



Universiteit Utrecht

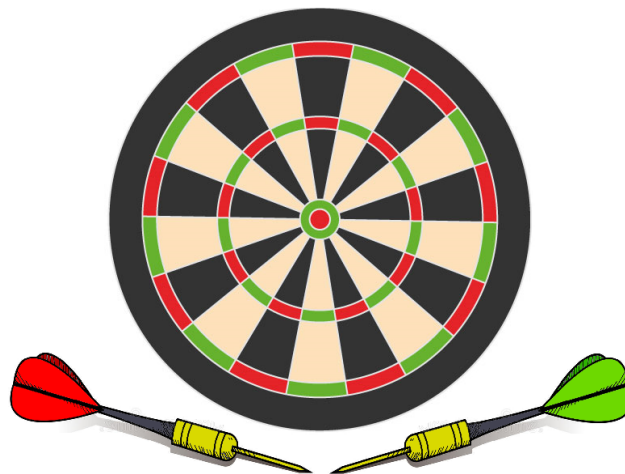
Bachelorscriptie Wiskunde

De Statistiek achter Darts

Auteur: W.N. van Veluw

Begeleider: S. Dirksen

Einddatum: 5 juni 2020



Inhoudsopgave

1	Inleiding	2
2	Een wiskundig model van het dartbord	4
2.1	Een exact dartboard	4
2.2	Worpen en scores	4
2.3	De verwachtingswaarde van de score	5
2.4	Ellipsen bij bivariate normale stochasten	6
3	De covariantie matrix volgens een simpel model	9
3.1	Het EM-algoritme	9
3.2	De covariantie matrix	12
3.3	Overzicht	15
4	De covariantie matrix volgens een geavanceerd model	16
4.1	De covariantie matrix	16
4.2	Importance sampling	20
4.2.1	(Pseudo)willekeurige getallen	20
4.2.2	Monte Carlo benadering van integralen	20
4.2.3	Importance sampling	21
4.3	Importance Sampling voor het dartmodel	22
4.3.1	Willekeurige posities genereren	22
4.3.2	Toepassing van Importance Sampling	23
4.4	Overzicht	24
5	Hypothesetoetsen	25
5.1	T- en Z-testen	25
5.2	Test voor dartmodel	26
6	Resultaten	28
6.1	Het simpele model	28
6.2	Het geavanceerde model	30
6.3	Hypothesetoetsen	31
7	Discussie	39
8	Conclusie	40
9	Dankbetuigingen	41

1 Inleiding

Iedereen heeft wel eens een dartpijl gegooid of weet wat een spelletje darten inhoud. Het populaire caféspel dateert al uit 1900 en heeft zich sindsdien ontwikkeld tot een professionele tak waarbij er voor duizenden euro's aan prijzengeld ligt te wachten op de winnaar van een toernooi. Hoewel het spel vooral in Groot-Brittannië wordt gespeeld, heeft Nederland ook zeker zijn darts-iconen: denk maar eens aan Raymond van Barneveld en Michael van Gerwen.

Darts wordt gespeeld met twee spelers die hun dartpijlen richting een cirkelvormig bord gooien. Hoewel er veel verschillende spelvormen bestaan, is het meestal het doel om een aantal van 501 zo snel als mogelijk en precies uit te gooien. De Professional Darts Corporation (PDC) hanteert hierbij nog extra regels als de "double out" (de laatste pijl moet een "double" zijn) en een "best of" knock-out systeem tijdens zijn toernooien.

Het dartbord is verdeel in twintig evengrote taartvormige stukken en aan alle stukken is een waarde tussen de één en de twintig toegekend. Elk stuk is op zijn beurt weer onderverdeeld in 'singles', 'doubles' en 'triples'. Deze drie geven recht op respectievelijk één, twee of zelfs drie keer de waarde van een taartvormig stuk van het bord. Een speciale waarde is toegekend aan het centrum van het dartbord: de "bullseye". De single bullseye levert 25 punten op en een pijl in de double bullseye (of eenvoudigweg bullseye) levert het dubbele op. Merk op dat dit echter niet het vakje is met de hoogste score: de meeste darters mikken op de triple 20, die zestig punten oplevert. Zie figuur 1.

Hierin schuilt ook meteen de hoofdvraag van deze bachelorscriptie: is de triple 20 wel het mikpunt met de hoogste verwachte score? Een blik op het bord leert dat er rond de triple 20 veel lage waardes voorkomen, zoals de één en de vijf. Een inaccuraat pijl gemikt op de triple 20 kan daarom zomaar in een vakje terecht komen met een lage waarde. Kan zo'n onervaren darter niet beter ergens anders mikken om zijn kansen op een hoge score te vergroten?

Om deze vraag te beantwoorden zal het artikel "A Statistician Plays Darts"[1] centraal staan. In dit artikel worden twee modellen rondom het dartbord en de worp van een darter gebouwd om het optimale mikpunt voor een darter te berekenen. Het zal blijken dat het voor deze berekening nodig is om de precisie (lees: covariantie matrix) van een darter te schatten. Voor deze schatting zullen we twee veelgebruikte statistische technieken, namelijk het EM-algoritme en een Monte Carlo simulatie genaamd "Importance Sampling", moeten toepassen. Daarna nemen we de proef op de som: we gooien zelf een steekproef en bereken aan de hand van het artikel het optimale mikpunt. Afgaande op deze berekening bekijken we of er daadwerkelijk hogere scores worden behaald als we op dit optimale punt gaan mikken. We doen dit laatste door middel van het uitvoeren van een hypothesetoets.



Figuur 1: Het dartbord.

2 Een wiskundig model van het dartbord

In deze sectie gaan we behandelen hoe we het dartbord door wiskundige ogen kunnen zien, welke stochastische variabelen we definiëren en hoe we de verwachtingswaarde van de score van een worp kunnen maximaliseren.

2.1 Een exact dartboard

Om de posities op een dartboard goed te kunnen beschrijven, tekenen we een assenstelsel op het dartbord waarbij de oorsprong precies in het middelpunt van de double bullseye ligt. Omdat ons object cirkelvormig is, kunnen we handig gebruik maken van *poolcoördinaten*. Voor zo'n poolcoördinaat geldt

$$\begin{aligned}x &= r \cos(\theta) \\ y &= r \sin(\theta)\end{aligned}$$

waarbij r de afstand tot de oorsprong is en θ de hoek die de x-as en de lijn door het punt en de oorsprong maken. Het is eenvoudig te zien dat we een cartesisch coördinaat (x, y) om kunnen schrijven naar een poolcoördinaat met

$$\begin{aligned}r &= \sqrt{x^2 + y^2} \\ \theta &= \arctan\left(\frac{y}{x}\right).\end{aligned}$$

We kunnen nu elk vakje van het dartboard beschrijven door middel van deze poolcoördinaten, zie de figuren aan het eind van deze sectie. In figuur 1 is een normaal dartboard afgebeeld, in figuur 3 een dartboard met de hoeken (in graden) voor elk taartvormig stuk. Tenslotte staan in tabel 1 de afstanden tussen de oorsprong en de verschillende ringen.

Uit deze figuren en de tabel is bijvoorbeeld af te leiden dat voor de triple 20 geldt $r \in (99, 107)$ en $\theta \in (81^\circ, 99^\circ)$. Voor de double 7 geldt dan $r \in (162, 170)$ en $\theta \in (225^\circ, 243^\circ)$.

2.2 Worpen en scores

Laat ons de vakjes van het dartbord als volgt noteren. We koppelen telkens een letter (S,D of T) aan een getal, waar bij de S voor 'single' staat, de D voor 'double' en de T voor 'triple'. Een double 10 geven we dus aan met D10, een single 2 met S2 en de double bullseye met D25.

Verder noteren we met $Z = \begin{pmatrix} z_x \\ z_y \end{pmatrix}$ een stochastische variabele die de tweedimensionale positie van een pijl in het dartbord aangeeft en berekenen we de score van een pijl met de functie

$$s : \mathbb{R}^2 \rightarrow \mathbb{N}_0 : Z \mapsto X.$$

Als een pijl bijvoorbeeld in de oorsprong gegooid wordt, dan geldt $Z = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in D25$ en dus $s(Z) = 50$. Merk op dat de score ook een stochastische variabele is: we noteren deze variabele met X .

Als een darter een pijl werpt, mikt hij op een bepaald punt. Deze zogenoemde "mikpositie" noteren we met $\mu \in \mathbb{R}^2$. Het is echter zo dat de pijl niet altijd daar land waar hij op gemikt wordt: er zit een afwijking in de precisie van de worp. Dit kan komen doordat de darter net een andere houding heeft, iets harder gooit dan nodig of de pijl te vroeg los laat. De afwijking noteren we met ε . We kunnen een worp dan modelleren als

$$Z = \mu + \varepsilon,$$

waarbij we aannemen dat $\varepsilon \sim \mathcal{N}(0, \Sigma)$, voor een covariantie matrix Σ . Een belangrijke eigenschap van deze covariantie matrix is dat het een symmetrische matrix is. Vanuit de manier waarop we Z gemodelleerd hebben

kunnen we afleiden dat Z dus een bivariate, normaal verdeelde toevalsvariabele is met een kansdistributie beschreven door

$$f(z) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}(\mathbf{z} - \mu)^T \Sigma^{-1}(\mathbf{z} - \mu)\right). \quad (1)$$

2.3 De verwachtingswaarde van de score

Nu we weten wat de kansverdeling van Z is, kunnen we ook praten over de verwachtingswaarde van de score. Er geldt dan

$$\begin{aligned} E[s(Z)] &= \int_{-\infty}^{\infty} s(z) \cdot f(z) dz \\ &= \int_{-\infty}^{\infty} s(z) \cdot \frac{1}{2\pi\sqrt{\det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}(z - \mu)^T \Sigma^{-1}(z - \mu)\right) dz \end{aligned} \quad (2)$$

We willen deze (tweevoudige) integraal berekenen en dan maximaliseren voor μ . Er zijn echter twee problemen die we dan tegenkomen.

1. De integraal is (te) moeilijk te analyseren.
2. De kansdichtheidsfunctie, en daarmee de schatting van het mikpunt met de hoogste verwachte score, is afhankelijk van Σ .

Voor het eerste probleem merken we op dat de integraal te schrijven is als een *convolutie* en dat een convolutie relatief eenvoudig uit te rekenen is met een zogeheten *Fast Fourier Transform (FFT)*. Computers kunnen snel rekenen met zo'n transformatie en daarom is het handig om de verwachte te schrijven als zo'n transformatie.

Definitie 2.1. *Gegeven twee functies $g, h : \mathcal{V} \rightarrow \mathcal{W}$. Dan wordt de operatie convolutie, genoteerd met $*$, gegeven door*

$$(g * h)(t) = \int_{-\infty}^{\infty} h(\tau) \cdot g(t - \tau) d\tau.$$

Stelling 2.1 (Convolutie stelling). *Gegeven twee functies $g, h : \mathcal{V} \rightarrow \mathcal{W}$ met convolutie $(g * h)$. Noteer met \mathcal{F} de Fast Fourier Transform, dan geldt*

$$(g * h) = \mathcal{F}^{-1}(\mathcal{F}(g) \cdot \mathcal{F}(h)).$$

Met deze definitie en deze stelling kunnen we de verwachte score uit (2) schrijven als

$$\begin{aligned} E_{\mu}[s(Z)] &= (g * s)(\mu) \\ &= \mathcal{F}^{-1}(\mathcal{F}(g) \cdot \mathcal{F}(s))(\mu), \end{aligned}$$

waarbij

$$g(x) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}x^T \Sigma^{-1}x\right).$$

We kiezen deze g zodat geldt $g(\mu - z) = f(z)$. Op deze manier verkrijgen we een expliciete uitdrukking voor de verwachte score en kunnen we de verwachte score maximaliseren. Het optimale mikpunt, genoteerd met μ_{optimaal} , kunnen we dan uitdrukken als

$$\mu_{\text{optimaal}} = \arg \max_{\mu} E[s(Z)].$$

Er is al genoemd dat de schatting van μ_{optimaal} afhankelijk is van Σ , het tweede probleem waar we omheen zullen werken. We gaan deze Σ dan ook schatten in sectie 3 en 4, berustend op verschillende aannames,

zodat we μ_{optimaal} kunnen berekenen.

Voor de schatting van Σ hebben we uiteraard een steekproef nodig van n worpen, waarbij $n \geq 50$. We zorgen er daarnaast voor dat die n worpen gericht zijn op de oorsprong, zodat geldt $Z = \varepsilon$ en dus $Z \sim \mathcal{N}(0, \Sigma)$. In de praktijk zal het nemen van deze steekproef echter lastig zijn: er is namelijk nooit met zekerheid te zeggen wat de exacte coördinaten van een geworpen pijl zijn (de waarde van Z), dus is het lastig om iets te zeggen over de kansverdeling van die worpen. Het is daarentegen wél eenvoudig om de score van een geworpen pijl te noteren: de variabele X . In de komende twee secties zal blijken hoe we uit de scores (X_1, \dots, X_n) van n pijlen, gericht op de oorsprong, iets kunnen zeggen over de covariantie matrix van de exacte posities Z .

2.4 Ellipsen bij bivariate normale stochasten

Als we eenmaal de covariantie matrix en het optimale mikpunt voor een darter hebben berekend, zullen we ook de ellips tekenen waarbinnen 95% van de geworpen pijlen waarschijnlijk terecht zullen komen. Om deze ellips, of niveaulijn, te construeren, bewijzen we eerst een stelling.

