

T-61.2010 Datasta tietoon, syksy 2011

professori Erkki Oja

Tietojenkäsittelytieteen laitos, Aalto-yliopisto

7.11.2011

Tämän luennon sisältö

1 Vektoridatan tiivistäminen ja dekorrelointi

- Datamatriisi
- Pääkomponenttianalyysi (PCA)
- PCA-esimerkkejä: ominaiskasvot
- DSS-menetelmä

Datamatriisi

- Palautetaan mieleen *datamatriisi* $\mathbf{X}$:

$$\mathbf{X} = \begin{array}{c} \overbrace{\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & & & \\ x_{31} & & \bullet & & \\ \vdots & & & \ddots & \\ \vdots & & & & \bullet \\ x_{d1} & & & & x_{dn} \end{bmatrix}}^{n \text{ vektoria (havaintoa)}} \end{array} \left. \vphantom{\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & & & \\ x_{31} & & \bullet & & \\ \vdots & & & \ddots & \\ \vdots & & & & \bullet \\ x_{d1} & & & & x_{dn} \end{bmatrix}} \right\} d \text{ vektorelementtia (dimensio)}$$

- Usein matriisin sarakkeilla (vektoreilla $\mathbf{x}(t)$, $t = 1, \dots, n$) ei ole mitään määrättyä järjestystä vaan ne ovat vain otos jostakin perusjoukosta (esim. tietyn kokoisten digitaalikuvien joukko).

Datamatriisi (2)

- Joskus vektoreilla $\mathbf{x}$ on kuitenkin selkeä ajallinen järjestys; ks. DSS-menetelmä luvun lopussa.
- Usein vaikeus on vektoreiden suuri dimensio (esim. tekstidokumenttien sanahistogrammit: dimensio voi olla 50.000; digitaalinen kuva rivi riviltä skannattuna: dimensio on rivit $\times$ sarakkeet)
- Kuitenkin vektorialkioiden välillä on voimakkaita riippuvuuksia (esim. kuvan vierekkäiset pisteet)

Datamatriisi (4)

- Menetelmiä: pääkomponenttianalyysi, joka poistaa korrelaatiot (2. kertaluvun tilastolliset riippuvuudet); riippumattomien komponenttien analyysi, joka pyrkii poistamaan kaikki riippuvuudet.

Pääkomponenttianalyysi (PCA)

- Pääkomponenttianalyysi (Principal Component Analysis, PCA; eli Hotelling-muunnos eli Karhunen-Loève -muunnos) on klassinen tekniikka datan ja signaalien analyysissä
- PCA *poistaa korrelaatiot* vektorialkioiden välillä ja samalla löytää kierron vektoriavaruudessa siten että uusilla (kierretyillä) koordinaateilla on maksimaalisen suuret varianssit (energiat).
- Tietyillä ehdoilla PCA löytää tilastollisesti riippumattomat komponentit (jos data normaalijakautunutta).
- Katsotaan seuraavassa matemaattisesti miten tämä tehdään.

Pääkomponenttianalyysi (PCA) (2)

- Lähtökohtana ovat datamatriisin sarakkeet: vektorit $\mathbf{x}$ joilla on d alkia. Datamatriisi sisältää otoksen $\mathbf{x}(1), \dots, \mathbf{x}(n)$ näitä vektoreita.
- Tyypillisesti $\mathbf{x}$:n alkioit voivat siis olla piirre-arvoja, kuvapikseleiden harmaatasoja, histogrammiarvoja tms. kuten edellisissä luvuissa on esitetty
- Pääkomponenttimuunnoksessa vektorit $\mathbf{x}$ nollakeskiarvoistetaan ensin vähentämällä keskiarvo:

$$\mathbf{x} \leftarrow \mathbf{x} - E\{\mathbf{x}\}$$

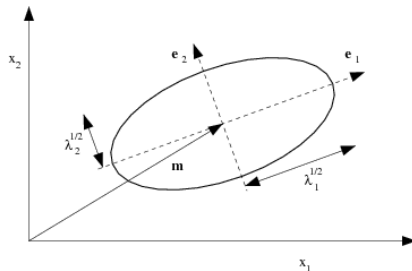
Pääkomponenttianalyysi (PCA) (3)

