Ydinmenetelmä riippumattomien komponenttien analyysiin

Lauri Heinonen

Pro gradu -tutkielma
Kesäkuu 2021

Tarkastajat:
FT J.V.
TkT N.L.

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Lauri Heinonen: Ydinmenetelmä riippumattomien komponenttien analyysiin
Pro gradu -tutkielma, 34 s., 1 liite.
Tilastotiede
Kesäkuu 2021

Tutkielmassa esitellään ja johdetaan uusi menetelmä, ydin-FOBI, joka on ydinmenetelmä riippumattomien komponenttien analyysiin. Lisäksi esitellään MDS-FOBI, jonka avulla FOBI-ratkaisu voidaan tuottaa pelkän havaintojen etäisyysmatriisin perusteella. Johdantona aiheeseen esitellään ja johdetaan pääkomponenttianalyysi, sen ydinversio ja moniulotteinen skaalaus sekä esitellään riippumattomien komponenttien analyysi ja johdetaan sen lineaarinen FOBI-ratkaisu. Lopuksi käsiteltyjä menetelmiä vertaillaan kolmella aineistolla.

Riippumattomien komponenttien analyysissä havaintovektorin muuttujien ajatellaan olevan riippumattomien satunnaismuuttujien lineaarikombinaatioita. Tarkoitus on palauttaa vaihtelu takaisin näihin komponentteihin. FOBI on eräs riippumattomien komponenttien ongelman ratkaisu ja se perustuu neljänsien momenttien muodostaman kurtoosimatriisin ominaisarvohajotelmaan.

Tutkielmassa esitetään tapa FOBIn laskemiseen käyttäen vain havaintojen sisätulomatriisia. Kun sisätulomatriisi korvataan ydinmatriisilla, saadaan ydin-FOBI ja kun se korvataan tietyllä etäisyysmatriisiin pohjautuvalla matriisilla, saadaan MDS-FOBI. Menetelmiä tutkittaessa havaitaan, että ydin-FOBI voidaan nähdä ydinpääkomponenttianalyysinä, jonka antamiin pistemääriin sovelletaan lineaarista FOBIa.

Simuloiduilla aineistoilla tehdyssä tarkastelussa havaitaan, että ydin-FOBIn tuottamia komponentteja voidaan käyttää ryhmien erotteluun aineistosta. Toisesta aineistolla tehdystä esimerkistä havaitaan, että ydin-FOBI soveltuu myös niin sanottujen ominaiskasvojen tuottamiseen. Sopivan ydinfunktion etuna on tällöin, että se erottaa kuvista reunat melko terävästi, vaikka kuvien välillä kasvot ovatkin hieman eri paikoissa. Kolmas esimerkki taas osoittaa, että MDS-FOBIa voidaan käyttää tavallisen moniulotteisen skaalauksen lailla. Tällöin MDS-FOBIn ominaisuutena on, että se erottelee pisteet hieman tavallista moniulotteista skaalausta voimakkaammin ryhmiin.

Asiasanat: riippumattomien komponenttien analyysi, pääkomponenttianalyysi, ydinmenetelmät, FOBI, MDS

Sisällys

1 Johdanto

Riippumattomien komponenttien analyysissä (independent component analysis, ICA) [1] havaittavien muuttujien ajatellaan olevan tuntemattomien satunnaisten lähteiden tuntemattomia lineaarikombinaatioita. Erityisesti oletetaan, että nämä lähteet ovat riippumattomia. Menetelmän tarkoituksena on palauttaa havaitut muuttujat taustalla oleviksi riippumattomiksi muuttujiksi. Riippumattomien komponenttien ongelman eräs ratkaisutapa on neljännen asteen sokkotunnistus (fourth-order blind identification, FOBI) [2], joka perustuu neljänsien momenttien muodostaman kurtoosimatriisin ominaisarvohajotelmaan.

Pääkomponenttianalyysi (principal component analysis, PCA) [3] on perinteinen ja erittäin käytetty menetelmä, jossa pyritään löytämään aineiston vaihtelua mahdollisimman hyvin selittäviä, korreloimattomia alkuperäisten muuttujien lineaarikombinaatioita. Usein näistä valitaan muutama ensimmäinen ja näin pyritään esittämään aineisto järkevällä tavalla pienemmällä määrällä muuttujia, esimerksi kaksiulotteisena kuvana. Tätä kutsutaan dimensionpienennykseksi. Pääkomponenttianalyysi perustuu aineiston kovarianssimatriisin ominaisarvohajotelmaan ja pääkomponenttien suunnat ovatkin juuri ominaisvektoreita.

Pääkomponenttianalyysi ja riippumattomien komponenttien analyysi ovat tavoitteiltaan samanlaisia, sillä niiden molempien tarkoituksena on esittää aineisto uusien, alkuperäisten muuttujien lineaarikombinaationa tuotettujen, komponenttien avulla. Pääkomponenttianalyysissä nämä komponentit ovat korreloimattomia, riippumattomien komponenttien analyysissä vielä vahvemmin riippumattomia. Pääkomponenttianalyysi perustuu kovarianssimatriisin ja FOBI tämän lisäksi vielä kurtoosimatriisin diagonalisointiin. Molempien menetelmien rajoitteena on, että ne ovat lineaarisia, koska ne perustuvat lineaarikombinaatioihin. Tästä syystä ne eivät tunnista aineistossa olevia epälineaarisia rakenteita.

On luonnollista miettiä, voitaisiinko menetelmistä luoda yleisempi epälineaarinen versio. Erään mahdollisuuden tähän tarjoavat tukivektorikoneiden [4] yhteydessä esitellyt ydinfunktiot (kernel function), joiden avulla tuotetut ydinmenetelmät ovat yleisesti käytössä modernissa tilastotieteessä ja koneoppimisessa. Schölkopf, Smola ja Müller julkaisivatkin 1998 ydinpääkomponenttianalyysin (kernel principal component analysis) [5]. Se perustuu havaintojen kuvaamiseen jollakin epälineaarisella funktiolla korkeaulotteisempaan avaruuteen ja pääkomponenttianalyysin tekemiseen tässä avaruudessa. Oleellista on kuitenkin, ettei menetelmää käytettäessä tarvitse laskea tai edes tietää tätä havaintojen kuvaamiseen käytettävää funktiota, vaan pelkkä sitä vastaava ydinfunktio. Tarkemmin tämä ydinfunktio vastaa sisätuloa havaintojen kuvien avaruudessa. Näin ydinmenetelmien johtaminen perustuukin mallin esittämiseen havaintojen pelkkien sisätulojen avulla.

Kun pääkomponenttianalyysi esitetään sisätulojen avulla, havaitaan myös, että sillä on vahva yhteys toiseen menetelmään, moniulotteiseen skaalaukseen (multidimensional scaling, MDS) [6, luku 14], joka voidaan nähdä pääkomponenttianalyysinä aineistolle, jossa havaintojen (tai niiden sisätulojen) sijaan tiedossa ovatkin vain havaintojen väliset etäisyydet. Moniulotteisessa skaalauksessa halutaan esittää sellainen, usein kaksi- tai kolmiulotteinen pistejoukko, jonka euklidiset etäisyydet vastaavat mahdollisimman hyvin annettuja etäisyyksiä. Esimerkki tästä on tuottaa

maailmankartta, kun tiedetään kaupunkien väliset etäisyydet maapallon pinnalla.

Tämän opinnäytteen tarkoituksena on johtaa kaksi uutta menetelmää: ydin-FOBI ja MDS-FOBI, joista ensimmäinen on ydinfunktioiden käytön avulla luotu epälineaarinen versio FOBIsta ja toinen havaintojen etäisyyksien perusteella laskettava FOBI. Tämä tehdään seuraavasti: Esitellään aluksi pääkomponenttianalyysi, sen ydinversio ja moniulotteinen skaalaus sekä osoitetaan näiden suhteet toisiinsa. Tämän jälkeen esitellään riippumattomien komponenttien analyysi ja sen FOBI-ratkaisu. Uutena tuloksena johdetaan keino FOBI-ratkaisun tuottamiseen vain havaintojen sisätulojen kautta. Kun tähän yhdistetään pääkomponenttianalyysin kohdalla käsitelty teoria, saadaan uudet menetelmät.

Luvussa 2 pohjustetaan työtä käymällä läpi pääkomponenttianalyysiä ja siihen liittyviä menetelmiä. Luvun alussa esitellään ja johdetaan pääkomponenttianalyysi sekä sen laskeminen havaintojen sisätulojen avulla. Lyhyesti tarkastellaan myös pääkomponenttien määrän valintaa. Tämän jälkeen esitellään ydinpääkomponenttianalyysin johto alkuperäisartikkelissa [5] esitellyllä tavalla. Tämän perään käsitellään lyhyesti mahdollisia ydinfunktioita. Sitten esitellään ja johdetaan moniulotteinen skaalaus ja näytetään sen yhteys pääkomponenttianalyysiin. Sen perään esitellään metrinen moniulotteinen skaalaus, jolla on yhteys ydinpääkomponenttianalyysiin. Tässä luvussa käsitellään paljon sellaista teoriaa (singulaariarvohajotelma, ydinfunktiot, MDS jne.), joka pätee suoraan myös ydin-FOBIn ja MDS-FOBIn tapaukseen.

Luvussa 3 esitellään riippumattomien komponenttien analyysi ja sen FOBI-ratkaisu sekä johdetaan ydin-FOBI. Aluksi esitellään riippumattomien komponenttien analyysin taustalla oleva perustehtävä ja perustellaan siihen liittyvät oletukset. Tämän jälkeen esitellään kurtoosimatriisi sekä FOBI-ratkaisu ja perustellaan tämän toiminta. Lisäksi näytetään, miten FOBI-ratkaisu lasketaan havaintomatriisista. Tämän jälkeen johdetaan FOBIn laskeminen havaintojen sisätulojen pohjalta ja näin esitellään ydin-FOBI ja MDS-FOBI. Lopuksi käydään läpi vielä joitain menetelmien piirteitä.

Luvussa 4 esitetään kolme laskennallista esimerkkiä työssä käsitellyistä menetelmistä: ryhmien erottelu simuloidulla aineistolla ydin-FOBIlla, ominaiskasvojen (eigenfaces) [7] tuottaminen eri menetelmillä sekä kaupunkien sijaintien estimoiminen etäisyysaineistosta MDS:llä ja MDS-FOBIlla.

2 Pääkomponenttianalyysi ja moniulotteinen skaalaus

Tässä luvussa esitellään pääkomponenttianalyysi, perustellaan sen laskeminen ominaisarvohajotelman avulla, näytetään, miten tämä lasku tehdään havaintomatriisista ja kerrotaan muutama tapa pääkomponenttien määrän valintaan. Sen jälkeen näytetään, miten pääkomponentit voi laskea vain havaintojen välisten sisätulojen avulla ja esitellään tästä ajatuksesta seuraava ydinpääkomponenttianalyysi. Lopulta vielä esitellään ja johdetaan moniulotteinen skaalaus ja näytetään sen yhteys pääkomponenttianalyysiin. Sitten esitellään lyhyesti metrinen moniulotteinen skaalaus, jolla taas on yhteys ydinpääkomponenttianalyysiin.

2.1 Pääkomponenttianalyysi

Ulottuvuuksien pienennysmenetelmien (dimensionality reduction) tarkoituksena on löytää jokin alempiulotteinen avaruus, jossa aineiston voi esittää niin, että hävitetään mahdollisimman vähän informaatiota. Pääkomponenttianalyysi (principal component analysis, PCA), joka esiteltiin alunperin artikkelissa [8], on perinteisin tällainen menetelmä ja siinä informaation mittana käytetään varianssia. Tarkoituksena on löytää järjestyksessä muuttujien lineaarikombinaatioita, joista jokaisen varianssi on suurin mahdollinen ja joista jokainen on korreloimaton kaikkien edellisten kanssa.

2.1.1 Menetelmän johto

Esitetään johto kirjan [3] luvun 1.1 mukaisesti. Olkoon $\mathbf{x} \in \mathbb{R}^p$ satunnaisvektori ja $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ sen kovarianssimatriisi. Oletetaan, että $\mathbf{\Sigma}$ on täysiasteinen, jolloin positiividefiniittinä sen kaikki ominaisarvot ovat positiivisia (jos aste $d < p$, tuottaa menetelmä vain d komponenttia). Halutaan löytää sellainen uusien muuttujien, pääkomponenttien $z_j = \mathbf{u}_j'\mathbf{x}$ ($\mathbf{u}_j \in \mathbb{R}^p$), järjestetty jono, jossa jokaisen varianssi $\mathbb{V}\mathrm{ar}[z_j] = \mathbf{u}_j'\mathbf{\Sigma}\mathbf{u}_j$ on mahdollisimman suuri ja jossa jokaisen kovarianssi aiempien (ja siten kaikkien myöhempienkin) kanssa on nolla eli $\mathbb{C}\mathrm{ov}[z_j, z_{j'}] = \mathbf{u}_j'\mathbf{\Sigma}\mathbf{u}_{j'} = 0$ kaikille $j' < j$. Tämän lisäksi vektorit $\mathbf{u}_j$ ovat yksikköpituisia eli $\mathbf{u}_j'\mathbf{u}_j = 1$. Ensimmäinen vektori siis ratkaisee optimointitehtävän

$$\begin{aligned} \max_{\mathbf{u}_1\in\mathbb{R}^p} \quad & \mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_1 \\ \text{s.t.} \quad & \mathbf{u}_1'\mathbf{u}_1 = 1. \end{aligned}$$

Käytetään Lagrangen kerroinmenetelmää eli etsitään stationaariset pisteet lausekkeelle $\mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_1 + \lambda_1(1-\mathbf{u}_1'\mathbf{u}_1)$, jossa λ_1 on Lagrangen kerroin. Näin ratkaisuksi saadaan

$$\begin{aligned} & \frac{\partial}{\partial \mathbf{u}_1}\left[\mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_1 + \lambda_1(1-\mathbf{u}_1'\mathbf{u}_1)\right] = \mathbf{0} \\ \iff \quad & 2\mathbf{\Sigma}\mathbf{u}_1 - 2\lambda_1\mathbf{u}_1 = \mathbf{0} \\ \iff \quad & \mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1. \end{aligned}$$

Nähdään, että λ_1 on matriisin $\mathbf{\Sigma}$ ominaisarvo ja $\mathbf{u}_1$ sitä vastaava ominaisvektori. Tutkitaan maksimoitavaa suuretta $\mathbb{V}\mathrm{ar}[z_1] = \mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1'\mathbf{u}_1 = \lambda_1$. Suurimman varianssin saamiseksi kerroinvektoriksi $\mathbf{u}_1$ tulee siis valita matriisin $\mathbf{\Sigma}$ suurinta ominaisarvoa vastaava ominaisvektori, jolloin varianssiksi saadaan kyseinen ominaisarvo.

Johdetaan nyt toinen pääkomponentti. Kun käytetään tietoa, että $\mathbf{u}_1$ on kovarianssimatriisin $\mathbf{\Sigma}$ ominaisvektori, saadaan $\mathbb{C}\mathrm{ov}[z_2, z_1] = \mathbf{u}_2'\mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_2'\mathbf{u}_1$. Kun oletetaan $\lambda_1 \neq 0$, voidaan ehto $\mathbb{C}\mathrm{ov}[z_2, z_1] = 0$ esittää muodossa

$$\mathbf{u}_2'\mathbf{u}_1 = 0.$$

Toinen pääkomponentti saadaan siis optimointitehtävästä

$$\begin{aligned} \max_{\mathbf{u}_2 \in \mathbb{R}^p} \quad & \mathbf{u}_2'\mathbf{\Sigma}\mathbf{u}_2 \\ \text{s.t.} \quad & \mathbf{u}_2'\mathbf{u}_2 = 1 \\ & \mathbf{u}_2'\mathbf{u}_1 = 0, \end{aligned}$$

jonka Lagrangen funktio on $\mathbf{u}_2'\mathbf{\Sigma}\mathbf{u}_2 + \lambda_2(1 - \mathbf{u}_2'\mathbf{u}_2) + \mu\mathbf{u}_2'\mathbf{u}_1$. Sen ratkaisu on puolestaan

$$\begin{aligned} & \frac{\partial}{\partial \mathbf{u}_2}\left[\mathbf{u}_2'\mathbf{\Sigma}\mathbf{u}_2 + \lambda_2(1 - \mathbf{u}_2'\mathbf{u}_2) + \mu\mathbf{u}_2'\mathbf{u}_1\right] = \mathbf{0} \\ \iff \quad & 2\mathbf{\Sigma}\mathbf{u}_2 - 2\lambda_2\mathbf{u}_2 + \mu\mathbf{u}_1 = \mathbf{0} \qquad (1) \\ \implies \quad & \mathbf{u}_1'(2\mathbf{\Sigma}\mathbf{u}_2 - 2\lambda_2\mathbf{u}_2 + \mu\mathbf{u}_1) = 0 \\ \iff \quad & 2\mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_2 - 2\lambda_2\mathbf{u}_1'\mathbf{u}_2 + \mu\mathbf{u}_1'\mathbf{u}_1 = 0 \\ \iff \quad & 2\mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_2 + \mu = 0. \end{aligned}$$

Kun tiedetään, että $\mathbf{u}_2'\mathbf{\Sigma}\mathbf{u}_1 = 0$ ja että $\mathbf{\Sigma}$ on symmetrinen, saadaan yhtälöstä $2\mathbf{u}_1'\mathbf{\Sigma}\mathbf{u}_2 + \mu = 0$ pääteltyä $\mu = 0$. Siten yhtälöstä (1) saadaan $\mathbf{\Sigma}\mathbf{u}_2 = \lambda_2\mathbf{u}_2$.

Vastaavalla tavalla voidaan johtaa loputkin pääkomponentit. Huomataan siis, että pääkomponenttien varianssit ovat kovarianssimatriisin ominaisarvoja (suurimmasta pienimpään) ja kerroinvektorit vastaavia ominaisvektoreita.

