Tampere University

Ellinoora Hetemaa

PÄÄKOMPONENTTIANALYYSI POIKKEAMIEN HAVAITSEMISEEN GAMMASPEKTROMETRISISSÄ AIKASARJOISSA

Diplomityö
Tekniikan ja luonnontieteiden tiedekunta
Tarkastajat: Assoc. Prof. Pasi Raumonen,
Laboratorionjohtaja Aleksi Mattila
Joulukuu 2023

TIIVISTELMÄ

Ellinoora Hetemaa: Pääkomponenttianalyysi poikkeamien havaitsemiseen gammaspektrometrisissä aikasarjoissa
Diplomityö
Tampereen yliopisto
Teknis-luonnontieteellinen DI-ohjelma
Joulukuu 2023

Ympäristön säteilyturvallisuuden valvonnassa on keskeistä tunnistaa luonnollisen radioaktiivisen säteilyn seasta ihmislähtöistä, keinotekoista säteilyä. Säteilyn havainnointi perustuu usein mitattujen spektrien visualisointiin, jolloin osa keinotekoista säteilyä sisältävistä spektreistä peittyy helposti taustasäteilyn sekaan. Lisäksi visuaalinen havainnointi on hyvin riippuvaista datan tulkitsijasta sekä analysointiin käytettävissä olevan ajan määrästä.

Luonnon taustasäteilyn tila vaihtelee jatkuvasti erilaisten sääilmiöiden, kuten sateen tai lumipeitteen vuoksi, jolloin säteilynhavannointiin käytettäviin järjestelmiin aiheutuu kohinan lisäksi vääriä hälytyksiä. Lisäksi liikkuvia mittauksia tehtäessä mittausaineistossa on yleensä matalaresoluutioisten ilmaisimien käytön vuoksi paljon kohinaa. Tällöin mielenkiintoiset säteilytapaukset peittyvät helposti kohinan alle.

Tässä työssä tarkastellaan pääkomponenttianalyysin (PCA) käyttöä poikkeamien havaitsemiseen gammaspektrometrisistä aikasarjoista. Työn aikana kehitellään menetelmä, jolla pääkomponenttianalyysiä voidaan hyödyntää poikkeamien tunnistuksessa sekä testataan menetelmää muutamille erillisille testiaineistoille.

PCA on hyvin yleisesti käytetty monimuuttujamenetelmä, jolla pyritään ensisijaisesti vähentämään datan dimensioiden määrää. Pääkomponenttianalyysia voidaan kuitenkin hyödyntää myös kohinan poistamiseen tai poikkeavien havaintojen tunnistamiseen. PCA:n käyttöä luonnonsäteilystä poikkeavien säteilytapauksien havaitsemiseen on tutkittu jonkin verrana aikaisemmin, mutta useimmat nykyään käytössä olevat algoritmit perustuvat edelleen joko yksittäisen havainnon analysointiin tai spektridatan visuaaliseen tarkasteluun.

Työn kirjallisuuskatsauksessa perehdytään usean muuttujien aineistojen peruskäsitteisiin sekä syvennytään pääkomponenttianalyysin teoriaan ja sovellusalueisiin. Lisäksi esitellään perusasioita ympäristön säteilynvalvonnasta ja gammaspektrometrisistä säteilymittauksista.

Työn soveltavassa osiossa esitellään työn aikana kehitetty PCA-pohjainen algoritmi, jolla voidaan tunnistaa taustasäteilyn joukosta poikkeavia säteilyhavaintoja ilman taustasäteilyn vaihtelusta johtuvia vääriä hälytyksiä. Algoritmille syötetään esitietona taustasäteilystä muodostuva harjoitusdata, jonka perusteella se määrittää, onko tuntematon spektri harjoitusdatan spektrien kaltainen vai poikkeava. Testien perusteella algoritmilla voidaan havaita poikkeavaa säteilyä tiettyyn aktiivisuuten asti, eikä taustasäteilyn seasta nouse esiin juurikaan vääriä hälytyksiä.

Avainsanat: pääkomponenttianalyysi, PCA, gammaspektrometria, poikkeaman tunnistus, radionuklidi

ABSTRACT

Ellinoora Hetemaa: Principal Component Analysis for Anomaly Detection in Time Series of Gamma Spectra
Master's Thesis
Tampere University
Master's Programme in Science and Engineering
December 2023

In environmental radiation safety monitoring, a key challenge is to distinguish human-made artificial radiation from natural radioactive radiation. Radiation detection often relies on visualizing measured spectra, where spectra containing artificial radiation can easily be obscured by the background radiation. Additionally, visual observation is highly dependent on the interpreter's skills and the amount of time available for analysis.

The state of natural background radiation constantly varies due to various weather phenomena, such as rain or snow cover, leading to significant false alarms in radiation detection systems. Furthermore, when conducting mobile measurements, the data often contains a considerable amount of noise due to low-resolution detectors, causing interesting radiation events to be easily overshadowed by the noise.

This study examines the application of Principal Component Analysis (PCA) for detecting anomalies in time series of gamma spectra. A method is developed to utilize PCA for anomaly detection, and the method is tested on several independent datasets.

PCA is a widely used multivariate technique primarily aimed at reducing the dimensions of data. However, PCA can also be employed for noise removal or identifying abnormal observations. While some research has investigated the use of PCA for detecting radiation events deviating from natural background radiation, most current algorithms still rely on the analysis of individual observations or visual inspection of spectral data.

The literature review section delves into the basic concepts of multivariate datasets and explores the theory and application areas of Principal Component Analysis. Additionally, fundamental aspects of environmental radiation monitoring and gamma spectroscopic radiation measurements are introduced.

In the applied section of the study, a PCA-based algorithm developed during the research is presented. This algorithm can identify radiation observations deviating from background radiation without generating false alarms due to variations in background radiation. The algorithm is trained on background radiation data, allowing it to determine whether an unknown spectrum is similar to the training data spectra or deviates from them. Test results indicate that the algorithm can detect abnormal radiation up to a certain level of activity without producing significant false alarms from the background radiation.

Keywords: principal component analysis, PCA, gammaspectrometry, outlier detection, radionuclide

ALKUSANAT

Aloitin Säteilyturvakeskuksessa eli STUKissa kesäharjoittelijana keväällä 2022 tehtävänäni kehittää gammaspektrien analyysiin tarkoitettua ohjelmistoa. Työn mielekkyys ja monipuolisuus matematiikan, fysiikan ja ohjelmoinnin saralla saivat minut haaveilemaan mahdollisesta jatkopaikasta ja päädyimmekin lopulta pohtimaan diplomityön aihetta seuraavalle kesälle. Työn kirjoittaminen ja kasaan saattaminen on ollut ennen kaikkea antoisa ja mielenkiintoinen tehtävä, mutta välillä myös uuvuttava ja kärsivällisyyttä koetteleva projekti.

Haluan kiittää koko STUKia mahdollisuudesta kasvaa ja kehittyä oman alani asiantuntijana sekä kaikkia työkavereitani miellyttävistä lounas- ja kahvihetkistä lähityöpäivinä. Erityisesti haluan kiittää ohjaajaani Tero Karhusta loputtomista neuvoista ja vastauksista kiperiin kysymyksiini sekä esimiestäni Aleksi Mattilaa rohkaisevasta asenteesta ja joustavasta johtamisesta. Kiitos yliopisto-ohjaajalleni Pasi Raumoselle sekä projektiryhmäni jäsenille Mats Erikksonille ja Mark Dowdallille hedelmällisistä keskusteluista Teams-kokouksissa.

Mieheni ja kurssikaverini Aapo ansaitsee kiitokset kärsivällisyydestä ja tuesta koko opiskeluaikanamme. Haluan kiittää vanhempiani kasvatus- ja kannustustyöstä. Kiitän kaikkia läheisiäni ja ystäviäni Tampereella, muualla Suomessa ja ulkomailla. Työn valmistumisen myötä on mukava lähteä joulunviettoon.

Tampereella, 22. joulukuuta 2023

Ellinoora Hetemaa

SISÄLTÖ

1 Johdanto 1

2 Usean muuttujan aineistot 4
2.1 Odotusarvovektori 4
2.2 Matriisin aste 5
2.3 Vektorien ja matriisien ortogonaalisuus 5
2.4 Kovarianssi- ja korrelaatiomatriisi 6
2.5 Vektorien välinen etäisyys 8
2.6 Ominaisarvohajotelma 9
2.7 Singulaariarvohajotelma 9

3 Pääkomponenttianalyysi 11
3.1 Pääkomponenttien muodostaminen 11
3.1.1 Pääkomponentit kovarianssimatriisista 12
3.1.2 Pääkomponentit korrelaatiomatriisista 17
3.1.3 Pääkomponentit singulaariarvohajotelmasta 19
3.2 Datan esittäminen pääkomponenttien avulla 20
3.2.1 Biplot-kuvaajat 21
3.2.2 PCA-rekonstruktio 22
3.3 Pääkomponenttien määrän valinta 25
3.4 Poikkemien havaitseminen 26
3.4.1 Poikkeamat Biplot-kuvaajassa 27
3.4.2 Mahalanobiksen etäisyys 28
3.4.3 Hotellingin T2 -testi 29
3.4.4 PCA-sovitus 30

4 Aineisto ja menetelmät 31
4.1 Ympäristön gammaspektrometriset säteilymittaukset 31
4.1.1 Keinotekoisten radionuklidien havaitseminen spektristä 32
4.1.2 Luonnon taustasäteilyn spektrit 33
4.2 Aineisto 34
4.2.1 $LaBr_3$(Ce), Nuorgam 35
4.2.2 $LaBr_3$(Ce), Rovaniemi 36
4.2.3 $LaBr_3$(Ce), Kotka ja Harjavalta 36
4.3 Aiempia sovelluksia 37

4.4 PCA-algoritmi 38
4.4.1 Aineiston esikäsittely 39
4.4.2 PCA-mallin luonti harjoitusdatasta 40
4.4.3 Hälytysrajan määritys 40
4.4.4 PCA-mallin sovitus testidataan 43

5 Tulokset ja pohdintaa 44
5.1 $LaBr_3$(Ce), Nuorgam 44
5.2 $LaBr_3$(Ce), Rovaniemi 49
5.3 $LaBr_3$(Ce), Kotka ja Harjavalta 53

6 Yhteenveto 58

Lähteet 61
Liite A: PCA-algoritmin lähdekoodi 65
Liite B: Residuaalien simulointi 71
Liite C: Residuaalien jakauman todistus 74

LYHENTEET JA MERKINNÄT

ARAD	Autoencoder Radiation Anomaly Detection
d_e	Euklidinen etäisyys
$\mathbb{E}$	Odotusarvo
d_{em}	Muunneltu euklidinen etäisyys
FN	False Negatives
FP	False Positives
GEVD	Generalized Extreme Value Distribution, yleistetty ääriarvojakauma
$\mathbf{I}$	Identiteettimatriisi
$LaBr_3(Ce)$-ilmaisin	Cesiumilla rikastettuun lantaanibromidikiteeseen perustuva tuikeilmaisin
$\mathcal{H}$	Hypoteesi
$\mathcal{N}$	Normaalijakauma
NaI(Tl)-ilmaisin	Talliumilla rikastettuun natriumjodidikiteeseen perustuva tuikeilmaisin
NORM	Naturally Occurring Radioactive Material
$\mathcal{P}$	Poisson-jakauma
$\mathbb{P}$	Todennäköisyys
PCA	Pääkomponenttianalyysi
PNS	Pienin neliösumma
$\mathbf{R}$	Otoskorrelaatiomatriisi
$\mathbf{S}$	Otoskovarianssimatriisi
$\mathbf{\Sigma}$	Populaation kovarianssimatriisi
$\mathbb{R}$	Reaalilukujen joukko
STUK	Säteilyturvakeskus
FEM	Fixed Effects Model, kiinteiden vaikutusten malli
NASVD	Noise Adjusted Singular Value Decomposition
SVD	Singular Value Decomposition, Singulaariarvohajotelma

’	Matriisin transpoosi
TN	True Negatives
TP	True Positives
TW	Tracy-Widom

1. JOHDANTO

Gammasäteilyn mittaaminen ja analysointi ovat keskeisessä roolissa useiden maiden säteilyturvallisuustilanteen seurannassa. Ympäristössä sijaitsevat säteilynilmaisimet ovat alttiita säätilan muutosten aiheuttamalle vaihtelulle. Sateiden aikana ylempää ilmakehästä kulkeutuu gammasäteilyä lähettäviä luonnon radionuklideja lähelle maanpintaa, mikä saa aikaan vaihtelua gammaspektrien aikasarjoihin vaikeuttaen keinotekoisen aktiivisuuden havaitsemista ja aiheuttaen vääriä hälytyksiä. Toisaalta etenkin Suomessa talvisin vaikuttava lumipeite voi vaikeuttaa keinotekoisten säteilylähteiden havaitsemista.

Paikallaan sijaitsevien ilmaisimien lisäksi haitalliselle vaihtelulle ovat alttiita erilaisiin liikkuviin kalustoihin – kuten autot, helikopterit ja dronet – asennetut säteilynilmaisimet. Tällaisissa liikkuvissa mittausasemissa eli *mobiilimittausasemissa* käytetään tyypillisesti hyvin matalaresoluutioisia gammasäteilyn ilmaisimia, kuten NaI(Tl)-tuikeilmaisimia. Lisäksi mittausaika on mobiilimittauksissa usein hyvin lyhyt, vain noin 1-2 sekuntia. Tällöin analysoitavaa spektridataa kertyy huomattavia määriä ja se sisältää usein paljon kohinaa. Mobiilimittausten analysointi vaatii siis joko hyvin nopeaa reaaliaikaista toimintaa tai myöhemmin aikaa datan jälkikäsittelyyn. Vaikka mobiilimittaukset ovatkin jo hyvin vakiintunut tekniikka säteilyturvallisuuden saralla, on sen asema entisestäänkin korostunut. Mobiilimittaukset ovat korvaamaton apuväline orpojen säteilylähteiden[1] löytämiseen ja kontrollointiin. Toisaalta mobiilimittaukset ovat teknologian kehittymisen myötä siirtymässä uudempiin ilmaisimiin ja mukautuvampiin kuljetusajoneuvoihin, mitkä vaativat perehtymistä myös uusiin analyysimenetelmiin.

Tyypillisesti mobiilimittaukset ja reaaliaikaiset ilmaisinsysteemit nojautuvat datan visualisointiin, kuten spektrin vesiputous-näkymiin. Vesiputous-näkymä paljastaa hyvin tehokkaasti korkeaenergiset lähteet, mutta matala-energiset lähteet peittyvät helposti taustasäteilyn sekaan. Lisäksi vesiputousnäkymiä voi olla puuduttavaa katsoa datan suuren määrän vuoksi ja havainnointi on hyvin riippuvaista datan tulkitsijasta. Edellä mainitut skenaariot ja muut muuttujat ympäristössä aiheuttavat

[1] Orpo lähde on hylätty, hukattu, väärin sijoitettu, varastettu tai muuten ilman asianmukaista valtuutusta hallussa pidetty riittävän voimakas radioaktiivinen lähde, joka ei ole koskaan ollut tai ei ole enää viranomaisvalvonnan alla. [4, s. 2]

keinotekoista säteilyä havaitseviin järjestelmiin runsaasti Tyypin I ja Tyypin II virheitä. Tämän työn tavoitteena on kehittää menetelmä, jolla erityisesti Tyypin I virheiden lukumäärä edellä mainittujen kaltaisissa systeemeissä saadaan minimoitua. Tyypin I virhe syntyy kyseisessä sovelluskohteessa silloin, kun luonnon taustasäteilyä luokitellaan virheellisesti keinotekoiseksi säteilyksi.

Tässä työssä annetaan lukijalle kattava katsaus hyvin yleisesti käytetyn monimuuttujamenetelmän pääkomponenttianalyysin teoriasta ja sovellusalueista sekä kuvataan PCA-pohjainen algoritmi poikkeavan säteilyn havaitsemiseen $LaBr_3(Ce)$-tuikeilmaisemalla mitatuista gammaspektreistä. Menetelmää testataan usealla eri havaintoaineistolla sekä simuloidulla että oikeasti mitatulla spektridatalla. Menetelmä on luotu vastaamaan Suomen säteily- ja ydinturvallisuusviranomaisen, Säteilyturvakeskuksen (STUK) tarpeita. PCA:n sovellusaloja on useita ja niihin lukeutuvat esimerkiksi koneoppiminen, kuvankäsittely, biotieteet, genomiikka, taloustieteet, spektroskopia, signaalinkäsittely, geofysiikka, psykologia, sosiaalitieteet ja kauppatieteet. Lähteissä [26] ja [28] kerrotaan hyvin kattavasti PCA:n teoriasta ja sovelluksista.

PCA:ta on sovellettu aikaisemminkin gammaspektrometriaan hieman erilaisilla näkökulmilla. Lähteessä [32] PCA:ta on vastikään hyödynnetty robottipohjaisten ilmaisimien hakustrategioiden optimointiin. Lähteessä [45] PCA:ta on käytetty luonnollisten gammasäteiden analysointiin ilmaspektrometriassa ja lähteessä [52] NaI(Tl)-ilmaisimella mitattujen gammaspektrien analyysiin. PCA:han perustuvia menetelmiä on kehitetty lähteissä [40] ja [57]. Näistä edeltävä menetelmä on tarkoitettu säteilylähteen paikantamiseen ja jälkimmäinen kohinan vähentämiseen gammaspektreistä. Tämän työn kannalta tärkeimmässä lähteessä [10] on kuvattu PCA:han perustuva menetelmä, jolla pystytään havaitsemaan keinotekoisia säteilylähteitä ympäristöstä, ilman taustasäteilystä aiheutuvia vääriä hälytyksiä.

PCA:han perustuvia menetelmiä on siis kehitetty jo aiemmin, mutta juuri STUKin käyttöön soveltuvaa menetelmää ei ole kirjallisuudesta suoraan löydettävissä. Artikkelin [10] menetelmä eroaa tässä työssä kuvatusta algoritmista monella eri osa-alueella, vaikka aihepiiri ja tavoite ovatkin samat. Ensinnäkin aineistojen esikäsittelymetodit ovat osittain erilaisia. Lisäksi tässä työssä keskitytään $LaBr_3(Ce)$-tuikeilmaisimella mitattuihin gammaspektreihin, kun edellä mainitussa artikkelissa menetelmällä tutkitaan NaI(Tl)-ilmaisimella mitattua dataa. Menetelmien hälytysrajan määrityksissä on myös eroavaisuutta, kun artikkelin menetelmässä hälytysraja asetetaan NORM-säteilyn yläpuolelle, niin tämän työn menetelmässä hälytysraja lasketaan sovitusvirheiden jakaumista. Lisäksi sovitusvirheiden laskentatavat ovat erilaiset. Artikkelissa sovitusvirheiden laskemiseen käytetään *Mahalanobiksen etäisyyttä* ja tässä työssä puolestaan *euklidista etäisyyttä*.

Työ alkaa kirjallisuuskatsauksella ja päättyy soveltavaan osuuteen. Luvussa 2 esi-

tellään usean muuttujan aineistoihin liittyviä peruskäsitteitä ja ominaisuuksia. Luvussa 3 käsitellään esimerkkien avulla pääkomponenttianalyysin teoriaa ja sovellusmahdollisuuksia. Luvussa 4 puolestaan siirrytään työn soveltavaan osuuteen ja käydään lyhyesti läpi ympäristön säteilyvalvontaa sekä esitellään poikkeamien havaitsemiseen kehitetty algoritmi ja testauksessa käytetyt aineistot. Luvussa 5 esitellään PCA-algoritmin tulokset testiaineistoille ja pohditaan algoritmin toimivuutta kyseiseen sovellusalaan.

2. USEAN MUUTTUJAN AINEISTOT

Tilastotieteessä analysoitava aineisto sisältää usein useamman kuin yhden satunnaismuuttujan. Tällöin monia muuttujien määritelmiä, ominaisuuksia ja tuloksia on laajennettava avaruuteen, jossa toisistaan riippumattomia tai riippuvia muuttujia on enemmän kuin yksi. Usean muuttujan aineistoihin perustuvat tilastolliset menetelmät luokitellaan *monimuuttujamenetelmiksi* ja yksi niiden keskeisimmistä osa-alueista on muuttujien välisten korrelaatioiden tutkiminen. Muita yleisesti käytettyjä menetelmiä ovat muun muassa monen muuttujan varianssianalyysi, faktorianalyysi sekä tämän työn tärkein teoreettinen sisältö – pääkomponenttianalyysi.

Usean muuttujan aineistoa kuvataan yleensä *satunnaisvektorilla*, eli vektorilla, jonka elementit ovat satunnaismuuttujia. Satunnaisvektorin realisoitunut otos on matriisi, joka koostuu realisoituneista vektoreista. Populaatioita tarkasteltaessa, tiedossa on oltava satunnaismuuttujien jakaumat. Seuraavissa kappaleissa esitetään muutamia usean muuttujan populaatioiden ja otosten peruskäsitteitä. [51]

2.1 Odotusarvovektori

Usean muuttujan aineistolle voidaan laskea keskiarvo joko satunnaismuuttujien tai havaintojen suhteen. Tällöin tuloksena saadaan vektori, joka sisältää joko muuttujien tai havaintojen lukumäärää vastaavan määrän keskiarvoja. Olkoon $\mathbf{x}$ satunnaismuuttujista koostuva vektori, jonka dimensio on p. Lisäksi x_i on i. satunnaismuuttuja vektorissa $\mathbf{x}$ ja $i = 1, 2, \ldots, p$. Jos aineistossa on n havaintoa, kukin otos x_i on n vektori

$$x_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in}. \end{bmatrix} \tag{2.1}$$

Satunnaismuuttujien *odotusarvovektori* on p-dimensioinen vektori:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix}. \tag{2.2}$$

Odotusarvovektori voidaan laskea myös populaatiolle, jos tiedetään satunnaismuuttujien $x_1, x_2, \ldots, x_p$ odotusarvot. Tällöin satunnaisvektorin odotusarvoksi saadaan $\mathbb{E}[\mathbf{x}] = (\mathbb{E}[\mathbf{x}_1], \ldots, \mathbb{E}[\mathbf{x}_2])'$. [51]

2.2 Matriisin aste

Matriisin $\mathbf{X}$ *aste* on sen suurin mahdollinen lineaarisesti riippumattomien rivien tai sarakkeiden määrä, jotka ovat aina yhtä suuret [1, s. 74]. Jos matriisi $\mathbf{X}$ on dimensioltaan $n \times p$, niin matriisin $\mathbf{X}$ suurin mahdollinen aste on pienempi dimensioista n ja p. Jos matriisin aste on suurin mahdollinen, sanotaan, että matriisi on *täysiasteinen*. [51]

Esimerkki 2.1. Olkoon $\mathbf{X}$ 2×3-dimensioinen matriisi

$$\mathbf{X} = \begin{bmatrix} 1 & -2 & 3 \\ 5 & 2 & 4 \end{bmatrix}. \tag{2.3}$$

Matriisin $\mathbf{X}$ aste on 2, sillä sen rivit ovat lineaarisesti riippumattomia (kumpaakaan niistä ei saada toisen lineaarikombinaationa). Siitä huolimatta matriisin sarakkeet ovat lineaarisesti riippuvia, koska niille voidaan löytää vakiokertoimet c_1, c_2 ja c_3 siten, että

$$c_1 \begin{bmatrix} 1 \\ 5 \end{bmatrix} + c_2 \begin{bmatrix} -2 \\ 2 \end{bmatrix} + c_3 \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.4}$$

Eräs ratkaisu yhtälölle 2.4 on vektori $\mathrm{c} = [14\ -11\ -12]'$.

2.3 Vektorien ja matriisien ortogonaalisuus

Vektorit $\mathbf{u}$ ja $\mathbf{v}$, joiden dimensiot ovat samat, ovat ortogonaalisia, jos

$$\mathbf{u}'\mathbf{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = 0. \tag{2.5}$$

Lisäksi, jos $\mathbf{u}'\mathbf{u} = 1$ ja $\mathbf{v}'\mathbf{v} = 1$, vektorit $\mathbf{u}$ ja $\mathbf{v}$ ovat *ortonormaaleja*. [51]

Ortogonaalimatriisi on sellainen matriisi $\mathbf{A}$, jonka sarakkeet ovat ortonormaaleja

keskenään. Tällöin matriisille $\mathbf{A}$ pätee

$$\mathbf{A}'\mathbf{A} = \mathbf{I} \tag{2.6}$$

ja toisaalta

$$\mathbf{A}\mathbf{A}' = \mathbf{I}, \tag{2.7}$$

mistä nähdään, että myös matriisin $\mathbf{A}$ rivit ovat ortonormaaleja. Lisäksi nähdään, että matriisi $\mathbf{A}'$ on matriisin $\mathbf{A}$ käänteismatriisi. [51]

2.4 Kovarianssi- ja korrelaatiomatriisi

Satunnaisvektorin *kovarianssimatriisi* kuvastaa sen satunnaismuuttujien välistä vuorovaikutusta. Kovarianssimatriisi on oikeastaan moniulotteisen satunnaismuuttujan varianssi. Sen diagonaalilla on jokaisen komponentin varianssit ja diagonaalin ulkopuolella eri komponenttien väliset kovarianssit. [11] Määritellään merkinnät $\mathrm{cov}(x, y)$ ja $\mathrm{var}(x)$ kuvaamaan kahden satunnaismuuttujan x ja y välistä kovarianssia sekä satunnaismuuttujan x varianssia.

Määritelmä 2.1 (Kovarianssimatriisi)**.** [11, s. 83] Olkoon $\mathbf{x}$ satunnaisvektori, jonka dimensio on p. Lisäksi x_i on i. satunnaismuuttuja vektorissa $\mathbf{x}$ ja $i = 1, 2, \ldots, p$ Vektorin $\mathbf{x}$ *kovarianssimatriisi* on $p \times p$ matriisi

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \ldots & \mathrm{cov}(x_1, x_p) \\ \mathrm{cov}(x_1, x_2) & \mathrm{var}(x_2) & \ldots & \mathrm{cov}(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_1, x_p) & \mathrm{cov}(x_2, x_p) & \ldots & \mathrm{var}(x_p) \end{bmatrix}. \tag{2.8}$$

Satunnaisvektorin kovarianssimatriisi voidaan esittää myös odotusarvojen avulla [17]. Tällöin satunnaisvektorin $\mathbf{x}$ varianssi on sen kovarianssimatriisi:

$$\mathbf{\Sigma} = \mathrm{var}(\mathbf{x}) = \mathrm{cov}(\mathbf{x}, \mathbf{x}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])'], \tag{2.9}$$

missä $\mathbb{E}[\mathbf{x}]$ on satunnaisvektorin $\mathbf{x}$ odotusarvovektori.

Otoskovarianssimatriisi määritellään satunnaismuuttujien otosvarianssien ja otoskovarianssien avulla määritelmän 2.2 mukaisesti. Jos otoksessa on n havaintoa, muuttujan x_i otosvarianssi on

$$s_{ii} = s_i^2 = \frac{1}{n-1}(\sum_{k=1}^{n} x_{ik}^2 - n\overline{x}_i^2) \tag{2.10}$$

ja muuttujien x_i ja x_j otoskovarianssi puolestaan

$$s_{ij} = \frac{1}{n-1}(\sum_{k=1}^{n} x_{ik}x_{jk} - n\overline{x}_i\overline{x}_j). \tag{2.11}$$

[35, s. 6]

Määritelmä 2.2 (Otoskovarianssimatriisi)**.** [35, s. 6] Olkoon $\mathbf{X}$ matriisi, jossa on p muuttujaa ja n havaintoa. Lisäksi s_{ii} on muuttujan x_i otosvarianssi, s_{ij} muuttujien x_i ja x_j otoskovarianssi, $i = 1, 2, \ldots, p$ ja $j = 1, 2, \ldots, p$. Matriisin $\mathbf{X}$ *otoskovarianssimatriisi* on $p \times p$ matriisi

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}. \tag{2.12}$$

Otoskorrelaatiomatriisi määritellään samoin kuin otoskovarianssimatriisi, mutta kovarianssien tilalla matriisissa on muuttujien väliset korrelaatiot. Kahden satunnaismuuttujan x_i ja x_j välinen korrelaatio on

$$r_{ij} = \frac{s_{ij}}{s_i s_j} \tag{2.13}$$

[35, s. 7].

