CWI Syllabi

Centrum voor Wiskunde en Informatica (CWI)

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Telephone + 31 - 20 592 9333
Telefax + 31 - 20 592 4199
Website http://www.cwi.nl/publications/

CWI is the nationally funded Dutch institute for research in Mathematics and Computer Science.

Proceedings of the fifty-second European Study Group with Industry

Amsterdam, The Netherlands, 31 January – 4 February 2005

J. Hulshof (editor)

Centrum voor Wiskunde en Informatica

CWI SYLLABUS 55

Study groups with industry are meetings where people from industry join forces with mathematicians to jointly tackle industrial problems. The first such meeting was held in the sixties at the University of Oxford. Nowadays study groups with industry are organised in many countries, e.g. Australia, the USA and several European countries.

2000 Mathematics Subject Classification:
00B25, 35QXX, 60JXX, 92D40, 92C05, 31AXX, 90B80, 90C10.

ISBN 90 6196 532 2
NUGI-code: 811

Contents

Preface

Mathematics has intrinsic beauty, but it can also be fruitfully applied to solve concrete problems in the real world. This is the goal in Study Groups Mathematics with Industry. They originated in Oxford in the sixties, and have by now spread around the globe. The *52nd European Study Group with Industry* (ESGI52), locally known as SWI2005, was held at the Vrije Universiteit Amsterdam from January 31 to February 4, 2005. The results obtained during this week have been materialised in the proceedings that are now lying before you.

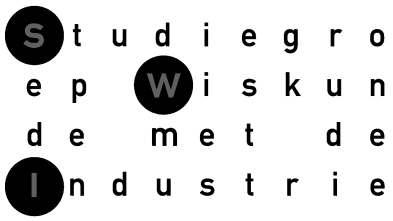
The problems where brought to the study group from many different quarters: the Netherlands Forensic Institute, the Academic Medical Center, Royal Dutch Airlines (KLM/Air France), the Dutch National Metrology Institute (NMi), the Software Improvement Group, and the Institute for Fundamental and Clinical Movement Sciences (IFKB). The main financial support for the study group came from the Technology Foundation STW of the Netherlands Organisation for Scientific Research (NWO), Division for Physical Sciences (EW), and the Ministry of Economic Affairs. Additional support was provided by the Department of Mathematics of the Vrije Universiteit Amsterdam, the European RTN Network "Fronts-Singularities", and the European Consortium for Mathematics in Industry (ECMI). Finally, many mathematicians contributed their time and effort. We are very grateful to everyone.

The mathematics in four of the six problems lean heavily on methods of optimisation, where both linear methods (in the form of a decomposition in principle components) and nonlinear optimisation techniques were employed. It is striking that from the mathematical point of view these problems are so similar, while they arise from vastly different applications.

All the problems have at least partially been solved during the week; of course some outcomes were more detailed than others, and sometimes the result was more of a pointer towards an avenue to be pursued. Overall, the problem presenters expressed great satisfaction about the progress made during the week. We are tempted to conclude that mathematics has once again proved itself useful in practical situations. Finally, many participants told us they would like to participate again. Well, the next study group in the Netherlands will be in Eindhoven from January 30 to February 3, 2006; for more information, visit http://www.win.tue.nl/swi2006/.

Organising committee

Jan Bouwe van den Berg
Sandjai Bhulai
Joost Hulshof
Ger Koole
Corrie Quant
JF Williams



Nederlandse Samenvattingen

Dutch Summaries

De Studiegroep Wiskunde met de Industrie 2005 vond plaats van 31 januari tot 4 februari aan de Vrije Universiteit Amsterdam. Tijdens deze week is door een groep wiskundigen gewerkt aan een aantal vragen die voortkomen uit praktische problemen:

- Hoe snel kan een patiënt na een hartoperatie het beste weer worden opgewarmd?
- Hoeveel water moet een vliegtuig meenemen aan boord zodat er voldoende is voor toiletbezoek en thee en koffie voor de passagiers?
- Wat maakt het voor de bewijswaarde in een rechtzaak uit hoe het bewijs is geselecteerd?
- Hoe kun je het beste bepalen wat precies een kilogram is?
- Wat is de beste manier om een groot computerprogramma in hanteerbare stukjes software op te delen?
- Kan wiskunde licht werpen op de aansturing van spieren door de hersenen?

De resultaten van het werk tijdens de Studiegroep zijn vastgelegd in deze *proceedings*. Om de toegankelijkheid hiervan te verhogen volgen nu eerst korte Nederlandstalige samenvattingen van de verslagen.

1 Modelvorming van het koel- en opwarmproces van patiënten tijdens openhartoperaties

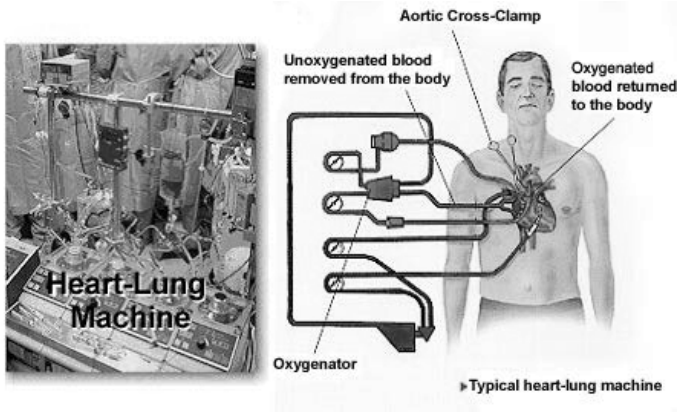
Natascha Severens en Mark Peletier

Voor hen die het leven te kort vinden is cryogene suspensie — conservering van het lichaam door bevriezing — altijd een fascinerende gedachte geweest. Voor de meeste wetenschappers en medici is het ook niet meer dan dat: tot nu toe is suspensie en opwarming alleen geslaagd bij de simpelste organismen, zoals nematode wormen. Dit weerhoudt sommige optimisten er niet van om tegen torenhoge kosten afspraken te maken met organisaties als de Alcor Life Extension Foundation, in de hoop over duizenden jaren te ontwaken in een betere wereld — of op z'n minst in een wereld waar de ouderdom succesvol verlengd kan worden.

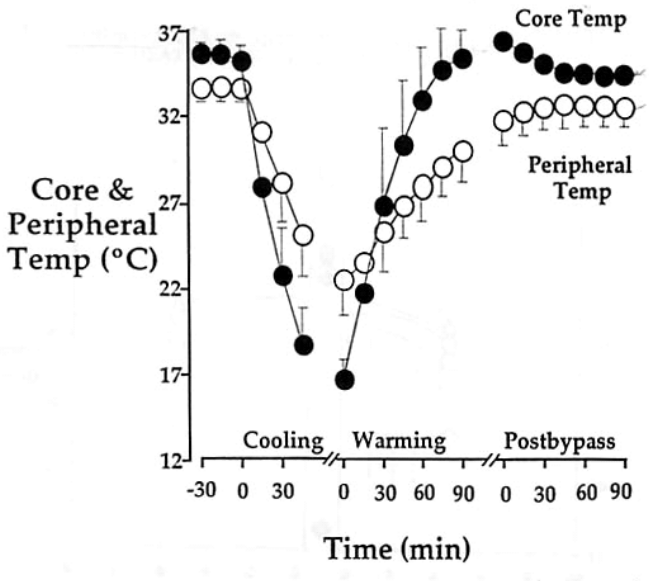
Weinigen beseffen dat afkoeling en opwarming routine is in hedendaagse ziekenhuizen. Tijdens **cardiopulmonaire** bypass operatie, een type operatie waarbij de functie van het hart en de longen tijdelijk door een *hart-longmachine* wordt overgenomen, wordt de patient vaak vijf of tien graden gekoeld, en soms zelfs tot twintig graden toe - tot een temperatuur van 17 graden Celsius. Vergelijk dit met de ervaring dat onderkoelde bergbeklimmers bij een kerntemperatuur van 25 graden al nauwelijks te reanimeren zijn. Zonder *life-support system* was bij dergelijke onderkoeling de patient al dood geweest.

Een deel van de afkoeling tijdens zo'n open-hartoperatie is niet eens opzettelijk. De patient verliest alleen al warmte door de vaatverwijding die de toegediende anesthetica met zich meebrengen, door de koude operatiekamer, door de opening van de borstkas, waardoor een groot, nat oppervlak vrijkomt, en doordat het bloed in contact komt met de koude hart-longmachine. De hart-longmachine bevat daarom een warmtewisselaar waarmee de perfusionist — de specialist verantwoordelijk voor de buitenlichamelijke bloedsomloop — de temperatuur van het bloed kan regelen.

Als de temperatuur van het bloed toch geregeld kan worden, waarom wordt de patient dan zo ver afgekoeld? Dat heeft te maken met risicobeperking. Bij een open-



Figuur 1.1: Hart-longmachine



Figuur 1.2: Temperatuurverloop tijdens een operatie (dichte bollen: kern, open bollen: periferie) [1]

hartoperatie leidt een onverwachte situatie (een gesprongen slagader, bijvoorbeeld) al snel tot een onderbreking van de bloedsomloop. De hersenen zijn de grootste zuurstofverbruiker van het lichaam, en bij gebrek aan zuurstof ontstaat snel hersenbeschadiging. De zuurstofbehoefte van cellen neemt echter sterk af als de temperatuur daalt. Door het lichaam te koelen heeft de chirurg veel meer tijd om een mogelijke crisis te bezweren.

Zolang de patient aan de hart-longmachine verbonden is gaat dit proces goed. De vitale organen, vooral de hersenen, zijn zeer goed doorbloed, en nemen de temperatuur van het bloed snel over. De problemen ontstaan pas wanneer de hart-longmachine wordt afgekoppeld, en de patient weer zijn eigen temperatuur moet gaan regelen. Dit leidt altijd tot een snelle daling van de kerntemperatuur met soms wel enkele graden, de zogeheten *afterdrop*.

Die afterdrop is erg ongelukkig, vooral bij een hartoperatie. De patient krijgt het koud en begint te rillen, waardoor veel energie en zuurstof verbruikt wordt, en dat betekent een grotere belasting voor het hart. Hoe kleiner de afterdrop, hoe beter het genezingsprces, zo is de ervaring, en de centrale vraag aan de Studiegroep was om deze afterdrop beter te begrijpen.

De oorsprong van de afterdrop ligt in het verschil in doorbloeding tussen de ‘kern’ van het lichaam, de romp en de hersenen, en de ‘periferie’, bestaande uit de armen en benen. Tijdens de urenlange operatie koelt de periferie af; na opwarming door de hart-longmachine is de temperatuur van de goed doorbloede kern wel weer op niveau, maar die van de minder goed doorbloede periferie loopt een paar graden achter. Na afkoppeling wordt de warmte herverdeeld: de kern warmt de periferie op, en wordt daardoor zelf kouder.

Modelvorming

Aan de basis van modelvorming staan de transportvergelijkingen die de warmtewisselingsprocessen beschrijven die plaatsvinden binnen het lichaam van de patiënt en tussen patiënt en omgeving. Tijdens de narcose spelen actieve regelmechanismen zoals zweten, rillen en vasoconstrictie geen rol; het lichaam gedraagt zich, wat warmte betreft, als passieve materie. Weefsels veranderen dan op twee manieren van temperatuur: door warmteuitwisseling met het bloed (perfusie) of door warmtegeleiding met naburige weefsels.

Waar alle weefsels warmte ongeveer even goed geleiden — we bestaan tenslotte voornamelijk uit water — zijn er grote verschillen in doorbloeding van verschillende typen weefsel. De 'kern' van het lichaam, de hersenen en de romp, is uitstekend doorbloed, zodat er nauwelijks verschil in temperatuur is tussen het hersenweefsel en het bloed dat na de warmtewisselaar het lichaam weer ingaat. Het 'perifeer' weefsel, voornamelijk de armen en benen, heeft een veel zwakkere bloedvoorziening, en is voor temperatuurveranderingen deels op geleiding aangewezen.

Het simpelste model voor de temperatuurveranderingen tijdens een operatie bevat dan twee compartimenten, kern en periferie. We hebben daar een derde compartiment aan toegevoegd; het rectum, het gebied rondom de anus, is belangrijk omdat tijdens de operatie de chirurg beslissingen neemt over opwarming en afkoeling op basis van de rectale temperatuur.

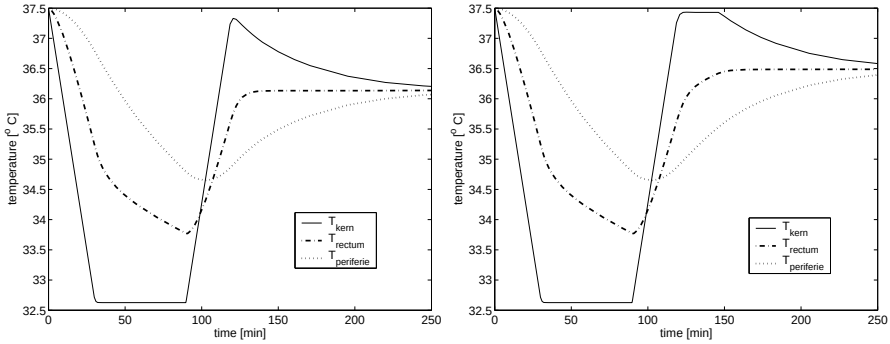
Resultaten

Met dit model is geprobeerd het verloop van experimenteel gemeten kern en rectale temperatuur te reproduceren. De operatie wordt opgesplitst in vier delen: afkoelen, constante (lage) temperatuur, opwarmen, en de periode na afkoppelen van de hart-longmachine; de vier delen zijn in de grafieken hieronder te herkennen. Dit model blijkt in staat te zijn om typische kenmerken van het temperatuurverloop tijdens een operatie te voorspellen, zoals de langzame reactie van de periferie ten opzichte van de kern en rectum en het optreden van afterdrop. Het blijkt dat de periferie zich in de afkoelfase als een warmtebron gedraagt en in de opwarmfase als een warmteput; deze warmteput is verantwoordelijk voor de afterdrop. Uit simulaties met het model blijkt dat de chirurg de afterdrop kan verkleinen door bijvoorbeeld

1. de periode van lage temperatuur zo kort mogelijk te maken (niet getoond), of
2. de patiënt aan het eind van de opwarmperiode wat langer aan de hart-longmachine gekoppeld te laten (zie grafieken in Figuur 1.3).

Conclusie

Wat is de medische stand met dit model opgeschoten? Sommige resultaten komen overeen met de ervaringen van de chirurgen. Dat is wel eens frustrerend ('ja, maar dat wisten we allang!') maar het is een bevestiging van de voorspellende kwaliteit van het model, en betekent daarom dat er goede kansen zijn om met dit model verder te kunnen komen dan menselijke ervaring en intuïtie. Vooral wijzen de resultaten de weg naar uitgebreider onderzoek: een goed begrip van perifeertemperatuur, bijvoorbeeld,



Figuur 1.3: Snel afkoppelen betekent hoge afterdrop: als patiënt meteen na het opwarmen van de hart-longmachine aan zichzelf wordt overgelaten, is de eindtemperatuur (uiteindelijke waarde in de grafiek links, 35.5 °C) hoger dan wanneer de patiënt een tijdje aangekoppeld blijft (rechts, uiteindelijke waarde 36.6 °C)

en hoe die verandert met de tijd en met de kerntemperatuur, blijkt essentieel te zijn in het beheersen van afterdrop.

In een samenwerking tussen het AMC, de TU Eindhoven, en de Universiteit Maastricht is inmiddels een promotieproject gaande dat zich richt op precies dit verdere onderzoek. Hierin zullen allerlei effecten nauwkeuriger bestudeerd en beschreven worden, zoals de verdeling van de temperatuur over de verschillende ledematen, het verlies aan warmte door de open borstkas, en het effect van simpele maatregelen zoals kruiken en warme dekens over de benen. Wie weet blijkt na een uitgebreide wiskundige analyse dat grootmoeder altijd al het beste wist hoe je je voeten weer warm krijgt?

Met dank aan Dirk-Jan Veldman en Bas de Mol van het AMC, en aan Joyce Aitchison, Christina Giannopapa, Vincent Guyonne, Miroslav Kramar, Simon van Mourik, Jasmina Panovska, en Marcus Tindall.

Referenties

[1] Tissue Heat Content and Distribution During and After Cardiopulmonary Bypass at 17 °C; Rajek, A., Lenhardt, R., Sessler, D.I., Grabenwger, Kastner, J., Mares, P., Jantsch, U., and Gruber, E., *Anesth. Analg.*, 88, 1220-5, 1999.

2 Planning van drinkwater voor vliegtuigen

Marco Bijvank, Menno Dobber, Quentin Botton, Eléonore de le Court, Jean-Christophe Van den Schrieck, Moïra de Viron, Maarten Soomer, Myriam Cisneros-Molina, Klaus Schmitz, Remco van der Hofstad, Ellen Jochemsz, Tim Mussche, Martin Summer, Maroescha Hoekstra, Jeroen Mulder en Mark Paelinck

Tijdens een vlucht wordt er drinkwater gebruikt voor consumpties en voor het doorspoelen van de toiletten. Het management van de Nederlandse luchtvaartmaatschappij KLM wil dat er genoeg drinkwater aan boord is op hun vluchten. Daarentegen, willen ze ook niet te veel water meenemen omdat dit, vanwege het gewicht, hogere kerosine kosten met zich meebrengt.

Voor een vlucht wordt de watertank van het vliegtuig gevuld. De apparatuur waar dit mee gebeurt, heeft alleen de mogelijkheid de tank te vullen tot een veelvoud van achtsten van de tank. De hoeveelheid water wordt van te voren bepaald aan de hand van de bestemming en het aantal passagiers van de vlucht. Overtollig water wordt na de vlucht uit de watertank geloosd. Het doel van dit onderzoek is om een optimale hoeveelheid water te bepalen waarmee het vliegtuig moet vertrekken. Er moet aan het eind van de vlucht zo weinig mogelijk water over zijn en tegelijkertijd moet de kans op een watertekort tijdens de vlucht erg klein zijn. Om het laatste vast te stellen moet er een geschikte definitie voor het service niveau worden gevonden.

Om een optimale hoeveelheid drinkwater te bepalen, maken we gebruik van historische data. Deze gegevens hebben helaas een aantal nadelen. Voordat het vliegtuig opstijgt, leest de purser het waterniveau van een display in de cabine. In de meeste vliegtuigen heeft deze display maar acht markeringen, waardoor het moeilijk is de exacte hoeveelheid water af te lezen. Daarom wordt deze waarde door de purser afgerond naar de dichtstbijzijnde markering. Omdat de tank ook gevuld wordt in achtsten beschouwen we deze waarde als exact. Na de landing wordt het watervolume weer genoteerd. Ook deze waarde is weer afgerond (en kan niet als exact worden beschouwd). Verder bevatten de datarecords gegevens over het type van het vliegtuig, het aantal passagiers, het vertrekpunt en de bestemming, de vertrektijd en vliegduur van de vlucht. De vraag is nu wat voor invloed de afgeronde data hebben en hoe dit verwerkt moet worden in het model.

Data-analyse

Op dit moment gebruikt KLM de historische data van vluchten met hetzelfde beginpunt en dezelfde bestemming, om het waterverbruik op een bepaalde vlucht te schatten. Het is interessant om te onderzoeken of vluchten met een ander beginpunt of bestemming, toch een vergelijkbaar patroon in waterverbruik hebben. Als dit het geval is, kunnen de data van vluchten geclusterd worden en kan er dus een betrouwbaardere schatting worden gegeven.

Voor de meeste bestemmingen is er een duidelijke correlatie tussen de bestemming en het waterverbruik per passagier. Dit betekent dat deze vluchten niet hetzelfde patroon in waterverbruik hebben. Vluchten met dezelfde vliegduur kunnen ook niet

samengenomen worden. Alleen tussen een aantal vluchten naar dezelfde regio, zoals vluchten van Amsterdam naar Aruba en van Amsterdam naar Bonaire, kon geen onderscheid worden gemaakt. De data van deze vluchten kunnen dus gecombineerd worden om de waterhoeveelheid te bepalen.

Service Level

De huidige definitie van service level, die KLM hanteert, is het percentage vluchten (met hetzelfde beginpunt en dezelfde bestemming) dat genoeg drinkwater aan boord heeft. Een service level van 95% betekent dus dat er op ten hoogste 5% van die vluchten een watertekort is. Dit betekent echter niet dat de kans voor een passagier om geconfronteerd te worden met een watertekort ook maar 5% is. Als een watertekort een structureel probleem is op vluchten met veel passagiers, dan heeft een passagier in een druk vliegtuig een grotere kans om met een watertekort geconfronteerd te worden. Dit is de reden dat wij een ander service level hebben gedefinieerd. Deze wordt hieronder besproken.

De totale waterconsumptie tijdens een vlucht met n passagiers, zal aangeduid worden met S_n. Op basis van de historische data hebben we alleen gegevens over afgeronde waardes van S_n. Om het bovenvermelde probleem te omzeilen, definiëren we nu het service level als de kans dat er genoeg water aanwezig is, gegeven dat er n passagiers aan boord zijn. Dit service level moet dan groter of gelijk zijn aan een vooraf (door het management van KLM) gedefinieerde waarde α. Dit betekent dat aan de volgende voorwaarde voldaan moet worden:

$$\mathbb{P}\left(S_n \leq \frac{j}{8}T\right) \geq \alpha, \qquad \forall n, \tag{2.1}$$

waarbij $j/8$, het percentage van de tank dat gevuld wordt ($j \in \{0, 1, \ldots, 8\}$) is en T staat voor de capaciteit van de tank (in liters). Deze definitie van service level wordt ook wel Quality of Service (QoS) genoemd, omdat het geldt voor elk passagiersaantal. Het is gedefinieerd vanuit het perspectief van de consument, die heeft nu onafhankelijk van het aantal medepassagiers hetzelfde risico om met een watertekort te worden geconfronteerd. Deze definitie was een openbaring voor de KLM en bracht nieuw inzicht in het probleem.

Uiteraard is de volgende stap om een zo klein mogelijke hoeveelheid water te bepalen dat aan boord aanwezig moet zijn, zodanig dat aan het service level voldaan wordt (d.w.z. de kleinste waarde voor j in vergelijking (2.1)). Om dit te bepalen hebben we een kansverdeling voor de totale waterconsumptie S_n nodig. We hebben drie verschillende methoden ontwikkeld om deze kansverdeling te bepalen. De methoden gebruiken de historische data en moeten daarom rekening houden met de afrondingen. De methoden worden in de volgende drie paragrafen besproken.

Empirische Benadering

De kans dat $j/8$ste van de watertank wordt gebruikt als er n passagiers aan boord zijn, wordt bij deze aanpak afgeleid uit de frequentie dat dit voorkomt in de data. Deze kansen kunnen worden gebruikt als een kansdichtheidsfunctie voor het watergebruik gedurende de vlucht. Om echter tot betrouwbare frequenties te komen, moeten er

genoeg vluchten in de data zijn met precies n passagiers. Helaas is dat niet het geval. Dit kan worden opgelost door ook waarnemingen mee te nemen waarbij het aantal passagiers in de buurt van n ligt.

We kunnen een kansverdeling schatten door een continue functie door deze negen punten te construeren en dan de oppervlakte onder de functie te normaliseren. Dit betekent dat we een interpolatiemethode nodig hebben. De meest gebruikte methode is een benadering door polynomen. In dit geval lijkt cubic spline interpolatie (Press et al. [1]) beter geschikt, omdat dit stabieler is dan het gebruik van polynomen. Het doel van cubic spline is een interpolatie formule te krijgen die continu is in de tweede afgeleide. Als dit gebeurd is, moet de verkregen functie genormaliseerd worden. Hierna kan het minimum waterniveau, waarvoor het service level op zijn minst α is, worden bepaald. Voor praktische toepassing zorgen we er voor dat dit waterniveau een monotoon stijgende functie is van het aantal passagiers. Omdat er weinig data beschikbaar zijn van vluchten met weinig passagiers, nemen we ook aan dat het gemiddelde waterverbruik per passagier per uur maximaal 1 liter is.

Normale Verdeling

De vorige aanpak gebruikte maar een gedeelte van de historische data om het waterverbruik te schatten (namelijk alleen de vluchten met ongeveer gelijke passagiersaantallen). In deze aanpak zal de totale waterconsumptie S_n voor een vlucht met n passagiers weergegeven worden door

$$S_n = \sum_{k=1}^{n} Y_k, \tag{2.2}$$

waarin Y_k staat voor de waterconsumptie van de $k$de passagier (in liters). De aanname die nu gemaakt wordt is dat het waterverbruik per passagier onafhankelijk en hetzelfde verdeeld is. Aangezien het aantal passagiers groot is, kan de centrale limietstelling (Ross [2]) toegepast worden. Als we deze stelling generaliseren door een constante toe te voegen aan het gemiddelde en de variantie van het waterverbuik, dan krijgen we

$$S_n \overset{d}{\sim} \mathcal{N}(\mu_0 + n\mu, \sigma_0^2 + n\sigma^2), \tag{2.3}$$

met μ en σ^2 het gemiddelde en de variantie van het waterverbruik per passagier en μ_0 en σ_0^2 het gemiddelde en variantie van de minimale hoeveelheid water die altijd gebruikt wordt. De waarden van deze vier parameters moeten geschat worden. Deze schatting wordt uitgevoerd met de maximum likelihood methode (Ross [2]).

Binomiale Verdeling

In de vorige aanpak werd geen aanname gemaakt over de exacte verdeling van het waterverbruik per passagier. Bij deze aanpak veronderstellen we dat een passagier een maximum hoeveelheid water (M) verbruikt met kans p of een minimum hoeveelheid water (m) met kans $1-p$. Dit betekent dat zowel het aantal passagiers dat de maximum hoeveelheid water verbruikt, als het aantal passagiers dat de minimale hoeveelheid verbruikt, binomiaal verdeeld is.

De waarden voor m, M en p moeten geschat worden om het gewenste waterniveau te bepalen. Dit kan op een intuïtieve manier gedaan worden, maar ook door meer geavanceerde methodes, zoals maximum likelihood.

Resultaten en Conclusies

In dit onderzoek hebben we een model ontwikkeld om de minimale hoeveelheid water te bepalen dat benodigd is tijdens een vlucht. Er moet echter wel aan een vooraf gesteld service niveau voldaan worden. Hierin is het service niveau gedefinieerd als de kans dat er genoeg water aanwezig is, gegeven het aantal passagiers aan boord. Het vaststellen van een kansverdeling, voor het totale watergebruik tijdens een vlucht, hebben we op drie verschillende manieren aangepakt. De grootste beperking waarmee rekening gehouden moest worden, was dat de data over het waterverbruik gegeven was in veelvouden van achtsten van de watertank.

Het gebruik van de empirische benadering is lastig omdat er in de praktijk voor veel vluchten niet genoeg data beschikbaar is voor een gegeven passagiersaantal. Daarom wordt ook de data met ongeveer gelijke passagiersaantallen gebruikt. Dit heeft tot gevolg dat de verdeling voor het totale waterverbruik dikkere staarten krijgt, waardoor het verkregen waterniveau te hoog ligt. Deze methode geeft dus eigenlijk een bovengrens voor de benodigde waterhoeveelheid.

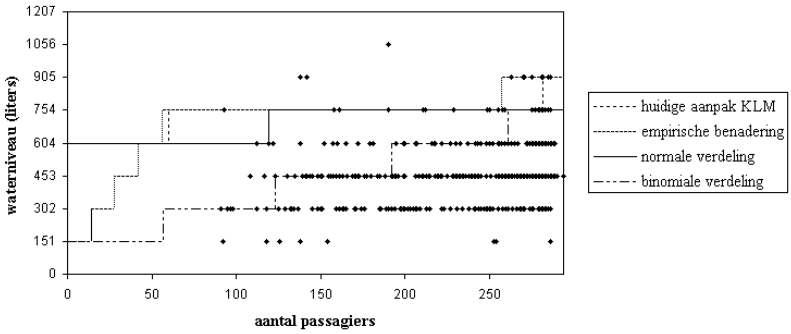
De methoden met de normale en binomiale verdeling gaan er vanuit dat het gemiddelde en de variantie van het totale waterverbruik lineair toenemen met het aantal passagiers. Het eerste klopt inderdaad met de data, het tweede helaas niet. Het verschil tussen deze twee methodes ligt in de geschatte parameters voor het gemiddelde en de variantie. Het binomiale model heeft als eigenschap dat het minimum en maximum waterverbruik begrensd is.

Voor de vlucht van Amsterdam (AMS) naar Bangkok (BKK) zullen we de drie methodes numeriek toelichten. De resultaten zijn samengevat in Tabel 2.1 en Figuur 2.1.

Onze voorkeur gaat uit naar de methode met de normale verdeling, waarbij de empirische benaderingsmethode als bovengrens gebruikt kan worden. De binomiale methode kan gebruikt worden om de invloed van het wijzigen van parameters te evalueren. De resultaten van de huidige methode van KLM, lijken op die van de methode met de normale verdeling. Deze laatste heeft echter een betere onderbouwing en gebruikt een betere definitie van het service level, daarom heeft deze methode ook de voorkeur van de KLM.

waterniveau (liters)	drempelwaardes voor het waterniveau (aantal passagiers)			
	huidige aanpak KLM	emprische benadering	normale verdeling	binomiale verdeling
151	-	0 - 13	-	0 - 56
302	-	14 - 27	-	57 - 122
453	-	28 - 41	-	123 - 191
604	0 - 59	42 - 55	0 - 118	192 - 260
754	60 - 281	56 - 256	119 - 294	261 - 294
905	282 - 294	257 - 294	-	-

Tabel 2.1: Aantal passagiers en aanbevolen waterhoeveelheid met een 95% Quality of Service, bepaald met de verschillende methodes, voor de vlucht AMS-BKK.



Figuur 2.1: Aanbevolen waterhoeveelheid met een 95% Quality of Service met de verschillende methoden en de afgeronde data voor de vlucht AMS-BKK

Referenties

[1] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., 1988, *Numerical Recipes in C*, Cambridge University Press, New York

[2] Ross, S.M., 2003, *Introduction to Probability Models, Eighth Edition*, Academic Press, San Diego

3 Selectie-effecten in de forensische wetenschap

Geert Jan Franx, Yves van Gennip, Peter Hochs, Misja Nuyens, Luigi Palla, Corrie Quant en Pieter Trapman

Inleiding en probleemstelling

Op het slachtoffer van een misdrijf treft men een rode vezel aan. Nadat de politie een verdachte heeft aangehouden, wordt er een rode trui in zijn klerenkast aangetroffen. De trui en de vezel worden vervolgens naar het forensisch laboratorium gebracht. Daar wordt gecontroleerd of ze van hetzelfde materiaal zijn en zo ja, hoe sterk dit bewijsmateriaal is. Een heel zeldzame trui zou namelijk als sterker bewijsmateriaal moeten gelden dan een die bij een grote kledingzaak is gekocht. Maar hoe hangt de sterkte van het bewijsmateriaal af van de omstandigheden waarin de trui is gevonden? Is het bewijsmateriaal bijvoorbeeld sterker als de verdachte geen andere truien in zijn kast had hangen, of maakt dat niet uit?

Dit is een voorbeeld van de volgende vraag die door het Nederlands Forensisch Instituut (NFI) werd gesteld aan de Studiegroep Wiskunde met de Industrie 2005: moet de forensisch expert weten op wat voor manier het bewijsmateriaal is geselecteerd? Deze vraag is ook relevant wanneer iemand verdacht wordt wanneer zijn of haar DNA in een DNA-database zit en overeenkomt met een stukje DNA dat op de plaats van een misdrijf is gevonden. Zulke DNA-databases bevatten natuurlijk niet het DNA van de hele bevolking, maar maakt dat wat uit?

Op dit moment wordt in zaken waar vezels als bewijsmateriaal worden gebruikt geen rekening gehouden met de manier waarop het bewijsmateriaal is verzameld. Dit is verontrustend, omdat het intuïtief duidelijk lijkt dat een match tussen de vezel en een verdachte die weinig truien heeft sterker bewijsmateriaal is dan een match met een 'truienverzamelaar'. Als dit inderdaad waar is, dan zouden statistische correcties gemaakt moeten worden wanneer dit soort bewijsmateriaal in de rechtzaal wordt gepresenteerd. De NFI-expert die de zaak behandelt, is echter in het algemeen geen statisticus. Hij heeft de mogelijkheid om de hulp van een statisticus in te roepen, maar doet dat alleen als hij dat noodzakelijk vindt. De vraag die aan de Studiegroep werd gesteld kan daarom gezien worden als een vraag om hulp in situaties die eenvoudig lijken, maar eigenlijk statistische correcties verlangen.

In dit stukje leggen we uit hoe kansrekening gebruikt kan worden om de sterkte van bewijsmateriaal te beoordelen. We maken een heel simpel model voor een vezelmatch, en trekken daar een aantal conclusies uit. Uiteraard is elk model ver verwijderd van de realiteit, en moet men extreem oppassen met het toepassen van theoretische resultaten toepassen op situaties in de echte wereld. Het is helemaal gevaarlijk wanneer statistiek wordt gebruikt om te bewijzen dat een misdaad heeft plaatsgevonden, zie [1]. Desondanks geloven we dat ons simpele model ons iets kan vertellen over hoe de selectie van bewijsmateriaal de sterkte van het bewijsmateriaal kan beïnvloeden.

De waarde van bewijsmateriaal beoordelen met kansrekening

Tijdens een rechtzaak moet de forensisch expert de sterkte van het bewijsmateriaal beoordelen. Een gedeelte van het bewijsmateriaal kan met de verdachten in verband wor-

den gebracht (dit geven we aan met E). De vraag is of de verdachte dit materiaal zelf heeft achtergelaten (deze gebeurtenis geven we aan met H), of dat de relatie tussen de verdachte en het bewijsmateriaal toevallig is. Deze gebeurtenis is het complement van H: H^c. De forensisch expert wordt verzocht om een zogenaamd aannemelijkheidquotiënt (Engels: *likelihood ratio*) te rapporteren: de kans op het bewijsmateriaal gegeven dat de verdachte schuldig is, gedeeld door de kans op het bewijsmateriaal gegeven dat de verdachte onschuldig is:

$$LR = \frac{P(E|H)}{P(E|H^c)}.$$

Het is verleidelijk om te denken dat een groot aannemelijkheidquotiënt impliceert dat de verdachte schuldig is, maar dit is in het algemeen niet waar. We zijn dan ook niet geïnteresseerd in het aannemelijkheidquotiënt zelf, maar in de kans dat de verdachte schuldig is gegeven dat het bewijsmateriaal overeenkomt, $P(H|E)$. Natuurlijk hebben deze twee met elkaar te maken: met regel van Bayes kunnen we $P(H|E)$ als volgt uitrekenen in termen van het aannemelijkheidquotiënt :

$$\begin{aligned} P(H|E) &= \frac{P(H \cap E)}{P(E)} = \frac{P(E|H)P(H)}{P(E \cap H) + P(E \cap H^c)} \\ &= \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)}. \end{aligned}$$

Wanneer we de teller en noemer delen door $P(E|H)P(H)$, krijgen we

$$P(H|E) = \left(1 + \frac{P(H^c)}{P(H)} \frac{P(E|H^c)}{P(E|H)}\right)^{-1} = \left(1 + \frac{P(H^c)}{P(H)} \frac{1}{LR}\right)^{-1}$$

De kans $P(H|E)$ heet een *posteriori kans*, $P(H)$ en $P(H^c)$ heten *a priori kansen*. De rechter heeft schattingen van deze laatste nodig om een idee te krijgen van $P(H|E)$.

Het truimodel

We gaan nu een heel simpel model bouwen voor in de inleiding beschreven geval van een gevonden vezel. Veronderstel dat we weten dat er een gebeurtenis heeft plaatsgevonden waarbij een onbekend persoon (de *donor*) een vezel van zijn trui heeft achtergelaten op een andere persoon (het *slachtoffer*). Het moment waarop de vezel werd overgedragen heet het transfermoment. Het type vezel noemen we Y en er zijn geen andere vezels op het slachtoffer gevonden. Om de donor te vinden, onderzoeken we de truien in de kast van een willekeurig persoon (de *verdachte*). Daar vinden we een trui gemaakt van hetzelfde type vezel.

We willen nu de kans uitrekenen dat de verdachte daadwerkelijk de donor van de vezel op het slachtoffer is, gegeven het bewijsmateriaal. De vraag is nu hoeveel we moeten weten van het gevonden bewijsmateriaal: volstaat het om te weten dat een van de truien in de kast van de verdachte overeenkwam met de vezel op het slachtoffer? Of moeten we bijvoorbeeld ook weten hoeveel truien de verdachte in zijn kast had?

Om de kans uit te rekenen dat de verdachte de donor was, nemen we aan dat de relatieve frequentie waarmee de hele bevolking truien draagt die bestaan uit vezels van

type Y geschat kan worden op g_Y; de kans dat de verdachte een trui aanhad van dit type noemen we f_Y. Tenslotte nemen we aan dat niemand truien heeft verborgen of weggegooid en dat alle truien uit één vezeltype bestaan.

We schrijven E_1 voor de gebeurtenis dat de vezel op het slachtoffer van type Y is, en E_2 voor de gebeurtenis dat we een vezel van type Y in de kast van de verdachte vinden. De gebeurtenis dat de verdachte de donor van de vezel op het slachtoffer is noemen we H. We nemen ook aan dat E_2 en H onafhankelijk zijn. Zoals uitgelegd in de vorige paragraaf, kunnen we nu $P(H|E_1 \cap E_2)$ uitrekenen:

$$\begin{aligned} P(H|E_1 \cap E_2) &= \left(1 + \frac{P(H^c)}{P(H)} \frac{P(E_1 \cap E_2|H^c)}{P(E_1 \cap E_2|H)}\right)^{-1} \\ &= \left(1 + \frac{P(H^c)}{P(H)} \frac{P(E_2|H^c)P(E_1|H^c \cap E_2)}{P(E_2|H)P(E_1|H \cap E_2)}\right)^{-1} \\ &= \left(1 + \frac{P(H^c)}{P(H)} \frac{P(E_2)g_Y}{P(E_2)f_Y}\right)^{-1} = \left(1 + \frac{P(H^c)}{P(H)} \frac{g_Y}{f_Y}\right)^{-1}. \end{aligned}$$

We zien dat als g_Y klein is (type Y vezels zijn zeldzaam), dan is $P(H|E_1 \cap E_2)$ relatief groot. Verder, als f_Y klein is (de verdachte draagt zijn trui met type Y vezels niet vaak), dan is $P(H|E_1 \cap E_2)$ klein. In het speciale geval dat de verdachte k truien heeft en die even vaak draagt, krijgen we $f_Y = 1/k$. We zien dan dat hoe meer truien de verdachte heeft, hoe kleiner de kans is dat hij de donor is. Dat lijkt redelijk: een persoon die duizend truien heeft, loopt een grote kans dat een van zijn truien overeenkomt met de vezel op het slachtoffer, maar de sterkte van die match is natuurlijk niet erg groot.

