraa, että kytkennällisissä malleissa $P(H^2 = h) = P(H^3 = h)$. Yksinkertainen summaus yli kaksosen 3 mahdollisten genotyyppien antaa nyt

$$P(X_3 = 1) \;=\; \sum_{h^3 \in \mathcal{H}} p(h^3)\, f(h^3) \;=\; K$$

ja

$$P(X_3 = 1, X_4 = 1) \;=\; \sum_{h^3 \in \mathcal{H}} p(h^3) f(h^3)^2.$$

Jo aiemmin osoitettiin, kuinka taudin yleisyys K voidaan laskea ajassa $\Theta(|\mathcal{G}|)$. On helppo nähdä, että yhteistodennäköisyys $P(X_3 = 1, X_4 = 1)$ voidaan laskea vastaavasti ja yhtä nopeasti.

5.1.3 Lapsiriski

Yleisissä malleissa lapsiriski K_O voidaan laskea samaan tapaa kuin identtisen kaksosen riski, jolloin aikavaativuudeksi saadaan sama $\Theta(|\mathcal{H}|^3)$. Seuraavaksi näytetään, kuinka kytkennällisen periytymisen malleissa tulosta voidaan jälleen parantaa.

Lapsiriskin määritelmän mukaan voidaan tarkastella sairastumisriskiä joko sairaan isän lapselle tai sairaan äidin lapselle. Seuraavassa tarkastellaan ensimmäistä vaihtoehtoa, tosin vanhemman sukupuolella ei ole merkitystä. Riippumattoman tasaisen periytymisen vuoksi lapsi jakaa isänsä kanssa aina täsmälleen puolet heidän järjestetyistä genotyypeistään. Toisen puolen he perivät riippumattomasti: lapsi äidiltään ja isä omilta vanhemmiltaan. Käytetään lapsen isältään perimästä alleelijonosta merkintää h', lapsen äidiltään perimästä alleelijonosta h'' sekä merkintää h''' siitä isän alleelijonosta, jota lapsi ei perinyt. Lapsen järjestetty genotyyppi on siten h'/h'', ja isän järjestetty genotyyppi on h'/h'''. Todennäköisyys, että sekä isä että lapsi ovat sairaita, on nyt

$$\begin{aligned} P(X_3 = 1, X_1 = 1) &= \sum_{h' \in \mathcal{A}} p(h') \sum_{h'' \in \mathcal{A}} p(h'')\, f(h'/h'') \sum_{h''' \in \mathcal{A}} p(h''')\, f(h'/h''') \\ &= \sum_{h' \in \mathcal{A}} p(h') (\sum_{h'' \in \mathcal{A}} p(h'')\, f(h'/h''))^2. \end{aligned}$$

Summaukset voidaan ottaa lokuksittain kaavan (5.1) tapaan, jolloin laskenta vie ajan $\Theta(|\mathcal{A}|^2) = \Theta(|\mathcal{H}|)$.

Lapsiriski K_O saadaan jakamalla edellä lausuttu yhteistodennäköisyys taudin yleisyydellä K. Perustelu on sama kuin identtisen kaksosen riskille. Pannaan merkille, että lokusten välinen kytkentä ei vaikuta lapsiriskin arvoon eikä laskemiseen.

5.1.4 Sisarusriski

Edellä käsiteltyihin riskeihin verrattuna sisarusriskin laskeminen vaikuttaa vaativalta. Lapsiriskiä laskettaessa riitti tarkastella genotyyppien jakautumista kolmella yksilöllä ja identtisen kaksosen tapauksessakin voitiin hyödyntää kaksosten genotyyppien identtisyyttä. Sisarusriskiä laskettaessa on huomioitava genotyyppien jakauma neljällä yksilöllä.

Aloitetaan kirjoittamalla sisarusriskille

$$K_S \;=\; P(X_4 = 1 \mid X_3 = 1) \;=\; \frac{P(X_4 = 1, X_3 = 1)}{P(X_3 = 1)}.$$

Nimittäjä $P(X_3 = 1)$ voidaan laskea vastaavasti kuin edellisissä osioissa sekä yleisissä että kytkennällisen periytymisen malleissa.

Osoittajan laskemiseksi summataan vanhempien 1 ja 2 sekä sisarusten 3 ja 4 genotyyppien yli, jolloin saadaan

$$\begin{aligned} P(X_3 = 1, X_4 = 1) \;&=\; \sum_{h^1 \in \mathcal{H}} p(h^1) \sum_{h^2 \in \mathcal{H}} p(h^2) \sum_{h^3 \in \mathcal{H}} f(h^3)\, p(h^3 \mid h^1, h^2) \\ &\quad\; \sum_{h^4 \in \mathcal{H}} f(h^4)\, p(h^4 \mid h^1, h^2) \\ &=\; \sum_{h^1 \in \mathcal{H}} p(h^1) \sum_{h^2 \in \mathcal{H}} p(h^2) \Big(\sum_{h^3 \in \mathcal{H}} f(h^3)\, p(h^3 \mid h^1, h^2) \Big)^2 . \end{aligned} \tag{5.2}$$

Summan suoraviivainen laskeminen vie ajan $\Theta(|\mathcal{H}|^3)$. Koska genotyyppien frekvenssit ja periytymistodennäköisyydet voivat olla mielivaltaisia, tämä on aikavaativuus yleisille malleille.

Kytkemättömän periytymisen malleissa laskentaa voidaan välittömästi tehostaa: järjestetyt genotyypit voidaan korvata järjestämättömillä genotyypeillä; voidaan hyödyntää isän 1 ja äidin 2 symmetrista asemaa; nähdään, että useille h^3 periytymistodennäköisyys $p(h^3 \mid h^1, h^2)$ on 0, kun h^1 ja h^2 ovat kiinnitetyt. Esimerkiksi kaksialleelisissa malleissa parannusten voidaan osoittaa johtavan aikavaativuuteen $\Theta(10^\ell)$, joka on hieman enemmän kuin neliöllinen suhteessa mallin parametrien määrään. Kaavan 5.2 suoraviivainen soveltaminen veisi sen sijaan kuutiollisen ajan $\Theta((3^\ell)^3) = \Theta(27^\ell)$, vaikka laskettaisiin järjestämättömillä genotyypeillä. Parannus on siis huomattava. Seuraavassa osiossa 5.2 näytetään, että toisenlaisella lähestymistavalla päästään vielä parempaan tulokseen – lähelle lineaarista vaativuutta. Edellä mainittuja tekniikoita ei siksi kuvata tässä tarkemmin.

5.2 Parannus sisarusriskin laskentaan

Tässä osiossa jatketaan sisarusriskin laskemisen tarkastelua kytkemättömän ja kytkennällisen periytymisen malleissa. Osoittautuu, että osiossa 5.1.4 annetun kaavan (5.2) summausjärjestyksen vaihtaminen mahdollistaa tehokkaan algoritmin laatimisen. Lähestymistapa johtaa osiossa 5.1 esitettyjä menetelmiä monimutkaisempaan laskentaan. Myöhemmin osoitetaan, että menetelmä yleistyy mielivaltaisille sukulaisuustyypeille.

Muistutetaan, että todennäköisyys $P(X_3 = 1)$ voidaan aikaisemmin esitetyn nojalla laskea yleisissä malleissa ajassa $\Theta(|\mathcal{H}|^3)$ ja kytkennällisissä malleissa ajassa $\Theta(|\mathcal{G}|)$. Seuraavassa keskitytään todennäköisyyden $P(X_3 = 1, X_4 = 1)$ laskemiseen.

Summaamalla sisarusten 3 ja 4 järjestettyjen genotyyppien yli saadaan

$$P(X_3 = 1, X_4 = 1) \;=\; \sum_{h^3 \in \mathcal{H}} \sum_{h^4 \in \mathcal{H}} p(h^3, h^4)\, f(h^3)\, f(h^4).$$

Määritellään nyt matriisi W^S kokoa $|\mathcal{H}| \times |\mathcal{H}|$, indeksoituna joukon $\mathcal{H}$ genotyypeillä, asettamalla

$$W^S(h^3, h^4) \;=\; p(h^3, h^4).$$

Kun ilmenemisyleisyysfunktio f tulkitaan sarakevektoriksi, matriisimerkinnöin voidaan kirjoittaa

$$P(X_3 = 1, X_4 = 1) \;=\; f^T W^S f, \tag{5.3}$$

missä yläindeksi T viittaa vektorin transpoosiin. Tätä lauseketta kutsutaan *neliömuodoksi.*

Jos matriisi W^S tunnettaisiin, kaavan (5.3) suoraviivainen laskeminen veisi ajan $\Theta(|\mathcal{H}|^2)$. Vaikeutena onkin matriisin W^S muodostaminen. Ongelman ratkaisemiseksi seuraavaksi rajoitutaan kytkennällisten mallien ja kytkemättömien mallien perheisiin. Osoitetaan, että hyvän hajotelman löytäminen matriisille W^S ei ainoastaan mahdollista matriisin nopeaa laskemista, vaan se myös tarjoaa tehokkaan keinon lausekkeen (5.3) laskemiseen. Ensin kuitenkin esitellään sovellettava matriisihajotelma ja sitä hyödyntävä algoritmi yleisesti.

5.2.1 Kroneckerin tulo ja Yatesin algoritmi

Olkoot m ja n luonnollisia lukuja. Matriisin $A = (a_{ij})_{m\times m}$ ja matriisin $B = (b_{kl})_{n\times n}$ *Kroneckerin tulo* määritellään $mn \times mn$ -matriisina $(C_{ij})_{m\times m}$, jolle $C_{ij} = a_{ij}\, B$. Kun esimerkiksi $m = n = 2$, on voimassa

$$A \otimes B \;=\; \begin{pmatrix} a_{11}\, B & a_{12}\, B \\ a_{21}\, B & a_{22}\, B \end{pmatrix} \;=\; \begin{pmatrix} a_{11}\, b_{11} & a_{11}\, b_{12} & a_{12}\, b_{11} & a_{12}\, b_{11} \\ a_{11}\, b_{21} & a_{11}\, b_{22} & a_{12}\, b_{21} & a_{12}\, b_{22} \\ a_{21}\, b_{11} & a_{21}\, b_{12} & a_{22}\, b_{11} & a_{22}\, b_{11} \\ a_{21}\, b_{21} & a_{21}\, b_{22} & a_{22}\, b_{21} & a_{22}\, b_{22} \end{pmatrix},$$

missä symboli $\otimes$ tarkoittaa Kroneckerin tuloa. Tätä tavallisesta matriisitulosta poikkeavaa operaatiota nimitetään toisinaan myös *suoraksi tuloksi* tai *tensorituloksi* (lisätietoja Kroneckerin tulon ominaisuuksista löytyy mm. teoksista [Sea66, MaM92]).

Tarkastellaan sitten ongelmaa, jossa tehtävänä on laskea matriisikertolasku

$$C\, v, \tag{5.4}$$

missä C on $r \times r$ -matriisi ja v on $r \times 1$ -vektori. Tuloksena on siis $r \times 1$ -vektori. Tavallisesti laskenta vaatii r^2 kertolaskua. Jos kuitenkin on erikoisesti voimassa

$$C \;=\; A \otimes B,$$

missä $A = (a_{ij})_{m\times m}$ ja $B = (b_{kl})_{n\times n}$, tätä erityistä rakennetta hyödyntämällä tulo (5.4) voidaan laskea $mn(m+n)$ operaatiolla $(mn)^2$ operaation sijaan.

Yatesin algoritmi on yksi tapa saavuttaa tämä säästö. Yates [Yat37] esitti klassisen algoritminsa alkuaan faktoristen kokeiden laskennalliseen analyysiin. Selkeän matriisimuotoon puetun yleistyksen ja tarvittavat todistukset antoi Good [Goo58]. Bureau et al. [BDS98] sovelsivat algoritmia erääseen kytkentäanalyysin laskennalliseen ongelmaan, mutta kuvasivat algoritmin yleisellä tasolla. Seuraava esitys noudattelee pääosin heidän muotoiluaan.

Yatesin algoritmissa matriisit A ja B muunnetaan harvoiksi, eli paljon nollia sisältäviksi, $mn \times mn$ -matriiseiksi. Määritellään

$$\widetilde{A} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \\ n+1 \\ n+2 \\ \vdots \\ \vdots \\ mn \end{array} \begin{pmatrix} a_{11} \dots a_{1m} & 0 & \cdots & 0 \\ 0 & a_{11} \dots a_{1m} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{11} \dots a_{1m} \\ a_{21} \dots a_{2m} & 0 & \cdots & 0 \\ 0 & a_{21} \dots a_{2m} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{m1} \dots a_{mm} \end{pmatrix} \tag{5.5}$$

ja $\widetilde{B}$ vastaavasti. Muunnoksen etu on, että Kroneckerin tulo voidaan nyt esittää tavallisena matriisitulona:

$$A \otimes B \;=\; \widetilde{A}\,\widetilde{B}. \tag{5.6}$$

Tämän todistamiseksi riittää verrata Kroneckerin tulon ja tavallisen matriisitulon tuloksena saatavien matriisien alkioita (tarkempi todistus sivuutetaan). Matriisitulon liitännäisyyden vuoksi voidaan nyt kirjoittaa

$$C\,v \;=\; \widetilde{A}\,(\widetilde{B}\,v),$$

ja siten riittää suorittaa kaksi peräkkäistä matriisikertolaskua, jossa ensimmäinen tekijä on harva matriisi ja toinen tekijä on vektori. Huomautettakoon, että $\widetilde{B}\,v$ on vektori, jonka pituus on sama kuin vektorin v pituus.

Kuvan 5.3 algoritmi selvittää yksityskohtaisesti kertolaskun askeleet. Aliohjelma YatesinTulo saa syötteenä alkuperäistä muotoa olevan matriisin ja kerrottavan vektorin. Algoritmin oikeellisuus on selvä metriisikertolaskun määritelmän ja edellä annetun harvan esityksen (5.5) nojalla. Helposti nähdään, että tarvittavien kerto- ja yhteenlaskujen lukumäärä on $mmn = |A||v|$, missä $|A|$ on matriisin A rivien määrä ja $|v|$ on vektorin v pituus (indeksien laskemista ei huomioida).

Ennen kuin esitetään yksityiskohtaisesti, kuinka aliohjelmaa YatesinTulo hyödynnetään iteratiivisesti, yleistetään edellä kuvattu menetelmä moninkertaiselle Kroneckerin tulolle. Tehtävänä on siis laskea muotoa

$$(A_1 \otimes A_2 \otimes \cdots \otimes A_k)\, v$$

```
YatesinTulo(A, v)
m := |A|
n := |v|/m
jokaiselle i := 1 , . . . , m :
    jokaiselle j := 1 , . . . , n :
        h := (i − 1)n + j
        u_h := 0
        jokaiselle k := 1 , . . . , m :
            u_h := u_h + v_h a_ik
palauta u
```

Kuva 5.3: Algoritmi muotoa $\widetilde{A}\, v$ olevan tulon laskemiseen. Matriisin A rivien määrää merkitään $|A|$ ja vektorin v pituutta $|v|$.

oleva tulo. Helpolla induktiolla nojautuen tulokseen (5.6) nähdään, että jokaiselle A_i on olemassa matriisi $\widetilde{A}_i$ siten, että

$$A_1 \otimes A_2 \otimes \cdots \otimes A_k \quad = \quad \widetilde{A}_1\, \widetilde{A}_2 \cdots \widetilde{A}_k .$$

Matriisikertolasku saa nyt iteratiivisen muodon

$$\widetilde{A}_1\, (\widetilde{A}_2 \cdots (\widetilde{A}_k\, v)).$$

Kuvan 5.4 algoritmi Yates laskee tämän suoraviivaisesti. Aiemmin esitetyn huomion nojalla rivin 1 tai 3 suorittaminen vaatii $|A_i||v|$ operaatiota, kun $i = 1, 2, \ldots, k$. Siten operaatioita suoritetaan yhteensä

$$|v| \sum_{i=1}^{k} |A_i| \quad = \quad \Big(\prod_{i=1}^{k} |A_i| \Big) \Big(\sum_{i=1}^{k} |A_i| \Big). \tag{5.7}$$

Jos erikoisesti $|A_i| = |v|^{1/k}$ kaikilla i, tarvittavien operaatioiden määrä voidaan kirjoittaa muodossa $k|v|^{1+1/k}$.

```
Yates(A_1, A_2, . . . , A_k, v)
1. u := YatesinTulo(A_k, v)
2. jokaiselle i := k − 1 , . . . , 1 :
3.     u := YatesinTulo(A_i, u)
4. palauta u
```

Kuva 5.4: Yatesin algoritmi k-kertaiselle Kroneckerin tulolle.

Pannaan vielä merkille, että aliohjelman YatesinTulo suorittamat laskut voidaan palauttaa tavanomaiseen matriisikertolaskuun, jossa toinen tekijä ei ole sarakevektori. Tämä tehdään tulkitsemalla $1 \times n$ -vektori v sellaisena $m \times n/m$ -matriisina V, jolle $V_{ij} = v_{i+(j-1)m}$. Vektori $u = \widetilde{A}v$ vastaa tällöin matriisia $U = AV$ vastaavuudella $U_{ij} = u_{i+(j-1)m}$. Voidaan siis soveltaa matriisien kertomiseen kehitettyjä algoritmeja. Esimerkiksi kahden $m \times m$ -matriin tulo voidaan laskea ajassa $O(m^{2.376})$ [CoW90]. Tällaisten asymptoottisten tulosten merkitys Yatesin algoritmin parantamiselle on kuitenkin teoreettinen – käytännössä suoraviivaisempi laskenta aliohjelman YatesinTulo tapaan vaikuttaa nopeammalta. Näin on erityisesti silloin, jos n/m on paljon suurempi kuin m.

5.2.2 Kytkemättömän periytymisen mallit

Palataan nyt neliömuotoon (5.3), ja näytetään, kuinka kytkemättömän periytymisen malleissa voidaan hyödyntää edellisessä osiossa esitettyä Yatesin algoritmia.

Kun lokusten välillä ei ole kytkentää, sisarusten genotyyppiparit (h_i^3, h_i^4) lokuksissa i ovat riippumattomia muiden lokusten genotyyppipareista. Todennäköisyys, että ensimmäisellä sisaruksella on genotyyppi $h^3 = (h_1^3, \ldots, h_\ell^3)$ ja toisella sisaruksella on genotyyppi $h^4 = (h_1^4, \ldots, h_\ell^4)$, voidaan tällöin kirjoittaa tulona:

$$W^S(h^3, h^4) \quad = \quad p(h^3, h^4) = \prod_{i=1}^{\ell} p(h_i^3, h_i^4). \tag{5.8}$$

Määritellään sitten matriisit W_i^S kokoa $|\mathcal{H}_i| \times |\mathcal{H}_i|$, indeksoituna joukkojen $\mathcal{H}_i$ genotyypeillä, asettamalla

$$W_i^S(h_i^3, h_i^4) \quad = \quad p(h_i^3, h_i^4).$$

Kaavan (5.8) nojalla matriisi W^R on matriisien W_i^R Kroneckerin tulo. Yhtälö (5.3) voidaan siten kirjoittaa muodossa

$$P(X_3 = 1, X_4 = 1) \quad = \quad f^T(W_1^S \otimes \cdots \otimes W_\ell^S)f. \tag{5.9}$$

Nähdään, että edellisessä osiossa esitetty Yatesin algoritmi mahdollistaa kaavan (5.9) tehokkaan laskemisen edellyttäen, että matrsiisit W_i^S tunnetaan. Tuloksen (5.7) nojalla tulon $(W_1^S \otimes \cdots \otimes W_\ell^S)f$ laskeminen vie ajan

$\Theta(|\mathcal{H}| \sum_i |\mathcal{H}_i|)$. Tämän jälkeen Yatesin algoritmin laskema vektori on vielä kerrottava transpoosilla f^T, mikä onnistuu ajassa $O(|\mathcal{H}|)$. Panemalla merkille, että järjestettyjen genotyyppien sijaan voidaan laskea järjestämättömillä genotyypeillä, aikavaativuudeksi saadaan $\Theta(|\mathcal{G}| \sum_i |\mathcal{G}_i|)$. Kaksialleelisissa malleissa aikavaativuus on siten $\Theta(\ell\, 3^\ell)$.

On vielä osoitettava, että matriisit W_i^S voidaan laskea nopeasti. Intuitiivisesti tämä vaikuttaa mahdolliselta, koska eri lokuksille i matriisit voidaan käsitellä toisistaan riippumatta. Matriisien laskemiseen annetaan mielivaltaisille sukulaisuustyypeille soveltuva menetelmä osiossa 5.3.1.

5.2.3 Kytkennällisen periytymisen mallit

Tarkastellaan sitten todennäköisyyden $P(X_3 = 1, X_4 = 1)$ laskemista kytkennällisen periytymisen malleissa. Koska lokusten välillä sallitaan kytkentä, kaavan (5.9) hajotelma ei ole käytettävissä. Toisenlainen Kroneckerin tulo -hajotelma osoittautuu kuitenkin hyödylliseksi.

Koska isän ja äidin alleelit periytyvät riippumattomasti, on voimassa

$$W^S(h^3, h^4) \;\; = \;\; p(h^3_{\cdot 1}, h^4_{\cdot 1})\, p(h^3_{\cdot 2}, h^4_{\cdot 2}).$$

Määritellään siksi matriisit $W^S_{\cdot 1}$ ja $W^S_{\cdot 2}$ kokoa $|\mathcal{A}| \times |\mathcal{A}|$, indeksoituna joukon $\mathcal{A}$ alleelijonoilla, asettamalla

$$W^S_{\cdot j}(h^3_{\cdot j}, h^4_{\cdot j}) \;\; = \;\; p(h^3_{\cdot j}, h^4_{\cdot j}),$$

missä $j = 1$ tai $j = 2$ vastaten sisarusten isää tai äitiä. Yhtälö (5.3) saa nyt muodon

$$P(X_3 = 1, X_4 = 1) \;\; = \;\; f^T(W^S_{\cdot 1} \otimes W^S_{\cdot 2})f. \tag{5.10}$$

Soveltamalla Yatesin algoritmia aikavaativuudeksi saadaan $\Theta(|\mathcal{H}|\,(|\mathcal{A}| + |\mathcal{A}|)) = \Theta(|\mathcal{H}|^{\frac{3}{2}})$. Tämä tulos tietenkin edellyttää, että matriisit $W^S_{\cdot j}$ tunnetaan.

Ennen lopullisen aikavaativuuden julistamista onkin osoitettava, että matriisit $W^S_{\cdot j}$ voidaan laskea riittävän nopeasti eli ajassa $O(|\mathcal{H}|^{\frac{3}{2}})$. Intuitiivisesti ei ole selvää, kuinka vaativaa matriisien laskeminen on. Seuraavaksi johdetaan riittävän nopea mutta melko monimutkainen algoritmi.

Koska periytymisessä isältä ja äidiltä ei ole eroa, matriisit $W^S_{\cdot 1}$ ja $W^S_{\cdot 2}$ ovat identtiset ja riittää tarkastella genotyyppien periytymistä isältä. Aloitetaan summaamalla yli sisarusten isän mahdollisten genotyyppien. Rajoittumalla vain oleellisiin genotyyppeihin g^1, kun alleelijonot $h^3_{\cdot 1}$ ja $h^4_{\cdot 1}$ on kiinnitetty, saadaan

$$W^S_{\cdot 1}(h^3_{\cdot 1}, h^4_{\cdot 1}) \;=\; \sum_{g^1 \in \mathcal{G}^1(h^3_{\cdot 1}, h^4_{\cdot 1})} p(g^1)\, p(h^3_{\cdot 1}, h^4_{\cdot 1} \mid g^1). \tag{5.11}$$

Merkintä $\mathcal{G}^1(h^3_{\cdot 1}, h^4_{\cdot 1})$ tarkoittaa joukkojen

$$\mathcal{G}^1_i(h^3_{i1}, h^4_{i1}) = \left\{g^1_i \in \mathcal{G}_i : p(h^3_{i1}, h^4_{i1} \mid g^1) > 0\right\}, \quad i = 1, \ldots, \ell$$

karteesista tuloa.

Käyttämällä ketjusääntöä, kytkentätasapainoa ja tietoa, että sisarusten alleelit lokuksessa i ovat riippumattomia isän genotyypeistä lokuksissa $i+1, \ldots, \ell$, saadaan

$$\begin{aligned} W^S_{\cdot 1}(h^3_{\cdot 1}, h^4_{\cdot 1}) \;=\; & \sum_{g^1_1 \in \mathcal{G}^1_1(h^3_{11}, h^4_{11})} p(g^1_1)\, p(h^3_{11}, h^4_{11} \mid g^1_1) \\ & \sum_{g^1_2 \in \mathcal{G}^1_2(h^3_{21}, h^4_{21})} p(g^1_2)\, p(h^3_{21}, h^4_{21} \mid g^1_{[2]}, h^3_{[1]1}, h^4_{[1]1}) \\ & \cdots \sum_{g^1_\ell \in \mathcal{G}^1_\ell(h^3_{\ell 1}, h^4_{\ell 1})} p(g^1_\ell)\, p(h^3_{\ell 1}, h^4_{\ell 1} \mid g^1_{[\ell]}, h^3_{[\ell-1]1}, h^4_{[\ell-1]1}). \end{aligned} \tag{5.12}$$

Tässä $h^3_{[\ell-1]1}$ on lyhennysmerkintä jonolle $(h^3_{11}, \ldots, h^3_{\ell-11})$, vastaavasti $g^1_{[2]}$ on lyhennysmerkintä jonolle $(g^1_1, \ldots, g^1_2)$ jne. Ongelmana on nyt etsiä käyttökelpoinen lauseke todennäköisyyksille $p(h^3_{i1}, h^4_{i1} \mid g^1_{[i]}, h^3_{[i-1]1}, h^4_{[i-1]1})$. Ratkaisu löytyy luokittelemalla erilaiset periytymisvaihtoehdot muutamaan oleelliseen tapaukseen. Hahmotellaan tätä esimerkkien avulla.

Esimerkki 5.1. Lasketaan todennäköisyys $p(h^3_{31}, h^4_{31} \mid g^1_{[3]}, h^3_{[2]1}, h^4_{[2]1})$, kun $h^3_{31} = c$, $h^4_{31} = C$, $g^1_{[3]} = AaBBCc$, $h^3_{[2]1} = AB$ ja $h^4_{[2]1} = AB$. Tehdään ensin huomio, että isällä on neljä mahdollista järjestettyä genotyyppiä: $ABC/aBc, ABc/aBC, aBC/ABc$ ja aBc/ABC. Jos oletetaan, että ensimmäinen vaihtoehto ($h^1 = ABC/aBc$) on voimassa, voidaan todeta seuraavat seikat:

1. Jotta sisarukselle 3 periytyisi alleeli c, on lokusten 1 ja 3 välillä tapahduttava tekijänvaihto.

2. Jotta sisarukselle 4 periytyisi alleeli C, ei lokusten 1 ja 3 välillä saa tapahtua tekijänvaihtoa.

Näiden kahden tapahtuman yhteistodennäköisyys on $\theta_{13}\,(1-\theta_{13})$. Tämä pätee myös silloin, jos isän järjestetty genotyyppi on symmetrisesti aBc/ABC. Jos isän järjestetty genotyyppi on ABc/aBC tai symmetrisesti aBC/ABc, niin vastaavalla päättelyllä saadaan periytymistodennäköisyydeksi jälleen sama $\theta_{13}\,(1-\theta_{13})$. Koska eri vaihtoehdot isän järjestetylle genotyypille ovat yhtä todennäköisiä, päädytään tulokseen

$$p(c, C \mid AaBBCc, AB, AB) \;=\; \theta_{13}\,(1-\theta_{13}).$$

□

Esimerkki 5.2. Tarkastellaan esimerkin 5.1 muunnelmaa, jossa $h^4_{31} = c$. Oletetaan jälleen ensin, että isän järjestetty genotyyppi on ABC/aBc. Jotta sisaruksille periytyisi alleeli c, on lokusten 1 ja 3 välillä tapahduttava tekijänvaihto molemmissa periytymistapahtumissa. Tämän todennäköisyys on θ^2_{13}. Jos isän järjestetty genotyyppi onkin ABc/aBC, niin lokusten 1 ja 3 välillä ei saa tapahtua tekijänvaihtoa. Tämän todennäköisyys on $(1-\theta_{13})^2$. Kummankin käsitellyn vaihtoehdon todennäköisyys on symmetria huomioiden $\frac{1}{2}$. Laskemalla yhteen saadaan

$$p(c, c \mid AaBBCc, AB, AB) \;=\; \frac{1}{2}\,\theta^2_{13} + \frac{1}{2}\,(1-\theta_{13})^2.$$

□

Esimerkkien 5.1 ja 5.2 huomioiden yleistämiseksi otetaan ensin käyttöön joitakin lyhentäviä ilmaisuja. Sanotaan, että j on *lokusta i edeltävä genotyypin g heterotsygootti lokus*, jos

$$j \;=\; \max\bigl(\{k : 0 < k < i \text{ ja } g_k \text{ on heterotsygootti}\} \cup \{0\}\bigr).$$

Erikoistapauksessa siis $j = 0$, jos i on genotyypin g ensimmäinen heterotsygootti lokus. Kirjoitetaan lisäksi $\underline{\text{tai}}$, kun tarkoitetaan poissulkevaa tai-operaatiota (engl. exclusive-or, lyh. xor).

Lemma 5.1. *Olkoot $h^3_{\cdot 1}$ ja $h^4_{\cdot 1}$ sisarusten isältä perityt alleelijonot ja olkoon $g^1 \in \mathcal{G}^1(h^3_{\cdot 1}, h^4_{\cdot 1})$ isän järjestämätön genotyyppi. Silloin kaikilla lokuksilla i on voimassa*

$$p(h^3_{i1}, h^4_{i1} \mid g^1_{[i]}, h^3_{[i-1]1}, h^4_{[i-1]1}) = \begin{cases} 1, & \text{jos } g^1_i \text{ on homotsygootti} \\ \frac{1}{4}, & \text{muuten jos } j = 0 \\ \theta_{ji}\,(1-\theta_{ji}), & \text{muuten jos } h^3_{j1} \neq h^4_{j1} \ \underline{\text{tai}}\ h^3_{i1} \neq h^4_{i1} \\ \frac{1}{2}\,\theta^2_{ji} + \frac{1}{2}\,(1-\theta_{ji})^2, & \text{muuten} \end{cases},$$

missä j on lokusta i edeltävä genotyypin g^1 heterotsygootti lokus.

Lemman todistus voidaan jakaa tapauksiin väitteen mukaisesti. Kaksi ensimmäistä tapausta ovat helppoja, lähes triviaaleja. Kaksi jälkimmäistä tapausta voidaan käsitellä esimerkkien 5.1 ja 5.2 tapaan. Täsmällinen todistus sivuutetaan tässä sen pituuden ja yksitotisuuden vuoksi.

Lemma 5.1 ei ainoastaan tarjoa keinoa ehdollisten periytymistodennäköisyyksien laskemiseen, vaan se myös mahdollistaa kaavassa (5.12) summattavien termien karsimisen. Nimittäin jokaiselle i todennäköisyys $p(h^3_{i1}, h^4_{i1} \mid g^1_{[i]}, h^3_{[i-1]1}, h^4_{[i-1]1})$ riippuu ainoastaan genotyypin g^1_i tyypistä (homotsygootti vai heterotsygootti) – ei genotyypin alleeleista. Summausindeksit g^1_i pelkistyvät seuraaviin tapauksiin:

1. Jos $h^3_{i1} \neq h^4_{i1}$, silloin välttämättä g^1_i on heterotsygootti $\{h^3_{i1}, h^4_{i1}\}$.
2. Jos $h^3_{i1} = h^4_{i1}$, silloin g^1_i voi olla joko homotsygootti tai heterotsygootti. Edellinen on totta todennäköisyydellä $p(h^3_{i1})^2$ ja jälkimmäinen todennäköisyydellä $p(h^3_{i1})\,(1 - p(h^3_{i1}))$. Riittääkin summata yli näiden kahden vaihtoehdon.

Kuvan 5.5 algoritmi hyödyntää edellistä huomiota. Kaavaan (5.12) nojautuen se laskee matriisin $W^S_{\cdot 1}$ alkiot käyttäen rekursiivista aliohjelmaa Alkio. Pienen ongelman lemman 5.1 soveltamisessa tuottaa tekijänvaihtotodennäköisyyksien θ_{ji} laskeminen. Aliohjelmassa Alkio tekijänvaihtotodennäköisyydet lasketaan iteratiivisesti Trow'n kaavaa käyttäen (ks. kaava (3.7)). Rivin 2 iteraatioaskeleen jälkeen muuttuja r kertoo lokuksen i ja sitä edeltävän heterotsygootin lokuksen tekijänvaihtotodennäköisyyden. Toinen tekninen ongelma on pitää yllä tietoa sisarusten alleelien välisistä suhteista heterotsygooteissa lokuksissa. Aliohjelmassa Alkio muuttujan b arvo on 'tosi', jos sisaruksilla on eri alleelit edellisessä heterotsygootissa lokuksessa, muuten arvo on 'epätosi'.

Alkio$(h^3_{\cdot 1}, h^4_{\cdot 1}, i, r, b)$
1. **jos** $i > \ell$ **niin palauta** 1
2. $r := \frac{1}{2}(1 - (1 - 2r)(1 - 2\theta_i)$
3. **jos** b <u>tai</u> $h^3_{i1} \neq h^4_{i1}$ **niin**
4. $t := r(1 - r)$
5. **muuten**
6. $t := \frac{1}{2}r^2 + \frac{1}{2}(1 - r)^2$
7. **jos** $h^3_{i1} \neq h^4_{i1}$ **niin** // $g^1_i = \{h^3_{i1}, h^4_{i1}\}$
8. $s := 2\,p(h^3_{i1})\,p(h^4_{i1})\,t$ Alkio$(h^3_{\cdot 1}, h^4_{\cdot 1}, i + 1, 0,$ tosi$)$
9. **muuten**
10. $s := 2\,p(h^3_{i1})\,(1 - p(h^3_{i1}))\,t$ Alkio$(h^3_{\cdot 1}, h^4_{\cdot 1}, i + 1, 0,$ epätosi$)$ // g^1_i heterots.
11. $s := s + p(h^3_{i1})^2$ Alkio$(h^3_{\cdot 1}, h^4_{\cdot 1}, i + 1, r, b)$ // g^1_i homots.
12. **palauta** s

Kuva 5.5: Algoritmi alkion $W^S_{\cdot 1}(h^3_{\cdot 1}, h^4_{\cdot 1})$ laskemiseen kytkennällisessä mallissa.