Määritelmä 2.3 (Otoskorrelaatiomatriisi)**.** [35, s. 7] Olkoon $\mathbf{X}$ matriisi, jossa on p muuttujaa ja n havaintoa. Lisäksi r_{ij} on muuttujien x_i ja x_j välinen korrelaatio, $i = 1, 2, \ldots, p$ ja $j = 1, 2, \ldots, p$. Matriisin $\mathbf{X}$ *otoskorrelaatiomatriisi* on $p \times p$ matriisi

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix}. \tag{2.14}$$

Odotusarvon avulla ilmaistuna satunnaisvektorin $\mathbf{x}$ korrelaatiomatriisi on

$$\mathbf{R} = \mathbb{E}[\mathbf{x}\mathbf{x}']. \tag{2.15}$$

[46]

Vakiovektoreista koostuvalle matriisille $\mathbf{A}$ ja satunnaisvektorille $\mathbf{X}$ pätee $\text{var}[\mathbf{AX}] = \mathbf{A}\text{var}(\mathbf{X})\mathbf{A}'$. Tämä voidaan todistaa transpoosin laskusääntöjä, kovarianssimatriisin määritelmää ja odotusarvon lineaarisuutta hyödyntäen. Lyhyesti ilmaistuna $\text{var}[\mathbf{AX}] = \mathbb{E}[(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])'] = \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])'] = \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])\mathbf{X} - \mathbb{E}[\mathbf{X}])'\mathbf{A}'] = \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])']\mathbf{A}' = \mathbf{A}\text{var}(\mathbf{X})\mathbf{A}'$.[51, s. 10, 48 ja 68].

2.5 Vektorien välinen etäisyys

Vektorien väliset etäisyydet antavat usein hyödyllistä tietoa esimerkiksi muuttujien tai havaintojen samankaltaisuuksista tai poikkeavuuksista. Yleisimmin käytetty avaruuden $\mathbb{R}^n$ vektorien välisistä etäisyyksistä on *euklidinen etäisyys*. Satunnaisvektoreiden tapauksessa voidaan ottaa huomioon niiden komponenttien varianssit ja korrelaatiot käyttämällä Mahalanobiksen etäisyyttä. Määritellään seuraavaksi edellä mainitut etäisyysmitat.

Euklidinen etäisyys kahden p-vektorin $\mathbf{u}$ ja $\mathbf{v}$ välillä lasketaan kaavalla

$$d_e = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_p - v_p)^2}. \tag{2.16}$$

Euklidinen etäisyys kaavassa 2.16 voidaan esittää myös muodossa

$$d_e = (\mathbf{u} - \mathbf{v})'(\mathbf{u} - \mathbf{v}). \tag{2.17}$$

Kuten edellä mainittiin, Mahalanobiksen etäisyys arvioi kahden datapisteen välisen etäisyyden ottaen huomioon satunnaismuuttujien väliset keskinäiset riippuvuudet. Se tarjoaa tehokkaan tavan määrittää, kuinka kaukana tietty datapiste on keskimääräisestä pisteestä moniulotteisessa avaruudessa, jossa muuttujat saattavat olla eri mittakaavoissa ja korreloituneita keskenään.

Kahden vektorin $\mathbf{u}$ ja $\mathbf{v}$ välisen Mahalanobiksen etäisyyden neliö on

$$d_m^2 = (\mathbf{u} - \mathbf{v})'\mathbf{S}^{-1}(\mathbf{u} - \mathbf{v}), \tag{2.18}$$

missä $\mathbf{S}$ on otoskovarianssimatriisi. Usein Mahalanobiksen etäisyydelle käytetään myös kaavaa

$$d_m^2 = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{2.19}$$

missä d_m on Mahalanobiksen etäisyys, x on tarkasteltava datapiste, $\boldsymbol{\mu}$ on datan keskiarvo ja $\boldsymbol{\Sigma}^{-1}$ on datan kovarianssimatriisin käänteismatriisi. Jälkimmäisellä kaavalla 2.19 lasketaan, kuinka paljon tietty vektori poikkeaa populaation keskiarvosta, kun taas edeltävällä kaavalla 2.18 tutkitaan kahden vektorin välistä etäisyyttä jossain otoksessa. [51]

2.6 Ominaisarvohajotelma

Olkoon $\mathbf{X}$ täysiasteinen $(n \times n)$ matriisi sekä $\mathbf{C}$ matriisi, joka sisältää matriisin $\mathbf{X}$ normalisoidut ominaisvektorit. Matriisi $\mathbf{X}$ voidaan esittää *ominaisarvohajotelmana (engl. eigenvalue decomposition tai spectral decomposition)* [51, s. 35]

$$\mathbf{X} = \mathbf{C}\mathbf{D}\mathbf{C}', \tag{2.20}$$

missä

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}. \tag{2.21}$$

Diagonaalimatriisin $\mathbf{D}$ diagonaalialkiot ovat matriisin $\mathbf{X}$ ominaisarvot.

Lisäksi matriisi $\mathbf{C}$, jonka sarakkeina ovat symmetrisen matriisin $\mathbf{X}$ normalisoidut ominaisvektorit, on ortogonaalinen [51, s. 35]. Tällöin $\mathbf{C}'\mathbf{C} = \mathbf{C}\mathbf{C}' = \mathbf{I}$ ja kun kerrotaan hajotelmaa 2.20 vasemmalta matriisilla $\mathbf{C}'$ ja oikealta matriisilla $\mathbf{C}$ saadaan

$$\mathbf{C}'\mathbf{X}\mathbf{C} = \mathbf{D}. \tag{2.22}$$

Toisin sanoen symmetrinen matriisi $\mathbf{X}$ voidaan *diagonalisoida* ortogonaalimatriisilla $\mathbf{C}$ (ts. matriisi $\mathbf{X}$ on *diagonalisoituva*), joka sisältää matriisin $\mathbf{X}$ normalisoidut ominaisvektorit. Tämän seurauksena muodostuva diagonaalimatriisi $\mathbf{D}$ sisältää siis matriisin $\mathbf{X}$ ominaisarvot.

2.7 Singulaariarvohajotelma

Singulaariarvohajotelma yleistää luvussa 2.6 esitetyn neliömatriisien ominaisarvohajotelman kaikille $(n \times p)$ matriiseille. Olkoon meillä $(n \times p)$ matriisi $\mathbf{X}$, joka sisäl-

tää n havaintoa ja p muuttujaa. $\mathbf{X}$ voidaan esittää *singulaariarvohajotelmana (engl. singular value decomposition, SVD)*

$$\mathbf{X} = \mathbf{ULA}', \tag{2.23}$$

missä $\mathbf{U}$ on $(n \times n)$ ortonormaalimatriisi, $\mathbf{A}$ on $(p \times p)$ ortonormaalimatriisi, $\mathbf{L}$ on $(n \times p)$-diagonaalimatriisi [28]. Seuraavat määritelmät ja lauseet tarkentavat singulaariarvohajotelman ominaisuuksia.

Lause 2.1. [31, s. 18] *Olkoon $\boldsymbol{X}$ $(n \times p)$-matriisi. Tällöin $n \times n$ matriisi $\boldsymbol{X}'\boldsymbol{X}$ on symmetrinen, diagonalisoituva ja sen ominaisarvot ovat ei-negatiivisia. Tod. [31, s. 15].*

Määritelmä 2.4 (Singulaariarvot)**.** [31, s. 18] Olkoon $\mathbf{X}$ $n \times p$ matriisi. Matriisin $\mathbf{X}$ *singulaariarvot* $\sigma_1, \sigma_2, \ldots, \sigma_n$ ovat matriisin $\mathbf{X}'\mathbf{X}$ ominaisarvojen $\lambda_1, \lambda_2, \ldots, \lambda_n$ neliöjuuria $\sigma_i = \sqrt{\lambda_i} \geq 0$.

Edellisen määritelmän 2.4 matriisi $\mathbf{X}'\mathbf{X}$ on aina symmetrinen ja reaalinen, mistä seuraa, että sen ominaisarvot ovat aina reaalisia.

Lause 2.2 (Singulaariarvohajotelma)**.** [31, s. 18] *Jokainen $n \times p$ matriisi $\mathbf{X}$ voidaan esittää muodossa*

$$\mathbf{X} = \mathbf{ULA}', \tag{2.24}$$

missä $(n \times n)$-matriisin $\mathbf{U}$ sarakevektorit u_i ovat matriisin $\mathbf{XX}'$ ominaisvektoreita. Lisäksi $(n \times p)$-diagonaalimatriisin $\mathbf{L}$ järjestyksessä $\sigma_1 \geq \cdots \geq \sigma_r > 0$ olevat diagonaalialkiot ovat matriisin $\mathbf{X}$ singulaariarvoja (aina positiivisia). Singulaariarvojen määrä on yhtä suuri kuin matriisin $\boldsymbol{X}$ aste r. Jos matriisi $\boldsymbol{X}$ ei ole täysiasteinen, matriisi $\boldsymbol{L}$ täydennetään riittävän suurilla nollamatriiseilla. $(p \times p)$-matriisin $\mathbf{A}$ sarakevektorit v_i ovat matriisin $\mathbf{X}'\mathbf{X}$ ominaisvektoreita. Lisäksi matriisit $\mathbf{U}$ ja $\mathbf{A}$ ovat molemmat ortogonaalisia. Tod. [31][34].

3. PÄÄKOMPONENTTIANALYYSI

Pääkomponenttianalyysi (*engl. Principal Component Analysis*, PCA) on tilastollinen monimuuttujamenetelmä, jolla pyritään pienentämään datan dimensionaalisuutta. Menetelmällä muodostetaan uusi muuttujajoukko alkuperäisten muuttujien lineaarikombinaatioina siten, että mahdollisimman vähän datan informaatiosta katoaa. Näitä uusia muuttujia sanotaan *pääkomponenteiksi* (*engl. principal component*, PC) ja ne on järjestetty sen mukaan, kuinka suuren osan datan varianssista ne selittävät. Ensimmäinen pääkomponentti selittää suurimman osan datan varianssista, toinen pääkomponentti toiseksi suurimman ja niin edelleen. Pääkomponenttianalyysiä käytetään hyvin laajasti eri tieteenaloilla, muun muassa biologiassa, kemiassa, tietojenkäsittelytieteissä ja kauppatieteissä. [28]

Datan dimensioiden määrää voidaan vähentää valitsemalla pääkomponenteista vain osa, usein niin monta komponenttia alusta, että ne yhdessä selittävät tarpeeksi suuren osan datan vaihtelusta. Luvussa 3.3 käydään läpi erilaisia tapoja, joiden avulla pääkomponenttien määrä voidaan valita. [28]

3.1 Pääkomponenttien muodostaminen

Olkoon $\mathbf{x}$ satunnaismuuttujista koostuva vektori, jonka dimensio on p. Olkoon lisäksi $\boldsymbol{\alpha}$ avaruuden $\mathbb{R}^p$ vektori. Ensimmäinen pääkomponentti on lineaarikombinaatio $\boldsymbol{\alpha}_1'\mathbf{x}$ siten, että sen varianssi on mahdollisimman suuri ja

$$\boldsymbol{\alpha}_1'\mathbf{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = \sum_{j=1}^{p} a_{1j}x_j. \tag{3.1}$$

Yhtälössä 3.1 vektori $\boldsymbol{\alpha}_1$ sisältää vakiokertoimet kullekin satunnaismuuttujalle x_1, $x_2, \ldots, x_p$.

Toinen pääkomponentti $\boldsymbol{\alpha}_2'\mathbf{x}$ on korreloimaton ensimmäisen pääkomponentin kanssa ja selittää jälleen mahdollisimman suuren osan datan jäljellä olevasta varianssista. Loput komponentit muodostetaan samalla tavalla, siten että k komponentin muodostamisen jälkeen komponentti $\boldsymbol{\alpha}_k'\mathbf{x}$ ei korreloi komponenttien $\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}, \ldots, \boldsymbol{\alpha}_{k-1}'\mathbf{x}$ kanssa ja sen varianssi on mahdollisimman suuri.

Pääkomponentit on määritelty kirjallisuudessa kahdella tapaa. Joissain lähteissä [28][51] pääkomponentit määritellään lineaarikombinaatioina $\boldsymbol{\alpha}_k'\mathbf{x}$ ja toisissa [3][53] taas ainoastaan kerroinvektoreina $\boldsymbol{\alpha}_k$. Lineaarikombinaatioista $\boldsymbol{\alpha}_k'\mathbf{x}$ käytetään usein myös nimitystä *pisteet (engl. scores)* ja niitä merkitään yleensä p-vektorilla

$$\mathbf{z} = \mathbf{A}'\mathbf{x}, \tag{3.2}$$

missä matriisin $\mathbf{A}$ sarakkeet ovat kerroinvektoreita $\boldsymbol{\alpha}_k$ ja vektori $\mathbf{x}$ sisältää alkuperäiset muuttujat. Käytetään tässä työssä pääkomponenteille muotoa $\boldsymbol{\alpha}_k'\mathbf{x}$, missä $\boldsymbol{\alpha}_k$ on kerroin- tai *latausvektori (engl. loadings)*. Käytetään lisäksi muunnosta 3.2 kaikkien pääkomponenttien esittämiseen samanaikaisesti.

Pääkomponentteja voidaan johtaa niin monta, kuin alkuperäisessä datassa on muuttujia. On kuitenkin toivottavaa, että jo muutamat ensimmäiset komponentit selittäisivät suurimman osan datan vaihtelusta. [28]

3.1.1 Pääkomponentit kovarianssimatriisista

Esitetään seuraavaksi eräs yleisesti käytetty tapa johtaa pääkomponentit, mukaillen lähdettä [28]. Muodostetaan tässä vain kaksi ensimmäistä komponenttia, sillä loput komponentit $\boldsymbol{\alpha}_3'\mathbf{x}, \boldsymbol{\alpha}_4'\mathbf{x}, \ldots, \boldsymbol{\alpha}_p'\mathbf{x}$ saadaan johdettua hyvin samalla tavalla, kuin toinen pääkomponentti $\boldsymbol{\alpha}_2'\mathbf{x}$.

Ensimmäisen pääkomponentin kohdalla kerroinvektori $\boldsymbol{\alpha_1}$ maksimoi lausekkeen $\text{var}[\boldsymbol{\alpha_1'x}] = \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1$, missä $\boldsymbol{\Sigma}$ on analysoitavan aineiston, satunnaisvektorin $\mathbf{x}$ kovarianssimatriisi. Tehtävä olisi mahdoton ilman jonkinlaista rajoitetta vektorille $\boldsymbol{\alpha}_1$, joten valitaan normalisaatiorajoitteeksi $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$. Ensimmäisen komponentin johto voidaan siis esittää optimointitehtävänä

$$\begin{aligned} &\max_{\boldsymbol{\alpha}_1 \in \mathbb{R}^p} \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1, \\ &\text{s.e. } \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1. \end{aligned} \tag{3.3}$$

Optimointitehtävä 3.3 voidaan ratkaista esimerkiksi Lagrangen kertoimia [29] käyttäen. Toisin sanoen maksimoidaan

$$\boldsymbol{\alpha_1'\Sigma\alpha_1} - \lambda_1(\boldsymbol{\alpha_1'\alpha_1} - 1), \tag{3.4}$$

missä λ_1 on Lagrangen kerroin.

Derivoidaan $\boldsymbol{\alpha}_1$ suhteen ja asetetaan derivaatta nollaksi, jolloin 3.4 saadaan muotoon

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda_1\boldsymbol{\alpha}_1 = 0 \tag{3.5}$$

tai

$$(\boldsymbol{\Sigma} - \lambda_1\mathbf{I}_p)\boldsymbol{\alpha}_1 = 0, \tag{3.6}$$

missä $\mathbf{I}_p$ on $(p \times p)$-identiteettimatriisi. Yhtälöstä 3.6 nähdään, että λ_1 on matriisin $\boldsymbol{\Sigma}$ ominaisarvo ja $\boldsymbol{\alpha}_1$ sitä vastaava ominaisvektori.

Optimointitehtävässä 3.3 maksimoitava lauseke saadaan reunaehdon ja yhtälön 3.5 perusteella muotoon

$$\boldsymbol{\alpha_1'\Sigma\alpha_1} = \boldsymbol{\alpha}_1'\lambda_1\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda_1, \tag{3.7}$$

mistä huomataan, että tehtävän ratkaisemiseksi riittää valita ominaisarvo λ_1 siten, että sen on ominaisarvoista on suurin. Tämän perusteella ensimmäisen pääkomponentin kerroinvektoriksi $\boldsymbol{\alpha}_1$ saadaan kovarianssimatriisin $\boldsymbol{\Sigma}$ suurinta ominaisarvoa λ_1 vastaava ominaisvektori.

Toinen pääkomponentti saadaan johdettua ensimmäisen komponentin avulla, kun otetaan huomioon uusi rajoite, jonka mukaan seuraavan komponentin täytyy olla korreloimaton edellisen komponentin kanssa. Optimoitavaksi saadaan siis tehtävä

$$\begin{aligned} &\max_{\boldsymbol{\alpha}_2 \in \mathbb{R}^p} \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_2, \\ &\text{s.t. } \operatorname{cov}[\boldsymbol{\alpha_1'x}, \boldsymbol{\alpha_2'x}] = 0 \qquad (3.8) \\ &\text{ja } \boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1. \qquad (3.9) \end{aligned}$$

Kovarianssirajoitteelle pätee

$$\operatorname{cov}[\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}] = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2'\lambda_1\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2. \tag{3.10}$$

Kun noudatetaan reunaehtoa 3.8 ja asetetaan yhtälön 3.10 viimeinen lauseke nollaksi, saadaan normalisaatiorajoite huomioon ottaen maksimoitavaksi suureeksi

$$\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1', \tag{3.11}$$

missä λ_2 ja ϕ ovat Lagrangen kertoimia. Asettamalla muuttujan $\boldsymbol{\alpha_2}$ suhteen derivoitu yhtälö nollaksi saadaan

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = 0. \tag{3.12}$$

Kun yhtälöä 3.12 kerrotaan vasemmalta puolelta vektorilla $\boldsymbol{\alpha'}_1$, se saadaan muotoon

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha_1}'\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0. \tag{3.13}$$

Yhtälön 3.10 ja reunaehtojen 3.8 ja 3.3 perusteella yhtälön 3.13 kaksi ensimmäistä termiä ovat nollia ja $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$. Siten nähdään, että $\phi = 0$. Nyt yhtälöstä 3.12 saadaan

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 = 0. \tag{3.14}$$

Kuten ensimmäisen pääkomponentin kohdalla, tästä nähdään, että λ_2 on matriisin $\boldsymbol{\Sigma}$ ominaisarvo ja $\boldsymbol{\alpha}_2$ sitä vastaava ominaisvektori. Kun oletetaan, että matriisilla $\boldsymbol{\Sigma}$ ei ole samansuuruisia ominaisarvoja, on oltava $\lambda_2 \neq \lambda_1$. Siten λ_2 on kovarianssimatriisin toiseksi suurin ominaisarvo ja $\boldsymbol{\alpha}_2$ sitä vastaava ominaisvektori. Loput pääkomponentit saadaan johdettua samaan tapaan siten, että pääkomponenttien kerroinvektorit $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_p$ ovat matriisin $\boldsymbol{\Sigma}$ suuruusjärjestyksessä olevia ominaisarvoja $\lambda_3, \lambda_4, \ldots, \lambda_p$ vastaavat ominaisvektorit.

Johdetaan seuraavaksi kovarianssimatriisin $\boldsymbol{\Sigma}$ ominaisarvojen ja pääkomponenttien varianssien suhde. Oletetaan, että edellä esitetyt normalisaatio- ja kovarianssirajoitteet ovat edelleen voimassa. Toisin sanoen kerroinvektoreille pätee $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$, ja korreloimattomille pääkomponenteille pätee $\boldsymbol{\alpha}_k'\Sigma\boldsymbol{\alpha}_i = 0$, missä $k \neq i$.

Nyt pääkomponenttien 3.2 kovarianssimatriisi on muotoa

$$\boldsymbol{\Sigma}_z = \begin{bmatrix} \text{var}(\boldsymbol{\alpha}_1'\mathbf{x}) & \text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) & \dots & \text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_p'\mathbf{x}) \\ \text{cov}(\boldsymbol{\alpha}_2'\mathbf{x}, \boldsymbol{\alpha}_1'\mathbf{x}) & \text{var}(\boldsymbol{\alpha}_2'\mathbf{x}) & \dots & \text{cov}(\boldsymbol{\alpha}_2'\mathbf{x}, \boldsymbol{\alpha}_p'\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\boldsymbol{\alpha}_p'\mathbf{x}, \boldsymbol{\alpha}_1'\mathbf{x}) & \text{cov}(\boldsymbol{\alpha}_p'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) & \dots & \text{var}(\boldsymbol{\alpha}_p'\mathbf{x}) \end{bmatrix} \tag{3.15}$$

$$= \begin{bmatrix} \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_p \\ \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_p \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\alpha}_p'\Sigma\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_p'\Sigma\boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_p'\Sigma\boldsymbol{\alpha}_p \end{bmatrix} \tag{3.16}$$

$$= \begin{bmatrix} \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\alpha}_p'\boldsymbol{\Sigma}\boldsymbol{\alpha}_p \end{bmatrix} \tag{3.17}$$

$$= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \tag{3.18}$$

Yhtälön 2.22 mukaan mikä tahansa symmetrinen neliömatriisi $\mathbf{X}$ voidaan diagonalisoida matriisilla $\mathbf{C}$, jos matriisin $\mathbf{C}$ sarakkeet ovat matriisin $\mathbf{X}$ normalisoidut ominaisvektorit. Tässä tapauksessa matriisia $\mathbf{C}$ vastaa matriisi $\mathbf{A}$ ($\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}$) ja matriisia $\mathbf{X}$ vastaa matriisi $\boldsymbol{\Sigma}$, joten saadaan

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \mathbf{D}, \tag{3.19}$$

missä matriisin $\mathbf{D}$ diagonaalialkiot ovat matriisin $\boldsymbol{\Sigma}$ ominaisarvoja. Yhdistämällä kohdat 3.17 ja 3.19 saadaan

$$\begin{bmatrix} \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\alpha}_p'\Sigma\boldsymbol{\alpha}_p \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \tag{3.20}$$

eli toisin sanoen pääkomponenteille pätee $\text{var}(z_k) = \lambda_k$.

Edellä esitetty algebrallinen johto on yksi mahdollinen lähestymistapa pääkomponenttien löytämiseksi. Toinen mahdollinen lähestymistapa on johtaa pääkomponentit geometrisesti, kuten esimerkiksi lähteissä [22, 36] on tehty.

Esimerkki 3.1. Muodostetaan esimerkin vuoksi pääkomponentit yleisesti käytetylle Kurjenmiekka-aineistolle *(engl. Iris)*, joka on kerätty ja julkaistu alkunperin lähteissä [6][18]. Aineisto sisältää mittaushavaintoja kolmen erityyppisen Kurjenmiekkakukan eri osista. Muuttujia aineistossa on yhteensä 5 ja mittaushavaintoja 150, 50 jokaisesta kolmesta eri kukkatyypistä. Taulukossa 3.1 on aineiston 6 ensimmäistä ja viimeinen havainto.

Taulukko 3.1. *Kurjenmiekka-aineisto.*

havainto	Verholehden pituus (cm)	Verholehden leveys (cm)	Terälehden pituus (cm)	Terälehden leveys (cm)	Laji
1	5.1	3.5	1.4	0.2	setosa *(lat.)*
2	4.9	3.0	1.4	0.2	setosa *(lat.)*
3	4.7	3.2	1.3	0.2	setosa *(lat.)*
4	4.6	3.1	1.5	0.2	setosa *(lat.)*
5	5.0	3.6	1.4	0.2	setosa *(lat.)*
6	5.4	3.9	1.7	0.4	setosa *(lat.)*
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3	5.1	1.8	virginica *(lat.)*

Otetaan analyysiin mukaan vain muuttujat, joiden arvot ovat numeerisia. Esimerkkiaineiston 3.1 otoskovarianssimatriisi neljälle ensimmäiselle muuttujalle on

$$\mathbf{S} = \begin{bmatrix} 0.6857 & -0.0424 & 1.2743 & 0.5163 \\ -0.0424 & 0.1900 & -0.3297 & -0.1216 \\ 1.2743 & -0.3297 & 3.1163 & 1.2956 \\ 0.5163 & -0.1216 & 1.2956 & 0.5810 \end{bmatrix}. \tag{3.21}$$

Muodostetaan matriisit $\mathbf{V}$ ja $\mathbf{D}$ siten, että matriisin $\mathbf{V}$ sarakkeet ovat matriisin $\mathbf{S}$ ominaisvektorit ja matriisin $\mathbf{D}$ diagonaalialkiot niitä vastaavat ominaisarvot. Matriisi $\mathbf{V}$ on tällöin

$$\mathbf{V} = \begin{bmatrix} 0.3614 & -0.6566 & 0.5821 & 0.3155 \\ -0.0845 & -0.7302 & -0.5977 & -0.3201 \\ 0.8567 & 0.1735 & -0.0761 & -0.4798 \\ 0.3583 & 0.0752 & -0.5461 & 0.7535 \end{bmatrix}, \tag{3.22}$$

ja matriisi $\mathbf{D}$ on

$$\mathbf{D} = \begin{bmatrix} 4.2283 & 0 & 0 & 0 \\ 0 & 0.2427 & 0 & 0 \\ 0 & 0 & 0.0782 & 0 \\ 0 & 0 & 0 & 0.0238 \end{bmatrix}. \tag{3.23}$$

Nyt matriisiseista $\mathbf{D}$ ja $\mathbf{V}$ nähdään, että matriisin $\mathbf{S}$ suurin ominaisarvo on $\lambda_1 = 4.2283$ ja sitä vastaava ominaisvektori $\boldsymbol{\alpha}_1 = (0.3614, -0.0845, 0.8567, 0.3583)$. Ensimmäisen pääkomponentin kerroinvektori on siis edellä määritelty $\boldsymbol{\alpha}_1$ ja sen varianssi on λ_1. Kun merkitään taulukon 3.1 aineistoa vektorilla $\mathbf{x}$, ensimmäinen pääkomponentti saadaan kertolaskulla $\boldsymbol{\alpha}_1\mathbf{x} = 0.3614 \cdot$ (verh. pit.) - $0.0845 \cdot$ (verh. lev.) + $0.8567 \cdot$ (ter. pit.) + $0.3583 \cdot$ (ter. lev.) = (2.8182, 2.7882, 2.6134, 2.757, 2.7736, 3.2215, ..., 6.8925). Edellä laskettu vektori sisältää nyt siis suurimman osan aineiston 3.1 varianssista. Kerroinvektorit lopuille pääkomponentteille nähdään samaan tapaan matriisista $\mathbf{V}$.

