Tilastotieteen pro gradu –tutkielma

# Puuttuvan tiedon käsittely aivosähkökäyrämittauksissa

Lauri Era

Jyväskylän yliopisto

Matematiikan ja tilastotieteen laitos

25. Toukokuuta 2016

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Era, Lauri: Puuttuvan tiedon käsittely aivosähkökäyrämittauksissa

Tilastotieteen pro gradu -tutkielma, 26 sivua 8 liitettä

11.7.2016

# Tiivistelmä

Aivosähkökäyrämittaukset ovat tyypillisesti hyvin työläitä ja pitkäkestoisia. Näiden seikkojen johdosta otoskoot ovat usein pieniä. Lisäksi osa koehenkilöistä päätyy lopettamaan mittaukset kesken, tai ei jaksa keskittyä annetun ohjeistuksen mukaisesti. Näin ollen valmiiksi hyvin rajalliseen aineistoon syntyy puuttuvuutta siten, että tietyt osakokeet jäävät kokonaan mittaamatta joidenkin henkilöiden osalta.

Aineistossa ilmenevä puuttuvuus on käsitelty aivojenkuvantamismittauksissa tavanomaisesti siten, että analyysiin otetaan mukaan vain niitä koehenkilöitä koskevat mittaukset, joilta ei puutu lainkaan tietoa. Tätä menetelmää kutsutaan täydellisten havaintorivien analyysiksi. Tällainen analyysi tuottaa harhattomia tuloksia vain siinä tilanteessa, että puuttuvuus on täysin satunnaista. Lisäksi päädytään haaskaamaan tehdyt mittaukset niiltä henkilöiltä, joilta jokin osakoe puuttuu.

Moni-imputointi on yleinen tapa käsitellä puuttuvuutta. Siinä jokainen puuttuva havainto korvataan joko mallin mukaisesta jakaumasta simuloidulla arvolla tai jollain havaitulla arvolla. Tässä työssä on valittu, että korvaavat arvot ovat havaittuja arvoja. Koehenkilöä, jonka arvoa käytetään korvaavana havaintona, kutsutaan luovuttajaksi. Työn ensimmäinen lähestymistapa luovuttajan valitsemiseen on mallipohjainen, jossa haetaan uskottavaa havaintoa puuttuvan tilalle mallintamalla puuttuvuutta sisältävä muuttuja. Käytettäessä mallipohjaista moni-imputointia saadaan harhattomia tuloksia myös satunnaisen puuttuvuuden tilanteessa, jolloin oletusta puuttuvuuden täydestä satunnaisuudesta ei tarvita. Toinen lähestymistapa on korvaavien arvojen valitseminen satunnaisesti. Tällöin vaaditaan edelleen oletus puuttuvuuden täydestä satunnaisuudesta, mutta tietoa ei jouduta haaskaamaan.

Aivosähkökäyrämittauksissa saadaan jokaiselle koehenkilölle kuhunkin eri koetilanteeseen liittyvä aikasarja jokaisesta mittaukseen käytetystä elektrodista. Mittauksista lasketaan vakiintuneita tunnuslukuja, joilla kuvataan koehenkilön reaktion voimakkuutta annettuun ärsykkeeseen. Moni-imputoinnissa voidaankin käyttää näitä tunnuslukuja korvaavan arvon valitsemiseen. Valittaessa aivosähkökäyrämittausten yhden aikasarjan eri tunnusluvuille korvaavat arvot kerralla samalta luovuttajalta voidaan ajatella, että imputoidaan kokonainen aikasarja. Tällöin voitaisiin puuttuvuuden käsittelyä seuraavissa analyysin vaiheissa hyödyntää valitun luovuttajan koko aikasarjaa.

Tässä työssä menetelmien välisen paremmuuden mittaamiseen käytetyt mittarit ovat estimaattoreiden estimoitu ja havaittu keskivirhe sekä havaittu harha. Katsottaessa kahta eri valittua tapaa tehdä moni-imputointia sekä täydellisten havaintorivien analyysia nähdään, että näiden menetelmien paremmuusjärjestys riippuu tilanteesta. Kun havaintoja puuttuu vain vähän, mallipohjainen moni-imputointi ei eroa tarkastelluilla mittareilla muista menetelmistä, mutta sen soveltamiseksi vaaditaan lievemmät oletukset. Puuttuvan tiedon osuuden kasvaessa suureksi satunnaisesti valittujen luovuttajien moni-imputointimenetelmä nousi tarkasteltujen mittareiden valossa parhaimmaksi. Käsitellyt estimaattorit ovat aikasarjojen tunnusluvuille. Alkuperäisten aikasarjojen käyttö moni-imputointia seuraavissa analyysin vaiheissa onkin merkittävin jatkokehityssuunta.

**Avainsanat**: EEG, ERP, Baeys-bootstrap, hot deck, jackknife, moni-imputointi, puuttuva tieto, täydellisten havaintorivien analyysi.

# Kiitokset

Tähän työhön valikoitui aiheeksi puuttuvan tiedon käsittely aivosähkökäyrämittauksissa. Idea syntyi yliopistotutkija Piia Astikaisen ja professori Juha Karvanen kanssa käytyjen keskustelujen pohjalta. Professori Astikainen myös koosti ja antoi käytetyn aineiston käyttööni, tästä suuret kiitokset hänelle. Suurimmat kiitokset kuuluvat professori Karvaselle hänen monista hyvistä ideoistaan, rakentavasta palautteestaan ja kannustuksestaan työn kaikissa vaiheissa.

Olen kiitollinen professori Jukka Nyblomille hänen antamastaan palautteesta tekstiin, joka sen seurauksena on huomattavasti selkeämpi. Kiitokset myös Salla Manniselle viimeisen tekstin tarkistuksen kanssa, jossa kieliasu otti viimeisen harppauksensa eteenpäin.

Tampereella 25.5. 2016

Lauri Era

# Sisältö

# 1 Johdanto

Aivosähkökäyrämittauksilla (englanniksi electroencephalography eli EEG) seurataan aivojen sähköistä aktivaatiota päähän kiinnitettävien elektrodien avulla. Perusteellisempi esitys EEG-mittauksista löytyy esimerkiksi Niedemeyerin ja Da Silvan teoksesta ”*Electroencephalography Basic Principles, Clinical Applications, and Related Fields*” (2005). Herätevastemittaukset (event related potential, ERP) tarkoittavat EEG-mittauksia, joissa tutkittavalle esitetään ärsyke, jonka aiheuttamasta aivojen aktivaation muutoksesta ollaan kiinnostuneita. Tyypillisesti nämä aktivaation muutokset ovat ääriarvoja tietyllä aikavälillä ärsykkeen esittämisen jälkeen. Esimerkki tällaisesta ääriarvosta on maksimiarvo aikavälillä 70 - 130 millisekuntia ärsykkeen esittämisestä.

Aivoissa sähköistä aktivaatiota on jatkuvasti jonkin verran kaikkialla aivoissa, lisäksi eri puolilla aivoja tapahtuvat aktivaatiot sekoittuvat toisiinsa pään pinnalta mitattaessa. Tämä tekee yksittäisten mittausten välisestä vaihtelusta suurta, vaikka mittaukset olisivat samalta koehenkilöltä ja samanlaisista tilanteista. Tämän takia kutakin eri tarkasteltavaa ärsykettä täytyy toistaa lukuisia kertoja, jotta saadaan juuri siihen liittyvä aktivaation muutos selville. Näitä tarkasteltavia ärsykkeitä saattaa olla yhdessä tutkimuksessa useita. Nämä seikat tekevät mittaustilanteista hyvin pitkiä. Henkilön keskittyminen tutkimuksen ohjeistuksen mukaisesti saattaakin herpaantua jossakin vaiheessa tutkimusta ohjeistuksen ollessa esimerkiksi katseen kohdistaminen keskelle tietokoneen näyttöä. Mikäli henkilö ei ole keskittynyt ohjeistuksen mukaisesti, ei hänen aivojensa aktivaatiokaan kuvasta haluttua ilmiötä. Näin ollen aineisto on puutteellinen siten, että tiettyä ärsykettä koskevat mittaustulokset puuttuvat kokonaan kyseisen henkilön kohdalta. Koetilanteiden kokonaiskesto luultavasti vaikuttaa myös siihen, että toisinaan koehenkilöt keskeyttävät mittaukset kokonaan, jolloin kaikki sillä hetkellä käymättömät koetilanteet jäävät puuttuviksi. Aivojenkuvantamistutkimuksissa otoskoot eivät tyypillisesti ole kovin suuria, jolloin saattaa olla haitallista, jos pienikin määrä havainnoista puuttuu. Tilanne vaatii siis puuttuvan tiedon järkevää ja tehokasta käsittelyä.

Puuttuvuutta on yleisesti käsitelty ERP-datassa jättämällä kokonaan pois sellaiset koehenkilöt, joista puuttuu tieto yhdestä tai useammasta muuttujasta. Tätä käsittelytapaa puoltaa sen yksinkertaisuus ja käsittelyn nopeuden tuoma tehokkuus. Menetelmän yhtenä huonona puolena on, että vahvan oletuksen puuttuvuuden täydestä satunnaisuudesta täytyy päteä, jotta yksinkertainen poisjättäminen tuottaisi harhattomia tuloksia. Toinen huono puoli on, että puuttuvuutta sisältävästä muuttujasta mahdollisesti saatavilla olevaa muuta informaatiota ei hyödynnetä mitenkään. Kolmas huono puoli on se, että jätettäessä kokonaisia koehenkilöitä tarkastelun ulkopuolelle joudutaan heidän osalta hylkäämään arvot myös havaittujen muuttujien osalta. Esimerkki kolmannesta huonosta puolesta on tilanne, jossa kymmenessä eri muuttujassa on puuttuvuutta riippumattomasti todennäköisyydellä 0.1. Esimerkissä odotusarvo osajoukon osuudelle, jossa ei ole lainkaan puuttuvuutta, on $0.9^{10} \approx 0.35$. Esimerkin tilanteessa päädytään hyödyntämään vain noin kolmasosaa saatavilla olevasta datasta, vaikka datasta puuttuu vain kymmenesosa. Tarkoituksena on vertailla tätä jo käytössä olevaa täydellisten havaintorivien analyysiksi kutsuttua metodia moni-imputointiin.

Tässä työssä käytetään puuttuvuuden käsittelyyn moni-imputointia. Imputointi tarkoittaa puuttuvan arvon korvaamista ja moni-imputoinnissa tämä toistetaan useaan kertaan. Tämä voidaan toteuttaa joko mallipohjaisesti hakien uskottavaa arvoa tai satunnaisesti. Tarkasteltava kysymys on, saadaanko tyypillisessä EEG-mittauksessa moni-imputoinnilla tarkempia estimaatteja kuin täydellisten havaintorivien analyysilla käytettäessä tarkkuuden mittarina estimaattien hajontaa ja harhaa. Valitut moni-imputointi-menetelmät ovat hot deck -menetelmiä, joissa puuttuva havainto korvataan useita kertoja jollakin havaitulla arvolla, niin sanotun luovuttajan havaintoarvolla. Vertailu toteutetaan tutkimalla, miten puuttuvan tiedon osuus vaikuttaa saatuihin tuloksiin eri metodeja käytettäessä. Ensimmäisessä hot deck -sovelluksessa tavoitteena

on muun tiedon käyttäminen, jotta puuttuvuutta sisältävästä muuttujasta saataisiin mahdollisimman paljon informaatiota. Muu tieto tarkoittaa tässä yhteydessä mittauksia toisesta tilanteesta, joilla pyritään mallintamaan puuttuvuutta sisältävässä tilanteessa saatavia havaittuja arvoja. Toinen valittu lähestymistapa hot deck -moni-imputointiin on Bayes-bootstrap, jossa luovuttaja valitaan satunnaisesti ilman mallinnusta. Tarkemman tuloksen lisäksi halutaan mahdollisimman harhaton arvio siitä epävarmuudesta, jota tiedon puuttuminen aina väistämättä tuottaa. Erityisesti puuttuvuuden aiheuttamaa informaation menetystä voidaan luotettavammin arvioida käyttämällä moni-imputointia kuin käyttämällä yksinkertaista imputointia tai täydellisten havaintorivien analyysia.

Käsiteltävä aineisto on Jyväskylän yliopiston psykologian laitoksen tutkimuksesta, jossa mitattiin koehenkilöiden aivoissa tapahtuvaa aktivaatiota erilaisissa koeasetelmissa. Kaikissa koeasetelmissa henkilölle näytetään ihmisten kasvoja ruudulta, tarkastelu rajoitetaan kolmeen koeasetelmaan. Vertailu toteutetaan simuloimalla puuttuvuutta havaitun aineiston pohjalta tuotettuun keinotekoiseen aineistoon ja tekemällä eri analyysit kullakin simulointikierroksella. Samantyyppisesti, mutta simuloiden puuttuvuutta suoraan havaittuun aineistoon, on edetty tämän metodin soveltuvuutta tutkittaessa fMRI:n yhteydessä (Vaden, Gebregziabher, Kuchinsky & Eckert, 2012).

Tässä työssä tutustutaan aluksi yleisesti puuttuvaan tietoon ja sen eri tyyppeihin. Seuraavaksi esitellään moni-imputointi yleisesti ja miten sitä sovelletaan tämän työn erityiseen tilanteeseen. Tämän jälkeen esitellään käytetty aineisto ja miten tähän aineistoon saadaan sovitetuksi esitelty moni-imputointimenetelmä. Sen jälkeen esitellään saadut tulokset ja esitetään näiden tulosten tulkinnat. Lopuksi käydään läpi pohdintaa siitä, mitä tulokset tarkoittavat ja nostetaan esiin mahdollisia seuraavia askeleita esitellyn kysymyksen parissa.

# 2 Puuttuva tieto

Puuttuvaa tietoa voi ilmetä joko kaikissa koehenkilöä koskevissa muuttujissa, jolloin voidaan puhua kadosta (nonresponse), tai vain joissain muuttujissa, jolloin voidaan puhua osittaiskadosta (*item nonresponse*) (Graham, 2012). Tässä työssä tarkastellaan tilannetta, jossa koehenkilöstä on aina joitain tietoja käytettävissä eli osittaiskadosta. Silloin on mahdollista mallintaa muuttujat, joista havainto puuttuu ja tehdä toisten muuttujien avulla niistä ennusteet myös niille koehenkilöille, joilta tietoa puuttuu. Loppuosassa tätä työtä puhutaan koehenkilöistä eikä missään vaiheessa yleisemmin havaintoyksiköistä. Tämä valinta on tehty puhtaasti esitystä selventämään eikä itsessään tarkoita menetelmien rajoittuvan vain tietynlaisiin aineistoihin.

Muuttujan puuttuvuus voi olla täysin satunnaista, jolloin todennäköisyys puuttumiselle on sama kaikille koehenkilöille riippumatta koehenkilöiden saamista arvoista missään tarkastellussa muuttujassa. Tällaisen puuttuvuuden käsittely on kaikkein yksinkertaisinta, koska puuttuvuus ei tuota harhaa. Täysin satunnainenkin puuttuvuus tarkoittaa informaation menetystä ja on sikäli haitallista, joten sekin on syytä pyrkiä tehokkaasti käsittelemään. Toisen ääripään muodostaa ei-satunnainen puuttuvuus. Tällöin se, puuttuuko havainto vai ei, riippuu havainnosta itsestään esimerkiksi siten, että suuremmat arvot puuttuvat todennäköisemmin kuin pienet. Tämän tyyppinen puuttuvuus tuottaa selvästi harhaa tuloksiin. Ei-satunnaisen puuttuvuuden tehokas käsittely vaatiikin yleisessä tapauksessa puuttuvuusmekanismin mallinnusta, jotta harhasta päästään eroon. (Little & Rubin, 2002.)

Edellä esitetyn kahden ääripään välissä on satunnainen puuttuvuus. Tällöin muuttujan puuttuvuuden todennäköisyys vaihtelee, mutta ei riippuen puuttuvuutta sisältävästä muuttujasta itsestään vaan yhdestä

tai useammasta havaitusta muuttujasta. Ei-satunnaisen puuttuvuuden lisäksi myös satunnainen puuttuvuus saattaa tuottaa harhaa tuloksiin. Harhaa syntyy joissakin tapauksissa, kun käytetään täydellisten havaintorivien analyysia (complete case analysis) tai saatavilla olevien havaintojen analyysia (available case analysis tai pairwise deletion) (van Buuren, 2012.). Nykyaikaiset puuttuvan tiedon käsittelymenetelmät, muun muassa mallipohjainen hot deck -moni-imputointi, yleensä olettavat puuttuvuuden olevan satunnaista (van Buuren, 2012). Nämä menetelmät toimivat myös täysin satunnaisen puuttuvuuden tilanteessa, koska se on satunnaisen puuttuvuuden erikoistapaus, jossa puuttuvuuden todennäköisyys on vakio yli havaittujen muuttujien mahdollisten arvojen joukon.

Puuttuvuuden satunnaisuudesta ei yleisessä tapauksessa kuitenkaan voida sanoa mitään pelkästään havaitun datan pohjalta (Graham, 2012). Siksi puuttuvuuden tyyppi täytyy osittain määritellä sen mukaan, mikä vaikuttaa uskottavalta tarkasteltavassa tilanteessa. Tosin vaikka puuttuvuus ei olisi satunnaista, tämä ei välttämättä ole käytännön kannalta merkityksellistä, jos puuttuvuusmekanismin ja puuttuvien havaintojen välinen yhteys ei ole kovin vahva ja käytetään moni-imputointia puuttuvuuden käsittelyyn (Graham, 2012; van Buuren, 2012).

# 3 Moni-imputointimenetelmiä

Yleinen tapa käsitellä puuttuvuutta on moni-imputointi, joka soveltuu myös tilanteisiin, joissa puuttuvuus ei ole täysin satunnaista. Moni-imputoinnissa puuttuva arvo korvataan *D* kertaa. Korvaava arvo voi olla, joko mallinmukaisesta jakaumasta simuloitu tai jokin havaittu arvo. Koehenkilöä, jonka arvoa käytetään, kutsutaan luovuttajaksi. Menetelmää, jossa käytetään korvaavana arvona toisen koehenkilön havaittua arvoa, kutsutaan hot deck -moni-imputoinniksi. (Little & Rubin, 2002.) Tässä työssä käytettävät moni-imputointimenetelmät ovat hot deck -moni-imputointia.

Käytettäessä hot deck -menetelmää ERP-mittausten tapauksessa, ERP-mittausten autokorrelaatiorakennetta ei tarvitse suoranaisesti mallintaa lainkaan, sillä luovuttajan valintaan voidaan käyttää ERP-mittauksissa laskettavia ääriarvoja. Näin ollen menetelmällä voidaan imputoida kokonaisia aikasarjoja pelkkien pistearvojen sijaan. Moni-imputoinnin hot deck -metodia on sovellettu aivokuvantamiseen liittyvän puuttuvan tiedon käsittelyyn aiemminkin (Vaden et al., 2012), tuolloin kuvantamismenetelmänä oli toiminnallinen magneettiresonanssikuvaus (functional magnetic resonance imaging eli fMRI). Koska hot deck -menetelmä on onnistuttu ottamaan käyttöön toisenlaisessa aivojenkuvantamisessa, voidaan ajatella sen olevan mahdollinen vaihtoehto myös ERP-datan kohdalla.