2.1.2 Pääkomponenttien laskeminen otoksesta

Otostasolla pääkomponenttianalyysi tehdään tarkastelemalla satunnaismuuttujien kovarianssimatriisin $\mathbf{\Sigma}$ sijaan havaintojen otoskovarianssimatriisia $\mathbf{S}$ ja sen ominaisarvohajotelmaa. Olkoon $\mathbf{X} \in \mathbb{R}^{n\times p}$ havaintomatriisi. Määritellään vektori $\mathbf{1} = (1, 1, \ldots, 1)' \in \mathbb{R}^n$ ja keskistämismatriisi $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$. Nyt havaintojen $\mathbf{X}$ keskiarvovektori on $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$ ja otoskovarianssimatriisi $\mathbf{S} = \frac{1}{n-1}(\mathbf{HX})'(\mathbf{HX}) = \frac{1}{n-1}\mathbf{X}'\mathbf{HX}$ [9, luku 3.3].

Olkoon $\mathbf{\Lambda}$ matriisin $\mathbf{S}$ ominaisarvojen diagonaalimatriisi ja $\mathbf{U}$ matriisi, jonka sarakkeina ovat vastaavat ominaisvektorit. Koska matriisi $\mathbf{S}$ on symmetrinen, sen ominaisvektorit ovat ortgonaalisia ja siten matriisi $\mathbf{U}$ on ortogonaalinen. Otoskovarianssimatriisin ominaisarvohajotelma on siis $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$. Olkoon $\mathbf{Z} \in \mathbb{R}^{n\times p}$ matriisi, jonka sarakkeina ovat havainnoista laskettujen pääkomponenttimuuttujien arvot. Tehdään yleinen valinta ja käytetään keskistettyjä muuttujia $\mathbf{HX}$. Silloin saadaan

$\mathbf{Z} = \mathbf{HXU}$. Koska matriisin $\mathbf{U}$ sarakkeet kertovat, millä painolla mikäkin alkuperäinen muuttuja on osa uutta pääkomponenttimuuttujaa, niiden alkioita kutsutaan *latauksiksi* ja matriisia $\mathbf{U}$ *latausmatriisiksi.*

2.1.3 Pääkomponenttien määrän valinta

Lähtökohtaisesti pääkomponenttianalyysi tuottaa yhtä monta pääkomponenttia kuin aineistossa on muuttujia (olettaen, ettei muuttujien välillä ole täyskorrelaatiota). Lähes aina kuitenkin pääkomponenttianalyysillä pyritään vähentämään muuttujien määrää. Tämä saavutetaan jättämällä joitain pääkomponentteja pois. Yleensä poisjätetyt komponentit vastaavat pienimpiä ominaisarvoja eli pienimpiä pääkomponenttien variansseja.

Jos tarkoituksena on havaintojen esittäminen kuvana, valitaan yleensä kaksi tai kolme ensimmäistä pääkomponenttia. Muissa tapauksissa pääkomponentteja valitaan jollakin menetelmällä sopiva määrä. Menetelmät perustuvat yleensä ominaisarvoihin ja ovat usein ainakin osittain subjektiivisia. Esitellään nyt lyhyesti kirjan [3] luvun 6 mukaisesti erilaisia menetelmiä.

Koska kovarianssimatriisin diagonaalilla ovat muuttujien x_j varianssit, saadaan niiden summa eli *kokonaisvarianssi* kovarianssimatriisin jälkenä (trace). Lisäksi tiedetään, että jälki saadaan ominaisarvojen summana. Siispä k ensimmäistä pääkomponenttia, joiden varianssit ovat $\lambda_1, \lambda_2, \ldots, \lambda_k$, selittävät osuuden $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$ kokonaisvarianssista. Yleisenä ohjeena pidetään, että tämän kumulatiivisen osuuden pitäisi olla n. $70\% - 90\%$. Kuitenkin, jos ensimmäiset yksi tai kaksi pääkomponenttia selittävät lähes kaiken vaihtelun, voi silti olla hyödyllistä ottaa mukaan muitakin komponentteja, sillä ne saattavat sisältää vaikeammin havaittavaa informaatiota. Toisaalta jos alkuperäisessä aineistossa on valtavasti muuttujia, voi olla että 70% rajaan vaaditaan tarpeettoman monta pääkomponenttia.

Ominaisarvojen käyttöä varianssin selittämisen mittarina voidaan perustella myös sen kautta, miten katkaistu ominaisarvohajotelma approksimoi kovarianssimatriisia. Olkoot $\mathbf{S} \in \mathbb{R}^{p \times p}$ kovarianssimatriisi, λ_i $(i = 1, 2, \ldots, p)$ sen ominaisarvot ja $\mathbf{u}_i$ vastaavat ominaisvektorit. Tällöin $\mathbf{S} = \sum_{i=1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i'$. Merkitään tästä saatavaa asteen $m \leq p$ approksimaatiota $\mathbf{S}_m = \sum_{i=1}^{m} \lambda_i \mathbf{u}_i \mathbf{u}_i'$. Määritellään vielä matriisin $\mathbf{A} = (a_{ij})$ euklidinen normi säännöllä $||\mathbf{A}|| = \sqrt{\sum_{ij} a_{ij}^2}$. Suoraviivaisella laskutoimituksella saadaan, että $||\mathbf{A}|| = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}')}$. Kun muistetaan, että ominaisvektorit ovat ortonormaaleja, saadaan

$$
\begin{aligned}
||\mathbf{S} - \mathbf{S}_m|| &= ||\sum_{i=m+1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i'|| \\
&= \sqrt{\mathrm{tr}(\sum_{i=m+1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i' \sum_{i=m+1}^{p} (\lambda_i \mathbf{u}_i \mathbf{u}_i')')} \\
&= \sqrt{\mathrm{tr}(\sum_{i=m+1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i' (\lambda_i \mathbf{u}_i \mathbf{u}_i')' + 2 \sum_{i<j} \lambda_i \mathbf{u}_i \mathbf{u}_i' (\lambda_j \mathbf{u}_j \mathbf{u}_j')')}
\end{aligned}
$$

$$= \sqrt{\sum_{i=m+1}^{p} \lambda_i^2 \operatorname{tr}(\mathbf{u}_i \mathbf{u}_i' \mathbf{u}_i \mathbf{u}_i') + 2 \sum_{i<j} \lambda_i \lambda_j \operatorname{tr}(\mathbf{u}_i \mathbf{u}_i' \mathbf{u}_j \mathbf{u}_j')}$$

$$= \sqrt{\sum_{i=m+1}^{p} \lambda_i^2 \operatorname{tr}(\mathbf{u}_i \mathbf{u}_i')}$$

$$= \sqrt{\sum_{i=m+1}^{p} \lambda_i^2}.$$

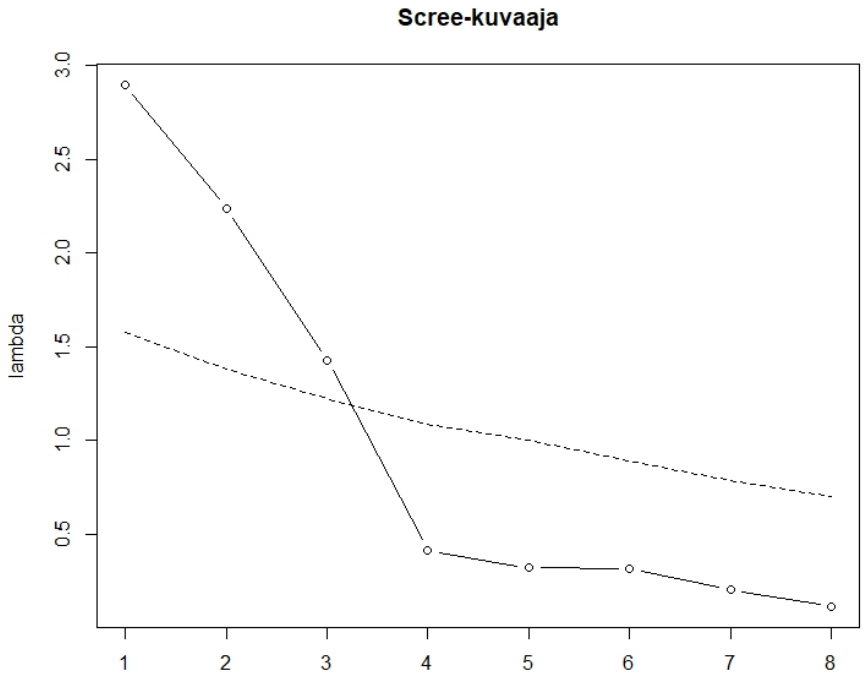
Tästä siis nähdään, että pieniä ominaisarvoja vastaavat termit summassa, ja siten komponentit $\mathbf{u}_i$, eivät sisällä merkitsevää informaatiota kovarianssimatriisista ja -rakenteesta.

Yksi vaihtoehto tarkastella ominaisarvoja on *scree-kuvaaja* (scree tarkoittaa vuoren kyljessä olevaa soraa). Siinä murtoviivalla yhdistetään pisteet, joiden pystykoordinaatti on ominaisarvo ja vaakakoordinaatti sen järjestysluku. Tarkoituksena on löytää jokin selkeä pudotus, jonka jälkeen tulevat ominaisarvot ovat huomattavasti aiempaa pienempiä. Toinen mahdollinen etsittävä piirre on kohta, jonka jälkeen peräkkäisten ominaisarvojen väliset erot ovat pieniä.

Scree-kuvaajaa voidaan tulkita vertaamalla saatuja ominaisarvoja niihin, joita saataisiin satunnaisista matriiseista, joissa ei ole muuttujien välisiä yhteyksiä. *Paralleelianalyysissä (Parallel analysis)* tuotetaan satunnaisia havaintomatriiseja, joissa on yhtä monta havaintoa ja muuttujaa kuin tutkittavassa oikeassa aineistossa ja joiden muuttujat ovat useimmiten riippumattomasti normaalijakautuneita. Näistä lasketaan pääkomponettianalyysi. Oikean aineiston k:nneksi suurinta pääkomponenttivarianssia verrataan satunnaisten matriisien k:nsien pääkomponenttivarianssien jakaumaan. Tässä menetelmässä käytetään kaikkien pääkomponenttien laskemiseen kovarianssimatriisin sijaan korrelaatiomatriisia, jotta ominaisarvot ovat vertailukelpoisia. Tällöin nimittäin ominaisarvojen summa eli alkuperäisten muuttujien varianssien summa on muuttujien määrä, koska kaikki muuttujat on skaalattu niin, että niiden varianssi on yksi. Menetelmä on alunperin kehitetty faktorianalyysiä varten, mutta sitä voidaan käyttää myös pääkomponenttianalyysin yhteydessä. Perussääntönä on, että pääkomponentti nähdään merkityksellisenä, jos sen varianssi on suurempi kuin 95% satunnaismatriiseista saaduista variansseista. Pääkomponentteja valitaan alusta siihen asti, kunnes jonkun komponentin varianssi ei ylitä tätä rajaa. Esimerkkinä menetelmän käytöstä on scree-kuvaajaan (kuva 1) piirretty viiva, jonka määrää paralleelianalyysin (500 toistoa) antama raja. Kuvan aineistossa havaintojen määrä $n = 100$ ja kahdeksasta muuttujasta kolme on riippumattomia ja normaalijakautuneita, loput näiden lineaarikombinaatioita. Paralleelianayysi ja silmämääräinenkin scree-kuvaajan tarkastelu paljastavat tämän hyvin.

2.2 Ydinpääkomponenttianalyysi

Seuraavaksi esitellään pääkomponenttianalyysin laskeminen, kun tiedetään vain havaintojen sisätulomatriisi. Tämä tilanne on itsessään sovelluksissa harvinainen, mutta oleellista onkin, että se tarjoaa mahdollisuuden ydinpääkomponenttianalyysiin,



Kuva 1: Scree-kuvaaja, jossa piirrettynä paralleelianalyysin määräämä raja

kun sisätulot korvataan ydinfunktion arvoilla. Ydinpääkomponenttianalyysistä esitetään myös alkuperäiseen artikkeliin [5] pohjautuva johto. Lisäksi esitetään lyhyesti ydinfunktioiden taustalla oleva ajatus ja esitellään yleisimpiä ydinfunktioita.

2.2.1 PCA sisätulojen avulla

Olkoon matriisin $\mathbf{HX} \in \mathbb{R}^{n \times p}$ singulaariarvohajotelma

$$\mathbf{HX} = \mathbf{V}\boldsymbol{\Delta}\mathbf{U}', \tag{2}$$

jossa $\mathbf{V} \in \mathbb{R}^{n \times n}$ ja $\mathbf{U} \in \mathbb{R}^{p \times p}$ ovat ortogonaalisia ja $\boldsymbol{\Delta} \in \mathbb{R}^{n \times p}$ (oletetaan, että $n \geq p$) muotoa

$$\boldsymbol{\Delta} = \begin{pmatrix} \delta_1 & 0 & \dots & 0 \\ 0 & \delta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \delta_p \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Muodostetaan hajotelma niin, että singulaariarvot δ_i ovat kaikki epänegatiivisia. Oletetaan lisäksi yksinkertaisuuden vuoksi, että singulaariarvot δ_i ovat erisuuria. Jos näin ei ole, tulosten sisältö ei oleellisesti muutu, mutta niiden esittäminen hankaloituu koska tällöin ominais- ja singulaarivektorit eivät enää ole yksikäsitteisiä (edes merkkiä vaille).

Silloin otoskovarianssimatriisi $\mathbf{S}$ voidaan esittää muodossa

$$\begin{aligned}\mathbf{S} &= \frac{1}{n-1}(\mathbf{HX})'(\mathbf{HX}) \\ &= \frac{1}{n-1}\mathbf{U}\boldsymbol{\Delta}'\mathbf{V}'\mathbf{V}\boldsymbol{\Delta}\mathbf{U}' \\ &= \frac{1}{n-1}\mathbf{U}\boldsymbol{\Delta}'\boldsymbol{\Delta}\mathbf{U}',\end{aligned}$$

joka on matriisin $\mathbf{S}$ ominaisarvohajotelma, kun merkitään $\boldsymbol{\Lambda} = \frac{1}{n-1}\boldsymbol{\Delta}'\boldsymbol{\Delta}$. Siis matriisin $\mathbf{S}$ ominaisarvohajotelman ja matriisin $\mathbf{HX}$ singulaariarvohajotelman matriisit $\mathbf{U}$ ovat todella sama matriisi. Singulaariarvojen ja ominaisarvojen välillä on yhteys $\lambda_i = \delta_i^2/(n-1)$. Lasketaan vielä singulaariarvohajotelmaa (2) käyttäen pääkomponenttimatriisi

$$\mathbf{Z} = \mathbf{HXU} = \mathbf{V}\boldsymbol{\Delta}\mathbf{U}'\mathbf{U} = \mathbf{V}\boldsymbol{\Delta}.$$

Mietitään nyt tilannetta, jossa tunnetaan vain havaintojen välinen *sisätulomatriisi* $\mathbf{XX}'$, jonka kohdassa (i, j) oleva alkio on muotoa $\mathbf{x}_i'\mathbf{x}_j$. Tätä matriisia nimitetään myös havaintojen *Gramin matriisiksi*. Sen avulla voidaan laskea matriisi $\mathbf{HXX}'\mathbf{H}$ joka on singulaariarvohajotelman (2) perusteella muotoa

$$\begin{aligned}\mathbf{HXX}'\mathbf{H} &= \mathbf{V}\boldsymbol{\Delta}\mathbf{U}'\mathbf{U}\boldsymbol{\Delta}'\mathbf{V}' \\ &= \mathbf{V}\boldsymbol{\Delta}\boldsymbol{\Delta}'\mathbf{V}'\end{aligned}$$

Nyt siis matriisin $\mathbf{HXX}'\mathbf{H}$ ominaisarvohajotelmasta voidaan selvittää matriisit $\mathbf{V}$ ja $\boldsymbol{\Delta}\boldsymbol{\Delta}'$ sekä jälkimmäisen neliöjuurena matriisi $\boldsymbol{\Delta}$. Näiden perusteella voidaan edelleen laskea pääkomponentit.

Koska havaintojen väliset sisätulot sisältävät aidosti vähemmän informaatiota kuin alkuperäiset havainnot (joista sisätulotkin olisi helppo laskea), ei ole intuitiivisesti suuri yllätys, että latausmatriisia $\mathbf{U}$ ei sisätulomatriisiin pohjautuvalla menetelmällä saada selville. Latausmatriisin sijaan voidaan alkuperäisten muuttujien ja saatujen pääkomponenttien välisiä yhteyksiä tutkia myös näiden välisestä kovarianssimatriisista $\mathbf{C}$ (tai usein paremmin korrelaatiomatriisin avulla). Laskettaessa kovarianssimatriisi (täysin tavallisesti) on toki tunnettava alkuperäiset havainnot (ei vain sisätuloja). Kovarianssia voidaankin käyttää erityisesti kohta esiteltävän ydinpääkomponenttianalyysin tapauksessa, jossa havainnot tunnetaan, mutta pääkomponenttianalyysi suoritetaan silti ikään kuin sisätulojen avulla (ja siten latausmatriisia $\mathbf{U}$ ei tunneta). Tunnettaessa latausmatriisi $\mathbf{U}$ voitaisiin kovarianssimatriisi laskea helposti laskulla $\mathbf{C} = \frac{1}{n-1}\mathbf{Z}'\mathbf{HX} = \frac{1}{n-1}(\mathbf{HXU})'\mathbf{HX} = \mathbf{U}'\frac{1}{n-1}\mathbf{X}'\mathbf{HX} = \mathbf{U}'\mathbf{S} = \mathbf{U}'\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' = \boldsymbol{\Lambda}\mathbf{U}'$. Tästä syystä laskemalla matriisi $\mathbf{C}$ saadaan myös jossakin mielessä korvattua latausmatriisin puutetta.