3.1.2 Pääkomponentit korrelaatiomatriisista

Jos pääkomponentit muodostetaan korrelaatiomatriisista, ne voidaan esittää muunnoksen 3.2 tavoin muodossa

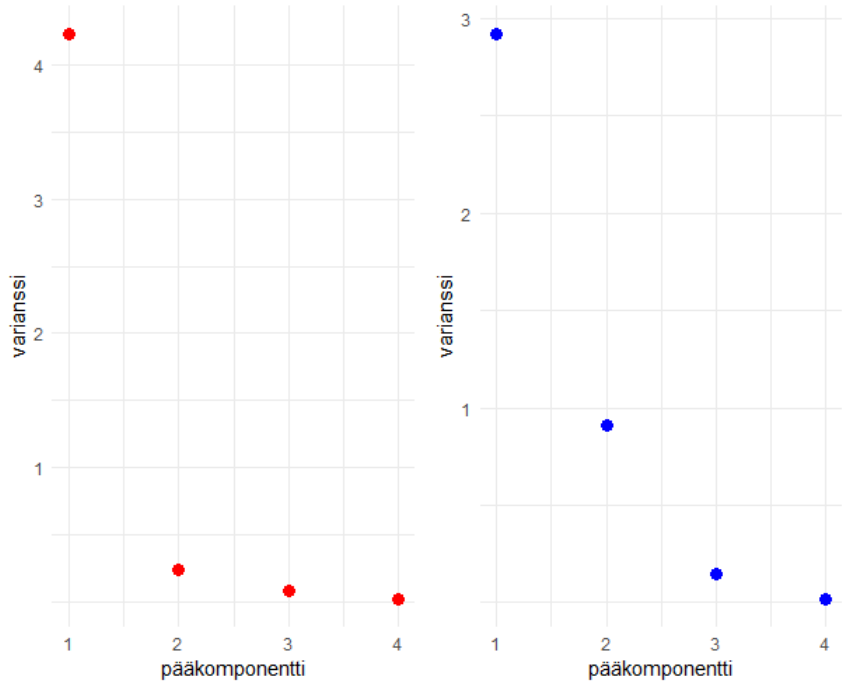
$$\mathbf{z} = \mathbf{A}'\mathbf{x}^*, \tag{3.24}$$

missä matriisin $\mathbf{A}$ sarakkeet ovat korrelaatiomatriisin ominaisvektoreita ja satunnaisvektori $\mathbf{x}^*$ sisältää tarkasteltavan aineiston, satunnaisvektorin $\mathbf{x}$ standardisoidut muuttujat. Muuttujat x_j^* vektorissa $\mathbf{x}^*$ on standardisoitu siten, että $x_j^* = x_j/\sigma_{jj}$, missä $j = 1, 2, \ldots, p$, x_j on vektorin $\mathbf{x}$ j. muuttuja, ja σ_{jj}^2 on muuttujan x_j varianssi. Siten vektorin $\mathbf{x}^*$ kovarianssimatriisi on vektorin $\mathbf{x}$ korrelaatiomatriisi ja pääkomponentit voidaan muodostaa suoraan alkuperäisten muuttujien korrelaatiomatriisin avulla. [28]

Pääkomponenttien muodostaminen korrelaatiomatriisista on tilanteesta riippuen usein järkevämpää kuin kovarianssimatriista. Suurin puoltava tekijä on se, että korrelaatiomatriisia käyttämällä PCA-tulokset erilaisille muuttujajoukoille ovat helpommin vertailtavissa keskenään. Kovarianssimatriisin iso haittapuoli on sen pääkomponenttien sensitiivisyys muuttujien mittayksikköjen varianssille. Toisin sanoen sellaiset muuttujat, joiden varianssi on suuri verrattuna muihin, hallitsevat helposti ensimmäisiä pääkomponentteja. Toisaalta kovarianssimatriisin käyttäminen voi olla pe-

rusteltua silloin, kun muuttujat ovat samassa mittayksikössä ja muuttujien mittasuhteissa ei ole suurta eroa. Tämäkään ei aina riitä, sillä usein myös samassa mittayksikössä olevien muuttujien välillä voi olla suuria skaalaeroja. Tarkastellaan esimerkin vuoksi tilannetta, jossa yksi muuttuja ilmaisee ihmisen pituutta ja toinen ranteen paksuutta. Vaikka molemmat muuttujat mitattaisiin senttimetreinä, pituus-muuttujan arvot olisivat selvästi suurempia, ja siten niissä olisi myös enemmän vaihtelua. Tällöin kovarianssimatriisi antaa enemmän painoarvoa muuttujalle, jonka arvot ovat paljon suurempia, tässä tapauksessa pituudelle. [28].

Korrelaatiomatriisin käyttö onkin perusteltua, jos aineisto koostuu hyvin erilaisista muuttujista eri mittayksiköissä. Myös tilanteissa, joissa halutaan vertailla eri analyysien tuloksia keskenään, pääkomponenttien muodostaminen korrelaatiomatriisista on välttämätöntä. Kuvassa 3.1 on esitettynä aineistolle 3.1 sekä kovarianssi- että korrelaatiomatriisista muodostettujen pääkomponenttien varianssit. Kuvaajista nähdään, että kovarianssimatriisin tapauksessa ensimmäinen pääkomponentti dominoi selvästi pääkomponenttien kokonaisvarianssia. Sen sijaan korrelaatiomatriisista muodostettujen pääkomponenttien varianssit ovat huomattavasti lähempänä toisiaan. Vaikka muuttujat ova samassa mittayksikössä, tämä ero näillä kahdella eri tavalla muodostettujen pääkomponenttien välillä kertoo luultavasti siitä, että skaalaerojen vuoksi jonkin muuttujan varianssi on selvästi muita suurempi. Tämä saattaa johtaa joidenkin analyysin tulosten vääristymiseen. [28]



Kuva 3.1. *Iris -aineistosta 3.1 muodostettujen pääkomponenttien varianssit. Vasemmanpuoleisessa kuvaajassa pääkomponentit on otettu kovarianssimatriisista ja oikeanpuoleisessa kuvaajassa korrelaatiomatriisista.*

3.1.3 Pääkomponentit singulaariarvohajotelmasta

Kolmas yleinen tapa muodostaa pääkomponentit on tehdä se singulaariarvohajotelman avulla. Hajotelma on määritelty kappaleessa 2.7 ja sen mukaan matriisi $\mathbf{X}$ voidaan esittää hajotelmana $\mathbf{ULA}'$.

Olkoon $\mathbf{S}$ otoksen $\mathbf{X}$ otoskovarianssimatriisi. Jos $\mathbf{X}$:n muuttujien keskiarvot ovat tuntemattomia, otoskovarianssimatriisin alkio (j, k) on

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k), \tag{3.25}$$

missä $\overline{x}_j$ on otoskeskiarvo ja x_{ij} on i. havainnon ja j. muuttujan arvo matriisissa $\mathbf{X}$. Nyt matriisi $\mathbf{S}$ voidaan esittää muodossa

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X}. \tag{3.26}$$

Matriisien $\frac{1}{1-n}\mathbf{X}'\mathbf{X}$ ja $\mathbf{X}'\mathbf{X}$ ominaisvektorit ovat samat ja matriisin $\frac{1}{1-n}\mathbf{X}'\mathbf{X}$ ominaisarvot ovat $\frac{1}{1-n}$ kertaa matriisin $\mathbf{X}'\mathbf{X}$ ominaisarvot. [28]

Olkoon $l_1, \ldots, l_r$ ja $\mathbf{a}_1, \ldots, \mathbf{a}_r$ matriisin $\mathbf{X}'\mathbf{X}$ ominaisarvot ja ominaisvektorit. Matriisin $\mathbf{X}'\mathbf{X}$ ominaisarvohajotelma on

$$(n-1)\mathbf{S} = \mathbf{X}'\mathbf{X} = l_1\mathbf{a}_1\mathbf{a}_1' + l_2\mathbf{a}_2\mathbf{a}_2' + \cdots + l_r\mathbf{a}_r\mathbf{a}_r', \tag{3.27}$$

missä r on matriisin $\mathbf{X}'\mathbf{X}$ aste. Määritellään nyt matriisit **A, U** ja **L** siten, että matriisin $\mathbf{A}$ $(p \times r)$ k. sarake on $\mathbf{a}_k$, matriisin $\mathbf{U}$ $(n \times r)$ k. sarake on $\mathbf{u}_k = l_k^{-1/2}\mathbf{X}\mathbf{a}_k$, diagonaalimatriisin $\mathbf{L}$ $(r \times r)$ k. diagonaalialkio on $l_k^{1/2}$ ja $k = 1, 2, \ldots, r$. Sen lisäksi, että edellä määritellyt matriisit toteuttavat ehdot 1, 2 ja 3 määritelmästä 2.23, voidaan osoittaa, että $\mathbf{X} = \mathbf{ULA}'$. Todistusta ei käydä tässä työssä tarkemmin läpi, mutta se löytyy esimerkiksi lähteestä [28, s. 44].

PCA:n kannalta on tärkeää huomata, että samalla kun matriisit $\mathbf{A}$ ja $\mathbf{L}$ antavat matriisin $\mathbf{X}'\mathbf{X}$ ominaisvektorit ja ominaisarvojen neliöjuuret, ne antavat myös pääkomponenttien kerroinvektorit ja keskihajonnat. Lisäksi matriisista $\mathbf{U}$ saadaan kertoimella $\frac{1}{n-1}$ skaalatut versiot pääkomponenteille. [28]

Singulaariarvohajotelman käytölle pääkomponenttien muodostuksessa on useita perusteluja. Ensinnäkin se tarjoaa laskennallisesti tehokkaan tavan löytää pääkomponentit [28]. Lisäksi useissa eri lähteissä [15, 19, 37, 50] mainitaan, että SVD antaa aiempiin menetelmiin verraten lisätietoa pääkomponenttianalyysistä sekä hyödyllisiä keinoja PCA-tulosten esittämiseen. Jos tarkastellaan esimerkiksi hajotelman 2.23

alkioita

$$x_{ij} = \sum_{k=1}^{r} u_{ik} l_k^{1/2} a_{jk}, \ k = 1, 2, \ldots, r \tag{3.28}$$

missä u_{ik} ja a_{jk} ovat alkiot (i, k) ja (j, k) matriiseissa $\mathbf{U}$ ja $\mathbf{A}$ ja $l_k^{1/2}$ on k. diagonaalialkio matriisissa $\mathbf{L}$. Summan 3.28 alkiot vastaavat r ensimmäistä pääkomponenttia. Jos valitaan vain $m < r$ ensimmäistä pääkomponenttia, matriisin $\mathbf{X}$ alkioille kaavassa 3.28 saadaan approksimaatio

$$\hat{x}_{ij} = \sum_{k=1}^{m} u_{ik} l_k^{1/2} a_{jk}. \tag{3.29}$$

Voidaan itse asiassa osoittaa, että kaava3.29 antaa parhaan mahdollisen m-asteisen approksimaation yhtälön 3.28 alkiolle x_{ij} minimoimalla euklidisen etäisyyden niiden välillä. [19][23]

3.2 Datan esittäminen pääkomponenttien avulla

Pääkomponenttianalyysin päätarkoitus on pienentää datan dimensioiden määrää, mutta sen lisäksi se antaa myös joitain hyväksi todettuja työkaluja moniulotteisen datan visualisointiin. Etenkin tilanteissa, joissa $p > 2$, PCA:n kyky projisoida data moniulotteisesta avaruudesta 2-ulotteiseen on hyvin hyödyllinen datan rakenteen hahmottamisen kannalta.

Eräs yksinkertainen tapa esittää aineisto 2-ulotteisessa avaruudessa on projisoida data kahdelle ensimmäiselle pääkomponentille ja sitten piirtää havaintopisteet näille kahdelle akselille. Toinen yleisesti käytetty tapa on käyttää niin kutsuttua *biplot*-kuvaajaa, joka esittää datan havainnot pisteinä ja alkuperäiset muuttujat vektoreina kahdella ensimmäisellä pääkomponentilla. Vektorien pituus ja suunta ilmaisevat kunkin muuttujan vaikutuksen kokonaisvarianssiin. Eri suuntaan osoittavat muuttujat korreloivat negatiivisesti ja samaan suuntaan osoittavat muuttujat positiivisesti.

Joskus moniulotteisen aineiston rakenne antaa mahdollisuuden esittää koko data 2-ulotteisessa kuvaajassa. Tyypillisiä esimerkkejä tällaisille aineistoille ovat erilaiset aikasarjat, joissa kukin havaintopiste kuvaa tietyin aikavälein tehtyä mittausta ja joissa muuttujat ovat samassa mittayksikössä. Näin mitatusta aineistosta voidaan piirtää jokaiselle havainnolle oma kuvaaja. Kun jokaisen havainnon rekonstruoi pääkomponenttien avulla, saadaan tulokseksi alkuperäisen aineiston approksimaatio, joissa havainnot on selitetty jonkin varianssiosuuden mukaan. Näitä havaintokuvaajia ja rekonstruoidun aineiston havaintokuvaajia vertailemalla voidaan saada olen-

naista tietoa esimerkiksi poikkeavista havainnoista.

3.2.1 Biplot-kuvaajat

Biplot-kuvaaja on tehokas työkalu datan visualisointiin moniulotteisessa avaruudessa. Se mahdollistaa sekä muuttujien että havaintopisteiden visualisoinnin samalla kuvaajalla. Tässä kuvaajassa muuttujat esitetään vektoreina, joiden suunnat ja kulmat kertovat olennaista tietoa niiden välisistä korrelaatioista. Havaintopisteet puolestaan projisoidaan tähän avaruuteen kahden ensimmäisen pääkomponentin avulla.

Biplot -kuvaaja perustuu luvussa 2.7 esiteltyyn singulaariarvohajotelmaan, jonka mukaan matriisi $\mathbf{X}$ $(n \times p)$ voidaan esittää muodossa

$$\mathbf{X} = \mathbf{ULA}'. \tag{3.30}$$

Nyt diagonaalimatriisin $\mathbf{L}$ $(r \times r)$ elementit ovat $l_1^{1/2} \geq l_2^{1/2} \geq \ldots l_r^{1/2}$ ja r on matriisin $\mathbf{X}$ aste. Määritellään nyt $\mathbf{L}^\alpha$ ja $\mathbf{L}^{1-\alpha}$ diagonaalimatriiseina, joiden diagonaalialkiot ovat $l_1^{\alpha/2}, l_2^{\alpha/2}, \ldots, l_r^{\alpha/2}$ ja $l_1^{(1-\alpha)/2}, l_2^{(1-\alpha)/2}, \ldots, l_r^{(1-\alpha)/2}$, kun $0 \leq \alpha \leq 1$. Olkoon lisäksi $\mathbf{G} = \mathbf{UL}^\alpha$ ja $\mathbf{H}' = \mathbf{L}^{1-\alpha}\mathbf{A}'$. Nyt saadaan

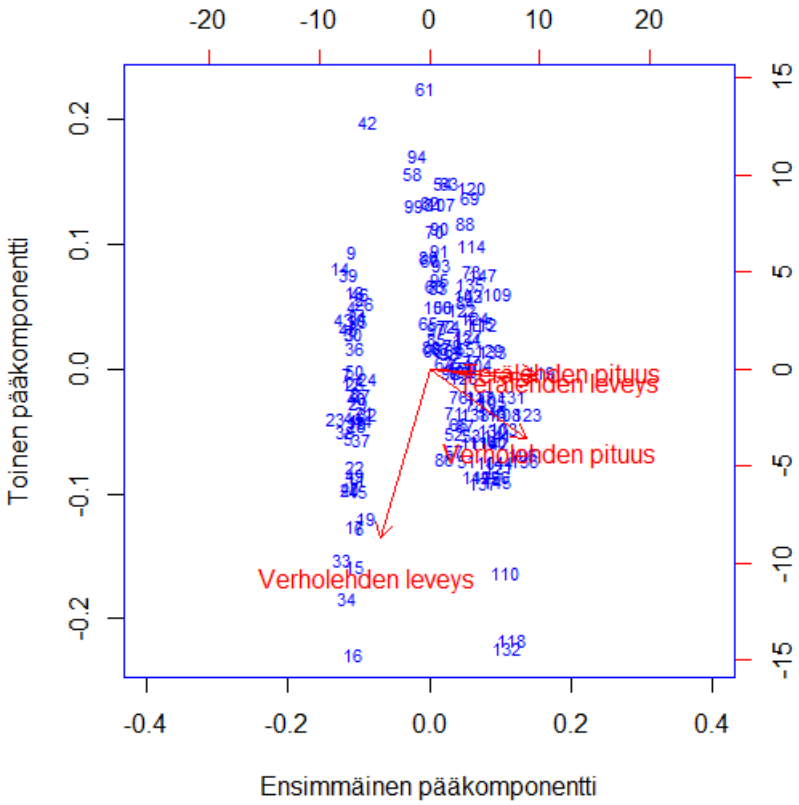
$$\mathbf{GH}' = \mathbf{UL}^\alpha\mathbf{L}^{1-\alpha}\mathbf{A}' = \mathbf{ULA}' = \mathbf{X}, \tag{3.31}$$

minkä mukaan matriisin $\mathbf{X}$ alkio (i, j) on

$$x_{ij} = \mathbf{g}_i'\mathbf{h}_j \tag{3.32}$$

ja $\mathbf{g}_i'$ ja $\mathbf{h}_j'$ ovat matriisien $\mathbf{G}$ ja $\mathbf{H}$ i. ja j. rivit. Toisin sanoen $\mathbf{g}_i'$ ja $\mathbf{h}_j'$ kuvaavat matriisin $\mathbf{X}$ i. havaintoa ja j. muuttujaa projisoituna pääkomponenteille. Nyt $\mathbf{g}_i'$ ja $\mathbf{h}_j'$ sisältävät molemmat r alkiota. Jos $r = 2$ kaikki nämä alkiot voidaan piirtää kaksiulotteisen avaruuden pisteinä, mistä muodostuu haluttu biplot-kuvaaja. Yleisemmin tapauksissa, joissa $r > 2$, voidaan $\mathbf{g}_i'$ ja $\mathbf{h}_j'$ approksimoida ottamalla niistä tarkasteluun vain $m < r$ ensimmäistä alkiota. [28]

Kuvassa 3.2 on biplot piirrettynä taulukon 3.1 aineistolle. Kuvassa punaisella näkyvät alkuperäiset muuttujat ja niiden vektorit sekä sinisellä aineiston havainnot. Ensinnäkin kuvasta nähdään, että kaksi havaintoryhmää poikkeaa toisistaan ensimmäisen pääkomponentin suhteen. Lisäksi vektoreiden avulla voidaan päätellä muuttujien välillä olevan osittain positiivista ja osittain negatiivista korrelaatiota.



Kuva 3.2. *Biplot taulukon 3.1 aineistolle. Akselit vasemmalla ja alhaalla ovat alkuperäisten havaintojen esittämiseen ja akselit ylhäällä ja oikealla alkuperäisten muuttujien esittämiseen pääkomponenttiavaruudessa*

3.2.2 PCA-rekonstruktio

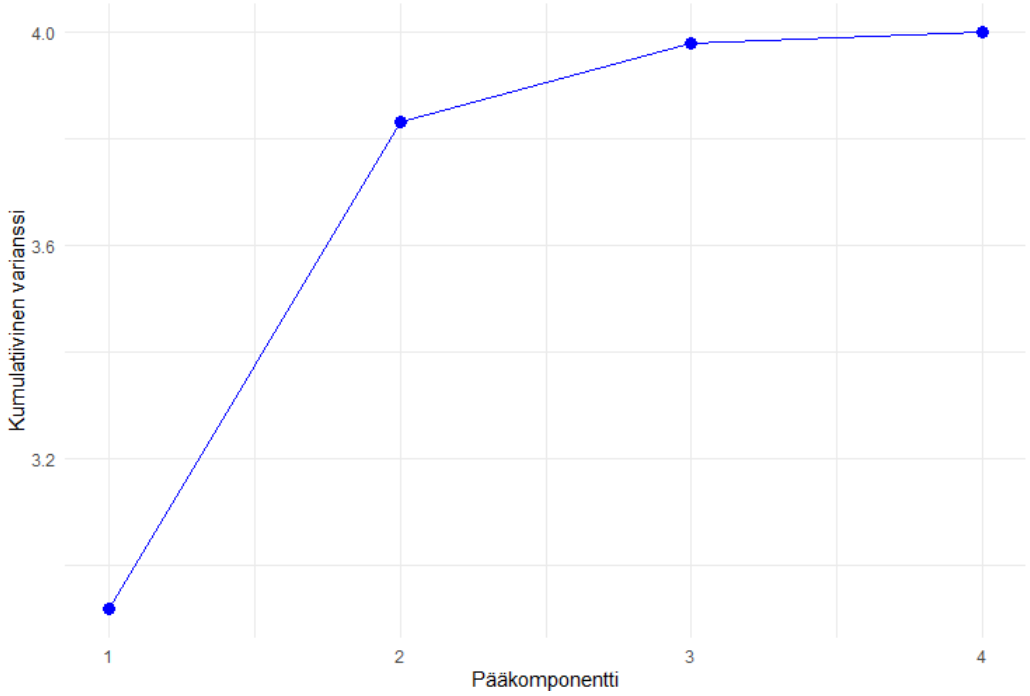
Pääkomponenttien avulla alkuperäiselle datalle voidaan muodostaa approksimaatio, niin kutsuttu PCA-rekonstruktio, jonka dimensiot vastaavat alkuperäisen datan dimensioita. Rekonstruktio pyrkii kuvaamaan alkuperäistä dataa mahdollisimman tarkasti, mutta kuitenkin siten, että siitä ilmenee vain haluttu osa sen vaihtelusta. Usein rekonstruktio muodostetaan muutamien ensimmäisten pääkomponenttien avulla siten, että suurin osa datan vaihtelusta säilyy. [9] PCA-rekonstruktio liitetään usein samaan kontekstiin kiinteiden vaikutusten mallin (*fixed effects model, FEM*) kanssa. [2]

Olkoon $\mathbf{X}$ $(n \times p)$-matriisi, jossa on n havaintoa ja p muuttujaa. Olkoon lisäksi $\mathbf{A}_M$ matriisi, jonka sarakkeet ovat matriisin $\mathbf{X}$ otoskovarianssimatriisin M ensimmäistä ominaisvektoria järjestyksessä suurimman ominaisarvon mukaan. Datapisteiden projektiot PCA-avaruuteen ovat pääkomponentteja $\mathbf{Z} = \mathbf{A}'\mathbf{X}$. Nämä projektiot voidaan palauttaa takaisin alkuperäisen aineiston avaruuteen kertomalla niitä ominaisvektoreilla

$$\hat{\mathbf{X}} = \mathbf{A}_M \mathbf{Z} = \mathbf{A}_M \mathbf{A}'_M \mathbf{X}. \tag{3.33}$$

Jos $M = p$, niin $\mathbf{A}_M \mathbf{A}'_M \mathbf{X} = I$ ja $\hat{\mathbf{X}} = \mathbf{X}$. [2]

Esimerkki 3.2. Lasketaan PCA-rekonstruktio taulukon 3.1 aineistolle korrelaatiomatriisia käyttäen. Kuvaajasta 3.3 nähdään, että kaksi ensimmäistä pääkomponenttia selittävät lähes kaiken esimerkkiaineiston kokonaisvaihtelusta. Kumulatiivisen varianssin perusteella olisi sopivaa käyttää kahta ensimmäistä pääkomponenttia aineiston analysointiin.



Kuva 3.3. *Kumulatiivinen varianssi pääkomponenttien funktiona taulukon 3.1 aineistolle.*

(a) *Approksimaatio Iris-aineistolle*

h	Sep.L (cm)	Sep.W (cm)	Pet.L (cm)	Pet.W (cm)
1	3.3399	4.0937	1.6599	1.8397
2	3.1530	3.5896	1.6644	1.8210
3	3.0807	3.7463	1.5416	1.7061
4	3.0808	3.6063	1.5913	1.7490
5	3.3193	4.1654	1.6154	1.7987
6	3.7047	4.4680	1.8671	2.0630
⋮	⋮	⋮	⋮	⋮
150	5.2315	3.1584	3.7520	3.877

(b) *Alkuperäinen Iris-aineisto*

h	Sep.L (cm)	Sep.W (cm)	Pet.L (cm)	Pet.W (cm)
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
⋮	⋮	⋮	⋮	⋮
150	5.9	3	5.1	1.8

Taulukko 3.2. *Iris-aineistolle muodostettu approksimaatio kahden ensimmäisen pääkomponentin mukaan taulukossa 3.2a ja sen alkuperäinen versio taulukossa 3.2b.*

Taulukossa 3.2a on kaavalla 3.33 laskettu approksimaatio Iris -aineistolle 3.2b käyttäen kahta ensimmäistä pääkomponenttia. Ensimmäisten rivien perusteella approksimaatio näyttäisi kuvaavan Iris-dataa melko hyvin.

3.3 Pääkomponenttien määrän valinta

Aineiston dimensioiden määrän pienentäminen pääkomponenttianalyysin avulla tarkoittaa käytännössä pääkomponenttien määrän pienentämistä. Lopulliseen analyysiin otetaan mukaan vain tietty määrä – usein M ensimmäistä – pääkomponenttia. Yksi yksinkertainen tapa valita M on tehdä se siten, että pääkomponenttien kumulatiivinen osuus kokonaisvarianssista on riittävän suuri. Tällä tarkoitetaan sitä, että valittujen pääkomponenttien varianssien summa suhteessa kokonaisvarianssiin on riittävän iso. Jos valitaan tarkasteluun kaikki pääkomponentit, kumulatiivisen varianssin osuus kokonaisvarianssista on 1. Kumulatiivista varianssia on havainnollistettu kuvassa 3.3. M ensimmäisen pääkomponentin kumulatiivinen prosenttiosuus kokonaisvarianssista voidaan laskea kaavalla

$$t_M = 100\frac{\sum_{k=1}^{m} l_k}{\sum_{k=1}^{p} l_k}, \tag{3.34}$$

missä l_k on k. pääkomponentin varianssi. [28]

Lähteessä [28] mainitaan, että hyvä tulos prosenttiosuudeksi on välillä 70 % ja 90% riippuen kuitenkin paljon sovellusalasta. Joskus voi olla hyväksyttävää valita vähemmän komponentteja. Yleisesti ottaen paras arvo prosenttiosuudelle laskee, kun muuttujien tai havaintojen määrä kasvaa. Jos muuttujien määrä on hyvin suuri, korkea prosenttiosuuden arvo tarkoittaa sitä, että analyysiin otettavien pääkomponenttien määrä on epäkäytännöllisen suuri. Tällöin pääkomponentteja on suotavaa ottaa vähemmän.

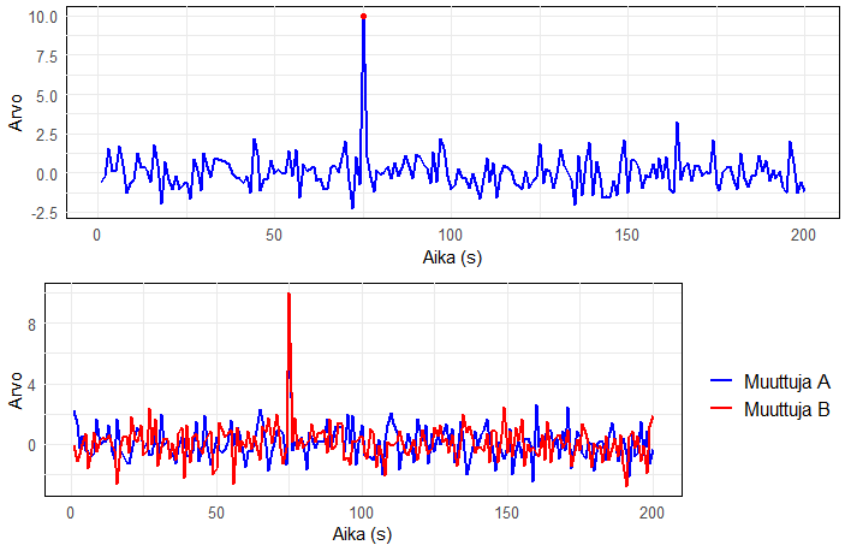
Toinen tapa valita pääkomponenttien määrä on katsoa pääkomponentteja vastaavien varianssien kuvaajaa. Kuvaajasta pyritään valitsemaan sellainen piste, joka erottaa vasemmalle puolelle jyrkän käyrän ja oikealle puolelle tasaisen käyrä eli niin sanottu *kulmapiste*. Tämä piste vastaa pääkomponenttien määrää. Oikealle puolelle jäävä käyrä ei välttämättä ole horisontaalinen, mutta selkeästi suorempi, kuin vasemmalle jäävä käyrä. Ensimmäinen piste suoralla käyrällä on viimeinen analyysiin mukaan otettava pääkomponentti.[28] Edellä kuvatusta menetelmästä käytetään tässä työssä nimitystä *kulmapistemenetelmä*

Lisää tapoja valita pääkomponenttien määrä esitetään lähteessä [28]. Näitä ovat muun muassa ristiinvalidointiin, hypoteesitestaukseen tai osittaisten korrelaatioiden tarkasteluun perustuvat menetelmät.

