Multivariate data analyse

BWI Werkstuk

augustus 2011

Marjolein Schipper

Begeleider: Martijn Onderwater

Vrije universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Master Business Mathematics & Informatics
De Boelelaan 1081a
1081 HV Amsterdam

Voorwoord

Het BWI-werkstuk is onderdeel van de master opleiding *Business Mathematics and Informatics* aan de Vrije Universiteit te Amsterdam. Voor het BWI-werkstuk wordt er zelfstandig een onderzoek uitgevoerd naar aanleiding van een praktisch probleem dat bedrijfsgerichte aspecten bevat, en wiskundige of informatica aspecten. Het werkstuk omvat grotendeels een literatuurstudie, aangevuld met praktijkstudie waarin de beschreven theorie in de praktijk wordt getoetst. De resultaten van het onderzoek zijn vastgelegd in dit verslag.

Dit werkstuk is geschreven onder begeleiding van Martijn Onderwater. Ik wil hem graag bedanken voor de enthousiaste hulp en ondersteuning die hij mij heeft gegeven bij het uitvoeren van het onderzoek en het schrijven van het verslag.

Marjolein Schipper
Augustus 2011

Management samenvatting

Door de groeiende complexiteit van de databases tegenwoordig is er veel vraag naar geautomatiseerde tools om data efficiënt en snel te analyseren. Het analyseren van deze grote hoeveelheden data wordt bemoeilijkt wanneer het aantal variabelen en het volume van de dataset erg groot is.

Een manier om deze grote, complexe datasbases te analyseren is door middel van multivariate data analyse technieken. Multivariate data analyse omschrijft gelijktijdige analyse van data met meer dan twee variabelen. Het voornaamste voordeel van multivariate data analyse is het vermogen om tussen meerdere variabelen tegelijk complexe relaties op te sporen.

In dit onderzoek is gefocust op twee multivariate technieken, namelijk principale componenten analyse (PCA) en hiërarchische clusteranalyse. Beide technieken hebben als voornaamste doel van de analyse het inzichtelijk maken van de structuur in de data. PCA is een multivariate analyse techniek, die door een groot aantal variabelen te beschrijven in een kleiner aantal lineaire combinaties, een beter inzicht geeft in de structuur van de data. Hiërarchische clusteranalyse is een multivariate techniek die groepen van waarnemingen of variabelen met soortgelijke karakteristieken vormt. Waar PCA zich richt op variabelen, zal clusteranalyse zich richten op het groeperen/clusteren van waarnemingen. Deze twee veelgebruikte statistische technieken zullen in dit onderzoek uitgelegd worden.

De bruikbaarheid van principale componenten analyse en hiërarchische clusteranalyse zal worden weergeven aan de hand van analyses toegepast op de Baseball dataset. De data bestaat uit 337 observaties; 337 Major League Baseball spelers die minstens 1 wedstrijd hebben gespeeld in de seizoenen van 1991 en 1992.

PCA biedt de onderzoeker een krachtige techniek in het bereiken van een beter begrip van de structuur van de data, maar ook een manier om de dataset te reduceren zodat het makkelijker bruikbaar wordt voor vervolganalyses.

Hiërarchische clusteranalyse is een techniek met een breed scala aan toepassingen. Tijdens het gebruik van deze techniek dient de onderzoeker rekening te houden met de juiste toepassing van de onderliggende principes. Wanneer deze keuzes op een juiste manier genomen worden, heeft de techniek het potentieel om onderliggende structuren in de data zichtbaar te maken die anders niet waren ontdekt.

Inhoudsopgave

1 Inleiding

Door de groeiende complexiteit van de databases tegenwoordig is er veel vraag naar geautomatiseerde tools om data efficiënt en snel te analyseren. Het analyseren van deze grote hoeveelheden data wordt bemoeilijkt wanneer het aantal variabelen en het volume van de dataset erg groot is. De univariate en bivariate data analyse technieken kunnen deze complexe analyses niet voldoende uitvoeren.

Dankzij de uitgebreide computer technologieën en de vele statistische programma"s van deze tijd is het mogelijk om zelfs de grootste, complexe datasets snel en relatief makkelijk te analyseren met multivariate data analyse technieken. Multivariate data analyse omschrijft "gelijktijdige analyses van data met meer dan twee variabelen" of "meten, verklaren en voorspellen van de mate van samenhang in gewogen combinaties van variabelen" (Hair et al., 1998). Deze statistische technieken worden op grote schaal toegepast in het bedrijfsleven, door de overheid en voor verschillende onderzoeken.

Multivariate data analyse bestaat uit een zeer breed gebied met vele toepassingen. In dit onderzoek zullen wij twee veelgebruikte statistische technieken bespreken, namelijk principale componenten analyse (PCA) en hiërarchische clusteranalyse. PCA is een multivariate analyse techniek, die door een groot aantal variabelen te beschrijven in een kleiner aantal lineaire combinaties, een beter inzicht geeft in de structuur van de data. Hiërarchische clusteranalyse is een multivariate techniek die groepen van waarnemingen of variabelen met soortgelijke karakteristieken vormt. Waar PCA zich richt op variabelen, zal hiërarchische clusteranalyse zich richten op het groeperen/clusteren van waarnemingen. Deze twee veelgebruikte statistische technieken zullen in dit onderzoek uitgelegd worden.

Het doel van dit onderzoek is inzicht verkrijgen in de eerder genoemde multivariate data analyse technieken. Eerst zal er literatuurstudie gedaan worden naar deze technieken. Vervolgens zal de bruikbaarheid van de technieken getoetst worden aan de hand van een praktische dataset in het programma R. Hierdoor zal nog een beter inzicht in de technieken verkregen worden.

In het volgende hoofdstuk zal gerelateerd werk besproken worden. In deze literatuurstudie zullen we beginnen met het bespreken van veelgebruikte toepassingen om één dimensionale data te analyseren, univariate data analyse technieken. Vervolgens zal er een introductie worden gegeven over de bivariate data analyse technieken voor twee dimensionale data. Tevens zal een korte introductie gegeven worden over multivariate data analyse technieken, PCA en hiërarchische clusteranalyse.

In hoofdstuk 3 zal een beschrijving gegeven worden van PCA en hiërarchische clusteranalyse. In het volgende hoofdstuk zal het experiment besproken worden. De gekozen dataset zal beschreven worden en de resultaten van de analyse van deze dataset in het programma R zullen weergeven worden. In het laatste hoofdstuk worden conclusies en eventuele aanbevelingen gegeven.

2 Gerelateerd werk

2.1 Univariate data analyse

Univariate data analyse is het analyseren van één dimensionale data. Het voornaamste doel van univariate data analyse is het beschrijven van de data. De meest gebruikte methodes zullen wij kort bespreken.

