

DOB valmennus

Data-analyysi

Esiprosessointi

24.08.2017

Jyrki Rasku cc by 4.0

<https://creativecommons.org/licenses/by/4.0/deed.fi>

6Aika



Uudenmaan liitto
Nylands förbund



Euroopan unioni
Euroopan aluekehitysrahasto

Vipuvoimaa
EU:lta
2014–2020

Data ja informaatio

Data liittyy johonkin asiayhteyteen ja se saattaa sisältää informaatiota

Datan sisältämä informaatio ei ole yksikäsitteistä, vaan sillä on erilainen merkitys informaation käyttäjälle

Dataan liittyvä oivallus on sen sisältämän itselle hyödyllisen informaation löytämisessä ja sen hyödyntämisessä

Esimerkkidatoja ja joitain tulkintoja

- Konekirjoitettu tekstidokumentti (asia)
- Käsinkirjoitettu tekstidokumentti (tunne)
- Ääni (varoitusta, assosiaatio)
- Haju (varoitusta, assosiaatio)
- Valokuva (kohde, kuvausmenetelmä)

Data-analyysin vaiheet

- Datan kerääminen->Aihealueen kattava data
- Datan esiprosessointi->Havaintomatriisi
- Tutkiva data-analyysi->Mitä datasta voi löytyä?
- Mallien valinta ja muodostaminen
- Mallin hyvyyden tarkastelu
- Mallin käyttö ja tilastotiedon ylläpito

Datan kerääminen

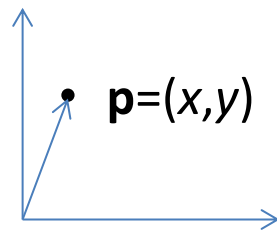
- Ennen datan keräämistä mieti tarkkaan, mihin kysymykseen etsit datan avulla vastausta
- Selvitä onko dataa saatavilla valmiina
- Missä muodossa data on? (Vanhhat levyt ja tiedostomuodot voivat aiheuttaa lisäkuluja)
- Tarvitseeko kaikkea dataa tallettaa? (Big Data)
- Miten saatat datan sellaiseen muotoon, että sitä voi tutkia

Datan luonne

- Riippumattomat havaintoyksiköt
- Riippuvat havaintoyksiköt
- Aikariippuvuus
- Tilariippuvuus

Havaintoyksikkö

- Yksittäistä havaintoa kuvataan havaintoavaruudessa vektorilla, joka lähtee havaintoavaruuden origosta ja päättyy johonkin havaintoavaruuden pisteeseen



Havaintomatriisi

- Havaintomatriisi sisältää riveinään kaikki tarkasteltavat havintoyksiköt
- Havaintomatriisin sarakkeet sisältävät tarkasteltavan ilmiön muuttujat
- $N \times P$ matriisissa on N havaintoa ja jokaisen havainnon dimensio on P
- Eheä havaintomatriisi saadaan datan esiprosessointivaiheen tuloksena
- Havaintomatriisin avulla tehdään kaikki analyysit

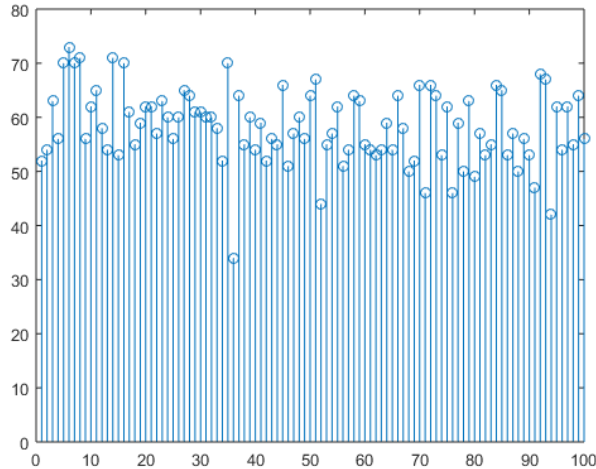
Muuttujien tyypit (1/2)

- Muuttujan tyyppi kertoo, mitä laskutoimituksia muuttujalla voidaan tehdä
- Nominaaliasteikko nimeää muuttujat. Yhtäsuuruus/erisuuruus ja moodi
- Järjestysasteikko järjestää muuttujat jonkin ominaisuuden mukaan. Suurempi, pienempi, lyhyempi jne...

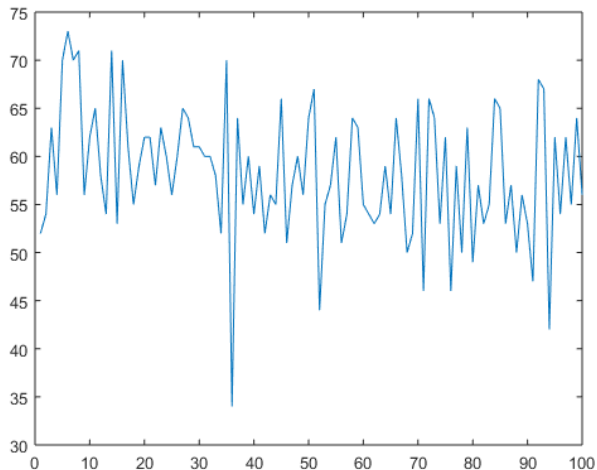
Muuttujien tyypit (2/2)

- Intervalliasteikko antaa järjestyksen lisäksi myös täsmällisen välimatkan muuttujien arvoille. Lämpötila Celsiusasteina
- Suhdeasteikko antaa täsmällisen välimatkan lisäksi myös tulkinnan sille, että jokin ominaisuus häviää kokonaan. Esim pisteeseen vaikuttava voima

Diskreetti ja jatkuva



Diskreetti data saa arvoja vain tietyissä kohdissa ja Sen arvoalue on rajoitettu. Digitaaliset järjestelmät ovat tyypillisesti diskreettejä.

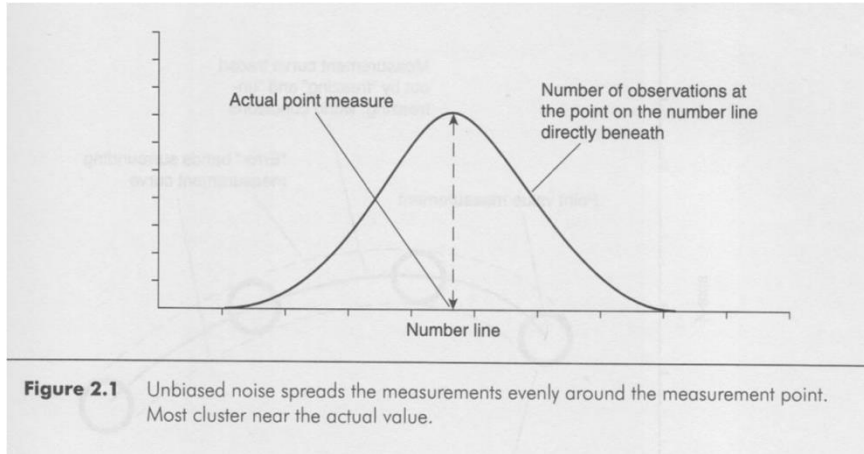


Jatkuva data voi saada mielivaltaisen tarkkoja arvoja missä tahansa kohdassa. Analogiset järjestelmät ovat tyypillisesti jatkuvia.

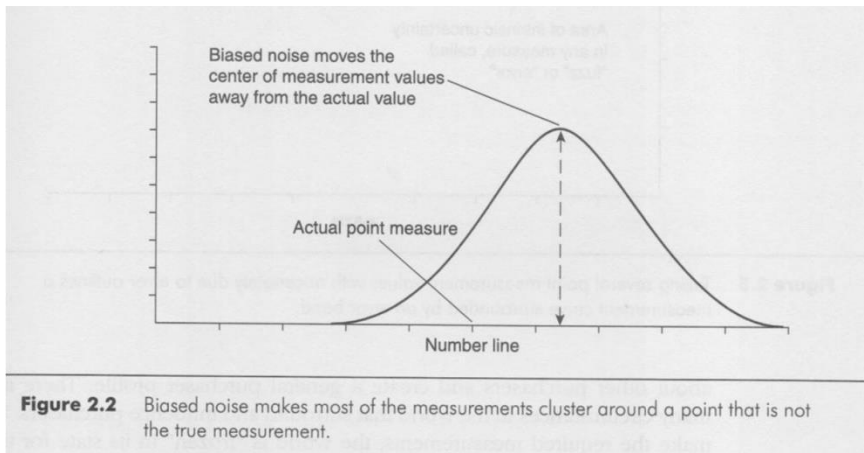
Mittausvirheet (1/6)

- Lähes kaikessa mittaamisessa on mukana virhettä, jota ei voida täysin poistaa
- Mittausvirhe on joko systemaattista tai satunnaista
- Systemaattinen virhe on hallittavissa, mutta satunnaismittausvirhettä voidaan hallita vain johonkin rajaan saakka. (Parempi mittalaite)

Mittausvirheet (2/6)



Satunnaisvirhe levittää havaintoja oikean mittauspisteen ympäristöön

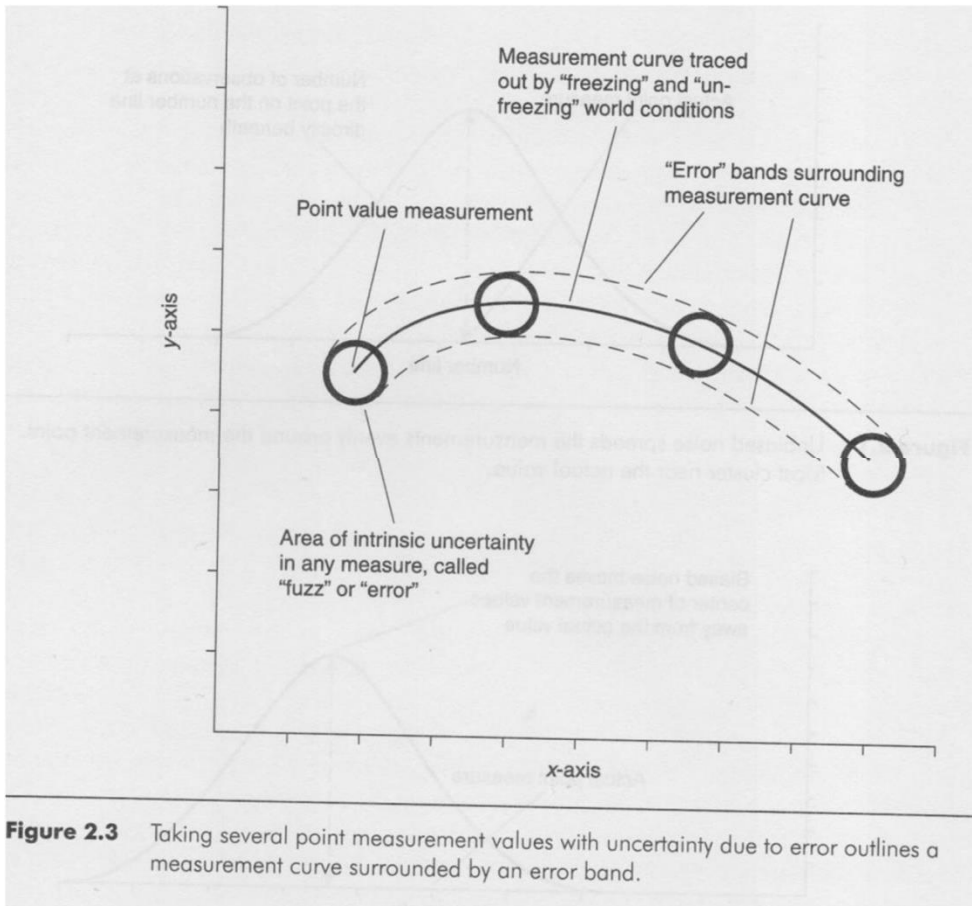


Systemaattinen virhe siirtää Satunnaisvirheen muodostamaa kuvaajaa oikean mittauspisteen suhteen

