Modelselectie: AIC en BIC

Project Toegepaste Wiskunde 2

Als onderdeel van de

Bacheloropleiding Wiskunde
Radboud Universiteit Nijmegen

Marianne Coenen Tom Huls

10 juli 2008

Voorwoord

Een vraag waar statistici in de praktijk vaak mee te maken krijgen, is welk model (in ons geval een kansverdeling) het meest waarschijnlijk is voor een bepaald verschijnsel, gegeven een eindige hoeveelheid gegevens. Dit probleem staat bekend als het probleem van de modelselectie. Een keuze maken tussen meerdere mogelijke modellen kan gebeuren aan de hand van een zogeheten criterium. Bekende criteria die hiervoor gebruikt worden, zijn het Akaike InformatieCriterium (AIC) en het Bayesiaanse InformatieCriterium (BIC). Men zou zich kunnen afvragen hoe deze criteria werken en welke van de twee het beste gebruikt kan worden.

In ons project hebben wij theoretisch onderzoek gedaan naar aanleiding van vragen van het Algemeen Burgerlijk Pensioenfonds (ABP), tegenwoordig Algemene Pensioen Groep (APG) geheten, over het probleem van modelselectie. Onze contactpersoon bij dat bedrijf, dhr. Henk Angerman, vertelde ons dat het APG voor problemen omtrent modelselectie veelal gebruikmaakt van het BIC, zonder precies te weten waarop het gebaseerd is of in welke zin dit criterium optimaal is, als het dat al is. Hij vindt het persoonlijk erg onbevredigend een criterium toe te passen zonder kennis te hebben van deze zaken. Onze opdracht was dan ook het beantwoorden van de vragen die hij had, teneinde hem (en het APG) een beter inzicht te geven in het probleem van modelselectie.

De vragen die ons gesteld werden, zijn de volgende:

1. Hoe zijn de BIC en de AIC afgeleid? Welke principes liggen eraan ten grondslag?
2. Is het mogelijk om, als je extra structuur oplegt (bijvoorbeeld dat het gezochte model met zekerheid een lineair regressiemodel is), betere criteria af te leiden?
3. Is er een algemeen principe dat je kunt gebruiken om in een specifiek geval een criterium af te leiden? Op die manier zou je een stricte scheiding kunnen aanbrengen tussen het postuleren van dit algemene principe, wat een zekere mate van subjectiviteit bevat, en het vervolgens analytisch en objectief afleiden van het ermee corresponderende criterium.
4. De bovenstaande vragen zijn puur theoretisch. Het zou wellicht ook interessant zijn om verschillende selectiecriteria te testen met behulp van computersimulaties waarbij de onderliggende modellen bekend zijn.

Wij hebben contact gezocht met dhr. Angerman en met hem gesproken over bovenstaande vragen. Na overleg is besloten dat ons onderzoek zich in het bijzonder zal richten op de volgende vragen:

1. Hoe zijn het AIC en het BIC afgeleid? Welke principes liggen eraan ten grondslag?
2. In het Akaike InformatieCriterium wordt de zogenoemde Kullback-Leiblerdiscrepantie gebruikt. Wat is een discrepantie, waarom wordt juist deze gebruikt in het AIC en in welke situaties zijn andere discrepanties wellicht beter?
3. Welk criterium is beter als je weet dat het onderliggende model lineair moet zijn? Kun je een criterium bedenken dat in dat geval beter is dan zowel het AIC als het BIC?
4. Bekijk de werking van de criteria met behulp van simulaties en onderbouw hiermee je conclusies.

Tijdens ons onderzoek ervaarden we dat het uitwerken van bovenstaande onderzoekspunten lang niet eenvoudig was. Zo is veel literatuur beschikbaar over het Akaike InformatieCriterium, maar viel in vrijwel geen boek of artikel een heldere afleiding te vinden. Over het Bayesiaanse InformatieCriterium konden we veel minder informatie vinden. Een heldere afleiding bleek ook daar te veel gevraagd. Ten slotte moesten we ook op het gebied van de Kullback-Leiblerdiscrepantie concluderen dat het vinden van de aannamen waarop die gebaseerd is meer tijd en moeite zou gaan kosten dan we vooraf hadden gedacht.

Desondanks zijn we tijdens onze zoektocht veel interessante dingen tegengekomen. Dit eindverslag beoogt een zo goed mogelijk beeld te geven van die resultaten.

Marianne Coenen
Tom Huls

Inhoudsopgave

Hoofdstuk 1

Introductie

Om goed te begrijpen wat het Akaike InformatieCriterium en het Bayesiaanse InformatieCriterium zijn en de afleidingen te kunnen volgen, is het nodig eerst enkele termen en definities te begrijpen. Deze zullen in dit hoofdstuk worden behandeld. De informatie is deels ontleend aan [9].

1.1. Het operating model en de operating family

Met het **operating model** bedoelen we het werkelijke model, in ons geval dus een kansverdeling, waarop alle data gebaseerd zijn. In het ideale geval is dit het model aan de hand waarvan voorspellingen worden gedaan en waarop conclusies worden gebaseerd. Kennis over het operating model wordt verkregen door wat we weten van het onderwerp. Zo is soms bekend dat de stochasten alleen niet-negatieve waarden kunnen aannemen (bijvoorbeeld in het geval dat de waarnemingen een hoeveelheid water voorstellen die in een bepaalde tijd langs een zeker punt komt), dat bepaalde gebeurtenissen even waarschijnlijk zijn, dat zekere gebeurtenissen onafhankelijk zijn, enzovoorts.

Slechts in uitzonderlijke gevallen is er genoeg informatie bekend om het operating model helemaal vast te leggen. Veelal is het echter alleen mogelijk een familie van mogelijke modellen aan te geven, waar het operating model toe behoort. Deze familie wordt de **operating family** genoemd. Hoe meer er bekend is over het operating model, hoe kleiner de familie zal zijn.

1.2. Approximating family

Per conventie wordt de omvang van een familie bepaald aan de hand van het aantal onafhankelijke parameters, dat wil zeggen, het aantal gegevens dat bekend moet zijn om een enkel element van de familie te kunnen selecteren. Deze parameters moeten worden geschat door waarnemingen te doen. De nauwkeurigheid waarmee dat kan, hangt af van de hoeveelheid gegevens die beschikbaar is in verhouding tot het aantal parameters dat geschat moet worden. Hoe meer gegevens bekend zijn, of hoe minder parameters geschat moeten worden, hoe nauwkeuriger we de parameters kunnen schatten.

Het zou mooi zijn als we altijd een operating family konden vinden met niet al te veel parameters. Dit zou idealiter kunnen met enkele gefundeerde aannamen, die de omvang van de familie

verkleinen. In dat geval is het proces van modelselectie niet al te ingewikkeld. In de praktijk is dat echter niet vaak het geval en zijn er meer parameters dan je op grond van de beschikbare gegevens goed kunt schatten. In dat geval heb je te maken met **overfitting**: de modellen die je zo vindt zijn sterk afhankelijk van de specifieke steekproef, en verschillende steekproeven leiden tot sterk verschillende modellen. In paragraaf 4.2 komen we hier nog op terug.

Op basis van het bovenstaande zou men kunnen vermoeden dat de kunst van modelselectie eenvoudigweg het vinden van voldoende gegevens is. In de praktijk ligt de steekproefgrootte echter veelal vast door praktische problemen bij het verzamelen van de gegevens en de kosten die daarmee gepaard gaan. Een oplossing voor dit probleem wordt gevonden door te werken met een **approximating family** van modellen. De modellen in die familie zijn veelal simpeler dan die uit de operating family. Merk op dat nu het operating model niet in de approximating family hoeft te zitten en dat het kan voorkomen dat je bij de modelselectie nooit op het operating model uit kunt komen.

Ter verduidelijking: er zal nu bij modelselectie een keuze gemaakt worden voor een model uit de approximating family. Deze approximating family wordt gekozen door de onderzoeker en hoeft daarom niet het operating model te bevatten. Veelal zal dat zelfs niet het geval zijn, omdat je laatstgenoemde niet kent. Het is aan de onderzoeker zijn approximating family zodanig te kiezen dat verwacht mag worden dat een geschikt model erin zit.

Soms zijn er meerdere approximating families waaruit een keuze kan worden gemaakt. Ter vereenvoudiging zullen we hier aannemen dat er slechts één familie is.

1.3. Kiezen van een approximating model middels discrepanties

Een vraag die direct voortkomt uit bovenstaande alinea, is hoe je de approximating family kiest. We zullen ons daar nu echter niet op richten. We nemen aan dat al een keuze is gemaakt voor de approximating family en dat het probleem dat resteert, is om een keuze te maken voor een model uit de approximating family.

Bij een keuze gebaseerd op een discrepantie is het idee dat je het model kiest waarvan je schat dat die het meest van toepassing is in de gegeven omstandigheden, te weten de steekproefgrootte, de aannamen die de onderzoeker doet en zijn specifieke eisen. Je geeft als eerste aan in welke zin je wil dat het model aan het operating model voldoet, dat wil zeggen, je kiest een discrepantie die aangeeft hoe groot de afwijking van het model tot het operating model is. Daarna ga je kijken welk approximating model de discrepantie minimaliseert. Die kies je vervolgens.

Een groot voordeel van deze manier van kiezen is dat hij erg flexibel is: je kunt zelf aangeven hoe belangrijk je het vindt dat een zeker aspect van het operating model goed benaderd wordt. Je kiest dan namelijk een discrepantie die extra gewicht toekent aan dat aspect.

Het grote probleem dat nu nog resteert, is het schatten van de verwachte discrepantie. Dat is zelden een eenvoudige zaak. Dit onderdeel van de modelselectie op basis van discrepanties is, evenals het kiezen van een goede discrepantie, een zwakke schakel in het hele selectieproces.

1.4. Discrepanties

Er zijn verschillende soorten discrepanties. Ze hebben gemeen dat ze altijd groter dan of gelijk aan nul zijn en precies gelijk aan nul als het approximating model gelijk is aan het operating

model. De meest gebruikte discrepantie is de **Kullback-Leiblerdiscrepantie**, die gedefinieerd is als

$$I(f,g) = \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) \mathrm{d}\mathbf{x}.$$

Deze formule wordt door Kullback en Leibler besproken in hun artikel **On information and sufficiency** [7]. Zij leggen een verband met de Shannonentropie, die een maat is voor de onzekerheid die bij een kansverdeling hoort. Het is ons niet geheel duidelijk geworden op basis waarvan Kullback en Leibler tot deze definitie gekomen zijn.

Een andere discrepantie die wel gebruikt wordt, is de L_2-norm. Bij regressieanalyse (waar we het verderop over zullen hebben) is deze gedefinieerd als $L_2 = \frac{1}{n}\|\mathbf{X}_*\beta_* - \mathbf{X}\hat{\beta}\|^2$.

Er bestaan nog vele andere bekende discrepanties, zoals de Kolmogorovdiscrepantie, de Pearson chi-kwadraatdiscrepantie en de Neyman chi-kwadraatdiscrepantie [10]. Dit zijn simpelweg andere uitdrukkingen in f en g en vormen, zoals iedere discrepantie, een hulpmiddel om de mate van gelijkheid van f en g te kunnen benoemen.

1.5. Criteria

Met een criterium kun je beslissen welk model het meest van toepassing is en daarom het beste gekozen kan worden. Een criterium is een functie waaraan je de gegevens van je steekproef en een model uit de approximating family meegeeft. Zo krijg je een waarde terug voor dat model. Bij het vergelijken van twee modellen gebruik je dezelfde steekproef (je wilt immers op basis van die steekproef bepalen met welk onderliggend model je te maken hebt). Omdat het model verschilt, geeft je criterium bij verschillende modellen verschillende uitkomsten.

Om een keuze voor een model te maken, ga je deze uitkomsten vergelijken. Een logische keuze is dat het criterium het meegegeven model met het echte, onderliggende model vergelijkt. Het bepaalt dan de afstand (op basis van een discrepantie) van het approximating model tot het echte model. Hoe kleiner de afstand hoe beter, dus we zoeken dát model dat de kleinste waarde van het criterium heeft.

Twee bekende voorbeelden van criteria zijn het AIC (Akaike InformatieCriterium) en het BIC (Bayesiaanse InformatieCriterium). In ons onderzoek hebben we voornamelijk gekeken naar deze modellen.

1.6. Likelihoodfuncties en maximum-likelihoodschatters

Bij modelselectie wil je, gegeven de steekproef, het model kiezen dat de onderliggende verdeling van de data weergeeft. Maar je hebt maar een deel van de gegevens, namelijk de steekproef, en kunt dus niet met zekerheid zeggen wat het onderliggende model is. Daarom zoekt je een model dat gegeven je steekproef het meest waarschijnlijk is. Dit is de achterliggende gedachte van de **likelihoodtheorie** [11].

Wanneer je uit een populatie met dichtheidsfunctie f een steekproef van lengte n trekt en deze trekking noteert als $\mathbf{x} = (x_1, \ldots, x_n)$, dan wordt de **likelihoodfunctie** gedefinieerd door

$$\mathcal{L}(x_1, \ldots, x_n) := f(x_1) \cdots f(x_n),$$

waar $\mathcal{L} : \mathbb{R}^n \to [0, +\infty)$ en $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Als deze functie in een bepaald gebied $A \subset \mathbb{R}^n$ alleen maar kleine waarden aanneemt, dan is het niet waarschijnlijk dat de uitkomst van de steekproef in A zal liggen. Dit komt doordat we kunnen schrijven

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int \cdots \int_A \mathcal{L}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n =: \int_A \mathcal{L}(x_1, \ldots, x_n)\, \mathrm{d}\mathbf{x}.$$

De laatste stap is een definitie, om meervoudige integralen eenvoudiger te kunnen noteren.

De steekproeven $(x_1, \ldots, x_n)$ waarin $\mathcal{L}$ een kleine waarde aanneemt noemen we onwaarschijnlijk (unlikely). En daarmee noemen we zo'n steekproef waar $\mathcal{L}$ een grote waarde aanneemt waarschijnlijk (likely). De functie $\mathcal{L}$ geeft dus aan hoe waarschijnlijk de uitkomst van een steekproef is, gegeven een bepaalde verdeling.

Bij modelselectie worden vaak verschillende verdelingen beschouwd. Stel, je hebt een populatie met een dichtheidsfunctie die van een parameter θ afhangt: $f(x, \theta)$. Dan hangt ook de likelihoodfunctie van θ af. Daarom noteren we haar als $\mathcal{L}_\theta$. Als we een uitkomst van een steekproef hebben, dan willen we bepalen uit welke verdeling deze data komen, met andere woorden, welke θ voor een steekproef met deze uitkomst zorgt. Deze θ vinden we door de functie $\theta \mapsto \mathcal{L}_\theta(x_1, \ldots, x_n)$ te maximaliseren. We kiezen dus die θ waarvoor de uitkomst van deze steekproef het meest waarschijnlijk is.

We gaan ervan uit dat er maar één maximale θ is. Deze noemen we $\hat{\theta}$: de maximum likelihood estimator (ofwel **maximum-likelihoodschatter**). Omdat de likelihoodfunctie van de $x_1, \ldots, x_n$ afhangt, zal voor een andere steekproef doorgaans ook een andere $\hat{\theta}$ gelden. Desondanks noteren we $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$, want uit de context blijkt vaak om welke steekproef het gaat.

We kunnen dit alles in een klein voorbeeld samenvatten. Stel, we hebben een exponentieel verdeelde populatie met parameter θ en een steekproef $(x_1, \ldots, x_n)$ uit deze populatie. We bepalen nu de maximum-likelihoodschatter van θ.

De dichtheidsfunctie wordt voor $x \geq 0$ gegeven door

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} \mathrm{e}^{-x/\theta} & \text{als } x \geq 0, \\ 0 & \text{anders.} \end{cases}$$

De likelihoodfunctie wordt dan voor $\theta > 0$ gegeven door

$$\mathcal{L}_\theta(x_1, \ldots, x_n) = \frac{1}{\theta^n} \mathrm{e}^{-(x_1 + \ldots + x_n)/\theta}.$$

Differentiëren we deze functie naar θ, dan krijgen we

$$\frac{\partial}{\partial \theta} \mathcal{L}_\theta(x_1, \ldots, x_n) = \left(\frac{x_1 + \ldots + x_n}{\theta^{n+2}} - \frac{n}{\theta^{n+1}} \right) \mathrm{e}^{-(x_1 + \ldots + x_n)/\theta}.$$

Op 0 stellen geeft dat de eerste factor gelijk aan 0 moet zijn, want de e-macht is altijd groter dan 0. Simpel uitschrijven geeft dan

$$\theta = \frac{x_1 + \ldots + x_n}{n}.$$

Dit geldt voor alle steekproeven, dus de maximum-likelihoodschatter van θ is

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) = \frac{x_1 + \ldots + x_n}{n} = \overline{\mathbf{x}}.$$

Hoofdstuk 2

Afleiding Akaike InformatieCriterium

Het **Akaike InformatieCriterium**, ook wel Akaike Information Criterion (en daarvoor An Information Criterion) genoemd, is een maat om te bepalen hoe goed een statistisch model van toepassing is. De formule voor het AIC is eenvoudig, namelijk

$$\mathrm{AIC} = -2 \log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) + 2d.$$

Hier is $\hat{\boldsymbol{\theta}}$ de **maximum-likelihoodschatter**, berekend uit de steekproefgegevens $\mathbf{x}$ onder een approximating model g, $\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}}) = \prod_{i=1}^{n} g_{\boldsymbol{\theta}}(\mathbf{x})$ de **likelihoodfunctie** en d het aantal vrije parameters in het statistische model.