Een histogram is een univariate techniek die grafisch de frequentieverdeling van gegroepeerde data weergeeft. Een histogram geeft een beeld van de kansdichtheid waaruit de data afkomstig zijn.

Een stem-and-leaf plot is een zelfde soort methode als een histogram. Deze methode geeft de vorm van de verdeling van de data weer.

De boxplot is wederom een grafische weergave van metingen. Deze techniek laat de verdeling van de data zien. De mediaan wordt omgeven door de „box" met de kwartielen (de 25e en de 75e percentielen) die met een lijn is verbonden met de uiterste waarden van de metingen. De „box" vertegenwoordigt 50% van de waarnemingen (Tukey, 1977).

2.2 Bivariate data analyse

Bivariate data analyse is het analyseren van twee dimensionale data. Terwijl univariate data beschreven wordt, zal bivariate data meer verklaard worden.

Een bivariate data analyse techniek is de scatterplot. Deze techniek wordt al sinds de achttiende eeuw gebruikt. Het is een grafische weergave van bivariate data (twee variabelen) die inzicht geeft in de eventuele lineaire of kwadratische relatie tussen de variabelen. Een scatterplot geeft tevens informatie over het gemiddelde, de vorm van de verdeling, de extreme waarden van de data (outliers) en eventuele clustervormingen. Scatterplots zijn ook te gebruiken voor het plotten van trivariate data, data met drie variabelen.

Een andere techniek is om data samen te vatten is de *kxr* frequentie tabel. Deze techniek geeft voor beide variabelen alle mogelijke waarden van waarnemingen weer.

Voor verdere uitleg van univariate of bivariate methodes wordt aangeraden Tukey (1977) te lezen.

2.3 Multivariate data analyse

Univariate en bivariate data analyse technieken zijn makkelijk bruikbaar om de structuur en samenhang van de data te beschrijven. Van de besproken technieken zijn de meeste zeer nuttig voor de modellering stap binnen de data analyse, maar geven geen volledig beeld van de dataset. Een reden hiervoor is dat de besproken technieken zich niet focussen op hogere dimensies van de data, iets wat multivariate data analyse technieken wel doen (Hardle, 2007). Het voornaamste voordeel van multivariate data analyse is het vermogen om tussen meerdere variabelen tegelijk complexe relaties op te sporen.

Multivariate data analyse is een zeer breed gebied met vele toepassingen. Dit gebied is op te spitsen in twee hoofdgroepen: afhankelijkheids (dependence)- technieken waar afhankelijke en onafhankelijke variabelen onderscheiden worden en onderlinge afhankelijksheids (interdependence)- technieken waarbij geen onderscheid tussen afhankelijke en onafhankelijke variabelen wordt gemaakt. Dit onderzoek zal zich richten op PCA en hiërarchische clusteranalyse, welke beide tot de laatste groep behoren.

Het brede scala van multivariate technieken kent verschillende doeleinden van de data analyse zoals datareductie, groeperen van data, voorspelling, hypothese constructie/testen en onderzoek naar afhankelijkheid van variabelen (Johsson en Wichern, 2002).

PCA is in ontelbaar veel applicaties gebruikt en op vele gebieden toepasbaar. Hieronder volgen er een aantal:

- Quant et al. (2003) hebben met behulp van PCA de relatie tussen dagelijkse sterfte en belangrijke bronnen van luchtverontreiniging onderzocht.
- Zhao et al. (1998) hebben PCA toegepast voor gezichtsherkenning.
- Raychaudhuri et al. (2000) onderzoeken met behulp van PCA hoe genreacties variëren onder verschillende omstandigheden.
- Langley et al. (2010) hebben met behulp van PCA een algoritme voor het analyseren van ECG-morfologie getest.

Ook hiërarchische clusteranalyse is een breed toepasbare techniek. Hieronder volgen een aantal voorbeelden van toepassingen:

- Kok Sørensen et al. (2006) gebruiken hiërarchisch clusteren om de mate van financiële integratie in voornamelijk de bankensector van de eurolanden te meten.
- Green et al.(1967) gebruiken hiërarchische clusteranalyse om verschillende steden te groeperen voor test marketing doeleinden.
- Jain et al. (1992) hebben met behulp van hiërarchische clusteranalyse een methode ontwikkeld om tekst te identificeren in documenten.
- Moore et al. (2009) hebben door gebruik van hiërarchische clusteranalyse nieuwe fenotypes van astma geïdentificeerd.

3 Principale componenten analyse en hiërarchische clusteranalyse

3.1 Introductie

In dit hoofdstuk zullen de multivariate technieken principale componenten analyse en hiërarchische clusteranalyse besproken worden. Beide analyse technieken hebben als voornaamste doel het geven van meer inzicht in de structuur van de data. Hierdoor zal de data begrijpelijker worden en bruikbaar in vervolg analyses.

3.2 Principale componenten analyse

Principale componenten analyse (PCA) is een veelgebruikt statistische techniek voor het vinden van patronen in datasets met hoge dimensies. Het is een multivariate analysemethode die een grote hoeveelheid variabelen beschrijft door middel van een kleiner aantal lineaire combinaties, ook wel hoofdcomponenten of principale componenten genoemd. De techniek is in 1901 bedacht door Karl Pearson en verder ontwikkeld in 1930 door Harrold Hotelling (Chatfield, 1980).

De voornaamste doelstellingen van deze techniek zijn de structuur van de data inzichtelijk maken en datareductie. Wanneer er patronen in de data zijn gevonden, zal het aantal variabelen worden gereduceerd naar een kleiner aantal principale componenten. Er zal dan (bijna) evenveel informatie in deze componenten aanwezig zijn als in het origineel aantal variabelen. Verder laat PCA eventuele relaties tussen de variabelen zien die niet eerder waren opgevallen (Johnson, 2002).

3.2.1 Procedure voor PCA

Gegeven zijn de p variabelen in een dataset met random vector $X' = (X_1, X_2, \ldots, X_p)$. In PCA wordt er gezocht naar p lineaire combinaties $Y_1, Y_2, \ldots, Y_p$ die niet gecorreleerd zijn:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \text{ voor } i = 1, 2, \ldots, p. \qquad (3.1)$$

Y_i zijn de principale componenten. Van deze niet gecorreleerde principale componenten zijn de varianties zo groot mogelijk en geldt: $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \geq \text{Var}(Y_p)$. De eerste principale component is dus de lineaire combinatie met de grootste variantie, $\text{Var}(Y_1)$.