Stelling 2.2. *Gegeven een n -dimensionale stochastische variabele $Z \sim \mathcal{N}(\mu, \Sigma)$. Dan geldt dat*

$$(Z - \mu)^T \Sigma^{-1} (Z - \mu) \sim \chi_{(n)}^2.$$

Bewijs. Vanuit de eigenschappen van de normale verdeling volgt

$$\begin{aligned} Z &\sim \mathcal{N}(\mu, \Sigma) \\ Z - \mu &\sim \mathcal{N}(0, \Sigma) \\ \Sigma^{-\frac{1}{2}}(Z - \mu) &\sim \mathcal{N}(0, \mathbb{I}), \end{aligned} \tag{3}$$

waarbij \mathbb{I} de tweedimensionale eenheidsmatrix is. Nemen we nu $\mathbf{X} = (X_1, \dots, X_n)$ zodat $X_i = \left(\Sigma^{-\frac{1}{2}}(Z - \mu) \right)_i$, voor $1 \leq i \leq n$, dan geldt dus $X_i \sim \mathcal{N}(0, 1)$. Van de som van kwadraten van standaard normaal verdeelde stochasten is bekend dat deze een chi-kwadraat verdeling volgt met het aantal vrijheidsgraden gelijk aan de bovengrens van de som. Dat wil zeggen, omdat $X_i \sim \mathcal{N}(0, 1)$, geldt dat

$$\sum_{i=1}^n X_i^2 \sim \chi_{(n)}^2.$$

Merken we nu op dat deze som te schrijven is als

$$\begin{aligned} \sum_{i=1}^n X_i^2 &= \mathbf{X}^T \mathbf{X} \\ &= (\Sigma^{-\frac{1}{2}}(Z - \mu))^T (\Sigma^{-\frac{1}{2}}(Z - \mu)) \\ &\downarrow \text{Omdat } \Sigma \text{ symmetrisch is} \\ &= (Z - \mu)^T \Sigma^{-1} (Z - \mu), \end{aligned}$$

dan volgt het resultaat. ■

De niveaulijnen van de stochast Z hebben dan de vorm

$$(Z - \mu)^T \Sigma^{-1} (Z - \mu) = c,$$

waarbij de c een waarde is die bepaald wordt aan de hand van de kansdistributie van $(Z - \mu)^T \Sigma^{-1} (Z - \mu)$. Gebruiken we hiervoor de stelling en het feit dat we een tweedimensionale stochast gebruiken, dan zien we dat geldt

$$\begin{aligned} c &= \mathbb{P}(\chi_{(2)}^2 \leq 0.95) \\ &= 5.99 \dots \end{aligned}$$

De vergelijking van de ellips die het 95%-interval aangeeft is dus gegeven door

$$(z - \mu)^T \Sigma^{-1} (z - \mu) = 5,99 \dots$$

In \mathbb{R}^2 , de verzameling waarin het dartmodel zich bevindt, vinden we het volgende als we de linkerkant uitvermenigvuldigen

$$\frac{1}{\det(\Sigma)} \left(\Sigma_{22}(z_1 - \mu_1)^2 + \Sigma_{11}(z_2 - \mu_2)^2 - (\Sigma_{12} + \Sigma_{21})(z_1 - \mu_1)(z_2 - \mu_2) \right) = 5,99 \dots$$

Omdat Σ symmetrisch is en dus $\Sigma_{12} = \Sigma_{21}$, kunnen we ook schrijven

$$\frac{1}{\det(\Sigma)} \left(\Sigma_{22}(z_1 - \mu_1)^2 + \Sigma_{11}(z_2 - \mu_2)^2 - 2\Sigma_{12}(z_1 - \mu_1)(z_2 - \mu_2) \right) = 5,99 \dots \quad (4)$$

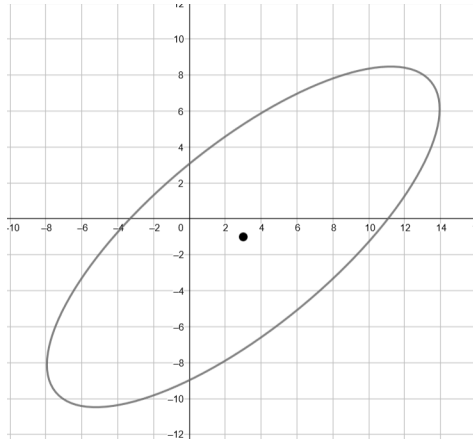
Ter illustratie van deze formule berekenen we een voorbeeld. Stel dat $Z \sim \mathcal{N}(\mu, \Sigma)$, waarbij

$$\begin{aligned} \mu &= \begin{pmatrix} 3 \\ -1 \end{pmatrix}, \text{ en} \\ \Sigma &= \begin{pmatrix} 20 & 13 \\ 13 & 15 \end{pmatrix}. \end{aligned}$$

Dan geldt $\det(\Sigma) = 131$ en dus wordt de ellips (in het xy -vlak) gegeven door

$$\frac{1}{131} \left(15(x - 3) + 20(y + 1) - 26(x - 3)(y + 1) \right) = 5,99 \dots$$

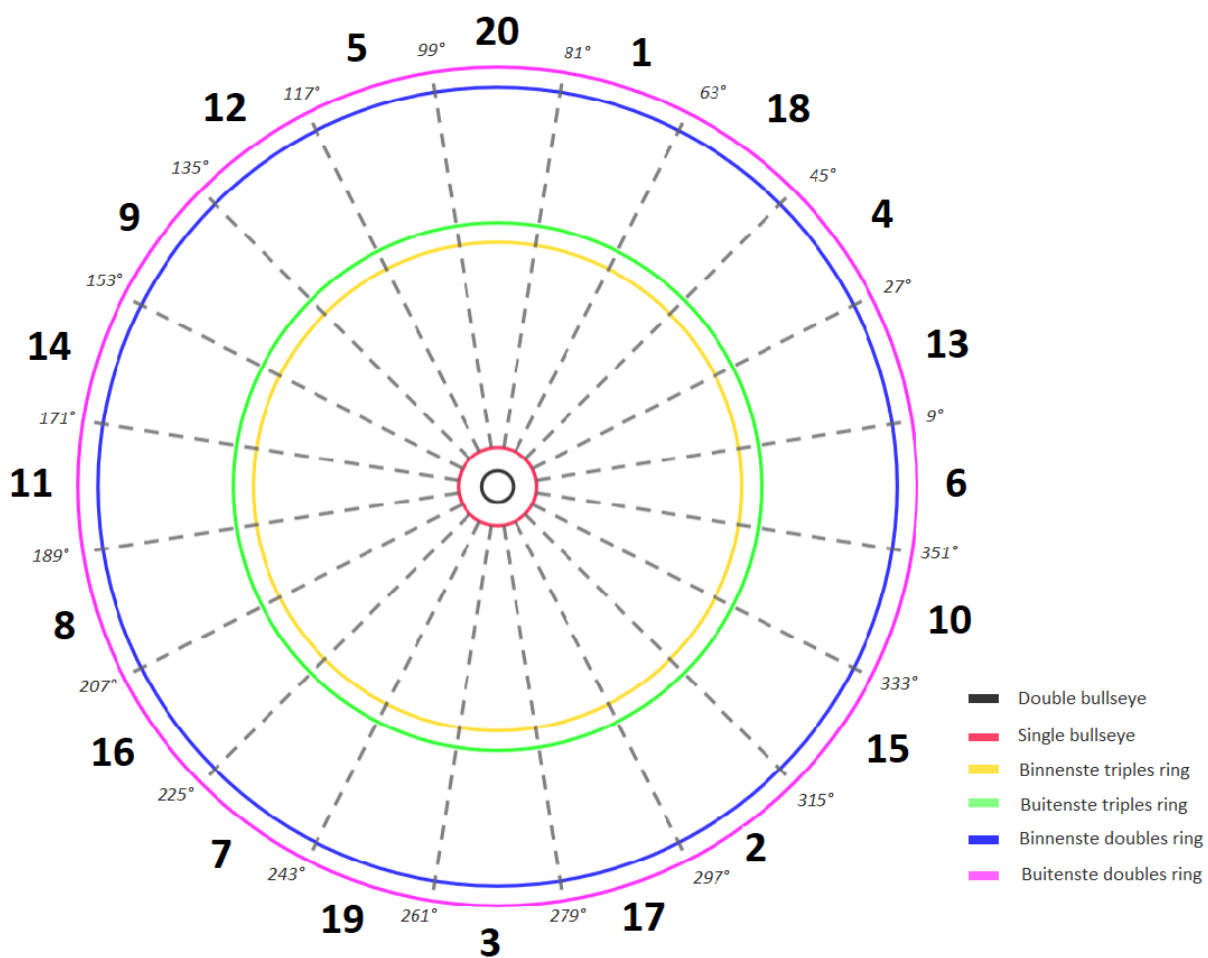
Deze ellips kunnen we plotten zoals gedaan is in figuur 2.



Figuur 2: De ellips die het 95%-interval aangeeft.

Oorsprong tot double bullseye	6.35
Oorsprong tot single bullseye	15.9
Oorsprong tot binnenste triples ring	99
Oorsprong tot buitenste triples ring	107
Oorsprong tot binnenste doubles ring	162
Oorsprong tot buitenste doubles ring	170

Tabel 1: Afstanden van de oorsprong tot aan verschillende ringen in millimeters.



Figuur 3: Het exacte dartbord.

3 De covariantie matrix volgens een simpel model

In sectie 2 stuitte we op het probleem dat de covariante matrix onbekend is, maar dat we deze kunnen schatten door alleen de scores van n pijlen te bekijken. In deze sectie nemen we aan dat een worp horizontaal evenveel afwijkt als verticaal. Met andere woorden, we nemen aan dat $\Sigma = \sigma^2 \mathbb{I}$, waar \mathbb{I} de twee-dimensionale eenheidsmatrix is, en gaan de waarde van σ schatten.

Het probleem van het schatten van een niet-geobserveerde variabele Z door middel van een geobserveerde variabele X leent zich voor een vaak gebruikte statistische methode: het EM-algoritme, die zeer lijkt op de 'maximum likelihood' schatter. Eerst bekijken we de algemene theorie over dit algoritme en daarna bekijken we wat dit betekend voor ons dart-experiment.

3.1 Het EM-algoritme

Voor het EM-algoritme beschouwen we twee stochastische vectoren: de vector $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ met geobserveerde variabelen en de vector $\mathbf{Z} = (Z_1, Z_2, \dots, Z_m)^T$ met niet-geobserveerde variabelen. Voor deze twee variabelen moet het volgende gelden:

- X_i en Z_j zijn paarsgewijs onafhankelijk, voor alle $i \in \{1, \dots, n_1\}$ en $j \in \{1, \dots, n_2\}$.
- Alle X_i zijn onafhankelijk en identiek verdeeld met kansdichtheidsfunctie $f(x | \theta)$.
- De gezamenlijke kansdichtheid van \mathbf{X} is $g(x | \theta)$.
- De gezamenlijke kansdichtheid van \mathbf{X} en \mathbf{Z} is $h(x, z | \theta)$.
- De voorwaardelijke pdf van de niet-geobserveerde data gegeven de geobserveerde data is $k(z | \theta, \mathbf{x})$.

Hierbij is (zijn) θ de parameter(s) van belang. We kunnen voor het verband tussen k, h en g handig gebruik maken van de definitie van een voorwaardelijke kans:

$$k(z | \theta, x) = \frac{h(x, z | \theta)}{g(x | \theta)}. \quad (5)$$

We bekijken nu twee waarschijnlijkheidsfuncties die gebruikt worden bij het algoritme.

Definitie 3.1. *De geobserveerde waarschijnlijkheidsfunctie (GWF) is*

$$L(\theta | x) = g(x | \theta).$$

De complete waarschijnlijkheidsfunctie (CWF) is

$$\begin{aligned} L^c(\theta | x, z) &= h(x, z | \theta) \\ &\downarrow \text{gebruik (5)} \\ &= g(x | \theta) \cdot k(z | \theta, x) \\ &= L(\theta | x) \cdot k(z | \theta, x) \end{aligned}$$

Zoals vaak wordt gedaan bij dit soort maximaliserings processen bekijken we het logaritme van de waarschijnlijkheidsfuncties.

$$\begin{aligned}
\log(L(\theta)) &= \int_z \log(L(\theta)) \cdot k(z | \hat{\theta}_0, x) dz \\
&= \int_z \log(g(x | \theta)) \cdot k(z | \hat{\theta}_0, x) dz \\
&\downarrow \text{gebruik (5)} \\
&= \int_z \log\left(\frac{h(x, z | \theta)}{k(z | \theta, x)}\right) \cdot k(z | \hat{\theta}_0, x) dz \\
&= \int_z [\log(h(x, z | \theta)) - \log(k(z | \theta, x))] k(z | \hat{\theta}_0, x) dz \\
&= \int_z \log(h(x, z | \theta)) k(z | \hat{\theta}_0, x) dz - \int_z \log(k(z | \theta, x)) k(z | \hat{\theta}_0, x) dz \\
&= \mathbb{E}[\log(L^c(\theta | x, z)) | \hat{\theta}_0, x] - \mathbb{E}[\log(k(z | \theta, x)) | \hat{\theta}_0, x] \tag{6}
\end{aligned}$$

Hierbij nemen we $\hat{\theta}_0$ vast. Dit is slechts een eerste (willekeurige) schatting van de parameter van belang.

Vanuit (6) definiëren we

$$Q(\theta | \hat{\theta}_0, x) = \mathbb{E}[\log(L^c(\theta | x, z)) | \hat{\theta}_0, x]$$

en zien we dat het, voor het maximaliseren van de *GWF* ten opzichte van θ , genoeg is om deze Q te maximaliseren. Dit is vanwege het feit dat de tweede verwachtingswaarde geen functie van θ is en dus beschouwd kan worden als een constante.

De maximalisatie gebeurt in twee stappen: de *E*- en de *M*-stap. Dit zijn precies de stappen van het EM-algoritme.

Definitie 3.2 (EM-algoritme). *Zij $\hat{\theta}_m$ de m -de schatter van parameter θ . Voor de $(m+1)$ -de schatter, voer de volgende stappen uit.*

1. *E-stap: bereken $Q(\theta | \hat{\theta}_m, \mathbf{x})$.*
2. *M-stap: maximaliseer $Q(\theta | \hat{\theta}_m, \mathbf{x})$ en neem*

$$\hat{\theta}_{m+1} = \arg \max_{\theta} [Q(\theta | \hat{\theta}_m, \mathbf{x})].$$

We kunnen laten zien dat de schatter van de parameter θ bij het gebruik van dit algoritme steeds beter wordt. Met andere woorden, de likelihood wordt steeds hoger:

$$L(\hat{\theta}_{m+1} | \mathbf{x}) \geq L(\hat{\theta}_m | \mathbf{x}).$$

We bewijzen hiervoor eerst een lemma.

Lemma 3.1. *Er geldt*

$$\mathbb{E}[\log(k(z | \hat{\theta}_{m+1}, x)) | \hat{\theta}_m, x] \leq \mathbb{E}[\log(k(z | \hat{\theta}_m, x)) | \hat{\theta}_m, x].$$

Bewijs. We gebruiken de ongelijkheid van Jensen. Dan vinden we

$$\begin{aligned} \mathbb{E}\left[\log\left(\frac{k(z | \hat{\theta}_{m+1}, x)}{k(z | \hat{\theta}_m, x)}\right) | x, \hat{\theta}_m\right] &\leq \log\left(\mathbb{E}\left[\frac{k(z | \hat{\theta}_{m+1}, x)}{k(z | \hat{\theta}_m, x)} | x, \hat{\theta}_m\right]\right) \\ &= \log\left(\int_z \frac{k(z | \hat{\theta}_{m+1}, x)}{k(z | \hat{\theta}_m, x)} k(z | \hat{\theta}_m, x) dz\right) \\ &= \log\left(\int_z k(z | \hat{\theta}_{m+1}, x) dz\right) \\ &= \log(1) \\ &= 0 \end{aligned}$$

We vinden dus

$$\mathbb{E}\left[\log\left(\frac{k(z | \hat{\theta}_{m+1}, x)}{k(z | \hat{\theta}_m, x)}\right) | x, \hat{\theta}_m\right] \leq 0$$

en kunnen de linkerkant van de ongelijkheid uitwerken tot

$$\mathbb{E}\left[\log(k(z | \hat{\theta}_{m+1}, x)) | x, \hat{\theta}_m\right] - \mathbb{E}\left[\log(k(z | \hat{\theta}_m, x)) | x, \hat{\theta}_m\right] \leq 0.$$

■

We kunnen met dit resultaat de volgende stelling bewijzen.

Stelling 3.2. *De schatter van θ wordt steeds beter:*

$$L(\hat{\theta}_{m+1} | x) \geq L(\hat{\theta}_m | x).$$

Bewijs. We weten dat $\hat{\theta}_{m+1}$ de functie $Q(\theta | \hat{\theta}_m, x)$ maximaliseert. Dat betekent

$$Q(\hat{\theta}_{m+1} | \hat{\theta}_m, \mathbf{x}) \geq Q(\hat{\theta}_m | \hat{\theta}_m, \mathbf{x}), \text{ voor } \hat{\theta}_m \neq \hat{\theta}_{m+1},$$

oftewel

$$\mathbb{E}[\log(L^c(\hat{\theta}_{m+1} | x, z)) | \hat{\theta}_m, x] \geq \mathbb{E}[\log(L^c(\hat{\theta}_m | x, z)) | \hat{\theta}_m, x].$$

Gebruiken we nu (6), dan zien we dat

$$\begin{aligned} \log(L(\hat{\theta}_{m+1} | x)) + \mathbb{E}[\log(k(z | \hat{\theta}_{m+1}, \mathbf{x})) | \hat{\theta}_m, x] \\ \geq \log(L(\hat{\theta}_m | \mathbf{x})) + \mathbb{E}[\log(k(z | \hat{\theta}_m, x)) | \hat{\theta}_m, x] \end{aligned}$$

Passen we nu lemma 3.1 toe, dan vinden we dat moet gelden

$$\log(L(\hat{\theta}_{m+1} | x)) \geq \log(L(\hat{\theta}_m | x)).$$

Nemen we nu de ϵ -macht aan beide kanten dan vinden we het resultaat.