Aliohjelman Alkio ensimmäisellä kutsukerralla $i = 1$, mutta r ja b saavat olla mielivaltaisia. Nimittäin sopimuksen $\theta_1 = \frac{1}{2}$ nojalla r saa rivillä 2 joka tapauksessa arvon $\frac{1}{2}$. Koska rivit 4 ja 6 johtavat nyt samaan sijoitukseen $t := \frac{1}{4}$, ei muuttujan b arvolla ole merkitystä. Rivin 11 vuoksi muuttuja b vaikuttaa seuraavaan rekursiotasoon, mutta tässä rekursiohaarassa g^1_i on homotsygootti. Muuttuja b saa heti oikean arvon, kun edetään haaraan, jossa g^1_i on heterotsygootti (rivit 8 ja 10).

Aliohjelman Alkio tarvitsemien operaatioiden määrää riippuu alleelijonoista $h^3_{\cdot 1}$ ja $h^4_{\cdot 1}$. Rekursiivisten kutsujen määrä on summa kullakin rekursiotasolla $1, 2, \ldots, \ell$ tehtävistä kutsuista. Tasolla k rekursiivisia kutsuja tarvitaan

$$\prod_{i=1}^{k}(1 + 1_{\{h^3_{i1} = h^4_{i1}\}}).$$

Rekursiivisten kutsujen määrä on siten yhteensä

$$\sum_{k=1}^{\ell}\prod_{i=1}^{k}(1 + 1_{\{h^3_{i1} = h^4_{i1}\}}).$$

Aliohjelmaa Alkio kutsutaan kuvan 5.6 aliohjelmasta Matriisi, joka laskee matriisin $W^S_{\cdot 1}$ alkiot yksitellen. Aliohjelman Alkio kutsujen kokonaismääräksi tulee

$$\begin{aligned}
\sum_{h^3_{\cdot 1}\in\mathcal{A}}\sum_{h^4_{\cdot 1}\in\mathcal{A}}\sum_{k=1}^{\ell}\prod_{i=1}^{k}(1+1_{\{h^3_{i1}=h^4_{i1}\}}) &= \sum_{k=1}^{\ell}\prod_{i=1}^{k}\sum_{h^3_{i1}\in\mathcal{A}_i}\sum_{h^4_{i1}\in\mathcal{A}_i}(1+1_{\{h^3_{i1}=h^4_{i1}\}}) \\
&= \sum_{k=1}^{\ell}\prod_{i=1}^{k}(2\,|\mathcal{A}_i|+|\mathcal{A}_i|^2-|\mathcal{A}_i|) \\
&= \sum_{k=1}^{\ell}\prod_{i=1}^{k}|\mathcal{A}_i|^2\left(1+\frac{1}{|\mathcal{A}_i|}\right).
\end{aligned}$$

Koska $\prod_{i=1}^{\ell}|\mathcal{A}_i|^2 = |\mathcal{H}|$ ja koska $|\mathcal{A}_i| \geq 2$ kaikilla i, kutsujen määrän ylärajaksi saadaan

$$\begin{aligned}
\sum_{k=1}^{\ell}\left(\frac{1}{4}\right)^{\ell-k}|\mathcal{H}|\left(1+\frac{1}{2}\right)^{k} &= \left(\frac{1}{4}\right)^{\ell}|\mathcal{H}|\sum_{k=1}^{\ell}6^k \\
&< \frac{6}{5}\left(\frac{3}{2}\right)^{\ell}|\mathcal{H}|.
\end{aligned}$$

Siten aliohjelman Matriisi aikavaativuus on $O((\frac{3}{2})^{\ell}\,|\mathcal{H}|)$. Tämä on vähemmän kuin Yatesin algoritmin vaativuus $\Theta(|\mathcal{H}|^{3/2})$, sillä $|\mathcal{H}|^{1/2} = |\mathcal{A}| \geq 2^{\ell} > (\frac{3}{2})^{\ell}$.

Matriisi
1. **jokaiselle** $h^3_{\cdot 1} \in \mathcal{A}$ **:**
2. **jokaiselle** $h^4_{\cdot 1} \in \mathcal{A}$ **:**
3. $W^S_{\cdot 1}(h^3_{\cdot 1}, h^4_{\cdot 1}) :=$ Alkio$(h^3_{\cdot 1}, h^4_{\cdot 1}, 1, \frac{1}{2},$ epätosi$)$
4. **palauta** $W^S_{\cdot 1}$

Kuva 5.6: Algoritmi matriisin $W^S_{\cdot 1}$ laskemiseen kytkennällisen periytymisen malleissa.

Ollaan johdettu menetelmä, jolla sisarusriski voidaan kytkennällisen periytymisen malleissa laskea ajassa $\Theta(|\mathcal{H}|^{3/2})$.

5.2.4 Laajennus: kaksilapsinen ydinperhe

Sisarusriskin laskemiseen edellä johdettu menetelmä voidaan helposti laajentaa sukupuuhavaintojen uskottavuuden laskemiseen kaksilapsiselle ydinperheelle. Numeroidaan jälleen isä, äiti, vanhempi sisar ja nuorempi sisar luvuilla 1, 2, 3 ja 4, tässä järjestyksessä. Tavoitteena on nyt laskea todennäköisyys

$$p(x_1,x_2,x_3,x_4) = P(X_1=x_1, X_2=x_2, X_3=x_3, X_4=x_4)$$

missä jono $x = (x_1, x_2, x_3, x_4)$ yksilöiden havaittuja ilmiasuja on annettu. Tässä siis vaaditaan, että jokainen perheenjäsen on havaittu sairaaksi tai terveeksi. Rajoitus on näennäinen ja se tehdään ainoastaan merkintöjen yksinkertaistamiseksi.

Aloitetaan määrittelemällä funktio f_{x_j}, joka kertoo yksilön j ilmiasun x_j todennäköisyyden, kun yksilön genotyyppi tunnetaan. Ilmenemisyleisyyksien avulla asetetaan $f_{x_j}(h) = f(h)$, jos $x_j = 1$, ja $f_{x_j}(h) = 1 - f(h)$, jos $x_j = 0$. (Jos yksilön j ilmiasu olisi tuntematon, voitaisiin asettaa $f_{x_j}(h) = 1$ kaikilla $h \in \mathcal{H}$.)

Muodostetaan sitten neliömuotoa (5.10) vastaava lauseke. Määritellään siksi matriisi W_{x_1,x_2} kokoa $|\mathcal{H}| \times |\mathcal{H}|$, indeksoituna joukon $\mathcal{H}$ genotyypeillä, asettamalla

$$W_{x_1,x_2}(h^3, h^4) \;=\; P(H^3 = h^3, H^4 = h^4, X_1 = x_1, X_2 = x_2)$$

sekä matriisit $W^{(j)}_{x_j}$ kokoa $|\mathcal{A}| \times |\mathcal{A}|$, indeksoituna joukon $\mathcal{A}$ genotyypeillä, asettamalla

$$W^{(j)}_{x_j}(h^3_{\cdot j}, h^4_{\cdot j}) \;=\; P(H^3_{\cdot j} = h^3_{\cdot j}, H^4_{\cdot j} = h^4_{\cdot j}, X_j = x_j).$$

Nyt seuraa, että

$$W_{x_1,x_2} \;=\; W^{(1)}_{x_1} \otimes W^{(2)}_{x_2},$$

ja siten

$$p(x_1, x_2, x_3, x_4) \;=\; f^T_{x_3}(W^{(1)}_{x_1} \otimes W^{(2)}_{x_2}) f_{x_4}.$$

Kaava (5.13) voidaan laskea Yatesin algoritmia käyttäen ajassa $\Theta(|\mathcal{H}|^{\frac{3}{2}})$ edellyttäen, että matriisit $W^{(j)}_{x_j}$ tunnetaan.

Matriisit $W^{(1)}_{x_1}$ ja $W^{(2)}_{x_2}$ eivät ole yleisesti samoja, koska voi olla $x_1 \neq x_2$. Laskentamenetelmää ajatellen niitä voidaan kuitenkin kohdella yhdenmukaisesti ja siksi seuraavaksi tarkastellaan ainoastaan isään liittyvää matriisia. Kaavan (5.11) tapaan matriisin $W^{(1)}_{x_1}$ alkiot voidaan esittää summauksena yli isän mahdollisten genotyyppien:

$$\begin{aligned} W^{(1)}_{x_1}(h^3_{\cdot 1}, h^4_{\cdot 1}) &= \sum_{g^1 \in \mathcal{G}^1(h^3_{\cdot 1}, h^4_{\cdot 1})} p(g^1)\, p(h^3_{\cdot 1}, h^4_{\cdot 1}, x_1 \mid g^1) \\ &= \sum_{g^1 \in \mathcal{G}^1(h^3_{\cdot 1}, h^4_{\cdot 1})} p(g^1)\, p(h^3_{\cdot 1}, h^4_{\cdot 1} \mid g^1)\, f_{x_1}(g^1). \end{aligned}$$

Kaavaa (5.12) vastaten saadaan edelleen

$$\begin{aligned} W_{x_1}^{(1)}(h_{\cdot 1}^3, h_{\cdot 1}^4) \;=\; & \sum_{g_1^1 \in \mathcal{G}_1^1(h_{11}^3, h_{11}^4)} p(g_1^1)\, p(h_{11}^3, h_{11}^4 \mid g_1^1) \\ & \cdots \sum_{g_\ell^1 \in \mathcal{G}_\ell^1(h_{\ell 1}^3, h_{\ell 1}^4)} p(g_\ell^1)\, p(h_{\ell 1}^3, h_{\ell 1}^4 \mid g_{[\ell]}^1, h_{[\ell-1]1}^3, h_{[\ell-1]1}^4)\, f_{x_1}(g^1). \end{aligned}$$

Summan viimeisen termin, $f_{x_1}(g^1)$, vuoksi nyt ei voida rajoittua tarkastelemaan ainoastaan genotyyppien g^1 tyyppejä kuten kaavassa (5.12). On summattava yli kaikkien mahdollisten genotyyppien. Algoritmi summalausekkeen laskemiseen esitetään kuvassa 5.7. ALKIO-2 on aliohjelman ALKIO laajennus.

ALKIO-2$(h_{\cdot 1}^3, h_{\cdot 1}^4, x_1, i, r, b)$
1. **jos** $i > \ell$ **niin palauta** $f_{x_1}(g^1)$
2. $r := \frac{1}{2}(1-(1-2r)(1-2\theta_i))$
3. **jos** b tai $h_{i1}^3 \neq h_{i1}^4$ **niin**
4. $\quad t := r(1-r)$
5. **muuten**
6. $\quad t := \frac{1}{2}r^2 + \frac{1}{2}(1-r)^2$
7. **jos** $h_{i1}^3 \neq h_{i1}^4$ **niin** // $g_i^1 = \{h_{i1}^3, h_{i2}^4\}$
8. $\quad s := 2\,p(h_{i1}^3)\,p(h_{i1}^4)\,t$ ALKIO-2$(h_{\cdot 1}^3, h_{\cdot 1}^4, x_1, i+1, 0,$ tosi$)$
9. **muuten**
10. $\quad s := p(h_{i1}^3)^2$ ALKIO-2$(h_{\cdot 1}^3, h_{\cdot 1}^4, x_1, i+1, r, b)$ // $g_i^1 = \{h_{i1}^3\}$
11. $\quad$ **jokaiselle** $a \in \mathcal{A} \setminus \{h_{i1}^3\}$ **:** // $g_i^1 = \{h_{i1}^3, a\}$
12. $\quad\quad s := s + 2\,p(h_{i1}^3)\,p(a))\,t$ ALKIO-2$(h_{\cdot 1}^3, h_{\cdot 1}^4, x_1, i+1, 0,$ epätosi$)$
13. **palauta** s

Kuva 5.7: Algoritmi alkion $W_{x_1}^1(h_{\cdot 1}^3, h_{\cdot 1}^4)$ laskemiseen kytkennällisen periytymisen mallissa.

Merkille pantavaa on, että rekursion loppuessa (rivillä 1) tarvitaan genotyyppiä g^1, jonka laskemista ei algoritmissa kuvata. Laskeminen ei kuitenkaan ole ongelmallista: Jokaisella rekursiotasolla i kutakin riveistä 8, 10 ja 12 vastaa tietty genotyyppi g_i^1. Esimerkiksi rivillä 12 on voimassa $g_i^1 = \{h_{i1}^3, a\}$. Jonot $(g_1^1, \ldots, g_i^1)$ voidaan kuvata indeksiksi sopivaan taulukkoon niin, että rivin 1 suoritus vie vakioajan (vrt. algoritmiin YLEISYYS). Luettavuuden vuoksi nämä tekniset seikat on jätetty kirjoittamatta aliohjelmaan ALKIO-2.

Nähdään, että aliohjelman ALKIO-2 rekursiivisten kutsujen määrä on

$$\sum_{k=1}^{\ell} \prod_{\substack{i=1 \\ h_{i1}^3 = h_{i1}^4}}^{k} |\mathcal{A}_i|.$$

Aliohjelmaa ALKIO-2 kutsuu kuvan 5.8 algoritmi MATRIISI-2, joka laskee matriisin $W_{x_1}^{(1)}$ alkiot yksitellen. Aliohjelman ALKIO-2 kutsujen yhteismääräksi tulee

$$\begin{aligned}\sum_{h_{\cdot 1}^3 \in \mathcal{A}} \sum_{h_{\cdot 1}^4 \in \mathcal{A}} \sum_{k=1}^{\ell} \prod_{\substack{i=1 \\ h_{i1}^3 = h_{i1}^4}}^{k} |\mathcal{A}_i| &= \sum_{k=1}^{\ell} \prod_{i=1}^{k} (|\mathcal{A}|\,|\mathcal{A}_i| + |\mathcal{A}_i|^2 - |\mathcal{A}_i|) \\ &= \sum_{k=1}^{\ell} \prod_{i=1}^{k} |\mathcal{A}_i|^2 \left(2 - \frac{1}{|\mathcal{A}_i|}\right).\end{aligned}$$

Ylärajaksi saadaan

$$\begin{aligned}\sum_{k=1}^{\ell} \left(\frac{1}{4}\right)^{\ell-k} |\mathcal{H}|\, 2^k &= \left(\frac{1}{4}\right)^{\ell} |\mathcal{H}| \sum_{k=1}^{\ell} 8^k \\ &< \frac{8}{7} 2^{\ell} |\mathcal{H}|.\end{aligned}$$

Algoritmin MATRIISI-2 aikavaativuus on siten $O(2^{\ell} |\mathcal{H}|)$, joka on $O(|\mathcal{H}|^{3/2})$.

MATRIISI-2(x_1)
1. **jokaiselle** $h_{\cdot 1}^3 \in \mathcal{A}$ **:**
2.   **jokaiselle** $h_{\cdot 1}^4 \in \mathcal{A}$ **:**
3. $W_{x_1}^{(1)}(h_{\cdot 1}^3, h_{\cdot 1}^4) :=$ ALKIO-2($h_{\cdot 1}^3, h_{\cdot 1}^4, x_1, 0, 1, \frac{1}{2}$, epätosi)
4. **palauta** $W_{x_1}^{(1)}$

Kuva 5.8: Algoritmi matriisin $W_{x_1}^{(1)}$ laskemiseen.

Ollaan johdettu menetelmä, jolla havaintojen uskottavuus kaksilapsiselle ydinperheelle voidaan kytkennällisen periytymisen malleissa laskea ajassa $\Theta(|\mathcal{H}|^{3/2})$.

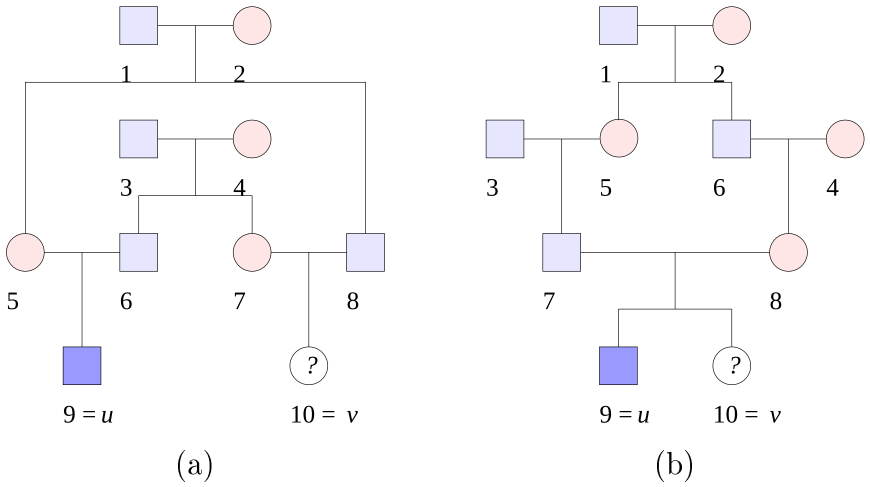
5.3 Riski mielivaltaista tyyppiä olevalle sukulaiselle

Edellisissä osioissa käsiteltiin taudin yleisyyden ja ydinperheriskien laskentaa eri malliperheissä. Nämä yksinkertaiset riskit ovat mallintamismenetelmän kannalta keskeisimpiä, mutta ne eivät paljasta sukulaisriskien laskennan koko vaikeutta. Tässä osiossa tarkastellaan sukulaisriskin laskemista sairaan yksilön mielivaltaista tyyppiä olevalle sukulaiselle.

Palautetaan mieleen sukulaisriskin määritelmä tyyppiä R olevalle sukulaiselle:

$$K_R = P(X_v = 1 \mid X_u = 1),$$

missä u ja v ovat tyyppiä R vastaavasta sukupuusta osoitettuja yksilöitä. Kuvassa 5.9 on kaksi esimerkkiä.



Kuva 5.9: Kaksinkertaiset serkut (a) ja sisarukset, joiden vanhemmat ovat serkuksia (b). Kysymysmerkillä on osoitettu yksilö, jonka sairastumisriskiä tarkastellaan.

Kuvan 5.9 sukupuut kuvaavat selvästi erityyppistä sukulaisuutta. Laskennan kannalta oleellinen ero on, että sukupuussa (a) yksilön u (myös yksilön v) vanhemmat eivät ole sukulaisia, mutta sukupuussa (b) vanhemmat ovat sukua toisilleen, serkuksia. Sukupuussa (b) yksilöiden u ja v sanotaan olevan *sukusiittoisia* (engl. inbred). Sukusiittoisuuden yleinen määritelmä on yksinkertainen: yksilö on sukusiittoinen, jos hänen vanhempansa ovat sukua toisilleen. Sukusiittoisuudella on oleellinen vaikutus sukulaisriskien laskentaan: jos yksilön vanhemman ovat sukua toisilleen, heidän genotyyppinsä eivät ole riippumattomia, mikä vaikeuttaa laskentaa.

Laskennallisia kysymyksiä käsitellään tarkemmin osioissa 5.3.1–5.3.3. Osiossa 5.3.1 sallitaan sukusiittoisuus, mutta tarkastelu rajoitetaan kytkemättömien mallien perheeseen. Esitettävä menetelmä on yleistys osiossa 5.2.2 sisarusriskin laskemiseen johdetulle menetelmälle. Osiossa 5.3.2 perehdytään lyhyesti toisenlaiseen lähestymistapaan, joka ei salli sukusiittoisuutta mutta soveltuu myös kytkennällisille malleille. Yksityiskohtaista esitystä ei anneta, ja avoimeksi jäävien kysymysten vuoksi aikavaativuusarviot esitetään konjektuureina. Lopuksi osiossa 5.3.3 palataan laskentaan yleisissä malleissa. Sukulaisriskien laskemiseen sovelletaan yleisten sukupuuhavaintojen uskottavuuksien laskemiseen tarkoitettua Elstonin ja Stewartin algoritmia [ElS71].

5.3.1 Matriisihajotelma

Osiossa 5.2.2 näytettiin, kuinka sisarusriski voidaan kytkemättömän periytymisen malleissa laskea tehokkaasti. Tähän päästiin soveltamalla Yatesin algoritmia matriiseille, jotka lokuksittain kuvaavat sisarusten genotyyppien yhteistodennäköisyysjakauman. Seuraavaksi osoitetaan, että menetelmää voidaan yleistää menestyksellisesti. Samalla annetaan yleinen keino lokuksittaisten matriisien laskemiseen ja vastataan näin osiossa 5.2.2 avoimeksi jätettyyn ongelmaan.

Kuten aiemminkin, lähtökohtana on jälleen kaava

$$K_R \quad = \quad \frac{P(X_v = 1, X_u = 1)}{P(X_u = 1)}.$$

Aloitetaan osoittajan $P(X_v = 1, X_u = 1)$ tarkastelulla, ja käsitellään nimittäjän $P(X_u = 1)$ laskemista vasta lopuksi.

Kaavaa (5.3) vastaten voidaan kirjoittaa

$$P(X_v = 1, X_u = 1) \quad = \quad f^T W^R f, \tag{5.13}$$

kun matriisi W^R, indeksoituna joukon $\mathcal{G}$ genotyypeillä, määritellään asettamalla

$$W^R(g^u, g^v) \quad = \quad p(g^u, g^v).$$

Määritellään sitten matriisit W_i^R, indeksoituna joukon $\mathcal{G}_i$ genotyypeillä, asettamalla

$$W_i^R(g_i^u, g_i^v) \quad = \quad p(g_i^u, g_i^v).$$

Koska lokuksen oletetaan kytkemättömiksi, genotyypit eri lokuksissa ovat riippumattomia ja saadaan, että matriisi W^R on matriisien W_i^R Kroneckerin tulo.

(Tarkkaan ottaen tähän tarvitaan myös muita kytkemättömän periytymisen mallien riippumattomuusoletuksia, mutta asian intuitiivisuuden vuoksi asiaan ei tässä syvennytä tarkemmin.)

Seuraavaksi esitettävä menetelmä matriisien W_i^R laskemiseen nojautuu *alkuperä-identtisyyden* (engl. indentity-by-descend, lyh. IBD) käsitteeseen. Kahden alleelin sanotaan olevan alkuperältään identtiset, jos toinen on toisen fyysinen kopio tai molemmat ovat kopioita samasta esivanhemman alleelista [Sha98, s. 98]. Alleelien joukon sanotaan olevan IBD, jos joukon kaikki alleelit ovat pareittain IBD.

Kun tarkastellaan kahta sukulaista, ollaan kiinnostuneita heidän alleeliensa suhteista jossakin lokuksessa. Olkoon siksi a^u ja b^u, a^v ja b^v kahden sukulaisen u and v alleelit lokuksessa i. Vaihtoehtojen täydellinen luetteleminen tuottaa yhteensä 15 erilaista hahmoa. Jos isän ja äidin symmetria huomioidaan nämä *yksityiskohtaiset identtisyystilat* (engl. detailed identity state), kuten Lange [Lan97, s. 74] kutsuu, pelkistyvät yhdeksään *typistettyyn identtisyystilaan* (engl. condensed identity state):

$$
\begin{aligned}
S^{(1)} &: a^u \equiv b^u \equiv a^v \equiv b^v \\
S^{(2)} &: a^u \equiv b^u \not\equiv a^v \equiv b^v \\
S^{(3)} &: a^u \equiv b^u \equiv a^v \not\equiv b^v \\
S^{(4)} &: a^u \equiv b^u \not\equiv a^v \not\equiv b^v \\
S^{(5)} &: a^u \not\equiv b^u \equiv a^v \equiv b^v \\
S^{(6)} &: a^u \not\equiv b^u \not\equiv a^v \equiv b^v \\
S^{(7)} &: a^u \equiv a^v \not\equiv b^u \equiv b^v \\
S^{(8)} &: a^u \equiv a^v \not\equiv b^u \not\equiv b^v \\
S^{(9)} &: a^u \not\equiv a^v \not\equiv b^u \not\equiv b^v
\end{aligned}
$$

(samanlaisen esityksen antaa Sham [Sha98, s. 99]). Symboli $\equiv$ tarkoittaa tässä alkuperä-identtisyyttä. Esimerkiksi $a^u \equiv b^v$, jos alleleelit a^u ja b^v ovat IBD.

Yhdeksän hahmoa vaihtelevat tilanteesta, jossa kaikki neljä alleelia ovat IBD ($S^{(1)}$), tapaukseen jossa mikään alleeleista ei ole IBD yhdenkään toisen alleelin kanssa ($S^{(9)}$). Hahmojen esiintymistodennäköisyyksiä kutsutaan *typistetyn identtisyyden kertoimiksi* (engl. condensed indentity coefficients [Lan97, s. 74]).

Yksilöiden sukulaisuussuhde määrä kertoimien arvon.

Kun yksilöiden alleeleja ajatellaan alkuperänsä muistavina satunnaismuuttujina, sukulaisten u ja v typistetty identtisyystila lokuksessa i on satunnaismuuttuja, merkitään S_i^{uv}. Todennäköisyyttä, että S_i^{uv} saa arvon $S^{(k)}$ merkitään tavallisesti Δ_k^R tai lyhyemmin Δ_k. Pannaan merkille joitakin näiden kertoimien ominaisuuksia. Yksilön alleelit voivat olla keskenään IBD vain, jos yksilön vanhemmat ovat keskenään sukua. Tästä seuraa, että kertoimien Δ_1, Δ_2, Δ_3, ja Δ_4 arvo on 0, jos u ei ole sukusiittoinen. Samoin huomataan, että kertoimien Δ_1, Δ_2, Δ_5 ja Δ_6 arvo on 0, jos v ei ole sukusiittoinen. Todennäköisyys, että yksilön u alleelit ovat IBD, on $\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$.

Esimerkki 5.3. Sisarukset perivät vanhemmiltaan samat alleelit todennäköisyydellä $\Delta_7^S = \frac{1}{4}$. Todennäköisyys, että he ovat perineet identtiset alleelit vain toiselta vanhemmalta, on $\Delta_8^S = \frac{1}{2}$. Täysin eri alleelit periytyvät sisaruksille todennäköisyydellä $\Delta_9^S = \frac{1}{4}$. Muut typistetyn identtisyyden kertoimet ovat nollia, koska sisarusten vanhemmilla ei voi olla IBD-alleeleja. (Vanhemmat eivät ole sukusiittoisia eivätkä sukua toisilleen.) $\square$

Matriisin W_i^R alkiot voidaan nyt kirjoittaa summauksena yli tilojen $S^{(k)}$:

$$W_i^R(g_i^u, g_i^v) = \sum_{k=1}^{9} \Delta_k^R \, p(g_i^u, g_i^v \mid S^{(k)}).$$

Kun tila $S^{(k)}$ tunnetaan, sukulaisuustyyppi R ei vaikuta todennäköisyyteen, että sukulaisilla u ja v on genotyypit g_i^u ja g_i^v. Jos nimittäin kahden alleelin tiedetään olevan IBD, niin ne ovat peräisin joltakin sukupuun perustajista. Vaihtoehtoisia perustajia voi olla useita, kukin jollakin todennäköisyydellä, mutta oli perustaja kuka tahansa, IBD-alleelit edustavat alleelia $a \in \mathcal{A}_i$ todennäköisyydellä $p(a)$. Jos sen sijaan tiedetään, että alleelit eivät ole IBD, ne ovat peräisin eri perustajilta ja tämän vuoksi riippumattomia. (Vastaavan perustelun antaa mm. Bulmer [Bul80, s. 41].)

Määritelläänkin sukulaisuustyypistä R riippumattomat matriisit $W_i^{(k)}$ asettamalla

$$W_i^{(k)}(g_i^u, g_i^v) = p(g_i^u, g_i^v \mid S^{(k)}),$$

jolloin saadaan

$$W_i^R = \sum_{k=1}^{9} \Delta_k^R W_i^{(k)}.$$

Seuraava esimerkki havainnollistaa matriisien $W_i^{(k)}$ määräytymistä.

Esimerkki 5.4. Tarkastellaan kaksialleelista lokusta, jonka alleelit ovat A ja a. Merkitään $p = p(A)$ ja $q = p(a) = 1 - p$. Olkoon tarkasteltavat sukulaiset u ja v. Jos sukulaisten u ja v identtisyystila on $S^{(1)}$, heidän kaikki neljä alleeliansa ovat IBD. Seuraa heti, että molempien genotyyppien on oltava samoja ja homotsygootteja. Tilaa $S^{(1)}$ vastaava matriisi on siten yksinkertaisesti

$$W_i^{(1)} = \begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{pmatrix} p & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & q \end{pmatrix} \end{array}.$$

Jos kumpikaan sukulaisista u ja v ei ole sukusiittoinen, ainoastaan tilat $S^{(7)}$, $S^{(8)}$ ja $S^{(9)}$ ovat mahdollisia. Tiloja vastaavat matriisit ovat

$$W_i^{(7)} = \begin{pmatrix} p^2 & 0 & 0 \\ 0 & 2pq & 0 \\ 0 & 0 & q^2 \end{pmatrix},$$

$$W_i^{(8)} = \begin{pmatrix} p^3 & p^2q & 0 \\ pq^2 & pq & p^2q \\ 0 & pq^2 & q^3 \end{pmatrix},$$

$$W_i^{(9)} = \begin{pmatrix} p^4 & 2p^3q & p^2q^2 \\ 2p^3q & 4p^2q^2 & 2pq^3 \\ p^2q^2 & 2pq^3 & q^4 \end{pmatrix}.$$

□

Matriisien $W_i^{(k)}$ laskeminen on suoraviivaista kuten esimerkin 5.4 matriiseista saattaa nähdä. Esimerkkinä olkoon kuvan 5.10 algoritmi, joka laskee tilaa $S^{(3)}$ vastaavan matriisin $W_i^{(3)}$. Tässä tapauksessa useat matriisin alkiot ovat nollia, koska kolmen alleelin vaaditaan olevan keskenään IBD. Neljäs alleeli voi olla riippumattomasti sama tai eri kuin kolme muuta. Huomautettakoon, että matriisi $W_i^{(3)}$ ei ole symmetrinen.

Matriisi-S3(i)

jokaiselle $\{a^u, b^u\} \in \mathcal{G}_i$ **:**
  jokaiselle $\{a^v, b^v\} \in \mathcal{G}_i$ **:**
    jos $a^u \neq b^u$ tai $(a^u \neq a^v$ ja $a^u \neq b^v)$ **niin**
      $W_i^{(3)}(\{a^u, b^u\}, \{a^v, b^v\}) := 0$
    muuten
      $W_i^{(3)}(\{a^u, b^u\}, \{a^v, b^v\}) := p(a^v)\, p(b^v)$
palauta $W_i^{(3)}$

Kuva 5.10: Algoritmi matriisin $W_i^{(3)}$ laskemiseen.

On helppo nähdä, että matriisit $W_i^{(k)}$ voidaan laskea ajassa $O(|\mathcal{G}_i|^2)$. Jos kertoimet Δ_k tunnetaan, myös matriisin W_i^R laskeminen vie ajan $O(|\mathcal{G}_i|^2)$. Yatesin algoritmi on siten aikavaativuudeltaan hallitseva. Tavallisimmille sukulaisuustyypeille typistetyn identtisyyden kertoimet löytyvät kirjallisuudesta (esim. [Lan97, s. 71], [Sha98, s. 209]). Yleisimmillään kertoimet joudutaan kuitenkin laskemaan sukulaisuustyypin määritelmään perustuen. Tämä vaatii sukulaisuustyyppiä vastaavan sukupuun analysointia, minkä tutkiminen jätetään tämän työn ulkopuolelle.

Oletetaankin tässä, että kertoimet voidaan laskea ajassa $O(\gamma(R))$. Yhteistodennäköisyyden $P(X_v = 1, X_u = 1)$ laskeminen edellä kuvatulla menetelmällä vie silloin yhteensä ajan $O(\gamma(R) + |\mathcal{G}| \sum_i |\mathcal{G}_i|)$. Koska sukulaisuustyyppi R ei riipu geneettisestä mallista, termi $\gamma(R)$ voidaan perustellusti pudottaa pois. Aikavaativuudeksi voidaan silloin kirjoittaa $\Theta(|\mathcal{G}| \sum_i |\mathcal{G}_i|)$.

Edellä käsiteltiin vasta yhteistodennäköisyyden $P(X_v = 1, X_u = 1)$ laskemista. Matriisin W^R avulla marginaalitodennäköisyydelle $P(X_u = 1)$ voidaan kirjoittaa neliömuodon (5.13) kaltainen lauseke:

$$P(X_u = 1) \quad = \quad f^T W^R \mathbf{1}.$$

Tässä $\mathbf{1}$ on vektori, jonka jokainen komponentti on 1. Yatesin algoritmia soveltamalla marginaalitodennäköisyys voidaan laskea yhtä tehokkaasti kuin yhteistodennäköisyys. Todetaan antamatta perustelua, että marginaalitodennäköisyys voidaan laskea selvästi nopeamminkin, ajassa $\Theta(|\mathcal{G}|)$. Jos u ei ole sukusiittoinen, on erikoisesti voimassa $P(X_u = 1) = K$.

On syytä mainita, että tässä osiossa käsitellyt matriisit W_i^R ovat hyvin samankaltaisia kuin Lin ja Sacksin [LiS54] esittelemät stokastiset matriisit. Lin ja Sacksin matriisit esittivät sukulaisen genotyyppien ehdollista jakaumaa yhteisjakauman sijaan. He käsittelivät vain yhtä lokusta eivätkä sallineet sukusiittoisia sukulaisuussuhteita.

Campbell ja Elston [CaE71] yleistivät Lin ja Sacksin matriisit usealle lokukselle Kroneckerin tuloa käyttäen. He käsittelivät myös joitakin sukusiittoisia sukulaisuussuhteita. Tässä kirjoittajien lähestymistapa poikkesi edellä esitetystä. Campbell ja Elston nimittäin pitäytyivät kolmessa matriisissa, jotka vastaavat identtisyystiloja $S^{(7)}$, $S^{(8)}$ sekä $S^{(9)}$, ja antoivat sukusiittoisuuden vaikuttaa ainoastaan matriisien kertoimiin. Esimerkiksi kuvan 5.9.b sukupuun esittämille sukulaisille he saivat yhden lokuksen tapauksessa tuloksen

$$p(g^v \mid g^u) \;=\; \frac{1}{4}\,p(g^v \mid g^u, S^{(9)}) + \frac{17}{32}\,p(g^v \mid g^u, S^{(8)}) + \frac{7}{32}\,p(g^v \mid g^u, S^{(7)})$$

[CaE71, s. 228]. Tämä vaikuttaa virheeltä: Kerroin $\frac{1}{4}$ on todennäköisyys sille, että yksilöillä u ja v ei ole keskenään yhtään alleelia IBD. Tämä sallii yksilön u alleelien olevan IBD, mikä ei vastaa tilaa $S^{(9)}$.

Mainitaan vielä, että Campbell ja Elston käsittelivät myös kahden lokuksen välistä kytkentää. Heidän lähestymistapansa ei kuitenkaan vaikuta lupaavalta laskennallista vaativuutta ajatellen.