Conclusies

In ons basale truimodel hebben we laten zien dat veel dingen gerapporteerd zouden moeten worden om het bewijsmateriaal goed te kunnen interpreteren. Niet alleen dat er een trui in de kast van de verdachte is gevonden die overeenkomt met een vezel op de plaats van de misdaad, maar bijvoorbeeld ook hoeveel truien hij in zijn kast had, en hoe vaak de verdachte die bepaalde trui draagt.

Hoewel ons model ver van realistisch was, denken we dat in het algemeen informatie over de manier waarop bewijsmateriaal is vergaard zo compleet mogelijk moet zijn. Alle informatie moet in een groot model gepropt worden. In zo'n model moeten ook andere stukken bewijsmateriaal gepropt worden, net als 'negatief bewijs', d.w.z., stukken bewijsmateriaal die niet met de verdachte in verband kunnen worden gebracht. We realiseren ons dat we zelfs in dat geval slechts een model hebben, en dat elk model tot veel discussies aanleiding zal geven.

We concluderen dat selectie-effecten in de forensische wetenschap een belangrijke rol spelen en dat de moeite gedaan zou moeten worden om de statistische interpretatie van bewijsmateriaal in de rechtzaal te verbeteren.

Referenties

[1] Van Lambalgen, M., and Meester, R., On the (ab)use of statistics in the legal case against the nurse Lucia de B, preprint, available from `http://www.few.vu.nl/~rmeester/pre.html` (2005).

4 Optimale weegschema's

Sandjai Bhulai, Thomas Breuer, Eric Cator en Fieke Dekkers

Inleiding

De kilogram is de laatste fysische grootheid die nog gedefinieerd is in termen van een tastbaar object: het is de massa van een platinum-iridium cilinder, vervaardigd in 1889, die bij het 'Bureau des Poids et des Mesures' in Frankrijk bewaard wordt. Om precies te zijn is de kilogram gedefinieerd als de massa van dit object net nadat het gewassen is. Platinum-iridium adsorbeert koolhydraten uit de atmosfeer, waardoor de cilinder regelmatig gewassen moet worden om het gewicht te verwijderen dat erop neergeslagen is.

In Nederland is het Nederlands Meetinstituut (NMi) in letterlijke zin verantwoordelijk voor de kilogram: de Nederlandse standaard kilogram, de platinum-iridium cilinder nummer 53, wordt op NMi terrein onder zorgvuldig gereguleerde omstandigheden bewaard. De cilinder reist af en toe naar Parijs om daar vergeleken te worden met de internationale standaard. Uiteraard draagt het NMi zorg voor meer dan alleen de veilige opslag van de kilogram: de afdeling 'massameting' is, onder meer, ook verantwoordelijk voor zeer nauwgezette kalibratie van gewichten. Deze gewichten worden gebruikt om de massa's van gewichten van lagere standaarden te bepalen, die bijvoorbeeld door supermarkten gebruikt worden om hun weegschalen te kalibreren, of door doping laboratoria, waar zeer kleine massa's met grote nauwkeurigheid vastgesteld moeten worden.

Voor de kalibratie van gewichten gebruikt het NMi gewichten van roestvrij staal, waarvan de massa's met zeer grote nauwkeurigheid bepaald moeten worden. Om resultaten met voldoende precisie te verkrijgen, moeten de metingen gecorrigeerd worden voor effecten als de opwaartse luchtdruk en de verschillen in de positie van het massamiddelpunt voor verschillende gewichten!

Het probleem

De Nederlandse kilogram wordt gebruikt als startpunt in het bepalen van de massa's van de gewichten van de hoogste kwaliteit roestvrij staal: eerst wordt de massa van een roestvrij stalen kilogram bepaald door directe vergelijking met de nationale standaard. In de tweede stap wordt de twee roestvrij stalen kilogram gebruikt en niet de nationale platinum-iridium kilogram, opdat externe invloeden minimaal effect hebben op de nationale standaard. De roestvrij stalen kilogram wordt vervolgens gebruikt om een roestvrij stalen set van gewichten te kalibreren bestaande uit een gewicht met nominale massa van 500 gram, twee van 200 gram en twee van 100 gram, zodat de honderdvouden van 1000 gram naar 100 gram gedekt zijn. Het verschil tussen de nominale massa en de echte massa van de gewichten is extreem klein. De gewichten van deze set worden vervolgens gebruikt om massa's van andere sets met andere ranges te bepalen.

Om de massa van een individueel gewicht te bepalen, kunnen we het vergelijken met het standaard gewicht met gelijke nominale massa waarvan de feitelijke massa bekend is met voldoende nauwkeurigheid – tenzij we natuurlijk de massa op het hoogste

niveau van precisie willen bepalen. Massa metrologie instituten gebruiken zogenaamde *weegschema's* om dit probleem op te lossen. Een weegschema voor de set van gewichten van 1000g naar 100g van het NMi bestaat uit paren van combinaties van gewichten uit de collectie van de 6 gewichten. Een schema kan bijvoorbeeld een vergelijking van een van de 200g gewichten met de twee 100g gewichten in zich hebben. Voor elk paar in het schema worden de verschillen in massa tussen de twee paren bepaald door middel van de STS-procedure, die hieronder beschreven zal worden. Om voldoende precisie te garanderen worden de paren in een weegschema zo gekozen dat ze gelijke nominale massa hebben.

Het is in theorie mogelijk om de vijf onbekende massa's te bepalen door vijf geschikte metingen van massaverschillen uit te voeren. In de praktijk worden er echter meetfouten gemaakt en moeten er meer metingen – niet noodzakelijk allemaal met verschillende combinaties van gewichten – verricht worden om tot nauwkeurigere schattingen van de ware massa's te komen. De schattingen van de massa's van de gewichten en de onzekerheid in deze massa's kunnen verkregen worden uit een overbepaald stelsel met behulp van de kleinste kwadraten analyse.

Gedurende de 'Studiegroep Wiskunde met de Industrie' die in februari 2005 aan de Vrije Universiteit gehouden werd, werd de volgende vraag door het NMi gesteld: *Wat is een optimaal weegschema voor de set van gewichten van het NMi, d.w.z. een weegschema dat de onzekerheid in de geschatte massa's minimaliseert onder de voorwaarde dat het aantal metingen kleiner is dan een gegeven getal.*

De STS procedure

Zoals eerder vermeld is, wordt voor elk paar in een weegschema het verschil in massa bepaald door middel van de STS-procedure. Neem, om de STS procedure te illustreren, twee sets van gewichten met ongeveer dezelfde massa. De balans om de massaverschillen te meten is een balans met enkelvoudige arm, waarmee het massaverschil tussen twee sets gewichten als volgt bepaald wordt. De eerste set, die we de standaard set (S) noemen, wordt geplaatst op de balans, die daarna op 0 wordt gesteld. De set S wordt verwijderd, en daarna nog eens op de balans geplaatst, waarna de eerste meting, x_0, afgelezen wordt.

Helaas is x_0 in het algemeen niet gelijk aan nul; in de praktijk wordt er een drift geobserveerd tussen opeenvolgende metingen, ook worden er meetfouten gemaakt. Vervolgens wordt de set S verwijderd en de tweede set, die we de test set (T) noemen, wordt op de balans geplaatst, waarna de tweede meting, x_1, afgelezen wordt. De metingen worden nu door S en T te alterneren voorgezet; dit verklaart de naam *STS procedure*. Als de drift niet al te wild fluctueert tussen opeenvolgende metingen, dan kan dit geëlimineerd worden door gebruik te maken van de STS procedure.

De metingen worden zo veel mogelijk herhaald om tot betrouwbare resultaten te komen. In de huidige opstelling die het NMi nu gebruikt, worden de gewichten handmatig op elkaar gestapeld, wat een tijdrovende procedure is. In de praktijk is het daarom zelden mogelijk om meer dan 20 metingen te verrichten zonder onderbrekingen.

Het modelleren van STS metingen

Stel dat we $k+1$ STS metingen verrichten met de sets S, met massa m_S, en T, met massa m_T. Laat x_i de i-de meting zijn, voor $0 \leq i \leq k$. We veronderstellen dat

$$x_i = 1_{\{i \text{ oneven}\}}(m_T - m_S) + D(i) + V_i,$$

waar V_i een meetfout is, die we proportioneel aan de totale massa op de balans veronderstellen. Om precies te zijn nemen we aan dat

$$V_i \sim N(0, \alpha^2 m_S^2), \tag{4.1}$$

waarbij $\alpha \in \mathbb{R}$ een onbekende constante is. De term $D(i)$ beschrijft de drift van de balans. We gaan er vanuit dat

$$\frac{D(i+1) + D(i-1)}{2} - D(i) \approx 0, \tag{4.2}$$

voor alle $1 \leq i \leq k-1$, wat consistent is met aannamen die doorgaans door het NMi gemaakt worden. Aan deze eis wordt natuurlijk voldaan wanneer D lineair is.
Definieer voor $1 \leq i \leq k-1$

$$\Delta\mu_i = \frac{(-1)^{i+1}}{m_S}\left(x_i - \frac{x_{i+1} + x_{i-1}}{2}\right) \approx \frac{m_T - m_S}{m_S} + E_i, \tag{4.3}$$

waarbij

$$E_i = \frac{1}{m_S}\left(V_i - \frac{V_{i+1} + V_{i-1}}{2}\right) \sim N(0, \alpha^2),$$

onafhankelijk van m_S. Observeer dat, gebruikmakend van (4.2), de drift geëlimineerd wordt.

In de procedure die nu door het NMi gebruikt wordt, wordt het gemiddelde van de $\Delta\mu_i$ gebruikt als een benadering voor $\frac{m_T - m_S}{m_S}$. We stellen een procedure voor die deze tussenliggende stap overbodig maakt.

Optimale weegschema's

Na het modelleren van de STS metingen kunnen de optimale weegschema's voor de set van gewichten van het NMi bepaald worden. De balans in de STS procedure kan gebruikt worden om nauwkeurig hele kleine verschillen in massa's te meten, maar kan niet met voldoende precisie voor grote massaverschillen gebruikt worden. Dit impliceert dat de test en de standaard massa's in een STS meting dezelfde nominale massa moeten hebben, wat een sterkte restrictie in het aantal mogelijke combinaties met zich meebrengt. In feite zijn er voor de set van gewichten van het NMi slechts 10 mogelijke combinaties (of: *weegvergelijkingen*) mogelijk. Deze worden in Tabel 4.1 weergegeven. De notatie 200 en 200• wordt gebruikt om het onderscheid tussen de twee gewichten met nominale massa 200g te maken. (Analoog geldt dit ook voor de twee gewichten met nominale massa 100g).

Bij een gegeven weegschema kan er een reeks van STS metingen worden gedaan

Weegvergelijking	Standaard	Test
1	1000	500, 200, 200•, 100
2	1000	500, 200, 200•, 100•
3	500	200, 200•, 100
4	500	200, 200•, 100•
5	200, 100	200•, 100•
6	200, 100•	200•, 100
7	200	200•
8	200	100, 100•
9	200•	100, 100•
10	100	100•

Tabel 4.1: Mogelijke combinaties van gewichten (in nominale massa's genoteerd)

voor elke weegvergelijking in het schema. Zodoende krijgen we een reeks van $\Delta\mu_i$, zoals gedefinieerd in (4.3), voor elk van deze combinaties. De $\Delta\mu_i$ kunnen gecombineerd worden in een standaard lineair model. Met behulp van statistische technieken kan er een schatting gegeven worden van de ware massa's van de gewichten uitgedrukt in de gemeten waarden van de $\Delta\mu_i$, alsmede een schatting van de onzekerheid. In deze procedure worden de $\Delta\mu_i$ gebruikt, en niet het (ongewogen) gemiddelde, zoals in de huidige procedure bij het NMi gedaan wordt. Deze aanpassing leidt tot betere schattingen.

Veronderstel nu dat we een optimaal schema willen berekenen met een gegeven aantal weegvergelijkingen, die niet allemaal noodzakelijkerwijs verschillend zijn: een meting herhalen van dezelfde vergelijking in een weegschema kan nuttig zijn, omdat herhalingen extra, onafhankelijke informatie verschaffen. Als deze vergelijkingen uit Tabel 4.1 moeten komen, dan is het aantal weegschema's bestaande uit N vergelijkingen begrensd door 10^N (deze grens is uiteraard niet scherp). Het gevolg hiervan is dat het mogelijk is om de onzekerheid in de massa's van de gewichten door te rekenen voor alle mogelijke weegschema's bestaande uit maximaal 14 vergelijkingen – het maximum aantal opgelegd door het NMi – met behulp van Matlab op een gewone computer.

Het huidige weegschema van het NMi kent 8 weegvergelijkingen uit Tabel 4.1. De vergelijkingen worden gegeven door

$$1, 2, 3, 4, 5, 6, 8, 9.$$

Het is onduidelijk hoe dit weegschema, dat dateert uit een tijd waarin alle mogelijke schema's doorrekenen niet mogelijk was, tot stand is gekomen. Met behulp van de hedendaagse moderne computers is het mogelijk om aan te tonen dat dit schema niet optimaal is. Dus het is mogelijk een ander weegschema met 8 weegvergelijkingen te construeren zodanig dat de onzekerheid in de geschatte massa's van de gewichten kleiner is.

Tabel 4.2 geeft de optimale schema's bestaande uit N combinaties ($8 \leq N \leq 14$) weer, waarbij aangenomen wordt dat voor elke weegvergelijking er 20 STS metingen verricht worden.

De onzekerheid in het weegschema van het NMi is $1.1812\,\alpha^2$, waarbij α de con-

N	optimaal weegschema
8	1, 1, 2, 4, 7, 8, 9, 10
9	1, 1, 2, 3, 4, 7, 8, 9, 10
10	1, 1, 2, 2, 3, 4, 7, 8, 9, 10
11	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10
12	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10, 10
13	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 9, 10, 10
14	1, 1, 1, 2, 2, 3, 4, 4, 7, 8, 8, 9, 10, 10

Tabel 4.2: Optimaal weegschema bestaande uit vergelijkingen uit Tabel 4.1

stante is in (4.1), terwijl de onzekerheid in het optimale schema met 8 vergelijkingen $0.8468\,\alpha^2$ is. Dit betekent dus dat er een reductie van ongeveer 28% in onzekerheid bewerkstelligd kan worden zonder extra metingen te verrichten. Als het aantal weegvergelijkingen in het schema verhoogd wordt tot 14, wat overigens extra werk met zich meebrengt, dan kan de onzekerheid tot $0.4655\,\alpha^2$ teruggebracht worden, een reductie van ongeveer 63%. Zoals wel te verwachten is, neemt de onzekerheid af naarmate het aantal weegvergelijkingen in het schema toeneemt.

Het is opvallend dat in geen van de optimale schema's de vergelijkingen 5 en 6 voorkomen, de enige vergelijkingen waarbij de standaard set uit meer dan één gewicht bestaat. Het schema van het NMi bevat deze vergelijkingen wel. In feite bevat geen enkele oplossing binnen 1% van de optimale de vergelijkingen 5 en 6, wat impliceert dat dit geen effect van afrondingsfouten is. In plaats van de vergelijkingen 5 en 6, bevat het optimale schema vergelijkingen 7 en 10. Ga na dat deze samen dezelfde informatie verschaffen als de vergelijkingen 5 en 6 samen. In praktische situaties leidt dit tot een reductie in onzekerheid, vanwege het feit dat vergelijkingen 7 en 10 minder opstapelingen van gewichten vereisen.

Het verhogen van het aantal weegvergelijkingen is niet de enige manier om de onzekerheid te reduceren in de schattingen van de ware massa's. In de nabije toekomst stapt het NMi over op automatische balansen die de STS metingen minder afhankelijk maken van handmatige procedures. Bovendien kunnen er zo ook meer metingen verricht worden in een STS serie.

Het verhogen van het aantal STS metingen per weegvergelijking kan effectiever zijn dan het verhogen van het aantal weegvergelijkingen: het optimale schema bestaande uit 10 vergelijkingen met 30 STS metingen per vergelijking heeft een kleinere onzekerheid ($0.4358\,\alpha^2$) dan het schema met 12 vergelijkingen en 25 STS metingen ($0.4458\,\alpha^2$). In beide gevallen is het aantal metingen in totaal 300. Het verhogen van het aantal STS metingen per weegvergelijking is echter niet altijd de beste aanpak: 10×28 STS metingen leiden tot betere schattingen dan 8×35 metingen.

Andere sets van gewichten

Verschillende massa metrologie instituten gebruiken verschillende sets van gewichten. Het resultaat voor de set van gewichten van het NMi, dat bestaat uit gewichten met nominale massa 1000g, 500g, 200g (tweemaal) en 100g (tweemaal), kan gegeneraliseerd worden naar sets die door andere massa metrologie instituten gebruikt worden.

Het Duitse metrologie instituut gebruikt bijvoorbeeld een set bestaande uit acht gewichten met 104 mogelijke combinaties. Het berekenen van de onzekerheid voor alle mogelijke schema's was te tijdrovend om binnen de studiegroep door te rekenen. Echter, gebaseerd op de inzichten met de Nederlandse weegschema's werd er besloten om alleen de weegvergelijkingen mee te nemen waarbij de test set uit een enkel gewicht bestond. Het optimale schema in deze vereenvoudigde setting had een onzekerheid die ongeveer 28% kleiner was dan die van het corresponderende schema in gebruik.

Conclusie

Het weegschema dat door het NMi gebruikt wordt is suboptimaal. Door over te gaan naar een ander weegschema, kan de onzekerheid in de massa's van de nationale standaard gewichten gereduceerd worden met ongeveer 63%. Dit weegschema gebruikt weliswaar meer metingen dan het huidige schema. Indien de hoeveelheid werk gelijk gehouden wordt, dan kan er een reductie van 28% gerealiseerd worden.

Het onderzoek, waarover dit artikel rapporteert, is uitgevoerd door Sandjai Bhulai, Thomas Breuer, Eric Cator en Fieke Dekkers. Onze dank gaat uit naar Inge van Andel van het NMi voor de door haar geleverde informatie.

5 Automatische partitionering van softwaresystemen

Rob Bisseling, Jarosław Byrka, Selin Cerav-Erbas, Nebojša Gvozdenović, Mathias Lorenz, Rudi Pendavingh, Colin Reeves, Matthias Röger en Arie Verhoeven

Kostenbesparende gepartitioneerde softwaresystemen

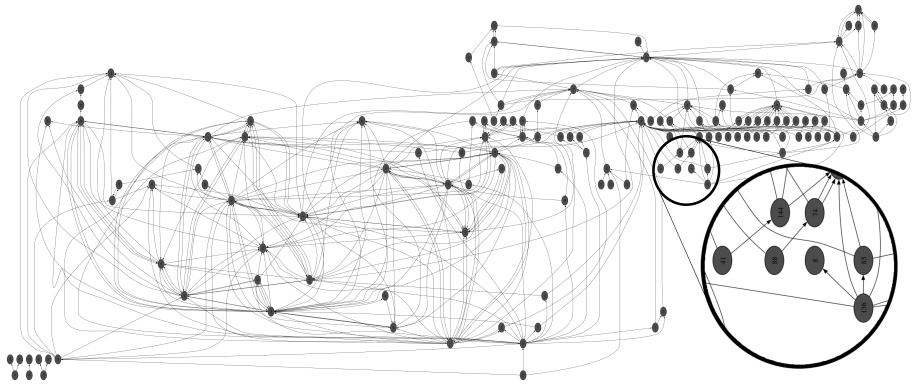
De afgelopen jaren heeft de vooruitgang in de informatietechnologie een enorme invloed gehad. Steeds meer organisaties, zoals banken en overheidsinstellingen, zijn afhankelijk van immer groeiende softwaresystemen om hun taken te kunnen uitvoeren. De grote complexiteit van deze systemen maakt het systeemonderhoud erg duur. Een mogelijke oplossing voor dit probleem bestaat in het opsplitsen van deze systemen in kleinere minder complexe modules. Iedere module wordt dan onderhouden door een kleine groep systeembeheerders. Zo'n gepartitioneerd systeem heeft echter een *interface* nodig voor de communicatie tussen de modules. Dit betekent dat ieder programma dat door andere modules gebruikt wordt, ook wel een interfaceprogramma genoemd, voorzien wordt van extra communicatiefaciliteiten. De kosten voor het extra onderhoud van deze interface zijn evenredig met de grootte van de interface ofwel het totaal aantal interfaceprogramma's. Voor een goede partitionering is de omvang van de interface minimaal en zijn de modules voldoende klein. In de praktijk zijn voor het vinden van zo'n splitsing ervaren deskundigen nodig. Voor erg grote softwaresystemen is het echter bijzonder handig om gebruik te maken van automatische partitioneeralgoritmen, die een goede suggestie kunnen geven. Het bedrijf Software Improvement Group (SIG, `http://www.sig.nl`) gevestigd in Diemen ontwikkelt dit soort hulpmiddelen.

Partitieprobleem van een call-graaf

Een *graaf* is een bekend wiskundig concept voor het representeren van een netwerk. Een graaf bestaat uit een aantal punten, *knopen* genaamd, die verbonden zijn door lijnen. Als de verbindingen ook een richting hebben, worden deze *pijlen* genoemd en spreken we van een gerichte graaf. Een softwaresysteem kan worden gemodelleerd door een call-graaf, waarvan de knopen de programma's voorstellen. Als programma v programma w aanroept (een call) is dit equivalent met een pijl van knoop v naar knoop w. Een korte notatie hiervoor is $v \to w$. Een module kan nu worden gerepresenteerd door een deelverzameling van de knopen van de call-graaf. Dit betekent dat de omvang van een module gelijk is aan het aantal knopen van de bijbehorende deelverzameling. We noemen knoop v een *interfaceknoop* als hij een inkomende pijl heeft vanuit een knoop in een andere module. De grootte van de interface is nu gelijk aan het aantal interfaceknopen. Zodoende kan het vinden van een goede splitsing voor het softwaresysteem worden geformuleerd als het volgende partitioneringsprobleem van een call-graaf:

Gegeven: een call-graaf en de gehele getallen K, L.

Vind: een partitionering van de knopen van de call-graaf in L deelverzamelingen zodat iedere deelverzameling maximaal K knopen heeft en het totaal aantal interfaceknopen minimaal is.



Figuur 5.1: Call-graaf van het softwaresysteem `Java1` met 158 knopen (programma's). De inzet rechtsonder geeft een gedetailleerde blik op een deel van de graaf waar de pijlen tussen de programma's duidelijk zichtbaar zijn.

Figuur 5.1 laat een call-graaf zien van een softwaresysteem uit de praktijk met de naam `Java1`.

Verschillende oplosmethoden

Twee verschillende probleemformuleringen blijken veelbelovend voor het probleem van SIG. Allereerst modelleren we het probleem als een *geheeltallig lineair programmeringsprobleem* (in het Engels: *Integer Linear Programming*, ILP). We gebruiken daarbij beslissingsvariabelen y_{vl} en x_{vl}, waarvan een waardering als volgt moet worden opgevat:

$$y_{vl} = \begin{cases} 1 & \text{knoop } v \text{ zit in deelverzameling } l \\ 0 & \text{anders} \end{cases}$$

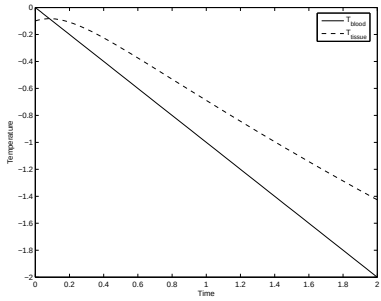
$$x_{vl} = \begin{cases} 1 & \text{knoop } v \text{ is een interfaceknoop} \\ & \text{bevat in deelverzameling } l \\ 0 & \text{anders.} \end{cases}$$

De eis dat module l niet meer dan K knopen mag bevatten kan dan geformuleerd worden als een lineaire ongelijkheid:

$$y_{v_1 l} + y_{v_2 l} + \cdots + y_{v_N l} \leq K.$$

En evenzo kan met lineaire vergelijkingen worden afgedwongen dat elke knoop in precies één module zit, etc. Het doel om zo min mogelijk interfaceknopen te gebruiken komt neer op het minimaliseren van de som van alle x_{vl}.

Deze formulering is een goede eerste stap naar een praktische oplossing. Een ILP probleem heeft een sterk meetkundig karakter, en er is een effective oplossingsmethode die de meetkundige aard van dit probleem uitbuit. Zodoende kan ook ons partitioneringsprobleem exact worden opgelost, maar de rekentijd loopt snel op als de grootte van het softwaresystemen toeneemt. Efficiënte herformuleringen en slimme hulpalgoritmen worden cruciaal om de oplossing te vinden binnen een redelijke tijd.



Figuur 5.2: Verbindingsmatrix (158×158) van het probleem `Java1`.

Als alternatief hebben we het probleem ook geformuleerd als een *hypergraafpartitioneringsprobleem*. Een hypergraaf is een generalisatie van een gewone ongerichte graaf met netten in plaats van lijnen. Een *net* verbindt een willekeurig aantal knopen, terwijl een lijn slechts twee knopen verbindt. Net n_j bestaat uit knoop j en alle knopen i met $i \to j$. Als de knopen in net n_j in verschillende deelverzamelingen vallen, is het net *gebroken*. Na toevoeging van de pijl $j \to j$ voor alle knopen j, blijkt het aantal interfaceknopen gelijk te zijn aan het aantal gebroken netten. Het call-graafpartitioneringsprobleem kan dus ook worden geformuleerd als een hypergraafpartitioneringsprobleem (HP).

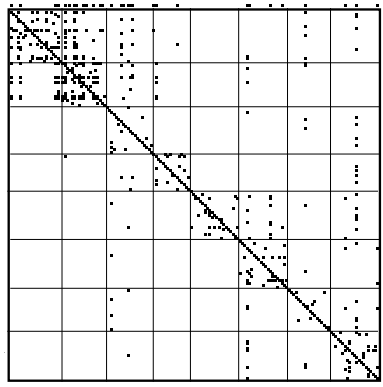
Een gerichte graaf met N knopen wordt gewoonlijk beschreven door haar verbindingsmatrix $A = (a_{ij})_{i,j=1,\ldots,N}$, gedefiniëerd door

$$a_{ij} = \begin{cases} 1 & \text{als } i \to j, \\ 0 & \text{anders.} \end{cases}$$

Figuur 5.2 laat de verbindingsmatrix zien die hoort bij de call-graaf `Java1`. De verbindingsmatrix kan worden gepartitioneerd door zogenoemde multi-level methoden, zoals het matrix-partitioneringsprogramma Mondriaan. Dit programma werd aangepast, zodat het toegepast kon worden op de voorbeelden van SIG. Het is *heuristisch* (niet exact) van aard, maar desondanks bleek het erg snel en betrouwbaar zodat de resultaten dicht bij de optimale oplossing liggen.

Vergelijking van de twee algoritmen

We hebben twee methoden ontwikkeld voor het probleem, namelijk een optimaal ILP- en een heuristisch HP-algoritme. Het is een interessante vraag hoe de kwaliteit en rekentijd zich verhouden tussen deze twee varianten. Daarom zijn voor een aantal praktijkproblemen, geleverd door SIG, de beide methoden gebruikt om een optimale partitionering in $L = 8$ modules te vinden, waarbij de maximale omvang van een module gelijk is aan $K = 1.2\frac{N}{L}$. Het softwarepakket CPLEX is gebruikt op een 400 MHz Sun Ultra-Sparc II computer om de optimale oplossing te vinden, terwijl Mondriaan een 867 MHz Apple PowerBook G4 computer gebruikte. Figuur 5.3 laat het resultaat van het hypergraafpartitioneringsalgoritme zien voor het probleem `Java1`. Elke kolom correspondeert met een net; gebroken netten zijn bovenaan gemarkeerd.



Figuur 5.3: Gepartitioneerde verbindingsmatrix (158×158) van het probleem `Java1`. De rijen en kolommen zijn zodanig gepermuteerd dat ieder rij-blok bij één module hoort.

In Tabellen 5.1 en 5.2 is de interfaceomvang $|I|$ en de rekentijd weergegeven voor beide methoden. De verschillende call-grafen hebben N knopen en $|A|$ pijlen. Als we de resultaten vergelijken zien we dat de heuristische methode hele goede resultaten levert. De ILP methode leidt tot de optimale interface; de interface verkregen door de heuristische HP methode is groter, echter hoogstens een factor 1.63. Aan de andere kant is de rekentijd van de HP methode veel kleiner en lineair evenredig met $|A|$. De rekentijd van de ILP methode is over het algemeen veel hoger en heeft een onvoorspelbaar karakter.

Conclusie

Het splitsen van een groot softwaresysteem in kleinere eenheden is voor veel organisaties een uitdaging geworden. Tijdens de "Study Group Mathematics with Industry"maakten we kennis met dit probleem, waar het werd geïntroduceerd door SIG. Tijdens de week van de studiegroep en erna hebben we gewerkt aan een duidelijke wiskundige formulering van het probleem. Tevens hebben we twee verschillende aanpakken gevonden, die beide erg geschikt en waardevol kunnen zijn voor de bedoelde

Probleem	N	$\|A\|$	K	L	opt. $\|I\|$	$\|I\|/N$ (%)	rekentijd (s)
`Java1`	144	422	23	8	26	16.5	103
`Java3`	837	5252	127	8	242	28.4	246456^a
`Java4`	15	39	2	8	11	68.8	0.22
`Cobol1`	947	1900	209	8	13	0.9	118
`Cobol2`	449	659	81	8	6	1.1	351
`Cobol3`	1145	2686	203	8	51	3.8	6452
`Cobol4`	1100	2951	167	8	32	2.9	742172^a

Tabel 5.1: Resultaten met 0-1 lineair programmeren. De met 'a' gemarkeerde tijden zijn afkomstig van een andere computer.

Probleem	N	$\|A\|$	beste $\|I\|$	gem. $\|I\|$	$\|I\|/N$ (%)	rekentijd (s)
`Java1`	158	580	30	30.7	19.0	0.06
`Java3`	851	6103	275	283.2	32.3	0.54
`Java4`	16	55	11	11.2	68.8	0.001
`Cobol1`	1398	3298	17	22.4	1.2	0.33
`Cobol2`	545	1204	10	11.5	1.8	0.12
`Cobol3`	1357	4043	69	74.6	5.1	0.34
`Cobol4`	1116	4067	52	56.5	4.7	0.41

Tabel 5.2: Resultaten met hypergraafpartitionering.

toepassingen. De wiskundige formuleringen reduceerden het probleem tot standaardproblemen uit de discrete optimalisering. Zodoende konden moderne algoritmen gebruikt worden om de praktijkproblemen van SIG op te lossen. Onze twee methoden kunnen het relatieve aantal interfaceprogramma's drastisch verlagen. Voor Cobolsystemen is het zelfs mogelijk dat de omvang van de interface minder dan 5% wordt. Voor kleine problemen bevelen wij aan de optimale ILP methode te gebruiken. Voor grote problemen, met duizenden knopen in de call-graaf, is multilevel hypergraafpartitionering de enige realistische optie.

Het bedrijf SIG waardeerde de voorgestelde oplossingen, evenals het inzicht dat verkregen kon worden door middel van ons onderzoek. Aan de andere kant hebben wij geprofiteerd van dit goed gestelde probleem wat ons leidde tot nieuwe interessante vragen zoals de vergelijking tussen een exacte en een heuristische methode in een realistische situatie.

6 Mathematische technieken voor neuromusculair onderzoek

JF Williams, Geertje Hek, Alistair Vardy, Vivi Rottschäfer, Jan Bouwe van den Berg en Joost Hulshof

Binnen de neurowetenschappen zijn er tegenwoordig ontzettend veel manieren om onderzoek te doen. Ethische en praktische factoren beperken echter de toepasbaarheid van veel technieken bij onderzoek naar het zenuwstelsel. In veel spieronderzoek is het niet wenselijk en soms zelf vervelend voor de proefpersoon om direct aan het zenuwstelsel zelf metingen te verrichten. Metingen aan de buitenkant van het lichaam geven een ratjetoe van alle informatie die zich onder de huid afspeelt. Het ontrafelen van deze informatie is waar we in dit artikel op uit zijn.

α-motorneuronen en motor-eenheden

We gebruiken onze spieren om te bewegen. Ons zenuwstelsel activeert onze spieren door er elektrische pulsen naartoe te sturen. Deze elektrische pulsen gaan over het spieroppervlak waardoor de spier samentrekt. Als we beter kijken, zien we dat spieren zijn opgebouwd uit kleinere eenheden, spiervezels genaamd. Eén spier wordt aangestuurd door een heleboel zenuwen. Deze zenuwen heten alpha-motorneuronen. Elk van deze zenuwen stuurt een aantal spiervezels aan. Een alpha-motorneuron en zijn spiervezels vormen een geheel, dat we een motor-eenheid zullen noemen. Kleine spieren hebben 10 van deze motor-eenheden. In grote spieren kan dit aantal oplopen tot 300.

Als we onze spieren aanspannen, zijn er motor-eenheden actief. Niet alle motor-eenheden zijn tegelijk actief. Als we onze spieren vrijwillig aanspannen worden eerst de kleinere motor-eenheden geactiveerd en pas bij zwaardere inspanning ook de grotere. Bij vrijwillige inspanning is de recruteringsvolgorde dus van klein naar groot. Op deze manier kunnen we op verschillende inspanningsniveaus de kracht goed doseren. Naast vrijwillige inspanning is het ook mogelijk een spier kunstmatig te activeren, Door kleine elektrische schokjes toe te dienen Als er op deze manier inspanning wordt geleverd, dan is de recruteringsvolgorde van groot naar klein. Bovendien worden bij kunstmatige stimulatie van een spier alle motor-eenheden op hetzelfde moment geactiveerd, terwijl bij vrijwillige inspanning de motor-eenheden op verschillende tijdstippen worden geactiveerd.

Bij spierziekten zoals polio of Duchenne spierdystrofie verzwakken de spieren. Bij polio komt dit doordat de verbindingen tussen de alpha-motorneuronen en de spiervezels afsterven. Als een verbinding afsterft, dan raakt de spier een motor-eenheid kwijt. Bij Duchenne spierdystrofie zorgt het ontbreken van een specifiek spiereiwit ervoor dat spiercellen afsterven. Ook dit leidt tot een vermindering van motor-eenheden. Om de progressie van spierziekten bij te kunnen houden, zouden we graag weten hoeveel motor-eenheden een spier bevat.

In dit artikel bekijken wij een techniek om het aantal motor-eenheden in een spier te bepalen. Een veelgebruikte techniek bij neurologisch onderzoek naar spieren is het meten van de elektrische activiteit op de huid. Een probleem hierbij is dat de gemeten

signalen niet alleen van de spieren komen. Elektrische signalen van het hart en de andere organen worden ook door de huid geleid. De geleidende eigenschappen van de huid worden ook benut door bijvoorbeeld leugendetectoren. Het totaal dat we meten, is een combinatie en vervorming van signalen. Gelukkig is de activiteit van spieren krachtig. Als we vlak boven een spier meten, is dit signaal krachtig genoeg om de overige signalen te overstemmen. Het meten van de elektrische activiteit van de spieren wordt elektromyografie (EMG) genoemd. Onze interesse gaat uit naar het meten van de activiteit van de motor-eenheden. Om nauwkeurige metingen te verrichten, worden de spieren bedekt met een matje van elektroden. Bij een vrijwillige of (met elektrische schokjes) gestimuleerde inspanning zien we activiteit die is opgebouwd uit de activiteit van verschillende motor-eenheden. Helaas liggen de spiervezels van motor-eenheden niet netjes gebundeld, maar kriskras door elkaar. Dit betekent dat de elektrische activiteit die binnenkomt bij elke elektrode op het matje een mengelmoes van signalen is.

Om het probleem te illustreren, voeren we een gedachtenexperiment uit. Stel we zouden elk van de motor-eenheden één voor één kunnen activeren. Elke motor-eenheid zou dan zijn eigen 'vingerafdruk' achterlaten. Als we elke vingerafdruk zouden kennen, dan is het probleem gereduceerd tot het bepalen van het aantal verschillende vingerafdrukken in de gemeten signalen. Helaas geven de motor-eenheden hun vingerafdrukken niet zomaar prijs. Een manier om dit probleem te omzeilen, is om een serie van elektrische schokjes toe te dienen in oplopende sterkte. Bij de eerste schokjes is slechts de grootste motor- eenheid actief. Daarna komen één voor één de kleinere motor-eenheden erbij. We zien op deze manier de optelsom van de vingerafdrukken. Door van het gemengde signaal de vorige af te trekken, krijgen we de individuele vingerafdrukken. Deze aanpak werkt mits het totaal aan signalen niet te vervuild is. De vervuiling die optreedt, noemen we ruis. Het is noodzakelijk dat de activiteit van de motor-eenheden groter is dan de ruis om ze nog te kunnen onderscheiden. Op dit moment wordt dit proces met de hand uitgevoerd. Dit is echter een tijdrovende bezigheid die moet worden uitgevoerd door een expert, en waarvoor geavanceerde apparatuur nodig is. Ons doel is een manier te vinden om het aantal motor-eenheden te bepalen op een automatische, eenvoudig uit te voeren wijze. Als we aannemen dat alle vingerafdrukken uniek zijn, hoeven we alleen te weten hoeveel verschillende componenten er aanwezig zijn in het totaal van signalen. Hierbij biedt de wiskunde een uitkomst.

