TAMPEREEN
YLIOPISTO

PRO GRADU -TUTKIELMA

Matematiikan, tilastotieteen ja filosofian laitos
Tilastotiede
Lokakuu 2007

SUVI KÄÄRIÄ

Odotusarvon ja kovarianssirakenteen estimointi splinien avulla

Tampereen yliopisto

Matematiikan, tilastotieteen ja filosofian laitos

KÄÄRIÄ, SUVI: Odotusarvon ja kovarianssirakenteen estimointi splinien avulla

Pro gradu -tutkielma, 44 s.

Tilastotiede

Lokakuu 2007

Tiivistelmä

Tutkimusta tehtäessä saatetaan päätyä tilanteeseen, jossa analysoitavaa aineistoa ei voida tyydyttävästi mallintaa parametristen mallien, kuten parametrisen regression avulla. Tällöin aineiston mallinnusta voidaan lähestyä epäparametristen menetelmien kautta. Epäparametrisia aineiston mallinnusmenetelmiä on olemassa useita, joista tässä tutkielmassa käsitellään regressiospliniä ja tasoittavaa spliniä. Työssä osoitetaan, kuinka splinien avulla voidaan joustavasti mallintaa sekä aineiston odotusarvokäyrää että kovarianssimatriisia.

Odotusarvon estimointi niin poikittais-, kuin pitkittäisaineiston tapauksessa on perinteinen sovellusala, johon splinejä on käytetty. Tässä tutkielmassa esitetään, kuinka regressiosplini ja tasoittava splini voidaan muodostaa riippumattomien havaintojen aineistosta. Kyseisten spliniestimaattorien käytännön toimivuutta havainnollistetaan puun runkokäyrän esimerkkiaineiston avulla. Tutkielmassa osoitetaan myös, kuinka tasoittavalla kuutiosplinillä voidaan poikittaisaineiston lisäksi mallintaa myös pitkittäisaineiston odotusarvokäyrää. Koska tasoittavan kuutiosplinin ja lineaarisen sekamallin välillä on löydettävissä yhteys, myös lineaarinen sekamalli esitellään lyhyesti työssä ja osoitetaan, kuinka kyseinen yhteys muodostuu.

Pitkittäisaineiston ominaispiirteenä on havaintojen sisäinen korreloituneisuus, minkä johdosta kovarianssimatriisin onnistunut mallinnus on olennainen osa käyrän sovitusta. Perinteisesti kovarianssimatriisia on mallinnettu esimerkiksi erilaisten kovarianssirakenteiden avulla, jotka kuitenkin käytännön sovelluksissa osoittautuvat usein liian rajoittuneiksi. Tässä tutkielmassa osoitetaan, kuinka tasoittavia kuutiosplinejä voidaan odotusarvon mallinnuksen lisäksi hyödyntää myös kovarianssimatriisin parametrien estimoimisessa. Lähtökohtana toimii modifioitu Choleskyn hajotelma, jonka termien alkioita voidaan estimoida tasoittavilla kuutiosplineillä ja saavuttaa näin aineistoon hyvin sopiva kovarianssimatriisi. Menetelmän toimivuutta havainnollistetaan 40 Suomenkarja-rotuisen sonnin painon kasvukäyräaineiston avulla.

Asiasanat regressiosplini, luonnollinen tasoittava kuutiosplini, kasvukäyrämalli, lineaarinen sekamalli, modifioitu Choleskyn hajotelma

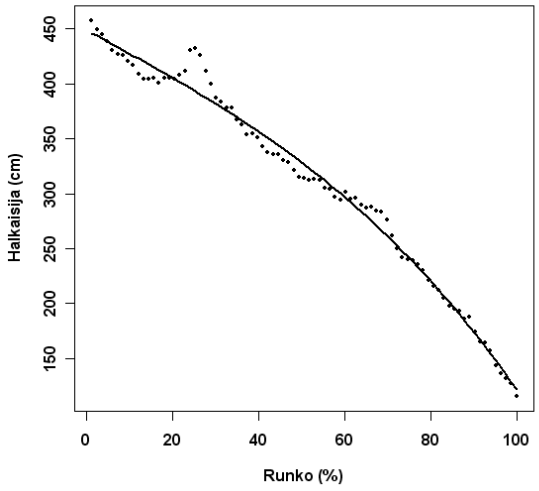
Sisältö

1 Johdanto

1.1 Tutkielman lähtökohta: epäparametrinen regressiomalli

Parametrinen regressiomalli on klassinen ja laajalti käytetty tilastollinen mallinnusmenetelmä. Siinä oletuksena on, että regressiofunktion muoto on täysin tunnettu lukuun ottamatta äärellistä määrää tuntemattomia parametreja. Nämä tuntemattomat parametrit voidaan estimoida sopivan tilastollisen menetelmän, kuten pienimmän neliösumman menetelmän avulla ja saavuttaa näin mallin odotusarvoestimaatit. (Eubank 1999.)

Kuitenkin monissa sovelluksissa parametrinen regressio saattaa osoittautua liian rajoittuneeksi ja usein ongelmaksi muodostuu tarpeeksi vähäparametrisen mallin löytäminen, joka kuitenkin estimoi aineiston odotusarvokäyrän riittävän hyvin. Mikäli aineistoon sovitetaan epäsopiva regressiomalli, päädytään helposti harhaanjohtaviin johtopäätöksiin. (Müller, 1980.) Tällainen tilanne on nähtävissä kuviossa 1.1, jossa alaluvussa 2.1 esiteltävän Puu-aineiston esimerkkipuulle on sovitettu kolmannen asteen polynomi. Kuviosta on helppo havaita, ettei estimoitu käyrä mallinna aineistoa tyydyttävästi.



Kuvio 1.1. Kolmannen asteen polynomin sovitus esimerkkipuulle.

Epäparametrinen regressiomalli tarjoaa vaihtoehtoisen lähestymistavan, kun tavoitteena on mallintaa kuvion 1.1 kaltaista aineistoa, johon parametrinen regressiomalli sopii heikosti. Epäparametrisessa regressiossa funktion muotoa ei ole rajoitettu parametrien kautta, vaan ainoana oletuksena on, että regressiokäyrä kuuluu johonkin funktioavaruuteen. Tällöin esimerkiksi oletetaan, että funktio on k kertaa jatkuvasti differentioituva (Müller 1980.) Kyseinen lähestymistapa on aineistopohjainen, jolloin estimoitava aineisto määrittelee täysin odotusarvofunktion muodon. Tavallisesti mallin monimutkaisuutta voidaan säädellä yhden tai kahden niin kutsutun tasoitusparametrin avulla. (Wu & Zhang 2006.)

On olemassa monia epäparametrisia regressiomenetelmiä, joista suosittuja ovat muun muassa kernel-tasoitus, paikalliset polynomisovitteet (local polynomial fitting), regressiosplinit ja tasoittavat splinit. Tässä työssä käsitellään menetelmistä regressiospliniä ja tasoittavaa spliniä.

1.2 Splinimenetelmien sovelluksia

Kirjallisuudessa on laaja valikoima teoksia, joissa poikittaisaineiston odotusarvon estimointia splinitasoitusmenetelmin on käsitelty. Regressiosplineistä esityksen antaa esimerkiksi Eubank (1998, 1999) ja tasoittavia splinejä ovat käsitelleet muun muassa Wahba (1990) sekä Green ja Silverman (1994). Kuitenkin vasta viime vuosina on havaittu epäparametristen mallinnusmenetelmien käyttökelpoisuus myös pitkittäisaineiston tapauksessa ja menetelmien kehittely onkin ollut kasvavan kiinnostuksen kohteena. Müller (1988) oli ensimmäinen, joka sovelsi epäparametrisia menetelmiä pitkittäisaineistoon. Kuitenkin hänen teoksessaan jokaisen yksilön odotusarvokäyrää on mallinnettu erikseen, minkä johdosta menetelmä on yhtenevä poikittaisaineiston mallinnusmenetelmien kanssa. Myöhemmin pitkittäisaineiston regressiosplinitasoitusta ovat käsitelleet muun muassa Shi, Weiss ja Taylor (1996), Rice ja Wu (2001) sekä Huang, Wu ja Zhou (2002). Tasoittavia splinejä puolestaan ovat käyneet läpi muun muassa Brumback ja Rice (1998), Wang (1998a,b) sekä Lin ja Zhang (1999). Viimeisimmistä julkaisuista kattavan esityksen eri menetelmistä tarjoavat muun muassa Wu ja Zhang (2006).

Odotusarvokäyrän mallintaminen on perinteinen, muttei ainoa sovellus johon splinejä on käytetty. Viime vuosina splinipohjaisia menetelmiä on sovellettu tuloksekkaasti muun muassa puun runkokäyrän ennustamisessa. Ensimmäisenä tasoittaviin kuutiosplineihin perustuvan puun runkokäyrän ennustamismenetelmän esittivät Möttönen ja Nummi (2002) ja myöhemmin menetelmää ovat käsitelleet ja kehittäneet eteenpäin muun muassa Nummi ja Möttönen (2004a) sekä Koskela, Nummi, Wenzel ja Kivinen (2006).

Tässä työssä esitellään vielä eräs tutkimuksen alla oleva sovellusala, jossa tasoittava kuutiosplini osoittautuu käyttökelpoiseksi välineeksi. Kovarianssimatriisin mallinnus on perinteisesti pohjautunut erilaisiin rakenteellisiin malleihin, jotka kuitenkin käytännön tilanteissa osoittautuvat usein liian rajoittuneiksi.

Tässä tutkielmassa esitellään aineistopohjainen menetelmä, jossa modifioitua Choleskyn hajotelmaa ja tasoittavia kuutiosplinejä hyödyntämällä saavutetaan aineistoon hyvin sopiva kovarianssimatriisi.

1.3 Tutkielman tavoitteet ja rakenne

Tutkielman tavoitteena on osoittaa, kuinka splinien avulla voidaan joustavasti ja tehokkaasti mallintaa samanaikaisesti sekä odotusarvokäyrää että kovarianssimatriisia. Käsiteltävät menetelmät ovat regressiosplini ja tasoittava splini ja päähuomion työssä saa erityisesti tasoittava kuutiosplini. Teorian ja käytännön esimerkkien avulla lukija perehdytetään splinifunktioiden tärkeimpiin ominaisuuksiin ja osoitetaan, kuinka tasoittavan kuutiosplinin avulla voidaan saavuttaa paitsi tehokkaat odotusarvoestimaatit, myös joustavasti estimoida kovarianssimatriisin parametreja.

Johdannon ja loppupäätelmien lisäksi työ rakentuu kolmesta pääluvusta. Luvussa 2 esitellään aluksi aineistot, joiden avulla odotusarvon ja kovarianssimatriisin mallinnusta on käytännössä pyritty havainnollistamaan. Alaluvussa 2.1 esitellään odotusarvon estimoimisessa käytetty Puu-aineisto ja alaluvussa 2.2 kovarianssimatriisin mallinnusta havainnollistava Sonni-aineisto.

Luku 3 käsittelee odotusarvon mallintamista splinifunktioden avulla ja käsittelyyn on otettu sekä poikittais- että pitkittäisaineiston odotusarvokäyrä. Alaluvussa 3.1 esitellään aluksi regressiosplini, kun mallinnettava aineisto koostuu riippumattomista havainnoista. Pitkittäisaineiston regressiospliniä ei työssä käsitellä, sillä regressiosplinin tarkoituksena on toimia ainoastaan johdatteluna työn varsinaiseen kiinnostuksen kohteeseen, tasoittavaan spliniin. Poikittaisaineiston tasoittava splini esitellään alaluvussa 3.2, jonka jälkeen alaluvussa 3.3 käydään läpi tasoittava splini pitkittäisaineistolle. Koska tasoittavan kuutiosplinin ja lineaarisen sekamallin välillä on löydettävissä yhteys, myös lineaarinen sekamalli esitellään alaluvussa 3.4 ja osoitetaan, kuinka kyseinen yhteys muodostuu.

Luvussa 4 käydään läpi kovarianssimatriisin mallinnusta. Alaluvussa 4.1 lukija perehdytetään aluksi pitkittäisaineistossa käytettäviin klassisiin kovarianssimatriisin mallinnuskeinoihin, jonka jälkeen alaluvussa 4.2 esitellään uusi, modifioituun Choleskyn hajotelmaan perustuva menetelmä mallintaa kovarianssimatriisia splinien avulla. Alaluvussa 4.3 menetelmää havainnollistetaan käytännön aineistolla ja osoitetaan kolmea eri mallia vertailemalla menetelmän toimivuus. Alaluvun 4.3 tuloksia on käytetty myös julkaisussa Nummi, Kääriä ja Pan (2007).

2 Esimerkkiaineistot

2.1 Puu-aineisto

Odotusarvon mallintamista on havainnollistettu esimerkkiaineistolla, jossa metsämännyn rungon halkaisija on mitattu 83 eri runkopisteessä. Alkuperäinen aineisto, josta esimerkkipuu on poimittu, koostuu yhteensä 25 metsämännystä. Jokaisesta männystä mitattiin rungon halkaisija manuaalisesti 10 cm välein tyvestä latvaan päin, jonka jälkeen mittaukset tasoitettiin eksponentiaalisesti ja otettiin mukaan vain 30 cm välein tehdyt mittaukset. Myös tyvimittaukset väliltä 0–140 cm jätettiin aineistosta pois. Lopullinen aineisto koostuu siis 25:stä, tasavälein mitatusta rungosta. Mittauksia rungoilla on eri määrä, joka vaihtelee välillä 13–83. (Nummi & Möttönen 2004b.)

Tässä tutkielmassa 25 puun datasta on poimittu esimerkkiaineistoksi puu numero 6, jolla on yhteensä 83 tasavälein tehtyä runkomittausta. Työssä esimerkkipuun avulla on pyritty havainnollistamaan riippumattomien havaintojen odotusarvokäyrän mallintamista. Riippumattomuusoletus ei tietenkään voi täysin pitää paikkaansa aineistossa, jossa samalle puulle on tehty toistuvia mittauksia. Koska esimerkin tarkoituksena on kuitenkin ainoastaan antaa graafinen esitys splinin sopivuudesta aineistoon, ei korrelaatiorakenteen poisjättöä voida pitää ongelmana.

2.2 Sonni-aineisto

Luvussa 4 havainnollistetaan kovarianssimatriisin mallinnusta 40 sonnin esimerkkiaineiston avulla, joka on osa suurempaa, 2712 sonnin aineistoa. Alkuperäinen aineisto koostuu 2136 Ayshire-, 338 Suomenkarja- ja 238 Frisian-rotuisesta sonnista, joiden painon kehitystä on seurattu vuosien 1965–1977 välisenä aikana Suomessa. Vuosina 1965–1969 mittaukset sonneille suoritettiin 30–365 päivän iässä 30 päivän välein (yhteensä 12 aikapistettä). Vuosien 1970–1974 välisenä aikana testausajanjakso oli 60–365 päivää ja eläinten painot mitattiin ensin 60 ja 80 päivän iässä, jonka jälkeen loput mittaukset tehtiin 30 päivän välein 365 päivän ikään asti (yhteensä 8 mittauspistettä). Lopuksi vuosina 1975–1977 sonnit mitattiin ainoastaan 4 aikapisteessä: ikäpäivinä 60, 180, 270 ja 365. (Liski 1987.)

Alkuperäinen aineisto ei ole täydellinen, vaan erityisesti suuri määrä 30 päivän iässä tehtyjä mittauksia vuosien 1966–1969 välisenä aikana puuttuu. Useita puuttuvia mittauksia esiintyy myös vuosien 1970–1974 ikäpäivinä 90, 120 ja

150. Vuosien 1965–1967 välisenä aikana aineisto sisältää ainoastaan Ayshire- ja Suomenkarja-rotuisia sonneja ja vuonna 1968 ainoastaan 4 Frisian-rotuista sonnia. Yksityiskohtaisempaa tietoa aineistosta on saatavilla julkaisuista Linström ja Maijala (1970) sekä Liski (1987).

Tämän työn esimerkkiaineistoon on otettu mukaan vuonna 1966 syntyneet Suomenkarja-rotuiset sonnit, joita on yhteensä 40 kappaletta. Mittauspisteitä on yhteensä 12 ja ne on tehty tasaisin välein ja kaikille yksilöille samoina ajanhetkinä. Myöskään puuttuvia havaintoja ei esiinny, joten aineisto on tasapainoinen ja täydellinen. Kyseessä on siis niin kutsuttu kasvukäyräaineisto (ks. alaluku 3.3.4).