5.3.2 Varianssihajotelma

Tässä osiossa käsitellään lyhyesti lähestymistapaa, joka poikkeaa edellisissä osioissa kuvatuista menetelmistä. Se nojautuu Fisherin [Fis18] varhaisiin geneettistä varianssia käsitteleviin matemaattisiin tuloksiin, joita myöhemmin laajensi mm. Kempthorne [Kem57]. Yhteyden sukulaisriskeihin esitti James [Jam71]. Asian monimutkaisuuden vuoksi tässä joudutaan tyytymään vain suuntaa antavaan esitykseen. Aikavaativuuksia koskevat tulokset perustuvat osittain arveluihin, minkä vuoksi ne muotoillaan konjektuureiksi.

Aloitetaan Jamesin [Jam71] esittämällä kaavalla. Olkoon R mielivaltainen sukulaisuustyyppi, johon ei liity sukusiittoisuutta, ja olkoon v tyypin R sukulainen yksilölle u. Jamesia mukaillen kirjoitetaan

$$\mathrm{Cov}(X_u, X_v) = E(X_u X_v) - E(X_u)\, E(X_v) = P(X_u = 1, X_v = 1) - K^2.$$

Tässä tarvittiin oletusta, että yksilöt u ja v eivät ole sukusiittoisia, jolloin $P(X_u = 1) = P(X_v = 1) = K$. Muistutetaan, että $K_R = P(X_u = 1, X_v = 1)/K$, ja siksi voidaan kirjoittaa

$$K_R = K + \frac{1}{K}\,\mathrm{Cov}(X_u, X_v).$$

Tehtävänä on nyt löytää kovarianssille $\mathrm{Cov}(X_u, X_v)$ sopiva lauseke ja näyttää, kuinka se voidaan laskea.

Kun geneettinen malli tunnetaan, kovarianssin $\mathrm{Cov}(X_u, X_v)$ arvo voidaan määrätä. Ajatuksena on tarkastella ilmenemisyleisyyksien varianssia ja jakaa se varianssikomponentteihin $V_A, V_D, V_{AA}, \ldots$. Yleisemmin merkitään $V_{A^sD^t}$, kun tarkoitetaan varianssia aiheutuen s additiivisen komponentin vuorovaikutuksista ja t dominantin komponentin vuorovaikutuksista. Näiden komponenttien täsmälliset määritelmät joudutaan tässä sivuuttamaan niiden monimutkaisuuden vuoksi (ks. [Kem57, Bul80]). Oleellinen tulos on lauseke

$$\mathrm{Cov}(X_u, X_v) \;=\; \sum_{1 \le s+t \le \ell} \alpha^R(s,t)\, V_{A^sD^t}, \tag{5.14}$$

missä s ja t ovat ei-negatiivisia kokonaislukuja. Yksilöiden sukulaisuussuhde ja lokusten välinen kytkentä vaikuttavat ainoastaan termeihin $\alpha^R(s,t)$.

Seuraavaksi tarkastellaan lyhyesti varianssien laskentaa. Sitten annetaan lausekkeet kertoimille $\alpha^R(s,t)$ olettaen, että lokusten välillä ei ole kytkentää. Pannaan merkille, että kaavan (5.14) suoraviivainen laskenta vie vain neliöllisen ajan suhteessa lokusten lukumäärään, mikäli varianssikomponentit $V_{A^sD^t}$ ja kertoimet $\alpha^R(s,t)$ tunnetaan. Varianssikomponenttien laskeminen näyttää kuitenkin olevan selvästi vaativampaa. Tämä ei ole yllättävää, kun muistetaan, että mallien parametrien lukumäärä on eksponentiaalinen suhteessa lokusten lukumäärään.

Varianssikomponentit voidaan laskea Bulmerin [Bul80, s. 46–51] antaman määritelmän mukaan. Tässä käsitellään vain yhtä määritelmän osaa, *ilmenemisyleisyyksien marginaalien* määrittelyä. Muokataan jonkin verran Bulmerin muotoilua ja määritellään alleelipaikkojen joukko L karteesisena tulona

$\{1, \ldots, \ell\} \times \{1, 2\}$. Esimerkiksi alleelipaikka $(i, 1) \in L$ esittää isältä perityn alleelin paikkaa lokuksessa i. Jokaiselle $S \subseteq L$ määritellään ilmenemisyleisyyksien marginaali f_S asettamalla

$$f_S \;=\; E_{L \setminus S}(f),$$

missä $E_{L \setminus S}$ tarkoittaa odotusarvoa yli alleelipaikkojen $(i, j) \in L \setminus S$ alleelien. Jos esimerkiksi $\ell = 2$ ja $S = \{(1, 1), (2, 2)\}$, niin kaikille $a \in \mathcal{A}_1$ ja $b \in \mathcal{A}_2$ on voimassa

$$f_S(a, b) \;=\; \sum_{c \in \mathcal{A}_1} \sum_{d \in \mathcal{A}_2} p(c)\, p(d)\, f(a, c, d, b).$$

Mainittakoon, että varianssikomponentit saadaan ottamalla ilmenemisyleisyyksien marginaalien sopivia summia ja laskemalla näiden varianssit.

Varianssikomponenttien tehokas laskenta näyttää vaativan ilmenemisyleisyyksien marginaalien laskemista. Tämän vaiheen osuus komponenttien laskennassa vaikuttaa myös aikavaativuudeltaan hallitsevalta. Vaativuuden arvioimiseksi tehdään muutama huomio. Ensinnäkin pannaan merkille, että kaikkia osajoukkoja $S \subseteq L$ ei tarvitse ottaa huomioon, koska ilmenemisyleisyydet ovat järjestämättömien genotyyppien funktioita. Toiseksi huomataan, että ilmenemisyleisyyksien marginaalit voidaan laskea tehokkaasti, jos laskentajärjestys valitaan sopivasti. Joukon L osajoukkojen osittaisen järjestyksen seuraaminen lähtien suurista joukoista vaikuttaa järkevältä. Näihin huomioihin ja aikavaativuuden hieman yksinkertaistettuun analyysin (ei esitetä tässä) perustuen saadaan seuraava konjektuuri.

Konjektuuri 5.2. *Jos alleelien lukumäärä on sama jokaisessa ℓ lokuksessa, varianssikomponentit $V_{A^s D^t}$ voidaan laskea ajassa $O(|\mathcal{A}|^{1/\ell}\,(|\mathcal{G}|^{1/\ell}{+}|\mathcal{A}|^{1/\ell}{+}1)^\ell)$.*

Siirrytään sitten tarkastelemaan kertoimia $\alpha^R(s, t)$. Alkuperäisessä muodossaan [Kem57, s. 420], kaavassa (5.14) oletettiin, että lokusten välillä ei ole kytkentää. Tässä erikoistapauksessa

$$\alpha^R(s, t) \;=\; (2\,\phi^R)^s\,(\Delta_7^R)^t, \tag{5.15}$$

missä ϕ^R on *sukulaisuuskerroin* (engl. coefficient of kinship). Sukulaisuuskerroin määritellään todennäköisyytenä, että kaksi sukulaisilta satunnaisesti valittua alleelia, yksi kummaltakin sukulaiselta, ovat IBD. Identtisyyskertoimia käyttäen voidaan kirjoittaa

$$\phi^R = \frac{1}{2}\Delta_7^R + \frac{1}{4}\Delta_8^R$$

Nähdään, että kerroin $\alpha^R(s,t)$ voidaan laskea helposti, kun kertoimet Δ_7^R ja Δ_8^R tunnetaan.

Bulmer käsitteli termien $\alpha^R(s,t)$ esitystä myös kytkennällisille malleille [Bul80, s. 91] ja antoi esimerkin kahden lokuksen tapauksessa. Käsittely jäi kuitenkin sen verran korkealle tasolle, että suoranaista laskentamenetelmää ei hänen esitykseensä nojautuen ole mahdollista laatia. Vaikuttaa kuitenkin siltä, että varianssikomponentteihin verrattuna kertoimien $\alpha^R(s,t)$ laskeminen on nopeaa. Tilanne on samankaltainen kuin identtisyyskertoimien Δ_i^R laskemisessa: varianssikomponentteihin vaikuttaa, lukuun ottamatta tekijänvaihtotodennäköisyyksiä, kaikki mallin parametrit mutta kertoimiin $\alpha^R(s,t)$ ainoastaan lokusten lukumäärä, tekijänvaihtotodennäköisyydet ja sukulaisuustyypin R määrittävä sukupuu. Muotoillaan esitetyt arvelut konjektuuriksi.

Konjektuuri 5.3. *Jos alleelien lukumäärä on sama jokaisessa ℓ lokuksessa, ja R ei ole sukusiittoinen, sukulaisriski K_R voidaan kytkennällisten mallien perheessä laskea ajassa $O(|\mathcal{A}|^{1/\ell}(|\mathcal{G}|^{1/\ell} + |\mathcal{A}|^{1/\ell} + 1)^\ell)$.*

Myös esitystä sukusiittoisten sukulaisten kovarianssille on tutkittu. Harris [Har64] onnistui yleistämään kaavan (5.14). Hänen tuloksena on samaa muotoa kuin (5.14), mutta termien määrältään huomattavasti suurempi ja rakenteeltaan monimutkaisempi. Harrisin esitykseen ei ole mahdollista perehtyä tässä työssä.

5.3.3 Sukupuun kuorinta

Edellisissä osioissa käsiteltiin menetelmiä, jotka hyödyntävät kytkennällisten ja kytkemättömien mallien tekemiä oletuksia tehokkaasti. Nyt palataan yleisten mallien perheeseen ja tarkastellaan, kuinka sukupuuhavaintojen uskottavuuden laskemiseen tarkoitettuja menetelmiä voidaan soveltaa sukulaisriskien laskentaan.

Aloitetaan johtamalla Elstonin ja Stewartin [ElS71] mallin mukaisesti mielivaltaisen sukupuuhavainnon uskottavuus yleisessä geneettisessä mallissa. Olkoon tarkasteltava sukupuu $\mathcal{S}$, jonka yksilöiden joukko on $\mathcal{V}$ ja perustajien joukko $\mathcal{F} \subseteq \mathcal{V}$. Palautetaan mieleen osiossa 3.1 johdettu kaava ilmiasujen ja genotyyppien yhteisjakaumalle:

$$\begin{aligned} P(X_{\mathcal{V}}, H^{\mathcal{V}}) &= \Big(\prod_{u \in \mathcal{F}} P(H^u)\Big) \\ &\quad \Big(\prod_{u \in \mathcal{V}\setminus\mathcal{F}} P(H^u \mid H^{F(u)}, H^{M(u)})\Big)\Big(\prod_{u \in \mathcal{V}} P(X_u \mid H^u)\Big). \end{aligned} \tag{5.16}$$

Olkoon sitten x jono joukon $\mathcal{U} \subseteq \mathcal{V}$ yksilöiden havaittuja ilmiasuja. Jotta sukupuun yksilöitä voitaisiin käsitellä yhtenäisesti huolimatta siitä, onko yksilön ilmiasu havaittu tai mikä ilmiasu on, määritellään

$$f_{(u)}(h) = \begin{cases} 1, & \text{jos } u \notin \mathcal{U} \\ f(h), & \text{muuten jos } x_u = 1 \\ 1 - f(h), & \text{muuten} \end{cases}.$$

Havaintojonon x todennäköisyys saadaan kaavasta (5.16) summaamalla yli yksilöiden tuntemattomien genotyyppien ja ilmiasujen:

$$\begin{aligned} P(X_{\mathcal{U}} = x) &= \sum_{h^{\mathcal{V}} \in \mathcal{H}^{|\mathcal{V}|}} \Big(\prod_{u \in \mathcal{F}} p(h^u)\Big) \\ &\quad \Big(\prod_{u \in \mathcal{V}\setminus\mathcal{F}} p(h^u \mid h^{M(u)}, h^{F(u)})\Big)\Big(\prod_{u \in \mathcal{V}} f_{(u)}(h^u)\Big). \end{aligned} \tag{5.17}$$

Pannaan merkille, että summaus yli tuntemattomien ilmiasujen sisältyy termeihin $f_{(u)}(h^u)$, kun $u \notin \mathcal{U}$, sillä

$$\sum_{x_u \in \{0,1\}} P(X_u = x_u \mid H^u = h^u) = 1 = f_{(u)}(h^u).$$

Nähdään, että kaavan (5.17) suoraviivainen laskenta vie ajan $\Theta(|\mathcal{V}|\,|\mathcal{H}|^{|\mathcal{V}|})$, joka on eksponentiaalinen suhteessa yksilöiden lukumäärään. Jos sukupuu sisältään identtisiä sisaruksia, niin kaavoihin (5.16) ja (5.17) joudutaan tekemään pieniä muutoksia (ks. osio 3.1).

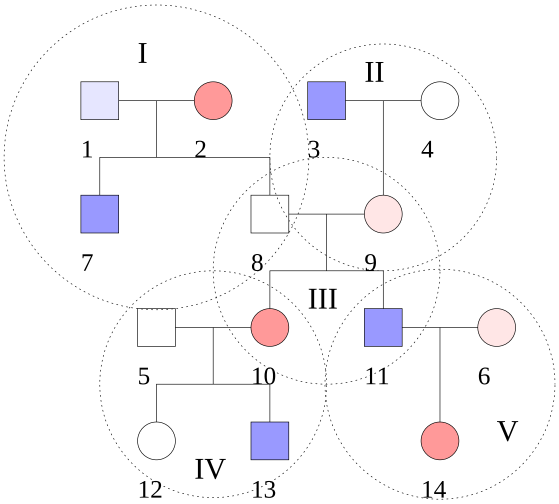
Geneettisen mallin mukainen Sukulaisriski voidaan laskea muotoa (5.17) olevien summausten osamääränä. Jos nimittäin yksilö v on yksilölle u tyypin R sukulainen, sukulaisriski K_R saadaan kaavasta

$$K_R = P(X_v = 1 \mid X_u = 1) = \frac{P(X_u = 1, X_v = 1)}{P(X_u = 1)}.$$

Sukulaisriski voidaan siten laskea ajassa $\Theta(|\mathcal{V}|\,|\mathcal{H}|^{|\mathcal{V}|})$, kun sukulaisuustyyppiä R vastaavan sukupuun yksilöiden joukko on $\mathcal{V}$.

Siirrytään sitten tarkastelemaan, kuinka kaavan (5.17) laskentaa voidaan tehostaa. Elston ja Stewart havaitsivat, että sopivasti rajoitetuille sukupuille kaava (5.17) voidaan laskea lineaarisessa ajassa suhteessa yksilöiden lukumäärään. Heidän esittämiään rajoituksia on myöhemmin lievennetty, mutta monimutkaisten sukupuiden tapauksessa parannetutkin menetelmät vievät eksponentiaalisen ajan suhteessa sukupuun yksilöiden määrään [BGS98]. Aikavaativuudet riippuvat oleellisesti sukupuun rakenteesta, mikä on ehkä vaikuttanut siihen, että tarkkoja vaativuustuloksia ei ole esitetty. Tässä tyydytään laajaan esimerkkiin (esimerkki 5.5), joka selvittää lineaarisessa ajassa toimivien algoritmien periaatteen. Samalla esimerkki antaa mahdollisuuden menetelmän aikavaativuuden tarkempaan arviointiin.

Esimerkki 5.5. Johdetaan menetelmä kuvan 5.11 suupuuhavaintojen laskemiseen yleisessä geneettisessä mallissa. Sukupuun yksilöiden joukko on siis $\{1, 2, \ldots, 14\}$, havaittujen yksilöiden joukko on $\{2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 14\}$, ja vastaava havaintojono on $x = (1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1)$. Esitettävän menetelmän ajatuksena on laskea summa (5.17) sopivia välituloksia hyödyntäen.



Kuva 5.11: Sukupuu, jolle havaintojen uskottavuuden laskeminen yleisessä geneettisessä mallissa on aikavativuudeltaan lineaarista suhteessa yksilöiden lukumäärään. Viisi ydinperhettä on numeroitu luvuilla I, II, III, IV ja V.

Havaitaan, että sukupuu sisältää kolme kaksilapsista ydinperhettä (I, III, IV) ja kaksi yhden lapsen ydinperhettä (II, V). Perheet I, II, IV ja V ovat erillisiä, mutta perhe III jakaa muiden perheiden kanssa yhden yksilön. Muut neljä perhettä ovat toisistaan riippuvia (geneettisesti) vain perheen III kautta.

Havaintojen x uskottavuuden laskemiseksi todetaan, että aina on voimassa

$$p(x) \;=\; \sum_{h^8,h^9,h^{10},h^{11}} p(x, h^8, h^9, h^{10}, h^{11}),$$

kun h^8, h^9, h^{10} ja h^{11} käyvät läpi yksilöiden $8, 9, 10$ ja 11 mahdolliset järjestetyt genotyypit. Ketjusäännön ja edellä esitettyjen perheiden riippumattomuutta koskevien huomioiden nojalla voidaan kirjoittaa

$$\begin{aligned}
p(x) \;&=\; \sum_{h^8,h^9,h^{10},h^{11}} p(x_2, x_7, x_8, h^8)\, p(x_3, x_4, h^9) \\
&\qquad p(x_5, x_{10}, x_{12}, x_{13}, h^{10} \mid h^8, h^9)\, p(x_{11}, x_{14}, h^{11} \mid h^8, h^9) \\
&=\; \sum_{h^8,h^9} p(x_2, x_7, x_8, h^8)\, p(x_3, x_4, h^9) \\
&\qquad \Big(\sum_{h^{10}} p(h^{10} \mid h^8, h^9)\, p(x_5, x_{10}, x_{12}, x_{13} \mid h^{10}) \Big) \\
&\qquad \Big(\sum_{h^{11}} p(h^{11} \mid h^8, h^9)\, p(x_{11}, x_{14} \mid h^{11}) \Big).
\end{aligned}$$

Jos nyt arvot $p(x_2, x_7, x_8, h^8)$, $p(x_3, x_4, h^9)$, $p(x_{11}, x_{14} \mid h^{11})$ ja $p(x_{11}, x_{14} \mid h^{11})$ tunnetaan kaikille $h^8, h^9, h^{10}, h^{11} \in \mathcal{H}$, niin havaintojen uskottavuus $p(x)$ voidaan laskea ajassa $\Theta(|\mathcal{H}|^3)$.

Seuraavaksi annetaan summakaavat tarvittaville termeille $p(x_2, x_7, x_8, h^8)$, $p(x_3, x_4, h^9)$, $p(x_{11}, x_{14} \mid h^{11})$ ja $p(x_{11}, x_{14} \mid h^{11})$. Kun lasketaan suoraviivaisesti

$$\begin{aligned}
p(x_2, x_7, x_8, h^8) \;&=\; \sum_{h^1} p(h^1)\, f_{(1)}(h^1) \sum_{h^2} p(h^2)\, f_{(2)}(h^2) \\
&\qquad p(h^8 \mid h^1, h^2)\, f_{(8)}(h^8) \sum_{h^7} p(h^7 \mid h^1, h^2)\, f_{(7)}(h^7)
\end{aligned}$$

kaikille $h^8 \in \mathcal{H}$, tarvittavien operaatioiden määrä on $\Theta(|\mathcal{H}|^4)$. Kolmelle muulle termille voidaan kirjoittaa

$$\begin{aligned}
p(x_3, x_4, h^9) \;&=\; \sum_{h^3} p(h^3)\, f_{(3)}(h^3) \sum_{h^4} p(h^4)\, f_{(4)}(h^4) \\
&\qquad p(h^9 \mid h^3, h^4)\, f_{(9)}(h^9),
\end{aligned}$$

$$
\begin{aligned}
p(x_5, x_{10}, x_{12}, x_{13} \mid h^{10}) \;=\;& f_{(10)}(h^{10}) \sum_{h^5} p(h^5)\, f_{(5)}(h^5) \\
& \prod_{u=12,13} \Big(\sum_{h^u} p(h^u \mid h^5, h^{10})\, f_{(u)}(h^u) \Big).
\end{aligned}
$$

ja

$$
\begin{aligned}
p(x_{11}, x_{14} \mid h^{11}) \;=\;& f_{(11)}(h^{11}) \sum_{h^6} p(h^6)\, f_{(6)}(h^6) \\
& \Big(\sum_{h^{14}} p(h^{14} \mid h^6, h^{11})\, f_{(14)}(h^{14}) \Big).
\end{aligned}
$$

Nähdään, että kunkin summauksen laskeminen genotyyppien h^9, h^{10} ja h^{11} eri arvoille vie yhteensä ajan $\Theta(|\mathcal{H}|^3)$. Esitetyn menetelmän aikavaativuus on siten $\Theta(|\mathcal{H}|^4)$.

Kyseiselle sukupuulle on olemassa nopeampikin laskujärjestys. Edellä kuvatussa menetelmässä sukupuu *kuorittiin* (engl. peel) aloittaen ydinperheistä I, II, IV ja V. Seuraavaksi esitettävä menetelmä aloittaa nuorimmista ydinperheistä IV ja V sekä perheestä II ja etenee sitten perheen III kautta perheeseen I.

Yhteistodennäköisyys $p(x)$ puretaan kirjoittamalla ensin

$$
\begin{aligned}
p(x) \;=\;& \sum_{h^1} p(h^1) \sum_{h^2} p(x_2, h^2) \Big(\sum_{h^7} p(h^7 \mid h^1, h^2) f_{(7)}(h^7) \Big) \\
& \Big(\sum_{h^8} p(h^8 \mid h^1, h^2) p(x_{\{3,4,5,8,10,11,12,13,14\}} \mid h^8) \Big)
\end{aligned}
$$

ja jatkamalla sitten

$$
\begin{aligned}
p(x_{\{3,4,5,8,10,11,12,13,14\}} \mid h^8) \;=\;& f_{(8)}(h^8) \sum_{h^9} p(x_3, x_4, h^9) \\
& \Big(\sum_{h^{10}} p(h^{10} \mid h^8, h^9)\, p(x_5, x_{10}, x_{12}, x_{13} \mid h^{10}) \Big) \\
& \Big(\sum_{h^{11}} p(h^{11} \mid h^8, h^9)\, p(x_{11}, x_{14} \mid h^{11}) \Big).
\end{aligned}
$$

Kuorinta aloitetaan arvojen $p(x_3, x_4, h^9)$, $p(x_5, x_{10}, x_{12}, x_{13} \mid h^{10})$ ja $p(x_{11}, x_{14} \mid h^{11})$ laskemisella aiemmin esitettyjä summakaavoja käyttäen. Nähdään, että aikavaativuudeksi saadaan $\Theta(|\mathcal{H}|^3)$. On kuitenkin syytä panna merkille, että vaativuudeksi tulisi $\Theta(|\mathcal{H}|^4)$, jos perheessä II olisi enemmän kuin yksi lapsi. $\square$

Edellä esitty esimerkki näytti, että sukupuuhavaintojen uskottavuus voidaan laskea tehokkaasti kuorimalla. Täsmällisemmin sanottuna kuorimisella tarkoitetaan muotoa $p(x_{\mathcal{U}'}, h^u)$ olevien yhteistodennäköisyyksien ja muotoa $p(x_{\mathcal{U}'} \mid h^u)$ olevien ehdollisten todennäköisyyksien laskemista sopivassa järjestyksessä. Yhteistodennäköisyydet viittaavat "kuorimiseen esivanhempista alkaen" ja ehdolliset todennäköisyydet "kuorimiseen jälkeläisistä alkaen". Nähdään, että yhteistodennäköisyyksille kuoriminen yhden sukupolven verran voidaan yleisesti laskea ajassa $\Theta(|\mathcal{H}|^4)$ ja ehdollisille todennäköisyyksille ajassa $\Theta(|\mathcal{H}|^3)$.

Elston ja Stewart [ElS71] rajoittuivat sukupuihin, joissa riittää kuorinta jälkeläisistä alkaen. Esimerkin 5.5 sukupuussa jouduttiin käyttämään kuorintaa sekä jälkeläisiin että esivanhempiin päin. Monimutkaisissa sukupuissa kuorinta ei ole mahdollista. Tällaisia ovat esimerkiksi aiemmin esitetyn kuvan 5.9 sukupuut: sukupuussa 5.9.a kuorinnan estää *pariutumissilmukka* (engl. marriage loop) ja sukupuussa 5.9.b kuorinnan tekee mahdottomaksi sukusiittoisuus.

Useat kirjoittajat [CTS78, LaB83, BGS98] ovat todenneet, että monimutkainenkin sukupuu voidaan esittää – epätäsmällisesti sanoen – ydinperheiden verkkona. Tällöin sukupuun kuorittavuus on yhtäpitävää verkon syklittömyyden kanssa. Becker et al. [BGS98] panivat myös merkille, että saatava verkko on *Bayes-verkko* (ks. [MyT98]). Edellä selostettu kuorinta-algoritmi on tällöin erikoistapaus yleisestä Bayes-verkkojen viestinvälitysalgoritmista [Pea88]. Monimutkaisten sukupuiden käsittelyyn Becker et al. sovelsivat yleisille Bayes-verkoille tarkoittamaansa syklinpoistoalgoritmia. Lange ja Boehnke [LaB83] esittivät jo aiemmin samankaltaisen menetelmän, joka yleisessä muodossaan oli tarkoitettu sopivan summausjärjestyksen löytämiseen laskettaessa yhteen tuloja, joiden tekijät ovat moniulotteisten taulukoiden alkioita.

Sukulaisriskien laskennassa voidaan hyödyntää kaikkia edellä mainittuja menetelmiä. Koska sukulaisriskit koskevat taudin ilmenemistä yksilöpareilla, yleensä riittää tarkastella kohtuullisen yksinkertaisia sukupuita. Esimerkiksi ydinperheriskien laskemiseksi voitiin käsitellä pieniä ydinperheitä, joissa riitti kuoriminen yhden sukupolven verran (ks. luku 5.1).

5.4 Yhteenveto

Edellisissä osioissa kuvatuilla menetelmillä saavutettavat aikavaativuudet eri suure-malliperhe -pareille on koottu taulukkoon 5.1. Nähdään, että laskenta nopeutuu, kun siirrytään yleisistä sukulaisriskeistä näiden erikoistapauksiin. Havaitaan myös, että aikavaativuudet paranevat merkittävästi siirryttäessä yleisistä malleista kytkennällisen ja kytkemättömän periytymisen malleihin. Esimerkki 5.6 havainnollistaa mahdollista parannusta numeerisesti.

suure	*laskennan vaativuus*					
	$\mathcal{M}_{[4]}$		$\mathcal{M}_{[3]}$		$\mathcal{M}$	
K	+		$\Theta(\lvert\mathcal{G}\rvert)$	(5.1.1)	$\Theta(\lvert\mathcal{H}\rvert)$	(5.1.1)
K_{MZ}	+		$\Theta(\lvert\mathcal{G}\rvert)$	(5.1.2)	+	
K_O	+		$\Theta(\lvert\mathcal{H}\rvert)$	(5.1.3)	+	
K_S	+		$\Theta(\lvert\mathcal{H}\rvert^{3/2})$	(5.2.3)	$\Theta(\lvert\mathcal{H}\rvert^3)$	(5.1.4)
K_R $^{(1}$	+		?	(5.3.2)	+	
K_R $^{(2}$	$\Theta(\lvert\mathcal{G}\rvert \sum_i \lvert\mathcal{G}_i\rvert)$	(5.3.1)	+		$\Theta(\lvert\mathcal{V}\rvert\, \lvert\mathcal{H}\rvert^{\lvert\mathcal{V}\rvert})$	(5.3.3)

Taulukko 5.1: Sukulaisriskien laskennan aikavaativuudet kytkemättömien ($\mathcal{M}_{[4]}$), kytkennällisten ($\mathcal{M}_{[3]}$) ja yleisten ($\mathcal{M}$) mallien perheissä. Aikavaativuuksien yhteyteen on merkitty viite tämän luvun alilukuun, jossa vastaava menetelmä kuvataan. Merkintä '+' tarkoittaa, että aikavaativuus on sama kuin paras samalla rivillä oikealla tai samassa sarakkeessa alempana. Merkintä '?' viittaa konjektuuriin 5.3. $^{1)}$ Sukulaisuustyyppi R ei salli sukusiittoisuutta. $^{2)}$ Sukulaisuustyyppi R voi olla mielivaltainen; $|\mathcal{V}|$ on yksilöiden lukumäärä tyyppiä R vastaavassa sukupuussa.

Esimerkki 5.6. Tarkastellaan sisarusriskin laskemista kytkemättömän periytymisen mallissa, jossa on kolme kaksialleelista lokusta. Jos sisarusriskin laskemiseen käytetään naiivisti yleisille malleille soveltuvaa menetelmää, niin tarvittavien operaatioiden määrä on suuruusluokkaa $(4^3)^3 = 262\ 144$. Jos sen sijaan käytetään osioissa 5.2.2 ja 5.3.1 johdettua menetelmää, on tarvittavien operaatioiden määrä luokkaa $3 \cdot 3 \cdot 3^3 = 243$ (ks. taulukko 5.2). □

Verrataan sitten konjektuureissa 5.2 ja 5.3 esitettyjä aikavaativuuksia taulukon 5.1 tuloksiin. Jos konjektuuri 5.2 pitää paikkansa, sukulaisriski mielivaltaiselle sukulaisuustyypille R voidaan kytkemättömien mallien perheessä laskea ajassa $O(|\mathcal{A}|^{1/\ell}\,(|\mathcal{G}|^{1/\ell}+|\mathcal{A}|^{1/\ell}+1)^{\ell})$. Konjektuurissa vaadittiin, että R ei sisällä

m	ℓ	$\lvert\mathcal{H}\rvert^{3/2}$	$\lvert\mathcal{A}\rvert^{1/\ell}(\lvert\mathcal{G}\rvert^{1/\ell}+\lvert\mathcal{A}\rvert^{1/\ell}+1)^{\ell}$	$\ell\,\lvert\mathcal{G}\rvert^{1+1/\ell}$	$\lvert\mathcal{H}\rvert$	$\lvert\mathcal{G}\rvert$
2	3	512	432	243	64	27
2	4	4 096	2 592	972	256	81
3	3	19 683	3 000	3 888	729	216
3	4	531 441	30 000	31 104	6 561	1 296

Taulukko 5.2: Aikavaativuuksia kuvaavia lukuja, kun lokuksissa on m alleelia.

sukusiittoisuutta ja että alleelien lukumäärä jokaisessa ℓ lokuksessa on sama. Aikavaativuutta on verrattava tulokseen $\Theta(\ell\,|\mathcal{G}|^{1+1/\ell})$. Voidaan osoittaa, että jälkimmäinen tulos on parempi tai yhtä hyvä kuin edellinen, jos alleelien lukumäärä lokuksissa on pieni (2, 3) ja lokusten lukumäärä suuri $(3, 4, \ldots)$. Jos alleeleja on useita, konjektuurissa esitetty vaativuus on parempi. Vaativuuksusta kertovia lukuja on taulukossa 5.2 (4. ja 5. sarake).

Konjektuurissa 5.3 esitettiin sama vaativuus kuin konjektuurissa 5.2 mutta kytkennällisten mallien perheelle. Tätä tulosta voidaan verrata sisarusriskin laskennalle esitettyyn vaativuuteen $\Theta(|\mathcal{H}|^{3/2})$. Jos konjektuuri pitää paikkansa, voidaan sisarusriski laskea esitettyä menetelmää nopeammin varianssihajotelmaa käyttäen. Suuntaa antavia lukuja on taulukossa 5.2 (3. ja 4. sarake). Avoimeksi jää kysymys siitä, voidaanko varianssihajotelma yleistää menestyksellisesti kaksilapsisen ydinperheen tapaukseen.

Käsiteltyjä sukulaisriskien laskentamenetelmiä voidaan monelta osin pitää yleisempien menetelmien sovelluksina. Esimerkiksi Yatesin algoritmi on jo sellaisenaan yleinen menetelmä. Se on lisäksi läheistä sukua nopealle Fourier muunnokselle (engl. fast Fourier transform, lyh. FFT) [Goo58]. Sukupuiden kuorinta-algoritmien kehittämisessä voidaan puolestaan soveltaa Bayes-verkkojen teoriaa ja Pearlin viestinvälitysalgoritmia [Pea88], kuten osiossa 5.3.3 jo mainittiin. Voidaan tarkastella vieläkin yleisemmin: Aji ja McEliece [AjM00] osoittivat, että edellä mainitut ja niiden kaltaiset laskentamenetelmät perustuvat lopulta osittelulain $ab + ac = a(b + c)$ hyödyntämiseen. He myös antoivat yleisen algoritmin, jonka erikoistapauksina mm. FFT ja Pearlin viestinvälitysalgoritmi saadaan. Toimiakseen hyvin yleiset menetelmät vaativat luonnollisestikin riippumattomuusoletusten tyhjentävää kuvaamista.

Luku 6

Stokastinen simulointi

Luvussa 4 esitettiin bayesiläinen lähestymistapa, joka mahdollistaa monipuolisen tilastollisen päättelyn geneettisten mallien posteriorijakaumaan nojautuen. Tässä luvussa tarkastellaan avoimeksi jätettyä ongelmaa: mallien posterioritodennäköisyyksillä painotettujen summien laskemista (ks. kaava (4.3)). Tarkastelussa rajoitutaan Markovin ketjuja hyödyntävään Monte Carlo -integrointiin, lyhyemmin MCMC-menetelmään, jota on viime vuosina laajasti käytetty integroitaessa yli moniulotteisten todennäköisyysjakaumien.

MCMC:n bayesiläisissä sovelluksissa tarkasteltava todennäköisyysjakauma on mallien posteriorijakauma. MCMC-menetelmän voidaan tällöin ajatella koostuvan kahdesta vaiheesta. Ensimmäinen vaihe on edustavan satunnaisotoksen muodostaminen mallien posteriorijakaumasta. Tämä tehdään simuloimalla sopivaa geneettisten mallien Markovin ketjua. Koska Markovin ketjut ovat ovat stokastisia prosesseja, puhutaan stokastisesta simulaatiosta. Toisessa vaiheessa hyödynnetään simuloinnilla tuotettua otosta. Kun kohteena olevaa odotusarvoa arvioidaan likimääräisesti satunnaisotoksen keskiarvona, on kyse Monte Carlo -integroinnista.

Luku aloitetaan esittelemällä MCMC:n teorian perusteet lyhyesti. Tämän jälkeen perehdytään menetelmän soveltamiseen tutkielman aiheena olevaan mallintamisongelmaan. Esityksen lähtökohtana on menetelmän rajoitettu toteutus, ohjelma nimeltään GeME (lyh. engl. Genetic Model Explorer). Ohjelma suorittaa MCMC:n ensimmäisen vaiheen eli stokastisen simulaation.

6.1 Teoriaa

MCMC:n keskeisiä käsitteitä ovat Monte Carlo -integrointi ja Markovin ketjut. Seuraavaksi esitellään lyhyesti näiden käsitteiden matemaattista teoriaa erityisesti bayesiläistä soveltamista ajatellen. Yksinkertaisuuden vuoksi tarkastellaan diskreettejä jakaumia; teoria yleistäminen jatkuville ei vaadi suuria käsitteellisiä muutoksia [Tie96]. Esitys nojautuu pääosin lähteisiin [GRS96, s. 1–19] ja [Rob95].

