

Principale Componentenanalyse en 'hockeystick-short centring'

Peter Grünwald
HOVO 31-10 2011



Principale Componentenanalyse

- bedacht door Karl Pearson in 1901



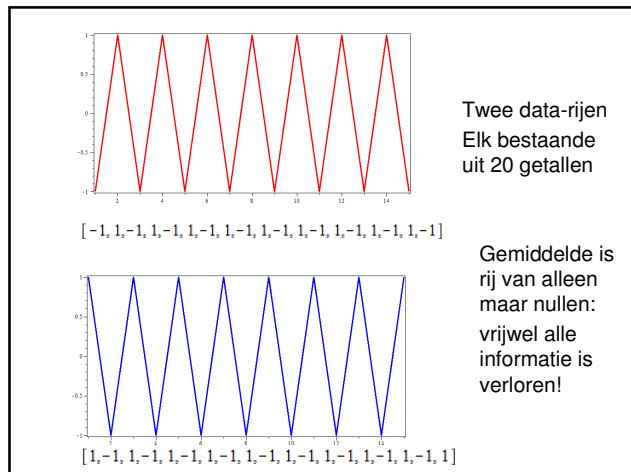
- Stel we hebben een grote hoeveelheid data. Elk "datum" is zelf weer een vector (reeks getallen)
- We willen deze data **samenvatten** tot een enkel **representatief "datum"** (of een paar representatieve "data") op een manier die **zoveel mogelijk "relevante" informatie behoudt**

Principale Componentenanalyse

- We willen deze data **samenvatten** tot een of een paar representatieve "data" op een manier die zoveel mogelijk "relevante" informatie behoudt
- We doen eerst alles met data die bestaat uit **twee rijen van getallen, die we tot een reduceren**. Dat kunnen we nl. mooi tekenen!
- Pas later kijken we naar wat we werkelijk willen:
 - data met **100en rijen die we tot een paar rijen (zeg 3, 4 of 5)** willen reduceren

Principale Componentenanalyse

- We willen twee rijen **samenvatten** in één rij
- Eerste idee: neem het **gemiddelde** van beide rijen
- **Dat gaat niet werken...**



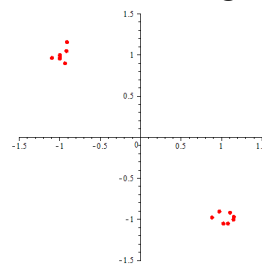
Principale Componentenanalyse

- We willen twee data rijen *samenvatten* in één rij
- Eerste idee: neem het gemiddelde van beide rijen- werkt niet!
- Tweede idee: gemiddelde van de twee rijen
 $[x_1, x_2, \dots, x_n] \quad [y_1, x_2, \dots, y_n]$
 is $[0.5x_1 + 0.5y_1, \dots, 0.5x_n + 0.5y_n]$
- Misschien dat een **andere lineaire combinatie**
 $[ax_1 + by_1, \dots, ax_n + by_n]$
 beter werkt? Dus een **soort gewogen gemiddelde**, maar dan **anders** (want gewichten kunnen bijv. negatief zijn)

Gecentreerde data

- We willen 2 rijen samenvatten in 1 rij zdd zoveel mogelijk relevante informatie behouden blijft
- We kijken voortaan naar “gecentreerde”data: **data waarbij iedere rij zelf (horizontaal opgeteld) gemiddelde waarde 0 heeft.**
- We kunnen deze data mooi tekenen in een twee-dimensionale grafiek (elke as staat voor een rij)

Voorbeeld: anti-gecorreleerde data



Dit is de data die we net ook al gezien hadden maar met een beetje “ruis” toegevoegd.

x -as: $[-1 + \epsilon_1, 1 + \epsilon_2, -1 + \epsilon_3, \dots, 1 + \epsilon_{20}]$
 y -as: $[1 + \delta_1, -1 + \delta_2, 1 + \delta_3, \dots, -1 + \delta_{20}]$

Roteren en As Verwijderen

- We kunnen gemiddelde nemen. Dat is niet goed.
- We kunnen gewoon een van de twee rijen weggooien. Dat is natuurlijk ook helemaal niet goed
- Maar wat als we eerst ons assenstelsel op een slimme wijze **roteren**, en dan een rij ("as") weggooien?

Maximale Variantie

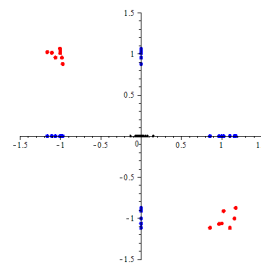
- Het gemiddelde van n punten is $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
(in ons geval: 0 in elke rij!)