2.2.2 Ydin-PCA:n johto

Pisteiden sisätulomatriisin sijaan pääkomponenttianalyysissä voitaisiin käyttää jotakin muuta matriisia $\mathbf{K} \in \mathbb{R}^{n\times n}$, jonka kohdassa (i, j) olisi pisteiden $\mathbf{x}_i$ ja $\mathbf{x}_j$ avulla laskettu *ydinfunktion (kernel function)* k arvo $k(\mathbf{x}_i, \mathbf{x}_j)$. Tällaista matriisia $\mathbf{K}$ kutsutaan ytimen k määräämäksi *ydinmatriisiksi* tai Gramin matriisiksi. Näin saatava menetelmä on nimeltään *ydinpääkomponenttianalyysi* ja sen ajatus ydinfunktion käytöstä on sama kuin esimerkiksi tukivektorikoneissa [4].

Ydinpääkomponenttianalyysissä siis lasketaan ensin ydinfunktion k avulla ydinmatriisi $\mathbf{K}$. Sen jälkeen tuotetaan ydinmatriisin $\mathbf{K}$ (oikeastaan sen keskistetyn version $\mathbf{HKH}$) ominaisarvohajotelma $\mathbf{V\Delta\Delta'V'}$. Lopuksi lasketaan tästä komponenttipistemäärät $\mathbf{Z} = \mathbf{V\Delta}$.

Esitetään menetelmälle johto pohjautuen artikkeliin [5], jossa se alunperin esiteltiin. On hyvä huomata, että tässä luvussa käytetään aikaisemmista luvuista tuttuja merkintöjä tarkoittamaan havaintojen kuvien avaruuden alkioita ja niistä johdettuja suureita, kuten otoskovarianssimatriisia $\mathbf{S}$.

Olkoon $(F, \langle\cdot,\cdot\rangle)$ reaalinen sisätuloavaruus ja $\phi\colon \mathbb{R}^p \to F$ kuvaus euklidisesta avaruudesta $\mathbb{R}^p$ tähän avaruuteen. Ydin-PCA:n tarkoituksena on tehdä pääkomponenttianalyysi avaruudessa F. Oletetaan, että havaintojen kuvat avaruudessa F ovat keskistettyjä eli että $\sum_{i=1}^n \phi(\mathbf{x}_i) = 0$. Kuvien keskistämistä käsitellään tarkemmin myöhemmin. Jos F on äärellisulotteinen, voidaan kuvien $\phi(\mathbf{x}_i)$ otoskovarianssimatriisi $\mathbf{S}$ kirjoittaa tavalliseen tapaan $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)'$ (käytetään jakajana lukua n merkintöjen helpottamiseksi). Kovarianssimatriisia vastaava lineaarikuvaus saadaan myös jos F on epäeuklidinen, erityisesti jos se on ääretönulotteinen, tulkitsemalla ulkotulo $\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)'$ lineaarikuvauksena, joka kuvaa vektorin $\mathbf{v} \in F$ vektoriksi $\langle\phi(\mathbf{x}_i), \mathbf{v}\rangle\phi(\mathbf{x}_i)$. Käytettäessä merkintää $\mathbf{S}$ tarkoitetaan tätä kuvausta.

Tarkoituksena on selvittää tavanomaiseen tapaan kuvauksen $\mathbf{S}$ ominaisvektorit $\mathbf{u} \in F$ ja ominaisarvot λ, jotka siis toteuttavat ehdon $\mathbf{Su} = \lambda\mathbf{u}$. Kun tähän sijoitetaan kuvauksen $\mathbf{S}$ määritelmä ja jaetaan molemmat puolet luvulla λ (oletetaan $\lambda \neq 0$), saadaan $\mathbf{u} = \frac{1}{\lambda n}\sum_{i=1}^n \langle\phi(\mathbf{x}_i), \mathbf{u}\rangle\phi(\mathbf{x}_i)$. Ominaisvektorit $\mathbf{u}$ kuuluvat siis havaintojen kuvien $\phi(\mathbf{x}_i)$ virittämään äärellisulotteiseen aliavaruuteen ja ne voidaan esittää muodossa

$$\mathbf{u} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i), \tag{3}$$

jossa $\alpha_i = \frac{1}{\lambda n}\langle\phi(\mathbf{x}_i), \mathbf{u}\rangle$. Ominaisvektorit määrittelevä ehto $\mathbf{Su} = \lambda\mathbf{u}$ voidaan näin kirjoittaa ekvivalentisti

$$\langle\phi(\mathbf{x}_k), \mathbf{Su}\rangle = \lambda\langle\phi(\mathbf{x}_k), \mathbf{u}\rangle \text{ kaikilla } k = 1, 2, \ldots, n. \tag{4}$$

Tämä ekvivalenssi on voimassa kun tehdään alkuperäisessä artikkelissakin [5] implisiittisesti käytetty lisäoletus, että avaruus F on äärellisulotteinen ja että kuvavektorit $\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)$ virittävät sen. Tällöin voidaan karsia kuvien joukko avaruuden kannaksi ja tutkia ehtoa (4) käyttäen vektorin $\mathbf{u}$ kantaesitystä. Näin jokaisella i määrätään yksi vektorin komponentti.

Kirjoitetaan nyt ehto (4) käyttäen kuvauksen $\mathbf{S}$ määritelmää ja esitystä (3):

$$\begin{aligned}
& \langle \phi(\mathbf{x}_k), \mathbf{S}\mathbf{u} \rangle = \lambda \langle \phi(\mathbf{x}_k), \mathbf{u} \rangle \\
\iff \quad & \left\langle \phi(\mathbf{x}_k), \frac{1}{n} \sum_{j=1}^{n} \langle \phi(\mathbf{x}_j), \mathbf{u} \rangle \phi(\mathbf{x}_j) \right\rangle = \lambda \langle \phi(\mathbf{x}_k), \mathbf{u} \rangle \\
\iff \quad & \left\langle \phi(\mathbf{x}_k), \frac{1}{n} \sum_{j=1}^{n} \langle \phi(\mathbf{x}_j), \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) \rangle \phi(\mathbf{x}_j) \right\rangle = \lambda \langle \phi(\mathbf{x}_k), \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) \rangle \\
\iff \quad & \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n} \langle \phi(\mathbf{x}_k), \phi(\mathbf{x}_j) \rangle \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle = n\lambda \sum_{i=1}^{n} \alpha_i \langle \phi(\mathbf{x}_k), \phi(\mathbf{x}_i) \rangle \qquad (5) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{kaikilla } k = 1, 2, \ldots, n.
\end{aligned}$$

Olkoon $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ kerrointen vektori ja $\mathbf{K}$ kuvien sisätulomatriisi, jonka alkiot saadaan siis säännöllä $k_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. Ehto (5) voidaan nyt kirjoittaa

$$\mathbf{K}^2 \boldsymbol{\alpha} = n\lambda \mathbf{K} \boldsymbol{\alpha}. \qquad (6)$$

Osoitetaan seuraavaksi, että ehdon (6) toteuttavat parit $(\lambda, \boldsymbol{\alpha})$ saadaan ominaisarvoongelmasta

$$\mathbf{K}\boldsymbol{\alpha} = n\lambda \boldsymbol{\alpha}. \qquad (7)$$

Tehdään tämä hieman alkuperäisestä artikkelista poikkeavalla tavalla.

Selvästi kaikki yhtälön (7) ratkaisut toteuttavat myös yhtälön (6) (kerrotaan yhtälön molemmat puolet matriisilla $\mathbf{K}$), mutta toisin päin näin ei yleisesti ole. Olkoon $(\lambda, \boldsymbol{\alpha})$ yhtälön (6) ratkaisu. Oletetaan, että $\lambda \neq 0$, mikä on perusteltua, sillä etsitään pääkomponentteja, joiden varianssi ei ole nolla. Osoitetaan, että tällöin $\boldsymbol{\alpha}$ on muotoa

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \mathbf{h},$$

jossa $(\lambda, \boldsymbol{\alpha}_0)$ on yhtälön (7) ratkaisu ja $\mathbf{K}\mathbf{h} = \mathbf{0}$. Osoitetaan myös, että kaikki tätä muotoa olevat vektorit toteuttavat yhtälön (6).

Oletetaan ensin, että $(\lambda, \boldsymbol{\alpha})$ on yhtälön (6) ratkaisu. Silloin $\boldsymbol{\alpha}$ voidaan kirjoittaa muodossa

$$\boldsymbol{\alpha} = \frac{1}{n\lambda} \mathbf{K}\boldsymbol{\alpha} + (\boldsymbol{\alpha} - \frac{1}{n\lambda} \mathbf{K}\boldsymbol{\alpha}).$$

Laskulla $\mathbf{K}\frac{1}{n\lambda}\mathbf{K}\boldsymbol{\alpha} = \frac{1}{n\lambda}\mathbf{K}^2\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha} = n\lambda\frac{1}{n\lambda}\mathbf{K}\boldsymbol{\alpha}$ nähdään, että $(\lambda, \frac{1}{n\lambda}\mathbf{K}\boldsymbol{\alpha})$ toteuttaa yhtälön (7), ja laskulla $\mathbf{K}(\boldsymbol{\alpha} - \frac{1}{n\lambda}\mathbf{K}\boldsymbol{\alpha}) = \mathbf{K}\boldsymbol{\alpha} - \frac{1}{n\lambda}\mathbf{K}^2\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\alpha} = 0$, että toinenkin ehto täyttyy.

Oletetaan nyt, että $\boldsymbol{\alpha}$ on muotoa $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \mathbf{h}$, jossa $(\lambda, \boldsymbol{\alpha}_0)$ on yhtälön (7) ratkaisu ja $\mathbf{K}\mathbf{h} = \mathbf{0}$. Silloin

$$\begin{aligned}
& \mathbf{K}^2\boldsymbol{\alpha} = n\lambda\mathbf{K}\boldsymbol{\alpha} \\
\iff \quad & \mathbf{K}^2\boldsymbol{\alpha}_0 + \mathbf{K}^2\mathbf{h} = n\lambda\mathbf{K}\boldsymbol{\alpha}_0 + n\lambda\mathbf{K}\mathbf{h} \\
\iff \quad & n\lambda\mathbf{K}\boldsymbol{\alpha}_0 + \mathbf{K}\mathbf{0} = (n\lambda)^2\boldsymbol{\alpha}_0 \\
\iff \quad & (n\lambda)^2\boldsymbol{\alpha}_0 = (n\lambda)^2\boldsymbol{\alpha}_0,
\end{aligned}$$

eli $(\lambda, \boldsymbol{\alpha})$ on yhtälön (6) ratkaisu.

Kun havainnoille lasketaan pääkomponenttipistemäärä, tapahtuu se laskulla $\mathbf{K}\boldsymbol{\alpha}$ (tämä perustellaan seuraavaksi), joka antaa selvästi saman tuloksen kuin vastaava $\mathbf{K}\boldsymbol{\alpha}_0$. Näin siis voidaan todeta, että jokaista yhtälön (6) ratkaisua vastaa joku yhtälön (7) ratkaisu, joka tuottaa samat pääkomponenttipistemäärät. Täten vain yhtälön (7) ratkaisu riittää. Tarkemmin yhtälöjen ratkaisuja voidaan luonnehtia seuraavasti: yhtälön (6) ratkaisut $(\lambda, \boldsymbol{\alpha}) = (\lambda, \boldsymbol{\alpha}_0 + \mathbf{h})$ jakautuvat ekvivalenssiluokkiin niin, että samaan luokkaan kuuluvissa ratkaisuissa vakio λ ja yhtälön (7) ratkaisuvektori $\boldsymbol{\alpha}_0$ ovat yhteisiä (eli vektorien $\boldsymbol{\alpha}$ erotus on $\mathbf{h}$, jolle pätee $\mathbf{K}\mathbf{h} = \mathbf{0}$). Yhtälön (7) ratkaisut muodostavat erään edustajiston tälle partitiolle ekvivalenssiluokkiin ja sen löytäminen riittää.

Merkitään matriisin $\mathbf{K}$ ominaisarvoja $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ (nämä ovat yhtälön (7) ratkaisut $n\lambda$) ja vastaavia ominaisavektoreita $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$. Normalisoidaan ne ominaisvektorit, joita vastaava ominaisarvo ei ole nolla, säännöllä $\lambda_k \boldsymbol{\alpha}_k' \boldsymbol{\alpha}_k = 1$, jolloin yhtälön (3) avulla saadaan

$$\begin{aligned}
\langle \mathbf{u}_k, \mathbf{u}_k \rangle &= \left\langle \sum_{i=1}^{n} \alpha_{ki} \phi(\mathbf{x}_i), \sum_{i=1}^{n} \alpha_{ki} \phi(\mathbf{x}_i) \right\rangle \\
&= \sum_{i,j=1}^{n} \alpha_{ki} \alpha_{kj} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\
&= \sum_{i,j=1}^{n} \alpha_{ki} \alpha_{kj} \left(\mathbf{K}\right)_{ij} \\
&= \boldsymbol{\alpha}_k' \mathbf{K} \boldsymbol{\alpha}_k \\
&= \lambda_k (\boldsymbol{\alpha}_k' \boldsymbol{\alpha}_k) = 1
\end{aligned}$$

Nyt pisteen $\mathbf{x}$ projektio avaruudessa F olevalle pääkomponentille $\mathbf{u}_t$ on

$$\langle \mathbf{u}_t, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{n} \alpha_{ti} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle = \sum_{i=1}^{n} \alpha_{ti} k(\mathbf{x}_i, \mathbf{x}),$$

joka havainnon $\mathbf{x}_j$ tapauksessa saadaan muodossa $\sum_{i=1}^{n} \alpha_{ti} K_{ij}$ eli j:ntenä komponenttina vektorista $\boldsymbol{K}\boldsymbol{\alpha}_t$. Näin siis pystytään laskemaan ydinpääkomponenttipistemäärät, vaikka latausvektoreita $\mathbf{u}_t$ ei tunneta.

Alkuperäisten havaintojen keskistäminen on helppoa, mutta on vaikeaa varmistaa että niiden kuvat ovat keskistettyjä. Keskistettyjen kuvien $\phi(\mathbf{x}_i) - \frac{1}{n} \sum_{l=1}^{n} \phi(\mathbf{x}_l)$

sisätulomatriisi $\tilde{\mathbf{K}}$ voidaan kuitenkin esittää matriisin $\mathbf{K}$ avulla:

$$\begin{aligned}(\tilde{\mathbf{K}})_{ij} &= \left\langle \phi(\mathbf{x}_i) - \frac{1}{n}\sum_{l=1}^{n}\phi(\mathbf{x}_l), \phi(\mathbf{x}_j) - \frac{1}{n}\sum_{l=1}^{n}\phi(\mathbf{x}_l) \right\rangle \\ &= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle - \frac{1}{n}\sum_{l=1}^{n}\langle \phi(\mathbf{x}_l), \phi(\mathbf{x}_j)\rangle \\ &\quad - \frac{1}{n}\sum_{l=1}^{n}\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_l)\rangle + \frac{1}{n^2}\sum_{l=1}^{n}\sum_{t=1}^{n}\langle \phi(\mathbf{x}_l), \phi(\mathbf{x}_t)\rangle \\ &= \left(\mathbf{K} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{K}\mathbf{1}\mathbf{1}'\right)_{ij}.\end{aligned}$$

Saadaan siis $\tilde{\mathbf{K}} = \mathbf{HKH}$. Käytetäänkin ydinmenetelmissä aina tätä matriisia keskistämättömän ydinmatriisin $\mathbf{K}$ paikalla.

2.2.3 Mahdolliset ydinfunktiot

Ydinfunktioksi voitaisiin oikeastaan valita mikä tahansa funktio, mutta usein halutaan, että sitä todella vastaa jonkin avaruuden sisätulo. Esitellään seuraavaksi lyhyesti ehto tälle artikkelin [10] luvun 2.2. mukaan.

Olkoon X epätyhjä joukko ja $k : X \times X \to \mathbb{R}$ funktio. Määritellään pisteille $x_1, x_2, \ldots, x_n \in X$ funktion k määräämä ydinmatriisi (Gramin matriisi) $\mathbf{K} \in \mathbb{R}^{n\times n}$ tavallisesti säännöllä $(\mathbf{K})_{ij} = k(x_i, x_j)$. Jos matriisi $\mathbf{K}$ on positiivisemidefiniitti kaikilla luvun $n \in \mathbb{N}$ ja pisteiden $x_1, x_2, \ldots, x_n \in X$ valinnoilla, funktiota k kutsutaan *positiivisemidefiniitiksi ytimeksi*. Jos funktio k on määritelty jonkin avaruuden sisätulon avulla muodossa $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$, saadaan

$$\mathbf{a}'\mathbf{K}\mathbf{a} = \sum_{i,j=1}^{n} a_i a_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle = \left\langle \sum_{i=1}^{n} a_i \phi(\mathbf{x}_i), \sum_{j=i}^{n} a_j \phi(\mathbf{x}_j) \right\rangle \geq 0 \text{ kaikilla } \mathbf{a} \in \mathbb{R}^n.$$

Tästä nähdään, että kaikki sisätuloa vastaavat ytimet ovat positiivisemidefiniittejä.

Tulos pätee myös toisin päin: Mooren-Aronzjnin lauseen [11] mukaan jokaista positiivisemidefiniittiä ydintä vastaa yksikäsitteisesti niin sanottu *jäljentävän ytimen Hilbertin avaruus (reproducing kernel Hilbert space, RKHS)*. Tämä avaruus saadaan seuraavalla periaatteella: Olkoon $\mathbb{R}^X$ kuvausten $X \to \mathbb{R}$ joukko ja k positiivisemidefiniitti ydin. Määritellään kuvaus $\Phi\colon X \to \mathbb{R}^X, \Phi(x) = k_x$, jossa $k_x(y) = k(y, x)$. Muodostettavan avaruuden alkiot ovat muotoa $f = \sum_i a_i k_{x_i}$. Määritellään funktioiden f ja $g = \sum_j b_j k_{x_j}$ sisätulo kaavalla $\langle f, g\rangle = \sum_i \sum_j a_i b_j k(x_i, x_j)$. Voidaan osoittaa, että kyseessä todella on sisätulo [10, luku 2.2.1] ja lisäksi, että avaruus (kuten kaikki sisätuloavaruudet) voidaan täydellistää Hilbertin avaruudeksi [12, esim 10.7.2].