Moni-imputoinnissa ratkaistaan ensin puuttuvan tiedon ongelma imputoimalla puuttuvat havainnot ja vasta tämän jälkeen siirrytään estimoimaan haluttuja suureita. Tämän järjestyksen mukaan edettäessä voidaan puuttuvan tiedon ongelma erottaa varsinaisesta tutkimuskysymyksestä ja tarkastella näitä kahta erillään. Tällöin voidaan rakentaa oma metodiikkansa käsittelemään puuttuvuuden tuottavaa mekanismia ilman, että se muuttaa alkuperäistä analyysia. Puuttuvuus ei aseta rajoitteita käytettävälle metodiikalle, kun se käsitellään moni-imputoimalla. (van Buuren, 2012.)

## 3.1 Mallipohjainen hot deck

Loppuosassa tekstiä tullaan käyttämään seuraavia merkintöjä. $Y$ on alkuperäinen havaintoarvojen vektori, joka sisältää sekä saatavilla olevat että puuttuvat havainnnot. $Y^{obs}$ on vektori, joka sisältää tarkasteltavan muuttujan kaikki saatavilla olevat havainnot. $Y^{mis}$ on vektori, jonka arvot ovat tuntemattomia ja jonka pituus on sama kuin puuttuvien havaintojen lukumäärä tarkasteltavassa muuttujassa. $Y^{mis,(d)}$ on vektori, joka sisältää imputoidut arvot puuttuville havainnoille moni-imputoinnin kierroksella *d*. $\hat{y}^{mis}$ on vektori, joka sisältää imputointimallin antamat ennusteet puuttuville havainnoille tarkasteltavasta muuttujasta annetulla moni-imputoinnin kierroksella. $\hat{y}^{obs}$ on vektori, joka sisältää imputointimallin antamat ennusteet havaituille havainnoille tarkasteltavasta muuttujasta. Olkoon esimerkiksi $Y = [3, NA, 4, NA, 2.5, 1.5]^T$ tällöin $Y^{mis} = [NA, NA]^T$ ja $Y^{obs} = [3, 4, 2.5, 1.5]^T$, olkoon meillä esimerkin tapauksessa myös tietoa muista muuttujista ja merkitään niitä

$$X = \begin{bmatrix} 9.6 & -5.7 \\ 6.7 & -4.0 \\ 12.8 & -9.6 \\ 7.1 & -3.4 \\ 7.5 & -3.2 \\ 4.7 & -2.8 \end{bmatrix}.$$

Tähän esimerkkiin palataan läpi osion 3 loppuosan.

Hot deck -moni-imputoinnissa on monta kierrosta. Jokaisella moni-imputoinnin kierroksella käytetään puuttuvan arvon sijaan jotain havaittua arvoa. Sitä koehenkilöä, jonka arvo asetetaan puuttuvan havainnon tilalle, kutsutaan luovuttajaksi. Luovuttajan valitsemista varten puuttuvuutta sisältävä muuttuja voidaan mallintaa. Tässä työssä valittu mallinnus on tavanomaista lineaarista regressiota. Mallista saadaan ennusteet niin puuttuville havainnoille kuin havaituille havainnoillekin. Mahdollisiksi luovuttajiksi valitaan ne koehenkilöt, joiden ennusteet ovat lähinnä puuttuvan havainnon ennustetta. Annetulle puuttuvalle havainnolle $Y_i^{mis}$ , voidaan valita mahdolliset luovuttajat laskemalla

$$\left|\hat{y}_i^{mis} - \hat{y}_j^{obs}\right|, \qquad (3.1)$$

missä $\hat{y}_j^{obs}$ on aina yhden havaitun havainnon ennustettu arvo ja näistä valitaan pienimmät yli kaikkien $\hat{y}^{obs}$. Tässä työssä pyritään kuitenkin ottamaan huomioon puuttuvuutta sisältävän muuttujan jakauman mahdollinen vinous. Tämä toteutetaan valitsemalla mahdollisiksi luovuttajiksi sekä koehenkilöitä, jotka ovat ennustearvoiltaan puuttuvaa havaintoa pienempiä, että koehenkilöitä, joille ennustearvot ovat suurempia. Kaksi koehenkilöä, joiden ennusteet ovat pienemmät kuin puuttuvan havainnon, saadaan laskemalla

$$\hat{y}_i^{mis} - \hat{y}_j^{obs}\,, \qquad (3.2)$$

missä erotukset lasketaan yli kaikkien niiden $\hat{y}_j^{obs}$, jotka ovat pienempiä kuin $\hat{y}_i^{mis}$. Mahdollisiksi luovuttajiksi valitaan ne kaksi *j*-havaintoa, joita vastaavat etäisyydet ovat kaikkein pienimmät. Kaksi koehenkilöä, joiden ennusteet ovat suuremmat tai yhtäsuuret kuin puuttuvan havainnon, saadaan laskemalla

$$\hat{y}_j^{obs} - \hat{y}_i^{mis}. \quad (3.3)$$

Erotukset lasketaan yli kaikkien niiden $\hat{y}_j^{obs}$, jotka ovat suurempia tai yhtäsuuria kuin $\hat{y}_i^{mis}$ ja mahdollisiksi luovuttajaksi valitaan ne kaksi *j*-havaintoa, joita vastaavat etäisyydet ovat kaikkein pienimmät. Kaikkiaan mahdollisia luovuttajia valitaan siis korkeintaan neljä. Välttämättä ei kuitenkaan löydy kahta sellaista $\hat{y}_j^{obs}$, joille pätisi $\hat{y}_j^{obs} \geq \hat{y}_i^{mis}$, jolloin mahdollisia luovuttajia ovat kaksi $\hat{y}_j^{obs}$, joille pätee $\hat{y}_j^{obs} < \hat{y}_i^{mis}$ ja

mahdollisesti yksi $\hat{y}_j^{obs}$, jolle pätee $\hat{y}_j^{obs} \geq \hat{y}_i^{mis}$. Samaten välttämättä ei löydy kahta $\hat{y}_j^{obs}$, joille pätisi $\hat{y}_j^{obs} < \hat{y}_i^{mis}$, jolloin toimitaan vastaavasti kuin edellä.

Yhdellä imputointikierroksella mahdollisiksi luovuttajiksi päätyneiden koehenkilöiden joukosta täytyy valita yksi varsinaiseksi luovuttajaksi. Tämä tehdään kaksivaiheisesti. Ensin katsotaan, löytyikö ennustearvoltaan sekä pienempiä että suurempia mahdollisia luovuttajia. Jos löytyi, siirrytään todennäköisyydellä 0.5 valitsemaan luovuttaja ennustearvoiltaan puuttuvaa havaintoa pienempien mahdollisten luovuttajien joukosta ja todennäköisyydellä 0.5 ennustearvoiltaan suurempien tai yhtäsuurten joukosta. Jos ei löytynyt, valitaan luovuttaja niiden mahdollisten luovuttajien joukosta, jotka on voitu määrittää. Toisessa vaiheessa valitaan mahdollisten luovuttajien valitusta osajoukosta luovuttaja laskemalla kullekin valitun osajoukon mahdolliselle luovuttajalle *i* luovuttajaksi päätymisen todennäköisyys

$$P_i = \frac{\frac{1}{d_i}}{\frac{1}{d_1} + \frac{1}{d_2}}, \qquad (3.4)$$

missä *i* = 1,2 $d_i$ on mahdollisen luovuttajan *i* ennustearvon etäisyys puuttuvan havainnon ennustearvosta. Mikäli ensimmäisessä vaiheessa valitussa osajoukossa on vain yksi mahdollinen luovuttaja, on tämän luovuttajaksi päätymisen todennäköisyys 1. Esimerkki edellisestä olisi tilanne, jossa valitaan mahdollisten luovuttajien osajoukoksi ennustearvoltaan pienemmät ja ennustearvoiltaan pienempiä olisi vain yksi.

Hot deck -moni-imputointia varten tehtiin valinnat siitä, miten mallinnus tehdään ja miten tämän mallinnuksen pohjalta saatujen ennusteiden avulla valitaan mahdolliset luovuttajat sekä miten mahdollisten luovuttajien joukosta valitaan luovuttaja. Mallinnukseksi valittiin lineaarinen regressio ja ennusteiden etäisyysmitaksi erotuksen itseisarvo. Mahdollisten luovuttajien valitsemiseksi otettiin jakauman mahdollisen vinouden huomioiva ennustearvoiltaan suurempien ja pienempien havaintojen valinta. Mahdollisten luovuttajien valitsemistodennäköisyyksiksi valittiin

$$\begin{cases} \frac{1}{2}P_i \text{ , jos löytyi ennustearvoiltaan sekä pienempiä että suurempia havaintoja} \\ P_i \text{ , muutoin.} \end{cases}$$

Edellä $P_i$ on kuten määriteltiin kaavassa (3.4).

## 3.2 Puuttuvuutta sisältävien muuttujien mallintaminen

Puuttuvia havaintoja sisältäviä muuttujia mallinnettaessa halutaan löytää yksinkertainen malli. Ollakseen hyödyllinen mallin on onnistuttava kuvaamaan riittävän suuri osa puuttuvuutta sisältävän muuttujan vaihtelusta. Tässä työssä on päädytty lineaarisiin malleihin. Näiden katsottiin olevan riittäviä, koska tarkasteltaessa malleja, jotka sisälsivät vain kahden selittäjän päävaikutukset, päästiin jo korkeisiin selitysasteisiin kaikille puuttuvuutta sisältäville muuttujille. Näin ollen tarvetta monimutkaisemmille malleille ei ollut. Nämä selitysastetarkastelut tehtiin käyttäen niiden henkilöiden havaintoarvoja, joilla ei puuttuvuutta ollut lainkaan. Niinpä yksinkertaiset lineaariset mallit näyttävät olevan riittäviä haluttujen muuttujien mallintamiseen.

Laskettaessa ennusteita havainnoille, jotka eivät puutu, otetaan ennusteet suoraan mallin antamina populaatiotason odotusarvoina vastaaville selittävien muuttujien arvoille. Toisin sanoen ennusteissa ei oteta huomioon mallin parametrien epävarmuutta eikä estimoitua yksilöidenvälistä vaihtelua, vaan ennusteet ovat samat jokaisella moni-imputoinnin *D*:llä kierroksella. Laskettaessa ennusteita puuttuville havainnoille

pyritään kaikki epävarmuus ottamaan huomioon, niin mallin parametreihin liittyvä virhe kuin populaatiossa ilmenevä vaihtelukin. Tämä toteutettiin van Buurenin (2012, s. 58) esittämän algoritmin mukaisesti seuraavasti:

1. Lasketaan $S = X^{obs\prime}X^{obs}$, missä $X^{obs}$ on matriisi, joka sisältää vain täydellisten havaintorivien selittävien muuttujien arvot mukaan lukien vakiota vastaavan ykkösten sarakkeen.
2. Lasketaan $V = (S + diag(S)\kappa)^{-1}$, käytetty $\kappa$ on 0.0001 ja $diag(S)$ on matriisi, jossa diagonaalillla on matriisin *S* diagonaalialkiot ja muut alkiot ovat nollia.
3. Lasketaan regressiokertoimet $\dot{\beta} = VX^{obs\prime}y^{obs}$, missä $y^{obs}$ on täydellisten havaintorivien selitettävän muuttujan arvot sisältävä vektori.
4. Arvotaan satunnaisluku $\dot{g} \sim \chi^2_v$, missä $v$ on täydellisten havaintorivien määrä miinus mallissa olevien parametrien määrä.
5. Lasketaan $\dot{\sigma}^2 = \frac{(y^{obs} - X^{obs}\hat{\beta})'(y^{obs} - X^{obs}\hat{\beta})}{\dot{g}}$.
6. Arvotaan vektoriin $z_1$ riippumattomia N(0,1) jakautuneita satunnaislukuja, joita on yhtä monta kuin regressiokertoimia mallissa.
7. Lasketaan $V^{1/2}$ Cholesky hajotelmana, jolloin siis $V = V^{\frac{1}{2}}\left(V^{\frac{1}{2}}\right)'$.
8. Lasketaan $\dot{\beta} = \hat{\beta} + \dot{\sigma}V^{\frac{1}{2}}z_1$.
9. Arvotaan vektoriin $z_2$ riippumattomia N(0,1) jakautuneita satunnaislukuja, joita on yhtä monta kuin puuttuvia havaintoja.
10. Lasketaan puuttuvien havaintojen ennusteet $\dot{y} = X^{mis}\dot{\beta} + z_2\dot{\sigma}$, missä $X^{mis}$ on selittävien muuttujien matriisi, joka sisältää vain puuttuvia havaintoja vastaavat selittävien muuttujien arvot mukaan lukien vakiota vastaavan ykkösten sarakkeen.

Kohta kahdeksan on van Buurenin (2012) esityksessä $\dot{\beta} = \hat{\beta} + \dot{\sigma}V^{\frac{1}{2}}z_1$. Tämä on ilmeinen virhe, sillä $z_1$ on *p* mittainen vektori ja $V^{\frac{1}{2}}$ on *p* ulotteinen neliömatriisi. Edellisessä *p* on imputointimallin parametrien lukumäärä, joka pääsääntöisesti on enemmän kuin yksi. Tämä algoritmi toistetaan kullakin moni-imputoinnin *D*:stä kierroksesta, jolloin eri kierroksilla saadut puuttuvia havaintoja vastaavat ennusteet poikkeavat toisistaan. Siihen, kuinka paljon ennusteet samalle puuttuvalle havainnolle poikkeavat toisistaan, vaikuttaa mallin tuottamien ennusteiden poikkeama havaituista arvoista sekä mallissa käytettävien parametrien määrä. Mitä pienempi on poikkeama mallin antamien ennusteiden ja havaittujen arvojen välillä ja toisaalta mitä pienemmällä määrällä parametreja tämä saadaan, sitä vähemmän ennusteet tietylle puuttuvalle havainnolle keskimäärin poikkeavat toisistaan.

Kohdissa 1-3 lasketaan lineaarisen regressiomallin regressiokertoimet tavanomaiseen tapaan, siinä lasketut suureet ovat samat jokaisella moni-imputoinnin kierroksella. Kohdissa 4-8 pyritään ottamaan huomioon mallin parametreihin liittyvä epävarmuus. Kohdissa 9-10 pyritään ottamaan huomioon populaation sisäinen vaihtelu. Satunnaislukujen käyttäminen kohdissa neljä ja yhdeksän mahdollistaa erilaisten tulosten saamisen moni-imputoinnin eri kierroksilla. Viimeistä kohtaa ei van Buurenin (2012) ohjeistuksessa tehdä hot deck -moni-imputoinnissa vaan pelkästään sellaisessa mallipohjaisessa moni-imputoinnissa, jossa imputoidut arvot ovat mallin mukaisesta jakaumasta simuloituja arvoja. Toteutettaessa hot deck -moni-imputointi sekä ilman viimeistä kohtaa että sen kanssa havaittiin alustavissa simulointikokeissa, että viimeisen osan lisääminen paransi tuloksia tässä yhteydessä. Tulos saattaa johtua tarkasteltavassa tilanteessa olevasta hyvin pienestä otoskoosta ja samanaikaisesta hyvin pieneksi estimoituvasta mallin parametreihin liittyvästä epävarmuudesta. Mallin parametreihin liittyvän epävarmuuden $\dot{\sigma}^2$ estimoituessa pieneksi saadaan eri imputointikierroksilla hyvin samanlaiset

regressiokertoimet $\dot{\beta}$ ja estimoitu populaatiovaihtelu $z_2\dot{\sigma}$ jää pieneksi, tällöin eri kierrosten ennusteet $X^{mis}\dot{\beta} + z_2\dot{\sigma}$ ovat hyvin lähellä toisiaan. Saataessa samanlaisia ennusteita eri kierroksilla myös mahdollisten luovuttajien joukko ja niiden luovuttajiksi päätymisen todennäköisyydet pysyvät samana, jolloin ei saada luoduksi riittävää vaihtelua moni-imputointikierrosten välille.

Palataan luvun 3.1 esimerkkiin ja tehdään mallinnus käyttäen molempia saatavilla olevia matriisissa *X* olevia muuttujia, joissa ei ole puuttuvuutta. Tällöin

$$X^{obs} = \begin{bmatrix} 1 & 9.6 & -5.7 \\ 1 & 12.8 & -9.6 \\ 1 & 7.5 & -3.2 \\ 1 & 4.7 & -2.8 \end{bmatrix}, X^{mis} = \begin{bmatrix} 1 & 6.7 & -4.0 \\ 1 & 7.1 & -3.4 \end{bmatrix} \text{ ja } y^{obs} = \begin{bmatrix} 3 \\ 4 \\ 2.5 \\ 1.5 \end{bmatrix} \text{ jolloin saadaan } \hat{\beta} = \begin{bmatrix} 0.015 \\ 0.347 \\ 0.050 \end{bmatrix}$$

ja $v = 1$. Saatu estimaatti $\hat{\beta}$ on sama kaikilla moni-imputoinnin kierroksilla, sen avulla saadaan ennusteet havaituille havainnoille:

$$\hat{y}^{obs} = \begin{bmatrix} 1 & 9.6 & -5.7 \\ 1 & 12.8 & -9.6 \\ 1 & 7.5 & -3.2 \\ 1 & 4.7 & -2.8 \end{bmatrix} * \begin{bmatrix} -0.198 \\ 0.395 \\ 0.108 \end{bmatrix} = \begin{bmatrix} 3.06 \\ 3.98 \\ 2.46 \\ 1.51 \end{bmatrix}.$$

Lasketaan esimerkin datalla suurin osa yhdestä moni-imputointikierroksesta. Generoidaan satunnaisluku $\dot{g} = 0.6$ ja satunnaisluvut $z_1 = [-0.464, 0.781, -0.690]^T$ jolloin saadaan

$$\dot{\sigma}^2 = \frac{\left(y^{obs} - \hat{y}^{obs}\right)^T \left(y^{obs} - \hat{y}^{obs}\right)}{0.6} \text{ ja } \dot{\beta} = \begin{bmatrix} 0.015 \\ 0.347 \\ 0.050 \end{bmatrix} + \dot{\sigma} * V^{1/2} * \begin{bmatrix} -0.464 \\ 0.781 \\ -0.690 \end{bmatrix} \approx \begin{bmatrix} -0.089 \\ 0.339 \\ 0.044 \end{bmatrix}.$$

Tämän jälkeen otetaan satunnaisluvut $z_2 = [-0.675, -0.370]^T$ jolloin saadaan ennusteet puuttuville havainnoille tälle kierrokselle

$$\hat{y}^{mis} = \begin{bmatrix} 1 & 6.7 & -4.0 \\ 1 & 7.1 & -3.4 \end{bmatrix} * \begin{bmatrix} -0.089 \\ 0.339 \\ 0.044 \end{bmatrix} + \begin{bmatrix} -0.675 \\ -0.370 \end{bmatrix} * \dot{\sigma} = \begin{bmatrix} 1.94 \\ 2.13 \end{bmatrix}.$$

Nyt $\hat{y}_j^{ob} < \hat{y}_i^{mis}$ pätee vain alkiolle $\hat{y}_4^{obs}$, jolloin se päätyy mahdolliseksi luovuttajaksi molemmille puuttuville havainnoille. $\hat{y}_j^{obs} \geq \hat{y}_i^{mis}$ pätee lopuille $\hat{y}_j^{obs}$ missä $j = 1,2,3$, näistä kaksi lähintä on molemmille puuttuville havainnoille 1 ja 3. Lasketaan mahdollisten luovuttajien todennäköisyydet päätyä luovuttajaksi ensimmäiselle puuttuvalle havainnolle. Tätä varten tarvitaan etäisyydet sekä mahdolliselle luovuttajalle 1 että 3, jotka saadaan seuraavasti $d_1 = |1.94 - 3.06| = 1.12$ ja $d_3 = |1.94 - 2.46| = 0.52$. Tällöin $P_1 = \frac{1/1.12}{\frac{1}{1.12}+\frac{1}{0.52}} \approx 0.317$ ja $P_3 = \frac{1/0.52}{\frac{1}{1.12}+\frac{1}{0.52}} \approx 0.683$. Edellisestä saadaan laskettua luovuttajaksi päätymisen todennäköisyydet $0.5 * P_1 \approx 0.159$, $0.5 * P_3 \approx 0.342$ ja ainoalle $\hat{y}_j^{obs}$, jolle pätee $\hat{y}_j^{obs} < \hat{y}_i^{mis}$, todennäköisyys on suoraan 0.5. Samaan tapaan saataisiin laskettua todennäköisyydet myös toisen puuttuvan havainnon mahdollisten luovuttajien valitsemiseksi ja näiden todennäköisyyksien mukaisesti valiten saataisiin $Y^{mis,(1)}$ ja ensimmäinen moni-imputoinnin kierros suoritettua.