Mittausvirheet (3/6)

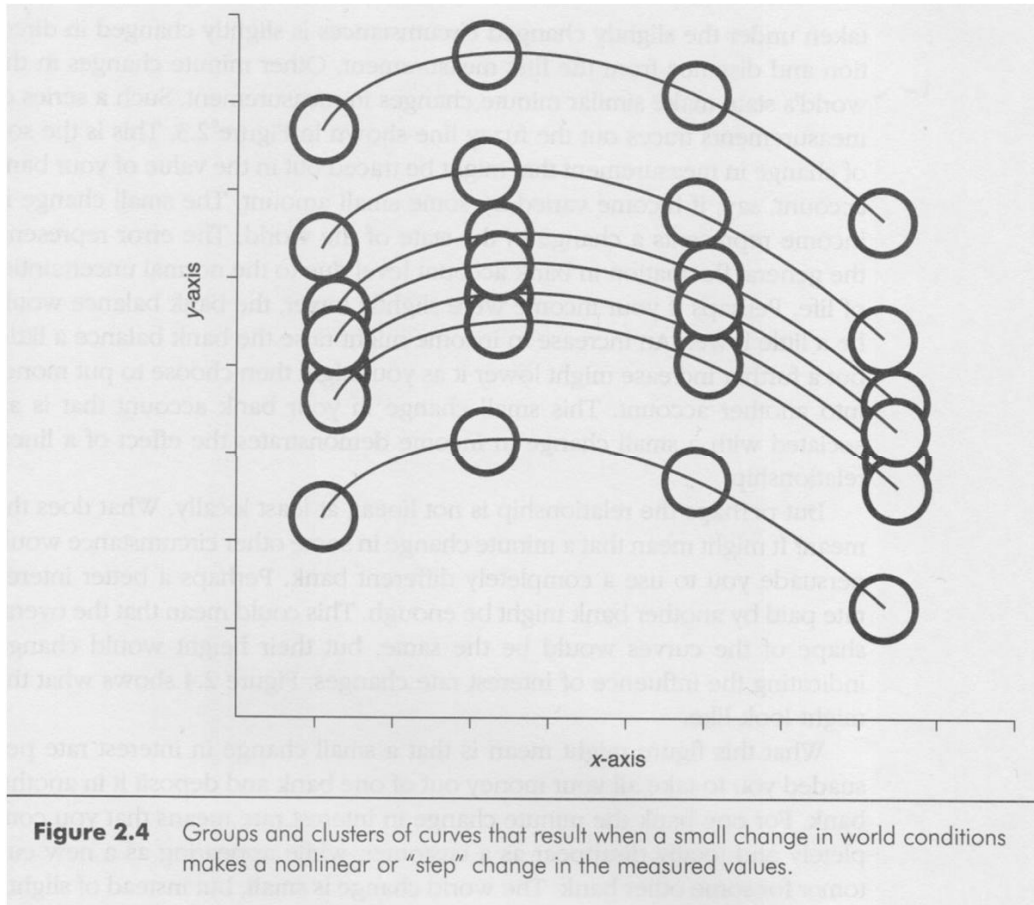
- Yksittäinen mittaus liittyy aina johonkin ajanhetkeen
- Analysoitava data voi vaihdella sellaisten olosuhteiden vaikutuksesta, joita ei datasta saa suoraan selville

Mittausvirheet (4/6)



Ajalliset ympäristön muutokset voivat vaikuttaa mittaustuloksiin. Esimerkiksi metallin työstökoneen ympäristön lämpeneminen

Mittausvirheet (5/6)

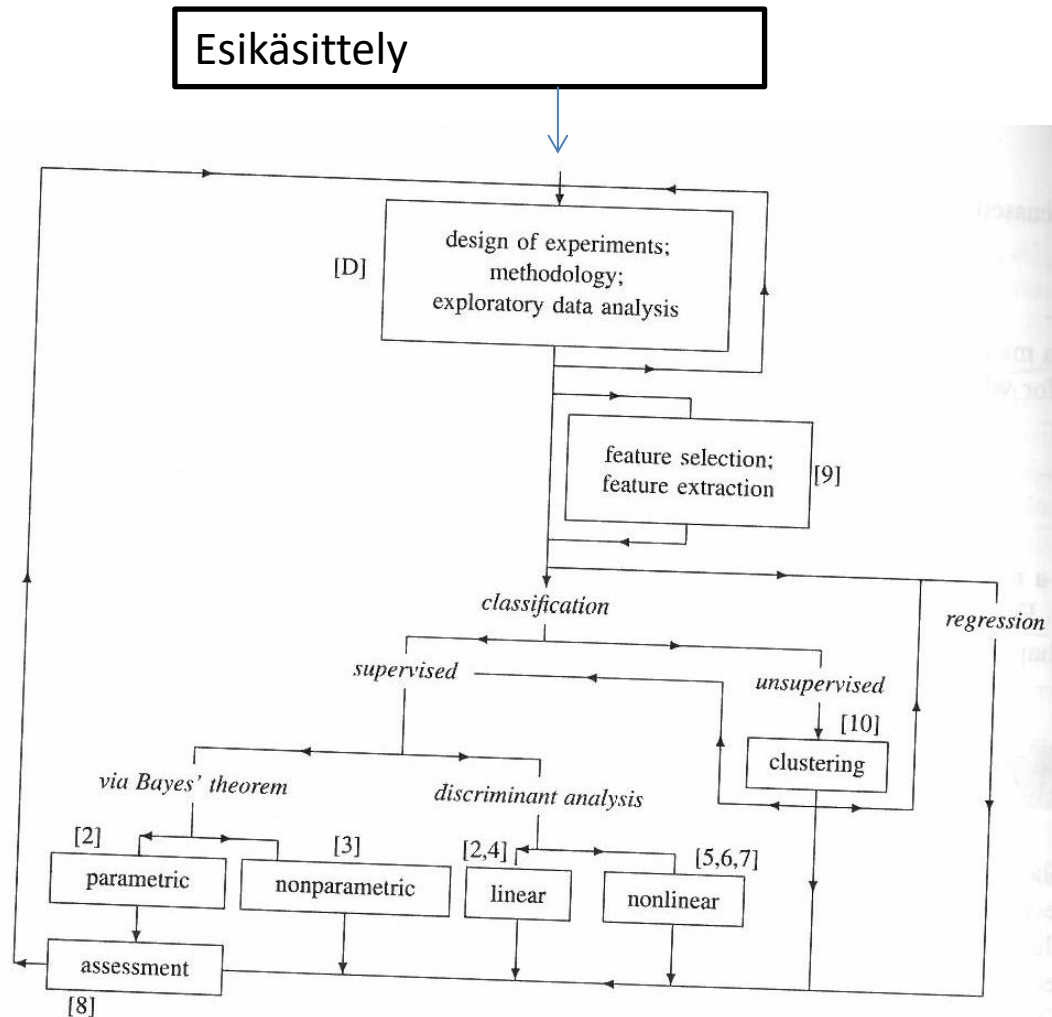


Rinnakkaiset havainnot samasta ilmiöstä sisältävät ajallisen ja paikallisen muutoksen aiheuttaman virheen. Esim. säähavaintoasemat

Mittausvirheet (6/6)

- Yksittäinen mittaus koostuu kahdesta osasta z ja $\hat{z}$. Tässä z on teoreettinen oikea mittausarvo ja $\hat{z}$ on mittausarvon estimaatti
- Nyt mittausvirhe $v = E(\hat{z}) - z$
- Näin ollen täsmällisen mittausvirheen määrittäminen on teoreettinen käsite ja sitä hallitaan valitsemalla tilanteeseen parhaiten sopiva mittalaite

Data-analyysin vaiheet



Kirjallisuudessa lähdetään usein liikkeelle valmiista datasta

Datan esikäsittely

- Puuttuvat arvot
- Poikkeavat havainnot
- Metriikat ja skaalaus
- Suodatus
- Muunnokset
- Piirteiden irroitus
- Datapisteiden esitysavaruuden ulottuvuuksien vähentäminen

Puuttuvat arvot (1/3)

- Usein joistakin havaintoyksiköistä yksi tai useampi muuttuja puuttuu kokonaan
- Muuttujan puuttuminen voi olla satunnaista tai sillä on syy, joka pitää ottaa huomioon analyysiä tehdessä
- Jos dataa on runsaasti, sellaiset havaintoyksiköt poistetaan datasta, joista muuttujia puuttuu (listwise deletion)

Puuttuvat arvot (2/3)

- Jos dataa on niukasti, niin puuttuvia muuttujien arvoja pitää korvata. Tätä menettelyä sanotaan imputoinniksi
- Keinotekoinen datan imputointi tuottaa epävarmuutta, mutta se mahdollistaa analyysin tekemisen

Puuttuvat arvot (3/3)

- Puuttuva muuttujan arvo voidaan korvata muiden havaintoyksiköiden ko. muuttujan keskiarvolla
- Muodostamalla regressiomalli muista muuttujista ja ennustamalla mallilla puuttuvaa arvoa
- EM (Expectation Maximization) algoritmilla. R Amelia II paketti