Esitellään kirjan [13] pohjalta kaksi yleistä avaruudessa $\mathbb{R}^n$ käytettyä epälineaarista ydintä: *Polynomiydin* määritellään kaavalla

$$k(\mathbf{x}_i, \mathbf{x}_j) = (c + \mathbf{x}_i'\mathbf{x}_j)^p,$$

jossa luvut $p \in \mathbb{N}$ ja $c \in \mathbb{R}$ ovat parametreja. Jos $c = 0$, polynomiydintä kutsutaan *homogeeniseksi*. Yksinkertainen esimerkki saadaan kun havainnot ovat muotoa $\mathbf{x}_i =$

$(x_{i1}, x_{i2}) \in \mathbb{R}^2$ ja valitaan $p = 2$. Tällöin ydin voidaan esittää sisätulona avaruudessa $\mathbb{R}^6$, kun käytetään kuvausta $\phi(\mathbf{x}) = (c, \sqrt{2c}x_1, \sqrt{2c}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$. Polynomiydin siis tuottaa havaintovektoreiden komponenttien asteen p polynomin. Se sisältää myös alempiasteiset termit, jos $c \neq 0$.

Toinen yleinen ydin on *(gaussinen) sädekantafunktioydin*, joka määritellään kaavalla

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2}||\mathbf{x}_i - \mathbf{x}_j||^2\right).$$

Oleellista tässä ytimessä on, että sen arvo riippuu vain pisteiden $\mathbf{x}_i$ ja $\mathbf{x}_j$ etäisyydestä. Tästä tulee myös nimi sädekantafunktio (radial basis function, RBF). Syy lausekkeen muotoon ja esimerkiksi valittuun parametrisointiin on, että ytimen halutaan muistuttavan normaalijakauman tiheysfunktiota. Eräs perusteltu valinta parametriksi σ^2 onkin havaintojen neliöityjen etäisyyksien keskiarvo $\frac{1}{n(n-1)}\sum_{i\neq j}||\mathbf{x}_i - \mathbf{x}_j||^2$. Tämä muistuttaa normaalijakaumassa esiintyvää varianssiparametria ja sen voi ajatella samalla lailla standardoivan eksponentissa olevan lausekkeen. Ajattelemalla eksponenttifunktiota äärettömänä polynomina, voidaan osoittaa, että sädekantafunktioydintä vastaavat kuvat $\phi(\mathbf{x})$ ovat ääretönulotteisia. Tämä perustelee, miksi ydin-PCA:ta käsittelevässä luvussa ja yleisesti ydinfunktioiden tapauksessa kuva-avaruutta F ei haluta olettaa äärellisulotteiseksi

2.3 Moniulotteinen skaalaus

Joskus aineisto koostuu vain etäisyystiedoista, esimerkiksi kaupunkien etäisyyksistä maapallon pinnalla tai eläinlajien havaituista eroista. Tällöin kuvailun kannalta olisi toivottavaa pystyä esittämään aineisto pisteinä jossakin, usein kaksi- tai kolmiulotteisessa, avaruudessa niin, että pisteiden väliset etäisyydet vastaisivat mahdollisimman hyvin aineiston etäisyystietoja. Moniulotteinen skaalaus on juuri tähän kehitetty menetelmä. Esitetään sen johto kirjan [6] luvun 14.2 mukaan.

Symmetristä matriisia $\mathbf{D} \in \mathbb{R}^{n\times n}$ sanotaan *etäisyysmatriisiksi*, jos sen alkiot d_{ij} toteuttavat ehdot

$$d_{ii} = 0 \text{ ja } d_{ij} \geq 0 \text{ kaikilla } i \neq j.$$

Etäisyysmatriisi on *euklidinen*, jos jollakin p on olemassa joukko pisteitä $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, jossa kaikilla i, j pisteiden $\mathbf{x}_i$ ja $\mathbf{x}_j$ euklidinen etäisyys on d_{ij}.

Esitellään sitten kaksi tarpeellista matriisia: Olkoon $\mathbf{D} = (d_{ij})$ etäisyysmatriisi. Määritellään matriisi $\mathbf{A} = (a_{ij})$ kaavalla

$$a_{ij} = -\frac{1}{2}d_{ij}^2 \tag{8}$$

ja matriisi $\mathbf{B} = (b_{ij})$ kaavalla

$$\mathbf{B} = \mathbf{HAH}, \tag{9}$$

jossa $\mathbf{H}$ on keskistämismatriisi.

Näytetään seuraavaksi, että etäisyysmatriisi $\mathbf{D}$ on euklidinen jos ja vain jos siitä johdettu matriisi $\mathbf{B}$ on positiivisemidefiniitti. Samalla esitetään tapa löytää moniulotteisessa skaalauksessa halutut pisteet. Näiden pisteiden euklidinen etäisyysmatriisi on tarkalleen $\mathbf{D}$ vain jos $\mathbf{D}$ todella on euklidinen, mutta menetelmää voidaan käyttää muulloinkin.

Aloitetaan näyttämällä, että jos $\mathbf{D}$ on pisteiden $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ euklidinen etäisyysmatriisi, niin $b_{ij} = (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$. Tällöin matriisi $\mathbf{B}$ on siis positiivisemidefiniitti:

Matriisin $\mathbf{A}$ alkiot ovat tässä tilanteessa

$$a_{ij} = -\frac{1}{2}d_{ij}^2 = -\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = -\frac{1}{2}(\mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j).$$

Esitetään nyt matriisi $\mathbf{B}$ muodossa

$$\mathbf{B} = \mathbf{HAH} = \mathbf{A} - \frac{1}{n}\mathbf{A11}' - \frac{1}{n}\mathbf{11}'\mathbf{A} + \frac{1}{n^2}\mathbf{11}'\mathbf{A11}'.$$

Summan matriisit ovat alkioittain

$$\frac{1}{n}\mathbf{A11}' = \begin{pmatrix} \bar{a}_{1\cdot} & \ldots & \bar{a}_{1\cdot} \\ \vdots & \ddots & \vdots \\ \bar{a}_{n\cdot} & \ldots & \bar{a}_{n\cdot} \end{pmatrix}, \quad \frac{1}{n}\mathbf{11}'\mathbf{A} = \begin{pmatrix} \bar{a}_{\cdot 1} & \ldots & \bar{a}_{\cdot n} \\ \vdots & \ddots & \vdots \\ \bar{a}_{\cdot 1} & \ldots & \bar{a}_{\cdot n} \end{pmatrix}, \quad \frac{1}{n^2}\mathbf{11}'\mathbf{A11}' = \begin{pmatrix} \bar{a}_{\cdot\cdot} & \ldots & \bar{a}_{\cdot\cdot} \\ \vdots & \ddots & \vdots \\ \bar{a}_{\cdot\cdot} & \ldots & \bar{a}_{\cdot\cdot} \end{pmatrix},$$

jossa

$$\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}.$$

Siis

$$b_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}. \tag{10}$$

Voidaan laskea seuraavat tulokset:

$$\begin{aligned}
\bar{a}_{i\cdot} &= \frac{1}{n}\sum_{j=1}^{n} a_{ij} = -\frac{1}{2}\frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j) = -\frac{1}{2}(\mathbf{x}_i'\mathbf{x}_i + \overline{\mathbf{x}'\mathbf{x}} - 2\mathbf{x}_i'\bar{\mathbf{x}}) \\
\bar{a}_{\cdot j} &= \frac{1}{n}\sum_{i=1}^{n} a_{ij} = -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j) = -\frac{1}{2}(\overline{\mathbf{x}'\mathbf{x}} + \mathbf{x}_j'\mathbf{x}_j - 2\bar{\mathbf{x}}'\mathbf{x}_j) \\
\bar{a}_{\cdot\cdot} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} = -\frac{1}{2}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j) \\
&= -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i'\mathbf{x}_i + \overline{\mathbf{x}'\mathbf{x}} - 2\mathbf{x}_i'\bar{\mathbf{x}}) = -\frac{1}{2}(2\overline{\mathbf{x}'\mathbf{x}} - 2\bar{\mathbf{x}}'\bar{\mathbf{x}}).
\end{aligned}$$

Näiden avulla saadaan $b_{ij} = \mathbf{x}_i'\mathbf{x}_j - \mathbf{x}_i'\bar{\mathbf{x}} - \bar{\mathbf{x}}'\mathbf{x}_j + \bar{\mathbf{x}}'\bar{\mathbf{x}}$ eli

$$b_{ij} = (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}). \tag{11}$$

Kun pisteet asetetaan matriisin $\mathbf{X}$ riveiksi, saadaan $\mathbf{B} = (\mathbf{HX})(\mathbf{HX})'$. Matriisi $\mathbf{B}$ on siis symmetrinen. Lisäksi, koska kaikilla $\mathbf{v} \in \mathbb{R}^n$ pätee $\mathbf{v}'(\mathbf{HX})(\mathbf{HX})'\mathbf{v} = ||(\mathbf{HX})'\mathbf{v}||^2 \geq 0$, se on positiivisemidefiniitti. Näin siis saatiin osoitettua, että euklidista etäisyysmatriisia $\mathbf{D}$ vastaava matriisi $\mathbf{B}$ on aina positiivisemidefiniitti.

Todistetaan seuraavaksi päinvastainen implikaatio esittämällä moniulotteisen skaalauksen taustalla varsinaisesti oleva ajatus siitä, että jos $\mathbf{B}$ on positiivisemidefiniitti ja astetta p, voidaan seuraavalla tavalla valita pisteet $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$,

joiden euklidinen etäisyysmatriisi on $\mathbf{D}$: Olkoon $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ matriisin $\mathbf{B}$ positiiviset ominaisarvot (oletetaan yksinkertaisuuden vuoksi niiden erisuuruus) ja $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)} \in \mathbb{R}^n$ vastaavat ominaisvektorit skaalattuna säännöllä $\mathbf{x}'_{(i)}\mathbf{x}_{(i)} = \lambda_i,\ i = 1, 2, \ldots, p$. Kun nämä vektorit asetetaan matriisin $\mathbf{X}$ sarakkeiksi, voidaan riveiltä lukea halutut pisteet $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Näytetään tämä seuraavaksi.

Koska $\mathbf{B}$ on positiivisemidefiniitti ja astetta p, sillä on p positiivisista ominaisarvoa ja muut ovat nollia. Olkoot $\mathbf{\Lambda} \in \mathbb{R}^{p\times p}$ matriisi, jonka diagonaalilla ovat sen positiiviset ominaisarvot $\lambda_1 > \lambda_2 > \cdots > \lambda_p$, ja $\mathbf{X} \in \mathbb{R}^{n\times p}$ matriisi, jonka sarakkeina ovat vastaavat ominaisvektorit $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)}$ skaalattuna säännöllä $\mathbf{x}'_{(i)}\mathbf{x}_{(i)} = \lambda_i,\ i = 1, 2, \ldots, p$. Määritellään sitten matriisi $\mathbf{\Gamma} = \mathbf{X}\mathbf{\Lambda}^{-\frac{1}{2}}$, jossa ominaisvektorit on skaalattu yksikköpituuteen. Nyt matriisin $\mathbf{B}$ ominaisarvohajotelma on

$$\mathbf{B} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}' = \mathbf{X}\mathbf{X}'. \tag{12}$$

Siis matriisin $\mathbf{B}$ alkiot $b_{ij} = \mathbf{x}'_i\mathbf{x}_j$ ovat matriisin $\mathbf{X}$ rivien sisätuloja.

Näytetään vielä, että $\mathbf{D}$ on saatujen pisteiden $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ etäisyysmatriisi. Lausekkeiden (10) ja (8) avulla saadaan

$$\begin{aligned} ||\mathbf{x}_i - \mathbf{x}_j||^2 &= \mathbf{x}'_i\mathbf{x}_i + \mathbf{x}'_j\mathbf{x}_j - 2\mathbf{x}'_i\mathbf{x}_j \\ &= b_{ii} + b_{jj} - 2b_{ij} = a_{ii} + a_{jj} - 2a_{ij} \\ &= -\frac{1}{2}(d^2_{ii} + d^2_{jj} - 2d^2_{ij}) = d^2_{ij}. \end{aligned}$$

Saatiin siis osoitettua toivottu tulos: Kun asetetaan matriisin $\mathbf{B}$ skaalatut ominaisvektorit matriisin $\mathbf{X}$ sarakkeiksi, saadaan riveiltä luettua halutut pisteet.

Kun verrataan pääkomponenttianalyysiä sisätulomatriisin ominaisarvohajotelman avulla ja moniulotteista skaalusta, erityisesti lauseketta (11), nähdään, että kyseessä on oleellisesti sama asia: Jos pistejoukon euklidisen etäisyysmatriisin avulla tehdään moniulotteinen skaalaus, päädytään tekemään juuri samoin kuin sisätulojen avulla tehtävässä pääkomponenttianalyysissä. Moniulotteisen skaalauksen antamat akselit ovat siis pääkomponentteja. Lisäksi yhtälön (11) mukaisesti pisteet ovat keskistettyjä. Moniulotteinen skaalaus siis antaa keinon tehdä pääkomponenttianalyysi kovarianssimatriisin ja sisätulojen lisäksi myös etäisyyksien avulla. Tämän voi tulkita niin, että jos tiedetään kaikkien pisteiden väliset euklidiset etäisyydet, tiedetään myös niiden keskistetyt sisätulot (toisin päin tämä on ilmeistä, koska $||\mathbf{x}_i - \mathbf{x}_j||^2 = ||(\mathbf{x}_i - \bar{\mathbf{x}}) - (\mathbf{x}_j - \bar{\mathbf{x}})||^2 = (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}}) + (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}) - 2(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$).

Moniulotteiselle skaalaukselle voidaan esittää kaksi optimaalisuustulosta [6]: Olkoon $\mathbf{X} \in \mathbb{R}^{n\times p}$ matriisi, jonka rivit vastaavat jotakin pistejoukkoa, $\mathbf{D} = (d_{ij})$ pisteiden euklidinen etäisyysmatriisi ja $\mathbf{B} = (b_{ij})$ siitä kaavalla (9) johdettu matriisi. Olkoon $(\mathbf{L}_1, \mathbf{L}_2) \in \mathbb{R}^{p\times p}$ ortogonaalimatriisi (eli jonkin kierron ja/tai peilauksen matriisi), jossa $\mathbf{L}_1 \in \mathbb{R}^{p\times k}$ $(k \leq p)$. Muodostetaan matriisi $\hat{\mathbf{X}} = \mathbf{X}\mathbf{L}_1$, jonka rivit ovat alkuperäisten pisteiden projektioita matriisin $\mathbf{L}_1$ sarakkeiden virittämään k-ulotteiseen avaruuteen. Merkitään näiden pisteiden euklidista etäisyysmatriisia $\hat{\mathbf{D}} = (\hat{d}_{ij})$ ja sisätulomatriisia $\hat{\mathbf{B}} = (\hat{b}_{ij})$. Tiedetään, että $\hat{d}^2_{ij} \leq d^2_{ij}$. Lisäksi pätee, että

1. Kaikista tällä tavalla saaduista pisteiden projektioista k-ulotteiseen avaruuteen moniulotteinen skaalaus minimoi virheen $\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^2 - \hat{d}_{ij}^2)$. [6, lause 14.4.1]

2. Kaikista tällä tavalla saaduista pisteiden projektioista k-ulotteiseen avaruuteen moniulotteinen skaalaus minimoi virheen $\sum_{i=1}^{n}\sum_{j=1}^{n}(b_{ij} - \hat{b}_{ij})^2$. Tämä pätee myös vaikka käytetty etäisyysmatriisi $\mathbf{D}$ olisi epäeuklidinen. [6, lause 14.4.2]

2.3.1 Metrinen moniulotteinen skaalaus

Esitellään seuraavaksi moniulotteisen skaalauksen yleistys ja sen yhteys ydinpääkomponenttianalyysiin artikkelin [14] mukaisesti. Edellä esitetyssä tavallisessa tai klassisessa moniulotteisessa skaalauksessa halutaan löytää pisteet, joiden euklidiset etäisyydet $\hat{d}_{ij}$ vastaavat mahdollisimman hyvin haluttuja etäisyyksiä d_{ij}. Yleisempi tehtävä on, että näiden kahden välille tahdotaan jokin tietty, halutun funktion f määräämä, yhteys $\hat{d}_{ij} \approx f(d_{ij})$. Tätä kutsutaan *metriseksi moniulotteiseksi skaalaukseksi*. Perinteinen moniulotteinen skaalaus on metrisen moniulotteisen skaalauksen erikoistapaus, jossa $f(d_{ij}) = d_{ij}$. Joskus mielenkiintoista on vain etäisyyksien järjestys ja tällaista menetelmää kutsutaan *ordinaaliseksi* tai *epämetriseksi moniulotteiseksi skaalaukseksi*.

Yksinkertainen tapa toteuttaa metristä moniulotteista skaalausta on muuntaa etäisyydet halutulla tavalla $d_{ij} \mapsto f(d_{ij})$ ja käyttää näitä muunnettuja etäisyyksiä klassisessa moniulotteisessa skaalauksessa. Toinen tapa on muodostaa virhe- tai *stressifunktio*

$$S = \frac{\sum_{ij} \omega_{ij}(\hat{d}_{ij} - f(d_{ij}))^2}{\sum_{ij} \hat{d}_{ij}},$$

jossa luvut ω_{ij} ovat jollakin lailla valittuja painoja. Stressifunktio voidaan minimoida etsimällä lausekkeen osittaisderivaatat etäisyydet $\hat{d}_{ij}$ tuottavien pisteiden suhteen ja käyttämällä gradienttipohjaisia optimointimenetelmiä. Tätä kutsutaan *pienimmän neliösumman skaalaukseksi*.