Het is interessant te weten hoe deze formule is afgeleid, aangezien dit de gebruiker een beter inzicht kan geven in waar hij mee bezig is. Met meer kennis van de achterliggende theorie kan de gebruiker beter kiezen tussen het AIC en de vele andere informatiecriteria die er beschikbaar zijn. In dit hoofdstuk zullen we de afleiding van het Akaike InformatieCriterium behandelen.

We volgen de grote lijnen van [5], maar hebben aanpassingen doorgevoerd in het detailniveau en de notaties. De gekozen literatuur blinkt uit in haar relatieve eenvoud. Een gevolg hiervan is dat de afleiding algemeen blijft, wat betekent dat we niet alle aannamen uitvoerig behandelen. Deze zijn overigens alle vrij technisch van aard; zo moet de parameterruimte Θ compact zijn, en $\boldsymbol{\theta_0}$ uniek zijn en in het inwendige van Θ liggen. Resultaten gelden voor steekproeven met steekproefomvang $n \to \infty$. Voor grondigere afleidingen (waarin ook alle aannamen uitvoerig worden behandeld), kan men onder andere [1, 4, 6, 13, 14, 15, 17] raadplegen.

2.1. Achtergrond

De afleiding zal het beste begrepen worden als de gebruikte notaties geheel duidelijk zijn. Deze zullen we in deze paragraaf behandelen. Daarnaast beogen we enkele theorie op te halen die in de afleiding is verwerkt.

Notatie 2.1. Vectoren zullen worden weergegeven met vet gedrukte, rechtopstaande letters, zoals $\mathbf{x}$ en $\mathbf{y}$. Uitzonderingen vormen vectoren die worden aangegeven door Griekse letters; zij zullen schuin gedrukt zijn. Daarnaast stellen vet gedrukte cijfers ook vectoren voor. Dit zijn dan veelal speciale vectoren, zoals $\mathbf{0} = (0,\ 0,\ \ldots,\ 0)$ en $\mathbf{1} = (1,\ 1,\ \ldots,\ 1)$. Een vector met n elementen wordt een n-vector genoemd.

Notatie 2.2. In de afleiding heeft $\mathbf{x}$ een dubbele betekenis, namelijk enerzijds die van vector met steekproefgegevens, en anderzijds die van integratievariabele (wanneer we integreren in een n-dimensionale ruimte, wat bij iedere integratie het geval zal zijn, tenzij expliciet anders vermeld). In het geval beide in dezelfde regel voorkomen, zullen we $\mathbf{y}$ gebruiken voor de steekproefgegevens. Het zou juist verwarrend werken als consequent werd geprobeerd een andere notatie voor een van de voorgaande zaken in te voeren, aangezien soms van de ene betekenis in de andere wordt overgegaan. Wel is het belangrijk op te merken dat de steekproef, of ze nu genoteerd wordt met $\mathbf{x}$ of met $\mathbf{y}$, voortgekomen is uit het operating model f, en niet uit een approximating model g (tenzij natuurlijk $f = g$). Om soortgelijke redenen zullen we de likelihoodfunctie soms aanduiden met $g(\mathbf{x} \mid \boldsymbol{\theta})$.

De algemene afleiding maakt gebruik van de tweede-orde Taylorontwikkeling. Deze theorie komt aan de orde in een eerstejaars college integraal- en differentiaalrekening en wordt later precies gemaakt in colleges analyse. In het kort komt ze neer op het volgende. Zij $h(\boldsymbol{\theta})$ een reëelwaardige functie op d dimensies. Dan wordt de tweede-orde Taylorontwikkeling in het punt $\boldsymbol{\theta_0}$ rond $\boldsymbol{\theta}$ gegeven door

$$\begin{aligned} h(\boldsymbol{\theta}) &= h(\boldsymbol{\theta_0}) + \left[\frac{\partial h(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}}\right]' [\boldsymbol{\theta} - \boldsymbol{\theta_0}] + \frac{1}{2}[\boldsymbol{\theta} - \boldsymbol{\theta_0}]' \left[\frac{\partial^2 h(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}^2}\right] [\boldsymbol{\theta} - \boldsymbol{\theta_0}] + \mathrm{Re} \\ &\approx h(\boldsymbol{\theta_0}) + \left[\frac{\partial h(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}}\right]' [\boldsymbol{\theta} - \boldsymbol{\theta_0}] + \frac{1}{2}[\boldsymbol{\theta} - \boldsymbol{\theta_0}]' \left[\frac{\partial^2 h(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}^2}\right] [\boldsymbol{\theta} - \boldsymbol{\theta_0}], \end{aligned} \tag{2.1}$$

waar $\boldsymbol{\theta}$ en $\boldsymbol{\theta_0}$ twee verschillende punten uit de d-dimensionale ruimte zijn waarover de functie h is gedefinieerd. De term Re is de restterm. Over deze term is veel bekend, maar het belangrijkste is dat hij 'snel naar nul gaat' voor $\boldsymbol{\theta} \to \boldsymbol{\theta_0}$. Dit 'snel naar nul gaan' kunnen we wiskundig precies maken. We introduceren hiervoor de volgende notaties.

Notatie 2.3. Met de notatie $f(n) = O(g(n))$ wordt bedoeld dat er constantes c en N bestaan zodanig dat $f(n) \leq c \cdot g(n)$ voor alle $n > N$.

Notatie 2.4. Voor d-vectoren $\mathbf{w}$ en $\mathbf{z}$ noteren we de Euclidische afstand ertussen met

$$\|\mathbf{z} - \mathbf{w}\| = \sqrt{\sum_{i=1}^{d} (z_i - w_i)^2}.$$

Nu blijkt te gelden [2] dat $\mathrm{Re} = O(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^3)$. Er geldt dus dat hoe dichter $\boldsymbol{\theta}$ en $\boldsymbol{\theta_0}$ bij elkaar liggen, hoe kleiner de fout wordt. We zullen in het vervolg met $\boldsymbol{\theta}$ een algemene vector aanduiden (veelal met steekproefgegevens), met $\hat{\boldsymbol{\theta}}$ de maximum-likelihoodschatter voor $\boldsymbol{\theta}$ en met $\boldsymbol{\theta_0}$ de optimale waarde voor $\boldsymbol{\theta}$, dat wil zeggen, de waarde waarvoor de Kullback-Leiblerdiscrepantie minimaal is. Dan geldt $\mathbb{E}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta_0}$ voor grote n.

Aangezien we voor de modelselectie voor h een loglikelihoodfunctie zullen invullen en ervan uitgaan dat $\boldsymbol{\theta}$ en $\boldsymbol{\theta_0}$ dicht bij elkaar liggen (hetgeen het geval is als $n \to \infty$), hoeven we ons over de restterm geen zorgen meer te maken.

Notatie 2.5. Met

$$\left[\frac{\partial h(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}}\right] = \begin{bmatrix} \frac{\partial h(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial h(\boldsymbol{\theta})}{\partial \theta_d} \end{bmatrix}_{\boldsymbol{\theta} = \boldsymbol{\theta_0}}$$

bedoelen we de kolomvector bestaande uit de d eerste-orde partiële afgeleiden van $h(\boldsymbol{\theta})$ naar $\theta_1, \ldots, \theta_d$ geëvalueerd in het punt $\boldsymbol{\theta} = \boldsymbol{\theta_0}$.

Notatie 2.6. Met

$$\left[\frac{\partial^2 h(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}^2}\right] = \left[\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}, \qquad 1 \leq i, j \leq d, \tag{2.2}$$

bedoelen we de matrix bestaande uit de $d \times d$ gemengde tweede-orde partiële afgeleiden van $h(\boldsymbol{\theta})$ naar $\theta_1, \ldots, \theta_d$ geëvalueerd in het punt $\boldsymbol{\theta} = \boldsymbol{\theta_0}$.

We beschouwen een approximating model als goed wanneer het erg weinig verschilt van het operating model. Hoe kleiner dit verschil, hoe beter het model. De grootte van het verschil wordt gegeven door de Kullback-Leiblerdiscrepantie, die we in Hoofdstuk 1 bekeken hebben. Deze wordt gegeven door

$$I(f, g) := \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x} \mid \boldsymbol{\theta_0})}\right) \mathrm{d}\mathbf{x},$$

waar f het operating model is en g het approximating model waarvan we de afstand tot f willen berekenen. We integreren over alle mogelijke steekproeven. Als we $f = g$ nemen, dan staat er $\log(1) = 0$ in de uitdrukking, wat betekent dat $I(f, f) = 0$. Dit is ook intuïtief geheel duidelijk.

Onder de nodige aannamen [9] voldoet $\boldsymbol{\theta_0}$ aan de vergelijking

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x} \mid \boldsymbol{\theta_0})}\right) \mathrm{d}\mathbf{x} = \mathbf{0}.$$

Gebruiken we nu dat $\log(a/b) = \log(a) - \log(b)$, dan volgt dat

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}) \log(f(\mathbf{x})) \, \mathrm{d}\mathbf{x} - \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}) \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \, \mathrm{d}\mathbf{x} = \mathbf{0}.$$

Omdat $\boldsymbol{\theta}$ niet voorkomt in f, is de eerste term hierboven $\mathbf{0}$. De tweede term kan geschreven worden als

$$\begin{aligned} \int f(\mathbf{x}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta}))\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} &= \mathbb{E}_f \left[\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta}))\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}\right] \\ &= \mathbb{E}_f \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))\right] \\ &= \mathbf{0}. \end{aligned} \tag{2.3}$$

In de afleiding van het AIC maken we gebruik van de verwachtingswaarde van stochasten. Deze verwachtingswaarden zijn gewoon integralen. De reden dat we niet overal de integraal uitschrijven, is dat de notatie van verwachtingswaarde makkelijker en eenvoudiger te lezen is. Toch is het goed altijd de achterliggende integralen in gedachten te houden. Daarmee valt een hoop te verklaren, zoals het omwisselen van verwachtingswaarden:

$$\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}}(h(\mathbf{x}, \mathbf{y})) = \mathbb{E}_{\mathbf{y}} \mathbb{E}_{\mathbf{x}}(h(\mathbf{x}, \mathbf{y})),$$

waar $h(\cdot\,, \cdot)$ een willekeurige functie is en $\mathbf{x}$ en $\mathbf{y}$ stochasten. Uit de stelling van Fubini volgt dat deze gelijkheid geldt ongeacht of $\mathbf{x}$ en $\mathbf{y}$ dezelfde kansverdeling hebben, en ongeacht of ze wel of niet onafhankelijk zijn.

We gaan nu nader kijken naar (2.2). Als we voor h de loglikelihoodfunctie invullen, dat wil zeggen, de logaritme van de likelihoodfunctie $g(\mathbf{x} \mid \boldsymbol{\theta})$, dan vinden we

$$\left[\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta}))}{\partial \theta_i \partial \theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}},$$

een matrix die duidelijk verband houdt met de **Fisher-informatiematrix**, die gegeven wordt door

$$\mathcal{I}(\boldsymbol{\theta_0}) = \mathbb{E}_g \left[-\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}.$$

Als g het echte model is, wat alleen het geval is als $f = g$ of als f een speciaal geval is van g,[1] dan is de covariantiematrix Σ van de maximum-likelihoodschatter (voor grote steekproefomvang) gelijk aan $\Sigma = [\mathcal{I}(\boldsymbol{\theta_0})]^{-1}$. Dat wil zeggen, $\Sigma = \mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' = \mathcal{I}(\boldsymbol{\theta_0})$. Als g niet het echte model is, dus als een ander model ten grondslag ligt aan de gemeten waarden $\mathbf{x}$, dan geldt in het algemeen dat $\Sigma \neq [\mathcal{I}(\boldsymbol{\theta_0})]^{-1}$.

Bij de afleiding van het AIC nemen we echter de verwachting met betrekking tot de functie f, en niet met betrekking tot de functie g, zoals hierboven. Het heeft dus zin nog een matrix in te voeren, en wel

$$I(\boldsymbol{\theta_0}) = \mathbb{E}_f \left[-\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} = \mathbb{E}_f \left[-\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial\boldsymbol{\theta}^2} \right]. \tag{2.4}$$

Er geldt dat $\mathcal{I}(\boldsymbol{\theta_0}) = I(\boldsymbol{\theta_0})$ dan en slechts dan als $f = g$ of als f een speciaal geval is van g.

Daarnaast kunnen we een soortgelijke notatie invoeren voor de (onbekende) matrix

$$\hat{I}(\boldsymbol{\theta_0}) = \left[-\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} = -\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial\boldsymbol{\theta}^2}.$$

Het is duidelijk dat $\mathbb{E}_f(\hat{I}(\boldsymbol{\theta_0})) = I(\boldsymbol{\theta_0})$. Als $\mathbf{x}$ een steekproef is uit de verdeling met dichtheidsfunctie f, dan convergeert $\hat{I}(\boldsymbol{\theta_0})$ naar $I(\boldsymbol{\theta_0})$ voor $n \to \infty$. We kunnen dit noteren als

$$\hat{I}(\boldsymbol{\theta_0}) = I(\boldsymbol{\theta_0}) + \mathrm{Re},$$

waar Re weer een restterm is. In het slechtste geval geldt $\mathrm{Re} = O(1/\sqrt{n})$.

Een schatter van $I(\boldsymbol{\theta_0})$ is $\hat{I}(\hat{\boldsymbol{\theta}})$, waar

$$\hat{I}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))}{\partial\boldsymbol{\theta}^2}. \tag{2.5}$$

Omdat $\hat{\boldsymbol{\theta}}$ de maximum-likelihoodschatter onder het model $g(\mathbf{x} \mid \boldsymbol{\theta})$ is, convergeert $\hat{\boldsymbol{\theta}}$ naar $\boldsymbol{\theta_0}$ als $n \to \infty$, dus convergeert $\hat{I}(\hat{\boldsymbol{\theta}})$ naar $I(\boldsymbol{\theta_0})$. Dus is $\hat{I}(\hat{\boldsymbol{\theta}}) \approx I(\boldsymbol{\theta_0})$, en de fout is maximaal $O(1/\sqrt{n})$. Merk op dat de Fisher-informatiematrix in het punt $\hat{\boldsymbol{\theta}}$ niet altijd naar $I(\boldsymbol{\theta_0})$ convergeert.

Stel dat $f = g$. Omdat het approximating model een kansdichtheidsfunctie is, geldt dat

$$\int g(\mathbf{x} \mid \boldsymbol{\theta})\, \mathrm{d}\mathbf{x} = 1,$$

en daarom geldt ook dat

$$\int \frac{\partial g(\mathbf{x} \mid \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\, \mathrm{d}\mathbf{x} = \mathbf{0}.$$

[1] Als we zeggen dat f een speciaal geval is van g, dan bedoelen we dat f genest is in g. Dat betekent dat we altijd model f kunnen krijgen uit model g door een geschikte keuze van parameters, maar niet andersom. Denk bijvoorbeeld aan een situatie waarin model f en g beide een histogram voorstellen, maar waar de staafjes van model f tweemaal zo breed zijn als van model g. Door steeds paarsgewijs staafjes even hoog te kiezen in model g, kunnen we ieder model f maken. In dit geval is model f dus genest in model g.

Omdat geldt dat

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) = \frac{1}{g(\mathbf{x} \mid \boldsymbol{\theta})} \left[\frac{\partial g(\mathbf{x} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right],$$

volgt dat

$$\int g(\mathbf{x} \mid \boldsymbol{\theta}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right] \mathrm{d}\mathbf{x} = \mathbf{0}. \tag{2.6}$$

Als we nu de vector met partiële afgeleiden van (2.6) naar $\boldsymbol{\theta}$ nemen, dan volgt met de kettingregel en enkele bovenstaande resultaten dat

$$\int g(\mathbf{x} \mid \boldsymbol{\theta}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right]' \mathrm{d}\mathbf{x} + \int g(\mathbf{x} \mid \boldsymbol{\theta}) \frac{\partial^2 \log g(\mathbf{x} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \, \mathrm{d}\mathbf{x} = O, \tag{2.7}$$

waar O de $d \times d$-nulmatrix is. Herschrijven van (2.7) geeft nu

$$\mathbb{E}_g \left[\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right]' \right] = \mathbb{E}_g \left[-\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2} \right],$$

of

$$\mathbb{E}_g \left[\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right]' \right] = \mathcal{I}(\boldsymbol{\theta}).$$

We definiëren nu $\mathcal{J}(\boldsymbol{\theta})$ als het linkerlid van bovenstaande vergelijking, dus

$$\mathcal{J}(\boldsymbol{\theta}) = \mathbb{E}_g \left[\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right]' \right].$$

Dus $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$.