Laat $\boldsymbol{\Sigma}$ de covariantie matrix van X' met eigenwaarde-eigenvector paren $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ waarbij $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Dan is de i^{de} principale component gegeven door

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p \text{ voor } i = 1, 2, \ldots, p. \qquad (3.2)$$

De eigenwaardes λ_i van $\boldsymbol{\Sigma}$ geven de varianties van de principale componenten en de eigenvectoren e_i de coëfficiënten van de componenten. De totale variantie van de dataset is $\lambda_1 + \lambda_2 + \ldots + \lambda_p$, en dus het deel van de totale variantie verkregen door de k^{de} principale component is gelijk aan:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} \text{ voor } k = 1, 2, ..., p. \quad (3.3)$$

Als het grootste deel van de totale variantie kan worden toegeschreven aan de eerste één, twee of drie componenten, dan kunnen deze componenten de originele *p* variabelen van de dataset vervangen met minimaal verlies van informatie.

Wanneer de grootte van de variabelen erg veel verschillen, zullen de variabelen vooraf gestandaardiseerd worden. Dit gebeurt om dominantie van variabelen met een grote variantie te voorkomen. PCA is vrij gevoelig voor verschillen van grootte tussen variabelen.

Het is ook mogelijk PCA toe te passen op de correlatiematrix ρ in plaats van op de covariantie matrix **Σ**. In dit geval hoeven de variabelen vooraf niet gestandaardiseerd te worden. Voor de *p* geldt dan:

$$Z = \left(Z_1 = \frac{(X_1 - \mu_1)}{\sigma_1}, Z_2 = \frac{(X_2 - \mu_2)}{\sigma_2}, ..., Z_p = \frac{(X_p - \mu_p)}{\sigma_p} \right) \quad (3.4)$$

met Z_i de principale componenten. De eigenwaarden en eigenvectoren zullen berekend worden uit de correlatie matrix ρ van X. De eigenwaarde-eigenvector paren (λ_i, e_i), verkregen van de correlatie matrix ρ, zijn dus niet hetzelfde als degene verkregen uit de covariantie matrix **Σ**. In dit geval is de totale variantie

$$\sum_{i=1}^{p} Var(Z_i) = p \quad (3.5)$$

aangezien de variantie van elke variabele Z_i 1 is. Het deel van de totale gestandaardiseerde variantie verkregen door de k^{de} principale component is nu

$$\frac{\lambda_k}{p} \quad \text{voor } k = 1, 2, ..., p \quad (3.6)$$

met λ_k"s de eigenwaardes van ρ.

3.2.2 Stappen binnen PCA

Een belangrijke keuze binnen het PCA proces is het bepalen van het aantal principale componenten. Een veelgebruikt criterium is om de componenten te negeren op het punt dat het volgende component voor weinig toename in de totale verklaarde variantie zorgt. Een ander manier is om de principale componenten te kiezen die een bepaald deel van de variantie vertegenwoordigen, bijvoorbeeld 80% van de totale variantie. Een derde criterium is om de componenten met een variantie van minder dan 1 (wanneer de correlatiematrix gebruikt is) te negeren. Het idee hierachter is dat de genegeerde componenten weinig waardevolle informatie bevatten.

3.3 Hiërarchische clusteranalyse

Hiërarchische clusteranalyse is een verzameling van multivariate data analyse technieken die data groeperen en sorteren. Hierbij worden groepen/clusters van waarnemingen of variabelen met soortgelijke karakteristieken gevormd (Hair, 1998). Waar PCA zich richt op de variabelen, zal hiërarchische clusteranalyse zich richten op het groeperen van de waarnemingen. De data kan door hiërarchische clusteranalyse gesimplificeerd worden. Alle waarnemingen worden in een groep met gelijke kenmerken geplaatst, en worden zodoende als groep bekeken in plaats van als unieke observaties. De data is nu makkelijker bruikbaar voor vervolg analyses. Tevens kunnen onderliggende relaties onderscheiden worden dankzij hiërarchische clusteranalyse.

3.3.1 Stappen binnen het cluster proces

Hiërarchische clusteranalyse bestaat uit drie belangrijke stappen. Eerst zal de gelijkenis tussen de waarnemingen worden bepaald. Vervolgens zullen op basis van deze gelijkenis metingen de waarnemingen toegewezen worden aan een cluster met als doel dat de verschillen tussen de clusters zo groot mogelijk zijn en de waarnemingen in de cluster zo dicht mogelijk bij elkaar liggen. Tenslotte zal het aantal clusters in de eindoplossing bepaald worden.

3.3.2 Gelijkenis meten

Wanneer waarnemingen geclusterd zijn, kunnen de gelijkenissen tussen de waarnemingen op verschillende manieren worden bepaald. In de meeste gevallen wordt dit gemeten aan de hand van de afstand tussen twee *p*-dimensionale waarnemingen met $x' = [x_1, x_2, ..., x_p]$ en $y' = [y_1, y_2, ..., y_p]$. Er zijn verschillende definities van afstand mogelijk. Vergelijkingen (3.7) t/m (3.9) geven de meest gebruikte afstandsmaten weer.

Euclidische afstand:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} \qquad (3.7)$$

Gestandaardiseerde euclidische afstand:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^{\,2}} + ... + \frac{(x_p - y_p)^2}{\sigma_p^{\,2}}} \qquad (3.8)$$

met $\sigma_i^{\,2}$ de variantie voor $i = 1,...,p$.

Minkowinski afstand:

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{1/m} \qquad (3.9)$$

Voor m =1 geeft het de *city block afstand*. Voor m = 2 is het de *euclidische afstand*.

De meest gebruikte afstandmaat is de *euclidische afstand*. Andere gebruikte afstandsmaten zijn Canberra en Czekanowiski (Johnsson, 2002).

Er zijn dus verschillende manieren om gelijkenis te meten. De keuze van welke gelijkenis meting te gebruiken binnen het clusterproces is belangrijk. Deze keuze hangt sterk af van de gebruikte data en de keuze van de onderzoeker. In sommige gevallen is het beter om clusters van variabelen te vormen. Voor het meten van gelijkenis tussen variabelen worden correlatie metingen gedaan. Deze techniek wordt echter nauwelijks gebruikt (Johnson, 2002).

De aard van de variabelen speelt ook een belangrijke rol in de keuze van de gelijkenismeting. In Hardle(2007) worden gelijkenismaten specifiek voor binaire en continue variabelen omschreven.

Aan de hand van de gelijkenis metingen zullen de waarnemingen toegewezen worden aan een cluster. Hier kunnen verschillende technieken voor worden geselecteerd. Dit is tevens een belangrijke keuze binnen het clusterproces. Alle technieken zullen ervoor zorgen dat de verschillen tussen de clusters zo groot mogelijk zijn en de waarnemingen in de cluster zo dicht mogelijk bij elkaar liggen. In dit onderzoek zullen de hiërarchische technieken worden besproken.