Havainnollistetaan nyt metrisen moniulotteisen skaalauksen yhteyttä ydinpääkomponenttianalyysiin. Ydinfunktiota $k(\mathbf{x}_i, \mathbf{x}_j)$ kutsutaan *stationaariseksi*, jos se riippuu vain pisteiden erotusvektorista $\mathbf{x}_i - \mathbf{x}_j$, ja *isotrooppiseksi*, jos se riippuu vain pisteiden etäisyydestä $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$. Isotrooppinen ydin on siis muotoa $k(\mathbf{x}_i, \mathbf{x}_j) = r(d_{ij})$ jollakin funktiolla r. Oletetaan, että ydin on skaalattu niin, että $r(0) = 1$. Esimerkkinä isotrooppisesta ydinfunktiosta on sädekantafunktio, jossa $r(d_{ij}) = \exp(-\frac{1}{2\sigma^2}d_{ij}^2)$.

Olkoon k isotrooppinen ydin ja ϕ sitä vastaava kuvaus. Nyt kuva-avaruudessa määritellylle etäisyydelle $\tilde{d}_{ij}$ saadaan

$$\begin{aligned}
\tilde{d}_{ij}^2 &= ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 \\
&= \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_i)\rangle + \langle\phi(\mathbf{x}_j), \phi(\mathbf{x}_j)\rangle - 2\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle \\
&= r(d_{ii}) + r(d_{jj}) - 2r(d_{ij}) \\
&= 2(1 - r(d_{ij}))
\end{aligned}$$

Kaavalla (8) määritellyn matriisin $\mathbf{A}$ alkiot ovat nyt muotoa $a_{ij} = r(d_{ij}) - 1$. Kun käytetään ydinfunktioiden arvojen matriisia $\mathbf{K}$ matriisi $\mathbf{A}$ voidaan esittää $\mathbf{A} = \mathbf{K} - \mathbf{1}\mathbf{1}'$ ja siten keskitetty matriisi $\mathbf{B} = \mathbf{HAH} = \mathbf{HKH}$.

Huomataan siis, että ydinpääkomponenttianalyysi isotrooppisella ydinfunktiolla $k(\mathbf{x}_i, \mathbf{x}_j) = r(d_{ij})$ vastaa metristä moniulotteista skaalausta, jossa etäisyyksien välinen yhteys määritellään $\hat{d}_{ij} \approx \sqrt{2(1 - r(d_{ij}))}$. Kun funktio r on vähenevä, kuten sädekantafunktion tapauksessa, on lauseke $\sqrt{2(1 - r(d_{ij}))}$ kasvava etäisyyden d_{ij} funktio, mikä on intuitiivisesti järkevä tilanne.

Jos ydin ei ole isotrooppinen, saadaan edelleen $\tilde{d}_{ij}^2 = k(\mathbf{x}_i, \mathbf{x}_i) + k(\mathbf{x}_j, \mathbf{x}_j) - 2k(\mathbf{x}_i, \mathbf{x}_j)$. Näin saadaan $(\mathbf{A})_{ij} = (\mathbf{K})_{ij} - \frac{1}{2}k(\mathbf{x}_i, \mathbf{x}_i) - \frac{1}{2}k(\mathbf{x}_j, \mathbf{x}_j)$. Käyttämällä keskistysmatriisin $\mathbf{H}$ määritelmää, huomataan, että kaksi jälkimmäistä termiä häviävät ja saadaan taas $\mathbf{B} = \mathbf{HKH}$. Matriisin $\mathbf{K}$ avulla voidaan siis tehdä moniulotteinen skaalaus ytimen kuva-avaruudessa, mutta epäisotrooppisella ytimellä etäisyyksien välillä ei ole selkeää funktionaalista yhteyttä.

2.4 Yhteenveto

Tässä luvussa osoitettiin, miten pääkomponenttianalyysi lasketaan pelkän sisätulomatriisin $\mathbf{X}\mathbf{X}'$ kautta. Korvaamalla sisätulomatriisi ydinmatriisilla $\mathbf{K}$ saadaan ydinpääkomponenttianalyysi (menetelmästä esitettiin myös alkuperäisen artikkelin mukainen johto) ja korvaamalla se etäisyysmatriisista tuotetulla matriisilla $\mathbf{B}$ saadaan moniulotteinen skaalaus. Kaikki esitellyt menetelmät muodostavat siis yhden perheen pääkomponenttianalyysin ympärille. Työn varsinainen ydin on seuraavaksi esittää tämä sama perhe, mutta riippumattomien komponenttien analyysiin käytetyn FOBI-menetelmän ympärille. Tässä käytetään hyväksi paljon luvun 2 tuloksia ja huomioita, jotka toistuvat samanlaisina.

3 Riippumattomien komponenttien analyysi

Riippumattomien komponenttien (independent component, IC) malli [15] määritellään satunnaismuuttujatasolla lausekkeella

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{s}.$$

Havainnot koostuvat p-ulotteinen satunnaisvektorin $\mathbf{x}$ saamista arvoista. Kutsutaan vektoria havaintovektoriksi, vaikka se onkin satunnaismuuttuja. *Lähdevektori* $\mathbf{s}$ on p-ulotteinen satunnaisvektori, kun taas *sekoitusmatriisi* $\mathbf{A} \in \mathbb{R}^{p\times p}$ ja odotusarvovektori $\boldsymbol{\mu} \in \mathbb{R}^p$ eivät ole satunnaisia. Kaikki kolme ovat kuitenkin tuntemattomia. Vektorista $\mathbf{s}$ oletetaan, että kaikilla i pätee $\mathbb{E}[s_i] = 0$, $\mathbb{V}\mathrm{ar}[s_i] = 1$. Tehdään myös mallin nimessäkin esiintyvä oletus, että kaikki satunnaismuuttujat s_i ovat riippumattomia, ja oletetaan lisäksi, että ne eivät ole normaalijakautuneita. Malli on erikoistapaus yleisemmästä hajontamallista, jossa satunnaisvektorin $\mathbf{s}$ voitaisiin olettaa olevan esimerkiksi multinormaalijakautunut.

Osoitetaan artikkelin [16] korollaarin 3.1. todistusta mukaillen, että näillä oletuksin matriisi $\mathbf{A}$ voidaan yksilöidä sarakkeiden järjestystä ja etumerkkiä lukuun ottamatta. Mainitaan alkuun todistuksessa käytettävä Skitovichin-Darmoisin lause (ks. esim [17]): Olkoon $s_1, s_2, \ldots, s_k$ riippumattomia satunnaismuuttujia ja $a_1, a_2, \ldots a_k$, $b_1, b_2, \ldots b_k$ nollasta eroavia reaalilukuja. Jos $\sum_{i=1}^k a_i s_i$ ja $\sum_{i=1}^k b_i s_i$ ovat riippumattomia, niin $s_1, s_2, \ldots, s_k$ ovat normaalijakautuneita.

Olkoot $\mathbf{A}^* = \mathbf{A}\mathbf{C}^{-1}$ ja $\mathbf{s}^* = \mathbf{C}\mathbf{s}$, jossa $\mathbf{C} \in \mathbb{R}^{p\times p}$ on kääntyvä ja $\mathbf{s}^*$ toteuttaa yllä annetut lähdevektorille asetetut ehdot. Osoitetaan, että matriisin $\mathbf{C}$ jokaisessa sarakkeessa (ja rivissä) on tasan yksi nollasta poikkeava alkio, joka on ± 1. Kääntyvyydestä seuraa suoraan, että jokaisessa sarakkeessa on vähintään yksi nollasta poikkeava alkio. Osoitetaan, että alkioita on enintään yksi. Olkoot $s_i^* = \sum_k c_{ik} s_k$ ja $s_j^* = \sum_k c_{jk} s_k$ $(i \neq j)$ vektorin $\mathbf{s}^*$ komponentteja. Koska ne ovat riippumattomia, myös summat

$$\sum_{k:\ c_{ik}c_{jk}\neq 0} c_{ik}s_k \quad \text{ja} \quad \sum_{k:\ c_{ik}c_{jk}\neq 0} c_{jk}s_k$$

ovat riippumattomia ja lisäksi näissä summissa kertoimet c ovat nollasta eroavia. Nyt Skitovichin-Darmoisin lauseen mukaan kaikki summissa olevat satunnaismuuttujat s_k ovat normaalijakautuneita. Koska lähdevektorista oletetaan, että sen komponentit eivät ole normaalijakautuneita, täytyy summan olla tyhjä. Siis kaikilla lukujen i, j $(i \neq j)$ ja k arvoilla pätee, että $c_{ik}c_{jk} = 0$ eli jokaisessa sarakkeessa on vain yksi nollasta poikkeava alkio. Lopuksi oletuksesta $\mathbb{V}\mathrm{ar}[\mathbf{s}^*] = 1$ seuraa, että nollasta poikkeavien alkioiden on oltava itseisarvoltaan yksi. On hyvä huomata, että oletus lähdevektorin varianssista tehdään lopulta vain sekoitusmatriisin yksikäsitteisyyden takia. Yllä olevasta nähdään, että ilman sitä matriisin $\mathbf{C}$ sarakkeita voitasiin kertoa vakiolla ilman, että tämä vaikuttaisi vektoriin $\mathbf{x}$.

Tarkoituksena riippumattomien komponenttien analyysissä on muuntaa havainnot $\mathbf{x}$ takaisin lähteen $\mathbf{s}$ arvoiksi. Pääkomponenttianalyysiin verrattuna riippumattomien komponenttien analyysissä etsittävät komponentit eivät siis ole vain korreloimattomia, vaan myös aidosti satunnaismuuttujina riippumattomia.

Näytetään, että haettu muunnos löytyy estimoimalla havaintovektorin $\mathbf{x}$ kovarianssimatriisi, standardoimalla sen avulla vektori $\mathbf{x}$ ja etsimällä sitten sopiva orto-

gonaalimatriisi [18, lause 1]: Koska $\mathbb{E}[\mathbf{s}] = \mathbf{0}$ ja $\mathbb{C}\mathrm{ov}[\mathbf{s}] = \mathbf{I}$, niin odotusarvo $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ ja kovanrianssimatriisi $\boldsymbol{\Sigma} = \mathbb{C}\mathrm{ov}[\mathbf{x}] = \mathbf{A}\mathbf{A}'$. Olkoon matriisin $\mathbf{A}$ singulaariarvohajotelma $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$. Tällöin $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{D}^2\mathbf{U}'$. Määritellään $\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$. Tällöin $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I}$. Määritellään standardoitu havaintovektori $\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})$. Tällöin $\mathbb{E}[\mathbf{x}_{st}] = \mathbf{0}$ ja $\mathbb{C}\mathrm{ov}[\mathbf{x}_{st}] = \mathbf{I}$.

Laskelmilla

$$\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{s} = \mathbf{U}\mathbf{V}'\mathbf{s}$$

ja $\mathbf{U}\mathbf{V}'(\mathbf{U}\mathbf{V}')' = \mathbf{I}$ huomataan, että lähdevektori $\mathbf{s}$ saadaan standardoidusta havaintovektorista $\mathbf{x}_{st}$ ortogonaalisella muunnoksella $\mathbf{O}' = (\mathbf{U}\mathbf{V}')'$, jota toki ei voida näin suoraan laskea. Riippumattomien komponenttien mallin ratkaisu on siis etsiä sellainen ortogonaalinen matriisi $\mathbf{O}'$, joka muuntaa standardoidun havaintovektorin komponentit riippumattomiksi.

Pääkomponenttianalyysi ja riippumattomien komponenttien analyysi ovat perusasetelmiltaan hyvin samanlaisia menetelmiä, koska molemmissa tarkoitus on jakaa aineiston vaihtelu jollakin lailla erillisiin osiin. PCA:ssa nämä ovat korreloimattomia ja ICA:ssa riippumattomia. Koska riippumattomuus on korreloimattomuutta vahvempi oletus, on jossakin mielessä arvattavaa, että ICA voidaan ajatella jalostettuna pääkomponenttianalyysinä: Satunnaisvektorin $\mathbf{x}$ pääkomponenttipistemäärä on $\mathbf{U}'\mathbf{x}$ ja riippumattomien komponenttien pistemäärä $\mathbf{V}\mathbf{U}'\mathbf{x}_{st} = \mathbf{V}\mathbf{U}'\mathbf{U}\mathbf{D}^{-1}\mathbf{U}'\mathbf{x} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}'\mathbf{x}$. Riippumattomien komponenttien pistemäärät saadaan siis pääkomponenttipistemääristä standardoivalla skaalauksella $\mathbf{D}^{-1}$ ja rotaatiolla $\mathbf{V}$.

Riippumattomien komponenttien analyysiin on kehitetty erilaisia tehokkaita menetelmiä, kuten ominaismatriisien yhteisdiagonalisointi (joint approximate diagonalization of eigenmatrices, JADE) [19], fastICA [20] ja symmetrisoitujen hajontamatriisien samanaikainen diagonalisointi (simultaneous diagonalization of symmetrized scatter matrices) [21]. Tässä työssä kuitenkin käsitellään yksinkertaisempaa menetelmää, jota kutsutaan lyhenteellä FOBI, ja johdetaan siitä ydinversio.

3.1 FOBI

Neljännen asteen sokkotunnistus (fourth-order blind identification, FOBI) on eräs ratkaisumenetelmä riippumattomien komponenttien analyysiin ja se pohjautuu nimensä mukaisesti havaintovektorin neljännen momentin käyttöön. Menetelmä esiteltiin alunperin artikkelissa [2].

Skalaarisatunnaismuuttujan x n:s keskusmomentti m_n määritellään odotusarvona $m_n = \mathbb{E}[(x - \mathbb{E}[x])^n]$. Selvästi m_1 on aina nolla ja $m_2 = \mathbb{V}\mathrm{ar}[x]$. Satunnaismuuttujan x *kurtoosi (huipukkuus,* engl. *kurtosis)* määritellään

$$\beta[x] = \frac{m_4}{m_2^2}.$$

Normaalijakautuneelle satunnaismuuttujalle x pätee $\beta[x] = 3$ ja siksi usein kurtoosiksi kutsutaankin lukua $\kappa[x] = \beta[x] - 3$, joka mahdollistaa helpon vertailun normaalijakaumaan. Jos $\mathbb{E}[x] = 0$ ja $\mathbb{V}\mathrm{ar}[x] = 1$, kurtoosin arvo $\beta[x] = \mathbb{E}[x^4]$. Tämän pohjalta voidaan suoraan määritellä ehdot $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ ja $\mathbb{C}\mathrm{ov}[\mathbf{x}] = \mathbf{I}$ toteuttavalle

p-ulotteiselle satunnaisvektorille $\mathbf{x}$ kovarianssimatriisin kaltainen kurtoosimatriisi

$$\mathbf{B}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}'\mathbf{x}\mathbf{x}'] \in \mathbb{R}^{p\times p}.$$

Tämän matriisin kohdassa (i,j) on $\mathbb{E}[x_i x_j \sum_k x_k^2]$.

Tutkitaan nyt p-ulotteista satunnaisvektoria $\mathbf{s}$, jonka komponentit ovat lisäksi riippumattomia. Jos $i \neq j$, niin $\mathbb{E}[s_i s_j \sum_k s_k^2] = 0$. Syynä tähän on, että kaikilla k, joko $i \neq k$ tai $j \neq k$ ja siten (oletetaan $i \neq k$) $\mathbb{E}[s_i] = 0$ antaa $\mathbb{E}[s_i s_j s_k^2] = \mathbb{E}[s_i]\mathbb{E}[s_j s_k^2] = 0$. Tapauksessa $i = j$ pätee

$$\mathbb{E}[s_i^2 \sum_k s_k^2] = \mathbb{E}[s_i^4] + \sum_{k\neq i} \mathbb{E}[s_i^2 s_k^2] = \beta[s_i] + \sum_{k\neq i} \mathbb{V}\mathrm{ar}[s_i]\,\mathbb{V}\mathrm{ar}[s_k] = \beta[s_i] + p - 1.$$

Siis tälle satunnaisvektorille $\mathbf{s}$ kurtoosimatriisi

$$\mathbf{B}[\mathbf{s}] = \mathrm{diag}(\beta[s_1] + p - 1, \beta[s_2] + p - 1, \ldots, \beta[s_p] + p - 1).$$

Palataan nyt riippumattomien komponenttien ongelmaan. Olkoon $\mathbf{O} \in \mathbb{R}^{p\times p}$ se ortogonaalimatriisi, jolla $\mathbf{x}_{st} = \mathbf{O}\mathbf{s}$. Tällöin

$$\mathbf{B}[\mathbf{x}] = \mathbb{E}[\mathbf{x}_{st}\mathbf{x}_{st}'\mathbf{x}_{st}\mathbf{x}_{st}'] = \mathbb{E}[\mathbf{O}\mathbf{s}\mathbf{s}'\mathbf{O}'\mathbf{O}\mathbf{s}\mathbf{s}'\mathbf{O}'] = \mathbf{O}\mathbb{E}[\mathbf{s}\mathbf{s}'\mathbf{s}\mathbf{s}']\mathbf{O}' = \mathbf{O}\mathbf{B}[\mathbf{s}]\mathbf{O}'.$$

Koska $\mathbf{B}[\mathbf{x}]$ on symmetrinen ja $\mathbf{B}[\mathbf{s}]$ diagonaalinen, ICAn ratkaisemiseksi tarvittu matriisi $\mathbf{O}$ saadaan siis standardoidun havaintovektorin $\mathbf{x}_{st}$ kurtoosimatriisin ominaisarvohajotelmasta. Kurtoosimatriisin olemassaoloon tarvitaan toki oletus, että kaikilla i pätee $\mathbb{E}[s_i^4] < \infty$, josta seuraa myös muiden odotusarvojen äärellisyys. Lisäksi oletetaan, että lähdevektorin komponenttien kurtoosit $\beta[s_i]$ ovat erisuuria. Tämä takaa sen, että kurtoosimatriisin ortogonaalinen ominaisarvomatriisi on yksikäsitteinen $\mathbf{O}$. Jos joillekin i, j pätee $\beta[s_i] = \beta[s_j]$, niin vastaavia komponentteja s_i, s_j ei voida erotella toisistaan, mutta muut komponentit voidaan.