■

3.2 De covariantie matrix

We hebben gezien dat het EM-algoritme gebruikt kan worden om een schatting te krijgen van een parameter van belang en dat deze schatting ook steeds beter wordt. Hoe passen we dit algoritme toe op ons probleem?

Met de aanname die we hebben gemaakt op de vorm van Σ kunnen we zien dat geldt

$$\begin{aligned}\sqrt{\det(\Sigma)} &= \det(\Sigma)^{\frac{1}{2}} \\ &= (\sigma^4)^{\frac{1}{2}} = \sigma^2, \text{ en} \\ \Sigma^{-1} &= \frac{1}{\sigma^4} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \frac{1}{\sigma^2} \mathbb{I}\end{aligned}$$

Hiermee kunnen we de kansdichtheidsfunctie van Z herschrijven. Breng in herinnering dat we voor het schatten van Σ de oorsprong als mikpunt hadden genomen en dat er daarom geldt $Z = \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$. Vanuit (1) kunnen we dan het volgende herleiden.

$$\begin{aligned}f(z) &= \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{1}{2}z^T\left(\frac{1}{\sigma^2}\mathbb{I}\right)z\right) \\ &= \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{1}{2\sigma^2}z^Tz\right) \\ &= \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{1}{2\sigma^2}\|z\|^2\right)\end{aligned}\tag{7}$$

We gaan het EM-algoritme dus gebruiken om de standaardafwijking σ te schatten.

Voor het toepassen van het algoritme hebben we de complete waarschijnlijkheidsfunctie nodig. Vanuit definitie 3.1 weten we dat we hiervoor de gezamenlijke kansdichtheid van $\mathbf{X} = (X_1, \dots, X_n)$ en $\mathbf{Z} = (Z_1, \dots, Z_n)$ gegeven σ^2 nodig hebben, genoteerd met $h(\mathbf{z}, \mathbf{x} \mid \sigma^2)$. Gebruiken we nu dat alle worpen, dus alle Z_i voor $1 \leq i \leq n$, onafhankelijk zijn, dan is de kansdichtheid voor Z_i gegeven door (7).

Voor de gezamenlijke kansdichtheid geldt het volgende.

$$\begin{aligned}L^c(\sigma^2 \mid \mathbf{Z}, \mathbf{X}) &= h(\mathbf{z}, \mathbf{x} \mid \sigma^2) \\ &= \mathbb{P}(\mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x} \mid \sigma^2) \\ &\downarrow \text{Alle } Z_i \text{ en } X_i \text{ zijn paarsgewijs onafhankelijk} \\ &= \prod_{i=1}^n \mathbb{P}(Z_i = z_i, X_i = x_i \mid \sigma^2) \\ &= \prod_{i=1}^n \mathbb{P}(Z_i = z_i, s(Z_i) = x_i \mid \sigma^2)\end{aligned}$$

Merk nu op dat $\mathbb{P}(Z_i = z_i, s(Z_i) = x_i \mid \sigma^2) = 0$ als $s(z_i) \neq x_i$. Omdat we het product nemen over $i \in \{1, \dots, n\}$, moet gelden dat $s(z_i) = x_i$ voor alle i , anders is het product gelijk aan nul. Met andere woorden, we kunnen de complete waarschijnlijkheidsfunctie schrijven als

$$L^c(\sigma^2 \mid \mathbf{z}, \mathbf{x}) = \begin{cases} \prod_{i=1}^n \mathbb{P}(Z_i = z_i \mid \sigma^2) & \text{als } s(z_i) = x_i \text{ voor alle } i \in \{1, \dots, n\}, \text{ ??} \\ 0 & \text{als er een } i \text{ is zodat } s(z_i) \neq x_i \end{cases}\tag{8}$$

(9)

$$\downarrow \text{gebruik (7)}\tag{10}$$

(11)

$$= \begin{cases} \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \|z_i\|^2\right) & \text{als } s(z_i) = x_i \text{ voor alle } i \in \{1, \dots, n\}, \\ 0 & \text{als er een } i \text{ is zodat } s(z_i) \neq x_i. \end{cases}$$

Het logaritme van de waarschijnlijkheidsfunctie is dan gegeven door

$$\log \left(L^c(\sigma^2 \mid \mathbf{z}, \mathbf{x}) \right) = \begin{cases} -n \log(2\pi) - n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \|z_i\|^2 & \text{als } s(z_i) = x_i \text{ voor alle } i \in \{1, \dots, n\}, \\ -\infty & \text{als er een } i \text{ is zodat } s(z_i) \neq x_i. \end{cases}$$

Nu we weten hoe het logaritme van de complete waarschijnlijkheidsfunctie eruit ziet, kunnen we de functie Q beschrijven. We nemen daartoe een beginschatting $\hat{\sigma}_0^2$ en zien dat geldt

$$\begin{aligned} Q(\sigma^2 \mid \hat{\sigma}_0^2, x) &= \mathbb{E} \left[\log \left(L^c(\sigma^2 \mid x, z) \right) \mid \hat{\sigma}_0^2, \mathbf{X} \right] \\ &= \begin{cases} -n \log(2\pi) - n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \mathbb{E} \left[\|Z_i\|^2 \mid X_i, \hat{\sigma}_0^2 \right] & \text{als } s(z_i) = x_i \text{ voor alle } i \in \{1, \dots, n\}, \\ -\infty & \text{als er een } i \text{ is zodat } s(z_i) \neq x_i. \end{cases} \end{aligned}$$

Voor de M-stap van het algoritme moeten we deze Q maximaliseren voor σ^2 . Voor deze maximalisatie merken we op dat de verwachtingswaarden in de som niet van σ^2 (maar in plaats daarvan van $\hat{\sigma}_0^2$) afhangen. We stellen daarom

$$C = \sum_{i=1}^n \mathbb{E} \left[\|Z_i\|^2 \mid X_i, \hat{\sigma}_0^2 \right].$$

We willen nu dus een σ^2 die $-n \log(2\pi) - n \log(\sigma^2) - \frac{C}{2\sigma^2}$ maximaliseert. We stellen hiervoor de afgeleide gelijk aan nul en lossen op voor σ^2 .

$$\begin{aligned} \frac{dQ}{d(\sigma^2)} &= -\frac{n}{\sigma^2} + \frac{C}{2(\sigma^2)^2} \\ 0 &= -\frac{n}{\sigma^2} + \frac{C}{2(\sigma^2)^2} \\ \frac{n}{\sigma^2} &= \frac{C}{2(\sigma^2)^2} \\ 0 &= \sigma^2(C - 2n\sigma^2) \\ \sigma^2 &= 0 \text{ of } \sigma^2 = \frac{C}{2n} \end{aligned}$$

Merk op dat voor $\sigma^2 = 0$ de functie Q niet gedefinieerd is, vanwege delen door σ^2 . De uitkomst van de M-stap van het algoritme is dus

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \mathbb{E} \left[\|Z_i\|^2 \mid X_i, \hat{\sigma}_0^2 \right]}{2n}.$$

Er rest ons nu nog om een expliciete uitdrukking te vinden van de som in de teller van de breuk.

We weten dat bij elke score X_i er een (aantal) vlak(ken) is/zijn waar onze pijl in heeft kunnen landen. Zo geldt bijvoorbeeld

- $X_i = 10$, dan zit de pijl in S10 (2 vlakken) of D5.
- $X_i = 18$, dan zit de pijl in S18 (2 vlakken), D9 of T6.
- $X_i = 50$, dan zit de pijl in DB.

We schrijven de vereniging van deze vlakken als $S_i = \bigcup_j A_j$, waarbij A_j de losse vlakjes zijn. Voor de voorbeelden van hierboven geldt dus

- $X_i = 10$, dan $S_i = A_1 \cup A_2 \cup A_3 = S10 \cup S10 \cup D5$.

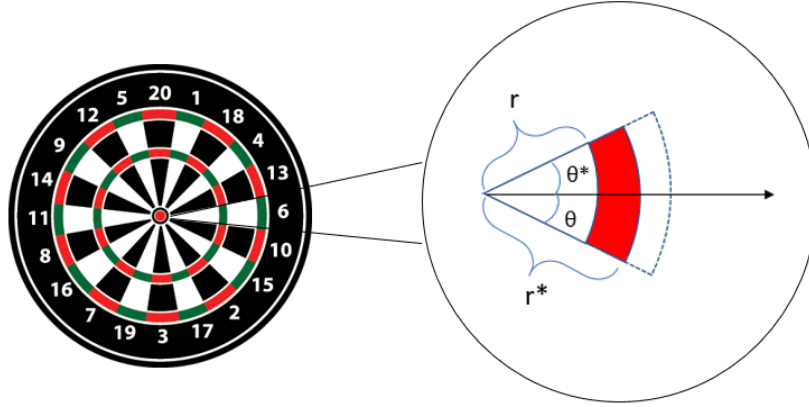
- $X_i = 18$, dan $S_i = A_1 \cup A_2 \cup A_3 \cup A_4 = S18 \cup S18 \cup D9 \cup D6$.
- $X_i = 50$, dan $S_i = A_1 = DB$.

Nu kunnen we de eigenschappen van een conditionele verwachtingswaarde gebruiken. Stellen we even $Z_i = \begin{pmatrix} z_{i,x} \\ z_{i,y} \end{pmatrix}$, dan geldt

$$\begin{aligned} \mathbb{E}[\|Z_i\|^2 \mid X_i, \hat{\sigma}_0^2] &= \mathbb{E}[Z_{i,x}^2 + Z_{i,y}^2 \mid Z_i \in \bigcup_j A_j, \hat{\sigma}_0^2] \\ &= \frac{\sum_j \int \int_{A_j} (z_{i,x}^2 + z_{i,y}^2) \exp\left(-\frac{z_{i,x}^2 + z_{i,y}^2}{2\hat{\sigma}_0^2}\right) dz_{i,x} dz_{i,y}}{\sum_j \int \int_{A_j} \exp\left(-\frac{z_{i,x}^2 + z_{i,y}^2}{2\hat{\sigma}_0^2}\right) dz_{i,x} dz_{i,y}} \end{aligned}$$

Om deze integraal uit te werken, veranderen we de coördinaten naar poolcoördinaten, zie sectie 2. We kunnen dan elk vlak A_j schrijven als $[r_j, r_j^*] \times [\theta_j, \theta_j^*]$, waarbij de parameters afhangen van de afmetingen van het dartbord. Zo geldt

- $A_j = DB$, dan $A_j = (0, 6.35) \times (0, 360^\circ)$.
- $A_j = T6$, dan kunnen we A_j voorstellen als $A_j = (99, 107) \times (-9^\circ, 9^\circ)$, zie figuur 4.



Figuur 4: Voorstelling van het triple 6 vlakje in poolcoördinaten.

We kunnen de conditionele verwachtingswaarde dan uitschrijven als

$$\begin{aligned} \mathbb{E}[\|Z_i\|^2 \mid X_i, \hat{\sigma}_0^2] &= \frac{\sum_j \int_{r_j}^{r_j^*} \int_{\theta_j}^{\theta_j^*} r^3 \cdot \exp\left(-\frac{r^2}{2\hat{\sigma}_0^2}\right) d\theta dr}{\sum_j \int_{r_j}^{r_j^*} \int_{\theta_j}^{\theta_j^*} r \cdot \exp\left(-\frac{r^2}{2\hat{\sigma}_0^2}\right) d\theta dr} \\ &= \frac{\sum_j \int_{r_j}^{r_j^*} (\theta_j - \theta_j^*) r^3 \cdot \exp\left(-\frac{r^2}{2\hat{\sigma}_0^2}\right) dr}{\sum_j \int_{r_j}^{r_j^*} (\theta_j - \theta_j^*) r \cdot \exp\left(-\frac{r^2}{2\hat{\sigma}_0^2}\right) dr} \\ &= \frac{\sum_j \int_{r_j}^{r_j^*} r^3 \cdot \exp\left(-\frac{r^2}{2\hat{\sigma}_0^2}\right) dr}{\sum_j \int_{r_j}^{r_j^*} r \cdot \exp\left(-\frac{r^2}{2\hat{\sigma}_0^2}\right) dr} \\ &\downarrow \text{ partiële integratie over } r \\ &= \frac{\sum_j \left((r_j^2 + 2\hat{\sigma}_0^2) \exp\left(-\frac{r_j}{2\hat{\sigma}_0^2}\right) - (r_j^{*2} + 2\hat{\sigma}_0^2) \exp\left(-\frac{r_j^*}{2\hat{\sigma}_0^2}\right) \right)}{\sum_j \exp\left(-\frac{r_j + r_j^*}{2\hat{\sigma}_0^2}\right)} \end{aligned}$$

3.3 Overzicht

In deze sectie hebben we gezien hoe het EM-algoritme gebruikt kan worden om een (steeds betere) schatting te krijgen van de covariantie matrix $\Sigma = \sigma^2 \mathbb{I}$. Hiervoor namen we een steekproef $\mathbf{X} = (X_1, \dots, X_n)$ van n scores, waarbij gericht werd op de oorsprong van het bord. We konden daarmee de volgende uitdrukking voor de E-stap van het algoritme maken.

$$Q(\sigma^2 \mid \hat{\sigma}_m^2, \mathbf{x}) = \begin{cases} -n \log(2\pi) - n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \mathbb{E} \left[\|Z_i\|^2 \mid X_i, \hat{\sigma}_m^2 \right] & \text{als } s(z_i) = x_i \forall i, \\ -\infty & \text{als } s(z_i) \neq x_i. \end{cases}$$

Voor de M-stap van het algoritme vonden we dat geldt

$$\hat{\sigma}_{m+1}^2 = \frac{\sum_{i=1}^n \mathbb{E} \left[\|Z_i\|^2 \mid X_i, \hat{\sigma}_m^2 \right]}{2n}.$$

Als laatst hebben we de verwachtingswaarde in de teller expliciet uit kunnen drukken als

$$\mathbb{E}[\|Z_i\|^2 \mid X_i, \hat{\sigma}_m^2] = \frac{\sum_j \left((r_j^2 + 2\hat{\sigma}_m^2) \exp\left(-\frac{r_j}{2\hat{\sigma}_m^2}\right) - (r_j^{*2} + 2\hat{\sigma}_m^2) \exp\left(-\frac{r_j^*}{2\hat{\sigma}_m^2}\right) \right)}{\sum_j \exp\left(-\frac{r_j + r_j^*}{2\hat{\sigma}_m^2}\right)},$$

waarbij we vlakken A_j konden vinden, gedefinieerd door $[r_j, r_j^*] \times [\theta_j, \theta_j^*]$, die een score van X_i opleverde.