Maximale Variantie

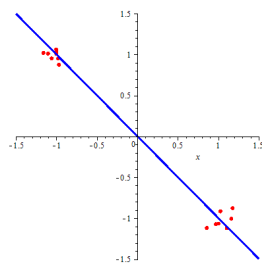
- Het gemiddelde van n punten is $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- We roteren het stelsel totdat de punten op de x-as een zo groot mogelijke **variantie** hebben:
(gemiddelde kwadratische afwijking van gemiddelde)

$$\text{VAR} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$$

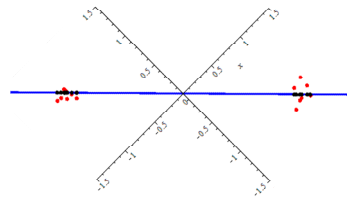
In ons geval!



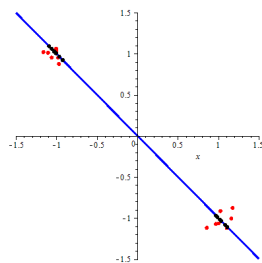
Zwart: gemiddelde puntenrij
Blauw: resultaat van weggooien x-as of y-as



De Principale Componentenas



De Geroteerde Principale Componentenas
De zwarte punten krijg je als je (nieuwe) y-as weggooit



De Terug-Geroteerde Principale Componentenas
De zwarte punten zijn nu je benadering van de oorspronkelijke punten

De Principale Component

- We roteren het stelsel totdat de punten op de x-as een zo groot mogelijke **variantie** hebben
Deze nieuwe x-as is de **principale component**
- Als we ook de rotatiehoek onthouden kunnen we de corresponderende y-as weggooien en het stelsel weer terugdraaien.
- Door de slimme keuze van de rotatiehoek is de verloren gegane informatie beperkt gebleven!

Andere interpretatie / meer dimensies

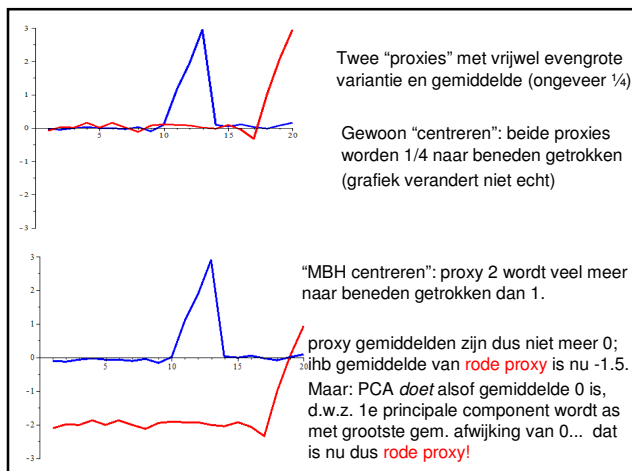
- PCA roteert de assen totdat data **ongecorreleerd** zijn.
 - equivalent: "PCA projecteert data op lijn door oorsprong die het meeste variantie verklaart"
 - let op: ongecorrleerd is niet per se ongerelateerd!
- Dit kun je natuurlijk met meer dan twee rijen (dimensies) doen.
- 1e principale component is de (geroteerde) as die het meeste van de "variantie in de data" verklaart,
- 2e principale component verklaart het "op een na meeste" variantie, en zo voort.

PCA en de Hockeystickgrafiek

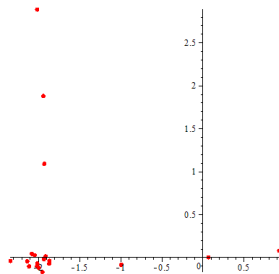
- Proxies: lange datarijen van bijv. boom-ring-diktes (bijv. van 1400 tot nu, een per jaar)
- Van sommige gebieden zijn veel meer 'proxies' beschikbaar dan van andere.
- MBH98 besluiten die proxies 'samen te vatten' zodat er van elk gebied ongeveer evenveel proxies zijn
- Ze nemen bijv. de 2 principale componenten van de NoAmer database (114 rijen) (later nemen ze er 5)

PCA is fout geprogrammeerd

- Het computerprogramma dat PCA implementeert bij Mann bevat een fout met serieuze consequenties: om PCA te kunnen gebruiken, moet je elke rij (proxy) eerst **centreren**: per rij het **gemiddelde** van **alle** punten aftrekken (zodat **nieuwe gemiddelde = 0**) in plaats daarvan trekken MBH98 het gemiddelde **in de jaren 1900-1980 van alle punten af**
- **Gevolg:**
als er een proxy met een hockeystickvorm tussen zit, dan wordt dat de 1e principale component!!!



De hockeystick proxy (x-as) wint!