3.1.1 FOBIn laskeminen otoksesta

Neljännen asteen erottelu määritellään aina satunnaismuuttujatasolla, sillä siinä oletetaan lähdevektorin komponenttien riippumattomuus, jolle ei ole olemassa otosversiota. Esitellään nyt tapa mallin laskemiseen otosmatriisista $\mathbf{X} \in \mathbb{R}^{n\times p}$.

Estimoidaan ensin havaintojen kovarianssimatriisi $\mathbf{S}$ ja muodostetaan sen ominaisarvohajotelman $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$ avulla matriisi $\mathbf{S}^{-\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'$. Määritellään nyt standardoitu havaintomatriisi $\mathbf{Y} = \mathbf{X}_{st} = \mathbf{H}\mathbf{X}\mathbf{S}^{-\frac{1}{2}}$, jonka riveinä ovat standardoidut havainnot $\mathbf{y}_i'$

Lasketaan tämän avulla otoskurtoosimatriisi

$$\begin{aligned}
\mathbf{B} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i'\mathbf{y}_i\mathbf{y}_i' \\
&= \frac{1}{n}\left(\mathbf{y}_1|\ldots|\mathbf{y}_p\right)\begin{pmatrix} \mathbf{y}_1'\mathbf{y}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{y}_2'\mathbf{y}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{y}_p'\mathbf{y}_p \end{pmatrix}\begin{pmatrix} \mathbf{y}_1' \\ \hline \vdots \\ \hline \mathbf{y}_p' \end{pmatrix} \\
&= \frac{1}{n}\mathbf{Y}'\mathrm{diag}(\mathbf{Y}\mathbf{Y}')\mathbf{Y}
\end{aligned}$$

Tästä matriisista ratkaistaan ominaisarvohajotelma $\mathbf{B} = \mathbf{W}\mathbf{\Pi}\mathbf{W}'$. Siitä saadaan matriisi $\mathbf{Z} = \mathbf{Y}\mathbf{W}$, jonka riveinä ovat lähdevektorin $\mathbf{s}$ estimoidut arvot.

3.2 Ydin-FOBI ja MDS-FOBI

Pääkomponenttianalyysi perustuu kovarianssimatriisin ominaisarvohajotelmaan, kun taas FOBI perustuu kurtoosimatriisin ominaisarvohajotelmaan. Koska pääkomponenttianalyysi voidaan esittää sisätulojen avulla, herää kysymys voidaanko näin tehdä myös FOBIn kohdalla. Esitetään tässä luvussa singulaariarvohajotelmaan perustuva johto FOBIn laskemiselle sisätulojen avulla.

Olkoon keskistetyn havaintomatriisin $\mathbf{X} \in \mathbb{R}^{n\times p}$ (oletetaan $n \geq p$) singulaariarvohajotelma $\mathbf{X} = \mathbf{V}\mathbf{\Delta}\mathbf{U}'$. Silloin sen kovarianssimatriisi voidaan esittää muodossa $\mathbf{S} = \frac{1}{n-1}\mathbf{U}\mathbf{\Delta}'\mathbf{\Delta}\mathbf{U}'$ ja sen käänteinen neliöjuuri $\mathbf{S}^{-\frac{1}{2}} = \sqrt{n-1}\mathbf{U}(\mathbf{\Delta}'\mathbf{\Delta})^{-\frac{1}{2}}\mathbf{U}'$. Tästä saadaan standardoidulle havaintomatriisille esitys

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\mathbf{S}^{-\frac{1}{2}} \\ &= \mathbf{V}\mathbf{\Delta}\mathbf{U}'\sqrt{n-1}\mathbf{U}(\mathbf{\Delta}'\mathbf{\Delta})^{-\frac{1}{2}}\mathbf{U}' \\ &= \sqrt{n-1}\mathbf{V}\mathbf{\Delta}(\mathbf{\Delta}'\mathbf{\Delta})^{-\frac{1}{2}}\mathbf{U}' \\ &= \sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{U}',\end{aligned}$$

jossa $\mathbf{I}_{n\times p} = \begin{pmatrix}\mathbf{I}_p \\ \mathbf{0}_{(n-p)\times p}\end{pmatrix}$.

Tällöin kurtoosimatriisille saadaan esitys

$$\begin{aligned}\mathbf{B} &= \frac{1}{n}\mathbf{Y}'\mathrm{diag}(\mathbf{Y}\mathbf{Y}')\mathbf{Y} \\ &= \frac{1}{n}\sqrt{n-1}\mathbf{U}\mathbf{I}'_{n\times p}\mathbf{V}'\mathrm{diag}(\sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{U}'\sqrt{n-1}\mathbf{U}\mathbf{I}'_{n\times p}\mathbf{V}')\sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{U}' \\ &= \frac{(n-1)^2}{n}\mathbf{U}\mathbf{I}'_{n\times p}\mathbf{V}'\mathrm{diag}(\mathbf{V}\mathbf{I}_{n\times p}\mathbf{I}'_{n\times p}\mathbf{V}')\mathbf{V}\mathbf{I}_{n\times p}\mathbf{U}'.\end{aligned}$$

Muodostetaan nyt symmetriselle matriisille $\mathbf{I}'_{n\times p}\mathbf{V}'\mathrm{diag}(\mathbf{V}\mathbf{I}_{n\times p}\mathbf{I}'_{n\times p}\mathbf{V}')\mathbf{V}\mathbf{I}_{n\times p}$ ominaisarvohajotelma $\mathbf{E}\mathbf{D}\mathbf{E}'$, jossa matriisi $\mathbf{E}$ on ortogonaalinen. Nyt kun asetetaan $\mathbf{W} = \mathbf{U}\mathbf{E}$ ja $\mathbf{\Pi} = \frac{(n-1)^2}{n}\mathbf{D}$, huomataan, että tämä vastaa kurtoosimatriisin ominaisarvohajotelmaa $\mathbf{B} = \mathbf{W}\mathbf{\Pi}\mathbf{W}'$. Lähdevektorit saadaan nyt kaavalla

$$\mathbf{Z} = \mathbf{Y}\mathbf{W} = \sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{U}'\mathbf{U}\mathbf{E} = \sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{E}.$$

Koska havaintojen sisätulomatriisille on esitys $\mathbf{X}\mathbf{X}' = \mathbf{V}\mathbf{\Delta}\mathbf{U}'\mathbf{U}\mathbf{\Delta}'\mathbf{V}' = \mathbf{V}\mathbf{\Delta}\mathbf{\Delta}'\mathbf{V}'$, voidaan pelkästä sisätulomatriisista selvittää $\mathbf{V}$, sen perusteella $\mathbf{E}$ ja sitten näitä käyttäen lähdevektorien matriisi $\mathbf{Z}$.

Samalla lailla kuin sisätulojen avulla lasketussa pääkomponenttianalyysissä, menetelmä ei tässäkään anna latausmatriisia $\mathbf{W}$. Kuitenkin myös FOBIn tapauksessa alkuperäisten muuttujien ja tuotettujen riippumattomien komponenttien välistä yhteyttä voidaan tutkia näiden välisen korrelaatiomatriisin avulla. Tätä havainnollistetaan luvun 4.2 esimerkissä.

Samoin kuin pääkomponenttianalyysin tapauksessa, myös nyt voidaan sisätulomatriisi korvata ydinfunktion k arvoista koostuvalla matriisilla $\mathbf{K}$. Tämän matriisin ominaisarvohajotelmasta voidaan sitten laskea halutut matriisit $\mathbf{V}$, $\mathbf{E}$ ja siten $\mathbf{Z}$. Näin saadaan *ydin-FOBI*.

Pääkomponenttianalyysin ja moniulotteisen skaalauksen välinen yhteys taas saatiin kun huomattiin, että suoraan lasketun sisätulomatriisin sijasta voidaan käyttää pisteiden etäisyysmatriisista $\mathbf{D} = (d_{ij})$ johdettua matriisia $\mathbf{B} = \mathbf{HAH}$, jossa $\mathbf{A} = (-\frac{1}{2}d_{ij}^2)$ (Matriisia $\mathbf{B}$ ei tule sekoittaa kurtoosimatriisiin). Samalla lailla voidaan toimia nytkin ja näin saadaan keino laskea FOBI etäisyysmatriisin avulla. Kutsutaan sitä lyhenteellä *MDS-FOBI*.

Jotta menetelmän antamat lähdevektorit vastaisivat tulkinnallisesti riippumattomien komponenttien analyysiä, tulee niiden ainakin olla korreloimattomia ja niiden kurtoosimatriisin olla diagonaalinen. Tämän toteamiseen vaaditaan, että saatujen lähdevektorien matriisi $\mathbf{Z}$ on keskistetty: Koska $\mathbf{H1} = \mathbf{0}$, saadaan $\mathbf{HKH1} = \mathbf{0}$. Siten vektori $\mathbf{1}$ kuuluu symmetrisen matriisin $\mathbf{HKH}$ ytimeen (nollavektoriksi kuvautuvat vektorit) ja on kohtisuora matriisin kuva-avaruuden kanssa. Koska kaikki matriisin $\mathbf{HKH}$ ominaisvektorit (joita vastaava ominaisarvo on positiivinen) eli erityisesti matriisin $\mathbf{V}$ sarakkeet kuuluvat kuva-avaruuteen, pätee $\mathbf{1}'\mathbf{Z} = \sqrt{n-1}\mathbf{1}'\mathbf{V}\mathbf{I}_{n\times p}\mathbf{E} = \mathbf{0}$. Täten matriisi $\mathbf{Z}$ on keskistetty. Näin voidaan laskea kovarianssimatriisi

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\mathbf{Z}] &= \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} \\
&= \frac{1}{n-1}\sqrt{n-1}\mathbf{E}'\mathbf{I}'_{n\times p}\mathbf{V}'\sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{E} \\
&= \mathbf{E}'\mathbf{I}'_{n\times p}\mathbf{I}_{n\times p}\mathbf{E} \\
&= \mathbf{E}'\mathbf{I}_p\mathbf{E} \\
&= \mathbf{I}_p
\end{aligned}$$

ja kurtoosimatriisi

$$\begin{aligned}
\mathbf{B} &= \frac{1}{n}\mathbf{Z}'\mathrm{diag}(\mathbf{Z}\mathbf{Z}')\mathbf{Z} \\
&= \frac{1}{n}\sqrt{n-1}\mathbf{E}'\mathbf{I}'_{n\times p}\mathbf{V}'\mathrm{diag}(\sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{E}\sqrt{n-1}\mathbf{E}'\mathbf{I}'_{n\times p}\mathbf{V}')\sqrt{n-1}\mathbf{V}\mathbf{I}_{n\times p}\mathbf{E} \\
&= \frac{(n-1)^2}{n}\mathbf{E}'\mathbf{I}'_{n\times p}\mathbf{V}'\mathrm{diag}(\mathbf{V}\mathbf{I}_{n\times p}\mathbf{I}'_{n\times p}\mathbf{V}')\mathbf{V}\mathbf{I}_{n\times p}\mathbf{E} \\
&= \frac{(n-1)^2}{n}\mathbf{E}'\mathbf{E}\mathbf{D}\mathbf{E}'\mathbf{E} \\
&= \frac{(n-1)^2}{n}\mathbf{D} \\
&= \mathbf{\Pi}.
\end{aligned}$$

Tästä nähdään, että saadut komponentit ovat korreloimattomia ja niiden kurtoosimatriisi on diagonaalinen.

Esitetään ydin-FOBI vielä algoritmimuodossa. Ei käytetä tässä matriisia $\mathbf{I}_{n\times p}$, joka poimii p suurinta ominaisarvoa vastaavat vektorit. Sen sijaan valitaan suoraan matriisiin $\mathbf{V} \in \mathbb{R}^{n\times d}$ haluttu määrä d matriisin $\mathbf{K}$ ominaisvektoreita. Näin

siis jätetään valittavaksi parametriksi d, kuinka monta ominaisarvoa otetaan huomioon eli kuinka monta riippumatonta komponenttia menetelmä antaa. Jos ydinfunktiona käytettäisiin tavallista sisätuloa, olisi ydinmatriisin $\mathbf{K}$ aste, eli riippumattomien ominaisvektorien määrä, (lineaarisesti riippumattomien) muuttujien määrä p, jolloin voimassa olisi rajoite $d \leq p$. Yleisesti kuitenkin pätee vain selvä yläraja $d \leq \text{rank}(\mathbf{HKH}) \leq \text{rank}(\mathbf{H}) = n - 1$.

Ydin-FOBI

1. Valitse käytettävä ydinfunktio k, siihen liittyvät mahdolliset parametrit ja tuotettavien komponenttien määrä d.

2. Laske ydinfunktion k avulla havaintojen ydinmatriisi $\mathbf{K} \in \mathbb{R}^{n \times n}$.

3. Laske d keskistetyn matriisin $\mathbf{HKH}$ suurinta ominaisarvoa vastaavaa ominaisvektoria ja muodosta näistä matriisi $\mathbf{V} \in \mathbb{R}^{n \times d}$, jossa vektorit ovat järjestyksessä laskevien ominaisarvojen mukaisesti.

4. Muodosta matriisi
$$\mathbf{V}'\text{diag}(\mathbf{V}\mathbf{V}')\mathbf{V}.$$

5. Selvitä tämän matriisin ominaisvektoreiden matriisi $\mathbf{E} \in \mathbb{R}^{d \times d}$, jossa ominaisvektorit ovat järjestyksessä laskevien ominaisarvojen mukaan.

6. Riippumattomien komponenttien pistemäärät saadaan laskulla
$$\mathbf{Z} = \sqrt{n-1}\mathbf{V}\mathbf{E}.$$

4 Esimerkkejä

Tarkastellaan vielä esiteltyjä menetelmiä kolmen laskennallisen esimerkin kautta: Ensimmäisessä esimerkissä erotellaan ryhmiä simuloidusta aineistosta käyttäen ydin-FOBIa. Tämä perustuu tavallisen FOBIn kykyyn etsiä sellainen muuttujien lineaarikombinaatio, jonka suhteen ryhmät erottuvat hyvin. Toisessa esimerkissä sovelletaan ydin-FOBIa ja ydin-PCA:ta niin sanottujen ominaiskasvojen tuottamiseen. Perinteisesti pääkomponenttianalyysiin perustuvia ominaiskasvoja voidaan käyttää kasvojen tunnistuksessa piirteiden löytämiseen. Kolmannessa esimerkissä käytetään tavallista moniulotteista skaalausta ja MDS-FOBIa kaupunkien asettamiseen kartalle niiden välisten lentoetäisyyksien perusteella.

4.1 Ryhmien erottelu

Kurtoosimatriisin ominaisvektoreita, joihin FOBI perustuu, voidaan käyttää ryhmärakenteiden tunnistamiseen [22]. Tämä perustuu ajatukseen, että jos satunnaismuuttujan jakauma koostuu kahdesta selkeästi erottuvasta huipusta, satunnaismuuttujan kurtoosi on pieni. Intuitiivisesti voisi ajatella, että kurtoosi kuvaisi vaihtelun vaihtelua, joka on pieni, jos todennäköisyysmassa on kahdessa tiiviissä kasassa symmetrisesti odotusarvon molemmilla puolilla.

Tarkemmin asiaa voidaan lähestyä seuraavasti: Olkoon x satunnaismuuttuja, jonka oletetaan merkintöjen yksinkertaisuuden vuoksi olevan standardoitu. Tällöin Jensenin epäyhtälön perusteella ($x \mapsto x^2$ on konveksi funktio) kurtoosille pätee $\beta[x] = \mathbb{E}[x^4] = \mathbb{E}[(x^2)^2] \geq \mathbb{E}[x^2]^2 = \mathbb{V}\mathrm{ar}[x]^2 = 1$. Olkoon nyt $x \sim Bernoulli(\frac{1}{2})$. Tällöin $\mathbb{E}[x] = \frac{1}{2}$ ja $\mathbb{V}\mathrm{ar}[x] = \frac{1}{4}$. Lasketaan nyt standardoidun satunnaismuuttujan $2x-1$ kurtoosi $\beta[2x-1] = \mathbb{E}[(2x-1)^4] = 1$. Havaitaan siis, että Bernoulli-jakaumalla, joka on ääriesimerkki kaksihuippuisesta jakaumasta, on pienin mahdollinen kurtoosi.

Esitetään tässä luvussa ydin-FOBIn käyttöä ryhmärakenteiden tunnistamiseen simuloidussa, kahdesta ryhmästä koostuvassa tilanteessa, jossa on p muuttujaa ja $n = 200$ havaintoa: Luodaan ensin ryhmäindikaattori $y \sim Bernoulli(\frac{3}{5})$ ja sitten sen pohjalta muuttuja x_1, jonka jakauma on $x_1 \mid y \sim N(5y - 2.5, 1)$. Seuraavaksi luodaan haluttu määrä $p-1$ muita muuttujia $x_2, \ldots x_p \sim N(0, 1)$ ja kaikki muuttujat $x_1, \ldots, x_p$ kootaan matriisiin $\mathbf{X}_0 \in \mathbb{R}^{200 \times p}$. Lopuksi luodaan havainnot sekoittamalla muuttujia satunnaisella matriisilla $\mathbf{A} = (a_{ij})$, jossa $a_{ij} \sim N(0, 1)$ riippumattomasti, ja saadaan aineisto $\mathbf{X} = \mathbf{X}_0\mathbf{A}$.

Aineistoon käytetään työssä kuvattua ydin-FOBIa. Ydinfunktiona on gaussinen sädekantafunktio $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2\sigma^2}||\mathbf{x}_i - \mathbf{x}_j||^2)$, jossa parametriksi σ^2 asetetaan havaintojen neliöityjen etäisyyksien keskiarvo. Olkoon jonkun muuttujan x kohdalla $\bar{x}_1$ pisteiden, joilla $y = 1$, keskiarvo ja s_1 keskihajonta sekä $\bar{x}_0$ ja s_0 toisen ryhmän vastaavat tunnusluvut. Mitataan sitä, kuinka hyvin ryhmät erottuvat tämän muuttujan kohdalla, luvulla $h = \frac{|\bar{x}_1 - \bar{x}_0|}{\frac{1}{2}(s_1 + s_0)}$.