Bepaling van principale componenten voor data met vertragingen

Gegeven een verzameling data $X \in \mathbb{R}^{m,n}$ is het een klassieke taak om het aantal principale componenten te bepalen. Over de gemeten data worden typisch aangenomen dat ze van de vorm

$$X_{ij} = \sum_{k=1}^{N} C_{ik} v_k(t_j) + \eta_{ij} \tag{6.1}$$

zijn, waar $X, \eta \in \mathbb{R}^{m,n}$, $C \in \mathbb{R}^{m,N}$, en $\{t_j\}_{j=1}^{m}$ zijn de tijdstippen waarop het signaal gesampled is. Dat wil dus zeggen dat de n signalen kunnen worden uitgedrukt als lineaire combinaties van een klein aantal N ($N \ll n$) basisvectoren, plus ruis η_{ij}. In dit probleem is het voornamelijk van belang om de dimensie N van de opspannende

verzameling $\{v_k\}$ te bepalen, corresponderend met het aantal actieve motor-eenheden.

Bij afwezigheid van ruis kan het probleem worden opgelost door de data te ontbinden in hun *principale componenten* door middel van een *singuliere waarden decompositie*: we vinden matrices $U \in \mathbb{R}^{m,m}, V \in \mathbb{R}^{n,n}, \sigma \in \mathbb{R}^m$ zodat

$$U^T X V = [\text{diag}(\sigma) 0_{n,m-n}].$$

Bovendien zijn U en V orthogonale matrices en $\sigma_i \geq \sigma_{i+1} \geq 0$. Van belang hierbij is dat als de matrix rang $N < n$ heeft, dan $\sigma_{N+1} = \ldots = \sigma_n = 0$. In dat geval zijn er precies N principale componenten. In het algemeen zijn echte meetgegevens niet zo netjes, en experimentele afwijkingen zorgen ervoor dat $\sigma_i > 0$ voor alle i. Maar de relatieve grootte van de principale waarden σ_i kan ons nog steeds veel informatie geven. In het bijzonder is σ_{k+1} gelijk aan de afstand van X tot de verzameling van matrices met rang k:

$$\sigma_{k+1} = \min_{\text{rang}(Y)=k} \|X - Y\|_2.$$

Gegeven een drempelwaarde ε kunnen we de *ε-rang* r_ε van een matrix te definiëren door te eisen dat

$$\sigma_{r_\varepsilon} > \varepsilon \geq \sigma_{r_\varepsilon+1}.$$

In het voorliggende probleem kunnen we deze ideeën niet direct topassen omdat er *twee* verschillende bronnen van fouten zijn:

1. Experimentele fouten van onbekend type en grootte *in* elk kanaal ("ruis").
2. Vertragingen van onbekende grootte *tussen* alle paren van signalen.

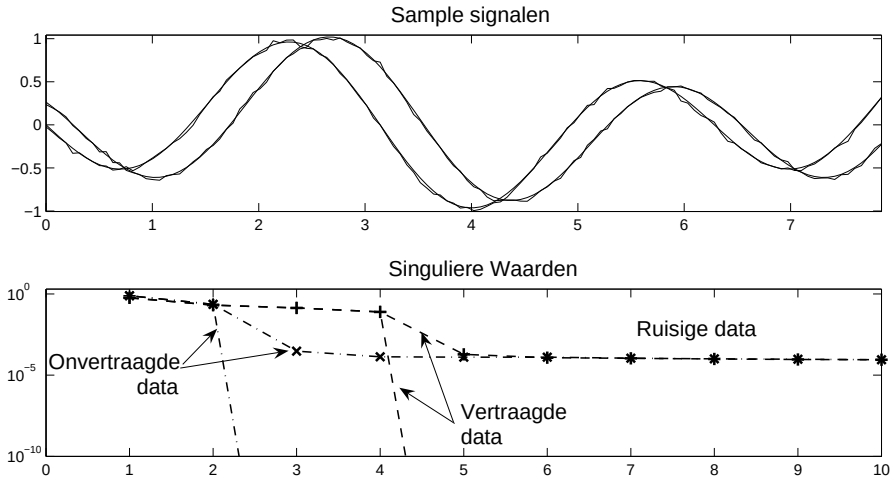
Fouten van de eerste soort zijn wijdverbreid en niet bijzonder problematisch; fouten van het tweede type moeten we nader bekijken. De metingen beginnen niet allemaal op hetzelfde tijdstip in relatie tot het begin van het signaal in de spieren. Hierdoor zijn de metingen ten opzichte van elkaar enigszins vertraagd. Laten we aannemen dat het tijdvertraagde signaal gegeven wordt door (cf. (6.1))

$$\widehat{X}_{ij} = \sum_{k=1}^{N} C_{ik} v_k(t_j + s_i) + \eta_{ij},$$

waarbij s_i de tijdsvertraging in de i-de meting is. Dan

$$\begin{aligned}
\widehat{X}_{ij} &= \sum_{k=1}^{N} C_{ik} \left(v_k(t_j) + s_i v_k'(t_j) + \frac{s_i^2}{2} v_k''(t_j) + \ldots \right) + \eta_{ij} \\
&\simeq \sum_{k=1}^{N} C_{ik} \left(v_k(t_j) + s_i v_k'(t_j) \right) + \eta_{ij} \qquad \text{(bij kleine vertragingen)} \\
&= \sum_{k=1}^{N} C_{ik} v_k(t_j) + \sum_{k=1}^{N} \tilde{C}_{ik} \tilde{v}_k(t_j) + \eta_{ij}.
\end{aligned}$$

De functies $\tilde{v}_k$ kunnen nu niet als lineaire combinaties van de v_k geschreven worden, dus zelfs zonder ruis zouden we de rang van de matrix en de ogenschijnlijke dimensie van de basis van principale componenten hebben verdubbeld.



Figuur 6.1: Boven: voorbeeldsignaal met typische vertraging en kleine ruis. Onder: de singuliere waarden voor de data zonder ruis nemen af tot $\sigma_i = 10^{-40}$ na $N = 2$ respectievelijk $N = 4$ voor de onvertraagde en vertraagde data. Kiezen we in dit voorbeeld $\varepsilon = 10^{-3}$ dan doet de toevoeging van ruis niets af aan de bepaling van de ε-rang, maar de vertragingen verdubbelen de schatting voor het aantal principale componenten.

We geven een klein voorbeeld om dit te illustreren. Beschouw matrices van de vorm

$$X_{ij} = \alpha_i \sin(t_j) + \beta_i \sin(2t_j), \qquad i = 1, \ldots, 10, t_j \in [0, 5\pi]$$
$$\widehat{X}_{ij} = \alpha_i \sin(t_j + s_i) + \beta_i \sin(2t_j + s_i)$$

waar α_i, β_i, s_i willekeurig gekozen zijn in $[-1, 1]$, en $\varepsilon = 10^{-3}$. In dit voorbeeld hebben we $N = 2$ en $n = 10$ genomen. In Figuur 6.1 tonen we een signaal met en zonder vertragingen en ruis en daaronder de bijbehorende singuliere waarden.

Dit voorbeeld toont dat we de singuliere waarde decompositie niet zomaar kunnen toepassen, zelfs niet met een drempelwaarde, om een goede schatting te krijgen voor de dimensie van de basis van principale componenten voor de vertraagde data met ruis. Om het aantal principale componenten te bepalen gaan we in twee stappen te werk. Eerst nemen we aan dat we de dimensie van de basis weten en zoeken we de beste/juiste verschuivingen, die de tijdsvertragingen "opheffen". Daarna zullen we proberen de juiste dimensie te bepalen.

Het komt er dus op neer dat we de verschuivingen zo zouden willen kiezen dat ze σ_{N+1} minimaliseren, even aannemende dat we N weten. In de praktijk blijkt dat echter een zeer gedegenereerd en moeilijk probleem. In plaats daarvan bekijken we

$$V = -\frac{\sum_{i=1}^{N} \sigma_i}{\sum_{i=1}^{n} \sigma_i} = -\sum_{i=1}^{N} \sigma_i \qquad \text{(door normalisatie)}. \tag{6.2}$$

Hierbij nemen we dus aan dat N een vast getal is. Als de singuliere waarden snel

afnemen voor $i > N$ dan

$$-\sum_{i=1}^{N} \sigma_i \simeq -\sum_{i=1}^{n} \sigma_i + \sigma_{N+1} \simeq -1 + \sigma_{N+1}$$

, en in zulke gevallen is de minimalisatie van (6.2) een goede benadering voor σ_{N+1}.

Het is goed om hier op te merken dat singuliere waarde decompositie een *lineair* probleem is, terwijl het vinden van de beste verschuivingen (om de vertragingen te compenseren) een essentieel *niet-lineair* probleem is. We hebben drie methoden getest om de functionaal in (6.2) te minimaliseren met betrekking tot de verschuivingen:

A1. Een pseudo-Newton methode voor het vinden van een nulpunt van ∇V met behulp van `minunc` in de Optimalisatie Toolbox van Matlab.

A2. Een gradiënt evolutie van V met betrekking tot een artificiële tijd door middel van

$$\frac{ds_i}{dt} = -\frac{\partial V}{\partial s_i}$$

via de code `ode113` in Matlab.

A3. Direct zoeken met de Matlab routine `fminsearch`.

De resultaten van de algoritmes A1 en A2 zijn vrijwel identiek, maar A2 kost ongeveer tien keer zoveel tijd. De resultaten van A3 zijn minder bevredigend en zijn tien keer zo langzaam als A2.

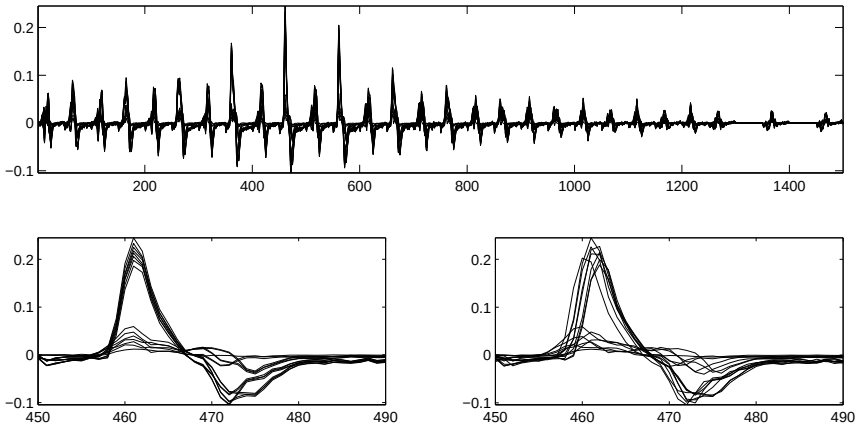
We nemen nu aan dat we redelijke schattingen hebben voor het minimale aantal $N_{\min}$ en het maximale aantal $N_{\max}$ principale componenten. Dan voeren we de minimalisatie procedure uit met betrekking tot de verschuivingen. Dit leidt, voor vaste N, tot de beste teruggeschoven data.

In het ideale geval zouden we graag een duidelijk stap willen zien van grote naar kleine singuliere waarden (zoals in het voorbeeld in Figuur 6.1), zodat we een duidelijke ε-rang kunnen bepalen. Helaas lijkt het niet mogelijk om op goede gronden een waarde van ε te kiezen. In plaats daarvan bepalen we voor elke (vaste) $N \in [N_{\min}, N_{\max}]$ de maximale verhouding tussen opeenvolgende singuliere waarden:

$$m(N) = \max_k \frac{\sigma_k}{\sigma_{k+1}}. \tag{6.3}$$

Laat $N_r(N)$ de waarde van k zijn waarvoor dit maximum wordt aangenomen. $N_r(N)$ is dan de schatting voor het aantal principale componenten als we de verschuivingen geoptimaliseerd hebben voor de N grootste componenten. Voor consistentie is duidelijk vereist dat voor het werkelijke aantal principale componenten geldt dat $N_r(N) = N$. Natuurlijk kunnen er meerdere waardes van N zijn waarvoor aan deze consistentie eis is voldaan. Als beste schatting voor het werkelijk aantal principale componenten, kiezen we de N die de verhouding $m(N)$ maximaliseert. Conceptueel vindt deze benadering de dimensie waarin er de duidelijkste stap van "grote" naar "kleine" componenten is. Aan de andere kant, dit sluit niet uit dat de data in elke meting bestaat uit één hele grote component en verscheidene kleinere. In dat geval blijkt het erg moeilijk om de kleine signalen van de ruis te onderscheiden omdat de vertragingen het klassieke signaal-versus-ruis probleem versterken.

Samenvattend is het totale algoritme als volgt:



Figuur 6.2: Artificiële data. Boven: het complete signaal in alle kanalen. Linksonder: detail van de onvertraagde data. Rechtsonder: detail van de vertraagde data. Na optimalisatie van de verschuivingen zijn de originele data en de teruggeschoven data in essentie niet te onderscheiden.

1. Minimaliseer V in (6.2) met betrekking tot verschuivingen, $N \in [N_{\min}, N_{\max}]$.
2. Bepaal $m(N)$ in (6.3) voor de teruggeschoven data.
3. Maximalisatie van $m(N)$ met betrekking tot N leidt tot de schatting van het aantal principale componenten.

In Figuur 6.2 laten we een test met artificiële data zien. Deze zijn "gesimuleerd" door een van de experimentatoren. De resulaten van het algoritme zijn samengevat in Tabel 6.1.

Conclusies

We hebben gezien dat het algoritme succesvol is in het reduceren van het effect van de vertragingen in de data, zodat het een betrouwbare schatting voor het aantal principale componenten geeft. Na voltooiing van dit verslag zijn we voorzien van "echte"

N	N_r	$\sigma_{N_r}/\sigma_{N_r+1}$	Rekentijd (s)
6	6	125	47
5	5	98	45
4	4	2731	33
3	4	113	39
2	4	31	31

Tabel 6.1: Resultaten van het algoritme met verschafte "gesimuleerde" data. We concluderen hieruit terecht dat er vier principale componenten zijn.

data, en onze voorlopige conclusie is dat de benadering met het beschreven algoritme hierbij ook goed lijkt te werken, maar dat de gradiënt evolutie (algoritme A2) het meest geschikt is. Hier moet nog diepgaander naar gekeken worden. Er is nog een punt waarop progressie kan worden geboekt. We hebben namelijk geen informatie gebruikt over de structuur van de signalen, terwijl dat ons toch zou moeten kunnen helpen bij het ontrafelen van het juiste aantal principale componenten.

De vraag die we hier bekeken hebben bestaat eigenlijk uit twee componenten. Hoe kunnen we de data terugschuiven zodat de vertragingen worden opgeheven? En hoe kunnen we daarna het aantal principale componenten bepalen? We denken dat we de eerste vraag beter kunnen beantwoorden. Dat is niet verrassend, want de tweede vraag is een oud probleem dat niet volledig is opgelost in veel praktische situaties.

Een andere manier om inzicht te krijgen in het werken van motor-eenheden is te kijken naar het gedrag van de zenuwen die de spieren aansturen, de alpha-motorneuronen. In ditzelfde onderzoek hebben we ook een stochastisch model voor een alpha-motorneuron opgesteld om het activatiegedrag te onderzoeken. Een zenuw krijgt invoer van andere zenuwen in de vorm van elektrische pulsjes. De zenuw reageert op het totaal van deze pulsen door zelf ook pulsen te genereren. Een alpha-motorneuron krijgt zijn invoer van het centrale zenuwstelsel en stuurt zijn pulsen naar zijn spiervezels. Het model is ontworpen om pulsen van het centrale zenuwstelsel op te vangen en daarop zelf pulsen te sturen. Simulaties van het model geven een realistisch beeld van het pulsgedrag van een alpha-motorneuron.

The Mathematical Modelling of Cooling and Rewarming Patients during Cardiac Surgery

Marcus Tindall[*] *Mark Peletier*[†] *Joyce Aitchison*[*]
Simon van Mourik[‡] *Natascha Severens*[†]

Abstract

The process of cooling bodies, by the use of a heart lung machine (HLM), is utilised in a number of surgical procedures primarily to reduce the metabolic rate of the organs and hence their consumption of oxygen. On completion of surgery the blood is rewarmed by the HLM. A major consequence at the end of this process is afterdrop: a rapid decrease in the core organ temperature as a result of spatial temperature differences between the core organs and remainder of the body, which can lead to post-operative complications. This report details two mathematical models developed to understand heat transfer processes between the core organs, rectal region and peripheral body parts (primarily skin, muscle and fat). A one compartment spatially independent model, describing the temperature distribution of a single tissue type through which blood perfuses, shows that temperature dependent perfusion reproduces the observed differences in blood and tissue temperatures, whilst temperature independent perfusion does not. The model is extended to account for heat transfer between the blood pool (core), rectal regions and periphery. This three compartmental model is able to qualitatively reproduce the observed temperature differences in the three regions. Analysis of the model shows that a period of constant warming at the end of the rewarming period has a positive effect in reducing afterdrop.

1 Introduction

The cooling of the body of patients during surgical procedures has been used for a number of years in a variety of operations [1]. Cooling is primarily used to reduce the metabolic rate of organs within the body and thus the amount of oxygen they consume. This has two purposes: (1) it is less likely that irreparable damage to vital organs will occur due to oxygen deficiency; and (2) it allows the surgeon more time should some unusual complications occur during the surgical procedure.

In this report we are concerned with the application of the procedure to patients undergoing cardiac surgery. During cardiac surgery with cardiopulmonary bypass - the majority of cardiac surgical interventions - cooling is performed by means of a heart lung machine (HLM). The process consists of six distinct phases as detailed below and shown in Figure 1.

[*]University of Oxford
[†]Technische Universiteit Eindhoven
[‡]Universiteit Twente

1. The patient is anaesthetised whereby the body temperature drops naturally by approximately 2°C (not shown in Figure 1).

2. The first stage of the surgery begins. This consists of the chest cavity being opened and the area around the heart (muscle and tissue) being prepared for the main surgical procedure.

3. The body is connected to a HLM whereby the blood is circulated through the machine and the temperature of the blood lowered at the instruction of the cardiac surgeon in agreement with the perfusionist. Blood from the HLM generally enters the body through a tube inserted in the aorta. A HLM contains a simple heat-exchanger whereby the cooling or warming fluid is generally water.

4. The main cardiac surgical procedure takes place, during which the body is kept at a constant cooled temperature. The temperature during surgery depends on the surgical intervention e.g. for aorta valve replacements and coronary artery bypass grafts 30°C is a common temperature, whilst during surgery on the aortic arch the patient is cooled to 16-18 °C.

5. Following completion of the surgical procedure the blood is warmed at a steady rate. Rewarming must not take place too quickly for large spatial or temporal differences in temperature can cause damage to cells and, on the larger scale, irreversible damage to organs.

6. Once the core organs have reached a certain temperature the patient is disconnected from the HLM and the temperature of the body allowed to self-equilibrate. This often results in a phenomena known as afterdrop: an observed decrease in the temperature around the core organs. The afterdrop effect is considered to be a result of the large temperature difference between the core and peripheral regions. Patients who experience a large afterdrop in temperature often take longer to recover and may experience further post-operative complications. Thus clinicians try to minimise the afterdrop effect as much as possible.

The current protocol for both cooling and rewarming patients is very much an *ad hoc* procedure, i.e. it relies on the expertise and experience of the surgeon, perfusionist, and anaesthetist. For instance, in the course of recent years the target temperature for average operations, such as aortic valve replacements and coronary artery bypass grafts, has risen from 28°C to 30°C, as the result of clinical experience.

Previous mathematical modelling in the area has focused on developing models describing the cooling/rewarming process in specific body parts as well as the temperature distribution throughout the whole body, the models varying in both detail and complexity. Pennes [4] described the temperature distribution in the tissue and arterial blood of the human forearm using a standard formulation of the heat equation which takes account of perfusion and possible heat sources and/or sinks. Curtis and Trezek [1] have formulated a five compartmental model which accounts for heat transfer between the core organs, muscle, fat, skin and the blood.

Fiala et al. [2, 3] have developed a computational model for predicting human thermal body regulation. They consider the body to consist of two parts: the passive system and the active system. The passive system describes heat transfer both

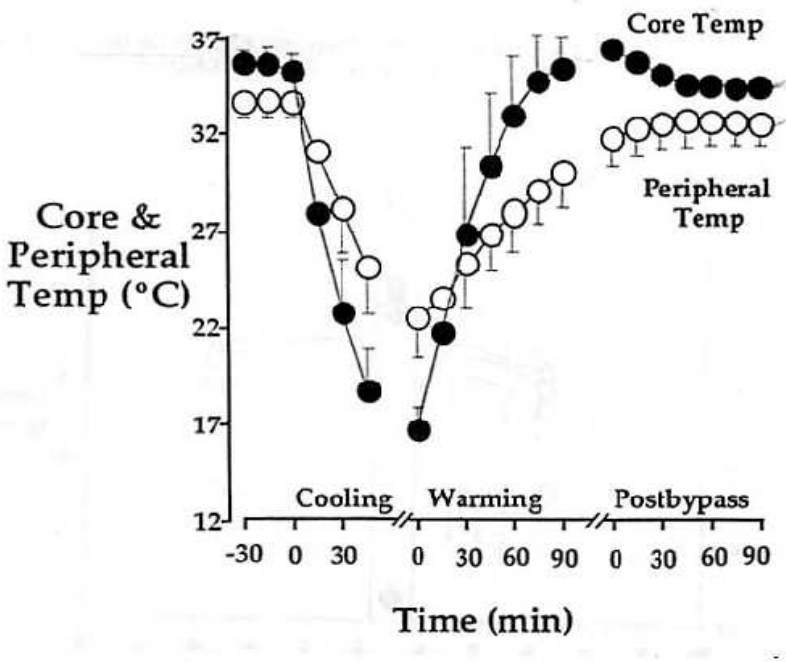


Figure 1: A typical temperature curve for the cooling and rewarming process. Taken from [5].

within the body and between the body and the external environment. The active system describes the thermoregulatory responses of the body when the body temperature deviates from its normal temperature e.g. vasodilation, vasoconstriction, shivering or sweating. They construct a whole-body model of an average human (73.5kg, 14% body fat) which consists of 15 spherical or cylindrical segments representing the head, arms, thorax, etc. Pennes' equation [4] is applied to each of the segments with the appropriate boundary and initial conditions.

The Study Group was asked to focus on developing suitable mathematical models which would:

1. help explain the observed differences in temperatures recorded in different parts of the body (mainly the core organs, rectum and peripheral body regions - skin, fat and muscle);

2. assist in developing more robust and physiologically based cooling and warming protocols for different size patients, e.g. fat versus thin, tall versus short; and

3. provide insight into possible causes or reasons for the afterdrop effect occurring and how this effect may be reduced or remedied.

This report details two models developed during the Study Group to help answer these questions. Section 2 details experimental data available on the problem. Sections 3 and 4 detail two models: the first is a simple model developed in order to

Body part	Perfusion rate - W [l/(s · m^3)]	Tissue conductivity - D $\times 10^{-7}$ [m^2/s]
Main Organs	4-10	1.0-2.0
Muscle	0.5	1.6-4.0
Skin	1-10	2.0
Fat	3.6 $\times 10^{-3}$	0.64

Table 1: Typical perfusion and conductivity coefficients for tissue in different parts of the human body. Taken from [2].

reproduce some of the observations obtained in the cooling and rewarming procedure. The second model or 'Amsterdam' model extends the first, but focuses on developing a model specifically targeted at explaining the observed temperature differences between the core organs (including the brain), rectum and peripheral regions. Comparisons between the models are made and their application and areas of possible further development are discussed in Section 5.

2 Clinical and experimental data

A number of data sources exist on the temperature of bodies during both the cooling and rewarming process for different surgical procedures. Given the detail of the model developed by Fiala et al. [2, 3] their papers provide a good source of data on rates of heat transfer (via both perfusion and conduction) and the capacitance of specific body parts and organs as shown in Table 1. Here perfusion refers to the amount of fluid (blood) moving through a certain volume of tissue per unit time. By means of this blood flow, convective heat transfer takes place. This is in contrast to conductive heat transfer, which refers to the transport of heat through tissue or the surrounding medium (air) which is dependent upon the conductivity of the medium.

The large variation in perfusion rates for different parts of the body detailed in Table 1 causes us to think carefully about how this affects heat transfer through the body in the context of the surgical cooling/rewarming procedure. As blood from the HLM enters the aorta it is pumped around the body, reaching the core organs first. Thus given the high rate of perfusion in this area the temperature quickly rises to that of the blood temperature. However, in regions which are not so well perfused such as the peripheral regions, in particular fat, the rate of heat transfer will be greatly reduced. Indeed heating such regions will take considerably longer although the tissue conductivity of heat between the core and periphery is approximately the same.

Given this data, in what follows the body is broadly considered to consist of three lumped regions:

1. the core organs - the main organs of the body found between the lower abdomen and up to and including the brain;
2. the rectum - including the large and small intestines and the rectal area; and
3. the periphery - muscle, skin and fatty tissue.

As the thermoregulatory responses of the body (the active system) are impaired during cardiac surgery, we will only consider passive heat processes.

3 A one-compartment model

We begin by formulating a simple one-compartment model which accounts for the transfer of temperature between the blood and tissue. In this model and the one detailed in Section 4 we assume that the temperature distribution is spatially homogenous and we are only interested in time dependent heat transfer between the defined compartments. The objective here is to see how well such a simplified model reproduces some of the qualitative features seen in Figure 1.

We consider a single human body in which heat transfer between the blood and the body tissue is dominated by convection through the rate of perfusion of the tissue. It is assumed that no heat losses take place between the outlet of the HLM and the aortic inlet of the patient. For simplicity we measure all temperatures in this model from a reference value of the normal blood temperature.

Let $f(t)$ be the temperature of the blood, which is determined by the HLM. Given the linear behaviour in the fall and rise of the core temperature seen in Figure 1, we define $f(t)$ by setting $f(0) = 0$ and

$$f'(t) = \begin{cases} -a, & 0 \leq t < t_1 \\ 0, & t_1 \leq t < t_2 \\ a, & t_2 \leq t < t_3, \end{cases} \tag{1}$$

where the cooling procedure commences at time $t = 0$, the main surgical procedure commences at $t = t_1$ (and the temperature is then kept constant), surgery finishes and the rewarming process is started at $t = t_2$, and the rewarming process finishes at $t = t_3$. Here a represents the rate of cooling and rewarming of the blood.

The temperature $T(t)$ in the tissue is then governed by

$$\frac{dT}{dt} = W^*(T)(f(t) - T) \tag{2}$$

where $W^*(T)$ is proportional to the perfusion rate of blood through the tissue.

We solve equation (2) with an initial tissue temperature $T(0) = T_0 < 0$. This matches the situation in Figure 1, where the peripheral temperature is initially below that of the core.

We have considered two scenarios for $W^*(T)$, namely

$$W^*(T) = W_0 \quad \text{and} \quad W^*(T) = W_0 e^{kT}, \tag{3}$$

where W_0 and k are both constants. This second scenario relates to the observations of Stolwijk [6] that the rate of perfusion of tissue varies with temperature as

$$W(T) \propto 2^{\frac{T - T_{ref}}{10}} \tag{4}$$

For simplicity we have thus assumed an exponential function to see what effect this has on the temperature distribution within the tissue.

In the case of $W^*(T) = W_0$ equation (2) reduces to

$$\frac{dT}{dt} = W_0(f(t) - T) \tag{5}$$

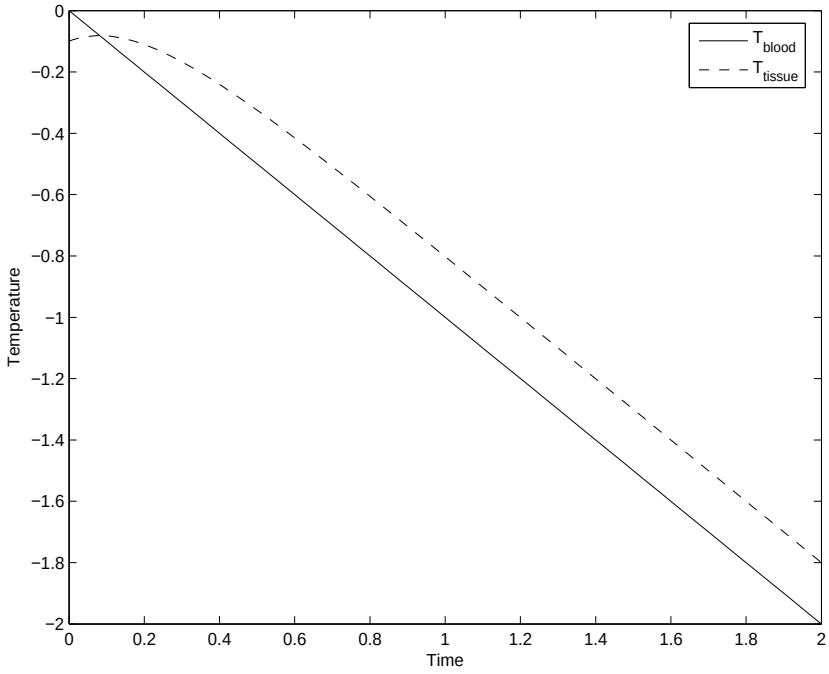


Figure 2: The effect of temperature reduction on the body tissue as predicted by the one compartment model. Here $W(T) = 5$ and $f'(t) = -1$.

which has an explicit solution. The part that is defined on $0 < t < t_1$ is

$$T(t) = -at + \frac{a}{W_0} + \Big(T_0 - \frac{a}{W_0}\Big)e^{-W_0 t} \qquad \text{for} \qquad 0 \leq t < t_1. \tag{6}$$

We note that for large t, $T(t) \to -a(t - \frac{1}{W_0})$. Hence the temperature decreases linearly, the difference between the blood and tissue temperature given by $\frac{a}{W_0}$. This is confirmed by a plot of the cooling part of the solution as shown in Figure 2. We note that although the model can reproduce the observed cross-over between blood and tissue temperature, the two curves remain parallel, given the linear dependence of T for long time-scales, and hence the two curves do not show the divergent behaviour evident in the cooling part of Figure 1.

We next include a temperature dependent perfusion rate to see if this produces any of the observed behaviour. In this second case of $W^*(T) = W_0 e^{kT}$ equation (2) is now given by

$$\frac{dT}{dt} = W_0 e^{kT}(f(t) - T), \tag{7}$$

where $f(t)$ is still given by equation (1). Figure 3 shows a numerical solution to this equation, again for the cooling part of the procedure. We note that this curve shows a more divergent behaviour between the temperature of the blood and that of the body tissue – a result which more closely resembles the behaviour shown in Figure 1.

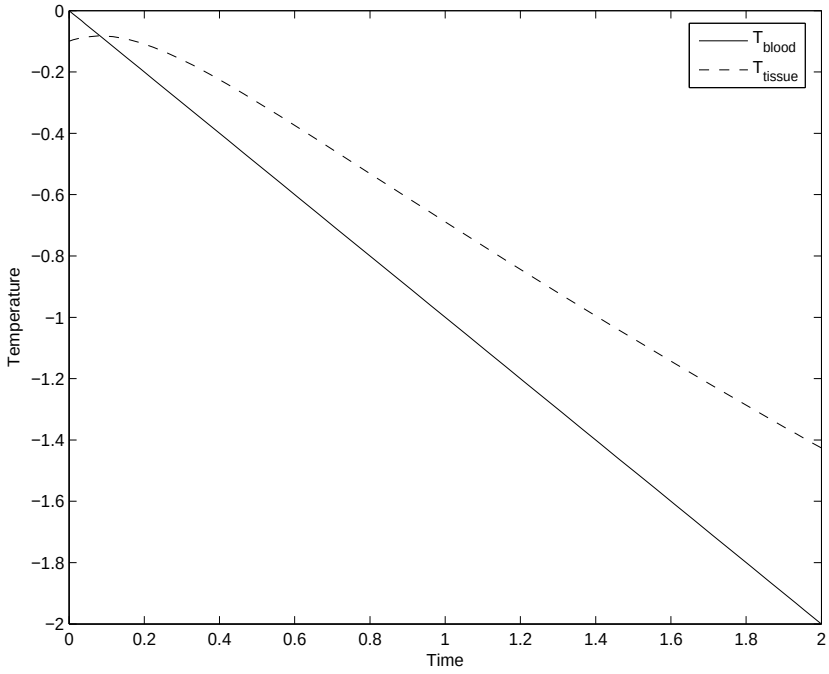


Figure 3: The effect of temperature reduction on the body tissue as predicted by the one compartment model. Here $W(T) = 5e^T$ and $f'(t) = -1$.

4 A three-compartment model – The 'Amsterdam' model

Whilst the simple one-compartment model formulated in the previous section has provided insight into the heat transfer process between the blood pool and the body tissue, we now turn our attention to the issue of understanding what causes afterdrop.

In order to do so we consider a three-compartment model consisting of a blood pool, rectal region (small and large intestines) and peripheral parts of the body. The blood pool is considered to be that of the core region of the body, whereby the heating of the core regions is assumed to be instantaneous. We further assume that heat transfer is dominated by the effects of perfusion of the blood through the rectal and peripheral compartments, except for the heat transport between the rectum and the periphery.

Whilst model data is available elsewhere in the literature for temperature recordings taken in other regions of the body, our focus here is on developing a model which accounts for temperatures recorded in both the nasal cavity, which we have taken to be part of the core, and the rectum. Given the low perfusion rate of the periphery, in particular fat, and its relatively high conductivity, the periphery may actually act as a heat source during the cooling procedure and a heat sink during rewarming of the body. Hence we include it here to see what effect it will have on the overall blood (core) temperature.

The equations governing the heat transfer process are given by

$$V_C\rho_C c_C\frac{dT_B}{dt} = \begin{cases} -a, & 0\le t<t_1,\\ 0, & t_1\le t<t_2,\\ a, & t_2\le t<t_3,\\ -\rho_B c_B V_R W_R(T_B-T_R) & \\ \qquad -\rho_B c_B V_P W_P(T_B-T_P), & t_3\le t\le t_4, \end{cases} \quad (8)$$

$$V_R\rho_R c_R\frac{dT_R}{dt} = \rho_B c_B V_R W_R(T_B-T_R)-k_{RP}(T_R-T_P), \quad 0\le t\le t_4, \quad (9)$$

$$V_P\rho_P c_P\frac{dT_P}{dt} = k_{RP}(T_R-T_P)+\rho_B c_B V_P W_P(T_B-T_P), \quad 0\le t\le t_4, \quad (10)$$

where

- T_B, T_R and T_P are the temperatures of the blood, rectum and periphery;
- a is the rate of cooling and rewarming of the blood;
- W_R and W_P are the blood perfusion rates of the rectum and periphery;
- k_{RP} is the heat transport coefficient between the rectum and periphery;
- ρ_i and c_i $(i=B,C,R,P)$ represent the density and heat capacity of the blood, the core organs, the rectum, and the periphery; and
- V_i $(i=C,R,P)$ represent the volume of the core, rectum and periphery.

In comparison to the protocol discussed in Section 3 there is an additional period, $t_3<t<t_4$, during which the HLM is switched off and the body regulates its own temperature via the heat balance of the fourth part of equation (8).

We see that in this final period the model confirms that the total heat content of the body

$$V_C\rho_C c_C T_C + V_R\rho_R c_R T_R + V_P\rho_P c_P T_P$$

is conserved.

We assume that all parts of the body are approximately at the same temperature at the start of the cooling procedure such that

$$T_B(0)=T_R(0)=T_P(0)=T_0. \quad (11)$$

Data from Fiala et al. [2] gives that the ratios of densities and heat capacities in the different regions are approximately equal to unity. Hence we can simplify the above model by dividing throughout by the respective $\rho_i c_i$ on the left-hand side of each equation to yield

$$V_C\frac{dT_B}{dt} = \begin{cases} -a^*, & 0\le t<t_1,\\ 0, & t_1\le t<t_2,\\ a^*, & t_2\le t<t_3,\\ -W_R^*(T_B-T_R)-W_P^*(T_B-T_P) & t_3\le t\le t_4, \end{cases} \quad (12)$$

$$V_R\frac{dT_R}{dt} = W_R^*(T_B-T_R)-k_{RP}^*(T_R-T_P) \quad 0\le t\le t_4 \quad (13)$$

$$V_P\frac{dT_P}{dt} = k_{RP}^*(T_R-T_P)+W_P^*(T_B-T_P) \quad 0\le t\le t_4, \quad (14)$$

where W_i^*, k_{RP}^* and a^* are rescaled perfusion, transport and cooling rates given by

$$W_R^* = V_R W_R, \quad W_P^* = V_P W_P, \quad k_{RP}^* = k_{RP}/(\rho c) \quad \text{and} \quad a^* = a/(\rho_C c_C).$$

The presence of the pre-factor of the different volumes for each region allows us to see what effect such regions, in particular the periphery, may have on the temperature of the blood and rectum.

4.1 Parameter Values

The model was solved for estimated values of the volume of the core, peripheral, and rectal regions and the change in blood temperature as dictated by the HLM. Here we have assumed that the volume of the periphery and core are approximately the same and that of the rectal region is considerably less, i.e.

$$V_C = 40l = V_P \text{ and } V_R = 10l. \tag{15}$$

We assume that the rate of perfusion of the rectum and the rescaled coefficient of heat transport between the rectum and the periphery are approximately equal and for simplicity take

$$W_R^* = 1/s = k_{RP}^*.$$

The perfusion rate of the periphery is considerably less, and we take $W_P^* = 1 \times 10^{-3}/s$.

4.2 Model solutions and results

Equations (12)-(14) can be solved for a number of different scenarios, showing how the time length of each particular phase affects the heat distribution in the different regions.

We begin by considering a normal surgical procedure whereby the predicted temperature of the core, rectum, and periphery are shown in Figure 4(a). It is noted that the peripheral temperature drops at a much lower rate than that of the core and rectal temperature. This produces many of the qualitative features of the experimentally recorded temperatures shown in Figure 1. Here the afterdrop effect is quite considerable and the time taken for the three regions to reach an equilibrium temperature is relatively long.

Figure 4(b) shows the change in temperature in the three regions for a case in which the surgical procedure did not take place, i.e. some emergency may have interrupted the operation. Whilst the change in temperature between each of the three compartments is similar, the afterdrop in temperature is greatly reduced. This is because the periphery has not been allowed to cool for a long enough period; therefore during the rewarming procedure it is a reduced heat sink in comparison to the case of the normal surgical procedure.