6.1.1 Ongelma

Olkoon X diskreetti satunnaismuuttuja, jonka arvojoukko on $\mathcal{X}$. Yksinkertaisuuden vuoksi oletetaan, että $\mathcal{X}$ on äärellinen. Olkoon satunnaismuuttujan X jakauma π. Bayesiläisissä sovelluksissa X koostuu mallin parametreista ja mahdollisesti puuttuvista havainnoista, π on mallien posteriorijakauma. Bayesiläisessä päättelyssä tarkasteltavat posteriorijakauman piirteet, kuten momentit, kvantiilit ja suuren todennäköisyyden alueet sekä parametrien marginaalijakaumat, voidaan esittää mallien funktioiden odotusarvona. Näin päädytään ongelmaan, jossa on laskettavana odotusarvo

$$E(f(X)) = \sum_{x\in\mathcal{X}} \pi(x)f(x), \tag{6.1}$$

missä f on funktio joukolta $\mathcal{X}$ reaaliluvuille. Usein odotusarvo ei ole analyyttisesti laskettavissa tai laskenta on hyvin vaikeaa, minkä vuoksi turvaudutaan numeerisiin menetelmiin.

Kun esitettävää teoriaa jäljempänä sovelletaan osiossa 4.2.1 muotoiltuun ongelmaan, X vastaa geneettistä mallia ja f funktiota, joka kuvaa jonkin mallin mielenkiintoisen piirteen. (X ja f eivät siis tässä viittaa taudin ilmiasuun ja ilmenemisyleisyyksiin kuten aikaisemmissa luvuissa.)

6.1.2 Monte Carlo -integrointi

Moniulotteisten integraalien numeeriseen arvioimiseen voidaan käyttää satunnaiseen otantaan perustuvaa Monte Carlo -integrointia. Odotusarvoon (6.1) sovellettuna Monte Carlo -integroinnissa muodostetaan jakaumasta π satunnaisotos $\{X_i \mid i = 1, \ldots, n\}$ ja lasketaan likiarvo

$$E(f(X)) \approx \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Otoksen satunnaisuuden vuoksi likiarvo on satunnaismuuttuja. Jos otokset ovat riippumattomia, saadaan likiarvo suurten lukujen lain nojalla mielivaltaisen tarkaksi kasvattamalla otoksen kokoa n.

Useissa tapauksissa riippumattomien otosten muodostaminen ei ole käytännöllistä. Tyypillisesti jakauma π on epätasainen keskittyen joihinkin moniulotteisen avaruuden osiin. Pisteitä satunnaisesti arpomalla vain harvoin osutaan niihin osiin, joissa π on merkittävästi nollaa suurempi. Riippumattomuus ei kuitenkaan ole välttämätöntä: väljästi sanoen riittää, että otokset muodostava prosessi tuottaa arvoja jakauman π mukaisesti sopivissa suhteissa. Yksi keino sopivan otoksen muodostamiseen hyödyntää Markovin ketjua, jonka *tasapainojakauma* on π.

6.1.3 Markovin ketjut

Markovin ketju on diskreettiaikainen stokastinen prosessi $\{X_0, X_1, \ldots\}$, jolle kaikilla $t = 1, 2, \ldots$ satunnaismuuttujan X_t jakauma ehdollistettuna prosessin historialle, eli muuttujille $X_0, X_1, \ldots, X_{t-1}$, riippuu ainoastaan muuttujasta X_{t-1}. Matemaattisemmin tämä *Markov-ominaisuus* kirjoitetaan

$$P(X_t \mid X_0, X_1, \ldots, X_{t-1}) = P(X_t \mid X_{t-1}).$$

Satunnaismuuttujaa X_t kutsutaan prosessin *tilaksi askeleella* t. Markovin ketju on täysin määrätty, kun annetaan alkutilan X_0 jakauma ja tilasiirtymien jakauma $P(X_t \mid X_{t-1})$.

Oletetaan sitten, että jokaisella askeleella todennäköisyys siirtyä tilasta x tilaan y on kiinteä. Tällöin Markovin ketjun sanotaan olevan *stationäärinen* ja

Markovin ketjun *tilasiirtymämatriisi* $\mathbb{P}$ määritellään asettamalla

$$\mathbb{P}(x,y) = P(X_t = y \mid X_{t-1} = x)$$

kaikilla $x, y \in \mathcal{X}$. Tilasiirtymämatriisin potenssi $\mathbb{P}^m$ sisältää tällöin m askeleen tilasiirtymien todennäköisyydet:

$$\mathbb{P}^m(x,y) = P(X_{t+m} = y \mid X_t = x).$$

Markovin ketjun sanotaan unohtavan alkutilansa, jos tila X_m "tulee riippumattomaksi" alkutilasta X_0, kun m kasvaa rajatta. Tällöin on olemassa yksikäsitteinen *tasapainojakauma* π, jolle

$$\lim_{m\to\infty} \mathbb{P}^m(x,y) = \pi(y)$$

kaikilla tiloilla $x, y \in \mathcal{X}$. Voidaan helposti osoittaa, että π toteuttaa yhtälön

$$\pi = \pi\mathbb{P}.$$

Mitä tahansa todennäköisyysjakaumaa, jolle tämä yhtälö on voimassa, kutsutaan tasapainojakaumaksi. Jos tiloja on äärellinen määrä, on tasapainojakaumia aina olemassa. Jotta tasapainojakauma olisi yksikäsitteinen, matriisin $\mathbb{P}$ on toteuttava seuraavat *ergodisuusehdot*:

Pelkistymättömyys: Markovin ketju on pelkistymätön (engl. irreducible), jos jokaisesta tilasta mikä tahansa toinen tila voidaan saavuttaa jollakin askelmäärällä positiivisella todennäköisyydellä. Pelkistymättömyys vastaa sellaisen verkon yhtenäisyyttä, jonka solmujen joukko on tila-avaruus $\mathcal{X}$ ja solmusta x solmuun y on kaari, jos $\mathbb{P}(x,y) > 0$. Pelkistymättömyys on triviaalisti voimassa, jos $\mathbb{P}(x,y) > 0$ kaikille tiloille x ja y.

Jaksottomuus: Markovin ketju on jaksoton (engl. aperiodic), jos joukon

$$\{m \geq 1 : \mathbb{P}^m(x,x) > 0\}$$

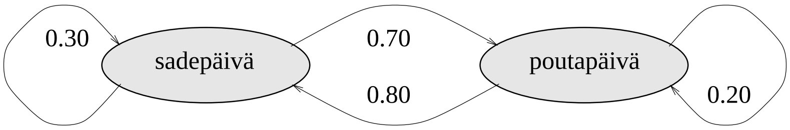
suurin yhteinen jakaja on 1 kaikille tiloille x. Jaksottomuus mm. estää prosessin juuttumisen vuorottelemaan joidenkin tilojen välillä. Jaksottomuus on triviaalisti voimassa, jos $\mathbb{P}(x,x) > 0$ kaikille tiloille x.

Pelkistymättömyys ja jaksottomuus takaavat, että äärellistilaisella Markovin ketjulla on yksikäsitteinen tasapainojakauma. Jotta ääretön- tai jatkuvatilaisella Markovin ketjulla olisi yksikäsitteinen tasapainojakauma, on niiden lisäksi toteutettava eräitä lisäehtoja [Rob95, Tie96].

Esimerkki 6.1. Tarkastellaan sadepäivien ja poutapäivien muodostamaa ketjua. Tilojen joukko olkoon $\mathcal{X} = \{x_1, x_2\}$, missä x_1 vastaa sadepäivää ja x_2 poutapäivää. Olkoon säätilojen muutosten todennäköisyydet kuvan 6.1 mukaiset. Sade- ja poutapäivien jono on tällöin Markovin prosessi, jonka tilasiirtymämatriisi on

$$\mathbb{P} = \begin{pmatrix} 0.30 & 0.70 \\ 0.80 & 0.20 \end{pmatrix}.$$

Nähdään, että pelkistymättömyys ja jaksottomuus ovat triviaalisti voimassa, koska ketjua vastaava verkko on täysin kytketty. Markovin ketjulla on siten yksikäsitteinen tasapainojakauma $\pi = (\pi(x_1), \pi(x_2))$, joka määräytyy yhtälöistä $\pi\,\mathbb{P} = \pi$ ja $\pi(x_1) + \pi(x_2) = 1$. Ratkaisemalla yhtälöryhmä saadaan $\pi \approx (0.533, 0.467)$. Tämä tarkoittaa, että pitkällä aikavälillä sadepäivien osuus on noin 53% ja poutapäivien osuus noin 47%. □



Kuva 6.1: Sade- ja poutapäivien väliset siirtymätodennäköisyydet. Nuoli osoittaa siirtymän suunnan ja sen yhteyteen kirjoitettu luku siirtymän todennäköisyyden.

Tasapainojakauma saavutetaan teoriassa vasta äärettömän monen askeleen jälkeen. Kohtuullisen hyvä otos tasapainojaukaumasta on $\{X_m, X_{m+1}, \ldots, X_n\}$, jos *konvergointijakson* pituus m ja *otantajakson* pituus n ovat riittävän suuria. Likiarvo odotusarvolle (6.1) saadaan tällöin *ergodisena keskiarvona*

$$\overline{f} = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t). \qquad (6.2)$$

Tämä on eräänlainen Monte Carlo -keskiarvo otoksesta, jonka näytteet ovat riippuvia. Jos m on pieni, seurauksena voi olla otoksen vinoutuminen ja siten huono likiarvo. Pieni n voi puolestaan aiheuttaa pyöristysvirheen (Monte Carlo -virhe). Peräkkäisten näytteiden riippuvuutta on luvallista heikentää ottamalla otokseen mukaan vain joka k:nnes tila jollakin sopivalla k.

6.1.4 Markovin ketjun rakentaminen ja simulointi

Edellisessä osiossa määritelty ergodinen keskiarvo on satunnaismuuttuja, jonka arvo määräytyy otoksen $\{X_m, X_{m+1}, \ldots, X_n\}$ realisaation mukaan. Kun Markovin ketju (aloitusjakauma ja tilasiirtymämatriisi) on annettu, realisaatio voidaan tuottaa suoraviivaisella ketjun simuloinnilla.

Esimerkki 6.2. Simuloidaan esimerkin 6.1 markovin ketjua. Olkoon aloitusjakaumassa molemmat tilat yhtä todennäköisiä. Arpomalla aloitustilaksi voidaan saada joko x_1 tai x_2 yhtä suurilla todennäköisyyksillä. Olkoon ensimmäinen tila x_2. Toinen tila on nyt x_2 todennäköisyydellä 0.20 ja x_1 todennäköisyydellä 0.80. Olkoon arpomisen tulos tila x_1. Jatkamalla näin 10 askelta saatu realisaatio voi olla esimerkiksi tilojen jono

$$(x_2, x_1, x_2, x_1, x_1, x_2, x_2, x_1, x_2, x_1).$$

Lasketaan tämän realisaation mukainen ergodinen keskiarvo funktiolle f, jolle $f(x_1) = 1$ ja $f(x_2) = 0$. Kun konvergointijakson pituudeksi valitaan $m = 3$ ja otantajakson pituudeksi $n = 10$, ergodinen keskiarvo, oikeammin sen realisaatio, on kaavan (6.2) mukaisesti

$$\overline{f} = \frac{1}{10-3}\sum_{t=4}^{10} f(x_t) = \frac{1}{7}(1+1+0+0+1+0+1) = \frac{4}{7}.$$

Tämä arvo on lähellä esimerkissä 6.1 laskettua arvoa $\pi(x_1)$ eli tilan x_1 suhteellista osuutta äärettömän pitkässä simulaatiossa. □

Alkuperäisessä ongelmassa ainoastaan tasapainojakauma π on annettu, ei aloitusjakaumaa eikä tilasiirtymämatriisia. Jos otoksen muodostamisessa halutaan hyödyntää Markovin ketjua, on sellainen ensin rakennettava. Ainoa vaatimus määriteltävälle Markovin ketjulle on, että sillä on yksikäsitteisenä tasapainojakaumanaan π. Tästä seuraa, että useat erilaiset siirtymätilamatriisit voivat olla kelvollisia. Lisäksi aloitusjakauma voi olla mielivaltainen. Seuraavaksi esitetään Hastingsin [Has70] ehdottama menetelmä sopivan Markovin ketjun muodostamiseen. Se on yleistys Metropolisin et al. [Met53] tulokselle.

Metropolisin ja Hastingsin algoritmi kuvaa tilasiirtymät kaksivaiheisesti. Ensimmäisessä vaiheessa ehdotetaan uutta tilaa; toisessa vaiheessa päätetään, hyväksytäänkö ehdotettu tila vai pysytäänkö edellisessä tilassa. Molemmat vaiheet ovat stokastisia. Tarkastellaan algoritmia tarkemmin kuvassa 6.2 esitettyyn muotoiluun pohjautuen. Algoritmi saa syötteenään jakauman π, aloitustilan X_0 jakauman ϕ ja ehdotusjakauman q sekä iteraatioiden lukumäärän n. Rivillä 1 arvotaan aloitustila X_0 jakauman ϕ mukaisesti. Tämän jälkeen siirrytään kaksivaiheisiin tilasiirtymiin. Rivillä 3 ehdokastila Y valitaan ehdotusjakauman q mukaisesti; askeleella t ehdokastila on ehdollinen tilalle X_t. Kun ollaan tilassa x, saadaan ehdokastilaksi y todennäköisyydellä

$$P(Y_t = y \mid X_t = x) \quad = \quad q(x, y).$$

Rivit 4–8 ilmaisevat, että ehdokastila hyväksytään todennäköisyydellä $\alpha(X_t, Y)$, jonka Hastings [Has70] määrittelee asettamalla

$$\alpha(X_t, Y) \quad = \quad \min\left\{1, \frac{\pi(Y)q(Y, X_t)}{\pi(X_t)q(X_t, Y)}\right\}.$$

Jos ehdokastila hylätään, pysytään edellisessä tilassa (rivi 8).

METROPOLIS-HASTINGS(π, ϕ, q, n)
1. valitse $X_0 \sim \phi(.)$
2. **jokaiselle** $t := 1, \ldots, n$ **:**
3. valitse $Y \sim q(X_t, .)$
4. valitse $U \sim \mathrm{Tas}(0, 1)$
5. **jos** $U \leq \alpha(X_t, Y)$ **niin**
6. $X_{t+1} := Y$
7. **muuten**
8. $X_{t+1} := X_t$
9. **palauta** $(X_0, X_1, \ldots, X_n)$

Kuva 6.2: Metropolisin ja Hastingsin algoritmi. Algoritmi saa syötteenään tasapainojakauman π, aloitusjakauman ϕ, ehdotusjakauman q ja iteraatioiden lukumäärän n. Algoritmi palauttaa simuloinnin tuottaman tilojen jonon.

Markovin ketjun tilasiirtymätodennäköisyydet saadaan ehdotustodennäköisyyksien ja hyväksymistodennäköisyyksien yhdistelmänä:

$$\mathbb{P}(x, y) \quad = \quad \begin{cases} q(x, y)\, \alpha(x, y), & \text{kun } x \neq y \\ 1 - \sum_{z \neq x} q(x, z)\, \alpha(x, z), & \text{kun } x = y \end{cases}.$$

Voidaan osoittaa, että $\pi = \pi\,\mathbb{P}$, mikä tarkoittaa, että π on kyseisen Markovin ketjun tasapainojakauma. Tasapainojakauman yksikäsitteisyyden takaavat ergodisuusehdot palautuvat ehdotusjakaumalta vaadittaviin ominaisuuksiin (tarkempi käsittely sivuutetaan).

Metropolisin ja Hastingsin algoritmissa on oleellista, että osamäärä

$$\frac{\pi(y)q(x,y)}{\pi(x)q(y,x)}$$

osataan laskea kaikille tiloille x ja y. Bayesiläisessä mallintamisessa on tyypillistä, että todennäköisyysjakauma π tunnetaan vakiokerrointa vaille. Osamäärää laskettaessa tämä kuitenkin riittää, sillä tuntematon kerroin supistuu pois.

Jos ehdotusjakauma on monimutkainen, todennäköisyyksien $q(x, y)$ laskeminen voi olla vaikeaa tai hidasta. Laskennan kannalta on mukavaa, jos ehdotusjakauma on symmetrinen, eli $q(x, y) = q(y, x)$ kaikilla $x, y \in \mathcal{X}$. Tällöin ehdotustodennäköisyyksiä ei tarvitse laskea, sillä hyväksymistodennäköisyys sievenee muotoon

$$\alpha(X_t, Y) \quad = \quad \min\left\{1, \frac{\pi(Y)}{\pi(X_t)}\right\}.$$

Tämä on Metropolisin et al. [Met53] esittämä hyväksymistodennäköisyys.

6.1.5 Konvergoinnista

Edellä todettiin, että Metropolisin ja Hastingsin algoritmilla voidaan muodostaa Markovin ketju, joka konvergoi tasapainojakaumaan iteraatioiden määrän kasvaessa rajatta. Käytännössä iteraatioita voidaan suorittaa vain rajallinen määrä ja olisikin hyödyllistä, jos osattaisiin sanoa jotain konvergoinnin asteen ja iteraatioiden määrän suhteesta. Tässä konvergoinnin asteella tarkoitetaan väljästi konvergoinnin hyvyyttä eli saavutetun otoksen sopivuutta tasapainojakauman arvointiin.

Kun valitaan otokselta vaadittava tarkkuus, voidaan teoriassa ennalta laskea riittävä pituus kovergointijaksolle ja otantajaksolle. Saatavat arviot ovat yleensä käytännön kannalta liian suuria [Rob95] ja siksi on pyritty kehittämään keinoja, joilla konvergoinnin astetta voidaan tarkastella simulaation kuluessa tai sen jälkikäsittelynä. Laskennallisia menetelmiä on esitetty, mutta käytännöllistä voi olla myös simulaation tulosteen visuaalinen tarkastelu. Laskennallisten menetelmien käsittely sivuutetaan tässä; tulosteiden visualisointiin palataan tämän luvun jälkimmäisissä osioissa ja myöhemmin luvussa 7.

6.2 Simulaatio-ohjelma GeME

Edellä esitettyä yleistä MCMC-menetelmää voidaan soveltaa bayesiläiseen päättelyyn geneettisten mallien perheessä, jolloin tavoitteena on luvussa 4 esitetyn kaavan 4.3 likimääräinen laskeminen. Markovin ketjun tilat ovat tällöin geneettisiä malleja ja ketjun tasapainojakauma on mallien posteriorijakauma. Tässä osiossa käsitellään menetelmän rajoitettua toteutusta, simulaatio-ohjelmaa GeME. Esitys noudattelee pääosin ohjelmassa tehtyjä rajoituksia. Toteutuksen suoranaiseen dokumentointiin ei kuitenkaan pyritä.

6.2.1 Havaintoaineistot

GeME:n keskeisin syöte on tautia kuvaava havaintoaineisto, joka koostuu taudin yleisyyttä ja ydinperheriskejä kuvaavista lukupareista (N_1, N), (N_1^O, N^O), (N_1^S, N^S) ja (N_1^{MZ}, N^{MZ}). Tällaisten havaintoaineistojen etuja käsiteltiin luvussa 2. Luvussa 5 annettiin menetelmiä, jotka mahdollistavat havaintoja vastaavien riskien nopean laskennan, kun geneettinen malli tunnetaan.

6.2.2 Mallit

GeMEin tuloste koostuu geneettisten mallien jonosta sekä joukosta aputulosteita. Ohjelma rajoittuu käsittelemään kaksialleelisten kytkemättömän periytymisen malleja, toisin sanoen malliperheen $\mathcal{M}_{[5]}$ jäseniä.

Markovin ketjun simuloiminen tila-avaruudessa $\mathcal{M}_{[5]}$ on ongelmallista, koska mallien parametrien lukumäärät vaihtelevat. Esimerkiksi yhden lokuksen malleissa parametreja on 4, kun niitä kolmen lokuksen malleissa on 30. Siirtyminen mallista toiseen voi siis vaatia hyppyä 4-ulotteisesta avaruudesta 30-ulotteiseen avaruuteen. Teoriassa tämä ei aiheuta ongelmaa ja esitetty Metropolisin ja Hastingsin algoritmi toimii moitteettomasti. Käytännössä ongelmia aiheuttaa sopivien ehdotusjakaumien valitseminen.

GeME välttää ongelman simuloimalla Markovin ketjua rajoittuen malliluokkaan $\mathcal{M}_{[5]}^{[\ell]}$, lyhyemmin $\mathcal{M}^{[\ell]}$. Lokusten lukumäärä ℓ annetaan ohjelmalle syötteenä. Tällöin tuloksena saatava mallien jono on satunnaisotos jakaumasta, joka on ehdollinen paitsi havaintoaineistolle, myös lokusten lukumäärälle. Rajoituksen merkittävä haitta on, että lokusten lukumäärän arvioiminen todennäköisyyksiin perustuen ei ole mahdollista. Erityisesti ei voida soveltaa esimerkissä 4.1 esitettyä lähestymistapaa. Jonkinlainen kuva lokusten määrän vaikutuksesta saadaan tarkastelemalla posteriorijakaumia eri malliluokissa $\mathcal{M}^{[\ell)}$, missä lokusten lukumäärä ℓ vaihtelee. Koska mallien parametrien lukumäärä kasvaa eksponentiaalisesti suhteessa lokusten lukumäärään, on simulaatiot käytännössä rajoitettava yhden, kahden ja kolmen lokuksen malliluokkiin. Lähestymistapaa tarkastellaan lähemmin luvussa 7. Vaihtoehtoisia tapoja pohditaan luvussa 8.

GeME käyttää osiossa 3.6 kuvattuja parametrien arvojen rajoituksia: symmetrian eliminointia ja ilmenemisyleisyyksien monotonisuutta. Kuten jo aiemmin todettiin, rajoituksia ei voi käyttää yhdessä. Ohjelman käyttäjän onkin valittava, kumpi rajoitus pidetään voimassa.

6.2.3 Markovin ketjun simulointi

GeME simuloi geneettisten mallien Markovin ketjua soveltaen Metropolisin ja Hastingsin algoritmia (kuva 6.2). Simulointia varten on valittava aloitusjakauma ϕ, tasapainojakauma π ja ehdotusjakauma q.

Aloitusjakauma ja tasapainojakauma

Aloitusjakauma voidaan valita mielivaltaisesti. Ainoa vaatimus on, että jakaumasta osataan arpoa näytteitä kohtuullisen helposti. GeME ei käytä varsinaista aloitusjakaumaa, vaan simulaatio aloitetaan aina mallista, jossa alleelifrekvenssien ovat 0.25 ja ilmenemisyleisyyksien arvot 0.50. Tämän ratkaisun ongelma on, että simulaation konvergointuminen voi kestää pitkään. Parempi olisi valita aloitusjakauma, joka on lähellä tasapainojakaumaa, mutta tämä on vaikeaa.

Tasapainojakauma on mallien posteriorijakauma. Kuten aiemmin esitettiin, posteriorijakauma määräytyy mallien priorien ja havaintojen uskottavuuden avulla vakiokerrointa vaille:

$$\pi(M) \;=\; p(M \mid D) \;\propto\; p(M)\, p(D \mid M) \;=\; p(M)\, L(M; D).$$

Kun käytetään sukulaisriskihavaintojen uskottavuudelle luvussa 4 esitettyä kaavaa, saadaan

$$\pi(M) \;\propto\; p(M) \prod_{R \in \mathcal{R}} \binom{N^R}{N_1^R} K_R^{N_1^R} (1 - K_R)^{N^R - N_1^R},$$

missä $\mathcal{R}$ sisältää taudin yleisyyttä, lapsiriskiä, sisarusriskiä ja identtisen kaksosen riskiä vastaavat sukulaisuustyypit.

Metropolisin ja Hastingsin algoritmin kannalta on riittävää, että $\pi(M)$ osataan laskea vakiokerrointa vaille jokaisella mallilla M. Sukulaisriskien K_R laskemista käsiteltiin aiemmin luvussa 5; aiheeseen palataan vielä osiossa 6.2.4. Prioritodennäköisyyksiä $p(M)$ tarkasteltiin aiemmin luvussa 4, jossa priorijakaumaksi ehdotettiin tasaista jakaumaa. Tällä valinnalla prioritodennäköisyydet voidaan jättää huomiotta. Toisena vaihtoehtona on ohjelmassa GeME kokeiltu prioria, jolle

$$p(M) \;\propto\; \prod_{i=1}^{\ell} (0.50 - |0.50 - p_i|), \tag{6.3}$$

missä p_i on i:nnen lokuksen tautialleelin frekvenssi. Tämä priori suosii malleja, joiden alleelifrekvenssit eivät ole äärimmäisen suuria (lähellä arvoa 1) tai äärimmäisen pieniä (lähellä arvoa 0).

Ehdotusjakauma ja iterointi komponenteittain

Teorian ja käytännön ero tulee näkyviin valittaessa sopivaa ehdotusjakaumaa q. Vaikka teoriassa lähes mikä tahansa ehdotusjakauma on kelvollinen, käytännössä simulaation konvergoinnin nopeus riippuu oleellisesti valitusta ehdotusjakaumasta.

Tarkastellaan ensin yksinkertaista tilannetta, jossa ehdotusjakauma on tasainen ja riippumaton edellisestä tilasta. Siis kaikilla $M, M' \in \mathcal{M}^{(\ell)}$ on voimassa

$$q(M, M') \;=\; \frac{1}{|\mathcal{M}^{(\ell)}|},$$

missä $|\mathcal{M}^{(\ell)}|$ on malliluokan koko. Tasaisen jakauman heikkous on, että ehdokastila hyväksytään harvoin. Tyypillisessä tapauksessa lähes kaikki geneettiset mallit sopivat havaintoaineistoon huonosti. Vastaavasti havaintoaineistoon hyvin sopivien mallien osuus on häviävän pieni.

Toinen äärimmäisyys saavutetaan, jos kaikilla $M, M' \in \mathcal{M}^{(\ell)}$ asetetaan

$$q(M, M') \;=\; \pi(M').$$

Tällöin jokainen ehdotettu tila hyväksytään. Ongelmana on luonnollisesti se, että ehdokastilojen generoiminen jakaumasta π ei onnistu helposti – tämähän on juuri tarkasteltavana oleva ongelma!

Sen sijaan, että yritettäisiin määritellä sopiva ehdotusjakauma kokonaisille malleille, luontevan tavan ehdotusjakauman muodostamiseen tarjoaa *komponenteittainen Metropolisin ja Hastingsin algoritmi* (aluperäinen ajatus Metropolisin et al. [Met53]), jossa ehdotusjakaumat määritellään mallin komponenteille erikseen. Komponentit ovat joko yksittäisiä parametreja tai näiden *ryppäitä*. Havainnollisuuden ja merkinnällisen yksinkertaisuuden vuoksi oletetaan seuraavaksi, että komponentit ovat yksittäisiä parametreja (alleelifrekvenssejä ja genotyyppien ilmenemisyleisyyksiä). Algoritmin muotoilu perustuu pääosin Gilksin et al. esitykseen [GRS96, s. 10–11]

Kirjoitetaan mallin M komponentit joukkona $\{M_{.1}, M_{.2}, \ldots, M_{.h}\}$, missä $h = \ell + 3^{\ell}$. Merkinnällä $M_{.j}$ tarkoitetaan mallin j:nnettä parametria viitaten tarpeen mukaan parametrin arvoon tai vastaavaan satunnaismuuttujaan. Otetaan lisäksi käyttöön merkintä

$$M_{.-j} \;=\; \{M_{.1}, \ldots, M_{.j-1}, M_{.j+1}, \ldots, M_{.h}\}.$$

Komponenteittaisessa Metropolisin ja Hastingsin algoritmissa (kuva 6.3) yksi askel on h iteraatiosta koostuva *iteraatiokierros*. Käytetään komponentin $M_{.j}$ iteraatiokierroksen t jälkeisestä arvosta merkintää $M_{t.j}$. Seuraavan iteraatiokierroksen j:nnessä iteraatiossa tilaa muutetaan käyttäen Metropolisin ja Hastingsin menetelmää. Todennäköisyys, että ehdokasarvoksi valitaan $M'_{.j}$ on $q_j(M_{t.j}, M_{t.-j}, M'_{.j})$, missä merkinnällä $M_{t.-j}$ tarkoitetaan $M_{.-j}$:n arvoa $j-1$ iteraation jälkeen iteraatiokierroksella $t+1$:

$$M_{t.-j} \;=\; \{M_{t+1.1}, \ldots, M_{t+1.j-1}, M_{t.j+1}, \ldots, M_{t.h}\}.$$

Ehdokasarvo hyväksytään todennäköisyydellä

$$\alpha(M_{t.-j}, M_{t.j}, M'_{.j}) \quad = \quad \min\left\{1, \frac{\pi(M'_{.j}, M_{t.-j})\, q_j(M'_{.j}, M_{t.-j}, M_{t.j})}{\pi(M_{t,j}, M_{t.-j})\, q_j(M_{t.j}, M_{t.-j}, M'_{.j})}\right\}.$$

METROPOLIS-HASTINGS-KOMPONENTEITTAIN(π, ϕ, q, n)

1. valitse $M_0 \sim \phi(.)$
2. **jokaiselle** $t := 1\ ,\dots,\ n$ **:**
3.   **jokaiselle** $j := 1\ ,\dots,\ h$ **:**
4.   valitse $M'_{.j} \sim q_j(M_{t.j}, M_{t.-j}, .)$
5.   valitse $U \sim \mathrm{Tas}(0,1)$
6.   **jos** $U \leq \alpha(M_{t.-j}, M_{t.j}, M'_{.j})$ **niin**
7. $M_{t+1.j} := M'_{.j}$
8.   **muuten**
9. $M_{t+1.j} := M_{t.j}$
10. **palauta** $(M_0, M_1, \dots, M_n)$

Kuva 6.3: Komponenteittainen Metropolisin ja Hastingsin algoritmi. Algoritmi saa syötteenään tasapainojakauman π, aloitusjakauman ϕ ja iteraatiokierrosten lukumäärän n. Ehdotusjakaumat kuvataan algoritmin syöteparametrissa q. Algoritmi palauttaa simuloinnin tuottaman geneettisten mallien jonon.

Algoritmin toteuttaminen ei ole ongelmatonta: on valittava sopivat ehdotusjakaumat; ehdotustodennäköisyydet on osattava laskea; hyväksymistodennäköisyyksien laskemisen on oltava numeerisesti vakaata. Ohjelmassa GeME nämä ongelmat on ratkaistu käyttämällä symmetrisiä ehdotusjakaumia ja laskemalla hyväksymistodennäköisyydet logaritmisina. Näin algoritmin rivillä 6 esitetty ehto voidaan esittää yhtäpitävästi muodossa

$$\log U \quad \leq \quad \log p(M') + l(M'; D) - \log p(M) - l(M; D).$$

Tässä M on simulaation edellinen tila (malli) ja M' ehdotettu uusi tila. Termit $l(M'; D)$ ja $l(M; D)$ ovat havaintojen D logaritmiset pseudouskottavuudet malleissa M' ja M (logaritmista pseudouskottavuutta käsiteltiin luvussa 4, ks. kaava (4.2)). Priorien (6.3) logaritmit voidaan laskea myös numeerisesti vakaasti logaritmien summina. Jos priorijakauma on tasainen, termit supistuvat pois. Symmetrian vuoksi ehdotusjakaumat supistuvat pois; ehdotusjakaumien muodostamista käsitellään tarkemmin jäljempänä.

Ehdotusjakaumat ohjelmassa GeME

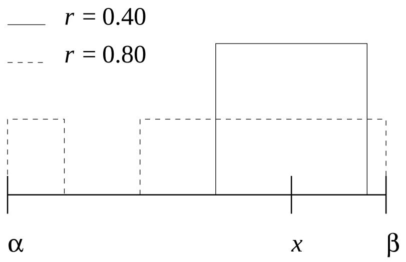
Ehdotusjakaumiin vaikuttavat luonnollisesti parametrien arvojoukoille valitut rajoitukset. Tarkastellaan seuraavaksi komponentin $M_{.j}$ ehdotusjakaumaa, kun ollaan tilassa M_t. Oletetaan ensin, että $M_{.j}$ on yksittäinen parametri.

Mikäli ilmenemisyleisyyksien ei vaadita olevan monotonisia, käytetään symmetrian eliminointiehtoa. Jos komponentti $M_{.j}$ on i:nnen lokuksen alleelifrekvenssi $p(A_{i1})$, ehdotusjakauma on $\mathrm{Tas}(0, \frac{1}{2})$. Jos komponentti $M_{.j}$ on genotyypin g ilmenemisyleisyys $f(g)$, ehdotusjakauma on $\mathrm{Tas}(0, 1)$. Pannaan merkille, että ehdotusjakaumat ovat symmetrisiä.

Mikäli ilmenemisyleisyyksien vaaditaan olevan monotonisia, ehdotusjakaumassa on huomioita kyseinen rajoitus. Oletetaan ensin, että komponentti $M_{.j}$ on genotyypin g ilmenemisyleisyys $f(g)$. Monotonisuusvaatimuksen vuoksi ei voida ehdottaa mielivaltaisia arvoja väliltä $[0, 1]$. Sen sijaan ehdotusjakaumaksi voidaan valita $\mathrm{Tas}(\alpha, \beta)$, missä alaraja α ja yläraja β määräytyvät mallin M_t ilmenemisyleisyyksistä. Jos komponentti $M_{.j}$ on alleelifrekvenssi $p(A_{i1})$, ehdotusjakauma on $\mathrm{Tas}(0, 1)$. Tässäkin tapauksessa ehdotusjakaumat ovat symmetrisiä.

Ohjelma GeME mahdollistaa parametrien *ryvästämisen* eli useista parametreista koostuvien komponenttien käytön. Jos sallitaan ilmenemisyleisyyksien epämonotonisuus, niin komponentin ehdotusjakauma muodostetaan yksittäisten parametrien jakaumien avulla. Ehdotusjakauma on silloin yksinkertaisesti yksittäisten parametrien jakaumien tulo. Jos ilmenemisyleisyyksien vaaditaan olevan monotonisia, ei ilmenemisyleisyyksiä voida ehdottaa riippumattomasti eri genotyypeille. Monotonisuuden säilyttävää parametriryppäiden ehdotusjakaumaa ei ohjelmassa GeME ole toteutettu.