Tehdään simulaatio, jonka tarkoituksena on selvittää, mikä komponentti erottaa ryhmät parhaiten. Tuotetaan edellä kuvatulla tavalla aineistoja muuttujien määrillä $p = 5, 10, \ldots, 50$. Näistä lasketaan ydin-FOBIlla $d = 1, 2, \ldots, 10$ komponenttia. Simulaatio toistetaan 200 kertaa jokaisella yhdistelmällä ja lasketaan arvo h. Simulaatiosta havaitaan, että suurimman luvun h eli parhaan erottelukyvyn tuottaa

viimeinen komponentti. Taulukossa 1 on esitetty, kuinka suuressa osassa tilanteita viimeinen komponentti tuottaa suurimman h:n arvon. Koska ydin-FOBI voidaan ajatella tavallisena FOBIna jossakin ydinfunktion määräämässä avaruudessa ja komponentit on järjestetty ominaisarvojen eli tämän avaruuden kurtoosien perusteella, voidaan ajatella, että parhaan erottelukyvyn antava komponentti vastaa pienintä kurtoosia tässä avaruudessa. Tulos on siis yhtäpitävä aiemman teorian kanssa.

	2	3	4	5	6	7	8	9	10
5	0.95	0.94	0.94	0.92					
10	0.99	0.96	0.98	0.96	0.94	0.85	0.73	0.60	0.46
15	1.00	1.00	0.99	0.98	0.99	0.94	0.84	0.70	0.57
20	1.00	1.00	0.99	0.98	0.98	0.96	0.97	0.85	0.67
25	0.99	0.99	0.99	0.98	0.94	0.92	0.90	0.86	0.73
30	1.00	1.00	0.98	0.98	0.92	0.90	0.86	0.80	0.72
35	1.00	0.98	0.98	0.95	0.91	0.88	0.85	0.80	0.70
40	1.00	0.98	0.98	0.96	0.94	0.89	0.81	0.78	0.67
45	0.98	0.96	0.94	0.94	0.89	0.80	0.78	0.72	0.65
50	1.00	1.00	0.96	0.92	0.82	0.78	0.74	0.60	0.54

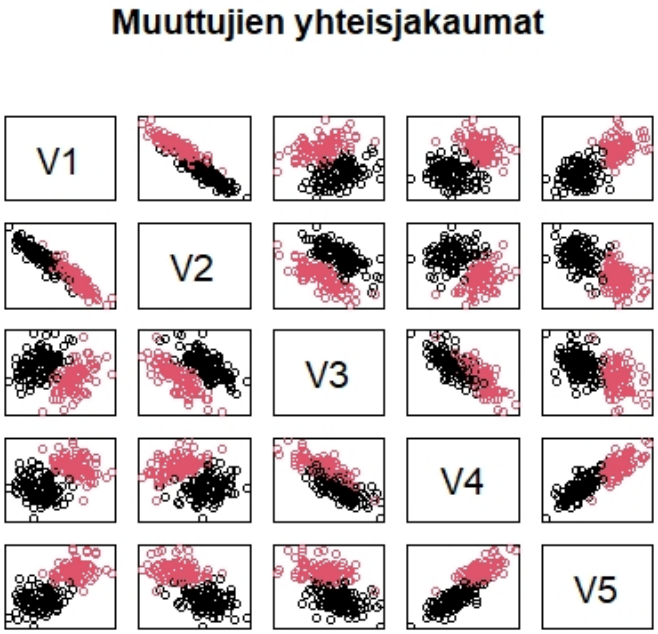
Taulukko 1: Toistojen, joissa viimeinen komponentti tuotti parhaan tuloksen, osuus. Rivillä on muuttujien määrä p ja sarakkeessa tuotettujen komponenttien määrä d.

Taulukossa 2 on esitetty jokaisessa tilanteessa viimeisen komponentin antamien arvojen h keskiarvo. Tästä havaitaan, että paras komponenttien määrä d on yksi. Tässä tilanteessa matriisi $\mathbf{V}$ on vain pystyvektori ja matriisi $\mathbf{E}$ skalaari. Tällöin saatava riippumaton komponentti $\mathbf{Z}$ on sama kuin ydin-PCA:n antama ensimmäinen pääkomponentti ja ydin-FOBI palautuu täysin ydin-PCA:han. Tässä simulaatiossa parhaiten ryhmät erotteleva suunta on siis ydintä vastaavassa kuva-avaruudessa suurimman varianssin tuottava suunta. Saatu tulos myös vahvistaa, että ydin-FOBIn antamat komponenttien joukot eivät ole sisäkkäisiä, kuten esimerkiksi (ydin)pääkomponenttianalyysissä: Kun tuotetaan kahden sijaan kolme komponenttia, näiden kolmen joukkoon ei yleensä sisälly aiempaa kahta. Jos näin olisi, niin komponenttien määrän d kasvattaminen ei voisi huonontaa erottelukykyä. Nyt kuitenkin havaitaan näin tapahtuvan.

	1	2	3	4	5	6	7	8	9	10
5	4.81 (1.8)	4.62 (1.8)	4.48 (1.7)	3.97 (1.5)	3.44 (1.3)					
10	3.86 (1.1)	3.77 (1.1)	3.66 (1.1)	3.53 (1.0)	3.32 (1.0)	2.98 (1.1)	2.55 (1.1)	2.07 (1.1)	1.68 (1.0)	1.35 (1.0)
15	4.00 (0.8)	3.93 (0.9)	3.81 (0.8)	3.65 (0.8)	3.45 (0.8)	3.23 (0.8)	2.79 (0.9)	2.29 (1.0)	1.89 (1.0)	1.47 (0.9)
20	3.68 (0.7)	3.57 (0.8)	3.46 (0.8)	3.29 (0.8)	3.14 (0.8)	2.97 (0.8)	2.73 (0.8)	2.45 (0.7)	1.98 (0.7)	1.54 (0.8)
25	3.64 (0.6)	3.48 (0.7)	3.38 (0.7)	3.25 (0.8)	3.09 (0.8)	2.81 (0.9)	2.62 (0.9)	2.37 (0.8)	2.09 (0.8)	1.72 (0.8)
30	3.42 (0.5)	3.30 (0.6)	3.17 (0.6)	3.00 (0.7)	2.82 (0.7)	2.60 (0.8)	2.38 (0.8)	2.13 (0.8)	1.90 (0.8)	1.57 (0.8)
35	3.39 (0.5)	3.23 (0.5)	3.05 (0.6)	2.87 (0.6)	2.66 (0.7)	2.42 (0.7)	2.22 (0.8)	2.01 (0.7)	1.78 (0.7)	1.55 (0.7)
40	3.27 (0.4)	3.11 (0.5)	2.91 (0.6)	2.78 (0.6)	2.58 (0.6)	2.35 (0.7)	2.14 (0.7)	1.88 (0.7)	1.68 (0.7)	1.44 (0.7)
45	3.17 (0.4)	3.00 (0.6)	2.77 (0.7)	2.58 (0.7)	2.41 (0.7)	2.18 (0.7)	1.88 (0.8)	1.70 (0.7)	1.52 (0.7)	1.33 (0.7)
50	3.17 (0.4)	3.03 (0.5)	2.83 (0.5)	2.57 (0.6)	2.28 (0.7)	1.99 (0.7)	1.76 (0.7)	1.58 (0.7)	1.34 (0.7)	1.17 (0.7)

Taulukko 2: Viimeisten komponenttien h-arvojen keskiarvot. Rivillä on muuttujien määrä p, sarakkeessa tuotettujen komponenttien määrä d, solussa h-arvojen keskiarvo ja suluissa keskihajonta.

Näytetään yksi aineistoesimerkki simulaatiosta ja käytetään muuttujien määrää $p = 5$ ja komponenttien määrää $d = 2$. Kuvassa 2 ovat alkuperäisten muuttujien parittaiset hajontakuviot. Ydin-FOBIn käytöllä saadaan tuotettua komponentti, jo-



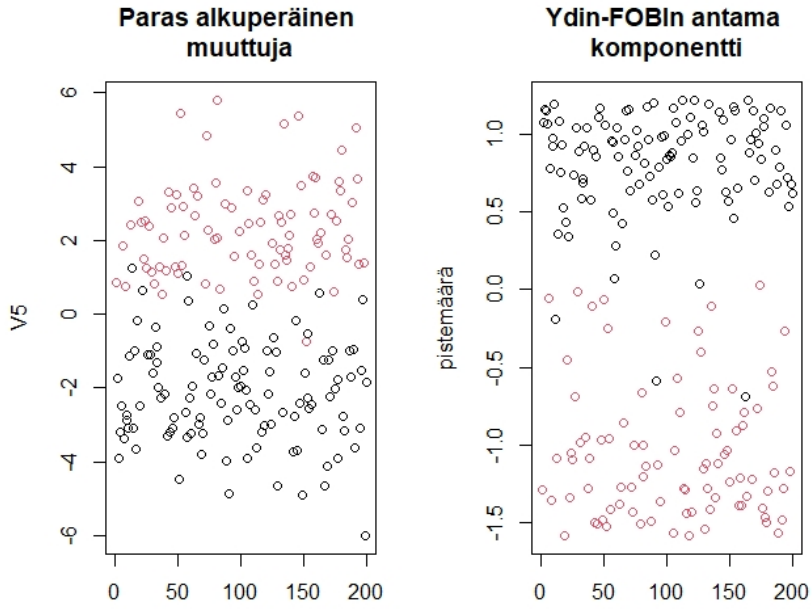
Kuva 2: Simulaation muuttujien parittaisen hajontakuviot

ka erottelee ryhmät selkeämmin ($h = 4.80$) kuin alkuperäisen aineiston parhaiten erotteleva muuttuja ($h = 3.36$). Tämä nähdään kuvan 3 hajontakuvasta. Kuvan 4 tiheysfunktioestimaateista nähdään, että ydin-FOBIn antaman komponentin jakauma on selkeästi kaksihuippuisempi, kuin parhaiten ryhmät erottelevan alkuperäisen muuttujan. On erityisen hyvä huomata, että ydin-FOBIssa käytettävä komponentti (viimeinen) on valittu ohjaamattomasti yleisen säännön pohjalta, päin vastoin kuin aineiston muuttuja (suurin h-arvo).

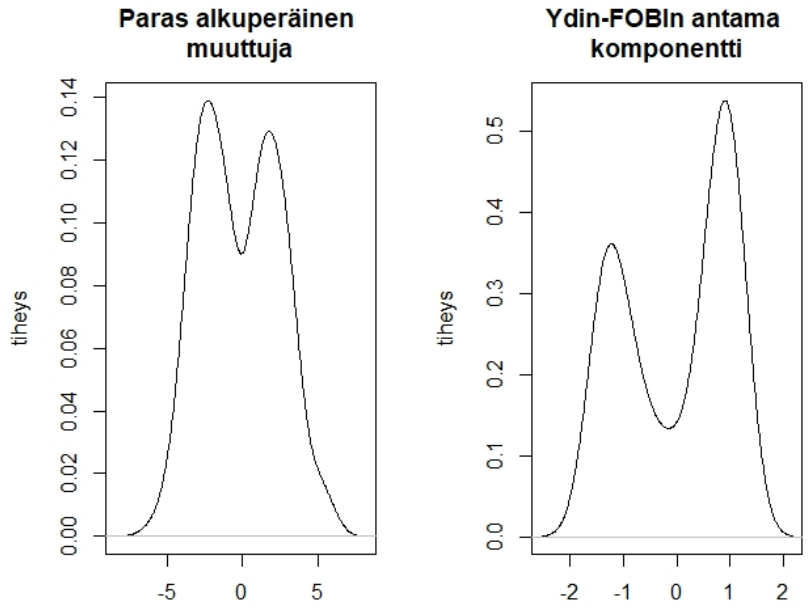
4.2 Ominaiskasvot

Pääkomponenttianalyysiä käytetään kasvojentunnistuksessa niin sanottujen *ominaiskasvojen (eigenfaces)* tuottamiseen [7]. Tämä tehdään tulkitsemalla (mustavalko-) kuvan pikselien tummuudet muuttujien arvoiksi, vektoroimalla jokainen kuva yhdeksi havaintomatriisin riviksi ja tekemällä tähän aineistoon pääkomponenttianalyysi. Näin voidaan tuottaa pääkomponentteja vastaavia kuvia, joissa pikselin tummuutena on sitä vastaavan muuttujan korrelaatio kyseisen pääkomponentin kanssa. Ajatuksena on, että tietyn pääkomponentin kanssa korreloivat pikselit vastaavat tiettyä ominaisuutta ja tämä on helppo todeta kuvista.

Tutkitaan nyt ydin-FOBIn käyttöä vastaavien kuvien tuottamiseen ja vertaillaan sitä muihin työssä esiteltyihin menetelmiin. Aineistona käytetään Yale Face



Kuva 3: Hajontakuvio ryhmien erottelusta ydin-FOBIlla

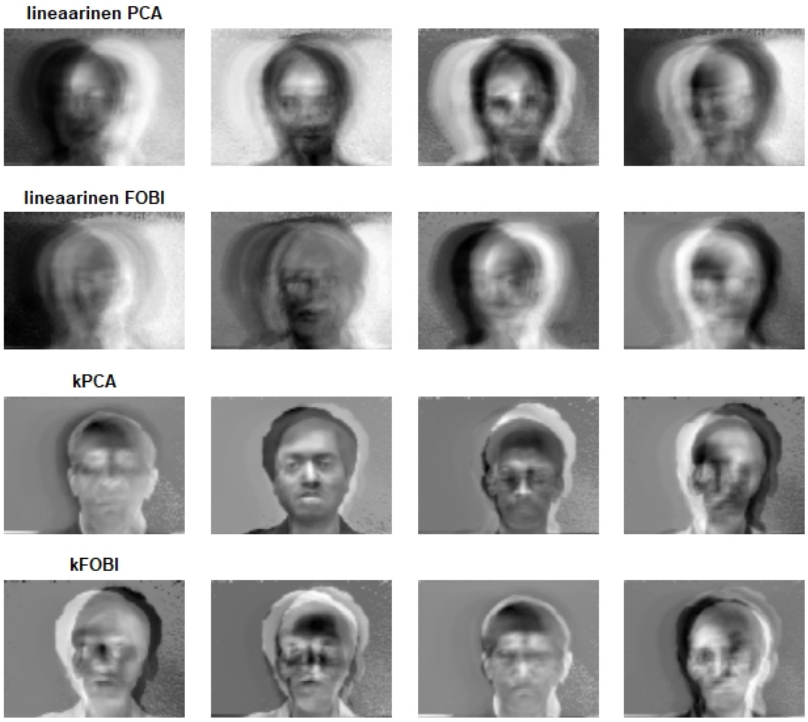


Kuva 4: Estimoidut tiheysfunktiot ryhmien erottelusta ydin-FOBIlla

Databasen kasvokuvia [23].

Kuvassa 5 esitetään neljän ensimmäisen komponentin tuottamat kuvat tavallisesta pääkomponenttianalyysistä ja FOBIsta sekä näiden sädekantafunktioytimen avulla toteutetuista ydinversioista eli kaikki komponenttien määrällä $d = 4$. Säde-

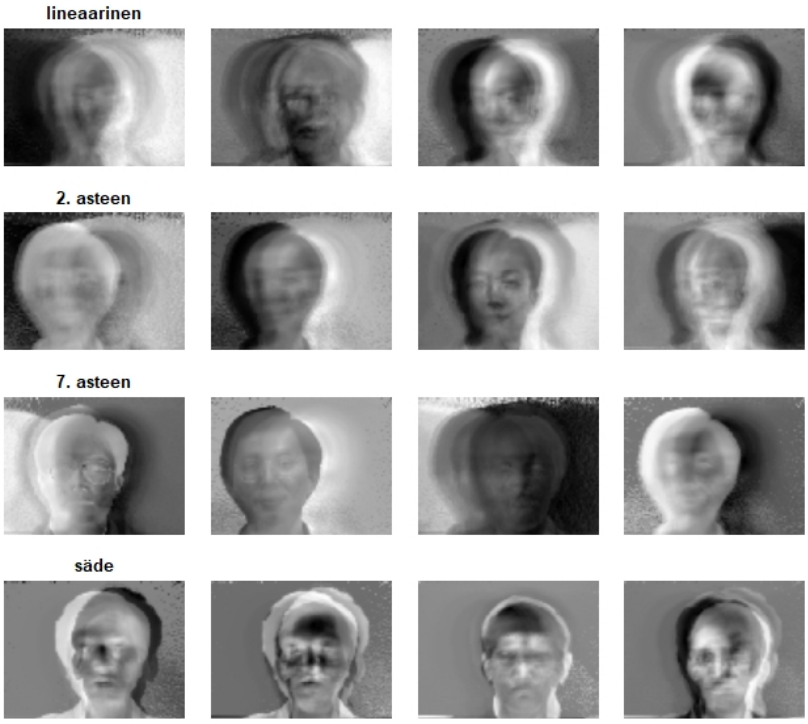
kantafunktion parametriksi σ^2 on valittu havaintojen neliöityjen etäisyyksien keskiarvo. Kuvasta huomataan, että lineaariset menetelmät ovat erittäin herkkiä sille, että kasvot ovat kuvissa hieman eri kohdissa. Ydinversiot taas tuottavat terävämpiä kuvia. Syynä saattaa olla se, että havaintojen välisen etäisyyden laskemiseen perustuva ydinfunktio tunnistaa paremmin yhteyden kahden samanlaisen, mutta hieman eri kohdassa olevan alueen välillä.



Kuva 5: Eri menetelmillä tuotettuja ominaiskasvoja

Kuvassa 6 on vertailtu ydin-FOBIn komponenttien määrällä $d = 4$ tuottamia kuvia eri ytimillä: lineaarisella (eli tavallisia sisätuloja käyttävällä), toisen ja seitsemännen asteen polynomiytimellä sekä sädekantafunktioytimellä. Tämä vertailu vahvistaa havaintoa, että juuri sädekantafunktio tuottaa tarkkarajaisimpia piirteitä.