3 Odotusarvon estimointi

Oletetaan aluksi poikkileikkausaineiston tilanne, jossa satunnaisotannalla poimituista yksilöistä on saatu kustakin yksi mittaus. Aineistossa on siis n riippumatonta havaintoa ja tavoitteena on estimoida mallin

$$y_j = f(t_j) + \epsilon_j, \qquad j = 1, ..., n, \tag{3.1}$$

odotusarvokäyrä $f(t)$. Mikäli käyrää $f(t)$ ei voida tyydyttävästi mallintaa parametrisen regressiomallin avulla, voidaan ongelmaa lähestyä epäparametristen menetelmien kautta. Malli (3.1) on nyt niin kutsuttu *yksinkertainen epäparametrinen regressiomalli (simple nonparametric regression model)*, missä vastemuuttujan y_j, $j = 1, ..., n$ arvot on mitattu erillisissä mittauspisteissä t_j, $j = 1, ..., n$. Lisäksi virhetermien ϵ_j, $j = 1, ..., n$ oletetaan olevan riippumattomia ja normaalisti jakautuneita $\epsilon_j \sim N(0, \sigma^2)$. Matemaattisesti funktio $f(t)$ on vastemuuttujan y_j ehdollinen odotusarvo, kun t_j oletetaan tunnetuksi

$$f(t) = \mathrm{E}(y_j | t_j = t), \qquad j = 1, ..., n.$$

(Wu & Zhang 2006.)

Pitkittäisaineisto puolestaan koostuu usean yksilön toistuvista mittauksista, joiden sijainti ja määrä saattavat vaihdella eri yksilöiden välillä. Lisäksi pitkittäisaineiston ominaispiirteenä on, että yksilöiden väliset mittaukset oletetaan toisistaan riippumattomiksi, mutta sisäisten mittausten ei voida olettaa olevan riippumattomia. Nyt epäparametrinen regressiomalli voidaan määritellä seuraavasti

$$y_{ij} = f(t_{ij}) + \epsilon_i(t_{ij}), \qquad j = 1, 2, ..., n_i; \quad i = 1, ..., n, \tag{3.2}$$

missä y_{ij} on i. yksilön j. vastemuuttujan arvo, kun $i = 1, ..., n$, $j = 1, ..., n_i$ ja t_{ij} on vastaava mittausajankohta sekä $\epsilon_i(t_{ij})$ vastaava virhetermi. Lisäksi n_i ilmaisee i. yksilön mittausten määrän. Kuten mallissa (3.1), nyt myös $f(t)$ on kiinteä, mutta tuntematon populaation odotusarvofunktio, joka voidaan estimoida aineiston perusteella. Wu ja Zang (2006) kutsuvat mallia (3.2) *epäparametriseksi populaation odotusarvomalliksi (nonparametric population mean model)*. Malli (3.2) on verrattavissa yksinkertaiseen epäparametriseen regressiomalliin (3.1) sillä erotuksella, että mallin (3.2) mittausvirheet eivät ole riippumattomia, vaan $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \boldsymbol{R}_i$, missä $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})'$ ja $\sigma^2 \boldsymbol{R}_i$ on yksilön i kovarianssimatriisi. (Wu & Zhang 2006.)

Tässä luvussa esitellään kaksi epäparametrista lähestymistapaa mallintaa odotusarvofunktiota $f(t)$. Käsiteltävät menetelmät ovat alaluvussa 3.1 esitettävä regressiosplini sekä alaluvuissa 3.2 ja 3.3 käsiteltävä tasoittava splini. Erityisen mielenkiinnon kohteena on tasoittava kuutiosplini. Koska tasoittavan kuutiosplinin ja lineaarisen sekamallin välillä on löydettävissä yhteys, myös lineaarinen sekamalli esitellään alaluvussa 3.4.

3.1 Regressiosplini

Regressiosplini on eräs tasoitusmenetelmä, jolla voidaan estimoida aineiston odotusarvokäyrää. Tässä alaluvussa osoitetaan, kuinka odotusarvokäyrää voidaan mallintaa regressiosplinin avulla aineiston koostuessa riippumattomista havainnoista. Päähuomion saa erityisesti kuutiollinen regressiosplini, jonka muodostus käydään luvussa läpi yksityiskohtaisesti. Koska regressiosplinissä solmukohtien määrän ja sijainnin määrittäminen osoittautuu keskeiseksi tehtäväksi, esitellään luvussa myös joitakin yleisesti tunnettuja menetelmiä, joiden avulla solmuongelma voidaan ratkaista.

Pitkittäisaineiston regressiospliniä ei tässä työssä käsitellä, mutta kattavan esityksen kyseisestä aiheesta tarjoavat muun muassa Wu ja Zhang (2006).

3.1.1 Regressiosplinin muodostus

Oletetaan siis riippumattomien havaintojen aineisto, jolloin päämääränä on estimoida mallin (3.1) odotusarvokäyrä $f(t)$. Olkoot pisteet

$$\tau_0, \tau_1, \tau_2, ..., \tau_K, \tau_{K+1}$$

määriteltyjä välillä $[a, b]$ siten, että $a = \tau_0 < \tau_1 < ... < \tau_K < \tau_{K+1} = b$. Näitä pisteitä on tapana kutsua *solmuiksi*. Solmut jakavat välin $[a, b]$ osaväleihin

$$[\tau_r, \tau_{r+1}), \qquad r = 0, 1, ..., K,$$

siten, että jokaisen vierekkäisen solmukohdan τ_r ja τ_{r+1}, $r = 0, 1, ..., K$, välissä kulkee p-asteinen polynomi. Solmukohdissa käyrä ja sen $p - 1$ derivaattaa ovat jatkuvia, mutta viimeinen derivaatta voi olla epäjatkuva. Astetta p oleva regressiosplini solmukohdilla τ_r, $r = 0, 1, ..., K$ on siis p-asteinen paloittain määritelty polynomi, jonka polynomisegmentit liittyvät yhteen solmuissa $1, ..., K$. (Wu & Zhang 2006.)

Regressiosplini voidaan muodostaa niin kutsutun p asteisen *typistetyn potenssikannan (truncated power basis)*

$$1, t, t^2, ..., t^p, (t - \tau_1)_+^p, (t - \tau_2)_+^p, ..., (t - \tau_K)_+^p \tag{3.3}$$

avulla, missä $w_+ = \max(0, w)$ sekä $w_+^p = [w_+]^p$ on w:n positiivinen osa. Voidaan havaita, että kannan (3.3) ensimmäiset $(p+1)$ funktiota ovat polynomeja

asteeseen p asti ja loput kantafunktiot ovat p asteen typistettyjä potenssifunktioita. Nyt p asteinen regressiosplini voidaan lausua muodossa

$$f(t) = \sum_{s=0}^{p} \beta_s t^s + \sum_{r=1}^{K} \beta_{p+r}(t - \tau_r)_+^p, \tag{3.4}$$

missä $\tau_1, ..., \tau_K$ ovat regressiosplinin solmukohdat ja $\beta_0, \beta_1, ..., \beta_{p+K}$ vastaavat kertoimet. Tavallisesti valitaan $p = 1, 2$, tai 3, jolloin tuloksena saadaan lineaarinen, neliöllinen tai kuutiollinen regressiosplini. (Wu & Zhang 2006.)

Kantafunktiota (3.3) kutsutaan myös "+"-funktioksi (Smith 1979) ja sen avulla regressiosplini on helppo muodostaa. Toinen suosittu kantafunktio on B-splinikanta (de Boor 1978) ja myös esimerkiksi kopioivan ytimen Hilbertin avaruutta (reproducing kernel Hilbert space) voidaan käyttää regressiosplinin kantana (Wahba 1990). Monimutkaisemmat kannat ovat usein laskennallisesti kantaa (3.3) tehokkaampia, mutta kannan (3.3) etuna on sen antama selvä tilastollinen tulkinta regressiosplinille: regressiosplini voidaan sovittaa aineistoon käyttämällä pienimmän neliösumman menetelmää (Smith 1979). Tähän palataan tarkemmin seuraavassa alaluvussa.

3.1.2 Kuutiollinen regressiosplini

Kuutiollisten splinien on väitetty olevan matala-asteisimpia splinifunktioita, joissa solmukohtien epäjatkuvuutta ei voida silmin havaita. Toisaalta tuskin koskaan on syytä käyttää yli kolmannen asteen splinejä, minkä johdosta kuutiolliset splinit ovatkin käytännön sovelluksissa suosittuja valintoja. (Hastie, Tibshirani & Friedman 2001.) Jatkossa keskitytäänkin kuutiolliseen regressiospliniin, jolloin (3.4) saa muodon

$$f(t) = \sum_{s=0}^{3} \beta_s t^s + \sum_{r=1}^{K} \beta_{3+r}(t - \tau_r)_+^3. \tag{3.5}$$

Olkoon kannalla (3.3) nyt vektoriesitys

$$\boldsymbol{\Phi}_k(t) = \left[1, t, t^2, t^3, (t - \tau_1)_+^3, ..., (t - \tau_K)_+^3\right]',$$

missä $k = K + p + 1 = K + 4$ on kantafunktioiden lukumäärä. Vastaavasti merkitään malliin liittyviä kertoimia

$$\boldsymbol{\beta} = [\beta_0, ..., \beta_3, \beta_4, ..., \beta_{3+K}]'.$$

Kuutiollinen regressiosplini (3.5) voidaan nyt lausua muodossa

$$f(t) = \boldsymbol{\Phi}_k(t)'\boldsymbol{\beta},$$

jolloin mallille (3.1) saadaan lineaarisen mallin matriisiesitys

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.6}$$

missä

$$\begin{aligned}\boldsymbol{y} &= (y_1, ..., y_n)',\\ \boldsymbol{X} &= (\boldsymbol{\Phi}_k(t_1), ..., \boldsymbol{\Phi}_k(t_n))',\\ \boldsymbol{\epsilon} &= (\epsilon_1, ..., \epsilon_n)'.\end{aligned}$$

Oletetaan, että $n \geq k$ ja että $\boldsymbol{X}$:llä on täysi sarakeaste. Tällöin myös $\boldsymbol{X'X}$ on täysiasteinen ja siis ei-singulaarinen. Luonnollinen estimaattori lineaarisen mallin (3.6) kertoimille $\boldsymbol{\beta}$ on tavallinen pienimmän neliösumman (ordinary least squares, OLS) estimaattori

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}.$$

Vastaava funktion $f(t)$ sovite on

$$\hat{f}_k(t) = \boldsymbol{\Phi}_k(t)'(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}, \tag{3.7}$$

jota tavallisesti kutsutaan f:n *regressiosplinitasoittimeksi (regression spline smoother)*. Erityisesti mittauspisteissä t_j, $j = 1, ..., n$ saadaan

$$\hat{\boldsymbol{y}}_k = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \boldsymbol{A}_k\boldsymbol{y},$$

missä $\hat{\boldsymbol{y}}_k = (\hat{y}_1, ..., \hat{y_n})'$, kun $\hat{y}_j = \hat{f}_k(t_j)$, $j = 1, ..., n$ ja matriisia

$$\boldsymbol{A}_k = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} \tag{3.8}$$

kutsutaan *regressiosplinin tasoittajamatriisiksi (regression spline smoother matrix)*. Matriisi (3.8) on idempotentti ja symmetrinen, eli se toteuttaa ehdot $\boldsymbol{A}_k^2 = \boldsymbol{A}_k$ ja $\boldsymbol{A}_k' = \boldsymbol{A}_k$. Lisäksi df $= \text{tr}(\boldsymbol{A}_k) = k$, eli matriisin $\boldsymbol{A}_k$ jälki kertoo sekä regressiosplinin (3.7) kantafunktioiden, että vapausasteiden lukumäärän. (Wu & Zhang 2006.)

3.1.3 Solmukohtien valinta

Solmukohtien määrän ja sijainnin oikeanlainen määrittäminen on tärkein osa onnistunutta regressiosplinin sovitusta. Mikäli solmut sijoitetaan huonosti, saatetaan menettää joitakin yksityiskohtia käyrästä. Liian monen solmukohdan valinta taas johtaa sovitetun splinin suureen paikalliseen vaihteluun. (Smith & Kohn 1996.)

Erilaisia menetelmiä solmukohtien määrän ja sijainnin määrittämiseksi on kehitetty useita. Wun ja Zhangin (2006) esitystä seuraten, tässä alaluvussa esitellään lyhyesti kolme laajalti käytettyä menetelmää määrittää solmukohtien sijainti. Enemmän solmujen valintamenetelmistä ovat kirjoittaneet muun muassa Friedman ja Silverman (1989), Friedman (1991) sekä Smith ja Kohn (1996).

Solmut tasavälein. Tämä menetelmä ottaa solmukohdiksi K tasavälein määriteltyä pistettä kiinnostuksen kohteena olevalta väliltä $[a, b]$. Solmut määritellään seuraavasti:

$$\tau_r = a + (b - a)r/(K + 1), \qquad r = 1, 2, ..., K.$$

Menetelmä on riippumaton mittauspisteistä t_j, $j = 1, ..., n$ ja toimii yleensä hyvin silloin, kun mittauspisteet ovat tasaisesti jakautuneita välillä $[a, b]$.

Solmuina otoskvantiilit. Tämä menetelmä käyttää tasaisin välein määriteltyjä, mittauspisteiden t_j, $j = 1, 2, ..., n$ otoskvantiileja solmuina. Olkoot $t_{(1)}, ..., t_{(n)}$ mittauspisteiden järjestystunnusluvut. Tällöin K solmut määritellään kaavalla:

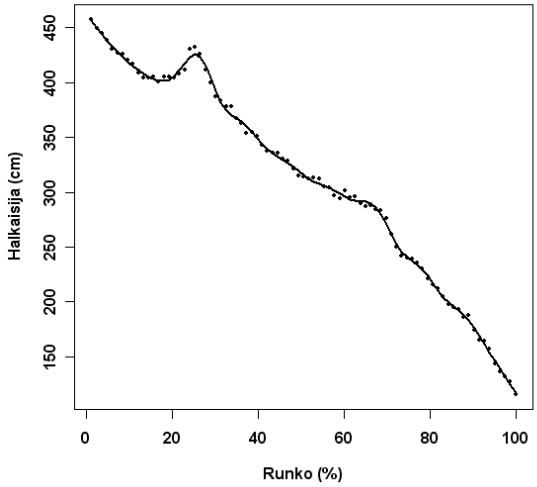
$$\tau_r = t_{(1+[rn/(K+1)])}, \qquad r = 1, 2, ..., K.$$

Oheinen menetelmä on mukautuva, sillä se asettaa enemmän solmuja sinne, missä mittauspisteitä sijaitsee enemmän. Mikäli mittauspisteet ovat tasaisesti jakautuneita, antaa menetelmä suurin piirtein samat solmukohdat kuin tasavälinen menetelmä.

Mallinvalintapohjainen menetelmä. Tämä menetelmä käyttää kaikkia mittauspisteitä solmuehdokkaina. Kun regressiosplinin muodostuksessa käytetään typistettyä potenssikantaa (3.3), tarkoittaa yhden solmun poisjättö samalla solmukohtaan liittyvän typistetyn potenssikannan funktion poisjättöä, mikä taas on yhtäpitävää lineaarisen mallin (3.6) yhden kovariaatin poistamisen kanssa. Tästä syystä solmujen valinta voidaan tehdä käyttämällä mallinvalinnan menetelmiä, kuten askeltavaa regressiota, taaksepäin eliminointia tai eteenpäin valintaa.

Yleisesti voidaan sanoa, että solmujen lukumäärän K tulisi olla pienempi kuin otoskoon n. Jos käytetään esitetyistä menetelmistä viimeistä, tulee solmujen määrä valituksi samalla kertaa, kun määritellään solmukohtien sijainti. Kaksi edellistä menetelmää vaativat kuitenkin solmujen määrän K päättämistä erikseen. Tämä voidaan tehdä erilaisten tasoitusparametrin valintamenetelmien, kuten ristiinvalidoinnin avulla, joihin palataan tarkemmin alaluvussa 3.2.3. (Wu & Zhang 2006.)

Kuviossa 3.1 on esimerkkipuulla havainnollistettu, kuinka kuutiollinen regressiosplini käytännössä estimoi aineistoa. Koska esimerkkiaineistossa runkokäyrän mittauspisteet ovat tasaisesti jakautuneita, on solmukohtien sijainti määritelty tasavälisellä menetelmällä. Solmujen lukumäärä on puolestaan valittu ristiinvalidointia käyttäen. Solmujen määräksi on tällöin saatu $K = 18$ otoskoon ollessa $n = 83$. Voidaan havaita, että kyseisillä menetelmillä saadaan melko hyvin aineistoon sopiva käyrä.



Kuvio 3.1. Esimerkkipuulle sovitettu regressiosplini kun K=18.