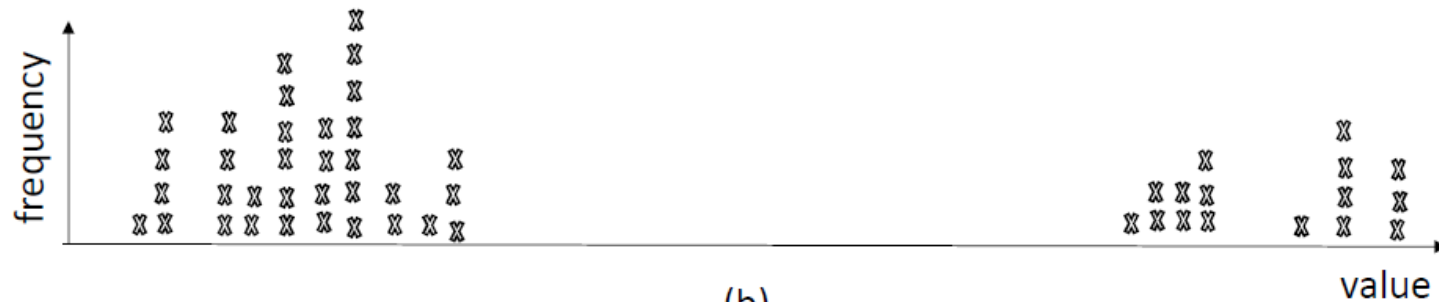
Poikkeavat havainnot (1/5)

- Poikkeava havainto on sellainen, jossa yksi tai useampi muuttuja saa sellaisia arvoja, jotka poikkeavat ”paljon” keskimääräisestä muuttujan arvosta (outliers)
- Poikkeava havainto voi olla järkevä tai virheellinen. Esim pulssi 30 ja 0
- Poikkeava havainto voi olla artefakti, eli keinotekoisesti mukana datassa

Poikkeavat havainnot (2/5)

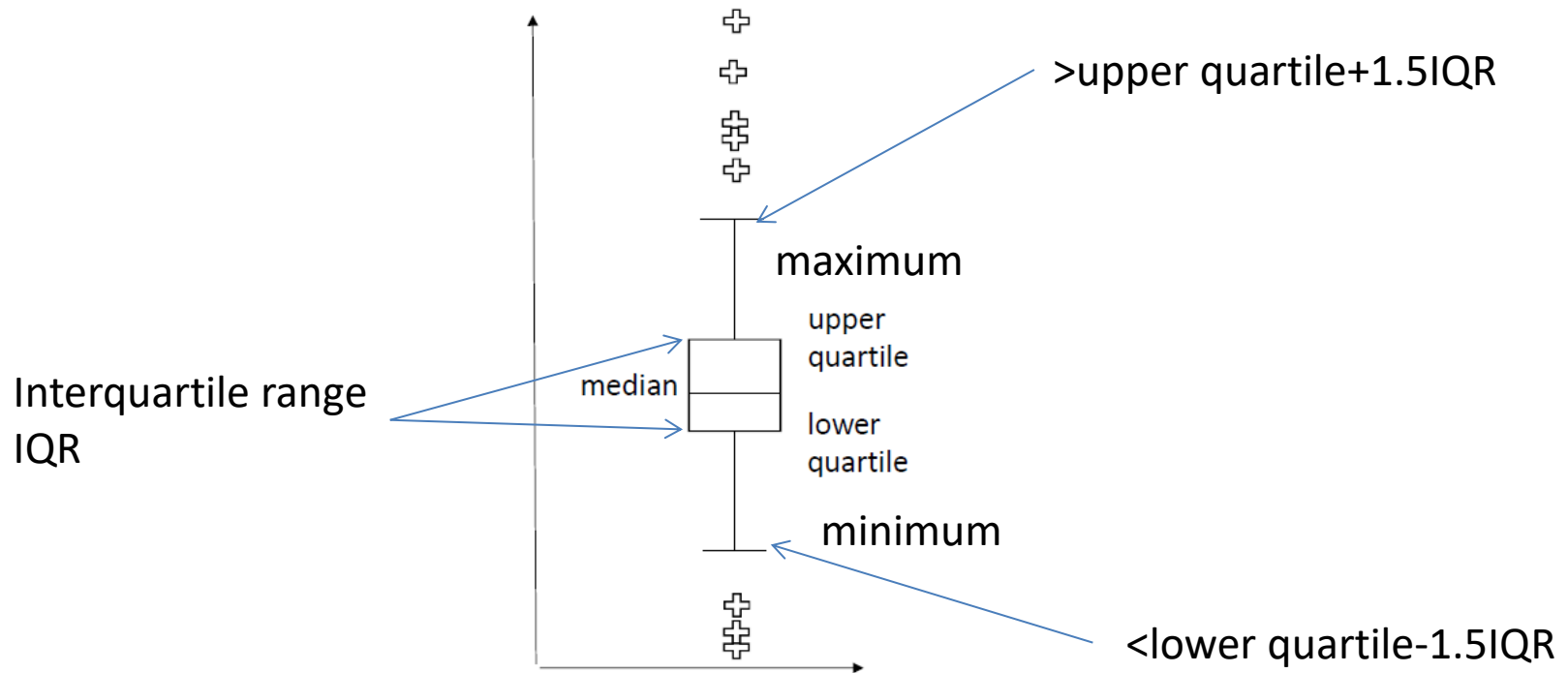


(a)



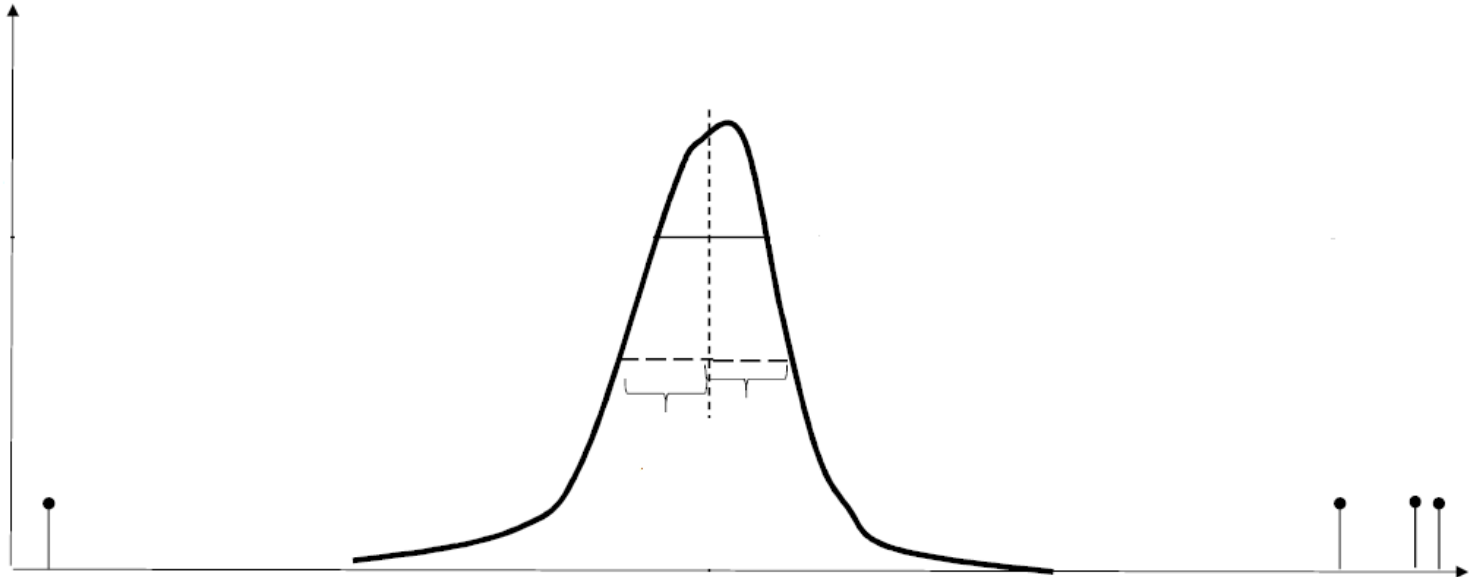
(b)

Poikkeavat havainnot (3/5)



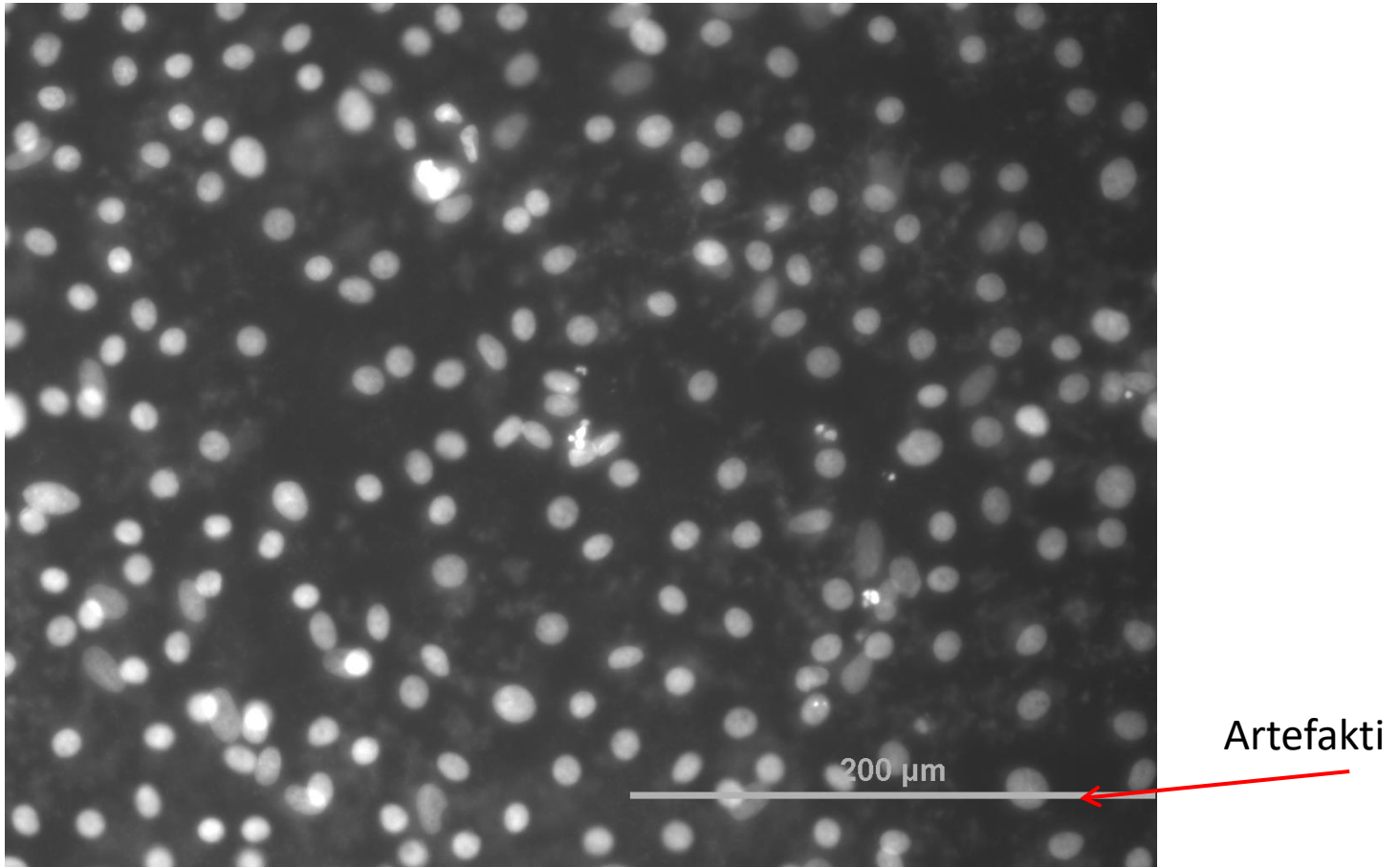
Laatikkokuvaaja (Box plot) on
hyödyllinen poikkeavien havaintojen
visualisoinnissa