Tämän luvun yleisessä esityksessä mallipohjaisesta moni-imputoinnista tehtiin vain kaksi tähän työhön liittyvää erityistä valintaa. Ensimmäiseksi rajoitettiin puuttuvuutta sisältävän muuttujan mallinnus tavanomaiseen lineaariseen regressioon. Tämä tehtiin, koska valittu menetelmä oli riittävä ja toisaalta laskennallisesti tehokas. Toinen valinta oli ottaa ennusteissa huomioon estimoitujen regressiokertoimien lisäksi populaatiovaihtelu, mitä ei van Buurenin (2012) esityksessä tehdä. Jälkimmäinen valinta tehtiin alustavien simulointikokeiden tulosten perusteella, joissa havaittiin että eri imputointikierroksilla saatavien luovuttajien joukko pysyi liian suurelta osin samana ilman valittua lisäystä.

## 3.3 Bayes-bootstrap hot deck

Vaihtoehtoinen tapa tehdä moni-imputointia on Bayes-bootstrap (BB) hot deck. Käytettäessä BB hot deck -menetelmää puuttuvuutta sisältävää muuttujaa ei mallinneta lainkaan, mikä yksinkertaistaa moni-imputointiprosessia huomattavasti, silti säilyttäen moni-imputoinnin tuoman arvion puuttuvuuden tuomasta epävarmuudesta. Toisaalta mallintamattomuus johtaa myös siihen, että ei ole mahdollista hyödyntää muiden muuttujien sisältämää informaatiota puuttuvuutta sisältävästä muuttujasta, mikä mahdollisesti toisi tarkempia tuloksia tarkasteltavasta ilmiöstä. Mikäli muuttujan puuttuvuus riippuu jostain havaitusta muuttujasta, ei tätä pystytä huomioimaan, jolloin riippuvuuden mahdollisesti tuoma harha säilyy edelleen imputoidussa aineistossa. Jotta BB-moni-imputointi toimisi harhattomasti, vaaditaan MCAR-oletus. Tässä BB-moni-imputoinnin esityksessä ei ole tehty valintoja liittyen tämän työn erityispiirteisiin, vaan edetään yleisen esityksen mukaisesti.

BB ja tavallinen bootstrap eroavat siinä, että kullakin BB:n kierroksella mahdollisten eri arvojen $x_i$ valintatodennäköisyydet $g_i$ eivät ole yhtä suuret vaan ne lasketaan erikseen jokaisella kierroksella kullekin $x_i$. Tällöin kahden eri arvon valintatodennäköisyydet ovat erisuuret ja saman arvon valintatodennäköisyydet eri kierroksilla ovat erisuuret. Tämä toteutetaan Rubinin (1981) esittelemänä seuraavasti:

1. Olkoon mahdollisten arvojen joukko $x_1, \dots, x_n$
2. Otetaan *n* – 1 satunnaislukua $u_1, \dots, u_{n-1}$ väliltä [0, 1].
3. Järjestetään satunnaisluvut suuruusjärjestykseen ja lisätään joukkoon ensimmäiseksi $u_0 = 0$ ja viimeiseksi $u_n = 1$.
4. Lasketaan saadun lukujonon peräkkäisten lukujen väliset erotukset $g_i = u_i - u_{i-1}$, missä i = 1, …, n.
5. Käytetään valintatodennäköisyytenä kullekin $x_i$ lukua $g_i$.

Tätä voidaan soveltaa moni-imputointiin valitsemalla mahdollisiksi arvoiksi $x_i$ havaitut arvot $y^{obs}$ ja korvaamalla havainnot $y^{mis}$ BB-otoksella Rubinin ja Schenkerin (1986) osoittamalla tavalla seuraavasti. Kullekin havainnolle mahdollisten arvojen joukossa $y^{obs}$ määritetään luovuttajaksi päätymisen todennäköisyys edellä määriteltynä valintatodennäköisyytenä. Tämän jälkeen otetaan joukon $y^{mis}$ kokoinen BB-otos joukosta $y^{obs}$ ja käytetään tätä otosta imputoituina arvoina joukolle $y^{mis}$. Tällä tavoin edeten toistetaan kohdat 1-5 jokaisella moni-imputoinnin kierroksella. Vaikkakin Rubin ja Schenker puhuvat BB-moni-imputoinnista vain diskreetin muuttujan tapauksesta, muiden muassa Andridge ja Little (2010) esittelevät sen moni-imputointimenetelmänä yleisessä tapauksessa.

Palataan jälleen kohdassa 3.1 esiteltyyn esimerkkiin. BB-moni-imputoinnissa voidaan matriisi *X* unohtaa kokonaan, sillä minkäänlaista mallinnusta ei tehdä. Nyt $y^{obs} = [3, 4, 2.5, 1.5]^T$, joukko $y^{mis}$ on kahden alkion kokoinen ja $n = 4$. Toteutetaan yksi kierros moni-imputointia. Otetaan $n - 1 = 3$ satunnaislukua väliltä 0 – 1 ja laitetaan nämä suuruusjärjestykseen. Lisätään saatuun joukkoon ensimmäiseksi 0 sekä viimeiseksi 1, saadaan $u = [0, 0.182, 0.369, 0.691, 1]^T$. Lasketaan peräkkäisten lukujen erotukset $g_i$ saadaan $g = [0.182, 0.187, 0.322, 0.309]^T$. Seuraavaksi otetaan lukuja $g_i$ todennäköisyyksinä käyttäen kahden alkion kokoinen otos takaisinpalauttaen joukosta $y^{obs}$ ja näin on saatu $Y^{mis,(1)}$ ja yksi kierros BB-moni-imputointia suoritettua. Koska kullakin kierroksella otetaan uudet satunnaisluvut $u_i$, niin eri imputointikierroksilla saadut todennäköisyydet päätyä luovuttajaksi vaihtelevat.

## 3.4 Moni-imputointitulosten yhdistäminen

Useissa moni-imputointimenetelmissä pyritään mallintamaan muuttujaa, jossa puuttuvuutta ilmenee. Tämän jälkeen puuttuvat havainnot korvataan joko suoraan estimoidusta mallista tai saman muuttujan havaitulla arvolla. Tässä työssä on valittu, että korvaavat arvot ovat havaittuja arvoja. Tämä toistetaan *D* kertaa, jolloin saadaan *D* kappaletta hieman toisistaan poikkeavia alkuperäisen kokoisia aineistoja, joissa ei enää ole puuttuvuutta. Haluttu alkuperäinen analyysi tehdään kullekin näistä aineistoista erikseen. Lopuksi eri aineistoista saatujen analyysien tulokset yhdistetään ns. Rubinin kaavoilla. (Little & Rubin, 2002.) Seuraava esitys Rubinin kaavoista koskee yleistä tilannetta, siinä ei ole tehty mitään valintoja koskien tämän työn käsittelemää moni-imputoinnin sovellusta ja sen erityispiirteitä.

Olkoon $\theta$ haluttu suure, jota halutaan estimoida. Tällöin Rubinin kaava kaikkien *D* kierroksen analyysien tulosten yhdistämiseksi on

$$\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d\,, \qquad (3.5)$$

missä $\hat{\theta}_d$ on yhdellä imputointikierroksella $d$ saatu estimaatti. Tulokset yhdistävän estimaatin varianssi rakentuu kahdesta osasta. Ensiksi estimoidaan imputointikierrosten keskimääräinen sisäinen varianssi

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d\,, \qquad (3.6)$$

missä $W_d$ on yhdellä imputointikierroksella laskettu estimaatin varianssi. $W_d$:n tarkka muoto on tilannekohtaista riippuen estimaatista $\hat{\theta}_d$. Toiseksi estimoidaan imputointikierrosten välisen vaihtelun komponentti

$$B_D = \frac{D+1}{D(D-1)}\sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta}_D)^2\,, \qquad (3.7)$$

mikä voidaan edelleen hajottaa kahteen osaan. Yleiseen tapaan laskettavaan varianssiin yli saatujen estimaattien ja sen painokertoimeen $\frac{D+1}{D}$, jolla pyritään huomioimaan imputointikierrosten rajallisuus. Edellisten summana saadaan estimaatin $\bar{\theta}_D$ koko varianssi

$$T_D = \bar{W}_D + B_D. \qquad (3.8)$$

$B_D$ kuvaa nimenomaan puuttuvuudesta tulevaa informaation katoa ja haluttaessa puuttuvan tiedon aiheuttamaa informaation menetystä voidaankin arvioida osuutena

$$\hat{\gamma}_D = \frac{(1+1/D)B_D}{T_D}. \qquad (3.9)$$

(Little & Rubin 2002 s. 86 – 87). On ilmeistä, että tätä informaation menetystä kuvaavaa estimaattia ei saada täydellisten havaintoriven analyysilla tai imputoimalla vain yksi havainto jokaiselle puuttuvalle havainnolle.

## 3.5 Jackknife-sovellus hot deck -moni-imputointiin

Toteutettaessa moni-imputointi luvuissa 3.1 – 3.4 esitetyllä tavalla havaittiin, että moni-imputoinnin antama valitun estimaattorin varianssin estimaattori $T_D$ oli harhainen antaen jopa kolminkertaisia arvoja simuloinnissa havaittuun varianssiin nähden. Tämä seikka teki menetelmästä täydellisten havaintorivien analyysiin verrattuna tältä osin vaillinaisen. Moni-imputointiestimaattorilla havaittu varianssi oli selvästi ja säännönmukaisesti pienempi kuin täydellisten havaintorivien analyysilla, jolloin tässä työssä valituilla hajontamittareilla katsoen moni-imputointi olisi parempi, mikäli varianssin estimaattorista saataisiin vähemmän harhainen. Tästä johtuen päätettiin tehdä moni-imputointiin jackknife-sovellus, jolloin estimaattorin varianssi voidaan estimoida jackknife-menetelmän mukaisesti.

Jackknife-menetelmässä jaetaan havaintoaineisto *k*:hon osaan ja lasketaan haluttu estimaatti *k* kertaa. Kullakin *k*:sta kierroksesta *k*:s osa havaintoaineistosta jätetään pois laskettaessa halutun suureen $\theta$ estimaattia. Tällöin kaikki koehenkilöt ovat olleet poissa laskennasta täsmälleen yhden kerran. Lopuksi nämä eri tulokset yhdistetään ottamalla niiden keskiarvo ja laskemalla yhdistetyn estimaatin varianssi summana *k* kappaleen estimaattien neliöpoikkeamina keskiarvostaan. Tässä työssä *k* on koehenkilöiden lukumäärä *n*, jolloin kullakin jackknife-kierroksella jätetään vain yksi koehenkilö pois. Efronin ja Steinin (1981) merkinnöin:

1. Olkoon $X_1, X_2, \dots, X_n$ riippumattomia samoinjakautuneita muuttujia ja $S(X_1, X_2, \dots, X_n)$ kiinnostuksen kohteena olevan suureen estimaattori.
2. Olkoon $S_i \equiv S(X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ S:n saama arvo kun $X_i$ on poistettu otoksesta.
3. Lasketaan $S_i$ kaikille *i* = 1, 2, …, *n*.
4. Saadaan yhdistetty estimaatti $S_{(.)}$ ja sen varianssi $\widehat{VAR}\, S_{(.)}$

$$S_{(.)} \equiv \sum_{i=1}^{n} \frac{S_i}{n}, \qquad (3.10)$$

$$\widehat{VAR}\, S_{(.)} \equiv \frac{n-1}{n} \sum_{i=1}^{n} \left[S_i - S_{(.)}\right]^2, \qquad (3.11)$$

Hot deck -moni-imputointiin sovellettuna jackknife tulkitaan siten, että $y^{obs}$ ovat havainnot, joista vuorotellen jätetään yksi pois. Estimaattori *S* on moni-imputoinnin kautta saatu parametrin $\theta$ estimaattori $\bar{\theta}_D$, jolloin koko moni-imputointiprosessi käydään läpi kullakin jackknife-sovelluksen kierroksella. Tämä moni-imputoinnin päälle tehty jackknife-sovellus tehdään kaikilla valituilla puuttuvan tiedon käsittelymenetelmillä. Jackknife-menetelmä moni-imputointiin voidaan siis ilmaista kaavojen (3.5), (3.10) ja (3.11) avulla seuraavasti:

1. Aineisto $y$ on *N* riippumattoman samoinjakautuneen muuttujan otos, $\theta$ kiinnostuksen kohteena oleva suure, $S(y)$ kiinnostuksen kohteena olevan suureen estimaattori ja $y = (y^{obs}, y^{mis})$, missä joukon $y_{obs}$ koehenkilöiden lukumäärä on *n*.
2. $S_{(.)}$ saadaan moni-imputoinnin tapauksessa muotoon

$$S_{(.)} \equiv \sum_{i=1}^{n} \frac{S_i}{n} = \sum_{i=1}^{n} \frac{\bar{\theta}_D{}^{(i)}}{n} = \sum_{i=1}^{n} \frac{1}{n} \left[\frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d\right]^{(i)}, \qquad (3.12)$$

   missä $\hat{\theta}_d$ lasketaan käyttäen (*n* - 1):tä koehenkilöä joukosta $y^{obs}$ ja kaikkia joukosta $y^{mis}$. Edellisessä *i* kertoo jackknife-sovelluksen kierroksen ja samalla havainnon $y^{obs}{}_i$, joka jätetään kyseisellä kierroksella pois.

   Täydellisten havaintorivien tapauksessa päädytään tavanomaiseen jackknife-menetelmään joukon $y^{obs}$ sisällä

$$S_{(.)} \equiv \sum_{i=1}^{n} \frac{S_i}{n} = \sum_{i=1}^{n} \frac{\hat{\theta}^{(i)}}{n}, \qquad (3.13).$$

Esitelty jackknife-menetelmän sovellus moni-imputointiin on tässä työssä rakennettu erityissovellus. Samoin joukon $y^{obs}$ valinta joukoksi, josta jätetään aina yksi pois, on tässä työssä tehty valinta ja erityispiirre.

# 4 Hot deck -moni-imputoinnin soveltaminen ERP-dataan

Varsinainen mielenkiinnon kohde on moni-imputoinnin soveltaminen ERP-dataan, jossa tarkastellaan annetun ärsykkeen tuottamaa välitöntä aivojen sähköistä reaktiota tilanteessa, jossa reaktion voimakkuutta voidaan kuvata yksinkertaistetusti parilla voimakkuusluvulla yhtä aivojen aktivaatiota mittaavaa elektrodia kohden. Näitä reaktion voimakkuuslukuja kutsutaan jatkossa piirteiksi. Esitetyt ärsykkeet ovat kuvia eri tunnetiloja ilmentävistä kasvoista, näitä kutsutaan jatkossa ilmeiksi. Tarkasteltavaksi suureeksi on valittu yhden koeasetelman eri ilmeiden aiheuttaman aktivaation ero. Mielenkiinnon kohteena olevat estimaattorit ovat erotuksen odotusarvon estimaattorina erotuksien keskiarvo

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_{i,ilme1} - y_{i,ilme2}\,, \qquad (4.1)$$

missä *i* viittaa koehenkilöön ja *n* koehenkilöiden lukumäärään, sekä tämän keskivirhe

$$se(\hat{\theta}) = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \left(\hat{\theta} - \left(y_{i,ilme1} - y_{i,ilme2}\right)\right)^2}, \qquad (4.2)$$

missä $y_{i,ilme1}$ on piirteen arvo koehenkilöllä *i,* kun esitetty ilme on ollut *ilme*1, $y_{i,lme2}$ on saman piirteen arvo koehenkilöllä *i*, kun esitetty ilme on ollut *ilme*2 ja $\hat{\theta}$ on kuten edellä on määritelty. Otetaan edellisestä esimerkki: tarkastellaan koehenkilöä 1 ja ilmeitä *ilme*1=iloinen sekä *ilme*2=neutraali, tällöin laskettu erotus $y_{1,ilme1} - y_{1,lme2}$ on annetun piirteen saamien arvojen erotus ilmeiden iloinen ja neutraali välillä.

Piirteet kuvastavat annetun ärsykkeen aiheuttamaa keskimääräistä aktivaation muutosta, jolloin laskettava erotus kertoo, miten paljon eri ilmeiden aiheuttamat keskimääräiset aktivaatiomuutokset eroavat toisistaan. Valittaessa tämä erotus tarkastelun kohteeksi nähdään, miten hyvin tämä ero pystytään havaitsemaan toisaalta täydellisten havaintorivien analyysilla ja toisaalta moni-imputoinnilla. Erotukset ovat elektrodi- ja piirrekohtaisia ja ne lasketaan erikseen kullekin koehenkilölle. Käsiteltävät erotukset lasketaan mittauksista, jotka ovat samalta koehenkilöltä. Kunkin yksittäisen annettua elektrodia koskevan piirteen kohdalla havaitaan kaksi riippuvaa otosta, yksi kummastakin ilmeestä. Erotuksia lasketaan vain koeasetelmista, joissa näytetään täsmälleen kahta eri ilmettä.