3.4 Poikkemien havaitseminen

Poikkeavien havaintojen *(engl. outlier)* tunnistaminen on ollut jo pitkään mielenkiintoinen ongelma, sillä muusta aineistosta eroavat pisteet ovat usein joko hyvällä tai huonolla tavalla erityisiä. Poikkeaville havainnoille ei ole olemassa tarkkaa määritelmää, mutta usein niiden ajatellaan olevan sellaisia havaintoja, jotka jollain merkittävällä tavalla eroavat suurimmasta osasta muita havaintoja. Sellaiselle aineistolle, jossa on p muuttujaa, poikkeavat havainnot ovat kaukana p-dimensionaalisen avaruuden muista pisteistä. Tällaiset havainnot saattavat vaikuttaa tai olla vaikuttamatta pääkomponenttianalyysin tuloksiin, riippuen niiden sijainnista mittaussarjassa. Usein onkin hyödyllistä määritellä, mitkä poikkeamat ovat niitä niin kutsuttuja *vaikuttavia havaintoja* (engl. influential observations).

Kuvassa 3.4 on kaksi satunnaisesti generoitua aikasarjaa, joista molemmista voidaan huomata selkeästi muista pisteistä poikkeava havainto noin 75 sekunnin kohdalla. Ylempi kuva demonstroi yhden ja alempi kahden muuttujan aikasarjaa.



Kuva 3.4. *Kaksi satunnaisesti generoitua aikasarja, jotka sisältävät poikkeavan havainnon. Ylemmässä aikasarjassa on yksi ja alemmassa kaksi muuttujaa.*

Poikkeamien tunnistus -ongelma voidaan määrittää esimerkiksi tilastollisena hypoteesitestinä tai luokitteluongelmana. Tilastollisessa testissä nollahypoteesinä olisi, että jokin havainto noudattaa tietyn populaation tai otoksen jakaumaa. Vaihtoehtoinen hypoteesi taas väittää, että havainto poikkeaa kyseisen populaation tai otoksen jakaumasta. Testi i. havainnon poikkeavuudelle voidaan muotoilla seuraavasti:

$$f(x) = \begin{cases} \mathcal{H}_0 : x_i \sim f_0 \\ \mathcal{H}_1 : x_i \nsim f_0, \end{cases}$$

missä x_i on jokin havainto datassa ja f_0 on populaation jakauma. Kun halutaan kontrolloida Tyypin I virheiden lukumäärää, voidaan testille määrittää merkitsevyystaso α:

$$p(x_i) = \mathbb{P}(\text{reject } \mathcal{H}_0 | \mathcal{H}_0 \text{ is true }) \leq \alpha, \tag{3.35}$$

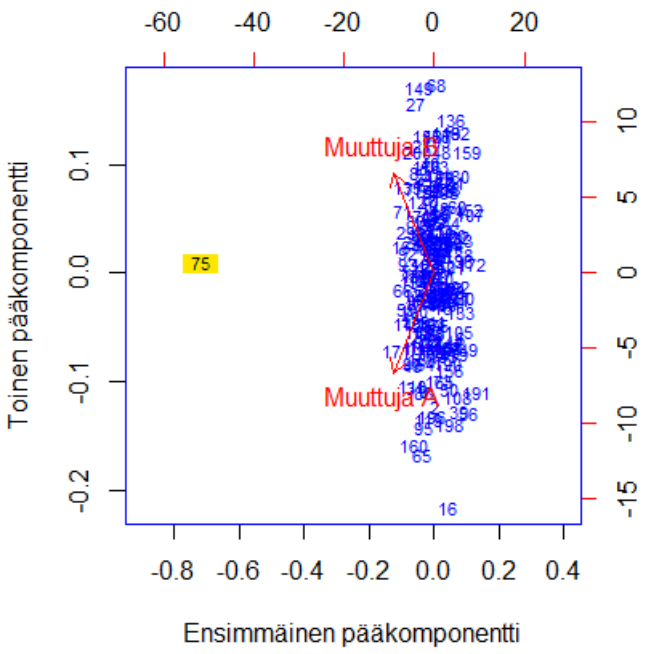
missä $p(x_i)$ on todennäköisyys sille, että havainto x_i on poikkeama. [13]

Poikkeamien tunnistus-ongelma voidaan määritellä myös yksiluokkaisena luokitteluongelmana, jossa pyritään selvittämään, kuuluuko tietty havainto poikkeamien luokkaan vai ei. Toisaalta voidaan tarkastella sitä, kuuluuko havainto populaation jakaumaa noudattavien havaintojen joukkoon vai ei.

3.4.1 Poikkeamat Biplot-kuvaajassa

Eräs yksinkertainen tapa tunnistaa poikkeavia havaintoja aineistosta on tarkastella muutamaa ensimmäistä tai viimeistä pääkomponenttia. Ensimmäisissä pääkomponenteissa näkyvät usein sellaiset poikkeamat, jotka vaikuttavat merkittävästi koko datan vaihteluun. Sen sijaan sellaiset poikkeamat, joilla ei ole suurta vaikutusta aineiston kokonaisvarianssiin, tulevat tavallisesti esiin vasta myöhemmissä pääkomponenteissa. Tarkastellaan tietojoukkoa, joka sisältää useita ihmisen terveyteen liittyviä ominaisuuksia, kuten ikä, paino, sukupuoli ja niin edelleen. Suurin osa tietojoukon tapauksista edustaa terveitä yksilöitä, joiden terveyteen liittyvät arvot ovat samassa linjassa. Kuitenkin tietojoukossa on pieni määrä yksilöitä, jotka ovat jonkin ominaisuuden osalta erilaisia kuin muut. Nämä poikkeavat yksilöt näkyvät todennäköisesti vasta myöhemmissä komponenteissa, sillä yksi pieni poikkeava joukko yhdessä muuttujassa ei edusta kovinkaan hyvin aineiston kokonaisvaihtelua. [28]

Poikkeamia ensimmäisten tai viimeisten pääkomponenttien suhteen voidaan tunnistaa muun muassa luvussa 3.2.1 esitellyn biplotin avulla. Jos jokin havainnoista poikkeaa selvästi kuvaajan muista pisteistä, on kyseessä poikkeama. Kuvassa 3.5 on biplot kuvan 3.4 kahden muuttujan aikasarjalle. Biplotista nähdään, että havaintopiste 75 sekunnin kohdalla eroaa selvästi muista pisteistä esimmäisen pääkomponentin suunnassa. Myös havainto 16 eroaa hiukan muista pisteistä toisen pääkomponentin suhteen, mutta muuten pisteet näyttävät sijoittuvan samaan rykelmään.



Kuva 3.5. *Kaksi ensimmäistä pääkomponenttia kahden muuttujan satunnaiselle aikasarjalle, joka on esitettynä kuvassa 3.4.*

Vaikka biplotin käyttö onkin joskus hyvin informatiivinen ja käytännöllinen tapa tunnistaa poikkeamia, vaatii se tietynlaiset olosuhteet luotettavan analyysin saavuttamiseksi. Biplotin käyttöä poikkeamien havaitsemiseksi voi rajoittaa erityisesti sen vaikeaselkoisuus ja pisteiden päällekkäisyys korkeaulotteisten aineistojen tapauksissa. Lisäksi kohina aineistossa saattaa vaikuttaa negatiivisesti poikkeamien näkyvyyteen yksittäisissä pääkomponenteissa.

3.4.2 Mahalanobiksen etäisyys

Mahalanobiksen etäisyyttä 2.19 voidaan hyödyntää myös PCA-analyysissä, kun halutaan löytää mahdollisia poikkeamia datasta. Se mittaa havaintojen poikkeavuutta otoksen keskiarvosta ottaen huomioon sekä niiden sijainnin pääkomponenttiavaruudessa että muuttujien väliset korrelaatiot. Käytännössä Mahalanobiksen etäisyys lasketaan projisoimalla datapiste pääkomponenttiavaruuteen ja arvioimalla sen etäisyys pääkomponenttien jakauman keskiarvosta. Suuret Mahalanobiksen etäisyydet viittaavat havaintoihin, jotka eroavat merkittävästi normaalista datan varianssista ja joiden voidaan olettaa olevan potentiaalisia poikkeamia.

Mahalanobiksen etäisyys i. havainnon ja otoskeskiarvon välillä on $(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$, missä $\mathbf{S}$ on otoskovarianssimatriisi ja $\overline{\mathbf{x}}$ otoskeskiarvo. Pääkomponenttien avulla tämä voidaan laskea kaavalla

$$D_i^2 = \sum_{k=1}^{p} \frac{z_{ik}^2}{l_k}, \tag{3.36}$$

missä D_i on Mahalanobiksen etäisyys *i.* havainnon ja otoskeskiarvon välillä, p on alkuperäisten muuttujien lukumäärä, z_{ik} on *k.* PCA-komponentin arvo *i.* muuttujalle ja l_k on *k.* pääkomponenttia vastaava varianssi. Viimeksi mainittu väite seuraa siitä, että otoskovarianssimatriisille $\mathbf{S}$ pätee $\mathbf{S} = \mathbf{A}\mathbf{L}^2\mathbf{A}'$ (ominaisarvohajotelma), missä $\mathbf{L}^2$ on diagonaalimatriisi, jonka *k.* diagonaalialkio on ominaisarvo l_k ja $\mathbf{A}$ on ominaisvektoreista a_k koostuva matriisi. Lisäksi $\mathbf{S}^{-1} = \mathbf{A}\mathbf{L}^{-2}\mathbf{A}'$, $\mathbf{x}_i' = \mathbf{z}_i'\mathbf{A}'$ ja $\overline{\mathbf{x}}' = \overline{\mathbf{z}}'\mathbf{A}'$, jolloin etäisyyden neliölle D_i^2 pätee

$$\begin{aligned}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) &= (\mathbf{z}_i - \overline{\mathbf{z}})'\mathbf{A}'\mathbf{A}\mathbf{L}^{-2}\mathbf{A}'\mathbf{A}(\mathbf{z}_i - \overline{\mathbf{z}}) \\ &= (\mathbf{z}_i - \overline{\mathbf{z}})'\mathbf{L}^{-2}(\mathbf{z}_i - \overline{\mathbf{z}}) \\ &= \sum_{k=1}^{p} \frac{z_{ik}^2}{l_k}.\end{aligned}$$

jos halutaan laskea Mahalanobiksen etäisyys vain r ensimmäisen pääkomponentin avulla, missä $r < p$, kaava saadaan muotoon

$$D_i^2 = \sum_{k=1}^{r} \frac{z_{ik}^2}{l_k}, \tag{3.37}$$

3.4.3 Hotellingin T2 -testi

Hotellingin T^2 -testi (*Hotelling's T-squared test*) on tilastollinen menetelmä, jolla pyritään havaitsemaan poikkeamia moniulotteisissa datamatriiseissa. Testi laajentaa perinteistä yhden muuttujan T-testiä *(T-test)* [38] kahden tai useamman toisistaan riippuvan muuttujan avaruuteen. [27]

Edellä mainittu testi perustuu Hotellingin T^2 -jakaumaan, joka on monen muuttujan todennäköisyysjakauma. Jos jokin vektori d on normaalijakautunut ja M on $p \times p$ yksikkömatriisi, joka noudattaa Wishartin jakaumaa vapausasteella m, niin silloin matriisin M neliöllinen muoto noudattaa Hotellingin T^2 -jakaumaa ja sitä kautta myös F-jakaumaa parametreilla $F_{p,m-p+1}$. [56]

Hotellingin T^2 -testissä testisuureena käytetään edellisessä kappaleessa 3.4.2 määriteltyä Mahalanobiksen etäisyyttä. Otoskovarianssimatriisin muoto $(n-1)\mathbf{S}$ noudattaa Wishartin jakaumaa, parametreilla p ja $n-1$. Otoskeskiarvon otoskovarianssimatriisi on $\mathbf{S}/n$ esti on erityisen hyödyllinen silloin, kun halutaan selvittää, poikke-

aako tietyt havainnot tilastollisesti merkitsevästi datan yleisestä keskiarvosta. Hotellingin T^2-testimuuttuja voidaan ilmaista muodossa

$$T^2 = (\overline{x} - \mu)'(S/n)^{-1}(\overline{x} - \mu), \tag{3.38}$$

missä $\overline{x}$ on otoskeskiarvo, $\mathbf{S}$ on otoskovarianssimatriisi ja μ on populaation jakauman odotusarvo. Huomataan, että yhtälö 3.38 on oikeastaan Mahalanobiksen etäisyys, joka esiteltiin edellisessä luvussa. Testin $p-arvo$ saadaan siis $F_{p,n-p}$-jakaumasta [39]

Hotellingin T^2 - testi asettaa tiettyjä oletuksia testattavalle aineistolle. Ensinnäkin sen havaintojen tulee olla otos multinormaalijakautuneesta popuulaatiosta, jolla on odotusarvovektori μ ja kovarianssimatriisi Σ. Lisäksi havaintojen tulee olle riippumattomia keskenään. [44][49] Oletetaan, että edellä määritellyt oletukset toteutuvat. Tällöin Hotellingin T^2 -testiä voidaan soveltaa PCA-analyysiin, kun lasketaan testisuureita pääkomponenttiavaruudessa. Suuri Hotellingin T^2 -arvo osoittaa, että tietty havainto poikkeaa merkittävästi pääkomponenttitilassa määritetystä keskiarvosta. Tämä voi viitata poikkeamiin eli havaintoihin, jotka eivät ole yhteensopivia datan yleisen rakenteen kanssa.

T^2-testisuure i. havainnolle on pääkomponenttien avulla ilmaistuna

$$T_i^2 = \sum_{k=1}^{r} \frac{z_{ik}^2}{l_k}, \tag{3.39}$$

missä $r \leq p$, p on alkuperäisten muuttujien lukumäärä, z_{ik} on k. PCA-komponentin arvo i. muuttujalle ja l_k on k. pääkomponenttia vastaava varianssi. [43]

3.4.4 PCA-sovitus

Eräs keino löytää poikkeamia aineistosta on laskea residuaali alkuperäisten ja pääkomponenttien avulla estimoitujen havaintojen välillä. Jos saatavilla on riittävästi harjoitusdataa, joka ei sisällä merkittäviä poikkeamia, voidaan PCA-sovitus havainnolle tehdä tästä harjoitusdatasta muodostettujen pääkomponenttien avulla. Tuntemattoman havainnon ja sen sovituksen välistä erotusta tarkastelemalla voidaan selvittää kuinka paljon uusi havainto poikkeaa harjoitusdatasta. Jos erotus on suuri, kyseessä on mahdollinen poikkeama. Edellä kuvattua menetelmää poikkeamien havaitsemiseen on esitelty muun muassa lähteissä [10][21][52].

4. AINEISTO JA MENETELMÄT

4.1 Ympäristön gammaspektrometriset säteilymittaukset

Ympäristössä esiintyvät radionuklidit ovat ionisoivaa säteilyä lähettäviä radioaktiivisia aineita, joita muodostuu joko luonnostaan tai ihmisen toiminnan seurauksena. Suurin osa näistä nuklideista on peräisin luonnollisista lähteistä, kuten maaperässä olevista primordiaalisista radionuklideista tai kosmisen taustasäteilyn seurauksena syntyvistä kosmogeenisista radionuklideista. Näiden nuklidien lähettämää säteilyä kutsutaan *luonnon taustasäteilyksi* ja sen joukosta on oleellista tunnistaa ihmisen toiminnan seurauksena syntyneitä keinotekoisia radionuklideja. [47, s.12]

Keinotekoisia radionuklideja voi päätyä ympäristöön esimerkiksi ydinasekokeilujen, ydinvoimaloiden tai lääketieteellisen käytön seurauksena. Tsernobylin ydinvoimalaonnettumuus (1986) ja kylmän sodan (1950-1991) aikaan tehdyt ydinasekokeet aiheuttivat radioaktiivisten aineiden laskeumaa ympäri maailmaa. Ihmislähtöiset radionuklidit aiheuttavat alle viidesosan suomalaisen keskimääräisestä vuosittaisesta säteilyannoksesta. [47, s. 28]

Ympäristön radioaktiivista säteilyä voidaan mitata muun muassa *tuikeilmaisimilla*. Tuikeilmaisimien käyttö perustuu *tuikeaineeseen* (skintilaattori) ja *valomonistinputkeen*. Tuikeaineet voivat olla joko kiinteää, nestettä tai kaasua, ja eri tuikeaineita yhdistämällä voidaan samanaikaisesti mitata eri säteilylajeja. Gammasäteilyn energian ja sen intensiteetin mittaamiseen käytetään usein epäorgaanisia tuikeaineita. Säteilyenergian absortoituminen virittää tuikeaineen atomin energiatilalle, jonka purkautuessa emittoituvaa valoa voidaan mitata. Mittaus saadaan muunnettua digitaaliseen muotoon käyttämällä spektrometriä ja gammasäteilyn kohdalla tarkemmin gammaspektrometriä. Mitatun gammaspektrin digitaalista muotoa kutsutaan *pulssinkorkeusjakaumaksi*. Kun näitä mittauksia tehdään peräkkäin jollain aikavälillä, saadaan useista spektreistä muodostuva *aikasarja*. [33]

Ulkoisen säteilyn annosnopeutta mitataan Suomessa jatkuvasti ulkoisen säteilyn valvontaverkon avulla. Verkko koostuu noin 260:sta eri puolilla maata sijaitsevasta mittausasemasta. Niiden säteilyanturit ovat suojakuorien sisällä olevia geigerputkia, joiden mittausalue kattaa annosnopeudet 0.01 μSv/h – 10 Sv/h. Osa asemista

on varustettu $LaBr_3(Ce)$ -spektrometrillä. [5, s. 342]

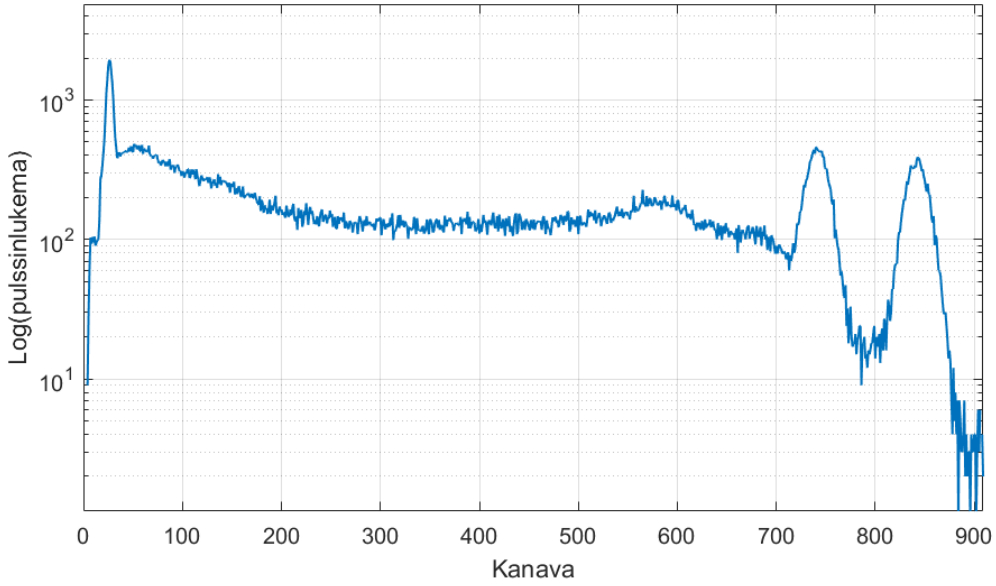
4.1.1 Keinotekoisten radionuklidien havaitseminen spektristä

Jokainen gammasäteilyä lähettävä radionuklidi synnyttää gammaspektriin sille tyypillisen *vasteen*, joka ilmenee spektrissä esimerkiksi yhtenä tai useampana *fotopiikkinä* tai *Compton-reunana* ja *sironneen säteilyn jatkumona*. Fotopiikin sijainti spektrissä määräytyy emittoituvan gammafotonin energiasta ja sen pinta-ala puolestaan säteilyn intensiteetistä ja ilmaisimen tehokkuudesta. [5] Compton-reunan energia ja takaisinsirontareunan energia saadaan Comptonin yhtälöstä ja jatkumon muoto puolestaan Kleinin ja Nishinan yhtälöstä [48].

Vaste on spektrometrin mittaama kullekin radionuklidille ominainen gammasäteilyn energiajakauma. Näitä jakaumia analysoimalla ja pulssien lukumääriä tutkimalla ilmaisimeen tulleen säteilyn lähteenä oleva nuklidi voidaan tunnistaa. [5] Kuvassa 4.1 on nuklidin koboltti-60 gammaspektri. Kuvaajassa vaaka-akselilla on energiakanavat ja pystyakselilla kanavia vastaavien pulssinlukemien logaritminen muoto. Kukin kanava vastaa jotakin tiettyä energia-aluetta ja tälle energia-alueelle osuneet fotonit kasvattavat kanavan pulssinlukemaa. Kanavien energia-alueet ovat nousevassa järjestyksessä ja jokainen energia-alue on yhtä pitkä. Koboltti-60 lähettää hajotessaan kaksi gammafotonia, joiden energiat ovat 1173 keV 1332 keV. Nämä fotonit näkyvät tässä gammaspektrissä fotopiikkeinä kanavien 700 ja 900 välissä. Spektrissä on myös molempia gammaemissioita vastaavat Compton-reunat ja sironneen säteilyn jatkumot, joista ainakin toiset voidaan nähdä selkeästi noin kanavan 600 ja sitä edeltävien kanavien kohdalla.

Säteilytilanteen muutoksen havaitsemiseen käytettyjä algoritmeja on erilaisia, mutta useimmat niistä käyttävät jotakin ennakkotietoa johtopäätösten tekemiseen. Jotkut niistä analysoivat mittauksen laskentataajuutta (*engl. cross-count-rate*), eli ilmaisimeen osuneiden ja havaittujen fotonien kokonaismäärää aikayksikköä kohden. Nämä algoritmit käyttävät hälytysrajan määrittämiseen taustasäteilystä määritettyä tarkkaa pulssien laskentataajuutta. Tällöin mikä tahansa hälytysrajan ylittävä säteily johtaa hälytykseen, jolloin intensiteettitason nousut taustasäteilyssä aiheuttavat järjestelmään runsaasti vääriä hälytyksiä.

Toiset algoritmit analysoivat tietylle ennakkoon määritetylle energia-alueelle tulevien pulssien lukumäärää. Tällaisen menetelmän käyttäjältä odotetaan tietämystä siitä, mitä radionuklidia halutaan etsiä, joten metodi ei ole kovinkaan yleispätevä. Lisäksi ilmaisimeen ja sen lähistölle tulleet fotonit siroavat ilmassa tai muussa väliaineessa olevista aerosolihiukkasista ja säteilyn ollessa tarpeeksi intensiivistä sironnut säteily aiheuttaa pulssinkorkeusjakaumaan laajaankin energia-alueeseen vaikuttavan jatkumon. Tällöin ROI:n (*Region Of Interest*) valinta nuklidin aiheuttamien piikkien



Kuva 4.1. *Nuklidin koboltti-60 gammapspektri. Kanavien 700 ja 900 välissä voidaan nähdä kaksi fotopiikkiä ja aiemmilla kanavilla sironneen säteilyn jatkumo, jonka kulma on noin kanavan 600 kohdalla.*

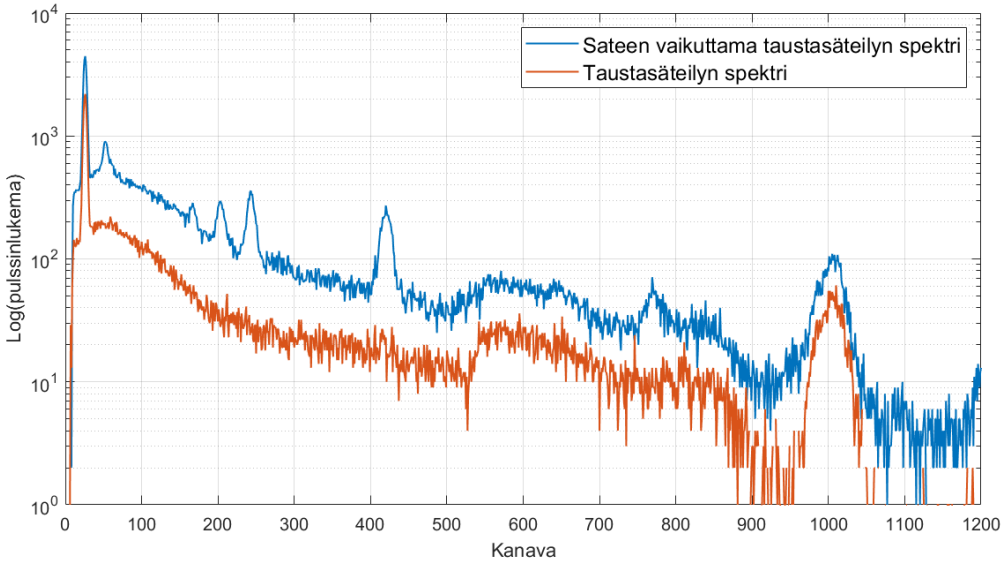
perusteella jättää suuren osan mielenkiintoisesta spektristä huomiotta.

Jotkin algoritmit käyttävät puolestaan koko gammasäteilyn energiajakaumaa poikkeamien havaitsemiseen. Nämä algorimit pyrkivät usein sovittamaan tunnettuja radionuklidien energiajakaumia mitattuun spektriin, jolloin metodi on hyvin riippuvainen siitä, kuinka tarkasti jonkin tietyn nuklidin teoreettinen jakauma vastaa sen säteilystä mitatun spektrin jakaumaa. Usein mitatun spektrin muodostumiseen vaikuttavat monet muutkin tekijät, kuten mittausympäristö, mittalaitteen tarkkuus tai ilmaisimen ja säteilylähteen väliset säteilyä absorboivat aineet. Tällöin mitattu spektri saattaa olla hyvinkin monimutkainen verrattuna teoreettiseen tulokseen. [20]

4.1.2 Luonnon taustasäteilyn spektrit

Luonnon taustasäteilyn lähteet tuottavat mittauksissa aina samankaltaisen spektrin, mutta niissä voi esiintyä kuitenkin luonnon sääilmiöiden, kuten sateen tai lumipeitteen, aiheuttamaa vaihtelua. Vaihtelua sisältävät spektrit vaikeuttavat usein keinotekoisten nuklidien havaitsemista aiheuttamalla systeemiin vääriä hälytyksiä tai peittämällä alleen merkityksellisiä havaintoja. Taustasäteilyn merkittävimmät lähteet ovat kalium-40, torium-232, uraani-235, uraani-238 sekä toriumin ja uraanin pitkien hajoamisketjujen eri nuklidit. Sateen aikana uraanin ketjuun kuuluvan radonin hajoamistuotteita kulkeutuu veden mukana ilmasta maan pinnalle, jolloin taustasäteilyn taso nousee hetkellisesti tietyillä energia-alueilla.

Kuvassa 4.2 on kaksi $LaBr_3(Ce)$ -ilmaisimella mitattua luonnonsäteilyn spektriä. Oranssi spektri on tavanomaisen taustasäteilyn mittauksesta muodostettu spektri ja sininen spektri on sateen vaikuttama luonnon taustasäteilyn spektri. Sinisen spektrin piikit energiakanvien 150–270 ja 400–450 alueilla ovat syntyneet nuklidin uraani-238 hajoamisketjuun kuuluvien tytärnuklidien lyijy-212, lyijy-214 sekä vismutti-214 säteilystä, joita on sateen mukana kulkeutunut ilmasta lähelle maan pintaa. Sateen vaikuttaman spektrin pulssinlukemat ovat lisäksi kauttaaltaan suuremmat kuin tavanomaisessa spektrissä, mikä on myös seurausta sateen hetkellisesti nostattamasta taustasäteilyn tasosta. [47]



Kuva 4.2. *Kaksi luonnon taustasäteilystä mitattua gammaspektriä. Sinisessä spektrissä esiintyy sateen aiheuttamia energiapiikkejä nuklidin uraani-238 hajoamisketjuun kuuluvien nuklidien energia-alueilla.*

4.2 Aineisto

Työn aikana kehitetyn menetelmän testaamiseksi on kerätty ja valmisteltu mittausaineistoa, joka pyrkii mahdollisimman hyvin kattamaan erilaiset säteilytapaukset. Luonnon taustasäteilyä on kerätty Suomessa sijaitsevan ulkoisen säteilyn valvontaverkon [7] mittausasemilta, Nuorgamista ja Rovaniemeltä. Ilmaisimien sijainnit ja aineistojen mittausaika on pyritty valitsemaan niin, että mittausdatan seassa on sekä sateisia että poutaisia päiviä. Lisäksi lumipeitteen vaikutus taustasäteilyyn on pyritty ottamaan huomioon valitsemalla mukaan päiviä myös talvisilta kuukausilta. Näiden lisäksi on kerätty mittausdataa Harjavallasta ja Kotkasta, aikajaksoilta, joissa tiedetään olleen mukana poikkeavaa säteilyä.