3.2 Tasoittava splini

Kuten alaluvun 3.1.3 perusteella voidaan todeta, regressiosplinin solmukohtien määrän ja sijainnin määrittäminen ei aina ole yksiselitteinen ja helppo tehtävä. Ongelma kuitenkin ratkeaa, kun siirrytään käyttämään tasoittavaa spliniä, sillä tasoittavassa splinissä kaikki aineiston mittauspisteet toimivat solmuina. Sovitetun käyrän tasaisuutta kontrolloidaan nyt sakkotermin avulla ja ainoa parametri, joka täytyy erikseen määritellä, on tasoitusparametri λ. (Wu & Zhang 2006.)

Tässä luvussa keskitytään erityisesti luonnolliseen tasoittavaan kutiospliniin ja osoitetaan, kuinka kyseinen splini voidaan muodostaa aineiston koostuessa riippumattomista havainnoista. Lisäksi käydään läpi menetelmiä, joiden avulla tasoitusparametri voidaan helposti määritellä.

3.2.1 Sakotettu pienimmän neliösumman kriteeri

Tarkastellaan edelleen yksinkertaista epäparametrista regressiomallia (3.1). Olkoot $t_1, ..., t_n$ välin $[a, b]$ nousevaan järjestykseen asetetut mittauspisteet siten, että $a < t_1 < ... < t_n < b$. Olkoon lisäksi f mallin (3.1) funktio, joka on p kertaa derivoituva ja määritelty välillä $[a, b]$. Nyt käyrän f estimaattori $\hat{f}(t)$ minimoi niin kutsutun *sakotetun pienimmän neliösumman kriteerin (penalized least square (PLS) criterion)*

$$S(f) = \sum_{j=1}^{n} [y_j - f(t_j)]^2 + \lambda \int_a^b \left[f^{(p)}(t)\right]^2 \mathrm{dt} \tag{3.9}$$

yli p asteisen Sobolevin avaruuden $\mathcal{W}_2^p[a,b]$:

$$\left\{ f : f^{(r)} \text{on absoluuttisesti jatkuva kun } 0 \leq r \leq p-1, \quad \int_a^b \{f^p(t)\}^2 \, \mathrm{dt} < \infty \right\}.$$

Mittauspisteet $t_1, ..., t_n$ ovat nyt kaikki kriteerin (3.9) minimoivan käyrän $\hat{f}(t)$ solmuja $\tau_1, ..., \tau_K$ siten, että $t_1 = \tau_1, ..., t_n = \tau_K$. (Wu & Zhang 2006.)

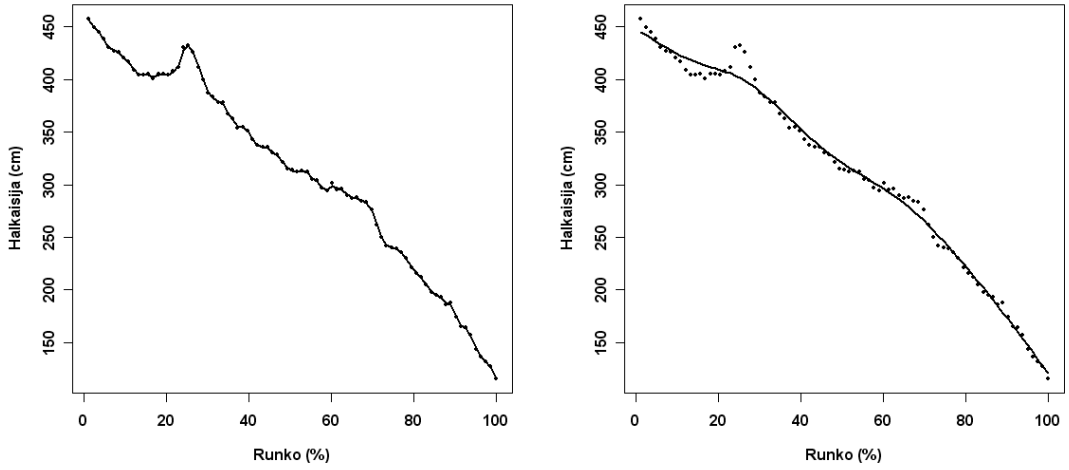
PLS-kriteerin (3.9) ensimmäinen termi

$$\sum_{j=1}^{n} [y_j - f(t_j)]^2 ,$$

on jäännösneliösumman termi, joka minimoimalla saataisiin aineiston interpoloiva käyrä. Tilastollisesti tällainen malli ei kuitenkaan ole erityisen järkevä käyrän nopean heilahtelun vuoksi. Nopeaa paikallista vaihtelua voidaan tasoittaa *rosoisuuden sakkotermillä (roughness penalty term)*

$$\lambda \int_a^b \left[f^{(p)}(t) \right]^2 \mathrm{dt}, \tag{3.10}$$

missä integraali $\int_a^b \left[f^{(p)}(t) \right]^2 \mathrm{dt}$ edustaa käyrän rosoisuutta ja $\lambda > 0$ on tasoitusparametri, joka säätelee rosoisuuden määrää. Kun λ saa pienen arvon, kulkee sovitettu käyrä tarkasti havaintopisteiden kautta. Vastaavasti suurilla λ:n arvoilla painoarvon saa rosoisuuden termi, jolloin käyrä lähenee lineaarista regressiota. Tasoitusparametrin arvoa säätelemällä voidaan löytää tasapaino näiden kahden ääripään väliltä. (Silverman 1985; Green & Silverman 1994.)



Kuvio 3.2. Esimerkkipuulle sovitettu käyrä tasoitusparametrin arvoilla $\lambda = 0.1$ ja $\lambda = 1500$.

Kuviossa 3.2 on havainnollistettu tasoitusparametrin vaikutusta sovitettuun käyrään. Voidaan havaita, kuinka pienellä tasoitusparametrin arvolla käyrä kulkee tarkasti mittauspisteiden kautta, kun taas suurella tasoitusparametrin arvolla estimoitu käyrä lähestyy suoraa.

3.2.2 Luonnollinen tasoittava kuutiosplini

PLS-kriteerin (3.9) minimointia ei voida pitää laskennallisesti kovinkaan helppona tehtävänä, sillä se edellyttää rosoisuuden sakkotermin (3.10) integraalin laskemista sekä parametrien estimointia. Green ja Silverman (1994) ovat kuitenkin osoittaneet, että mikäli $p = 2$, helpottuu käyrän estimointi huomattavasti. Palataan tähän myöhemmin ja tarkastellaan ensin hieman estimaattorin $\hat{f}(t)$ ominaisuuksia.

Oletetaan siis, että $p = 2$, jolloin (3.9) saa muodon

$$S(f) = \sum_{j=1}^{n} [y_j - f(t_j)]^2 + \lambda \int_a^b \left[f''(t)\right]^2 \mathrm{dt}. \tag{3.11}$$

Reinsch (1967) on osoittanut, että kriteerin (3.11) minimoivalla estimaattorilla $\hat{f}(t)$ on seuraavat ominaisuudet:

1. Funktio on kullakin välillä (t_j, t_{j+1}), $j = 1, ..., n-1$ kolmannen asteen polynomi.
2. Mittauspisteissä t_j käyrä itse ja sen ensimmäiset kaksi derivaattaa ovat jatkuvia, mutta kolmas derivaatta voi olla epäjatkuva.
3. Välillä $(-\infty, t_1)$ ja (t_n, ∞) toinen derivaatta on nolla, joten $\hat{f}(t)$ on lineaarinen määrittelyalueensa ulkopuolella.

Jokaista käyrää, joka toteuttaa ehdot 1 ja 2 kutsutaan kuutiospliniksi solmukohdilla t_j, $j = 1, ..., n$. Lisäksi käyrää, joka toteuttaa myös ehdon 3, kutsutaan luonnolliseksi kuutiospliniksi. Saavutettu estimaattori $\hat{f}(t)$ on siis *luonnollinen tasoittava kuutiosplini (natural cubic smoothing spline)*. Jatkossa käytetään usein lyhyempää muotoa tasoittava kuutiosplini, jolla kuitenkin aina tarkoitetaan nimenomaan luonnollista tasoittavaa kuutiospliniä.

Osoitetaan nyt Greenin ja Silvermanin (1994) esitystä seuraten, kuinka tasoittava kuutiosplini voidaan saavuttaa yksinkertaisin matriisioperaatioin. Olkoot $t_1 < t_2 < ... < t_n$ aineiston nousevaan järjestykseen asetetut mittauspisteet, jotka samalla toimivat tasoittavan kuutiosplinin $\hat{f}(t)$ solmukohtina. Merkitään

$$f_j = f(t_j) \quad \text{ja} \quad \gamma_j = f''(t_j), \qquad j = 1, ..., n.$$

Luonnollisen kuutiosplinin määritelmästä seuraa, että f:n toinen derivaatta pisteissä t_1 ja t_n on nolla, joten $\gamma_1 = \gamma_n = 0$. Määritellään nyt vektorit $\boldsymbol{f} = (f_1, ..., f_n)'$ ja $\boldsymbol{\gamma} = (\gamma_2, ..., \gamma_{n-1})'$ ja olkoot lisäksi

$$h_r = t_{r+1} - t_r, \qquad r = 1, 2, ..., n-1.$$

Olkoon nyt $\boldsymbol{Q}$ $n \times (n-2)$ matriisi, jonka kaikki alkiot saavat arvon 0 lukuun ottamatta alkioita

$$q_{r,r} = h_r^{-1}, \qquad q_{r+1,r} = -(h_r^{-1} + h_{r+1}^{-1}), \qquad q_{r+2,r} = -h_{r+1}^{-1},$$

kun $r = 1, 2, ..., n-2$.
Olkoon $\boldsymbol{B}$ vastaavasti symmetrinen $(n-2) \times (n-2)$ matriisi, jonka kaikki alkiot saavat arvon 0 lukuun ottamatta alkioita

$$b_{11} = (h_1 + h_2)/3, \qquad b_{21} = h_2/6,$$

ja

$$b_{n-3,n-2} = h_{(n-2)}/6, \qquad b_{n-2,n-2} = (h_{(n-2)} + h_{(n-1)})/3,$$

sekä r:n arvoilla $r = 1, 2, ..., n-4$

$$b_{r,r+1} = h_{(r+1)}/6, \qquad b_{r+1,r+1} = (h_{(r+1)} + h_{(r+2)})/3, \qquad b_{r+2,r+1} = (h_{(r+2)}/6).$$

Lineaarialgebran avulla voidaan osoittaa (esim. Todd 1962, kappale 8.19), että matriisi $\boldsymbol{B}$ on positiivisesti definiitti. Tämän vuoksi voidaan määritellä $n \times n$ niin kutsuttu *rosoisuusmatriisi (roughness matrix)* $\boldsymbol{G}$ siten, että

$$\boldsymbol{G} = \boldsymbol{Q}\boldsymbol{B}^{-1}\boldsymbol{Q}'. \tag{3.12}$$

Edellä esitetyt tulokset johtavat nyt hyvin käyttökelpoiseen lauseeseen.

Lause 3.1. *Vektorit $\boldsymbol{f}$ ja $\boldsymbol{\gamma}$ määrittelevät luonnollisen kuutiosplinin f jos ja vain jos yhtälö*

$$\boldsymbol{Q}'\boldsymbol{f} = \boldsymbol{B}\boldsymbol{\gamma} \tag{3.13}$$

pätee. Mikäli yhtälö (3.13) pätee, niin

$$\int_a^b f''(t)^2 dt = \boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma} = \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f}.$$

Lauseen todistuksen ovat esittäneet Green ja Silverman (2004, s. 24–25).

On selvää, että jäännösneliösumma voidaan kirjoittaa muodossa

$$\sum_{j=1}^{n} [y_j - f(t_j)]^2 = (\boldsymbol{y} - \boldsymbol{f})'(\boldsymbol{y} - \boldsymbol{f}),$$

jolloin PLS-kriteeri (3.11) saa matriisiesityksen

$$\begin{aligned} S(f) &= (\boldsymbol{y} - \boldsymbol{f})'(\boldsymbol{y} - \boldsymbol{f}) + \lambda \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f} \\ &= \boldsymbol{f}'(\boldsymbol{I} + \lambda\boldsymbol{G})\boldsymbol{f} - 2\boldsymbol{y}'\boldsymbol{f} + \boldsymbol{y}'\boldsymbol{y}. \end{aligned} \tag{3.14}$$

Matriisi $\lambda \boldsymbol{G}$ on ei-negatiivisesti definiitti, mistä seuraa, että matriisi $(\boldsymbol{I} + \lambda \boldsymbol{G})$ on positiivisesti definiitti. Näin ollen lauseke (3.14) saavuttaa yksikäsitteisen miniminsä, kun asetetaan

$$\hat{\boldsymbol{f}} = (\boldsymbol{I} + \lambda \boldsymbol{G})^{-1} \boldsymbol{y}. \tag{3.15}$$

Nyt siis $\hat{\boldsymbol{f}}$ on solmukohdissa t_j, $j = 1, ..., n$ estimoitu tasoittava kuutiosplini. Edelleen sovitettu vastevektori kaikissa mittauspisteissä on

$$\hat{\boldsymbol{y}} = \boldsymbol{A} \boldsymbol{y},$$

missä

$$\boldsymbol{A} = (\boldsymbol{I} + \lambda \boldsymbol{G})^{-1} \tag{3.16}$$

tunnetaan nimellä *tasoittavan kuutiosplinin tasoittajamatriisi (cupic smoothing spline smoother matrix)*. (Green & Silverman 1994; Wu & Zhang 2006.)

Kuten regressiosplinin tapauksessa, myös nyt pätee $\mathrm{df} = \mathrm{tr}(\boldsymbol{A})$. Tasoitusparametri määrittelee nyt tasoittavan kuutiosplinin asteen ja samalla mallin monimutkaisuuden siten, että

$$\begin{aligned} &\text{kun} \quad \lambda \to 0, \quad \mathrm{df} \to n, \\ &\text{kun} \quad \lambda \to \infty, \quad \mathrm{df} \to 2. \end{aligned}$$

(Hastie ym. 2001). Huomattakoon vielä, että toisin kuin regressiosplinin, tasoittavan splinin aste ei yleensä ole kokonaisluku.

3.2.3 Tasoitusparametrin valinta

Tasoitusparametri λ kontrolloi tasapainoa käyrän tasaisuuden ja yhteensopivuuden välillä, minkä vuoksi sen oikeanlainen valinta on luonnollisesti olennainen osa onnistunutta käyrän sovitusta. Monissa käytännön sovelluksissa voi riittää, että tasoitusparametri valitaan subjektiivisesti piirtämällä joitakin käyriä ja valitsemalla käyristä se, joka ”näyttää parhaalta”. Myös tutkimuksellisesta näkökulmasta tämä lähestymistapa saattaa osoittautua hyödylliseksi, sillä se suuntaa huomion sellaisiin kiinnostaviin piirteisiin, jotka paljastuvat vain tietyillä tasoitusparametrien arvoilla. (Silverman 1985.)

Kuitenkin löytyy myös lukuisia syitä, miksi tasoitusparametrin valinta on hyödyllistä suorittaa automaattisin menetelmin. Automaattiset menetelmät muun muassa helpottavat kokemattoman käyttäjän työtä ja antavat lähtöpisteen silloinkin, kun myöhemmät arviot tehdään subjektiivisesti. Tulosten raportoinnin ja vertailtavuuden kannalta standardoidut menetelmät ovat myös tärkeitä. Lisäksi mikäli tasoitusmenetelmää käytetään rutiininomaisesti moniin aineistoihin tai osana suurempaa proseduuria, ovat automaattiset menetelmät jopa välttämättömiä. (Silverman 1985.)

Tässä alaluvussa käsitellään kahta suosittua tasoitusparametrin valintamenetelmää, jotka ovat ristiinvalidointi (cross-validation) ja yleistetty ristiinvalidointi (generalized cross-validation). Muita menetelmiä ovat muun muassa

Akaiken informaatiokriteeri (AIC) ja Bayesin informaatiokriteeri (BIC) joiden käytön tasoitusparametrin valinnassa esittelevät muun muassa Wu ja Zhang (2006).