Moni-imputointimenetelmistä on valittu nimenomaan hot deck, koska valittaessa puuttuvalle havainnolle luovuttaja saadaan sille imputoitua koko aikasarja. Näin ollen ei tarvitse mallintaa alkuperäisten aikasarjojen autokorrelaatiorakenteita, vaikka haluttaisiin hyödyntää imputoinnin tuloksia muutenkin kuin imputointimallissa käytettyjen piirteiden osalta. Tässä työssä ei tarkastella tällä tavoin saadun aikasarja-aineiston käsittelyä. Tämä jääkin kenties merkittävimmäksi jatkokehityssuunnaksi.

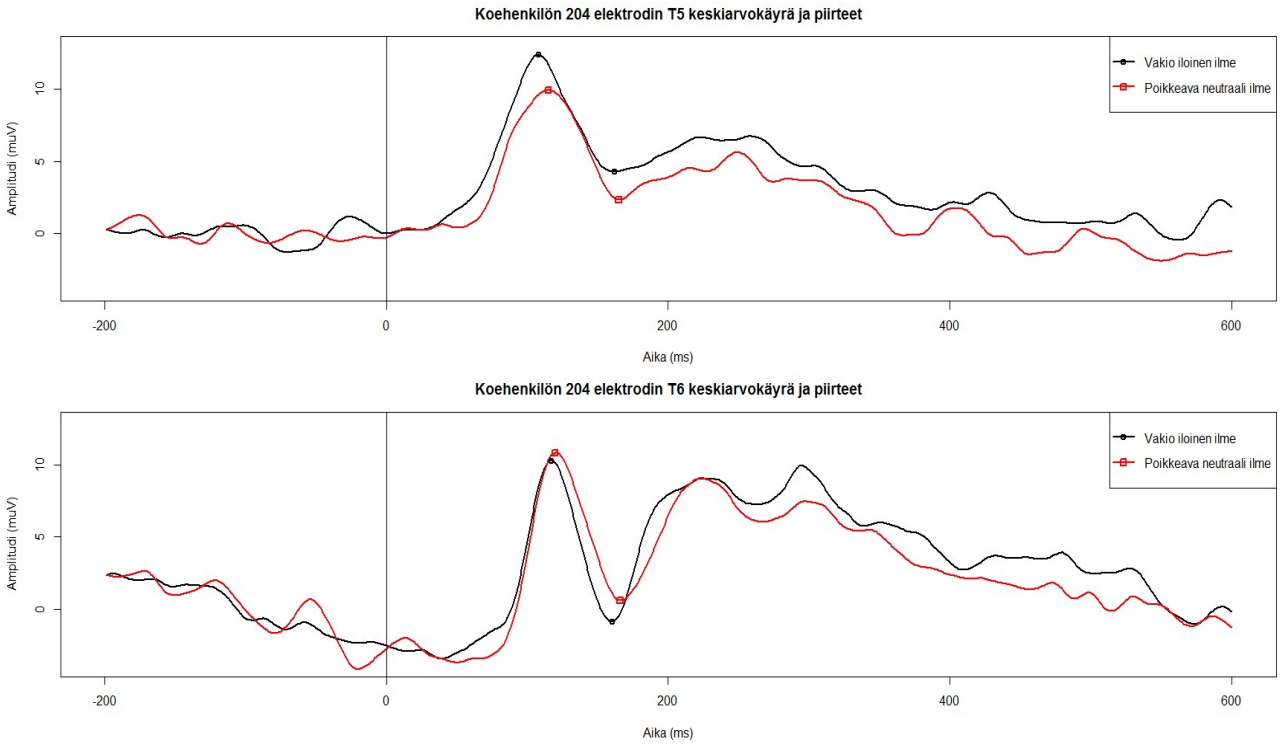
## 4.1 Aineisto

Aineisto on saatu Jyväskylän yliopiston psykologian laitokselta tutkimuksesta, jossa tarkasteltiin eri ilmeiden herättämien reaktioiden voimakkuutta EEG-mittauksissa. Kyseisessä tutkimuksessa oltiin erityisen kiinnostuneita mahdollisista eroista aktivaatioissa kahden ryhmän välillä, joista toisessa olevat koehenkilöt olivat masennusdiagnoosin saaneita ja toisessa olevista ei kellään ollut diagnoosia. Tätä kysymystä ei tarkastella lainkaan, vaan rajoitutaan tarkastelemaan pelkästään kontrolliryhmän asemassa olevaa diagnosoimattomien ryhmää. Tämä rajaus tehtiin puhtaasti tilanteen yksinkertaistamiseksi, jotta voidaan keskittyä menetelmien vertailuun havaintoaineiston muodostavan ryhmän ollessa puuttuvuuteen tiettävästi liittyvien seikkojen osalta mahdollisimman yhdenmukainen.

Kokeen aikana osallistuja istui mukavassa tuolissa himmeästi valaistussa huoneessa. Kukin osallistuja oli mittausten ajan huoneessa yksin ja heitä tarkkailtiin videokameran avulla. Osallistujia ohjeistettiin kuuntelemaan kuunnelmaa, jota toistettiin osallistujan pään yläpuolella olevasta kaiuttimesta normaalin puheäänen voimakkuudella. Visuaaliset ärsykkeet esitettiin tietokoneen näytöllä, jonka keskellä oli rasti. Osallistujia ohjeistettiin kohdistamaan katseensa rastiin. Osallistujille kerrottiin, että visuaaliset ärsykkeet tulisivat olemaan kasvoja. Osallistujia kehotettiin keskittymään kuunnelmaan ja olemaan huomioimatta kasvoja. Kasvot olivat neljän eri mallin, kahden naisen ja kahden miehen. Kasvot esitettiin harmaasävykuvina.

Alkuperäisestä tutkimuksesta on valittu mittaukset kolmesta eri koeasetelmasta. Näistä kaksi ovat niin sanottuja oddball-asetelmia, joissa näytetään kahta ilmettä vakioksi kutsuttua 400 kertaa ja poikkeavaksi kutsuttua 80 kertaa satunnaisissa väleissä siten, että kahden poikkeavan ärsykkeen välissä on vähintään kaksi vakioärsykettä. Näistä oddball-asetelmista ensimmäisessä vakio oli iloinen ilme ja toisessa vakio oli surullinen ilme, molemmissa poikkeavana oli neutraali ilme. Juuri näissä asetelmissa ilmenevää puuttuvuutta käsitellään, jolloin nämä toimivat selitettävinä tilanteina. Kolmannessa valitussa asetelmassa näytettiin kaikkia seitsemää perusilmettä (neutraali, iloinen, surullinen, vihainen, inho, pelko ja yllättynyt) satunnaisesti, mutta jokaista ilmettä 80 kertaa siten, että samaa ilmettä ei tullut koskaan kahta kertaa peräkkäin. Tästä seitsemän ilmeen asetelmasta on hyödynnetty vain tietoja iloisen, surullisen ja neutraalin ilmeen mittauksista. Kolmannen asetelman havainnoilla pyritään selittämään havaintoja puuttuvuutta sisältävistä koeasetelmista. Kyseinen valinta tehtiin kahdesta syystä: ensinnäkin asetelmassa ei ole lainkaan puuttuvuutta, joten puuttuvuuden rakenne valitussa osa-aineistossa on hyvin yksinkertainen. Toiseksi, koska asetelma sisältää havainnot kaikista eri tunnetiloista, sitä voidaan käyttää selittäjänä kaikille muille asetelmille. Metodiikasta haluttiin tällä tavoin tehdä mahdollisuuksien mukaan yleistettävä.

Jokaista ilmettä on varsinaisessa mittaustilanteessa näytetty kullekin koehenkilölle lukuisia kertoja, kussakin eri koeasetelmassa. Näin saadaan juuri esitettyyn ilmeeseen annetussa tilanteessa liittyvä keskimääräinen aktivaatio erotetuksi muusta aivojen aktivaatiosta. Mittauksista on otettu millisekuntikohtainen keskiarvo kullekin koehenkilölle, jokaiselle ilmeelle ja jokaiselle elektrodille. Näin saadaan jokaisen henkilön mittausten perusteella laskettu keskiarvokäyrä koeasetelman eri ilmeille kaikille elektrodeille. Tutkimuksessa elektrodeja oli 128, mutta näistä tarkastellaan vain kahta, T5:tä ja T6:ta. Nämä kaksi elektrodia kuvaavat eri aivopuoliskoja. Aivojen eri puolet haluttiin ottaa mukaan, koska kasvojen tunnistuksen katsotaan painottuvan tyypillisesti oikealle puolelle aivoja (Purves, Augustine, Fitzpatrick, Hall, LeMantia & White, 2012). Näin ollen otettaessa mittaukset eri puolilta aivoja saadaan oletettavasti laadullisesti erilaiset tilanteet, joista ensimmäisessä nähdään voimakkaita eroja eri ilmeiden välillä ja toisessa korkeintaan vähäisiä eroja. Loput elektrodit jätettiin pois tarkasteluista puhtaasti tilanteen yksinkertaistamisen vuoksi, jotta varsinaisessa analyysissa voidaan keskittyä nimenomaan metodien toimivuuden tarkasteluun.




**Kuva 1 Esimerkkikoehenkilön mittaukset.**

**Piirre P1 on ensimmäinen merkitty kohta kussakin käyrässä ja piirre N170 on jälkimmäinen merkitty kohta kussakin käyrässä.**

Yksittäinen mittaus, jossa näytetään koehenkilölle ilme kestää 800 millisekuntia. Ensimmäiset 200 millisekuntia ovat ennen kuin kasvot ilmestyvät näkyviin. Ajanhetkeksi nolla määritellään hetki, jolloin ärsyke tulee näkyviin. Tutkimuksessa ollaan siis kiinnostuneita refleksinomaisesta hyvin nopeasti tapahtuvasta reaktiosta, joka on nähtävissä jo alle sekunnissa. Tämän reaktion voimakkuutta voidaan kuvata kullekin koehenkilölle ottamalla maksimiarvo aikaväliltä 70 - 130 millisekuntia ärsykkeen esittämisen jälkeen. Tämä on piirre P1, ja minimiarvo aikaväliltä 130 - 210 millisekuntia ärsykkeen esittämisen jälkeen on piirre N170. Tarkasteluissa käytetään kyseisiä reaktion voimakkuutta kuvaavia piirteitä alkuperäisten keskiarvokäyrien sijaan, sillä nämä ovat aivojen sähköisessä aktivaatiossa tapahtuvan herätevasteen yleisesti käytettyjä mittareita. Herätevasteella tarkoitetaan esitetyn ärsykkeen aiheuttamaa muutosta aivojen sähköisessä aktivaatiossa. Piirteet on laskettu keskiarvokäyristä. Piirre P1 on yleisesti näköärsykkeiden yhteydessä esiintyvä vaste, joka ei niinkään riipu siitä, mikä on esitetty ärsyke. N170 on nimenomaan kasvojen tuottamaa aktivaatiota kuvaava piirre (Bentin, Allison, Puce & Perez, 1996). Kuvassa 1 on esitetty esimerkkikoehenkilön 204 keskiarvokäyrät kahdelle oddball-koeasetelmalle. Kuvassa on merkittynä myös ääriarvot P1 ja N170, P1 on kunkin käyrän ensimmäinen merkitty kohta ja N170 on kunkin käyrän jälkimmäinen merkitty kohta.

Aineisto saatiin Jyväskylän yliopiston psykologian laitokselta keskiarvoistettuina aikasarjoina. Saaduissa aikasarjoissa on yksi mittausarvo per millisekunti, ne alkavat 200 millisekuntia ennen ärsykkeen esittämistä ja päättyvät 600 millisekuntia ärsykkeen esittämisen jälkeen. Aikasarjoja on yksi kunkin koehenkilön kummallekin tarkastellulle elektrodille koskien kunkin koeasetelman kutakin ärsykettä. Näin saadaan seitsemän ilmeen tilanteessa yhdelle koehenkilölle kaksi elektrodia kertaa seitsemän ärsykettä eli 14 aikasarjaa. Näistä aikasarjoista piirteiden poimiminen on tehty R-koodilla. Piirteiden poiminnasta on seitsemän ilmeen tilanteesta R-koodi liitteessä H, muille koeasetelmille poiminta on tehty vastaavasti.

Tarkasteltaessa kahta tunnetilaa kahdella elektrodilla ja näiden elektrodien tuloksista kahta piirrettä saadaan siis yhteensä kahdeksan muuttujaa tehdyistä mittauksista per koeasetelma. Koeasetelmia oli alun

perin kaikkiaan kuusi. Näistä kuudesta neljä oli oddball-asetelmia, joissa aina toisena esitettävänä ilmeenä oli neutraali ilme ja toisena iloinen tai surullinen. Kahdessa muussa asetelmassa esitettiin kaikkia seitsemää perustunnetta ilmentäviä kasvoja, joko kasvot oikein tai väärin päin. Näin ollen tässä tarkastellut kolme koeasetelmaa sisältävät molemman tyyppisiä koeasetelmia, joita alkuperäisessä tutkimuksessa oli.

Kaikkiaan valittuun ryhmään kuului 19 koehenkilöä. Tästä ryhmästä kahdeksallatoista ei ollut puuttuvuutta yhdessäkään valitussa koeasetelmassa ja analyysi perustuukin näiden kahdeksantoista koehenkilön mittauksiin. Liitteesä B on R-koodi koehenkilöiden valitsemiseksi. Tämä rajaus haluttiin tehdä, jotta voidaan tarkastella myös, kuinka eri metodien tuottamat estimaatit poikkeavat estimaatista, joka on laskettu käyttäen täydellistä aineistoa.

## 4.2 Puuttuvaan tietoon liittyvät oletukset

Tässä käsitellyn aineiston puuttuvuuteen vaikuttaa oletettavasti lähinnä koehenkilöiden masentuneisuus. Masentuneiden koehenkilöiden on oletettavasti hieman vaikeampi jaksaa ylläpitää keskittymistään annetun ohjeistuksen mukaisesti. Toinen mahdollisesti puuttuvuuteen vaikuttava seikka on koeasetelmien järjestys, sillä myöhemmin tuleviin koeasetelmiin voi olla kasvavan väsymyksen vuoksi vaikeampi keskittyä. Toinen järjestykseen liittyvä seikka on se, että koehenkilöt saattavat lopettaa mittaukset kesken. Kesken lopettamisesta johtuva puuttuminen on myöhemmin tuleville koeasetelmille vähintään yhtä suuri kuin aiemmin tulleille koeasetelmille. Järjestys ei kuitenkaan tee systemaattista eroa koeasetelmien välille, sillä niiden järjestys ei ollut vakio vaan se tasapainotettiin järjestysefektin huomioimiseksi. Järjestystä ja sen mahdollista vaikutusta ei näin ollen tarkastella tässä työssä. Usein ihmisiä ja ihmisten toimintaa tutkittaessa huomioidaan koehenkilöiden ikä ja sukupuoli. Voi olla, että myös nämä seikat vaikuttavat ERP-mittauksissa esiintyvään puuttuvuuteen. Valitussa osa-aineistossa on puuttuvia havaintoja vain yksi, ja tämä koehenkilö on iältään 61-vuotias vaihteluvälin ollessa kontrollihenkilöiden keskuudessa 22 – 66 vuotta ja hänen edustamaansa sukupuolta on suurin osa muistakin valitun osa-aineiston koehenkilöistä. Puuttuva havainto ei siis ole äärimmäinen näiden muuttujien suhteen, jolloin ei ole syytä olettaa puuttuvuuden johtuvan näistä tekijöistä. Ikää ja sukupuolta ei olekaan huomioitu tulevissa analyyseissa.

Koehenkilöiden masentuneisuudesta on olemassa tietoa, jolloin kyseessä on satunnainen puuttuvuus. Tämän lisäksi tämä seikka voidaan jättää tulevissa tarkasteluissa kokonaan huomiotta, koska tarkasteltavan osajoukon sisällä se on vakio. Koska valittu ryhmä on diagnosoimattomien ryhmä, ei masentuneisuus luo eroa tämän ryhmän jäsenten välille. Rajoituttaessa käsittelemään vain diagnosoimattomien ryhmää voidaan olettaa, että kyseessä on täysin satunnainen puuttuvuus ja puuttuvuuden aiheuttamaa mekanismia ei näin ollen käsitellä tulevissa analyyseissa. Yleisesti ottaen ERP-mittauksissa myös sukupuoli ja ikä ovat tiedossa, jolloin voitaisiin mallipohjaista moni-imputointia edelleen soveltaa ja saada harhattomia tuloksia, vaikka kyseessä olisi satunnainen puuttuvuus.

## 4.3 Hot deck -sovellus

Valitussa aineistossa on jokaista ilmettä ja elektrodia kohden kaksi piirrettä. Näin ollen on valittava luovuttaja puuttuvalle havainnolle samaksi näille molemmille muuttujille kullakin moni-imputoinnin kierroksella. Muutoin ei voitaisi ajatella, että imputoidaan kokonaisia aikasarjoja. Nyt valittaessa jokaiselle ilmeelle ja elektrodille oma luovuttaja voitaisiin tältä luovuttajalta ottaa pelkkien piirteiden sijaan kyseisen ilmeen ja elektrodin koko keskiarvokäyrä ja käyttää koko sen tarjoamaa informaatiota seuraavissa analyyseissa. Toisaalta yhdistettäessä vielä useamman muuttujan imputointi voidaan näitä kaikkia koskevaa tietoa hyödyntää luovuttajan valinnassa. Alustavien tulosten perusteella ei kuitenkaan kannata valita luovuttajaa kaikille koeasetelman muuttujille kerralla, koska näin toimittaessa keskivirhe estimoituu liian pieneksi. Alustavissa tarkasteluissa kokeiltiin myös yhden luovuttajan valitsemista kerralla yhden koeasetelman kaikille muuttujille käyttäen mallipohjaista hot deck -moni-imputointia. Tässä tilanteessa havaittiin, että puuttuvuuden kasvaessa riittävän suureksi eri imputointikierroksilla *d* valituiksi luovuttajiksi päätyivät samat koehenkilöt. Tällöin eri imputointikierrosten välinen vaihtelu jäi hyvin pieneksi ja siten keskivirhe estimoitui jopa pienemmäksi kuin tilanteessa, jossa ei ole puuttuvuutta lainkaan. Suuremmilla puuttuvuuden määrillä tuloksissa alkoi olemaan myös selvästi harhaa, mikä johtui luultavasti siitä, että jotkin koehenkilöt päätyivät luovuttajiksi useamman kerran myös yksittäisten imputointikierrosten sisällä ollen luovuttajina useammalle puuttuvalle havainnolle.

Edellä esitellyt ongelmat johtuvat kenties siitä, että tämän kokoluokan aineistossa mahdollisia luovuttajia on vähän. Tässä tilanteeessa yksittäiselle luovuttajalle saattaa kertyä todennäköisyyttä liian paljon sen ollessa kääntäen verrannollinen imputointimallin selittäjinä toimivien muuttujien avaruudessa määritettyyn euklidiseen etäisyyteen puuttuvan havainnon ja mahdollisen luovuttajan välillä. Euklidinen etäisyys käy hyvin suureksi yhä useammalle mahdolliselle luovuttajalle kun ulottuvuuksien määrä kasvaa, ilmiö tunnetaan nimellä ”the curse of dimensionality” (Hastie, Tibshirani & Friedman, 2009, s. 22). Siksi luovuttaja valitaan kerralla vain niille kahdelle piirteelle, jotka tulevat koeasetelman samasta ilmeestä ja elektrodista. Kun valitaan kullekin elektrodille oma luovuttajansa, menetelmä yleistyy helposti useamman elektrodin tilanteeseen. Molemmat koeasetelmat, joihin simuloidaan puuttuvuutta, käsitellään erikseen.