Kokonaisuutena voidaan sanoa, että pikseleittäin toimivien kuvantunnistusmenetelmien suurin heikkous on herkkyys pieniä siirtoja kohtaan, mutta sädekantafunktion käyttö parantaa tätä tilannetta. Ydin-FOBIn ja ydin-PCA:n välillä ei ole kuitenkaan selkeää eroa. Lineaaristen menetelmien etuna taas on, että kaikki kasvot voidaan tuottaa ominaiskasvojen lineaarisena summana. Epälineaaristen ydinmenetelmien tapauksessa tämä ei ole (ainakaan helposti) mahdollista. Kasvoista tuotettuja riippumattomia komponentteja voitaisiin kuitenkin käyttää kasvojen erotteluun. Eräs mahdollisuus olisi esimerkiksi ryhmitellä havainnot jonkin tietyn komponentin (ja siis sitä mahdollisesti vastaavan piirteen) mukaan.

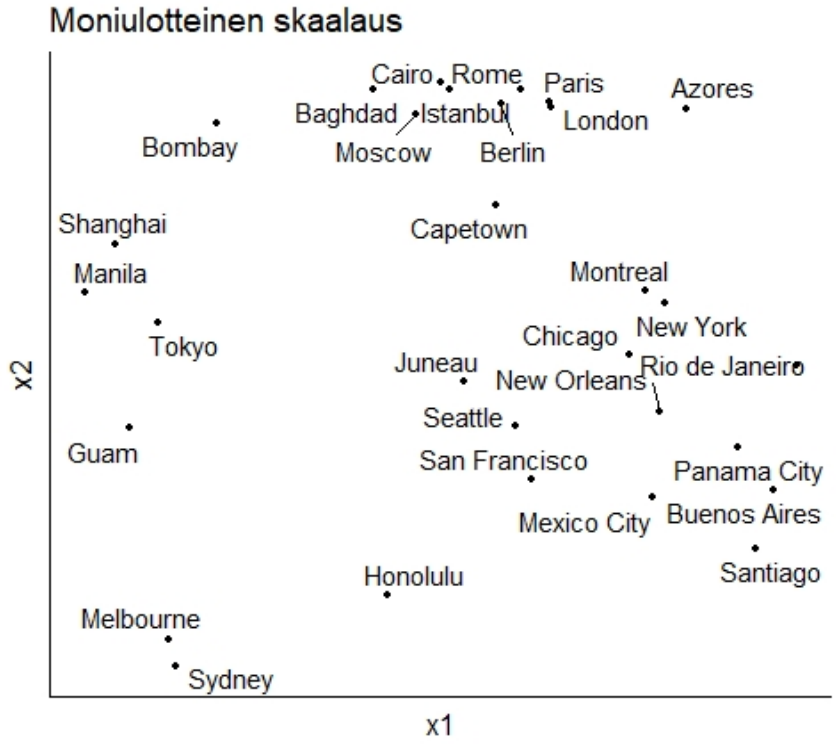


Kuva 6: Erilaisia ydinfunktioita käyttäen ydin-FOBIlla tuotettuja ominaiskasvoja

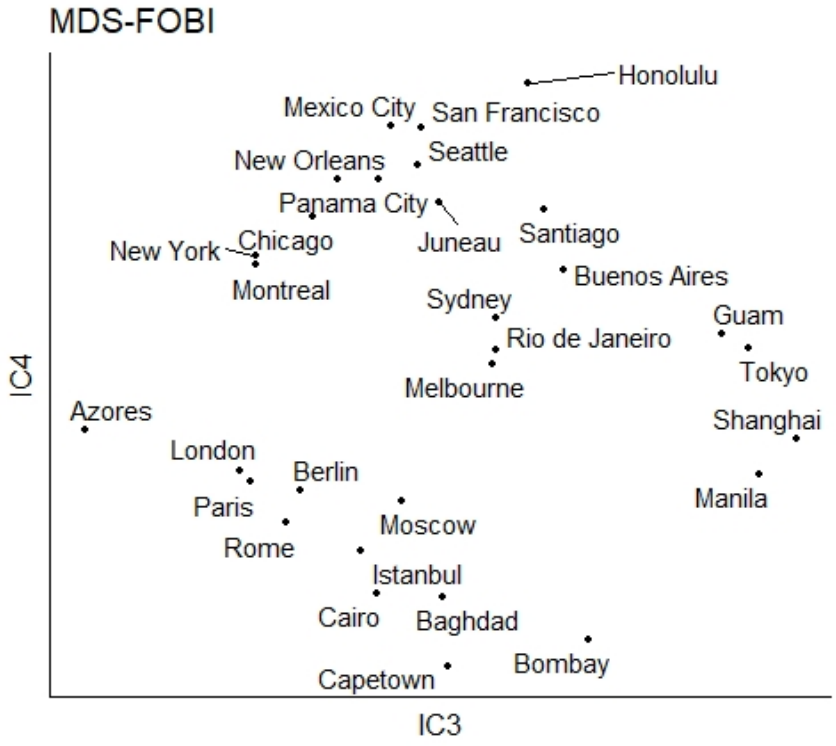
4.3 Kaupunkien paikat MDS-FOBIlla

Esitellään vielä MDS-FOBIn käyttöä ja vertaillaan sitä perinteiseen moniulotteiseen skaalaukseen. Aineisto (alunperin kirjassa [24]) koostuu kaupunkien välisistä lentäen mitatuista välimatkoista. Etäisyydet eivät ole euklidisia, sillä lentäminen tapahtuu maapallon pinnan mukaisesti, ei tasossa. Tämä edustaakin erästä perinteisintä moniulotteisen skaalauksen sovellusta (esimerkiksi paperisen maailmankartan tekeminen).

Kuvassa 7 on esitetty tavallisen moniulotteisen skaalauksen tuottama kaupunkien sijoittelu. Se vastaa pääpiirteittäin todellista maailmankarttaa. Kuvassa 8 taas on tehty MDS-FOBI komponenttien määrällä $d = 4$ ja piirretty kartta kahden viimeisen komponentin avulla. Tässä kuvassa kaupungit ovat jakautuneet jossain määrin kahteen ryhmään: Eurooppaan, Afrikkaan ja Lähi-itään sekä Amerikkaan, Aasiaan ja Australiaan. Tällainen jako ryhmiin on hyvä ominaisuus, sillä usein moniulotteisen skaalauksen ja muiden visualisointimenetelmien tarkoituksena on juuri löytää aineistosta ryhmiä. Tämä myös yhtenee aiempiin havaintoihin pääkomponenttianalyysin (jota moniulotteinen skaalaus edustaa) ja FOBIn eroista.



Kuva 7: Kaupunkien sijainnit lentoetäisyyksien perusteella moniulotteisella skaalauksella



Kuva 8: Kaupunkien sijainnit lentoetäisyyksien perusteella MDS-FOBIlla

5 Pohdinta ja johtopäätökset

Tutkielmassa esitettiin kaksi uutta menetelmää: ydin-FOBI ja MDS-FOBI. Näiden taustana toimi pääkomponenttianalyysin ja siihen liittyvien menetelmien sekä riippumattomien komponenttien analyysin ja lineaarisen FOBIn esittely. Keskeisin tulos oli FOBIn estimointi vain havaintojen sisätulojen avulla, sillä korvaamalla sisätulomatriisi ydinmatriisilla saatiin ydin-FOBI ja korvaamalla se etäisyysmatriisilla saatiin MDS-FOBI.

Ensimmäisessä, simuloiduilla aineistoilla tehdyssä tutkimuksessa havaittiin, että ydin-FOBI soveltuu ryhmien erotteluun. Komponenttien valintaperiaatteeksi tässä tilanteessa saatiin, että lähes aina viimeinen komponentti erottelee ryhmät parhaiten. Tämä on yhtenevää aiemman lineaarista FOBIa koskevan tutkimuksen kanssa. Ainakin tässä käytetyllä simulaatioaineistolla havaittiin kuitenkin, että ydin-FOBI toimii parhaiten kun komponenttien määrä on yksi, joka vastaa ydinpääkomponenttianalyysiä. Toisesta aineistolla tehdystä esimerkistä havaittiin, että ydin-FOBI (ja ydinpääkomponenttianalyysi) soveltuu myös niin sanottujen ominaiskasvojen tuottamiseen. Sädekantafunktion etuna on tällöin, että se erottaa kuvista reunat melko terävästi, vaikka kuvien välillä kasvot ovatkin hieman eri paikoissa. Kolmas esimerkki taas kertoo, että MDS-FOBIa voidaan käyttää tavallisen, pääkomponenttianalyysiperheeseen kuuluvan, moniulotteisen skaalauksen lailla. Tällöin sen ominaisuutena on, että se erottelee moniulotteiseen skaalaukseen verrattuna pisteet voimakkaammin ryhmiin.

Lineaarinen ICA voidaan ajatella pääkomponenttianalyysin yleistyksenä, koska riippumattomien komponenttien pistemäärät saadaan standardoimalla ja kiertämällä pääkomponenttipistemääriä. Kun katsotaan ydin-FOBI-algoritmia, havaitaan, että sekin koostuu kahdesta osasta: ydinpääkomponenttianalyysistä ja saaduilla pistemäärillä lasketusta tavallisesta FOBIsta. Tämä vahvistaa havaintoa menetelmien selkeästi sisäkkäisestä rakenteesta. Jää myös pohdittavaksi, voisiko muista ICA-menetelmistä, tai kokonaan muista menetelmistä, tuottaa ydinversioita tällä samalla ajatuksella.

Tässä tutkielmassa valittiin käsiteltäväksi riippumattomien komponenttien menetelmäksi FOBI, koska se on jossain mielessä yksinkertaisin ICA-menetelmä ja vastaa aivan selvällä tavalla pääkomponenttianalyysiä, jonka ydinversio tunnetaan. ICA-menetelmiä on kuitenkin runsaasti lisää (esimerkiksi ominaismatriisien yhteisdiagonalisointi (JADE) [19] ja symmetrisoitujen hajontamatriisien samanaikainen diagonalisointi [21]) ja näitä menetelmiä pidetään usein FOBIa tehokkaampina, joten ilmeinen jatkotutkimuksen aihe on kehittää ydinversioita myös näistä muista menetelmistä.

Ydin-FOBI on tavallisen FOBIn selkeä yleistys ja saakin nimensä siitä, että käytettäessä ytimenä tavallista euklidista sisätuloa, se antaa saman tuloksen kuin lineaarinen FOBI. Lineaarinen FOBI esiteltiin tässä tutkielmassa ratkaisuna tiettyyn lineaariseen riippumattomien komponenttien $\mathbf{s}$ sekoitusongelmaan $\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{s}$. Ydin-FOBIn tapauksessa ei kuitenkaan ole selvää, vastaako se jotakin tällaista komponenttien sekoitusongelmaa alkuperäisessä avaruudessa ja erityisesti mikä on ydinfunktion ja sitä vastaavan sekoitusongelman suhde. Jatkotutkimuksen aihe olisikin tutkia voitaisiinko ydin-FOBI (ja muut ydin-ICA-menetelmät) ymmärtää ratkaisuna

johonkin muotoa $\mathbf{x} = \Psi(\mathbf{s})$ olevaan ongelmaan sopivalla (ja riittävän säännöllisellä) Ψ.

Kiinnostava jatkotutkimuksen aihe on myös tutkia ydin-FOBIn ja muiden ydin-ICA-menetelmien asymptoottisia ominaisuuksia. Koska otoskoon n kasvattaminen kasvattaa myös $n \times n$ ydinmatriisia $\mathbf{K}$, käytetään tilanteen $n \to \infty$ tutkimiseen yleensä jäljentävän ytimen Hilbertin avaruuksien (RKHS) teoriaa. Siinä tutkitaan käsiteltävää menetelmää vastaavaa operaattoria ydintä vastaavassa RKHS:ssä. Esimerkkinä tästä on kanonisten korrelaatioiden analyysin ydinversion tarkentuvuutta käsittelevä artikkeli [25].

Viitteet

[1] Hyvärinen, A., Karhunen, J., Oja, E. (2001) *Independent Component Analysis*. John Wiley & Sons

[2] Cardoso, J. (1989) *Source Separation Using Higher Order Moments*. International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, 4: 2109-2112

[3] Jolliffe I. (2002) *Principal Component Analysis*, 2nd ed. Springer-Verlag New York

[4] Cortes, C., Vapnik, V. (1995) *Support-vector Networks*. Machine Learning 20(3): 273–297

[5] Schölkopf, B., Smola, A., Müller, K., (1998) *Nonlinear Component Analysis as a Kernel Eigenvalue Problem*. Neural Computation, 10(5): 1299-1319

[6] Mardia, K., Kent, J., Bibby, J., (1979) *Multivariate Analysis*. Academic Press, London-New York-Toronto-Sydney-San Francisco

[7] Sirovich, L., Kirby, M. (1987) *Low-dimensional Procedure for the Characterization of Human Faces*. Journal of the Optical Society of America A, 4: 519–524

[8] Pearson, K. (1901) *On Lines and Planes of Closest Fit to Systems of Points in Space*. Philosophical Magazine, 2: 559-572

[9] Härdle, W., Simar, L. (2007) *Applied Multivariate Statistical Analysis*. Springer-Verlag Berlin Heidelberg

[10] Hofmann, T., Schölkopf, B., Smola, A. J. (2008) *Kernel Methods in Machine Learning*. Annals of Statistics, 36(3): 1171–1220

[11] Aronszajn, N. (1950) *Theory of Reproducing Kernels*. Transactions of the American Mathematical Society, 68(3): 337–404

[12] Muscat, J. (2014) *Functional Analysis*. Springer International Publishing

[13] Hastie, T., Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer

[14] Williams, K. (2002) *On a Connection Between Kernel PCA and Metric Multidimensional Scaling*. Machine Learning, 46: 11–19, Kluwer Academic Publishers

[15] Nordhausen, K, Oja, H. (2018) *Independent Component Analysis: A Statistical Perspective*. WIREs Comput Stat., 10:e1440

[16] Virta, J., Nordhausen, K., Oja H. (2016) *Projection Pursuit for non-Gaussian Independent Components*. arXiv:1612.05445

[17] Ibragimov, I.A. (2014) *On the Ghurye–Olkin–Zinger Theorem*. Journal of Mathematical Sciences 199: 174–183

[18] Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H. (2015) *Fourth Moments and Independent Component Analysis*. Statistical Science 30(3): 372–390

[19] Cardoso, J.-F. ja Souloumiac, A. (1993) *Blind Beamforming for Non-Gaussian Signals*. In IEE Proceedings F-Radar and Signal Processing, 140: 362–370

[20] Hyvärinen, A. (1999) *Fast and Robust Fixed-point Algorithms for Independent Component Analysis*. IEEE Transactions on Neural Networks, 10(3): 626–634

[21] Taskinen, S., Sirkiä, S., ja Oja, H. (2007) *Independent Component Analysis Based on Symmetrised Scatter Matrices*. Computational Statistics & Data Analysis, 51(10): 5103–5111

[22] Peña, D., Prieto, F. J., Viladomat, J. (2010) *Eigenvectors of a Kurtosis Matrix as Interesting Directions to Reveal Cluster Structure*. Journal of Multivariate Analysis, 101(9): 1995-2007

[23] Belhumeur, P., Hespanha, J., Kriegman, D. (1997) *Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 711–720

[24] Hartigan, J. (1975) *Clustering Algorithms*. Wiley

[25] Fukumizu, K., Bach, F. R., ja Gretton, A. (2007). *Statistical Consistency of Kernel Canonical Correlation Analysis.* Journal of Machine Learning Research, 8(2)

[26] Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). *kernlab – An S4 Package for Kernel Methods in R.* Journal of Statistical Software, 11(9): 1–20

[27] Qiu, Y., Mei, J. (2019) *RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems*. https://CRAN.R-project.org/package=RSpectra

A R-koodit

R-koodi ydin-FOBIn ja MDS-FOBIn tekemiseen. Toteutuksissa käytetään R-paketteja `kernlab` [26] ja `RSpectra` [27].

```
# Ladataan käytettävät R-paketit

library(kernlab)
library(RSpectra)

# kfobi-funktio tuottaa annetusta matriisista ydin-FOBI pistemäärämatriisin
# X on käytettävä datamatriisi
# kernel on käytettävä ydinfunktio kernlab-paketista
# p on haluttu riippumattomien komponenttien määrä
# scores on saatujen riippumattomien komponenttien pistemäärien matriisi
# kurt on riippumattomien komponenttien kurtoosien vektori

kfobi = function(X, kernel=vanilladot(), p = ncol(X)){
  n = nrow(X)
  H = diag(rep(1,n)) - (1/n)*(rep(1,n) %*% t(rep(1,n))) # keskistysmatriisi
  K_pre = kernelMatrix(kernel, X) # keskistämätön ydinmatriisi
  K = H %*% K_pre %*% H # keskistetty ydinmatriisi
  eig = eigs_sym(K, p)
  V = eig$vectors
  C = (((n-1)^2)/n)*t(V) %*% diag(diag(V %*% t(V))) %*% V
  eig2 = eigen(C)
  E = eig2$vectors
  Pi = eig2$values
  Z = sqrt(n-1)*V %*% E
  return(list(scores = Z, kurt = Pi))
}

# MDSfobi-funktio tuottaa annetusta
#  etäisyysmatriisista MDS-FOBI pistemäärämatriisin
# D on käytettävä etäisyysmatriisi
# p on haluttu riippumattomien komponenttien määrä
# scores on saatujen riippumattomien komponenttien pistemäärien matriisi
# kurt on riippumattomien komponenttien kurtoosien vektori

MDSfobi = function(D, p = 2){
  n = nrow(D)
  H = diag(rep(1,n)) - (1/n)*(rep(1,n) %*% t(rep(1,n)))
  A = -0.5*D^2
  B = H %*% A %*% H
  eig = eigs_sym(B, p)
  V = eig$vectors
  C = (((n-1)^2)/n)*t(V) %*% diag(diag(V %*% t(V))) %*% V
  eig2 = eigen(C)
  E = eig2$vectors
  Pi = eig2$values
  Z = sqrt(n-1)*V %*% E
  return(list(scores = Z, kurt = Pi))
}
```