- Tässä merkintä $E\{\mathbf{x}\}$ tarkoittaa keskiarvoa: käytännössä se lasketaan seuraavasti:

$$E\{\mathbf{x}\} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}(i)$$

- Oletetaan seuraavassa että tämä on tehty ja vektorit $\mathbf{x}$ ovat siis nollakeskiarvoisia.
- Sitten $\mathbf{x}$ muunnetaan lineaarimuunnoksella toiseksi vektoriksi $\mathbf{y}$ jossa on m alkia, $m \leq d$, niin että korrelaatioista johtuva redundanssi poistuu.
- Samalla kertaa $\mathbf{x}$:n uusille koordinaattiakseleille otettujen projektoiden varianssit maksimoituvat

Pääkomponenttianalyysi (PCA) (4)



Kuva: Vektorijakauman (kuvattu ellipsillä) pääakselit kahdessa dimensiossa: e_1 ja e_2 . Keskiarvo on m . Ominaisarvot λ_1 ja λ_2 antavat varianssit pääakselien suunnassa

Pääkomponenttianalyysi (PCA) (5)

- Ensimmäinen akseli e_1 kuvassa 2 (pääkomponentin suunta) vastaa suurinta varianssia, toinen on suurin varianssi ensimmäistä vasten kohtisuorassa suunnassa jne.
- Viimeisillä pääkomponenteilla on jo niin pieni varianssi, että ne voidaan kokonaan jättää pois. Tähän perustuu vektoreiden tiivistäminen (dimension alentaminen).

► Esimerkki 3D-, 2D- ja 1D-datasta

Pääkomponenttianalyysi (PCA)

Matemaattinen johtaminen maksimivarianssikriteerillä

- Matemaattisesti: ajatellaan lineaarikombinaatiota

$$y_1 = \sum_{k=1}^d w_{k1} x_k = \mathbf{w}_1^T \mathbf{x}$$

vektorin $\mathbf{x}$ alkioista $x_1, \dots, x_n$ (projektiota suunnalle $\mathbf{w}_1$)

- Summa y_1 on nimeltään $\mathbf{x}$:n *ensimmäinen pääkomponentti*, jos y_1 :n varianssi $E\{y_1^2\}$ on mahdollisimman suuri. Taas varianssi voidaan käytännössä laskea summana yli otoksen $\mathbf{x}(i)$. Vektori $\mathbf{w}_1$ on vastaava *pääkomponenttivektori*.
- Jotta $E\{y_1^2\}$ ei kasvaisi äärettömän suureksi, täytyy vektorin $\mathbf{w}_1$ pituutta jotenkin rajoittaa. Kätevin rajoite on että sen normi on vakio, käytännössä 1

Pääkomponenttianalyysi (PCA) (2)

Matemaattinen johtaminen maksimivarianssikriteerillä

- Siten PCA-kriteeri on seuraava: maksimoi

$$J_1^{PCA}(\mathbf{w}_1) = E\{y_1^2\} = E\{(\mathbf{w}_1^T \mathbf{x})^2\} = \mathbf{w}_1^T E\{\mathbf{x}\mathbf{x}^T\} \mathbf{w}_1 = \mathbf{w}_1^T \mathbf{C}_x \mathbf{w}_1$$

missä $\|\mathbf{w}_1\| = 1$

- Matriisi $\mathbf{C}_x = E\{\mathbf{x}\mathbf{x}^T\}$ on $d \times d$ ja nimeltään $\mathbf{x}$:n *kovarianssimatriisi*
- Se lasketaan käytännössä nolla-keskiarvoistetusta datamatriisista $\mathbf{X}$ kaavalla $\mathbf{C}_x = \frac{1}{n} \mathbf{X}\mathbf{X}^T$
- Ratkaisu on

$$\mathbf{w}_1 = \mathbf{e}_1$$

missä $\mathbf{e}_1$ on $\mathbf{C}_x$:n ominaisvektori vastaten suurinta ominaisarvoa λ_1 .

- Miksi?