Poikkeavat havainnot (4/5)



Normaalisti jakautuneen datan yhteydessä käytetään poikkeavan havainnon rajana kahta tai kolmea keskihajonnan mittaa keskiarvosta

Poikkeavat havainnot (5/5)



Maksasolujen tumia kuvattu mikroskoopin läpi

Laskennan perusideoita (1/2)

- Datapisteiden $\mathbf{x}=(x_1,\dots,x_n)$ ja $\mathbf{y}=(y_1,\dots,y_n)$ vertaileminen on mahdollista vertailemalla niiden samankaltaisuutta tai erilaisuutta
- Yleisesti tunnettu samankaltaisuuden mitta on pisteiden $\mathbf{x}$ ja $\mathbf{y}$ välinen Euklidinen etäisyys
- Jos datapisteiden samanlaisuus on $s(\mathbf{x},\mathbf{y})$, niin niiden erilaisuus voidaan ilmaista vaikkapa muodossa $ds=1-s(\mathbf{x},\mathbf{y})$

Laskennan perusideoita (2/2)

- Samankaltaisuudesta voidaan käyttää myös nimitystä läheisyys (proximity)
- Metriikka on hyvin määritelty kahden tai useamman pisteen välisen etäisyyden käsite

Metriikka

- (1) $d(\mathbf{x}, \mathbf{y}) > 0$ if $\mathbf{x} \neq \mathbf{y}$ positivity
- (2) $d(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} = \mathbf{y}$ reflexivity
- (3) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ symmetry
- (4) $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ for all vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ triangle inequality

Etäisyysmittoja (1/3)

Manhattan $d_1(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^p |x_k - y_k|$ Euclidean $d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^p (x_k - y_k)^2}$

Tshebyshev $d_3(\mathbf{x}, \mathbf{y}) = \max_k (|x_k - y_k|)$

Minkowski $d_4(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^p |x_k - y_k|^\lambda \right)^{1/\lambda}$

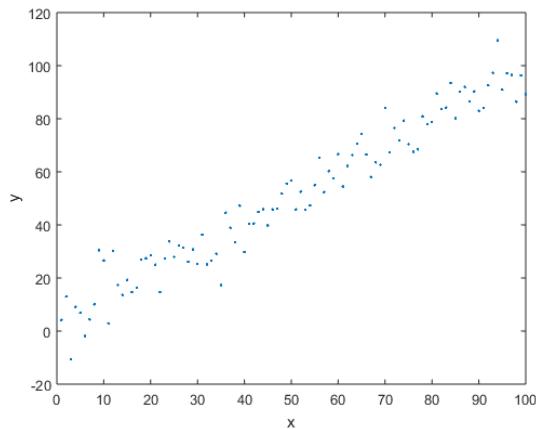
weighted Euclidean $d_{WE}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^p w_k (x_k - y_k)^2}$

Mahalanobis $d_{MH}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})}$

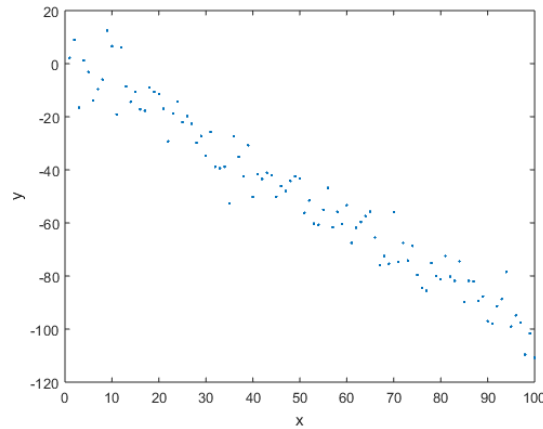
Etäisyysmittoja (2/3)

$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^p (x_k - \bar{x}_k)(y_k - \bar{y}_k)}{\sqrt{\sum_{k=1}^p (x_k - \bar{x}_k)^2 \sum_{i=1}^p (y_k - \bar{y}_k)^2}}$$

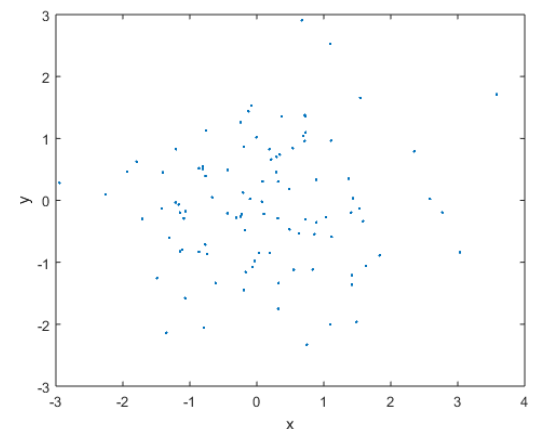
Korrelaatiokerroin r kuvaa vektoreihin $\mathbf{x}$ ja $\mathbf{y}$ sisältyvien yksittäisten muuttujien välistä lineaarista yhteisvaihtelua. (Ei mittaa riippuvuutta) Voi saada arvot $[-1 \ 1]$



$r \approx 1$



$r \approx -1$



$r \approx 0$

Etäisyysmittoja (3/3)

$$\text{cm}(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{k=1}^p x_k y_k}{\sqrt{\sum_{k=1}^p x_k^2 \sum_{k=1}^p y_k^2}} = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| |\mathbf{y}|}$$

Kosinimittaa käytetään usein tekstidokumenttien vertailussa

Samankaltaisuusmittoja (1/2)

Pällekkäisyys
mitta

$$s_1(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^p x_k y_k}{\min\left(\sum_{k=1}^p x_k^2, \sum_{k=1}^p y_k^2\right)}$$

Esimerkiksi vektorien $\mathbf{X}=(0,1,2)$ ja $\mathbf{y}=(1,0,0)$
pällekkäisyys on 0

$$s_2(\mathbf{x}, \mathbf{y}) = \frac{2 \sum_{k=1}^p x_k y_k}{\sum_{k=1}^p x_k^2 + \sum_{k=1}^p y_k^2}$$

Dice

$$s_3(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^p x_k y_k}{\sum_{k=1}^p x_k^2 + \sum_{k=1}^p y_k^2 - \sum_{k=1}^p x_k y_k}$$

Jaccard

$$s_4(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^p x_k y_k}{\sqrt{\sum_{k=1}^p x_k^2 \sum_{k=1}^p y_k^2}}$$

Kosini

Samankaltaisuusmittoja (2/2)

Gower similarity

$$s_5(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^p w_k u_k}{\sum_{k=1}^p w_k},$$

where $u_k = 1 - |x_k - y_k| / R_k$, $R_k = \max_k - \min_k$ for a quantitative variable and for a binary or nominal variable u_k is 1 if $x_k = y_k$ and 0 if $x_k \neq y_k$

w_k on 1, jos x_k ja y_k on olemassa ja 0, jos toinen puuttuu. Tämä pystyy siis käsittelemään puuttuvaa tietoa

Binaaridatan samankaltaisuus

- Binaariarvoisten havaintoyksiköiden samankaltaisuutta voidaan soveltaa, jos kaikki datan muuttujat ovat binaarisia

	$k=1$	$k=0$
v_1	$n_{1,1}$	$n_{1,0}$
v_2	$n_{0,1}$	$n_{0,0}$

Binaarimittoja (1/4)

$$H(\mathbf{x}, \mathbf{y}) = n_{1,0} + n_{0,1}$$

Hamming etäisyys kuvaa
binaarivektoreiden komponenttien
erilaisuuksien lukumäärää.

$$M(j, k) = \frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}}$$

Sokal ja Michener antaa täsmäävien kohtien suhteellisen
osuuden

Binaarimittoja (2/4)

$$J(j, k) = \frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}$$

Jaccardin täsmäyskerroin antaa täsmäävien ykkösten osuuden ja jättää täsmäyvät nollat pois

$$D(j, k) = \frac{2n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}} \quad \text{Dice}$$

$$RR(j, k) = \frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}} \quad \text{Russel ja Rao}$$

$$SS(j, k) = \frac{n_{1,1}}{n_{1,1} + 2(n_{1,0} + n_{0,1})} \quad \text{Sokal ja Sneath}$$

Binaarimittoja (3/4)

$$KR(j, k) = \frac{n_{1,1}}{n_{1,1} + n_{0,1}} \quad \text{Kulzinsky ja Rogers}$$

$$T(j, k) = \frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{0,0} + 2(n_{1,0} + n_{0,1})} \quad \text{Tanimoto}$$

Binaarimittoja (4/4)

$$Y(j, k) = \frac{n_{1,1}n_{0,0} - n_{1,0}n_{0,1}}{n_{1,1}n_{0,0} + n_{1,0}n_{0,1}}$$

Yulen
binaarikorrelaatio

$$BC(j, k) = \frac{n_{1,1}n_{0,0} + n_{1,0}n_{0,1}}{\sqrt{(n_{1,1} + n_{1,0})(n_{0,1} + n_{0,0})(n_{1,1} + n_{0,1})(n_{1,0} + n_{0,0})}}$$

Binaarikorrelaatio

Sekatyyppisten datojen mittoja

$$HEOM(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^p d_k(x_k, y_k)^2}$$

Heterogenous Euclidean overlap measure

$$d_k(x_k, y_k) = \begin{cases} 1, & \text{if } x_k \text{ or } y_k \text{ or both are missing} \\ \frac{|x_k - y_k|}{R_k}, & \text{where } R_k = \max_k - \min_k \\ \text{overlap}(x_k, y_k) = \begin{cases} 0, & \text{if } x_k = y_k \\ 1, & \text{if } x_k \neq y_k \end{cases} \end{cases}$$

Normalisointi ja skaalaus (1/3)

- Usein havaintoyksiköiden muuttujien arvoalueet poikkeavat huomattavasti toisistaan
- Autoa voidaan esimerkiksi kuvata vektorilla $A=(\text{paino}, \text{kulutus})$. Yksittäinen havainto voisi olla vaikkapa $A=(1300, 5)$ ja $B=(1500, 7)$
- Jos autojen A ja B samankaltaisuutta mitataan niiden välisellä Euklidisella etäisyydellä, niin saadaan $d(A, B) = \sqrt{(200)^2 + (2)^2}$

Normalisointi ja skaalaus (2/3)

- Etäisyyden $d(A,B)$ laskennassa painojen erotus saa tyypillisesti paljon suurempia arvoja, kuin kulutusten erotus. Tällön kulutuksen vaikutus etäisyyteen on pieni. Painon suuri arvo dominoi etäisyyden laskennassa.
- Jotta jokaisella muuttujalla olisi yhtä suuri vaikutus etäisyyden laskennassa, on muuttujien arvoalueet saatava samansuuruisiksi

Normalisointi ja skaalaus (3/3)

$$x_{ik}^{Norm} = \frac{x_{ik} - \min_{x_k}}{\max_{x_k} - \min_{x_k}}, \quad i = 1, \dots, n$$

i :nnen havaintoyksikön k :s muuttuja skaalataan välille $[0,1]$

$$x_{ik}^{St} = \frac{x_{ik} - \bar{x}_k}{s_k}, \quad i = 1, \dots, n$$

Normalisoinnissa jokaisen havaintoyksikön i :nnestä muuttujasta vähennetään ko. muuttujan keskiarvo ja erotus jaetaan ko. muuttujan keskihajonnalla. Normalisoitujen arvojen keskiarvo on 0 ja keskihajonta on 1.