Ristiinvalidointi

Olkoon λ jokin kiinteä tasoitusparametrin arvo ja olkoon $\hat{f}^{-j}(t_j;\lambda)$ sovitettu arvo mittauspisteessä t_j, kun käyrän $\hat{f}^{-j}(t;\lambda)$ estimoinnissa on käytetty kaikkea muuta aineistoa, paitsi (t_j, y_j), kun $j = 1, ..., n$. Nyt ristiinvalidoinnin avulla valittu tasoitusparametri on se λ:n arvo, joka minimoi funktion

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left[y_j - \hat{f}^{-j}(t_j;\lambda)\right]^2. \tag{3.17}$$

Funktion (3.17) suora laskeminen on melko työlästä, sillä jokaiselle j:n arvolle on erikseen määriteltävä lokaali sovite $\hat{f}^{-j}(t_j;\lambda)$. Näin ollen (3.17) määrittäminen jokaiselle λ:n arvolle vaatii yhteensä n lokaalia sovitetta. Voidaan kuitenkin osoittaa (Green & Silverman 1994, s. 31–33), että (3.17) on mahdollista kirjoittaa muodossa

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j - \hat{f}(t_j)}{1 - \boldsymbol{A}_{jj}(\lambda)}\right)^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j - \hat{y}_j}{1 - \boldsymbol{A}_{jj}(\lambda)}\right)^2, \tag{3.18}$$

missä $\boldsymbol{A}_{jj}$ on tasoittajamatriisin (3.16) j. diagonaalielementti ja $\hat{f}$ on koko havaintoaineistosta estimoitu splinitasoittaja tasoitusparametrin arvolla λ. Täten (3.17) voidaan määritellä jokaiselle λ:n arvolle ainoastaan yhden lokaalin sovitteen avulla. (Green & Silverman 1994.)

Yleistetty ristiinvalidointi

Yleistetty ristiinvalidointi on ristiinvalidoinnin approksimaatio, jonka ensimmäisinä esittivät Wahba (1977) sekä Craven ja Wahba (1979). Yleistetyn ristiinvalidoinnin esitys saadaan ristiinvalidoinnin lausekkeesta (3.18) korvaamalla jokainen tasoittajamatriisin elementti $\boldsymbol{A}_{jj}$ keskiarvolla

$$\frac{1}{n}\sum_{j=1}^{n} a_{jj} = \frac{1}{n}\mathrm{tr}(\boldsymbol{A}) = \frac{\mathrm{df}}{n}.$$

Saadaan siis

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\sum_{j=1}^{n}\left[y_j - \hat{y}_j\right]^2}{\left[1 - \mathrm{tr}(\boldsymbol{A})/n\right]^2}. \tag{3.19}$$

Voidaan havaita, että lausekkeen (3.19) osoittaja edustaa käyrän yhteensopivuutta, kun taas nimittäjä ilmaisee käyrän monimutkaisuuden. Tästä seuraa, että valitsemalla λ siten, että lauseke (3.19) minimoituu, saavutetaan tasapaino käyrän yhteensopivuuden ja tasaisuuden välillä. (Wu & Zhang 2006.)

3.3 Tasoittava kuutiosplini pitkittäisaineistolle

Oletetaan nyt malli (3.2), jossa usealle yksilölle on tehty toistuvia mittauksia, jolloin yksilön sisäiset mittaukset ovat keskenään korreloituneita. Tasoittavan kuutiosplinin estimaattori korreloituneille havainnoille voidaan johtaa kahden eri menetelmän kautta. Muun muassa Wu ja Zhang (2006) esittelevät sakotettuun yleistettyyn pienimmän neliösumman kriteeriin (penalized generalized least squares criterion) perustuvan lähestymistavan. Verbyla, Cullis, Kenward ja Welham (1999) sekä Nummi ja Koskela (2007) käyttävät puolestaan sakotettua logaritmoitua uskottavuusfunktiota (penalized log-likelihood function) spliniestimaattorin saavuttamiseksi.

Tässä alaluvussa käydään läpi molemmat menetelmät ja osoitetaan lisäksi, kuinka pitkittäisaineiston tasoitusparametri voidaan valita. Lisäksi esitellään eräs pitkittäisaineiston erikoistapaus, kasvukäyrämalli ja johdetaan sille spliniestimaattori.

3.3.1 Sakotettu yleistetty pienimmän neliösumman kriteeri

Oletetaan mallin (3.2) mukaisesti, että n on aineiston riippumattomien yksilöiden lukumäärä ja olkoon $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})'$ i. yksilön vastearvojen vektori, kun $i = 1, ..., n$. Oletetaan lisäksi, että mittauspisteet $t_{i1}, ..., t_{in_i}$ on määritelty välillä $[a, b]$ siten, että $a < t_{i1} < ... < t_{in_i} < b$. Mallin (3.2) funktio $f(t)$ on tasoitettu, kahdesti derivoituva käyrä, jonka satunnaisvirheet ϵ_{ij} oletetaan normaalijakautuneiksi odotusarvolla nolla ja kovarianssimatriisilla $\sigma^2\boldsymbol{R}_i$. Vektoriesityksenä malli (3.2) voidaan lausua muodossa

$$\boldsymbol{y}_i = \boldsymbol{f}_i + \boldsymbol{\epsilon}_i, \tag{3.20}$$

missä $\boldsymbol{f}_i = [f(t_{i1}), ..., f(t_{in_i})]'$ ja $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})' \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{R}_i)$. (Nummi & Koskela 2007.)

Nyt käyrän $f(t)$ estimaattori $\hat{f}(t)$ minimoi *sakotetun yleistetyn pienimmän neliösumman kriteerin (penalized generalized least squares criterion)*

$$(\boldsymbol{y}_i - \boldsymbol{f}_i)'\boldsymbol{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{f}_i) + \lambda \int_a^b \left[f''(t)\right]^2 \mathrm{dt} \tag{3.21}$$

yli toisen asteen Sobolevin avaruuden $\mathcal{W}_2^2[a, b]$. Kriteerissä (3.21) λ on (kuten alaluvussa 3.2.1) käyrän yhteensopivuutta (ensimmäinen termi) ja tasaisuutta (toinen termi) kontrolloiva tasoitusparametri ja estimaattori $\hat{f}(t)$ on tasoittava kuutiosplini solmukohdilla $t_{i1}, ..., t_{in_i}$. Edelleen voidaan sakotettu yleistetty pienimmän neliösumman kriteeri kaikille pitkittäisaineiston yksilöille $1, ..., n$ esittää muodossa

$$\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{f}_i)'\boldsymbol{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{f}_i) + \lambda \int_a^b \left[f''(t)\right]^2 \mathrm{dt}, \tag{3.22}$$

missä kriteerin (3.22) minimoiva, funktion $f(t)$ estimaattori $\hat{f}(t)$ on tasoittava kuutiosplini, jossa solmuina toimivat kaikkien yksilöiden erilliset mittauspisteet $\tau_1, \tau_2, ..., \tau_M$. (Wu & Zhang 2006.)

Kuten alaluvussa 3.2.2 osoitettiin, sakotetun pienimmän neliösumman kriteerin (3.21) rosoisuustermi voidaan kirjoittaa muodossa

$$\int_a^b \left[f''(t)\right]^2 \mathrm{dt} = \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f},$$

missä nyt $[f(\tau_1), ..., f(\tau_M)]' = \boldsymbol{f}$ on vektori, joka saavutetaan estimoimalla odotusarvofunktio $f(t)$ kaikissa solmukohdissa $\tau_1, ..., \tau_M$ ja $\boldsymbol{G}$ on vastaava identiteetillä (3.12) määritelty rosoisuusmatriisi. Käyrän $f(t)$ arvo $f(t_{ij})$ jokaisessa mittauspisteessä t_{ij} voidaan ilmaista $\boldsymbol{f}$:n avulla siten, että

$$\begin{aligned} f(t_{ij}) &= \boldsymbol{x}_{ij}'\boldsymbol{f}, \qquad \boldsymbol{x}_{ij} = (x_{ij1}, ..., x_{ijM})', \\ x_{ijr} &= 1, \quad \text{jos } t_{ij} = \tau_r \text{ ja 0 muuten kaikilla } r = 1, ..., M. \end{aligned}$$

Toisin sanoen $\boldsymbol{x}_{ij}$ on indikaattorivektori, joka ilmaisee kulloistakin mittauspistettä t_{ij} vastaavan solmukohdan τ_r. Asettamalla

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i})', & \boldsymbol{X} &= (\boldsymbol{X}_1', ..., \boldsymbol{X}_n')', \\ \boldsymbol{y}_i &= (y_{i1}, ..., y_{in_i})', & \boldsymbol{y} &= (\boldsymbol{y}_1', ..., \boldsymbol{y}_n')', \end{aligned}$$

voidaan kriteeri (3.22) lausua muodossa

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{f})' \boldsymbol{R}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{f}) + \lambda \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f},$$

tai yhtäpitävästi

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{f})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{f}) + \lambda \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f}, \tag{3.23}$$

missä $\boldsymbol{R} = \mathrm{diag}(\boldsymbol{R}_1, ..., \boldsymbol{R}_n)$. Estimaattori, joka minimoi lausekkeen (3.23), on

$$\hat{\boldsymbol{f}}_{gss} = \left(\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} + \lambda\boldsymbol{G}\right)^{-1} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{3.24}$$

ja erityisesti mittauspisteissä $t_{ij}, j = 1, 2, ..., n_i; i = 1, 2, ..., n$ saadaan

$$\hat{\boldsymbol{y}}_{gss} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} + \lambda\boldsymbol{G}\right)^{-1} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} = \boldsymbol{A}_{gss}\boldsymbol{y}.$$

(Wu & Zhang 2006.)

Wu & Zhang (2006) kutsuvat yllä kuvattua menetelmää *yleistetyksi splinitasoitusmenetelmäksi (generalized smoothing spline (GSS) method)* ja estimaattoria (3.24) *kuutiolliseksi GSS-estimaattoriksi (cupic GSS estimator)*. Matriisia $\boldsymbol{A}_{gss}$ voidaan vastaavasti kutsua *GSS-estimaattorimatriisiksi (GSS estimator matrix)*.

Wang (1998b) esitti ensimmäisenä yllä kuvatun GSS-menetelmän, mutta jo ennen tätä Rice ja Silverman (1991) sekä Hoover, Rice, Wu & Yang (1998) osoittivat, kuinka pitkittäisaineistoa voidaan mallintaa tasoittavalla kuutiosplinillä, kun aineiston korreloituneisuutta ei oteta huomioon. Wu ja Zhang (2006) kutsuvat tätä *yksinkertaiseksi splinitasoitusmenetelmäksi (naive smoothing spline (NSS) method)*. Voidaan olettaa, että NSS-menetelmä toimii tilanteissa, joissa pitkittäisaineisto ei ole vahvasti korreloitunut. Tällöin saamme niin kutsutun *kuutiollisen NSS-estimaattorin (cupic NSS estimator)*

$$\hat{\boldsymbol{f}}_{nss} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{G})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{3.25}$$

ja erityisesti mittauspisteissä $t_{ij}, j = 1, 2, ..., n_i;\ i = 1, 2, ..., n$ saadaan

$$\hat{\boldsymbol{y}}_{nss} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{G}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}_{nss}\boldsymbol{y}.$$

Matriisia $\boldsymbol{A}_{nss}$ voidaan nyt kutsua *NSS-tasoittajamatriisiksi (NSS smoother matrix)*.

Rice ja Silverman (1991) toteavat, että mikäli jokaisella yksilöllä on samat mittauspisteet $t_{ij} = \tau_j, j = 1, ..., M$ ja $n_i = M, i = 1, ..., n$ (kasvukäyräaineisto), on kuutiollinen NSS-estimaattori $\hat{f}(t)$ yhtäpitävä tasoittavan kuutiosplinin kanssa, joka saavutetaan, kun tasoitettavana aineistona toimivat mittauspisteissä $\tau_1, ..., \tau_M$ lasketut vasteiden $\boldsymbol{y}_i$ keskiarvot, kun $i = 1, ..., n$. Tämä voidaan osoittaa, kun huomioidaan, että kyseiselle aineistolle $\boldsymbol{X}_i = \boldsymbol{I}_M$. Tällöin NSS estimaattori (3.25) sievenee muotoon

$$\begin{aligned}
\hat{\boldsymbol{f}}_{nss} &= \left(\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{X}_i + \lambda\boldsymbol{G}\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{y}_i \\
&= (n\boldsymbol{I}_M + \lambda\boldsymbol{G})^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i \\
&= (\boldsymbol{I}_M + n^{-1}\lambda\boldsymbol{G})^{-1}\bar{\boldsymbol{y}}, \qquad (3.26)
\end{aligned}$$

missä $\bar{\boldsymbol{y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i$, kun $i = 1, 2, ..., n$. (Wu & Zhang 2006.)

3.3.2 Tasoitusparametrin valinta

Alaluvussa 3.2.3 osoitettiin, kuinka tasoitusparametri λ voidaan valita yleistetyn ristiinvalidoinnin avulla havaintojen ollessa riippumattomia. Vastaava menetelmä voidaan laajentaa myös pitkittäisaineistolle. Tässä alaluvussa esitellään menetelmän pääpiirteet. Yksityiskohtaisemman esityksen todistuksineen antavat Wu ja Zhang (2006, s. 153–154, 158).

GSS-estimaattorille (3.24) voidaan yleistetyn ristiinvalidoinnin lauseke kirjoittaa muodossa

$$\mathrm{GCV}_{gss}(\lambda) = \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{f}}_{gss})'\boldsymbol{R}_i^{-1/2}\boldsymbol{W}_i\boldsymbol{R}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{f}}_{gss})}{\left[1 - \mathrm{tr}(\boldsymbol{A}_{gss})/N\right]^2}, \tag{3.27}$$

missä $\boldsymbol{W} = \text{diag}(\boldsymbol{W}_1, ..., \boldsymbol{W}_n)$ on painomatriisi ja $N = \sum_{i=1}^{n} n_i$. Vastaavasti NSS-estimaattorille (3.25) saadaan

$$\text{GCV}_{nss}(\lambda) = \frac{\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{f}}_{nss})' \boldsymbol{W}_i (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{f}}_{nss})}{[1 - \text{tr}(\boldsymbol{A}_{nss})/N]^2}. \tag{3.28}$$

Painomatriisi $\boldsymbol{W}$ voidaan spesifioida kolmen erilaisen mallin mukaan.

1. Asetetaan $\boldsymbol{W}_i = N^{-1} \boldsymbol{I}_{n_i}$, jolloin kaikki mittaukset ovat samoin painotettuja.
2. Asetetaan $\boldsymbol{W}_i = (n n_i)^{-1} \boldsymbol{I}_{n_i}$, jolloin yksilöiden sisäiset mittaukset ovat samoin painotettuja, mutta yksilöiden välillä on painotuseroja sen mukaan, montako mittausta kullakin yksilöllä on.
3. Asetetaan $\boldsymbol{W}_i = \boldsymbol{R}_i^{-1}$, missä $\boldsymbol{R}_i = \text{Cov}(\boldsymbol{y}_i)$, jolloin ryhmien sisäinen korrelaatio otetaan huomioon.

NSS-estimaattorin (3.25) tapauksessa on luonnollista määritellä $\boldsymbol{W}$ mallien (1) tai (2) mukaan. GSS-estimaattorille (3.24) malli (3) on paras vaihtoehto. (Wu & Zhang 2006.)

3.3.3 Sakotettu logaritmoitu uskottavuusfunktio

Esitetään seuraavaksi *sakotettuun logaritmoituun uskottavuusfunktioon (penalized log-likelihood)* pohjautuva korreloituneen aineiston splinitasoitusmenetelmä. Oletetaan edelleen alaluvussa 3.3.1 määritelty malli (3.20) vastaavin oletuksin. Nyt tavoitteena on valita $\boldsymbol{f}_i$ siten, että (vakiota vaille) sakotettu logaritmoitu uskottavuusfunktio

$$2l = -\log\left[\det(\sigma^2 \boldsymbol{R}_i)\right] - \frac{1}{\sigma^2}\left[(\boldsymbol{y}_i - \boldsymbol{f}_i)' \boldsymbol{R}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{f}_i) + \lambda \int \{f_i''(t)\}^2 \, \text{dt}\right]$$

maksimoituu. Myös nyt λ on käyrän rosoisuutta kontrolloiva tasoitusparametri. Kiinteillä λ ja $\sigma^2 \boldsymbol{R}_i$ arvoilla maksimoinnin tuloksena saavutettu estimaattori $\hat{\boldsymbol{f}}_i$ on tasoittava kuutiosplini solmukohdissa $t_1, ..., t_{n_i}$ ja se voidaan määritellä lausekkeena

$$\hat{\boldsymbol{f}}_i = (\boldsymbol{R}_i^{-1} + \lambda \boldsymbol{G}_i)^{-1} \boldsymbol{R}_i^{-1} \boldsymbol{y}_i. \tag{3.29}$$

Edelleen $\boldsymbol{G}_i$ on identiteetillä (3.12) määritelty rosoisuusmatriisi $\boldsymbol{G}_i = \boldsymbol{Q}_i \boldsymbol{B}_i^{-1} \boldsymbol{Q}_i'$, kun solmukohdat ovat $t_1, ..., t_{n_i}$. (Verbyla ym. 1999.)