The results of Figure 4(b) indicate that if the body can be re-warmed for a longer period of time then the effect of the peripheral and rectal regions acting as heat sinks could be reduced. In order to see whether this is the case we investigated holding the blood (core) temperature constant for a period following the rewarming process. The results for two different times (30 and 60 minutes respectively) of holding the

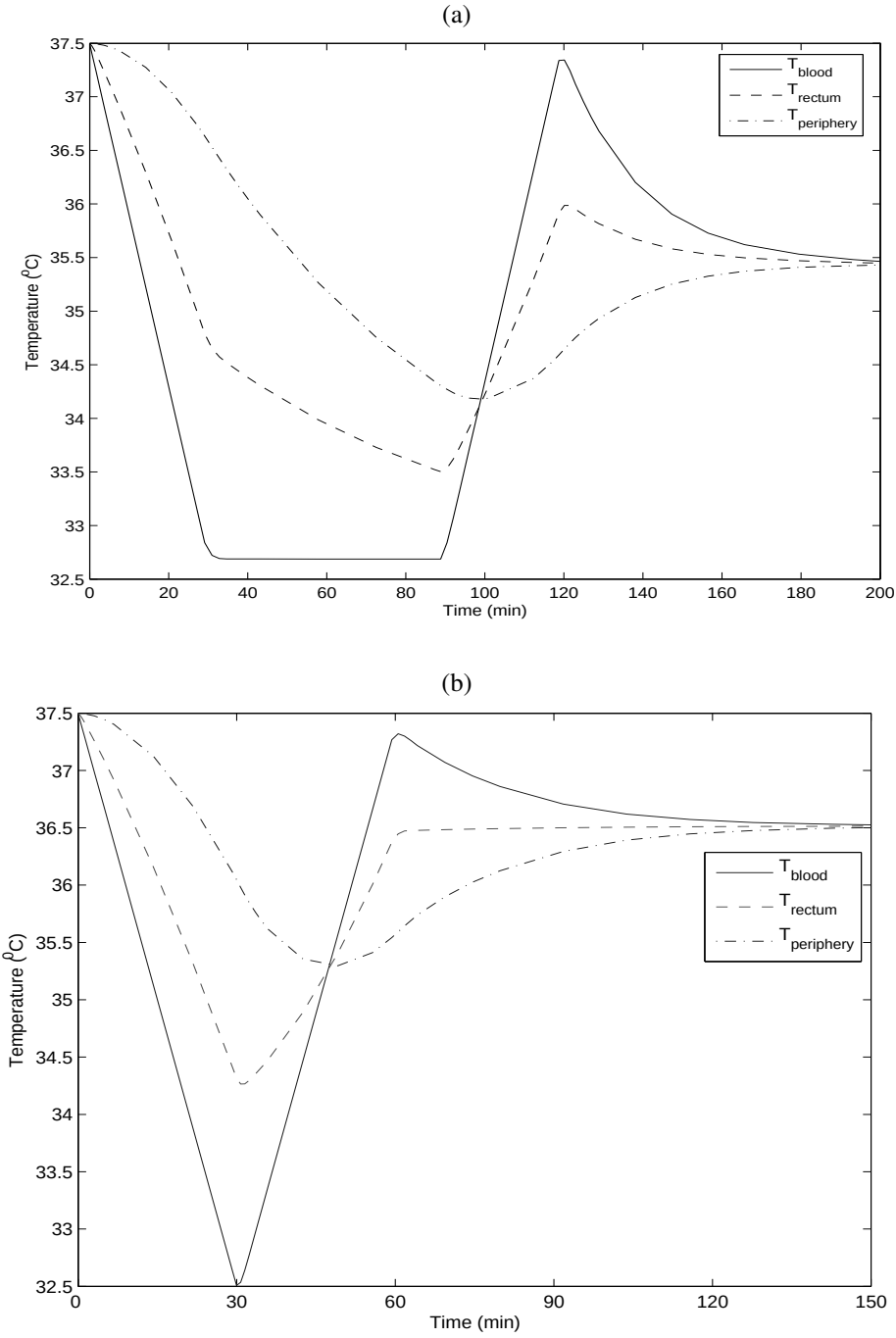


Figure 4: (a) The cooling and rewarming procedure for a surgical time of 60 minutes and (b) The effect of a short surgical time on the rewarming procedure. Note the small afterdrop in temperature.

temperature constant is shown in Figures 5(a) and (b). Here we note a significant reduction in afterdrop in each case whereby the 30 minute warming appears to halve the afterdrop effect and that of 60 minutes has a slightly greater effect.

5 Summary and discussion

We have formulated a number of models to understand cooling and rewarming of patients during cardiac surgery. A simple one-compartment model has reproduced some of the differences in qualitative behaviour between the core organ and peripheral (muscle, fat and skin) body regions. The assumption that the perfusion rate of the tissue is dependent upon the tissue temperature reproduced the correct behaviour in temperature change between the tissue and blood temperatures, specifically during the cooling process.

A three-compartment model was then formulated in order to understand the heat transfer process between the core organs, rectal region, and peripheral body parts and the effect that varying perfusion rates have. This model reproduced much of the observed behaviour in the temperature differences between each region and showed that a procedure of keeping the patient connected longer to the HLM at a high blood temperature, after the rewarming phase, has a considerable effect on reducing afterdrop. An initial analysis of the effect of excess body fat, i.e. fat versus thin people, on the rewarming procedure shows that fatter patients may not necessarily need to be rewarmed for longer given that the peripheral regions do not have enough time to drop to near the core temperature. However, this result appears to be at odds with clinical experience and requires further investigation. *Maybe also include explanation of Mark here...* One effect that has not been considered in this model is that of the temperature-dependent rate of perfusion considered in the one-compartment model. This may have important consequences for the rewarming of the peripheral body regions.

The 'Amsterdam' model has helped answer a number of questions raised during the Study Group, in particular how temperature distribution varies throughout the different body regions and most importantly how the effect of afterdrop can be reduced.

Suggested areas of future work include:

- establishing the importance of metabolism and its possible role during rewarming;
- understanding the affect that excess body fat has on the re-warming process;
- experimental and clinical testing of the results of the Amsterdam model;
- assessing the effect of temperature-dependent perfusion on the Amsterdam model;
- accounting for the effects of thermal cooling and heating. The effect of the use of thermal blankets during or at the end of the surgical procedure has not been taken into account. This effect may have important consequences on some of the rewarming recommendations detailed above;
- heat losses, for instances large evaporative heat losses in the open thorax during surgery.

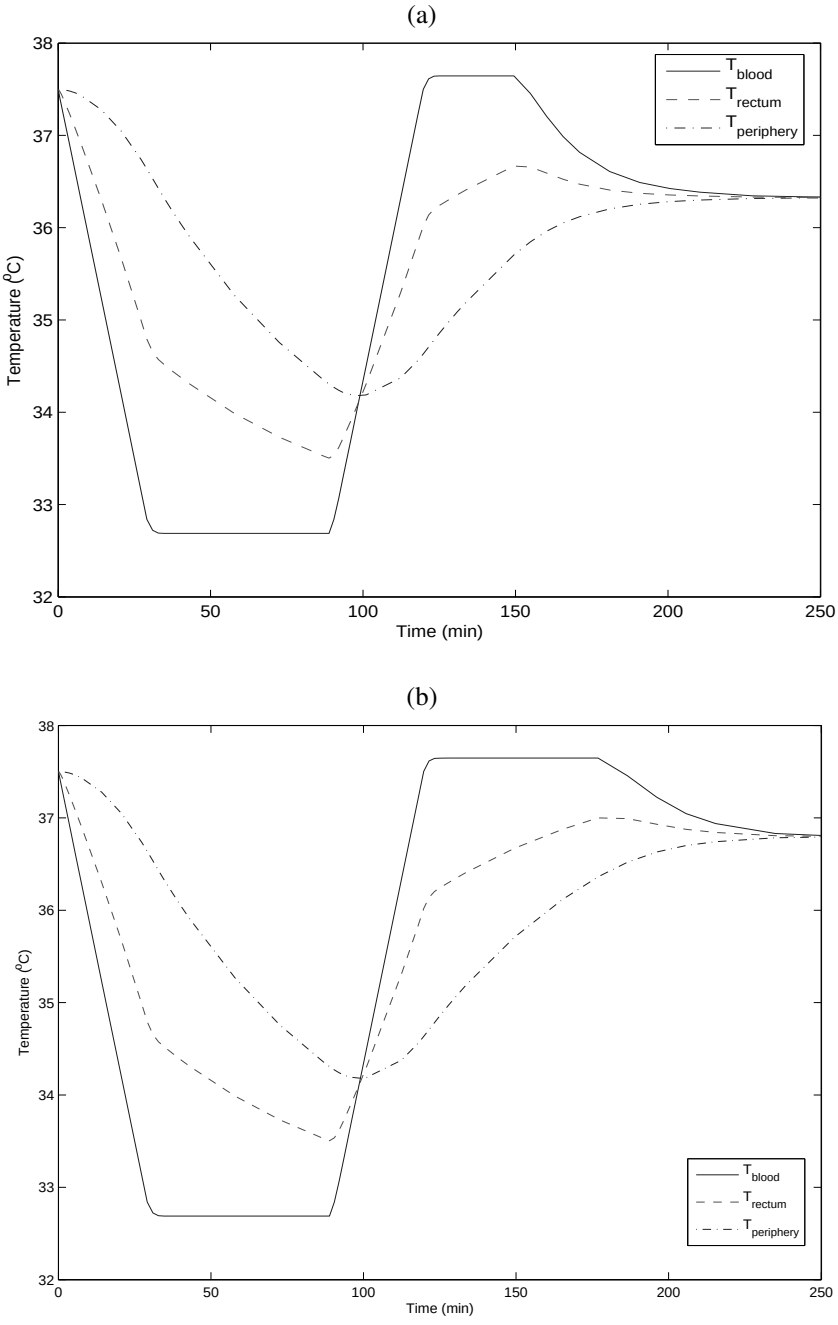


Figure 5: The effect of maintaining the core blood pool at a constant temperature at the end of the rewarming procedure using the HLM reduces the afterdrop effect. Here two cases are considered: (a) in which the core temperature is held constant for 30 minutes and (b) whereby the core temperature is held constant for 60 minutes.

Contributors

Dr. Joyce Aitchison (University of Oxford, UK), Dr. Christina Giannopapa (Technische Universiteit Eindhoven, Netherlands), Mr. Vincent Guyonne (Vrije Universiteit, Amsterdam, Netherlands), Mr. Miroslav Kramar (Vrije Universiteit, Amsterdam, Netherlands), Mr. Simon van Mourik (Universiteit Twente, Netherlands), Ms Jasmina Panovska (University of Edinburgh, UK), Dr. Mark Peletier (Technische Universiteit Eindhoven, Netherlands), Ms. Natascha Severens (Technische Universiteit Eindhoven & AMC, Amsterdam, Netherlands) and Dr. Marcus Tindall (University of Oxford, UK).

Acknowledgements

The Study Group participants are particularly grateful to Prof. dr. mr. Bas de Mol and his cardiac surgery team for allowing them to visit him in his surgery during surgical hours and for answering questions pertinent to the cooling and rewarming protocol.

References

[1] Analysis of Heat Exchange During Cooling and Rewarming in Cardiopulmonary Bypass Procedures, Curtis, R.M and Trezek, G.J. in *Heat Transfer in Medicine and Biology*, Shtizer, A. and Eberhard, R.C. (eds), 261-286., Plenum, New York, 1985.

[2] A Computer Model of Human Thermoregulation for a Wide Range of Environmental Conditions: the Passive System, Fiala, D., Lomas, K.J. and Stohrer, M. J. Appl. Physiol., 87 (5), 1957-1972, 1999.

[3] Computer Prediction of Human Thermoregulatory and Temperature Responses to a Wide Range of Environmental Conditions, Fiala, D., Lomas, K.J. and Stohrer, M. Int. J. Biometeorol. 45, 143-159, 2001.

[4] Analyis of Tissue and Arterial Blood Temperatures in the Resting Human Forearm, Pennes, H.H. J. Appl. Physiol., 1 (2), 93-122.

[5] Tissue Heat Content and Distribution During and After Cardiopulmonary Bypass at 17°C, Rajek, A., Lenhardt, R., Sessler, D.I., Grabenwöger, Kastner, J., Mares, P., Jantsch, U. and Gruber, E. Anesth. Analg., 88, 1220-5, 1999.

[6] A Mathematical Model of Physiological Temperature Regulation in Man, NASA contractor report CR-1855, NASA, Washington DC, 1971.

Planning Drinking Water for Airplanes

Marco Bijvank[*] *Menno Dobber*[*] *Maarten Soomer*[*]
Quentin Botton[†] *Eléonore de le Court*[†]
Jean-Christophe Van den Schrieck[†] *Moïra de Viron*[†]
Myriam Cisneros-Molina[‡] *Klaus Schmitz*[‡]
Remco van der Hofstad[§] *Ellen Jochemsz*[§] *Tim Mussche*[§]
Martin Summer[¶] *Maroescha Hoekstra*[‖] *Jeroen Mulder*[‖]
Mark Paelinck[‖]

Abstract

The management of the Dutch national airline company KLM intends to bring a sufficient amount of water on board of all flights to fulfill customer's demand. On the other hand, the surplus of water after a flight should be kept to a minimum to reduce fuel costs. The service to passengers is measured with a service level. The objective of this research is to develop models, which can be used to minimize the amount of water on board of flights such that a predefined service level is met. The difficulty that has to be overcome is the fact that most of the available data of water consumption on flights are rounded off to the nearest eighth of the water tank. For wide-body aircrafts this rounding may correspond to about two hundred litres of water. Part of the problem was also to define a good service level. The use of a service level as a model parameter would give KLM a better control of the water surplus.

The available data have been analyzed to examine which aspects we had to take into consideration. Next, a general framework has been developed in which the service level has been defined as a Quality of Service for each flight: The probability that a sufficient amount of water is available on a given flight leg. Three approaches will be proposed to find a probability distribution function for the total water consumption on a flight. The first approach tries to fit a distribution for the water consumption based on the available data, without any assumptions on the underlying shape of the distribution. The second approach assumes normality for the total water consumption on a flight and the third approach uses a binomial distribution. All methods are validated and numerically illustrated. We recommend KLM to use the second approach, where the first approach can be used to determine an upper bound on the water level.

[*] Vrije Universiteit Amsterdam
[†] Université Catholique de Louvain
[‡] University of Oxford
[§] Technische Universiteit Eindhoven
[¶] Vorarlberger Gesellschaft für angewandte Mathematik
[‖] KLM

1 Introduction

During flights people use drinking water for different purposes (e.g. consumption and going to the toilet). This water comes from a single water tank. During the preparation of each intercontinental flight, the remaining water of the previous flight is drained from the water tank and then filled up again to a predetermined level with a regulator. This regulator can only fill up the tank to multiples of 1/8th of the particular water tank. The determination of the water level is explained at the end of this section. We start with a detailed description of the water filling process.

Before the flight takes off, the purser reads the water level from a display in the cabin. On most wide-body aircrafts this display only has eight marks, which makes it difficult to determine the exact water volume in the tank. Consequently, the purser rounds off the water level to the nearest mark. The water level before take off is not by definition equal to the the amount of water pumped in the aircraft, since some water could be left behind after the draining process. We explicitly assume that the water tank is filled with high precision, such that the rounding error before take off is negligible. The same person reads the water volume in the tank again after the landing. These data records (including the aircraft type, the number of passengers, the origin and destination, the time of departure and the flight duration) are available for all flights. The only data that differ are the one from the MD-11 aircraft type. For this type, the water level is read by the purser in percentages (multiples of one hundreds). However, the regulator which is used to fill up the water tank is less precise: It fills the tank to multiples of 1/5th of the water tank.

Since the data have been gathered by operational personnel, it is very likely that some of the data are wrong. Therefore, a cleaning procedure is used to remove bad data records. First, data records with a negative water consumption are removed, as well as records with an average water consumption of more than one litre per passenger per hour. Also data records with zero water usage on long distance flights are removed.

The question becomes how the situation can be modelled such that the rounding errors are taken under consideration. This model should incorporate a suitable definition of service level as a control parameter. Based upon the model the amount of drinking water should be minimized, because any surplus of water will result in extra fuel costs.

In the current situation KLM defines the service level as the percentage of flights on the same flight leg[1] that has enough drinking water on board. So a service level of 95% for a certain flight leg means that only 5% of the flights on that flight leg will have a water shortage. A weakness of this definition is that this does not mean that a passenger will only be confronted with a water shortage in 5% of the flights. If water shortage is a structural problem on crowded flights, a passenger is more likely to be confronted with a shortage on such flights. Therefore, we recommed that the service level should be defined from a passenger's point of view. This leads to the concept of Quality of Service, as will be discussed in more detail in Section 3.

In the current approach used by KLM, the optimization of the amount of drinking water is done by calculating a regression line through the measurements of a particular flight leg. This line shows the relationship between the number of passengers and the

[1] A flight leg is a unique origin-destination and aircraft type combination.

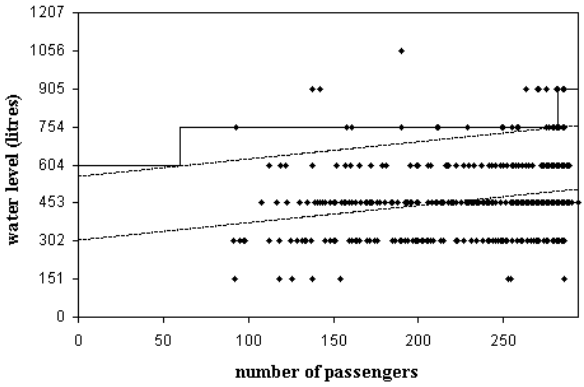


Figure 1: The water level based on the current approach used by KLM, including the regression line and the shifted regression line (dotted lines).

average water consumption (see Figure 1). The regression line is then shifted upward until a certain percentage of the measurements are below this line. Finally, all the values of the shifted line are rounded upward to the nearest multiple of one eighth of the water tank. This is performed for each flight leg. This model is based on three important assumptions. The first is that the water usage depends linearly on the number of passengers. The second is that the variance in the water usage does not depend on the number of passengers. The third is that linear regression can still be applied with rounded data. The second assumption is validated in Section 5.2. The third assumption would need further investigation, because it is not clear which of the measurements are rounded upward, and which are rounded downward. Take for instance two measurements corresponding to two similar flights A and B with 150 and 220 passengers respectively. Suppose the recorded water consumption for both flights is 5/8th of a tank. Assuming that the water consumption depends on the number of passengers on board, the water consumption of flight B is more likely to have been rounded downward than of flight A. This suggests that the probability of a value having been rounded upward or downward will most likely depend on the ratio between the number of passengers and the recorded water consumption.

This paper is organized as follows. Section 2 focuses on the data analysis of the historical data of KLM. We have evaluated possibilities to cluster different flights and studied other aspects that should be taken under consideration. A general framework has been developed for the problem in Section 3. In this section the definition of the service level is also presented. In each of the next three sections, different approach are discussed to solve the problem. Each method has been validated and illustrated on common examples. Conclusions and ideas for further research are presented in Section 7.

2 Data Analysis

In the current approach used by KLM, the historical data from flights with the same origin and destination and aircraft type are used to estimate the water usage for a flight. There are several reasons to investigate whether a larger set of flights can be used. If these flights have the same behaviour in water consumption, then using more data will give a better estimate. Furthermore, for predictions, understanding of similaraties of flights is crucial. In case of a new destination, it will be necessary to use data from flights to other destinations because there is no data available for the new destination.

Considering the whole data set (over 40.000 "valid" records), there is significant correlation between total water usage and the number of passengers (0.49), the flight duration (0.65), and the aircraft type (0.47). The first two correlations were expected. The last correlation follows from the other two correlations; First, the influence from the aircraft types can be explained by the various tank sizes and their rounding errors, and second, the same type of aircraft is used for flights with the same duration and number of passengers. Because of the high correlation with flight duration, it is interesting to investigate whether flights with the same duration can be clustered.

To compare two flights with almost the same duration but with different destinations, the correlation between the destination and the average water usage per passenger has been calculated. A significant correlation indicates that the destination determines the average water usage, and consequently, that it is not possible to cluster the flights.

Moreover, flights with different durations were compared. The correlation between those destinations and the average water usage per person per hour was considered. To avoid the effect of different aircraft types in our analysis, the calculations have been performed considering different flights to different destinations with the same aircraft type (MD-11). The results are summarized in Table 1.
It can be concluded that there is a correlation between most destinations and the water usage per passenger per hour. Therefore, these destinations cannot be clustered. However, there are particular cases for which the correlation is negligible. Not surprisingly, this very often happens on flights with destinations in the same region. As an example, clustering flights from Amsterdam (AMS) to Aruba (AUA) and from Amsterdam (AMS) to Bonaire (BON) looks reasonable (see also Figure 2). Day and night flights can also be clustered, since no correlation appears from the data. The study should be extended to other destinations and aircraft types as well. Aspects of data analysis that concern the validation of the different approaches are discussed in the corresponding sections.

AMS -	BON	DEL	DXB	JFK	LOS	MIA	MSP	NBO	SFO	YUL	YVR	YYZ
AUA	0.024	0.181	0.097	0.058	-0.114	0.071	0.011	0.143	0.125	0.050	0.121	0.003
BON		0.108	0.054	0.017	-0.144	0.026	-0.014	0.083	0.068	0.019	0.049	-0.021
DEL			-0.024	-0.094	-0.227	-0.131	-0.180	-0.019	-0.086	-0.102	-0.067	-0.191
DXB				-0.045	-0.142	-0.058	-0.094	0.009	-0.028	-0.049	-0.020	-0.100
JFK					-0.107	-0.005	-0.066	0.069	0.035	-0.006	0.052	-0.064
LOS						0.171	0.103	0.195	0.214	0.144	0.157	0.115
MIA							-0.063	0.095	0.058	-0.004	0.061	-0.073
MSP								0.139	0.116	0.039	0.150	-0.009
NBO									-0.054	-0.076	-0.040	-0.150
SFO										-0.046	0.010	-0.130
YUL											0.045	-0.049
YVR												-0.139

Table 1: Correlation between MD-11 flights from Amsterdam.

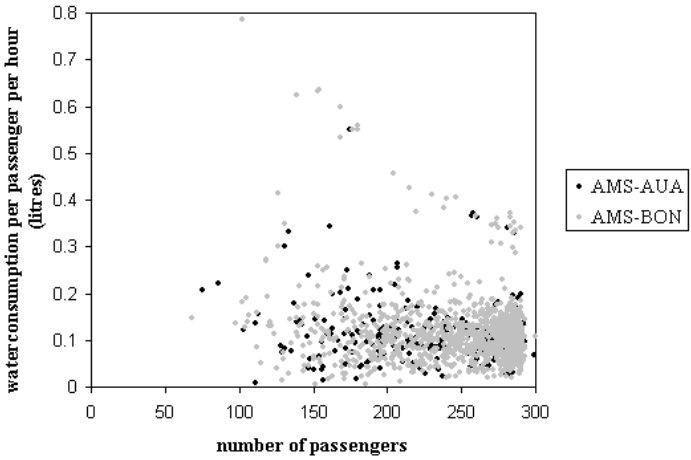


Figure 2: The water usage per passenger per hour on the flights from Amsterdam (AMS) to Aruba (AUA) and Bonaire (BON) looks similar.

3 General Framework

Consider a flight with n passengers to a certain destination. The total water consumption S_n on this flight equals

$$S_n = \sum_{k=1}^{n} Y_k, \tag{1}$$

where Y_k is the water consumption of the k-th passenger (in litres). Based on the data, there is only something known about the rounded values of S_n for each level of n and for different flights [2].

For a given flight leg (for which the number of passengers n is known), the service level is defined as the probability that a sufficient amount of water is available. The service level should at least be equal to some predefined value α, which is established by the management of KLM. So, we should have

$$\mathbb{P}\left(S_n \leq \frac{j}{8}T\right) \geq \alpha, \tag{2}$$

where $j/8$ is the percentage of tank capacity filled before take off ($j \in \{0, 1, \ldots, 8\}$) and T is the tank capacity (in litres). Since the service level for all flights should satisfy the constraint formulated in Equation (2) independently of the number of passengers on the flight, it is called a Quality of Serivce (QoS).

We are interested in finding the smallest water level for which the service constraint is satisfied (i.e. the smallest value of j in Equation (2)). Therefore, we need to find a probability distribution function for the total water consumption S_n on a flight.

[2] We know the rounded water level of the tank before take off and landing, the difference is the rounded water consumption S_n.

In the following sections, three different approaches are proposed to find such a distribution. All methods use the available data to estimate this distribution. Therefore, they have to take the rounding effect into account.

4 Curve Fitting Approach

The probability that $j/8$th of the water tank is used on a flight with n passengers, is derived from the data by looking at the frequencies how often this occurs. These probability are denoted by p_j:

$$p_j = \mathbb{P}\left(\left(\frac{j}{8} - \frac{1}{16}\right)T < S_n \leq \left(\frac{j}{8} + \frac{1}{16}\right)T\right), \quad j = 0, 1, \ldots, 8 \tag{3}$$

This can be seen as a probability mass function (pmf) of the water consumption during a flight. Based on these nine probabilities, a probability density function (pdf) of the total water usage on a flight can be estimated by fitting a curve through the pmf and then normalizing this curve such that the mass below the continuous function adds up to one.

The procedure described above has to be performed for the water consumption of a known number of passengers n. There is, however, a limited amount of measurements available on a particular flight leg for this fixed number of passengers n. In order to find enough data records to base the pmf on, the measurements for the surrounding number of passengers are used as well. We assume that at least 100 measurements are required to find a representative pmf.

The next step is to find an interpolation formula between these points. Therefore, an analytic expression for $f(x)$ (where x is the total water consumption on a flight in litres) has to be formulated, where

$$f\left(\frac{j}{8}T\right) = p_j, \quad j = 0, 1, \ldots, 8 \tag{4}$$

such that the value of $f(x)$ can be calculated at any arbitrary point.

Interpolation schemes must model the function by some plausible functional form. By far most common among the functional forms used are polynomials. One of them is Lagrange's classical formula. Since we have nine known values, this results in a high order polynomial. A characteristic of high ordered polynomials is that they tend to have a wild oscillation behaviour between the values (Press et al. [4]). This is not desirable, since we assume a smooth form for the density function for the total water usage.

Another possibility is cubic spline interpolation. Splines tend to be more stable than polynomials. The goal of cubic spline is to get an interpolation formula that is smooth in the first derivative, and continuous in the second derivative. Roughly, the idea is to take the first three data points and fit a second degree polynomial. The same has to be done for the 2^{nd}, 3^{th} and 4^{th} data point etc. Finally, these polynomials have to be concatenated together such that a continuous function appears. The exact formulation is

$$f(x) = Ap_j + Bp_{j+1} + Cf''(x_j) + Df''(x_{j+1}), \quad x_j \leq x \leq x_{j+1} \tag{5}$$

where $x_j = \frac{j}{8}T$ and A, B, C and D are defined as

$$A = \frac{x_{j+1} - x}{x_{j+1} - x_j} \qquad B = 1 - A$$
$$C = \frac{1}{6}(A^3 - A)(x_{j+1} - x_j)^2 \qquad D = \frac{1}{6}(B^3 - B)(x_{j+1} - x_j)^2 \tag{6}$$

The only problem now is that we supposed the $f''(x_j)$'s to be known, when, actually, they are not. The key idea of a cubic spline is to require a continuous interpolation scheme. This is realized by getting equations for the second derivatives $f''(x_j)$, given by

$$\frac{x_j - x_{j-1}}{6} f''(x_{j-1}) + \frac{x_{j+1} - x_{j-1}}{3} f''(x_j) + \frac{x_{j+1} - x_j}{6} f''(x_{j+1}) = \frac{p_{j+1} - p_j}{x_{j+1} - x_j} - \frac{p_j - p_{j-1}}{x_j - x_{j-1}} \tag{7}$$

for $j = 1, 2, \ldots, 7$. This equation gives seven linear equations and nine unknowns, therefore we set $f''(x_0) = f''(x_8) = 0$. For more details see Press et al. [4]. Note that $C = 0$ and $D = 0$ results in a piecewise linear interpolation scheme.

This continuous function $f(x)$ has to be normalized such that the function becomes a probability density function $g(x)$.

$$g(x) = \frac{\big(f(x)\big)^+}{\int_0^T \big(f(y)\big)^+ dy}, \quad x \in [0, T] \tag{8}$$

where the value of the integral can be found with the use of numerical integration.

The service level is given by

$$\mathbb{P}\left(S_n \leq \frac{j}{8}T\right) = \int_0^{\frac{j}{8}T} g(x)dx, \quad j \in \{0, 1, 2, \ldots, 8\} \tag{9}$$

The next step is to find the minimum water level for which this service level is at least α, as shown in Equation (2). In practice, the data can give rise to the fact of taking less drinking water on board when there are more passengers. This is not logical. Therefore, we adjust the water level, such that it becomes monotonically increasing with the number of passengers on a flight. A second modification is required for flights with small number of passengers, since there are no data available in those situations. When the assumption is made that a person consumes at most one litre of water per hour, we get the following upper bound on the water level

$$\text{water level for } n \text{ passengers} \leq nD, \tag{10}$$

where D is the scheduled duration of the flight. The outcome of Equation (10) has to be rounded upward to a multiple of an eighth of the water tank.

4.1 Validation

The MD-11 data are used for validating the approach, because the tank volume is read in percentages and, therefore, more precise. Since the number of data records for the

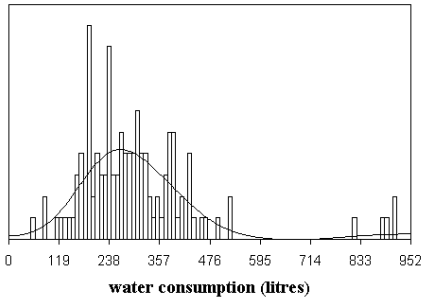


Figure 3: The histogram of the true water consumption and the estimated density function for the total water consumption on a flight leg with 285 passengers.

flights from Amsterdam (AMS) with destinations Bonaire (BON) and Aruba (AUA) is numerous, this cluster of flights is examined. As explained in Section 2, these flights are equivalent and therefore can be grouped together.

In the curve fitting approach two assumption are made. The first assumption is that the density of the water consumption does not vary too much for the same order of number of passengers. So, if 100 flights are grouped together this gives a good impression of the true density. Secondly, the probability mass function of Equation (3) can be translated into a probability density function with the use of an interpolation scheme. In particular, the density of the water consumption needs to be sufficiently regular.

The first assumption depends upon the data. When there is enough data available, this assumption can be justified. Otherwise, the tails of the distribution are too thick. This is explained in more detail in Section 4.3. The second assumption can be checked with the use of the MD-11 dataset. Figure 3 shows the frequencies of water consumption for the flights to the Antilles with 285 passengers. It also shows the estimated density function of the water consumption, when the data are rounded to multiples of eighths and subsequently cubic spline is applied. The estimated density function coincides with the form of the true density function, which is represented by the histogram. Based on this result, we might conclude that cubic spline interpolation schemes give a representative estimation for the probability distribution functions of the total water consumption on a flight leg.

4.2 Numerical example

The curve fitting approach is illustrated for the flight from Amsterdam (AMS) to Bangkok (BKK) using a Boeing 74E aircraft. For this particular flight leg, the dataset contains 548 records. The number of passengers ranges from 91 untill 294, with an average of 243 passengers. The management of KLM decided to use a service level requirement of 95%. The results for the curve fitting approach, including the data points, are given in Figure 4. The outcome may seem strange, because the required water level for 175 passengers is equal to the water level when 250 passengers are on a flight. Figure 5 shows the estimated density functions of the total water consumption for a flight with 175 and 250 passengers. Based on these two densities we can

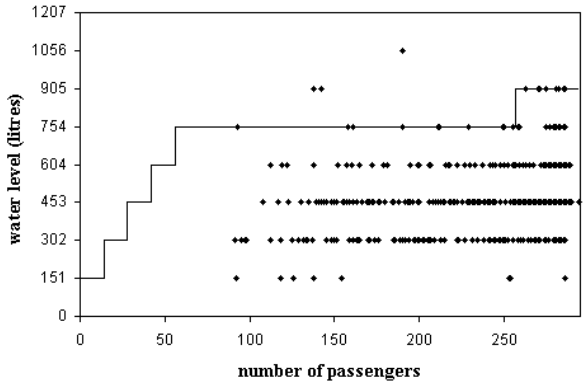


Figure 4: The water level based on the curve fitting approach for the AMS-BKK flight with the 74E aircraft and a 95% service level constraint.

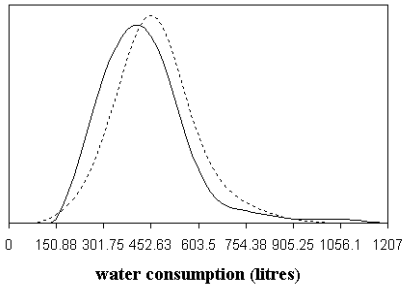


Figure 5: The estimated density function for the total water consumption when 175 passengers (straight line) and 250 passengers (dotted line) are on board.

hardly make any difference in the tail behaviour. Hence, the outcome of the method is the same. Figure 5 shows nicely a shift in the average water consumption when the number of passengers increases.

4.3 Performance

The curve fitting approach makes no assumptions about the distribution for the total water consumption on a flight, other than that it has a smooth form without any wild oscillation behaviour. When the water level for a flight with n passengers has to be determined, this method looks at all data records on flights with n passengers. If there are not enough data available on those flights, we use a subset of the data records surrounding n passengers. Consequently, the distribution for the total water usage on a flight gets thicker tails and the determined water level is too high. Hence, it becomes very relevant to find possibilities to aggregate data from different flight legs. Especially if the service level should be high, since the tail behaviour of the water consumption is becoming more relevant in those situations. Another reason why we can conclude that the curve fitting approach results in an upper bound on the water

level, is the fact that errors in the data strongly influence the outcome. Figure 3 and Figure 5 show that the water usage distribution has a skewness to the right (even a strange increase). This could be because of errors in the data, which result in a higher required water level.

In conclusion, the curve fitting approach should give the best results since there is no assumption made about the distribution for the total water usage. However, this method can only be applied when enough data records are available. Otherwise the outcome is an upper bound on the required water level. For a lot of flight legs there is not much data available. Therefore, we need to develop another more suitable method.

5 Normality Approach

The previous approach uses only a subset of the data to estimate a distribution for the total water consumption of n passengers. However, when the assumption is made that the water consumption of each passenger for a particular flight is an independent and identically distributed (i.i.d.) random variable and since the number of passengers is typically large, the central limit theorem can be applied (Ross [6]). When we generalize this theorem by adding a constant to the average and variance of the water consumption, we get

$$S_n \overset{d}{\sim} \mathcal{N}(\mu_0 + n\mu, \sigma_0^2 + n\sigma^2), \tag{11}$$

where μ is the average water usage of a person on a flight, μ_0 is a constant representing the water usage that is always required, σ^2 is the variance of this water usage and σ_0^2 is a constant added to the variance. All four parameters are expressed in litres. Nonetheless, the assumption of independence and identical distribution for the water usage per person is not really needed. The central limit theorem behaviour for sums of random variables holds more generally than under the assumption of i.i.d. summands (Feller [2]). The other assumptions will be verified in Section 5.2.

The values for the four parameters μ_0, μ, σ_0^2 and σ^2 have to be estimated. This estimation is done using the maximum likelihood method. In contrast to the previous approach, this approach uses all data to perform the estimation. When the estimates for the parameters are determined, we have an estimation for the density function of the total water usage (S_n) and we can calculate the service level for any tank volume by Equation (2).

Vardeman and Lee [7] give a systematic study of statistical analysis with rounded data. They also suggest a maximum likelihood approach if the distance between the largest and the smallest rounded value is not larger than the rounding unit. The normality approach is different in the sense that in this problem the water usage depends on the number of passengers.

5.1 Maximum likelihood estimation

The likelihood of the data is the probability of observing the data for certain parameter values (Oosterhoff and Van der Vaart [5]) and is expressed by Equation (12).

$$L(\mu_0, \mu, \sigma_0^2, \sigma^2; x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p_{n_i}(x_i \mid \mu_0, \mu, \sigma_0^2, \sigma^2), \tag{12}$$

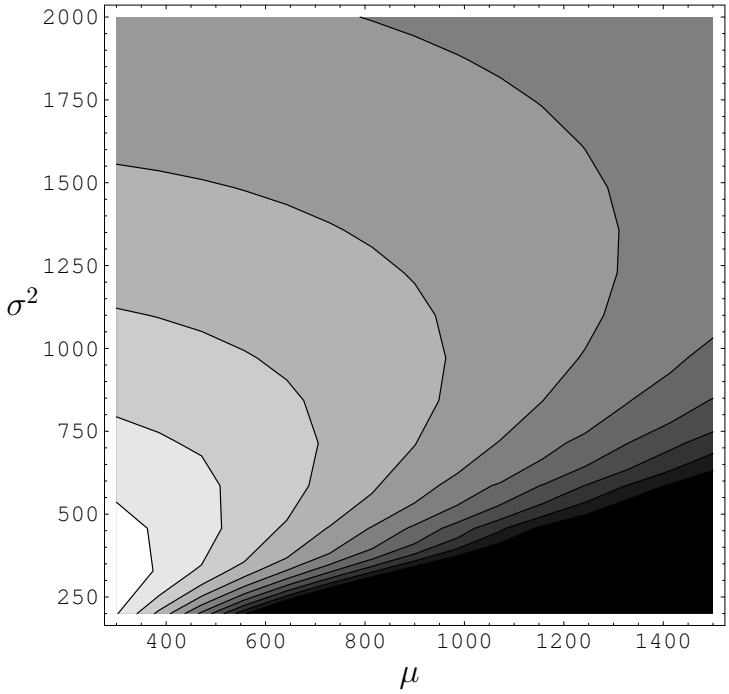


Figure 6: The contour plot of the log-likelihood function for the flight AMS-BKK.

where N is the total number of flights in the data set, n_i is the number of passengers in the i-th flight, x_i is the amount of water used during flight i (in litres)[3] and $p_{n_i}(x_i|\mu_0, \mu, \sigma_0^2, \sigma^2)$ is the probability of observing a water usage of x_i on a flight with n passengers and with μ_0, μ, σ_0^2 and σ^2 given:

$$\begin{aligned} p_n(x \mid \mu_0, \mu, \sigma_0^2, \sigma^2) &= \mathbb{P}\Big(x - \frac{1}{16}T \leq S_n \leq x + \frac{1}{16}T\Big) \\ &= \int_{x-\frac{1}{16}T}^{x+\frac{1}{16}T} \frac{1}{\sqrt{2\pi(\sigma_0^2 + n\sigma^2)}} e^{\frac{-(y-\mu_0-n\mu)^2}{2(\sigma_0^2+n\sigma^2)}} dy, \quad (13) \end{aligned}$$

since the distribution of S_n is given by Equation (11).