Jokainen aineisto koostuu mitattujen spektrien aikasarjasta. Kukin spektri sisältää 2048 kanavaa, joita vastaa jokin pulssinlukema. Spektreistä on poistettu alusta 41

kanavaa, sillä analyysissä huomattiin, että alkupään epäoleelliset kanavat aiheuttivat mittausvirheiden vuoksi vääristymiä tuloksiin.

Taustasäteilyaineistojen pohjalta Rovaniemen ja Nuorgamin aineistoille on tuotettu testiaineistoa lisäämällä spektreihin keinotekoisten nuklidien vasteita. Vasteet on luotu synteettisesti spektrianalyysiohjelmistolla. Ohjelmisto lisää vasteeseen fotopiikit tunnettujen gammafotonien energioiden sekä käyttäjän syöttämän referenssiemission piikin pinta-alan avulla. Referenssiemissioksi on kaikissa tapauksissa valittu se emissio, jota nuklidi lähettää suurimmalla todennäköisyydellä. Vasteisiin lisätään myös Compton-reuna ja sironneen säteilyn jatkumo, jotka saadaan Comptonin yhtälön sekä Klein-Nishina -yhtälön avulla. Vasteissa on huomioitu lisäksi kutakin ilmaisinta vastaavat energia- ja tehokkuuskalibraatiot [12][16].

Testeissä käytetyt keinotekoiset nuklidit, joita ei esiinny luonnossa ilman ihmisen toimintaa, ovat amerikium-241, koboltti-60, cesium-134, cesium-137, jodi-131, ksenon-133 ja ksenon-135. Lisäksi jokaiselle testinuklidille on määritelty kolme eri aktiivisuustasoa, jotka vastaavat referenssiemission pinta-alalla määritetyn vasteen aktiivisuutta. Kullekin nuklidille on käytetty referenssiemission piikin pinta-alana lukuja 200, 400 ja 600. Pinta-alojen yksikkö on $kanava \cdot pulssinlukema$. Kaikkien aineistojen testinuklidien aktiivisuudet on määritetty samojen referenssipinta-alojen avulla. Tällöin eri aineistojen testinuklidien aktiivisuuksissa voi olla pientä eroavaisuutta, sillä eri ilmaisimilla on omat tehokkuus- ja energiakalibraatioiden yhtälönsä.

4.2.1 $LaBr_3$(Ce), Nuorgam

Nuorgamissa sijaitsevalla $LaBr_3$(Ce) 1.5” -tuikespektrometrillä mitattua luonnon taustasäteilyn spektridataa on kerätty yhden kokonaisen vuoden (2022) ajalta käsittäen 56 691 erillistä spektriä. Aineistoon on pyritty sisällyttämään säävaihteluita koko vuoden ajalta, jotta taustasäteilyn vaihtelut saataisiin mahdollisimman kattavasti testeihin mukaan. Aikasarja sekoitetaan satunnaiseen järjestykseen, jotta harjoitus-, validointi- ja testidataan saadaan mahdollisimman kattava joukko erilaisia mittauksia. Sekoitettu data jaetaan siten, että 50% siitä nimetään testidataksi, 35% harjoitusdataksi ja 15% validointidataksi. Testiaineistosta 210 spektriin lisätään vaste, 10 vastetta kullekin taulukon 4.1 radionuklidin eri aktiivisuuksille. Saman nuklidin samalle aktiivisuudelle lisätään vaste useampaan eri spektriin, jotta yksittäisen spektrin vaikutus poikkeaman havaitsemiseen saadaan huomioitua.

Taulukko 4.1. *Nuklidit, joille generoitu $LaBr_3(Ce)$-spektrometrin vaste on lisätty Nuorgamin testispektreihin sekä niiden aktiivisuudet.*

Nuklidi	Aktiivisuus 1 ($\frac{Bq}{m^3}$)	Aktiivisuus 2 ($\frac{Bq}{m^3}$)	Aktiivisuus 3 ($\frac{Bq}{m^3}$)
^{241}Am	49.4 ± 38.5%	98.8 ± 21.4%	148.2 ± 16.3%
^{60}Co	29.5 ± 12.1%	59.0 ± 11.1%	88.5 ± 10.7%
^{134}Cs	20.6 ± 15.3%	41.2 ± 12.0%	61.8 ± 11.2%
^{137}Cs	24.9 ± 14.9%	49.8 ± 11.9%	74.7 ± 11.1%
^{131}I	17.8 ± 17.6%	35.7 ± 12.8%	53.5 ± 11.6%
^{133}Xe	39.3 ± 40.5%	78.7 ± 22.3%	118.1 ± 16.9%
^{135}Xe	13.5 ± 21.8%	27.0 ± 14.6%	40.5 ± 12.7%

4.2.2 $LaBr_3(Ce)$, Rovaniemi

Rovaniemen aineisto on saatu Rovaniemellä sijaitsevasta kerääjästä, jonka avulla voidaan mitata radioaktiivisten aineiden laskeumaa. Rovaniemen aineistossa luonnon taustasäteilyn spektridataa on Nuorgamin aineiston tavoin yhden kokonaisen vuoden ajalta käsittäen 56 335 erillistä spektriä. Myös Rovaniemem aikasarja sekoitetaan satunnaiseen järjestykseen ja sekoitettu aineisto jaetaan harjoitus-, validointi- ja testidatoihin samassa suhteessa, kuin Nuorgamin aineisto. Testiaineistosta 210 spektriin lisätään vaste, 10 vastetta kullekin taulukon 4.2 lähteen eri aktiivisuudelle.

Taulukko 4.2. *Nuklidiaktiivisuudet, joille generoitu $LaBr_3(Ce)$-spektrometrin vaste on lisätty Rovaniemen testispektreihin.*

Nuklidi	Aktiivisuus 1 ($\frac{Bq}{m^3}$)	Aktiivisuus 2 ($\frac{Bq}{m^3}$)	Aktiivisuus 3 ($\frac{Bq}{m^3}$)
^{241}Am	50.2 ± 25.6%	100.4 ± 15.8%	150.6 ± 13.1%
^{60}Co	30.0 ± 12.1%	60.0 ± 11.1%	90.0 ± 10.7%
^{134}Cs	20.9 ± 13.8%	41.9 ± 11.6%	62.8 ± 11.0%
^{137}Cs	25.3 ± 13.4%	50.6 ± 11.5%	76.0 ± 10.9%
^{131}I	18.1 ± 14.9%	36.2 ± 11.9 %	54.4 ± 11.2%
^{133}Xe	40.0 ± 26.1 %	80.0 ± 16.1%	120.0 ± 13.3%
^{135}Xe	13.7± 17.2 %	27.5 ± 13.0%	41.2 ± 11.9%

4.2.3 $LaBr_3(Ce)$, Kotka ja Harjavalta

PCA-algoritmia testattiin myös kahdelle oikeasti esiintyneelle keinotekoiselle säteilytapaukselle. Aineistot ovat peräisin Kotkasta ja Harjavallasta. Kotkan ilmaisimella on mitattu röntgensäteilyä ja aineistossa on yksi spektri, jossa se ilmenee jatkumona. Loput 25 516 spektriä on tavanomaista taustasäteilyä. Harjavallan ilmaisimella on puolestaan mitattu nuklidin jodi-131 gammasäteilyä, ja sitä esiintyy aineistossa tun-

nistettavasti neljässä eri spektrissä. Taustasäteilyn spektrejä Harjavallan aineistossa on 25 629 kappaletta.

4.3 Aiempia sovelluksia

Gammaspektrometristä säteilydataa on aikaisemminkin analysoitu pääkomponentteihin perustuvilla menetelmillä. Lähteessä [52] tutkitaan SVD-hajotelmaan perustuvan NASVD-menetelmän *(Noise-adjusted Singular Value Decomposition)* käyttöä gammaspektrometristen mittausten aikasarjojen analyysissä. Lähteessä selvitetään, miten NASVD-metodin avulla voidaan tunnistaa keinotekoisia säteilylähteitä luonnon taustasäteilyn seasta.

NASVD-menetelmä käyttää PCA-analyysistä tuttuja pääkomponentteja kohinan vähentämiseen aineistosta. Menetelmässä oleellista on, että se muodostaa pääkomponentit SVD-hajotelman avulla sekä pyrkii vähentämään aineistosta kohinaa. NASVD-menetelmän käytölle onkin ominaista se, että analysoitavan datan tiedetään sisältävän kohinaa ja sille kohinalle on olemassa jokin ennalta tiedetty malli – a priori -malli. A priori -mallia käytetään yleensä analysoitavan datan esikäsittelyssä. Yleensä edellä mainittu tarkoittaa sitä, että analysoitava data normalisoidaan jakamalla kukin havainto pienimmän neliösumman sovituksella kyseisen havainnon ja keskiarvoisen havainnon välillä ennen SVD-hajotelman muodostamista. [40][41] SVD-hajotelmasta saatujen pääkomponenttien avulla spektreille muodostetaan PCA-rekonstruktiot. Spektreistä saadaan vähennettyä kohinaa, kun rekonstruktiot muodostetaan vain muutamien ensimmäisten pääkomponenttien – niiden, jotka sisältävät mahdollisimman vähän kohinaa – avulla. Lähteessä [52] käytetään edellä kuvailtua NASVD-menetelmää ennen varsinaista nuklidin tunnistusalgoritmia.

Lähteessä [10] esitellään toinen pääkomponenttianalyysiä hyödyntävä menetelmä poikkeamien havaitsemiseen datasta. Menetelmän alku on hyvin samankaltainen edellä esitetyn NASVD-menetelmän kanssa, mutta SVD-hajotelman sijaan se käyttää datan korrelaatiomatriisin ominaisarvohajotelmaa pääkomponenttien muodostamiseen. Lisäksi datan normalisointi tehdään hieman eri tavalla. Tässä menetelmässä olennaista on, että pääkomponentit muodostetaan harjoitusdatasta, joka sisältää vain luonnon taustasäteilyä. Harjoitusdatassa on siis myös luonnon sääolosuhteista johtuvaa vaihtelua. Pääkomponenttien avulla uusi tuntematon spektri rekonstruoidaan niin, että rekonstruktio selittää mahdollisimman hyvin spektrin sisältämän taustasäteilyn. Rekonstruktiolle ja spektrille lasketaan Mahalanobiksen etäisyys ja ennalta määritetyn hälytysrajan perusteella päätetään, onko spektri poikkeava vai ei. Metodi mukailee luvussa 3.4.4 esiteltyä PCA-sovitukseen perustuvaa poikkeamien tunnistusmenetelmää.

4.4 PCA-algoritmi

Työn soveltavaa osuutta varten on kehitetty algoritmi, jolla pyritään havaitsemaan keinotekoisten nuklidien säteilyä gammaspektrien aikasarjoissa. Algoritmissa luodaan taustasäteilydatan pohjalta PCA-malli, jota hyödynnetään tuntemattoman havainnon luokittelussa. Algoritmissa on karkeasti jaoteltuna neljä erillistä vaihetta: aineiston esikäsittely, PCA-mallin luominen harjoitusdatan avulla, hälytysrajan laskeminen validointidatan avulla sekä mallin sovittaminen testidataan. Algoritmi ottaa syötteenä harjoitus-, validointi- ja testidatat ja hälyttää, jos testidatassa on mukana taustasäteilystä poikkeavia tapahtumia.

Algoritmin suorituskykyä voidaan testata esimerkiksi luvussa 4.2 esiteltyjen aineistojen avulla. Yksi tapa tarkastella algoritmin toimivuutta on selvittää sen havaitsemisprosentit keinotekoisille nuklideille. Lisäksi testiaineiston avulla voidaan tarkastella algoritmin sensitiivisyyttä ja spesifisyyttä jonkin parametrin suhteen. Sensitiivisyys ja spesifisyys lasketaan yhtälöiden 4.1 ja 4.2 avulla:

$$sensitiivisyys = \frac{TP}{TP + FN} \tag{4.1}$$

$$spesifisyys = \frac{TN}{TN + FP} \tag{4.2}$$

Edellä esitetyissä yhtälöissä käytettävät kirjainyhdistelmät merkitsevät aina jotakin tiettyä havaintojoukkoa. Sensitiivisyyden yhtälössä 4.1 TP on oikein määritettyjen positiivisten havaintojen joukko ja FN väärin määritettyjen negatiivisten havaintojen joukko. Positiiviset havainnot käsittävät tässä tapauksessa ne spektrit, joihin on lisätty jonkin keinotekoisen säteilyn vaste ja negatiiviset havainnot ne spektrit, joissa ei esiinny mitään taustasäteilystä poikkeavaa. Käytännössä TP kattaa siis kaikki keinotekoista säteilyä sisältävät spektrit, jotka algoritmilla havaitaan ja FN puolestaan kattaa kaikki keinotekoista säteilyä sisältävät spektrit, joita algoritmilla ei havaita. Spesifisyyden yhtälössä 4.2 TN on oikein määriteltyjen negatiivisten havaintojen joukko, eli kaikki ne pelkkää taustasäteilyä sisältävät spektrit, joista algoritmi ei aiheuta hälytystä. Joukko FP puolestaan kattaa kaikki taustasäteilyn spektrit, joista algoritmi virheellisesti hälyttää. [42]

Tässä työssä joukosta TP käytetään nimitystä *oikeat hälytykset* ja joukosta FP nimitystä *väärät hälytykset*. Lisäksi nimityksiä *oikea tausta* ja *väärä tausta* käytetään kuvaamaan joukkoja TN ja FN.

4.4.1 Aineiston esikäsittely

Työssä tutkittavan algoritmin suorituskykyä voidaan parantaa erilaisilla esikäsittelymenetelmillä. Helpoin tapa lyhentää algoritmin suoritusaikaa on vähentää spektrin kanavien määrää. Yhdellä vähennyskerralla spektrin vierekkäiset kanavat summataan yhteen, jolloin kanavien määrä puolittuu. Vähennyskertoja toistetaan niin monta kertaa, että algoritmin laskenta saadaan riittävän nopeaksi. Vierekkäisten kanavien summaamminen yhteen n kertaa ei vaikuta oleellisesti PCA-algoritmin tuloksiin, sillä spektrien hahmot säilyvät samoina.

Luonnon vaihtelun vuoksi spektrien intensiteettitasot voivat nousta tai laskea. Sen vaikutuksen minimoimiseksi spektreihin tehdään lisäksi niin sanottu *pienimmän neliösumman (PNS) -normalisointi*, minkä tavoitteena on saada spektrien muodot mahdollisimman lähelle toisiaan. PNS-normalisointi esitellään algoritmissa 1. Metodilla pyritään muokkaamaa spektrit sellaiseen muotoon, että niiden ja keskimääräisen spektrin välinen residuaali on mahdollisimman pieni. Spektrit halutaan mahdollisimman lähelle keskiarvoista spektriä, jotta taustasäteilyn intensiteettien vaihtelujen vaikutus pääkomponentteihin saadaan minimoitua.

Algorithm 1 PNS-normalisointi

1: annettu data
2: **for** spektri $\in$ *data* **do**
3: $\quad \min_{\alpha \in \mathbb{R}} \sum_i (\mu[i] - \alpha \times \text{spektri}[i])^2$
4: $\quad$ spektri $= \alpha \times$ spektri
5: **end for**

Ratkaistaan seuraavaksi algoritmissa 1 esiintyvä α. Derivoidaan ensin pienimmän neliösumman lauseke

$$\frac{d}{d\alpha} \sum_i (\mu_i - \alpha x_i)^2 = \sum_i 2x_i^2 \alpha - 2x_i \mu_i. \tag{4.3}$$

Kun asetetaan derivaatta nollaksi, saadaan

$$\sum_i 2x_i^2 \alpha - 2x_i \mu_i = 0 \tag{4.4}$$

$$\Leftrightarrow \alpha = \frac{\sum_i 2x_i \mu_i}{\sum_i 2x_i^2} \tag{4.5}$$

Lopuksi aineiston muuttujat keskitetään ja standardisoidaan kaavalla

$$x_j = \frac{x_j - \mu_j}{\sigma_j}, \tag{4.6}$$

missä x_j on aineiston j. muuttuja ja μ_j ja σ_j sitä vastaavat odotusarvo ja keskihajonta. Aineiston normalisoinnissa on tärkeää huomata, että muuttujien odotusarvot ja keskihajonnat lasketaan harjoitusdatan perusteella ja validointi- ja testidatat normalisoidaan niitä käyttäen. Testi- tai validointidatan hyödyntäminen keskiarvon tai keskihajonnan laskemisessa voisi johtaa virheellisiin tuloksiin.

4.4.2 PCA-mallin luonti harjoitusdatasta

Algoritmin toisessa vaiheessa käyttäjän valitsemasta harjoitusdatasta muodostetaan pääkomponenttianalyysin avulla projektiomatriisi

$$\mathbf{B} = \mathbf{A}_M \mathbf{A}'_M, \tag{4.7}$$

missä matriisin A sarakkeet ovat treenidatasta muodostetut pääkomponentit ja M on valittujen pääkomponenttien määrä. Projektiomatriisilla kertomalla saadaan laskettua halutulle spektrille PCA-sovitus yhtälön 3.33 tavoin.

Pääkomponentit päädyttiin muodostamaan tässä sovelluksessa singulaariarvohajotelman kautta, sillä sen käyttö kovarianssi- tai korrelaatiomatriisin ominaisarvohajotelman sijaan parantaa algoritmin tehokkuutta. Singulaariarvohajotelman muodostamisen asymptoottinen yläraja on yleisessä tapauksessa $O(mn^2)$ ja ominaisarvohajotelmalle vastaava on $O(n^3)$, kun m ja n ovat matriisin rivien ja sarakkeiden määrät [54]. Lisäksi singulaariarvohajotelmalla on enemmän ominaisuuksia syvempää analyysia varten, kuten myös luvussa 3.1.3 mainittiin. [28]

4.4.3 Hälytysrajan määritys

Hälytysraja poikkeavan havainnon luokittelua varten määritellään algoritmissa käyttäjän syöttämien väärien hälytysten todennäköisyyden ja validointidatan avulla. Spektrien luokittelussa käytetään euklidista etäisyyttä, joka lasketaan spektrin ja sen PCA-sovituksen välille. Jos etäisyys ylittää tämän luvun menetelmällä lasketun hälytysrajan, suoritetaan hälytys. Työn aikana tarkastellaan pääosin kahta etäisyyttä, joista toinen on tavallinen euklidinen etäisyys ja toinen siitä muunneltu versio. Muunnellussa versiossa neliöjuuren sisällä olevaan summaan otetaan mukaan vain niiden kanavien erotukset, joissa alkuperäisen spektrin pulssinlukema on suurempi kuin sovituksen pulssinlukema. Tässä työssä tavallista euklidista etäisyyttä merkitään tunnuksella d_e ja muunneltua versiota tunnuksella d_{em}. Näistä etäisyyksistä käytetään työn aikana yleisesti nimitystä *residuaali*.

Hälytysraja pyritään siis määrittämään validointidatan avulla. Ensin validointidatan spektreille lasketaan PCA-sovitukset kertomalla kutakin spektriä yhtälön 4.7 pro-

jektiomatriisilla $\mathbf{B}$. Seuraavaksi lasketaan jokaisen spektrin x_i ja sen PCA-sovituksen $\mathbf{B}x_i$ välinen euklidinen etäisyys, joka saadaan kaavalla

$$d_e = ||\mathbf{B}x_i - x_i||. \tag{4.8}$$

Lopuksi pyritään selvittämään euklidisten etäisyyksien jakauma, jonka avulla tiettyä väärien hälytysten todennäköisyyttä vastaava hälytysraja on helppo laskea.

Testiaineistojen validointidataa vastaaville residuaaleille sovitettiin erilaisia jakaumia MATLABin [25] valmiita funktioita käyttäen. Lähimpänä residuaaleja vaikutti olevan näistä jakaumista yleistetty ääriarvojakauma *(Generalized Extreme Value Distribution, GEVD)*, mutta senkin sopiessa huonosti joihinkin tapauksiin, jakaumaa alettiin etsiä teoreettisesti. Lopulta todistusluonnosten kautta residuaaleihin päädyttiin sovittamaan erästä Tracy-Widom -jakaumaa [8], toiselta nimeltään suurimman ominaisarvon jakaumaa, jonka todettiin sopivan testiaineistojen tapauksille erittäin hyvin. Residuaaleja yritettiin sitten simuloida satunnaisesti luotujen matriisin $\mathbf{B}$ ja vektoreiden x_i tulon sekä vektoreiden itsensä välille. Seuraavien lauseiden ja huomioiden pohjalta yhälön 4.8 residuaaleja pystyttiin simuloimaan satunnaisesti luoduille gammaspektrometrisille aikasarjoille.

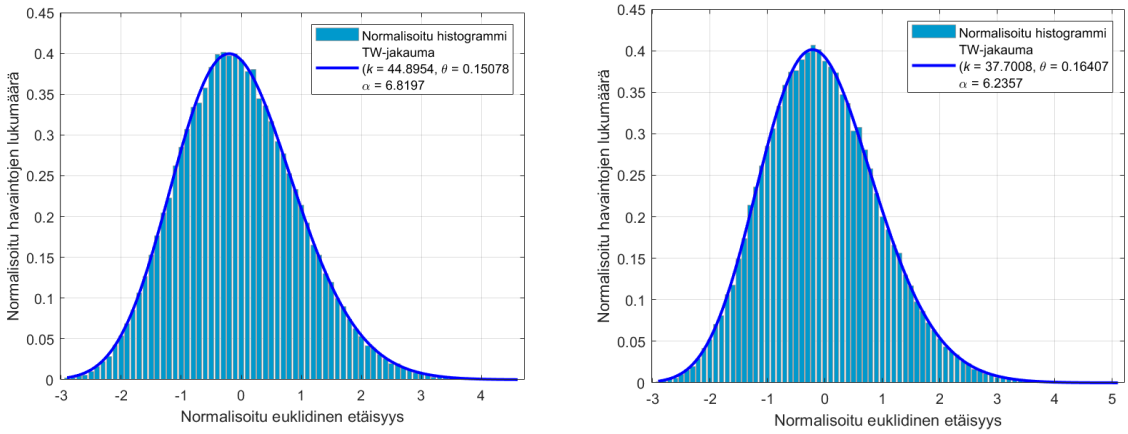
Gammaspektrometristen aikasarjojen kanavien tiedetään olevan Poisson-jakautuneita jollain parametrilla λ [30]. Poisson-jakautuneet satunnaismuuttujat taas ovat aina summia toisista Poisson-jakautuneista satunnaismuuttujista, kuten seuraavassa lauseessa 4.1 todetaan.

Lause 4.1. *[55, s. 195] Olkoon $X_1, X_2, \ldots, X_n$ i.i.d Poisson-jakautuneita satunnaismuuttujia siten, että $X_k \sim \mathcal{P}(\lambda_k)$ ja $k = 1, 2, \ldots, n$. Tällöin $S_n = X_1 + X_2 + \cdots + X_n \sim \mathcal{P}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$.*

$\mathcal{P}(\lambda)$ -jakautuneen satunnaismuuttujan voidaan lauseen 4.1 perusteella ajatella olevan summa λ määrästä i.i.d $\mathcal{P}(1)$-jakautuneita satunnaismuuttujia. Tällöin keskeisen raja-arvolauseen perusteella satunnaismuuttujien summan $S_n = \sum_1^\lambda \mathcal{P}(1)$ asymptoottinen jakauma on $\mathcal{N}(\lambda, \lambda)$ [55, s. 301].

Pääkomponentit noudattavat likimain normaalijakaumaa, jos myös alkuperäiset muuttujat ovat normaalisti jakautuneita [28, s.48]. Tällöin edellä mainitun perusteella gammaspektrometrisestä aikasarjasta muodostetut pääkompontit noudattavat myös likimain normaalijakaumaa, jos aikasarjan pulssien lukumäärät ovat tarpeeksi suuria. Tällöin yhtälön 4.8 residuualeja voidaan simuloida matriisille $\mathbf{B} = \mathbf{A}\mathbf{A}'$, missä $\mathbf{A}$ on matriisi, jonka sarakkeet ovat satunnaisesti luotuja ja normaalijakautuneita sekä vektoreille x_i, jotka on otettu satunnaisesti luodusta matriisista $\mathbf{x}$. Myös matriisin $\mathbf{x}$ sarakkeet ovat normaalijakautuneita. MATLAB-koodi simulaatiolle on liitteessä B.

Kun simulaatiota toistettiin useita kertoja eridimensioisille matriiseille ja vektoreille huomattiin, että residuaalit noudattavat joka kerta likimain edellä mainittua TW-jakaumaa. Simulaatiota toistettiin myös etäisyyksille d_{em}, joille jakauman todettiin sopivan myös hyvin. Kuvassa 4.3 on MATLAB-koodilla simuloitujen etäisyyksien d_e ja d_{em} histogrammit. Simulaatiossa otoksen koko on 100 000 ja vektorien pituus on 500. Matriisin $\mathbf{A}$ koko on (500×9).



Kuva 4.3. *Satunnaisesti luotujen matriisien* $\mathbf{Bx}$ *ja* $\mathbf{x}$ *rivien välisten etäisyyksien* d_e *(vas.) ja* d_{em} *(oik.) histogrammit. Histogrammeihin on sovitettu TW-jakauma.*

Lopulta jakauma saatiin johdettua myös teoreettisesti kun etäisyys 4.8 jaettiin vielä vektorin x_i normilla ja todistus sille on liitteessä C. Gammaspektrometrisen datan pääkomponenteista muodostuvan reaalimatriisin $\mathbf{A}'\mathbf{A}$ tiedetään olevan symmetrinen, jolloin se on myös Hermiittinen [14, s. 115]. Spektreistä x_i saadaan poistettua nollat, kun niiden kanaviin lisätään luku 1. Tällöin voidaan hyödyntää liitteen C todistusta sille, että Hermiittiselle matriisille $\mathbf{B}$ ja satunnaisille vektoreille x residuaalit $||\mathbf{B}x - x||/||x||$ noudattavat likimain siirrettyä TW-jakaumaa. Vaikka jakauman teoreettinen todistus onkin normalisoiduille residuaaleille, niin tässä työssä edellä mainittua todistusta ja liitteen B simulaatiota käytetään perusteena jakauman käytölle myös etäisyyksiä d_e ja d_{em} tarkasteltaessa.

Validointidata koostuu pelkästään taustasäteilyn spektreistä ja hälytysraja asetetaan sellaiselle tasolle, että STUKin säteilyvalvontaverkon 30 mittalaitteelle saa tulla yhteensä yksi väärä hälytys kuukaudessa. Mittalaitteille kertyy päivässä 144 havaintoa jokaista laitetta kohden eli yhteensä 129 600 havaintoa kuukaudessa. Tällöin *väärien hälytysten osuus (False Alarm Rate, FAR)* on $7,7 \times 10^{-6}$. Algoritmissa hälytysraja määritetään siis siirretystä TW-jakaumasta, jonka parametrit on laskettu validointidataa vastaavista residuaaleista.