We definiëren verder

$$J(\boldsymbol{\theta}) = \mathbb{E}_f \left[\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta})) \right]' \right].$$

Het zal alleen zo zijn dat $J(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$ als $f = g$ of f een speciaal geval is van g. Hoewel $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$ is er in het algemeen geen gelijkheid tussen $I(\boldsymbol{\theta})$ en $J(\boldsymbol{\theta})$ als g slechts een benadering van f is, dus als $I(f, g) > 0$. We kunnen echter wel verwachten dat $I(\boldsymbol{\theta_0}) \approx J(\boldsymbol{\theta_0})$, $\mathcal{I}(\boldsymbol{\theta_0}) \approx I(\boldsymbol{\theta_0})$ en $\mathcal{J}(\boldsymbol{\theta_0}) \approx J(\boldsymbol{\theta_0})$ als $I(f, g) \approx 0$, dus als g een goed approximating model is.

Er geldt verder dat

$$I(\boldsymbol{\theta_0})\Sigma = J(\boldsymbol{\theta_0})[I(\boldsymbol{\theta_0})]^{-1}, \tag{2.8}$$

en dus ook

$$\Sigma = [I(\boldsymbol{\theta_0})]^{-1} J(\boldsymbol{\theta_0})[I(\boldsymbol{\theta_0})]^{-1}. \tag{2.9}$$

Hier is Σ de werkelijke covariantiematrix van de maximum-likelihoodschatter van $\boldsymbol{\theta}$ voor grote steekproeven, afgeleid van model g met f als operating model.

Als we de likelihoodvergelijkingen geëvalueerd in $\boldsymbol{\theta_0}$ uitschrijven als eerste-orde Taylorontwikkelingen rond de maximum-likelihoodschatter, dan vinden we

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) \approx \frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) + \left[\frac{\partial^2 \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}^2} \right] (\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}).$$

De maximum-likelihoodschatter voldoet aan

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) = \mathbf{0},$$

zodat volgt

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) \approx \left[-\frac{\partial^2 \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))}{\partial^2} \right] (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = \hat{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \approx I(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}).$$

Daaruit volgt weer

$$[I(\boldsymbol{\theta_0})]^{-1} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) \right] \approx (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}). \tag{2.10}$$

Als we (2.10) transponeren, dan volgt

$$[I(\boldsymbol{\theta_0})]^{-1} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) \right]' [I(\boldsymbol{\theta_0})]^{-1} \approx (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})'.$$

Nemen we de verwachting naar f, dan krijgen we (zie (2.8)) dat

$$[I(\boldsymbol{\theta_0})]^{-1} J(\boldsymbol{\theta_0}) [I(\boldsymbol{\theta_0})]^{-1} \approx \mathbb{E}_f(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})' = \Sigma;$$

dit geeft ons (2.9) (wederom voor steekproeven met een grote omvang, dus als $n \to \infty$).

Zie [20] voor rigoureuze afleidingen van bovenstaande resultaten.

Om verwachtingen van de kwadratische vormen in (2.1) te kunnen nemen, zullen we een gelijkwaardige uitdrukking van die vorm gebruiken:

$$\mathbf{z}' A \mathbf{z} = \mathrm{tr}[A \mathbf{z}\mathbf{z}'].$$

Hier staat tr voor de spoorfunctie, die aan een vierkante matrix de som van de diagonaalelementen toekent. Deze operator is lineair, wat betekent dat geldt

$$\mathbb{E}_{\mathbf{z}}[\mathbf{z}' A \mathbf{z}] = \mathrm{tr}[\mathbb{E}_{\mathbf{z}}[A \mathbf{z}\mathbf{z}']].$$

Als A vast gekozen is, of in ieder geval niet van $\mathbf{z}$ afhangt, dan geldt dat $\mathbb{E}_{\mathbf{z}}[A\mathbf{z}\mathbf{z}'] = A\mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}']$. Als de verwachtingswaarde van $\mathbf{z}$ gelijk is aan nul, hetgeen het geval is bij bijvoorbeeld $\mathbf{z} = \hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}})$, dan geldt dat $\mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}'] = \Sigma$. Dan geldt dat

$$\mathbb{E}_{\mathbf{z}}[\mathbf{z}' A \mathbf{z}] = \mathrm{tr}[A\Sigma].$$

Als A een stochastische matrix is, maar wel onafhankelijk van $\mathbf{z}$, dan kunnen we gebruiken dat

$$\mathbb{E}_A \mathbb{E}_{\mathbf{z}}[\mathbf{z}' A \mathbf{z}] = \mathrm{tr}[\mathbb{E}_A \mathbb{E}_{\mathbf{z}}[A \mathbf{z}\mathbf{z}']] = \mathrm{tr}[\mathbb{E}_A(A)\mathbb{E}_A(\mathbf{z}\mathbf{z}')].$$

Ten slotte wijzen we erop dat

$$\begin{aligned}
\mathbb{E}_f \left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x}))) \right] &= \int f(\mathbf{x}) \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x}))) \, \mathrm{d}\mathbf{x} \\
&= \int f(\mathbf{y}) \log(g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))) \, \mathrm{d}\mathbf{y} \\
&= \mathbb{E}_f \left[\log(g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))) \right].
\end{aligned}$$

Het wisselen van notatie met betrekking tot $\mathbf{x}$ en $\mathbf{y}$ heeft dus geen effect op de uitkomst. In de afleiding hieronder zullen we dit gebruiken, aangezien we daar op sommige plaatsen werken met twee onafhankelijke steekproeven. In de afleiding zelf spelen steekproefgegevens geen enkele rol. We werken daar slechts met punten in een n-dimensionale uitkomstenruimte.

2.2. De afleiding

We gaan uit van de Kullback-Leiblerdiscrepantie en geven een afleiding van het AIC. De Kullback-Leiblerdiscrepantie wordt gegeven door

$$I(f,\ g(\cdot \mid \boldsymbol{\theta_0})) = \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x} \mid \boldsymbol{\theta_0})}\right) \mathrm{d}\mathbf{x}.$$

Merk op dat we $\boldsymbol{\theta}$ niet kennen, maar omdat we $\boldsymbol{\theta_0}$ schatten kunnen we de uitdrukking evalueren in $\boldsymbol{\theta_0}$. We krijgen zo bovenstaande uitdrukking. Daarnaast valt op dat de Kullback-Leiblerdiscrepantie $I(f, g)$ in het algemeen gezien kan worden als afhankelijk van de onbekende parameterwaarde, gegeven het approximating model. Maar $I(f, g)$ is onafhankelijk van steekproefgegevens, want $\mathbf{x}$ is slechts de variabele waarnaar we integreren.

Laat gegeven zijn een steekproef $\mathbf{y}$, die voortkomt uit het operating model f. De logische stap is dan het vinden van de maximum-likelihoodschatter $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ en vervolgens een schatter van de Kullback-Leiblerdiscrepantie te bepalen met

$$I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))) = \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))}\right) \mathrm{d}\mathbf{x}.$$

Die uitdrukking kunnen we niet expliciet uitrekenen (we kennen immers f niet), maar we zullen ermee verder redeneren en zien waar we uitkomen.

Als we $\boldsymbol{\theta_0}$ kunnen bepalen, dat wil zeggen, de waarde die de Kullback-Leiblerdiscrepantie voor een gegeven model g minimaliseert, dan zou een perfect model voldoen aan $I(f,\ g) = 0$. Op basis daarvan kunnen we ieder mogelijk model g beoordelen, want we prefereren dan een model met een kleinere Kullback-Leiblerdiscrepantie boven een ander model.

We hebben echter alleen een schatter van $\boldsymbol{\theta}$. Zelfs al zou ons model g exact gelijk zijn aan het operating model f, wat betekent dat $g(\mathbf{x} \mid \boldsymbol{\theta_0}) = f(\mathbf{x})$, dan zou onze schatter $\hat{\boldsymbol{\theta}}(\mathbf{y})$ nog altijd pas gelijk zijn aan $\boldsymbol{\theta_0}$ voor continue verdelingen bij een oneindig grote steekproefomvang. Voor discrete verdelingen geldt dat de gelijkheid geldt met een kans $\ll 1$. Als $\hat{\boldsymbol{\theta}}(\mathbf{y}) \neq \boldsymbol{\theta_0}$, dan zal $I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))) > I(f,\ g(\cdot \mid \boldsymbol{\theta_0}))$. Dit betekent dus dat, hoe goed het model ook is, in het algemeen altijd zal gelden dat de Kullback-Leiblerdiscrepantie strikt groter dan nul is. Dit roept de vraag op of we niet liever een andere uitdrukking willen minimaliseren.

We verwachten dat de Kullback-Leiblerdiscrepantie gemiddeld gelijk is aan

$$\mathbb{E}_{\mathbf{y}}\left[I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right].$$

Het ligt daarom voor de hand een model g_1 beter te vinden dan een model g_2 als geldt

$$\mathbb{E}_{\mathbf{y}}\left[I(f,\ g_1(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right] < \mathbb{E}_{\mathbf{y}}\left[I(f,\ g_2(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right].$$

Bedenk dat de verwachtingen naar f genomen worden en dat de notatie van de steekproef onbelangrijk is, dat wil zeggen, $\mathbf{x}$ of $\mathbf{y}$ is irrelevant. Omdat we $\boldsymbol{\theta}$ moeten schatten, zullen we werken met het criterium

"kies het model g dat $\mathbb{E}_{\mathbf{y}}\left[I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right]$ minimaliseert".

Ons doel is nu dus om de verwachte waarde van deze schatting van de Kullback-Leiblerdiscrepantie te minimaliseren. Als we de waarde van $\boldsymbol{\theta_0}$ zouden kunnen berekenen, dan zouden we de Kullback-Leiblerdiscrepantie zelf kunnen minimaliseren. Er geldt

$$\mathbb{E}_{\mathbf{y}}\left[I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right] - I(f,\ g(\cdot \mid \boldsymbol{\theta_0})) = \frac{1}{2}\mathrm{tr}\left[J(\boldsymbol{\theta_0})I(\boldsymbol{\theta_0})^{-1}\right];$$

dit verschil hangt niet van de steekproefgrootte af.

We kunnen de te minimaliseren term omschrijven. Dit geeft

$$\mathbb{E}_{\mathbf{y}}\left[I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right] = \int f(\mathbf{x})\log(f(\mathbf{x}))\,\mathrm{d}\mathbf{x} - \mathbb{E}_{\mathbf{y}}\left[\int f(\mathbf{x})\log[g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))]\,\mathrm{d}\mathbf{x}\right],$$

waaruit volgt dat

$$\mathbb{E}_{\mathbf{y}}\left[I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\right] = \text{constante} - \mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{x}}\left[\log[g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))]\right]. \tag{2.11}$$

Het blijkt dat we $\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{x}}\left[\log[g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))]\right]$ kunnen schatten, en derhalve een model kunnen kiezen dat (2.11) minimaliseert.

Er is ook een tweede aanpak die we kunnen volgen om, uitgaande van de Kullback-Leiblerdiscrepantie, het AIC af te leiden. We beginnen met

$$I(f,\ g(\cdot \mid \boldsymbol{\theta_0})) = \text{constante} - \mathbb{E}_{\mathbf{x}}\left[\log g(\mathbf{x} \mid \boldsymbol{\theta_0})\right]$$

en we proberen $\mathbb{E}_{\mathbf{x}}\left[\log g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))\right]$ te berekenen (of te schatten) met behulp van Taylorontwikkelingen. Het blijkt dat

$$\mathbb{E}_{\mathbf{x}}\left[\log g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))\right] \approx \mathbb{E}_{\mathbf{x}}\left[\log g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x}))\right] - \frac{1}{2}\mathrm{tr}\left[J(\boldsymbol{\theta_0})I(\boldsymbol{\theta_0})^{-1}\right] - \frac{1}{2}(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta_0})' I(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta_0}).$$

De laatste term van het rechterlid van bovenstaande vergelijking kunnen we op geen enkele manier berekenen of schatten. We kunnen echter wel aan beide kanten de verwachting naar $\mathbf{y}$ nemen. De uitdrukking die we dan krijgen,

$$\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{x}}\left[\log g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))\right] \approx \mathbb{E}_{\mathbf{x}}\left[\log g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x}))\right] - \mathrm{tr}\left[J(\boldsymbol{\theta_0})I(\boldsymbol{\theta_0})^{-1}\right],$$

kunnen we wel schatten.

Welke aanpak we ook kiezen, het komt er altijd op neer dat we, gegeven g, niet langer een model kiezen op basis van het minimaliseren van de Kullback-Leiblerdiscrepantie met bekende $\boldsymbol{\theta_0}$, maar een model kiezen met de geschatte $\boldsymbol{\theta}$ gebaseerd op het minimaliseren van een verwachte Kullback-Leiblerdiscrepantie. Het is nog altijd het geval dat slechts een relatief minimum kan worden gevonden, aangezien $\mathbb{E}_{\mathbf{x}}[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))]$ berekend noch geschat kan worden [3, 4, 13, 17].

We schatten de relatieve waarde van $\mathbb{E}_{\mathbf{y}}\left[I(f,\ g(\cdot \mid \hat{\boldsymbol{\theta}}(\mathbf{y}))\right]$ over de approximating family. Dat wil zeggen, we willen voor elk model in de approximating family de waarde van

$$T = \int f(\mathbf{y})\left[\int f(\mathbf{x})\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{y})))\,\mathrm{d}\mathbf{x}\right]\mathrm{d}\mathbf{y}$$

schatten.

In een iets vereenvoudigde, maar duidelijke notatie is het probleem nu om een nuttige uitdrukking voor en schatter van

$$T = \mathbb{E}_{\hat{\boldsymbol{\theta}}}\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))\right] \tag{2.12}$$

te vinden. Hier is $\hat{\boldsymbol{\theta}}$ de maximum-likelihoodschatter gebaseerd op de steekproef $\mathbf{y}$ en zijn beide verwachtingen wederom naar het operating model f. We kunnen T beschouwen als de dubbele verwachting gebaseerd op twee onafhankelijke steekproeven. Daaruit volgt dat modelselectie met het AIC asymptotisch gelijkwaardig is met crossvalidatie[2] [16]. Deze laatste is een algemeen geaccepteerd middel voor modelselectie.

De eerste stap is nu om (2.1) toe te passen op $\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))$ rond $\boldsymbol{\theta_0}$ voor elke $\mathbf{x}$. Dan volgt

$$\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) \approx \log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) + \left[\frac{\partial \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial \boldsymbol{\theta}}\right]' [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] + \frac{1}{2}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]' \left[\frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial \boldsymbol{\theta}^2}\right] [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]. \tag{2.13}$$

De fout die hier optreedt gaat naar 0 als $n \to \infty$. Om de uitdrukkingen (2.13) en (2.12) met elkaar in verband te brengen, nemen we bij uitdrukking (2.13) de verwachtingswaarde naar $\mathbf{x}$. Dit geeft ons

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))\right] \approx \mathbb{E}_{\mathbf{x}}[\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))] + \mathbb{E}_{\mathbf{x}}\left[\frac{\partial \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial \boldsymbol{\theta}}\right]' [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] + \\ + \frac{1}{2}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]' \left[\mathbb{E}_{\mathbf{x}} \frac{\partial^2 \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial \boldsymbol{\theta}^2}\right] [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}].\end{aligned}$$

De factor voor de $[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]$ in de eerste regel hierboven is precies (2.3). Met de notatie $\mathbb{E}_{\mathbf{x}}$ bedoelen we namelijk precies $\mathbb{E}_f$ over de steekproef $\mathbf{x}$ (en bedenk dat $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ onafhankelijk is van $\mathbf{x}$). De lineaire term is dus gelijk aan nul, want met (2.3) volgt

$$\mathbb{E}_{\mathbf{x}}\left[\frac{\partial \log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}.$$

Verder kunnen we (2.4) toepassen. Daarmee volgt dat

$$\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))\right] \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))\right] - \frac{1}{2}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]' I(\boldsymbol{\theta_0})[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]. \tag{2.14}$$

We kunnen nu de verwachting van (2.14) naar $\hat{\boldsymbol{\theta}}$ (dus naar $\mathbf{y}$) nemen. Daarmee volgt dat

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))\right] \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))\right] - \frac{1}{2}\mathrm{tr}\left[I(\boldsymbol{\theta_0})\mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]'\right]\right].$$

Het linkerlid van bovenstaande vergelijking is precies T uit (2.12). De term

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]'\right] = \Sigma$$

is de correcte theoretische steekproefvariantie van de maximum-likelihoodschatter voor grote steekproeven, omdat de verwachting hier genomen wordt naar f en niet naar g. Dan volgt

$$T \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))\right] - \frac{1}{2}\mathrm{tr}[I(\boldsymbol{\theta_0})\Sigma]. \tag{2.15}$$

[2]Bij crossvalidatie splits je de steekproefgegevens in twee delen, een deel om het model te selecteren en een deel om het geselecteerde model te controleren. Men herhaalt daarbij het proces van modelselectie en verificatie vele malen, steeds met een andere verdeling over de twee groepen.