Suodatus (1/2)

- Suodatus riippuu voimakkaasti sovellusalueesta, mutta perusajatuksena on poistaa, tai vähentää datassa olevaa häiritsevää ilmiötä
- Tekstidatan analyysissä voidaan poistaa tietyt sanat
- Digitaalisen signaalin käsittelyssä voidaan vaimentaa jotain signaalin ominaisuutta

Suodatus (2/2)

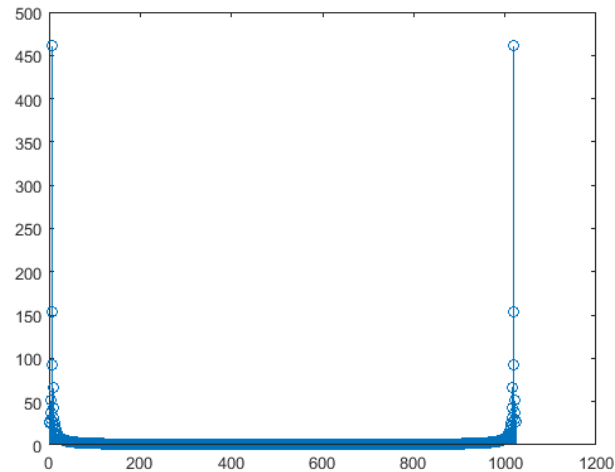
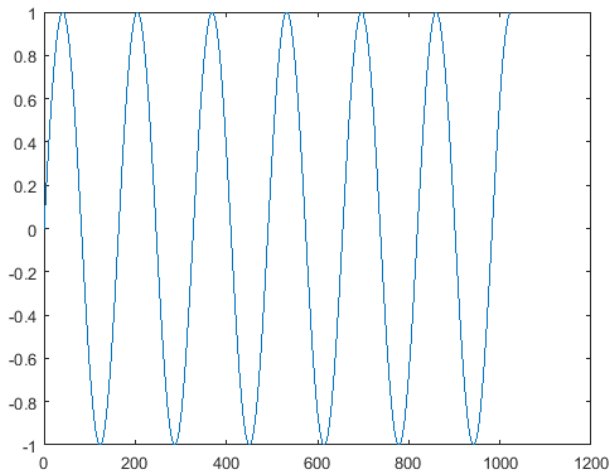
- Digitaalisen signaalin käsittelyssä helpoin suodatin on ns. liukuva keskiarvo
- Paremmin tilanteeseen sopivia suotimia voidaan suunnitella joko aika- tai taajuustasossa
- Usein digitoitava analoginen signaali suodatetaan valmiiksi jollakin laitetason ratkaisulla

Muunnokset (1/3)

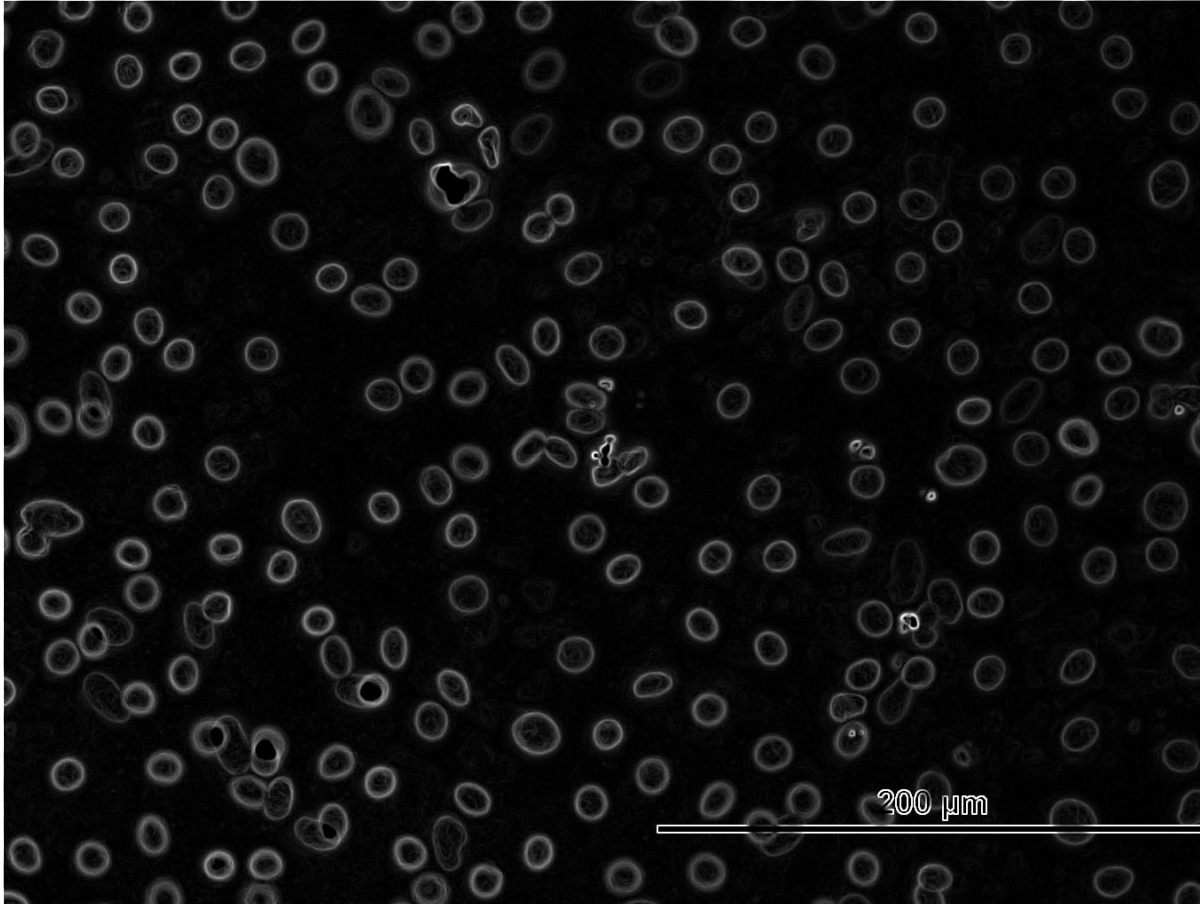
- Tilastollisissa menetelmissä dataan tehdään muunnoksia yleensä sen vuoksi, että datan jakauma saadaan halutuksi. Menetelmät olettavat yleensä jonkin jakauman, mutta monesti data ei sovi hyvin oletettuihin jakaumiin
- Toisaalta jokin muunnos voi tuoda datassa olevan mielenkiintoisen ominaisuuden paremmin esille

Muunnokset (2/3)

- Signaalinkäsittelyssä ja aikasarja-analyysissä yleisin muunnos lienee Fourier muunnos, jonka avulla pyritään selvittämään minkälaisia taajuuksia tarkasteltava signaali sisältää



Muunokset (3/3)

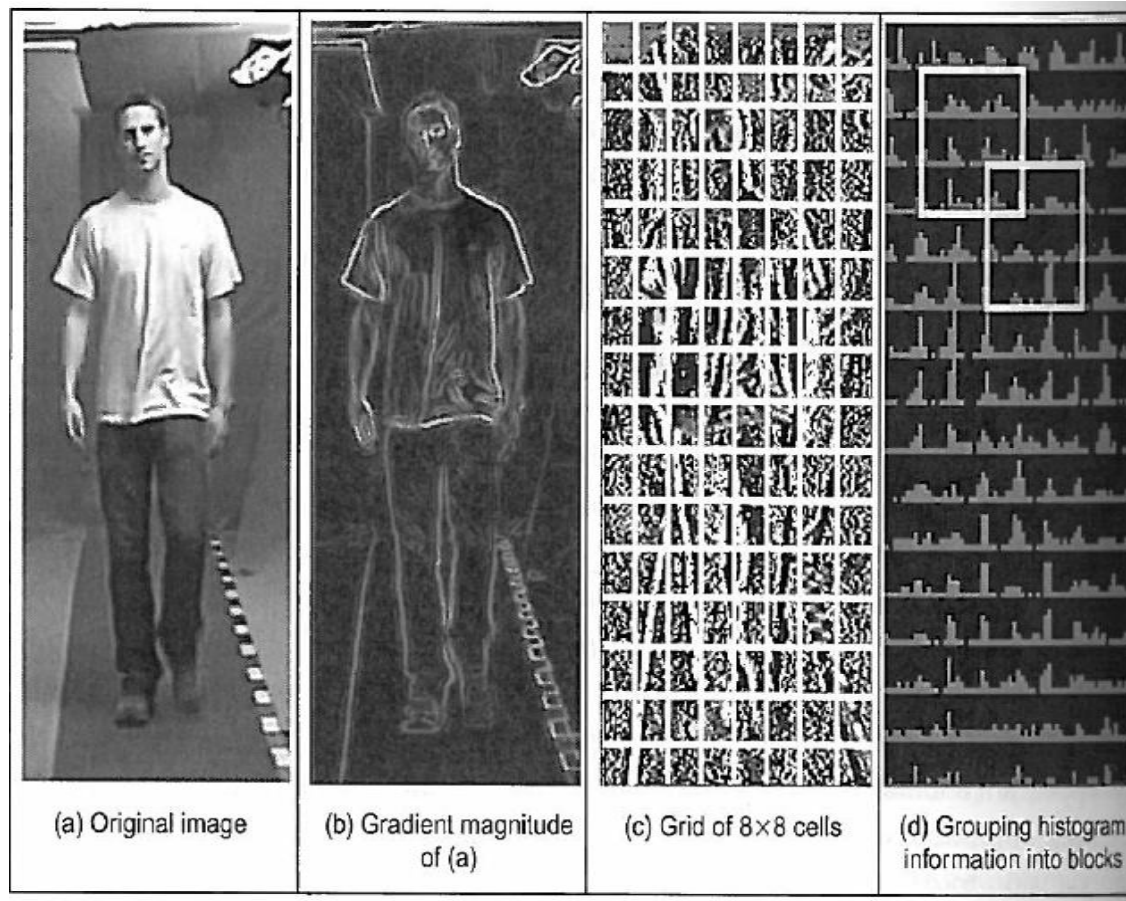


Maksasolujen tumien reunat vieräkkäisten harmaasävyjen muutosten voimakkuuksien avulla