4 De covariantie matrix volgens een geavanceerd model

In de vorige sectie zijn we er van uit gegaan dat de worp van een speler zowel in de x -richting als de y -richting met dezelfde waarde afwijkt. Bij de meeste darters, vooral bij de wat meer ervaren, is dit in de praktijk echter niet het geval. Als zo'n darter op de triple 20 mikt, gooit hij vaker te hoog of te laag (zodat hij een single 20 scoort) dan te ver naar links of rechts (zodat hij in de vijf of de één terecht komt). Voor het geavanceerde model gaan we er vanuit dat dit het geval is en doen we geen verdere aannames omtrent de covariantie matrix. Op die manier kunnen de varianties in beide richtingen verschillen.

4.1 De covariantie matrix

Het schatten van Σ gaat net als in sectie 2 met behulp van het EM-algoritme. We gaan eerst kijken wat het algoritme in dit geval betekent.

We gaan weer mikken op de oorsprong van het bord. Als we dan de stochastische vector $\mathbf{Z} = (Z_1, \dots, Z_n)$ bekijken, dan weten we

$$f(z_i) = \frac{1}{2\pi \det(\Sigma)^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} z_i^T \Sigma^{-1} z_i\right).$$

Omdat alle Z_i onafhankelijk en identiek verdeeld zijn, geldt dat

$$\begin{aligned} f(\mathbf{Z}) &= \prod_{i=1}^n f(z_i) \\ f(\mathbf{Z}) &= \left(\frac{1}{2\pi \det(\Sigma)^{\frac{1}{2}}}\right)^n \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^n z_i^T \Sigma^{-1} z_i\right). \end{aligned} \quad (12)$$

Omdat alleen de kansverdeling van Z_i verschilt, maar verder alle andere voorwaarden nog steeds van toepassing zijn, is de complete waarschijnlijkheidsfunctie hetzelfde als in sectie 2.2, zie (??). Combineren we deze waarschijnlijkheidsfunctie met (12), dan vinden we

$$L^c(\Sigma | \mathbf{X}, \mathbf{Z}) = \begin{cases} \left(\frac{1}{2\pi \det(\Sigma)^{\frac{1}{2}}}\right)^n \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^n z_i^T \Sigma^{-1} z_i\right) & \text{als } s(z_i) = x_i \text{ voor alle } i, \\ 0 & \text{als er een } i \text{ is zodat } s(z_i) \neq x_i. \end{cases}$$

Nemen we nu het logaritme, dan vinden we

$$\log\left(L^c(\Sigma | \mathbf{X}, \mathbf{Z})\right) = \begin{cases} -n \log(2\pi) - \frac{n}{2} \log(\det(\Sigma)) - \frac{1}{2} \sum_{i=1}^n z_i^T \Sigma^{-1} z_i & \text{als } s(z_i) = x_i \text{ voor alle } i, \\ -\infty & \text{als er een } i \text{ is zodat } s(z_i) \neq x_i. \end{cases} \quad (13)$$

We kunnen de uitdrukking in de som vereenvoudigen met het volgende lemma.

Lemma 4.1. $\sum_{i=1}^n z_i^T \Sigma^{-1} z_i = \text{tr}\left(\Sigma^{-1} \sum_{i=1}^n z_i z_i^T\right)$.

Bewijs. Er geldt dat $\sum_{i=1}^n z_i^T \Sigma^{-1} z_i$ een scalar is, geen vector of matrix. Dat betekent dat deze som hetzelfde is als het spoor (engels: trace) van de som. Dat wil zeggen

$$\sum_{i=1}^n z_i^T \Sigma^{-1} z_i = \text{tr}\left(\sum_{i=1}^n z_i^T \Sigma^{-1} z_i\right).$$

Het spoor van een matrix heeft de eigenschap dat de volgorde van vermenigvuldigen niet uitmaakt. Neem bijvoorbeeld twee matrices A en B , dan geldt $\text{tr}(A^T B A) = \text{tr}(B A A^T)$. Combineren we deze eigenschap met de lineariteit van het spoor, dan kunnen we bovenstaande uitdrukking schrijven als

$$\text{tr}\left(\sum_{i=1}^n z_i^T \Sigma^{-1} z_i\right) = \text{tr}\left(\sum_{i=1}^n \Sigma^{-1} z_i z_i^T\right).$$

Omdat de Σ niet van de index van de som afhangt, kunnen we hem uit de som halen. Dit geeft het lemma. ■

Gebruiken we het lemma, dan kunnen we (13) dus schrijven als

$$\log \left(L^c(\Sigma \mid \mathbf{X}, \mathbf{Z}) \right) = \begin{cases} -n \log(2\pi) - \frac{n}{2} \log(\det(\Sigma)) - \frac{1}{2} \text{tr} \left(\Sigma^{-1} \sum_{i=1}^n z_i z_i^T \right) & \text{als } s(z_i) = x_i, \\ -\infty & \text{als } s(z_i) \neq x_i. \end{cases}$$

Gegeven deze complete waarschijnlijkheidsfunctie kunnen we weer bekijken wat de E-stap van het algoritme wordt. We nemen weer een beginschatting $\hat{\Sigma}_0$, dan geldt

$$\begin{aligned} Q(\Sigma \mid \hat{\Sigma}_0, \mathbf{x}) &= \mathbb{E} \left[\log \left(L^c(\Sigma \mid \mathbf{x}, \mathbf{z}) \right) \mid \mathbf{X} \right] \\ &= -n \log(2\pi) - \frac{n}{2} \log(\det(\Sigma)) - \frac{1}{2} \text{tr} \left(\Sigma^{-1} \sum_{i=1}^n \mathbb{E} \left[Z_i Z_i^T \mid X_i, \hat{\Sigma}_0 \right] \right) \end{aligned}$$

Voor de M-stap willen de Q weer maximaliseren. Daartoe differentiëren we Q , stellen de afgeleide gelijk aan nul en lossen we de zo onstane vergelijking op voor Σ . Voor het differentiëren gebruiken we twee lemma's.

Lemma 4.2. *Gegeven twee matrices $A \in \mathbb{R}^{n \times m}$ en $B \in \mathbb{R}^{m \times n}$. Dan geldt $\frac{d}{dA} \text{tr}(AB) = B^T$.*

Bewijs. We bewijzen het lemma door middel van het uitschrijven van het spoor, zie [4].

$$\begin{aligned} \text{tr}(AB) &= \text{tr} \begin{pmatrix} \leftarrow & \vec{a}_1 & \rightarrow \\ \leftarrow & \vec{a}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \vec{a}_n & \rightarrow \end{pmatrix} \cdot \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{b}_1 & \vec{b}_2 & \cdots & \vec{b}_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \\ &= \text{tr} \begin{pmatrix} \vec{a}_1 \vec{b}_1 & \vec{a}_1 \vec{b}_2 & \cdots & \vec{a}_1 \vec{b}_n \\ \vec{a}_2 \vec{b}_1 & \vec{a}_2 \vec{b}_2 & \cdots & \vec{a}_2 \vec{b}_n \\ \vdots & \vdots & \ddots & \vdots \\ \vec{a}_n \vec{b}_1 & \vec{a}_n \vec{b}_2 & \cdots & \vec{a}_n \vec{b}_n \end{pmatrix} \\ &= \vec{a}_1 \vec{b}_1 + \vec{a}_2 \vec{b}_2 + \cdots + \vec{a}_n \vec{b}_n \\ &= \sum_{i=1}^m a_{1i} b_{i1} + \sum_{i=1}^m a_{2i} b_{i2} + \cdots + \sum_{i=1}^m a_{ni} b_{in} \end{aligned}$$

Nemen we nu de afgeleide naar a_{ij} , dan geldt

$$\frac{\partial}{\partial a_{ij}} \text{tr}(AB) = b_{ji}.$$

Hieruit volgt

$$\frac{d}{dA} \text{tr}(AB) = B^T.$$

■

Lemma 4.3. Gegeven een matrix $A \in \mathbb{R}^{n \times n}$. Dan geldt $\frac{d}{dA} \log(\det(A)) = (A^{-1})^T$.

Bewijs. Om dit lemma te bewijzen, gebruiken we twee andere matrices, zie [5].

1. De cofactor matrix van A , genoteerd met C .
2. De adjugate matrix van A , genoteerd met $\text{adj}(A)$. Merk op dat $\text{adj}(A) = C^T$.

We weten dat de inverse matrix van A dan te schrijven is als $A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$ en dus

$$(A^{-1})^T = \frac{1}{\det(A)} C. \quad (14)$$

We weten ook dat we de $\det(A)$ kunnen schrijven met behulp van de cofactor matrix: neem een $i \in \{1, \dots, n\}$, dan geldt

$$\det(A) = \sum_{k=1}^n A_{ik} C_{ik}.$$

We bekijken nu de partiële afgeleide naar element A_{ij} . We zien dat geldt

$$\begin{aligned} \frac{\partial}{\partial A_{ij}} \det(A) &= \sum_{k=1}^n \frac{\partial}{\partial A_{ij}} (A_{ik} C_{ik}) \\ &\downarrow \text{gebruik productregel voor differentiëren} \\ &= \sum_{k=1}^n \left(\frac{\partial A_{ik}}{\partial A_{ij}} C_{ik} + A_{ik} \frac{\partial C_{ik}}{\partial A_{ij}} \right) \end{aligned}$$

Welnu, als $k \neq j$, dan geldt $\frac{\partial A_{ik}}{\partial A_{ij}} = 0$. Anders is deze partiële afgeleide gelijk aan één. Er geldt dus

$$\frac{\partial}{\partial A_{ij}} \det(A) = C_{ij} + \sum_{k=1}^n A_{ik} \frac{\partial C_{ik}}{\partial A_{ij}}.$$

Verder gebruiken we dat alle elementen uit A die het element C_{ik} beïnvloeden, niet op rij i of kolom k liggen. Daarom geldt dat $\frac{\partial C_{ik}}{\partial A_{ij}} = 0$ voor alle k . We houden dus over

$$\frac{\partial}{\partial A_{ij}} \det(A) = C_{ij}.$$

Gebruiken we nu (14), dan zien we dat geldt

$$\frac{d}{dA} \det(A) = \left(\det(A) \cdot (A^{-1})^T \right). \quad (15)$$

Nu kunnen we het logaritme van de determinant bekijken. We differentiëren dit logaritme met behulp van de kettingregel.

$$\begin{aligned} \frac{d}{dA} \left(\log(\det(A)) \right) &= \frac{1}{\det(A)} \cdot \frac{d}{dA} \det(A) \\ &\downarrow \text{gebruik (15)} \\ &= \frac{1}{\det(A)} \cdot \det(A) \cdot (A^{-1})^T \\ &= (A^{-1})^T. \end{aligned}$$

■

Met deze twee lemma's kunnen we de M-stap van het algoritme uitwerken. Voordat we beginnen met differentiëren merken we nog een gevolg van de rekenregels voor logaritme op:

$$-\frac{n}{2} \log(\det(\Sigma)) = \frac{n}{2} \log(\det(\Sigma^{-1})).$$

We zien dan dat geldt

$$\begin{aligned} \frac{d}{d\Sigma^{-1}} Q &= \frac{d}{d\Sigma^{-1}} \left(-n \log(2\pi) + \frac{n}{2} \log(\det(\Sigma^{-1})) - \frac{1}{2} \text{tr} \left(\Sigma^{-1} \sum_{i=1}^n \mathbb{E} \left[Z_i Z_i^T | X_i, \hat{\Sigma}_0 \right] \right) \right) \\ &= \frac{n}{2} \frac{d}{d\Sigma^{-1}} \left(\log(\det(\Sigma^{-1})) \right) - \frac{1}{2} \frac{d}{d\Sigma^{-1}} \left(\text{tr} \left(\Sigma^{-1} \sum_{i=1}^n \mathbb{E} \left[Z_i Z_i^T | X_i, \hat{\Sigma}_0 \right] \right) \right) \\ &\downarrow \text{gebruik lemma 4.2 en 4.3} \\ &= \frac{n}{2} ((\Sigma^{-1})^{-1})^T - \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_0 \right] \\ &\downarrow \Sigma \text{ is symmetrisch} \\ &= \frac{n}{2} \Sigma - \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_0 \right] \end{aligned}$$

We stellen dit gelijk aan nul en lossen op voor Σ .

$$\begin{aligned} 0 &= \frac{n}{2} \Sigma - \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_0 \right] \\ \frac{n}{2} \Sigma &= \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_0 \right] \\ \Sigma &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_0 \right]. \end{aligned}$$

Al met al kunnen we dus concluderen dat het EM-algoritme de volgende resultaten geeft.

- De E-stap van het EM-algoritme is

$$Q(\Sigma | \hat{\Sigma}_m, \mathbf{x}) = \begin{cases} -n \log(2\pi) - \frac{n}{2} \log(\det(\Sigma)) - \frac{1}{2} \text{tr} \left(\Sigma^{-1} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_m \right] \right) & \text{als } s(z_i) = x_i \forall i, \\ -\infty & \text{als } \exists i \text{ zodat } s(z_i) \neq x_i. \end{cases}$$

- De M-stap van het EM-algoritme geeft

$$\hat{\Sigma}_{m+1} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[Z_i^T Z_i | X_i, \hat{\Sigma}_m \right]$$

Het enige wat we nu nog zoeken is een uitdrukking voor de verwachtingswaarde in de som, gegeven een schatter $\hat{\Sigma}_m$. Hiervoor gebruiken we een andere veelgebruikte statistische techniek: een Monte Carlo simulatie in de vorm van *Importance Sampling*. We bekijken eerst de theorie en daarna bekijken we hoe we de techniek kunnen toepassen.

4.2 Importance sampling

De verwachtingswaarde die we willen berekenen is uit te drukken met een integraal. Voor notationele redenen beschouwen we voor deze subsectie de volgende verwachtingswaarde, waarbij we de indices i en m en het dakje van de schatter weglaten.

$$\mathbb{E}\left[Z^T Z|X, \Sigma\right] = \int_z z^T z \cdot f(z | x, \Sigma) dz.$$

Deze integraal is erg lastig uit te werken met de methoden die bekend zijn. Daarom gaan we deze integraal numeriek benaderen: met het genereren van (een groot aantal) willekeurige getallen en *importance sampling*.

4.2.1 (Pseudo)willekeurige getallen

Het genereren van (pseudo)willekeurige getallen wordt gebruikt bij het simuleren van een uniforme toevalsvariabele op het interval $[0,1]$. De meest gebruikte methode om zo'n (pseudo)willekeurig getal te genereren, is het gebruik van een recursieve uitdrukking met beginwaarde x_0 . Met de recursieve uitdrukking gegeven door

$$x_{n+1} = ax_n \pmod{m},$$

waarbij a en m twee constanten zijn, kunnen we een rij getallen genereren. Deze rij bevat een willekeurige volgorde van getallen tussen 0 en $m - 1$. Vanuit deze rij kunnen we de (pseudo)willekeurige getallen uit het interval $[0,1]$ berekenen als $\frac{x_i}{m}$. Een (pseudo)willekeurige n -dimensionale vector is dan te simuleren door n (pseudo)willekeurige getallen te genereren.

Het is echter niet zo dat deze getallen volledig willekeurig zijn: de rij zal zich op een gegeven moment gaan herhalen. Dit is de reden dat we de getallen *pseudowillekeurig* noemen. De keuzes voor de waarde van a en m moeten er voor zorgen dat het zo veel mogelijk iteraties vergt voordat de rij zich herhaalt. De beste keuze voor m blijkt het grootste priemgetal te zijn dat past in het geheugen van een computer. Bij een 32-bits computer is dat bijvoorbeeld $m = 2^{31} - 1$. Voor a geldt dan $a = 7^5$.