Bij de volgende stap realiseren we ons dat we nog niet afgeleid hebben wat we willen, te weten een verband tussen T en $\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})(x))\right]$; de laatste term is de verwachting van de feitelijke loglikelihoodfunctie met de maximum-likelihoodschatter ingevuld. We maken nu een tweede Taylorontwikkeling, deze keer van $\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))$ rond het punt $\hat{\boldsymbol{\theta}}(\mathbf{x})$. We beschouwen nu $\mathbf{x}$ als de steekproefdata en we krijgen zo dus de maximum-likelihoodschatter van $\boldsymbol{\theta}$ voor $\mathbf{x}$. Deze procedure is acceptabel, omdat we alleen geïnteresseerd zijn in een verwachtingswaarde, wat betekent dat we de integraal nemen over alle mogelijke waarden in de steekproefruimte. Welke notatie we voor de steekproef gebruiken, $\mathbf{x}$ of $\mathbf{y}$, maakt dan ook niet uit. Passen we de Taylorontwikkeling toe (met dien verstande dat we de rol van $\hat{\boldsymbol{\theta}}$ en $\boldsymbol{\theta_0}$ omwisselen en er geldt dat $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$), dan krijgen we

$$\log(g(\mathbf{x} \mid \boldsymbol{\theta_0})) \approx \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) + \left[\frac{\partial \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}}\right]' [\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}] + \frac{1}{2}[\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]' \left[\frac{\partial^2 \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}^2}\right] [\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]. \tag{2.16}$$

De maximum-likelihoodschatter $\hat{\boldsymbol{\theta}}$ is de oplossing van, en voldoet dus aan, de vergelijking

$$\frac{\partial \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Daarom verdwijnt de lineaire term in (2.16). Nemen we nu de nodige verwachtingen, dan volgt

$$\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))\right] \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))\right] - \frac{1}{2}\mathrm{tr}\left[\mathbb{E}_{\mathbf{x}}\left[\hat{I}(\hat{\boldsymbol{\theta}})\right][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]'\right]. \tag{2.17}$$

Hier is $\hat{I}(\hat{\boldsymbol{\theta}})$ de Hessiaan van de loglikelihoodfunctie in de maximum-likelihoodschatter; zie (2.5). We gaan nu gebruiken dat $\hat{I}(\hat{\boldsymbol{\theta}}) \approx I(\boldsymbol{\theta_0})$. Dan volgt dat

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\left[\hat{I}(\hat{\boldsymbol{\theta}})\right][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]' &\approx [I(\boldsymbol{\theta_0})]\left[\mathbb{E}_{\mathbf{x}}[\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]'\right] \\
&= [I(\boldsymbol{\theta_0})]\left[\mathbb{E}_{\mathbf{x}}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]'\right] \\
&= [I(\boldsymbol{\theta_0})]\,\Sigma.
\end{aligned} \tag{2.18}$$

De fout die bij deze benadering hoort, is vaak $O(1/n)$ en daarom is deze benadering toegestaan voor $n \to \infty$. Het kan echter voorkomen dat de fout groter is, maar dan kunnen we alsnog het gewenste resultaat bereiken met

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\left[\hat{I}(\hat{\boldsymbol{\theta}})\right][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]' &\approx \left[\mathbb{E}_{\mathbf{x}}\left[\hat{I}(\hat{\boldsymbol{\theta}})\right]\right]\left[\mathbb{E}_{\mathbf{x}}[\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}]'\right] \\
&= [[I(\boldsymbol{\theta_0})]]\,\Sigma.
\end{aligned} \tag{2.19}$$

De bovenstaande benadering wordt beter naarmate de steekproefomvang toeneemt, maar de fout hier is niet eenvoudig te karakteriseren. Hoe dan ook, gebruikmakend van (2.18) of (2.19) in combinatie met (2.17) krijgen we

$$\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \boldsymbol{\theta_0}))\right] \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x})))\right] - \frac{1}{2}\mathrm{tr}[I(\boldsymbol{\theta_0})\Sigma].$$

Als we dit resultaat invullen in (2.15), dan volgt

$$T \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x})))\right] - \mathrm{tr}[I(\boldsymbol{\theta_0})\Sigma].$$

In de literatuur vinden we overigens vaak een iets andere uitdrukking, gebaseerd op (2.8):

$$T \approx \mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}(\mathbf{x})))\right] - \operatorname{tr}\left[J(\boldsymbol{\theta_0})[I(\boldsymbol{\theta_0}]^{-1}\right]. \tag{2.20}$$

We gebruiken hier de notatie $\hat{\boldsymbol{\theta}}(\mathbf{x})$ in plaats van $\hat{\boldsymbol{\theta}}$ om te benadrukken dat het rechterlid van bovenstaande vergelijking van slechts één steekproef afhangt, namelijk $\mathbf{x}$. Verder kunnen we afleiden dat een modelselectiecriterium (een bijna zuivere schatter van T) van de vorm

$$\hat{T} \approx \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) - \widehat{\operatorname{tr}}[I(\boldsymbol{\theta_0})\Sigma]$$

of

$$\hat{T} \approx \log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) - \widehat{\operatorname{tr}}[J(\boldsymbol{\theta_0})[I(\boldsymbol{\theta_0})]^{-1}] \tag{2.21}$$

is.

Het is niet mogelijk Σ eenvoudig en direct te schatten op basis van één waarneming, omdat er slechts één $\hat{\boldsymbol{\theta}}$ beschikbaar is. Zowel $J(\boldsymbol{\theta_0})$ als $I(\boldsymbol{\theta_0})$ zijn wel direct te schatten. Merk op dat het voor (2.21) nodig is dat $I(\boldsymbol{\theta_0})$ een matrix van volle rang is; dat wil zeggen, alle kolommen zijn onafhankelijk. Daaruit volgt namelijk dat de inverse bestaat. Zonder verlies van algemeenheid kunnen we echter aannemen dat alle parameters van de mogelijke modellen onafhankelijk van elkaar zijn, waaruit de volle rang volgt.

De gemaximaliseerde loglikelihood $\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))$ in (2.20) is een zuivere schatter van zijn eigen verwachting $\mathbb{E}_{\mathbf{x}}\left[\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))\right]$ (maar een onzuivere schatter van T). Het enige probleem dat ons nu nog in de weg staat, is de vraag hoe we een betrouwbare (dus een zuivere of vrijwel zuivere) schatter van de spoorterm kunnen vinden. Dan is het beste model dat model waarvoor de waarde van $\hat{T}$ zo groot mogelijk is, want dat model leidt dan tot de kleinste verwachte Kullback-Leiblerdiscrepantie. Per conventie wordt dit vaak geformuleerd als het minimaliseren van

$$-2\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) + 2\operatorname{tr}[J(\boldsymbol{\theta_0})[I(\boldsymbol{\theta_0})]^{-1}].$$

Als $f = g$ of f een speciaal geval van g is, dan geldt $I(\boldsymbol{\theta_0}) = \mathcal{I}(\boldsymbol{\theta_0}) = \mathcal{J}(\boldsymbol{\theta_0}) = J(\boldsymbol{\theta_0}) = \Sigma^{-1}$, waaruit volgt dat

$$\operatorname{tr}[I(\boldsymbol{\theta_0})\Sigma] = d.$$

Ook als g slechts een goede benadering is van f, maar geenszins gelijk aan of een algemeen geval van f is, blijkt te gelden dat de beste schatter waarschijnlijk $\widehat{\operatorname{tr}}[I(\boldsymbol{\theta_0})\Sigma] = d$ is [14].

Als het model te beperkt is om goed te kunnen zijn, dan zal de term $-2\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}}))$ erg groot zijn (vergeleken met de waarde die deze term heeft voor een beter model). We zullen dat model dan niet kiezen. Het is dan niet heel belangrijk of de schatter van de spoorterm goed is. Het Akaike InformatieCriterium, dat zoals bekend gegeven wordt door

$$\mathrm{AIC} = -2\log(g(\mathbf{x} \mid \hat{\boldsymbol{\theta}})) + 2d,$$

werkt het beste als er enkele goede modellen in de approximating family zitten, maar niet veel goede modellen met veel parameters. Hierbij vinden we een model goed als de Kullback-Leiblerdiscrepantie-waarde die het heeft klein is. Dan ligt het model dicht bij het operating model. Zie verder [5].

Hoofdstuk 3

Afleiding Bayesiaanse InformatieCriterium

Voor de afleiding van het Bayesiaanse InformatieCriterium (BIC) volgen we de grote lijnen van [8] en zullen we die op enkele punten uitbreiden.

We werken in een kansruimte $(\Omega,\ \mathcal{F},\ \mathbb{P})$. Dus Ω is een verzameling, $\mathcal{F}$ een σ-algebra, en $\mathbb{P}$ een kansfunctie. Een van de formules die ten grondslag ligt aan het BIC is de **regel van Bayes**. Deze luidt:

$$\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A) \cdot \frac{\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{3.1}$$

A en B zijn verzamelingen (gebeurtenissen) uit de gegeven kansruimte. De geldigheid van deze regel komt uit de definitie van een voorwaardelijke kans: $\mathbb{P}(A \mid B) = \mathbb{P}(A \text{ en } B)/\mathbb{P}(B)$ als $\mathbb{P}(B) \neq 0$. Bekijken we de breuk $\mathbb{P}(A \mid B)/\mathbb{P}(B \mid A)$ dan kunnen we voor beide voorwaardelijke kansen de definitie invullen. Vereenvoudigen levert:

$$\frac{\mathbb{P}(A \mid B)}{\mathbb{P}(B \mid A)} = \frac{\mathbb{P}(A \text{ en } B)/\mathbb{P}(B)}{\mathbb{P}(A \text{ en } B)/\mathbb{P}(A)} = \frac{1/\mathbb{P}(B)}{1/\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

en beide kanten met $\mathbb{P}(B \mid A)$ vermenigvuldigen geeft dan (3.1).

De Bayesianen vervangen in (3.1) nu de A door een H (van hypothese) en de B door een D (van data). De regel van Bayes zegt dan dat de kans op een hypothese gegeven de data gelijk is aan de kans op de data gegeven de hypothese maal een factor. Deze factor is $\mathbb{P}(H)/\mathbb{P}(D)$ en staat voor de kans dat de hypothese juist is voordat naar data gekeken is, gedeeld door de kans dat de data voorkomen (dus als je geen enkele theorie hebt). Deze twee kansen worden **priors** genoemd. Een belangrijke opmerking moet nu gemaakt worden. In de definitie van de regel van Bayes staan de symbolen A en B voor gebeurtenissen. Maar wanneer je deze door D en H vervangt, met de bovenstaande betekenis, heb je niet langer met gebeurtenissen te maken. De symbolen hebben dus een andere betekenis gekregen, en dat betekent dat (in de context van data en hypothese) de regel van Bayes niet langer een definitie (of identiteit) is. Dit is een zwak punt in de afleiding van het BIC.

Nu willen we twee verschillende hypotheses H_1 en H_2 met elkaar vergelijken, waarbij we willen zeggen welk van de twee het meest waarschijnlijk is. Een voor de hand liggende keuze is het berekenen van de breuk $\mathbb{P}(H_1 \mid D)/\mathbb{P}(H_2 \mid D)$. Is deze namelijk (veel) groter dan 1 dan vinden we H_1 waarschijnlijker dan H_2. Is de breuk (veel) kleiner dan 1, dan is H_2 waarschijnlijker.

We kunnen de verhouding van voorwaardelijke kansen uitschrijven met de regel van Bayes:

$$\begin{aligned}\frac{\mathbb{P}(H_1 \mid D)}{\mathbb{P}(H_2 \mid D)} &= \frac{\mathbb{P}(D \mid H_1)\mathbb{P}(H_1)/\mathbb{P}(D)}{\mathbb{P}(D \mid H_2)\mathbb{P}(H_2)/\mathbb{P}(D)} \\ &= \frac{\mathbb{P}(D \mid H_1)\mathbb{P}(H_1)}{\mathbb{P}(D \mid H_2)\mathbb{P}(H_2)} \\ &= \frac{\mathbb{P}(D \mid H_1)}{\mathbb{P}(D \mid H_2)} \cdot \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}.\end{aligned}$$

Volgens de Bayesianen zijn alle hypotheses even waarschijnlijk als je nog geen data hebt, en dus is $\mathbb{P}(H_1) \approx \mathbb{P}(H_2)$. Maar dan volgt

$$\frac{\mathbb{P}(H_1 \mid D)}{\mathbb{P}(H_2 \mid D)} \approx \frac{\mathbb{P}(D \mid H_1)}{\mathbb{P}(D \mid H_2)}. \tag{3.2}$$

Als g_1 en g_2 de kansverdelingsfuncties van H_1 respectievelijk H_2 zijn en $\mathbf{x}$ de waargenomen data dan kunnen we (3.2) schrijven als

$$\frac{\mathbb{P}(g_1 \mid \mathbf{x})}{\mathbb{P}(g_2 \mid \mathbf{x})} \approx \frac{\mathbb{P}(\mathbf{x} \mid g_1)}{\mathbb{P}(\mathbf{x} \mid g_2)} = \frac{\mathcal{L}(\mathbf{x} \mid g_1)}{\mathcal{L}(\mathbf{x} \mid g_2)},$$

waarbij de Bayesianen $\mathcal{L}(\mathbf{x} \mid g)$ definiëren als

$$\mathcal{L}(\mathbf{x} \mid g) = \int \mathbb{P}(\boldsymbol{\theta})\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})\,\mathrm{d}\boldsymbol{\theta}. \tag{3.3}$$

$\mathbb{P}(\boldsymbol{\theta})$ is hier een **a priori** kansverdeling over de parameters en $\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ de likelihoodfunctie. De integraal wordt genomen over het bereik van de parameters $\boldsymbol{\theta}$. We kunnen deze integraal ook wel benaderen met de methode van Laplace, genaamd **steepest descent** [22]. Deze methode wordt vooral gebruikt om integralen van de vorm $\int_a^b \mathrm{e}^{Mf(\mathbf{x})}\,\mathrm{d}\mathbf{x}$ te benaderen. Daar de integraal in (3.3) ook (ongeveer) van die vorm is, kunnen we deze methode toepassen. Daarvoor nemen we $\mathbf{x}$ vast en definiëren $S(\boldsymbol{\theta}) := -\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$. Dan is

$$\int \mathbb{P}(\boldsymbol{\theta})\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})\,\mathrm{d}\boldsymbol{\theta} = \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-S(\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta}$$

een positieve functie, want $\mathbb{P}(\boldsymbol{\theta})$ ligt tussen 0 en 1 en ook de e-macht neemt geen negatieve waarden aan.

We weten dat $\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ voor gegeven $\mathbf{x}$ een maximum heeft in $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, want het is een likelihoodfunctie. Om de methode van steepest descent te mogen gebruiken, moeten we aannemen dat $\hat{\boldsymbol{\theta}}$ binnen het integratiegebied ligt, en niet op de rand of erbuiten. Bovendien moet de likelihoodfunctie een vorm van continuïteit hebben, zodat (voor vaste $\mathbf{x}$) $\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ en $\mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})$ alleen dicht bij elkaar liggen als $\boldsymbol{\theta}$ en $\hat{\boldsymbol{\theta}}$ dicht bij elkaar liggen. In wiskundige notatie:

$$\forall_{\varepsilon>0}\ \exists_{\delta>0}\ [|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}| < \delta \ \Rightarrow\ |\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}}) - \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})| < \varepsilon].$$

Een derde aanname is dat $[\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})]'' < 0$. We maken nu een Taylorontwikkeling van $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ rond $\hat{\boldsymbol{\theta}}$ en bekijken de functie tot en met de tweede orde. Dit levert voor een vector $\boldsymbol{\theta}$ van dimensie 1:

$$\begin{aligned}\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}}) = \log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) &+ [\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})]'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \\ &+ \frac{1}{2}[\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})]''(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2 + O((\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^3).\end{aligned}$$

Omdat $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ een (lokaal) maximum heeft in $\hat{\boldsymbol{\theta}}$, en $\hat{\boldsymbol{\theta}}$ binnen het integratiegebied ligt, geldt $[\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})]' = 0$. Dan volgt dat we de volgende benadering hebben voor $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ als $\boldsymbol{\theta}$ dicht bij $\hat{\boldsymbol{\theta}}$ ligt:

$$\begin{aligned}\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}}) &\approx \log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) + \frac{1}{2}[\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})]''(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2 \\ &= \log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) - \frac{1}{2}|\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})|''(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2.\end{aligned}$$

Met deze benadering van $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ kunnen we ook de integraal uit (3.3) benaderen:

$$\begin{aligned}\int \mathbb{P}(\boldsymbol{\theta})\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})\,\mathrm{d}\boldsymbol{\theta} &= \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{\log \mathcal{L}(\mathbf{x}|g_{\boldsymbol{\theta}})}\,\mathrm{d}\boldsymbol{\theta} \\ &\approx \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{\log \mathcal{L}(\mathbf{x}|g_{\hat{\boldsymbol{\theta}}}) - \frac{1}{2}|\log \mathcal{L}(\mathbf{x}|g_{\hat{\boldsymbol{\theta}}})|''(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^2}\mathrm{d}\boldsymbol{\theta} \\ &= \mathrm{e}^{\log \mathcal{L}(\mathbf{x}|g_{\hat{\boldsymbol{\theta}}})} \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-\frac{1}{2}|\log \mathcal{L}(\mathbf{x}|g_{\hat{\boldsymbol{\theta}}})|''(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^2}\mathrm{d}\boldsymbol{\theta} \\ &= \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-\frac{1}{2}|\log \mathcal{L}(\mathbf{x}|g_{\hat{\boldsymbol{\theta}}})|''(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^2}\mathrm{d}\boldsymbol{\theta} \\ &= \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-\frac{1}{2}|S(\boldsymbol{\theta})|''(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^2}\mathrm{d}\boldsymbol{\theta}.\end{aligned}$$

Dit geldt, zoals gezegd, alleen als $\boldsymbol{\theta}$ een 1-dimensionale vector is. In het geval dat $\boldsymbol{\theta}$ een grotere dimensie heeft dan 1 is de tweede afgeleide van $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ (en dus ook van $S(\boldsymbol{\theta})$) namelijk een matrix van afgeleiden. Per definitie van de Taylorreeks laten we A de (bij aanname positief-definiete) minusmatrix in de kwadratische term van de Taylorexpansie van $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ (ofwel $S(\boldsymbol{\theta})$) rond $\hat{\boldsymbol{\theta}}$ zijn. Dan is

$$A_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} S(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

en dan is de tweede term van de Taylorreeks van $\log \mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})$ rond $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ dus $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}},\ A(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}))$ dat gelijk is aan de matrix met op plek (i, j) de waarde van $\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}})(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$.