Piirteiden irroitus (1/3)

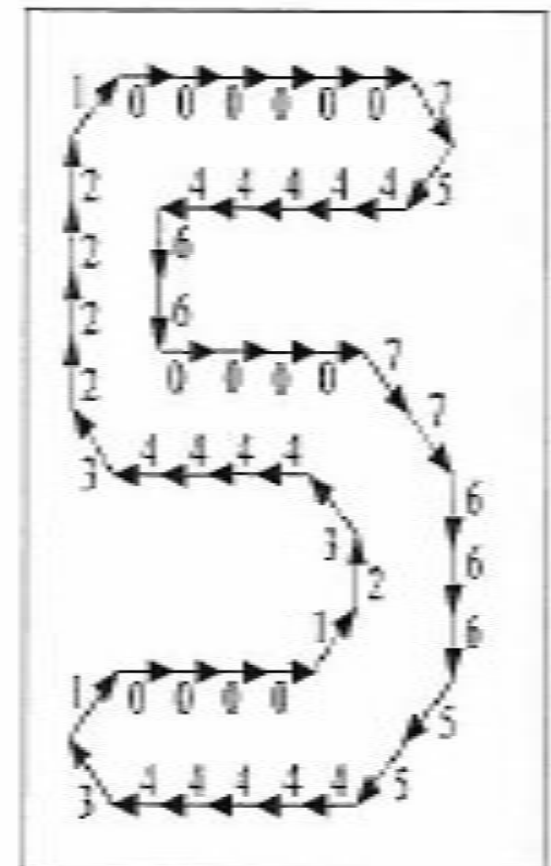
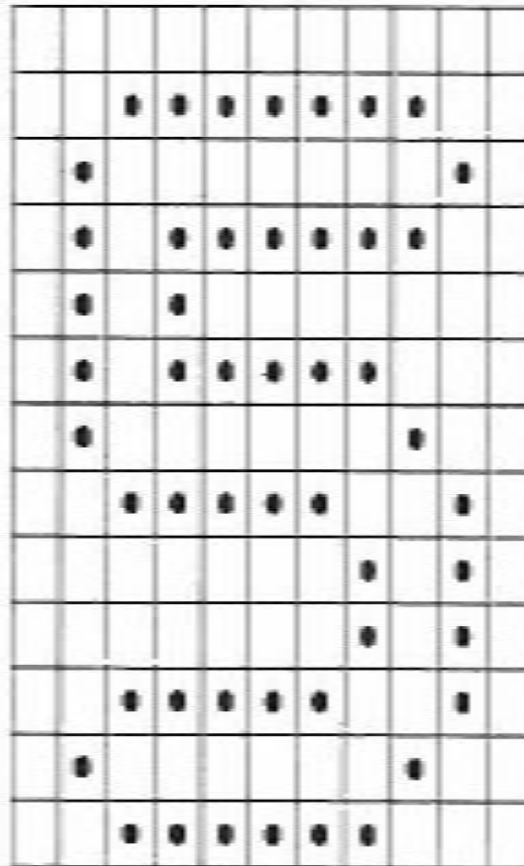
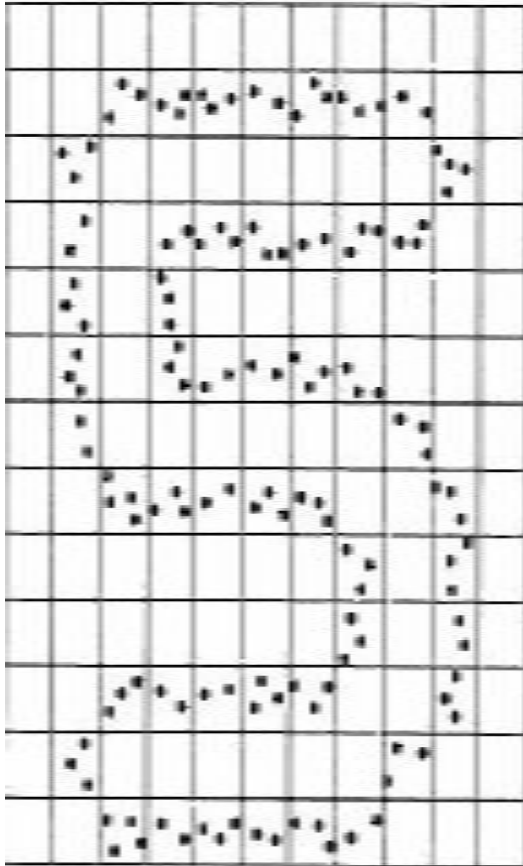
- Piirteiden irroitus tarkoittaa sitä, että havaintoyksiköistä valitaan sellaiset muuttujat, jotka ovat tehtävän analyysin kannalta oleellisia
- Valituille muuttujille voidaan edelleen tehdä tarvittavia muunnoksia
- Tässä vaiheessa on tärkeää kysyä neuvoa aihealueen asiantuntijalta. Toisaalta sovelluksen loppukäyttäjä voi tietää vielä paremmin

Piirteiden irroitus (2/3)



Digikuvan esittäminen paikallisilla histogrammeilla

Piirteiden irroitus (3/3)



Freemanin ketjukoodi kuvalle

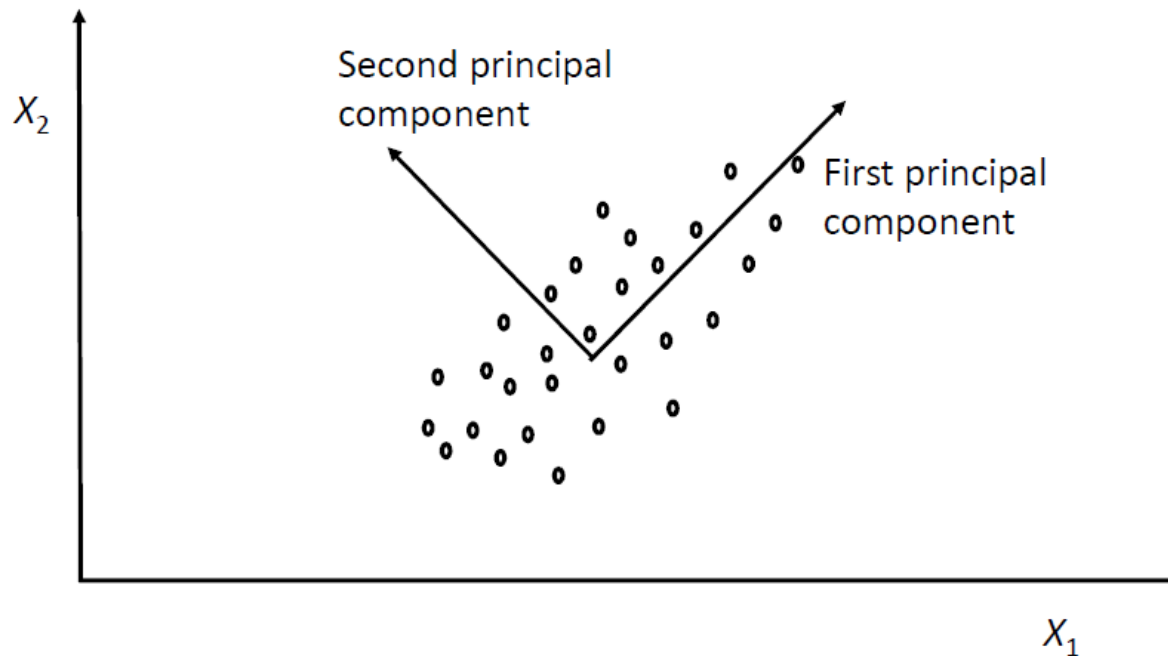
Datapisteiden tarkasteluavaruuden ulottuvuuksien vähentäminen (1/5)

- Kaikille analyysimentelmille on yhteistä se, että havaintoyksiköiden (havantovektorien) määrä N on suuri ja niiden muuttujien lukumäärä P on pieni
- Sovelluksen kannalta on usein selvää, että jotkin muuttujista voidaan jättää pois, koska ne eivät sisällä analyysin kannalta oleellista informaatiota

Datapisteiden tarkasteluavaruuden ulottuvuuksien vähentäminen (2/5)

- Jos jollain muuttujalla on sama arvo (vakio) kaikilla havaintoyksiköillä, niin se jätetään pois
- Jos jonkin muuttujan hajonta on ”pieni”, niin sen voi jättää pois
- Keskenään voimakkaasti korreloivista muuttujista voidaan osa jättää pois (mittaavat osittain samaa ilmiötä)

Datapisteiden tarkasteluavaruuden ulottuvuuksien vähentäminen (3/5)