Deel 2: De Normaal-Verdeling (ook wel: “Gauss-Verdeling”)



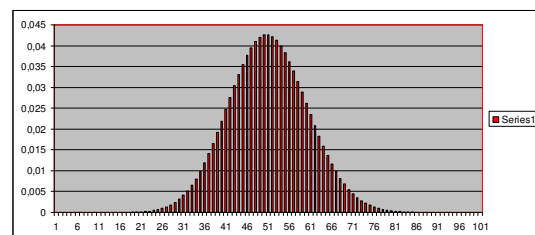
Leest allen:

“**Die Vermessung der Welt**”
van Daniel Kehlmann!

“Doe eens **Normaal**, Man!”

- In statistische analyses wordt heel vaak gedaan alsof meetbare hoeveelheden normaal verdeeld zijn
 - Lengtes
 - Meetfouten
 - Ruis bij data-communicatie
- Vaak klopt dat min of meer, vaak ook niet
 - Voorbeeld waarbij het klopt: de dobbelsteen
 - Voorbeeld waarbij het niet klopt: uitbijters

Herinnert u zich deze nog?

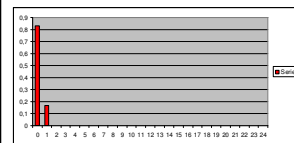


Die was dus eigenlijk **niet** precies normaalverdeeld!

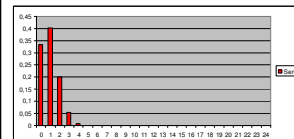
$\Pr(X \text{ zessen in } N \text{ worpen}) =$

$$\text{aantal rijen van } N \text{ worpen met } X \text{ zessen} \cdot \left(\frac{1}{6}\right)^X \cdot \left(\frac{5}{6}\right)^{N-X}$$

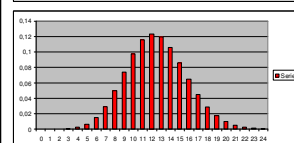
$$= \binom{N}{X} \cdot \left(\frac{1}{6}\right)^X \cdot \left(\frac{5}{6}\right)^{N-X}$$



1 keer gooien



6 keer gooien



75 keer gooien

Centrale Limietstelling

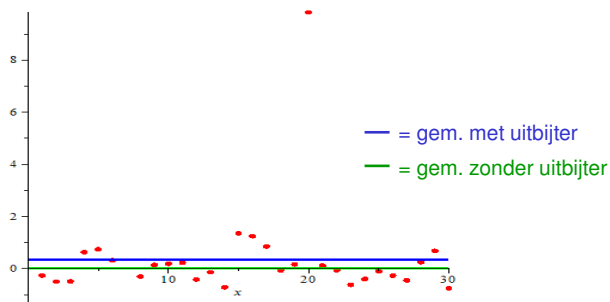
- Als X een optelsom is van vele kleine **onafhankelijke** bijdragen....dan is X "ongeveer" normaal (Gaussisch) verdeeld
 - de bijdragen zelf mogen op vrijwel alle mogelijke manieren verdeeld zijn, het hoeven bijv. echt geen dobbelstenen te zijn
- ...een van de belangrijkste stellingen van de kansrekening, of zelfs de gehele wiskunde
- Voor dit soort** X is het modelleren met de normaalverdeling een goed idee

Hoe **fit** ik een normaalverdeling?

$$f(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}(\bar{x}-\mu)^2}$$

- Standaard methodes (bijv. maximum likelihood): zet parameter μ ("theoretisch gemiddelde") op waarde van "empirisch gemiddelde", d.w.z. het gemiddelde dat je hebt waargenomen
- Probleem: data met **uitbijter(s)** (bijv. komma verkeerd...)

Een enkele Uitbijter trekt het gemiddelde flink omhoog



Visualisatie

- fitten van normaalverdelingen is extreem gevoelig voor de aanwezigheid van uitbijters
- Belang van **eyeballing**: data moet eigenlijk **altijd gevisualiseerd** worden, als **sanity check** of je geavanceerde statistische methode niet gewoon iets onzinnigs doet
 - Geldt niet alleen voor fitten van normaalverdeling, natuurlijk

Dunne Staarten

- Meer algemeen:
Normaalverdeling heeft twee extreem "dunne" (**exponentieel kleine**) staarten.
Bijv. kans dat je een uitkomst krijgt die meer dan 7x zo groot is als de "gemiddelde afwijking" is al **kleiner dan 1 op 100 miljard**
- Er zijn verdelingen die op het oog erg "normaal" lijken, maar veel "dikkere" (**polynomiale**) staarten hebben
 - Volgens sommigen is de Aziatische financiële crisis van 1998 mede veroorzaakt door speculatie op de beurs door computerprogramma's die "normale" aannames maken

Dikke Staarten

- Het voorbeeld van verdelingen met dikke staarten: **power laws**
- die zijn nu een hype, maar hebben ook zo hun problemen...
- ... Ze bieden nl. erg veel "flexibiliteit": je kunt vrijwel *elke* data set redelijk goed fitten met een power law (kleine kwadratische afwijkingen...)