Hiërarchische methoden zijn onder te verdelen in agglomerende methoden en splitsingsmethoden. Agglomerende methoden starten met evenveel clusters als individuele waarnemingen. Alle waarnemingen zitten als het ware in een eigen cluster. Eerst wordt er een cluster gevormd van de twee waarnemingen met de kortste afstand. Vervolgens wordt er gekeken naar de volgende kleinste afstand voor de vorming van een nieuwe cluster of verbreiding van een bestaande cluster. Dit gaat zo door totdat alle waarnemingen zijn toegewezen aan één cluster.

De splitsingsmethoden werken precies andersom. De methode start met een grote cluster met alle observaties erin. Het splitst dit cluster op in steeds kleiner wordende clusters totdat er evenveel clusters als waarnemingen zijn. Omdat agglomerende methoden het meest gebruikt worden, en splitsingsmethoden bijna hetzelfde werken, maar dan andersom, zullen wij ons in dit onderzoek focussen op agglomerende methoden.

Er zijn een aantal veelgebruikte agglomerenende methodes om clusters te ontwikkelen, namelijk:

- Single linkage (nearest neighbor): Groepen worden gevormd door het mergen van dichtstbijzijnde waarnemingen.
- Complete linkage: de afstand tussen de clusters wordt bepaald door de afstand tussen twee waarnemingen, beide in een andere cluster, die het meest van elkaar af liggen. Deze techniek zorgt er dus voor dat er een maximale afstand tussen de waarnemingen in een cluster zit.
- Average linkage: de afstand tussen twee clusters is de gemiddelde afstand tussen alle paren van waarnemingen.
- Ward‟s methode: minimaliseert het verlies van informatie
- Centroid methode: de afstand tussen de clusters wordt bepaald door de afstand tot de gemiddelde waarneming in de cluster.

Deze methoden verschillen in hoe de afstand tussen de clusters wordt bepaald. Er is geen enkele methode die de voorkeur heeft. De keuze voor een methode wordt beïnvloed door de gebruikte dataset.

In dit onderzoek zullen wij Ward"s methode gebruiken. Deze methode is ontwikkeld door Joe Ward(1963). Het doel van deze methode is het maken van clusters zodanig dat de variatie binnen deze groepen niet al te drastisch verhogen. De resulterende clusters worden zo homogeen mogelijk gemaakt.

De keuze om twee clusters samen te voegen wordt in elke stap bepaald door de „som van de kwadraten"-index. In elke stap wordt deze index voor alle te vormen clusters bepaald, en de clusters met de kleinste „som van de kwadraten" –index worden samengevoegd.

Ward"s methode is een veelgebruikte methode. Het ontwikkelt clusters die ongeveer even groot zijn. De methode heeft wel een intensieve rekentijd.

Voor meer informatie over de andere methodes voor het ontwikkelen van cluster verwijzen wij naar Johnsson (2002).

3.3.4 Bepalen van het aantal clusters in de eindoplossing

Bij hiërarchisch clusteren is er elke stap één cluster minder. Wanneer er voorafgaand aan hiërarchisch clusteren geen aantal te vormen clusters wordt bepaald, zal er één cluster gevormd worden door de hiërarchische methoden. Er zal dus vooraf een gewenst aantal clusters bepaald moeten worden. Het is de bedoeling om een balans te vinden tussen het aantal clusters en de mate van gelijkenis binnen de clusters.

Voor meer informatie over het bepalen van het aantal cluster in de eindoplossing wordt aangeraden Milligan en Cooper (1996) te lezen. Zij hebben een uitgebreide vergelijkende studie gedaan naar dertig methodes om het aantal clusters te bepalen.

3.3.5 Validatie van de oplossing

Het is belangrijk om na te gaan of de clusteroplossing representatief en betrouwbaar is voor de algemene populatie. Onderzoekers die deze stap overslaan, nemen het risico om een clusteroplossing te accepteren specifiek voor die dataset en die verder niet generaliseerbaar is. Een manier om de oplossing te valideren is door de dataset op te splitsen in twee delen, beide delen te clusteren, en de resultaten te vergelijken.

3.3.6 Overwegingen

In hiërarchische clusteranalyse zijn nog een aantal andere kritische punten waar rekening mee gehouden dient te worden (Milligan, 1996). De beslissingen over deze punten zullen wederom afhangen van de gekozen dataset.

- Selecteren van variabelen: hiërarchische clusteranalyse kan dramatisch worden beïnvloed door de opname van een of twee ongepaste variabelen. Hiërarchische clusteranalyse is alleen zinvol als er variabelen zijn die verschillen tussen

waarnemingen of groepen van waarnemingen veroorzaken. Ook de aard van de variabelen is van belang.

- Data standaardisatie: Er zal een keuze gemaakt worden of de data gestandaardiseerd wordt voordat de afstanden tussen de waarnemingen berekend worden of niet. Sommige afstandsmaten zijn vrij gevoelig voor grote verschillen tussen de grootte van variabelen. In dat geval is het verstandig om de data van te voren te standaardiseren.
- Extreme waarden (outliers): Er dient rekening gehouden te worden met extreme waarden is de dataset. Hiërarchische clusteranalyse is erg gevoelig voor outliers. De outliers kunnen de structuur verstoren en maken de gevormde clusters niet representatief voor de populatie.

4 Experiment

In dit hoofdstuk zal geïllustreerd worden hoe PCA en hiërarchische clusteranalyse toegepast worden op een meerdimensionale dataset. De analyse zal uitgevoerd worden in R, een geïntegreerde suite van software faciliteiten voor data manipulatie, statistische berekeningen en grafische weergave.

4.1 Baseball dataset

De bruikbaarheid van PCA en hiërarchische clusteranalyse zal worden weergeven aan de hand van analyses toegepast op de Baseball dataset. De dataset bestaat uit 337 observaties en 18 variabelen. Er waren 337 Major League Baseball spelers ondervraagd die minstens 1 wedstrijd hadden gespeeld in de seizoenen van 1991 en 1992. De data zijn geanalyseerd door Watnik (1998) om te onderzoeken of prestaties van Major League Baseball spelers invloed hadden op de hoogte van het salaris van de baseballer.