Hieruit volgt nu, door alle bovenstaande conclusies te combineren, dat

$$\begin{aligned}\mathcal{L}(\mathbf{x} \mid g) &= \int \mathbb{P}(\boldsymbol{\theta})\mathcal{L}(\mathbf{x} \mid g_{\boldsymbol{\theta}})\,\mathrm{d}\boldsymbol{\theta} \\ &= \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-S(\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta} \\ &\approx \mathcal{L}(\mathbf{x} \mid g_{\hat{\boldsymbol{\theta}}}) \int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta}, A\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta}.\end{aligned}$$

Omdat de prior $\mathbb{P}(\boldsymbol{\theta})$ volgens de Bayesianen bij gebrek aan voorkennis constant is (zeg c), kunnen we deze buiten de integraal halen. Vervolgens kunnen we de integraal uitbreiden naar $\mathbb{R}^d$. Dit mag omdat $\boldsymbol{\theta}$ op die uitbreiding, dankzij de e-macht, kleine waarden aanneemt; hij heeft dan immers een grote afstand tot het maximum. Vervolgens maken we gebruik van het feit dat A symmetrisch is. We kunnen A dan via een orthogonale transformatie met O diagonaliseren tot een matrix met diagonaalelementen $a_1, \ldots, a_d$. Voor de orthogonale matrix O geldt $|\det O| = 1$ (want $O^\mathsf{T}O = OO^\mathsf{T} = I$) en dus is de determinant van A gelijk aan het product van de diagonaalelementen: $\det A = \prod_{i=1}^d a_i$. We kunnen $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}},\ A(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}))$ dan schrijven als $\sum_{i=1}^d a_i \tilde{\theta}_i^2$,

waar $\tilde{\theta}_i$ de (coördinaat)getransformeerde θ_i is. We kunnen de hele integraal dan splitsen in d Gaussische integralen (met factor $a_i/2$ voor de $\boldsymbol{\theta}^2$), en daarvan weten we de uitkomst, namelijk $\sqrt{2\pi}/\sqrt{a_i}$. Uit dit alles volgt dan

$$\begin{aligned}
\int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-(\boldsymbol{\theta},A\boldsymbol{\theta})/2}\,\mathrm{d}\boldsymbol{\theta} &= c\int \mathrm{e}^{-(\boldsymbol{\theta},A\boldsymbol{\theta})/2}\,\mathrm{d}\boldsymbol{\theta}\\
&= c\int_{\mathbb{R}^d} \mathrm{e}^{-(\boldsymbol{\theta},A\boldsymbol{\theta})/2}\,\mathrm{d}\boldsymbol{\theta}\\
&= c\int_{\mathbb{R}^d} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^d a_i\tilde{\theta}_i^2}\,\mathrm{d}\boldsymbol{\theta}\\
&= c\int_{\mathbb{R}} \mathrm{e}^{-\frac{1}{2}a_1\tilde{\theta}_1^2}\,\mathrm{d}\theta_1\cdots\int_{\mathbb{R}} \mathrm{e}^{-\frac{1}{2}a_d\tilde{\theta}_d^2}\,\mathrm{d}\theta_d\\
&= c\left(\frac{\sqrt{2\pi}}{\sqrt{a_1}}\cdots\frac{\sqrt{2\pi}}{\sqrt{a_d}}\right)\\
&= c(2\pi)^{d/2}\frac{1}{\sqrt{a_1\cdots a_d}}\\
&= c(2\pi)^{d/2}(\det A)^{-1/2},
\end{aligned}$$

en dus is

$$\begin{aligned}
\log\int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-(\boldsymbol{\theta},A\boldsymbol{\theta})/2}\,\mathrm{d}\boldsymbol{\theta} &\approx \log\left(c(2\pi)^{d/2}(\det A)^{-1/2}\right)\\
&= \log c + \log(2\pi)^{d/2} + \log(\det A)^{-1/2}\\
&= \log c + \frac{d}{2}\log 2\pi - \frac{1}{2}\log(\det A).
\end{aligned}$$

Omdat $\lim\limits_{n\to\infty} \det A \sim n^d$ (met de centrale limietstelling), krijgen we de benadering

$$\begin{aligned}
-2\log\mathcal{L}(\mathbf{x}\mid g) &= -2\log\left(\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}})\int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta},A\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta}\right)\\
&= -2\log\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}}) - 2\log\int \mathbb{P}(\boldsymbol{\theta})\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta},A\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta}\\
&\approx -2\log\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}}) - 2\left(\log c + \frac{d}{2}\log 2\pi - \frac{1}{2}\log(\det A)\right)\\
&\approx -2\log\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}}) - 2\left(\log c + \frac{d}{2}\log 2\pi - \frac{1}{2}\log n^d\right)\\
&= -2\log\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}}) - 2\log c - 2\frac{d}{2}\log 2\pi + 2d\frac{1}{2}\log n\\
&= -2\log\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}}) - 2\log c - d\log 2\pi + d\log n.
\end{aligned}$$

We kunnen de constante termen in bovenstaande uitdrukking weglaten. We gebruiken het BIC namelijk om voor verschillende kandidaatmodellen te bepalen welk model de kleinste BIC-waarde heeft. Omdat de constante termen dus in elk van die waarden evenveel bijdrage hebben, hebben ze geen invloed op het bepalen van het model met de kleinste BIC-waarde. Zonder de constantes krijgen we dan de volgende uitdrukking:

$$\mathrm{BIC} = -2\log\mathcal{L}(\mathbf{x}\mid g_{\hat{\boldsymbol{\theta}}}) + d\log n.$$

Hoofdstuk 4

Lineaire regressie

In dit verslag behandelen we het AIC en het BIC en in het bijzonder de verschillen tussen deze criteria. Een belangrijke vraag is dan niet welke van de twee het beste is, maar welke het beste is gegeven een bepaalde situatie. Het zou natuurlijk mooi zijn als een van de twee altijd beter presteert dan de ander, maar dit is helaas niet zo. Omdat je als onderzoeker meestal wel een idee hebt uit wat voor soort populatie de data komen, kun je op basis daarvan voor het AIC of BIC kiezen. We bekijken in dit hoofdstuk het geval van lineaire regressie. Als je weet dat je met een lineair regressiemodel te maken hebt, welk criterium presteert dan het beste? In [10] vinden we informatie hierover.

4.1. Achtergrond

Voor we met de analyse beginnen zetten we kort de theorie van lineaire regressie uiteen. We hebben n stochasten $Y_1, Y_2, \ldots, Y_n$ die we willen voorspellen (waar we een lineair model voor willen opstellen). Voor deze n stochasten zijn er $n \cdot p$ verklarende variabelen $X_{11}, X_{12}, \ldots, X_{1p}$, $X_{21}, X_{22}, \ldots, X_{2p}, X_{31}, \ldots, X_{np}$. Hierbij verklaren $X_{i1}, X_{i2}, \ldots, X_{ip}$ de stochast Y_i. We kunnen nu een lineair model maken als geldt: er zijn $\beta_1, \beta_2, \ldots, \beta_p \in \mathbb{R}$ zodat voor alle i met $1 \leq i \leq n$ geldt

$$Y_i = \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \cdots + \beta_p \cdot X_{ip} + \varepsilon_i.$$

Hierbij zijn $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ onafhankelijk en normaal verdeeld met verwachting 0 en variantie σ^2.[1] Dus $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. De ε_i zijn dus random afwijkingen, ook wel **storingstermen** genoemd. Het is echter gebruikelijk om een constante term toe te voegen in het model. Dit zal $\beta_0 \in \mathbb{R}$ zijn, en heet de **intercept** (deze naam komt van het snijpunt op de y-as, namelijk als alle andere β_i's gelijk aan 0 zijn). We kunnen het ook opvatten als een extra verklarende variabele X_{i0} voor alle Y_i die altijd gelijk aan 1 is. We krijgen dan het volgende lineaire model:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \cdots + \beta_p \cdot X_{ip} + \varepsilon_i. \tag{4.1}$$

Het model is van **orde** $k = p + 1$ omdat er (bij dit model) $p + 1$ variabelen van invloed zijn op de Y_i.

[1] Dit kan ook genoteerd worden als $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ waar 0 de n-dimensionale nulvector is en I de $n \times n$-eenheidsmatrix.

Een andere, en kortere, manier om deze vergelijkingen te beschrijven is met behulp van matrices. We definiëren er vier:

$$X := \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}, \quad Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \beta := \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon := \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Hierbij zijn de ε_i onafhankelijk verdeeld volgens $\mathcal{N}(0, \sigma^2)$. We kunnen het stelsel van lineaire vergelijkingen dan schrijven als

$$Y = X \cdot \beta + \varepsilon. \tag{4.2}$$

We kunnen nu drie modelstructuren onderscheiden: het **operating model**, het **algemene model** en het **fitted model**, ook wel **approximating model** genoemd. Het operating model (het echte regressiemodel) is

$$Y = \mu_* + \varepsilon_* \tag{4.3}$$

met $\varepsilon_* \sim \mathcal{N}(0, \sigma_*^2 I)$, $Y = (Y_1, \ldots, Y_n)^\mathsf{T}$, $\mu_* = (\mu_{*1}, \ldots, \mu_{*n})^\mathsf{T}$ waar μ_{*i} de echte onbekende functie is en $\varepsilon_* = (\varepsilon_{*1}, \ldots, \varepsilon_{*n})^\mathsf{T}$. We nemen aan dat de ε_{*i} onafhankelijk en identiek normaal verdeeld zijn met variantie σ_*^2. De μ_{*i} zijn dus functies. In het ideale geval zijn het lineaire combinaties van de X_{ik} en β_k, want dan is μ te schrijven als $X\beta$ en hebben we een lineair model. Maar de werkelijkheid kan iets afwijken van deze lineaire situatie. Het beste wat we dan met regressie kunnen doen, is de werkelijkheid benaderen met een lineair model. Het operating model beschrijft de werkelijkheid, en is dus niet echt een model. Maar het heeft de vorm van (en is tot op zekere hoogte) een lineair (regressie) model. We zullen, net als in [10], alles wat met het operating model te maken heeft noteren met een x_*.

Het algemene model is

$$Y = X\beta + \varepsilon, \tag{4.4}$$

waar Y, X, β en ε gedefinieerd zijn zoals voor vergelijking (4.2).

Het approximating model is van de vorm van het algemene model, en daarvan willen we bestuderen of het de werkelijkheid goed benadert. We hebben dus een hele verzameling van approximating models, en daaruit willen we, volgens het doel van modelselectie, de beste kiezen.

We kunnen de verzameling van approximating models in twee deelverzamelingen splitsen: de groep van **underfitting models** en de groep van **overfitting models**. Bij underfitting heeft een model te weinig parameters (ten opzichte van het operating model) en bij overfitting heeft het model er te veel. Zie ook paragraaf 1.2.

Dit kunnen we wiskundig op de volgende manier noteren. Elk approximating model heeft een matrix X en een vector β die de parameters bevatten voor dat betreffende model. Dus een model van de vorm $Y_i = \beta_0 + \beta_2 X_{i2}$ heeft een $n \times 2$-matrix X (met in de eerste kolom allemaal enen) en een vector β van de vorm $(\beta_0, \beta_2)^\mathsf{T}$.

We splitsen de $n \times k$-matrix X van het algemene model in drie delen: $X = (X_0, X_1, X_2)$, waar X_i ($i \in \{0, 1, 2\}$) een $n \times k_i$ matrix is. Ook de vector β splitsen we: $\beta = (\beta_0, \beta_1, \beta_2)^\mathsf{T}$, waar β_i ($i \in \{0, 1, 2\}$) een $k_i \times 1$ vector is. Stel nu dat het operating model bestaat uit die parameters zodat $\mu_* = X_*\beta_*$ met $X_* = (X_0, X_1)$ en $\beta_* = (\beta_0^\mathsf{T}, \beta_1^\mathsf{T})$. Dan hebben we voor een approximating model met matrix X underfitting als $\text{rang}(X) < \text{rang}(X_*)$ en overfitting als $\text{rang}(X_*) < \text{rang}(X)$. De rang is het aantal onafhankelijke kolommen van een matrix. Bovendien kunnen we een underfitted model nu schrijven als

$$Y = X_0\beta_0 + \varepsilon,$$

en een overfitted model als

$$Y = X\beta + \varepsilon = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Het overfitted model heeft de precies dezelfde vorm als het algemene model, het underfitted model heeft minder parameters en dus kleinere matrices X.

We nemen aan dat de kleinste-kwadratenmethode gebruikt wordt om de β_i's te bepalen uit de data, en dat de approximating models van orde k zijn. De maximum-likelihoodschatter van β is $\widehat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$. Dit kunnen we op de volgende manier inzien. We weten volgens de aanname dat $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Zij U nu het bereik van de $n \times k$-matrix X, met $k \leq n$. Stel X heeft een maximale rang, dus er zijn geen afhankelijke kolommen. Zij $u \in U$. We kunnen dan schrijven $u = X(B_u)$ waar B_u de kolom van β_i's is die bij u horen (schatting). Omdat B een lineaire afbeelding is van U naar $\mathbb{R}^k$ (en misschien wel $\mathbb{R}^n$) volgt

$$X^\mathsf{T}u = X^\mathsf{T} \cdot X(B_u) = (X^\mathsf{T}X) \cdot B_u,$$

en dus

$$B_u = (X^\mathsf{T}X)^{-1} \cdot X^\mathsf{T}u,$$

en dus wanneer we Y willen weten,

$$\widehat{\beta} = B_Y = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y.$$

Omdat X een $n \times k$ matrix is, $X^\mathsf{T}X$ een $k \times k$ matrix, B_u een vector van lengte k en u een vector van lengte n volgt dat we op deze manier een goed gedefinieerde uitdrukking hebben om de β_i's te schatten. Dit gaat echter alleen op als $(X^\mathsf{T}X)^{-1}$ bestaat. Dus als $X^\mathsf{T}X$ inverteerbaar is, dan is er een unieke oplossing voor B_u (en dus voor B_Y).

We kunnen ook de **maximum-likelihoodschatter** van σ^2 bepalen. Stel dat de Y_i onafhankelijk en normaal verdeeld zijn met verwachting μ_i en variantie σ^2 (dus $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$) en $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$.