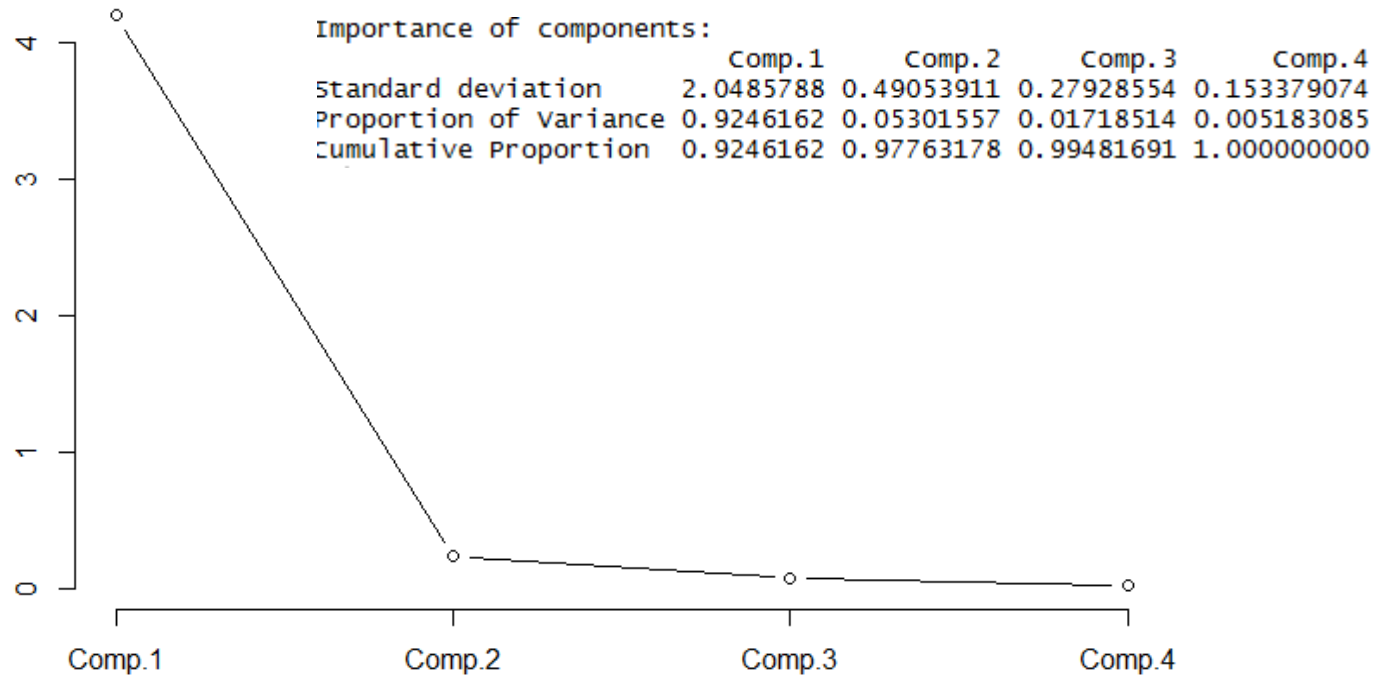


Pääkomponenttianalyysissä etsitään datalle uusi koordinaatisto siten, että ensimmäinen suunta on datan suurimman vaihtelun suunnassa

Datapisteiden tarkasteluavaruuden ulottuvuuksien vähentäminen (4/5)

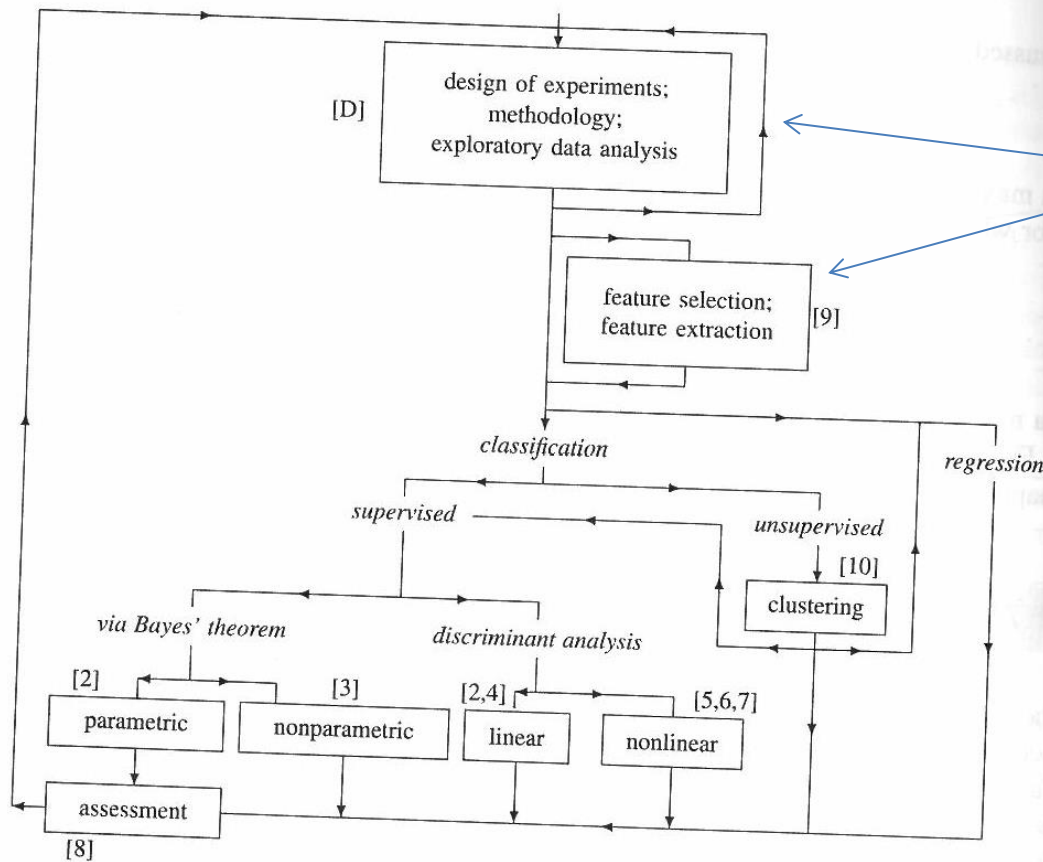
```
#Ladataan iris data
d<-iris[,1:4] #valitaan neljä ensimmäistä saraketta.
cm1<-cor(d)   #muuttujien korrelatio.
cm2<-cov(d)   #muuttujien kovarianssi.
pka<-princomp(d,cor=FALSE) #Tehdän PCA kovarianssimatriisista.
names(pka)    #Listaa pca:n tuottamat tiedot.
pka$sdev      #kovarianssimatriisin ominaisarvot
pka$loadings  #kovarianssimatriisin ominaisvektorit.
pka$center    #Alkuperäisten muuttujien keskiarvot.
pka$scores    #Muuttujien sijainti uudessa koordinaatistossa.
plot(pka$scores[,1])
plot(pka$scores[,2])
eigen(cm2)    #Kovarianssimatriisin ominaisvektorit.
summary(pka)  #pääkomponenttien osuudet varianssissa.
plot(pka,type="l")
```

Datapisteiden tarkasteluavaruuden ulottuvuuksien vähentäminen (5/5)



Tutkiva analytiikka (1/4)

Havaintomatriisi on nyt
valmis ja data korjattu

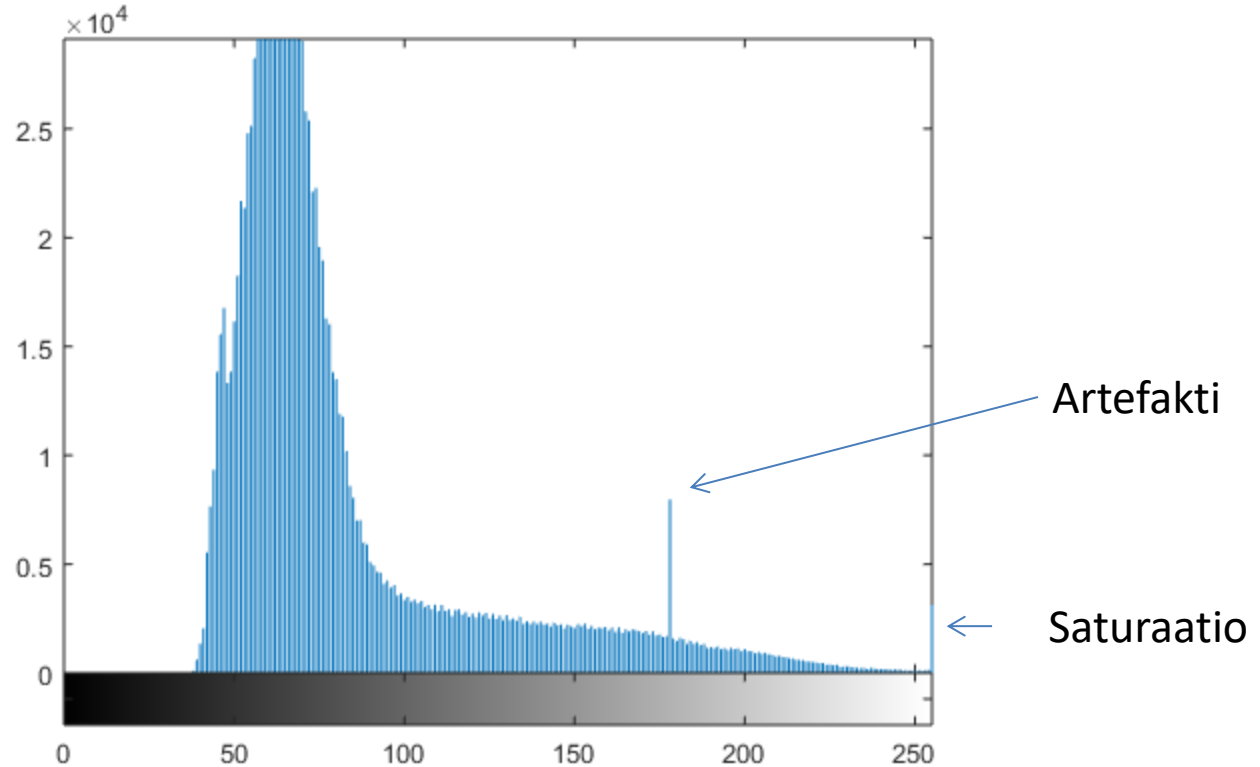


Iteratiivinen prosessi on
myös tässä kohdassa

Tutkiva analytiikka (2/4)

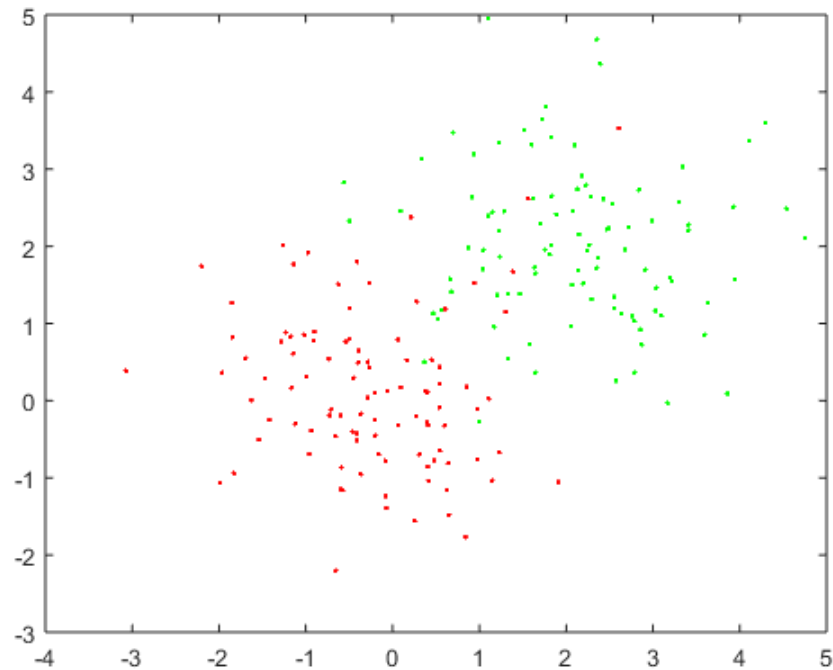
- Kun havaintomatriisi on valmis, dataa kannattaa visualisoida ja siitä voi laskea tyypillisimpiä tilastollisia tunnuslukuja. Keskiarvo, mediaani, moodi jne...
- Tutkivan analyysin tarkoituksena on miettiä mitä informaatiota datasta voi ylipäänsä löytää
- Tutkivan analyysin jälkeen tehdään varsinainen ennustava analyysi