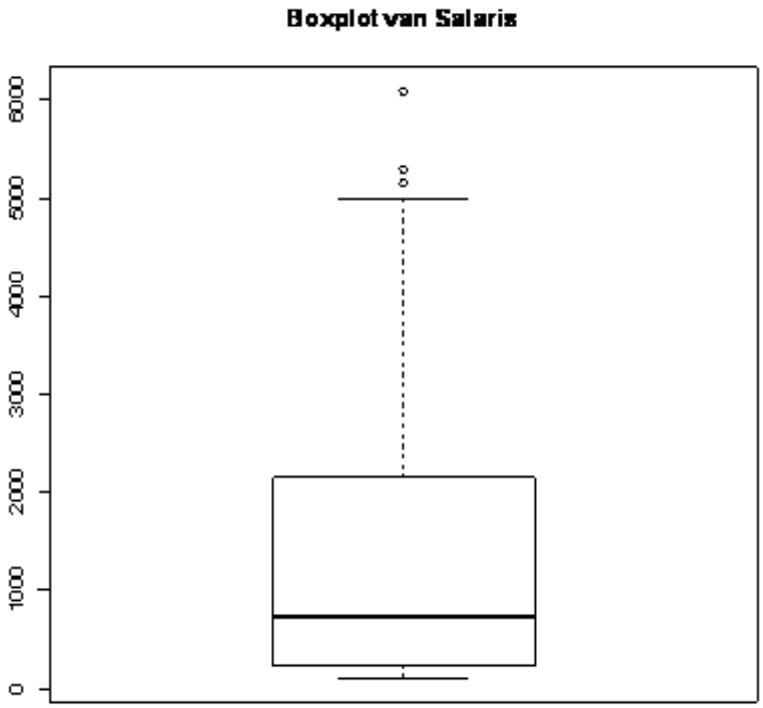
De variabelen zien er als volgt uit:

X_1: Salary (in thousands of dollars)
X_2: BA: Batting average
X_3: OBP: on-base percentage
X_4: Runs: number of runs scored
X_5: hits: number of hits
X_6: 2B: number of doubles
X_7: 3B: number of triples
X_8: HR: number of home runs
X_9: RBI: number of runs batted in
X_{10}: BB: number of bases on balls or walks
X_{11}: SO: number of strikeouts
X_{12}: SB: number of stolen bases
X_{13}: E: number of error made
X_{14}: FAE: indicator of freeagent eligibility
X_{15}: FA: indicator of free agent in 1991/92
X_{16}: AE: indicator of arbitration eligibility
X_{17}: A: indicator of arbitration in 1991/92
X_{18}: Name: name of the baseball player

X_1 is de output variabele en X_2 tot en met X_{18} zijn de input variabelen. X_{14} tot en met X_{17} zijn indicator variabelen. Dit zijn binaire variabelen. Deze variabelen geven aan hoe vrij een speler is om te veranderen van club. In dit onderzoek zullen wij deze variabelen beschouwen als numerieke variabelen, al is dit niet noodzakelijk de juiste keuze. Voor meer informatie over hoe er omgegaan dient te worden met binaire variabelen verwijzen we naar Hardle(2007).Voor de analyses zal de variabele “Salary” verwijderd worden, omdat dit een output variabele is. Ook zal de variabele “naam” verwijderd worden.

Wij zullen onderzoeken of het salaris van de spelers afhangt van de geleverde prestaties. Dit zullen wij aan de hand van PCA en hiërarchische clusteranalyse onderzoeken.

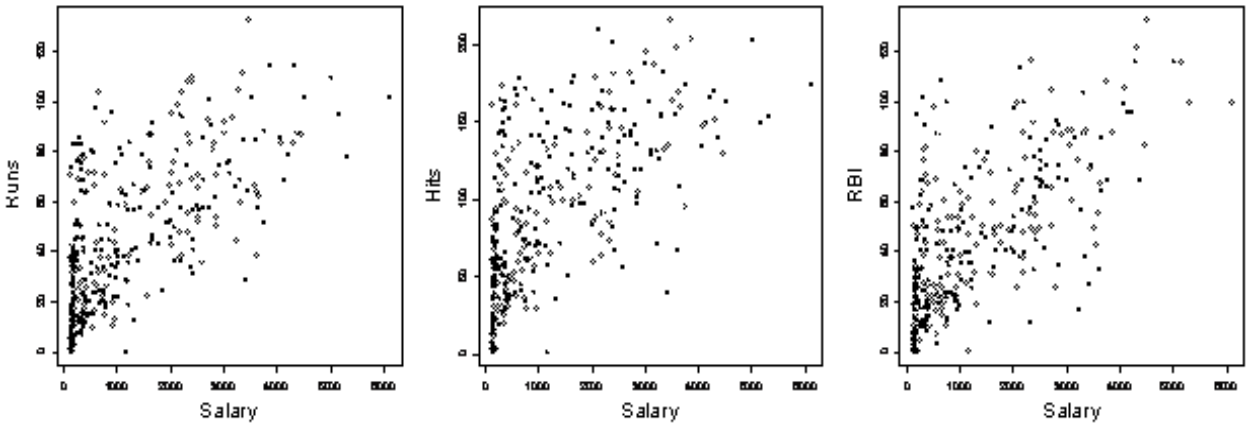
Er zal nu een korte verkennende analyse volgen van de data voordat we overgaan naar PCA en hiërarchische clusteranalyse.



Figuur 4.1: Boxplot van salaris

Figuur 4.1 laat een boxplot van salaris zien. Dit figuur geeft de verdeling van de variabele "Salary" weer. Te zien is dat er drie extreme waarden zijn. Het gaat hier om de spelers Bobby Bonilla, Danny Tartabull en Barry Bonds.

Figuur 4.2 geeft scatterplots van de drie variabelen die het sterkst correleren met de variabele "Salary" uitgezet tegen "Salary" weer. De drie variabelen die het sterkst positief correleren met de variabele "Salary" zijn "Runs", "Hits" en "RBI". De correlatiecoëfficiënten zijn respectievelijk 0.6429, 0.6212 en 0.6684. De scatterplots laten zien dat er enig verband is tussen de drie variabelen en de variabele "Salary", maar kan nog niks uit sluiten.



Figuur 4.2: Scatterplots van "Salary" uitgezet tegen "runs", "hits" en "RBI"

In tabel 4.1 en 4.2 zijn de meest en minst verdienende spelers van de Major League seizoen 1991/1992 weergeven met hun salaris.

Speler	Salaris
Bobby Bonilla	$6,100.00
Danny Tartabull	$5,300.00
Barry Bonds	$5,150.00
Ruben Sierra	$5,000.00
Cecil Fielder	$4,500.00

Tabel 4.1: Meest verdienende spelers met hun salaris

Speler	Salaris
Gary Pettis	$109.00
Royce Clayton	$109.00
Ted Wood	$109.00
Tim McIntosh	$109.00
John Ramos	$109.00

Tabel 4.2: Minst verdienende spelers met hun salaris

Voor een uitgebreide uitleg en analyse van de data wordt aangeraden Wanik(1998) of Izenman(2008) te lezen.

4.2 Resultaten principale componenten analyse

In deze paragraaf zal PCA uitgevoerd worden op de dataset. Wij zullen gaan onderzoeken of de prestaties van de baseball spelers invloed hebben op het salaris. PCA kan uitgevoerd worden met de functie *princomp* in R. Deze functie maakt bij het berekenen van de principale componenten gebruik van de eigenwaarden en eigenvectoren uit de correlatiematrix. Omdat de correlatiematrix gebruikt wordt, hoeven wij de data vooraf niet te standaardiseren.