Uskottavuusfunktion ominaisuuksien vuoksi yksittäisenkin parametriarvon suuria muutoksia hyväksytään harvoin. Jotta ehdotukset hyväksyttäisiin usein, on uutta arvoa ehdotettava edellisen arvon läheltä. Ohjelmassa GeME tämä seikka huomioidaan kaventamalla edellä esitettyjä tasaisia jakaumia erityisellä *kavennuskertoimella.* Ehdotusjakauman symmetrisyyden säilyttämiseksi menetellään seuraavasti. Jos parametrin ehdotusjakauma alunperin on $\mathrm{Tas}(\alpha, \beta)$, parametrin edellinen arvo on x ja kavennuskerroin on r, niin ehdokasarvo valitaan jakaumasta $\mathrm{Tas}(x - r(\beta - \alpha)/2, x + r(\beta - \alpha)/2)$. Jos x on lähellä arvoa α tai β, on ehdotusjakaumaa muokattava, jotta ei ehdotettaisi alarajaa α pienempiä tai ylärajaa β suurempia arvoja. Koska ehdotusjakauman halutaan olevan symmetrinen, jakaumaa ei vain katkaista ala- tai ylärajaan, vaan sitä "kierretään". Tätä havainnollistetaan kuvassa 6.4.



Kuva 6.4: Kavennetun tasaisen jakauman kierto. Yhtenäisellä viivalla kuvatussa ehdotusjakaumassa kavennuskerroin on 0.40. Katkoviivalla kuvatussa ehdotusjakaumassa kavennuskerroin on kaksinkertainen. Molemmissa tapauksissa parametrin edellinen arvo on x, parametrin arvoalueen alaraja on α ja yläraja on β.

6.2.4 Laskennan vaativuus

Markovin ketjun simulointi on laskentaintensiivistä. Hyvän konvergoinnin saavuttamiseksi simulaatiossa suoritettavien iteraatiokierrosten määrän n on käytännössä oltava luokkaa 10^6–10^8. Jokaisella iteraatiokierroksella geneettisen mallin mukaiset sukulaisriskit joudutaan laskemaan h kertaa, kun h on komponenttien määrä. Aikavaativuuden kannalta on oleellista, että sukulaisriskit lasketaan nopeasti.

Tarkastellaan laskennan vaativuutta olettaen ensin, että käytetään komponenteittaista Metropolisin ja Hastingsin algoritmia, jossa jokainen komponentti vastaa yhtä mallin parametria. Tällöin $h = \ell + 3^\ell$, kun simuloidaan malliluokassa $\mathcal{M}^{(\ell)}$. Luvussa 5 esitetyillä menetelmillä ydinperheriskit voidaan malliluokassa $\mathcal{M}^{(\ell)}$ laskea ajassa $\Theta(\ell\, 3^\ell)$. Jos riskit lasketaan jokaisella iteraatiokierroksen askeleella uudelleen, saadaan yhden iteraatiokierroksen aikavaativuudeksi $\Theta(\ell\, 9^\ell)$.

Näytetään sitten, kuinka hieman parempaan aikavaativuuteen päästään soveltamalla lisäyksellistä laskentaan. Palautetaan ensin mieleen taudin yleisyyden lauseke

$$K = \sum_{g^u \in \mathcal{G}} p(g^u) f(g^u).$$

Muistutetaan sitten, että sukulaisriskeille voidaan kirjoittaa yhtälö

$$K\,K_R = \sum_{g^u \in \mathcal{G}} \sum_{g^v \in \mathcal{G}} f(g^u) f(g^v) W^R(g^u, g^v),$$

jos R ei ole sukusiittoinen. (Näin on erityisesti, kun R on joko O, S tai MZ.)

Lisäyksellistä laskentaa voidaan nyt hyödyntää seuraavalla tavalla. Olkoon laskettuina taudin yleisyys $K(M)$ ja sukulaisriski $K_R(M)$ mallille $M = \langle p; f \rangle$. Olkoon malli $M' = \langle p'; f' \rangle$ ehdokastila, joka poikkeaa mallista M ainoastaan yhden parametrin suhteen. Oletetaan, että tämä parametri on ilmenemisyleisyys genotyypille $g \in \mathcal{G}$. Taudin yleisyys ehdokasmallissa M' on silloin

$$K(M') = \sum_{g^u \in \mathcal{G}} p'(g) f'(g^u) = K(M) + p(g)(f'(g) - f(g)).$$

Koska ilmenemisyleisyydet eivät vaikuta matriisin W^R alkioihin, sukulaisriskille saadaan yhtälö

$$\begin{aligned} K(M')\,K_R(M') &= \sum_{g^u \in \mathcal{G}} \sum_{g^v \in \mathcal{G}} f'(g^u) f'(g^v) W^R(g^u, g^v) \\ &= K(M)\,K_R(M) + 2(f'(g) - f(g)) \sum_{g^v \in \mathcal{G}} f(g^v) W^R(g, g^v) \\ &\quad + (f'(g)^2 - f(g)^2) W^R(g, g). \end{aligned}$$

Lisäyksen nähdään vievän ajan $\Theta(3^\ell)$.

Jos eroava parametri on alleelifrekvenssi, ei edellä esitettyä lisäyksellistä laskentaa voida soveltaa. Yksi iteraatiokierros vie siis yhteensä ajan $\Theta(3^\ell\,3^\ell + \ell\,\ell\,3^\ell) = \Theta(9^\ell)$. Nopeutus on siten ℓ-kertainen.

Käytettäessä parametrien ryvästystä ehdokasmalli voi poiketa simulaation edellisestä mallista usean parametrin osalta. Jos nämä parametrit ovat ilmenemisyleisyyksiä, lisäyksellinen laskenta voidaan suorittaa parametri kerrallaan. Tämä on kannattavaa kuitenkin vain, jos ryppään koko on pieni.

6.2.5 Simulaatioparametrit

Ohjelman GeME toteutuksessa on tehty useita rajoituksia, joiden muuttaminen vaatii uudelleen ohjelmointia. Ohjelman käyttäjällä on kuitenkin mahdollisuus vaikuttaa simulaation toimivuuteen *simulaatioparametreja* käyttämällä. Simulaatioparametrien arvot annetaan ohjelman parametreina ja niiden arvot pysyvät muuttumattomina ohjelman ajon aikana. Alla ovat oleellisimpien simulaatioparametrien kuvaukset.

Lokusten lukumäärä: Määrää malliluokan, josta tulosteena oleva otos simuloidaan.

Monotonisuus: Parametrilla on kaksi arvoa: Jos valitaan 'kyllä', niin rajoitutaan monotonisiin malleihin. Jos valitaan 'ei', niin mallien epämonotonisuus sallitaan.

Priorijakauma: Parametrilla on kaksi arvoa: Jos valitaan 'kyllä', käytetään aiemmin esitettyä "informatiivista" prioria, joka sakottaa mallia pienistä alleelifrekvensseistä. Jos valitaan 'ei', käytetään "epäinformatiivista" tasaista prioria.

Parametrien päivitysjärjestys: Määrää, miten parametreja ryvästetään ja missä järjestyksessä ryppäät päivetään käytettäessä komponenteittaista Metropolisin ja Hastingsin algoritmia.

Kavennuskerroin: Määrää, kuinka paljon alleelifrekvenssien ehdotusjakaumaa kavennetaan.

Iteraatiokierrosten lukumäärä: Määrää, kuinka monta tilaa Markovin ketjua simuloidaan. Jokainen tila eli geneettinen malli tulostetaan.

Diskretoinnin tiheys: Määrää mallien parametrien arvojoukon. Esimerkiksi tiheydellä 100 mallin parametrien arvojoukko on $\{0.00, 0.01, \ldots, 0.99, 1.00\}$. Jos tiheydeksi valitaan 0, käytetään satunnaislukugeneraattorin tarkkuutta (noin 2^{15}).

6.3 Simulaation tulosten käsittely

Edellä esiteltiin MCMC-menetelmä geneettisten mallien posteriorijakauman muodostamiseen. Todettiin, että simulaation tuloste on geneettisten mallien jono, jonka pituus voi olla miljoonia hyvän konvergoinnin saavuttamiseksi. Avoimeksi jätettiin, miten simulaation tuloste tiivistetään mielenkiintoiseksi tautia kuvaavaksi tiedoksi. Tulosteen jälkikäsittelyn tavoitteena on tällaisen tietämyksen lisäksi antaa tietoa simulaatiosta itsestään. Erityisesti konvergointiongelmien havaitseminen ja analysointi on välttämätöntä sopivampien simulaatioparametrien arvojen valitsemiseksi ja ehdotusjakaumien muuttamiseksi. Seuraavassa mainitaan muutamia yksinkertaisia jälkikäsittelymenetelmiä.

Konvergoinnin edellytyksenä on simulaation hyvä *sekoittuminen* (engl. mixing). Sekoittumisella tarkoitetaan parametrien arvojen vaihtelua simulaation aikana. Sekoittuminen on huonoa, jos ehdotettavat arvot hyväksytään harvoin tai jos simulaation peräkkäisillä askelilla parametrien arvot ovat hyvin lähellä toisiaan. Ensin mainittuun voi olla syynä arvojen ehdottaminen liian laajalta arvoalueelta. Jälkimmäinen voi olla seurausta liian kapeasta arvoalueesta. Sekoittumista voidaan tarkastella seuraamalla yksittäisten parametrien arvojen käyttäytymistä simulaation tuottamassa ketjussa.

Simulaation konvergointia vaikeuttaa tyypillisesti posteriorijakauman monihuippuisuus. Yhtenä syynä monihuippuisuuteen on geneettisten mallien symmetrisyys lokusten suhteen (kun lokuksia on vähintään kaksi). Vaikka symmetria vaikeuttaa konvergointia, se toimii toisaalta mittana konvergoitumisen asteelle: konvergoinnin edellytys on, että parametrien arvot ovat simulaation tuottamassa otoksessa jakautuneet lokusten suhteen likimäärin symmetrisesti. Symmetrisyys ei kuitenkaan takaa, että simulaatio olisi konvergoitunut hyvin. Symmetriaa voidaan tarkastella helposti piirtämällä kuvia alleelifrekvenssien pareittaisista yhteisjakaumista.

Simulaation tuottaman otoksen edustavuuden tutkimiseksi voidaan menetellä myös seuraavasti: Diskretoidaan parametrien arvojoukot sopivasti ja lasketaan kullekin parametrien arvoyhdistelmälle posteriorijakauman mukainen arvo (vakiokerrointa vaille). Näin saadaan hahmoteltua eräänlainen pinta, jonka korkeita kohtia pitäisi vastata tiheämmässä olevat otoksen näytteet. Laskennallisen vaativuuden vuoksi diskretointia ei voida käytännössä pitää tiheänä kuin muutamalle parametrille, minkä vuoksi varsinaiset huippukohdat jäävät pinnalla näkymättä.

Jos simulaation arvellaan konvergoineen riittävän hyvin, voidaan siirtyä MCMC-menetelmän toiseen vaiheeseen, Monte Carlo -integrointiin. Kuten teoriaosuudessa (ks. osio 6.1) esitettiin, Monte Carlo -integroinnilla pyritään laskemaan likimääräisesti odotusarvo sopivalle geneettisten mallien funktiolle. Bayesiläisessä päättelyssä odotusarvo otetaan yli mallien posteriorijakauman (ks. luku 4). Käytettävä funktio valitaan sen mukaan, mistä päättelyssä ollaan kiinnostuneita. Muistutetaan, että Monte Carlo -integrointi Markovin ketjun realisaatiota käyttäen tarkoittaa ergodisen keskiarvon laskemista.

Esimerkki 6.3. Olkoon mielenkiinnon kohteena todennäköisyys, että geneettisen mallin j:nnen parametrin on välillä $[y_1, y_2]$. Laskemalla ergodinen keskiarvo funktiolle z, jolle $z(M_t) = 1_{\{y_1 \leq M_{t.j} \leq y_2\}}$, saadaan todennäköisyydelle arvio

$$\overline{z} \;=\; \frac{1}{n-m} \sum_{t=m+1}^{n} z(M_t).$$

Tässä m on valitun konvergointijakson ja n otantajakson pituus. □

Luvussa 4 näytettiin, että lokusten lukumäärä voidaan päätellä bayesiläisittäin mallien posteriorijakaumaan nojautuen. Myöhemmin (osiossa 6.2.2) kuitenkin todettiin, että Genesist -ohjelmalla tuotettavat otokset eivät tähän riitä, koska otoksia ovat ehdollisia lokusten lukumäärälle. Jonkinlaista tietoa lokusten lukumäärän vaikutuksesta saadaan simuloimalla otokset yhden, kahden ja kolmen lokuksen malliluokille. Malliluokkien $\mathcal{M}^\ell$ sopivuutta havaintoaineistoon voidaan tällöin arvioida vertaamalla aineistojen keskimääräistä uskottavuutta eri malliluokissa. Jos esimerkiksi uskottavuus kahden ja kolmen lokuksen malleilla on suunnilleen yhtä suurta keskimäärin ja yhden lokuksen malleilla selvästi näitä pienempää, aineiston taustalla voidaan otaksua olevan kaksi vaikuttavaa lokusta. Tätä ajatusta havainnollistetaan koetuloksien yhteydessä luvussa 7. On syytä korostaa, että edellä mainittu päättely on heuristista ja karkeaa. Se ei vastaa lokusten lukumäärän bayesiläistä päättelemistä.

Luku 7

Koetuloksia

Geneettisten mallien bayesiläistä tarkastelua varten edellisessä luvussa 6 kuvattiin MCMC-menetelmä ja sen toteutus, ohjelma GeME. Tässä luvussa esitetään joitakin menetelmään ja toteutukseen liittyviä kysymyksiä, ja pyritään vastaamaan niihin kokeellisen tutkimuksen avulla. Samalla havainnollistetaan joitakin edellisessä luvussa esitettyjä simulaation jälkikäsittelykeinoja.

7.1 Kokeiden tavoitteet

Simulaatiokokeilla pyrittiin löytämään vastauksia seuraaviin kysymyksiin:

1. Mitkä tekijät vaikuttavat simulaation konvergointiin?
2. Minkälaisia ovat geneettisten mallien posteriorijakaumat? Miten yliparametrisointi ja symmetriat tulevat näkyviin simuloidussa otoksessa?
3. Voidaanko simulaation tuottamista otoksista päätellä, kuinka monta lokusta havaintoaineistoon liittyvään tautiin oleellisesti vaikuttaa?

Kattavan vastauksen saaminen asetettuihin kysymyksiin vaatisi huomattavia määriä erilaisia koeaineistoja ja simulaatioparametrien systemaattista vaihtelua. Tässä esiteltävissä kokeissa kattavuuden tavoittelusta luovuttiin ja tyydyttiin kohtuulliseen määrään kokeita. Tarkastelun ulkopuolelle jätettiin useita mielenkiintoisia kysymyksiä liittyen havaintoaineiston laatuun, laskennan nopeuteen ja simulaatioparametrien vaikutukseen. Joitakin jatkotutkimuksen aiheita hahmotellaan luvussa 8.

7.2 Koejärjestelyt

Seuraavaksi kuvataan simulaatiokokeiden koepisteet. Kukin koepiste vastaa yhtä simulaatiota, jossa mallien lokusten lukumäärä, havaintoaineisto ja simulaatioparametrit on kiinnitetty. Koejärjestelyissä keskityttiin vaihteleviin havaintoaineistoihin. Sen sijaan simulaatioparametreille kokeiltiin vain kahta erilaista arvoyhdistelmää.

7.2.1 Keinotekoiset aineistot

Keinotekoisten havaintoaineistojen luonti mahdollistaa aineiston piirteiden hallinnan. Ensinäkemältä aineiston muodostaminen vaikuttaa yksinkertaiselta, koska jokaisen havaintoaineiston määrää yksikäsitteisesti jo kahdeksan kokonaislukua. Kokeiden tavoitteiden kannalta edustavan havaintoaineistojoukon luomiseksi on kuitenkin tarkasteltava geneettisten mallien piirteitä. Ongelmana on mallien moniulotteisuus.

Tässä esiteltävissä kokeissa mallien piirteet pelkistettiin kahteen ulottuvuuteen. Ensimmäinen ulottuvuus on lokusten lukumäärä (yksi, kaksi tai kolme). Toisessa ulottuvuudessa pyrittiin kuvaamaan mallien muita oleellisia piirteitä lähtien vastakkainasettelusta peittyvän ja vallitsevan muodon välillä. Peittyvät ja vallitsevat mallit ovat kuitenkin samankaltaisia täydellisten (0/1-arvoisten) ilmenemisyleisyyksiensä vuoksi. Siksi kokeissa tarkasteltiin lisäksi muotoa, jossa ilmenemisyleisyydet ovat epätäydellisiä. Kahden ulottuvuuden yhdistelmänä saadaan yhdeksän erilaista mallityyppiä.

Mallityyppi ei vielä määrää geneettistä mallia yksikäsitteisesti. Mallityypin edustajan valitsemiseksi on kiinnitettävä alleelien frekvenssit ja genotyyppien ilmenemisyleisyydet. Alleelifrekvenssit laskettiin kahden ehdon perusteella. Ensinnäkin jokaisen tautialleelien frekvenssit valittiin yhtäsuuriksi. Lisäksi vaadittiin, että mallin kuvaaman taudin yleisyys on noin 1%, joka on samaa suuruusluokkaa kuin esimerkiksi diabeteksen ja skitsofrenian yleisyys (ks. osio 7.2.2). Epätäydellisten mallien edustajille on alleelifrekvenssien lisäksi valittava genotyyppien ilmenemisyleisyydet. Jos mallityypille valittaisiin useita edustajia, ilmenemisyleisyyksien satunnainen valitseminen sopivasta jakaumasta olisi järkevää. Kokeissa tyydyttiin vain yhteen edustajaan ja ilmenemisyleisyyksien arvojen asettamiseen niin, että malli vaikuttaa biologisesti järkevältä ja on riittävän erilainen kuin vallitsevat ja peittyvät mallit. Poiketen peittyvistä ja vallitsevista malleista epätäydellisten mallien alleelifrekvenssit valittiin keskenään erisuuriksi.

Näin saatiin yhteensä yhdeksän mallia vastaten eri muotoja ja lokusten lukumääriä. Näitä malleja kutsutaan jäljempänä *generoiviksi* malleiksi. Malleista käytetään merkintöjä $M_{i,\ell}$, missä i on joko R, D tai C viitaten vastaavasti peittyvään, vallitsevaan tai epätäydelliseen periytymismuotoon, ja ℓ on lokusten lukumäärä. Generoivien mallien parametrien arvot luetellaan liitteessä A.1.

Keinotekoiset havaintoaineistot muodostettiin laskemalla jokaiselle generoivalle mallille taudin yleisyys ja ydinperheriskit sekä valitsemalla havaintojen tarkkuus. Aineiston tarkkuus määrättiin valitsemalla aineiston kokoa kuvaavat luvut N, N^O, N^S, N^{MZ}. Havaintoaineistot muodotettiin optimaaliseksi suhteessa havaittujen yksilöiden määrään. Niinpä sairaiden havaittujen yksilöiden lukumäärät $N_1, N_1^O, N_1^S, N_1^{MZ}$ määrättiin pyöristämällä kokonaisluvuiksi generoivien mallin mukaiset odotusarvot KN, $K_O N^O$, $K_S N^S$ ja $K_{MZ} N^{MZ}$.

Havaintoaineistojen tarkkuus muodostaa koeaineistojen kolmannen ulottuvuuden. Kokeissa käytettiin kolmea tarkkuutta. *Heikko tarkkuus* saatiin valitsemalla $N = 10000$, $N^O = 100$, $N^S = 100$ ja $N^{MZ} = 10$, *kohtuullinen tarkkuus* valitsemalla kymmenkertaiset ja *hyvä tarkkuus* valitsemalla satakertaiset arvot. Generoivaa mallia $M_{i,\ell}$ vastaavasta hyvän tarkkuuden aineistosta käytetään jäljempänä merkintää $D_{i,\ell}$, kohtuullisen tarkkuuden aineistosta merkintää $D'_{i,\ell}$ ja heikon tarkkuuden aineistosta merkintää $D''_{i,\ell}$. Taulukossa 7.1 luetellaan hyvän tarkkuuden yhdeksän koeaineistoa. Kohtuullinen ja heikko tarkkuus saadaan jakamalla luvut vastaavasti kymmenellä ja sadalla.

Nimi	N	N_1	N^O	N_1^O	N^S	N_1^S	N^{MZ}	N_1^{MZ}
$D_{R,1}$	1 000 000	10 000	10 000	1 000	10 000	3 025	1 000	1 000
$D_{D,1}$	1 000 000	10 000	10 000	5 038	10 000	5 044	1 000	1 000
$D_{C,1}$	1 000 000	10 000	10 000	236	10 000	283	1 000	56
$D_{R,2}$	1 000 000	10 000	10 000	1 000	10 000	1 876	1 000	1 000
$D_{D,2}$	1 000 000	10 000	10 000	2 896	10 000	2 960	1 000	1 000
$D_{C,2}$	1 000 000	10 035	10 000	607	10 000	673	1 000	162
$D_{R,3}$	1 000 000	10 000	10 000	1 000	10 000	1 539	1 000	1 000
$D_{D,3}$	1 000 000	10 000	10 000	1 991	10 000	2 115	1 000	1 000
$D_{C,3}$	1 000 000	10 010	10 000	384	10 000	399	1 000	110

Taulukko 7.1: Keinotekoisia hyvän tarkkuuden havaintoaineistoja. N_1 yksilön on havaittu olevan sairas N tutkitusta yksilöstä. Merkinnät N^O, N_1^O, N^S, N_1^S, N^{MZ} ja N_1^{MZ} viittaavat vastaavasti jälkeläisriskiin, sisarusriskiin ja identtisten kaksosten riskiin.

7.2.2 Todelliset aineistot: diabetes ja skitsofrenia

Keinotekoisten aineistojen lisäksi kokeissa tarkasteltiin kahta todellista havaintoaineistoa. Varhaisiän diabetes on yleinen ja paljon tutkittu tauti, jonka uskotaan olevan perinnöllinen. Taulukossa 7.2 luetellut riskihavainnot perustuvat lähteeseen [Onk99a] (samansuuntaisia lukuja esitti Rich [Ric90]). Toisena tautina tarkasteltiin skitsofreniaa. Taulukossa 7.2 luetellut riskihavainnot perustuvat Rischin [Ris90] esittämiin lukuihin. Lähteiden riskihavainnot esitettiin eri muodossa, minkä vuoksi taulukon riskiarviot ovat eräänlaisia keskiarvoja, ja tarkkuutta kuvaavat luvut ovat väljiä arvioita. Keinotekoisiin aineistoihin suhteutettuna sekä *Diabetes*- että *Skitsofrenia*-aineiston tarkkuutta voidaan pitää kohtuullisena.

Nimi	N	N_1	N^O	N_1^O	N^S	N_1^S	N^{MZ}	N_1^{MZ}
Diabetes	50 000	250	5 000	265	5 000	500	32	10
Skitsofrenia	10 000	85	1 000	85	1 000	73	100	44

Taulukko 7.2: Diabeteksen ja skitsofrenian havaittuja ydinperheriskejä ja taudin yleisyyttä kuvaavat luvut. N_1 yksilön on havaittu olevan sairas N tutkitusta yksilöstä. Merkinnät N^O, N_1^O, N^S, N_1^S, N^{MZ} ja N_1^{MZ} viittaavat vastaavasti jälkeläisriskiin, sisarusriskiin ja identtisen kaksosen riskiin.

7.2.3 Simulaatioparametrien arvot ja kokeiden suoritus

Jokaiselle havaintoaineistolle suoritettiin simulaatiot yhden, kahden ja kolmen lokuksen malleilla. Muut simulaatioparametrien arvot pidettiin kiinnitettyinä seuraavasti:

- Mallien vaadittiin olevan monotonisia.
- Käytettiin "informatiivista" priorijakaumaa
- Parametrien päivitys tehtiin yksitellen ryvästämättä (ks. liite A.3).
- Iteraatiokierroksia suoritettiin 10^8.
- Ehdotusjakaumien kavennuskerroin oli 0.10.
- Diskretointi oli tihein mahdollinen.

Kokeet vaativat yhteensä $3 \cdot 3 \cdot 3 \cdot 3 = 81$ simulaatiota keinotekoisilla aineistoilla ja $2 \cdot 3 = 6$ simulaatiota todellisilla aineistoilla. Jokainen simulaatio vastaa yhtä koepistettä, joita siten on yhteensä 87. Koska simulaatiot ovat stokastisia, useiden ajojen suorittaminen kullekin koepisteelle olisi perusteltua. Kokeissa tyydyttiin kuitenkin yhteen pitkään simulaatioon jokaisessa koepisteessä. Huomautettakoon, että vaativimmillaan (kolmen lokuksen malleilla) yksi simulaatio vie tyypillisellä Pentium III -koneella aikaa kymmeniä tunteja.

Jokaisen simulaation tuloksena saatiin 10^8 mallin jono. Lopullinen otos muodostettiin jättämällä 10^7 ensimmäistä mallia pois (konvergointijakso) ja valitsemalla otokseen jäljelle jäävistä malleista tasavälein joka 10 000. malli. Jokainen otos koostui siten 9 000 mallista. Seuraavassa osiossa esitettävät tarkastelut perustuvat näihin otoksiin. Joissakin tapauksissa otos harvennettiin edelleen 1 000 malliin.

Edellä kuvatut peruskokeet muodostivat kokeiden ensimmäisen vaiheen. Koska tulokset vallitsevan muodon aineistoilla osoittautuivat heikoiksi, kokeita täydennettiin lisäkokeilla. Toisessa vaiheessa vallitsevan muodon aineistoille suoritettiin simulaatiot yhden, kahden ja kolmen lokuksen malleilla. Simulaatioparametrien asetukset poikkesivat peruskokeissa käytetyistä arvoista merkittävästi:

- Mallien sallittiin olevan epämonotonisia.
- Parametrien päivitys tehtiin ryvästystä hyödyntäen (ks. liite A.3).
- Käytettiin "epäinformatiivista" priorijakaumaa
- Iteraatiokierroksia suoritettiin 10^7.
- Ehdotusjakaumien kavennuskerroin oli 1.00 (eli ei kavennusta).

Lisäkokeissa tarkasteltiin siten 27 koepistettä. Kunkin simulaation tuottamasta otoksesta lopullinen 9 000 mallin otos muodostettiin jättämällä pois 10^6 ensimmäistä mallia ja valitsemalla sitten otokseen joka 1 000. malli.

7.3 Tulosten tarkastelu

Yli sadan simulaation tuloksena saatiin yhteensä miljoonia lukuja.[1] Näiden tulosten kattavan analysoinnin sijaan tässä tavoitteena on tuoda valikoivasti esille kokeissa havaittuja menetelmälle ja aineistoille tyypillisiä piirteitä sekä havainnollistaa osiossa 6.3 mainittuja visuaalisia jälkikäsittelykeinoja. Pääosin pitäydytään peruskokeiden tulosten selvitämisessä. Kun tarkastellaan lisäkokeissa tuotettuja otoksia, se mainitaan erikseen. Tällöin puhutaan lyhyemmin ja hieman epätarkasti epämonotonisista malleista, kun tarkoitetaan epämonotonisuuden sallivan malliluokan malleja.

Ensin rajoitutaan keinotekoisiin aineistoihin ja tarkastellaan mallien posteriorijakaumia simulaation tuottamiin otoksiin nojautuen. Tarkasteluun yhdistetään simulaation konvergoinnin asteen arviointi – otosten edustavuus edellyttää simulaation hyvää sekoittumista. Tämän jälkeen siirrytään vertaamaan yhden, kahden ja kolmen lokuksen mallien sopivuutta keinotekoisiin koeaineistoihin. Lopuksi esitellään diabetes- ja skitsofrenia-aineistoilla saatuja tuloksia.

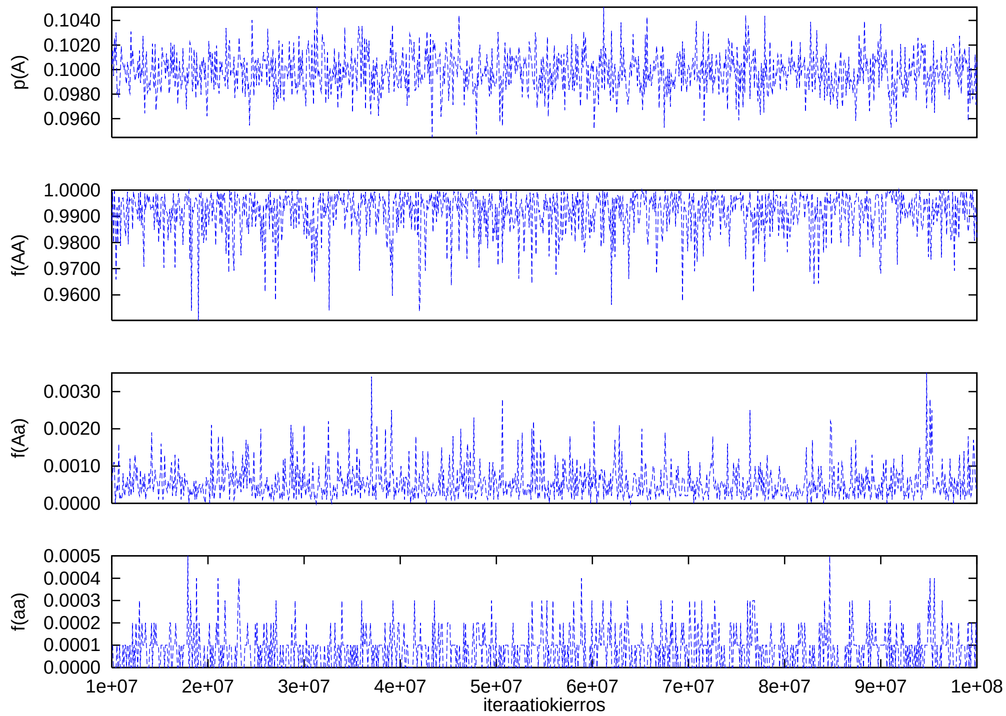
[1] Simulaatiotulokset ovat saatavissa WWW-osoitteessa
`http://www.cs.helsinki.fi/~mkhkoivi/GeME/simulation_results/`

7.3.1 Posteriorijakaumat ja konvergointi

Seuraavassa edetään havaintoaineistojen taustalla olevan generoivan mallin lokusten lukumäärän mukaan. Kunkin havaintoaineiston kohdalla keskitytään pääosin sellaisten mallien tarkasteluun, joissa lokusten lukumäärä on sama kuin aineiston generoivassa mallissa.

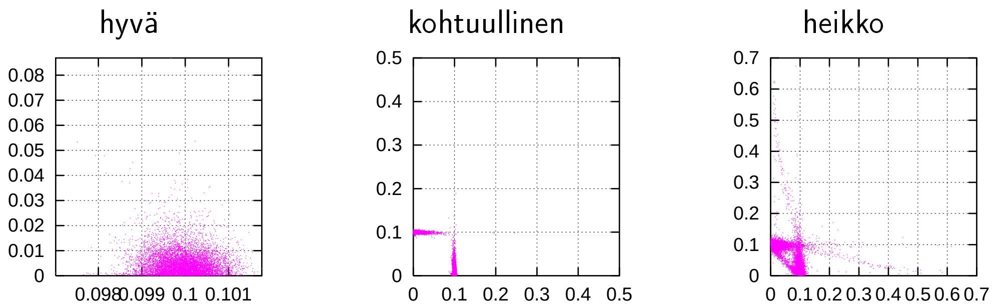
Yhden lokuksen aineistot

Aloitetaan tarkastelu peittyvän muodon aineistosta $D'_{R,1}$. Kuvasta 7.1 nähdään, kuinka yhden lokuksen mallien parametrien arvot muuttuvat simulaation kuluessa. Suurimman osan simulaatioajasta mallit ovat hyvin lähellä generoivaa mallia $M_{R,1} = \langle 0.1; 1.0, 0.0, 0.0 \rangle$. Esimerkiksi alleelifrekvenssi $p(A)$ on melkein jokaisella askeleella lähellä arvoa 0.1.



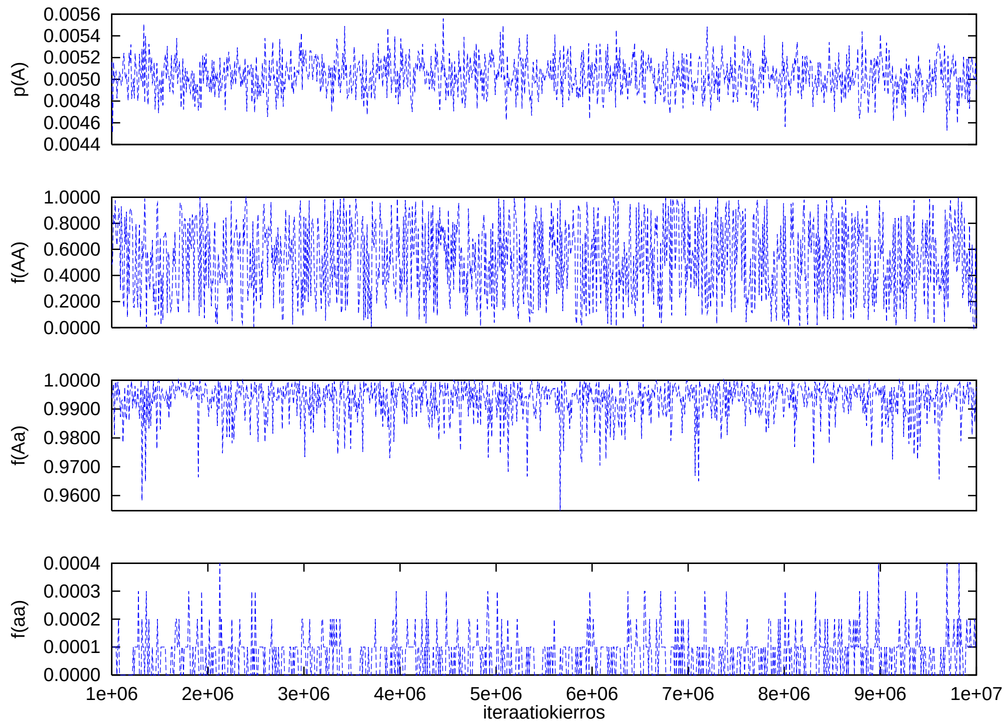
Kuva 7.1: Yhden lokuksen mallien parametrien arvojen muuttuminen simulaation aikana, kun aineistona on $D'_{R,1}$. Kuvaa varten 9 000 mallista on poimittu tasavälein 1 000 mallia.

Jatketaan yhden lokuksen peittyvän muodon havaintoaineistoilla tarkastelemalla nyt kahden lokuksen malleja. Kuva 7.2 esittää ensimmäisen ja toisen lokuksen alleelifrekvenssien yhteisjakaumia hyvän, kohtuullisen ja heikon tarkkuuden aineistoilla. Kuvan mukaan toisen lokuksen alleelifrekvenssin ollessa lähellä generoivan mallin alleelifrekvenssiä 0.1 toisen alleelin frekvenssi voi vaihdella. Tämä on ymmärrettävää, koska aineiston selittämiseksi riittää yksi vaikuttava lokus, jolloin toinen lokus voi olla ilmenemisyleisyyksiensä puolesta vaikutukseltaan mitätön. Merkille pantavaa on, että aineistoon sopivia kahden lokuksen malleja, joilla molemmat alleelifrekvenssit poikkeavat arvosta 0.1, ei juuri ole. Poikkeuksena on heikon tarkkuuden aineisto, johon sopivia alleelifrekvenssejä löytyy pisteiden $(0.1, 0.0)$ ja $(0.0, 0.1)$ välisen janan läheltä. Konvergointi havaitaan puutteelliseksi hyvän tarkkuuden aineistolla: vain toinen huippu (tai harjanne) on löytynyt.