4.4.4 PCA-mallin sovitus testidataan

Algoritmin viimeisessä vaiheessa PCA-malli sovitetaan testidatan spektreihin ja alkuperäisten ja sovitettujen spektrien välille lasketaan sovitusvirheet. Sovitusvirheen ylittäessä hälytysraja spektri luokitellaan poikkeavaksi. Algoritmin toisessa vaiheessa muodostettu PCA-malli voidaan sovittaa tuntemattomaan spektriin x kertomalla spektriä yhtälön 4.7 projektiomatriisilla seuraavasti:

$$\hat{x} = Bx. \tag{4.9}$$

Sovitukselle ja alkuperäiselle spektrille lasketaan euklidinen etäisyys

$$d_e = ||\hat{x} - x||, \tag{4.10}$$

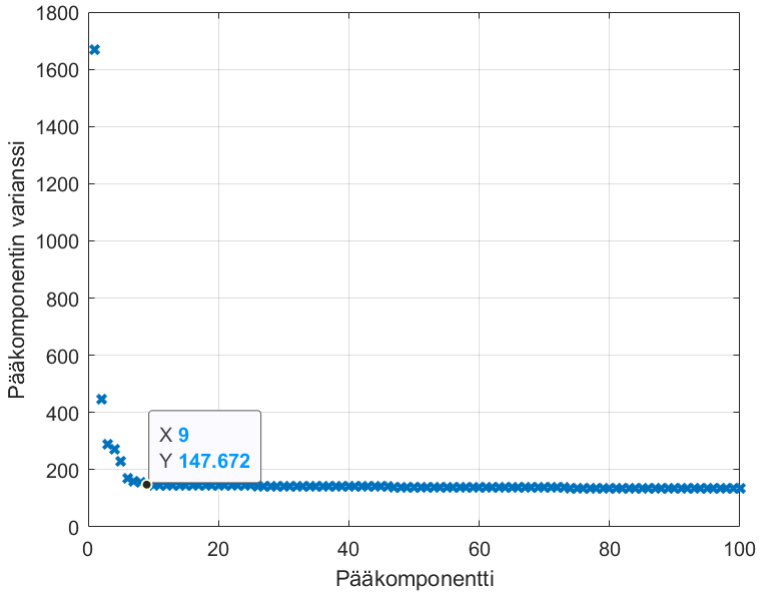
jota verrataan lopuksi luvussa 4.4.3 laskettuun hälytysrajaan.

Erilaisten etäisyysmittojen käyttöä residuaalin laskennassa testattiin ja niistä valikoitui käytettäväksi euklidinen etäisyys. Euklidisille etäisyyksille ohjelman laskentaaika oli huomattavasti lyhyempi, kuin esimerkiksi Mahalanobiksen etäisyyksille ja ne tuottivat taustasäteilyn spektreille hyvin tasaisen jakauman, ilman esiin ponnahtavia vääriä hälytyksiä. Lisäksi euklidisten etäisyyksien jakauma (normalisoituna) saatiin selville ensin kokeellisesti testaamalla ja myöhemmin teoreettisesti todistamalla, jolloin hälytysraja voidaan helposti määrittää riskitason perusteella.

5. TULOKSET JA POHDINTAA

5.1 $LaBr_3(Ce)$, Nuorgam

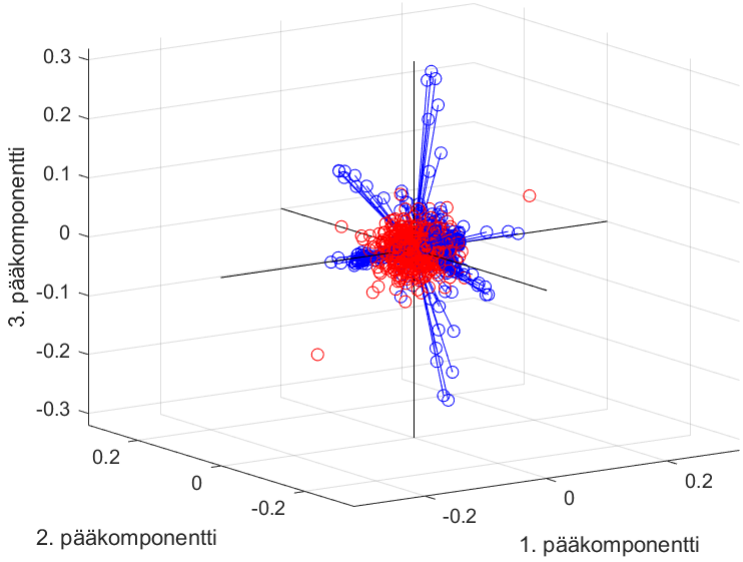
Luvussa 4.4 esitetty algoritmi ajettiin luvun 4.2.1 aineistolle. Harjoitusdatan, validointidatan ja testidatan lopulliset koot olivat 19 841 spektriä, 8 504 spektriä sekä 28 346 spektriä. Harjoitusdatalle laskettujen sadan ensimmäisen pääkomponentin varianssit ovat kuvassa 5.1. Kuvasta nähdään, että ensimmäisen pääkomponentin varianssi on ylivoimaisesti suurin. Yhdeksännen pääkomponentin kohdalla voidaan nähdä, että muita pääkomponentteja vastaavat varianssit muodostavat sen oikealle puolelle hyvin tasaisesti laskevan suoran, kun taas vasemmalle puolelle jyrkän käyrän (ks. luku 3.3). Tämän perusteella pääkomponenttien määräksi voidaan valita yhdeksän.



Kuva 5.1. *Nuorgamin harjoitusdatasta lasketut varianssit sadalle ensimmäiselle pääkomponentille.*

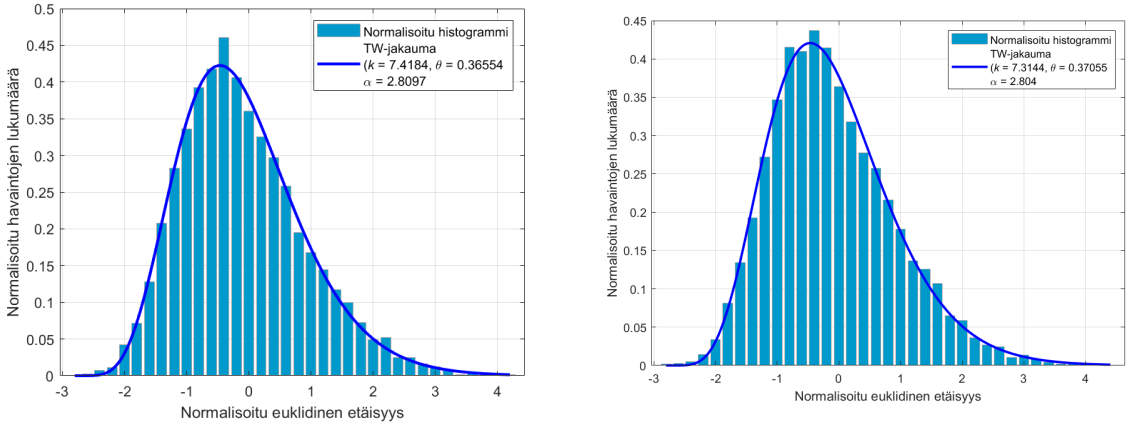
Kuvassa 5.2 on harjoitusdatan biplot-kuvaaja kolmen ensimmäisen pääkomponentin avaruudessa. Biplotin perusteella PCA ryhmittelee taustasäteilyn spektrit samankaltaisiksi huolimatta säätilojen muutosten aiheuttamasta vaihtelusta spektreissä.

Kaksi havaintoa poikkeavat hieman enemmän tuosta ryhmästä, mikä johtuu luultavasti yksittäisistä mittausvirheistä spektreissä.



Kuva 5.2. *Nuorgamin harjoitusdatalle piirretty biplot-kuvaajaa kolmen ensimmäisen pääkomponentin avaruudessa. Kuvassa punaisilla ympyröillä on merkattu harjoitusdatan havainnot ja sinisillä viiva-ympyrä-yhdistelmillä harjoitusdatan muuttujat.*

Hälytysrajan laskemista varten validointidatalle tehtiin PCA-sovitus yhdeksän ensimmäisen pääkomponentin avulla. Alkuperäisille ja sovitetuille spektreille laskettiin etäisyydet d_e ja niiden histogrammiin sovitettiin Tracy-Widom jakauma MATLABin lsqcurvefit-funktion avulla. Valmis funktio käyttää käyrän sovittamisessa trust-region-reflective -optimointialgoritmia [24]. Funktion käyttö edellytti histogrammidatan normalisointia ja normalisoitu histogrammi on piirrettynä sovitetun jakauman kanssa kuvassa 5.3. Jakauman perusteella hälytysrajaksi riskitasolla $7,7 \cdot 10^{-6}$ saatiin 350,02. Alkuperäisille ja sovitetuille spektreille laskettiin myös etäisyydet d_{em} ja niiden perusteella hälytysrajaksi saatiin 254,19. Kuvasta 5.3 nähdään, että myös nämä residuaalit noudattavat likimain TW-jakaumaa.



Kuva 5.3. *Nuorgamin validointidataa vastaavien etäisyyksen d_e (vas.) ja d_{em} (oik.) histogrammit.*

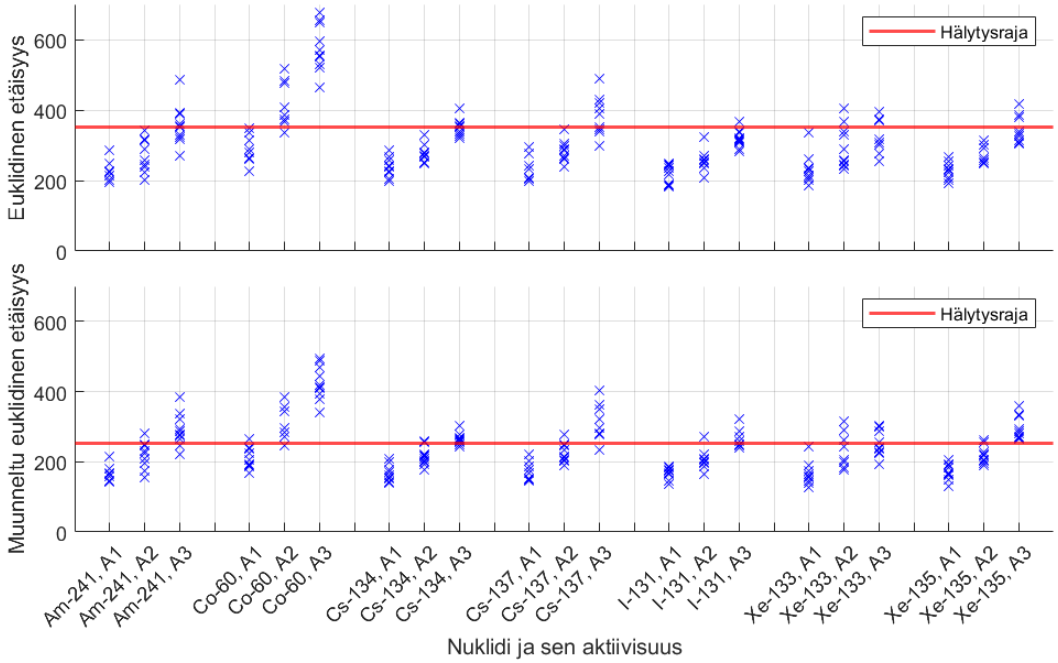
Algoritmin suorituskykyä arvioitiin sovittamalla PCA-malli testidataan. Etäisyyttä d_e käyttämällä oikeita hälytyksiä edellä mainitulla hälytysrajalla saatiin 45/210. Etäisyyttä d_{em} käyttämällä saatiin huomattavasti enemmän oikeita hälytyksiä, 76/210. Lisäksi taustasäteilyn spektrejä vastaavat lukuarvot näyttivät jäävän kauemmaksi hälytysrajan alle. Vääriä hälytyksiä molemmilla tavoilla saatiin 0/28 136.

Taulukossa 5.1 on kullekin testinuklidille etäisyyksiä d_e ja d_{em} vastaavat havaitsemisprosentit. Alin aktiivisuus ei näkynyt oikeastaan minkään nuklidin kohdalla kummankaan etäisyysmitan tapauksessa. Keskimmäisen aktiivisuuden kohdalla muunneltu etäisyys näyttää tuottavan jo selkeästi enemmän oikeita hälytyksiä, kun taas tavallinen jää lähes kaikkien nuklidien tapauksessa nollille. Yksi selkeästi poikkeava nuklidi on koboltti-60, joka nähdään lähes 100% varmuudella sekä keskimmäisen että korkeimman aktiivisuuden kohdalla. Melkein joka aktiivisuusalueella huonoiten erottuva nuklidi on jodi-131. Uraanin hajoamisketjuun kuuluvan nuklidin lyijy-214 intensiivisimmän gammaemission energia on 352 keV, mikä on lähellä jodin intensiivisintä emissioenergiaa 364 keV. Piikit osuvat spektrissä osittain päällekkäin, minkä vuoksi osa jodin piikistä selittyy PCA-sovituksessa lyijyn piikin avulla. Tällöin jodipiikin residuaali sovituksen kanssa jää suhteellisen alhaiseksi muihin testinuklideihin verrattuna.

Nuklidi	A1 d_e	A1 d_{em}	A2 d_e	A2 d_{em}	A3 d_e	A3 d_{em}
^{241}Am	0%	0%	0%	10%	60%	80%
^{60}Co	0%	10%	90%	90%	100%	100%
^{134}Cs	0%	0%	0%	20%	50%	80%
^{137}Cs	0%	0%	0%	10%	60%	90%
^{131}I	0%	0%	0%	10%	10%	70%
^{133}Xe	0%	0%	20%	30%	30%	40%
^{135}Xe	0%	0%	0%	20%	30%	100%

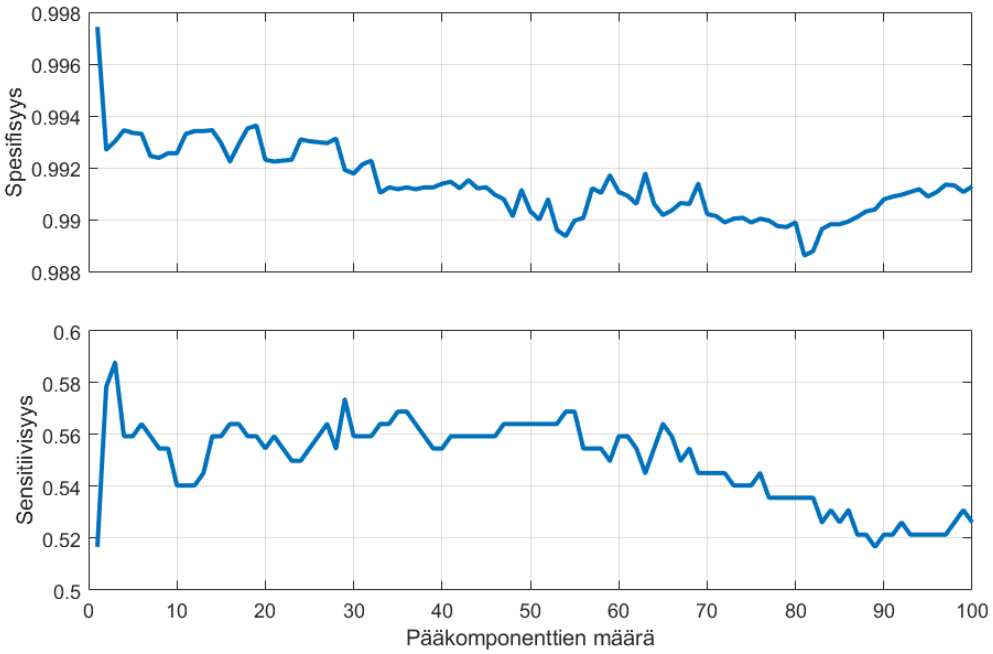
Taulukko 5.1. *Testisepktreihin generoitujen nuklidien ja niiden eri aktiivisuuksien havaitsemisprosentit.*

Kuva 5.4 näyttää kutakin nuklidia ja sen aktiivisuutta vastaavat etäisyydet d_e ja d_{em} ryhmiteltyinä. Vaaka-akselilla kuvassa on nuklidi ja sitä vastaava aktiivisuus ja pystyakselilla residuaali. Hälytysraja on taas merkattu punaisella viivalla. Kuten taulukosta 5.1, myös tästä kuva nähdään, että koboltti-60 nousee selkeästi muiden nuklidien yläpuolelle. Koboltin erottuminen johtunee tavasta, jolla vasteiden aktiivisuudet määriteltiin. Vaste laskettiin referenssipiikin perusteella, jolloin jokaiselle nuklidille annettiin sama referenssipiikin pinta-ala ja nuklidin muut fotopiikit laskettiin sen perusteella. Tällöin nuklidit, joilla on useita gammaemissioita eri energioilla, saavat suuremman yhteispinta-alan kuin sellaiset joilla niitä on vähemmän. Nuklidilla koboltti-60 on kaksi selkeästi erottuvaa fotopiikkiä. Lisäksi erottumiseen vaikuttanee luonnonsarjojen piikkien sijoittuminen kuhunkin testinuklidiin nähden. Cesium-134 ja jodi-131 erottuvat huonoiten. Nuklidin ksenon-133 tapauksessa on hieman erikoista, että korkein aktiivisuus erottuu lähes yhtä huonosti, kuin keskimäinen aktiivisuus. Tässä kohtaa on tärkeää muistaa, että spektrillä, johon nuklidin vaste lisätään, on suuri painoarvo siinä, kuinka hyvin nuklidi kyetään algoritmilla erottamaan. Ksenonin tapauksessa on luultavasti käynyt vain huono tuuri, kun sille on arvottu spektrejä.



Kuva 5.4. *Nuorgamin testidatan poikkeamien etäisyydet d_e (ylh.) ja d_{em} (alh.) ryhmiteltyinä kullekin aktiivisuudelle erikseen ja validointidatasta määritetty hälytysraja.*

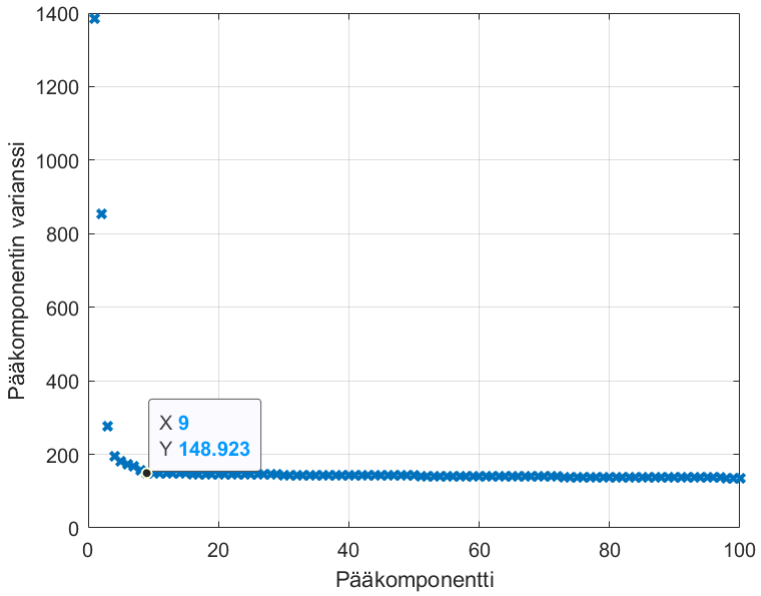
Sopivaa pääkomponenttien määrää tutkittiin lopuksi laskemalla testidatalle spesifisyys ja sensitiivisyys riskitasolla $7,7 \cdot 10^{-3}$ eri pääkomponenttien määrän funktiona. Väärien hälytysten todannäköisyyttä nostettiin, jotta pääkomponenttien määrän vaikutus validointidataa vastaavien etäisyyksien d_e jakauman yläpäähän tulisi mahdollisimman hyvin ilmi. Todennäköisyyden ollessa liian pieni spesifisyys saavuttaa erinomaisen arvon 1 lähes kaikilla pääkomponenttien määrillä. Spesifisyys ja sensitiivisyys pääkomponenttien määrille 1-100 ovat piirrettyinä kuvassa 5.5. Spesifisyyden perusteella paras pääkomponenttien määrä olisi 1, kun taas sensitiivisyyden perusteella se olisi 3. Yhden pääkomponentin avulla saatua residuaalien jakaumaa tutkittiin ja huomattiin, että sen oikeanpuoleinen häntä ei noudattanut kovinkaan hyvin TW-jakaumaa. Sen vuoksi jakaumasta laskettu hälytysraja on todellista rajaa korkeampi, mistä johtuu erityisen hyvä spesifisyys. Kyseiset kuvaajat on muodostettu etäisyyttä d_e käyttäen.



Kuva 5.5. *PCA-algoritmin sensitiivisyys ja spesifisyys Nuorgamin testidatalle pääkomponenttien määrän funktiona.*

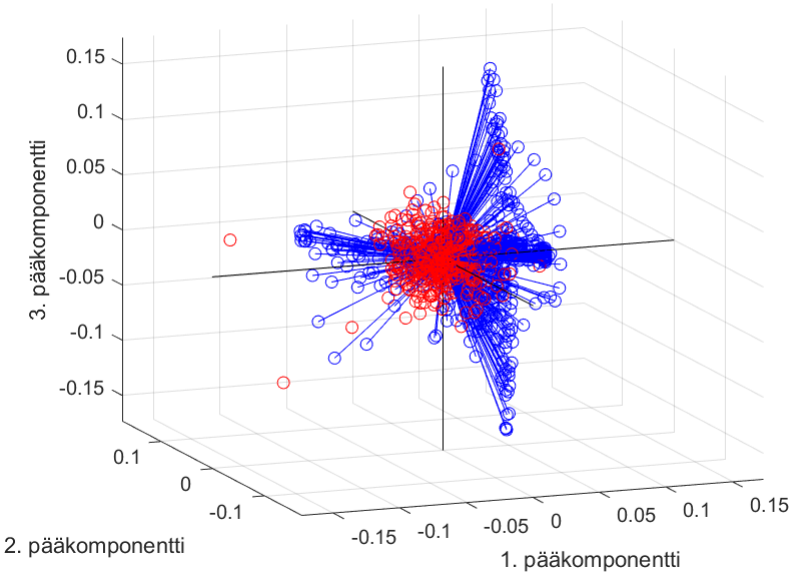
5.2 $LaBr_3(Ce)$, Rovaniemi

PCA-algoritmi ajettiin samaan tapaan luvussa 4.2.2 esitetylle Rovaniemen aineistolla, kuin Nuorgamin aineistolle luvussa 4.2.1. Harjoitusdatan, validointidatan ja testidatan lopulliset koot olivat 19 716 spektriä, 8 451 spektriä sekä 28 168 spektriä. Harjoitusdatalle laskettujen pääkomponenttien varianssit ovat piirrettyinä kuvassa 5.6. Rovaniemen tapauksessa kahden ensimmäisen pääkomponentin varianssit ovat selkeästi suurimpia. Pääkomponenttien määräksi voidaan valita yhdeksän samalla perusteella kuin Nuorgamin tapauksessa aiemmin.



Kuva 5.6. *Rovaniemen harjoitusdatasta lasketut pääkomponentteja vastaavat varianssit.*

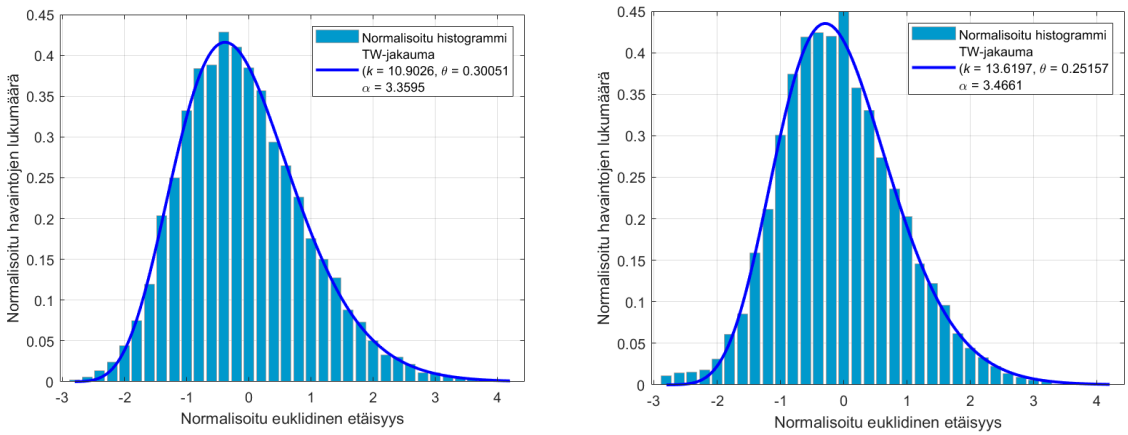
Harjoitusdatan biplot-kuvaaja 5.7 on hyvin samankaltainen, kuin Nuorgamin aineistolle piirretty. Myös tässä nähdään, että harjoitusdata sijoittuu aikalailla yhteen ryhmään lukuunottamatta muutamaa hieman poikkeavaa havaintoa.



Kuva 5.7. *Rovaniemen harjoitusdatalle piirretty biplot-kuvaajaa kolmen ensimmäisen pääkomponentin avaruudessa. Kuvassa punaisilla ympyröillä on merkattu harjoitusdatan havainnot ja sinisillä viiva-ympyrä-yhdistelmillä harjoitusdatan muuttujat.*

Validointidataa vastaavien residuaalien jakaumat ovat kuvassa 5.8. Myös Rovanie-

men tapauksessa molemmat residuaalit noudattavat likimain TW-jakaumaa. Etäisyyksien d_{em} jakauman vasemmanpuoleinen häntä ei vastaa ihan täysin TW-jakaumaa. Sen sijaan hälytysrajan kannalta oleellinen osuus, eli jakauman yläpää näyttää noudattavan jakaumaa hyvin. Luvussa 4.4.3 määritellyn riskitason perusteella etäisyyksien d_e jakaumasta hälytysrajaksi saatiin 325,61 ja etäisyyksien d_{em} jakaumasta 232,81.



Kuva 5.8. *Rovaniemen validointidataa vastaavien etäisyyksien d_e (vas.) ja d_{em} (oik.) histogrammit.*

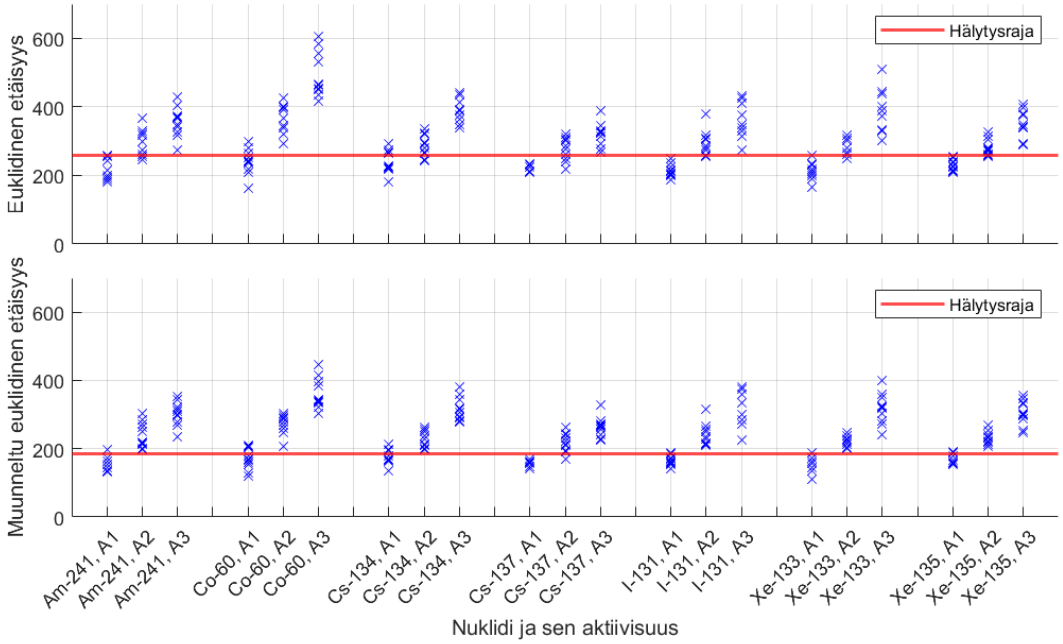
Taulukossa 5.2 on taulukon 4.2 aktiivisuuksia vastaavat havaitsemisprosentit. Myös Rovaniemen tapauksessa d_{em} tuottaa selkeästi paremmat havaitsemisprosentit kuin d_e. Etäisyydellä d_e vääriä hälytyksiä tuli 0/27 958 ja oikeita hälytyksiä 69/210. Etäisyyttä d_{em} käyttämällä testinuklideja havaittiin puolestaan 102/201. Vääriä hälytyksiä tuli myös sillä 0/27 958.