The objective is to find the values of the four estimators, which maximize this likelihood funtion. The log-likelihood function of equation (12) has quite a regular behaviour. Although we do not give a general proof of concavity, we illustrate concavity empirically by applying the function to the water consumption data. Note that a function $f : \mathbb{R}^N \to \mathbb{R}$ is concave if the contour set $C = \{(x, v) \in \mathbb{R}^{N+1} : v \leq f(x), x \in \mathbb{R}^N\}$ is convex (Mas-Colell et al. [3]). Figure 6 shows that the contours of the likelihood function for the data of the AMS-BKK flight are convex. Hence, the log-likelihood function is concave. Applying a numerical optimization method to the log-likelihood function of Equation (12) will result in a unique maximum. Since the log-likelihood function is also differentiable in this case, the maximizers can be

[3]based on the data $x_i \in \{0, \frac{1}{8}T, \ldots, T\}$

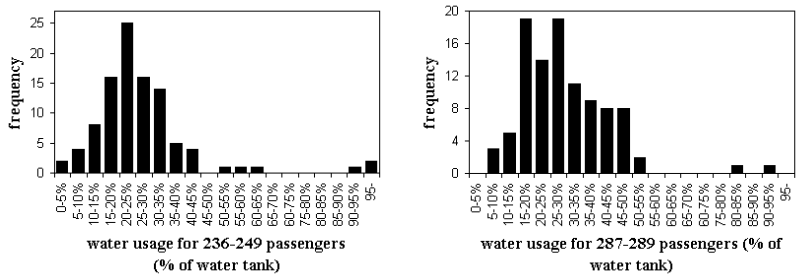


Figure 7: The frequencies of the water usage in percentages of the water tank for respectively 236-249 passengers and 287-289 passengers on the AMS-AUA and AMS-BON flights.

found by solving the first order conditions for the four parameters. This procedure gives a system of four quite complicated nonlinear equations. Therefore, this indirect method is computationally more costly than numerically optimizing the log-likelihood function.

5.2 Validation

This approach relies on the assumption that the water usage per fixed number of passengers n for a particular flight leg (or equivalent flight legs, as described in Section 2) has a normal distribution with parameters $\mu_0 + n\mu$ and $\sigma_0^2 + n\sigma^2$. To validate these assumptions, the same setting is used as in the validation section of the curve fitting approach (the flights AMS-BON and AMS-AUA for the MD-11 aircraft).

In order to say something about the distribution for a fixed number of passengers, the data have been divided into small ranges of passenger numbers. Each range contains 100 measurements. Figure 7 shows the water usage in percentages of the water tank (from $0-5\%$, $5-10\%$, etc.) plotted against the frequency. At first impression, a normal distribution does not seem farfetched. In most of the ranges, there are some strange values with a high water consumption. This seems to contradict a normal distribution, but these could also be the result of errors in the data (as mentioned in Section 4.3). To quantify normality, some statistical tests, like the Shapiro-Wilk and the Kolmogorov-Smirnov test, are performed.

The Shapiro-Wilk test (De Gunst and Van der Vaart [1]) is commonly used for testing normality of a given set of data points, where

H_0: the data come from a normal distribution (null hypothesis)

H_1: the data do not come from a normal distribution (alternative hypothesis)

The Kolmogorov-Smirnov test (De Gunst and Van der Vaart [1]) enables to compare the distributions of two datasets v and w, in which

H_0: v and w have the same distribution

H_1: v and w do not have the same distribution

For v the actual flight data are used, while for w random numbers are drawn from a normal distribution with a mean equal to the sample average of v and a variance equal to the sample variance of v. Besides normality, a logistic distribution could be used as well, to find out whether the actual distribution of the water consumption has

passenger range	all data			without outliers		
	Shapiro-Wilk	Kolomogorov Smirnov		Shapiro-Wilk	Kolomogorov Smirnov	
		normality	logistic		normality	logistic
68-160	1.80E-12	0.00010	0.00011	0.00006	0.0387	0.0702
160-187	1.53E-12	0.00043	0.00083	0.01588	0.5460	0.3408
187-204	1.03E-03	0.31730	0.40750	0.00103	0.3173	0.4075
204-220	1.07E-09	0.01938	0.03544	0.00678	0.4551	0.3126
221-236	1.09E-09	0.07637	0.11520	0.02192	0.4509	0.5142
236-249	3.74E-11	0.00309	0.01168	0.00284	0.3450	0.2995
249-259	1.10E-11	0.00381	0.01426	0.00992	0.4516	0.2865
259-268	9.08E-10	0.02498	0.04621	0.07290	0.5240	0.6547
268-274	1.94E-10	0.02069	0.05334	0.70550	0.6632	0.3295
275-280	4.04E-08	0.13770	0.26860	0.06546	0.6680	0.3275
280-284	1.13E-12	0.00135	0.00375	0.66710	0.8474	0.6001
284-287	2.70E-10	0.01262	0.03900	0.69250	0.8698	0.5270
287-289	8.91E-07	0.17030	0.15920	0.04831	0.4812	0.2129
289-300	4.17E-08	0.04494	0.11460	0.20670	0.4556	0.8048

Table 2: The p-values for the different normal distribution tests.

thicker tails. The p-value of a test refers to the probability of wrongly rejecting the null hypothesis if it is in fact true. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. The p-value is compared with a significance level. If it is smaller, the result is significant enough to reject the null hypothesis, otherwise it is not rejected (which does not imply that it must be true).

Table 2 shows the actual p-value per data range for the two specifics tests. Since Figure 7 already shows that the data contain some strange large water usages, we also tested each passenger range without these so called outliers. These results are also presented in Table 2. Globally, for the intervals with a large number of passengers (for which the ranges are also narrow) it can be concluded that the assumption of normality is supported. A logistic distribution seems, however, to fit the data more closely.

Another assumption made in the normality approach is linearity of the average water consumption with the number of passengers and linearity of the variance. For all ranges of number of passengers, the mean and variance of the water usage has been computed and plotted to see if there might be a linear trend. As can be seen in the left hand side of Figure 8, the average total water consumption could be interpreted as being linear in the number of passengers. However, this can not be said for the variance (the right hand side of Figure 8).

5.3 Numerical example

For the numerical example, the same setting is used compared to the curve fitting approach as explained in Section 4.2. Based on the implementation, we find $\hat{\mu_0} = 309.6$, $\hat{\mu} = 0.68$, $\sigma_0^2 = 129.4$ and $\hat{\sigma} = 0$. The trendline through the absolute value of the residuals, after subtracting the regression line from the data points, is constant. Therefore, the variation of the total waterusage on a flight can be seen as a constant, independent of the number of passengers.

Because the estimators for the parameters are known, we have found an estimate for the probability density function of the total water consumption S_n on a flight. The

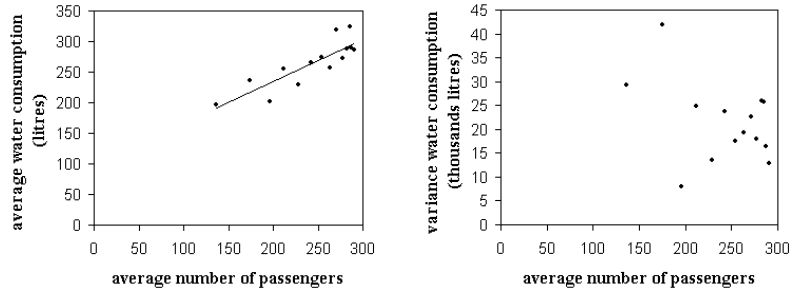


Figure 8: The average water consumption on a flight seems to have a linear trend with the number of passengers, while the variance does not.

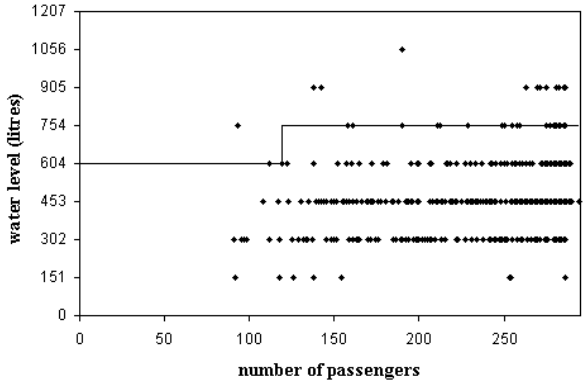


Figure 9: The water level based on the normality approach for the AMS-BKK flight with the 74E aircraft and a 95% service level constraint.

service level can be determined for a given water volume in the tank by Equation (2). The lowest values of this tank volume that satisfy the Quality of Service constraint of 95%, are given in Figure 9 for each number of passengers n.

5.4 Performance

The normality approach uses all data to say something about the water consumption for each passenger (the average and the variance). Therefore, errors in the data do not have a great effect on the outcome. However, the normality approach uses the assumption of a linear growth of the variance with the number of passengers. This is not supported by the data. Therefore, the variance becomes indifferent of the number of passengers. This is the same as the assumption made in the current approach used by KLM (see Section 1 for an explanation of the regression line approach used by KLM). We have to stress here, that these findings depend on the data and are therefore not generic. The values for $\hat{\mu_0}$ and $\hat{\mu}$ are almost similar to the coefficients for the regression line through the data points. Hence, this approach looks a lot like the regression line approach. The only difference is the way how the regression line is

shifted upward. In the normality approach, the line is shifted based on the normality distribution $\mathcal{N}(0, \sigma_0^2)$. Where in the regression line approach, the line is shifted based on some linear relationship. The variance of the regression line approach is larger compared to the findings of the normality approach.

6 Binomial Approach

In the previous approach no assumption was made on the distribution of the individual water consumption per passenger (only i.i.d.). In this section, the extra assumption is made that a passenger can either consume a maximum amount of water (M) with probability p or a minimum amount (m) with probability $1-p$. So, for each passenger k the consumption (in litres) is distributed as follows:

$$Y_k = \begin{cases} M & \text{with probability } p \\ m & \text{with probability } 1-p \end{cases} \tag{14}$$

This has the nice characteristic that the number of passengers (l) that uses the maximum amount of water has a binomial distribution (for the minimum amount of water usage this holds as well).

The total water consumption on a flight with n passengers will be given by $S_n = lM + (n-l)m = l(M-m) + nm$, such that

$$\frac{S_n - nm}{M-m} \overset{d}{\sim} bin(n,p) \tag{15}$$

and the service level equals

$$\begin{aligned} \mathbb{P}\left(S_n \leq \frac{j}{8}T\right) &= \mathbb{P}\left(\frac{S_n - nm}{M-m} \leq \frac{\frac{j}{8}T - nm}{M-m}\right) \\ &= \sum_{l=0}^{\left\lfloor \frac{\frac{j}{8}T - nm}{M-m} \right\rfloor} \binom{n}{j} p^l (1-p)^{n-l}, \quad j \in \{0,1,\ldots,8\} \end{aligned} \tag{16}$$

where l is the number of passengers asking for their maximum allowance.

In order to determine the smallest number of eighths required such that the service requirement is satisfied, the values for M, m and p have to be estimated. The first parameter can be based on the data of a particular flight leg. We look at the maximum water usage per passenger per hour and multiply this with the scheduled duration of the flight.

$$\widehat{M} = \max_{i=1,\ldots,N} \left\{ \frac{x_i}{n_i d_i} \right\} D, \tag{17}$$

where d_i is the actual duration of flight i and D is the scheduled duration of the flight. We assume $\hat{m}$=0. The estimation of the third parameter p is very important. In general, it can be seen as a measure of the dispersion of the data within the minimum m and the maximum M values. Let $\hat{\mu}$ be the estimator for the average water consumption

per passenger, given by the sample average

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} d_i n_i} D \tag{18}$$

A natural proposal for the estimator of p equals

$$\hat{p} = \frac{\hat{\mu} - \hat{m}}{\widehat{M} - \hat{m}}, \tag{19}$$

since $\hat{\mu} = \widehat{M}\hat{p} + \hat{m}(1 - \hat{p})$.

This way of estimating the parameters is more or less using intuition. We could however also use maximum likelihood, as we did in the previous approach. Equation (12) can still be applied, but now the probability of observing a water usage of x on a flight with n passengers and with M, m and p becomes

$$\begin{aligned} p_n(x \mid m, M, p) &= \mathbb{P}\Big(\left\lceil \frac{x - \frac{1}{16}T - nm}{M - m} \right\rceil \leq \frac{S_n - nm}{M - m} \leq \left\lfloor \frac{x + \frac{1}{16}T - nm}{M - m} \right\rfloor \Big) \\ &= \sum_{l=A}^{B} \binom{n}{l} p^l (1-p)^{n-l}, \end{aligned} \tag{20}$$

where

$$A = \left\lceil \frac{x - \frac{1}{16}T - nm}{M - m} \right\rceil \tag{21}$$

and

$$B = \left\lfloor \frac{x + \frac{1}{16}T - nm}{M - m} \right\rfloor \tag{22}$$

since the distribution of S_n is given in Equation (15).

6.1 Numerical example

The same setting is used for the numerical example as in the previous approaches. The average water consumption per passenger is 1.96 litres (this corresponds with $\hat{\mu}$ expressed in Equation (18)), with a maximum of $\widehat{M}$ = 8.36 litres per passenger. The probability of using the maximum quantity of water $\widehat{M}$ is calculated by Equation (19)

$$\hat{p} = \frac{1.96}{8.36} = 23.42\% \tag{23}$$

When we use the values for the parameters m, M and p which maximize the log-likelihood of Equation (20), we get $\hat{m}$ = 0, $\widehat{M}$ = 8.17 and $\hat{p}$ = 0.239. These values are somewhat similar as we expected based upon intuition. The resulting thresholds for the tank volume are presented in Figure 10.

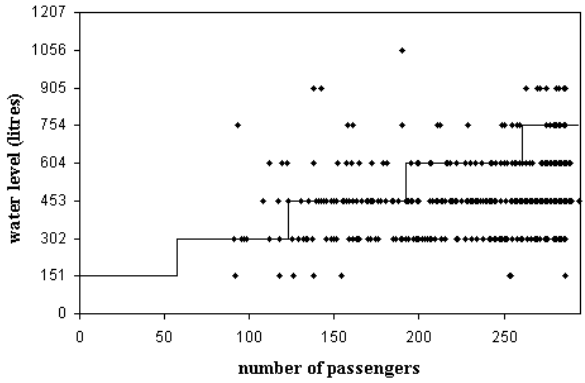


Figure 10: The water level based on the binomial approach for the AMS-BKK flight with the 74E aircraft and a 95% service level constraint.

6.2 Performance

The binomial approach can be seen as a special case of the normaltiy approach, since the binomial distribution gets close to a normal distribution for flights with many passengers. There are however some extra assumptions, which do not seem realistic, but they are an easy way to model the problem. It has the nice characteristic that the minimum and maximum water consumption is bounded.

The difference with the normality approach can be expressed in the average and variance of the total water consumption for both approaches:

	mean	variance
normality	$\mu_0 + n\mu$	$\sigma_0^2 + n\sigma^2$
binomial	$(M-m)np + nm$	$(M-m)^2np(1-p)$

In the binomial approach we use $\hat{p} = \frac{\hat{\mu}-\hat{m}}{\widehat{M}-\hat{m}}$, which comes down to $(\widehat{M} - \hat{m})n\hat{p} + n\hat{m} = n\hat{\mu}$ for the average water consumption on a flight with n passengers. Figure 8 shows that the average water consumption is indeed linear with the number of passengers. However, the linear relationship does not go through the origin of the graph. The variance for the binomial approach is also assumed to be linear with the number of passengers. This is also not supported by the data.

7 Conclusions and Future Research

In this paper we developed a framework to determine the minimal amount of drinking water on board of flights such that a predefined service level is met. We expressed the service level as the probability that a sufficient amount of water is available to fulfill passengers demand, wich is a Quality of Service. This way of formulating the problem was an eye-opener for KLM and brought new insights to the problem.

The next step was to estimate the probability density function of the total water usage on a flight. Since the available data only give information about the water

water level (litres)	the thresholds for the water level (number of passengers)			
	KLM approach	curve fitting approach	normality approach	binomial approach
151	-	1 - 13	-	1 - 56
302	-	14 - 27	-	57 - 122
453	-	28 - 41	-	123 - 191
604	0 - 59	42 - 55	0 - 118	192 - 260
754	60 - 281	56 - 256	119 - 294	261 - 294
905	282 - 294	257 - 294	-	-

Table 3: The ranges of the water level for which the Quality of Service is granted, for each approach applied to the AMS-BKK flight with the 74E aircraft.

consumption in multiples of $1/8$th of the water tank, three approaches were developed to tackle this problem. The curve fitting approach does not make any assumptions about the form of the distribution. This method, however, results in an upper bound on the required water level, because a subset of the data is used and not enough data are available and also because there are errors in the data. The normality approach uses all data to find an estimate for the distribution of the total water usage. This also reduces the effect of the errors in the data. The final approach (the binomial approach) uses extra assumptions on the water usage per passenger. The results from the different approaches are summarized in Table 3. In conclusion we recommend the normality approach, where the curve fitting approach is used as an upper bound. The binomial approach can be used to get a better understanding what the outcome will look like when parameters change. The regression line approach, which is currently used by KLM, looks very similar to the findings of the normality approach. The normality approach is based on a more general foundation and the method uses a better understanding of the service level. Hence, this method is prefered by KLM as well.

The normality approach could however be improved. As shown in Figure 8, the variance of the total water consumption on a flight does not grow linearly with the number of passengers. Therefore, other forms of relationships should be explored as well besides linearity. The numerical example of the normality approach showed that μ_0 and σ_0^2 play an important role. Therefore, these parameters should be modelled in more detail. At last, based on the results of the statistical tests in Table 2, the effect of a logistic distribution for the total water consumption on a flight should also be investigated closer.

The numerical examples are performed on a few flight legs. Therefore, we recommend to do the same calculations on several more flights. Larger data sets have to be checked also for a more complete overview of the validity of the assumptions made in the different approaches.

References

[1] De Gunst, M.C.M., Van der Vaart, A.W., 2001, *Statistische Data Analyse*, lecture notes, Faculty Exact Sciences, vrije Universiteit amsterdam

[2] Feller, W., 1970, *An introduction to Probability Theory and its Applications, Volume II, Second Edition*, Wiley

[3] Mas-Colell, A., Whinston, M.D., Green, J.R., 1995, *Microeconomic Theory (First Edition)*, Oxford University Press

[4] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., 1988, *Numerical Recipes in C*, Cambridge Univ. Press, New York

[5] Oosterhoff, J., Van der Vaart, A.W., 2000, *Algemene Statistiek*, lecture notes, Faculty Exact Sciences, vrije Universiteit amsterdam

[6] Ross, S.M., 2003, *Introduction to Probability Models (Eighth Edition)*, Academic Press, San Diego

[7] Vardeman, S. and Lee, C.S., January 2005, Likelihood-Based Statistical Estimation from Quantized Data, working paper, to appear in *IEEE Transactions on Instrumentation and Measurement*, http://www.public.iastate.edu/∼vardeman/homepage.html

Selection Effects in Forensic Science

Geert Jan Franx[*] *Yves van Gennip*[†] *Peter Hochs*[‡]
Misja Nuyens[*] *Luigi Palla*[*] *Corrie Quant*[*]
Pieter Trapman[*]

Abstract

In this report we consider the following question: does a forensic expert need to know exactly how the evidential material was selected? We set up a few simple models of situations in which the way evidence is selected may influence its value in court. Although reality is far from a probabilistic model, and one should be very careful when applying theoretical results to real life situations, we believe that the results in our models indicate how the selection of evidence affects its value. We conclude that selection effects in forensic science can be quite important, and that from a statistical point of view, improvements can be made to court room practice.

1 Introduction and problem statement

At a crime scene, a red fibre is found on the victim of a murder. After the police have found a suspect, a search of his wardrobe reveals a red jumper. The jumper and fibre are brought to the forensic lab, which has to check whether the fibre matches with the jumper and if so, how *strong* this evidence is. Obviously, a very rare jumper should be considered as stronger evidence than one bought at a large company like H&M. But does the strength of the evidence depend on how and in what circumstances the jumper was found? For example, is the evidence stronger if the suspect had no other jumpers? Or does it make no difference?

This is an example of the following question posed to the Study Group by the Netherlands Forensic Institute (NFI): does a forensic expert need to know how exactly the evidential material was/is selected? This question is also relevant to the use of video identification and DNA-databases. Suppose a video recording of a crime is shown on television, and a number of suspects are brought to the attention of the police by people who watched the broadcast. Can the same video material that was used to select the suspects also be used as evidence in court? And if so, does the strength of this video evidence, for example, depend on the number of people that know the suspect? For the second example, suppose that somebody becomes a suspect because his or her DNA is in a DNA-database, and matches a DNA-sample from a crime scene. Clearly such databases do not include the DNA of the entire population. But does this

[*]Vrije Universiteit Amsterdam
[†]Technische Universiteit Eindhoven
[‡]Radboud Universiteit Nijmegen

make a difference? And should, after the selection of a suspect via his or her DNA, this DNA match be discarded as evidence?

At the moment, the situation is usually as follows. When the evidence is presented, the fact that a suspect has been selected using, for example, video material, is considered not to influence the value of the video material as evidence. Neither is the number of clothes in someone's wardrobe taken into account when fibres found on a crime scene match with a suspect's clothes and are used as evidence. This is somewhat alarming, since the judge cannot be expected to have knowledge of statistics, and possible statistical corrections should be made before the evidence is presented to him. On the other hand, the NFI expert handling the case is also not a statistician. (S)he has the possibility to call in the help of statisticians, but does so only if (s)he feels the need to do that. So the question posed to the Study Group can be seen as a request for help in dealing with situations that seem straightforward, but in fact may need statistical corrections.

In this report we answer this question in the following way. We set up a few simple models of situations in which the way evidence is selected may influence its value in court. We then give expressions for the probability that the evidence found indeed originates from the suspect, given that lab tests, or a broadcast of video material, links the evidence to the suspect. From that we conclude if and how the selection of the evidence influences this probability. Obviously, reality is far from any probabilistic model, and one should be extremely careful when applying theoretical results to real life situations. This is especially dangerous when statistics is used to prove that a crime has been committed, see for example Van Lambalgen and Meester [4]. Therefore, in the models discussed in this report this situation is excluded by assuming that it is sure that a crime has been committed. Keeping these warnings in mind, we nevertheless believe that our simple models can teach us something about how the evidence selection procedure affects the value of the evidence.

The report will adhere to the following structure. In Section 2 we explain the notion of conditional probabilities and Bayes' rule, which are used later on. A judge might want to have an indication of the probability that a suspect is guilty given certain evidence. Bayes' rule allows us to express this probability in terms of the probability that the evidence is found when the suspect is guilty, the probability that the evidence is found when the suspect is innocent, and a so-called *prior*. In Section 3 we consider a simple model of the situation where the police find a fibre on a crime scene that matches a jumper from a suspect's wardrobe. We argue that the number of jumpers the suspect owns influences the evidential value of the fibre, and describe in what way it does. In Section 4 we look at video evidence. Under simplifying assumptions, we compute the probability that a suspect who is recognised on video, is actually the person on the video material. Finally, in Section 5 we consider a different aspect of the selection of evidence, namely the influence of non-matches between the suspect and the available evidence. We argue that these non-matches may count as negative evidence, and should therefore be taken into account as well. We finish the report with some conclusions.

2 Bayes' rule and likelihood ratios

During a trial, the scientific analysis of the evidence found on the crime scene is reported to the court by the forensic expert. In order to provide the judge with the means to evaluate such evidence, the law allows/requires the forensic scientist to summarise his expertise by means of a *likelihood ratio* (see [2]). Define

$$E := \text{“Evidence at crime scene is matched with the suspect”};$$
$$H := \text{“The suspect himself left the evidence”}.$$

The likelihood under the hypothesis that the suspect is guilty is given by the conditional probability $P(E \mid H)$; the likelihood under the (null) hypothesis that the suspect is innocent is given by $P(E \mid H^c)$. matching the uspect at the crime scene, then $P(E \mid H) = 1$. Hence, The *likelihood ratio* is defined as

$$LR = \frac{P(E \mid H)}{P(E \mid H^c)}.$$

Since the numerator is likely to be large, the crucial task of the forensic scientist boils down to assigning a value to the denominator of the above formula. This means evaluating the probability of a random match. In this evaluation procedure lies the difficulty in reporting evidence, since the likelihood under the null hypothesis depends on the assumptions made on the reference population, which might be hard to define.

A formalisation of the inferential process performed by the judge/juror is well expressed by Bayes' theorem ([1]), one of the basic formulae in probability theory. This theorem was first proposed by rev. Thomas Bayes in the 18th century and constitutes the basis for the 20th century Bayesian school of statistical inference ([5]), as opposed to the classical or frequentist school. Bayes' theorem (also named Bayes' rule for its simplicity) expresses the *posterior* probability in terms of likelihoods and *prior* probabilities:

$$\begin{aligned} P(H \mid E) &= \frac{P(H \cap E)}{P(E)} = \frac{P(E \mid H)P(H)}{P(E)} \\ &= \frac{P(E \mid H)P(H)}{P(E \mid H)P(H) + P(E \mid H^c)P(H^c)}. \end{aligned} \tag{1}$$

Here the denominator in the last term follows from the rule of total probability. The appeal of this formulation is due to a foundational inference argument, namely that the actual matter of interest in the inductive process is the posterior probability. This is the probability of the hypothesis (H) given the evidence (E), and not the likelihood or inverse probability, which is the probability of the evidence (E) given the hypothesis (H). In particular, it is the conditional event $H \mid E$ that the judge will, more or less consciously, probabilistically evaluate to give the sentence, while the scientist presents his probability estimate on $E \mid H$. Dividing the numerator and denominator of (1) by $P(E \mid H)P(H)$, we obtain the following formula:

$$P(H \mid E) = \left(1 + \frac{P(H^c)}{P(H)} \frac{P(E \mid H^c)}{P(E \mid H)}\right)^{-1}. \tag{2}$$

We can see how the posterior probability depends on the *prior odds*, i.e., the ratio $P(H^c)/P(H)$ between the a priori probabilities of the hypotheses and the inverse of the likelihood ratio as defined above. The prior odds give a formal way to incorporate non-statistical evidence into the model. in the evaluation judicial elements external to the scientific evidence. Observe that although it is tempting to think so, a large likelihood does not automatically imply a large probability that the suspect is guilty.

Bayesian inference is a generalisation of the exclusively likelihood based classical inference. The relevance of a Bayesian modelling approach is also more appropriate when dealing with unique cases and often limited pieces of evidence as in the forensic setting, where doing justice cannot be achieved via the long run philosophy underlying the classical approach ([3]). For this reason, although the forensic scientist cannot or is not allowed to specify priors on the suspect's guilt, it is useful for him to summarise the judicial inductive process before presenting results in court, as will be seen in the following section.

3 The jumper model

In this section we consider a simple example that shows that for computing orestimating the probability that a suspect is guilty, it is necessary that the information about the evidence and the way it is selected shouldbe as complete as possible. The example we give is not too realistic, but it teaches us a lot about more realistic models, and partly answers the NFI question.

3.1 Introduction

The case is as follows. Suppose an event has taken place at which someone, (the *donor*) has left a fibre of his jumper on some other person (the *victim*). We do not know anything about this donor. In order to find the donor, we investigate the jumpers in the wardrobe of an arbitrary person (for notational convenience, this person is called the *suspect*). We find a jumper made of a fibre of exactly the same type as was found on the victim. We want to compute the probability that our suspect is in fact the real donor of the fibre on the victim, given the evidence found, i.e., given that the fibres of the suspect and those on the victim match. The question we ask ourselves is how much we need to know about the evidence found: is it sufficient to know that one of the jumpers of the suspect matches the fibre on the victim? Or do we for example also need to know how many of his jumpers consist of fibres of other types?

In Section 3.2 we specify our assumptions. These allow us to compute the probability that our suspect is the donor in Section 3.3. In Section 3.4 we generalise this model to the case that more than one fibre is found on the victim. Finally, Section 3.5 deals with our conclusions and possible extensions of the model.

3.2 Assumptions and notation

We make the following assumptions to model the situation described above.

1. Only one fibre is found on the victim; this fibre is of type Y and does not belong to the victim. (We assume all fibres in the world can be categorised into

a number of different types that the forensic expert can tell apart.)

2. The fibre found on the victim was transferred during a meeting between the victim and the donor, and originates from the donor's jumper. We call the moment of this meeting the *transfer moment*.

3. Since the transfer moment nobody has thrown away or hidden any jumpers.

4. Every jumper consists of only one type of fibre.

5. The relative frequency of the total population wearing jumpers of a specified fibre type at the transfer moment is known. In particular, the relative frequency of the population wearing a jumper of fibre type Y is g_Y. This can be interpreted as the probability that some random person was wearing a jumper of fibre type Y at the transfer moment.

6. The probability that the suspect was wearing a jumper of fibre type Y at the transfer moment is known and denoted by f_Y.

7. The collection of jumpers of any person is independent of him being the donor or not.

Further, we write E_1 for the event that the fibre found on the victim is of type Y and E_2 for the event that we found a fibre of type Y in the collection of jumpers of the suspect. The event that the suspect is the donor of the fibre on the victim is denoted by D.

3.3 Computations

We are ready to compute the probability that our suspect is the donor of the fibre on the victim given the evidence found. Using (2), we compute

$$\begin{aligned} P(D \mid E_1 \cap E_2) &= \left(1 + \frac{P(D^c)}{P(D)} \frac{P(E_1 \cap E_2 \mid D^c)}{P(E_1 \cap E_2 \mid D)}\right)^{-1} \\ &= \left(1 + \frac{P(D^c)}{P(D)} \frac{P(E_2 \mid D^c) P(E_1 \mid D^c \cap E_2)}{P(E_2 \mid D) P(E_1 \mid D \cap E_2)}\right)^{-1} \\ &= \left(1 + \frac{P(D^c)}{P(D)} \frac{P(E_2) g_Y}{P(E_2) f_Y}\right)^{-1} = \left(1 + \frac{P(D^c)}{P(D)} \frac{g_Y}{f_Y}\right)^{-1}. \end{aligned}$$

For third equality we used assumptions 5 up to 7. Observe that if g_Y is smaller (the fibre is rare), the probability of the suspect being the donor is larger. Also, if f_Y is smaller (the suspect does not wear the jumper of fibre type Y often), the probability of the suspect being the donor is larger. In the special case that the suspect owns k jumpers and wears those with equal frequency, we have $f_Y = 1/k$. Hence, the more jumpers the suspect has, the smaller is the probability that he is the donor - in this report we use male pronouns for the suspect and the donor. This seems quite reasonable: a person owning thousand jumpers is very likely to match with the fibre on the victim, but the strength of this match is, of course, very low.

3.4 The jumper model with more than one fibre

In this section we consider the situation that a crime is committed at which the offender *possibly* donated a fibre of his jumper to the victim. We find a number of types of fibres at the crime scene, which surely originate from some jumpers. Based on the place at the crime-scene where a particular fibre is found, we may assign some probability to the event that that particular fibre was donated at the moment of the crime. As in the previous example, we investigate the wardrobe of a suspect to see if there are jumpers with fibres that match with fibres found on the crime scene.

We are interested in the questions "does the probability that the suspect is the offender, given that there is a match, depend on the number of fibres found at the crime scene?" and "does this probability depend on the number of matches found?"

Our assumptions are the same as in the previous section with the following exceptions.

1'. There are n types of fibres found at the victim, the ith type of fibre is called Y_i for $i \in \{1, \ldots, n\}$.

2'. The *a priori* probability that the ith type of fibre is connected to the crime is h_i. These h_is are independent of the identity of the donor, if we do not take his collection of jumpers into account.

8. At most one person donated a fibre at the crime, so there is only one offender; if there is a donor, he donated at most one fibre.

From these assumptions we see that $\sum_{i=1}^{n} h_i \leq 1$. Note that it is possible that none of the fibres is left by the offender, the probability of that event is $1 - \sum_{i=1}^{n} h_i$.

Write $E_1^{(i)}$ for the event that the i-th type of fibre on the crime scene is of type Y_i and $E_2^{(i)}$ for the event that a fibre of type Y_i is in the collection of jumpers of the suspect. The event that the i-th type of fibre was donated at the moment of the crime is $C^{(i)}$ and the event that the suspect is the donor of the fibre donated at the moment of the crime is D^*. Write V for the collection of fibre types at the crime scene and W for the collection of fibre types from jumpers of the suspect. Denote the intersection of V and W by A, so A is the set of types of fibres that are both in the collection of the suspect and at the crime scene (the matches).

We are interested in $P(D^* \mid \bigcap_{j\in V} E_1^{(j)}, \bigcap_{j\in W} E_2^{(j)})$, the probability that the suspect donated the fibre at the moment of the crime, given the evidence. Because the events $C^{(i)}$ are disjoint, and $D^* = \cup_i C^{(i)}$, we can write

$$\begin{aligned} P(D^* \mid \bigcap_{j\in V} E_1^{(j)}, \bigcap_{j\in W} E_2^{(j)}) &= P(\bigcup_{i\in V} \{D^* \cap C^{(i)}\} \mid \bigcap_{j\in V} E_1^{(j)}, \bigcap_{j\in W} E_2^{(j)}) \\ &= \sum_{i\in V} P(D^* \cap C^{(i)} \mid \bigcap_{j\in V} E_1^{(j)}, \bigcap_{j\in W} E_2^{(j)}) \\ &= \sum_{i\in A} P(D^* \cap C^{(i)} \mid E_1^{(i)} \cap E_2^{(i)}). \end{aligned}$$

Here we used that the event $\{D^* \cap C^{(i)}\}$ is independent of $E_1^{(j)}$ and $E_2^{(j)}$ for $i \neq j$ and that the summand is 0 if the fibre at the crime scene is not in the collection of jumpers

of the suspect. We can use the results of the previous section to conclude that:

$$\begin{aligned} P(D^* \mid \bigcap_{j\in V} E_1^{(j)}, \bigcap_{j\in W} E_2^{(j)}) &= \sum_{i\in A} P(C^{(i)} \mid E_1^{(i)} \cap E_2^{(i)}) P(D^* \mid C^{(i)} \cap E_1^{(i)} \cap E_2^{(i)}) \\ &= \sum_{i\in A} h_i \left(1 + \frac{P(D^{*c} \mid C^{(i)})}{P(D^* \mid C^{(i)})} \frac{g_{Y_i}}{f_{Y_i}}\right)^{-1} \\ &= \sum_{i\in A} h_i \left(1 + \frac{P(D^{*c})}{P(D^*)} \frac{g_{Y_i}}{f_{Y_i}}\right)^{-1}, \end{aligned}$$

where the second equation holds since under the condition that a certain fibre was donated at the crime, everything is the same as in the situation where only one fibre is found. The last equation holds since the probability of being the donor does not depend on the fibre found if nothing is said about the collection of jumpers of the suspect. Note that the probability that the suspect is the offender may be larger than the probability computed, because he may have committed the crime, while he did not donate any fibre.

If it is certain that one of the fibres found was donated at the moment of the crime and if all fibres have the same probability to have been donated at the moment of the crime, $1/n$, we get

$$P(D^* \mid \bigcap_{i\in V} E_1^{(i)}, \bigcap_{i\in W} E_2^{(i)}) \quad = \quad \sum_{i\in A} \frac{1}{n} \left(1 + \frac{P(D^{*c})}{P(D^*)} \frac{g_{Y_i}}{f_{Y_i}}\right)^{-1}.$$

So the probability that the suspect is the offender decreases when the fraction of fibres on the victim that match with jumpers in the collection of the suspect decreases.

3.5 Conclusions and possible extensions of the model

In our (very basic) model we have shown that many things should be reported in order to interpret the evidence well. Not only that a jumper in the collection of the suspect and a fibre found at the crime scene match, but also how often the suspect wears that particular jumper and how many fibres are found at the crime scene. We have shown that the number of jumpers in the wardrobe and the number of fibres at the crime scene influence the probability that the suspect is the offender.

In this section we have analysed a very basic example dealing with evidence in our model. We were forced to make many assumptions in order to get some results. In the future, efforts can be made to relax some of the assumptions. For example, one could introduce uncertainty in the matching of two fibres. This seems reasonable since the forensic expert could make a mistake when comparing two fibres. One may also think of dealing with the possibility that more than one type of fibre is donated at the transfer moment, for example, jumper fibres and jeans fibres. Finally, in this example we had only once piece of evidence, namely the match of some fibre. In the case that there is more evidence, e.g., blood stains, or footprints, these other pieces of evidence should also be incorporated in the model.

4 The video recognition model

In this section, we construct a model for the following situation. A crime has been committed and thanks to camera surveillance some video material of the criminal is available. This material is then shown to the general public via a television show, like *Opsporing verzocht*.

Obviously, if a person has more acquaintances, the probability that this person is reported is larger. A question that arises naturally is the following. Given a person has been reported, does the probability that he is guilty depend on his number of acquaintances? In other words, should the forensic expert of the judge take into account that the suspect was a very social person, or a very solitary one? In order to answer this, we consider a simple model with the following assumptions:

- The criminal is known to be Dutch and the video material is only shown on Dutch television.
- There is a group of l look-alikes in the Netherlands. These are people who cannot be distinguished from the person on the video by any means.
- One of the look-alikes, called ξ, has n acquaintances.
- Each of these acquaintances reports ξ, independently, with a probability p.