Helposti nähdään, että lauseke (3.29) voidaan esittää muodossa

$$\hat{\boldsymbol{f}}_i = (\boldsymbol{I} + \lambda \boldsymbol{R}_i \boldsymbol{Q}_i \boldsymbol{B}_i^{-1} \boldsymbol{Q}_i')^{-1} \boldsymbol{y}. \tag{3.30}$$

Nyt, mikäli $\boldsymbol{R}_i$ täyttää ehdon

$$\boldsymbol{R}_i \boldsymbol{Q}_i = \boldsymbol{Q}_i, \tag{3.31}$$

lauseke (3.30) sievenee muotoon

$$\hat{\boldsymbol{f}}_i = (\boldsymbol{I} + \lambda \boldsymbol{G}_i)^{-1} \boldsymbol{y}_i. \tag{3.32}$$

Tästä seuraa, että mikäli satunnaisvirheiden kovarianssimatriisilla on tietty, yhtälön (3.31) toteuttava rakenne, voidaan painotettu estimaattori (3.29) korvata yksinkertaisemmalla, painottamattomalla estimaattorilla (3.32). Nummi ja Koskela (2007) ovat osoittaneet, että tällaisia yhtälön (3.31) toteuttavia rakenteita ovat riippumaton-, tasa- ja lineaarinen rakenne, eli $\boldsymbol{R}_i = \boldsymbol{I}$, $\boldsymbol{R}_i = \boldsymbol{I} + \sigma_d^2 \boldsymbol{1}\boldsymbol{1}'$ ja $\boldsymbol{R}_i = \boldsymbol{I} + \boldsymbol{X}_i \boldsymbol{D}_i \boldsymbol{X}_i'$, missä $\boldsymbol{X}_i$ on $n_i \times 2$ matriisi, jossa ensimmäinen sarake koostuu ykkösistä ja toinen yksilön mittauspisteistä.

3.3.4 Kasvukäyrämallin spliniestimaattori

Eräs pitkittäisaineiston erikoistapaus on täydellinen ja tasapainoinen aineisto, jolloin puuttuvia tietoja ei esiinny ja mittaukset on tehty kaikille yksilöille samoina perättäisinä ajanhetkinä. Kyseessä on niin kutsuttu *kasvukäyrämalli (generalized multivariate analysis of variance, GMANOVA)*, jonka alkujaan esittivät Potthoff ja Roy (1964). Malli voidaan kirjoittaa muodossa

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{A}' + \boldsymbol{E}, \tag{3.33}$$

missä $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ on $q \times n$ havaintojen matriisi, $\boldsymbol{T}$ on $q \times p$ yksilöiden sisäinen suunnittelumatriisi ($p < q$, $\text{rank}(\boldsymbol{T}) = p$), $\boldsymbol{A}$ $n \times m$ yksilöiden välinen suunnittelumatriisi ($\text{rank}(\boldsymbol{A}) = m$) ja $\boldsymbol{B}$ on $p \times m$ tuntemattomien parametrien matriisi. Lisäksi $\boldsymbol{E}$ on $q \times n$ satunnaisvirheiden matriisi, jossa sarakkeet oletetaan riippumattomiksi odotusarvovektorilla $\boldsymbol{0}$ ja kovarianssimatriisilla $\sigma^2 \boldsymbol{R}$. Usein matriisi $\boldsymbol{A}$ muodostuu m indikaattorivektorista, jotka jakavat yksilön johonkin m ryhmästä (esim. tytöt ja pojat). Kuitenkin $\boldsymbol{A}$ voi lisäksi sisältää myös kovariaatteja, kuten tupakointi tai ikä. (Nummi & Koskela 2007.)

Mallin (3.33) yksilöiden odotusarvokäyrää on perinteisesti mallinnettu matala-asteisilla polynomeilla. Kuitenkin, kuten jo aiemmin todettiin, ei polynomien avulla saavuteta aina tyydyttäviä estimaattiarvoja. Tällöin on syytä korvata yksilöiden sisäinen osa mallissa (3.33) tasoittavilla kuutiosplineillä. Tällainen malli voidaan esittää muodossa

$$\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{A}' + \boldsymbol{E},$$

missä $\boldsymbol{F} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_m)$ on $q \times m$ odotusarvosplinien matriisi. (Nummi & Koskela 2007.)

Matriisin $\boldsymbol{F}$ odotusarvosplinit voidaan estimoida maksimoimalla (vakiota vaille) sakotettu logaritmoitu uskottavuusfunktio

$$\begin{aligned} 2l &= -\frac{1}{\sigma^2}\text{tr}\left[(\boldsymbol{Y}' - \boldsymbol{A}\boldsymbol{F}')\boldsymbol{R}^{-1}(\boldsymbol{Y}' - \boldsymbol{A}\boldsymbol{F}')' + \lambda_{gcm}(\boldsymbol{A}\boldsymbol{F}')\boldsymbol{G}(\boldsymbol{A}\boldsymbol{F}')'\right] - n\log|\sigma^2\boldsymbol{R}| \\ &= -\frac{1}{\sigma^2}\text{tr}\left[(\boldsymbol{Y}' - \boldsymbol{F}_*')\boldsymbol{R}^{-1}(\boldsymbol{Y}' - \boldsymbol{F}_*')' + \lambda_{gcm}(\boldsymbol{F}_*')\boldsymbol{G}(\boldsymbol{F}_*')'\right] - n\log|\sigma^2\boldsymbol{R}|, \end{aligned}$$

missä $\boldsymbol{G} = \boldsymbol{Q}\boldsymbol{B}^{-1}\boldsymbol{Q}'$ on jälleen identiteetillä (3.12) määritelty rosoisuusmatriisi ja $\boldsymbol{F}_* = \boldsymbol{F}\boldsymbol{A}'$. Nummi ja Koskela (2007) ovat osoittaneet, että kiinteillä σ^2, λ_{gcm} ja $\boldsymbol{R}$ arvoilla maksimi saavutetaan, kun

$$\tilde{\boldsymbol{G}} = (\boldsymbol{R}^{-1} + \lambda_{gcm}\boldsymbol{G})^{-1}\boldsymbol{R}^{-1}\boldsymbol{Y}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}, \tag{3.34}$$

ja mikäli $\boldsymbol{R}\boldsymbol{Q} = \boldsymbol{Q}$, lauseke (3.34) sievenee muotoon

$$\tilde{\boldsymbol{G}} = (\boldsymbol{I} + \lambda_{gcm}\boldsymbol{G})^{-1}\boldsymbol{Y}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}. \tag{3.35}$$

Kuten edellisessä alaluvussa todettiin, toteutuu yhtälö $\boldsymbol{R}\boldsymbol{Q} = \boldsymbol{Q}$, kun kovarianssimatriisilla $\boldsymbol{R}$ on riippumaton-, tasa- tai lineaarinen rakenne. (Nummi & Koskela 2007.)

Mikäli aineisto koostuu ainoastaan yhdestä ryhmästä, eli $\boldsymbol{A}$ on $n \times 1$ matriisi, lauseke (3.35) on yhtäpitävä alaluvussa 3.3.1 määritellyn NSS-estimaattorin lausekkeen (3.26) kanssa, kun $\lambda_{gcm} = \frac{1}{n}\lambda$.

3.4 Lineaarinen sekamalli

Lineaarinen sekamalli (linear mixed effects (LME) model) on tavallisen lineaarisen mallin laajennus, joka sisältää sekä kiinteitä- että satunnaisvaikutuksia. Satunnaisvaikutusten avulla on mahdollista mallintaa yksilöiden sisäistä korrelaatiorakennetta, minkä vuoksi LME-malli tarjoaakin joustavan ja tehokkaan työkalun pitkittäisaineiston käsittelyssä.

Lineaarisen sekamallin ja tasoittavan splinin välillä on löydettävissä yhteys, jonka ensimmäisenä esitti Speed (1991) ja jota myöhemmissä analyyseissään hyödynsivät muun muassa Brumback ja Rice (1998) sekä Wang (1998a,b). Tässä kappaleessa esitellään aluksi sekamallin tärkeimmät ominaisuudet, jonka jälkeen osoitetaan, kuinka tasoittava kuutiosplini voidaan lausua lineaarisen sekamallin avulla.

3.4.1 Lineaarisen sekamallin määrittely

Harville (1976, 1977) sekä Laird ja Ware (1982) esittivät ensimmäisinä lineaarisen sekamallin yleisen muodon

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, ..., n, \tag{3.36}$$

missä $\boldsymbol{y}_i$ on i. yksilön $n_i \times 1$ vastevektori, $\boldsymbol{X}_i$ on kiinteiden vaikutusten suunnittelumatriisi ja $\boldsymbol{\beta}$ vastaava $m \times 1$ kiinteiden vaikutusten vektori. Vastaavasti $\boldsymbol{Z}_i$ on satunnaisvaikutusten suunnittelumatriisi ja $\boldsymbol{u}_i$ satunnaisvaikutusten $k \times 1$ vektori. Lisäksi $\boldsymbol{\epsilon}_i$ on $n_i \times 1$ satunnaisvirheiden vektori. Satunnaisvirheet ja satunnaisvaikutukset ovat keskenään riippumattomia ja lisäksi niiden oletetaan olevan normaalisti jakautuneita

$$\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \qquad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i).$$

Oletuksista seuraa, että $\boldsymbol{y}_i$ noudattaa moniulotteista normaalijakaumaa odotusarvolla $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ ja kovarianssimatriisilla $\mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i = \boldsymbol{V}_i$, eli

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i).$$

3.4.2 Kiinteiden- ja satunnaisvaikutusten estimointi

Normaalijakaumaoletuksen tapauksessa, estimaattorit mallin kiinteille vaikutuksille $\boldsymbol{\beta}$ ja kovarianssimatriisin $\boldsymbol{V}_i$ parametreille saadaan minimoimalla (vakiota vaille) logaritmoitu uskottavuusfunktio

$$2l(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{R}_i) = \sum_{i=1}^{n} \left[(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) + \ln|\boldsymbol{V}_i| \right]. \tag{3.37}$$

Mikäli matriisit $\boldsymbol{D}$ ja $\boldsymbol{R}_i$ oletetaan tunnetuiksi, funktion (3.37) minimointi on yhtäpitävää niin kutsutun sekamalliyhtälön ratkaisemisen kanssa (Harville 1976, Robinson 1991)

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \tilde{\boldsymbol{D}}^{-1} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

missä

$$\begin{aligned} \boldsymbol{y} &= (\boldsymbol{y}_1', ..., \boldsymbol{y}_n')', \\ \boldsymbol{u} &= (\boldsymbol{u}_1', ..., \boldsymbol{u}_n')', \\ \boldsymbol{\epsilon} &= (\boldsymbol{\epsilon}_1', ..., \boldsymbol{\epsilon}_n')', \\ \boldsymbol{X} &= (\boldsymbol{X}_1', ..., \boldsymbol{X}_n')', \end{aligned} \qquad \begin{aligned} \boldsymbol{Z} &= \mathrm{diag}(\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n), \\ \tilde{\boldsymbol{D}} &= \mathrm{diag}(\boldsymbol{D}, ..., \boldsymbol{D}), \\ \boldsymbol{R} &= \mathrm{diag}(\boldsymbol{R}_1, ..., \boldsymbol{R}_n). \end{aligned}$$

Sekamalliyhtälöstä saadaan ratkaistua yleisesti tunnetut estimaattorit $\boldsymbol{\beta}$:lle ja $\boldsymbol{u}_i$:lle, jotka ovat *paras lineaarinen ja harhaton estimaattori (best linear unbiased estimator, BLUE)*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}^{-1})\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{3.38}$$

missä $\boldsymbol{V} = diag(\boldsymbol{V}_1, ..., \boldsymbol{V}_n)$, sekä *paras lineaarinen ja harhaton ennuste (best linear unbiased predictor, BLUP)*

$$\hat{\boldsymbol{u}}_i = \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}), \qquad i = 1, ..., n. \tag{3.39}$$

(Wu & Zhang 2006.)

Tavallisesti matriisit $\boldsymbol{D}$ ja $\boldsymbol{R}_i$ eivät ole tunnettuja, vaan ne on estimoitava ja korvattava BLUE:n (3.38) ja BLUP:n (3.39) lausekkeiden $\boldsymbol{D}$ ja $\boldsymbol{R}_i$ vastaavilla estimaattiarvoilla $\widehat{\boldsymbol{D}}$ ja $\widehat{\boldsymbol{R}}_i$. Kaksi suosittua menetelmää joilla estimaatit $\widehat{\boldsymbol{D}}$ ja $\widehat{\boldsymbol{R}}_i$ voidaan määritellä, ovat suurimman uskottavuuden -menetelmä sekä REML-menetelmä, joista esityksen antavat muun muassa Davidian ja Giltinan (1993) sekä Wu ja Zhang (2006).

3.4.3 Sekamallin ja tasoittavan kuutiosplinin yhteys

Kuten aiemmin todettiin, tasoittavan splinin ja lineaarisen sekamallin välillä on löydettävissä yhteys. Tämän alaluvun esitys pohjautuu Verbylan ym. (1999) esittämään todistukseen, jossa tasoittavan kuutiosplinin estimaattori (3.29) esitetään lineaarisen sekamallin BLUP-estimaatteina. Yhteys voidaan kuitenkin johtaa myös muuta kautta ja esimerkiksi Wu ja Zhang (2006) ovat esittäneet rosoisuusmatriisin (3.12) singulaariarvohajotelmaan pohjautuvan todistuksen. Nummi ja Koskela (2007) ovat puolestaan osoittaneet, kuinka kyseinen yhteys voidaan löytää kasvukäyräaineiston tapauksessa.

Olkoon matriisi $\boldsymbol{X}_i = (\boldsymbol{1}, \boldsymbol{t}_i)$, missä ensimmäinen sarake koostuu ykkösistä ja toinen yksilön i mittauspisteistä $\boldsymbol{t}_i = (t_1, ..., t_{n_i})'$. Määritellään lisäksi matriisit $\boldsymbol{Z}_i = \boldsymbol{Q}_i(\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}$ ja $\boldsymbol{V}_i = \sigma^2(\boldsymbol{R}_i + \lambda^{-1}\boldsymbol{Z}_i\boldsymbol{B}_i\boldsymbol{Z}_i')$, missä $\boldsymbol{Q}_i$ ja $\boldsymbol{B}_i$ saadaan identiteetiteetillä (3.12) määritellystä rosoisuusmatriisista $\boldsymbol{G}_i = \boldsymbol{Q}_i\boldsymbol{B}_i^{-1}\boldsymbol{Q}_i'$. Voidaan osoittaa (Verbyla ym. 1999, s. 297–298), että spliniestimaattori (3.29) voidaan esittää muodossa

$$\tilde{\boldsymbol{f}}_i = \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}_i + \boldsymbol{Z}_i\tilde{\boldsymbol{u}}_i, \tag{3.40}$$

missä

$$\tilde{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i \tag{3.41}$$

ja

$$\tilde{\boldsymbol{u}}_i = (\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i + \lambda\boldsymbol{B}_i^{-1})^{-1}\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}_i). \tag{3.42}$$

Estimaatit (3.41) ja (3.42) voidaan nähdä lineaarisen sekamallin

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i, \tag{3.43}$$

BLUP-ratkaisuina, missä $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \sigma_{u_i}^2\boldsymbol{B}_i)$ ja $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{R}_i)$. Lisäksi tasoitusparametri voidaan määritellä varianssisuhteena $\lambda = \sigma^2/\sigma_{u_i}^2$ ja logaritmoitu uskottavuusfunktio on $\boldsymbol{y}_i$:n ehdollinen logaritmoitu uskottavuusfunktio, kun $\boldsymbol{u}_i$ on annettu. (Verbyla ym. 1999.)

Mikäli yhtälö $\boldsymbol{R}_i\boldsymbol{Q}_i = \boldsymbol{Q}_i$ toteutuu, sievenee estimaattori (3.40) muotoon

$$\hat{\boldsymbol{f}}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i + \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i,$$

missä

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{y}_i$$

ja

$$\hat{\boldsymbol{u}}_i = (\boldsymbol{Z}_i'\boldsymbol{Z}_i + \lambda\boldsymbol{B}_i^{-1})^{-1}\boldsymbol{Z}_i'\boldsymbol{y}_i,$$

jolloin $\boldsymbol{u}_i \sim N(0, \sigma_{u_i}^2\boldsymbol{B}_i)$ ja $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2\boldsymbol{I})$. Huomattakoon vielä, että yhtälö (3.43) voidaan aina kirjoittaa muodossa

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{Z}_{i*}\boldsymbol{u}_{i*} + \boldsymbol{\epsilon}_i, \tag{3.44}$$

missä $\boldsymbol{Z_{i*}} = \boldsymbol{Z}_i \boldsymbol{B}_i^{1/2}$ ja $\boldsymbol{u_{i*}} = \boldsymbol{B}_i^{-1/2} \boldsymbol{u}_i$, kun $\boldsymbol{u_{i*}} \sim N(0, \sigma^2_{u_i} \boldsymbol{I})$ ja $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \boldsymbol{I})$. (Nummi ym. 2007.)