Nuklidi	A1 d_e	A1 d_{em}	A2 d_e	A2 d_{em}	A3 d_e	A3 d_{em}
^{241}Am	0%	0%	20%	50%	70%	100%
^{60}Co	0%	0%	80%	90%	100%	100%
^{134}Cs	0%	0%	10%	40%	100%	100%
^{137}Cs	0%	0%	0%	30%	40%	80%
^{131}I	0%	0%	10%	60%	80%	90%
^{133}Xe	0%	0%	0%	30%	90%	100%
^{135}Xe	0%	0%	10%	50%	80%	100%

Taulukko 5.2. *Testisepktreihin generoitujen nuklidien ja niiden eri aktiivisuuksien havaitsemisprosentit Rovaniemen datalle.*

Kuva 5.9 näyttää kutakin nuklidia ja sen aktiivisuutta vastaavat sekä etäisyydet d_e

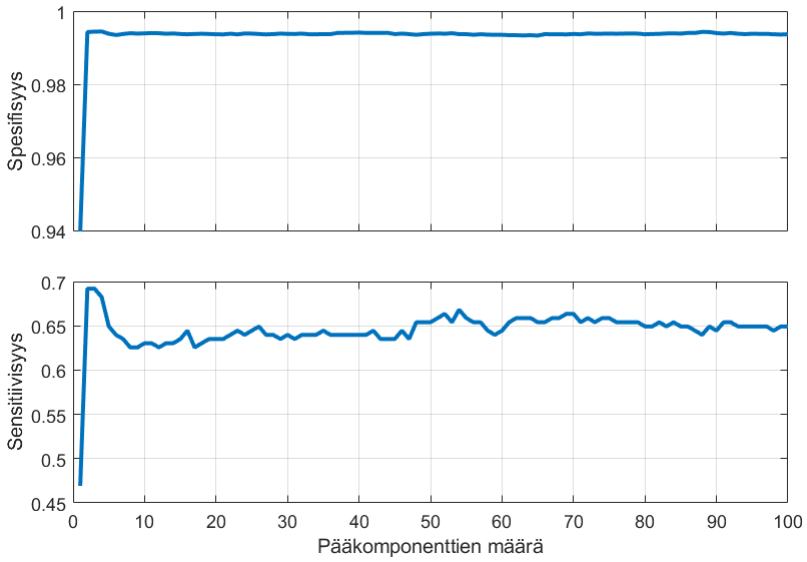
että etäisyydet d_{em} ryhmiteltyinä. Vaaka-akselilla kuvassa on nuklidi ja sitä vastaava aktiivisuus ja pystyakselilla residuaali. Hälytysraja on taas merkattu punaisella viivalla.



Kuva 5.9. *Rovaniemen testidatan poikkeamien etäisyydet d_e (ylh.) sekä d_{em} (alh.) ryhmiteltyinä kullekin aktiivisuudelle erikseen ja validointidatasta määritetyt hälytysrajat.*

Taulukon 5.2 ja kuvan 5.9 perusteella koboltti-60 näkyy parhaiten myös rovaniemen aineistolle. Nuklideista huonoiten havaittiin cesium-137, mikä poikkesi Nuorgamin testeistä, joissa se erottui muihin nuklideihin verrattuna hyvin. Tämä johtuu luultavasti siitä, että hiukkaskerääjällä mitatussa spektridatassa näkyy selkeämmin Tsernobylin ydinvoimalaonnettomuuden seurauksena luontoon vapautuneen nuklidin cs-137 säteily. Lisäksi jodi-131 näkyy huomattavasti selkeämmin kuin Nuorgamin tapauksessa. Muut nuklidit sijoittuvat aikalailla samoille prosenteille.

Myös Rovaniemen testidatalle laskettiin lopuksi spesifisyys ja sensitiivisyys riskitasolla $7,7 \cdot 10^{-3}$ eri pääkomponenttien määrän funktiona. Spesifisyys ja sensitiivisyys pääkomponenttien määrille 1-100 ovat piirrettyinä kuvassa 5.10. Spesifisyyden perusteella paras pääkomponenttien määrä olisi 3, kun taas sensitiivisyyden perusteella se olisi 2 tai 3. Kyseiset kuvaajat on muodostettu etäisyyttä d_e käyttäen.

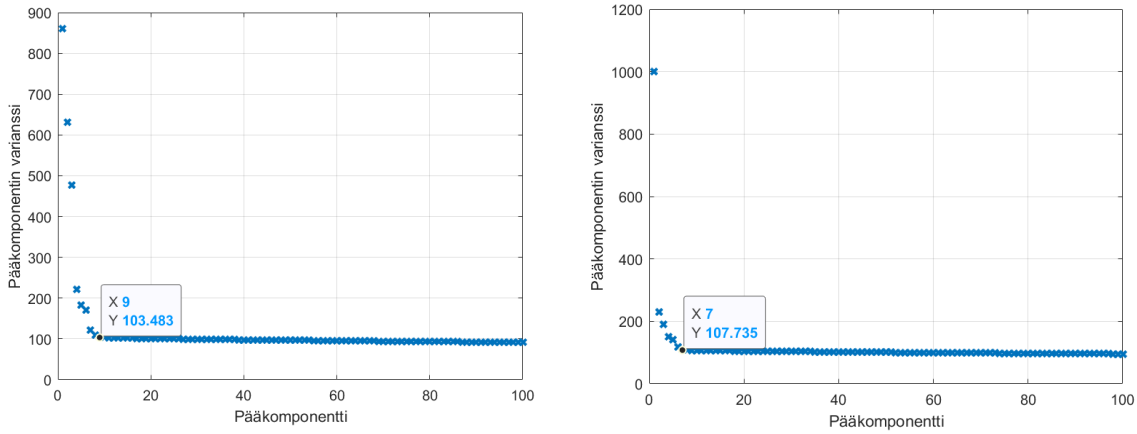


Kuva 5.10. *PCA-algoritmin sensitiivisyys ja spesifisyys Rovaniemen testidatalle pääkomponenttien määrän funktiona.*

5.3 $LaBr_3$(Ce), Kotka ja Harjavalta

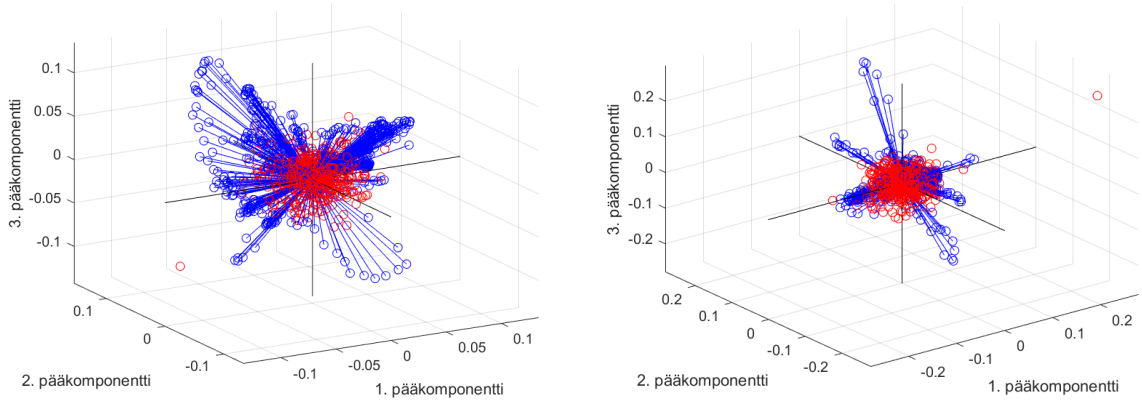
PCA-algoritmia testattiin lopuksi Kotkan ja Harjavallan oikeille säteilytapauksille. Kummassakin tapauksessa aineiston taustasäteilyn spektrit sekoitettiin satunnaiseen järjestykseen, kuten aiemmissakin testeissä tehtiin. Aineistojen loppuun sijoitettiin spektrit, joissa näkyy keinotekoisten säteilylähteiden vaikutus. Molemmat aineistot jaettiin harjoitus-, validointi- ja testidatoihin samassa suhteessa kuin Nuorgamin ja Rovaniemen aineistot aiemmin. Lopulta Kotkan aineistossa treenidataa oli 8 930 spektriä, validointidataa 3 828 spektriä ja testidataa 12 759 spektriä. Harjavallan aineisto koostui puolestaan 8 971 spektristä harjoitus-dataa, 3 845 spektristä validointidataa ja 12 817 spektristä testidataa.

Sekä Kotkan että Harjavallan harjoitusdatoille laskettujen sadan ensimmäisen pääkomponentin varianssit ovat piirrettyinä kuvassa 5.11. Kun käytetään luvussa 3.3 esiteltyä kulmapiste-menetelmää pääkomponenttien määrän valitsemiseksi, saadaan Kotkan harjoitusdatalle jälleen luku yhdeksän, mutta Harjavallan harjoitusdatalle sen sijaan luku seitsemän.



Kuva 5.11. *Kotkan (vas.) ja Harjavallan (oik.) harjoitusdatoista lasketut pääkomponentteja vastaavat varianssit sadalle ensimmäiselle pääkomponentille.*

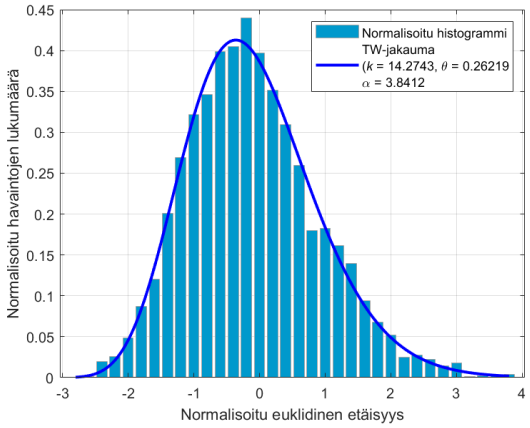
Harjoitusdatoille piirrettiin myös biplot-kuvaajat 5.12 kolmen ensimmäisen pääkomponentin avaruuteen. Kuvista nähdään, että taustadatat sijoittuvat suunnilleen yhteen samaan rykelmään, lukuunottamatta muutamaa hieman poikkeavaa havaintoa. Harjavallan biplotista nähdään yksi melko selkeästi erottuva havainto, joka voisi olla jokin poikkeama taustadatan seassa.



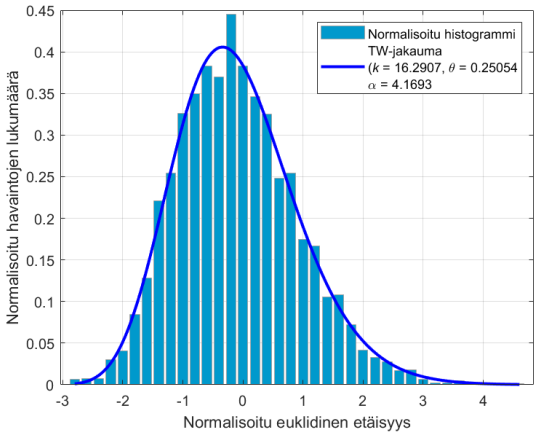
Kuva 5.12. *Kotkan (vas.) ja Harjavallan (oik.) harjoitusdatoille piirretyt biplot-kuvaajat kolmen ensimmäisen pääkomponentin avaruudessa. Kuvissa punaisilla ympyröillä on merkattu harjoitusdatan havainnot ja sinisillä viiva-ympyrä-yhdistelmillä harjoitusdatan muuttujat.*

Luvussa 4.4.3 määritellyn riskitason perusteella Kotkan aineistosta hälytysrajoiksi saatiin etäisyyksille d_e 490,25 ja etäisyyksille d_{em} 359,40. Harjavallan aineistosta vastaavat hälytysrajat olivat 393,71 ja 282,26. Kuvissa 5.13 ja 5.14 on kotkan etäisyyksien histogrammit ja kuvissa 5.15 ja 5.16 Harjavallan etäisyyksien histogrammit. Sekä Kotkan että Harjavallan aineiston molemmat etäisyydet näyttävät noudattavan hyvin TW-jakaumaa. Näissä kahdessa aineistossa validointidata on kuitenkin

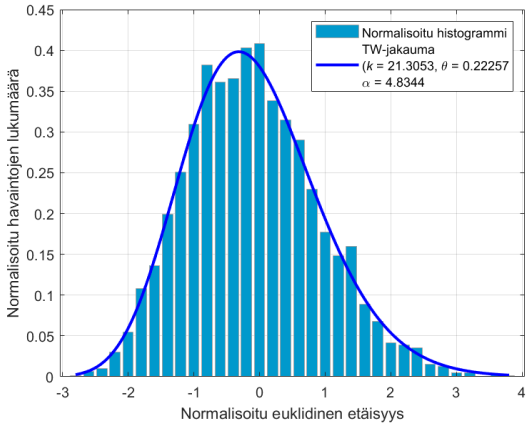
selkeästi pienempi kuin Nuorgamin ja Rovaniemen aineistoissa, minkä vuoksi histogrammeissa on nähtävissä pientä rosoisuutta.



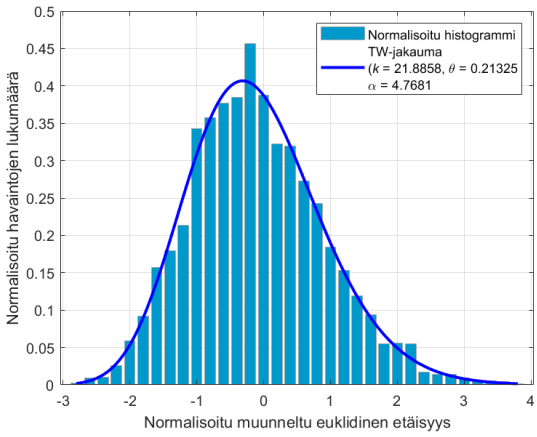
Kuva 5.13. *Kotkan validointidataa vastaavien etäisyyksien d_e histogrammi ja siihen sovitettu TW-jakauma.*



Kuva 5.14. *Kotkan validointidataa vastaavien etäisyyksien d_{em} histogrammi ja siihen sovitettu TW-jakauma.*



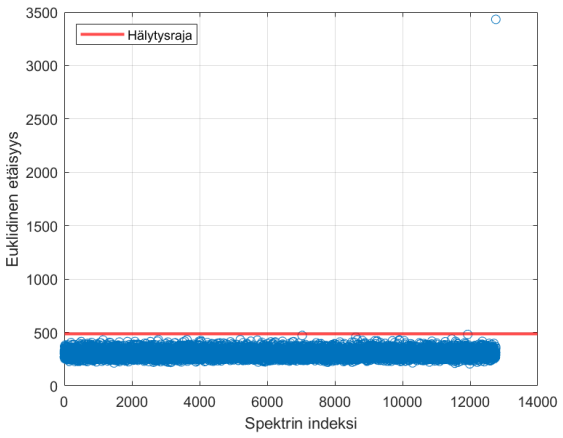
Kuva 5.15. *Harjavallan validointidataa vastaavien etäisyyksien d_e histogrammi ja siihen sovitettu TW-jakauma.*



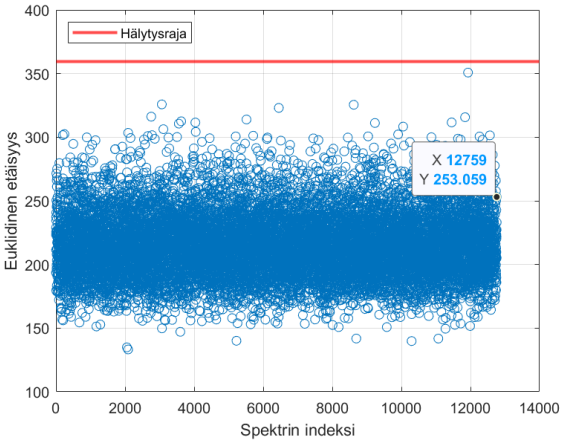
Kuva 5.16. *Harjavallan validointidataa vastaavien etäisyyksien d_{em} histogrammi ja siihen sovitettu TW-jakauma.*

Kuvassa 5.17 on Kotkan testidatan spektrejä vastaavat etäisyydet d_e sinisillä ympyröillä ja hälytysraja punaisella viivalla. Röntgensäteilyä vastaava spektri erottuu muun aineiston seasta selvästi ja vääriä hälytyksiä tällä hälytysrajalla tuli 2/12 759. Kuvassa 5.18 on testidatan spektrejä vastaavat etäisyydet d_{em}, joita käyttämällä röntgensäteilyn spektriä ei saatu havaittua. Edellä mainittu spektri on kuvassa korostettu mustalla pisteellä ja tietolaatikolla. Röntgensäteily ei muodosta spektriin samankaltaisia piikkejä kuin aiemmissa testeissä olleet nuklidit, vaan sen sijaan se muodostaa spektriin jatkumon. Röntgensäteilyn alkuperäistä ja sovitettua spektriä tutkittiin ja huomattiin, että jatkumon vuoksi luvussa 4.4 esitetty PNS-menetelmä

”painaa” alkuperäisen spektrin lähes kauttaaltaan alemmaksi kuin muut spektrit. Sen sijaan muutamilla ensimmäisillä pääkomponenteilla sovitettu spektri jää taustaspektrien tasolle. Siitä johtuen röntgensäteily ei näy ollenkaan etäisyyttä d_{em} käytettäessä, mutta näkyy selvästi etäisyyttä d_e käytettäessä. Etäisyydellä d_{em} ei havaittu kuitenkaan vääriä hälytyksiä.

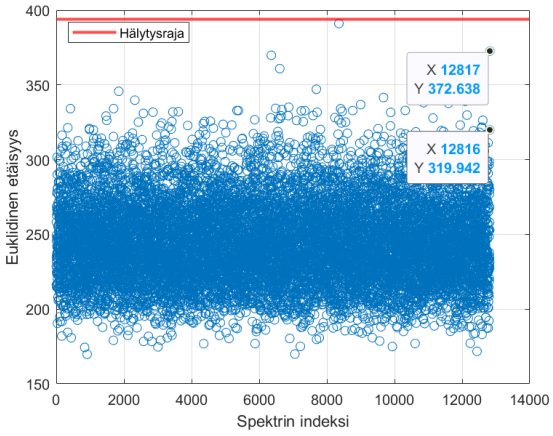


Kuva 5.17. *Kotkan testiaineiston havaintoja vastaavat etäisyydet d_e ja validointidatasta määritetty hälytysraja.*

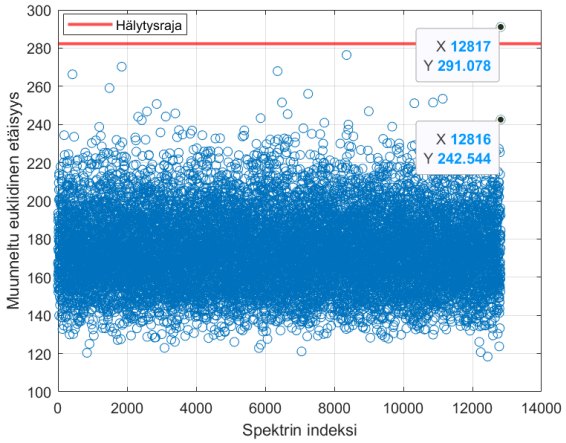


Kuva 5.18. *Kotkan testiaineiston havaintoja vastaavat etäisyydet d_{em} ja validointidatasta määritetty hälytysraja.*

Kuvassa 5.19 on Harjavallan testidatan spektrejä vastaavat euklidiset etäisyydet sinisillä ympyröillä ja hälytysraja punaisella viivalla. Tavallista euklidista etäisyyttä käyttämällä radionuklidin jodi-131 säteilyä ei havaittu. Kaksi selkeimmin erottuvaa jodia vastaavaa spektriä on korostettu kuvassa tietolaatikoilla. Kuvassa 5.20 on testidatan spektrejä vastaavat muunnellut euklidiset etäisyydet. Muunneltua euklidista etäisyyttä käyttämällä yksi nuklidin jodi-131 spektri havaittiin hälytysrajan yläpuolella. Loput spekrit jäivät selkeästi hälytysrajan alle. Kummassakaan tapauksessa spektrien joukosta ei noussut esiin vääriä hälytyksiä.



Kuva 5.19. *Harjavallan testiaineiston havaintoja vastaavat etäisyydet d_e ja validointidatasta määritetty hälytysraja.*



Kuva 5.20. *Harjavallan testiaineiston havaintoja vastaavat etäisyydet d_em ja validointidatasta määritetty hälytysraja.*

6. YHTEENVETO

Kirjallisuuskatsauksessa tutustuttiin usean muuttujan aineistoihin ja pääkomponenttianalyysiin. PCA on tehokas tilastollinen menetelmä, jonka avulla datasta voidaan saada esiin sen keskeisimmät tunnuspiirteet. PCA:lla pyritään yleensä pienentämään datan dimensioiden määrää, etsimään siitä samankaltaisia ryhmiä tai havaitsemaan poikkeamia. Jos saatavilla on riittävästi harjoitusdataa, PCA:lla voidaan kiteyttää harjoitusdatan merkittävintä tietoa sisältävät komponentit. Kun sovitetaan nämä komponentit tuntemattomaan datapisteeseen, voidaan tarkastella datapisteen ja harjoitusdatan välisiä eroavaisuuksia tai yhtäläisyyksiä.

Työn soveltavassa osiossa kehitettiin pääkomponenttianalyysiin perustuva algoritmi, jolla voidaan havaita keinotekoista säteilyä gammaspektrometrisistä aikasarjoista. Algoritmin yksi tärkeä tavoite oli minimoida taustasäteilystä aiheutuvat väärät hälytykset. Algoritmiin valikoitui testien perusteella tietyt ominaisuudet, kuten pääkomponenttien muodostaminen singulaariarvohajotelmasta, pääkomponenttien määrän valinta kulmapiste-menetelmällä, hälytysrajan valinta validointidatan avulla sekä PCA-sovitusvirheen laskeminen euklidisella etäisyydellä. Lisäksi datan esikäsittelyyn valittiin vierekkäisten kanavien yhdistäminen n kertaa, PNS-menetelmä sekä muuttujien keskitys ja skaalaus. Testeissä käytetyt parametrit ovat koottuna taulukossa 6.1.

Parametri	**Arvo**	**Huomioitavaa**
Spektrin kanavien yhdistämisen määrä	2	Tuloksena 512 kanavainen spektri
Pääkomponenttien määrä	9	Saatiin kulmapistemenetelmällä
Väärien hälytysten todennäköisyys	7,7e-6	Vastaa yhtä väärää hälytystä kuukaudessa 30 ilmaisinasemaa kohden
Ensimmäinen mukaan otettava kanava	42	Tätä kanavaa edeltävissä kanavissa on usein mittausepätarkkuuksia

Taulukko 6.1. *Algoritmin testauksessa käytetyt parametrit.*

Euklidiselle etäisyydelle kokeiltiin myös variaatiota, jossa vain positiiviset kanavien väliset erotukset kasvattivat residuaalia. Käytännössä sovitusvirheen laskennassa huomioitiin vain kanavat, joissa alkuperäinen spektri oli sovituksen yläpuolella. Tällöin huomattiin, että algoritmin sensitiivisyys piikkejä muodostavien säteilylähteiden suhteen parani merkittävästi ja väärien hälytysten lukumäärä väheni. Toisaalta tällöin sellaisia keinotekoisia säteilylähteitä, jotka tuottavat spektriin piikkien sijaan esimerkiksi jatkumon, ei havaittu.

Hälytysrajan löytämiseksi käytettiin harjoitus- ja analyysidatasta erillistä validointidataa. Validointidatalle laskettujen sovitusvirheiden jakaumaa tutkittiin ensin sovittamalla erilaisia jakaumia testiaineistoihin ja myöhemmin simuloimalla euklidisia etäisyyksiä satunnaisesti luotujen Hermiittisen matriisin $\mathbf{A}'\mathbf{A}$ ja vektorin $\mathbf{x}$ tulon sekä vektorin itsensä välille. Kokeellisen tarkastelun ja simulaatioiden perusteella etäisyydet näyttivät noudattavan aina likimain jonkin matriisin suurimman ominaisarvon jakaumaa, toiselta nimeltään Tracy-Widom -jakaumaa. Jakauma johdettiin myös teoreettisesti euklidisen etäisyyden normalisoidulle muodolle ja sen todistus on liitteenä C. Näiden tulosten perusteella väärien hälytysten todennäköisyyttä vastaava arvo haettiin validointidataa vastaaviin sovitusvirheisiin sovitetusta TW-jakaumasta.

Algoritmia testattiin neljällä eri aineistolla, joista kahteen lisättiin synteettisesti keinotekoisten radionuklidien vasteita ja kaksi sisälsi oikeiden keinotekoisten radionuklidien säteilyä. Ennalta määritetyllä hälytysrajalla testeissä vain Kotkan aineistosta nousi esiin vääriä hälytyksiä ja silloinkin vain kaksi tavallista euklidista etäisyyttä käytettäessä. Synteettisistä radionuklideista lähes kaikki havaittiin vähintään 70% todennäköisyydellä korkeimman aktiivisuusluokan kohdalla ja muunneltua euklidista etäisyyttä käyttämällä. Ainoastaan Nuorgamin aineiston ksenon-133 jäi tuolloinkin 40% havaitsemistodennäköisyyteen, mikä johtui luultavasti vain satunnaisesti valikoituneista taustaspektreistä. Koboltti-60 dominoi testejä saavuttaen vähintään 80% havaitsemistodennäköisyyden jokaisessa testissä aktiivisuustasoilla kaksi ja kolme. Nuorgamin tuloksissa huonoiten näkyi jodi-133, mikä johtuu luultavasti samankaltaisen luonnonnuklidin lyijy-214 esiintyvyydestä taustasäteilyn aineistossa. Rovaniemen tuloksissa jodi-133 näkyi jo yhtä hyvin kuin muutkin testinuklidit. Kotkan ja Harjavallan aineistoissa oli oikeita säteilytapauksia. Etäisyydellä d_{em} ei havaittu Kotkan aineistossa mukana ollutta röntgensäteilyä, mutta etäisyydellä d_e se havaittiin selvästi. Harjavallan poikkeamat havaittiin heikosti vain etäisyyttä d_{em} käyttämällä.

Datan esikäsittelyyn on olemassa useita erilaisia tapoja ja niitä tutkimalla algoritmia voisi vielä optimoida. Pääkomponenttien määrän valintamenetelmä valikoitui tässä työssä sen selkeyden ja yksinkertaisuuden vuoksi, mutta esimerkiksi sensitiivisyyden ja spesifisyyden perusteella laskettu määrä voisi antaa vielä parempia

tuloksia. Lisäksi erilaisten etäisyysmittojen, kuten Mahalanobiksen etäisyyden toimivuutta poikkeaman luokittelussa voisi olla syytä tarkastella enemmän. On myös kyseenalaistettava onko pääkomponenttianalyysi käyttötarkoitukseen parhaiten soveltuva menetelmä vai voisiko poikkeaman havaitsemisessa hyödyntää esimerkiksi neuroverkkoja. Artikkelissa [21] esitellään ensimmäinen autoenkooderia hyödyntävä ARAD-malli, joka on kehitetty havaitsemaan NaI(Tl)-ilmaisimella mitattua poikkeavaa keinotekoista säteilyä. Artikkelin perusteella ARAD-malli suoriutui joillain osa-alueilla paremmin kuin PCA, mutta joillain myös vähän huonommin.