Define the following events:

$$S := \text{“}\xi \text{ is the person on the video”},$$
$$R := \text{“}\xi \text{ is reported”}.$$

Applying Bayes' rule, we find

$$P(S \mid R) = P(S)\frac{P(R \mid S)}{P(R)} = P(S) = \frac{1}{l}, \tag{3}$$

where the second equality holds since all look-alikes look like the person on the video, no matter whether they really are him/her or not. The last identity holds true since there are l look-alikes in the Netherlands.

We see that in this model there is no dependence on either n or p, because we treat all look-alikes as indistinguishable. If we let go of this condition, we get a slightly more sophisticated model, with the following new assumptions.

- The criminal is known to be Dutch and the video material is only shown on Dutch television.
- Some person called ξ has n acquaintances.
- $P(\text{“one acquaintance reports } \xi\text{”} \mid S) = p$.
- $P(\text{“one acquaintance reports } \xi\text{”} \mid S^c) = q$.

Typically, we have $q \leq p$. Under these new assumptions, applying Bayes' rule (2) gives

$$P(S \mid R) = \Big(1 + \frac{P(S^c)}{P(S)} \frac{P(R \mid S^c)}{P(R \mid S)}\Big)^{-1}. \tag{4}$$

From

$$P(R \mid S) = 1 - (1 - P(\text{“one acquaintance reports } \xi\text{”} \mid S))^n = 1 - (1-p)^n,$$
$$P(R \mid S^c) = 1 - (1 - P(\text{“one acquaintance reports } \xi\text{”} \mid S^c))^n = 1 - (1-q)^n,$$

it follows that

$$P(S \mid R) = \Big(1 + \frac{P(S^c)}{P(S)} \frac{1-(1-q)^n}{1-(1-p)^n}\Big)^{-1}.$$

If $p, q \ll 1/n$, then using Taylor's formula yields

$$P(S \mid R) \approx \Big(1 + \frac{P(S^c)}{P(S)} \frac{q}{p}\Big)^{-1}.$$

Note that n does not play a role in this approximation. On the other hand, if $q \ll 1/n$, $q \ll p$, and $p \gg 1/n$, then approximating $(1-p)^n \approx 0$ yields

$$P(S \mid R) \approx \Big(1 + \frac{P(S^c)}{P(S)} qn\Big)^{-1}.$$

Assuming that $P(S^c) \gg P(S)$, the RHS decreases like $1/n$. So in this case, given that a person is reported, the more acquaintances he has, the less likely it is that he is indeed the person on the video. This outcome seems to be counter-intuitive, but can be explained as follows. The more acquaintances a person has, the more likely it is that he is reported. However, the probability that he is the person on the video, and is reported, namely $P(S)(1-(1-p)^n)$, does not change so much, as we assumed that $(1-p)^n \approx 0$. Hence, the probability that he really is the person on the video given that he is reported, decreases.

5 The influence of negative evidence

In criminal investigations not only positive matching results are found. Usually, negative results are not used in court. The question arises whether this is correct. For instance, in the basic jumper model considered in Section 3, what conclusion can be drawn if we do not find a match? Does this imply that the suspect is not very likely to be the offender, or does it hardly have any influence on the probability that the suspect is guilty?

Observe that the basic jumper model concentrates only on the probability that the suspect is the *donor* of the fibre found on the victim. In this section we extend the basic jumper model in order to concentrate on the probability that the suspect is the *offender*. In many crime scenes there are traces that *could* have been donated by the offender, like fibres. On the other hand, there may also be traces that are *very likely* to have been donated by the offender, like bullets or blood. Further, it is not only the

nature of the trace that is important here, but also the exact place where it was found. Fibres found on the neck of a strangled person are more likely to be an offender's trace than fibres found under the shoes.

Let us assume that based on this kind of information, for every trace T the forensic expert can assign a probability h_T that this trace was donated by the offender. In the extended model we define the following possible events:

$TY :=$ "Trace found on the victim is of material Y"
$TO :=$ "Trace found on the victim is donated by the offender"
$SO :=$ "Suspect is the offender"
$SY :=$ "Suspect can be linked to material Y."

The assumptions of the basic jumper model still apply to this extended model. However, there is no need to restrict the model to fibre matching. For instance, if we are considering a case of DNA-matching, all we have to do is assume that everybody has only one jumper (DNA-pattern). Again, f_Y and g_Y are the relative frequencies of the suspect respectively the whole population carrying material of type Y at the moment of the trace transfer. In the case of DNA we have $f_Y = 1$, since people are not able to change their DNA profile. For the probability that the suspect is guilty given the evidence of a positive match, we write, using (2):

$$\begin{aligned} P(SO \mid SY \cap TY) &= \left(1 + \frac{P(SO^c)}{P(SO)} \frac{P(SY \cap TY \mid SO^c)}{P(SY \cap TY \mid SO)}\right)^{-1} \\ &= \left(1 + \frac{P(SO^c)}{P(SO)} \frac{P(SY \mid SO^c)}{P(SY \mid SO)} \frac{P(TY \mid SO^c \cap SY)}{P(TY \mid SO \cap SY)}\right)^{-1}. \end{aligned}$$

Since we assumed that there is no dependence between someone's blood type (or type of fibre that (s)he is wearing) and his or her criminal intent, we have $P(SY \mid SO^c)/P(SY \mid SO) = 1$. Further,

$$\begin{aligned} P(TY \mid SO^c \cap SY) &= P(TY \cap TO \mid SO^c \cap SY) \\ &\qquad + P(TY \cap TO^c \mid SO^c \cap SY) \\ &= P(TO \mid SO^c \cap SY) P(TY \mid TO \cap SO^c \cap SY) \\ &\qquad + P(TO^c \mid SO^c \cap SY) P(TY \mid TO^c \cap SO^c \cap SY) \\ &= h_T g'_Y + (1 - h_T) g_Y, \end{aligned}$$

where g'_Y is the relative occurrence of type Y in the whole population minus our suspect. If the trace consist of extremely rare material (like large DNA samples), then g'_Y may differ substantially from g_Y. In fact, for unique material belonging to the suspect, $g'_Y = 0$.

$$\begin{aligned} P(TY \mid SO \cap SY) &= P(TY \cap TO \mid SO \cap SY) \\ &\qquad + P(TY \cap TO^c \mid SO \cap SY) \\ &= P(TO \mid SO \cap SY) P(TY \mid TO \cap SO \cap SY) \\ &\qquad + P(TO^c \mid SO \cap SY) P(TY \mid TO^c \cap SO \cap SY) \\ &= h_T f_Y + (1 - h_T) g_Y, \end{aligned}$$

then gives

$$P(SO \mid SY \cap TY) = \left(1 + \frac{P(SO^c)}{P(SO)} \frac{h_T g'_Y + (1-h_T) g_Y}{h_T f_Y + (1-h_T) g_Y}\right)^{-1}. \tag{5}$$

This result demonstrates the importance of h_T. If $h_T = 1$, we are sure that the donor of the trace is the offender. Therefore, if we substitute $h_T = 1$ in (5), we find the equivalent result for the basic jumper model discussed in Section 3. On the other hand, if $h_T = 0$, the likelihood ratio is 1, and the trace gives no information about the offender.

We turn to the case that the trace found on the victim cannot be matched to the suspect. As in the positive case, we find

$$P(SO \mid SY^c \cap TY) = \left(1 + \frac{P(SO^c)}{P(SO)} \frac{P(TY \mid SO^c \cap SY^c)}{P(TY \mid SO \cap SY^c)}\right)^{-1}. \tag{6}$$

Observe that $P(TY \mid SO^c \cap SY^c) = g'_Y$. For the denominator we find

$$\begin{aligned} P(TY \mid SO \cap SY^c) &= P(TO \mid SO \cap SY^c) P(TY \mid TO \cap SO \cap SY^c) \\ &\quad + P(TO^c \mid SO \cap SY^c) P(TY \mid TO^c \cap SO \cap SY^c) \\ &= (1-h_T) g'_Y. \end{aligned}$$

Here we used that $P(TY \mid TO \cap SO \cap SY^c) = 0$. Combining this with (6) yields

$$P(SO \mid SY^c \cap TY) = \left(1 + \frac{P(SO^c)}{P(SO)} \frac{1}{(1-h_T)}\right)^{-1}.$$

The likelihood ratio $(1-h_T) \leq 1$ depends only on h_T. We conclude that in general a suspect is less likely to be an offender if no trace can be matched to the suspect. If the police are quite sure that the trace originates from the offender, the likelihood ratio turns out to be very much in favour of the suspect. However, we have to bear in mind that in this model we assumed that no evidence has been destroyed by the suspect. For instance, if we know that the suspect has destroyed the clothes he was wearing during the crime, a negative result on fibre matching does not have any consequence for the likelihood ratio. This kind of complication is hard to incorporate in the model, since the probability that a suspect destroys his clothes depends highly on his innocence, and is hard to estimate. Luckily, this complication does not occur in case of blood traces, or any other traces that originate from the human body. Since these traces are used in many criminal investigations, the above model may still be useful.

Concluding this section, we remark that in many crime cases a lot of traces are collected. If only one of these traces can be linked to the suspect, this trace will be presented in court as evidence, and the other traces will be left out. This selection of evidence seems to be unfair, since the evidential value of the matching trace can be heavily weakened by all traces that cannot be linked to the suspect, especially if some of them were estimated in advance to have a high probability of being offender's traces. We conclude that selection effects in forensic science can be quite important, and from a statistical point of view, improvements can be made to court room practice.

6 Conclusions

To analyse the effect of the way in which evidence is selected, we have set up a very simple model for the matching of fibres found on the victim and clothes belonging to the suspect. In this model we have shown that many things should be reported in order to properly interpret the evidence. We have shown that, for example, the number of jumpers in the wardrobe and the number of fibres at the crime scene influence the probability that the suspect is the offender. Furthermore, in an extension of this jumper model, we have shown that the evidential value of a match can be heavily weakened by all traces that cannot be linked to the suspect.

We stress that we have proved these results in our probabilistic models, which are far from real life situations. Nevertheless, the results we obtained may guide our reasoning in this matter. We conclude that selection effects in forensic science play an important role, and that efforts should be made to improve the statistical interpretation of evidence in court room practice.

References

[1] Bayes T., An essay towards solving a problem in the doctrine of chances, Philosophical Transactions of the Royal Society of London 53, 370-418 (1763).

[2] Evett I., and Weir B., Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists, Sinauer Associates (1998).

[3] Hacking I., Logic of statistical inference, Cambridge University Press (1965).

[4] Van Lambalgen, M., and Meester, R., On the (ab)use of statistics in the legal case against the nurse Lucia de B, preprint, available from `http://www.few.vu.nl/~rmeester/pre.html` (2005).

[5] Robert C., The Bayesian choice: from decision theoretic foundations to computational implementation, Springer, New York (2001).

Optimal Weighing Schemes

Sandjai Bhulai[*] *Thomas Breuer*[†] *Eric Cator*[‡]
Fieke Dekkers[§]

Abstract

We study the problem of determining the masses of a set of weights, given one standard weight, based on comparing two disjoint subsets of those weights with approximately equal mass. The question is how to choose a weighing scheme, i.e., different pairs of subsets, such that the masses can be determined as accurately as possible within a given number of measurements. In this paper we discuss a new way of using the so-called STS method of comparing two approximately equal masses, and we will give optimal weighing schemes which turn out to outperform schemes that are currently used by national metrology institutes.

1 Introduction

In this paper we study the following problem presented to us by the Dutch metrology institute NMi (the 'Nederlands Meetinstituut'). Consider the following set of weights $M_0, \ldots, M_5$: one weight of 1000 g, one of 500 g, two of 200 g, and two of 100 g, respectively. The specified masses of these six weights are approximate, since their true masses, let us call them $\mu_0, \ldots, \mu_5$, are unknown. We want to estimate the true masses by comparing the weights in a certain way to each other and to the standard 1 kg, which is a platinum-iridium cylinder stored at the NMi. The comparison of weights is done using an electronic scale. This scale is capable of measuring small differences in mass quite accurately. Hence, we will only compare two sets of weights that have approximately the same total mass, and measure the difference in mass (a direct comparison). Consequently, we have to devise a weighing scheme that tells us which combination of weights we have to compare to which other combination of weights. For example, we could compare weight M_0 ($\mu_0 \approx 1000$) to the combination of weights M_1, M_2, M_3, and M_5 ($\mu_1 + \mu_2 + \mu_3 + \mu_5 \approx 1000$). For practical reasons, the two combinations in a direct comparison may not both contain the same weight. The problem now is to find a weighing scheme for a given number of measurements such that the relative error in the masses is as small as possible. To solve this problem, we will first take a closer look at the weighing procedure.

[*] Vrije Universiteit Amsterdam
[†] Vorarlberg University of Applied Sciences
[‡] Delft University of Technology
[§] Utrecht University

2 The STS weighing procedure

Suppose that we have selected two sets of weights with approximately the same total mass. How do we measure the difference in mass? The NMi uses what is called an STS-procedure. One set of weights is called the Standard set (S), and the other set is called the Test set (T). The set S is placed on the scale, at which time the scale is set to 0. Then the set S is removed and placed on the scale again, resulting in the first measurement, which we call x_1. After this, the set S is removed, and the set T is placed on the scale, resulting in measurement x_2. We then continue by alternating sets S and T, so the third measurement is of set S, the fourth of T, and so on; hence the name STS-procedure.

The data gathered by the STS-procedure will consist of k measurements $x_1, \ldots, x_k$. If we denote the mass of set S by μ_S, and the mass of set T by μ_T, we model the measurements x_i as realizations of the following random variables:

$$x_i = 1_{\{i=\text{even}\}}(\mu_T - \mu_S) + D(i) + V_i. \tag{1}$$

Here, V_i is a random effect (a measurement error), and we model the V_i's as independent and identically distributed random variables. The function $D(i)$ models the *drift* of the electronic scale, which is observed in practice: after each measurement the scale is set off by a small amount. We will now make some additional important assumptions:

(A1) We assume that the V_i's are independent, and that

$$V_i \sim N(0, \alpha^2 \mu_S^2).$$

Here $\alpha > 0$ is an unknown constant.

(A2) We assume that for all $1 \leq j \leq k-2$,

$$\tfrac{1}{2}\big(D(j) + D(j+2)\big) - D(j+1) \approx 0,$$

i.e., this quantity is negligibly small.

Assumption (A1) implies that the measurement error V_i is proportional to the mass being weighed. Here we use that $\mu_T \approx \mu_S$, so the difference is small compared to the total mass. Assumption (A2) is exactly fulfilled when the drift is linear. In fact, we only assume that the drift between two consecutive measurements is almost equal.

Having obtained the measurements $x_1, \ldots, x_k$, how should we use them to estimate the difference in mass $\Delta\mu$ given by

$$\Delta\mu := \mu_T - \mu_S?$$

To benefit from Assumption (A2), we define the following auxiliary variables for $1 \leq j \leq k-2$:

$$\Delta m_j = (-1)^{j+1}\big(x_{j+1} - \tfrac{1}{2}(x_j + x_{j+2})\big),$$

so

$$\begin{aligned} \Delta m_1 &= x_2 - \tfrac{1}{2}(x_3 + x_1), \\ \Delta m_2 &= \tfrac{1}{2}(x_4 + x_2) - x_3, \end{aligned}$$

and so forth. Using Equation (1) and Assumption (A2), we find that

$$\begin{pmatrix} \Delta m_1 \\ \Delta m_2 \\ \vdots \\ \Delta m_{k-2} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \Delta\mu + \begin{pmatrix} -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \dots & 0 & 0 & 0 \\ 0 & \frac{1}{2} & -1 & \frac{1}{2} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \mp\frac{1}{2} & \pm 1 & \mp\frac{1}{2} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_k \end{pmatrix}.$$

If we define, for $k \geq 3$

$$B_k = \begin{pmatrix} -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \dots & 0 & 0 & 0 \\ 0 & \frac{1}{2} & -1 & \frac{1}{2} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \mp\frac{1}{2} & \pm 1 & \mp\frac{1}{2} \end{pmatrix} \text{ and } \underline{\mathbf{1}} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

we see that

$$\Delta m = \underline{\mathbf{1}}\Delta\mu + B_k V. \tag{2}$$

Let Z' denote the transpose of an arbitrary matrix or vector Z. To illustrate how we should calculate an estimate for $\Delta\mu$ we choose an invertible $(k-2)\times(k-2)$ matrix D_k such that

$$D_k D_k' = B_k B_k' =: \Sigma_k.$$

This is always possible, because Σ_k is symmetric and positive definite. Now we multiply Equation (2) by D_k^{-1}:

$$D_k^{-1}\Delta m = D_k^{-1}\underline{\mathbf{1}}\Delta\mu + D_k^{-1}B_k V. \tag{3}$$

We know that

$$D_k^{-1}B_k V \sim N_{k-2}(0, \alpha^2\mu_S^2 D_k^{-1}B_k B_k' D_k'^{-1}) \stackrel{\mathrm{d}}{=} N_{k-2}(0, \alpha^2\mu_S^2 I),$$

where I is the identity matrix. This shows that Model (3) corresponds to a standard linear model

$$Y = X\beta + U,$$

with $Y = D_k^{-1}\Delta m$, $X = D_k^{-1}\underline{\mathbf{1}}$, $\beta = \Delta\mu$, and $U = D_k^{-1}B_k V$. For this model the least squares estimator is given by

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

which gives us

$$\widehat{\Delta\mu} = \frac{\underline{\mathbf{1}}'\Sigma_k^{-1}\Delta m}{\underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}}}. \tag{4}$$

This estimator differs from the estimator that is normally used in the STS-procedure, namely the average of all the Δm_i's. In fact, Estimator (4) is a weighted average of the Δm_i's, where the weights may be negative! For example, when $k = 10$ (i.e., there are ten measurements in the STS procedure), we get that

$$\widehat{\Delta\mu} = (0.30, -0.10, 0.25, 0.05, 0.05, 0.25, -0.10, 0.30)\cdot\Delta m.$$

We can show that in this case, the variation in the least squares estimator is almost 10% smaller than the variation in the average of the Δm_i's. It is true, however, that as k grows, the ratio of the two variances tends to 1.

We know that the variance of the least squares estimator is given by

$$\mathrm{Var}(\widehat{\Delta\mu}) = \frac{\alpha^2 \mu_S^2}{\underline{\mathbf{1}}' \Sigma_k^{-1} \underline{\mathbf{1}}}. \tag{5}$$

The usual unbiased estimator of this variance is given by

$$S^2 = \left(\Delta m' \Sigma_k^{-1} \Delta m - \frac{(\underline{\mathbf{1}}' \Sigma_k^{-1} \Delta m)^2}{\underline{\mathbf{1}}' \Sigma_k^{-1} \underline{\mathbf{1}}} \right) / (k-3).$$

This follows directly from the fact that Model (3) is a standard linear model.

3 Weighing schemes using the STS-procedure

Now that we know how to deal with the STS-procedure, we would like to have some idea of which weighing scheme we should use to accurately determine the masses $\mu_0, \ldots, \mu_5$. Since all the weights are unknown, we need to use the standard 1 kg weight in our weighing scheme. However, we do not want to use this precious weight very often, so we only use it to determine μ_0, which is approximately 1 kg. We proceed with the STS-procedure until we have determined μ_0 up to a certain accuracy. From then on, we will use different combinations of weights to determine $\mu_1, \ldots, \mu_5$ in terms of μ_0. In those combinations we will not use the standard 1 kg.

A combination of weights that can be used for the STS-procedure, can be described by a vector containing 5 entries, each of which is either -1, 0, or $+1$. Each entry corresponds to one of the weights $\mu_1, \ldots, \mu_5$ in the following way: a 0 indicates that the weight is not included, a -1 indicates that the weight is included in the Standard set of the STS-procedure, and a $+1$ indicates that the weight is included in the Test set of the STS-procedure. The reason that there is no entry for μ_0 is because the total mass of the Standard set has to be approximately equal to the total mass of the Test set. This means that the entry for μ_0 is determined by the other entries. Also, we define new parameters

$$\Delta\mu = \begin{pmatrix} \Delta\mu_1 \\ \Delta\mu_2 \\ \Delta\mu_3 \\ \Delta\mu_4 \\ \Delta\mu_5 \end{pmatrix} = \begin{pmatrix} \mu_1 - 0.5\,\mu_0 \\ \mu_2 - 0.2\,\mu_0 \\ \mu_3 - 0.2\,\mu_0 \\ \mu_4 - 0.1\,\mu_0 \\ \mu_5 - 0.1\,\mu_0 \end{pmatrix}.$$

Note that $\Delta\mu$ is small compared to μ.

Now suppose we have a combination of weights characterized by a (row) vector $w = (w_1, \ldots, w_5)$. Define for convenience

$$(M_0, M_1, M_2, M_3, M_4, M_5) = (1, 0.5, 0.2, 0.2, 0.1, 0.1),$$

so we have that $\mu_i \approx M_i$, for $0 \leq i \leq 5$. Define

$$w_0 = -\sum_{i=1}^{5} M_i w_i. \tag{6}$$

Then w_0 has to be either -1, 0 or 1. This means that there are essentially only ten possible choices of w, not taking into account interchanging the Standard set and the Test set (this corresponds to taking $-w$). These ten possible combinations are given in the following matrix W:

$$W = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ -1 & 1 & 1 & 1 & 0 \\ -1 & 1 & 1 & 0 & 1 \\ 0 & -1 & 1 & -1 & 1 \\ 0 & -1 & 1 & 1 & -1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Given a combination w and using Equation (6), we have for our STS-procedure that

$$\mu_S = \sum_{0 \le i \le 5 : w_i = -1} \mu_i \quad \text{and} \quad \mu_T = \sum_{0 \le i \le 5 : w_i = +1} \mu_i.$$

This means that

$$\mu_T - \mu_S = \sum_{i=0}^{5} w_i \mu_i = \sum_{i=1}^{5} w_i \Delta\mu_i = w\Delta\mu.$$

If we take k measurements in the STS-procedure, we see that Model (3) corresponds to

$$D_k^{-1}\Delta m = D_k^{-1}\underline{\mathbf{1}} w\Delta\mu + D_k^{-1} B_k V, \tag{7}$$

where

$$D_k^{-1} B_k V \sim N_{k-2}(0, \alpha^2 \mu_S^2 I). \tag{8}$$

It is of course impossible to estimate the full vector $\Delta\mu$ from these data, we can only estimate the linear combination $w\Delta\mu$. If we repeat the STS-procedure for a suitable set of different combinations, we can estimate $\Delta\mu$ as well. A choice of different combinations of weights is called a *weighing scheme*. A weighing scheme can be represented by a matrix A, consisting of different rows $A(l)$, which correspond to rows from the matrix W, i.e., the possible combinations of weights. For now we will assume that each row of A corresponds to $k = 20$ STS measurements (the number used by the NMi), using that particular combination of weights.

Equation (8) shows that we have to be a bit careful: changing the combination of weights might change μ_S, which would imply that in the full model, that takes all chosen combinations of weights into account, we would not have a measurement error with a constant variance. The way to handle this is quite straightforward: we rescale all STS measurements using a weight-combination w by dividing them by the total mass of the Standard set. With a slight abuse of notation, the total mass $\mu_S(w)$ of the Standard set given w is in good approximation given by

$$\mu_S(w) = \sum_{i=0}^{5} M_i 1_{\{w_i = +1\}}.$$

Model (7) then becomes

$$\frac{D_k^{-1}\Delta m}{\mu_S(w)} = \frac{D_k^{-1}\underline{\mathbf{1}}w\Delta\mu}{\mu_S(w)} + \frac{D_k^{-1}B_kV}{\mu_S(w)},$$

where

$$\frac{D_k^{-1}B_kV}{\mu_S(w)} \sim N_{k-2}(0,\alpha^2 I).$$

Now if our weighing scheme A consists of s rows $A(1),\ldots,A(s)$, then our full linear model becomes

$$\begin{pmatrix} \frac{(D_k^{-1}\Delta m^{(1)})_1}{\mu_S(A(1))} \\ \vdots \\ \frac{(D_k^{-1}\Delta m^{(1)})_{k-2}}{\mu_S(A(1))} \\ \frac{(D_k^{-1}\Delta m^{(2)})_1}{\mu_S(A(2))} \\ \vdots \\ \frac{(D_k^{-1}\Delta m^{(2)})_{k-2}}{\mu_S(A(2))} \\ \vdots \\ \frac{(D_k^{-1}\Delta m^{(s)})_1}{\mu_S(A(s))} \\ \vdots \\ \frac{(D_k^{-1}\Delta m^{(s)})_{k-2}}{\mu_S(A(s))} \end{pmatrix} = \begin{pmatrix} \frac{(D_k^{-1}\underline{\mathbf{1}})_1\, A(1)\Delta\mu}{\mu_S(A(1))} \\ \vdots \\ \frac{(D_k^{-1}\underline{\mathbf{1}})_{k-2}\, A(1)\Delta\mu}{\mu_S(A(1))} \\ \frac{(D_k^{-1}\underline{\mathbf{1}})_1\, A(2)\Delta\mu}{\mu_S(A(2))} \\ \vdots \\ \frac{(D_k^{-1}\underline{\mathbf{1}})_{k-2}\, A(2)\Delta\mu}{\mu_S(A(2))} \\ \vdots \\ \frac{(D_k^{-1}\underline{\mathbf{1}})_1\, A(s)\Delta\mu}{\mu_S(A(s))} \\ \vdots \\ \frac{(D_k^{-1}\underline{\mathbf{1}})_{k-2}\, A(s)\Delta\mu}{\mu_S(A(s))} \end{pmatrix} + U, \tag{9}$$

where $\Delta m^{(l)}$ is the vector of $k-2$ measurements from the STS-procedure using the weight combination $A(l)$ and $U \sim N_{s(k-2)}(0,\alpha^2 I)$. Now define

$$\Delta\tilde{m}^{(l)} = \frac{\Delta m^{(l)}}{\mu_S(A(l))} \text{ and } \tilde{A}(l) = \frac{A(l)}{\mu_S(A(l))}.$$

Use a matrix-block notation to see that Model (9) becomes

$$\begin{pmatrix} D_k^{-1} & & \\ & \ddots & \\ & & D_k^{-1} \end{pmatrix} \begin{pmatrix} \Delta\tilde{m}^{(1)} \\ \vdots \\ \Delta\tilde{m}^{(s)} \end{pmatrix} = \begin{pmatrix} D_k^{-1}\underline{\mathbf{1}} & & \\ & \ddots & \\ & & D_k^{-1}\underline{\mathbf{1}} \end{pmatrix} \tilde{A}\Delta\mu + U.$$

This is a standard linear model and the least squares estimator is given by

$$\widehat{\Delta\mu} = \left(\tilde{A}'\begin{pmatrix} \underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}} & & \\ & \ddots & \\ & & \underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}} \end{pmatrix}\tilde{A}\right)^{-1} \tilde{A}'\begin{pmatrix} \underline{\mathbf{1}}'\Sigma_k^{-1} & & \\ & \ddots & \\ & & \underline{\mathbf{1}}'\Sigma_k^{-1} \end{pmatrix}\Delta\tilde{m}.$$

The covariance matrix of this estimator is given by

$$\mathrm{Cov}(\widehat{\Delta\mu}) = \alpha^2\left(\tilde{A}'\begin{pmatrix} \underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}} & & \\ & \ddots & \\ & & \underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}} \end{pmatrix}\tilde{A}\right)^{-1}. \tag{10}$$

The diagonal of the covariance matrix gives us the variances of the separate estimators for $\Delta\mu_1, \ldots, \Delta\mu_5$. Clearly, we would like to choose our weighing scheme A such that these variances are minimized in some way. We believe that it is most sensible to minimize the relative error in each weight, which is why we choose the sum of squares of the relative errors as a measure of inaccuracy. Thus, we wish to find a weighing scheme A that minimizes the "loss function"

$$L(A) = \sum_{i=1}^{5} \frac{\mathrm{Var}(\widehat{\Delta\mu}_i)}{M_i^2}.$$

In the next sections we will discuss our findings and make a comparison with schemes that are actually used by national metrology institutes, such as the NMi, the Slovak metrology institute SMU, and the German metrology institute PTB. We already wish to note that Equation (10) can be easily generalized to the case where each row of A has a different number of STS measurements: simply use the appropriate k for each Σ_k in the block matrix.

For the sake of completeness, we will write down the estimate for α^2. Define

$$S = \begin{pmatrix} \underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}} & & \\ & \ddots & \\ & & \underline{\mathbf{1}}'\Sigma_k^{-1}\underline{\mathbf{1}} \end{pmatrix}.$$

Then

$$\hat{\alpha}^2 = \big(\Delta\tilde{m}'S\Delta\tilde{m} - \widehat{\Delta\mu}'\tilde{A}'S\tilde{A}\widehat{\Delta\mu}\big)/(n-1),$$

where n is the length of $\Delta\tilde{m}$.

4 Optimal weighing schemes for the NMi

The Dutch metrology institute NMi currently uses a weighing scheme with 8 combinations of weights (i.e., eight rows of W) and 20 STS measurements for each combination (i.e., $k = 20$). Let $W(l)$ denote the l-th row of the matrix W. Then the NMi weighing scheme A_{NMi} is given by

$$A_{\text{NMi}} = \big(W(1), W(2), W(3), W(4), W(5), W(6), W(8), W(9)\big),$$

with the following uncertainty associated with it

$$L(A_{\text{NMi}}) = 1.1812 \cdot \alpha^2 = 1.1812,$$

if we assume that $\alpha = 1$. We can do this without loss of generality, since we are comparing different weighing schemes with each other that all have the same factor α^2. Therefore, we shall assume that $\alpha = 1$ in the following.

In order to improve upon this weighing scheme we have certain degrees of freedom. Firstly, we can choose different weighing schemes A by choosing different combinations of weights from the matrix W. We shall describe matrix A by listing the indices of the chosen rows $W(l)$, e.g., the indices of A_{NMi} are 1, 2, 3, 4, 5, 6, 8, and 9. Secondly, we can change the number of combinations we choose from W, this

s	indices of $A = \big(W(1), \ldots, W(s)\big)$	$L(A)$
8	1, 1, 2, 4, 7, 8, 9, 10	0.8468
9	1, 1, 2, 3, 4, 7, 8, 9, 10	0.7242
10	1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.6441
11	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.5963
12	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10, 10	0.5507
13	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 9, 10, 10	0.5054
14	1, 1, 1, 2, 2, 3, 4, 4, 7, 8, 8, 9, 10, 10	0.4655

Table 1: Optimal weighing schemes for different s

number will, as before, be denoted by s. Finally, we can change the number of STS measurements per combination, this number was already denoted by k.

Let us now study what the optimal weighing scheme is for the NMi, and how this can be improved by increasing the number of combinations s in the weighing scheme. Table 1 summarizes these experiments (calculated using MatLab) and shows that the optimal weighing scheme with the same parameters used at the NMi (i.e., $s = 8$ and $k = 20$) already gives a reduction in uncertainty of around 28 percent. Note that there can be more weighing schemes with the same uncertainty which for simplicity we have not mentioned in Table 1. The uncertainty can be reduced even more by adding more combinations of weights to the scheme.

As a side remark, the Slovak metrology institute SMU (the 'Slovenský Metrologický Ústav') has the same set of weights. However, they use a weighing scheme with $s = 14$ of which the indices are given by 1, 2, 3, 4, 5, 6, 7, 7, 8, 8, 9, 9, 10, and 10. The associated uncertainty with this weighing scheme is given by $L(A_{\text{SMU}}) = 0.7285$. Table 1 shows that the optimal weighing scheme reduces the uncertainty with 36 percent compared to current practice.

Note that the optimal weighing schemes do not include measurements $W(5)$ and $W(6)$, which the NMi does include. Instead, the optimal schemes include measurements $W(7)$ and $W(10)$. One could ascribe this to rounding errors in the calculation, however, all solutions within 1 percent of the optimal solution have this property as well. This observation can be explained if one realizes that $W(5)$ and $W(6)$ provide the same information as $W(7)$ and $W(10)$, only with a greater uncertainty associated with them. This can be seen by adding and subtracting the rows: $W(5) + W(6) = 2 \times W(7)$, and $W(5) - W(6) = 2 \times W(10)$. Hence, both sets provide the same information, however, the set with $W(7)$ and $W(10)$ only uses one weight one the scale and thus adds less uncertainty to the measurements.

Although Table 1 lists optimal weighing schemes, it disregards the total number of measurements. For fixed s and k we need in total $s \times k$ measurements to determine the masses. The current weighing scheme needs $8 \times 20 = 160$ measurements. The measurements used to be done by hand, constraining the maximum number of measurements in a weighing scheme. However, due to the introduction of automatic weighing devices at the NMi the maximum number of measurements has been increased to around 280 to 300. In Table 2 we have computed the optimal weighing schemes as a function of the total number of measurements from 280 to 300. From the table we can see that with these parameters a reduction of around 62 percent in uncertainty can be achieved. We again mention that there might be more weighing

$s \times k$	s	k	indices of $A = \big(W(1), \ldots, W(s)\big)$	$L(A)$
280	14	20	1, 1, 1, 2, 2, 3, 4, 4, 7, 8, 9, 9, 10, 10	0.4655
280	10	28	1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.4659
280	8	35	1, 1, 2, 4, 7, 8, 9, 10	0.4940
286	13	22	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 9, 10, 10	0.4613
286	11	26	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.4633
288	9	32	1, 1, 2, 3, 4, 7, 8, 9, 10	0.4602
288	12	24	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10, 10	0.4623
288	8	36	1, 1, 2, 4, 7, 8, 9, 10	0.4799
290	10	29	1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.4513
294	14	21	1, 1, 1, 2, 2, 3, 4, 4, 7, 8, 9, 9, 10, 10	0.4461
296	8	37	1, 1, 2, 4, 7, 8, 9, 10	0.4679
297	9	33	1, 1, 2, 3, 4, 7, 8, 9, 10	0.4474
297	11	27	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.4479
299	13	23	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 9, 10, 10	0.4435
300	10	30	1, 1, 2, 2, 3, 4, 7, 8, 9, 10	0.4358
300	12	25	1, 1, 1, 2, 2, 3, 4, 7, 8, 9, 10, 10	0.4458

Table 2: Optimal weighing schemes for different s and k

schemes that achieve the same uncertainty, but which for simplicity we have not included into Table 2.

5 Improved weighing schemes for the PTB

Let us consider the set of weights used by the German metrology institute PTB (the 'Physikalisch-Technische Bundesanstalt'), as reported in Kochsiek and Gläser [1]. Apart from the standard 1000 g, this set has another weight of 1000 g, two of 500 g, two of 200 g, and two of 100 g. For this set there are 104 possible combinations of weights in the matrix W instead of 10. Identifying by full enumeration the optimal weighing scheme with say $s = 10$ out of all 104 combinations requires enormous computational resources. Therefore we only consider a reduced class of possible combinations, namely those combinations involving on one side only one weight. Note that this is a reasonable choice by the same argument we used to exclude $W(7)$ and $W(10)$ in the previous section. The matrix W now becomes

$$W = \begin{pmatrix}
-1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -1 & -1 & -1 & 0 \\
0 & -1 & 0 & -1 & -1 & 0 & -1 \\
0 & 0 & -1 & -1 & -1 & -1 & 0 \\
0 & 0 & -1 & -1 & -1 & 0 & -1 \\
1 & -1 & -1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & -1 & -1 & -1 & 0 \\
1 & -1 & 0 & -1 & -1 & 0 & -1 \\
1 & 0 & -1 & -1 & -1 & -1 & 0 \\
1 & 0 & -1 & -1 & -1 & 0 & -1 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & -1 & -1 & -1 & 0 \\
0 & 1 & 0 & -1 & -1 & 0 & -1 \\
0 & 0 & 1 & -1 & -1 & -1 & 0 \\
0 & 0 & 1 & -1 & -1 & 0 & -1 \\
0 & 0 & 0 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & -1 \\
0 & 0 & 0 & 0 & 1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 1 & -1
\end{pmatrix}.$$

	indices of $A = \big(W(1), \ldots, W(10)\big)$	$L(A)$
A_{opt1}	1, 3, 6, 6, 8, 13, 16, 18, 19, 20	1.1356
A_{opt2}	1, 4, 5, 5, 9, 14, 15, 18, 19, 20	1.1356
A_{PTB}	1, 2, 7, 12, 13, 16, 17, 18, 19, 20	1.6244

Table 3: Improved weighing schemes for the PTB set of weights

Having specified the matrix W, we can repeat the calculations for this set of weights using the parameters $s = 10$ and $k = 20$. Table 3 shows the weighing schemes A_{opt1} and A_{opt2} that are optimal in the reduced class, and compares them to the scheme A_{PTB} of the PTB. We can see that the optimal weighing schemes even in the reduced class leads to a 30 percent reduction in uncertainty compared to the PTB weighing scheme presently used.

Acknowledgements

We thank Inge van Andel from the NMi (the Dutch metrology institute) for the helpful discussions and for providing us with detailed information on current weighing practices.

References

[1] M. Kochsiek and M. Gläser (editors). *Comprehensive Mass Metrology*. John Wiley and Sons, Inc.-VCH, 2000.

Partitioning a Call Graph

Rob H. Bisseling[*] *Jarosław Byrka*[†] *Selin Cerav-Erbas*[‡]
Nebojša Gvozdenović[†] *Mathias Lorenz*[‡] *Rudi Pendavingh*[§]
Colin Reeves[¶] *Matthias Röger*[§] *Arie Verhoeven*[§]

Abstract

Splitting a large software system into smaller and more manageable units has become an important problem for many organizations. The basic structure of a software system is given by a directed graph with vertices representing the programs of the system and arcs representing calls from one program to another. Generating a good partitioning into smaller modules becomes a minimization problem for the number of programs being called by external programs. First, we formulate an equivalent integer linear programming problem with 0–1 variables. Theoretically, with this approach the problem can be solved to optimality, but this becomes very costly with increasing size of the software system. Second, we formulate the problem as a hypergraph partitioning problem. This is a heuristic method using a multilevel strategy, but it turns out to be very fast and to deliver solutions that are close to optimal.

1 Introduction

In recent years, the capabilities of information technology have increased tremendously. At the same time, large software systems in today's organizations such as banks, health care providers, or government agencies, have become costly to maintain. To reduce the maintenance costs, the systems need to be split into smaller, more manageable modules (typically 5–10 modules). Each module can then be assigned to a separate team of maintainers. A partitioned system needs interfaces for the communication between modules; the number of interfaces is the main cost factor. In a good partitioning, the size of each module is restricted and the total size of the interface is minimized. To find such a partitioning is a job for a trained expert, but when the system is large an automated suggestion for a partitioning into modules becomes useful.