Aiemmin on esitetty, että imputointikierrosten määrä kannattaa pitää melko pienenä (3 - 10), koska kierrosten lisäämisellä saavutettu hyöty on pieni verrattuna laskenta-ajan kasvuun (Little & Rubin, 2002). Toisaalta myöhemmin on näytetty, että tilastollisen testin teho saadaan selvästi lähemmäksi teoreettista maksimiaan, kun käytetään suuremmille puuttuvan tiedon osuuksille selvästi suurempia imputointikierrosten määriä (Graham, 2012). Graham esittääkin, että puuttuvan tiedon osuuden ollessa puolet tulisi kierroksia olla 40 (2012). Vanhemmissa lähteissä mainitut perustelut pienemmälle otoskoolle ovat liittyneet paljolti laskennan raskauteen, joka tekee menetelmästä hitaamman. Nykyään laskentaraskauteen liittyvät perustelut ovat relevantteja vain hyvin suurten aineistojen kohdalla, koska tietokoneiden laskentateho on kehittynyt niin paljon. Yhdessä koeasetelmassa 18 koehenkilön joukosta enimmillään 10 koehenkilön havainnot simuloidaan puuttuviksi. Puuttuvia havaintoja on enimmillään hieman yli puolet, joten imputointikierrosten lukumääräksi *D* on päätetty ottaa Grahamin ohjeistusta mukaillen viisikymmentä. Myös simulointitarkastelut näyttivät, että saavutettu lisähyöty kävi lähellä viittäkymmentä hyvin pieneksi.

Valittu kiinnostuksen kohteena oleva suure on samassa koeasetelmassa esitettyjen ilmeiden aiheuttamien aktivaatioiden erotus. Näin ollen kaava (3.1) saadaan muotoon

$$\bar{\theta}_{D=50} = \frac{1}{50}\sum_{d=1}^{50}\left(\frac{1}{18}\sum_{i=1}^{18}\left({y^{(d)}}_{i,ilme1} - {y^{(d)}}_{i,ilme2}\right)\right), \qquad (4.3)$$

missä vektorit $y_{ilme1}$ ja $y_{ilme2}$ sisältävät havaittujen arvojen lisäksi imputointikierroksella *d* imputoidut arvot. Kun $W_d$ määritellään kaavan (4.2) antaman tuloksen neliönä, saadaan kaavan (3.4) avulla estimaatin $\bar{\theta}_{D=50}$ varianssi seuraavasti

$$T_D = \frac{1}{50}\sum_{d=1}^{50}\left(\frac{1}{18(18-1)}\sum_{i=1}^{18}\left(\hat{\theta}_d - \left({y^{(d)}}_{i,ilme1} - {y^{(d)}}_{i,ilme2}\right)\right)^2\right) + \frac{50+1}{50(50-1)}\sum_{d=1}^{50}(\hat{\theta}_d - \bar{\theta}_{D=50})^2. \quad (4.4)$$

Nämä molemmat lasketaan erikseen kaikille yhden koeasetelman elektrodeille ja piirteille, jolloin estimaatteja $\bar{\theta}_D$ saadaan neljä yhtä koeasetelmaa kohden. Käytettäessä kaavaa (4.3) määrittämään haluttu estimaatti saadaan jackknife kaavat (3.11) ja (3.12) muotoon

$$S_{(.)} = \sum_{j=1}^{n}\frac{1}{n}\left[\frac{1}{50}\sum_{d=1}^{50}\left(\frac{1}{17}\sum_{i=1}^{17}\left({y^{(j,d)}}_{i,ilme1} - {y^{(j,d)}}_{i,ilme2}\right)\right)\right], \qquad (4.5)$$

missä ${y^{(j,d)}}_{ilme1}$ ja ${y^{(j,d)}}_{ilme2}$ ovat jackknife-kierroksella *j* toteutetun moni-imputoinnin kierroksella *d* sisältämät annetun piirteen arvot annetulla elektrodilla. Joukko, jonka yli erotusten keskiarvo lasketaan, on jackknife-sovelluksessa yhden pienempi, koska kullakin jackknife-kierroksella yhden koehenkilön mittaukset jätetään pois. Kaavassa (4.5) sisimmäisten sulkujen sisällä lasketaan erotus eri ilmeiden aiheuttamassa aktivaatiossa. Seuraavien sulkujen sisällä lasketaan tämän erotuksen keskiarvo yli aineiston, joka sisältää sekä havaitut että imputoidut arvot. Kolmansien sulkujen sisällä keskiarvo lasketaan kaikille imputointikierroksille ja otetaan näistä keskiarvo. Ylimmän tason keskiarvo lasketaan yli jackknife-kierrosten *n*, luku *n* on joukon $y_{obs}$ koko ja riippuu puuttuvien havaintojen lukumäärästä. Kaavan (3.11) antama varianssi saadaan muotoon

$$\widehat{VAR}\, S_{(.)} = \frac{n-1}{n}\sum_{j=1}^{n}\left[\frac{1}{50}\sum_{d=1}^{50}\left(\frac{1}{17}\sum_{i=1}^{17}\left({y^{(j,d)}}_{i,ilme1} - {y^{(j,d)}}_{i,ilme2}\right)\right) - S_{(.)}\right]^2, \qquad (4.6)$$

missä $S_{(.)}$on kuten määritelty kaavassa (4.5). Näiden jackknife-estimaattoreiden toteutus löytyy R-koodista liitteestä F.

Edellisiä laskutoimituksia varten tarvitaan imputoidut arvot ja niiden saamiseksi mallipohjaiselle moni-imputoinnille tarvitaan imputointimallit. Mallinnusta varten kukin muuttuja on normeerattu. Normeerauksen ensimmäisessä vaiheessa estimoidaan neljä ensimmäistä L-momenttia (Hosking, 1990), joiden avulla saadaan kvantiilisekoitusjakauma (Karvanen, 2006). Tämän jälkeen lasketaan estimoidun kvantiilisekoitusjakauman kertymäfunktion arvo kullekin muuttujan havaitulle arvolle. Kertymäfunktion antamiin arvoihin sovelletaan standardinormaalijakauman kertymäfunktion käänteisfunktiota. Käänteisfunktion antamat arvot ovat ne, joita mallinnuksessa ja ennusteiden laskemisessa käytetään. Nämä vaiheet toteutetaan erikseen kullekin muuttujalle. Havaitulle arvolle $y_{i,}$ saadaan

$$y_i^* = \Phi^{-1}\left(F\left(y_i; \widehat{L_1}, \widehat{L_2}, \widehat{L_3}, \widehat{L_4}\right)\right),$$

missä $\widehat{L_1}$, $\widehat{L_2}$, $\widehat{L_3}$ ja $\widehat{L_4}$ ovat estimoidut L-momentit ja *F* on sekoitusjakauman kertymäfunktio. Esitelty normeeraus toteutetaan kaikille puuttuvuutta sisältävän sekä imputointimallien selittäjät sisältävien koeasetelmien muuttujille. R-koodi normeerauksen toteuttamiseksi on liitteessä C.

Moni-imputointia varten tarvittava mallinnus on jaoteltu tässä siten, että kunkin elektrodin kullekin piirteelle on oma mallinsa vastaamaan kutakin ilmettä. Yhteensä malleja on kahdeksan yhtä koeasetelmaa

kohden. Selittäjinä kussakin mallissa toimivat seitsemän ilmeen koeasetelmasta saman tunnetilan ja elektrodin molemmat piirteet. Selittäviä muuttujia on kussakin mallissa kaksi. Mallien estimoinnin jälkeen lasketaan ennusteet kullekin arvolle, eli kutakin koehenkilöä kohden tulee kahdeksan eri ennustetta yhtä koeasetelmaa kohden, vastaten eri elektrodien eri piirteitä kummallekin ilmeelle. Nämä ennusteet lasketaan koehenkilöille käyttäen edellä määriteltyjä normeerattuja arvoja kappaleessa 3.5 määritellyllä tavalla. R-koodi ennusteiden laskemiseksi löytyy liitteestä C.

Mahdollisten luovuttajien valintaa varten tarvitaan kaksiulotteiseen tapaukseen soveltuva etäisyysmitta kaavoihin (3.7) ja (3.8), lisäksi samaista etäisyyttä tarvitaan mahdollisen luovuttajan valitsemistodennäköisyyden laskemiseksi kaavaan (3.9). Koska arvot otetaan samalta luovuttajalta molempiin tiettyä ilmettä ja elektrodia koskeviin muuttujiin annetussa koeasetelmassa, otetaan etäisyysmitaksi näiden eri muuttujien ennusteiden etäisyyksien summa. Olkoon $\hat{y}_i^{mis}$ koehenkilön *i*, jolta puuttuu arvot annetusta koeasetelmasta, yhtä ilmettä ja elektrodia koskevat ennusteet sisältävä vektori. Olkoon $\hat{y}_j^{obs}$ koehenkilön *j*, jolta ei puutu arvoja annetusta koeasetelmasta, samaa ilmettä ja elektrodia koskevat ennusteet sisältävä vektori. Tällöin ennusteiden etäisyys valitussa ilmeessä on

$$\sum_{k=1}^{2} \left| \hat{y}_{i,k}^{mis} - \hat{y}_{j,k}^{obs} \right|, \qquad (4.7)$$

missä *k* käy läpi piirteet. R-koodi etäisyyksien laskemiseksi on liitteessä C. Nämä etäisyydet lasketaan kullekin koehenkilölle, jolta havainto puuttuu, kutakin sellaista koehenkilöä kohden, jolta havaintoa ei puutu. Näin saatujen etäisyyksien avulla voidaan määrittää mahdolliset luovuttajat ja kullekin mahdolliselle luovuttajalle valintatodennäköisyys.

Mallipohjaisen moni-imputoinnin suorittava funktio on määritelty R-koodissa, joka löytyy liitteestä D, ja BB-moni-imputoinnin suorittava funktio on määritelty R-koodissa, joka löytyy liitteestä E. Molempien näiden menetelmien kohdalla on käytetty muita osia R-koodista samanlaisena.

## 4.4 Menetelmien vertailu simuloimalla

Mahdollisia eroja eri puuttuvuuden käsittelymetodien välillä tarkastellaan simuloimalla. Aluksi puuttuvuutta luotiin havaittujen kahdeksantoista koehenkilön joukkoon valitsemalla satunnaisesti määrätyn kokoinen joukko puuttuviksi. Tällä tavoin edetessä havaittu vaihtelu eri simulointikierrosten välillä jää kuitenkin varsin pieneksi, kahdeksastatoista havainnosta ei voida valita puuttumaan pientä määrää eri havaintoja kovinkaan monella eri tavalla. Sen lisäksi ettei puuttujia voida valita monella eri tavalla ovat myös mahdollisten luovuttajien joukot lähes samat. Tällä tavoin saadut tulokset jouduttiinkin hylkäämään ja lähestymistapaa muuttamaan.

Aineiston rajallisuuden tuottamat ongelmat simulointiin pyrittiin kiertämään simuloimalla joka simulointikierroksella myös aineisto. Aineiston simulointi tehtiin 24-ulotteisesta jakaumasta, jonka parametrit estimoitiin kahdeksantoista valitun koehenkilön aineistosta. Tässä estimoitiin yksi normaalijakauma koskemaan kaikkia kolmea eri koeasetelmaa, koska kaikkien koeasetelmien välillä on voimakasta korrelaatiota ja lisäksi tämä suoraviivaistaa menetelmää. Estimoidussa normaalijakaumassa on kahdeksan muuttujaa kutakin kolmea koeasetelmaa kohden, oma muuttujansa kummankin ilmeen kummallekin piirteelle kumpaakin elektrodia kohden. Aineiston simuloinnin jälkeen siihen simuloidaan puuttuvuutta täysin satunnaisesti, arpomalla *p* simuloitua koehenkilöä, joiden arvot asetetaan puuttuviksi valituissa muuttujissa. Puuttujien arpominen tehtiin erikseen koskien kumpaakin koeasetelmaa, joihin

puuttuvuutta haluttiin luoda, jolloin täydellisten havaintorivien analyysissa käytössä olevia rivejä voi siis olla vähemmän kuin (18-*p*)-kappaletta. Tarkasteluissa *p* käy läpi arvot yhdestä kymmeneen ja kutakin puuttuvuuden määrää simuloitiin 500 kertaa. Simulointikierrosten määrä jätettiin verrattain alhaiseksi, koska jackknife-sovelluksen lisääminen menetelmään teki laskennasta hyvin raskasta ja sikäli hidasta.

Puuttuvuuden luomisen jälkeen kullakin simulointikierroksella estimoitiin haluttujen erotusten odotusarvot ja näiden keskivirheet. Estimointi tehtiin käyttämällä sekä täydellisten havaintorivien analyysia että kuvailtuja moni-imputointimenetelmiä. Simuloinnin jälkeen on saatu 500 kappaletta erotusten odotusarvoja sekä näiden keskivirheiden estimaatteja kullekin puuttuvuuden lukumäärälle kaikille metodeille. Tämän jälkeen voidaan tarkastella, miten eri menetelmillä saadut tulokset vertautuvat toisiinsa ja miten eri menetelmien välinen paremmuus kenties vaihtelee puuttuvan tiedon osuuden kasvaessa.

Tarkasteltavia mittareita, joilla eri menetelmien hyvyyttä arvioidaan, on kolme. Ensimmäisenä on havaittu odotusarvon jackknife-estimaattorin keskivirhe

$$se_{obs}(\bar{\theta}) = \sqrt{\frac{1}{500-1}\sum_{i=1}^{500}\left(\bar{\theta}_i - \sum_{j=1}^{500}\frac{\bar{\theta}_j}{500}\right)^2}, \qquad (4.8)$$

missä $\bar{\theta}_i$ on simulointikierroksella *i* saatu odotusarvon estimaatti $S_{(.)}$ ja 500 simulointikierrosten määrä. Merkitään tällä tavoin laskettua BB-estimaattien havaittua keskivirhettä $se_{obs}(\bar{\theta})^{BB}$, mallipohjaisen moni-imputointi-estimaatin keskivirhettä $se_{obs}(\bar{\theta})^{MI}$ ja täydellisten havaintorivien estimaattien havaittua keskivirhettä $se_{obs}(\bar{\theta})^{CC}$. Toinen mittari on estimoitujen keskivirheiden keskiarvo, jota voidaan verrata havaittuun keskivirheeseen ja näin saada kuva sen oikeellisuudesta. Estimoitujen keskivirheiden keskiarvo on

$$\hat{E}\left(\widehat{se}(\bar{\theta})\right) = \sum_{i=1}^{500}\frac{1}{500}\left(\sqrt{\widehat{VAR}\,S_{(.)\,i}}\right), \qquad (4.9)$$

missä $\sqrt{\widehat{VAR}\,S_{(.)\,i}}$ on simulointikierroksella *i* saatu odotusarvon estimaatin keskivirheen estimaatti. Merkitään tällä tavoin laskettua BB-estimaattien estimoitujen keskivirheiden keskiarvoa $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$, mallipohjaisen moni-imputointi-estimaattien estimoitujen keskivirheiden keskiarvoa $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ ja täydellisten havaintorivien estimaattien estimoitujen keskivirheiden keskiarvoa $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$. Kolmanneksi katsotaan odotusarvon estimaattien keskiarvoa, koska tällä tavoin voidaan nähdä menetelmän mahdollisesti tuottama harha. Odotusarvon estimaattien keskiarvo on

$$\hat{E}(\bar{\theta}) = \sum_{i=1}^{500}\frac{\bar{\theta}_i}{500}, \qquad (4.10)$$

missä $\bar{\theta}_i$ on simulointikierroksella *i* saatu odotusarvon estimaatti. Koska aineisto on simuloitu, tiedetään oikea odotusarvo täsmälleen. Merkitään tällä tavoin laskettua BB-estimaattien keskiarvoa $E_{BB}(\bar{\theta})$, mallipohjaisen moni-imputointiestimaattien keskiarvoa $E_{MI}(\bar{\theta})$ ja täydellisten havaintorivien estimaattien keskiarvoa $E_{CC}(\bar{\theta})$. Odotusarvoissa ei kuitenkaan kiinnostavaa ole niiden absoluuttinen arvo vaan niiden poikkeama tunnetusta aineiston simuloinnissa käytetystä odotusarvosta eli harha. Odotusarvon estimaattien harhan keskiarvo on

$$\hat{E}(\bar{\theta})_{BIAS} = \hat{E}(\bar{\theta}) - \theta = \frac{1}{500}\sum_{i=1}^{500}(\bar{\theta}_i - \theta), \qquad (4.11)$$

missä $\theta$ on tunnettu simuloinnissa käytetty odotusarvo. Merkitään tällä tavoin saatua BB-estimaattien harhan keskiarvoa $E_{BB}(\bar{\theta})_{BIAS}$, mallipohjaisen moni-imputointiestimaattien harhan keskiarvoa $E_{MI}(\bar{\theta})_{BIAS}$ ja täydellisten havaintorivien estimaattien harhan keskiarvoa $E_{CC}(\bar{\theta})_{BIAS}$. R-koodi näiden simulointien suorittamiseksi löytyy liitteestä G, jossa funktiot joihin viitataan on määritelty mallipohjaisen moni-imputoinnin tapauksessa liitteissä C, D sekä F ja BB-moni-imputoinnin tapauksessa liitteissä E sekä F.

## 4.5 Simulointitulokset

Simuloidaan ensin havaintoaineisto kappaleen 4.3 mukaisesti ja toteutetaan moni-imputointi kappaleessa 4.2 esitellyllä tavalla. Lasketaan valitut mittarit kaavojen (4.8), (4.9) ja (4.11) mukaisesti yli simulointikierrosten kullekin elektrodille ja piirteelle erikseen kummassakin koeasetelmassa. Tehdään edellinen laskenta kullekin eri puuttuvuuden käsittelymenetelmälle. Erotuksen oikea odotusarvo tiedetään, merkitään sitä seuraavissa taulukoissa puuttuvien lukumäärän 0 kohdalla täydellisten havaintorivien analyysia koskevissa sarakkeissa.

Taulukoissa 1 – 8 esitetään yhtä koeasetelmaa koskevat simulointitulokset siten, että kussakin taulukossa on tulokset koskien koeasetelman sisällä yhtä elektrodia ja piirrettä. Valitussa koeasetelmassa iloinen ilme on vakio ärsyke ja iloinen ilme on poikkeava ärsyke.