Kuva 7.2: Kahden lokuksen monotonisten mallien alleelifrekvenssien yhteisjakaumat hyvän, kohtuullisen ja heikon tarkkuuden peittyvän muodon aineistoilla.

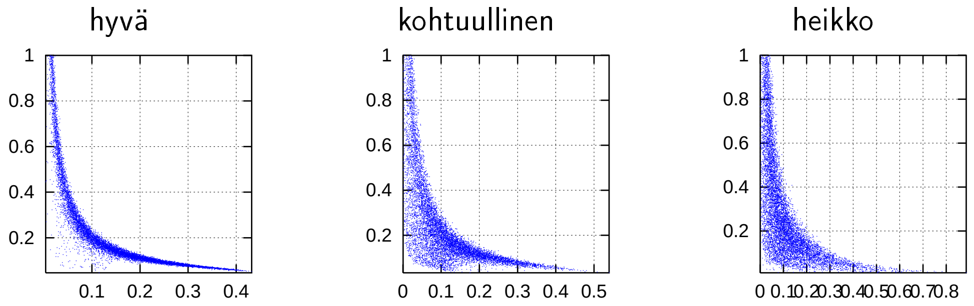
Siirrytään sitten tarkastelemaan vallitsevan muodon yhden lokuksen aineistoja. Simulaatiot monotonisilla malleilla tuottivat kaikilla kolmella tarkkuudella otoksen, jossa jokainen malli on likipitäen $\langle 0.1; 1.0, 0.0, 0.0\rangle$. Tämä ei ole toivottu tulos – aineistot generoiva mallihan oli $M_{D,1} = \langle 0.005; 1.0, 1.0, 0.0\rangle$! Tämä osoittaa, että jo neljän parametrin malleilla simulaatio voi juuttua vaikeasti ja tuottaa pahasti vääristyneen otoksen. Lisäkokeet epämonotonisilla malleilla antoivat halutun tuloksen. Kuvasta 7.3 nähdään, kuinka parametrien arvot simulaation aikana pysyvät lähellä generoivan mallin arvoja. Poikkeuksena on ilmenemisyleisyys $f(AA)$, jonka arvo vaihtelee arvoalueellaan vapaasti. Tämä on ymmärrettävää, koska alleelin A frekvenssin vuoksi genotyyppi AA on niin harvinainen, että sen ilmenemisyleisyydellä ei ole merkitystä taudin yleisyyteen ja sukulaisriskeihin.



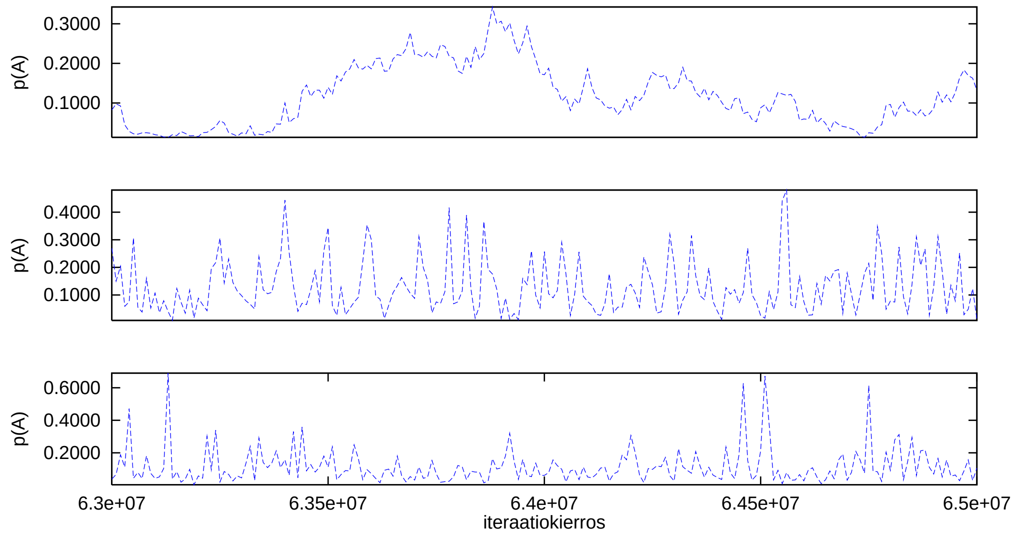
Kuva 7.3: Yhden lokuksen epämonotonisten mallien parametrien arvojen muuttuminen simulaation aikana, kun aineistona on $D'_{D,1}$. Kuvaa varten 9 000 mallista on poimittu tasavälein 1 000 mallia.

Tarkastellaan vielä epätäydellisen muodon yhden lokuksen aineistoja. Kuva 7.4 näyttää alleelifrekvenssin $p(A)$ ja ilmenemisyleisyyden $f(AA)$ riippuvuuden aineiston eri tarkkuuksilla. Toisin kuin täydellisen muodon aineistoilla parametrit saavat arvoja suurella arvoalueella. Osa otoksen malleista on lähellä generoivaa mallia $M_{C,1} = \langle 0.0592, 0.320, 0.040, 0.005\rangle$, mutta nämä mallit eivät ole otokselle erikoisen tyypillisiä. Tämä kertoo yliparametrisoinnista. Pannaan myös merkille, että tarkemmalla aineistolla parametrien riippuvuus on voimakkaampaa.

Tarkkuuden vaikutus simulaation sekoittumiseen paljastuu kuvassa 7.5. Havaitaan, että hyvän tarkkuuden aineistossa otoksen peräkkäisillä näytteillä on selvä autokorrelaatio ja siten simulaation sekoittuminen on hidasta. Kohtuullisen ja heikon tarkkuden aineistossa tällaista ei juurikaan ei havaita, ja ero näiden kahden tarkkuuden kesken on pieni. Nähdään myös, että tarkemmalla aineistolla parametrin arvojen varianssi on pienempi. Tämä seikka voitiin havaita jo kuvasta 7.4.



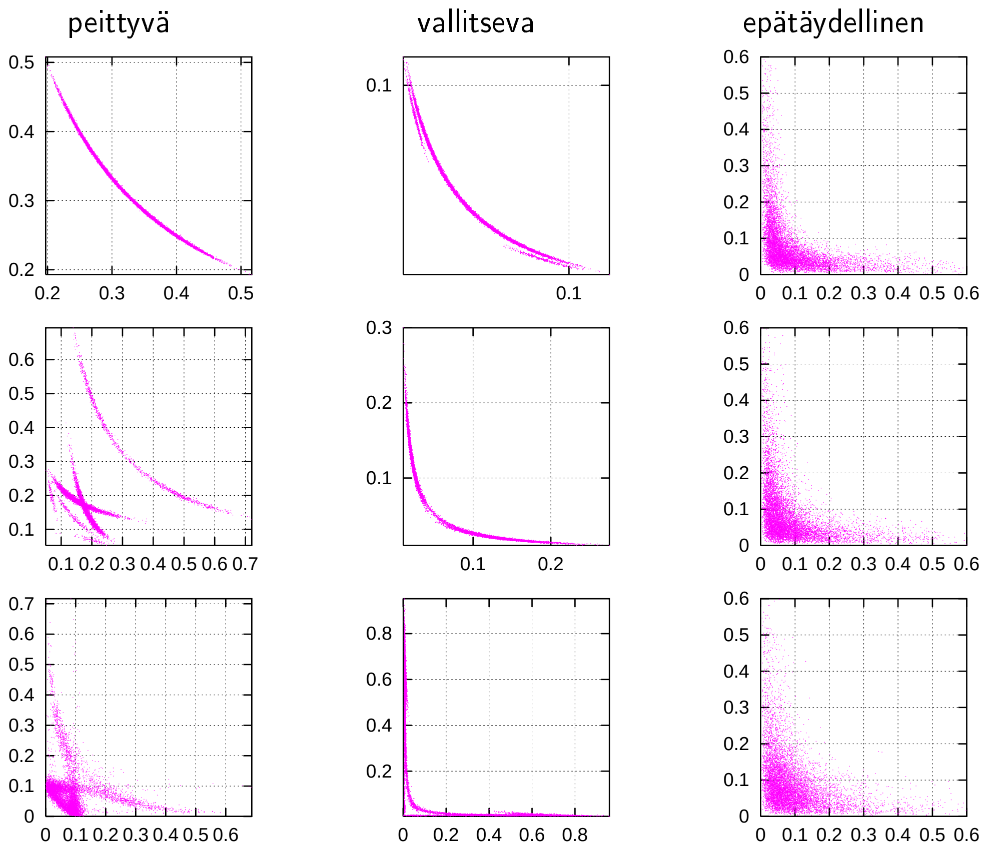
Kuva 7.4: Yhden lokuksen monotonisten mallien alleelifrekvenssin $p(A)$ (vaaka-akseli) ja ilmenemisyleisyyden $f(AA)$ (pystyakseli) yhteisjakaumat yhden lokuksen epätäydellisen muodon havaintoaineistolla.



Kuva 7.5: Yhden lokuksen monotonisten mallien alleelifrekvenssien arvojen muuttuminen simulaation keskivaiheilla, kun aineistoina ovat ylhäältä lukien $D_{C,1}$, $D'_{C,1}$ ja $D''_{C,1}$.

Kahden lokuksen aineistot

Mielenkiintoista tietoa kahden lokuksen mallien posteriorijakaumista antaa kuva 7.6. Jakaumien symmetria antaa olettaa, että simulaation konvergenssi on vähintään kohtuullista. Seuraavaksi posteriorijakaumien piirteitä tarkastellaan lähemmin. (Muutamia otoksia edustavia esimerkkimalleja annetaan liitteessä B.1.)

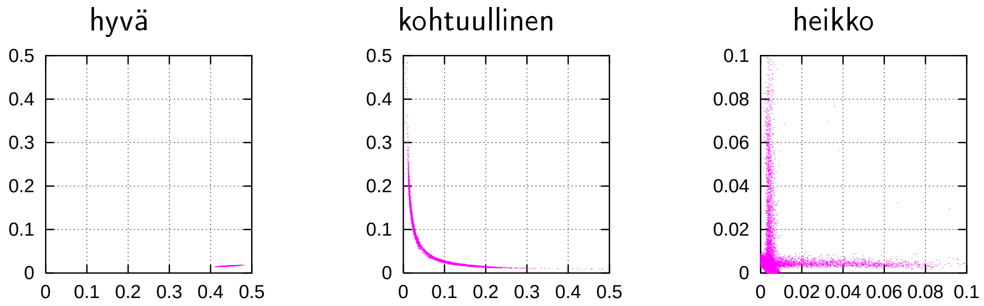


Kuva 7.6: Alleelifrekvenssien yhteisjakauma jakauma kahden lokuksen monotonisissa malleissa eri aineistoilla. Vaaka-akselilla on ensimmäisen lokuksen alleelin A frekvenssi ja pystyakselilla toisen lokuksen alleelin B frekvenssi. Ylimpänä on jakaumat hyvän tarkkuuden aineistoilla, keskellä kohtuullisen tarkkuuden aineistoilla ja alimpana heikon tarkkuuden aineistoilla.

Havaitaan, että kohtuullisen tarkkuuden aineistolla alleelifrekvenssien posterioritodennäköisyydet pisteen $(0.3, 0.3)$ läheltä kulkevalla käyrällä ovat suurehkoja. Sen sijaan heikon tarkkuuden aineistolla $D''_{R,2}$ todennäköisyysmassa keskittyy pienille alleelifrekvensseille ja otoksen jakauma muistuttaa yhden lokuksen peittyvän aineiston $D''_{R,1}$ tapausta (ks. kuva 7.2). Muistutetaan, että generoivassa mallissa $M_{R,2}$ molempien alleelifrekvenssit olivat 0.316. Kuvien nojalla generoivan mallin läheisyydessä posterioritodennäköisyys on hyvän tarkkuuden aineistolla suuri ja kohtuullisen tarkkuuden aineistolla melko suuri mutta heikon tarkkuuden aineistoilla pieni.

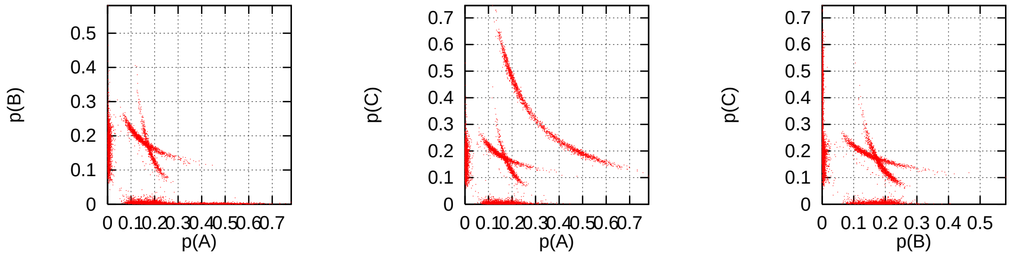
Vallitsevan muodon aineistoilla voidaan havaita tarkkuuden samankaltainen vaikutus. Heikon tarkkuuden aineistolla myös suuret alleelifrekvenssit ovat todennäköisiä, jos toisen alleelin frekvenssi on samanaikaisesti pieni. Tarkkuuden kasvaessa erottuu pisteen $(0.05, 0.05)$ läheltä kulkeva käyrä, jonka läheisyydessä posterioritodennäköisyydet ovat suuria. Muistutetaan, että generoivassa mallissa $M_{D,2}$ molempien alleelifrekvenssien arvot olivat 0.051. Epätäydellisen muodon aineistoilla ei tarkkuudella näytä olevan suurta merkitystä.

Lisäkokeet vallitsevan periytymisen aineistoilla tuottivat kuvan 7.7 alleelifrekvenssijakaumat. Hyvän tarkkuuden aineistolla konvergointi on selvästi puutteellista ja otos keskittyy arvoihin, jotka ovat kaukana peruskokeissa löydetystä harjanteesta. Tarkemmin tutkimalla nähdään kuitenkin, että otos sijaitsee peruskokeissa havaittujen käyrien jatkeella ja että otoksen mallit sopivat aineistoon hyvin. Kohtuullisen tarkkuuden aineistolla alleelifrekvenssien jakauma on oleellisesti sama kuin peruskokeissa saatu tulos. Heikon tarkkuuden aineistolla lisäkokeilla tuotettu otos ei tavoita alleelifrekvenssien riippuvuutta.



Kuva 7.7: Kahden lokuksen epämonotonisten mallien alleelifrekvenssien yhteisjakaumat hyvän, kohtuullisen ja heikon tarkkuuden vallitsevan muodon aineistoilla.

Kuva 7.8 esittää alleelifrekvenssien pareittaisia yhteisjakaumia kolmen lokuksen malleissa, kun havaintoaineistona on $D'_{R,2}$. Symmetriatarkastelu paljastaa konvergoinnin puutteelliseksi. Nähdään kuitenkin samankaltainen ilmiö kuin yhden lokuksen aineiston mallintamisessa kahden lokuksen malleilla (kuva 7.2): kun kahden lokuksen alleelifrekvenssit ovat toisistaan riippuvia, kolmannen lokuksen alleelifrekvenssi on tyypillisesti hyvin pieni.



Kuva 7.8: Alleelifrekvenssien yhteisjakaumia kolmen lokuksen monotonisissa malleissa kohtuullisen tarkkuuden peittyvällä havaintoaineistolla $D'_{R,2}$.

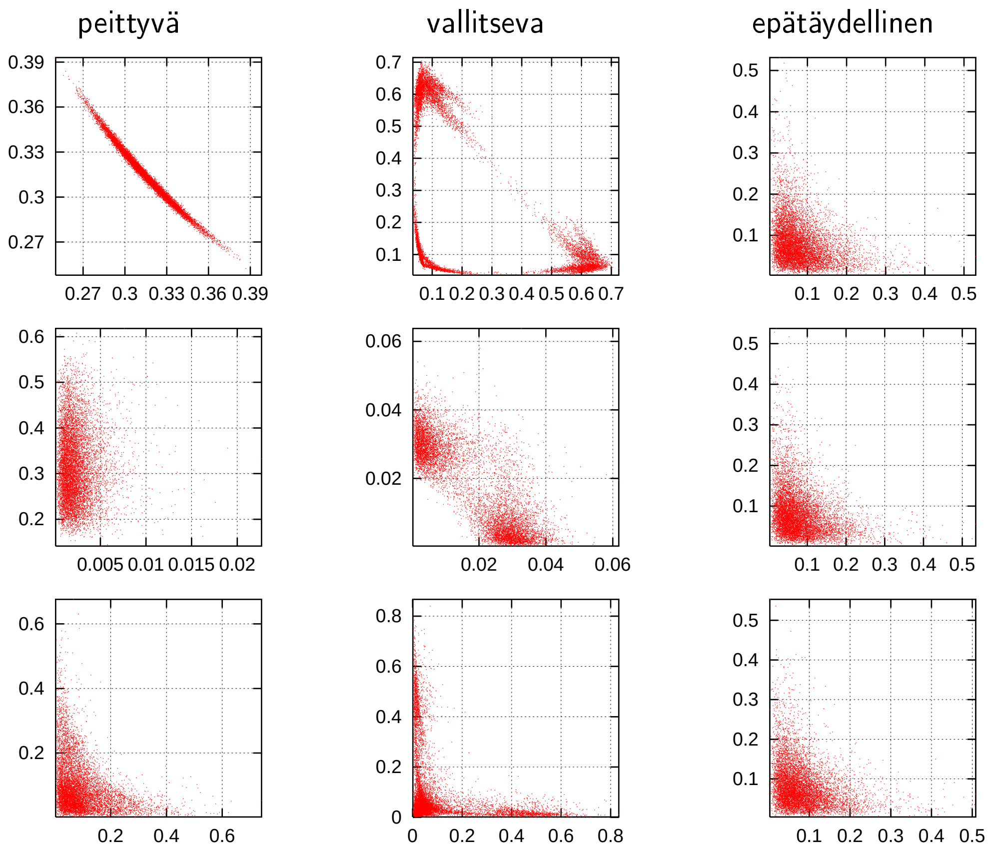
Kolmen lokuksen aineistot

Kuva 7.9 esittää kolmen lokuksen mallien alleelifrekvessien $p(A)$ ja $p(B)$ jakautumista peittyvän, vallitsevan ja epätäydellisen muodon kolmen lokuksen aineistoilla.

Peittyvän muodon generoivassa mallissa $M_{R,3}$ kunkin alleelifrekvenssin arvo oli 0.46. Kuvien nojalla generoivan mallin läheisyydessä todennäköisyysmassaa on vähän; otosten yksikään näyte ei ole lähellä generoivaa mallia. Vaikuttaakin siltä, että otoksen malleissa on oleellisesti kaksi vaikuttavaa lokusta kolmen sijaan.Tämä nähdään eriyisesti kohtuullisella tarkkuudella saadusta epäsymmetrisestä jakaumasta, jossa ensimmäinen alleelifrekvenssi on tyypillisesti hyvin pieni.

Vallitsevan muodon aineistoilla otokset alleelifrekvenssien posteriorijakaumasta ovat symmetrisiä kaikilla tarkkuuksilla. Tässäkin tapauksessa otosten mallien alleelifrekvenssit ovat kaukana generoivan mallin $M_{D,3}$ alleelifrekvenssien arvoista 0.11 (kukin). Poikkeuksen on heikon tarkkuuden aineisto, jolla osa näytteistä sijoittuu alleelifrekvessien osalta generoivan mallin läheisyyteen.

Epätäydellisen muodon aineistoilla alleelifrekvenssit ovat jakautuneet symmetrisesti. Edellisiin verrattavia kiinnostavia rakenteita ei havaita ja tarkkuudella



Kuva 7.9: Ensimmäisen ja toisen lokuksen alleelifrekvenssien yhteisjakauma jakauma kolmen lokuksen malleissa kolmen lokuksen eri aineistoilla. Vaaka-akselilla on ensimmäisen lokuksen alleelin A frekvenssi ja pystyakselilla toisen lokuksen alleelin B frekvenssi. Kolmannen lokuksen alleelifrekvenssiä ei tarkastella. Ylimpänä ovat jakaumat hyvän tarkkuuden aineistoilla, keskellä kohtuullisen tarkkuuden aineistoilla ja alimpana heikon tarkkuuden aineistoilla.

ei näytä olevan suurta vaikutusta. Generoivassa mallissa alleelifrekvenssit olivat 0.400, 0.075 ja 0.020, joita lähellä olevat arvoyhdistelmät eivät ole otoksessa erikoisen hyvin edustettuna. Sitä vastoin otos vaikuttaa painottuvan malleihin, jossa kaikki alleelifrekvenssit ovat pieniä (alle 0.15).

7.3.2 Mallien sopivuus aineistoihin

Tarkastellaan seuraavaksi geneettisten mallien sopivuutta keinotekoisiin havaintoaineistoihin. Tässä esitellään ainoastaan peittyvää muotoa koskevia tuloksia. Vastaavia tuloksia vallitsevalle ja epätäydelliselle muodolle esitetään liitteessä B.2.

Tarkastelu pohjautuu havaintoaineistojen uskottavuuden jakaumien vertailuun yhden kahden ja kolmen lokuksen malliluokkien kesken. Jakaumat esitetään havaintoaineistojen logaritmisen uskottavuuden histogrammeina. Malliluokkien sopivuutta aineistohin mitataan vertaamalla uskottavuutta teoreettiseen ylärajaan, joka saadaan laskemalla uskottavuus havaintoaineiston mukaisilla sukulaisriskeillä. Keinotekoisten aineistojen tapauksessa yläraja on siten likipitäen sama kuin aineiston uskottavuus generoivassa mallissa Uskottavuuksien ylärajat esitetään liitteessä A.2.

Yhden lokuksen aineistot

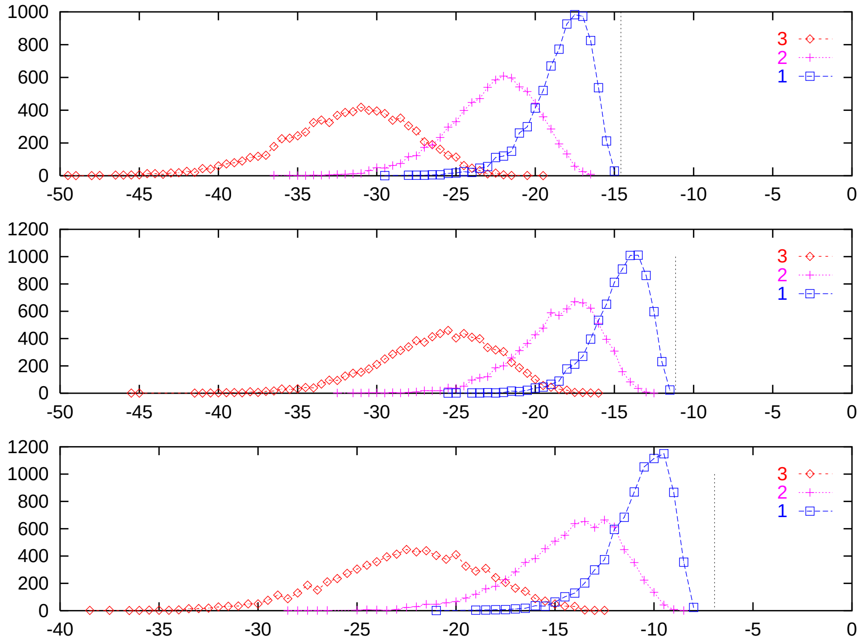
Kuva 7.10 esittää uskottavuuksien jakaumia yhden lokuksen peittyvillä havaintoaineistoilla. Nähdään, että yhden lokuksen malleilla aineistojen uskottavuus on keskimäärin suurinta ja kolmen lokuksen malleilla pienintä. Aineistojen tarkkuudella ei ole vaikutusta uskottavuusjakaumien suhteeseen. Parhaimmillaan otosten mallit saavuttavat uskottavuuden teoreettisen ylärajan.

Kahden lokuksen aineistot

Malliluokkien erot tulevat näkyviin kahden lokuksen aineistoilla. Kuva 7.11 osoittaa aineiston tarkkuuden vaikutuksen peittyvän muodon aineistoilla: Hyvän tarkkuuden aineistolla kahden lokuksen mallit sopivat aineistoon paremmin kuin kolmen lokuksen mallit ja selvästi paremmin kuin yhden lokuksen mallit. Heikon tarkkuuden aineistolla kuitenkin yhden lokuksen mallit ovat keskimäärin sopivimpia.

Kolmen lokuksen aineistot

Kuvan 7.12 histogrammit kertovat yhden, kahden ja kolmen lokuksen mallien sopivuudesta peittyvän muodon kolmen lokuksen havaintoaineistoihin. Heikon tarkkuuden aineistoissa yhden lokuksen mallit sopivat aineistoon karkeasti ot-



$D_{R,1}$

$D'_{R,1}$

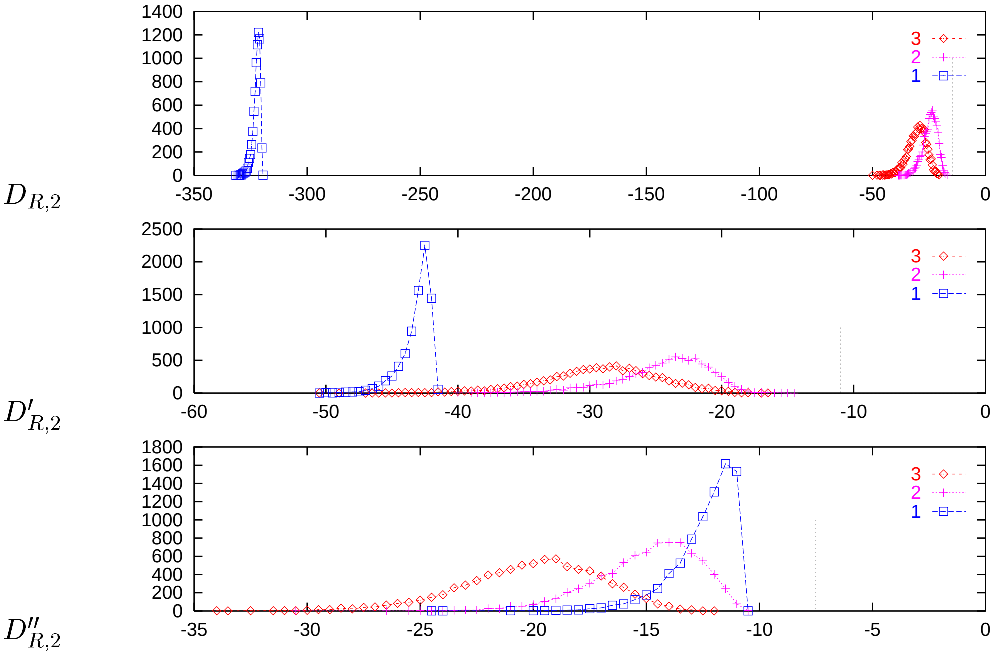
$D''_{R,1}$

Kuva 7.10: Histogrammit aineistojen $D_{R,1}$, $D'_{R,1}$ ja $D''_{R,1}$ uskottavuuden yhden, kahden ja kolmen lokuksen monotonisissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.

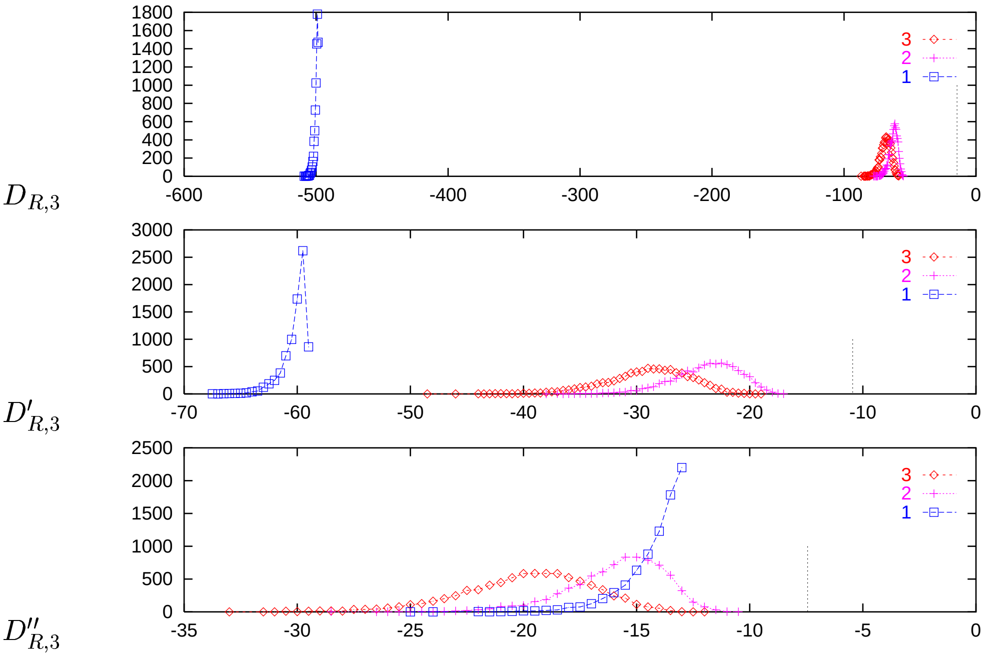
taen yhtä hyvin kuin kahden tai kolmen lokuksen mallit. Aineiston tarkkuuden kasvaessa yhden lokuksen mallien sopivuus heikentyy suhteessa useamman lokuksen malleihin. Samanlaista eroa ei nähdä kahden ja kolmen lokuksen mallien välillä, vaan hyvän tarkkuuden aineistoissakin kahden ja kolmen lokuksen mallit ovat likimäärin yhtä sopivia. Merkille pantavaa on, että tarkkojen aineistojen uskottavuus sopivimmillakin malleilla jää kauas ylärajasta. Heikon tarkkuuden aineistojen uskottavuus on kahden ja kolmen lokuksen sopivimmilla malleilla lähellä ylärajaa, mutta yhden lokuksen malleilla uskottavuudet ovat heikompia.

7.3.3 Diabetes ja skitsofrenia

Todellisilla havaintoaineistoilla *Diabetes* ja *Skitsofrenia* suoritettiin peruskokeet monotonisilla malleilla samaan tapaa kuin keinotekoisilla aineistoilla. Muistutetaan, että havaintojen tarkkuutta ei näissä tapauksissa vaihdeltu.



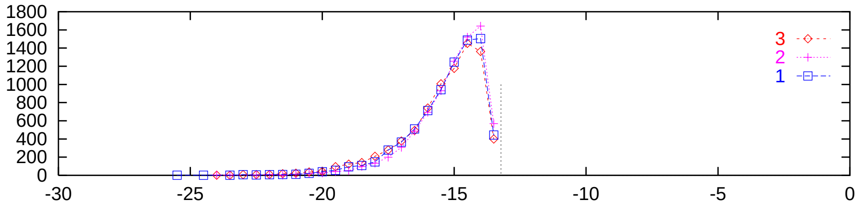
Kuva 7.11: Histogrammit aineistojen $D_{R,2}$, $D'_{R,2}$ ja $D''_{R,2}$ uskottavuuksisista yhden, kahden ja kolmen lokuksen monotonisissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.



Kuva 7.12: Histogrammit aineistojen $D_{R,3}$, $D'_{R,3}$ ja $D''_{R,3}$ uskottavuuksista yhden, kahden ja kolmen lokuksen monotonisissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.

Diabetes

Kuvassa 7.13 on histogrammit havaintoaineiston uskottavuuksien jakaumalle yhden, kahden ja kolmen lokuksen malleissa. Nähdään, että kunkin malliluokan malleissa päästään parhaimmillaan lähelle uskottavuuden ylärajaa. Jakaumien erot eri malliluokilla ovat pieniä. Näyttää siltä, että havaintojen selittämiseksi riittää yksi vaikuttava lokus.

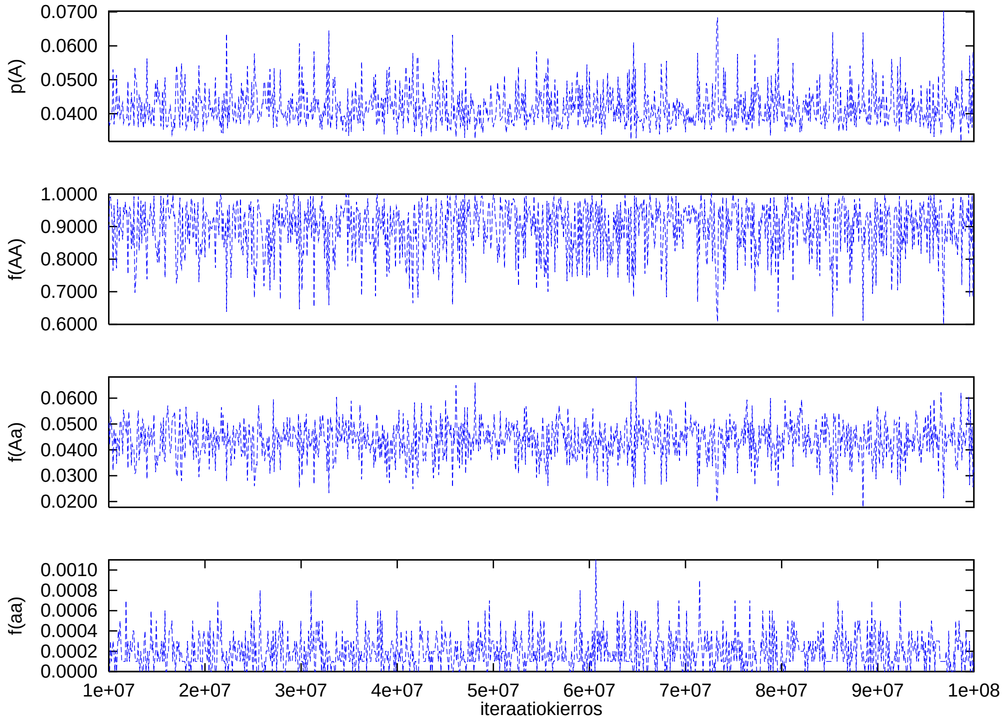


Kuva 7.13: Histogrammit diabetesaineiston uskottavuudesta yhden, kahden ja kolmen lokuksen malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.