Pääkomponenttianalyysi (PCA) (3)

Matemaattinen johtaminen maksimivarianssikriteerillä

- Katsotaan tätä harjoitustehtävänä, mutta eräs peruste on seuraava: $\mathbf{w}_1^T \mathbf{C}_x \mathbf{w}_1$ on vektoreiden $\mathbf{w}_1$ ja $\mathbf{C}_x \mathbf{w}_1$ sisätulo, joka on samalla niiden pituuksien (euklidisten normien) tulo kertaa niiden välisen kulman kosini. (Muista sisätulon määritelmä!)
- Sovittiin että $\|\mathbf{w}_1\| = 1$ joten jää kulman kosini kertaa $\|\mathbf{C}_x \mathbf{w}_1\|$. Kosini maksimoituu kun kulma on nolla, mutta silloin pätee $\mathbf{C}_x \mathbf{w}_1 = \lambda \mathbf{w}_1$ missä λ on jokin skalaarivakio.
- Tämä on matriisin $\mathbf{C}_x$ ominaisarvoyhtälö jonka ratkaisuina ovat sen d ominaisvektoria. Mikä niistä pitää valita?
- Silloin myös pätee $\|\mathbf{C}_x \mathbf{w}_1\| = \lambda$ joten λ :n pitäisi olla mahdollisimman suuri. Valitaan siis suurinta ominaisarvoa vastaava ominaisvektori pääkomponenttivektoriksi $\mathbf{w}_1$.

Pääkomponenttianalyysi (PCA) (4)

Matemaattinen johtaminen maksimivarianssikriteerillä

- Maksimivarianssikriteeri

$$\max E\{y_k^2\} = E\{(\mathbf{w}_k^T \mathbf{x})^2\}$$

voidaan yleistää m :lle pääkomponentille kun lisätään rajoitusehdot: joko

$$E\{y_m y_k\} = 0, \quad k < m \tag{1}$$

tai

$$\mathbf{w}_i^T \mathbf{w}_j = \delta_{ij}$$

- Ratkaisuna on että k :s pääkomponentti on $y_k = \mathbf{e}_k^T \mathbf{x}, \quad k = 1, \dots, n$

Pääkomponenttianalyysi (PCA)

Johtaminen pienimmän neliösumman virheeseen (MSE) perustuen

- Toinen tapa määritellä PCA on vektoreiden $\mathbf{x}$ *pienimmän neliösumman virhe* (MSE) kun ne esitetään *PCA-kehitemässä* jossa on mukana m termiä
- Merkitään joukkoa ortogonaalisia vektoreita $\mathbf{w}_1, \dots, \mathbf{w}_m$
- MSE-kriteeri:

$$J_{MSE}^{PCA} = E\left\{\left\|\mathbf{x} - \sum_{i=1}^m (\mathbf{w}_i^T \mathbf{x}) \mathbf{w}_i\right\|^2\right\} \quad (2)$$

Pääkomponenttianalyysi (PCA) (2)

Johtaminen pienimmän neliösumman virheeseen (MSE) perustuen

- On helppo näyttää että voimme kirjoittaa

$$J_{MSE}^{PCA} = E\{\|\mathbf{x}\|^2\} - E\left\{\sum_{i=1}^m (\mathbf{w}_i^T \mathbf{x})^2\right\} \quad (3)$$

$$= \text{trace}(\mathbf{C}_\mathbf{x}) - \sum_{i=1}^m \mathbf{w}_i^T \mathbf{C}_\mathbf{x} \mathbf{w}_i \quad (4)$$

- Voidaan osoittaa että kriteerin (4) minimi ortogonaalisuusehdon vallitessa on m :n ensimmäisen ominaisvektorin $\mathbf{e}_1, \dots, \mathbf{e}_m$ muodostama kanta
- PCA-kehitemä tietyllä vektorille $\mathbf{x}$ on silloin

$$\mathbf{x} = \sum_{i=1}^m (\mathbf{e}_i^T \mathbf{x}) \mathbf{e}_i \quad (5)$$

Pääkomponenttianalyysi (PCA) (3)

Johtaminen pienimmän neliösumman virheeseen (MSE) perustuen

- Pienimmän MSE-virheen arvo on silloin

$$J_{MSE}^{PCA} = \sum_{i=m+1}^n \lambda_i \quad (6)$$

- Tämä on niiden pienimpien ominaisarvojen summa, jotka vastaavat poisjätettyjä ominaisvektoreita $\mathbf{e}_{m+1}, \dots, \mathbf{e}_n$
- Tästä voi myös laskea kuinka monta pääkomponenttia täytyy ottaa mukaan kehittelmään (5) jotta päästään haluttuun kompressioon.