A software system can be described by a *call graph*. A call graph is a directed graph, where vertices represent programs, classes, or similar program units, and where an arc (v, w), i.e. $v \to w$, means that program v calls program w. Figure 1 shows a complete call graph of a Java software system named `Java1` and Figure 2 shows part of this graph in detail. Each vertex may have a weight, such as the number of lines

[*]Universiteit Utrecht
[†]Centrum voor Wiskunde en Informatica
[‡]Université Catholique de Louvain
[§]Technische Universiteit Eindhoven (`rudi@win.tue.nl`)
[¶]Coventry University

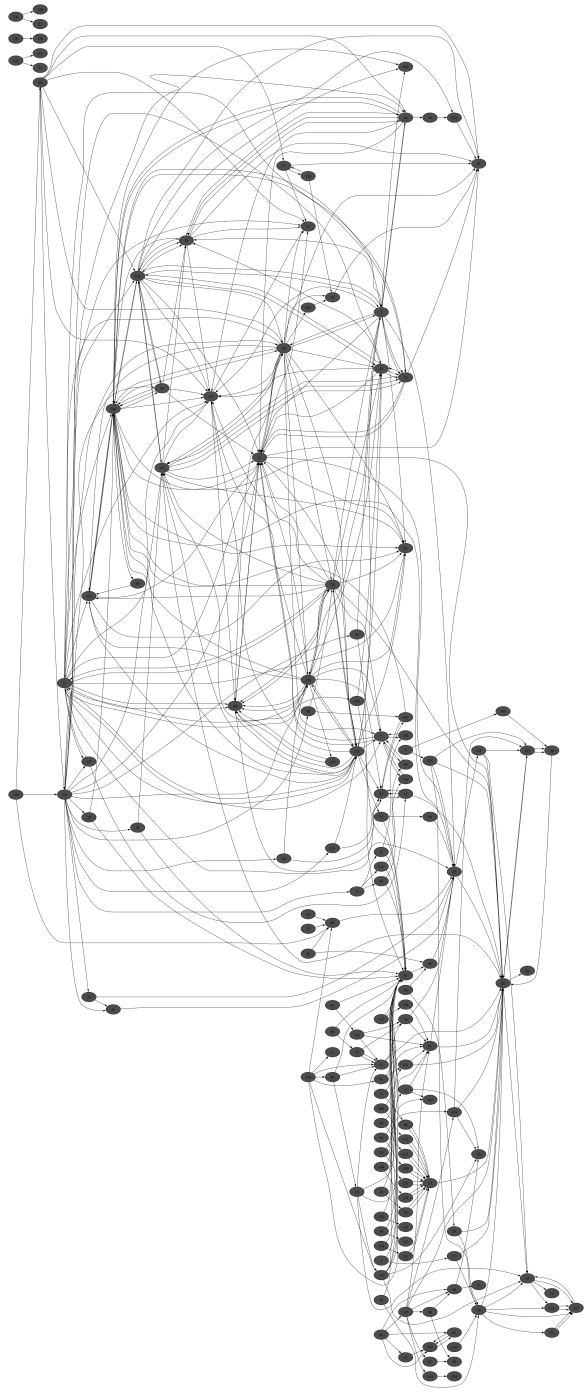


Figure 1: Call graph of the software system `Java1`, which was provided by SIG. The number of vertices (programs) is $N = 158$.

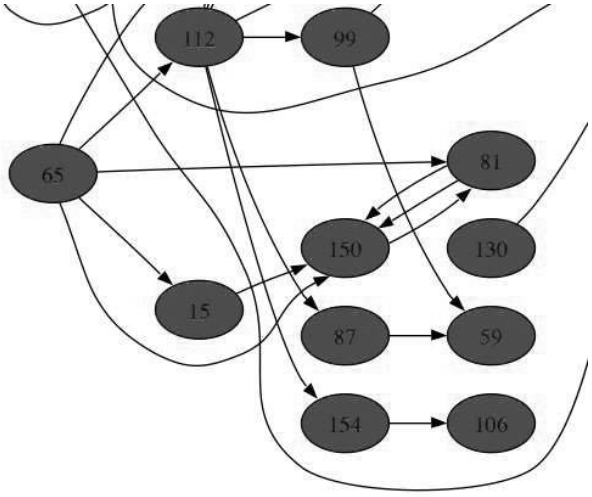


Figure 2: Detailed view of the bottom part of the call graph from Figure 1 showing the arcs between the programs. Note for instance that program 65 calls program 81. Program 81 calls 150 and vice versa. Duplicate calls may occur; these can be removed without affecting the problem.

of code of the corresponding program. Here, we assume that less detail is required so that all programs are equally costly to maintain, and hence all vertices have weight 1. A module contains a subset of the vertices, representing a subset of the programs. The size of a module equals the sum of the vertex weights in the corresponding subset; the size of its interface equals the number of vertices which have an incoming arc from a different module. Thus, the software splitting problem can be formulated as a partitioning problem of a call graph.

The Software Improvement Group (SIG, `http://www.sig.nl`), located in Diemen (the Netherlands), provides tools which help organizations to understand their software better. They are interested in partitioning algorithms which solely exploit the structure of the call graph to split the software and which are useful for various graph sizes, from hundreds of vertices to over a million.

The call-graph partitioning problem was posed by SIG at the opening day of the Study Group Mathematics with Industry 2005 in Amsterdam. We have studied the problem and in this report we propose our solutions. The following sections present different formulations of the same problem and different solution methods. In Section 2, the problem is mathematically formulated as a graph partitioning problem, and then translated into an integer linear programming (ILP) problem with variables taking values only in $\{0,1\}$, which can be solved by standard commercial software such as CPLEX [4]. Because the ILP problem is NP-hard, it cannot be solved to optimality for very large call graphs (more than 1500 vertices). Therefore, Section 3 describes a very fast heuristic method which is based on a multilevel approach to hypergraph partitioning. Section 4 compares the different methods for some real call graphs of software systems written in Java and COBOL, which were provided by SIG. Finally, Section 5 presents our conclusions and recommendations.

2 Solution by integer linear programming

We will first formulate the problem as a graph partitioning problem and then translate it into an ILP problem with 0–1 variables.

2.1 Graph partitioning problem

A call graph is a directed graph $D = (V, A)$, where the vertex set V is the set of programs of the software system and the arc set $A := \{(u, v) \in V \times V \mid u \text{ calls } v\}$. Given a subset of programs $U \subseteq V$, the *interface* of U in D is the set of all programs in U called by programs not in U:

$$I_D(U) := \{u \in U \mid (v, u) \in A \text{ for some } v \in V \setminus U\}. \tag{1}$$

We call $u \in I_D(U)$ an *interface vertex* of U. A *partition* of a set V is a collection of nonempty, pairwise disjoint subsets of V, such that the union of these subsets is V.

With these definitions, a mathematical formulation of the graph partitioning problem is:

Given: A directed graph $D = (V, A)$, $K \in \mathbb{N}$, $L \in \mathbb{N}$
Find: A partition $V_1, \ldots, V_L$ of V such that $|V_l| \leq K$ for each l, and such that $\sum_{l=1}^{L} |I_D(V_l)|$ is as small as possible.

2.2 Integer linear programming problem

Consider the following ILP:

$$\begin{array}{lll}
\text{minimize} & \sum_{l=1}^{L} \sum_{v \in V} x_{vl} & \\
\text{subject to} & \sum_{l=1}^{L} y_{vl} = 1 & \text{for all } v \in V \\
& \sum_{v \in V} y_{vl} \leq K & \text{for } l = 1, ..., L \\
& x_{vl} \leq y_{vl} & \text{for } l = 1, ..., L \text{ and for all } v \in V \\
& y_{vl} \leq y_{ul} + x_{vl} & \text{for } l = 1, ..., L \text{ and for all } (u, v) \in A \\
& x_{vl}, y_{vl} \in \{0, 1\} & \text{for } l = 1, ..., L \text{ and for all } v \in V
\end{array}$$

It is not difficult to see that if $V_1, \ldots, V_L$ is a proper solution to the graph partitioning problem, then by setting

$$y_{vl} = 1 \text{ if } v \in V_l, \text{ and 0 otherwise,} \tag{2}$$

$$x_{vl} = 1 \text{ if } v \text{ is an interface vertex of } V_l, \text{ and 0 otherwise,} \tag{3}$$

we obtain a feasible solution of the above ILP. Conversely, given an optimal solution to the ILP, taking

$$V_l := \{v \in V \mid y_{vl} = 1\}, \tag{4}$$

$$I := \{v \mid x_{vl} = 1 \text{ for some } l\}, \tag{5}$$

will yield an optimal solution $V_1, \ldots, V_L$ to the call-graph partitioning problem with a set of interface vertices I. (It is straightforward to adapt this formulation to the variant

of the call-graph partitioning problem where each program has a certain weight and the total weight of each module is bounded.)

The general ILP problem, which is to solve

$$\min\{c^{\mathrm{T}}x \mid Ax \leq b, x \in \mathbb{Z}^n\} \tag{6}$$

for a given matrix A and vectors b, c, is NP-hard, see [6]. The standard solution methods are efficient in practice when the polyhedron $P := \{x \in \mathbb{R}^n \mid Ax \leq b\}$ is close to the convex hull of $P \cap \mathbb{Z}^n$. We have attempted to create a formulation of our problem with this property — which is why we chose this formulation over others with less variables/constraints. For detailed information on integer programming theory and methods, see [7]. A textbook is [9].

In our formulation of the call-graph problem, each feasible partition $V_1, \ldots, V_L$ is represented at least $L!$ times, since each permutation of the subsets of the partition yields a different binary vector y. This decreases the efficiency of the standard solution methods, so some form of symmetry breaking is desired. To eliminate the abundance of representations of essentially the same feasible solution, we added, for certain vertices $s_1, \ldots, s_{L-1}$, the set of constraints

$$\sum_{i=1}^{l} y_{s_l i} = 1 \text{ for } l = 1, ..., L-1 \tag{7}$$

to our model. Thus, the first vertex is fixed in V_1, the second in $V_1 \cup V_2$, and so on. These constraints will allow at least one representation y of each feasible partition $V_1, \ldots, V_L$ in the feasible set (and exactly one for feasible partitions with all fixed vertices in different subsets). We chose $s_1, \ldots, s_{L-1}$ to be the set of $L-1$ vertices of largest *outdegree* in D (i.e., with the largest number of outgoing arcs). The choice of $s_1, ..., s_{L-1}$ and our method of symmetry breaking is still not optimal. Solving the symmetry problem properly, however, seems the key to solving the call-graph problem through an integer programming formulation. Further improvement of our method is up to future research.

3 Solution by multilevel hypergraph partitioning

We will reformulate the graph partitioning problem as a hypergraph partitioning problem and then present a heuristic solution method based on a multilevel approach.

3.1 Hypergraph partitioning problem

A hypergraph $H = (V, \mathcal{N})$ consists of a set of vertices $V = \{v_1, \ldots, v_N\}$ and a set $\mathcal{N}$ of *hyperedges*, or *nets*, which are subsets of V. A hypergraph is a generalization of an undirected graph: a hyperedge connects an arbitrary number of vertices, whereas an edge in a graph connects two vertices; an edge can be viewed as a subset $\{v_i, v_j\}$ of size two.

As before, let the structure of a software system be given by a directed graph $D = (V, A)$, where $V = \{v_1, ..., v_N\}$ represents the set of programs and $(v_i, v_j) \in A$

indicates that program i calls program j. We consider the hypergraph $H = (V, \mathcal{N})$ and choose the set of nets as

$$\mathcal{N} = \bigcup_{j=1}^{N} n_j, \tag{8}$$

where net n_j consists of program j and all programs that call j,

$$n_j \quad := \quad \{v_j\} \cup \{v_i \mid 1 \le i \le N \text{ and } (v_i, v_j) \in A\}. \tag{9}$$

A net is *broken* by a given partition $V_1, \ldots, V_L$ if its vertices are in different subsets of the partition, i.e., in different modules. A net n_j is broken if and only if program j is an interface program. This is because for a broken net n_j, at least one calling program i must be in a module different than the module of j. The software splitting problem has become a hypergraph partitioning problem where we are looking for a partition that minimizes the total interface size,

$$|I| \quad := \quad |\{j \mid 1 \le j \le N \text{ and } n_j \text{ is broken}\}|.$$

A convenient way of looking at a graph is to consider its *adjacency matrix*. We describe the calls between the N programs by the $N \times N$ adjacency matrix $A = (a_{ij})_{i,j=1,\ldots,N}$, with $a_{ij} \in \{0, 1\}$, defined by

$$a_{ij} \quad = \quad \begin{cases} 1 & \text{if program } i \text{ calls program } j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Thus, $a_{ij} = 1$ if and only if $(v_i, v_j) \in A$. Note that we identify the matrix A with the arc set A. We add a unit diagonal to the adjacency matrix. This way, the vertices of net n_j correspond exactly to the positions of nonzeros in column j of the matrix. Figure 3 shows an extended adjacency matrix.

A traditional application area of hypergraph partitioning is the design of electronic circuits. MLpart [2] is a hypergraph partitioner specifically developed for this purpose. Çatalyürek and Aykanat [3] introduced hypergraph partitioning for the purpose of distributing the work in multiplying a sparse matrix and a vector on a parallel computer, which is the core computation of iterative linear system solvers. They implemented the partitioning in software called PaToH. The package Mondriaan, recently developed at Utrecht University [8], is a two-dimensional sparse matrix partitioner which cuts the matrix recursively into smaller rectangular shapes, similar to the paintings of the Dutch painter Piet Mondriaan (1872–1944). Each cut is based on hypergraph *bi*partitioning, which is explained in the following.

Multilevel methods for graph or hypergraph partitioning reduce the size of the problem repeatedly by merging vertices with similar connectivity until the remaining problem is sufficiently small (a few hundred vertices), then solve the smaller problem for instance by a local heuristic such as the Kernighan–Lin [5] algorithm, and finally unmerge the merged vertices at the different levels, each time refining the solution by a simpler method such as trying to move interface vertices to the other subset of the partition. Typically, each level of merging halves the problem size. A good similarity criterion for merging, used in both PaToH and Mondriaan, is the inner product of the corresponding rows in the adjacency matrix. A large inner product for rows i and i'

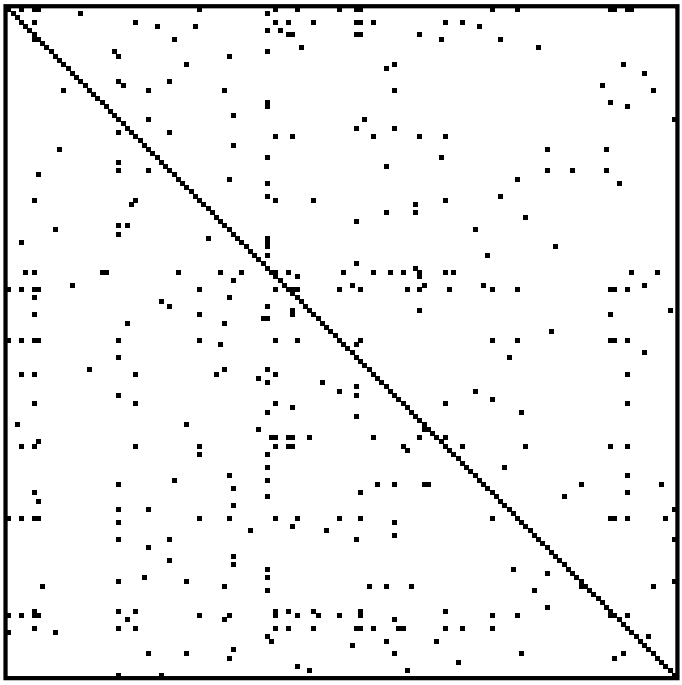


Figure 3: 158×158 adjacency matrix of the problem `Java1`, provided by SIG, which has 158 programs and 422 calls from programs to other programs. The matrix, extended by a unit diagonal, has 580 nonzero elements. It is *sparse*, since the vast majority of its elements is zero.

means that many nonzeros from those rows occur in the same positions, and hence programs i and i' often call the same program j. Multilevel methods, first proposed by Bui and Jones [1], have been very successful in graph and hypergraph partitioning. The Kernighan–Lin algorithm works by trying moves of a vertex to the other subset of the partition, each time accepting the move with the largest gain, i.e., reduction in number of broken nets. A temporary increase is also accepted if this leads to a reduction later on. Several passes are made through the whole set of vertices. This algorithm is local in nature and hence only works for a limited number of vertices; for larger problem sizes, the algorithm easily gets stuck in a local minimum.

To apply Mondriaan (version 1.01) to our problem, we had to make the following adjustments. First, we want to use the matrix partitioner Mondriaan only in one-dimensional (1D) mode, because the partitioning of the vertices we are looking for corresponds to a partitioning of the rows of the adjacency matrix. Our aim is to partition the rows such that as few columns (i.e. nets) as possible are broken. Fortunately, 1D partitioning is a standard option of Mondriaan. Second, Mondriaan penalizes every additional cut of a broken net. This is because each additional column part corresponds to an extra processor involved in the handling of the matrix column in a parallel computation. In the present application, however, the situation is different: once a program becomes an interface program, and serves at least one outside mod-

ule, it does not matter how many such modules it serves. Therefore, we can further break the already broken nets for free. Of course, this will have large effects on our minimization procedure. We modified Mondriaan in various places to take this into account. Third, the work load assumed by Mondriaan is just the number of nonzeros in the corresponding matrix part. In the present application, the workload is 1 for all programs, i.e., for all matrix rows. This was a relatively easy change, since internally Mondriaan already uses arbitrary weights (to enable merging vertices).

4 Comparison of the partitioning methods

4.1 Integer linear programming

We used CPLEX 6.3 (most recent version is 9.0, see [4]) on a 2.8 GHz Intel Xeon processor to solve the ILP model of the call-graph problem. To improve the running times, we provided CPLEX with a *branching order*, specifying that it should branch on x_{vl} before x_{wl} and on y_{vl} before y_{wl} if the outdegree of v in D is larger than the outdegree of w in D, and on y-variables before x-variables.

Table 1 shows the results. The numbers $|V|$ given are less than those for the original call graph, because vertices without outgoing arcs (corresponding to empty rows in the adjacency matrix) were removed beforehand. These programs can be assigned to any module without affecting the interface size $|I|$. The value $L = 8$ was chosen, because typically L is in the range 5–10, and because the current version of Mondriaan requires L to be a power of 2. The value of K was chosen such that no module would have more than 20% extra work compared to the average work of a module: $K = \lfloor 1.2|V|/L \rfloor$, where $|V|$ refers to the original call graph. We removed the small problem `Java2` with $V = 19$ and $|A| = 47$ from our test set, because it is infeasible for the parameter of 20% we chose; it would lead to $K = 2$, so that $KL = 16 < V$. (Of course, we can still find a solution if we are willing to accept more than 20% extra work.) The table gives the best result $|I|$ found, a lower bound on the best result possible, and the fraction $|I|/|V|$ of programs that are actually interface programs in the best solution, where $|V|$ refers to the original call graph.

We terminated problems `Java3` and `Cobol4` after about 4 days of CPU time; we expected that the gap between the best solution and the lower bound was going to be closed only at an extremely slow rate. Problem `Cobol2` aborted due to lack of memory; here, our current strategy for breaking symmetry apparently failed. But inspection of the last lower bound revealed that there could not be a solution better

Problem	$\|V\|$	$\|A\|$	K	L	best $\|I\|$	lower bound	best $\|I\|/\|V\|$(%)	running time (s)	remarks
`Java1`	144	422	23	8	26	26	16.5	2.04×10^2	
`Java3`	837	5252	127	8	251	230	29.5	3.59×10^5	terminated
`Java4`	15	39	2	8	11	11	68.8	0.22	
`Cobol1`	947	1900	209	8	13	13	0.9	1.06×10^3	
`Cobol2`	449	659	81	8	6	6	1.1	7.51×10^4	aborted
`Cobol3`	1145	2686	203	8	51	51	3.8	3.78×10^5	
`Cobol4`	1100	2951	167	8	32	28	2.9	3.60×10^5	terminated

Table 1: Integer linear programming results

than the one found anyway. Considering intermediate results, the best solution found for the larger problems after a day of CPU time was 251 for `Java3`, 57 for `Cobol3`, and 47 for `Cobol4`.

4.2 Multilevel hypergraph partitioning

Table 2 presents the results for 10 runs of the program Mondriaan, version 1.01, modified for this purpose. Mondriaan was run on an 867 MHz Apple PowerBook G4 computer running MacOS 10.2. Since Mondriaan uses a random number generator, we can run it with different random number seeds and get different solutions. The table shows the best result obtained in 10 runs, the average result, and also the average running time. The default settings of Mondriaan were used, except that the program was run in 1D mode and with random seed. An important default is that the multilevel algorithm moves over to the Kernighan–Lin algorithm when the number of vertices is 200 or less. This means that problems `Java1` and `Java4` were in fact solved by pure Kernighan–Lin. The number $|V|$ is the number of the original call graph; $|A|$ is the number of nonzeros of the matrix extended by a unit diagonal. The K and L values were chosen identical to those for the ILP solution. Therefore, the number of interface programs $|I|$ obtained by the two methods can be compared. (For the timings, the difference between the computers used in our experiments must be taken into account.)

Comparing Tables 1 and 2, we note that 5 out of the 7 feasible problems were solved to optimality by the ILP method and 1 by hypergraph partitioning (HP). The results of the ILP method are always better than those of HP, but never by more than a factor 1.63. The HP method on the other hand is much faster than the ILP method; the solution is almost instantaneous. The HP results can be improved by fine-tuning the Mondriaan parameters for the application at hand, instead of using the defaults which were chosen to obtain good performance for a wide range of applications. A quick trial of a few different parameter settings reduced $|I|$ for `Cobol4` from 52 to 47; further reduction should be possible.

Figure 4 compares the number of interface programs obtained by the ILP and HP methods. All ILP solutions are within a factor 1.14 from optimal; all HP solutions within a factor 1.86. Note that in fact they may even be closer, since the comparison is with a lower bound, not necessarily a known minimum. Figure 5 shows a solution obtained by the HP method for the problem `Java1`.

Problem	$\|V\|$	$\|A\|$	best $\|I\|$	avg $\|I\|$	best $\|I\|/\|V\|$ (%)	running time (s)
`Java1`	158	580	30	30.7	19.0	0.06
`Java3`	851	6103	275	283.2	32.3	0.54
`Java4`	16	55	11	11.2	68.8	0.001
`Cobol1`	1398	3298	17	22.4	1.2	0.33
`Cobol2`	545	1204	10	11.5	1.8	0.12
`Cobol3`	1357	4043	69	74.6	5.1	0.34
`Cobol4`	1116	4067	52	56.5	4.7	0.41

Table 2: Hypergraph partitioning results

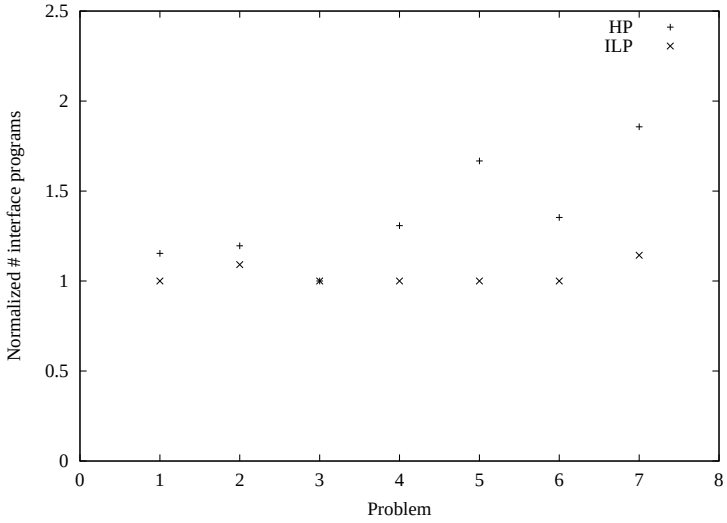


Figure 4: Number of interface programs obtained by the integer linear programming (ILP) method and the hypergraph partitioning (HP) method. The results are normalized by the lower bound provided by the ILP method. A value 1.0 means that the solution is guaranteed to be optimal. The problems are numbered 1–7, which corresponds to `Java1`, `Java3`, `Java4`, `Cobol1`–`Cobol4`.

5 Conclusions

Splitting a large software system into smaller and more manageable units has become an important problem and a challenging task for many organizations. We were introduced to this problem during the Study Group Mathematics with Industry 2005, where it was presented by the Software Improvement Group (SIG). Apart from the information specific to the application, the basic structure of a software system is given as a directed graph with vertices representing the programs of the system and arcs representing calls from one program to another. The question of generating a good partitioning into smaller modules becomes a minimization problem for the number of programs being called by external programs.

During the one week of the study group and in the six weeks of continuing investigations afterwards, we were able to give a clear mathematical description of the problem and to bring in some fresh ideas and new methods. We have presented two different solution strategies, which both seem to be a suitable and valuable tool for the intended applications. The formulations we gave reduce the problem to standard problems in discrete optimization; this makes it possible to apply some state-of-the-art software packages and to deal successfully with the real-world examples which were provided by SIG.

Two different approaches turned out to be promising to tackle the problem. First, we gave an equivalent formulation as an integer linear programming problem with 0–1 variables and used the software package CPLEX to implement this solution strategy. Theoretically, with this approach the problem can be solved to optimality, but this

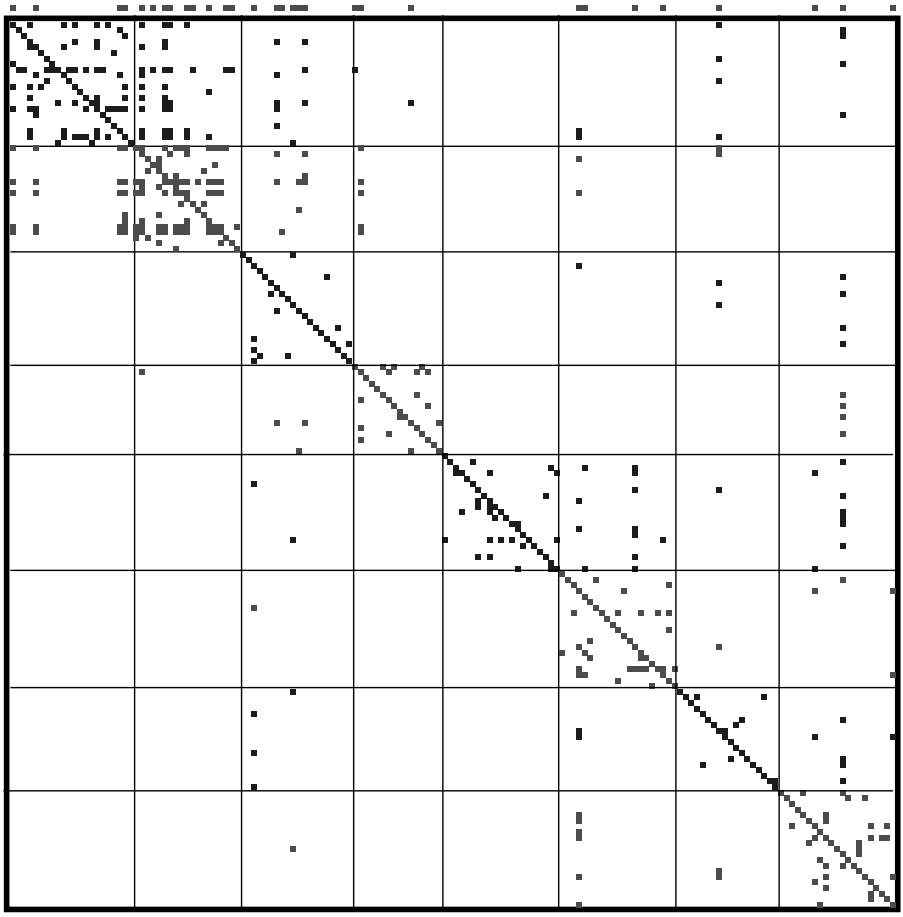


Figure 5: Permuted 158×158 adjacency matrix of the problem `Java1`. The rows were permuted such that rows (programs) belonging to the same module were brought together. Columns were permuted by the same permutation as the rows. Each block of rows corresponds to a subset of the vertices, and hence to a module. The number of modules is $L = 8$; the modules are shown by alternating coloring in red and blue. The number of programs in each module is 22, 19, 20, 16, 21, 21, 18, 21, respectively. The permutation corresponds to a solution with $|I| = 30$ interface programs produced by Mondriaan. Whether a program is an interface program or not can be read from the columns. Columns corresponding to interface programs are marked (above the matrix) and have at least one nonzero outside their diagonal block. Note that the solution method tries to confine all nonzeros to the diagonal blocks. Where this fails, the method does not attempt to limit the number of blocks involved; this explains the spread of the nonzeros over the different blocks.

becomes very costly with increasing size of the software system; obtaining efficient reformulations and a clever implementation becomes an important task. We succeeded to make this method work reasonably well for software systems of the size of the real-world examples.

Second, we have formulated the problem as a hypergraph partitioning problem. We have modified the package Mondriaan, recently developed at Utrecht University, and applied this to the examples given by SIG. This second approach is a heuristic method using a multilevel strategy, but it turns out to be very fast and to deliver solutions that are close to optimal.

Our two methods can drastically reduce the fraction of interface programs, in particular for Cobol systems, where the resulting fraction is at most 5.1%. For small problems, we recommend using the integer linear programming method, perhaps speeding up the solution process by starting with a heuristic solution produced by Mondriaan. For large problems, with thousands of vertices in the call graph, multilevel hypergraph partitioning such as done by Mondriaan is the only realistic option. Here, the performance can be improved by fine-tuning, perhaps aided by experience with smaller problems. Having knowledge of lower bounds or optimal solutions such as provided by the ILP method for smaller problems can be of tremendous help in the fine-tuning.

The problem presenter SIG apparently appreciated the solutions proposed in this report as well as the insight which they could gain from our investigations. Conversely, we have profited from this well-prepared problem which led us to new interesting questions such as for example the comparison between an exact and a heuristic method in a realistic situation. We would be pleased if the strategies presented here would have some impact on the applications and we hope that this collaboration stimulates further joint work between mathematics and industry.

References

[1] T. N. Bui and C. Jones, "A heuristic for reducing fill-in in sparse matrix factorization", in *Proceedings Sixth SIAM Conference on Parallel Processing for Scientific Computing*, pp. 445–452, SIAM, Philadelphia, 1993.

[2] A. E. Caldwell, A. B. Kahng, and I. L. Markov, "Improved Algorithms for Hypergraph Bipartitioning", in: *Proceedings Asia and South Pacific Design Automation Conference*, pp. 661–666, ACM Press, New York, 2000.

[3] Ü. V. Çatalyürek and C. Aykanat, "Hypergraph-Partitioning-Based Decomposition for Parallel Sparse-Matrix Vector Multiplication", *IEEE Transactions on Parallel and Distributed Systems*, **10** (7), pp. 673–693 (1999).

[4] `http://www.ilog.com/products/cplex/`

[5] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs", *Bell System Technical Journal*, **49**, pp. 291–307, (1970).

[6] C. H. Papadimitriou and K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Prentice-Hall, Englewood Cliffs NJ, 1982

[7] A. Schrijver, Theory of Linear and Integer Programming, Wiley, New York, 1986.

[8] B. Vastenhouw and R. H. Bisseling: “A two-dimensional data distribution method for parallel sparse matrix–vector multiplication”, *SIAM Review*, **47** (1), pp. 67–95 (2005).

[9] L. A. Wolsey, Integer Programming, Wiley, New York, 1998.

Mathematical Techniques for Neuromuscular Analysis

JF Williams[*] *Geertje Hek*[†] *Alistair Vardy*[‡]
Vivi Rottschäfer[§] *Jan Bouwe van den Berg*[‡] *Joost Hulshof*[‡]

Abstract

In the central nervous system, *alpha*-motor neurons play a key role in the chain that results in muscles producing force. A new non-invasive technique to measure the electrical activity involved with force production called High Density Surface Electromyography (HDsEMG) has been proven to be effective in providing novel clinical information on the way *alpha*-motor neurons control the muscles. This is important for the monitoring of the progression of certain neuromuscular disorders such as polio. The result of HDsEMG is, however, very difficult to interpret. In this paper we augment the usefulness of HDsEMG with automated mathematical techniques to aid the Motor Unit Number Estimation (MUNE) problem. Also, we create a stochastic model for the firing behavior of an *alpha*-motor neuron.

1 Introduction

The movement of parts of the body is an area studied by many disciplines. Combining the knowledge and techniques of multiple disciplines can help solve problems related to movement in a more fruitful way. Here, we will combine medical science, neuroscience and mathematics. First, we address two questions from medical science. Next we describe the techniques used in neuroscience with which we collect the relevant data. Finally, we describe the mathematics needed to process the data and reflect on the questions posed in this introduction.

Movement requires force produced by muscles. Before the muscles contract a chain of events take place. These events form the basis of the questions we will pose later. For descriptive purposes we assume that the origin of movement is activity in the brain (this is not the only starting point, eg. reflexes do not need intervention from the brain). The brain sends out signals and which consist of 'spikes' called *action potentials* (see Figure 1); signals are built up of trains of action potentials. The brain is part of the central nervous system which embodies all neural tissue in the body including the spinal cord. From the brain, signals travel to the spinal cord. The spinal cord can be divided into sections, called vertebrate discs, which are responsible for

[*]Simon Fraser University
[†]Universiteit van Amsterdam
[‡]Vrije Universiteit Amsterdam
[§]Universiteit Leiden

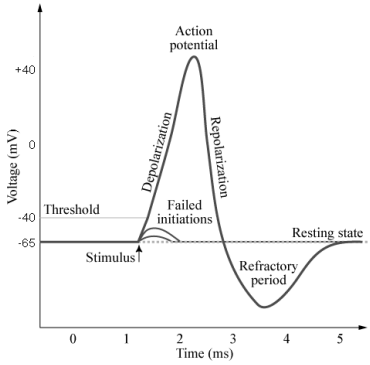


Figure 1: An action potential.

the control of a certain part of the body. The higher levels control the head and arms, the lower levels control the torso and legs. Neurons branch off at every vertebrae. At such a vertebrate disc in the spinal cord, the signal passes a number of intermediate points called interneurons which act as a switchboard redirecting the signals to all the further tissue that requires the information. The signals for movement reach their final (neural) destination at neurons which control the muscles (of which the cell body also lies in the vertebrate discs) called *alpha*-motor neurons (α-mn) (see Figure 2 (Left) for the anatomy of this neuron). These neurons will be the focus of this report.

An α-mn controls a set of muscle fibers. The collection of the α-mn and the fibers it innervates is called a *motor unit*. The number of fibers per motor unit varies from 10 to 300 in different muscles. It is not known a priori how many motor units a muscles has. To complicate matters, the fibers of different motor units are not neatly bundled, but intermingle with fibers of other motor units (see Figure 2 (Right)). When an α-mn fires, it produces a 'twitch' in the muscle fibers which is the result of electric current moving across the muscle fiber membrane. This current can be measured as

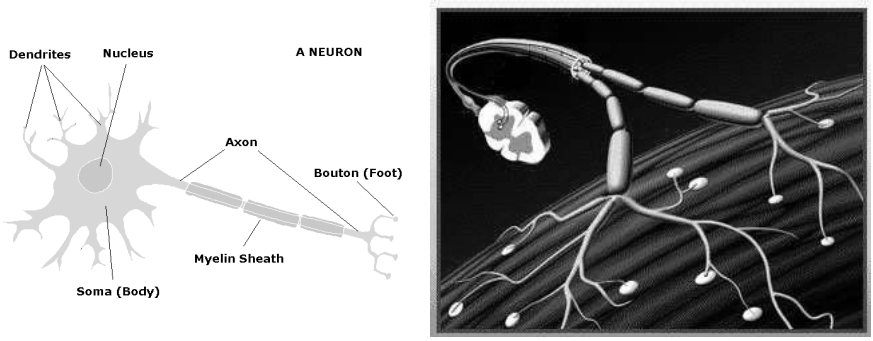


Figure 2: Left: A neuron and its components. Right: Connections between the Spinal cord, α-mns and muscle fibers.

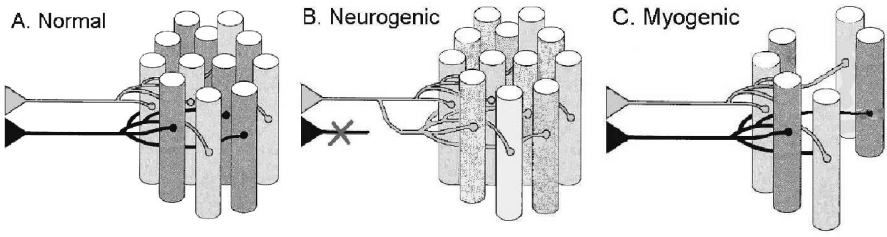


Figure 3: (A) Normal situation: Two motor units each controlled by separate α-mns. The fibers of the two motor units are intertwined. (B) Characteristics of a neurogenic disorder: reduced number of MUs and increased MU size. (C) Characteristics of a myogenic disorder: Decreased MU size. Eventually effective loss of MUs. Adapted from [3]

will be discussed later. The α-mn fires with a certain frequency resulting in repeated twitches of the fibers. The combined twitching of all fibers is what produces force and movement. Normally, the individual twitches are not perceived, only when one exerts a large amount of force can we see the effect of the underlying twitching mechanism.

We study two types of pathology which affect the distribution of fibers over the motor neurons. In the first pathology, *neurogenic disorders*, certain α-mns die off. Other neurons will take over the innervation of the 'disconnected' fibers. This causes the motor units to grow in size (ie. there are more fibers innervated by one α-mn). The increase in size of motor units results in loss of ability to perform fine tasks. In the second pathology, *myogenic disorders*, the muscle fibers die off. These fibers are not regenerated. The result of the disorder is an increasingly weaker muscle. The activity of the motor units decreases, resulting in an effective loss of motor units (see Figure 3). Diagnosis of these disorders can be done by muscle biopsy. Monitoring the progression and severity of the disorder is more difficult. Clinicians would benefit greatly from having an accurate estimate of the number of motor units in a muscle. This problem is know as Motor Unit Number Estimation (MUNE) and is the first question of medical science we will address.