4.2.2 Monte Carlo benadering van integralen

Beschouw de integraal

$$I = \int_0^1 g(x) dx.$$

We kunnen ons deze integraal ook voorstellen met behulp van een stochastische variabele U die uniform verdeeld is over $[0,1]$. Dan geldt $I = \mathbb{E}[g(U)]$. Als we nu de wet van de grote getallen (law of large numbers) gebruiken, die zegt dat

$$\sum_{i=1}^k \frac{g(U_i)}{k} \rightarrow \mathbb{E}[g(U)] \text{ voor } k \rightarrow \infty,$$

dan kunnen we I benaderen door heel veel willekeurige getallen u_k te genereren ($k \rightarrow \infty$) en dan het gemiddelde te nemen over $g(u_k)$.

Ook een integraal als

$$I = \int_a^b g(x) dx,$$

kan benaderd worden met deze Monte Carlo simulatie van willekeurige getallen. We passen daartoe een relatief eenvoudige transformatie toe: $y = \frac{x-a}{b-a}$. Dan is I te schrijven als

$$I = \int_0^1 h(y) dy,$$

met $h(y) = (b-a)g(a+(b-a)y)$. Nemen we nu een over $[0, 1]$ uniform verdeelde variabele Y , dan kunnen we I benaderen door de wet van grote getallen toe te passen.

Een oneigenlijke integraal is ook te benaderen met deze techniek. Bekijk de integraal

$$I = \int_0^{\infty} g(x) dx.$$

Passen we de substitutie $y = \frac{1}{x+1}$ toe, dan is de integraal te schrijven als

$$I = \int_0^1 h(y) dy$$

met $h(y) = \frac{g(\frac{1}{y}-1)}{y^2}$.

Ook meervoudige integralen zijn te benaderen met deze methode, zoals

$$I = \int_0^1 \int_0^1 \cdots \int_0^1 g(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n.$$

Om zo'n integraal te benaderen nemen we k keer een verzameling van n willekeurige getallen:

$$\begin{aligned} \text{Verzameling 1} &: u_1^1, \dots, u_n^1 \\ \text{Verzameling 2} &: u_1^2, \dots, u_n^2 \\ &\vdots \\ \text{Verzameling } k &: u_1^k, \dots, u_n^k \end{aligned}$$

De benadering van de integraal wordt dan het gemiddelde van $g(u_1, \dots, u_n)$ over alle k verzamelingen. Zo'n meervoudige integraal is op een zelfde wijze als bij de enkelvoudige integraal te generaliseren naar integralen met andere grenzen en oneigenlijke integralen.

4.2.3 Importance sampling

Importance sampling is een geavanceerde Monte Carlo benadering van integralen. Het werkt als volgt. Beschouw een stochastische vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ met een gezamenlijke kansdichtheid $f(\mathbf{x}) = f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ waarvan we een bepaalde verwachtingswaarde willen berekenen:

$$I = \mathbb{E}[h(\mathbf{X})] = \int_{\infty}^{\infty} h(\mathbf{x}) \cdot f(\mathbf{x}) d\mathbf{x}.$$

Het idee is dan om een derde kansdichtheidsfunctie, g , te vinden zó dat $f(\mathbf{x}) = 0$ als $g(\mathbf{x}) = 0$. Dan kunnen we de integraal I schrijven als

$$\begin{aligned} I &= \int_{\infty}^{\infty} h(\mathbf{x}) \cdot f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\infty}^{\infty} \frac{h(\mathbf{x}) \cdot f(\mathbf{x})}{g(\mathbf{x})} \cdot g(\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}_g \left[\frac{h(\mathbf{X}) \cdot f(\mathbf{X})}{g(\mathbf{X})} \right] \end{aligned}$$

Deze verwachtingswaarde kunnen we berekenen met de Monte Carlo benadering (zie sectie 3.3.2). We nemen daartoe k verzamelingen van n willekeurige getallen als $\mathbf{x}^i = \{x_1, x_2, \dots, x_n\}$, $1 \leq i \leq k$, waarbij \mathbf{x}^i

trekkingen zijn van een variabele met kansdichtheid g , en benaderen de verwachtingswaarde in kwestie als het gemiddelde van $\frac{h(\mathbf{x}^i) \cdot f(\mathbf{x}^i)}{g(\mathbf{x}^i)}$.

In het laatste stuk van deze subsectie bekijken we hoe importance sampling gebruikt kan worden voor conditionele verwachtingswaarden. We beschouwen weer een stochastische vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ met kansdichtheidsfunctie $f(\mathbf{x})$. Als we de verwachtingswaarde van een willekeurige functie $h(\mathbf{x})$ willen weten waarbij \mathbf{X} aan een bepaalde voorwaarde moet voldoen (d.w.z. $\mathbf{X} \in \mathcal{A}$, voor een verzameling \mathcal{A}), moeten we berekenen

$$I = \mathbb{E}[h(\mathbf{X}) \mid \mathbf{X} \in \mathcal{A}] = \int_{-\infty}^{\infty} h(\mathbf{x}) \cdot f(\mathbf{x} \mid \mathbf{X} \in \mathcal{A}) \, d\mathbf{x}.$$

Vanuit de definitie van voorwaardelijke kansen weten we dat geldt

$$f(\mathbf{x} \mid \mathbf{X} \in \mathcal{A}) = \frac{\mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{X} \in \mathcal{A})}{\mathbb{P}(\mathbf{X} \in \mathcal{A})}.$$

Merk op dat deze voorwaardelijke kans nul is als $\mathbf{x} \notin \mathcal{A}$, dus kunnen we schrijven

$$f(\mathbf{x} \mid \mathbf{X} \in \mathcal{A}) = \frac{f(\mathbf{x}) \mathbb{I}_{\mathcal{A}}(\mathbf{x})}{\mathbb{P}(\mathbf{X} \in \mathcal{A})}$$

waarbij $\mathbb{P}(\mathbf{X} \in \mathcal{A})$ de kans dat \mathbf{X} aan de gestelde voorwaarde voldoet en $\mathbb{I}_{\mathcal{A}}(\mathbf{X}) = 1$ als $\mathbf{X} \in \mathcal{A}$ en $\mathbb{I}_{\mathcal{A}}(\mathbf{X}) = 0$ als $\mathbf{X} \notin \mathcal{A}$.

We berekenen dit als de kans op \mathcal{A} en daarom kunnen we deze kans als constante beschouwen. We herschrijven integraal I dan als

$$\begin{aligned} I &= \int_{-\infty}^{\infty} h(\mathbf{x}) \cdot \frac{f(\mathbf{x}) \mathbb{I}_{\mathcal{A}}(\mathbf{x})}{\mathbb{P}(\mathbf{X} \in \mathcal{A})} \, d\mathbf{x} \\ &= \int_{\mathcal{A}} h(\mathbf{x}) \cdot \frac{f(\mathbf{x}) \mathbb{I}_{\mathcal{A}}(\mathbf{x})}{\mathbb{P}(\mathbf{X} \in \mathcal{A})} \, d\mathbf{x} \\ &= \frac{1}{\mathbb{P}(\mathbf{X} \in \mathcal{A})} \int_{\mathcal{A}} h(\mathbf{x}) \cdot f(\mathbf{x}) \mathbb{I}_{\mathcal{A}}(\mathbf{x}) \, d\mathbf{x} \\ &= \frac{1}{\mathbb{E}[\mathbb{I}_{\mathcal{A}}(\mathbf{X})]} \mathbb{E}[h(\mathbf{X}) \cdot \mathbb{I}_{\mathcal{A}}(\mathbf{X})], \end{aligned}$$

Op deze manier hebben we de conditionele verwachtingswaarde $E[h(\mathbf{X}) \mid \mathbf{X} \in \mathcal{A}]$ omgezet in twee niet-conditionele verwachtingswaarden. Deze laatste twee niet-conditionele verwachtingswaarden kunnen we benaderen met de methode die hiervoor genoemd is.

4.3 Importance Sampling voor het dartmodel

Bij het berekenen van de M-stap van het EM-algoritme vonden we

$$\hat{\Sigma}_{m+1} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i^T Z_i \mid X_i, \hat{\Sigma}_m].$$

Het probleem was daarbij dat we de verwachtingswaarde in de som niet eenvoudig te analyseren is. We gebruiken daarom importance sampling om een benadering te geven van de verwachtingswaarden.

4.3.1 Willekeurige posities genereren

In sectie 2 is beschreven dat we een dartbord handig voor konden stellen door middel van poolcoördinaten, met een hoek $\theta \in [0^\circ, 360^\circ]$ en een afstand tot de oorsprong $r \in [0, 170]$. Een willekeurige positie kan dan

gegeneert worden door een trekking uit een uniforme verdeling over $[0, 360]$ en een trekking uit een uniforme verdeling over $[0, 170]$. Dit gaat als volgt.

Zij x een trekking uit een uniforme verdeling over $[0, 1]$. Dan kan een trekking uit $[0, 360]$ gesimuleerd worden als $\theta = 360 \cdot x$. Op eenzelfde manier wordt een trekking uit $[0, 170]$ gesimuleerd als $r = 170 \cdot x$.

Om nu een uniforme verdeling over een vakje van het dartbord te simuleren, bijvoorbeeld de triple 13 met $\theta \in (9, 27)$ en $r \in (99, 107)$, trekken we willekeurig een θ en een r zoals beschreven in de vorige alinea, nemen $z = (\theta, r)$ en concluderen

$$g(z) = \begin{cases} \frac{1}{|T_{13}|} \approx \frac{1}{259} & \text{als } \theta \in (9, 27) \text{ en } r \in (99, 107), \\ 0 & \text{als } \theta \notin (9, 27) \text{ en/of } r \notin (99, 107). \end{cases}$$

4.3.2 Toepassing van Importance Sampling

Als we de verwachtingswaarde $\mathbb{E}[Z_i Z_i^T | X_i, \hat{\Sigma}_m]$ bekijken en de theorie over conditionele verwachtingswaarde toepassen, dan geldt

$$\begin{aligned} \mathbb{E}[Z_i^T Z_i | X_i, \hat{\Sigma}_m] &= \int_{-\infty}^{\infty} z_i^T z_i \cdot f_Z(z_i | x_i, \hat{\Sigma}_m) dz_i \\ &= \int_{-\infty}^{\infty} z_i^T z_i \cdot \frac{f_Z(z_i | \hat{\Sigma}_m) \mathbb{I}_{S_i}(z_i)}{\mathbb{P}(z_i \in S_i)} dz_i \\ &= \frac{1}{\mathbb{E}[\mathbb{I}_{S_i}(Z_i) | \hat{\Sigma}_m]} \mathbb{E}[Z_i^T Z_i \cdot \mathbb{I}_{S_i}(Z_i) | \hat{\Sigma}_m]. \end{aligned} \quad (16)$$

Hierbij geldt dat S_i de (vereniging van) regio('s) van het bord is die een score X_i oplevert en dus $\mathbb{I}_{S_i}(Z_i) = 1$ als $s(Z_i) = X_i$ en $\mathbb{I}_{S_i}(Z_i) = 0$ als $s(Z_i) \neq X_i$. Merk op dat er nog steeds een conditionele verwachtingswaarde staat, maar dat er geconditioneerd wordt op een (bekende) covariantie matrix in plaats van een variabele.

We kunnen de twee verwachtingswaarden nu benaderen met importance sampling. Daarvoor herschrijven we

$$\begin{aligned} \mathbb{E}[Z_i^T Z_i \cdot \mathbb{I}_{S_i}(Z_i) | \hat{\Sigma}_m] &= \int_{-\infty}^{\infty} z_i^T z_i \cdot \mathbb{I}_{S_i} \cdot f_Z(z_i | \hat{\Sigma}_m) dz_i \\ &= \int_{-\infty}^{\infty} |S_i| \cdot z_i^T z_i \frac{\mathbb{I}_{S_i}(z_i)}{|S_i|} f_Z(z_i | \hat{\Sigma}_m) dz_i \\ &= \int_{-\infty}^{\infty} \frac{z_i^T z_i \cdot f_Z(z_i | \hat{\Sigma}_m)}{\frac{1}{|S_i|}} \cdot \frac{\mathbb{I}_{S_i}(z_i)}{|S_i|} dz_i, \end{aligned}$$

waarbij $|S_i|$ de oppervlakte van S_i is. Merk op dat we dit kunnen zien als vermenigvuldigen en delen door een functie g die gedefinieerd is als $g(Z_i) = \frac{1}{|S_i|} = \frac{\mathbb{I}_{S_i}(Z_i)}{|S_i|}$ (en dus een uniforme verdeling over S_i is). Dan geldt

$$\begin{aligned} \mathbb{E}[Z_i^T Z_i \cdot \mathbb{I}_{S_i}(Z_i) | \hat{\Sigma}_m] &= \mathbb{E}\left[\frac{Z_i^T Z_i \cdot f_Z(Z_i | \hat{\Sigma}_m)}{g(Z_i)}\right] \\ &= \mathbb{E}[|S_i| \cdot V_i^T V_i f_Z(V_i) | \hat{\Sigma}_m] \\ &= |S_i| \cdot \mathbb{E}[V_i^T V_i f_Z(V_i) | \hat{\Sigma}_m], \end{aligned}$$

waarbij V_i uniform verdeeld zijn over S_i , voor $0 \leq i \leq n$. Op eenzelfde wijze geldt

$$\mathbb{E}[\mathbb{I}_{S_i}(Z_i) | \hat{\Sigma}_m] = |S_i| \cdot \mathbb{E}[f_Z(V_i | \hat{\Sigma}_m)]$$

en dus is (16) te schrijven als

$$\mathbb{E}[Z_i^T Z_i | X_i, \hat{\Sigma}_m] = \frac{\mathbb{E}[V_i^T V_i f_Z(V_i) | \hat{\Sigma}_m]}{\mathbb{E}[f_Z(V_i | \hat{\Sigma}_m)]}.$$

We genereren nu k pseudowillekeurige waarden v_i (zoals beschreven in 4.3.1) uit een uniforme verdeling over S_i en benaderen

$$\mathbb{E}_{\hat{\Sigma}_m}[Z_i^T Z_i | X_i] \approx \frac{1}{\frac{1}{k} \sum_{i=1}^k f_Z(v_i)} \cdot \frac{1}{k} \sum_{i=1}^k v_i^T v_i \cdot f_Z(v_i).$$

4.4 Overzicht

In deze sectie hebben we gezien hoe we de covariantie matrix van een darter kunnen schatten als we geen aannames op die matrix doen. Met behulp van het EM-algoritme vonden we een rij schatters recursief uitgedrukt als

$$\hat{\Sigma}_{m+1} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i^T Z_i | X_i, \hat{\Sigma}_m].$$

Daarna hebben we gezien hoe de verwachtingswaarde in de som benadert kan worden met importance sampling als

$$\mathbb{E}[Z_i^T Z_i | X_i, \hat{\Sigma}_m] \approx \frac{1}{\frac{1}{k} \sum_{i=1}^k f_Z(v_i | \hat{\Sigma}_m)} \cdot \frac{1}{k} \sum_{i=1}^k v_i^T v_i \cdot f_Z(v_i | \hat{\Sigma}_m),$$

waarbij V_i uniform verdeeld is over het gebied S_i . Dit gebied S_i geeft recht op een score X_i .

5 Hypothesetoetsen

In de vorige secties is aangetoond hoe we de covariantie matrix van een darter kunnen schatten (secties 3 en 4) en hoe we het optimale mikpunt kunnen berekenen als we die covariantie matrix eenmaal hebben geschat (sectie 2). Het is interessant om te onderzoeken of er ook significant hogere scores worden gehaald als een darter op dat optimale mikpunt mikt. Voor dit onderzoek voeren we een tweedimensionale hypothesetoets uit (een hypothesetoets met twee steekproeven). In deze sectie bespreken we welke test we gaan gebruiken en hoe zo'n test in zijn werk gaat.