De resultaten van het toepassen van *princomp* op de dataset zijn weergeven in tabel 4.3. De variantie van elke principale component is weergeven. De derde kolom geeft de totale variantie verkregen door de k^{de} component, het percentage van de variantie. In de laatste kolom is de cumulatieve variantie weergeven. Te zien is dat ruim 41% van de totale variantie van de dataset toegeschreven wordt aan de eerste component PC1. De eerste 5 componenten vertegenwoordigen samen bijna 78% van de totale variantie van de dataset en de eerste 6 ruim 83%. Voor een vervolgonderzoek zou gekozen kunnen worden de eerste 5 of 6 principale componenten te gebruiken. Dit zorgt voor een grote reductie van data zonder veel verlies van informatie.

	Variantie	% variantie	Cum(variantie)
PC 1	6.6194	0.4137	0.4137
PC 2	1.9542	0.1221	0.5359
PC 3	1.4871	0.0929	0.6288
PC 4	1.2821	0.0801	0.7089
PC 5	1.0985	0.0687	0.7776
PC 6	0.8474	0.0530	0.8305
PC 7	0.6253	0.0391	0.8696
PC 8	0.5523	0.0345	0.9042
PC 9	0.4244	0.0265	0.9307
PC 10	0.3778	0.0236	0.9543
PC 11	0.3066	0.0192	0.9734
PC 12	0.1893	0.0118	0.9853
PC 13	0.1062	0.0066	0.9919
PC 14	0.0642	0.0040	0.9959
PC 15	0.0401	0.0025	0.9984
PC 16	0.0251	0.0016	1.0000

Tabel 4.3: Variantie, percentage van de variantie en de cumulatieve variantie

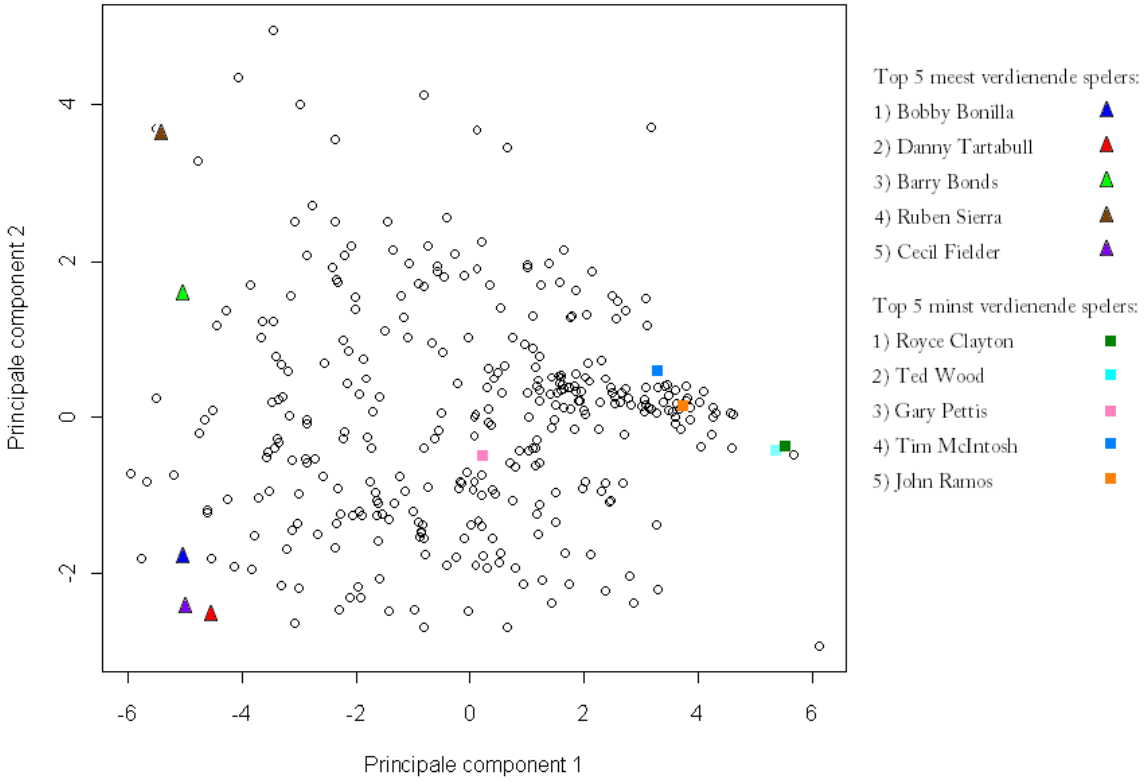
In tabel 4.4 zijn de correlaties tussen de originele variabelen en de eerste twee principale componenten weergeven. Te zien is dat de eerste component sterk (negatief) correleert met de Hits, 2B, RBI, BB en SO variabelen. Deze principale component kan geïnterpreteerd worden als de prestatie component. De variabelen FAE, FA, AE en A correleren sterk met de tweede component. Deze principale component geeft aan hoe vrij elke speler was om te veranderen naar een andere club.

Variabele	PC1	PC2
BA	-0.5122	0.2080
OBP	-0.5781	0.0442
Runs	-0.9606	0.0277
Hits	-0.9440	0.0628
2B	-0.8771	0.0632
3B	-0.5216	0.2823
HR	-0.7532	-0.2525
RBI	-0.9068	-0.1256
BB	-0.8363	-0.1601
SO	-0.7567	-0.1526
SB	-0.4514	0.2562
E	-0.3888	0.2181
FAE	-0.3261	-0.7391
FA	-0.0178	-0.5174
AE	-0.1571	0.7241
A	-0.1402	0.4906

Tabel 4.4: correlatiecoëfficiënt van de originele variabelen en de eerste twee componenten

Met behulp van PCA zal onderzocht worden hoe de 5 meest verdienende en 5 minst verdienende spelers zich verhouden tot de eerste twee principale componenten.

Figuur 4.3 geeft een twee dimensionale representatie van de projectie van de gehele dataset op de eerste twee componenten weer. Deze twee componenten vertegenwoordigen ruim 53% van de totale variantie van de dataset. De gekleurde driehoekjes geven de meest verdienende spelers aan en de gekleurde vierkantjes geven de minst verdienende spelers aan. Te zien is dat de meest verdienende spelers allemaal aan de linkerzijde van de plot zitten. Aan de rechterzijde van de plot zitten de minst verdienende spelers, met uitzondering van Gary Pettis. Er kan gezien worden dat de prestatie metingen dus iets te zeggen hebben over de hoogte van het salaris van de spelers.