Työssä saadut tulokset antavat lupaavan pohjan algoritmin käytölle myös todellisuudessa. On kuitenkin huomioitava, että testejä on tehty vasta hyvin rajalliselle määrälle erilaisia tapauksia. Algoritmin testausta voisi laajentaa koskemaan useampaa eri ilmaisinta useammasta eri maantieteellisestä sijainnista. Lisäksi algoritmia voisi testata kattavammalle radionuklidien joukolle useammalla eri aktiivisuudella. Kullekin radionuklidille voisi etsiä niin sanotun kynnysaktiivisuuden, jolloin nuklidin spektri juuri ja juuri vielä havaitaan. Lisäksi testejä voisi tehdä erityyppisille ilmaisimille, kuten Na(Tl)-tuikeilmaisimelle tai ajoneuvon kyydissä liikkuville ilmaisimille. Spektrien mittausaika on työn testispektreissä aina 10 minuuttia. Algoritmia voisi testata myös erilaisille mittausajoille, kuten spektreille, joita mitataan yhden sekunnin välein. Tällöin radionuklidin säteily näkyisi todennäköisesti useammassa peräkkäisessä spektrissä, jolloin ei olekaan niin olennaista saada kaikkia radionuklidia sisältäviä spektrejä näkyviin. Tällöin riittäisi tieto, että jossain tietyn aikavälin sisällä olevassa spektrissä havaitaan poikkeama. Lisäksi algoritmia testattiin vain sekoitetulle aikasarjalle. Olisi oleellista testata myös, kuinka hyvin joltain tietyltä aikaväliltä kerätyllä harjoitusdatalla pystytään ennustamaan erillisen aikavälin spektrejä.

LÄHTEET

[1] Karim Abadir. *Matrix Algebra*. Cambridge University Press, 2005. 433 pp.

[2] Herve Abdi and Lynne J. Williams. "Principal Component Analysis". In: *Wiley interdisciplinary reviews. Computational statistics* 2.4 (2010), pp. 433–459.

[3] J. Adewumi. "Understanding the Role of Eigenvectors and Eigenvalues in PCA Dimensionality Reduction". In: *Medium* (2019).

[4] IAEA (International Atomic Energy Agency). *Strengthening control over radioactive sources in authorized use and regaining control over orphan sources*. 2004. 108 pp.

[5] Lahtinen J. et. al. *Säteily ympäristössä, luku 8: Ulkoinen säteily*. Säteilyturvakeskus, 2003. 40 s.

[6] E. Anderson. "The Species Problem in Iris". In: *Annals of the Missouri Botanical Garden* 23.3 (1936), pp. 457–469+471–483+485–501+503–509.

[7] *Avoin data, STUK*. URL: https://stuk.fi/avoin-data.

[8] A. IU. Bejan. "Largest Eigenvalues and Sample Covariance Matrices. Tracy-Widom and Painlevé II: Computational Aspects and Realizations in s-plus with Applications". In: *Mathematics Subject Classification* (2006).

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning, Chapter 12: Continuous Latent Variables*. Springer, 2006. 738 pp.

[10] David et. al. Boardman. "Principal Component Analysis of Gamma-Ray Spectra for Radiation Portal Monitors". In: *IEEE Transactions of Nuclear Science* 59.1 (2012), pp. 154–160.

[11] F. Carlos. *Lecture notes for the course Probability and Statistics for Data Science*. Aug. 2017.

[12] R. et al. Casanovas. "Energy and resolution calibration of NaI(Tl) and LaBr3(Ce) scintillators and validation of an EGS5 Monte Carlo user code for efficiency calculations". In: *Nuclear Instruments and Methods in Physics Research A* 675 (2012), pp. 78–83.

[13] L. Chapel and C. Friquet. *Anomaly Detection with Score Functions Based on the Reconstruction Error of the Kernel PCA*. Springer, 2014. 668 pp.

[14] T. L. Chow. *Mathematical Methods for Physicists*. 2004. 555 pp.

[15] H.T. Eastment and W.J. Krzanowski. "Cross-validatory choice of the number of components from a principal component analysis". In: *Technometrics* 24 (1982), pp. 73–77.

[16] A. et al. Favalli. “Wide energy range efficiency calibration for a lanthanum bromide scintillation detector”. In: *Radiation Measurements* 43 (2007), pp. 506–509.

[17] William Feller. *Introduction to Probability Theory and Its Applications.* wiley, 1971.

[18] R. A. Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. In: *Annals of Human Genetics* 7.2 (1936), pp. 179–188.

[19] K.R. Gabriel. “Least Squares Approximation of Matrices by Additive and Multiplicative Models”. In: *J.R. Statist. Soc. B* 40 (1972), pp. 186–196.

[20] James Ghawaly. “A Datacentric Algorithm for Gamma-ray Radiation Anomaly Detection in Unknown Background Environments”. In: (2020).

[21] et. al. Ghawaly Jr. J. “Characterization of the Autoencoder Radiation Anomaly Detection (ARAD) model”. In: *Engineering Applications of Artifical Intelligence* 111 (2022).

[22] Harold Hotelling. “Analysis of a Complex of Statistical Variables Into Principal Components”. In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441.

[23] A.S. Householder and G. Young. “Matrix approximation and latent roots”. In: *Amer. Math. Mon.* 45 (1938), pp. 165–171.

[24] The MathWorks Inc. *lsqcurvefit, Solve nonlinear curve-fitting (data-fitting) problems in least-squares sense.* Natick, Massachusetts, United States, 2022. URL: https://se.mathworks.com/help/optim/ug/lsqcurvefit.html.

[25] The MathWorks Inc. *MATLAB version: 9.13.0 (R2022b).* Natick, Massachusetts, United States, 2022. URL: https://www.mathworks.com.

[26] G. et al. James. *An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, Incorporated.* 2014. 430 pp.

[27] R. A. Johnson and Wichern D. W. “Applied Multivariate Statistical Analysis”. In: *Prentice hall* 5 (2002).

[28] I. T. Jolliffe. *Principal Component Analysis, Second Edition.* Springer, 2002. 487 pp.

[29] Dan Kalman. “Leveling with Lagrange: An Alternative View of Constrained Optimization”. In: *Mathematics Magazine* 82.3 (2009), pp. 186–196.

[30] J. M. KirkPatrick and B. M. Young. “Poisson Statistical Methods for the Analysis of Low-Count Gamma Spectra”. In: *IEEE Transactions on Nuclear Science* 56 (2009), pp. 1278–1282.

[31] Jaakko Kirsilä. ”Matriisin singulaararvohajotelma” (2021).

[32] T. et. al. Kishimoto. “Path Planning for Localization of Radiation Sources Based on Principal Component Analysis”. In: *Applied Sciences* 11 (2021).

[33] S Klemola. *Säteily ja sen havaitseminen, luku 4: Säteilyn ilmaisimet.* Säteilyturvakeskus, 2002. 18 s.

[34] Steven J. Leon. *Linear Algebra with applications (9th ed.)* Pearson, 2015.

[35] E. Liski. "Monimuuttujamenetelmät" (2003).

[36] Luiz Gustavo D. et. al. Lopez. "A multivariate surface roughness modeling and optimization under conditions of uncertainty". In: *Measurement* 218.46 (2012), pp. 2555–2568.

[37] J. Mandel. "Principal Components, analysis of variance and data structure". In: *Statistica Neerlandia* 26 (1972), pp. 119–129.

[38] P. et. al Mashra. "Application of Student's test, Analysis of Variance, and Covariance". In: *Annals of Cardiac Anaesthesia* 22 (2019), pp. 407–411.

[39] et. al. Mashuri M. "PCA-based Hotelling's T2 chart with fast minimum covariance determinant (FMCD) estimator..." In: *Computers & Industrial Engineering* 158 (2021).

[40] B. Minty and Hovgaard J. "Reducing noise in gamma-ray spectrometry using spectra component analysis". In: *Exploration Geophysics* 33 (2002), pp. 172–176.

[41] B. Minty and Phil M. "Improved NASVD smoothing of airbone gamma-ray spectra". In: *Exploration Geophysics* 29 (1998), pp. 516–523.

[42] Thomas F. et al. Monaghan. "Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value". In: *Medicina(Kaunas)* 57.5 (2021).

[43] et. al. Mujica L.E. "Q-statistic and T2-statistic PCA-based measures for damage assessment in structures". In: *Structural Health Monitoring* 10.5 (2010), pp. 539–553.

[44] NCSS Team. *NCSS Documentation - Hotelling's One-Sample T2*. NCSS Statistical Software. 2023.

[45] R. Pires de Lima and K. Marfurt. "Principal component analysis and K-means analysis of airborne gamma-ray spectrometry surveys". In: *SEG Technical Program Expanded Abstracts* (2018).

[46] Hossein Pishro-Nik. *Introduction to probability, statistics, and random processes*. Blue Bell, PA: Kappa Research, LLC, 2014.

[47] R. Pöllänen. *Säteily ympäristössä, luku 1: Radioaktiiviset aikneet, säteily ja ympäristö*. Säteilyturvakeskus, 2003. 22 s.

[48] Chen-Kai et al. Qiao. "An Overview of the Compton Scattering Calculation". In: 11 (2021).

[49] E Randall. *Using R With Multivariate Statistics*. SAGE Publications, Inc, 2016.

[50] E.M et. al. Rasmusson. "Bienial variations in surface temperature over the United States as revealde by singular decomposition". In: *Mon. Weather Rev.* 109 (1981), pp. 587–598.

[51] Alvin Reincher and William Christensen. *Methods of multivariate analysis, third edition.* Hoboken: Wiley, 2012. 768 pp.

[52] Sascha Reinhardt. "Full Spectrum Analysis in Environmental Monitoring". In: *Radiation Protection Dosimetry* 160.4 (2014), pp. 311–317.

[53] M. Richardson. "Principal Component Analysis". In: (2009).

[54] B. Rieck. *A Note of the Relationship between PCA and SVD.* Oct. 2017.

[55] V. K. Rohatgi and A. K. MD. Ehsanes Saleh. *An Introduction to Probability and Statistics.* 2001. 716 pp.

[56] Eric W. Weisstein. *Hotelling T2 Distribution.* URL: https://mathworld.wolfram.com/HotellingT-SquaredDistribution.html.

[57] Bennet Williams. *Applications of Principal Component Analysis for Gamma-Ray Spectroscopy with Position-Sensitive Semiconductor Detectors.* 2019.

LIITE A: PCA-ALGORITMIN LÄHDEKOODI

Alla on PCA-algoritmin toteutus Java-lähdekoodina. Koodi sisältää luokan PCAAnomalyDetection, jossa on yksi julkinen metodi ja seitsemän yksityistä metodia. Algoritmin kannalta keskeisin metodi on runPCAAlgorithm, joka ottaa parametrinaan skaalatut, normalisoidut ja esikäsitellyt harjoitus- ja analyysidatat, pääkomponenttien määrän, hälytysrajan sekä harjoitusdatan keskiarvo- ja keskihajontavektorit. Metodissa luodaan PCA-malli harjoitusdatan perusteella ja analysoidaan sen avulla analyysi-datan spektrit. Algoritmista puuttuu hälytysrajan laskeminen validointidatan avulla, sen sijaan hälytysraja annetaan käyttäjän syötteestä.

Varsinainen PCA-malli luonti singulaariarvohajotelman avulla tapahtuu metodissa PCA ja varsinainen analysointi tapahtuu metodissa analyzeData. Metodi analyzeData kutsuu metodeja reconstructData, backScaleData, calculateL2norms ja classifySpectra. Ensimmäisessä näistä luodaan PCA-sovitukset halutulle spektrijoukolle. Metodissa backScaleData haluttu data skaalataan takaisin alkuperäiseen muotoon, jossa se oli ennen normalisointia ja skaalausta. Metodi calculateL2norms laskee euklidiset etäisyydet spektrien ja niiden sovitusten välille ja lopulta metodi classifySpectra luokittelee kunkin spektrin poikkeavaksi tai tavanomaiseksi.

Algoritmin toteutuksessa käytetään javan standardikirjastojen lisäksi Apachen kirjastoa Commons Math, jota hyödynnetään erityisesti singulaariarvohajotelman ja muiden matriisioperaatioiden laskemisessa.

```
import java.util.ArrayList;
import java.util.List;
import org.apache.commons.math3.exception.DimensionMismatchException;
import org.apache.commons.math3.linear.MatrixUtils;
import org.apache.commons.math3.linear.RealMatrix;
import org.apache.commons.math3.linear.RealVector;
import org.apache.commons.math3.linear.SingularValueDecomposition;

/**
 * Class for detecting anomalies in time series of gamma spectra. USer needs to
 * give data for training and analysing and also several parameters sich as
 * threshold, and number of PCs. With runPCAAlgorithm function of the class
```

```
 * user can run PCA based algorithm to find anomalies.
 * @author Ellinoora Vikman
 */
public class PCAAnomalyDetection{

    /** A PCA based function for detecting anomalies in spectral time series.
     * @param trainingData data for creating a model
     * @param analysingData data that is meant to be analyzed
     * @param numberOfPCs the number of PCs used in analysis
     * @param threshold the threshold where alarm is triggered
     */
    public void runPCAAlgorithm(List<double[]> trainingData, List<double[]>
            analysingData, int numberOfPCs, double threshold, double[] mu, double[]
                sigma) {

        int minPCs = trainingData.get(0).length;

        // Run PCA for training data
        List<RealMatrix> pca = PCA(trainingData);

        // Analyze given spectra
        analyzeData(analysingData, mu, sds, minPCs, numberOfPCs, threshold, pca);
    }

    /**
     * Analyses data with calculated PCA model
     * @param data data to be analyzed
     */
    private void analyzeData(List<double[]> analysingData, double[] mu,
            double[] sds, int min, int numberOfPCs, double threshold,
                List<RealMatrix> pca) {

        // Reconstruct analysing data
        RealMatrix recA = reconstructData(analysingData, numberOfPCs,
                pca, min);

        // Back scale recontruction
        recA = backScaleData(recA, mu, sds);

        // Back scale original data
        RealMatrix backScaledA = backScaleData(MatrixUtils
                .createRealMatrix(analysingData.toArray(new double[0][])),
```

```
                mu, sds);

        // Calculate euclidean distances between original and reconstructed data
        double[] l2normsAnalyse;

        l2normsAnalyse = calculateL2norms(backScaledA, recA);

        // classify spectra
        classifySpectra(l2normsAnalyse, threshold);
    }

    /**
     * Classifies spectra either to anomalies or usual, prints alarm if alarm
     * is triggered
     * @param lnormsAnalyse euclidean distances of analyzing spectra and their
     * reconstructions
     * @param threshold calculated threshold based on given risk level
     */
    private void classifySpectra(double[] lnormsAnalyse, double threshold) {

        for (int i = 0; i < lnormsAnalyse.length;i++) {
            if (lnormsAnalyse[i] > threshold) {
                System.out.println("Alarm triggered at spectrum " + i + ": " +
                        lnormsAnalyse[i]);
            }
        }
    }

    /**
     * Calculates PCA via SVD
     * @param data data for PCA analysis
     */
    private List<RealMatrix> PCA(List<double[]> data) {

        double [][] dataArray = data.toArray(new double[0][]);

        // Ccheck for missing values and replace them with zero
        for (double[] row : dataArray) {
            int i = 0;
            for (double point : row) {
                if (!Double.isFinite(point)) {
                    System.out.println("Infinite or missing values in data!");
```

```
                    row[i] = 1;
                }
                i++;
            }
        }

        // Perform Singular Value Decomposition
        RealMatrix matrix = MatrixUtils.createRealMatrix(dataArray);
        SingularValueDecomposition svd = new SingularValueDecomposition(matrix);

        // Get the singular value, U, and V matrices
        RealMatrix s = svd.getS();
        RealMatrix u = svd.getU();
        RealMatrix v = svd.getV();

        List<RealMatrix> pcaModel = new ArrayList<>();
        pcaModel.add(v);pcaModel.add(u);pcaModel.add(s);

        return pcaModel;
    }

     /**
     * Calculates PCA reconstruction for data
     * @param data data to be reconstructed
     * @return reconstructed data
     */
    private RealMatrix reconstructData(List<double[]> data, int numberOfPCs,
            List<RealMatrix> pca, int min) {
        double [][] dataArray = data.toArray(new double[0][]);
        RealMatrix matrix = MatrixUtils.createRealMatrix(dataArray);

        try {
            RealMatrix ev = pca.get(0).getSubMatrix(0, pca.get(0)
                .getColumnDimension()-1, 0, numberOfPCs-1);

            // Check for NAN values in datas
            checkNANValues(matrix);
            checkNANValues(ev);

            RealMatrix REC = ev.multiply(ev.transpose()).multiply(matrix
                    .transpose());
            return REC.transpose();
```

```
    } catch (NullPointerException | org.apache.commons.math3.exception
            .OutOfRangeException | DimensionMismatchException e) {

        System.err.println("Number of PCs exceed their count or try giving"
                + " min " + min + " spectras as training data!");

        return null;
    }
}

/**
 * Checks for NAN values in matrix and converts them to zero
 * @param matrix matrix to be checked
 */
private void checkNANValues(RealMatrix matrix) {
    for (int i = 0; i<matrix.getRowDimension(); i++) {
        for (int j = 0; j<matrix.getColumnDimension();j++ ) {
            if (Double.isNaN(matrix.getRow(i)[j])) {
                System.err.println("there is NaN in position (" + i
                        + ", " + j +")");
                matrix.setEntry(i, j, 0);
            }
        }
    }
}

/**
 * Back scales data
 * @param data data to be back scaled
 * @return back scaled data
 */
private RealMatrix backScaleData(RealMatrix data, double[] mu, double[] sds) {
    for (int i = 0; i < data.getRowDimension(); i++) {
        RealVector newRow;
        newRow = data.getRowVector(i).ebeMultiply(MatrixUtils
            .createRealVector(sds));


        double[] newRow2;
        newRow2 = newRow.add(MatrixUtils.createRealVector(mu))
            .toArray();
```

```
            data.setRow(i, newRow2);
        }
        return data;
    }

    /**
     * Calculates Euclidean distances between rows of two matrix
     * @param data actual data
     * @param pred predicted data
     * @return Euclidean distances in a  list
     */
    private double[] calculateL2norms(RealMatrix data, RealMatrix pred) {

        double[] l1norms = new double[data.getRowDimension()];
        for (int i = 0; i < data.getRowDimension(); i++) {
            RealVector j = data.getRowVector(i).subtract(pred.getRowVector(i));
            double lnorm = 0;
            for (double element : j.toArray()) {
                if (element > 0 ){
                    lnorm += element*element;
                }
            }
            l1norms[i] = Math.sqrt(lnorm);
        }

        return l1norms;
    }
}
```

LIITE B: RESIDUAALIEN SIMULOINTI

Alla on MATLAB-koodi satunnaisten Hermiten matriisin $\mathbf{A}'\mathbf{A}$ ja vektorin $\mathbf{x}$ tulon sekä vektorin itsensä välisten euklidisten etäisyyksien simulointia varten. Koodi simuloi yhden ajon aikana yhden satunnaisen Hermiten matriisin sekä n kappaletta satunnaisvektoreita. Kullekin vektorille lasketaan edellä mainittu etäisyys sekä tavallisessa, että muunnellussa muodossa. Lopuksi etäisyyksistä tehdään histogrammi ja histogrammiin sovitetaan Tracy-Widom -jakauma. Koodin lopussa esitellään funktio try_fit, jolla jakauman sovitus tehdään. Funktiossa käytetään MATLABin valmista sovitusmetodia lsqcurvefit, joka oletuksena hyödyntää trust-region-reflective -optimointialgoritmia [24].

```
% A code to simulate euclidean distances between real symmetric
% matrix. A'A and normally distributed random variables x. Also
% columns of A are normally distributed. TW-distribution is fitted
% to the histogram of residuals.

% Set the dimensions of the matrix and vector
n = 500; % Dimension 1 - or number of variables
p = 9; % Dimension 2 of the matrix
N = 100000; % Number of observations

% Initialize an array to store the residuals
distances = zeros(N, 1);

A = randn(p, n);

% Calculate A'A
A_transpose_A = A' * A;

% Generate N random vectors
x = randn(N, n);
distances = zeros(1,N);
```

```
% Calculate distances ||A'Ax - x|| for every observation in x
for i = 1:N
    vector = (x(i,:)'-A_transpose_A * x(i,:)');
    distance = sqrt((sum(vector.^2)));
    distances(i) = distance;
end

% Fit TW-distribution to residuals

data = (distances-mean(distances))./std(distances);

%tracy widom distribution
fun = @(p, x) 1./(gamma(p(1)).*p(2).^p(1)).*(x+p(3)).^(p(1)-1)...
        .*exp(-(x+p(3))./p(2)).*p(4);

p0 = [47,0.17, 7.8, 1];

figure(1)
p = try_fit(fun, data, p0, 42);

label = "TW-jakauma "+newline+"({\itk} = " +p(1) + ", " +"\theta = "
   +...
        p(2) + newline + "\alpha = "+p(3);
legend("Normalized histogram", label)
grid on;
title("")
xlabel('Normalized residual')
ylabel("Normalized count")
hold off;

function p = try_fit (fun, data, p0, nBins)
options = optimset('MaxFunEvals',100000);
options = optimset(options,'MaxIter',100000);

nBins = 42;
h = histogram(data,"Normalization", "pdf");
hold off;
YY = h.Values;
XX = h.BinEdges(1:length(YY));
```

```
startx = min(find(max(XX+3,0)));
XX = XX(startx:end);
YY = YY(startx:end);
x = XX(1):0.01:max(XX);
bar(XX, YY, "facecolor", [0, 0.60, 0.80], "edgecolor", [0.65 0.65
   0.65]);
hold on;

p = lsqcurvefit(fun, p0, XX, YY, [0,0,0,0],[50,1,20,2],options);

plot(x, fun(p, x), LineWidth=2, color="b")
```

LIITE C: RESIDUAALIEN JAKAUMAN TODISTUS

Alla olevassa todistuksessa johdetaan jakauma Satunnaisesti valitun Hermiten matriisin ja satunnaisesti luotujen reaalivektoreiden tulon sekä reaalivektoreiden itsensä välisille normalisoiduille euklidisille etäisyyksille. Päälause jakaumalle on lause 2, joka todistetaan lauseen 1 sekä korollaarien 1 ja 2 avulla.

Todistaaksemme, että Euklidiset etäisyydet $\|Bx - x\|/\|x\|$ noudattavat siirrettyä Tracyn ja Windomin jakaumaa, tarkastelemme ensin n.k. Von Mises iteraatiota, eli potenssimenetelmää suurinta ominaisarvoa vastaavan ominaisvektorin löytämiseksi. Tämän jälkeen osoitamme, että on olemassa jokin matriisi, jonka iteraatioina Euklidiset etäisyydet voidaan esittää. Iteraatioesityksen avulla osoitamme, että etäisyydet vastaavat jonkin matriisin suurimman ominaisarvon jakaumaa.

Lause 1 (Von Mises iteraatio) Ol. $x_0 \in \mathbb{R}^n$ vektori, jonka mikään komponentti ei ole 0. Ol. A matriisi ja v_1 sen suurinta ominaisarvoa λ_1 vastaava ominaisvektori. Iteraatio

$$x_1 = \frac{Ax_0}{\|Ax_0\|}$$

...

$$x_n = \frac{Ax_{n-1}}{\|Ax_{n-1}\|}$$

suppenee raja-arvoon v_1.
Tod. [1, s. 601-603]

Von Mises iteraatiosta seuraa suoraan seuraava, meille käyttökelpoisempi muoto.

Korollaari 1: Ol. λ_1 matriisin A ominaisvektoria v_1 vastaava ominaisarvo. Iteraatio

$$x_1 = \frac{Ax_0}{\|x_0\|}$$

...

$$x_n = \frac{Ax_{n-1}}{\|x_{n-1}\|}$$

suppenee raja-arvoon $\lambda_1 v_1$.

Tod.
Voimme kirjoittaa iteraation arvon x_n matriisin A ja alkuarvovektorin x_0 avulla muodossa

$$x_n = \frac{A^n x_0}{\|A^{n-1}x_0\|}.$$

Nyt

$$x_n = \frac{AA^{n-1}x_0}{\|A^{n-1}x_0\|} = A\left(\frac{A^{n-1}x_0}{\|A^{n-1}x_0\|}\right).$$

Koska sulkeiden sisäinen jono suppenee Lauseen 1 mukaan, saamme

$$\lim_{n\to\infty} x_n = Av_1 = \lambda_1 v_1.$$

Samaan tapaan kuin edellä, voimme saattaa Korollaarin 1 hyödyllisempään muotoon. Tarkastellaan nyt iteraatiota $x_n = (A^n x - A^{n-1}x)/\|A^{n-1}x\|$ ja sen normia.

Korollaari 2: Iteraatio

$$x_1 = \frac{Ax_0 - x_0}{\|x_0\|}$$

...

$$x_n = \frac{Ax_{n-1} - x_{n-1}}{\|x_{n-1}\|}$$

suppenee raja-arvoon $\lambda_1 v_1 - u$ missä u on yksikkövektori. Iteraation normi suppenee ominaisvektoria v_1 vastaavan ominaisarvon λ_1 ja reaaliluvun $-1 \leq r \leq 1$ erotukseen.

Tod.
Kuten edellä, voimme kirjoittaa iteraation arvon x_n matriisin A ja alkuarvovektorin x_0 avulla muodossa

$$x_n = \frac{A^n x_0 - A^{n-1} x_0}{\|A^{n-1} x_0\|}.$$

Nyt

$$x_n = \frac{AA^{n-1} x_0}{\|A^{n-1} x_0\|} - \frac{A^{n-1} x_0}{\|A^{n-1} x_0\|} = \frac{AA^{n-1} x_0}{\|A^{n-1} x_0\|} - u.$$

Korollaarista 1 ja raja-arvojen ominaisuuksista seuraa nyt suppeneminen

$$\lim_{n\to\infty} x_n = \lambda_1 v_1 - u.$$

Koska $\|v_1\| = 1$ ja $\|u\| = 1$, näemme että normille pätee

$$\lim_{n\to\infty} \|x_n\| = \|\lambda_1 - r\|,$$

missä r on reaaliluku ja -1 $\leq r \leq 1$.

Nyt voimme esitellä pääväitteemme ja sen todistuksen.

Lause 2 (Etäisyyksien jakauma) Ol. B hermiittinen satunnaismatriisi, ja x satunnainen reaalivektori, jonka mikään elementti ei ole 0 melkein varmasti. Tällöin normin $\|Bx - x\| / \|x\|$ jakauma on sama kuin jonkin matriisin suurimman ominaisvektorin jakauma (eli Tracyn ja Windomin jakauma) siirrettynä reaaliluvulla -1 $\leq r \leq 1$.

Tod.

Ol D_n matriisi, ja $D_n = \left(1 + \frac{1}{n}\right) B$. Ol myös p jokin reaaliluku. Nyt pätee

$$\mathbb{P}\left[\frac{\|Bx - x\|}{\|x\|} \leq p\right] = \mathbb{P}\left[\lim_{n\to\infty} \frac{\|D_n x - x\|}{\|x\|} \leq p\right].$$

Merkitään $x = D_n^{n-1} z$, ja ylempi yhtälö saadaan muotoon

$$\mathbb{P}\left[\frac{\|Bx - x\|}{\|x\|} \leq p\right] = \mathbb{P}\left[\lim_{n\to\infty} \frac{\|D_n x - x\|}{\|x\|} \leq p\right] = \mathbb{P}\left[\lim_{n\to\infty} \frac{\|D_n^n z - D_n^{n-1} z\|}{\|D_n^{n-1} z\|} \leq p\right].$$

Väite seuraa edellisestä yhtälöstä, sillä oikeanpuoleinen raja-arvo suppenee Korollaarin 2 mukaan, eli

$$\mathbb{P}\left[\frac{\|Bx - x\|}{\|x\|} \leq p\right] = \mathbb{P}\left[\lim_{n\to\infty} \frac{\|D_n^n z - D_n^{n-1} z\|}{\|D_n^{n-1} z\|} \leq p\right] = \mathbb{P}[\lambda_1 - r \leq p].$$

Lähteet:

[1]: John H. Mathews and Kurtis K. Fink, *Numerical Methods Using Matlab, 4th Edition,* Prentice-Hall Inc. Upper Saddle River, New Jersey, USA, 2004.