**Taulukko 1 Hajontaluvut surullinen ilme, elektrodi T5 ja piirre P1**

| Elektrodi T5, piirre P1 ja vertailuilme surullinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.141 | 0.141 | 0.135 | 0.140 | 0.139 | 0.171 |
| 2 | 0.149 | 0.144 | 0.134 | 0.158 | 0.145 | 0.199 |
| 3 | 0.174 | 0.162 | 0.140 | 0.180 | 0.156 | 0.219 |
| 4 | 0.188 | 0.167 | 0.145 | 0.204 | 0.163 | 0.239 |
| 5 | 0.238 | 0.195 | 0.154 | 0.238 | 0.179 | 0.251 |
| 6 | 0.244 | 0.204 | 0.159 | 0.259 | 0.194 | 0.260 |
| 7 | 0.255 | 0.221 | 0.158 | 0.282 | 0.206 | 0.274 |
| 8 | 0.284 | 0.254 | 0.166 | 0.304 | 0.229 | 0.273 |
| 9 | 0.313 | 0.283 | 0.177 | 0.360 | 0.253 | 0.274 |
| 10 | 0.331 | 0.311 | 0.184 | 0.357 | 0.272 | 0.281 |

**Taulukko 2 Odotusarvon harhan estimaatit surullinen ilme, elektrodi T5 ja piirre P1**

| Elektrodi T5, piirre P1 ja vertailuilme surullinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | -0.067 | |
| 1 | 0.000 | 0.001 | 0.010 |
| 2 | -0.001 | 0.005 | 0.021 |
| 3 | 0.000 | 0.001 | 0.011 |
| 4 | -0.001 | 0.003 | 0.008 |
| 5 | 0.004 | -0.012 | 0.008 |
| 6 | -0.006 | -0.008 | 0.002 |
| 7 | -0.005 | -0.015 | 0.007 |
| 8 | -0.012 | -0.006 | 0.009 |
| 9 | -0.002 | -0.023 | -0.014 |
| 10 | -0.012 | -0.012 | 0.011 |

**Taulukko 3 Hajontaluvut surullinen ilme, elektrodi T6 ja piirre P1**

| Elektrodi T6, piirre P1 ja vertailuilme surullinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.260 | 0.264 | 0.249 | 0.260 | 0.255 | 0.275 |
| 2 | 0.275 | 0.277 | 0.249 | 0.292 | 0.272 | 0.305 |
| 3 | 0.312 | 0.294 | 0.267 | 0.329 | 0.287 | 0.321 |
| 4 | 0.352 | 0.320 | 0.273 | 0.377 | 0.308 | 0.346 |
| 5 | 0.417 | 0.318 | 0.289 | 0.440 | 0.334 | 0.355 |
| 6 | 0.445 | 0.382 | 0.303 | 0.466 | 0.351 | 0.372 |
| 7 | 0.495 | 0.405 | 0.326 | 0.529 | 0.389 | 0.387 |
| 8 | 0.537 | 0.458 | 0.330 | 0.568 | 0.418 | 0.398 |
| 9 | 0.577 | 0.516 | 0.348 | 0.595 | 0.450 | 0.412 |
| 10 | 0.623 | 0.535 | 0.376 | 0.653 | 0.509 | 0.419 |

**Taulukko 4 Odotusarvon harhan estimaatit surullinen ilme, elektrodi T6 ja piirre P1**

| Elektrodi T6, piirre P1 ja vertailuilme surullinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | -0.149 | |
| 1 | 0.008 | 0.003 | 0.050 |
| 2 | -0.024 | -0.025 | 0.029 |
| 3 | 0.002 | 0.003 | 0.040 |
| 4 | 0.014 | 0.006 | 0.059 |
| 5 | 0.027 | 0.016 | 0.048 |
| 6 | 0.001 | -0.005 | 0.059 |
| 7 | -0.002 | -0.005 | 0.051 |
| 8 | 0.004 | 0.022 | 0.027 |
| 9 | -0.033 | -0.025 | 0.046 |
| 10 | -0.018 | 0.027 | 0.031 |

**Taulukko 5 Hajontaluvut ilme surullinen, elektrodi T5 ja piirre N170**

| Elektrodi T5, piirre N170 ja vertailuilme surullinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.168 | 0.169 | 0.157 | 0.173 | 0.171 | 0.218 |
| 2 | 0.188 | 0.186 | 0.179 | 0.190 | 0.178 | 0.257 |
| 3 | 0.211 | 0.190 | 0.173 | 0.221 | 0.194 | 0.293 |
| 4 | 0.222 | 0.201 | 0.181 | 0.247 | 0.205 | 0.316 |
| 5 | 0.269 | 0.220 | 0.190 | 0.284 | 0.225 | 0.340 |
| 6 | 0.292 | 0.246 | 0.198 | 0.307 | 0.234 | 0.348 |
| 7 | 0.331 | 0.271 | 0.209 | 0.352 | 0.254 | 0.357 |
| 8 | 0.344 | 0.314 | 0.213 | 0.353 | 0.280 | 0.355 |
| 9 | 0.383 | 0.341 | 0.230 | 0.412 | 0.304 | 0.370 |
| 10 | 0.396 | 0.378 | 0.233 | 0.444 | 0.345 | 0.364 |

**Taulukko 6 Odotusarvon harhan estimaatit surullinen ilme, elektrodi T5 ja piirre N170**

| Elektrodi T5, piirre N170 ja vertailuilme surullinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | -0.098 | |
| 1 | -0.007 | -0.003 | 0.007 |
| 2 | 0.011 | 0.015 | 0.019 |
| 3 | -0.007 | -0.003 | 0.017 |
| 4 | 0.002 | 0.006 | 0.028 |
| 5 | -0.009 | -0.011 | 0.003 |
| 6 | -0.017 | -0.016 | 0.010 |
| 7 | 0.008 | -0.007 | 0.021 |
| 8 | -0.023 | 0.016 | 0.024 |
| 9 | 0.018 | -0.014 | 0.004 |
| 10 | 0.016 | 0.041 | 0.012 |

**Taulukko 7 Hajontaluvut ilme surullinen, elektrodi T6 ja piirre N170**

| Elektrodi T6, piirre N170 ja vertailuilme surullinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.299 | 0.299 | 0.296 | 0.292 | 0.289 | 0.312 |
| 2 | 0.310 | 0.309 | 0.289 | 0.332 | 0.309 | 0.342 |
| 3 | 0.348 | 0.320 | 0.321 | 0.371 | 0.330 | 0.360 |
| 4 | 0.408 | 0.376 | 0.321 | 0.414 | 0.351 | 0.387 |
| 5 | 0.431 | 0.382 | 0.346 | 0.468 | 0.374 | 0.402 |
| 6 | 0.502 | 0.413 | 0.344 | 0.554 | 0.409 | 0.420 |
| 7 | 0.577 | 0.465 | 0.349 | 0.605 | 0.450 | 0.440 |
| 8 | 0.607 | 0.522 | 0.404 | 0.651 | 0.488 | 0.444 |
| 9 | 0.630 | 0.560 | 0.389 | 0.725 | 0.528 | 0.467 |
| 10 | 0.727 | 0.691 | 0.415 | 0.787 | 0.582 | 0.470 |

**Taulukko 8 Odotusarvon harhan estimaatit surullinen ilme, elektrodi T6 ja piirre N170**

| Elektrodi T6, piirre N170 ja vertailuilme surullinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | -0.564 | |
| 1 | 0.030 | 0.025 | 0.078 |
| 2 | -0.014 | -0.01 | 0.035 |
| 3 | 0.008 | 0.011 | 0.070 |
| 4 | 0.014 | 0.023 | 0.081 |
| 5 | -0.008 | 0.005 | 0.061 |
| 6 | -0.011 | 0.016 | 0.061 |
| 7 | -0.038 | 0.005 | 0.097 |
| 8 | 0.002 | 0.041 | 0.056 |
| 9 | 0.011 | 0.002 | 0.081 |
| 10 | -0.036 | -0.008 | 0.019 |

Taulukoista 2, 4, 6 ja 8 nähdään, ettei mikään menetelmä ole vahvasti harhainen. Mutta BB-moni-imputointi näyttäisi antavan säännönmukaisesti hieman oikeaa odotusarvoa suurempia arvoja. Puuttuvan tiedon osuuden kasvattaminen ei näytä muuttavan tilannetta minkään menetelmän osalta. Taulukoista 1, 3, 5 ja 7 nähdään menetelmien hajonnat. Havaitun keskivirheen osalta mallipohjainen moni-imputointi ja täydellisten havaintorivien analyysi näyttävät antavan samanlaisia tuloksia pienillä puuttuvuuden määrillä, mutta puuttuvuuden kasvaessa yli kolmeen on täydellisten havaintorivien analyysi parempi. Puuttuvan tiedon osuudesta riippumatta BB:llä on pienin havaittu keskivirhe kaikissa tarkastelluissa muuttujissa ja tämä ero BB:n hyväksi kasvaa puuttuvan tiedon osuuden kasvaessa. Tarkasteltaessa estimoitujen keskivirheiden keskiarvoja tilanne muuttuu huomattavasti BB:n osalta. BB:llä on suurin estimoitujen keskivirheiden keskiarvo kaikilla muuttujilla aluksi, tilanne kuitenkin kääntyy elektrodin T6 molempien piirteiden kohdalla BB:lle edulliseksi kun puuttuvia havaintoja on seitsemän tai enemmän. Estimoitujen keskivirheiden keskiarvossa täydellisten havaintorivien analyysi ja mallipohjainen moni-imputointi saavat samanlaisia arvoja, jos puuttuu korkeintaan kahden koehenkilön mittaukset, tämän jälkeen mallipohjainen moni-imputointi kääntyy selvästi huonommaksi tämänkin mittarin valossa kaikkien muuttujien kohdalla.

Simulointitulokset ilmeelle iloinen on esitetty liitteen A taulukoissa 9 – 16. Taulukoista 10, 12, 14 ja 16 nähdään, että myös iloisen ilmeen kohdalla BB antaa melko säännönmukaisesti hieman suurempia arvoja kuin muut menetelmät erotuksen odotusarvoa estimoitaessa, muuten menetelmissä ei ole nähtävissä harhaa riippumatta puuttuvuuden määrästä. Taulukoista 9, 11, 13 ja 15 nähdään iloista ilmettä koskevat hajontaluvut. Havaitut keskivirheet iloisen ilmeen kohdalla käyttäytyvät lähes samalla tavalla kuin surullisen, BB:llä on pienin keskivirhe ja mallipohjaisella moni-imputoinnilla saadaan suunnilleen samankokoisia arvoja kuin täydellisten havaintorivien analyysilla, jos puuttuvia havaintoja on korkeintaan kolme. Tämän jälkeen mallipohjainen moni-imputointi tuottaa selvästi ja säännönmukaisesti suurempia keskivirheitä. Huomionarvoinen ero on elektrodin T6 kohdalla, jos puuttuvia havaintoja on vain yksi, jolloin BB on niukasti huonompi menetelmä kuin muut tällä mittarilla katsottuna. Estimoitujen keskivirheiden keskiarvo on myös iloisen ilmeen kohdalla samanlainen mallipohjaisella moni-imputoinnilla ja täydellisten havaintorivien analyysilla, kun puuttuvia havaintoja on korkeintaan kaksi, minkä jälkeen moni-imputointi on jälleen selvästi ja säännönmukaisesti huonompi. BB iloisen ilmeen kohdalla antaa muita suurempia estimoitujen keskivirheiden keskiarvoja aluksi, mutta kääntyy kaikkien muuttujien kohdalla parhaaksi kun puuttuvia

havaintoja on paljon, elektrodin T5 piirteen N170 kohdalla vasta puuttuvien havaintojen lukumäärän ollessa 10.

Yhteenvetona tuloksista havaitaan seuraavat seikat. Taulukoista 2, 4, 6, 8, 10, 12, 14 ja 16 nähdään, että elektrodilla T6 saadaan selvästi itseisarvoltaan suurempia erotuksia vertailuilmeen ja neutraalin ilmeen välille kuin elektrodilla T5 molemmissa piirteissä. Näistä taulukoista nähdään myös, että piirre N170 antaa selvästi itseisarvoltaan suurempia erotuksia vertailuilmeen ja neutraalin ilmeen välille kummallakin elektrodilla. Molemmat tulokset käyvät yksiin ennakkotietojen kanssa. Ensimmäinen siksi, että kasvojen tunnistuksen ajatellaan olevan toiselle aivopuoliskolle painottunut toiminto (Purves et al., 2012). Toinen siitä syystä, että piirteen N170 ajatellaan kuvaavan nimenomaan kasvojen aiheuttamaa aktivaatiota (Bentin et al., 1996).

# 5 Pohdinta

Mallipohjainen moni-imputointi ja täydellisten havaintorivien analyysi vaikuttivat olevan yhtä tarkkoja menetelmiä havaittujen keskivirheiden ja estimoitujen keskivirheiden keskiarvojen avulla tarkasteltuna. Toisin kuin aluksi oletettiin, muun tiedon käyttämisellä ei saatu tarkempia tuloksia puuttuvuutta sisältävistä muuttujista. Tämä saattaa selittyä sillä, että käytetyssä menetelmässä tarvitaan puuttuville havainnoille luovuttajat, jotka saattavat kaikki olla imputointimallin selittävien muuttujien avaruudessa kaukana puuttuvasta havainnosta havaintojen määrän ollessa alle 20. Hyviä luovuttajia ei löydetä, koska niitä ei havaintoaineistossa ole. Mallipohjaista moni-imputointia pienillä puuttuvuuksien määrillä puoltaa kuitenkin sen asettamat väljemmät oletukset itse puuttuvuusmekanismille. Mallipohjainen moni-imputointi olisi edelleen harhaton, vaikka kyseessä olisi simuloinnissa käytetyn täysin satunnaisen puuttuvuuden sijaan satunnainen puuttuvuus.

BB-moni-imputointi oli menetelmistä paras havaitun keskivirheen näkökulmasta ja ero sen hyväksi kasvoi jatkuvasti puuttuvan tiedon osuuden kasvaessa. Sillä saatiin tarkempia tuloksia, vaikkei käytettykään tietoa muista muuttujista. Estimoitujen keskivirheiden keskiarvo osoitti, että hajonnan estimointi oli selvästi harhainen kaikilla puuttuvuuden määrillä. Keskivirhe estimoitui säännönmukaisesti liian suureksi, tällöin menetelmän käyttö ei lisää riskiä tyypin 1 virheen tekemiseen. BB-moni-imputointi oli paras menetelmä molemmilla hajontamittareilla molemmissa piirteissä tarkasteltaessa tunne-eroja paremmin erottelevaa elektrodia T6, kun puuttuvien havaintojen määrä oli noin puolet havainnoista. BB-estimaatit odotusarvolle olivat usein hieman liian suuria oikeaan odotusarvoon verrattuna, vaikutus oli hyvin vähäinen mutta säännönmukainen.

Alussa esitettiin kysymys, saadaanko moni-imputoimalla tarkempia tuloksia kuin täydellisten havaintorivien analyysilla tarkkuuden mittareiden ollessa estimaattien hajonta ja harha. Vastaus tähän vaikuttaa olevan hyvin tilannekohtainen. Menetelmien paremmuusjärjestys riippuu selvästi puuttuvuuden määrästä ja osittain myös tarkasteltavasta muuttujasta. Vähintään yhtä hyviä ja suuremman osan aikaa parempia tuloksia saadaan ainakin toisella tarkastelluista moni-imputointimenetelmistä. Mahdollisen tarkkuuden parantumisen lisäksi moni-imputointia käyttämällä säästytään informaation haaskaamiselta. Täydellisten havaintorivien analyysissa joudutaan jättämään kaikki tieto käyttämättä koehenkilöltä, jolta puuttuu tietoa jostain muuttujasta, jolloin päädytään haaskaamaan myös olemassa olevat mittaustulokset. Tämä käyttämättä jäävän datan osuus saattaa kasvaa hyvin merkittäväksi puuttuvuuden ilmetessä useammassa muuttujassa riippumattomasti.

Tässä työssä tarkasteltiin tilannetta, jossa otoskoko on lähtökohtaisesti hyvin pieni 18. Otoskoko ei kuitenkaan ole tämän tyyppiselle kokeelle erityisen pieni, sillä tavanomaisesti otoskoot ovat enimmilläänkin joitain kymmeniä ERP-mittauksissa. Mikäli otoksessa puuttuvien havaintojen määrä on vähäinen, mallipohjainen hot deck -moni-imputointi tuottaa harhattomia tuloksia myös MAR tilanteessa ilman keskivirhe-estimaattien merkittävää kasvua. Mikäli taas puuttuvien havaintojen osuus on suuri, niin BB-moni-imputointi saattaa tuottaa selvästi pienempiä keskivirheitä. BB-moni-imputointi vaatii kuitenkin MCAR-oletuksen tuottaakseen harhattomia tuloksia. Yleisessä tapauksessa ei puuttuvuuden aiheuttamasta mekanismista voida kuitenkaan sanoa mitään, sillä ei voida tietää mitä kaikkea ei tiedetä. Mallipohjaisen moni-imputoinnin yksi suuri vahvuus onkin puuttuvuuteen liittyvien oletusten lieventäminen.

Hot deck -moni-imputoinnissa voidaan ajatella, että onnistuttiin imputoimaan kokonaisia aikasarjoja yksittäisten arvojen sijaan. Tällöin imputoitua havaintoaineistoa voitaisiin analysoida paljon monipuolisemmin kuin tässä esitetyllä odotusarvotarkasteluilla. Tässä työssä ei kuitenkaan käsitelty sitä, miten kokonaisten aikasarjojen käyttö voitaisiin oikeuttaa. Tämä analyysin laajentaminen varsinaisesti aikasarjatasolle onkin kenties merkittävin jatkokehityssuunta. Toisena on molempien moni-imputointimenetelmien keskivirheiden estimoinnin harhan suuruus ja sen pienentäminen. Kolmantena on sen tutkiminen, miksi BB-moni-imputointi tuottaa lievästi harhaisia tuloksia odotusarvolle ja miten tämä voidaan korjata.

# Lähteet


Andridge, R. R. & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 40-64.

Bentin, S., Allison, T., Puce, A., & Perez, E. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 551-565.

Efron, B. & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 586-596.

Graham, J. W. (2012). *Missing data analysis and design.* New York: Springer.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning.* New York: Springer.

Hosking, J. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of Royal Statistical Society B*, 105-124.

Karvanen, J. (2006). Estimation of quantile mixtures via L-moments and trimmed L-moments. *Computational Statistics & Data Analysis*, 947–959.

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data.* New Jersey: John Wiley & sons Inc.

Niedemeyer, E. & Da Silva, F. L. (2005). *Electroencephalography Basic Principles, Clinical Applications, and Related Fields.* Philadelphia: Lippincott Williams & Wilkins.

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LeMantia, A.-S. & White, L. E. (2012). *Neuroscience.* Sunderland: Sinauer associates, inc.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 130-134.

Rubin, D. B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 366-374.

Vaden, K. I., Gebregziabher, M., Kuchinsky, S. E. & Eckert, M. A. (2012). Multiple imputation of missing fMRI data in whole brain analysis. *Neuroimage*, 1843-1855.

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton: Taylor & Francis group.