Yllä kuvattu idea, jossa tasoittava kuutiosplini esitetään LME-mallin BLUP-ratkaisuina, on monellakin tapaa kiinnostava. Nyt estimaattorin (3.40) ensimmäinen osa (kiinteä osa) $\boldsymbol{X}_i \tilde{\boldsymbol{\beta}}_i$ on regressiosuora ja toinen osa (satunnaisosa) $\boldsymbol{Z}_i \tilde{\boldsymbol{u}}_i$ tuo splinipiirteen mukaan käyrään. Lisäksi tasoitusparametrin valintaongelma vaihtuu nyt ongelmaksi määrittää varianssikomponentit σ^2 ja $\sigma^2_{u_i} = \sigma^2/\lambda$, mikä onnistuu helposti esimerkiksi EM-algoritmin avulla (esim. Laird ja Ware 1982). Kun mallina toimii (3.44), on varianssikomponentit lisäksi helppo estimoida tilastollisten ohjelmistojen funktioilla. S-PLUS:ssa ja R:ssä tämä onnistuu esimerkiksi funktiolla *lme* ja SAS:ssa proseduurilla *PROC MIXED*. (Nummi & Koskela 2007, Wu & Zhang 2006.)

4 Kovarianssirakenteen mallintaminen

Pitkittäisaineiston erityispiirteenä on, että samalle yksilölle on tehty toistuvia mittauksia, minkä johdosta yksilön sisäiset mittaukset ovat keskenään korreloituneita. Usein tämä korrelaatio on positiivista. Lisäksi korrelaatio vaimenee ajan kuluessa, eli kahden lähekkäin tehdyn mittauksen välinen korrelaatio on voimakkaampaa kuin mittausten, joiden välillä aikaa on kulunut enemmän. Käytännön tarkastelut ovat myös osoittaneet, että pitkittäisaineistossa varianssi tavallisesti kasvaa ajan kuluessa. (Fizmaurice, Laird & Ware 2004.)

Edellä kuvatut pitkittäisaineiston erityispiirteet johtavat siihen, että vaikka korrelaatiorakenne harvoin itsessään on varsinaisena kiinnostuksen kohteena, on se kuitenkin otettava huomioon aineistoa mallinnettaessa. Korrelaation sisällyttäminen malliin tavallisesti parantaa estimoitavien parametrien tehokkuutta tai tarkkuutta ja mikäli aineisto sisältää puuttuvia havaintoja, korrelaation oikeanlainen mallinnus on usein edellytys validien estimaattien saavuttamiseksi. Toisaalta väärin määritelty kovarianssirakenne voi niin ikään johtaa virheellisiin estimaattiarvoihin sekä vääriin tilastollisiin johtopäätöksiin. (Fitzmaurice ym. 2004.) Myös epäparametrisia menetelmiä käytettäessä kovarianssimatriisin valinnalla on vaikutusta estimoitaviin odotusarvoestimaatteihin, sillä tasoitusparametrin valintamenetelmät aliestimoivat tasoitusparametrin, mikäli kovarianssirakennetta ei oteta huomioon (Wang 1998b).

Tässä luvussa esitellään tapoja mallintaa pitkittäisaineiston kovarianssirakennetta. Alaluvussa 4.1 käydään aluksi läpi joitakin perinteisesti käytettyjä kovarianssimatriisin mallinnusmenetelmiä. Kappaleen varsinaisena tavoitteena on kuitenkin osoittaa, kuinka tasoittavia kuutiosplinejä voidaan odotusarvon estimoinnin lisäksi hyödyntää myös kovarianssimatriisin mallinnuksessa. Tässä lähtökohdan tarjoaa modifioitu Choleskyn hajotelma, joka esitellään alaluvussa 4.2. Alaluvussa 4.3 menetelmää havainnollistetaan käytännön aineistolla.

4.1 Kovarianssimatriisin klassiset mallit

Kovarianssimatriisin mallinnusmenetelmiä on kirjallisuudessa esitetty lukuisia ja tässä alaluvussa käsitellään niistä kolmea perinteistä ja laajalti sovellettua mallia. Luku perustuu Fitzmaurice ym. (2004) kirjan esitykseen.

4.1.1 Rakenteeton kovarianssimatriisi

Mikäli mittauspisteiden määrä on suhteellisen pieni ja kaikki yksilöt on mitattu samoina ajanhetkinä, on kenties perusteltua antaa kovarianssimatriisin rakentua mielivaltaisesti siten, että matriisin kaikki elementit ovat rajoittamattomia. Ainoa rajoite tällöin on vaatimus kovarianssimatriisin symmetrisyydestä ja positiivisdefiniittisyydestä. Tällaista kovarianssimatriisia kutsutaan rakenteettomaksi (unstructured) ja se voidaan ilmaista matriisimuodossa

$$\mathrm{Cov}(\boldsymbol{y}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n_i} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i1} & \sigma_{n_i2} & \dots & \sigma_{n_i}^2 \end{pmatrix}.$$

Rakenteettoman kovarianssimatriisin etuna on, ettei varianssista ja kovarianssista tehdä minkäänlaisia oletuksia. Tämä on tärkeää etenkin käytännön aineistoissa, joissa varianssi harvoin pysyy vakiona ajan kuluessa.

Kuitenkin, mikäli aineiston yksilölle i on tehty n_i määrä mittauksia, rakentuu kovarianssimatriisi yhteensä $n_i \times (n_i + 1)/2$ vapaasta parametrista. Tällöin estimoitavien parametrien määrä kasvaa nopeasti mittauspisteiden lukumäärän kasvaessa. Mikäli taas estimoitavien kovarianssiparametrien määrä on otoskokoon verrattuna suuri, tulee estimaatti usein epäluotettavaksi. Näin ollen rakenteeton kovarianssimatriisi on käyttökelpoinen ainoastaan tilanteissa, joissa yksilöiden määrä suhteessa mittauspisteiden määrään on suuri. Lisäksi rakenteetonta kovarianssimatriisia voidaan soveltaa vain tilanteisiin, joissa aineisto on tasapainoinen. Tämä taas on harvoin käytännössä realistinen oletus.

4.1.2 Kovarianssirakenteet

Kovarianssimatriisin rakennemallit johtavat usein vähäparametriseen kovarianssirakenteeseen, joka koostuu ainoastaan muutamasta vapaasta parametrista. Pitkittäisaineistoissa käytetyt kovarianssirakenteet on alun perin kehitetty aikasarja-aineistoille. Aikasarja- ja pitkittäisaineisto eroavat kuitenkin toisistaan, sillä aikasarja-analyysi keskittyy tavallisesti käsittelemään yhden yksilön pitkiä sarjoja, joissa mittauspisteet ovat lisäksi samoin jakautuneita. Erilaiset kovarianssirakenteet olettavat myös, että mittauspisteiden välinen korrelaatio joko pysyy vakiona tai vähenee nopeasti aikapisteiden välimatkan kasvaessa. Kumpikaan oletuksista ei kuitenkaan usein ole realistinen pitkittäisaineiston tapauksessa. Myöskään monen kovarianssirakenteen oletus vakiovarianssisuudesta harvoin pitää paikkansa.

Edellä esitettyjen syiden vuoksi pitkittäisaineistoon onkin usein joko mahdotonta tai tehotonta soveltaa kovarianssimatriisin rakennemalleja. Parhaiten ne soveltuvatkin aineistoille, jotka ovat tasapainoisia ja joissa mittauspisteet ovat tasaisesti jakautuneita.

Erilaisia kovarianssirakenteita on olemassa useita, joista tässä esitellään tasakorrelaatio- ja autoregressiivinen kovarianssirakenne. Muita mahdollisia ra-

kenteita ovat muun muassa Toeplitzin kovarianssirakenne, nauhamainen (banded) kovarianssirakenne sekä exponentiaalinen kovarianssirakenne. Näistä tarkemman esityksen antavat muun muassa Fizmaurice ym. (2004).

Tasakorrelaatiorakenne

Tasakorrelaatiorakenne on eräs ensimmäisistä kovarianssirakenteista, jota on käytetty toistomittausaineiston analysoimisessa. Tasakorrelaatiorakenne muodostuu, kun oletetaan, että varianssi σ^2 on vakio jokaisessa mittauspisteessä ja $\mathrm{Cor}(Y_{ij}, Y_{ik}) = \rho$, $\rho \geq 0$ kaikilla j ja k arvoilla. Toisin sanoen

$$\mathrm{Cov}(\boldsymbol{y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & ... & \rho \\ \rho & 1 & \rho & ... & \rho \\ \rho & \rho & 1 & ... & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & ... & 1 \end{pmatrix}.$$

Vähäparametrinen tasakorrelaatiorakenne sisältää vain kaksi estimoitavaa parametria. Kuitenkin tasakorrelaatio tekee vahvan oletuksen siitä, että mittausten välinen korrelaatio pysyy vakiona ajan kuluessa. Kuten todettu, tämä taas on harvoin tilanne pitkittäisaineistossa, jossa korrelaatio tavallisesti vaimenee kun mittaushetkien aikaväli kasvaa. Lisäksi oletus vakiovarianssisuudesta on myös usein epärealistinen. Tasakorrelaatiorakenteella pystytäänkin vain harvoin mallintamaan tyydyttävästi pitkittäisaineiston kovarianssimatriisia.

Autoregressiivinen rakenne

Autoregressiivinen (AR(1)) malli on suosittu ja usein sovellettu kovarianssirakenne. AR(1)-mallissa oletetaan, että varianssi σ^2 pysyy vakiona ajan kuluessa ja $\mathrm{Cor}(y_{ij}, y_{ij+k}) = \rho^k$ kaikilla j ja k kun $\rho \geq 0$. Toisin sanoen

$$\mathrm{Cov}(\boldsymbol{y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & ... & \rho^{n_i-1} \\ \rho & 1 & \rho & ... & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & ... & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & ... & 1 \end{pmatrix}.$$

Kuten tasakorrelaatiorakenne, myös AR(1)-rakenne on hyvin vähäparametrinen edellyttäen vain kahden parametrin estimointia. Tämä rakenne soveltuu pitkittäisaineiston kovarianssimatriisille usein kuitenkin tasakorrelaatiorakennetta paremmin, sillä korrelaation oletetaan vaimenevan mittaushetkien aikavälin kasvaessa. Kuitenkin monissa käytännön aineistoissa korrelaatio saman yksilön sisällä harvoin vaimenee yhtä nopeasti kuin autoregressiivinen malli olettaa. Myös varianssin vakioisuus on jälleen epärealistinen oletus.

4.1.3 Satunnaisvaikutusten kovarianssirakenne

Eräs tärkeä kovarianssimatriisin malli saadaan alaluvussa 3.4 esitetystä lineaarisesta sekamallista (3.36). Koska LME-mallissa sekä satunnaisvirheet $\boldsymbol{\epsilon}_i$, että satunnaisvaikutukset $\boldsymbol{u}_i$ oletetaan riippumattomiksi satunnaismuuttujiksi, saadaan kovarianssimatriisiksi

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{y}_i) &= \mathrm{Cov}(\boldsymbol{Z}_i\boldsymbol{u}_i) + \mathrm{Cov}(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\mathrm{Cov}(\boldsymbol{u}_i)\boldsymbol{Z}_i' + \mathrm{Cov}(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i + \boldsymbol{R}_i,\end{aligned}$$

missä $\boldsymbol{Z}_i$ on satunnaisvaikutusten suunnittelumatriisi sekä $\boldsymbol{D}$ ja $\boldsymbol{R}_i$ ovat satunaisvaikutusten ja virhetermien kovarianssimatriiseja. Matriiseille $\boldsymbol{D}$ ja $\boldsymbol{R}_i$ voidaan nyt määrittää jokin alaluvussa 4.1.2 esitetyistä kovarianssirakenteista. Usein tehdään oletus, että matriisi $\boldsymbol{R}_i$ on diagonaalimatriisi $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_{n_i}$.

Satunnaisvaikutusten mallin avulla voidaan kovarianssimatriisia mallintaa joustavasti ja monipuolisesti. Toisin kuin kovarianssimatriisin rakenteelliset mallit, satunnaisvaikutusten malli ei edellytä tasapainoista aineistoa. Lisäksi siinä missä kovarianssirakenteet usein olettavat varianssin pysyvän vakiona ajan kuluessa, sallii satunnaisvaikutusten kovarianssirakenne varianssin ja kovarianssin muuttuvan ajan funktiona.

4.2 Modifioitu Choleskyn hajotelma

Edellisessä alaluvussa käytiin läpi joitakin yleisesti käytettyjä tapoja mallintaa kovarianssimatriisia. Kuitenkin, kuten todettiin, täysin rakenteeton kovarianssimatriisi on usein melko epävakaa, kun taas virheellisesti määritelty rakennemalli saattaa johtaa harhaisiin estimaattiarvoihin. Tässä luvussa esitellään aineistopohjainen lähestymistapa, joka mahdollistaa tasapainottelun näiden kahden ääripään välillä. Lähtökohtana toimii modifioitu Choleskyn hajotelma, jonka käytön kovarianssimatriisin mallinnuksessa alkujaan esitti Pourahmadi (1999, 2000) ja jota myöhemmin ovat kehitelleet eteenpäin muun muassa Pan ja von Rosen (2005) sekä Ye ja Pan (2006).

Tähänastiset sovellukset ovat hyödyntäneet parametrisia menetelmiä modifioidun Choleskyn hajotelman termien estimoimisessa ja mallina ovat toimineet polynomit ajan suhteen. Tässä työssä osoitetaan, kuinka polynomit on mahdollista korvata joustavammilla tasoittavilla kuutiosplineillä.

4.2.1 Modifioidun Choleskyn hajotelman muodostus

Selkeyden ja yksinkertaisuuden vuoksi oletetaan Pourahmadin (1999, 2000) esitystä seuraten pitkittäisaineiston yksilöille homogeeniset kovarianssimatriisit, jolloin $\boldsymbol{R}_1 =, ..., = \boldsymbol{R}_n = \boldsymbol{R}$. Olkoon nyt $\boldsymbol{y} = (y_1, ..., y_m)'$ aikajärjestetty satunnaisvektori, jolla on odotusarvovektori $\boldsymbol{\mu}$ ja positiivisesti definiitti $m \times m$ kovarianssimatriisi $\boldsymbol{R}$. Voidaan osoittaa (Newton 1988, s. 359), että symmetrinen matriisi $\boldsymbol{R}$ on positiivisesti definiitti, jos ja vain jos löytyy yksikäsitteisesti

määritelty alakolmiomatriisi $\boldsymbol{T}$, jonka diagonaalielementit koostuvat ykkösistä, sekä yksikäsitteinen, positiivisista diagonaalielementeistä koostuva sellainen diagonaalimatriisi $\boldsymbol{H}$, että

$$\boldsymbol{TRT}' = \boldsymbol{H}, \quad \text{tai} \quad \boldsymbol{R}^{-1} = \boldsymbol{T}'\boldsymbol{H}^{-1}\boldsymbol{T}. \tag{4.1}$$

Hajotelma (4.1) on niin kutsuttu *modifioitu Choleskyn hajotelma (modified Cholesky decomposition, MCD)* ja matriisien $\boldsymbol{T}$ ja $\boldsymbol{H}$ elementeillä on nyt selvä, tilastollisesti mielekäs tulkinta. Matriisin $\boldsymbol{T}$ diagonaalin alapuoleiset elementit koostuvat autoregressiivisten kertoimien ϕ_{jk} vastaluvuista, jotka saadaan autoregressiivisesta mallista

$$\hat{y}_j = \mu_j + \sum_{k=1}^{j-1} \phi_{jk}(y_k - \mu_k).$$

Matriisin $\boldsymbol{H}$ diagonaalielementit ovat puolestaan ennustevirheiden variansseja $\sigma_j^2 = \mathrm{Var}(\epsilon_j) = \mathrm{Var}(y_j - \hat{y}_j)$, kun $1 \leq j \leq m$. Nimitetään jatkossa kertoimia ϕ_{jk} *yleistetyiksi autoregressiivisiksi parametreiksi* ja ennustevirheiden variansseja σ_j^2 *virhevariansseiksi*. (Pourahmadi 2000.)