Esimerkki: ominaiskasvot

- Eräs suosittu pääkomponenttianalyysin (PCA) sovellus on ns. *ominaiskasvot* (eigenfaces)
- Siinä kerätään joukko kasvokuvia, yleensä normeerattuja niin että kasvonpiirteet ovat suunnilleen samassa paikassa kuva-alueessa
- Kuvat skannataan vektoreiksi vaakariveittäin
- Nämä vektorit ovat nyt datamatriisin **X** sarakkeita.
- Lasketaan ominaisvektorit, jotka voidaan visualisoida ominaiskasvoina

Esimerkki: ominaiskasvot (2)



Kuva: Ominaiskasvoja

- Näitä voi käyttää kuvien tunnistamiseen projisoimalla kuvat mataladimensoiseen (2- tai 3-dim.) pääkomponenttiavaruuteen; saman henkilön eri kuvat osuvat siellä lähekkäin. ▶ Esimerkki projektioista T2-harjoituksissa

Esimerkki: ominaiskasvot (3)

- Tiivistettyä esitystä voi hyödyntää myös datan pakkaamisessa

► Tietokoneharjoitusten T2 esimerkki ominaiskasvoista

Esimerkki: käsinkirjoitetut merkit

- Seuraava esimerkki näyttää, kuinka datan rekonstruktio paranee kun yhä enemmän pääkomponentteja otetaan mukaan kehittelmään (5)
- Vasemmanpuoleisessa sarakkeessa on käsinkirjoitettuja numeroita digitoituna 32×32 kuvamatriisiin
- Seuraavassa sarakkeessa on niiden keskiarvovektori, joka siis vähennetään ennen PCA-analyysiä
- Vektorien $\mathbf{x}$ dimensio on siis 1024
- Kuva näyttää rekonstruktion PCA:lla kun 1, 2, 5, 16, 32, ja 64 pääkomponenttia on mukana kehittelmässä.

Esimerkki: käsinkirjoitetut merkit (2)



Kuva: Käsinkirjoitettuja merkkejä. Vasemmalta alkuperäinen, keskiarvo, rekonstruktioit käyttäen 1, 2, 5, 16, 32 ja 64 pääkomponenttia.

Esimerkki: ilmastodata

- Otetaan toinen esimerkki: ilmastodatan ns. *kokeelliset ortogonaalifunktiot* (*Experimental Orthogonal Functions, EOF*)
- Ilmastodata (säädata) tarkoittaa säännöllisin välein (tunneittain, päivittäin jne.) mitattuja lämpötiloja, ilmanpaineita, sademääriä jne. usealla paikkakunnalla
- Amerikkalainen järjestö NCAR (National Center for Atmospheric Research, Kansallinen ilmakehätutkimuksen keskus) julkaisee Webissä tällaisia mittaustietoja pitkältä ajanjaksolta ja koko maapallon peittävältä mittaushilalta (jossa puuttuvat arvot esim. keskellä valtameriä on käytännössä laskettu tietyillä kaavoilla).

Esimerkki: ilmastodata (2)

- Me käytimme heidän dataansa jossa ajanjakso on 1948–2004, mittaukset päivittäisiä, ja hila 2.5 asteen välein maapallon pinnalla
- Näistä voi tehdä datamatriisin $\mathbf{X}$ jossa vaakadimensio on aika (56 x 365 päivää = 20.440 päivää) ja pystydimensio on paikka (73 x 144 = 10.512 mittauspistettä). Matriisin koko siis 10.512 x 20.440.
- Säätieteilijät haluaisivat hajottaa (keskiarvoistetun) datamatriisin summaksi

$$\mathbf{X} = E\{\mathbf{X}\} + \sum_{i=1}^m \mathbf{w}_i \mathbf{y}_i^T$$

jossa sarakevektoreilla $\mathbf{w}_i$ on paikkadimensio ja rivivektoreilla $\mathbf{y}_i$ on aikadimensio

Esimerkki: ilmastodata (3)

- Tämän voi kätevästi tehdä PCA:lla: huomaa että PCA-kehitemä kaavasta (5) antaa koko datamatriisille

$$\mathbf{X} - E\{\mathbf{X}\} = \sum_{i=1}^m \mathbf{e}_i \mathbf{e}_i^T (\mathbf{X} - E\{\mathbf{X}\})$$

ja tässä sarakevektoreilla $\mathbf{e}_i$ on paikkadimensio, rivivektoreilla $\mathbf{e}_i^T (\mathbf{X} - E\{\mathbf{X}\})$ on aikadimensio.