# Liite A: simulointitulokset ilmeelle iloinen

**Taulukko 9 Hajontaluvut ilme iloinen, elektrodi T5 ja piirre P1**

| Elektrodi T5, piirre P1 ja vertailuilme iloinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.259 | 0.261 | 0.237 | 0.255 | 0.252 | 0.269 |
| 2 | 0.268 | 0.265 | 0.261 | 0.295 | 0.271 | 0.294 |
| 3 | 0.317 | 0.289 | 0.253 | 0.339 | 0.288 | 0.320 |
| 4 | 0.376 | 0.321 | 0.282 | 0.382 | 0.310 | 0.333 |
| 5 | 0.419 | 0.343 | 0.283 | 0.429 | 0.329 | 0.352 |
| 6 | 0.451 | 0.374 | 0.272 | 0.471 | 0.360 | 0.369 |
| 7 | 0.502 | 0.401 | 0.281 | 0.533 | 0.392 | 0.380 |
| 8 | 0.504 | 0.476 | 0.302 | 0.568 | 0.414 | 0.395 |
| 9 | 0.554 | 0.485 | 0.352 | 0.639 | 0.456 | 0.399 |
| 10 | 0.625 | 0.548 | 0.365 | 0.658 | 0.495 | 0.405 |

**Taulukko 10 Odotusarvon harhan estimaatit iloinen ilme, elektrodi T5 ja piirre P1**

| Elektrodi T5, piirre P1 ja vertailuilme iloinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | 0.268 | |
| 1 | -0.014 | -0.018 | 0.020 |
| 2 | -0.024 | -0.015 | 0.015 |
| 3 | 0.002 | -0.005 | 0.041 |
| 4 | -0.009 | 0.004 | 0.028 |
| 5 | 0.012 | -0.023 | 0.016 |
| 6 | 0.007 | -0.020 | 0.017 |
| 7 | -0.062 | -0.042 | 0.010 |
| 8 | 0.000 | -0.013 | 0.005 |
| 9 | 0.009 | -0.014 | 0.007 |
| 10 | -0.027 | -0.009 | -0.005 |

**Taulukko 11 Hajontaluvut ilme iloinen, elektrodi T6 ja piirre P1**

| Elektrodi T6, piirre P1 ja vertailuilme iloinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.193 | 0.193 | 0.199 | 0.201 | 0.198 | 0.225 |
| 2 | 0.220 | 0.209 | 0.201 | 0.226 | 0.211 | 0.252 |
| 3 | 0.250 | 0.232 | 0.196 | 0.254 | 0.219 | 0.279 |
| 4 | 0.273 | 0.239 | 0.199 | 0.285 | 0.235 | 0.292 |
| 5 | 0.308 | 0.253 | 0.221 | 0.323 | 0.254 | 0.310 |
| 6 | 0.348 | 0.274 | 0.230 | 0.368 | 0.278 | 0.321 |
| 7 | 0.354 | 0.299 | 0.239 | 0.397 | 0.308 | 0.337 |
| 8 | 0.410 | 0.383 | 0.258 | 0.444 | 0.324 | 0.333 |
| 9 | 0.449 | 0.374 | 0.284 | 0.482 | 0.345 | 0.347 |
| 10 | 0.477 | 0.461 | 0.282 | 0.513 | 0.382 | 0.352 |

**Taulukko 12 Odotusarvon harhan estimaatit iloinen ilme, elektrodi T6 ja piirre P1**

| Elektrodi T6, piirre P1 ja vertailuilme iloinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | -0.400 | |
| 1 | 0.002 | -0.003 | 0.054 |
| 2 | 0.000 | 0.001 | 0.048 |
| 3 | -0.027 | -0.025 | 0.052 |
| 4 | -0.001 | 0.004 | 0.045 |
| 5 | -0.003 | -0.015 | 0.049 |
| 6 | -0.004 | -0.014 | 0.043 |
| 7 | -0.031 | -0.009 | 0.053 |
| 8 | -0.009 | -0.023 | 0.043 |
| 9 | 0.023 | -0.010 | 0.048 |
| 10 | -0.016 | -0.008 | 0.043 |

**Taulukko 13 Hajontaluvut ilme iloinen, elektrodi T5 ja piirre N170**

| Elektrodi T5, piirre N170 ja vertailuilme iloinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.225 | 0.226 | 0.224 | 0.230 | 0.229 | 0.264 |
| 2 | 0.250 | 0.256 | 0.228 | 0.263 | 0.238 | 0.301 |
| 3 | 0.285 | 0.275 | 0.237 | 0.298 | 0.262 | 0.330 |
| 4 | 0.321 | 0.285 | 0.239 | 0.342 | 0.276 | 0.351 |
| 5 | 0.351 | 0.315 | 0.254 | 0.372 | 0.292 | 0.377 |
| 6 | 0.377 | 0.327 | 0.267 | 0.417 | 0.322 | 0.384 |
| 7 | 0.417 | 0.349 | 0.287 | 0.462 | 0.344 | 0.407 |
| 8 | 0.461 | 0.414 | 0.298 | 0.548 | 0.378 | 0.409 |
| 9 | 0.495 | 0.450 | 0.293 | 0.539 | 0.400 | 0.419 |
| 10 | 0.599 | 0.505 | 0.338 | 0.609 | 0.446 | 0.420 |

**Taulukko 14 Odotusarvon harhan estimaatit iloinen ilme, elektrodi T6 ja piirre P1**

| Elektrodi T5, piirre N170 ja vertailuilme iloinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | 0.004 | |
| 1 | -0.009 | -0.006 | 0.030 |
| 2 | -0.013 | -0.014 | 0.023 |
| 3 | 0.022 | 0.019 | 0.034 |
| 4 | 0.005 | 0.011 | 0.032 |
| 5 | 0.019 | 0.008 | 0.033 |
| 6 | -0.011 | -0.009 | 0.037 |
| 7 | -0.031 | -0.022 | 0.012 |
| 8 | 0.012 | 0.037 | 0.035 |
| 9 | -0.028 | -0.030 | 0.029 |
| 10 | 0.008 | 0.027 | 0.008 |

**Taulukko 15 Hajontaluvut ilme iloinen, elektrodi T6 ja piirre N170**

| Elektrodi T6, piirre N170 ja vertailuilme iloinen | | | | | | |
|---|---|---|---|---|---|---|
| Havaitut keskivirheet | | | | Estimoitujen keskivirheiden keskiarvot | | |
| Puuttuvien lkm. | $se_{obs}(\bar{\theta})^{MI}$ | $se_{obs}(\bar{\theta})^{CC}$ | $se_{obs}(\bar{\theta})^{BB}$ | $E_{MI}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{CC}\left(\widehat{se}(\bar{\theta})\right)$ | $E_{BB}\left(\widehat{se}(\bar{\theta})\right)$ |
| 0 | | | | | | |
| 1 | 0.208 | 0.214 | 0.224 | 0.231 | 0.230 | 0.245 |
| 2 | 0.239 | 0.235 | 0.229 | 0.258 | 0.238 | 0.263 |
| 3 | 0.285 | 0.267 | 0.236 | 0.295 | 0.258 | 0.284 |
| 4 | 0.316 | 0.286 | 0.244 | 0.341 | 0.275 | 0.305 |
| 5 | 0.332 | 0.286 | 0.260 | 0.375 | 0.297 | 0.320 |
| 6 | 0.382 | 0.321 | 0.261 | 0.420 | 0.323 | 0.333 |
| 7 | 0.423 | 0.370 | 0.266 | 0.467 | 0.350 | 0.350 |
| 8 | 0.458 | 0.400 | 0.281 | 0.529 | 0.387 | 0.343 |
| 9 | 0.508 | 0.451 | 0.299 | 0.571 | 0.410 | 0.364 |
| 10 | 0.547 | 0.501 | 0.327 | 0.591 | 0.447 | 0.362 |

**Taulukko 16 Odotusarvon harhan estimaatit iloinen ilme, elektrodi T6 ja piirre N170**

| Elektrodi T6, piirre N170 ja vertailuilme iloinen | | | |
|---|---|---|---|
| Odotusarvojen keskiarvoiset harhat | | | |
| Puuttuvien lkm. | $E_{MI}(\bar{\theta})_{BIAS}$ | $E_{CC}(\bar{\theta})_{BIAS}$ | $E_{BB}(\bar{\theta})_{BIAS}$ |
| 0 | | -0.579 | |
| 1 | 0.021 | 0.016 | 0.038 |
| 2 | -0.002 | -0.001 | 0.017 |
| 3 | -0.018 | -0.010 | 0.009 |
| 4 | 0.003 | -0.002 | 0.011 |
| 5 | -0.010 | -0.015 | 0.012 |
| 6 | 0.030 | 0.017 | 0.004 |
| 7 | 0.020 | -0.007 | 0.032 |
| 8 | -0.023 | 0.015 | 0.010 |
| 9 | 0.037 | 0.036 | 0.025 |
| 10 | -0.007 | -0.029 | 0.017 |

## Liite B: R-koodi osa-aineiston valitsemiseksi ja muuttujakohtaisten keskiarvojen sekä kovarianssimatriisin laskemiseksi

```
library(Lmoments)
yhdistetty<-as.data.frame(yhdistetty)
attach(yhdistetty)

tayd<-c(1:length(NDeSADStN_P1_58))[!is.na(NDeSADStN_P1_58)&!is.na(NDeHstN_P1_58)]
length(tayd)

NDeHstH_P1_58<-NDeHstH_P1_58[tayd]
NDeHstH_P1_96<-NDeHstH_P1_96[tayd]
NDeHstH_N170_58<-NDeHstH_N170_58[tayd]
NDeHstH_N170_96<-NDeHstH_N170_96[tayd]

NDeHstN_P1_58<-NDeHstN_P1_58[tayd]
NDeHstN_P1_96<-NDeHstN_P1_96[tayd]
NDeHstN_N170_58<-NDeHstN_N170_58[tayd]
NDeHstN_N170_96<-NDeHstN_N170_96[tayd]

NDeSADStSAD_P1_58<-NDeSADStSAD_P1_58[tayd]
NDeSADStSAD_P1_96<-NDeSADStSAD_P1_96[tayd]
NDeSADStSAD_N170_58<-NDeSADStSAD_N170_58[tayd]
NDeSADStSAD_N170_96<-NDeSADStSAD_N170_96[tayd]

NDeSADStN_P1_58<-NDeSADStN_P1_58[tayd]
NDeSADStN_P1_96<-NDeSADStN_P1_96[tayd]
NDeSADStN_N170_58<-NDeSADStN_N170_58[tayd]
NDeSADStN_N170_96<-NDeSADStN_N170_96[tayd]

N7ExpH_P1_58<-N7ExpH_P1_58[tayd]
N7ExpH_P1_96<-N7ExpH_P1_96[tayd]
N7ExpH_N170_58<-N7ExpH_N170_58[tayd]
N7ExpH_N170_96<-N7ExpH_N170_96[tayd]

N7ExpSAD_P1_58<-N7ExpSAD_P1_58[tayd]
N7ExpSAD_P1_96<-N7ExpSAD_P1_96[tayd]
N7ExpSAD_N170_58<-N7ExpSAD_N170_58[tayd]
N7ExpSAD_N170_96<-N7ExpSAD_N170_96[tayd]

N7ExpN_P1_58<-N7ExpN_P1_58[tayd]
N7ExpN_P1_96<-N7ExpN_P1_96[tayd]
N7ExpN_N170_58<-N7ExpN_N170_58[tayd]
N7ExpN_N170_96<-N7ExpN_N170_96[tayd]


library(MASS)
tayd<-c(1:length(tayd))
```

```
######################## Simuloinnin toteutus ########################
taydData<-cbind(NDeHstH_P1_58,NDeHstH_P1_96,NDeHstH_N170_58,NDeHstH_N170_96,
  NDeHstN_P1_58,NDeHstN_P1_96,NDeHstN_N170_58,NDeHstN_N170_96,
  NDeSADStSAD_P1_58,NDeSADStSAD_P1_96,NDeSADStSAD_N170_58,NDeSADStSAD_N170_96,
  NDeSADStN_P1_58,NDeSADStN_P1_96,NDeSADStN_N170_58,NDeSADStN_N170_96,
  N7ExpH_P1_58,N7ExpH_P1_96,N7ExpH_N170_58,N7ExpH_N170_96,
  N7ExpSAD_P1_58,N7ExpSAD_P1_96,N7ExpSAD_N170_58,N7ExpSAD_N170_96,
  N7ExpN_P1_58,N7ExpN_P1_96,N7ExpN_N170_58,N7ExpN_N170_96)


myy<-apply(X=taydData,MARGIN=2,FUN=mean)

CovMat<-cov(taydData)


### Alkuvalmistelut loppuvat ###
################################################################################
```

## Liite C: R-koodi, jossa määritellään apufunktiot muuttujien normeerausta, etäisyysmitan laskemista ja imputointimallien ennusteita varten

```
############### Apufunktioita ###################

norming <- function(x){
 # returns normalized values
 # if x[i]=NA then normx[i]=NA
 xapu<-x[!is.na(x)]
 misInd<-c(1:length(x))[is.na(x)]
 normx<-qnorm(pnormpoly(xapu,data2normpoly4(xapu)))
 if(!length(misInd)==0){
  nx<-vector(length=length(x))
  ind<-1
  for(i in 1:length(x)){
   if(is.na(x[i])){nx[i]<-NA}
   else{nx[i]<-normx[ind];ind<-ind+1}
  }
  normx<-nx
 }
 return(normx)
}

differ<-function(v1,v2){
 v<-rbind(v1,v2)
 dif<-colSums(abs(v))
 dif<-dif*sign(colSums(v))
 return(dif)
}

fitObs<-function(Y,X,obsInd, kappa=0.0001){
 X<-cbind(1,X)    #ilmeisesti algoritmi olettaa, että X:ssä on myös vakiota vastaava sarake
 S<-t(X[obsInd,]) %*% X[obsInd,]
 V<-solve( (S+kappa*diag(diag(S))) )
 betaHat<-V %*% (t(X[obsInd,])) %*% Y[obsInd]
 fitted<-X[obsInd,]%*%betaHat
 return(fitted)
}

ennusteet<-function(Y, X, obsInd, misInd, n, kappa=0.0001){
 X<-cbind(1,X)    #ilmeisesti algoritmi olettaa, että X:ssä on myös vakiota vastaava sarake
 S<-t(X[obsInd,]) %*% X[obsInd,]
 V<-solve( (S+kappa*diag(diag(S))) )
 betaHat<-V %*% (t(X[obsInd,])) %*% Y[obsInd]
 gDot<-rchisq(1,df= (n-length(misInd)-ncol(X)) )
 sigmaDot<-as.numeric( ( t(Y[obsInd]-X[obsInd,]%*%betaHat) %*%
                  (Y[obsInd]-X[obsInd,]%*%betaHat) ) )/gDot
```

```
  sigmaDot<-sqrt(sigmaDot)
  z1<-rnorm(ncol(X))
  Vsqrt<-chol(V)
  betaDot<-betaHat + sigmaDot*Vsqrt%*%z1
  z2<-rnorm(length(misInd))

  yhat<-X[misInd,]%*%betaDot +z2*sigmaDot
  return(yhat)
}
```

## Liite D: R-koodi, jossa määritellään mallipohjainen moni-imputointi

```
################################################################
############ Imptuointifunktio ####################

imp<-function(Y,X,D=50){
 obsInd<-c(1:nrow(Y))[!is.na(Y[,1])]
 misInd<-setdiff(1:nrow(Y),obsInd)
 n<-length(obsInd)+length(misInd)

 #C <- c(seq(0.8,0.9,by=0.1),seq(1.1,1.2,by=0.1))
 #Dupl<-dupl(cbind(Y,X),C)
 #YDupl<-Dupl[,1:ncol(Y)]
 #XDupl<-Dupl[,(ncol(Y)+1):ncol(Dupl)]
 #obsInd2 <- c(obsInd,seq(n+1,n+length(obsInd)*length(C)))

 XNorm<-apply(X=X,MARGIN=2,FUN=norming)
 YNorm<-apply(X=Y,MARGIN=2,FUN=norming)

 impData<-list()

 for(i in 1:ncol(Y)){
  impData<-c(impData,list(matrix(ncol=length(misInd),nrow=D)))
 }

 value<-values(YNorm[,c(1,2)],Y[,c(1,2)],XNorm[,c(1,5)],XNorm[,c(2,6)],D,obsInd,misInd)
 impData[[1]]<-value[[1]]
 impData[[2]]<-value[[2]]
 value<-values(YNorm[,c(3,4)],Y[,c(3,4)],XNorm[,c(3,7)],XNorm[,c(4,8)],D,obsInd,misInd)
 impData[[3]]<-value[[1]]
 impData[[4]]<-value[[2]]
 value<-values(YNorm[,c(5,6)],Y[,c(5,6)],XNorm[,c(5,1)],XNorm[,c(6,2)],D,obsInd,misInd)
 impData[[5]]<-value[[1]]
 impData[[6]]<-value[[2]]
 value<-values(YNorm[,c(7,8)],Y[,c(7,8)],XNorm[,c(7,3)],XNorm[,c(8,4)],D,obsInd,misInd)
 impData[[7]]<-value[[1]]
 impData[[8]]<-value[[2]]

return(impData)
}


values<-function(YNorm,Y,X1,X2,D,obsInd,misInd){

 imputations<-list(matrix(ncol=length(misInd),nrow=D))
 imputations<-c( imputations, list(matrix(ncol=length(misInd),nrow=D)) )
 fitObs1<-fitObs(YNorm[,1],cbind(X1[,1],X1[,2]),obsInd)
```

```
fitObs2<-fitObs(YNorm[,2],cbind(X2[,1],X2[,2]),obsInd)

for(d in 1:D){
 fit1<-ennusteet(YNorm[,1],cbind(X1[,1],X1[,2]),obsInd,misInd,n)
 fit2<-ennusteet(YNorm[,2],cbind(X2[,1],X2[,2]),obsInd,misInd,n)

 luovuttajatPos1 <- matrix(ncol=2,nrow=length(misInd))
 luovuttajatNeg1 <- matrix(ncol=2,nrow=length(misInd))

 for(r in 1:length(misInd)){
  dif1<-differ( (fit1[r]-fitObs1) , (fit2[r]-fitObs2) )

  for(l in 1:ncol(luovuttajatPos1)){
   ## donors for YNorm
   if(sum(dif1>=0)>=l){
    luovuttajatPos1[r,l]<-obsInd[c(1:length(dif1))[dif1>=0][rank(
                dif1[dif1>=0] )==l]]
    }
   if(sum(dif1<0)>=l){
    luovuttajatNeg1[r,l]<-obsInd[c(1:length(dif1))[dif1<0][rank(
                abs(dif1[dif1<0]) )==l]]
    }
   }
  }


 luovutusTnPos1<-matrix(ncol=2,nrow=length(misInd))
 distSumPos1<-matrix(ncol=2,nrow=length(misInd))
 luovutusTnNeg1<-matrix(ncol=2,nrow=length(misInd))
 distSumNeg1<-matrix(ncol=2,nrow=length(misInd))

 for(r in 1:length(misInd)){
  for(l in 1:ncol(luovuttajatPos1)){
   if(!is.na(luovuttajatPos1[r,l])){
    donor<-luovuttajatPos1[r,l]
    distSumPos1[r,l]<-abs(fit1[r]-fitObs1[obsInd==donor])+
     abs(fit2[r]-fitObs2[obsInd==donor])
   }


   if(!is.na(luovuttajatNeg1[r,l])){
    donor<-luovuttajatNeg1[r,l]
    distSumNeg1[r,l]<-abs(fit1[r]-fitObs1[obsInd==donor])+
     abs(fit2[r]-fitObs2[obsInd==donor])
   }
  }
 }
```

```
  for(r in 1:length(misInd)){
   for(l in 1:ncol(luovuttajatPos1)){
    if(!is.na(distSumPos1[r,l])){
     luovutusTnPos1[r,l]<-(1/distSumPos1[r,l]) / sum(1/distSumPos1[r,],na.rm=TRUE)}
    else{luovutusTnPos1[r,l]<-0}


    if(!is.na(distSumNeg1[r,l])){
     luovutusTnNeg1[r,l]<-(1/distSumNeg1[r,l]) / sum(1/distSumNeg1[r,],na.rm=TRUE)}
    else{luovutusTnNeg1[r,l]<-0}
   }
  }

  for(j in 1:length(misInd)){
   if(sum(!is.na(luovuttajatNeg1[j,]))>=1){
    if(sum(!is.na(luovuttajatPos1[j,]))>=1){
     p_n<-runif(1)
     if(p_n<=0.5){donor_d1<-sample(luovuttajatNeg1[j,],size=1,prob=luovutusTnNeg1[j,])}
     else{donor_d1<-sample(luovuttajatPos1[j,],size=1,prob=luovutusTnPos1[j,])}
    }
    else{donor_d1<-sample(luovuttajatNeg1[j,],size=1,prob=luovutusTnNeg1[j,])}
   }
   else{donor_d1<-sample(luovuttajatPos1[j,],size=1,prob=luovutusTnPos1[j,])}

   imputations[[1]][d,j]<-Y[donor_d1,1]
   imputations[[2]][d,j]<-Y[donor_d1,2]

  }
 }

 return(imputations)
}
```