4.2.2 Kovarianssimatriisin parametrien estimointi

Kovarianssimatriisin $\boldsymbol{R}$ ollessa rakenteeton, matriisien $\boldsymbol{T}$ ja $\log(\boldsymbol{H})$ ei-redundantit alkiot ϕ_{jk} ja $\log(\sigma_j^2)$ ovat rajoittamattomia. Mikäli näiden alkioiden sallitaan riippuvan joistakin tunnetuista kovariaateista, voidaan huomattavasti vähentää estimoitavien parametrien määrää. (Liu 2004.) Yksilön sisäinen kovarianssimatriisi voidaan siis uudelleenparametrisoida modifioidun Choleskyn hajotelman termien avulla, jolloin tuloksena saadaan yleistetyt autoregressiiviset parametrit sekä virhevarianssit. Tämän jälkeen voidaan saatuihin parametreihin sovittaa tilastollinen malli, jolloin estimoitujen uusien parametrien tuloksena saavutetaan vähäparametrisempi kovarianssirakenne. Menetelmän etuna on, että estimoinnin tuloksena saatu kovarianssimatriisi $\hat{\boldsymbol{R}}$ on positiivisesti definiitti. Lisäksi menetelmä on aineistopohjainen, joten kovarianssimatriisin todellinen rakenne on mahdollista saavuttaa. (Pan & von Rosen 2006.)

Pourahmadi (1999, 2000) mallinsi yleistettyjä autoregressiivisia parametreja ϕ_{jk} sekä virhevarianssin logaritmeja $\log(\sigma_j^2)$ lineaarisilla malleilla

$$\log(\sigma_j^2) = \boldsymbol{z}_j'\boldsymbol{\lambda}, \qquad \phi_{jk} = \boldsymbol{z}_{jk}'\boldsymbol{\gamma},$$

missä $\boldsymbol{z}_j$ ja $\boldsymbol{z}_{jk}$ ovat $q \times 1$ ja $d \times 1$ kovariaattien vektoreita, sekä $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_q)'$ ja $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_d)'$ vastaavia parametreja. Kovariaatit $\boldsymbol{z}_j$ ja $\boldsymbol{z}_{jk}$ voidaan ajatella esimerkiksi polynomeiksi ajan suhteen

$$\begin{aligned}\boldsymbol{z}_j &= (1, t_j, t_j^2, ..., t_j^{d-1})', \\ \boldsymbol{z}_{jk} &= (1, (t_j - t_k), (t_j - t_k)^2, ..., (t_j - t_k)^{q-1})',\end{aligned}$$

jolloin $\boldsymbol{\lambda}$ ja $\boldsymbol{\gamma}$ voidaan estimoida suurimman uskottavuuden menetelmän avulla (Pourahmadi 1999, s. 687–689).

Tässä työssä päämääränä on kuitenkin estimoida $\log(\sigma_j^2)$ ja ϕ_{jk} polynomien sijasta tasoittavilla kuutiosplineillä. Olkoon $\boldsymbol{\phi}_j = (\phi_{j,j-1}, ..., \phi_{j,1})'$ yleistettyjen autoregressiivisten parametrien vektori aikapisteessä j, kun $j = 2, ..., m$. Koska $\phi_{j,j-k}$, $k = 1, ..., j-1$ tarkoittaa autoregressiivisen prosessin parametria viiveellä k, voidaan olettaa, että $\phi_{j,j-k}$ on pieni j:n ollessa kiinteä ja k:n saadessa suuren arvon. Tästä seuraa, että vektori $\boldsymbol{\phi}_j$ on monotonisesti vähenevä. (Pourahmadi 1999.) Saadaan siis yhtälöryhmä

$$\begin{aligned}
\boldsymbol{\phi}_2 &= (\phi_{2,1})' \\
\boldsymbol{\phi}_3 &= (\phi_{3,2}, \phi_{3,1})' \\
&\vdots \\
\boldsymbol{\phi}_m &= (\phi_{m,m-1}, \phi_{m,m-2}, ..., \phi_{m,1})'.
\end{aligned}$$

Matriisin $\boldsymbol{H}$ diagonaalilta saadaan vastaavasti ennustevirheiden varianssit σ_j^2, kun $j = 1, ..., m$.

Nyt voidaan sovittaa tasoittava kuutiosplini sekä yleistetyille autoregressiivisille parametreille ϕ_{jk} että logaritmoiduille virhevariansseille $\log(\sigma_j^2)$ molemmille erikseen. Oletetaan yksinkertaisuuden vuoksi, että ϕ_{jk}, kun $j = 1, ..., m$, $k = 1, ..., j-1$ ovat keskenään riippumattomia. Samoin oletetaan $\log(\sigma_j^2)$, $j = 1, ..., m$ riippumattomuus. Yleistettyjen autoregressiivisten parametrien $\hat{\phi}_{jk}$ estimaatit saadaan nyt minimoimalla sakotettu yleistetty pienimmän neliösumman kriteeri (3.21), jolloin kriteeri saa muodon

$$\sum_{j=2}^{m} (\boldsymbol{\phi}_j - \boldsymbol{f}_j)'(\boldsymbol{\phi}_j - \boldsymbol{f}_j) + \lambda \int_a^b \left[f''(t)\right]^2 \mathrm{dt}. \tag{4.2}$$

Lausekkeen (4.2) minimoiva, käyrän $\boldsymbol{f}$ estimaattori on NSS estimaattori (3.25), eli

$$\hat{\boldsymbol{f}}_{nss} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{G})^{-1}\boldsymbol{X}'\boldsymbol{\phi}, \tag{4.3}$$

missä $\boldsymbol{\phi} = (\boldsymbol{\phi}_2', ..., \boldsymbol{\phi}_m')'$.

Virhevarianssin logaritmien $\log(\sigma_j^2)$ estimaatit $\log(\hat{\sigma}_j^2)$ saadaan vastaavasti, kun minimoidaan sakotettu pienimmän neliösumman kriteeri (3.9), jolloin kriteeri saa muodon

$$\sum_{j=1}^{m} \left[\log(\sigma_j^2) - f(t_j)\right]^2 + \lambda \int_a^b \left[f''(t)\right]^2 \mathrm{dt}. \tag{4.4}$$

Vastaava käyrän $\boldsymbol{f}$ estimaattori on tasoittava kuutiosplini (3.15), joka nyt saa muodon

$$\hat{\boldsymbol{f}} = (\boldsymbol{I}_m + \lambda \boldsymbol{G})^{-1}\log(\boldsymbol{\sigma}^2), \tag{4.5}$$

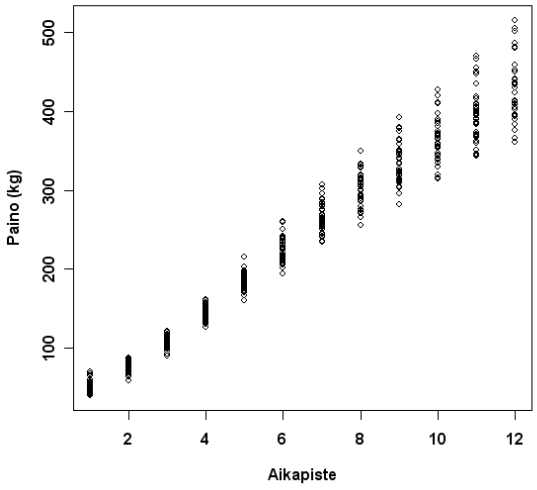
missä $\log(\boldsymbol{\sigma}^2) = [\log(\sigma_1^2), ..., \log(\sigma_m^2)]'$.

Myös tasoitusparametri λ voidaan valita erikseen molemmille estimaattoreille (4.3) ja (4.5). Estimaattorin (4.5) tasoitusparametri saadaan käyttämällä ristiinvalidointia (3.18) tai yleistettyä ristiinvalidointia (3.19). Vastaavasti estimaattorin (4.3) tasoitusparametri voidaan valita NSS-estimaattorin yleistetyn ristiinvalidoinnin (3.28) avulla.

4.3 Esimerkki kovarianssimatriisin mallintamisesta

4.3.1 Kovarianssimatriisin estimointi modifioidun Choleskyn hajotelman avulla

Havainnollistetaan nyt edellä kuvattua menetelmää alaluvussa 2.2 kuvatun Sonni-aineiston avulla. Tarkastellaan aluksi kuviota 4.1, johon on piirretty 40 Suomenkarja-rotuisen sonnin painon kehitys 12 aikapisteessä. Kuviosta voidaan heti havaita kaksi Fitzmaurice ym. (2004) mainitsemaa tyypillistä pitkittäisaineiston ominaispiirrettä: painon ja mittausajankohdan välillä on vahva positiivinen riippuvuus ja varianssi kasvaa selvästi ajan kuluessa. Jo kuvion 4.1 perusteella on siis helppo päätellä, että stationaarinen kovarianssirakenne ei sovellu aineistoon.



Kuvio 4.1. Suomenkarja-rotuisten sonnien painon kehitys.

Oletetaan nyt aineiston jokaisella yksilölle 12×12 kovarianssimatriisi $\boldsymbol{R}$ ja estimoidaan tätä matriisia tavallisella otoskovarianssimatriisilla

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})',$$

missä

$$\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i.$$

Modifioitu Choleskyn hajotelma (4.1) saadaan nyt helposti ratkaistua tavallisen Choleskyn hajotelman avulla

$$\boldsymbol{R} = \boldsymbol{S} = \boldsymbol{C}\boldsymbol{C}', \tag{4.6}$$

missä $\boldsymbol{C}$ on alakolmiomatriisi, jonka diagonaalielementit ovat positiivisia. Edelleen määritellään matriisi $\boldsymbol{L}$ siten, että

$$\boldsymbol{L} = \boldsymbol{C}(\mathrm{diag}(\boldsymbol{C})^{-1}),$$

millä $\boldsymbol{L}$ on sellainen alakolmiomatriisi, jonka diagonaalialkiot ovat ykkösiä. Nyt jos Choleskyn hajotelmassa (4.6) $\boldsymbol{C}$ korvataan matriisilla $\boldsymbol{L}(\mathrm{diag}(\boldsymbol{C}))$, niin saadaan

$$\begin{aligned}\boldsymbol{R} &= \boldsymbol{L}(\mathrm{diag}(\boldsymbol{C}))(\mathrm{diag}(\boldsymbol{C}))\boldsymbol{L}' \\ &= \boldsymbol{L}\boldsymbol{H}\boldsymbol{L}' \\ &= \boldsymbol{T}^{-1}\boldsymbol{H}(\boldsymbol{T}')^{-1},\end{aligned}$$

missä $\boldsymbol{H} = (\mathrm{diag}(\boldsymbol{C}))(\mathrm{diag}(\boldsymbol{C}))$.

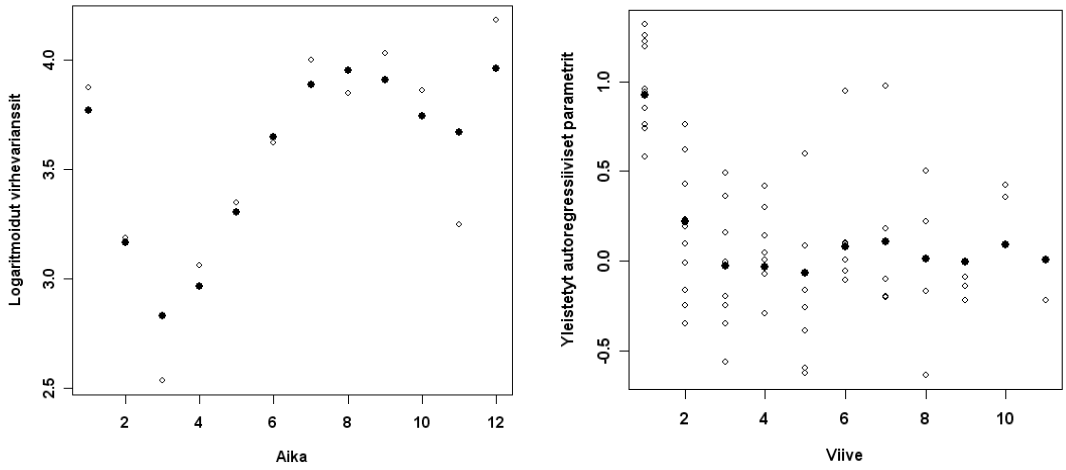
Taulukko 4.1. Sonni-aineistosta lasketut otosvarianssit (päädiagonaali), otoskorrelaatiot (päädiagonaalin yläpuoli), yleistetyt autoregressiiviset parametrit (päädiagonaalin alapuoli) ja virhevarianssit (viimeinen rivi).

t	1	2	3	4	5	6	7	8	9	10	11	12
1	48	0.64	0.66	0.58	0.61	0.59	0.54	0.47	0.38	0.34	0.33	0.31
2	0.58	41	0.88	0.83	0.77	0.76	0.71	0.71	0.70	0.69	0.65	0.66
3	0.19	0.95	61	0.86	0.86	0.79	0.72	0.67	0.64	0.60	0.55	0.55
4	-0.02	0.43	0.76	91	0.88	0.84	0.74	0.69	0.64	0.61	0.56	0.57
5	0.14	-0.20	0.62	0.76	154	0.91	0.87	0.81	0.77	0.72	0.67	0.65
6	0.09	0.42	-0.35	0.23	0.96	237	0.89	0.81	0.81	0.77	0.71	0.68
7	-0.06	0.60	-0.29	-0.56	0.76	0.74	332	0.94	0.90	0.86	0.82	0.79
8	-0.20	0.95	-0.60	-0.04	0.36	-0.35	1.20	491	0.94	0.89	0.87	0.85
9	-0.64	0.98	-0.11	-0.63	0.01	0.49	0.10	0.86	680	0.97	0.95	0.92
10	-0.14	0.50	-0.20	0.10	-0.16	-0.07	0.16	-0.25	1.26	917	0.98	0.96
11	0.35	-0.22	-0.17	0.18	0.00	-0.39	0.04	0.00	-0.01	1.23	1134	0.97
12	-0.22	0.42	-0.09	0.22	-0.10	0.10	-0.26	0.30	-0.25	-0.16	1.32	1498
	48	24	13	21	29	38	55	47	56	48	26	65

Taulukossa 4.1 on annettu yhteenveto Sonni-aineistosta lasketuista kovarianssiparametreista. Päädiagonaalilla ovat aineistosta lasketut otosvarianssit, päädiagonaalin yläpuolella otoskorrelaatiokertoimet ja alapuolella yleistetyt autoregressiiviset parametrit ϕ_{jk}. Viimeisellä rivillä ovat virhevarianssit σ_j^2. Taulukko 4.1 vahvistaa sen, mikä oli pääteltävissä jo kuvion (4.1) perusteella: korrelaatiokertoimet ovat kaikki positiivisia ja otosvarianssi kasvaa ajan myötä. Lisäksi voidaan havaita vielä eräs Fitzmaurice ym. (2004) mainitsema tyypillinen pitkittäisaineiston ominaispiirre: korrelaatiokertoimet pienenevät, kun havaintojen mittausvälimatka kasvaa.

Estimaatit $\log(\hat{\sigma}_j^2)$, kun $j = 1, ..., 12$ ja $\hat{\phi}_{jk}$, kun $j = 2, ..., 12$, $k = 1, ..., 11$ saadaan lausekkeista (4.3) ja (4.5). Tasoitusparametri estimaattorille (4.3) on valittu käyttämällä NSS-estimaattorin yleistettyä ristiinvalidointia (3.28) painomatriisilla $\boldsymbol{W} = N^{-1}\boldsymbol{I}_{n_i}$. Tasoitusparametrin arvoksi on tällöin saatu $\lambda = 0.7$. Estimaattorin (4.5) tasoitusparametri on puolestaan valittu tavallisen ristiinvalidoinnin (3.18) avulla, jolloin tasoitusparametrin arvoksi on saatu $\lambda = 0.4$.

Kuviosta 4.2 nähdään estimoidut tasoittavat kuutiosplinit. Näyttäisi siltä, että tasoittava kuutiosplini mallintaa molempia aineistoja hyvin. Voidaan lisäksi havaita, että sekä aineistosta lasketut autoregressiiviset parametrit että vastaavat estimaatit saavat suurehkon arvon viiveillä 1 ja 2, minkä perusteella voidaan päätellä kyseessä olevan AR(2)-prosessi.



Kuvio 4.2. Otosarvot (valkoiset pallot) ja sovitetut arvot (mustat pallot) virhevarianssien logaritmeille ja yleistetyille autoregressiivisille parametreille.