- Ominaisvektoreita $\mathbf{e}_i$, kun ne kuvataan graafisesti karttapinnalla, sanotaan kokeellisiksi ortogonaalifunktioksi (EOF).
- Oheinen kuva näyttää 3 ensimmäistä funktiota ja vastaavat aikakäyrät ilmanpainedatalle. (Ne on laskettu vain ajalle 1957–2000).

Esimerkki: ilmastodata

DSS-menetelmä: halutunlaisten aikakomponenttien etsiminen

- (DSS = denoising source separation)
- Selostetaan seuraavassa lyhyesti kuinka yo. menetelmää voi parantaa jos tiedetään minkälaisia aikakomponentteja halutaan
- Useat sääilmiöt ovat jaksollisia, esim. vuoden, 2 vuoden jne. jaksoja
- Tämä voidaan ottaa huomioon yhdistämällä PCA ja aikasuodatus
- Tehdään ensin säämittausmatriisille **X** pääkomponenttianalyysi kuten edellä selostetaan

Esimerkki: ilmastodata (2)

DSS-menetelmä: halutunlaisten aikakomponenttien etsiminen

- Skaalataan sitten aikakäyrät $\mathbf{e}_i^T(\mathbf{X} - E\{\mathbf{X}\})$ niin, että niiden varianssi (keskimääräinen energia) on yksi; käytännössä skaalaustekijä on $1/\sqrt{n\lambda_i}$ missä λ_i on ominaisvektoria $\mathbf{e}_i$ vastaava ominaisarvo
- Nimittäin

$$E\{\mathbf{e}_i^T(\mathbf{X} - E\{\mathbf{X}\})(\mathbf{X} - E\{\mathbf{X}\})^T \mathbf{e}_i\} = n\mathbf{e}_i^T \mathbf{C}_x \mathbf{e}_i = n\lambda_i$$

- Suodatetaan nyt skaalatut aikakäyrät taajuussuotimella jolla on haluttu taajuusominaisuus ja tehdään vielä toisen kerran pääkomponenttianalyysi
- Ensimmäinen pääkomponentti vastaa käyrää jolla suodatettuna on suurin varianssi (energia), ja jossa siis on eniten haluttuja taajuuksia.

Esimerkki: ilmastodata (3)

DSS-menetelmä: halutunlaisten aikakomponenttien etsiminen

- Näin löytyy hyvin El Niño -ilmiö; ks. kuvat.
- Ilmastodatasta saadaan näin paljastettua uutta tietoa
- Tämä saattaa olla uusi aluevaltaus ilmastomittausten analyysissä
- Ensimmäinen julkaisu (Ilin, Valpola, Oja) kesällä 2005

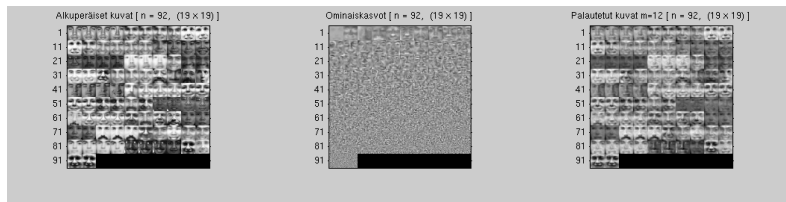
► (Ilin, Valpola, Oja 2005)

Yhteenveto

Tässä luvussa

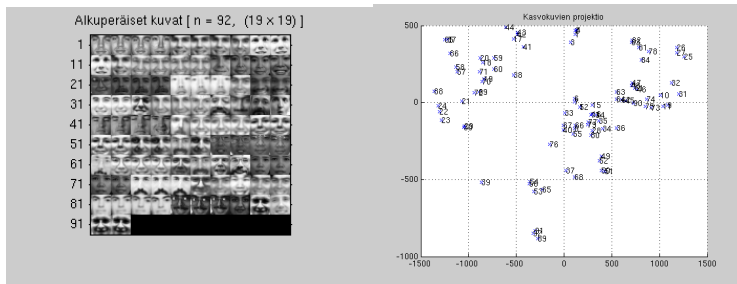
- esiteltiin pääkomponenttianalyysimenetelmä (PCA) datan dekorrelointiin ja koordinaatiston kiertoon maksimivarianssin suuntaan
- johdettiin ensimmäisen pääkomponenttivektorin suunta
- käytettiin PCA:ta lukuisissa tapauksissa mm. ominaiskasvojen (eigenfaces) tapauksessa