Yhden lokuksen mallien parametrien arvoja ja keskinäistä riippuvuutta voidaan arvioida Kuvasta 7.14. Arvojen vaihtelu simulaation kuluessa viittaa kelvolliseen konvergenssiin. Parametrien kesken havaitaan luontevia riippuvuuksia: kun alleelifrekvenssit ovat suuria (lähellä arvoa 0.05), genotyyppien AA ja Aa ilmenemisyleisyydet ovat vastaavasti pieniä. Aiemmin samankaltainen riippuvuus todettiin tarkasteltaessa epätäydellisen muodon yhden lokuksen havaintoaineistoa (kuva 7.4). Ilmenemisyleisyyksiensä puolesta mallit vaikuttavat jokseenkin peittyviltä. Laskemalla kuitenkin nähdään, että mallien mukaan suurimmalla osalla sairaista on genotyyppi Aa, mikä viittaa vallitsevaan periytymiseen. Toisin sanoen sairas alleeli periytyy tyypillisesti vain toiselta vanhemmalta.

Alleelifrekvenssien jakaumat kahden ja kolmen lokuksen malleilla (kuva 7.15) kertovat lokusten vähäisestä riippuvuudesta. Toisen alleelifrekvenssin ollessa lähellä arvoa 0.04–0.10 toisen alleelin frekvenssi voi vaihdella painottuen kuitenkin pieniin arvoihin.

Tulokset *Diabetes*-aineistolla osoittavat, että havaintoihin erinomaisesti sopivia yhden lokuksen malleja löytyy esimerkiksi mallin $\langle 0.04, 0.9, 0.045, 0.0002\rangle$

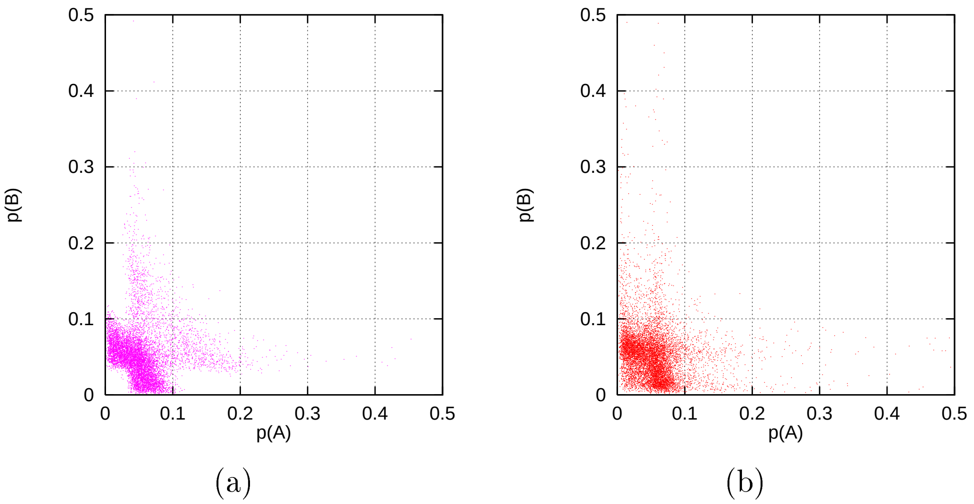


Kuva 7.14: Yhden lokuksen mallien parametrien arvojen muuttuminen simulaation aikana, kun aineistona on *Diabetes*. Kuvaa varten 9 000 mallista on poimittu tasavälein 1 000 mallia.

läheisyydestä. Tulokseen on suhtauduttava varauksella, sillä se ei sellaisenaan ole sopusoinnussa muilla tutkimusmenetelmillä hankitun tietämyksen kanssa. Tiedetään esimerkiksi, että ns. HLA-alueella on useita vahvasti kytkeytyneitä lokuksia, joiden alleelien eri yhdistelmille on mitattu keskenään erisuuria vaikutuksia sairauden ilmenemiseen [Onk99b, Ric90]. Kysymys onkin, voidaanko – ja onko järkevää – luokitella tällaiset genotyypit kahteen luokkaan (A ja a). Toinen ongelma on se, että tässä käytetyt sukulaisriskihavainnot eivät huomioi yksilöiden sukupuolen merkitystä. Vaikka yhden lokuksen mallit sopivat keskiarvoistettuun aineistoon erinomaisesti, ne eivät sopisi hyvin aineistoon, jossa riskit vaihtelevat sukupuoleen sidotusti.

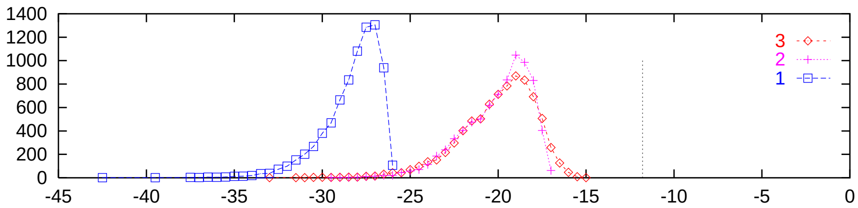
Skitsofrenia

Kuvassa 7.16 on histogrammit havaintoaineiston uskottavuuksien jakaumalle yhden, kahden ja kolmen lokuksen malliluokissa. Nähdään, että kunkin malli-



Kuva 7.15: Alleelifrekvenssien yhteisjakaumat (a) kahden ja (b) kolmen lokuksen malleilla aineistolla *Diabetes*. Havaitun symmetrian vuoksi kolmen lokuksen malleilla tarkastellaan ainoastaan ensimmäisen ja toisen lokuksen alleelifrekvenssejä.

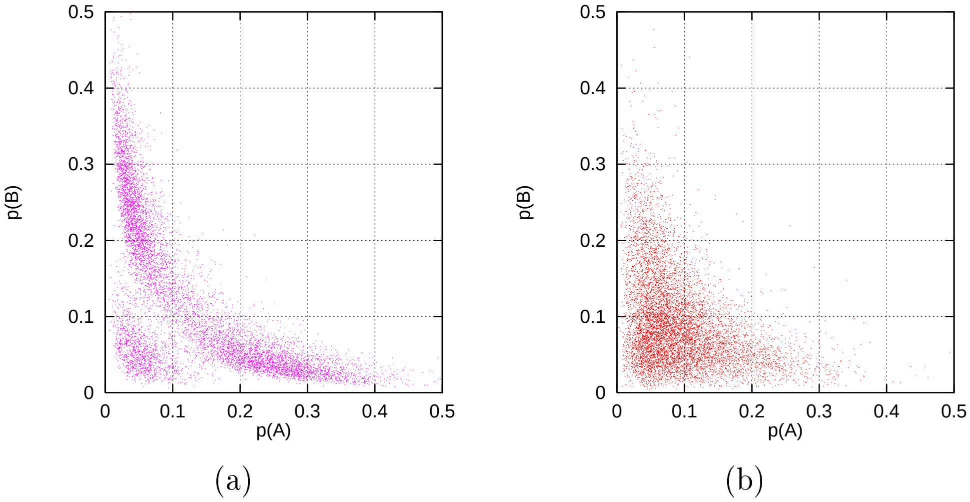
luokan malleissa jäädään kauas uskottavuuden ylärajasta. Jakaumien ero kahden ja kolmen lokuksen malliluokissa on pieni. Yhden lokuksen malleissa uskottavuus on keskimäärin selvästi heikointa.



Kuva 7.16: Histogrammit skitsofrenia-aineiston uskottavuudesta yhden, kahden ja kolmen lokuksen malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.

Alleelifrekvenssien jakaumat kahden ja kolmen lokuksen malliluokassa kertovat lokusten riippuvuudesta (kuva 7.17). Havaitaan, että havaintoaineistoon sopivat yhtä hyvin sellaiset mallit, joiden alleelifrekvenssit ovat likipitäen yhtäsuuret, kuin myös sellaiset mallit, joille toinen alleelifrekvenssi on suuri ja toinen pieni.

Skitsofrenia-aineistolla saatujen tulosten mukaan kahden lokuksen mallit sopivat aineistoon selvästi yhden lokuksen malleja paremmin. Kolmas lokus ei



Kuva 7.17: Alleelifrekvenssien yhteisjakaumat (a) kahden ja (b) kolmen lokuksen malleilla aineistolla *skitsofrenia*. Havaitun symmetrian vuoksi kolmen lokuksen malleilla tarkastellaan ainoastaan ensimmäisen ja toisen lokuksen alleelifrekvenssejä.

näytä auttavan havaintojen selittämisessä. Tämä tukee Rischin [Ris90] väitettä, jonka mukaan yksi lokus ei riitä selittämään sukulaisriskihavaintoja. Sen sijaan Campbell ja Elston [CaE71] ehdottivat yhden lokuksen mallia kahden lokuksen mallin sijaan. Kuten Risch huomautti, skitsofrenian erikoisuus on, että havaittu lapsiriski on selvästi korkeampi kuin sisarusriski (ks. taulukko 7.2). Tarkasteltavan malliperheen mallit eivät kykene selittämään tällaista eroa. Itseasiassa voidaan osoittaa, että kytkemättömän periytymisen malleissa sisarusriski on aina vähintään yhtä suuri kuin lapsiriski [KoM00]. Tämä on selvästi yksi syy siihen, miksi aineiston uskottavuus kahden ja kolmen lokuksen malliluokissa jää kauas uskottavuuden teoreettisesta ylärajasta.

7.4 Johtopäätöksiä

Koetulosten valossa simulaatioiden konvergoituminen on epävakaata. Konvergoitumiseen vaikuttavat havaintoaineiston piirteet, lokusten määrä tarkasteltavissa malleissa ja muut simulaatioparametrien arvot sekä sattuma.

Keskeisin havaintoaineiston ominaisuus näyttää olevan ”taudin periytyvyys” eli siihen sopivien mallien ilmenemisyleisyyksien täydellisyys (voidaan mitata identtisen kaksosen riskin suuruutena). Vaikuttaa siltä, että vallitsevan muodon aineistot ovat simulaation konvergoinnin kannalta ongelmallisimpia ja epätäydellisen muodon aineistot helpoimpia. Tätä voidaan yrittää selittää sillä, että täydellisen muodon aineistoilla posteriorijakaumat ovat "terävämmin monihuippuisia" kuin epätäydellisillä aineistoilla. Toinen konvergoinnin kannalta merkittävä havaintoaineiston ominaisuus on tarkkuus. Havaittiin odotetusti, että tarkkuuden kasvaessa simulaation konvergoituminen vaikeutui.

Malliluokan vaikutus simulaation konvergoitumiseen ei yllättänyt: mitä enemmän malliluokan malleilla on parametreja, sen heikompaa on konvergoituminen. Sen sijaan oli yllättävää, että simulaatio jumiutui yhden lokuksen monotonisilla malleilla, jopa yhden lokuksen heikon tarkkuuden vallitsevalla havaintoaineistolla. Lisäkokeet erilaisia simulaatioparametrien asetuksia käyttäen tuottivat kuitenkin hyvän tuloksen kaikilla aineiston tarkkuuksilla (vain kohtuullinen tarkkuus raportoitiin). Luultavimmin simulaatioiden sekoittuminen oli parametrien ryvästyksen ansiota. Lopullisen vastauksen saaminen ja optimaalisten simulaatioparametrien asetusten löytäminen vaatii kuitenkin lisätutkimuksia.

Alleelifrekvenssien pareittaisten jakaumien tarkastelu osoitti, että posteriorijakaumat ovat tyypillisesti monihuippuisia. Edelleen nähtiin, että huiput ovat oikeastaan moniulotteisia harjanteita. Tämä on ymmärrettävää, koska sukulaisriskihavainnot eivät teoriassakaan kykene määräämään yksikäsitteisiä suurimman uskottavuuden parametreja. Jo yhden lokuksen mallit ovat yliparametrisoituja, kun aineiston taustalla oleva sairaus on heikosti periytyvä. Luvussa 4 käsitelty neljän parametrin epäidentifioituvuus tulee esille siis myös käytännössä. Kun aineiston selittäminen edellyttää täydellisiä ilmenemisyleisyyksiä, ei yliparametrisoinnin ongelmaa yhden lokuksen malliluokassa kuitenkaan havaita.

Mallien sopivuuden tarkastelu havaintoaineistojen uskottavuuksien jakaumiin nojautuen paljasti odotetun ilmiön: kun aineiston tarkkuus paranee, liian vähäiseen lokusten määrään tukeutuvilla malleilla aineistojen uskottavuus on pieni. Nähtiin, että kahden lokuksen peittyvillä ja vallitsevilla aineistoilla yhden lokuksen mallien uskottavuus jäi selvästi heikommaksi. Kolmen lokuksen aineistoilla ei kuitenkaan voitu havaita vastaavaa ilmiötä: uskottavuuksien jakauman eroa kahden ja kolmen lokuksen malliluokissa. Tässä on myös huomioitava, että otosten sopivimmatkin mallit jäivät kauas teoreettisista uskottavuuden ylärajoista, jotka kuitenkin likimäärin saavutetaan generoivissa malleissa. Selitys voi olla simulaation puutteellinen konvergointi kolmen lokuksen malleilla ja simulaation jumiutuminen malleihin, joissa lopulta on vain kaksi vaikuttavaa lokusta. Toisaalta on myös mahdollista, että simulaatioiden tuottamat otokset todella ovat edustavia mutta silti edellä mainitun kaltaisia. Epätäydellisen muodon aineistoilla malliluokille ei saatu eroa uskottavuuksien jakaumia tarkastelemalla. Jopa kolmen lokuksen mallista generoidulla kohtuullisen tarkkuuden aineistoilla jakaumien muoto eri malliluokilla oli sama ja uskottavuuden yläraja saavutettiin.

Kyseiseen uskottavuuksien jakaumien tarkastelutapaan on suhtauduttava varauksella sen heuristisuuden ja epätäsmällisyyden vuoksi. Johtopäätös on silti ilmeinen: tautiin vaikuttavien lokusten lukumäärän arvioiminen lienee käytännössä vaikeaa, jos havainnot rajoittuvat sukulaisriskeihin. Näin on erityisesti silloin, jos sairaus on heikosti periytyvä tai aineistoa on vähän. Yhden lokuksen mallit ovat varsin joustavia, jolloin ne sopivat tyypillisiin aineistoihin, vaikka aineistojen taustalla olisi kaksi tai kolme lokuksta. Vielä paremmin sopivat kahden lokuksen mallit aineistoihin, joiden taustalla kolme lokusta. Luotettavat päätelmät ovat mahdollisia vain, jos aineisto on tarkka ja identtisen kaksosen riski on korkea. Vastaavan johtopäätöksen ovat tehneet mm. Smith [Smi71] ja Dizier et al. [DBC93] perustuen suurimman uskottavuuden menetelmään, samankaltaisiin malleihin ja generoituihin havaintoaineistoihin. Lokusten lukumäärä bayesiläinen arviointi antaisi luultavasti saman tuloksen mutta perustellummin.

Luku 8

Yhteenveto ja jatkotutkimus

Perinnöllisten ominaisuuksien mallintaminen on tutkimusalue, joka tarjoaa perinnöllisyystieteellisten ja matemaattisten kysymysten lisäksi tietojenkäsittelytieteellisiä haasteita. Tässä tutkielmassa tarkasteltiin kaksivasteisten sairauksien mallintamista, kun annettuna on yksinkertainen taudin esiintymistä kuvaava havaintoaineisto. Erityisesti perehdyttiin sukulaisriskeihin, jotka kertovat taudin esiintymisestä sairaan yksilön lähisukulaisilla.

Työn alkuosassa luotiin mallintamiselle laaja perusta. Sukulaisriskihavainnot määriteltiin yleisempien sukupuuhavaintojen avulla. Tämä lähestymistapa antoi mahdollisuuden soveltaa Elstonin ja Stewartin esittämää geneeristä todennäköisyysmallia, joka koostuu genotyyppien vaikutukset, frekvenssit ja periytymisen kuvaavista funktioista. Tärkeimpien yksinkertaistavia oletusten mukaisesti mallit jaoteltiin malliperheisiin ja malliluokkiin. Erityisesti keskityttiin malliperheisiin, joissa genotyyppien mielivaltaiset yhteisvaikutukset on mahdollista kuvata. Tutkielmassa myöhemmin käsiteltyjen asioiden kannalta oli merkittävää, että näihin perheisiin kuuluvien mallien parametrien lukumäärä kasvaa eksponentiaalisesti suhteessa lokusten lukumäärään.

Tilastollista päättelyä varten havainnot ja mallit sidottiin uskottavuusfunktion avulla. Cambellin ja Elstonin [CaE71] esittämää suurimman uskottavuuden menetelmää tarkasteltiin kriittisesti. Vaihtoehdoksi tarjottiin bayesiläinen lähestymistapa, joka teoriassa mahdollistaa suoraviivaisen todennäköisyyksiin nojautuvan päättelyn. Päättelyn kannalta oleellista identifioituvuusongelmaa käsiteltiin. Jo yksinkertaistenkin kytkemättömien mallien todettiin olevan yliparametrisoituja mikäli havainnot rajoittuvat sukulaispareihin. Bayesiläinen lähestymistapa kärsii ongelmasta kuitenkin vähemmän kuin suurimman uskottavuuden menetelmä.

Luku 5 omistettiin sukulaisriskien laskennalle. Keskeisin tulos oli, että naiiveja laskentamenetelmiä, jotka suoraan hyödyntävät yleisille malleille sopivia menetelmiä, voidaan huomattavasti tehostaa kytkennällisissä ja kytkemättömissä malleissa. Vaativuustuloksilla on merkitystä käytännössä: jos naiivit menetelmät pakottavat mallintajan tyytymään yhden ja kahden lokuksen mallien tarkasteluun, niin parannukset mahdollistavat kolmen ja neljän lokuksen mallien käsittelyn. Monelta osin luvun 5 tuloksia voidaan kuitenkin pitää teoreettisina. Esimerkiksi osioissa 5.3.1–5.3.2 esitettyjen matriisi- ja varianssihajotelmaan perustuvien menetelmien ero laskennallisessa vaativuudessa ei liene käytännössä merkittävä.

Tutkielman loppuosassa tarkasteltiin bayesiläisen päättelyn toteuttamista. Esiteltiin simulaatio-ohjelma GeME, joka tuottaa otoksen geneettisten mallien posteriorijakaumasta simuloimalla sopivaa Markovin ketjua. Bayesiläisen päättelyn vaatimia moniulotteisia integraaleja voidaan tämän jälkeen arvioida otokseen nojautuvalla Monte Carlo -integroinnilla. Tutkielmassa käsiteltyyn yleiseen mallinnusmenetelmään verrattuna GeME on varsin rajoittunut. Ohjelma hyödyntää ainoastaan havaintoaineistoja, jotka koostuvat taudin yleisyyttä, lapsiriskiä, sisarusriskiä ja identtien kaksosen riskiä kuvaavista luvuista. Mallien osalta GeME rajoittuu kytkemättömiin kaksialleelisiin malleihin. Koetulosten mukaan simulaatio konvergoituu posteriorijakaumaan luotettavasti, jos aineiston tarkkuus on korkeintaan kohtuullinen tai aineiston taustalla oleva malli ei edellytä täydellisiä ilmenemisyleisyyksiä. Toisaalta lokusten lukumäärän ja mallien parametrien suuntaa antava arviointi vaikuttaa mahdolliselta täsmälleen päinvastaisessa tapauksessa. Tämä paradoksaalisuus vaikuttaa väistämättömältä. Otokset mallien posteriorijakaumista vaikuttavat kuitenkin hyödyllisiltä. Ne antava tietoa parametrien arvoista ja keskinäisistä riippuvuuksista ja voivat siten auttaa taudin geneettisten mallien arvioinnissa.

Monia kysymyksiä jää avoimeksi. Konkreettisimmat jatkotutkimuksen aiheet liittyvät olemassa olevan ohjelman ominaisuuksien selvittämiseen. Toisaalta on tarvetta pohtia myös perustavampaa laatua olevia kysymyksiä liittyen havaintojen laatuun ja geneettisten mallien parametrisointiin. Seuraavaksi esitetään joitakin keskeisimpiä jatkotutkimuksen aiheita tarkemmin.

Luvussa 7 esitettyjä kokeita on tarpeen laajentaa. Satunnaisvaikutusten vuoksi kussakin koepisteessä on syytä tehdä useita simulaatioita mahdollisesti satunnaisista aloituspisteistä lähtien. Jos konvergointi on kunnollista simulaatioiden pitäisi antaa likipitäen samat tulokset. Myös simulaatioparametrien eri arvoyhdistelmiä on syytä kokeilla. Kiinnostavaa olisi tietää monotonisuusvaatimuksen ja epätasaisen priorin merkitys; tässä työssä tehdyt kokeet eivät olleet riittäviä tämän selvittämiseen. Jatkokehityksen kannalta erityisen tärkeää olisi selvittää ehdotusjakaumaa säätelevän parametrin ja parametrien ryvästyksen merkitys simulaation konvergoinnille. Jotta havaintoaineiston vaikutus posteriorijakaumiin ja konvergointiin selviäisi, keinotekoisia havaintoaineistoja on tarpeen generoida lukuisista epätäydellisistä malleista. Koska yksi simulaatio vie tyypillisesti aikaa useita tunteja, edellä mainittujen laajennusten täydellisten yhdistelmien kokeileminen ei ole mahdollista. Sopivien koejärjestelyjen laatiminen ja toteuttaminen onkin huomattava haaste.

Lokusten lukumäärän posterioritodennäköisyydelle annettiin lauseke esimerkissä 4.1. Myöhemmin todettiin, että GeME-ohjelman tulosteet eivät mahdollista lausekkeen suoraviivaista arvioimista, koska ohjelman tuottama otos on ehdollinen lokusten lukumäärälle. Vaikuttaa kuitenkin siltä, että lokusten lukumäärää voidaan bayesiläisittäin arvioida tällaisiin ehdollisiin otoksiin nojautuen. Keskeinen ongelma on, miten lasketaan havaintojen todennäköisyys $p(D \mid \ell)$, kun ainoastaan lokusten lukumäärä eli malliluokka on annettu. Vastaavalle yleiselle ongelmalle on esitetty mielenkiintoisia tärkeysotantaan (engl. importance sampling) perustuvia ratkaisuja [Raf95]. Yksi tällainen keino hyödyntää otosta mallien posteriorijakaumasta, jolloin tarkasteltava todennäköisyys saadaan havaintojen uskottavuuden harmonisena keskiarvona yli mallien posteriorijakaumaa edustavan otoksen. Tämän menetelmän soveltaminen lokusten lukumäärän arvioimiseen vaikuttaa lupaavalta. Se ei edellytä muutoksia olemassa olevaan ohjelmaan, vaan ohjelman tuloste on sellaisenaan hyödynnettävissä.

Elegantimpi ratkaisu olisi simuloida Markovin ketjua samanaikaisesti usealle malliluokalle yhden malliluokan sijaan. Toisin sanoen ketjussa sallittaisiin lokusten lukumäärän vaihtelu. Yksinkertainen komponenteittainen Metropolisin ja Hastingsin algoritmi ei tällöin tule kyseeseen. Sopivan kehyksen ulottuvuudeltaan vaihtelevien parametriavaruuksien hallintaan tarjoaa Greenin [Gre95] esittämä RJMCMC-menetelmä (engl. reversible jump Markov chain Monte Carlo). Esimerkiksi Sillanpää ja Arjas [SiA98] ovat hyödyntäneet RJMCMC-menetelmää menestyksellisesti lokusten lukumäärän bayesiläiseen arviointiin. Heidän käyttämänsä havaintoaineistot ja mallit poikkesivat kuitenkin oleellisesti tässä työssä tarkastelluista, eikä menetelmän soveltuvuudesta ole siten varmuutta. Vaikeimpia ongelmia RJMCMC-menetelmää sovellettaessa lienee ulottuvuuksien välisen hypyn toteuttaminen. Minkälaisia kolmen lokuksen malleja ehdotetaan, jos simulaatio on jotakin kahden lokuksen mallia vastaavassa tilassa?

Ohjelmaa GeME voidaan helposti yleistää niin, että se kykenee käsittelemään tavallisimpia sukulaistyyppejä. Havaintojen laajentaminen sukulaisriskien ulkopuolelle voi myös olla hyödyllistä. Laajemmat ydinperhehavainnot, kuten luvussa 4.3.2 esitettiin, voivat merkittävästi parantaa geneettisten mallien identifioitumista. Samalla ne näyttävät säilyttävän sukulaisriskien hyödyllisimmät ominaisuudet: havaintoaineistojen kohtuullisen helpon saannin ja havaintojen uskottavuuden riittävän tehokkaan laskennan.

Todelliset havaintoaineistot ja niillä saadut tulokset pakottavat miettimään, onko työssä tarkasteltujen, erityisesti kytkemättömän periytymisen, mallien kuvausvoima sopiva. Voisi olla järkevää mallintaa myös ympäristötekijöitä ja sukupuoleen sidottua periytymistä. Toisaalta mallit ovat yliparametrisoituja jo ilman laajennuksia. Toinen suunta onkin parametrien vähentäminen tai muunlainen rajoittaminen. Yksi mahdollisuus voisi olla varianssikomponenttien suora arvioiminen, jolloin parametrien lukumäärä olisi neliöllien suhteessa lokusten lukumäärään. Vaihtoehtoisesti voitaisiin genotyyppien ilmenemisyleisyyksille asettaa jonkinlainen riippuvuusrakenne, kuitenkin yleisempi kuin osiossa 3.3 esitetyt lokuskohtaisiin vaikutuksiin perustuvat riippuvuudet. Mallien yksinkertaistaminen vaikuttaa luonnollisesti myös sukulaisriskien laskentaan.

Lähteet

[AjM00] Aji, S. M., McEliece, R. J. (2000), The generalized distributive law. *IEEE Transactions of Information Theory*, 46: 325–343.

[BGS98] Becker, A., Geiger, D., Schäffer, A. A. (1998), Automatic selection of loop breakers for genetic linkage analysis. *Human Heredity*, 48: 49–60.

[BoT73] Box, G. E. P., Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading (MA).

[Bul80] Bulmer, M. G. (1980), *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

[BDS98] Bureau, A., Dudoit, S., Speed, T. P. (1998), Week 8, Tuesday lecture: Calculation of probabilities over pedigrees (2). Luentosarjassa Stat 260: Statistic in Genetics, toim. Speed, T., saatavissa osoitteessa http://stat-www.berkeley.edu/classes/s260/. [8.8.2000]

[CaE71] Campbell, M. A., Elston, R. C. (1971), Relatives of Probands: Models for Preliminary Genetic Analysis. *Annals of Human Genetics*, 35: 225–236.

[CTS78] Cannings, C, Thompson, E. A., Skolnick, M. H. (1978), Probability functions on complex pedigrees. *Advances in Applied Probability*, 10: 26–61.

[CoW90] Coppersmith, D., Winograd, S. (1990), Matrix multiplication via arithmetic programming. *Journal of Symbolic Computing*, 9: 251–280

[DBC93] Dizier, M. H., Bonaiti-Pellie, C., Clerget-Darpoux, F. (1993), Conclusions of segregation analysis for family data generated under two-locus models. *American Journal of Human Genetics*, 53: 1338–1346.

[ElS71] Elston, R. C., Stewart, J. (1971), A general model for the analysis of pedigree data. *Human Heredity*, 21: 523–542.

[Fis18] Fisher, R. A. (1918), The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of Royal Society of Edinburg*, 52: 399 – 433.

[FRS78] Fishman, P. M., Reich, T., Suarez, B., James J. W. (1978), A note on the essential parameters of the two-allele autosomal locus model. *American Journal of Human Genetics*, 30: 283–292.

[GRS96] Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, Lontoo.

[Goo58] Good, I. J. (1958), The interaction algorithm and practical Fourier analysis. *Journal of the Royal Statistical Society, Series B*, 20: 361–372.

[Gor92] Gorardia, T. M., Lange, K., Miller, P. L., Nadkarni, P., M. (1992), Fast computation of genetic likelihoods on human pedigree data. *Human Heredity*, 42: 42–62.

[Gre95] Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82: 711–732.

[Hal19] Haldane, J. B. S. (1919), The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics*, 8: 299–309.

[Har64] Harris, D. L. (1964), Genotypic covariances between inbred relatives. *Genetics*, 50: 1319–1348.

[Has70] Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57: 97–109.

[Jam71] James, J. W. (1971), Frequency in relatives for an all-or-none trait. *Annals of Human Genetics*, 35: 47–49.

[Kem57] Kempthorne, O. (1957), *An Introduction to Genetic Statistics*. John Wiley & Sons, New York.

[KoM00] Koivisto, M., Mannila, H. (2000), Sibling risk and offspring risk. *Human Heredity*, julkaistavana.

[Lai98] Laitinen, T., Räsänen, M., Kaprio, J., Koskenvuo, M., Laitinen, L. A. (1998), Importance of Genetic Factors in Adolescent Asthma: A Population-based Twin-family Study. *American Journal of Respiratory and Critical Care Medicine*, 157: 1073–1078.

[Lan97] Lange, K. (1997), *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York.

[LaB83] Lange, K., Boehnke, M. (1983), Extensions to pedigree analysis. *Human Heredity*, 33: 291–301.

[LiS54] Li, C. C., Sacks, L. (1954), The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics*, 10: 347–360.

[MaM92] Marcus, M., Minc, H. (1992), *A Survey of Matrix Theory and Matrix Inequalities*. Dover, New York.

[Met53] Metropolis et al. (1953), Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21: 1087–1091.

[MyT98] Myllymäki, P., Tirri, H. (1998), *Bayes-verkkojen mahdollisuudet*, Teknologiakatsaus 58/98, Teknologian kehittämiskeskus (Tekes), Helsinki.

[OlB94] Oliver, J., Baxter, R. (1994), MML and Bayesianism: Similarities and differences. Tekninen raportti 206, Department of Computer Science, Monash University.

[Onk99a] Onkamo, P. (1999), sähköpostiviesti, 18.11.1999.

[Onk99b] Onkamo, P. (1999), sähköpostiviesti, 13.12.1999.

[Pea88] Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo (CA), 441–452.

[Raf95] Raftery, A. E. (1995), Hypothesis testing and model selection. Teoksessa [GRS96], 163–187.

[Ric90] Rich, S. S. (1990), Mapping genes in diabetes: genetic epidemiological perspective. *Diabetes*, 39: 1315–1319.

[Ris90] Risch, N. (1990), Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics*, 46: 222–228.

[Rob95] Roberts, G. O. (1995), Markov chain concepts related to sampling algorithms. Teoksessa [GRS96], 45–57.

[Sea66] Searle, S. R. (1966), *Matrix Algebra for the Biological Sciences*, Wiley, New York.

[Sha98] Sham, P. (1998), *Statistics in Human Genetics*, Arnold, Lontoo.

[SiA98] Sillanpää, M. J., Arjas, E. (1998), Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, 148: 1373–1388.

[Smi71] Smith, C. (1971), Discriminating between different modes of inheritance in genetic disease. *Clinical Genetics*, 2: 303–314.

[Tie96] Tierney, L. (1996), Introduction to general state-space Markov chain theory. Teoksessa [GRS96], 59–74.

[Tir93] Tirri, R., Lehtonen, J., Lemmentyinen, R., Pihakaski, S., Portin, P. (1993), *Biologian sanakirja*, Otava, Keuruu.

[UVV95] Ulmanen, I., Valste, J., Viitanen, P. (1995), *Biologia: Geeni*, WSOY, Porvoo.

[Val95] Valste, J., Airamo, S., Holopainen, M., Koivisto, I., Suominen, T., Viitanen, P. (1995), *Biologia: Toimiva eliömaailma, perinnöllisyys ja evoluutio*, WSOY, Porvoo.

[ViL00] Vitányi, P. M. B., Li, M. (2000), Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions of Information Theory*, 46: 446–464.

[WKW91] Whittemore, A. S., Keller, J. B., Ward, M. J. (1991), Family data determine all parameters in Mendelian incomplete penetrance models. *Annals of Human Genetics*, 55: 175–177.

[Yat37] Yates, F. (1937), The design and analysis of factorial experiments. *Technical Communication 35*, Imperial Bureau of Soil Science, Harpenden.

Liite A

Koejärjestelyt

Tässä liitteessä selvitetään muutamia simulaatiokokeiden järjestelyyn liittyviä valintoja. Ensimmäisessä osiossa kuvataan keinotekoisten aineistojen tuottamisessa käytetyt generoivat mallit. Toisessa osiossa esitetään komponenteittaisen Metropolisin ja Hastingsin algoritmin käyttämät mallikomponenttien eli parametrien ryppäiden jonot. Kolmannessa osiossa luetellaan uskottavuuksien ylärajat kokeissa käytetyille aineistoille.

A.1 Generoivat mallit

Simulaatiokokeissa käytetyt generoivat mallit kuvataan alla seuraavasti. Jokaiseen genotyyppiin liitetään kolme lukua, joista kaksi on merkitty sulkeisiin. Ensimmäinen luku on genotyypin ilmenemisyleisyys eli varsinainen mallin parametri. Toinen luku ilmoittaa, kuinka suuri on genotyypin kantajien osuus populaatiosta. Kolmas luku kertoo, kuinka suuri osa sairaista on genotyypin kantajia. Taulukon alle kirjoitetaan ensin alleelin A frekvenssi ja sitten alleelin B frekvenssi jne. Tämän jälkeen sulkeissa ilmoitetaan, mikä on mallin mukaan taudin yleisyys, lapsiriski, sisarusriski ja identtisen kaksosen riski.