Figuur 4.3: projectie van de gehele dataset op de eerste twee componenten

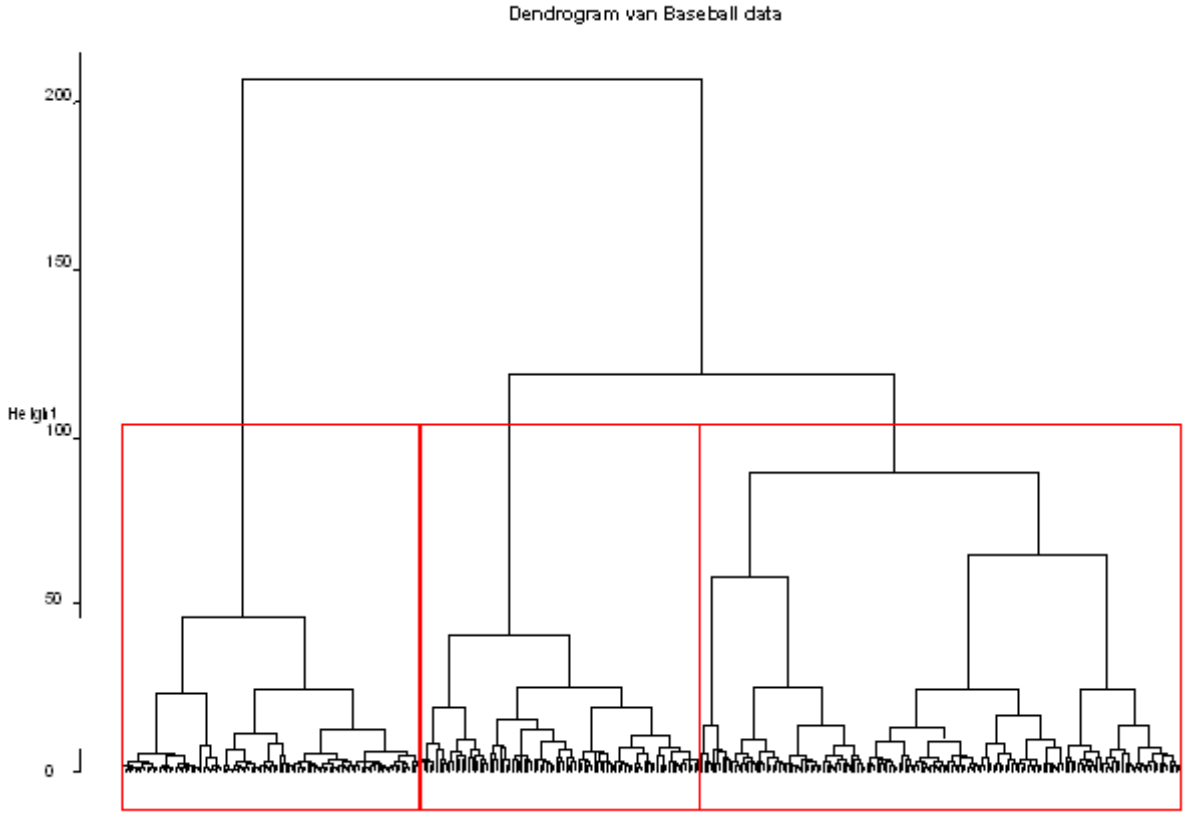
4.3 Resultaten hiërarchische clusteranalyse

In deze paragraaf zal een hiërarchische clusteranalyse worden toegepast op de dataset Baseball. Wij zullen gaan onderzoeken in welke clusters de meest en minst verdienende spelers ingedeeld worden.

Hiërarchische clusteranalyse kan gedaan worden door middel van de functie *hclust* in R. Deze functie zal clusters vormen aan de hand van agglomerende methodes. De input van deze functie is een afstandsmatrix. Er zijn verschillende opties om de afstandsmatrix te berekenen. Voor de Baseball dataset hebben wij gekozen om de afstand te berekenen door middel van de gestandaardiseerde euclidische afstand (zie 3.3.2 voor meer uitleg).

Vervolgens is er gekozen voor Ward's algoritme om clusters te ontwikkelen. Ward's algoritme is gebaseerd op het minimaliseren van "verlies van informatie" (zie 3.3.3 voor meer uitleg). Voor deze dataset geeft dit algoritme de beste resultaten in vergelijking met de andere methodes.

In figuur 4.4 is een dendrogram te zien. Dit twee dimensionale diagram laat de samenvoegingen op de verschillende niveaus zien. De dendrogram dient van onder naar boven gelezen te worden. De beginsituatie is dat alle waarnemingen aan een aparte cluster zijn toegewezen. Per stap worden er samenvoegingen gedaan welke in het figuur weergeven zijn. Wij hebben ervoor gekozen dat er drie clusters gevormd worden. Dit is in figuur 4.4 weergeven door de rode omranding.



Figuur 4.4: Dendrogram van Baseball data

Om een interpretatie van de clusters te geven, laat tabel 4.5 het aantal observaties, het gemiddelde en de standaarddeviatie van de variabele "Salary" per cluster zien. Te zien is dat de gemiddelden per variabelen per cluster significant verschillen. Cluster 1 is heeft de minste observaties, gevolgd door cluster 3 en 1. Het gemiddelde salaris van de eerste cluster is het hoogst. De standaarddeviaties per cluster zijn ook vrij groot.

	cluster 1	cluster 2	cluster 3
# observaties	89	153	95
Gemiddeld salaris	2169.03	1341.02	237.20
st. dev salaris	1346.03	1089.25	178.03

Tabel 4.5: Aantal observaties, gemiddelde en standaarddeviatie van "Salary" voor cluster 1, 2 en 3

In tabel 4.6 en 4.7 worden de top 5 meest en minst verdienende spelers weergeven. Aan de hand van hun prestaties zijn de spelers toegewezen aan bepaalde clusters. De meest verdienende spelers zijn allemaal aan cluster 1 en 2 toegewezen.

De minst verdienende spelers worden allemaal toegewezen in cluster 3, behalve Gary Pettis. Deze speler wordt aan de hand van zijn geleverde prestaties aan cluster 2 toegewezen. Te zien is dat de twee meest verdienende spelers ook aan deze cluster zijn toegewezen. Er zou geconcludeerd kunnen worden dat Gary Pettis te weinig heeft verdiend.

Speler	Cluster	Salaris
Bobby Bonilla	2	$6,100.00
Danny Tartabull	2	$5,300.00
Barry Bonds	1	$5,150.00
Ruben Sierra	2	$5,000.00
Cecil Fielder	1	$4,500.00

Tabel 4.6 Top 5 meest verdienende spelers met bijbehorend salaris en cluster

Speler	Cluster	Salaris
Gary Pettis	2	$109.00
Royce Clayton	3	$109.00
Ted Wood	3	$109.00
Tim McIntosh	3	$109.00
John Ramos	3	$109.00

Tabel 4.7 Top 5 minst verdienende spelers met bijbehorend salaris en cluster

5 Conclusie en aanbevelingen

In dit onderzoek zijn PCA en hiërarchische clusteranalyse besproken. Vervolgens zijn deze technieken toegepast op de dataset Baseball om de bruikbaarheid van de technieken te onderzoeken.