Dan krijgen we de volgende likelihoodfunctie:

$$\begin{aligned}
\mathcal{L}(\mu, \sigma^2) &= f_{\mu,\sigma^2}(y_1, y_2, \ldots, y_n) \\
&= f_{\mu_1,\sigma^2}(y_1) \cdot f_{\mu_2,\sigma^2}(y_2) \cdots f_{\mu_n,\sigma^2}(y_n) \\
&= \prod_{i=1}^{n} f_{\mu_i,\sigma^2}(y_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot \mathrm{e}^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \\
&= \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sigma^n} \cdot \mathrm{e}^{-\frac{\|y-\mu\|^2}{2\sigma^2}}.
\end{aligned}$$

Om σ te vinden kunnen we $\mathcal{L}$ naar σ differentiëren en die afgeleide op 0 stellen. Dit levert

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma} &= -n \cdot \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sigma^{n+1}} \cdot \mathrm{e}^{-\frac{\|y-\mu\|^2}{2\sigma^2}} - 2 \cdot \frac{\|y-\mu\|^2}{2\sigma^3} \cdot \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sigma^n} \cdot \mathrm{e}^{-\frac{\|y-\mu\|^2}{2\sigma^2}} \\
&= \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sigma^{n+1}} \cdot \mathrm{e}^{-\frac{\|y-\mu\|^2}{2\sigma^2}} \cdot (-n + \frac{\|y-\mu\|^2}{\sigma^2}).
\end{aligned}$$

Dit laatste lid stellen we op 0. Omdat de eerste drie termen geen nul worden, moet de laatste dit wel zijn. Dus we vinden σ^2 door deze op te lossen uit

$$\begin{aligned} & -n + \frac{\|y-\mu\|^2}{\sigma^2} = 0 \\ \Rightarrow \quad & \frac{\|y-\mu\|^2}{\sigma^2} = n \\ \Rightarrow \quad & \frac{\|y-\mu\|^2}{n} = \sigma^2. \end{aligned}$$

Dus de maximum-likelihoodschatter voor σ^2 is

$$\hat{\sigma}^2 = \frac{\|y-\mu\|^2}{n}. \tag{4.5}$$

Deze formule kunnen we natuurlijk alleen gebruiken als we μ kennen. Weten we μ niet, dan moeten we hiervoor de maximum-likelihoodschatter $\hat{\mu}$ van μ invullen. In Hoofdstuk 5 zullen we $\hat{\mu}$ bepalen. De maximum-likelihoodschatter van σ^2 is dus $\hat{\sigma}_k^2 = \frac{SSE_k}{n}$, waar $SSE_k = ||Y - \widehat{Y}||^2 = ||y - \hat{\mu}||^2$ de gebruikelijke som van kwadraten van afwijkingen is en $\widehat{Y} = X\widehat{\beta} = \hat{\mu}$.

Als van een onderliggend lineair regressiemodel uitgegaan wordt, dan is de belangrijkste vraag: welke (verklarende) variabelen zijn belangrijk? Zodra je deze variabelen weet, kun je goede schattingen maken van toekomstige waarnemingen.

In paragraaf 8 van [18] staat een voorbeeld van een situatie waar je uit verschillende lineaire modellen er een wil kiezen. Stel je hebt al p verklarende variabelen gevonden, maar je denkt dat een aantal daarvan niet in het model horen. Dit betekent dat in vergelijking (4.1) een aantal β's gelijk aan 0 is. De verzameling van approximating models bestaat dan uit de modellen $\mathcal{M}_S$ waar S een deelverzameling is van $\{0, 1, \ldots, p\}$. Voor model $\mathcal{M}_S$ geldt dat de β_i's in (4.1) gelijk aan 0 zijn, behalve de β_j met $j \in S$. Dus als $S = \{0, 1, 3\}$, dan is $\mathcal{M}_S$ het model $y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_3 \cdot x_{i3} + \varepsilon_i$. We hebben dus $2^{p+1} = 2^k$ verschillende modellen.

4.2. Efficiëntie en consistentie

Een gebruikelijk aanname in regressieanalyse volgens [10] is dat het operating model oneindig dimensionaal is. Dit betekent dat het operating model niet in de verzameling van approximating models zit. Het doel van modelselectie is dan uit de verzameling van (eindig dimensionale) approximating models $\mathcal{M}_S$ dat model te kiezen dat met zijn variabelen het dichtst in de buurt komt van het operating model. Een criterium dat voor grote steekproeven het model met de kleinste **mean square error distribution** kiest, wordt **asymptotisch efficiënt** genoemd. Voorbeelden van zulke criteria zijn het AIC en het Cp, dat wordt gegeven door

$$\text{Cp} = \frac{SSE_K}{s_k^2} - n + 2k,$$

met $K = k_* + o$ de som van de echte orde van het model en de mate van overfitting. Onderzoekers die geloven dat het systeem dat ze bestuderen oneindig gecompliceerd is, of dat niet alle belangrijke variabelen gemeten kunnen worden, kiezen een model op basis van efficiëntie.

Als men aanneemt dat het operating model eindig dimensionaal is, en dus bevat is in de verzameling van approximating models, dan is het doel van modelselectie het kiezen van het operating

model uit die verzameling. Een criterium dat (voor grote steekproeven) met kans 1 het operating model aanwijst heet **consistent**. SIC (de gecorrigeerde versie $\mathrm{SIC}_k = \log(\hat{\sigma}_k^2) + \frac{\log(n)k}{n}$ wordt in [10] bekeken) is zo'n criterium en is in zowel prestatie als in tijd van introductie gelijk aan het BIC (dus ook BIC hoort is consistent). Bij consistentie wordt ervan uitgegaan dat alle variabelen gemeten kunnen worden en dat genoeg bekend is over het "physical system" dat bestudeerd wordt om een lijst van alle belangrijke variabelen te maken. Dit zijn sterke aannamen, maar deze kunnen gelden in bijvoorbeeld de natuurkunde, waar veel theorieën zijn die deze aannamen rechtvaardigen.

Omdat de keuze voor een efficiënt of een consistent model erg afhankelijk is van de onderzoeker en de manier waarop hij de complexiteit en meetbaarheid inschat, is niet in het algemeen te zeggen welk van deze soorten criteria beter is.

Er is een klein verschil in definitie van underfitting en overfitting tussen consistentie en efficiëntie. Onder consistentie is overfitting gedefinieerd als het kiezen van een model met extra variabelen ten opzichte van het operating model. Underfitting is dan gedefinieerd als het kiezen van een model dat te weinig variabelen heeft ten opzichte van het operating model, of incompleet is. Er is geen term voor het kiezen van een model met het juiste aantal variabelen, maar met een verkeerde keuze van variabelen. Onder efficiëntie is overfitting gedefinieerd als het kiezen van een model met meer variabelen dan het approximating model. Underfitting is dan het kiezen van een model met te weinig variabelen vergeleken met het approximating model.

4.3. Signal-to-noise ratio

Bij het bepalen welk model je moet kiezen, heb je een maat nodig. Standaard is het nemen van een maat die de afstand tot het operating model bepaalt. Het model met de kleinste waarde vind je dan het beste. Je vergelijkt modellen dus door naar de verschillen in uitkomsten van de criteriumfunctie te kijken. Zij MSC het modelselectiecriterium dat je gebruikt. Als je twee modellen wilt vergelijken, één met k parameters en één met $k + L$, dan kies je voor het model met k parameters als $\mathrm{MSC}_k < \mathrm{MSC}_{k+L}$. Definieer nu het **signal** als $\mathbb{E}[\mathrm{MSC}_{k+L} - \mathrm{MSC}_k]$ (de verwachtingswaarde ten opzichte van de dichtheidsfunctie) en de **noise** als $\mathrm{sd}[\mathrm{MSC}_{k+L} - \mathrm{MSC}_k]$, de standaarddeviatie van het verschil. Dan wordt de **signal-to-noise** ratio gedefinieerd als

$$\frac{\mathbb{E}[\mathrm{MSC}_{k+L} - \mathrm{MSC}_k]}{\mathrm{sd}[\mathrm{MSC}_{k+L} - \mathrm{MSC}_k]}.$$

In [10] wordt deze verhouding gebruikt om criteria te vergelijken. Het signal hangt vooral af van de functie die te veel parameters straft (**penaltyfunctie** of **straffunctie**) en de noise hangt vooral af van de verdeling van de SSE en de verdeling van de verschillen in de SSE. Een kleine ratio is een indicatie van overfitting en een te grote ratio geeft mogelijke underfitting aan.

4.4. Vergelijking van criteria

In [10] is een heel hoofdstuk gewijd aan het vergelijken van modelselectiecriteria voor onder andere regressieanalyse. De auteurs hebben 16 criteria bekeken en die in een aantal verschillende situaties getest. Daarbij hebben ze verschillende modellen verkregen door de volgende parameters te variëren: het aantal observaties n, de fout in de variantie σ_*^2, de parameterstructuur

β_j, de orde van het operating model k_*, het niveau van overfitting o en de correlatie tussen de kolommen van X, ρ_x.

Om de resultaten te vergelijken gebruikt men **observed efficiency**. Deze geobserveerde efficiëntie is de afstand tussen het beste approximating model en het operating model. Er zijn twee maten gebruikt om deze afstand te meten, namelijk de Kullback-Leiblerdiscrepantie (K-L) en de L_2-norm. Als K-L gebruikt wordt, blijkt underfitting veel zwaarder gestraft te worden dan overfitting, terwijl L_2 het tegenovergestelde doet en overfitting zwaarder straft. Omdat een goed criterium niet te veel overfitting en underfitting mag hebben, zou het in beide situaties (K-L en L_2) goed moeten presteren. Daarom wordt aan elk criterium een rangnummer toegekend op basis van de K-L geobserveerde efficiëntie en een apart rangnummer op basis van L_2.

4.4.1. Twee speciale gevallen

In paragraaf 9.2.2 van [10] worden twee (speciale) lineaire regressiemodellen besproken. Het eerste model is

$$y_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + \varepsilon_{*i},$$

en het tweede model

$$y_i = 1 + x_{i1} + \frac{1}{2}x_{i2} + \frac{1}{3}x_{i3} + \frac{1}{4}x_{i4} + \frac{1}{5}x_{i5} + \varepsilon_{*i}.$$

Voor beide modellen geldt: $n = 25$, $\sigma_*^2 = 1$, $k_* = 6$, $\rho_* = 0$.

De auteurs van [10] hebben voor deze modellen en de 16 criteria veel simulaties uitgevoerd. We zullen alleen de resultaten behandelen. De tabellen met de cijfers en meer details zijn in [10] te vinden.

Als eerste zijn simulaties gedaan waarbij voor elk van de 16 criteria werd gekeken hoe ze presteerden in het vinden van het juiste model van de vorm van model 1. De verzameling van approximating models bestaat dus uit lineaire modellen waarvan de coëfficiënten allemaal gelijk aan 1 zijn, maar waarvan het aantal verklarende variabelen af kan wijken van de 6 verklarende variabelen (inclusief de intercept) van het operating model. Voor elk criterium zijn 10.000 simulaties gedaan. Daarvan werd onder andere het aantal keer dat het operating model werd aangewezen geteld, net als het aantal keer dat underfitting en overfitting voorkwam. Rangnummers zijn toegekend op basis van de K-L geobserveerde efficiëntie en apart de L_2-norm. De vijf beste criteria die daar uitrollen, zijn:

$$\begin{aligned}
\text{AICu} &= \log s_k^2 + \frac{n+k}{n-k-2}, \\
\text{HQc} &= \log \hat{\sigma}_k^2 + \frac{2\log\log(n)k}{n-k-2}, \\
\text{AICc} &= \log \hat{\sigma}_k^2 + \frac{n+k}{n-k-2}, \\
\text{GM} &= \frac{\text{SSE}}{s_K^2} + \log(n)k \qquad (K = k_* + o),
\end{aligned}$$

$$\text{FPEu} = \frac{n+k}{n-k} s_k^2.$$

Hierbij is de eerste term van de eerste drie criteria gelijk aan de eerste term in het AIC, dus de logaritme van de likelihoodfunctie.

Voor model 2 is hetzelfde gedaan; wederom 10.000 simulaties en de bijbehorende tellingen en geobserveerde efficiënties. Op basis van de rankings is weer een top vijf samen te stellen. Deze is op vier plaatsten hetzelfde als voor model 1: HQc, AICu, AICc en GM. FPEu zit nu niet in de top vijf, maar wel het criterium

$$\text{DCVB} = \frac{1}{Rn} \sum_{r=1}^{R} \sum_{i=1}^{n} \frac{v_{ir}^2}{(1-h_i)^2}.$$

Het DCVB (doubly cross-validated bootstrap) is een bootstrapcriterium, dat we hier niet verder zullen behandelen (zie [10]). Wel kan nog opgemerkt worden dat bij de simulaties voor model 2 elk criterium maar in (minder dan) 2% van de gevallen het operating model selecteerde. Voor model 1 varieerde dit van 12% tot 65%, en bovendien zit meer dan de helft van de criteria (rond of) boven de 40%. Ook presteren de bootstrapcriteria beter dan de crossvalidationcriteria.

4.4.2. Variatie van parameters

In een tweede simulatie worden modellen van de volgende vorm bekeken:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{k_*-1} x_{i,k_*-1} + \varepsilon_{*i}$$

met $\varepsilon_{*i} \sim \mathcal{N}(0, \sigma_*^2)$, de ε_{*i} onafhankelijk en $i = 1, \ldots, n$. Er worden 6 parameters gevarieerd, zie tabel 4.1, wat in totaal 540 verschillende modellen oplevert. Voor elk model wordt 100 keer gesimuleerd.

	keuze 1	keuze 2	keuze 3	keuze 4	keuze 5
steekproefgrootte n	15	25	35	50	100
fout in variantie σ_*^2	$0,1$	1	10		
parameterstructuur β_j	$1/j^2$	$1/j$	1		
echte orde k_*	3	6			
overfitting o	2	5			
ρ_x	0	$0,4$	$0,9$		

Tabel 4.1: Variatie in parameters

Bij de parameterstructuur betekent de $1/j^2$ dat de coëfficiënt β_j van variabele j gelijk is aan $1/j^2$. Omdat zoveel verschillende modellen bekeken worden, zijn deze simulaties geschikt voor het bestuderen van de criteria onder verschillende condities.

Voor elk model worden, zoals gezegd, 100 simulaties gedaan. Elk criterium selecteert elke keer een van de modellen en dan kan de geobserveerde efficiëntie van dat criterium berekend worden. Deze wordt vergeleken met de waarden van de andere criteria en de ranglijst kan dan gemaakt worden. Het criterium met de hoogste efficiëntie krijg rang 1. Als alle waarden voor alle criteria bekend zijn, dan wordt het gemiddelde rangnummer over alle simulaties voor elk criterium

bepaald. De beste vijf criteria op basis van deze simulaties zijn AICu, GM, HQc, FPEu, SIC en FPE4, waar FPE4 $= \hat{\sigma}^2 \frac{n+3k}{n-k}$. Het SIC en FPE4 eindigen op een gedeelde vijfde plek.

Omdat het onderscheid tussen consistentie en efficiëntie minder belangrijk is (dit blijkt uit de resultaten) dan de signal-to-noise ratio's staat het DCVB in de bovenste helft. Het Cp presteert hier niet zo goed, maar is wel heel goed in het selecteren van modellen met een andere (belangrijke) eigenschap. Namelijk het doen van voorspellingen voor toekomstige waarnemingen. Deze modellen liggen dan niet het dichtst bij het operating model, maar hebben een belangrijke waarde bij het doen van voorspellingen.

In het algemeen wordt geconcludeerd dat efficiënte criteria met een zwakke straffunctie (zoals het AIC) slecht presteren als kleine steekproeven bekeken worden. Dit komt door de overfitting van deze criteria. Omdat consistente criteria grotere straffuncties hebben dan efficiënte criteria presteren zij beter, zelfs met kleine steekproeven. AICu, HQc en FPEu hebben grote straffuncties, en we zagen al dat zij een heel goed resultaat hebben behaald voor de 540 modellen.

Belangrijk is hier op te merken dat de resultaten gebaseerd zijn op vele modellen en realisaties. Als een speciaal model gekozen is, dan is het goed mogelijk dat een criterium uit de top 5 voor dat model heel slecht presteert. Andersom kan ook een criterium dat onderaan staat in de ranglijst voor sommige situaties het beste criterium zijn. Daarom kan het een goed idee zijn om voor een criterium zowel de K-L geobserveerde efficiëntie te bekijken als die van de L_2. Als het geselecteerde model op basis van K-L afwijkt van dat op basis van L_2, dan moet goed naar de approximating models gekeken worden of daar niet iets te verbeteren/toe te voegen is.

Door parameters te variëren zijn 540 verschillende modellen gemaakt. Hierin zijn wat algemene trends te ontdekken. Als de steekproefgrootte n toeneemt, dan neemt de geobserveerde efficiëntie ook toe. Als σ_*^2 toeneemt dan neemt de geobserveerde efficiëntie af. Hetzelfde geldt als de correlatie tussen de kolommen van X, de orde van het operating model (k_*) of de overfitting (het aantal nutteloze parameters o) toeneemt. In al deze gevallen daalt de geobserveerde efficiëntie. De auteurs van [10] concluderen hieruit dat je makkelijker kunt werken met experimenten met weinig variabelen. Voor gecompliceerde experimenten heb je (logisch) veel werk.

4.4.3. Simulaties van grote steekproeven

We bekijken in deze paragraaf twee modellen, $A1$ en $A2$. Voor beide modellen geldt: $n = 25.000$, $k_* = 2$, $\beta_0 = 1$, $\beta_1 = 1$ en $\sigma_* = 1$. Het verschil tussen de modellen zit in de mogelijkheid voor overfitting. Bij model $A1$ geldt $o = 2$ en bij model $A2$ is $o = 5$. Omdat we nu een grote steekproef bekijken, moeten we asymptotische relaties kunnen afleiden. We zullen zien dat veel criteria, en in het bijzonder de efficiënte criteria, asymptotisch equivalent zijn.