5.1 T- en Z-testen

Bij het uitvoeren van tweedimensionale Z - of T -toetsen onderzoeken we of er verschil zit in een kenmerk tussen een twee groepen. De waarden die dat kenmerk kan aannemen zijn willekeurig. We kunnen het kenmerk daardoor beschouwen als stochastische variabele voor de twee groepen: namelijk de stochastische variabele X (kenmerk voor groep één) en Y (kenmerk voor groep twee). Voor het uitvoeren van de toets nemen we aan dat X en Y onafhankelijk zijn.

Om te toetsen of er een significant verschil is tussen de twee groepen, stellen we twee hypothesen op:

$$\begin{aligned}H_0 : \mu_x &= \mu_y \\ H_1 : \mu_x &\neq \mu_y\end{aligned}$$

De nulhypothese (H_0) impliceert dat er geen significant verschil zit tussen de groepen, terwijl de alternatieve hypothese (H_1) wel een significant verschil impliceert. Er zijn ook andere formuleringen van de alternatieve hypothese mogelijk, waarbij het \neq -teken vervangen kan worden door een van de ongelijkheidstekens. Zodoende kan men onderscheid maken tussen éénzijdige (bij gebruik van ongelijkheidstekens) en tweezijdige (bij gebruik van \neq) hypothesetoetsen.

Afgaande op de hypothesen is $\mu_x - \mu_y$ de parameter die we willen onderzoeken. Omdat deze (populatie) gemiddeldes niet bekend zijn, gaan we ze schatten. Daartoe nemen we twee steekproeven $\mathbf{x} = (x_1, \dots, x_n)$ en $\mathbf{y} = (y_1, \dots, y_m)$. Merk op dat de steekproeven niet even groot hoeven te zijn. Van deze steekproeven berekenen we het steekproefgemiddelde (\bar{x}_n en \bar{y}_m respectievelijk) en de steekproefvariantie (S_x^2 en S_y^2 respectievelijk). Samen met de steekproefgrootten (n en m) kunnen we de waarde van een *test-statistiek* voor $\mu_x - \mu_y$ berekenen. Deze test-statistiek Z heeft de vorm

$$Z := \frac{\bar{x}_n - \bar{y}_m - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}.$$

Mochten er gegronde redenen zijn om aan te nemen dat de varianties van X en Y gelijk zijn, dan is de bovenstaande test-statistiek te vereenvoudigen naar

$$Z = \frac{\bar{x}_n - \bar{y}_m - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

De waarde S_p^2 , de zogenaamde "pooled" variantie, wordt hierbij berekend als

$$S_p^2 = \frac{1}{n + m - 2} \left((n - 1)S_x^2 + (m - 1)S_y^2 \right).$$

Merk op dat onder H_0 geldt dat $\mu_x - \mu_y = 0$ en dat deze kwantiteit dus wegvalt bij het berekenen van de waarde van de test-statistiek.

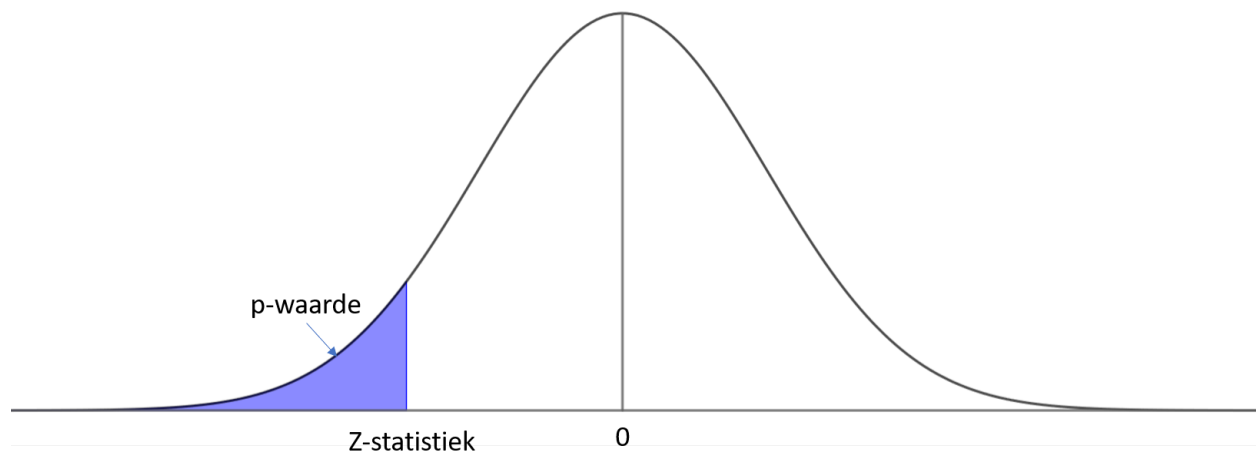
De test-statistiek is natuurlijk niet zomaar ontwikkeld: hij heeft een concrete kansverdeling. Onder H_0 , waarbij $\mu_x - \mu_y = 0$, geldt het volgende. Als de variabelen X en Y normaal verdeeld zijn, volgt de test-statistiek een student-t distributie met v vrijheidsgraden, waarbij

$$v = \left(\frac{S_x^2}{n} + \frac{S_y^2}{m} \right)^2 \left(\frac{\left(\frac{S_x^2}{n} \right)^2}{n-1} + \frac{\left(\frac{S_y^2}{m} \right)^2}{m-1} \right)^{-1}.$$

Als er aangenomen kan worden dat de populatievarianties gelijk zijn, dan vereenvoudigt v naar $v = n+m-2$. Zodoende wordt een T -test uitgevoerd.

Als de variabelen X en Y niet normaal verdeeld zijn, kunnen we de centrale limietstelling gebruiken om te constateren dat de test-statistiek standaard normaal verdeeld is. Als de test-statistiek deze verdeling heeft, noemen we de hypothesetoets een Z -test.

Er zijn naast de test-statistiek nog twee parameters nodig voordat we de nulhypothese kunnen toetsen. De eerste parameter is de p -waarde. De p -waarde geeft de kans dat de test-statistiek minstens zo extreem is als berekend aan de hand van de genomen steekproeven. We kunnen deze kans eenvoudig berekenen aan de hand van de kansverdelingen van de test-statistiek. Zie figuur 5.



Figuur 5: Kansverdeling Z -statistiek en p -waarde van een hypothesetoets met $H_1 : \mu_x < \mu_y$.

Naast deze p -waarde is het significantieniveau α van de toets een belangrijke parameter. De α geeft de grens aan tussen het accepteren dan wel verwerpen van H_0 .

- Als de p -waarde kleiner is dan α , dan verwerpen we H_0 .
- Als de p -waarde groter dan of gelijk is aan α , dan accepteren we H_0 .

We hebben nu gezien dat we met behulp van een steekproef de waarde van een test-statistiek kunnen berekenen, wat de kansverdeling van de test-statistiek is en hoe we daarmee de p -waarde kunnen berekenen. Als deze p -waarde kleiner is dan het significantieniveau α verwerpen we H_0 en concluderen we dat er verschil zit tussen de twee groepen.

5.2 Test voor dartmodel

Bij ons dartmodel is het te onderzoeken kenmerk de score per pijl. De twee groepen die we onderzoeken zijn de groep met scores waarbij gericht is op de triple 20 (de X) en een tweede groep scores waarbij gericht

is op het optimale mikpunt (de Y). We nemen steekproeven van grootte 102, oftewel $\mathbf{x} = x_1, \dots, x_{102}$ en $\mathbf{y} = (y_1, \dots, y_{102})$, zodat de steekproeven groot genoeg zijn. Merk op dat de eigenschap die we onderzoeken niet de (mik)positie is, zoals bij de vorige secties, maar de score van een pijl.

We definiëren de hypothesen als volgt.

$$\begin{aligned}H_0 &: \mu_x = \mu_y \\H_1 &: \mu_x < \mu_y\end{aligned}$$

De alternatieve hypothese impliceert hier dat de scores van groep Y significant hoger zijn dan die van groep X en dat de darter dus beter op zijn optimale mikpunt kan mikken. We voeren dus een éézijdige hypothesetoets uit en we gaan toetsen bij een significantieniveau van $\alpha = 0,05$.

Voordat de test-statistieken berekend gaan worden, bekijken we eerst hoe de test-statistiek eruit zien en welke kansverdeling deze heeft. Daarvoor bekijken we eerst de aannames die gemaakt kunnen/moeten worden.

- *X en Y zijn onafhankelijk:* omdat de steekproeven op verschillende dagen genomen zijn, en dus niet na elkaar, is het onwaarschijnlijk dat de worpen elkaar beïnvloed hebben. Dit lijkt dus een houdbare aanname.
- *De varianties zijn gelijk:* dit lijkt een moeilijker houdbare aanname. Omdat er bij de steekproeven op een ander punt gemikt is, zijn de kansen op bepaalde scores groter dan wel kleiner. Hierdoor liggen de behaalde scores op een andere manier ten opzichte van elkaar en kan de variantie dus ook verschillen.
- *X en Y zijn ruwweg normaal verdeeld:* in tegenstelling tot de posities van een pijl kunnen de scores niet oneindig veel waarden aannemen. De score van een pijl kan bijvoorbeeld niet lager dan nul zijn of hoger dan zestig. Het lijkt daarom waarschijnlijker dat een pijl een multinomiale distributie (generalisatie van een binomiale distributie) heeft, in plaats van een normale distributie.

Omdat we niet aannemen dat de varianties gelijk zijn, heeft de test-statistiek voor het dartmodel de vorm

$$Z := \frac{\bar{x}_n - \bar{y}_m - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}.$$

Daarnaast lijkt het onwaarschijnlijk dat de scores normaal verdeeld zijn en dus gebruiken we de centrale limietstelling om te concluderen dat de test-statistiek standaard normaal verdeeld is. Het significantieniveau waarbij deze Z -test getoetst zal gaan worden is $\alpha = 0.05$.

6 Resultaten

Voor het schatten van de standaardafwijking en de covariantie matrix en het berekenen van het optimale mikpunt is gebruik gemaakt van de code die beschikbaar is gesteld door de schrijvers van [1], zie [2]. De input van deze code is een lijst met 102 geworpen scores, waarbij gericht is op de oorsprong van het bord (double bullseye). Als output geeft de code een *heatmap* en een optimaal mikpunt.

Aan het onderzoek hebben veertien personen deelgenomen. Iedereen heeft drie datasets geworpen, allen op 2.37 meter afstand van een op 1.73 meter hoog hangend dartbord:

- De eerste dataset is gericht op de oorsprong van het bord. Met deze set is met behulp van de code een optimaal mikpunt berekend.
- Bij de tweede dataset werd er gericht op de triple 20.
- Bij de derde dataset werd er gericht op het berekende optimale mikpunt.

De tweede en de derde set zijn gebruikt op de hypothesetoets uit te voeren zoals beschreven is in sectie 4.

6.1 Het simpele model

In de figuren 8-13 aan het eind van deze sectie is de heatmap en het optimale mikpunt voor verschillende darters getoond. Hierbij moet vermeld worden dat alle personen die deelnamen aan het onderzoek geen ervaren maar wel aan elkaar gewaagde darters waren: de resultaten verschilden daarom niet zoveel onderling. Van de veertien darters presenteren we hierom de resultaten van vier darters. Om toch te onderzoeken hoe het model zich gedraagt bij meer accurate darters, zijn er twee darters gesimuleerd. Deze gesimuleerde darters worden aangegeven met s_1 en s_2 .

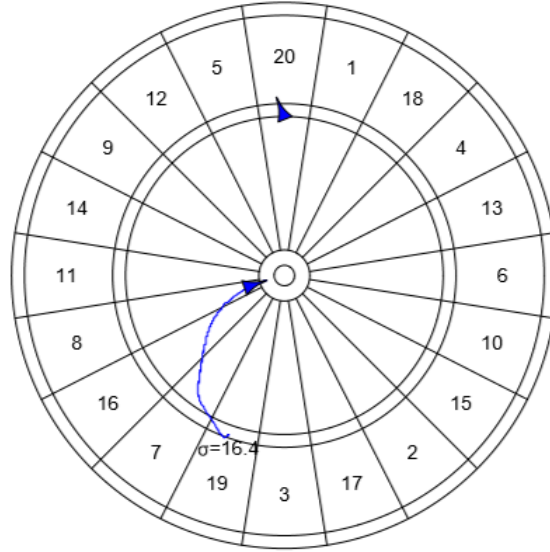
In de heatmap van elke darter (subfiguren 8a-13a) zijn de verwachte scores per pijl voor elk punt op en naast het dartbord weergegeven. Hierin is rood een lage verwachte score en geel/wit een hoge verwachte score per pijl. Het punt met de hoogste verwachte score per pijl is telkens weergegeven in subfiguren 8b-13b.

Er is mooi te zien dat voor hogere waarden van σ de verwachte scores een cirkelvormig patroon vormen (figuur 8a), terwijl lagere waarden van σ resulteren in kleinere gebieden met hogere verwachte scores. Als een darter erg goed kan gooien, zoals (de gesimuleerde) darter s_2 , dan is zelfs het dartboard te onderscheiden in de heatmap. De vakjes met lage scores zijn dan rood, terwijl de hoge triples duidelijk geel/wit zijn.

Uit subfiguren 8b-13b is een verband te zien tussen de waarde van σ en het optimale mikpunt. Er is te zien dat het optimale mikpunt dicht bij de bullseye ligt voor hoge waarden van σ (bijvoorbeeld voor darter 1), terwijl dit optimale mikpunt steeds dichterbij de triple 19 gaat naarmate de waarde van σ afneemt. We kunnen ook zien dat er een bepaalde waarde voor σ moet zijn waarbij het optimale mikpunt verspringt van de triple 19 naar de triple 20. Dit leiden we af uit het feit dat darter s_1 het best kan mikken op de triple 19 bij $\sigma = 24.21$, terwijl darter s_2 het best kan mikken op de triple 20 bij $\sigma = 5.18$.

We willen natuurlijk weten voor welke waarde van σ het optimale mikpunt verspringt tussen de triple 20 en de triple 19. Daartoe is de code duizend keer uitgevoerd: telkens voor een andere waarde van σ . Deze waarde varieerde tussen 0 en 100 en nam elke berekening met 0.1 toe. Het resultaat is te zien in figuur 6. Voor $\sigma < 16.4$ is te zien dat het optimale mikpunt in of dicht bij de triple 20 ligt. Zodra $\sigma = 16.4$ verspringt het optimale mikpunt naar de triple 19 en verplaatst daarna met een omtrekkende beweging richting de bullseye.