Alkuperäinen kovarianssimatriisi $\boldsymbol{R}$ koostuu yhteensä $n(n+1)/2 = 78$ vapaasta parametrista. Koska tasoittavan kuutiosplinin aste on tasoittajamatriisin jälki, saadaan tasoitusparametrin arvolla $\lambda = 0.7$ yleistettyjen autoregressiivisten parametrien estimaattorin (4.3) asteeksi 7.0. Vastaavasti tasoitusparametrin arvolla $\lambda = 0.4$ saadaan virhevarianssin logaritmien estimaattorin (4.5) asteeksi 6.2. Menetelmän avulla voidaan siis 78-parametrinen kovarianssimatriisi mallintaa käyttämällä vain 13 vapaata parametria. Kuten alaluvussa 3.2.2 todettiin, tasoittavan kuutiosplinin aste voi liikkua välillä $2 < \mathrm{df} < m$, missä m on mittauspisteiden lukumäärä. Näin ollen tasoitusparametrin arvoja säätelemällä saataisiin estimaattorin (4.3) asteeksi $2 < \mathrm{df}_{\phi_{jk}} < 11$ ja vastaavasti estimaattorin (4.5) asteeksi $2 < \mathrm{df}_{\sigma_j^2} < 12$. Estimoidun kovarianssimatriisin vapaiden parametrien lukumäärä voi siis liikkua välillä $4 < \mathrm{df} < 23$.

4.3.2 Mallien vertailu

Tässä alaluvussa vertaillaan kolmea sonni-aineistoon sovitettua mallia ja niissä käytettyjä kovarianssimatriiseja. Sonni-aineisto on ollut tutkimuksen kohteena useasti ennenkin ja ensimmäisen tässä vertailuun otettavan mallin esitti alkujaan Liski (1987), jonka monista sovitetuista malleista parhaaksi osoittautui malli

$$y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \ln(t) + \epsilon.$$

Mallissa t on ajan kovariaatti ja satunnaisvirheiden kovarianssimatriisi oletetaan rakenteettomaksi (malli A). Nummi (1997) osoitti edelleen, että vähäparametrisemmalla AR(1)-rakenteella voidaan tuloksekkaasti korvata mallin A rakenteeton kovarianssimatriisi (malli B). Liski (1987) ja Nummi (1997) käyttivät analyyseissään vuonna 1966 syntyneiden sonnien koko aineistoa (yht. 208 sonnia). Tässä työssä vastaavat mallit on sovitettu ainoastaan Suomenkarjarotuisille sonneille (yht. 40 sonnia).

Päämääränä on siis vertailla malleja A ja B malliin C, jossa aineistoon on sovellettu alaluvussa 3.3.1 esitettyä GSS-menetelmää, jossa odotusarvokäyrän estimaattorina toimii kuutiollinen GSS-estimaattori (3.24). Estimaattorin (3.24) kovarianssimatriisi on edellisessä alaluvussa saavutettu, modifioidun Choleskyn hajotelman avulla muodostettu kovarianssimatriisi.

Käytetään mallien vertailussa aikapisteissä laskettuja residuaalien keskiarvoja ja vastaavia RMSE-arvoja (root mean squared error). Residuaalien RMSE-arvot kussakin aikapisteessä t_j, $j = 1, ..., 12$ saadaan laskettua kaavalla

$$\text{RMSE}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i - \hat{\epsilon}_i)^2} = \sqrt{\frac{1}{40}\sum_{i=1}^{40}(\epsilon_i - \hat{\epsilon}_i)^2}. \tag{4.7}$$

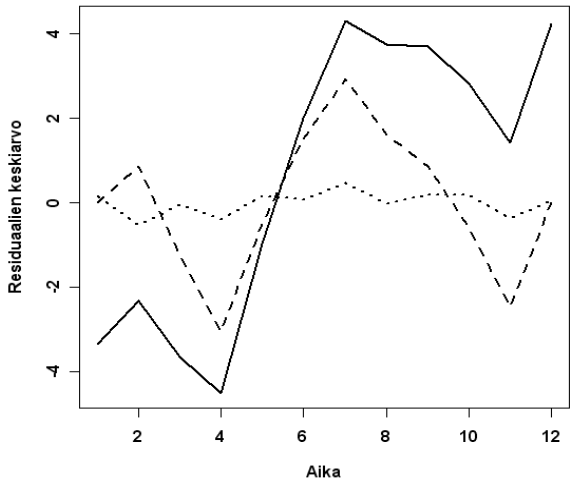
Kun oletetaan, että residuaalit ovat korreloimattomia odotusarvolla nolla, saa lauseke (4.7) muodon

$$\text{RMSE}_j = \sqrt{\frac{1}{40}\sum_{i=1}^{40}(-\hat{\epsilon}_i)^2}.$$

Kuviossa 4.3 on piirretty residuaalien keskiarvot kussakin aikapisteessä ja vastaavat arvot on esitetty taulukossa 4.2. Voidaan havaita, että residuaalien keskiarvo on selvästi pienin mallissa C. Mallissa A voidaan nähdä selvää aliestimointia periodin alussa ja yliestimointia loppupuolella periodia.

Taulukko 4.2. Residuaalien keskiarvot malleille A, B ja C.

Malli	1	2	3	4	5	6	7	8	9	10	11	12
A	-3.33	-2.32	-3.64	-4.49	-0.97	2.00	4.30	3.76	3.71	2.83	1.43	4.23
B	0.00	0.87	-1.24	-3.04	-0.50	1.52	2.93	1.60	0.86	-0.60	-2.46	0.00
C	0.17	-0.52	-0.05	-0.38	0.17	0.07	0.48	-0.03	0.18	0.19	-0.37	0.04



Kuvio 4.3. Residuaalien keskiarvot malleille A (kiinteä viiva), B (katkoviiva) ja C (pisteviiva).

Taulukossa 4.3 on kuvattu vastaavat residuaalien aikapisteittäin lasketut RMSE-arvot. Nyt ero mallin C hyväksi ei ole enää yhtä selkeä, vaan mallin C RMSE-arvot ovat vain hieman mallin B RMSE-arvoja pienempiä.

Taulukko 4.3. Residuaalien RMSE-arvot aikapisteissä malleille A, B ja C.

Malli	1	2	3	4	5	6	7	8	9	10	11	12
A	7.63	6.71	8.53	10.44	12.31	15.34	18.51	22.20	26.02	30.03	33.28	38.45
B	6.86	6.36	7.82	9.90	12.28	15.28	18.24	21.93	25.77	29.91	33.34	38.21
C	6.86	6.32	7.72	9.43	12.27	15.21	18.01	21.88	25.76	29.90	33.25	38.21

Oheisilla kriteereillä mitattuna malli C osoittautuu kolmesta vertailtavasta mallista parhaaksi. Tämän perusteella voidaan päätellä, että tasoittava kuutiosplini tarjoaa tehokkaan välineen estimoitaessa sekä odotusarvoa, että kovarianssimatriisia. Tässä työssä on kuitenkin tehty vielä monia yksinkertaistuksia estimoitaessa parametrien $\log(\sigma_j^2)$ ja $\phi_{j,k}$ arvoja tasoittavilla kuutiosplineillä ja täsmällisemmät sovitteet saavutettaisiinkin maksimoimalla Liun (2004) tapaan reunauskottavuusfunktio parametrien $\log(\sigma_j^2)$ ja ϕ_{jk} suhteen. Pitkittäisaineistolle tämä vaatii kuitenkin vielä lisää tutkimustyötä.

5 Loppusanat

Tutkielmassa osoitetaan, kuinka splinien avulla voidaan tehokkaasti estimoida sellaisen aineiston odotusarvokäyrää, jonka mallintamiseen perinteinen parametrinen regressiomalli ei sovellu. Voidaan havaita, kuinka erityisesti tasoittava kuutiosplini tarjoaa tähän tehokkaan työkalun. Lisäksi osoitetaan, miten tasoittavan kuutiosplinin käyttökelpoisuus ei rajoitu ainoastaan odotusarvon estimointiin, vaan sen avulla voidaan joustavasti estimoida myös kovarianssimatriisin parametreja.

Vaikka odotusarvon estimointia splinien avulla voidaan pitää suhteellisen tuoreena tilastollisena menetelmänä, on siitä kuitenkin kirjallisuudessa olemassa jonkin verran materiaalia. Sen sijaan kovarianssimatriisin mallintaminen käyttämällä apuna modifioitua Choleskyn hajotelmaa, jonka termit on estimoitu tasoittavilla kuutiosplineillä, on vasta tutkimuksen alla oleva menetelmä. Tässä työssä osoitetaan kyseisen menetelmän toimivan käytännössä ja tarjoavan täten potentiaalisen aihepiirin tuleville tutkimuksille.

Tutkielmassa modifioidun Choleskyn hajotelman termien parametrit on saavutettu käyttämällä lähtökohtana otoskovarianssimatriisia. Sekä yleistettyjen autoregressiivisten parametrien että logaritmoitujen virhevarianssien estimaatit on saavutettu estimoimalla molempia erikseen tasoittavilla kuutiosplineillä. Molempien spliniestimaattorien tasoitusparametrit on niin ikään valittu erikseen ristiinvalidoinnin avulla. Näin saavutettu kovarianssimatriisi on sijoitettu GSS-estimaattorin lausekkeeseen (3.24) ja saatu näin lopulta odotusarvokäyrän estimaatit.

Koska menetelmässä estimoidaan erikseen sekä kovarianssimatriisin parametrit että odotusarvo, ei saavutettuja estimaattiarvoja voida pitää optimaalisina. Pourahmadin (1999) esitystä seuraten, odotusarvon ja kovarianssimatriisin parametrit voitaisiin estimoida samanaikaisesti logaritmoidun uskottavuusfunktion avulla, mikäli mallina toimisivat polynomit ajan suhteen. Tutkielmassa kuitenkin sekä odotusarvoa, että kovarianssimatriisin parametreja estimoidaan tasoitetuilla kuutiosplineillä, jolloin lisää tutkimustyötä tarvitaan yhteisestimoinnin mahdollistavien algoritmien kehittämiseksi. Liun (2004) sakotetun logaritmoidun uskottavuusfunktion maksimointiin perustuva menetelmä tarjoaa tähän lähtökohdan.

Lopuksi haluaisin vielä lausua kiitoksen sanan ohjaajalleni dosentti Tapio Nummelle, joka ehdotti minulle tätä haastavaa, mutta kiinnostavaa aihetta, perehdytti minut ennestään tuntemattomaan aihepiiriin ja tarjosi pitkin työn etenemistä arvokkaita neuvoja ja tukea.

Lähdeluettelo

Brumback, B. & Rice, J. A. (1998), "Smoothing spline models for the analysis of nested and crossed samples of curves", *Journal of the American Statistical Association*, 93, 961–994.

Craven, P. & Wahba, G. (1979), "Smoothing noisy data with spline functions", *Numerische Matematik*, 31, 377–390.

Davidian, M. & Giltinan, D. M. (1993), "Some general estimation methods for nonlinear mixed effects models", *Journal of Biopharmaceutical Statistics*, 3, 23–55.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker: New York.

——— (1999), *Nonparametric Regression and Spline Smoothing*, Marcel Dekker: New York.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004), *Applied Longitudinal Analysis*, New Jersey: Wiley-Interscience.

Friedman, J. H. (1991), "Multivariate adaptive regression splines", *Annals of Statistics*, 19, 1–68.

Friedman, J. H. & Silverman B. W. (1989), "Flexible parsimonious smoothing and additive modeling", *Technometrics*, 31, 3–39.

Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman & Hall.

Harville, D. A. (1976), "Extension of the Gauss-Markov theorem to include the estimation of random effects", *Annals of Statistics*, 4, 384–395.

——— (1977), "Maximum likelihood approaches to variance component estimation and to related problems", *Journal of American Statistical Association*, 72, 320–340.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Hoover, D. R, Rice, J. A., Wu, C. O. & Yang, L. P. (1998) "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data", *Biometrika*, 85, 809–822.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002), "Varying-coefficient models and basis function approximations for longitudinal/clustered data", *Biometrika*, 89, 111–128.

Koskela, L., Nummi, T., Wenzel, S. & Kivinen, V. (2006), " On the analysis of cubic smoothing spline-based stem curve prediction for forest harvesters", *Canadian Journal of Forest Research*, 36, 2909–2919.

Laird, N. M. & Ware, J. H. (1982), "Random-effects models for longitudinal data", *Biometrics*, 38, 963–974.

Lin, X. & Zhang, D. (1999), "Inference in generalized additive mixed models by using smoothing splines", *Journal of Royal Statistical Society,* Series B, 61, 381–2.

Lindström, U. & Maijala, K. (1970), "Evaluation of performance test results for A.I. bulls", *Acta Agriculturae Scandinavica*, 20, 207–217.

Liski, E. P. (1987), "A growth curve analysis for bulls tested at station", *Biometrical Journal*, 29, 331–343.

Liu, N. (2004), "Covariance Selection and Estimation and the Value/Growth Spreads as Predictors of Returns.", Ph.D. dissertation, University of Pennsylvania.

Müller, H. G. (1988) *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics, New York: Springler-Verlag.

Möttönen, J. & Nummi, T. (2002), "Scots pine stem-curve predictions", in *Proceedings of the Woodfor Africa 2002*, Forest Engineering Conference: Forest Engineering Solutions for Achieving Sustainable Forest Resource Management - An International Perspective, Hilton College, Pietermanitzbury, South Africa, 2–3 July, 2002, eds. L. Kellog, B. Spong & P. Licth, Oregon State University, Department of Forest Engineering, Corvallis, Oregon, pp. 186–190.

Newton, H. J. (1988) *TIMESLAB: A Time Series Analysis Laboratory*, Pacific Grove, CA : Wadsworth & Brooks/Cole.

Nummi, T. (1997), "Estimation in random effects growth curve model", *Journal of Applied Statistics*, 24, 157–168.

Nummi, T. & Koskela, L. (2007), "Analysis of growth curve data using cubic smoothing splines", Submitted to *Journal of Applied Statistics*.

Nummi, T., Kääriä, S. & Pan, J. (2007), "Analysis of longitudinal data using cubic smoothing splines", unpublished manuscript.

Nummi, T. & Möttönen, J. (2004a), "Prediction of stem measurements of Scots pine", *Journal of Applied Statistics*, 31, 105–114.

——— (2004b), "Estimation and prediction for low degree polynomial models under measurement errors with an application to forest harvesters", *Applied Statistics*, 53, 495–505.

Pan, J. & von Rosen, D. (2005), "Modelling heterogeneous covariances in the growth curve models", Research Report 20, Probability and Statistics Group School of Mathematics, The University of Manchester.

Potthoff, R. F. & Roy, S. N. (1964), "A generalized multivariate analysis of variance model useful especially for growth curve problems", *Biometrika*, 86, 677–690.

Pourahmadi, M. (1999), "Joint mean-covariance models with applications to longitudinal data: Unconstrained parametrisation", *Biometrika*, 86, 677–690.

——— (2000), "Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix", *Biometrika*, 87, 425–435.

Reinsch, C. (1967), "Smoothing by spline functions", *Numerische Mathematik*, 10, 177–183.

Rice, J. A. & Silverman, B. W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves", *Journal of the Royal Statistical Society,* Series B, 53, 233–243.

Rice, J. A. & Wu, C. O. (2001), "Nonparametric mixed effects models for unequally sampled noisy curves", *Biometrics*, 57, 253–259.

Robinson, G. K. (1991), "That BLUP is a good thing: the estimation of random effects (with discussions)", *Statistics Science*, 6, 15–32.

Shi, M., Weiss, R. E. & Taylor, J. M. (1996), "An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves", *Applied Statistics*, 45, 151–163.

Silverman, B. W. (1985), "Some aspects of the spline smoothing approach to non-parametric regression curve fitting", *Journal of the Royal Statistical Society*, Series B, 47, 1–52.

Smith, P. L. (1979), "Splines as a useful and convenient statistical tool", *American Statistician*, 33, 57–62.

Smith, M. & Kohn, R (1996), "Nonparametric regression via Bayesian variable selection", *Journal of Econometrics*, 75, 317–344.

Speed, T. P. (1991), "Discussion of "That BLUP is a good thing: The estimation of random effects" by Robinson", *Statistics Science*, 6, 42–44.

Todd, J. (1962), *Survey of Numerical Analysis*, New York: McGraw-Hill.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. & Welham, S. J. (1999), "The analysis of designed experiments and longitudinal data by using smoothing splines", *Applied Statistics*, 48, 269–311.

Wahba, G. (1977), "A survey of some smoothing prblems and the method of generalized cross-validation for solving them", in *Applications of Statistics*, (P. R. Krisnaiah, ed.), 507–523. North Holland, Amsterdam.

——— (1990), *Spline Models for Observational Data.* SIAM, Philadephia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wang, Y. (1998a), "Mixed-effects smoothing spline ANOVA", *Journal of Royal Statistical Society,* Series B, 60, 159–174.

——— (1998b), "Smoothing spline models with correlated random errors", *Journal of the American Statistical Association*, 93, 341–348.

Wu, H. & Zhang, J.-T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis*, New Jersey: John Wiley & Sons.

Ye, H. & Pan, J. (2006), "Modelling of covariance structures in generalised estimating equations for longitudinal data.", *Biometrika*, 93, 927–941.