Tutkiva analytiikka (3/4)



Maksasolutumien kuvan harmaasävyjen histogrammi

Tutkiva analytiikka (4/4)



Pisteparvella voi havainnollistaa muuttujien sijoittumista piirreavaruudessa

Ennustava analytiikka (1/3)

- Ennustavassa analyysissä pyritään ennustamaan sellaisen havainnon arvoa tai luokkaa, jota ei olla aikaisemmin havaittu
- Regressioanalyysissä ennustetaan muuttujien arvoja (kts. mallintaminen)
- Luokitteluanalyysissä ennustetaan havainnon luokkaa

Ennustava analytiikka (2/3)

- Nimestään huolimatta tarkoittaa usein sitä, että pyritään hakemaan oikea luokka jo tapahtuneelle ilmiölle. Esimerkiksi mikä on jokin käsin kirjoitettu kirjain skannatussa dokumentissa
- Ennustava malli luodaan opetusdatan avulla

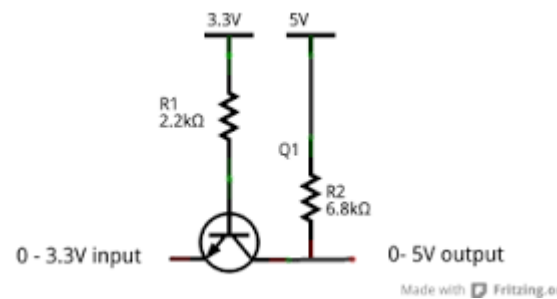
Ennustava analytiikka (3/3)

- Yleisimmin käytettyjä menetelmiä ovat päätöspuut, tukivektorikoneet ja neuroverkot

Mallintaminen (1/9)

- Mallintaminen tarkoittaa sitä, että tutkittavalle ilmiölle pyritään muodostamaan sellainen matemaattinen malli, joka ”mahdollisimman hyvin” kuvaa ilmiötä
- Ainoa oikeata mallia ei ole olemassa, mutta voi olla monia ”riittävän hyviä” malleja

Mallintaminen (2/9)



$$U=RI$$

Ideaalimalli

$$U=RI+e$$

Hieman todellisempi malli

Mitä on e ?

Mallintaminen (3/9)

- $U=RI+e$ on yksinkertainen lineaarinen malli, jossa osa RI kuvaa determinististä osaa ja e sellaista osaa, jota ei pystytä mallintamaan
- Usein ennustamatona osaa e kuvataan satunnaismuuttujana, joka noudattaa normaalijakaumaa
- Entä kaaosteoria?

Mallintaminen (4/9)

$$Y = aX + b + \varepsilon$$

Ensimmäisen asteen
yhtälö

$$y_1 = ax_1 + b + \varepsilon_1$$

$$y_2 = ax_2 + b + \varepsilon_2$$

$$\vdots$$

$$y_n = ax_n + b + \varepsilon_n$$

yhtälöryhmä

Minimoidaan virheiden ε_i neliöiden summa.
Paras malli saadaan, kun kokonaisvirhe on
mahdollisimman pieni.

Matrissimuodossa

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ a \\ + \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\bar{Y}$

$\bar{X}$

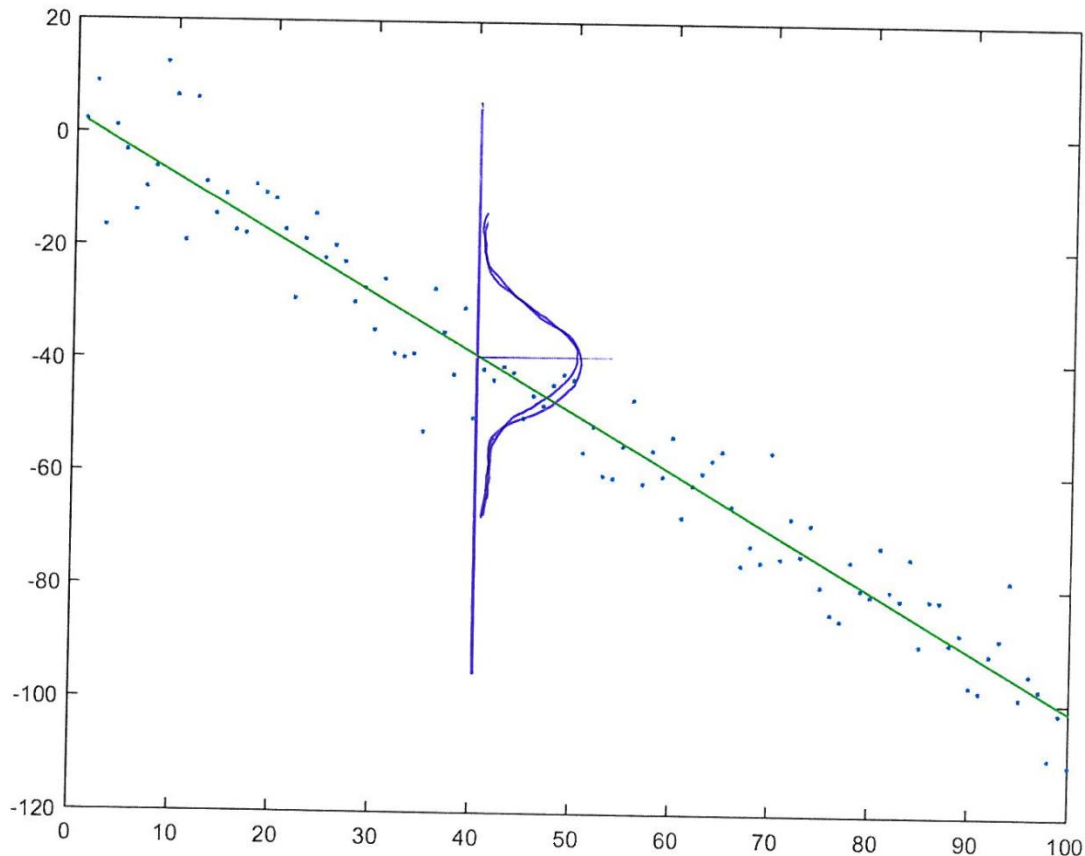
$\bar{b}$

Ratkaisu

$$\bar{b} = (X^T X)^{-1} X^T Y$$

Malli on parametrien
 a ja b suhteen
lineaarinen. Tässä
parit (x_i, y_i) ovat
tunnettuja.
Tuntemattomia ovat
parametrit a ja b .

Mallintaminen (5/9)



Satunnaisuuden normaalisuusoletus tekee vaihtelualueesta symmetrisen. Toisaalta monet ilmiöt ovat useiden muuttujien summia ja tämä tekee jakaumasta melko normaalin.

Mallintaminen (6/9)

Regressorimatrisi

$$y_i = ax_i + b$$

Yks. muuttuja.

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$y_i = ax_i + bz_i + c$$

Kaksi muuttujaa.

$$X = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{bmatrix}$$

Mallintaminen (7/9)

$$y_i = ax_i^2 + b$$

Toisen asteen malli.

$$X = \begin{bmatrix} 1 & x_1^2 \\ \vdots & \vdots \\ 1 & x_n^2 \end{bmatrix}$$

$$y_t = ay_{t-1}$$

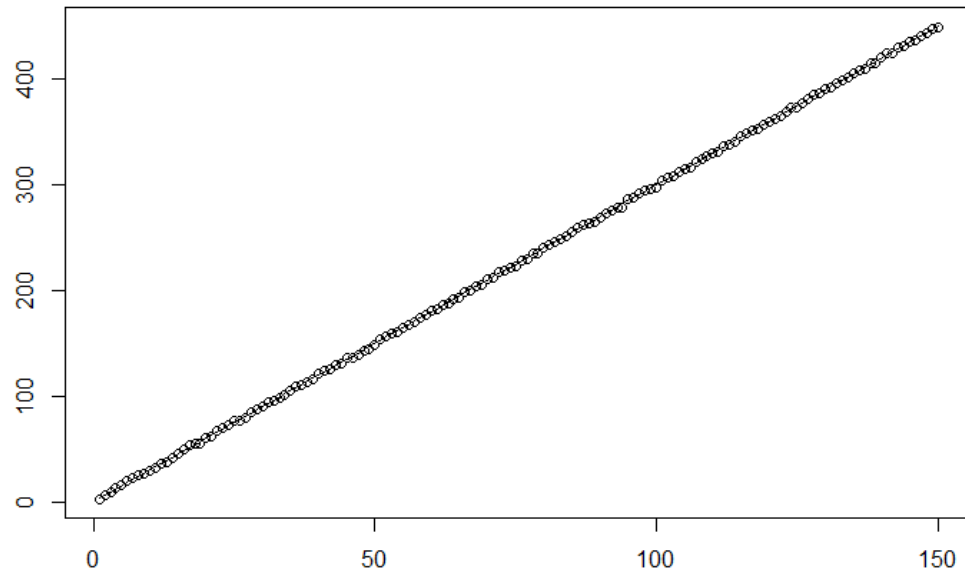
Aikasarja AR(1)

$$X = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Mallintaminen (8/9)

```
garb<-rnorm(150)
x<-seq(1:150)
y1<-3*x+garb|

mod<-lm(y1~x)
estimate<-fitted(mod)
res<-residuals(mod)
summary(mod)
```



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.01251	0.16535	0.076	0.94
x	3.00021	0.00190	1579.181	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.008 on 148 degrees of freedom
Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999
F-statistic: 2.494e+06 on 1 and 148 DF, p-value: < 2.2e-16

Mallintaminen (9/9)

```
garb<-rnorm(150)
x<-seq(1:150)
y1<-3*x+garb+5

mod<-lm(y1~x)
estimate<-fitted(mod)
res<-residuals(mod)
summary(mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.898393	0.175248	27.95	<2e-16	***
x	3.001709	0.002014	1490.77	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.068 on 148 degrees of freedom

Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999

F-statistic: 2.222e+06 on 1 and 148 DF, p-value: < 2.2e-16

Visualisointi ja raportointi

- Visualisointi ja raportointi kannattaa tehdä rinnakkain varsinaisen analyysin kanssa
- Tällöin raporttiin tulee kuvattua koko prosessi muuttujavalintoineen esiprossointimenetelmineen, analyysimenetelmineen ja tuloksineen. Vastaa alkuperäiseen tutkimuskysymykseen
- Kannattaa vaikkapa automatisoida
- Pitää huomioida raportin vastaanottaja