Als laatste onderwerp omtrent het simpele model, is ook onderzocht wat verwachte score is van het optimale mikpunt, de zogenaamde optimale score. Deze score hangt, net als het optimale mikpunt, af van de

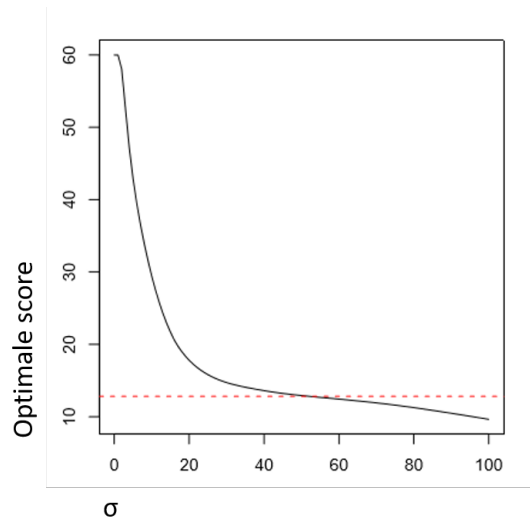


Figuur 6: Het verloop van het optimale mikpunt voor waarden van σ tussen 0 en 100 (bron: [1]).

covariantie matrix (in het simpele model dus eigenlijk van σ). We kunnen de optimale score dan voorstellen als

$$f(\sigma) = \max_{\mu} \mathbb{E}_{\mu}[s(Z) \mid \sigma].$$

Door de code wederom vaak uit te voeren over verschillende waarden van σ en telkens de optimale score te noteren, is de grafiek in figuur 7 geconstrueerd. De rode stippellijn geeft aan wat de optimale score is als de worp van een darter uniform verdeeld is: elke positie heeft evenveel kans en dus kan zo'n worp beschouwd worden als volledig willekeurig. Met andere woorden: de darter gooit blind zijn pijlen naar het bord.



Figuur 7: Optimale score voor verschillende waarden van σ (bron: [3]).

Het valt al meteen op dat ongeoeffende darters (met $\sigma \geq 60$) het niet snel beter zullen doen dan een darter die blind zijn pijlen gooit. Daarnaast is het opvallend dat de grafiek na $\sigma = 20$ langzaam daalt,

terwijl de grafiek snel daalt tussen $\sigma = 0$ en $\sigma = 20$. Op het domein $[0, 20]$ gedraagt de grafiek zich als $2^{-\sigma}$. De schrijvers van [1] weiden hier meer over uit in [3]: ze onderzoeken of deze verandering van gedrag rond $\sigma = 20$ door de indeling van het bord komt.

Een indirect gevolg van figuur 7 is dat het voor ongeoefende darters niet meteen loont om veel te gaan oefenen: hun precisie wordt wel beter, maar zolang $\sigma \geq 20$ zal de optimale score niet snel verbeteren. Voor geoefende darters zal trainen wel veel uitmaken: als zijn hun precisie verhogen zal dat resulteren in het snel oplopen van de optimale score.

6.2 Het geavanceerde model

In de figuren 14-20 volgen een paar resultaten die behaald zijn door toepassing van het geavanceerde model. De datasets die zijn gebruikt zijn dezelfde als de sets die zijn gebruikt voor het simpele model.

Uit de covariantie matrix is de variantie in x - en y -richting af te lezen. Deze varianties staan op de diagonaal: voor darter i is de variantie in x -richting het getal $\Sigma_{11}^{(i)}$ en de variantie in de y -richting het getal $\Sigma_{22}^{(i)}$. We noteren deze twee waarden met $\sigma_x^{(i)}$ respectievelijk $\sigma_y^{(i)}$. Het is goed om op te merken dat het idee achter Σ bevestigd lijkt te worden: de variantie in x -richting is bij iedere darter kleiner dan de variantie in de y -richting.

Als we de heatmaps bekijken, valt vooral op dat voor grote waarden van Σ de niveaulagen van de verwachte scores grotendeels cirkelvormig zijn. Naarmate deze waarde lager wordt, worden de niveaulagen grillig: geen mooie cirkels of ovaal meer, maar steeds kleinere gebieden met hoge verwachte scores. Bij accurate(re) darters worden "tentakels" gevormd: smalle, langgerekte stukken met hoge verwachtingswaarden (vooral mooi te zien in figuur 19a). Deze tentakels strekken zich uit van boven naar beneden, wat weer in overeenstemming is met de gedachte dat een worp preciezer is in x -richting dan in y -richting (anders hadden de tentakels van links naar rechts gelopen). Bij zeer lage waarden van Σ is het dartbord weer te onderscheiden in de heatmap, net zoals het geval was bij het simpele model, zie figuur 20.

In de figuren 14b-20b staan de optimale mikpunten gegeven, berekend aan de hand van het geavanceerde model. Ook is het tweedimensionale interval getekend waarin 95% van de geworpen pijlen zal vallen. Deze ellips is berekend aan de hand van de covariantie matrix en het optimale mikpunt van de worpen. Uit de figuren zien we ook weer eenzelfde verband tussen de waarde van Σ en het optimale mikpunt zoals bij het simpele model. Hoe kleiner de waarden van Σ worden, hoe dichter het optimale mikpunt van de bullseye naar de triple 19 gaat. Ook lijkt het erop dat er covariantie matrices zijn waarbij een kleine verandering resulteert in het verspringen van het optimale mikpunt (van de triple 20 naar de triple 19).

Verder merken we nog iets merkwaardigs op: dart(st)ers 1-3 hebben allen een positieve covariantie tussen de x - en y -coördinaat, terwijl darter 7 een negatieve covariantie heeft. Dit zien we mooi terug in het 95%-interval: deze ellips is bij darter 7 naar links georiënteerd, terwijl de andere dart(st)ers een naar rechts georiënteerd 95%-interval hebben. De invloed van de covariantie matrix is ook goed te zien bij darter 4. Zijn covariantie is nagenoeg nul. Dit resulteert in een 95%-interval die nagenoeg recht staat, waarbij de assen van de ellips evenwijdig zijn aan de x - en de y -as van het dartbord (zie figuur 17b). Daarnaast zien we ook dat het 95%-interval kleiner wordt en zich zal concentreren rond het optimale mikpunt naarmate de waarden in de covariantie matrix kleiner worden, zie figuur ??.

Als laatste resultaat voordat we de uitkomst van de hypothesetoetsen bespreken, willen we de resultaten van darter s_1 uitlichten. Vergelijken we de resultaten van het simpele model (figuur 12b) en het geavanceerde model (figuur 19b, dan valt ons iets speciaals op. Volgens het simpele model kan deze darter het beste richten op de triple 19, bij $\sigma^{(s_1)} = 24, 21$, terwijl het geavanceerde model hem adviseert om op de triple 20 te mikken bij relatieve standaardafwijkingen $\sigma_x^{(s_1)} = 5,92$ en $\sigma_y^{(s_1)} = 33,68$. Het verschil in deze adviezen komt dus doordat darter s_1 (erg) precies is in de x -richting, maar minder precies in de y -richting. Als hij op de triple 20 mikt, zal het volgens het simpele model wel eens voorkomen dat zijn pijl in de vijf of de één beland.

Dit verlaagt de verwachte score van een pijl gemikt op de triple 20. Volgens het geavanceerde model zal dit echter niet zo vaak het geval zijn: gezien de relatieve standaardafwijkingen van de darter zal hij eerder in de single 20 terecht komen dan in de vijf of de één. Dit bevordert de verwachte score van een pijl gericht op de triple 20. In figuur ?? is dit mooi te zien.

6.3 Hypothesetoetsen

Zoals reeds aangegeven is, zijn voor de hypothesetoetsen twee datasets gebruikt: een dataset $\mathbf{x} = (x_1, \dots, x_{102})$ waarbij gericht werd op de triple 20 en een dataset $\mathbf{y} = (y_1, \dots, y_{102})$ waarbij gericht werd op het optimale mikpunt. De twee gesimuleerde darters uit de vorige sectie zijn niet meegenomen in de toets. Deze waren namelijk alleen bedoeld om het gedrag van het model te onderzoeken, niet om te bekijken of er significant hoger wordt gegooid.

We hebben ervoor gekozen om de darters te laten richten op het mikpunt dat berekend is volgens het geavanceerde model. Dit lijkt een betere weerspiegeling te zijn van de realiteit: een worp wijkt meer af in de y -richting dan in de x -richting. Overigens maakt het voor de onervaren en zeer ervaren darters niet heel veel uit welk model we als basis nemen: de optimale mikpunten liggen relatief dicht bij elkaar. Daarentegen zou het voor een darter die een beetje ervaring heeft, met een σ van rond de 24 volgens het simpele model, wel interessant zijn om steekproeven te nemen voor beide modellen. De modellen gaven namelijk zeer verschillende optimale mikpunten, zie de resultaten van darter s_1 . Er zaten echter geen ervaren darters in de steekproef.

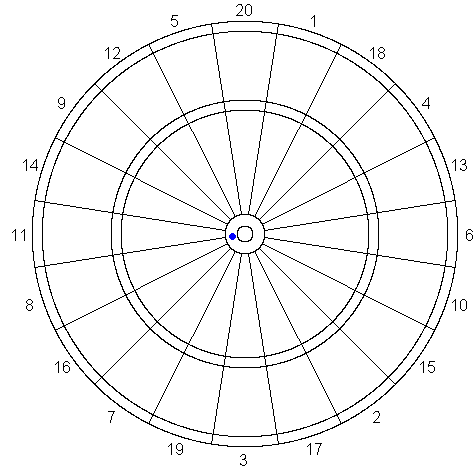
In tabel 2 staan de teststatistieken voor elke darter weergegeven. In de tabel is af te lezen dat alleen de p -waarden van dart(st)ers 2, 7 en 14 lager zijn dan het significantieniveau van 0.05. Daarom kunnen we voor deze darters de nulhypothese verwerpen en concluderen dat $\mu_x < \mu_y$. Voor dart(st)ers 2,7 en 14 geldt dus dat zij significant hoger scoren door op hun optimale mikpunt te mikken in plaats van op de triple 20. De resultaten van de andere darters geven echter geen significant hogere score aan: zij zouden alsnog kunnen overwegen om op de triple 20 te mikken.

Dart(st)er	Dataset X		Dataset Y		Z-waarde	p-waarde
	\bar{x}	S_x	\bar{y}	S_y		
1	13.22	11.82	11.29	10.30	1.23	0.8921
2	10.38	8.76	15.21	13.50	-3.03	0.0012
3	10.91	10.62	12.63	11.68	-1.10	0.1358
4	11.90	10.78	14.11	13.01	-1.32	0.0937
5	12.18	12.64	12.87	10.69	-0.42	0.3356
6	11.10	12.72	11.75	10.32	-0.40	0.3449
7	11.28	7.70	13.67	12.03	-1.68	0.0461
8	9.59	10.99	12.06	11.13	-1.60	0.0553
9	8.48	10.67	9.59	9.72	-0.77	0.2192
10	9.48	8.05	9.18	10.54	0.23	0.5915
11	14.31	14.35	16.53	13.02	-1.15	0.1241
12	12.37	12.82	11.75	8.63	0.40	0.6567
13	11.78	11.62	10.94	9.45	0.57	0.7152
14	12.22	12.30	15.38	11.54	-1.90	0.0290

Tabel 2: Teststatistieken van de hypothesetoets bij $\alpha = 0,05$. De gekleurde cellen zijn de p -waarden lager dan α .



(a) De heatmap.

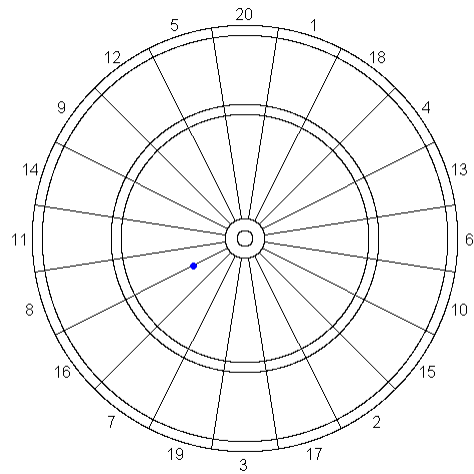


(b) Het optimale mikpunt.

Figuur 8: Resultaten van dartster 1 met $\sigma^{(1)} = 98,66$.



(a) De heatmap.

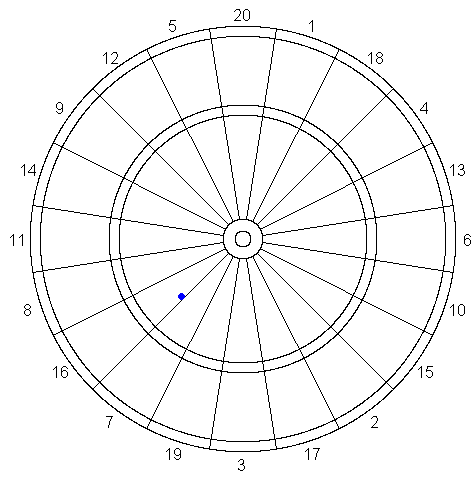


(b) Het optimale mikpunt.

Figuur 9: Resultaten van dartster 2 met $\sigma^{(2)} = 57,39$.



(a) De heatmap.

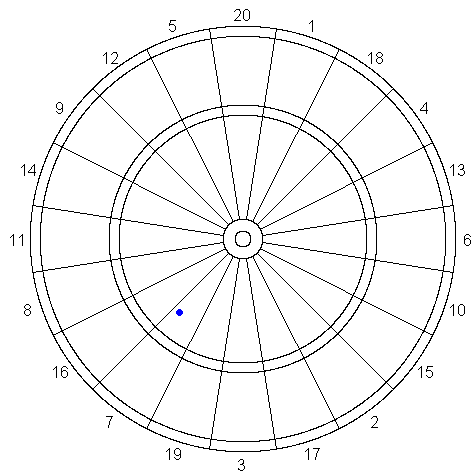


(b) Het optimale mikpunt.

Figuur 10: Resultaten van darter 3 met $\sigma^{(3)} = 48, 44$.

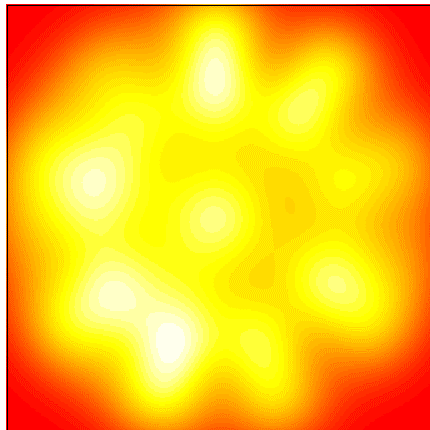


(a) De heatmap.

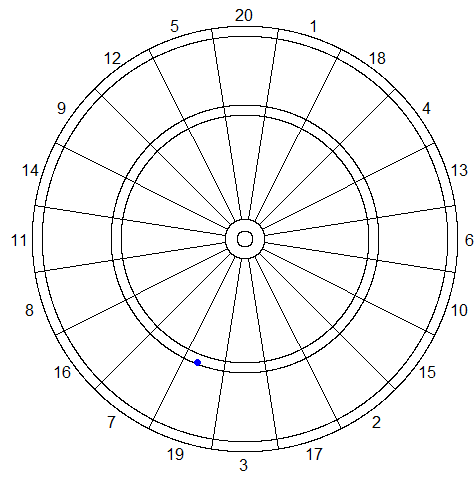


(b) Het optimale mikpunt.

Figuur 11: Resultaten van darter 4 met $\sigma^{(4)} = 43, 99$.

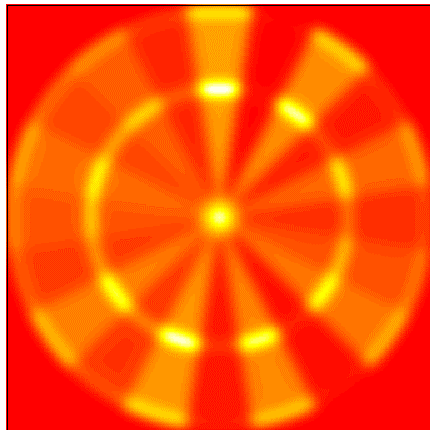


(a) De heatmap.