Peittyvä yhden lokuksen malli $M_{R,1}$:

```
    AA                      Aa                      aa
 1.0000 (0.0100 1.0000)  0.0000 (0.1800 0.0000)  0.0000 (0.8100 0.0000)

0.1000 (0.01000 0.10000 0.30250 1.00000)
```

Peittyvä kahden lokuksen malli $M_{R,2}$:

```
    AA                      Aa                      aa
 BB 1.0000 (0.0100 1.0000)  0.0000 (0.0432 0.0000)  0.0000 (0.0468 0.0000)
 Bb 0.0000 (0.0432 0.0000)  0.0000 (0.1870 0.0000)  0.0000 (0.2022 0.0000)
 bb 0.0000 (0.0468 0.0000)  0.0000 (0.2022 0.0000)  0.0000 (0.2186 0.0000)

0.3162 0.3162 (0.01000 0.10000 0.18759 1.00000)
```

Peittyvä kolmen lokuksen malli $M_{R,3}$:

```
 CC AA                      Aa                      aa
 BB 1.0000 (0.0100 1.0000)  0.0000 (0.0231 0.0000)  0.0000 (0.0133 0.0000)
 Bb 0.0000 (0.0231 0.0000)  0.0000 (0.0533 0.0000)  0.0000 (0.0308 0.0000)
 bb 0.0000 (0.0133 0.0000)  0.0000 (0.0308 0.0000)  0.0000 (0.0178 0.0000)

 Cc AA                      Aa                    aa
 BB 0.0000 (0.0231 0.0000)  (0.0000 0.0533 0.0000)  0.0000 (0.0308 0.0000)
 Bb 0.0000 (0.0533 0.0000)  (0.0000 0.1231 0.0000)  0.0000 (0.0710 0.0000)
 bb 0.0000 (0.0308 0.0000)  (0.0000 0.0710 0.0000)  0.0000 (0.0410 0.0000)

 cc AA                      Aa                      aa
 BB 0.0000 (0.0133 0.0000)  0.0000 (0.0308 0.0000)  0.0000 (0.0178 0.0000)
 Bb 0.0000 (0.0308 0.0000)  0.0000 (0.0710 0.0000)  0.0000 (0.0410 0.0000)
 bb 0.0000 (0.0178 0.0000)  0.0000 (0.0410 0.0000)  0.0000 (0.0237 0.0000)

0.4642 0.4642 0.4642 (0.01000 0.10000 0.15394 1.00000)
```

Vallitseva yhden lokuksen malli $M_{D,1}$:

```
    AA                     Aa                     aa
    1.0000 (0.0000 0.0025) 1.0000 (0.0100 0.9976) 0.0000 (0.9900 0.0000)

0.0050 (0.01000 0.50376 0.50438 1.00000)
```

Vallitseva kahden lokuksen malli $M_{D,2}$:

```
    AA                     Aa                     aa
 BB 1.0000 (0.0000 0.0007) 1.0000 (0.0003 0.0256) 0.0000 (0.0024 0.0000)
 Bb 1.0000 (0.0003 0.0256) 1.0000 (0.0095 0.9480) 0.0000 (0.0876 0.0000)
 bb 0.0000 (0.0024 0.0000) 0.0000 (0.0876 0.0000) 0.0000 (0.8100 0.0000)

0.0513 0.0513 (0.01000 0.28961 0.29602 1.00000)
```

Vallitseva kolmen lokuksen malli $M_{D,3}$:

```
 CC AA                     Aa                     aa
 BB 1.0000 (0.0000 0.0002) 1.0000 (0.0000 0.0034) 0.0000 (0.0001 0.0000)
 Bb 1.0000 (0.0000 0.0034) 1.0000 (0.0005 0.0535) 0.0000 (0.0021 0.0000)
 bb 0.0000 (0.0001 0.0000) 0.0000 (0.0021 0.0000) 0.0000 (0.0080 0.0000)

 Cc AA                     Aa                     aa
 BB 1.0000 (0.0000 0.0034) 1.0000 (0.0005 0.0535) 0.0000 (0.0021 0.0000)
 Bb 1.0000 (0.0005 0.0535) 1.0000 (0.0083 0.8290) 0.0000 (0.0321 0.0000)
 bb 0.0000 (0.0021 0.0000) 0.0000 (0.0321 0.0000) 0.0000 (0.1246 0.0000)

 cc AA                     Aa                     aa
 BB 0.0000 (0.0001 0.0000) 0.0000 (0.0021 0.0000) 0.0000 (0.0080 0.0000)
 Bb 0.0000 (0.0021 0.0000) 0.0000 (0.0321 0.0000) 0.0000 (0.1246 0.0000)
 bb 0.0000 (0.0080 0.0000) 0.0000 (0.1246 0.0000) 0.0000 (0.4829 0.0000)

0.1142 0.1142 0.1142 (0.01000 0.19913 0.21154 1.00000)
```

Epätäydellinen yhden lokuksen malli $M_{C,1}$:

```
    AA                      Aa                      aa
    0.3200 (0.0035 0.1120)  0.0400 (0.1113 0.4454)  0.0050 (0.8852 0.4426)

0.0592 (0.01000 0.02364 0.02829 0.05588  0.00)
```

Epätäydellinen kahden lokuksen malli $M_{C,2}$:

```
    AA                      Aa                      aa
 BB 0.9000 (0.0000 0.0002)  0.7500 (0.0000 0.0030)  0.4000 (0.0002 0.0073)
 Bb 0.6500 (0.0003 0.0191)  0.4000 (0.0053 0.2120)  0.0500 (0.0239 0.1193)
 bb 0.2200 (0.0097 0.2127)  0.0200 (0.1746 0.3481)  0.0010 (0.7859 0.0783)

0.1000 0.0150 0.01004 (0.06068 0.06728 0.16240)
```

Epätäydellinen kolmen lokuksen malli $M_{C,3}$:

```
 CC AA                      Aa                      aa
 BB 0.7000 (0.0000 0.0000)  0.7000 (0.0000 0.0001)  0.7000 (0.0000 0.0002)
 Bb 0.7000 (0.0000 0.0003)  0.6000 (0.0000 0.0023)  0.5000 (0.0001 0.0038)
 bb 0.4000 (0.0000 0.0017)  0.2500 (0.0003 0.0083)  0.0500 (0.0007 0.0033)

 Cc AA                      Aa                      aa
 BB 0.7000 (0.0000 0.0005)  0.6000 (0.0001 0.0037)  0.5000 (0.0001 0.0061)
 Bb 0.6500 (0.0003 0.0174)  0.4000 (0.0021 0.0858)  0.3000 (0.0043 0.1287)
 bb 0.0400 (0.0024 0.0094)  0.0100 (0.0188 0.0188)  0.0100 (0.0376 0.0376)

 cc AA                      Aa                      aa
 BB 0.4000 (0.0001 0.0043)  0.2500 (0.0009 0.0217)  0.0500 (0.0017 0.0087)
 Bb 0.0400 (0.0038 0.0152)  0.0100 (0.0305 0.0305)  0.0100 (0.0610 0.0610)
 bb 0.0150 (0.0334 0.0501)  0.0080 (0.2672 0.2136)  0.0050 (0.5345 0.2670)

0.2000 0.0540 0.0340 (0.01001 0.03837 0.03989 0.11033)
```

A.2 Uskottavuuksien ylärajat

Havaintoaineistolle, joka koostuu pareista (N_1, N), (N_1^O, N^O), (N_1^S, N^S) ja (N_1^{MZ}, N^{MZ}), saadaan uskottavuuden yläraja laskemalla uskottavuus kuvitteellisessa geneettisessä mallissa, jolle $K_R = N_1^R/N^R$. Huomautettakoon, että esimerkiksi kytkemättömän periytymisen malliperheessä ei tällaista mallia välttämättä ole olemassa. Taulukossa A.1 luetellaan ylärajat simulaatiokokeissa käytettyjen aineistojen uskottavuuksille.

aineisto	hyvä tarkkuus	koht. tarkkuus	heikko tarkkuus
$D_{R,1}$	-14.585	-11.133	-7.688
$D_{R,2}$	-14.423	-10.971	-7.534
$D_{R,3}$	-14.344	-10.892	-7.441
$D_{D,1}$	-15.181	-11.728	-8.279
$D_{D,2}$	-14.993	-11.540	-8.096
$D_{D,3}$	-14.758	-11.301	-7.854
$D_{C,1}$	-15.790	-11.239	-6.943
$D_{C,2}$	-17.128	-12.525	-8.080
$D_{C,3}$	-16.499	-11.901	-7.391
Diabetes	—	-13.226	—
Skitsofrenia	—	-11.785	—

Taulukko A.1: Uskottavuuksien ylärajat eri aineistoilla. Esitetyt luvut ovat uskottavuuksien 10-kantaisia logaritmeja. Todellisille aineistoille *Diabetes* ja *Skitsofrenia* on käytössä vain yksi tarkkuus.

A.3 Parametriryppäät

Parametrirypäs on joukko mallin parametreja, jotka komponenteittaisessa Metropolisin ja Hastingsin algoritmissa päivitetään yhtaikaisesti. GeME-ohjelmalle annetaan syötteenä jono parametriryppäitä eli parametrien *ryvästys*. Jokaisella iterointikierroksella syötteessä kuvatut ryppäät päivitetään annetussa järjestyksessä. Ryppäiden jono esitetään ohjelman syötetiedostossa tietyn esitysmuodon mukaisesti. Seuraavaksi esitellään käytetty muoto lyhyesti.

Parametreihin viitataan kokonaisluvuilla $0, 1, 2, \ldots$. Jos mallissa on ℓ lokusta, luvulla $i \in \{0, 1, \ldots, \ell - 1\}$ viitataan lokuksen $i - 1$ alleelifrekvenssiin. Esimerkiksi luku 1 viittaa parametriin $p(B)$, kun lokuksia on vähintään kaksi. Luvut

$j \in \{\ell, \ldots, \ell + 3^\ell - 1\}$ viittaavaat ilmenemisyleisyyksiin. Esimerkiksi kahden lokuksen malleissa järjestys on

$$f(AABB), f(AaBB), f(aaBB), f(AABb), \ldots, f(aabb),$$

jolloin luku 5 viittaa parametriin $f(AABb)$. Yksi rypäs kuvataan antamalla parametreihin viittaavat luvut suuruusjärjestyksessä (pienin luku ensin). Alleelifrekvensseihin ja ilemenemisyleisyyksiin viittaavien lukujen väliin merkitään erotin ";". Sama erotinmerkki kirjoitetaan ryppään kuvauksen perään.

Lisäksi käytetään kahta lyhennysmerkintää. Merkintä "`a ; a ;`" tarkoittaa rypästä, johon kuuluvat kaikki mallin parametrit. Merkintä "`. ; . ;`" on lyhennys $\ell + 3^\ell$ ryppään jonolle: jonon k:nteen ryppääseen kuuluu ainoastaan se parametri, johon viitataan luvulla $k - 1$.

Rypäsjonot kuvaavan tiedoston ensimmäisellä rivillä ilmaistaan tarkasteltava malliluokka. Jos esimerkiksi simuloidaan kahden lokuksen malleilla, ensimmäiselle riville kirjoitetaan "2".

Simulaatiokokeissa käytettiin ensin (peruskokeissa) kullekin kolmelle malliluokalle yksinkertaista ryvästystä, jossa jokainen rypäs koostui vain yhdestä parametrista. Edellä mainituin merkinnöin tämä ryvästys ilmaistaan kirjoittamalla "`. ; . ;`". Lisäkokeissa sallitiin mallien epämonotonisuus, ja voitiin siksi käyttää huomattavasti monipuolisempia ryvästyksiä – yhtä kullekin malliluokalle. Seuraavaksi luetellaan nämä ryvästykset käyttäen edellä selostettua esitystapaa.

Yhden lokuksen parametriryppäät

```
1
a ; a ;
. ; . ;
0 ; 1 ;
0 ; 2 ;
0 ; 3 ;
; 1 2 ;
; 2 3 ;
; 1 3 ;
```

Kahden lokuksen parametriryppäät

```
2
a ; a ;
. ; . ;
0 1 ; ;
0 1 ; 2 ;
0 1 ; 3 5 ;
0 1 ; 3 4 ;
0 1 ; 5 8 ;
0 1 ; 3 5 6 ;
0 1 ; 4 7;
0 1 ; 8 9;
0 1 ; 7 9;
0 1 ; 10 ;
0 1 ; 3 ;
0 1 ; 4 ;
0 1 ; 5 ;
0 1 ; 6 ;
0 1 ; 7 ;
0 1 ; 8 ;
0 1 ; 9 ;
```

Kolmen lokuksen parametriryppäät

```
3
a ; a ;
. ; . ;
0 1 2 ; ;
; 3 4 5 ;
; 6 7 8 ;
; 9 10 11 ;
; 12 13 14 ;
; 15 16 17 ;
; 18 19 20 ;
; 21 22 23 ;
; 24 25 26 ;
; 27 28 29 ;
0 1 2 ; 3 ;
0 1 2 ; 29 ;
0 1 2 ; 23 25 ;
0 1 2 ; 14 16 23 25 ;
0 1 2 ; 6 7 13 18 ;
0 1 2 ; 1 ;
0 1 2 ; 2 ;
0 1 2 ; 3 ;
0 1 2 ; 4 ;
0 1 2 ; 5 ;
0 1 2 ; 6 ;
0 1 2 ; 7 ;
```

```
0 1 2 ; 8 ;
0 1 2 ; 9 ;
0 1 2 ; 10 ;
0 1 2 ; 11 ;
0 1 2 ; 12 ;
0 1 2 ; 13 ;
0 1 2 ; 14 ;
0 1 2 ; 15 ;
0 1 2 ; 16 ;
0 1 2 ; 17 ;
0 1 2 ; 18 ;
0 1 2 ; 19 ;
0 1 2 ; 20 ;
0 1 2 ; 21 ;
0 1 2 ; 22 ;
0 1 2 ; 23 ;
0 1 2 ; 24 ;
0 1 2 ; 25 ;
0 1 2 ; 26 ;
; 0 1 3 4 9 10 12 13 ;
```

Liite B

Tuloksia

Tässä liitteessä esitellään valikoidusti simulaatiokokeiden tuloksia. Ensin annetaan joukko esimerkkimalleja, jotka edustavat peruskokeiden simulaatioissa tuotetettuja otoksia. Sitten esitellään histrogrammikuvia vallitsevan ja epätäydellisen muodon aineistojen uskottavuuksien jakaumista eri malliluokissa.

B.1 Esimerkkimalleja

Alla on kahden lokuksen malli (Malli 1), joka sopii hyvin aineistoon $D'_{R,1}$. Jokaiseen genotyyppiin on liittetty kolme lukua, joista kaksi on merkitty sulkeisiin. Ensimmäinen luku on genotyypin ilmenemisyleisyys. Toinen luku ilmoittaa, kuinka suuri on genotyypin kantajien osuus populaatiosta. Kolmas luku kertoo, kuinka suuri osa sairaista on genotyypin kantajia. Taulukon alle on kirjoitettu ensin alleelin A frekvenssi ja sitten alleelin B frekvenssi. Tämän jälkeen sulkeissa ilmoitetaan, mikä on mallin mukaan taudin yleisyys, lapsiriski, sisarusriski ja identtisen kaksosen riski. Viimeinen luku kertoo havaintoaineiston logaritmisen uskottavuuden kyseisessä mallissa.

```
Malli 1:
     AA                         Aa                         aa
 BB  0.9996 (0.0000 0.0000)  0.1707 (0.0000 0.0000)  0.1117 (0.0000 0.0000)
 Bb  0.9977 (0.0004 0.0035)  0.0638 (0.0006 0.0041)  0.0004 (0.0029 0.0006)
```

```
 bb  0.9968 (0.0097 0.9779)  0.0005 (0.1773 0.0090)  0.0000 (0.8095 0.0082)

0.0987 0.0018 (0.00989 0.09797 0.29606 0.97851  -13.50)
```

Alla on kolme mallia, jotka edustavat kahden lokuksen mallien posteriorijakaumaa aineistolla $D'_{R,2}$.

```
Malli 2:
     AA                        Aa                        aa
 BB  0.9954 (0.0101 0.9842)  0.0006 (0.0645 0.0038)  0.0002 (0.1031 0.0020)
 Bb  0.0005 (0.0277 0.0014)  0.0002 (0.1770 0.0035)  0.0001 (0.2830 0.0028)
 bb  0.0004 (0.0190 0.0007)  0.0000 (0.1214 0.0000)  0.0000 (0.1942 0.0000)

0.2382 0.4216 (0.01020 0.09905 0.19025 0.97956  -13.26)

Malli 3:
     AA                        Aa                        aa
 BB  0.9881 (0.0010 0.0952)  0.9619 (0.0090 0.8293)  0.0129 (0.0201 0.0249)
 Bb  0.0062 (0.0096 0.0057)  0.0008 (0.0856 0.0066)  0.0006 (0.1915 0.0110)
 bb  0.0048 (0.0228 0.0105)  0.0003 (0.2040 0.0059)  0.0003 (0.4563 0.0131)

0.1827 0.1735 (0.01042 0.10172 0.20393 0.89224  -24.07)

Malli 4:
     AA                        Aa                        aa
 BB  0.9905 (0.0004 0.0362)  0.9676 (0.0053 0.5150)  0.0009 (0.0193 0.0017)
 Bb  0.9399 (0.0039 0.3661)  0.0052 (0.0566 0.0295)  0.0006 (0.2058 0.0124)
 bb  0.0073 (0.0104 0.0076)  0.0021 (0.1506 0.0317)  0.0000 (0.5478 0.0000)

0.1209 0.1581 (0.00998 0.13226 0.19972 0.87855  -29.25)
```

Mallit 5 ja 6 edustavat kahden lokuksen mallien posteriorijakaumaa aineistolla $D'_{D,2}$.

```
Malli 5:
     AA                        Aa                        aa
 BB  0.9993 (0.0000 0.0007)  0.9970 (0.0004 0.0369)  0.0018 (0.0050 0.0009)
 Bb  0.9982 (0.0002 0.0171)  0.9936 (0.0092 0.9303)  0.0003 (0.1265 0.0038)
 bb  0.0369 (0.0011 0.0040)  0.0007 (0.0584 0.0041)  0.0000 (0.7993 0.0000)
```

```
0.0352 0.0733 (0.00987 0.28682 0.29316 0.97922  -13.77)

Malli 6:
     AA                       Aa                       aa
 BB  0.9903 (0.0000 0.0008)   0.9736 (0.0001 0.0078)   0.2369 (0.0002 0.0049)
 Bb  0.9803 (0.0009 0.0860)   0.9510 (0.0092 0.8644)   0.0022 (0.0239 0.0052)
 bb  0.0049 (0.0253 0.0122)   0.0003 (0.2619 0.0077)   0.0002 (0.6785 0.0133)

0.1618 0.0173 (0.01017 0.29388 0.30158 0.91588  -20.58)
```

Mallit 7 ja 8 edustavat yhden lokuksen mallien posteriorijakaumaa aineistolla *Diabetes*.

```
Malli 7:
     AA                       Aa                       aa
     0.9997 (0.0013 0.2516)   0.0522 (0.0684 0.7146)   0.0002 (0.9303 0.0372)

0.0355 (0.00500 0.05306 0.09999 0.28883  -13.27)

Malli 8:
     AA                       Aa                       aa
     0.7983 (0.0024 0.3640)   0.0344 (0.0934 0.6075)   0.0002 (0.9042 0.0342)

0.0491 (0.00529 0.04863 0.10350 0.31146  -14.96)
```

Mallit 9 ja 10 edustavat kahden lokuksen mallien posteriorijakaumaa aineistolla *Diabetes*.

```
Malli 9:
     AA                       Aa                       aa
 BB  0.9501 (0.0000 0.0003)   0.7960 (0.0001 0.0162)   0.7866 (0.0018 0.2797)
 Bb  0.9414 (0.0001 0.0121)   0.2242 (0.0045 0.2019)   0.0212 (0.0781 0.3342)
 bb  0.3001 (0.0007 0.0428)   0.0107 (0.0494 0.1068)   0.0000 (0.8654 0.0000)

0.0278 0.0432 (0.00495 0.05228 0.10093 0.31088  -13.29)

Malli 10:
```

```
     AA                       Aa                       aa
 BB  0.7250 (0.0000 0.0003)  0.3546 (0.0001 0.0044)  0.0716 (0.0005 0.0068)
 Bb  0.6837 (0.0002 0.0238)  0.2384 (0.0053 0.2504)  0.0058 (0.0401 0.0460)
 bb  0.5926 (0.0037 0.4320)  0.0063 (0.1112 0.1386)  0.0006 (0.8390 0.0996)

0.0622 0.0233 (0.00505 0.05074 0.10469 0.33546  -14.13)
```

Malli 11 edustaa yhden lokuksen mallien posteriorijakaumaa aineistolla *Skitsofrenia*.

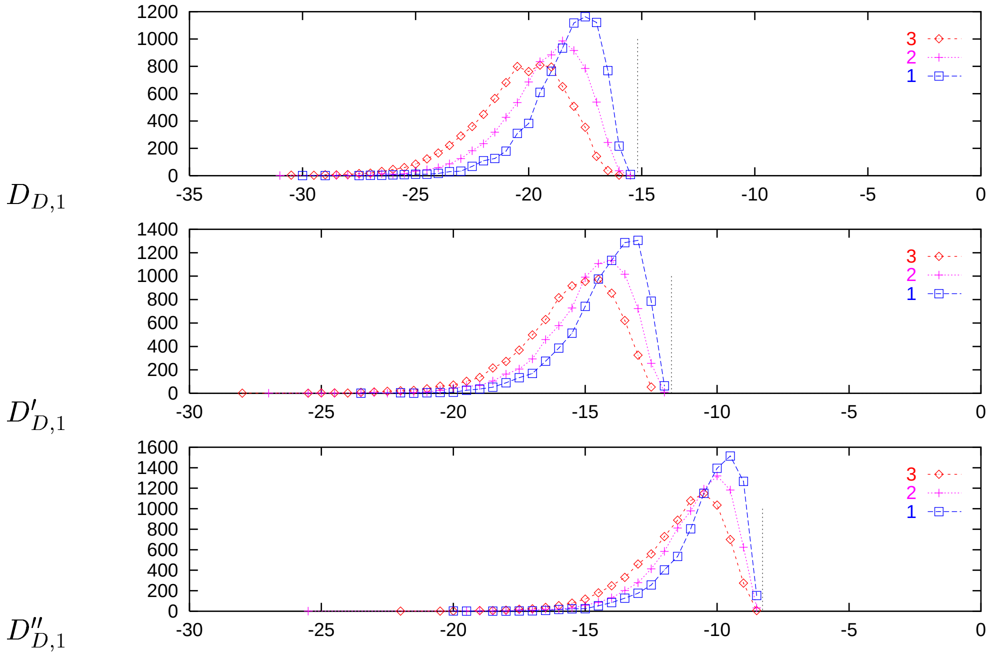
```
Malli 11:
     AA                       Aa                       aa
     0.9782 (0.0018 0.2067)  0.0813 (0.0814 0.7738)  0.0002 (0.9168 0.0214)

0.0425 (0.00856 0.07238 0.10462 0.26517  -25.88)
```

Malli 12 edustaa kahden lokuksen mallien posteriorijakaumaa aineistolla *Skitsofrenia*.

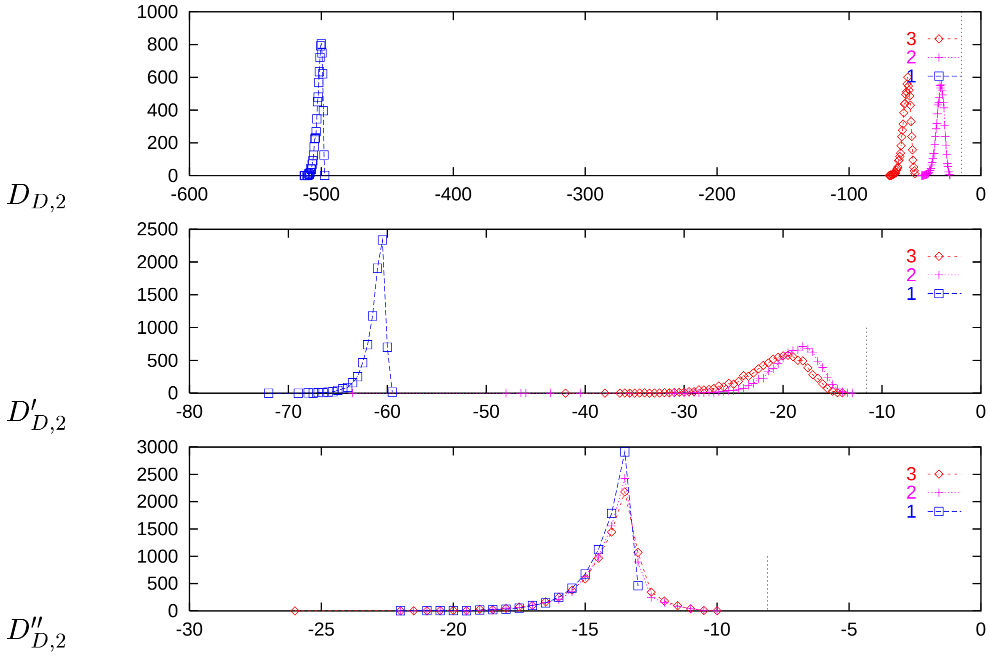
```
Malli 12:
     AA                       Aa                       aa
 BB  0.9654 (0.0000 0.0052)  0.8018 (0.0045 0.4221)  0.0119 (0.1082 0.1521)
 Bb  0.5426 (0.0002 0.0116)  0.0225 (0.0176 0.0469)  0.0062 (0.4282 0.3136)
 bb  0.2715 (0.0002 0.0058)  0.0027 (0.0174 0.0056)  0.0008 (0.4237 0.0400)

0.0202 0.3357 (0.00847 0.07657 0.09472 0.35619  -16.72)
```

B.2 Aineistojen uskottavuuksien jakaumia

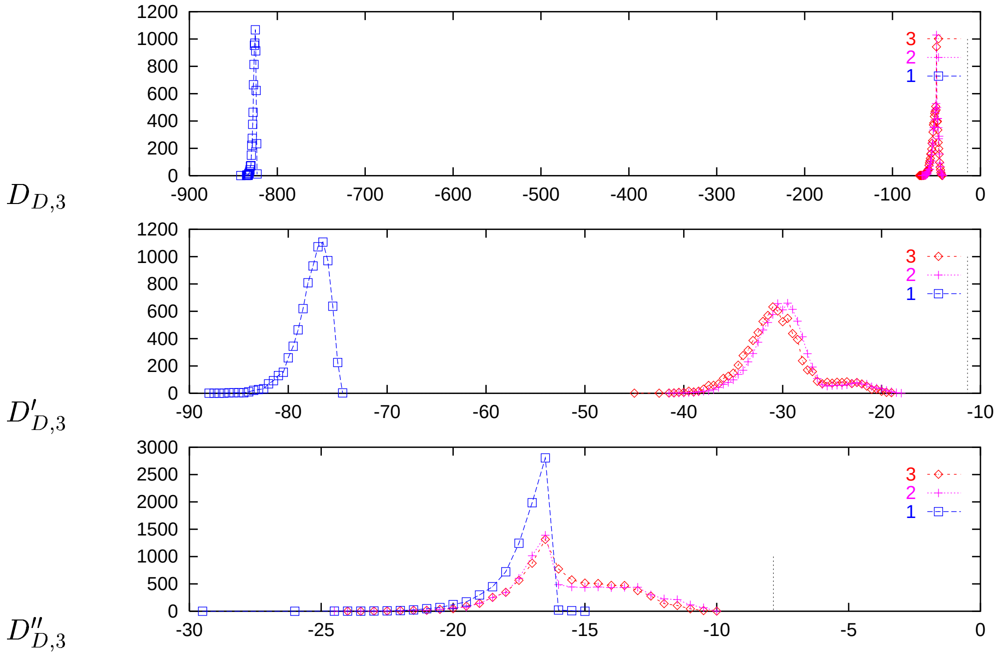
Kuvat B.1 – B.3 esittävät vallitsevan muodon aineistojen uskottavuuksien jakaumia eri malliluokkia, kun aineiston tarkkuus vaihtelee. Tulokset perustuvat lisäkokeisiin (ks. osio 7.2.3) ja mallien sallitan siten olevan epämonotonisia. Kuvat B.4, B.5 ja B.6 esittävät vastaavasti epätäydellisen muodon aineistojen uskottavuuksien jakaumia. Tulokset perustuvat peruskokeisiin (ks. osio 7.2.3) ja mallien vaaditaan siten olevan monotonisia.



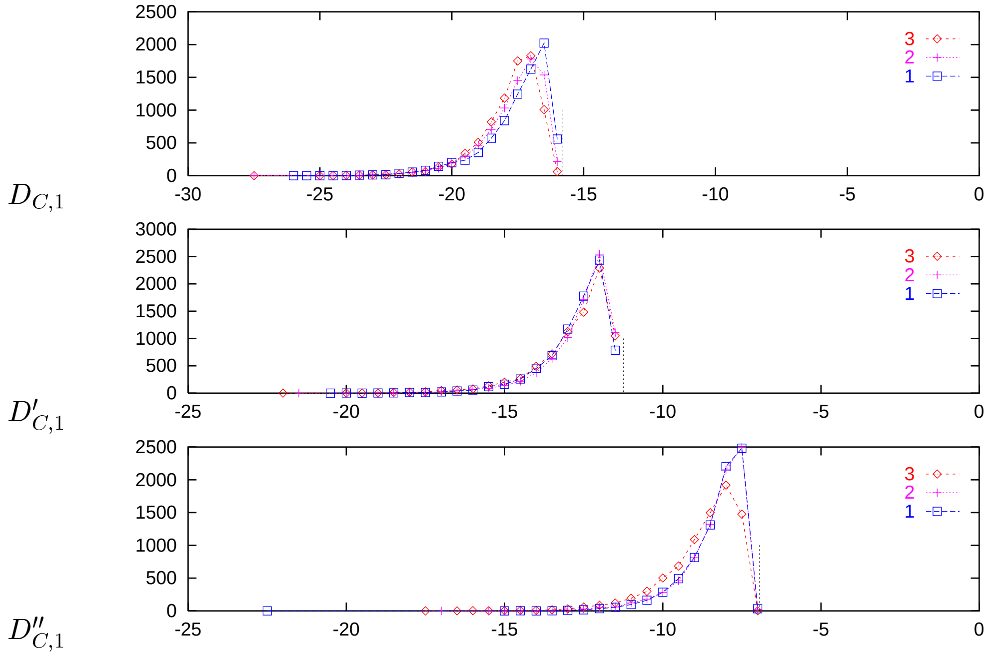
Kuva B.1: Aineistojen $D_{D,1}$, $D'_{D,1}$ ja $D''_{D,1}$ uskottavuuksien histogrammit yhden, kahden ja kolmen lokuksen epämonotonissa malleissa. Kustakin kolmesta simulaation tuloksesta on valittu 9 000 mallia. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä.



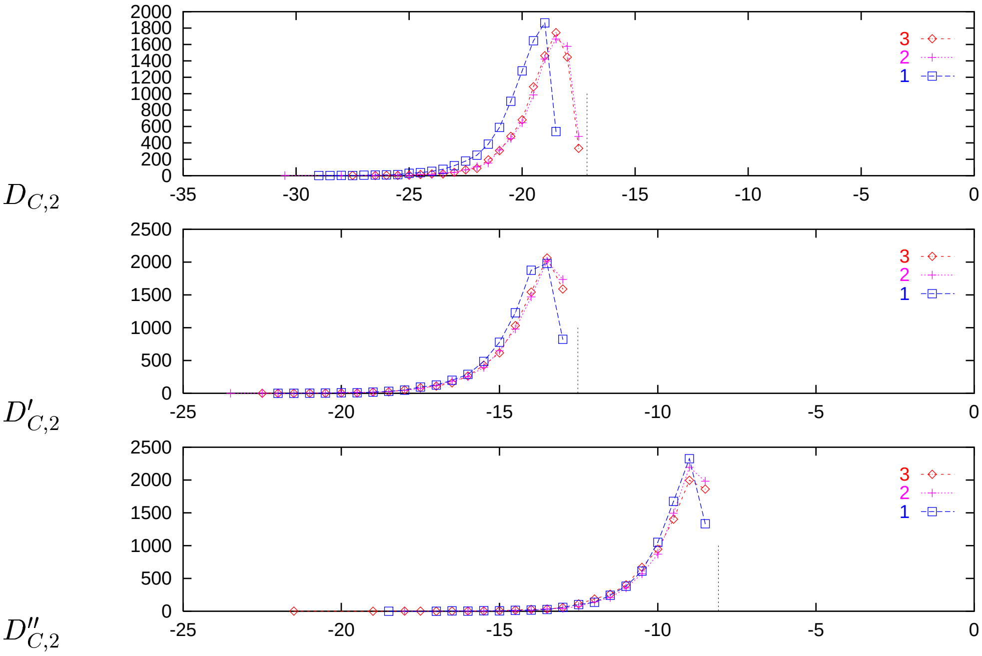
Kuva B.2: Aineistojen $D_{D,2}$, $D'_{D,2}$ ja $D''_{D,2}$ uskottavuuksien histogrammit yhden, kahden ja kolmen lokuksen epämonotonissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.



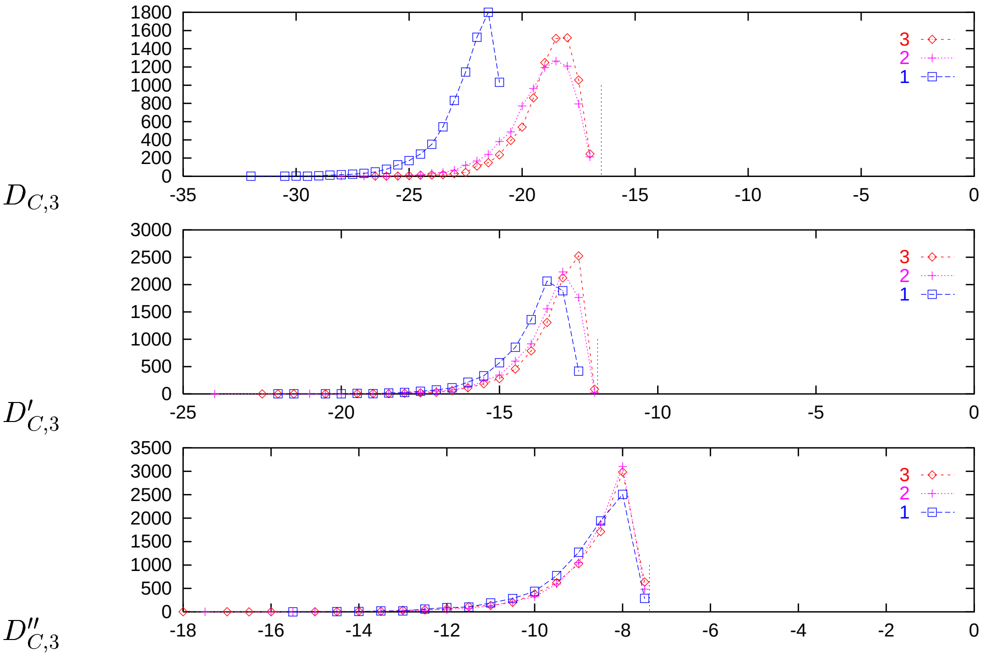
Kuva B.3: Aineistojen $D_{D,3}$, $D'_{D,3}$ ja $D''_{D,3}$ uskottavuuksien histogrammit yhden, kahden ja kolmen lokuksen epämonotonisissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.



Kuva B.4: Aineistojen $D_{C,1}$, $D'_{C,1}$ ja $D''_{C,1}$ uskottavuuksien histogrammit yhden, kahden ja kolmen lokuksen monotonisissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.



Kuva B.5: Aineistojen $D_{C,2}$, $D'_{C,2}$ ja $D''_{C,2}$ uskottavuuksien histogrammit yhden, kahden ja kolmen lokuksen monotonisissa malleilla. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.



Kuva B.6: Aineistojen $D_{C,3}$, $D'_{C,3}$ ja $D''_{C,3}$ uskottavuuksien histogrammit yhden, kahden ja kolmen lokuksen monotonisissa malleissa. Vaaka-akselilla on logaritminen uskottavuus ja pystyakselilla mallien lukumäärä. Otosten koko on 9 000 mallia. Uskottavuuden yläraja on osoitettu pystyviivalla.