PCA is een makkelijk te gebruiken datareductie techniek, die interessante relaties kan onderscheiden die niet eerder waren opgevallen. PCA heeft het vermogen om een groot aantal variabelen om te zetten in een kleiner aantal componenten zonder dat er (veel) informatie verloren gaat. Dit biedt de onderzoeker een krachtige techniek in het bereiken van een beter begrip van de structuur van de data, maar ook een manier om de dataset te reduceren zodat het makkelijker bruikbaar wordt voor vervolganalyses.

Hiërarchische clusteranalyse is een techniek met een breed scala aan toepassingen. Tijdens het gebruik van deze techniek dient de onderzoeker rekening te houden met de juiste toepassing van de onderliggende principes. Hiërarchische clusteranalyse heeft namelijk vele valkuilen waar tijdens het onderzoek met zorg aandacht aan besteed dient worden. Zo dienen er binnen het clusterproces een aantal keuzes gemaakt te worden waar de uiteindelijke oplossing sterk vanaf hangt. Er zal een keuze gemaakt moeten worden over hoe de afstand berekend wordt tussen de waarnemingen. In dit onderzoek is dat gedaan met behulp van de standaard euclidische afstand, maar er zijn nog een aantal andere technieken die de afstand kunnen berekenen.

Ook dient er een keuze gemaakt te worden over het aantal clusters dat ontwikkeld moet worden. Voor het bepalen van het aantal clusters zijn ook verschillende technieken ontwikkeld. Het onderzoeken van deze technieken valt buiten het bereik van dit onderzoek. Voor verdere analyses wordt aangeraden deze technieken wel toe te passen.

Wanneer deze keuzes op een juiste manier genomen worden, heeft de hiërarchische clusteranalyse het potentieel om onderliggende structuren in de data zichtbaar te maken die anders niet waren ontdekt.

Bijlagen: R-code

Dit is de code die gebruikt is om de analyses uit te voeren in R.

```
## Voor het installeren van MMST packages en het openen van de dataset Baseball
> install.packages("MMST")
> require(MMST)
> data(baseball)
## Maakt boxplot van variabele Salaris
>boxplot(baseball[,1], main="Boxplot van Salaris")
## Verwijder variabele naam
> bb <- subset(baseball, select = -c(Name))
> cor(bb)
## Verwijder variabelen Name en salary
> baseball1 <- subset(baseball, select = -c(Name, salary))

## Voert PCA uit. Correlatiematrix wordt gebruikt voor berekeningen door cor =TRUE
> pca1 <- princomp(x = baseball1, cor = TRUE)
> summary(pca1)
## Plot de eerste twee componenten
> plot(pca1$scores[,1],pca1$scores[,2], ylab="Principale component 2", xlab="Principale
component 1")
##Correlaties tabel
Correlatie_tabel <-cor(baseball1, pca1$scores)
## Geeft een lijst van spelers geordend op salaris
> bbOrdered = baseball[sort.list(baseball[,1]),];

## Schalen van de dataset
> base_scaled <- scale(baseball1, scale = TRUE, center =TRUE)
## Berekent van de gestandaardiseerde dataset de afstandsmatrix met de methode
## “euclidean”
> dis <- dist(base_scaled, method = "euclidean")
## Voert hiërarchische clusteranalyse uit. Clusters worden ontwikkeld met de methode
“ward”
> CA <- hclust(d=dis, method="ward")
## Plot dendrogram
> plot(CA, hang = -1)
> groep <-cutree(CA, k=3)
## Geeft gemiddeld salaris van cluster 1
> mean(baseball[groep==1, "salary"])
## Geeft standaarddeviatie van salaries van cluster 1
> sdev(baseball[groep==1, "salary"])
```

Referenties

Chatfield, C., & Collins, A.J. (1980). *Introduction to multivariate analysis*. London: Chapman & Hall.

Everitt, B. S., & Dunn, G. (2001). *Applied multivariate data analysis.* London: Arnold.

Everitt, B.S., Landau, S. and Leese, M. (2001). *Cluster Analysis, 4th Edition.* Oxford University Press, Inc., New York.

Green, P.E., Frank, R.E. & Robinson, P.J. (1967). Cluster analysis in test market selection. *Maanagement science,* april, 387-400.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W.C. (1998) *Multivariate data analysis*. Upper Saddle River, New Jersey: Prentice Hall.

Härdle, W., & Simar, L. (2007). *Applied Multivariate Statistical Analysis.* New York: Springer.

Izenman, A. J.(2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York: Springer.

Jain, A.K., Bhattacharjee, S. (1992) *Text segmentation using gabor filters for automatic document processing*. New York: Springer-Verlag.

Johnson, R.A., & Wichern, D.A. (2002). *Applied multivariate statistical analysis.* Upper Saddle River, New Jersey: Prentice Hill.

Kok Sørensen, C. & Gutiérrex, J.M.P (2006) Euro area banking sector integration, using hierarchical cluster analysis techniques. *Working paper series no. 267*.

Langley, P., bowers, E.J., & Murray, A. (2010) Principal component analysis as a tool for analyzing beat-to-beat changes in ECG features: application to ECG-derived respiration. *Biomedical Engineering IEEE Transactions,* 57(4), april, 821-829.

Lee, Myoung–Jae (1995) "Semi–parametric estimation of simultaneous equations with limited dependent variables: a case study of female labour supply", *Journal of Applied Econometrics*, 10(2), april–june, 187–200.

Manly, B.F.J.(2005) *Multivariate Statistical Methods: A Primer.* London: Chapman & Hall.

Milligan, G.W. (1996) *Clustering validation: results and implications for applied analyses.* Ohio: Columbus.

Moore, W.C., Meyers, D.A., Wenzel, S.E., Teague, W.G. (2009). Identification of asthma phenotypes using cluster analysis in the severe asthma research program.*Am. J. Respir. Crit. Care Med.,* 181(4):315.

Quant, C.M., Fisher, P.H., Buringh E., Ameling, C.B., Houthuijs, D.J.M. & Cassee F.R. (2003). Application of principal component analysis to time series of daily air pollution and mortality. *RIVM report.*

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposim on Biocomputing 5:* 452-463.

Tukey, J.W., (1977). Exploratory data analysis. Addison-Wesley.
Ward, J.H., Jr. (1963), Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 48, 236-244.

Watnik, M.R. (1998) Pay for play: are baseball salaries based on performance?. *Journal of Statitistics*, 6.

Zhao, W., Chellappa, R. & Krishnaswamy, A. (1998). Discriminant analysis of principal components for face recognition. *Proceedings, International Conference on Automatic Face and Gesture Recognitio:* 336–341.