Al eerder zagen we dat de belangrijkste aanname bij consistentie is dat het operating model in de verzameling van approximating models zit. Maar als dit het geval is, hoe presteren de efficiënte criteria dan? Omdat de modellen $A1$ en $A2$ een paar van de slechtste scenario's voor efficiëntie representeren, wordt dit in paragraaf 9.2.4 van [10] onderzocht. Het resultaat: efficiënte criteria zijn dan niet langer efficiënt. Zeker niet als het operating model van eindige orde k_* is. Daarnaast daalt de geobserveerde efficiëntie als de orde van het operating model daalt en als het operating model van eindige orde bij de approximating models zit, dan zijn de consistente criteria zowel consistent als efficiënt. Verder kan overfitting een willekeurig groot probleem worden omdat voor elke irrelevante variabele geldt dat hij met 15% kans wordt gekozen.

We bekijken weer de 16 criteria en voeren per criterium 1.000 simulaties uit. Voor model $A1$ volgt daaruit dat de efficiënte criteria (onder andere AIC, AICc, Cp en FPE) ongeveer

hetzelfde presteren en in meer dan 30% van de gevallen overfitten. De consistente criteria SIC en GM presteren zoals verwacht en selecteren bijna altijd het operating model. De criteria HQ ($= \log \hat{\sigma}_k^2 + \frac{\log(n)k}{n}$) en HQc zijn ook consistent, maar hebben een veel kleinere straffunctie ten opzichte van het SIC en hebben dus vaker een overfit. Desondanks selecteren ze in 92% van de gevallen het goede model. Geen van de criteria heeft ooit in die 1.000 simulties een model met te weinig parameters geselecteerd, dus underfitting komt niet voor.

Bij model $A2$ is de mogelijkheid voor overfitting groter, en dit is dan ook gebeurd. Maar als het operating model bij de approximating models zit en de steekproefgrootte groot is, dan hebben de irrelevante variabelen geen invloed op de consistente criteria. GM en SIC selecteren even vaak het operating model als in het geval van model $A2$. De criteria HQ en HQc hebben iets meer hinder van de kans op overfitting; zij selecteren twee keer zo vaak een model met te veel parameters ten opzichte van het aantal bij $A1$. Bij de efficiënte criteria is het verschil nog beter te zien. Nu komt overfitting in 50% van de gevallen voor, in plaats van 30%. Ook bij model $A2$ heeft geen van de criteria underfitting geconstateerd.

4.4.4. Echte data

Om de criteria in de praktijk te testen gebruikt [10] data uit andere bronnen, namelijk [19] en een ongepubliceerde masterscriptie van Carl Hoffstedt. De data gaan over grote snelwegen in Minnesota. In 1973 zijn 39 segmenten van die snelwegen geselecteerd en geobserveerd. Het doel was het aantal ongelukken per miljoen gereden mijl door de voertuigen te modelleren door 13 onafhankelijke variabelen. Daarbij zijn alle deelverzamelingen van de 13 variabelen bekeken.

Op twee na kozen alle criteria voor een van de volgende modellen:

Model 1 met 5 verklarende variabelen:
- de lengte van het bestudeerde segment van de snelweg,
- de maximale snelheid,
- het aantal wisselingen van verkeersteken per mijl van het segment,
- het aantal opritten per mijl van het segment
- of het een hoofdverkeersader (principal arterial highway) betreft.

Model 2 met 3 verklarende variabelen:
- de lengte van het bestudeerde segment van de snelweg,
- de maximale snelheid
- het aantal opritten per mijl van het segment.

De efficiënte criteria (zoals AIC, AICc, Cp en FPE) selecteren vooral model 1. Ook HQ selecteert model 1, wat ondersteunt dat HQ zich voor kleine steekproeven ongeveer hetzelfde gedraagt als het AIC. Criteria met grotere straffuncties kiezen vooral model 2, dat dan ook twee variabelen minder heeft. Omdat beide modellen ongeveer dezelfde karakteristieken van de residuen laten zien, is er geen reden het ene model boven het andere te verkiezen.

Model 2 is genest in model 1 en dus kunnen we een F-toets gebruiken om verder onderscheid te maken. In die toets wordt model 2 vergeleken met model 1. We toetsen daarbij welk model het meest waarschijnlijk is, en nemen aan dat dit model 2 is. Uit een F-toets komt een F-waarde (in dit geval 2.86) met een bijbehorende p-waarde (in dit geval 0.0715). Als de p-waarden onder de grens van 0.05 liggen, dan vinden we de aanname fout. Ligt de p-waarde erboven, dan vinden we de aanname goed (er is dan slechts een kans van maximaal 5% dat de aanname toch fout is, en model 1 dus het goede model is). Daaruit kunnen we concluderen dat model 2 beter is.

4.5. Conclusie

Vanwege de superlineaire straffuncies[2] van het AICc en AICu presteren deze criteria goed in de onderzoeken in [10]. Ze overfitten in het algemeen niet veel, in tegenstelling tot het AIC dat een lineaire strafffunctie heeft met betrekking tot de orde van het model. Daarom heeft het AIC veel overfitting bij kleine steekproeven. Hetzelfde geldt voor HQ; bij kleine simulatie doet de aangepaste versie HQc het veel beter.

Sommige klassieke criteria zoals het AIC, FPE en Cp presteerden slecht en worden door [10] afgeraden. Andere presteerden over het algemeen goed. Zo doen HQc, AICu en AICc het goed in het geval dat het operating model bij de approximating models zit, en ook als dat niet het geval is. Aangezien deze drie criteria alledrie verschillende asymptotische eigenschappen hebben, is het een goed idee de modellen die elk criterium kiest te vergelijken. Dit kan tot meer inzicht in het probleem leiden. Bovendien is het gebruiken van verschillende criteria een bredere aanpak in het kiezen van een model dan wanneer je je baseert op enkel één criterium.

Tot slot nog een opmerking over een eigenschap van het AIC. In [21] vinden we dat het BIC consistent en het AIC **minimax-rate optimaal** is voor het schatten van de regressiefunctie. Dit wil zeggen dat het AIC dat model kiest waarvoor het maximale risico (op fouten) minimaal is ten opzichte van de andere modellen. Als je weer aanneemt dat het operating model oneindig dimensionaal is, dan is minimax-rate optimaliteit een eigenschap die je belangrijk vindt bij regressie. Dit is nog een reden om bij regressieanalyse voor het AIC te kiezen.

[2]Superlineair in k betekent dat de eerste afgeleide positief is en dat de functie onbegrensd is voor een zekere $k \leq n$.

Hoofdstuk 5

Simuleren

Een manier om te kijken hoe goed het AIC en het BIC werken, is door middel van simulaties. Met een computerprogramma kun je een steekproef simuleren van een stochast waarvan de verdelingsfunctie je bekend is. Daarna kun je kijken of beide criteria ook daadwerkelijk die verdelingsfunctie aanwijzen als het beste model. In dit hoofdstuk zullen we een specifiek probleem behandelen, namelijk het herhaaldelijk opgooien van een munt. Allereerst geven we enige uitleg over het gebruikte computerprogramma. Daarna zullen we de resultaten van de simulatie bespreken.

5.1. Achtergrond

Voor het doen van de simulaties hebben we gebruikgemaakt van het computerprogramma R. Dit programma, dat vooral gebruikt wordt door statistici, is gratis verkrijgbaar voor een verscheidenheid van besturingssystemen. Het bevat een groot aantal voorgeprogrammeerde functies die nuttig zijn bij het bedrijven van statistiek en het simuleren van situaties. Zie [23] voor meer informatie.

In het vervolg van dit hoofdstuk is de code opgenomen die we gebruikt hebben bij het doen van de simulaties. Gelukkig laat de programmeertaal R zich eenvoudig lezen. Zo hier en daar hebben we extra uitleg over de syntax toegevoegd, maar vaak was dit (mede dankzij het bijgevoegde commentaar) niet eens nodig.

Belangrijk is verder nog te weten dat R een zogeheten imperatieve programmeertaal is. Dit betekent dat we direct opdrachten kunnen invoeren (achter een opdrachtprompt, `>`) en dat het resultaat meteen verschijnt zodra we op Enter drukken. Ook deze uitvoer is in dit hoofdstuk vermeld.

Het programma R werkt met vectoren. Uitkomsten van berekeningen waar getallen het resultaat zijn, worden daarom voorafgegaan door `[1]`. Dit geeft aan dat het eerstvolgende getal het eerste in de vector is. Getallen zijn simpelweg vectoren ter lengte één.

5.2. Probleem: herhaaldelijk opgooien van een munt

Stel dat we graag willen weten welk model beter van toepassing is op de verdeling van de stochast K, waar K het aantal keer kop bij het N keer opgooien van een munt is: een binomiale verdeling

of een normale verdeling. We kunnen dan het N keer opgooien van een munt vele keren (zeg n) simuleren en daarna kijken welk model beide criteria beter vinden.

Na de simulatie kunnen we vaststellen hoeveel keer k er kop is gegooid. Daarna vullen we simpelweg de formules voor beide criteria in. Dat gaat niet meteen, want om $\mathcal{L}(\mathbf{k} \mid g_{\hat{\boldsymbol{\theta}}})$ te kunnen berekenen, moeten we de likelihoodfunctie en de maximum-likelihoodschatters kennen. Het uitrekenen hiervan is echter snel gedaan.

We beschouwen zoals gezegd twee mogelijke modellen. Het eerste model is de binomiale verdeling met parameters N en $\hat{p}$. Het tweede model is de normale verdeling met parameters $\hat{\mu} \in \mathbb{R}$ en $\hat{\sigma}^2 \in \mathbb{R}$ (met $\hat{\sigma} \geq 0$). Hier zijn $\hat{p}$, $\hat{\mu}$ en $\hat{\sigma}^2$ de maximum-likelihoodschatters van p, μ en σ^2.

5.2.1. Werkelijke situatie

We beginnen met twee definities en een stelling uit [11]. Daarmee zullen we de werkelijke situatie kunnen analyseren. Het bewijs van de stelling is afkomstig uit [12].

Definitie 5.1. Een stochast X heet Bernoulli verdeeld met parameter $\theta \in (0, 1)$ als de stochast alleen de waarden 0 en 1 kan aannemen, en $\mathbb{P}(X = 1) = \theta$.

Definitie 5.2. Een stochast X heet binomiaal verdeeld met parameters $n \in \mathbb{N} \setminus \{0\}$ en $\theta \in [0, 1]$ als voor alle $k \in \mathbb{N}$ geldt dat $\mathbb{P}(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$.

Stelling 5.3. *Zij $X_1, \ldots, X_n$ een steekproef uit een Bernoulli verdeelde populatie met parameter θ. Dan is de stochast $S := \sum_{i=1}^{n} X_i$ binomiaal verdeeld met parameters n en θ.*

Bewijs. Kies $k \in \mathbb{N}$ willekeurig. Er geldt dat

$$\mathbb{P}(S = k) = \mathbb{P}(X_1 + \cdots + X_n = k) = \sum_{\boldsymbol{\alpha} \in A_k} \mathbb{P}(\mathbf{X} = \boldsymbol{\alpha}),$$

waar $\mathbf{X} := (X_1, \ldots, X_n)$ en $A_k := \{\boldsymbol{\alpha} \in \{0,1\}^n : |\boldsymbol{\alpha}| = k\}$ met $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_n$. We sommeren over alle $\boldsymbol{\alpha} \in A_k$, dus geldt dat $|\boldsymbol{\alpha}| = k$. Dan volgt dat

$$\begin{aligned}
\mathbb{P}(\mathbf{X} = \boldsymbol{\alpha}) &= \mathbb{P}(X_1 = \alpha_1, \ldots, X_n = \alpha_n) \\
&= \mathbb{P}(X_1 = \alpha_1) \cdots \mathbb{P}(X_n = \alpha_n) \\
&= \theta^{\alpha_1}(1-\theta)^{1-\alpha_1} \cdots \theta^{\alpha_n}(1-\theta)^{1-\alpha_n} \\
&= \theta^{|\boldsymbol{\alpha}|}(1-\theta)^{n-|\boldsymbol{\alpha}|} \\
&= \theta^k(1-\theta)^{n-k}.
\end{aligned}$$

Dan volgt dat

$$\begin{aligned}
\mathbb{P}(S = k) &= \sum_{\boldsymbol{\alpha} \in A_k} \theta^k(1-\theta)^{n-k} \\
&= \#(A_k) \times \theta^k(1-\theta)^{n-k} \\
&= \binom{n}{k}\theta^k(1-\theta)^{n-k}.
\end{aligned}$$

Maar dan is S per definitie 5.2 binomiaal verdeeld met parameters n en θ. □

Als we N keer een munt opgooien, dan kunnen we iedere gooi zien als een Bernoulli verdeelde stochast K_i met parameter p, waar p de kans is dat je kop gooit. De som $K := \sum_{i=1}^{N} K_i$ is dan precies het aantal successen, ofwel het aantal keer kop k. Met Stelling 5.3 volgt dan dat de stochast K binomiaal verdeeld is met parameters N en p.

Het werkelijke model is nu bekend. De vraag is: wijzen het AIC en het BIC ook dat model aan?

5.2.2. Likelihoodfunctie en maximum-likelihoodschatter voor het binomiale model

De likelihoodfunctie wordt voor de binomiale verdeling gegeven door [8]

$$\begin{aligned}\mathcal{L}(\mathbf{k} \mid g_{\boldsymbol{\theta}}) &= \prod_{i=1}^{n} g_{\boldsymbol{\theta}}(k_i) \\ &= \prod_{i=1}^{n} \binom{N}{k_i} p^{k_i}(1-p)^{N-k_i} \\ &= \binom{N}{k_1} \cdots \binom{N}{k_n} p^{k_1+\cdots+k_n}(1-p)^{(N-k_1)+\cdots+(N-k_n)}, \end{aligned} \tag{5.1}$$

waar N zoals boven het aantal keer werpen in één experiment is, n het aantal experimenten, $\mathbf{k}$ de vector met steekproefgegevens en k_i het aantal keer kop bij het i-de experiment (dus de i-de component van onze vector). Verder kent ons model slechts één vrije parameter, namelijk de kans op een succes p, dus $\boldsymbol{\theta} = (\ p\ )$. In het AIC wordt de maximum-likelihoodschatter $\hat{\boldsymbol{\theta}}$ van $\boldsymbol{\theta}$ ingevuld. We willen dus weten wat deze maximum-likelihoodschatter van $\boldsymbol{\theta}$ (en dus van p) is.

We weten dat $\hat{p}$ de waarde van p is die optimaal is, dat wil zeggen, waarvoor de likelihoodfunctie maximaal is. We kunnen deze waarde vinden door simpelweg (5.1) te differentiëren naar p, de resulterende uitdrukking nul te stellen en op te lossen naar p. Dit is in dit geval een eenvoudige zaak:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}p}\mathcal{L}(\mathbf{k} \mid g_p) &= \frac{\mathrm{d}}{\mathrm{d}p}\left[\binom{N}{k_1} \cdots \binom{N}{k_n} p^{k_1+\cdots+k_n}(1-p)^{(N-k_1)+\cdots+(N-k_n)}\right] \\ &= \binom{N}{k_1} \cdots \binom{N}{k_n}\Big[(k_1+\cdots+k_n)p^{(k_1+\cdots+k_n)-1} \cdot (1-p)^{(N-k_1)+\cdots+(N-k_n)} \\ &\qquad - p^{k_1+\cdots+k_n} \cdot ((N-k_1)+\cdots+(N-k_n))(1-p)^{(N-k_1)+\cdots+(N-k_n)-1}\Big].\end{aligned}$$

De bovenstaande afgeleide is gevonden na toepassen van de productregel; merk op dat bij de afgeleide van de tweede factor ook de kettingregel is gebruikt voor het nemen van de afgeleide van de binnenfunctie. Nul stellen geeft nu

$$\begin{aligned}0 = \binom{N}{k_1} \cdots \binom{N}{k_n} \cdot p^{(k_1+\cdots+k_n)-1} \cdot (1-p)^{(N-k_1)+\cdots+(N-k_n)-1} \times \\ \times \Big[(k_1+\cdots+k_n)(1-p) - p((N-k_1)+\cdots+(N-k_n))\Big].\end{aligned}$$

Omdat het product van twee factoren alleen nul is als minstens één van de factoren nul is en de eerste factor geen nul kan zijn (want $0 \neq p \neq 1$), geldt

$$\begin{aligned} 0 &= (k_1 + \cdots + k_n)(1-p) - p((N-k_1) + \cdots + (N-k_n)) \\ &= k_1 + \cdots + k_n - p(k_1 + \cdots + k_n) - p((N-k_1) + \cdots + (N-k_n)). \end{aligned}$$

Hieruit volgt dat

$$\begin{aligned} k_1 + \cdots + k_n &= p(k_1 + \cdots + k_n) + p((N-k_1) + \cdots + (N-k_n)) \\ &= p(k_1 + \cdots + k_n) + p(N + \cdots + N) - p(k_1 + \cdots + k_n), \end{aligned}$$

waaruit weer volgt dat

$$\begin{aligned} p &= \left(\textstyle\sum_{i=1}^n k_i\right) / \left(\textstyle\sum_{i=1}^n N\right) \\ &= \left(\textstyle\sum_{i=1}^n k_i\right) / (n \times N) \\ &= \overline{\mathbf{k}}/N, \end{aligned}$$

waar $\overline{\mathbf{k}}$ het gemiddelde van alle waarden in de vector $\mathbf{k}$ voorstelt. Dus $\hat{p} = \overline{\mathbf{k}}/N$.