Ominaiskasvoja



Kuva: T2-harjoitusten datasetti: 92 kasvokuvaa resoluutiolla $19 \times 19 = 361$. Vasemmalla alkuperäiset kuvat. Keskellä ominaisvektorit eli ominaiskasvot. Oikealla palautetut kasvokuvat 12 ominaiskasvon perusteella. Alkuperäisessä datassa siis 92×361 lukuarvoa, kun taas pakkausta ja purkua varten tarvitaan 12 ominaiskasvoa ja 12D-projektio pisteet $12 \times (361 + 92)$.

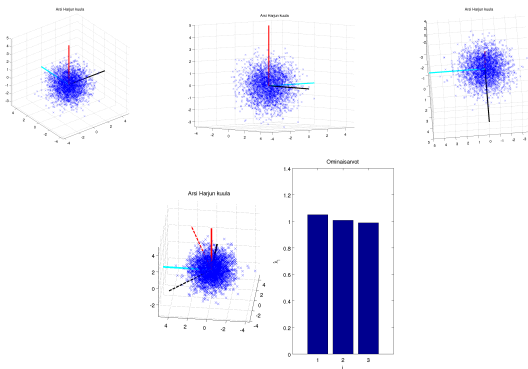
Kasvokuvien projektiio 2D-tasolle



Kuva: T2-harjoituksen esimerkki: (a) 92 alkuperäistä kasvokuvaa ja (b) niiden 2D-projektiot. Samasta henkilöstä otettujen valokuvien 2D-projektiopisteet ovat lähellä toisiaan.

PCA - koordinaattiakselien kierto

Arsi Harjun 3D-kuula ja geometrinen tulkinta



Kuva: Yläriivi: Arsi Harjun gaussinen kuula eri kulmista katsottuna. Alhaalla kuula PCA:n jälkeen. Ominaisarvot λ_1 , λ_2 , λ_3 melko samansuuruiset: yhtä satunnaista kaikissa suunnissa.

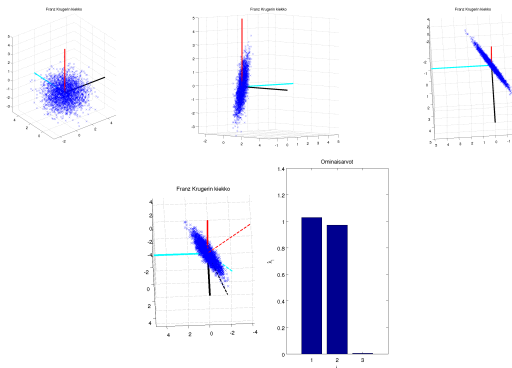
3D

2D

1D

PCA - koordinaattiakselien kierto

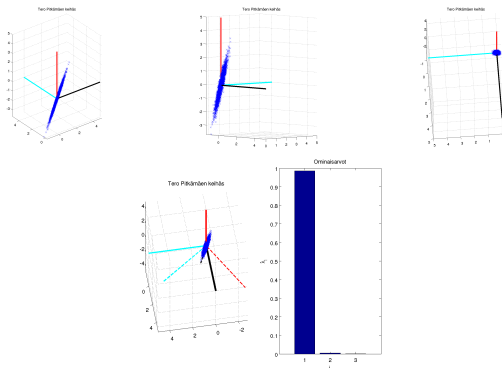
Frantz Krugerin 2D-kiekko ja geometrinen tulkinta



Kuva: Yläriivi: Frantz Krugerin gaussinen kiekko eri kulmista katsottuna. Alhaalla kiekko PCA:n jälkeen. PCA3:n suuntaan ominaisarvo λ_3 lähellä nollaa. ▶ 3D ▶ 2D ▶ 1D

PCA - koordinaattiakselien kierto

Tero Pitkämäen 1D-keihäs ja geometrinen tulkinta



Kuva: Yläriivi: Tero Pitkämäen gaussinen keihäs. Alhaalla keihäs PCA:n jälkeen PCA1:n suuntaisesti, jolloin vain ominaisarvo λ_1 on merkittävä. Dimension pudotus $3D \rightarrow 1D$.

▶ 3D

▶ 2D

1D

◀ Takaisin