## Liite E: R-koodi Bayes-bootstrap-moni-imputoinnin imputointifunktiolle

```
#################################################################
########### Imptuointifunktio ####################

imp<-function(Y,D=50){
 obsInd<-c(1:nrow(Y))[!is.na(Y[,1])]
 misInd<-setdiff(1:nrow(Y),obsInd)
 n<-length(obsInd)+length(misInd)

 impData<-list()

 for(i in 1:ncol(Y)){
  impData<-c(impData,list(matrix(ncol=length(misInd),nrow=D)))
 }

 value<-values(Y[,c(1,2)],D,obsInd,misInd)
 impData[[1]]<-value[[1]]
 impData[[2]]<-value[[2]]
 value<-values(Y[,c(3,4)],D,obsInd,misInd)
 impData[[3]]<-value[[1]]
 impData[[4]]<-value[[2]]
 value<-values(Y[,c(5,6)],D,obsInd,misInd)
 impData[[5]]<-value[[1]]
 impData[[6]]<-value[[2]]
 value<-values(Y[,c(7,8)],D,obsInd,misInd)
 impData[[7]]<-value[[1]]
 impData[[8]]<-value[[2]]

return(impData)
}

values<-function(Y,D,obsInd,misInd){

 imputations<-list(matrix(ncol=length(misInd),nrow=D))
 imputations<-c( imputations, list(matrix(ncol=length(misInd),nrow=D)) )

 for(d in 1:D){

  for(r in 1:length(misInd)){
            donorProb<-diff(sort(c(0,runif(n=(length(obsInd)-1),min=0,max=1),1)))
            donor<-sample(x=obsInd,size=1,replace=FALSE,prob=donorProb)
      imputations[[1]][d,r]<-Y[donor,1] ######  #######  ########
      imputations[[2]][d,r]<-Y[donor,2] ######  #######  ########
     }
    }
 return(imputations)
}
```

## Liite F: R-koodi, jossa määritellään tarkasteltava estimaattori ja sen jackknife-sovellus moni-imputointiin

```
#########################################################################################
##### Simuloinnissa tarkasteltava estimaattori  ###################

estimaattiTunneEro<-function(imputoinnitTunne1, havTunne1,imputoinnitTunne2,havTunne2){
 D<-nrow(imputoinnitTunne1)
 t<-vector(length=D)
 varianssit<-vector(length=D)
 n<-length(havTunne1)+ncol(imputoinnitTunne1)

 for(i in 1:D){
  erotukset<-c(havTunne1,imputoinnitTunne1[i,])-c(havTunne2,imputoinnitTunne2[i,])
  t[i]<-mean(erotukset)
  varianssit[i]<-var(erotukset)/n
 }
 B<-var(t)*((D+1)/D)  ##B<-((D+1)/(D*(D-1))) * sum((mean(t)-t)^2)
 U_line<-mean(varianssit)
 tulos<-c(mean(t),B,U_line)
 tulos
}

jackEstim<-function(y){
 if(sum(is.na(y))==length(y)){return(c(NA,NA))}
 na<-c(1:length(y))[is.na(y)]
 NoNa<-setdiff(c(1:length(y)),na)
 theta<-mean(y[NoNa])
 if(length(na)<=(length(y)-2)){
 variance<-sum((y[NoNa]-theta)^2) * ((length(y[NoNa])-1)/length(y[NoNa]))
 }else(variance<-NA)
 return(c(theta,variance))
}
```

## Liite G: R-koodi, jossa toteutetaan simulointi käyttäen liitteissä B – F määriteltyjä funktioita ja aineistoja

```
#############################################################################################
####### Simulointi ############################

set.seed(19.54)
nsim<-500
n<-length(tayd)

sim_tulos_H<-list()
sim_tulosTAU_H<-list()
sim_tulos_SAD<-list()
sim_tulosTAU_SAD<-list()

puuttuvatLKM<-c(1:10)

for(m in puuttuvatLKM){
  sim_tulos_H<-c(sim_tulos_H,list(matrix(nrow=nsim,ncol=8)))
  sim_tulosTAU_H<-c(sim_tulosTAU_H,list(matrix(nrow=nsim,ncol=8)))
  sim_tulos_SAD<-c(sim_tulos_SAD,list(matrix(nrow=nsim,ncol=8)))
  sim_tulosTAU_SAD<-c(sim_tulosTAU_SAD,list(matrix(nrow=nsim,ncol=8)))
}

X<-cbind(  N7ExpH_P1_58,N7ExpH_P1_96,N7ExpH_N170_58,N7ExpH_N170_96,
  N7ExpSAD_P1_58,N7ExpSAD_P1_96,N7ExpSAD_N170_58,N7ExpSAD_N170_96,
  N7ExpN_P1_58,N7ExpN_P1_96,N7ExpN_N170_58,N7ExpN_N170_96)

YH<-cbind(NDeHstH_P1_58,NDeHstH_P1_96,NDeHstH_N170_58,NDeHstH_N170_96,
  NDeHstN_P1_58,NDeHstN_P1_96,NDeHstN_N170_58,NDeHstN_N170_96)
YSAD<-cbind(
NDeSADStSAD_P1_58,NDeSADStSAD_P1_96,NDeSADStSAD_N170_58,NDeSADStSAD_N170_96,
  NDeSADStN_P1_58,NDeSADStN_P1_96,NDeSADStN_N170_58,NDeSADStN_N170_96)

count<-rep(0,length(puuttuvatLKM))

date()
for(p in 1:length(puuttuvatLKM)){
  print(p)
  for(k in 1:nsim){

    puuttuvatH<-sample(tayd,puuttuvatLKM[p],replace=FALSE)
    obsH<-setdiff(tayd,puuttuvatH)
    puuttuvatSAD<-sample(tayd,puuttuvatLKM[p],replace=FALSE)
    obsSAD<-setdiff(tayd,puuttuvatSAD)
    obsAll<-setdiff(tayd,c(puuttuvatH,puuttuvatSAD))
```

```
  simdata<-mvrnorm(n=length(tayd),mu=myy,Sigma=CovMat)
  simYH<-simdata[,1:8]
  simYSAD<-simdata[,9:16]
  simX<-simdata[,17:28]
  simYH[puuttuvatH,]<-NA
  simYSAD[puuttuvatSAD,]<-NA

  jackImpH<-matrix(ncol=12,nrow=length(obsH))
  jackTAUH<-matrix(ncol=8,nrow=length(obsH))
  for(jack in 1:length(obsH)){
   imputoinnitH<-imp(simYH[-obsH[jack],],simX[-obsH[jack],c(1:4,9:12)])
   jackImpH[jack,1:3]  <-estimaattiTunneEro(imputoinnitH[[1]],simYH[obsH[-
jack],1],imputoinnitH[[5]],simYH[obsH[-jack],5])
   jackImpH[jack,4:6]  <-estimaattiTunneEro(imputoinnitH[[2]],simYH[obsH[-
jack],2],imputoinnitH[[6]],simYH[obsH[-jack],6])
   jackImpH[jack,7:9]  <-estimaattiTunneEro(imputoinnitH[[3]],simYH[obsH[-
jack],3],imputoinnitH[[7]],simYH[obsH[-jack],7])
   jackImpH[jack,10:12]<-estimaattiTunneEro(imputoinnitH[[4]],simYH[obsH[-
jack],4],imputoinnitH[[8]],simYH[obsH[-jack],8])
          if( (length(obsAll)-(jack-1)) > 0){
      jackTAUH[jack,1:2]<-c(mean(simYH[obsAll[-jack],1]-simYH[obsAll[-jack],5]),
                  var(simYH[obsAll[-jack],1]-simYH[obsAll[-jack],5])/(length(obsAll[-jack])))

      jackTAUH[jack,3:4]<-c(mean(simYH[obsAll[-jack],2]-simYH[obsAll[-jack],6]),
                  var(simYH[obsAll[-jack],2]-simYH[obsAll[-jack],6])/(length(obsAll[-jack])))

      jackTAUH[jack,5:6]<-c(mean(simYH[obsAll[-jack],3]-simYH[obsAll[-jack],7]),
                  var(simYH[obsAll[-jack],3]-simYH[obsAll[-jack],7])/(length(obsAll[-jack])))

      jackTAUH[jack,7:8]<-c(mean(simYH[obsAll[-jack],4]-simYH[obsAll[-jack],8]),
                  var(simYH[obsAll[-jack],4]-simYH[obsAll[-jack],8])/(length(obsAll[-jack])))
                                                          }     else( jackTAUH[jack,]<-c(rep(NA,8)) )
  }
        sim_tulos_H[[p]][k,1:2]  <-jackEstim(jackImpH[,1])
        sim_tulos_H[[p]][k,3:4]  <-jackEstim(jackImpH[,4])
        sim_tulos_H[[p]][k,5:6]  <-jackEstim(jackImpH[,7])
        sim_tulos_H[[p]][k,7:8]  <-jackEstim(jackImpH[,10])

   sim_tulosTAU_H[[p]][k,1:2]<-jackEstim(jackTAUH[,1])
   sim_tulosTAU_H[[p]][k,3:4]<-jackEstim(jackTAUH[,3])
   sim_tulosTAU_H[[p]][k,5:6]<-jackEstim(jackTAUH[,5])
   sim_tulosTAU_H[[p]][k,7:8]<-jackEstim(jackTAUH[,7])

  jackImpSAD<-matrix(ncol=12,nrow=length(obsSAD))
  jackTAUSAD<-matrix(ncol=8,nrow=length(obsSAD))
  for(jack in 1:length(obsH)){
  imputoinnitSAD<-imp(simYSAD[-obsSAD[jack],],simX[-obsSAD[jack],5:12])
```

```
    jackImpSAD[jack,1:3]  <-estimaattiTunneEro(imputoinnitSAD[[1]],simYSAD[obsSAD[-
jack],1],imputoinnitSAD[[5]],simYSAD[obsSAD[-jack],5])
    jackImpSAD[jack,4:6]  <-estimaattiTunneEro(imputoinnitSAD[[2]],simYSAD[obsSAD[-
jack],2],imputoinnitSAD[[6]],simYSAD[obsSAD[-jack],6])
    jackImpSAD[jack,7:9]  <-estimaattiTunneEro(imputoinnitSAD[[3]],simYSAD[obsSAD[-
jack],3],imputoinnitSAD[[7]],simYSAD[obsSAD[-jack],7])
    jackImpSAD[jack,10:12]<-estimaattiTunneEro(imputoinnitSAD[[4]],simYSAD[obsSAD[-
jack],4],imputoinnitSAD[[8]],simYSAD[obsSAD[-jack],8])
          if(length(obsAll)-(jack-1)>0){
      jackTAUSAD[jack,1:2]<-c(mean(simYSAD[obsAll[-jack],1]-simYSAD[obsAll[-jack],5]),
                  var(simYSAD[obsAll[-jack],1]-simYSAD[obsAll[-jack],5])/(length(obsAll[-jack])))

      jackTAUSAD[jack,3:4]<-c(mean(simYSAD[obsAll[-jack],2]-simYSAD[obsAll[-jack],6]),
                  var(simYSAD[obsAll[-jack],2]-simYSAD[obsAll[-jack],6])/(length(obsAll[-jack])))

      jackTAUSAD[jack,5:6]<-c(mean(simYSAD[obsAll[-jack],3]-simYSAD[obsAll[-jack],7]),
                  var(simYSAD[obsAll[-jack],3]-simYSAD[obsAll[-jack],7])/(length(obsAll[-jack])))

      jackTAUSAD[jack,7:8]<-c(mean(simYSAD[obsAll[-jack],4]-simYSAD[obsAll[-jack],8]),
                  var(simYSAD[obsAll[-jack],4]-simYSAD[obsAll[-jack],8])/(length(obsAll[-jack])))
                                                          }
     else(jackTAUSAD[jack,]<-rep(NA,8))
  }
        sim_tulos_SAD[[p]][k,1:2]  <-jackEstim(jackImpSAD[,1])
        sim_tulos_SAD[[p]][k,3:4]  <-jackEstim(jackImpSAD[,4])
        sim_tulos_SAD[[p]][k,5:6]  <-jackEstim(jackImpSAD[,7])
        sim_tulos_SAD[[p]][k,7:8]  <-jackEstim(jackImpSAD[,10])

    sim_tulosTAU_SAD[[p]][k,1:2]<-jackEstim(jackTAUSAD[,1])
    sim_tulosTAU_SAD[[p]][k,3:4]<-jackEstim(jackTAUSAD[,3])
    sim_tulosTAU_SAD[[p]][k,5:6]<-jackEstim(jackTAUSAD[,5])
    sim_tulosTAU_SAD[[p]][k,7:8]<-jackEstim(jackTAUSAD[,7])

 }
}

date()
```

## Liite H: R-koodi, jossa poimitaan piirteiden arvot

```
data_7exp<-
matrix(ncol=9,dimnames=list(c(),c("ID","Kont","Emo","Sex","Age","P1_58","P1_96","N170_58","N170_96")
))

emotions<-c("A","D","F","H","N","SAD","S")

for(i in 1:7){
 rivi<-c(201,1,emotions[i],1,47,max(Ko201[270:330,i*2-
1]),max(Ko201[270:330,i*2]),min(Ko201[330:410,i*2-1]),min(Ko201[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(203,1,emotions[i],1,37,max(Ko203[270:330,i*2-
1]),max(Ko203[270:330,i*2]),min(Ko203[330:410,i*2-1]),min(Ko203[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(204,1,emotions[i],1,22,max(Ko204[270:330,i*2-
1]),max(Ko204[270:330,i*2]),min(Ko204[330:410,i*2-1]),min(Ko204[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(206,1,emotions[i],2,37,max(Ko206[270:330,i*2-
1]),max(Ko206[270:330,i*2]),min(Ko206[330:410,i*2-1]),min(Ko206[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(207,1,emotions[i],1,35,max(Ko207[270:330,i*2-
1]),max(Ko207[270:330,i*2]),min(Ko207[330:410,i*2-1]),min(Ko207[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(208,1,emotions[i],1,33,max(Ko208[270:330,i*2-
1]),max(Ko208[270:330,i*2]),min(Ko208[330:410,i*2-1]),min(Ko208[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(209,1,emotions[i],1,37,max(Ko209[270:330,i*2-
1]),max(Ko209[270:330,i*2]),min(Ko209[330:410,i*2-1]),min(Ko209[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
 rivi<-c(210,1,emotions[i],1,64,max(Ko210[270:330,i*2-
1]),max(Ko210[270:330,i*2]),min(Ko210[330:410,i*2-1]),min(Ko210[330:410,i*2]) )
 data_7exp<-rbind(data_7exp,rivi)
```

```
}
for(i in 1:7){
  rivi<-c(211,1,emotions[i],1,64,max(Ko211[270:330,i*2-
1]),max(Ko211[270:330,i*2]),min(Ko211[330:410,i*2-1]),min(Ko211[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(212,1,emotions[i],1,53,max(Ko212[270:330,i*2-
1]),max(Ko212[270:330,i*2]),min(Ko212[330:410,i*2-1]),min(Ko212[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(213,1,emotions[i],1,27,max(Ko213[270:330,i*2-
1]),max(Ko213[270:330,i*2]),min(Ko213[330:410,i*2-1]),min(Ko213[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(214,1,emotions[i],1,38,max(Ko214[270:330,i*2-
1]),max(Ko214[270:330,i*2]),min(Ko214[330:410,i*2-1]),min(Ko214[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(215,1,emotions[i],2,25,max(Ko215[270:330,i*2-
1]),max(Ko215[270:330,i*2]),min(Ko215[330:410,i*2-1]),min(Ko215[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(216,1,emotions[i],1,48,max(Ko216[270:330,i*2-
1]),max(Ko216[270:330,i*2]),min(Ko216[330:410,i*2-1]),min(Ko216[330:410,i*2])
 )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(217,1,emotions[i],2,30,max(Ko217[270:330,i*2-
1]),max(Ko217[270:330,i*2]),min(Ko217[330:410,i*2-1]),min(Ko217[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(218,1,emotions[i],1,56,max(Ko218[270:330,i*2-
1]),max(Ko218[270:330,i*2]),min(Ko218[330:410,i*2-1]),min(Ko218[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(219,1,emotions[i],2,57,max(Ko219[270:330,i*2-
1]),max(Ko219[270:330,i*2]),min(Ko219[330:410,i*2-1]),min(Ko219[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
```

```
  rivi<-c(220,1,emotions[i],1,66,max(Ko220[270:330,i*2-
1]),max(Ko220[270:330,i*2]),min(Ko220[330:410,i*2-1]),min(Ko220[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
for(i in 1:7){
  rivi<-c(221,1,emotions[i],1,61,max(Ko221[270:330,i*2-
1]),max(Ko221[270:330,i*2]),min(Ko221[330:410,i*2-1]),min(Ko221[330:410,i*2]) )
  data_7exp<-rbind(data_7exp,rivi)
}
data_7exp

##write.table(data_7exp,file="yhdistettyKont7exp.txt")

data_7exp<-read.table("yhdistetty7exp.txt",header=TRUE)
```