5.2.3. Likelihoodfunctie en maximum-likelihoodschatters voor het normale model

De likelihoodfunctie wordt voor de normale verdeling gegeven door [8]

$$\begin{aligned} \mathcal{L}(\mathbf{k} \mid g_{\boldsymbol{\theta}}) &= \prod_{i=1}^n g_{\boldsymbol{\theta}}(k_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \mathrm{e}^{-\frac{(k_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \mathrm{e}^{-\sum_{i=1}^n \frac{(k_i-\mu)^2}{2\sigma^2}}, \end{aligned} \tag{5.2}$$

waar n zoals boven het aantal experimenten is, $\mathbf{k}$ de vector met steekproefgegevens en k_i het aantal keer kop bij het i-de experiment (dus de i-de component van onze vector). Dit model kent twee vrije parameters, namelijk de verwachtingswaarde μ en de variantie σ^2. De variantie is het kwadraat van de standaarddeviatie σ. Hier is dus $\boldsymbol{\theta} = (\mu,\ \sigma^2)$. We berekenen weer de maximum-likelihoodschatter van $\boldsymbol{\theta}$ (en dus van μ en σ^2). Dit gaat weer door (5.2) naar respectievelijk μ en σ af te leiden, de resulterende uitdrukkingen nul te stellen en op te lossen naar respectievelijk μ en σ^2.

We bepalen eerst de maximum-likelihoodschatter $\hat{\mu}$ van μ.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mu}\mathcal{L}(\mathbf{k} \mid g_{\mu,\sigma^2}) &= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \mathrm{e}^{-\sum_{i=1}^n \frac{(k_i-\mu)^2}{2\sigma^2}}\right] \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n}\ \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\mathrm{e}^{-\sum_{i=1}^n \frac{(k_i-\mu)^2}{2\sigma^2}}\right] \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n}\ \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (k_i-\mu)^2}\right] \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \left[-\frac{1}{2\sigma^2}\left(2\sum_{i=1}^n (\mu - k_i)\right) \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (k_i-\mu)^2}\right]. \end{aligned}$$

Nul stellen geeft nu

$$0 = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \left[-\frac{1}{2\sigma^2} \left(2 \sum_{i=1}^{n} (\mu - k_i) \right) \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right].$$

Wederom geldt dat het product van twee factoren nul is als minstens één van de factoren nul is. Als we N keer met een munt gooien, dan krijgen we niet altijd hetzelfde aantal kop. Er geldt dus $\sigma > 0$. Maar dan is de factor voor de term tussen de rechte haken niet nul. Dus moet wel

$$0 = -\frac{1}{2\sigma^2} \left(2 \sum_{i=1}^{n} (\mu - k_i) \right) \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2}.$$

Met dezelfde argumentatie kunnen we meer factoren wegstrepen, waarna volgt dat

$$0 = \sum_{i=1}^{n} (\mu - k_i).$$

Hieruit volgt dat

$$n\mu = \sum_{i=1}^{n} k_i,$$

waaruit weer volgt dat

$$\mu = \frac{1}{n} \sum_{i=1}^{n} k_i = \overline{\mathbf{k}},$$

waar $\overline{\mathbf{k}}$ het gemiddelde van alle waarden in de vector $\mathbf{k}$ voorstelt. Dus $\hat{\mu} = \overline{\mathbf{k}}$.

We bepalen nu de maximum-likelihoodschatter $\hat{\sigma}^2$ van σ^2.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma} \mathcal{L}(\mathbf{k} \mid g_{\mu,\sigma^2}) &= \frac{\mathrm{d}}{\mathrm{d}\sigma} \left[\frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \mathrm{e}^{-\sum_{i=1}^{n} \frac{(k_i - \mu)^2}{2\sigma^2}} \right] \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\mathrm{d}}{\mathrm{d}\sigma} \left[\frac{1}{\sigma^n} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right] \\
&= \frac{1}{(2\pi)^{n/2}} \left(\frac{1}{\sigma^n} \frac{\mathrm{d}}{\mathrm{d}\sigma} \left[\mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right] + -n\sigma^{-n-1} \left(\mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right) \right) \\
&= \frac{1}{(2\pi)^{n/2}} \left(\frac{1}{\sigma^n} \left(-\frac{1}{2} \sum_{i=1}^{n} (k_i - \mu)^2 \cdot -2\sigma^{-3} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right) \right. \\
&\qquad \left. - n\sigma^{-n-1} \left(\mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right) \right).
\end{aligned}$$

Nul stellen geeft nu

$$\begin{aligned}
0 &= \frac{1}{(2\pi)^{n/2}} \left(\frac{1}{\sigma^n} \left(\sum_{i=1}^{n} (k_i - \mu)^2 \ \sigma^{-3} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right) - n\sigma^{-n-1} \left(\mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \right) \right) \\
&= \frac{1}{(2\pi)^{n/2}} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (k_i - \mu)^2} \left(\frac{1}{\sigma^n} \left(\sum_{i=1}^{n} (k_i - \mu)^2 \ \sigma^{-3} \right) - n\sigma^{-n-1} \right).
\end{aligned}$$

We kunnen weer enkele factoren wegstrepen die nooit nul worden.

$$
\begin{aligned}
0 &= \frac{1}{\sigma^n}\left(\frac{1}{\sigma^3}\sum_{i=1}^{n}(k_i-\mu)^2\right) - n\frac{1}{\sigma^{n+1}},\\
0 &= \frac{1}{\sigma^{n+1}}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(k_i-\mu)^2 - n\right),\\
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(k_i-\mu)^2 - n,\\
0 &= \sum_{i=1}^{n}(k_i-\mu)^2 - n\sigma^2.
\end{aligned}
$$

Daaruit volgt nu

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(k_i-\mu)^2.$$

Dus $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(k_i-\mu)^2$ met $\mu = \hat{\mu} = \overline{\mathbf{k}}$.

5.2.4. Resultaten van de simulatie

We gaan deze simulatie uitvoeren met R. Allereerst hebben we code nodig die ons gesimuleerde gegevens levert. Hiervoor hebben we de functie *ExperimentUitvoeren* geschreven. Deze functie verlangt twee argumenten, te weten n en N, en geeft als uitvoer een vector ter lengte n met daarin het aantal keer kop k_i (met $0 \leq k_i \leq N$) bij experiment i (met $1 \leq i \leq n$).

```
ExperimentUitvoeren <- function(n, N)
{
  # We maken een nulvector ter lengte n om de resultaten in bij te houden
  resultaten <- numeric(n);

  # We doen n experimenten waarin we N munten opgooien
  # runif(1) geeft een willekeurig getal tussen 0 en 1
  # Als het kleiner is dan 0.5 beschouwen we het als kop
  for (j in (1:n))
    for (k in (1:N))
      if (runif(1) < 0.5)
        resultaten[j] <- resultaten[j]+1;

  # En ten slotte geven we de vector terug
  resultaten;
}
```

We weten dat geldt $\mathrm{AIC} = -2\log\mathcal{L}(\mathbf{k} \mid g_{\hat{\boldsymbol{\theta}}}) + 2d$ en $\mathrm{BIC} = -2\log\mathcal{L}(\mathbf{k} \mid g_{\hat{\boldsymbol{\theta}}}) + d\log n$, waar d het aantal vrije parameters is. In het binomiale model is dus $d = 1$, namelijk p, en in het normale model is $d = 2$, namelijk μ en σ^2.

Omdat in de formules voor beide criteria de likelihoodfunctie voorkomt, is het handig extra functies te definiëren die de waarde hiervan berekenen. De functie zal afhangen van de maximum-likelihoodschatters. Die geven we daarom als argumenten mee. De functies die we zo krijgen, zijn de volgende.

```
LikelihoodBinomModel <- function(gegevens, p, n, N)
{
  termen <- numeric(n);

  for (i in (1:n))
    termen[i] <- choose(N, gegevens[i])*
                                  (p^(gegevens[i]))*((1-p)^(N-gegevens[i]));

  prod(termen);
}

LikelihoodNormModel <- function(gegevens, mu, sigmakw, n, N)
{
  termen <- numeric(n);
  constante <- 1/(sqrt(2*pi*sigmakw));

  for (i in (1:n))
    termen[i] <- constante*exp( - (((gegevens[i]-mu)^2)/(2*sigmakw)) );

  prod(termen);
}
```

Tot slot kunnen we dan de functies schrijven die de waarde van het AIC en het BIC voor het binomiale en het normale model berekenen. De waarde van de maximum-likelihoodschatters hebben we al gevonden. Die zetten we in de functie zelf. We krijgen zo de volgende functies.

```
AICBinomModel <- function(gegevens, n, N)
{
  # De maximum-likelihoodschatter is bekend:
  p <- mean(gegevens)/(N);

  # En de hoeveelheid vrije parameters ook:
  d <- 1;

  # Nu kunnen we de likelihoodfunctie invullen:
  likelihood <- LikelihoodBinomModel(gegevens, p, n, N);

  # En het AIC bepalen:
  AIC <- -2*log(likelihood)+2*d;
  AIC;
}
```

```
AICNormModel <- function(gegevens, n, N)
{
  # De maximum-likelihoodschatters zijn bekend:
  mu <- mean(gegevens);

  som <- 0;
  for (i in (1:n))
    som <- som + (gegevens[i] - mu)^2;
  sigmakw <- (1/n)*som;

  # En de hoeveelheid vrije parameters ook:
  d <- 2;

  # Nu kunnen we de likelihoodfunctie invullen:
  likelihood <- LikelihoodNormModel(gegevens, mu, sigmakw, n, N);

  # En het AIC bepalen:
  AIC <- -2*log(likelihood)+2*d;
  AIC;
}


BICBinomModel <- function(gegevens, n, N)           # Code vrijwel identiek
{                                                   # aan AICBinomModel
  p <- mean(gegevens)/(N);
  d <- 1;

  likelihood <- LikelihoodBinomModel(gegevens, p, n, N);

  BIC <- -2*log(likelihood)+d*log(n);
  BIC;
}


BICNormModel <- function(gegevens, n, N)            # Code vrijwel identiek
{                                                   # aan AICNormModel
  mu <- mean(gegevens);

  som <- 0;
  for (i in (1:n))
    som <- som + (gegevens[i] - mu)^2;
  sigmakw <- (1/n)*som;

  d <- 2;

  likelihood <- LikelihoodNormModel(gegevens, mu, sigmakw, n, N);
  BIC <- -2*log(likelihood)+d*log(n);
  BIC;
}
```

Nu we alle benodigde functies hebben gedefinieerd, kunnen we de simulaties gaan uitvoeren. We maken eerst een steekproef met $n = 10$ en $N = 50$ en slaan het resultaat op in de variabele `sample`. Daarna berekenen we de waarde van het AIC en het BIC. Dat geeft de volgende uitvoer.

```
> sample <- ExperimentUitvoeren(10,50);
> sample;
 [1] 26 20 32 19 25 31 29 22 26 26
> AICBinomModel(sample,10,50);
[1] 59.21597
> AICNormModel(sample,10,50);
[1] 60.7344
> BICBinomModel(sample,10,50);
[1] 59.51856
> BICNormModel(sample,10,50);
[1] 61.33958
```

We hebben tien keer achter elkaar vijftig munten opgeworpen en het aantal keer kop genoteerd. We zien aan de uitvoer dat zowel het AIC als het BIC de voorkeur geven aan het binomiale model; immers, voor het binomiale model geven ze een kleinere waarde. We kunnen ons afvragen of dit resultaat hetzelfde blijft als we de steekproefgrootte opvoeren. Als we bijvoorbeeld nemen $n = 200$ en $N = 100$, dan krijgen we de volgende resultaten.

```
> sample <- ExperimentUitvoeren(200,100);
> sample;
  [1] 44 46 44 52 52 52 60 52 59 51 51 59 43 50 52 45 51 39 50 47 48 54 49
 [24] 56 49 48 56 46 51 54 53 51 55 50 55 54 53 59 65 43 43 52 46 40 58 61
 [47] 45 58 42 56 57 55 50 53 40 54 49 49 44 45 48 52 46 49 59 48 45 53 60
 [70] 52 48 50 52 50 54 52 59 48 50 42 47 56 49 52 57 53 52 55 48 49 45 43
 [93] 41 46 54 45 56 54 47 46 51 52 53 42 48 50 58 49 49 47 49 50 51 52 50
[116] 56 53 53 56 57 52 46 48 50 59 53 59 50 43 55 46 41 48 40 50 53 48 49
[139] 49 49 50 55 43 51 54 57 50 40 47 52 46 51 45 48 52 43 43 47 56 54 46
[162] 47 54 50 57 53 48 43 48 50 45 47 52 50 54 50 52 47 54 54 55 50 51 47
[185] 49 51 49 52 44 50 50 53 52 48 46 54 50 46 50 59
> AICBinomModel(sample,200,100);
[1] 1194.082
> AICNormModel(sample,200,100);
[1] 1195.111
> BICBinomModel(sample,200,100);
[1] 1197.38
> BICNormModel(sample,200,100);
[1] 1201.708
```

We zien dat ons vorige resultaat overeind blijft: nog altijd wordt het binomiale model beter bevonden. Het normale model is echter bijna even goed. Deze resultaten zijn geheel in overeenstemming met de theorie die we in paragraaf 5.2.1 hebben behandeld. Dat het normale model bijna even goed wordt bevonden, heeft te maken met de **centrale limietstelling** [11]. Die zegt ons dat voor $n \to \infty$ de verdeling van de stochast $K = \sum_{i=1}^{n} K_i$ nadert naar een normale verdeling.

Bibliografie

[1] H. Akaike, **Information theory and an extension of the maximum likelihood principle**, Proceedings of the Second International Symposium on Information Theory (B. N. Petrov en F. Csaki, eds.), Akadémiai Kiadó, Budapest, 1973, pp. 267–281.

[2] T. M. Apostol, **Mathematical analysis: a modern approach to advanced calculus**, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1957.

[3] M. Bonneu en X. Milhaud, **A modified Akaike criterion for model choice in generalized linear models**, Statistics **25** (1994), no. 3, 225–238, Engelstalige samenvatting.

[4] H. Bozdogan, **Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions**, Psychometrika **52** (1987), no. 3, 345–370.

[5] K. P. Burnham en D. R. Anderson, **Model selection and multi-model inference**, tweede editie, Springer, New York, juli 2002.

[6] G. C. Chow, **A comparison of the information and posterior probability criteria for model selection**, Journal of Econometrics **16** (1981), 21–33.

[7] S. Kullback en R. A. Leibler, **On information and sufficiency**, Annals of Mathematical Statistics **22** (1951), 79–86.

[8] K. Landsman, **Modelselectie: AIC versus BIC**, collegedictaat geschreven voor de cursus Toegepaste Wiskunde 2 (bacheloropleiding wiskunde aan de Radboud Universiteit), 2008.

[9] H. Linhart en W. Zucchini, **Model selection**, Wiley series in probability and mathematical statistics. Applied probability and statistics, Wiley, New York, 1986.

[10] A. D. R. McQuarrie en C.-L. Tsai, **Regression and time series model selection**, World Scientific, Singapore, 1998.

[11] W. R. Pestman, **Mathematical statistics. An introduction**, Walter de Gruyter & Co., Berlin, 1998.

[12] W. R. Pestman en I. B. Alberink, **Mathematical statistics. Problems and detailed solutions**, Walter de Gruyter & Co., Berlin, 1998.

[13] T. Sawa, **Information criteria for discriminating among alternative regression models**, Econometrica **46** (1978), no. 6, 1273–1291.

[14] R. Shibata, **Statistical aspects of model selection**, From data to model (J. C. Willems, ed.), Springer-Verlag, London, 1989, pp. 215–240.

[15] C. J. Stone, **Local asymptotic admissibility of a generalization of Akaike's model selection rule**, Annals of the Institute of Statistical Mathematics **34** (1982), no. 1, 123–133.

[16] M. Stone, **An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion**, Journal of the Royal Statistical Society **39** (1977), no. 1, 44–47.

[17] N. Sugiura, **Further analysis of the data by Akaike's Information criterion and the finite corrections**, Communications in Statistics, Theory and Methods **A7** (1978), no. 1, 13–26.

[18] L. Wasserman, **Bayesian model selection and model averaging**, Journal of Mathematical Psychology **44** (2000), no. 1, 92–107.

[19] S. Weisberg, **Applied linear regression**, tweede editie, Wiley, New York, 1985.

[20] H. White, **Estimation, inference and specification analysis**, Cambridge University Press, Cambridge, UK, 1994.

[21] Y. Yang, **Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation**, Biometrika **92** (2005), no. 4, 937–950.

[22] V. A. Zorich, **Mathematical Analysis II**, vierde editie, Springer-Verlag, Berlijn, 2004.

[23] **The R Project for Statistical Computing**, `http://www.r-project.org/`.