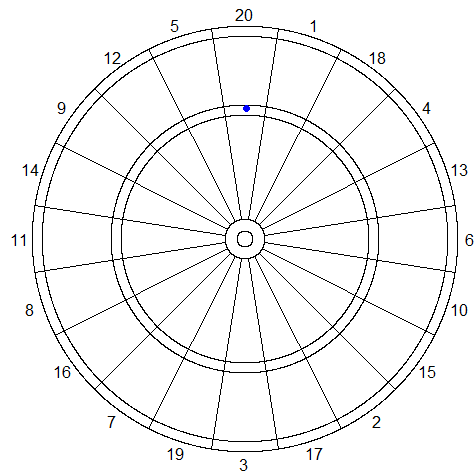


(b) Het optimale mikpunt.

Figuur 12: Resultaten van darter s_1 met $\sigma^{(s_1)} = 24, 21$.



(a) De heatmap.

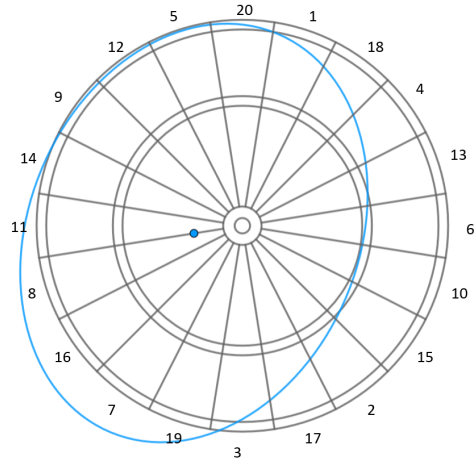


(b) Het optimale mikpunt.

Figuur 13: Resultaten van darter s_2 met $\sigma^{(s_2)} = 5, 15$.



(a) De heatmap.

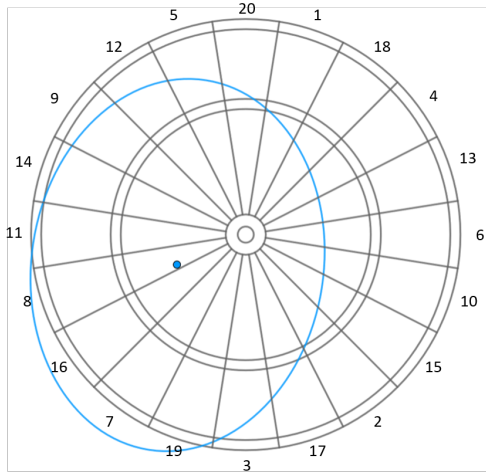


(b) Het optimale mikpunt.

Figuur 14: Resultaten van dartster 1 met $\Sigma^{(1)} = \begin{pmatrix} 3437 & 785 \\ 785 & 4982 \end{pmatrix}$.



(a) De heatmap.

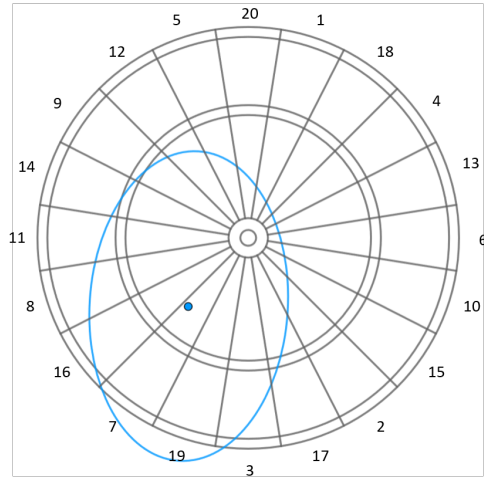


(b) Het optimale mikpunt.

Figuur 15: Resultaten van dartster 2 met $\Sigma^{(2)} = \begin{pmatrix} 2265 & 211 \\ 211 & 3597 \end{pmatrix}$.



(a) De heatmap.

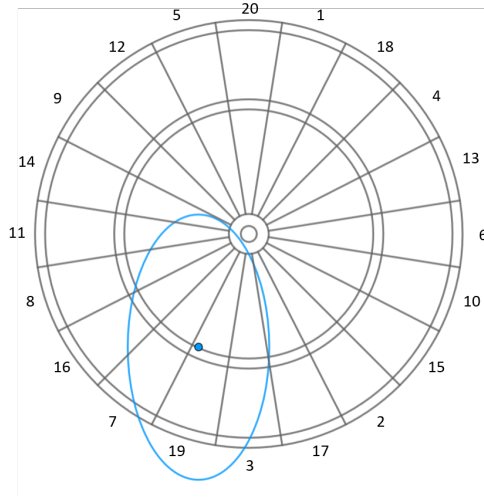


(b) Het optimale mikpunt.

Figuur 16: Resultaten van darter 3 met $\Sigma^{(3)} = \begin{pmatrix} 1070 & 84 \\ 84 & 2595 \end{pmatrix}$.



(a) De heatmap.

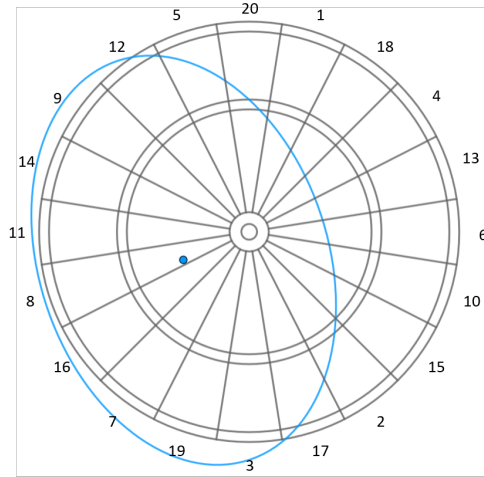


(b) Het optimale mikpunt.

Figuur 17: Resultaten van darter 4 met $\Sigma^{(4)} = \begin{pmatrix} 529 & -0.98 \\ -0.98 & 1855 \end{pmatrix}$.

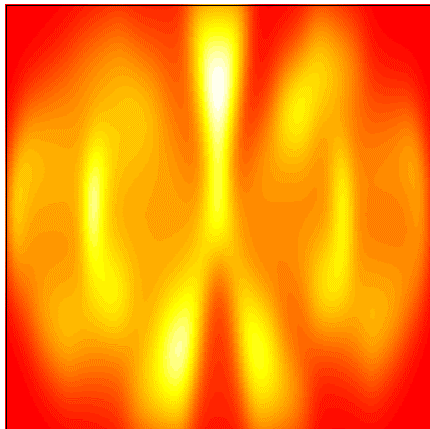


(a) De heatmap.

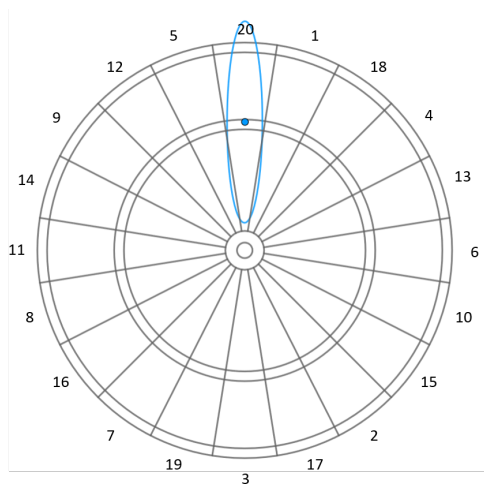


(b) Het optimale mikpunt.

Figuur 18: Resultaten van darter 7 met $\Sigma^{(4)} = \begin{pmatrix} 2319 & -691 \\ -691 & 4180 \end{pmatrix}$.

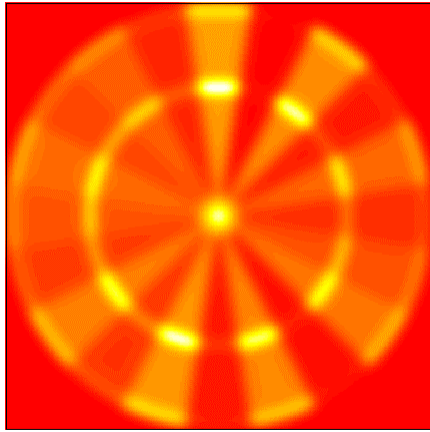


(a) De heatmap.

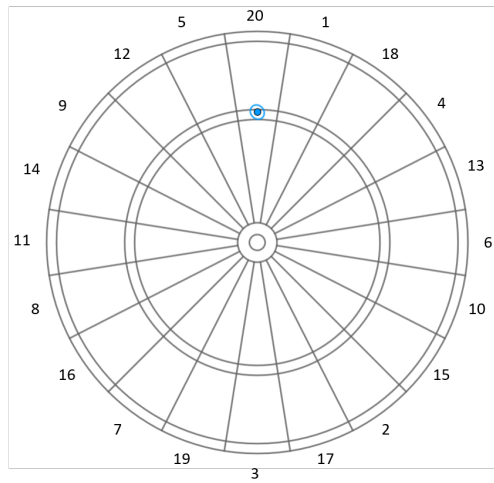


(b) Het optimale mikpunt.

Figuur 19: Resultaten van darter s_1 met $\Sigma^{(s_1)} = \begin{pmatrix} 35 & 2.25 \\ 2.25 & 1134 \end{pmatrix}$.



(a) De heatmap.



(b) Het optimale mikpunt.

Figuur 20: Resultaten van darter s_2 met $\Sigma^{(s_2)} = \begin{pmatrix} 5.193 & -0.366 \\ -0.366 & 5.387 \end{pmatrix}$.

7 Discussie

In deze scriptie is onderzocht of een darter zijn hoogste verwachte score behaalt als hij op de triple 20 of op zijn optimale mikpunt mikt. Daarvoor is een model ontwikkeld waarmee de precisie (variantie) van een darter berekend kon worden. Hiervoor werden voldoende pijlen gegooid, allemaal gericht op de (dubbele) bullseye, en werden vanwege praktische overwegingen de scores van de pijlen genoteerd. Hierdoor worden sommige worpen echter tekort (of juist ten goede) gedaan. Dit illustreren we met een voorbeeld.

Stel dat een darter op de double bullseye mikt en tien punten scoort met een pijl. Deze pijl kan dan in de single 10 zijn gekomen of in de double 5. Wat betreft precisie kan dit veel uitmaken: de double 5 is relatief ver weg van de double bullseye, terwijl een single 10 tegen de ring van de single bullseye kan zitten. Het laatstgenoemde geval is dan een veel preciezer worp dan het eerstgenoemde geval. Op deze manier komt een darter X (die telkens double 5 gooit) net zo goed uit het model als een darter Y (die telkens een single 10 gooit tegen de ring van de single bullseye aan). Dit geeft een scheef beeld: de darter Y zit veel dichterbij de double bullseye, gooit dus preciezer en zou een kleinere variantie moeten hebben dan darter X .

Het bovenstaande probleem kan wellicht opgelost worden door niet de score van een pijl te noteren, maar het vakje waarin de pijl terecht is gekomen. Er kan dan ook onderscheidt worden gemaakt op singles in de "inner" ring (tussen de bullseyes en de triples) en de "outer" ring (tussen de triples en de doubles). Op deze manier kun je darter X en Y uit de vorige alinea wel uit elkaar halen: darter X gooit telkens double 5, darter Y telkens inner 10.

Verder kan er natuurlijk een betere conclusie getrokken worden als er een grotere steekproef genomen wordt dan de veertien personen die meegewerkt hebben aan dit onderzoek. Op die manier worden de uitkomsten van het onderzoek betrouwbaarder.

8 Conclusie

In deze scriptie hebben we de volgende vraag proberen te beantwoorden: kan een (onervaren) darter het best mikken op de triple 20 of is er een ander mikpunt die een hogere verwachte score oplevert?

Om deze vraag te beantwoorden, hebben we de verwachtingswaarde van de score herschreven met behulp van Fast Fourier Transformaties en een convolutie. Daarna hebben we het EM-algoritme alleen dan wel in combinatie met Importance Sampling gebruikt om de covariantie matrix van een darter te schatten. Hiermee kon het optimale mikpunt (het punt met de hoogste verwachte score) berekend worden en daarna is met twee steekproeven gekeken of darters significant hogere scores gooien als ze op hun mikpunt mikken.

In tabel 2 is af te lezen dat in (slechts) drie van de veertien gevallen een dart(st)er significant hoger scoorde door op zijn/haar optimale mikpunt te gooien. Het antwoord op de onderzoeksvraag is daarom het volgende: waarschijnlijk niet. De resultaten van het onderzoek (drie van de veertien darters gooiden significant beter) geven het idee dat mikken op het mikpunt geen verbetering is ten opzichte van mikken op de triple 20.

Om hierover echter duidelijk uitsluitsel te krijgen, is een grotere steekproef nodig dan de veertien darters die mee hebben gewerkt aan dit onderzoek.

9 Dankbetuigingen

In de eerste plaats wil ik Ryan Tibshirani bedanken voor het toegankelijk stellen van de code die gebruikt is om het optimale mikpunt te berekenen voor beide modellen. Daarnaast wil ik iedereen bedanken die vrijwillig 306 pijlen op een dartboard heeft gegooid.

Dartster 1 - Inge Brouwer

Dartster 2 - Marilène de Rooij

Darter 3 - Willem van Veluw

Darter 4 - René de Rooij

Darter 5 - Mark Jan van Lieburg

Darter 6 - Bertram Flier

Darter 7 - Kees van Veluw

Darter 8 - Jeroen van Veluw

Darter 9 - Thijs van Veluw

Dartster 10 - Jasmijn de Boer

Darter 11 - Danny van Kooi

Darter 12 - Justus van Kooi

Dartster 13 - Monique van Kooi - Mellegers

Darter 14 - Jons Bolding

Als laatst wil ik mijn scriptiebegeleider, Sjoerd Dirksen, bedanken voor de prettige begeleiding en de handige tips tijdens het schrijven van de scriptie en het uitwerken van de resultaten.

Referenties

- [1] Tibshirani, R. J., Price, A., Taylor, J. (2011). A statistician plays darts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1), 213-226.
- [2] Tibshirani, R. J. (2011, 20 januari). Statistical Tools to Analyze Your Darts Game. [Webpagina]. Geraadpleegd van <https://cran.r-project.org/web/packages/darts/>
- [3] Tibshirani, R. J., Price, A., Taylor, J. (2011). Supplement to: A statistician plays darts rearranging the dartboard.
- [4] Duchi, J. (2007). Properties of the trace and matrix derivatives. Available electronically at https://web.stanford.edu/~jduchi/projects/matrix_prop.pdf.
- [5] Tay, K. J. Y. (2018). Derivative of $\log(\det X)$. [Webpagina]. Geraadpleegd van <https://statisticaloddsandends.wordpress.com/2018/05/24/derivative-of-log-det-x/>
- [6] Rencher, A. C. (2002) 'The Multivariate Normal Distribution' uit *Methods of Multivariate Analysis*. Hoboken, NJ: John Wiley Sons, Inc.
- [7] Anderson, C. J. (2017). Multivariate Normal Distribution [PowerPoint presentatie]. Geraadpleegd van <https://education.illinois.edu/docs/default-source/carolyn-anderson/edpsy584/lectures/MultivariateNormal-beamer-online.pdf>
- [8] Hogg, R. V., McKean, J., Craig, A. T. (2005). *Introduction to mathematical statistics*. Pearson Education.
- [9] Ross ?. Sheldon, M.(2002). *Simulation*. Academic Press, 3rd edition.