Question 1: *How many motor units are active during a particular contraction?*

The MUNE problem requires information generated by the muscle fibers. This can be done by measuring the current that travels along the muscle fiber membrane. An obvious problem in neuroscience is performing measurements on humans. This problem is more pronounced when measuring force production and movement, simply because things move around. When one wants to measure signals produced by the central nervous system and muscles there are many possibilities which can be divided into two categories. The first category consists of invasive techniques. For example, the recording of electrical signals produced by neural tissue with an electrode on the tip of a needle. The second category, non-invasive measurement techniques, uses sensors on the outside of the body, mainly on the skin. For many purposes invasive techniques give more information as the sensor is exactly where the experimenter wants it to be. However, invasive measurement is significantly more distressing to the

Figure 4: High-density sEMG stetup. This setup records up to 130 channels at once.

participant of the experiment than non-invasive techniques. Also, because ethics prohibit many invasive techniques many experimenters choose non-invasive techniques to measure signals produced by the central nervous system and muscles.The data we examine was obtained non-invasively.

A non-invasive technique which records muscle activity from the surface of the skin is called surface electromyography (sEMG). A number of electrodes are placed on the skin and the voltage difference between the electrodes and a reference electrode is measured. This particular high-density setup can record up to 130 channels of sEMG at once, as described in [11] (see also Figure 4). Muscle activity can be generated in two ways: voluntary or stimulated (the participant is given a small electric shock which triggers the α-mns) which differ in a fundamental way. As mentioned before, motor units differ in size, and not all motor units are active at each level of force. The recruitment of the motor units during voluntary contractions is such that the small motor units are recruited first and the larger motor units are added as more force is required (ie. small to large). This is called the *size principle* [6], and allows one to control movement at different force levels. Also, during voluntary contractions, the motor units are not triggered in unison; there are time delays between the individual firings. This property makes analysis of the data very hard as the time differences are not known. This effect is not present when the muscle is stimulated. During stimulated contractions, the motor units fire at the same time. However, they do not obey the size principle. During stimulated contractions the motor units are recruited large to small. This is not an issue as the larger motor units will generate higher levels of sEMG and can be detected more easily.

The MUNE problem now comes down to extracting the number of components in the total of 130 channels of sEMG. The mathematics for this extraction is based on *Principal Component Analysis (PCA)*. This technique is very common in signal analysis, but will have to be modified to suit our purpose. To get reliable estimates, we need to record muscle activity with similar content (ie. the same motor units firing) many times. These recordings have delays which are unknown. To compensate,

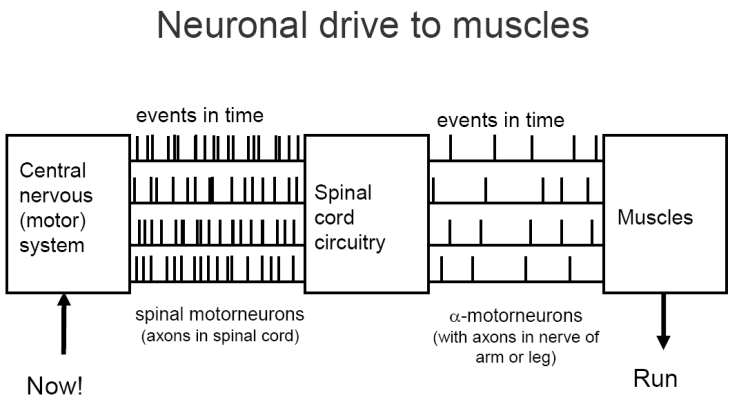


Figure 5: Neural drive from the central nervous system to the α-mn.

these delays, or shifts, will have to be estimated before we can use PCA. For the analysis we have say N recordings of 130 channels each. The high-density sEMG setup gives topological information of the muscle activity. This topological aspect has to be removed for the PCA. This is done by concatenating the information in the 130 channels of sEMG yielding N time series of which we still have to estimate the individual time delays. The N recordings are made according to a protocol which starts the first recording at a low stimulation level and continues the recordings with an increasing stimulation level. This ensures the experimenter that the largest motor units are present in almost all recordings and the smaller motor units become present in recordings with increased stimulation. The mathematics of the estimation of the shifts and the PCA are described in Section 2.

Question 2: *How is the output frequency distribution of a motor unit determined by the input frequency distribution?*

The second question we want to address concerns the response of an α-mn (see Figure 5). For this we derive a stochastic model based on certain assumptions of the nature of the signal received by the α-mn. We assume the input of an α-mn to be a Poisson process. Also, we will assume the time between action potentials arriving at the α-mn to have a Poisson(λ) distribution. It is often observed that the input to the muscle, and thus the output of the α-mn, has a characteristic frequency. Also, one can observe activity in the brain having specific frequencies. Our goal is to gain insight into the response of the α-mn to input of a known frequency.

2 Determining principle components for data with unknown shifts

Given a collection of data $X \in \mathbf{R}^{m,n}$ it is a common task to determine the principle components. Typically we model measured data as having the form

$$X_{ij} = \sum_{k=1}^{N} C_{ik} v_k(t_j) + \eta_{ij} \tag{1}$$

where $X, \eta \in \mathbf{R}^{m,n}$, $C \in \mathbf{R}^{m,N}$ and there are m times $\{t_j\}$ at which the signal was sampled. That is, the n signals can be expressed as a linear combination of a small collection of N ($N \ll n$) basis vectors and noise. In this problem we are interested in determining the dimension N of the spanning set $\{v_k\}$.

In the absence of noise this problem can be solved by decomposing the data onto it's singular values [5]; we find $U \in \mathbf{R}^{m,m}, V \in \mathbf{R}^{n,n}, \sigma \in \mathbf{R}^m$ such that

$$U^T X V = [\text{diag}(\sigma) 0_{n,m-n}]. \tag{2}$$

Additionally, U and V are orthogonal matrices and $\sigma_i \geq \sigma_{i+1} \geq 0$. Of immediate relevance is that if a matrix has rank $N < n$ then $\sigma_{N+1} = \ldots = \sigma_n = 0$. In this case there are precisely N principle components. Typically however real data is not so clearly delineated with fuzzy measurements ensuring that $\sigma_i > 0 \ \forall i$. However, the relative sizes of the principle values σ_i may still provide us with a great deal of information. In particular if we define

$$X_k = \sum_{i=1}^{k} \sigma_i U_i V_i^T$$

then σ_{k+1} is the 2-norm of the distance of X to all matrices of rank k:

$$\min_{\text{rank}(Y)=k} \|X - Y\|_2 = \|X - X_k\|_2 = \sigma_{k+1}. \tag{3}$$

Given a threshold ε this property can be used to define the ε-rank of a matrix, r_ε by requiring that

$$\sigma_{r_\varepsilon} > \varepsilon \geq \sigma_{r_\varepsilon + 1}.$$

If we have a known order of magnitude for errors ε we can define a suitable r_ε. For instance, if the entries of the matrix X_{ij} are known relatively to within $\pm 10^{-3}$ we could determine the ε-rank of the matrix by finding the minimal r_ε such that $\sigma_{r_\varepsilon+1} < 10^{-3}\|X\|_2$.

In the problem at hand we cannot directly apply these ideas as there are *two* distinct sources of errors:

1. Experimental errors of unknown type and magnitude occurring on each channel.
2. Shifts of unknown magnitude occurring *between* all pairs of channels.

Errors of the first type are not dramatically troublesome and in the next section we will discuss one algorithm for estimating the rank of a matrix in the presence of systematic noise. The second type of error needs to be examined more carefully. Denoting

$\hat{X}_{ij}$: measured data in the ith channel at time t_j,

s_i: time shift in the ith channel,

X_{ij}: data in the ith channel at time $t_j + s_i$,

η_{ij}: noise in the ith channel measured at time t_j.

Let us assume that

$$\hat{X}_{ij} = X_{ij} + \eta_{ij}$$

and that the unshifted data is spanned by N basis vectors

$$X_{ij} = \sum_{k=1}^{N} C_{ik} v_k(t_j)$$

then

$$\begin{aligned}
\hat{X}_{ij} &= \sum_{k=1}^{N} C_{ik} \left(v_k(t_j) + s_i v_k'(t_j) + \frac{s_i^2}{2} v_k''(t_j) + \dots \right) && (4)\\
&\simeq \sum_{k=1}^{N} C_{ik} \left(v_k(t_j) + s_i v_k'(t_j) \right) \qquad \text{(assuming small shifts)}\\
&= \sum_{k=1}^{N} C_{ik} v_k(t_j) + \sum_{k=1}^{N} \tilde{C}_{ik} \tilde{v}_k(t_j) && (5)
\end{aligned}$$

Generically the functions $\tilde{v}_k$ cannot be expressed as linear combinations of v_k and hence even without noise we would have doubled the rank of the matrix and the apparent dimension of the spanning set. We will know consider a small experiment to demonstrate these ideas.

Example 1 Consider four matrices of the form

$$\begin{aligned}
X_{ij} &= \alpha_i \sin(t_j) + \beta_i \sin(2t_j), \qquad i = 1, \dots, 10, t_j \in [0, 5\pi]\\
\tilde{X}_{ij} &= X_{ij} + \varepsilon \eta_{ij}\\
\hat{X}_{ij} &= \alpha_i \sin(t_j + s_i) + \beta_i \sin(2t_j + s_i)\\
&= \alpha_i (\sin(t_j)\cos(s_i) + \cos(t_j)\sin(s_i)) + \beta_i (\sin(2t_j)\cos(s_i) + \cos(2t_j)\sin(s_i))\\
&= \tilde{\alpha}_i \sin(t_j) + \tilde{\beta}_i \cos(t_j) + \tilde{\delta}_i \sin(2t_j) + \tilde{\gamma}_i \cos(2t_j)\\
\bar{X}_{ij} &= \hat{X}_{ij} + \varepsilon \eta_{ij}
\end{aligned}$$

where α_i, β_i, s_i and η_{ij} are randomly chosen in $[-1, 1]$ and $\varepsilon = 10^{-3}$. In this example we have that $N = 2$ and $n = 10$. In Figure 6 we present a sample signal with and without shifts and noise and also the singular values.

From this example we find that we cannot simply use the singular value decomposition – even with a known threshold – to estimate the dimension of the spanning set for the shifted noisy data. To determine the number of principle components for the given data we now proceed in two steps: first we estimate upper and lower bounds for N, N_u and N_l; then we search for the shifts assuming that we know the correct dimension.

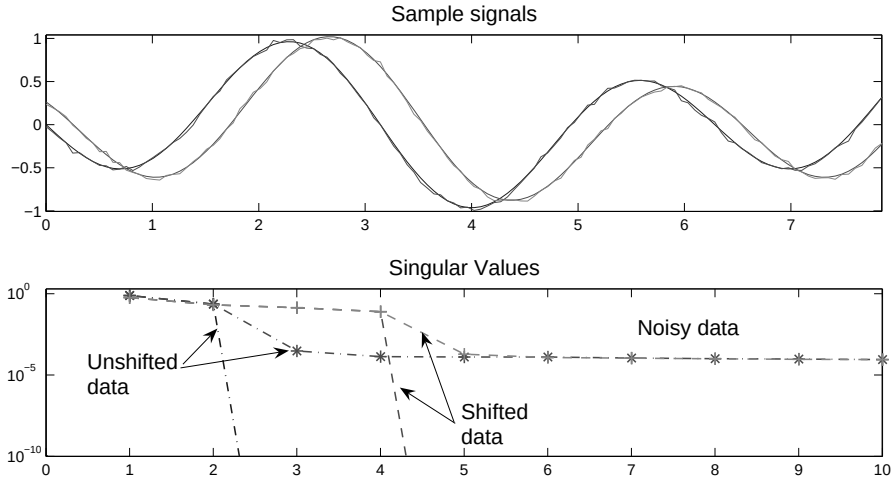


Figure 6: Example 1. Top: Sample signals from one channel showing a typical shift and small noise. Bottom: The singular values for the data with no noise fall off to $\sigma_i = 10^{-40}$ after $N = 2$ and $N = 4$ for the un-shifted and shifted data respectively. In this example the addition of the noise determines a clear ε-rank, but that the shifts double the estimate for the number of principle components.

2.1 The number of principle components

We begin by modelling our data by (1). To estimate the dimension of the spanning set we follow the philosophy of [2] and the algorithm of [8] who advocate a Bayesian approach. While the full details of the method are beyond the scope of this report, we will briefly describe the methodology.

The model is that in each channel the signal is generated from a small N-dimensional vector w via a linear transformation and an error term:

$$X_i = \sum_{k=1}^{N} C_{ik} w_k + \bar{m}_i + \eta_i \tag{6}$$

with $\langle C_i \rangle = 0$, ie. $\langle X_i \rangle = \bar{m}_i$. Critically it is assumed that both the noise vectors η_i and the principle component vectors C_i are sampled from spherical Gaussians (ie. are normally distributed). Note that here we are implicitly assuming that the noise in each channel is of the same magnitude and is thus additive. By construction, each sample X_i is also taken to be Gaussian. For given model parameters $C, \bar{m}$ and v this distribution takes the explicit form

$$P(X|C, \bar{m}, v) = \frac{1}{(2\pi)^{Nm/2}} \left|CC^T + vI\right|^{-N/2} \exp\left(-\frac{1}{2}\text{tr}(CC^T + vI)^{-1}S\right) \tag{7}$$

where S is the co-variance matrix $S_{ij} = (X_i - \bar{m}_i) \cdot (X_j - \bar{m}_j)$, and v is the variance of the noise assumed to be constant over all channels.

The evidence that the data fits the model with specific parameters can now be

determined by integrating over the parameter space

$$P(X|M) = \int_\theta P(X|\theta)P(\theta|M)\, d\theta \tag{8}$$

The model parameters which best fit the data with the highest probability are those which maximize this integral. The first difficult part is to find a parametrization of the family of matrices C in (7) such that the integral in (8) may be evaluated. In practice, it turns out that once this has been done the integral cannot be computed exactly but is instead approximated with Laplace's method [8]. Limiting the dimension of the spanning set to k we have that the evidence that our data X is of the form (6) with $C \in \mathbf{R}^{k,n}$ is [8]

$$P(X|k) = c_k \int \left|CC^T + vI\right|^{-N/2} \exp\left(-\frac{1}{2}\text{tr}(CC^T + vI)^{-1}S\right)\, dU\, dL\, dv \tag{9}$$

where $C = U(L - vI)^{1/2}R$, $U^TU = I$, $R^TR = I$ and c_k is a known function of n, k and m. The most probable dimension, $\tilde{N}$, is such that

$$P(X|\tilde{N}) = \max_k P(X|k). \tag{10}$$

The recent paper [8] compares several approximations to (9) and algorithms for determining the number of principle components. Based on multiple numerical tests the author asserts that Laplaces' approximation to (9) is both the least computationally intensive and the most reliable. An algorithm to determine the most likely dimension is to find $\tilde{N}$ satisfying (10) with $P(X|\tilde{N})$ approximated by Laplace's method (see [8] for details).

To determine the upper bound N_u we simply apply this algorithm and, based on the observation in (5), take $N_u = N/2$. Unfortunately this estimate is not sharp as the effects of real noise (ie. with different magnitudes and variances) tend to lead to an over estimation for N. Also, it is not clear which terms in the expansion will be recognized as signal rather than noise.

To compute the lower bound N_l we recall (4). We will assume that the shifts are small and hence that $C_{ik} \gg \tilde{C}_{ik}$. To find N_l we mask the effect of the shifts by amplifying the noise. We find a threshold ε such that $Y_i = X_i' + \varepsilon\eta_i$ is indistinguishable from noise. In terms of the above algorithm this means $N(\varepsilon) = 1\ \forall\delta \geq \varepsilon$. We now have a way to *artificially add noise to hide the effects of the shifts*, ie. compute

$$N_l = N(\hat{X} + \varepsilon\eta) \tag{11}$$

(In practice, this threshold may be too large. We will leave it's proper determination for later work.) One drawback of this approach is that it can *under-predict* N_l if some signals are much weaker than the others and as such can easily be drowned out by the added noise. One possible correction for this is to rescale all the signals before starting. While this rescaling does affect the particular singular values it in no way changes the shifts or the correct dimension of the spanning set.

Given the interval $[N_l, N_u]$ we now choose $N \in [N_l, N_u]$ and try to find the shifts *assuming* N.

2.2 Functional optimization

Given the geometry of the singular value decomposition we would like to choose the shifts to minimize σ_{N+1}. This maximizes the ε-rank of the matrix.

However, in practice this problem seems to be highly degenerate and difficult to solve. Instead we have chosen to consider

$$V = -\frac{\sum_{i=1}^{N} \sigma_i}{\sum_{i=1}^{n} \sigma_i} = -\sum_{i=1}^{N} \sigma_i \qquad \text{(by normalization)}. \tag{12}$$

If the singular values are decaying quickly for $i > N$ then

$$-\sum_{i=1}^{N} \sigma_i \simeq -\sum_{i=1}^{n} \sigma_i + \sigma_{N+1} \simeq -1 + \sigma_{N+1}$$

and in many cases the minimization of (12) is a good approximation of minimizing σ_{N+1}.

We have tested three approaches to minimize the functional (12) with respect to the shifts.

A1. A pseudo-Newton method to find a zero of ∇V using `minunc` in the Optimization Toolbox for Matlab.

A2. A gradient flow of V with respect to an artificial time with computation of

$$\frac{ds_i}{dt} = -\frac{\partial V}{\partial s_i}$$

using the code `ode113` in Matlab.

A3. Direct search using the Matlab routine `fminsearch`.

The results of algorithms A1 and A2 are almost identical but with A2 taking approximately ten times longer to terminate. The results of A3 are less satisfactory and take approximately ten times longer than A2.

Each step in all three algorithms requires both interpolation and the computation of singular values. Because the interpolation is being done onto a uniform grid cubic interpolation takes the form

$$X_i(t_j + s) = \alpha_{-1}(s)X_i(t_{j-1}) + \alpha_0(s)X_i(t_j) + \alpha_{+1}(s)X_i(t_{j+1}) + \alpha_{+2}(s)X_i(t_{j+2}),$$

for $s \in [0, 1]$. Here the α_i are cubic polynomials in s. This is a cheap operation taking only $\mathcal{O}(mn)$ operations.

In each algorithm, we approximate the gradient as

$$\frac{\partial V}{\partial s_i} \simeq \frac{V(s + \Delta s e_i) - V(s)}{\Delta s}$$

Here e_i is the unit vector in the $i-$th direction and $\Delta s = \sqrt{\epsilon} \simeq 10^{-8}$. Thus at each step we need to compute n singular value decompositions ($s_1 \equiv 0$) for a total

of $\mathcal{O}(mn^3)$ operations. Occasionally A1 also needs to compute the Hessian which requires $\mathcal{O}(mn^4)$ operations.

A3 is the cheapest per step but takes by far the most steps. A1 is the best as at step $n+1$

$$\left\| \frac{\partial V^{(n+1)}}{\partial s} \right\| \simeq C \left(\frac{\partial V^{(n)}}{\partial s} \right)^p, \qquad C_1 > 0, p > 1.$$

Whereas for A2 we can expect no better than

$$\left\| \frac{ds}{dt} \right\| = \left\| \frac{\partial V}{\partial s} \right\| \simeq C_2 e^{-C_3 t} \quad \text{as } t \to \infty \qquad C_2, C_3 > 0.$$

In order to reduce the number of function evaluations we use a high-order predictor corrector code but then the step sizes are limited by stability as $t \to \infty$. It is possible that hybrid method first using an explicit code and then an implicit one would make A2 competitive with A1.

2.3 A test for determining N

The minimization routine is run for the sequence of $N \in [N_l, N_u]$. Unfortunately there are still numerous problems preventing us from simply applying the evidence algorithm to determine the number of principle components. These include

1. The noise in the data does not strictly follow the model: it is not necessarily normally distributed, the variance varies between channels.
2. The shifts have not been completely eliminated.
3. There are small interpolation errors.
4. The minimization problem introduces additional numerical errors.
5. A mixture of signals of vastly different magnitudes in each channel. This makes some signals difficult to distinguish from noise and, more importantly, the effects of shifts in large signals can be on the same order of magnitude as some weak signals.

Ideally, we would like a nice obvious step from large singular values to small ones (as in Example 1) allowing for a clear ε-rank. Unfortunately, this combination of difficulties makes determining such an ε impossible. Instead, we solve for all $N \in [N_l, N_u]$ and then set

$$N_r \text{ such that } \frac{\sigma_{N_r}}{\sigma_{N_r+1}} = \max_k \frac{\sigma_k}{\sigma_{k+1}} \tag{13}$$

for each run. Then we take $\tilde{N}$ to be that which maximizes this ratio. Conceptually this approach finds the dimension such that there is a clear step from "large" components to "small" ones. But, it does not preclude the possibility that the data in each channel comprises one very large signal and several much smaller ones. In that case, for the reasons stated above, it may be very difficult to distinguish the small signals from noise as the shifts amplify the classic signal vs. noise problem.

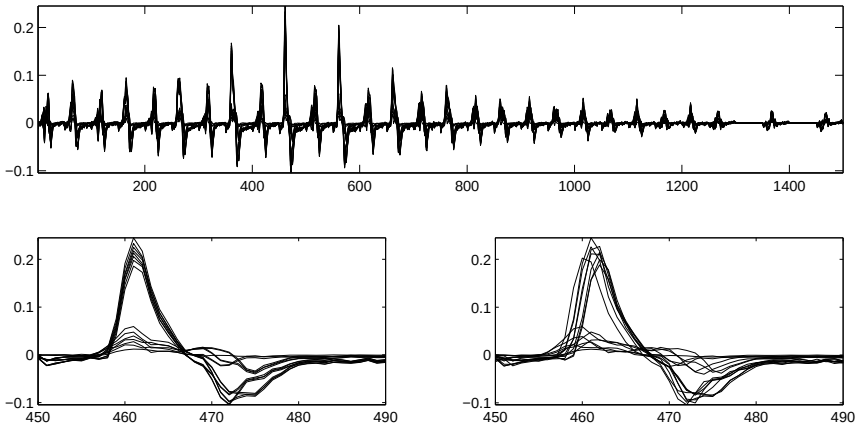


Figure 7: Artificial data. Top: Complete signals in all channels. Bottom left: Detail of unshifted data. Bottom right: Detail of shifted data. After optimization the original and output data are essentially indistinguishable.

N	N_r	$\sigma_{N_r}/\sigma_{N_r+1}$	Run time (s)
6	6	125	47
5	5	98	45
4	4	2731	33
3	4	113	39
2	4	31	31

Table 1: Results from running the algorithm with given data. From this we would correctly conclude that there are four principle components.

2.4 Example

To recap, the entire algorithm is

1. Determine N_u with (10).

2. Numerically find X' and find ε such that $X' + \varepsilon\eta$ is indistinguishable from pure noise.

3. Determine the N_l as the dimensionality of $X + \varepsilon\eta$.

4. Solve the optimization problem for $N = N_l, N_l + 1, \ldots N_u$.

We now consider a test on the artificially generated data as in Figure 7. This is "simulated" from one of our industrial collaborators. We begin with initially unshifted data to allow us to test the reliability of the algorithm. The results are summarized in Table 1.

2.5 Conclusions

This algorithm has been shown to be effective in reducing the effects of shifts in data to give a reliable estimate of the number of principle components. However, two areas for improvement remain. Firstly, no information about the structure of the signals has been incorporated and we leave this as an avenue for further work. Secondly, when optimizing with N greater than the number of true principle components the signals falsely seperate so as to fill all available spanning dimensions. This problem is easily remedied should it be required when dealing with real data.

The question we have looked at in this Section has two components. Firstly, how can we remove the shifts? Then, how do we determine the number of principle components? We believe that we can more reliably deal with the first question. This is not surprising as the second is a long-standing research question and not completely resolved in many practical application areas.

After the preparation of this report we were furnished with "real" data with which to work. Preliminary tests suggests that this approach works but that the ODE gradient flow approach may be more appropriate. We leave this is a direction for further research.

3 Discrete integrate and fire neurons

The second question posed was to model the frequency response of a motor unit. Clinical scientists are interested in an explanation of the 40 Hz components that are measured in sEMG, both at the brain and the muscles. We did not have sufficient information to answer this specific question, but have made a simple model that combines features of other, existing models, and may help to foster insight in the frequency response.

We present a simple, phenomenological, discrete model with which the average firing time or the length of the inter-spike intervals of neurons can be estimated, given a basic exponential potential function $F_0(t)$ and a series of incoming signals $g_i(t)$ from the brain. We choose a particular $F_0(t)$, but the discrete approach in this section can be applied to any function $F_0(t)$. Our potential $F_0(t)$ varies between its minimum F_0 and its saturation state $F_0 + \alpha$, and is assumed to be exponential:

$$F_0(t) = F_0 + \alpha(1 - e^{-\beta t}) \tag{14}$$

with $\alpha, \beta > 0$. The potential is then reset to the minimum F_0 after the neuron has fired. During an inter-spike interval, ie. as long as no firing has occurred, the potential increases monotonically.

The neuron fires if its total potential $F(t)$, that is a sum

$$F(t) = F_0(t) + \sum_i g_i(t) \tag{15}$$

of its own basic potential $F_0(t)$ and the incoming signals $g_i(t)$, reaches a certain threshold value θ.

The incoming signals $g_i(t)$ are in fact very short pulses, but they are often modeled as exponentials, as block-functions, or, since they should add up to build up the

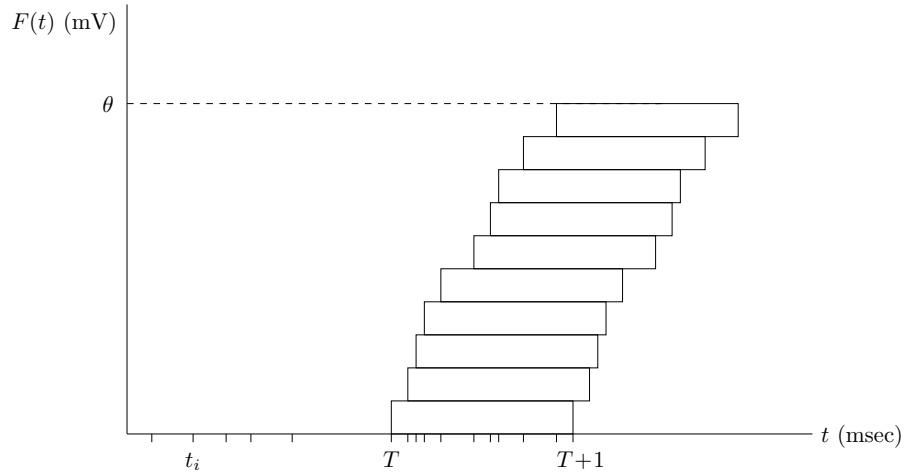


Figure 8: Blocks of length 1 arrive at arrival times t_i. Here 10 blocks have arrived within the time interval $[T, T+1]$.

potential $F(t)$, their derivatives are assumed to be delta-functions [4]. If the incoming signals $g_i(t)$ are modeled by $\frac{dg}{dt} = \delta(t - t_i)$, so as block functions that have value $g_i(t) = 0$ for all $t < t_i$ and $g_i(t) = 1$ for all $t > t_i$, they get an infinite length. There is a major drawback of this simplest assumption one can think of, since this way they all add up to build up the potential, whereas in real neurons the potential decreases as well. We therefore want our model to allow for a drop in the potential $F(t)$ if the incoming signals $g(t_i)$ are too far apart. Our approach is, to model the $g_i(t)$ by block functions, all of equal length and height. This form is of course still far from the real pulse-form of the signals, but is similar to the approximation mentioned above. If a stack of these blocks reaches the threshold θ within a certain time interval (that we scale to 1), the neuron will fire. If the threshold is not reached within this time interval, the total stack will decrease, and new incoming signals will start and build up a new stack (see Figures 8 and 9).

In fact, this model combines two features of the neuron-system that have been implemented separately before. As far as we know there are no existing models in which both are combined. Other models either use a constant function $F_0(t)$, which means in physiological terminology that no leaky dynamics are taken into account, or they use a function $F_0(t)$ with leaky part and use infinite-length $g_i(t)$-functions. If a constant $F_0(t)$ is chosen, then it is combined with the same infinite-length $g_i(t)$-functions or with more natural forms for $g_i(t)$, such as the above mentioned exponentials or simple pulse functions.

3.1 Stochastics

A standard assumption on the incoming signals $g_i(t)$ is that the arrival times t_i are distributed as a Poisson process. In other words, the time intervals $X_i = t_i - t_{i-1}$ are

random variables that are Poisson distributed, ie. they satisfy

$$P(X_i > x) = e^{-\lambda x} \tag{16}$$

for a positive parameter λ. For well-known results, see for instance [10].

If the interval lengths X_i satisfy the distribution (16), then the probability that there will be n arrival times t_i within any time interval of length 1 is equal to the the probability that there will be n arrival times t_i within the interval $[0, 1]$. With N the number of arrival times within $[0, 1]$, this probability is $P(N = n) = \frac{\lambda^n}{n!}e^{-\lambda}$. If the incoming blocks have length 1 and height 1, the probability that they will together at least reach an integer height h within a time interval of length 1 will thus be $P(N \geq h) = 1 - P(N \leq h - 1)$. The expected number of arrival times (expectation value) within an interval of length 1 is λ, with variance λ.

This is illustrated in Figure 8. There 10 blocks of length 1 and equal height have arrived within the time interval $[T, T + 1]$, and the threshold $\theta = 10$ is thus reached. The probability that exactly 10 blocks would have arrived within this interval is $P(N = 10) = \frac{\lambda^{10}}{10!}e^{-\lambda}$; the probability that at least 10 blocks would have arrived within this interval is $P(N \geq 10) = \sum_{k\geq 10} \frac{\lambda^k}{k!}e^{-\lambda}$.

The next step is to let the stack fall down and start a new one, if the stack has not reached the threshold height θ within the time interval $[T, T + 1]$. In modeling the function $F(t)$ (15), we impose that the level at which the stack in an interval $[T, T+1]$ starts to build up equals $F_0(T)$, with $F_0(t)$ given by (14) as shown in Figure 9. This means that if the stack does not reach the threshold θ within $[T, T + 1]$, it will fall down to $F_0(T + 1)$. If the stack has reached the threshold value within an interval $[T, T + 1]$, the neuron fires, the stack falls down, and the time t is reset to $t = 0$, so that the potential starts to build up from $F_0(0)$ again. To simplify the calculations, we let the neuron fire at the end of the interval, at $t = T + 1$.

3.2 Expectation value and variance of firing time

The above ingredients are sufficient to calculate the expectation value and variance of the firing time in this model. We take blocks of height 1 (notice that the scaling of the horizontal and vertical axes in Figures 8 and 9 is different).

Define p_n as the probability that the neuron fires in (or by the above assumption, at the end of) the interval $[n, n + 1]$. Then

$$p_n = P(N \geq \theta - F(n)) = P(N \geq \lceil \theta - F(n) \rceil) = \sum_{k\geq \lceil \theta - F(n) \rceil} \frac{\lambda^k}{k!}e^{-\lambda} = 1 - \sum_{k=0}^{\lfloor \theta - F(n) \rfloor} \frac{\lambda^k}{k!}e^{-\lambda}. \tag{17}$$

Here we used the notations $\lceil x \rceil$, which means the first integer larger than x, and $\lfloor x \rfloor$, which means the first integer smaller than x. They play a role, since $\theta - F(n)$ need not be an integer.

Now define the stochast Y as the firing time. Define f_n as the probability that the neuron has not fired at or before $t = n$, and does fire in the interval $[n, n + 1]$, or

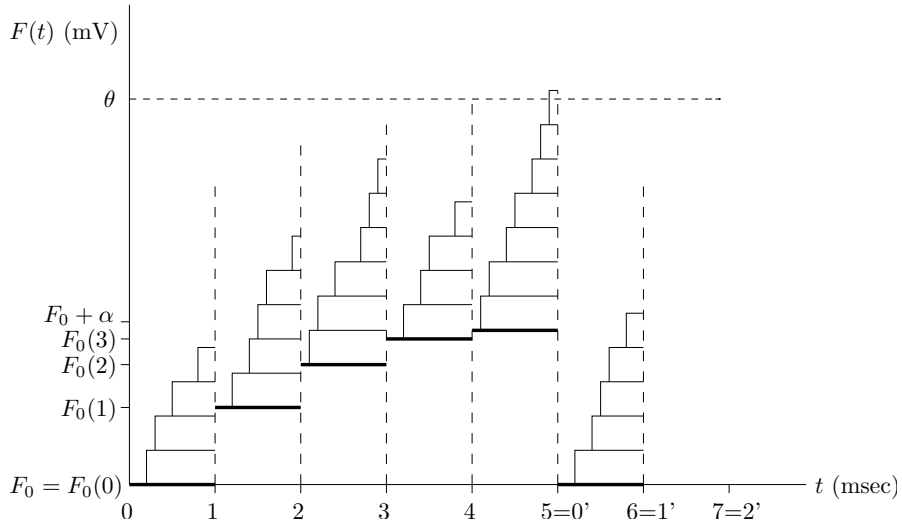


Figure 9: Stacks are being built up within intervals $[T, T+1]$. If a stack has not reached θ at $t = T+1$ it falls down at $t = T+1$, and the next block arrives at height $F_0(T+1)$. If the stack has θ at $t = T+1$, the neuron fires (not shown), the stack falls down, the time t is reset ($t = 5 = 0'$), and the next block arrives at height $F_0(0') = F_0(0)$ again.

better, at $t = n+1$. Then

$$f_n = P(Y = n+1) = (1-p_0)\dots(1-p_{n-1})p_n = p_n \prod_{j=0}^{n-1}(1-p_j). \qquad (18)$$

The expectation value for the firing time can now be calculated as

$$\begin{aligned} E(Y) &= 1.P(Y=1) + 2.P(Y=2) + 3.P(Y=3) + \dots \\ &= \sum_{l=1}^{\infty} f_{l-1} l = \sum_{l=1}^{\infty} p_{l-1} l \prod_{j=0}^{l-2}(1-p_j). \end{aligned}$$

With $e_n := \lceil \theta - F(n) \rceil$, this equals

$$E(Y) = \sum_{l=1}^{\infty} l \sum_{k \geq e_{l-1}} \frac{\lambda^k}{k!} e^{-\lambda} \prod_{j=0}^{l-2} (\sum_{k=1}^{e_{j-1}} \frac{\lambda^k}{k!} e^{-\lambda}). \qquad (19)$$

The variance in the expected firing time is $Var(Y) = E(Y^2) - E(Y)^2$, where $E(Y^2)$ is the so-called second moment

$$\begin{aligned} E(Y^2) &= 1^2.P(Y=1) + 2^2.P(Y=2) + 3^2.P(Y=3) + \dots \\ &= \sum_{l=1}^{\infty} f_{l-1} l^2 = \sum_{l=1}^{\infty} p_{l-1} l^2 \prod_{j=0}^{l-2}(1-p_j), \end{aligned}$$

which equals

$$E(Y^2) = \sum_{l=1}^{\infty} l^2 \sum_{k \geq e_{l-1}} \frac{\lambda^k}{k!} e^{-\lambda} \prod_{j=0}^{l-2} (\sum_{k=1}^{e_{j-1}} \frac{\lambda^k}{k!} e^{-\lambda}). \tag{20}$$

The n^{th} moment is likewise given by

$$E(Y^n) = \sum_{l=1}^{\infty} l^n \sum_{k \geq e_{l-1}} \frac{\lambda^k}{k!} e^{-\lambda} \prod_{j=0}^{l-2} (\sum_{k=1}^{e_{j-1}} \frac{\lambda^k}{k!} e^{-\lambda}). \tag{21}$$

3.3 Calculations

If the parameters F_0, α, θ and λ are chosen, the expected firing time and corresponding variance can be calculated from (19) and (20). The first three parameters vary for different types of neurons, and we will choose various values in our calculations. The parameter λ however is strongly related to the time-separation between the incoming pulses $g_i(t)$, since λ is the expected number of arrival times within an interval of length 1. We know that the pulses arrive with a frequency of about 10.000 Hz, so if we rescale our time t so that each interval of length 1 is an interval of 1 msec., the parameter λ should be set to $\lambda = 10$.

We calculated the expected firing time and variance for this model, with different (more or less realistic) values for F_0, α, β and θ. In [7] we find typical values for an α-motor neuron: the minimum F_0 of $F_0(t)$ is about $F_0 = -75$, the difference α between the minumum of $F(t)$ and its saturation value is about $\alpha = 5$, and the firing threshold θ is about $\theta = -55$. These values are all given in mV. In the standard book [1] we find values for other types of motor units as well; examples are $F_0 = -65$, $\alpha = 10$ or 15, and $\theta = -50$ or -45 (but of course always with $\alpha < \theta - F_0$). The parameter β is related to the saturation time of $F_0(t)$, which is typically 75 [7] to 100 msec. We assume that $F_0(t)$ has reached the saturation value $F_0 + \alpha$ if it is less than 1% off this value, so if $1 - e^{-\beta t} \leq 0.01$. This means that realistic values for β are for instance $\beta = 0.05$ or $\beta = 0.06$. The results of some calculations are listed in the Table 2.

To interpret these values, it is useful to realize that an expected firing time $E(Y) = \tau$ msec. corresponds to an expected firing frequency of $1000/\tau$ Hz. The last column in the table depicts the corresponding expected frequency (in Hz) in each of the calculated cases.

As realistic firing frequencies are between 8 and 20 Hz (15 Hz for an α-motor neuron) [1, 9], we conclude that for low threshold values and rather short saturation time ($\beta = 0.1$, *i.e.*, a saturation time of about 46 msec. according to the 1%-rule above) the model predicts firing frequencies that are far too high. However, for higher threshold values and a somewhat longer saturation time ($\beta = 0.06$, 77 msec.; or $\beta = 0.05$, 92 msec.), the predicted values for $\alpha = 5$ are very reasonable. For $\alpha = 10$, the predicted frequencies are (much) higher than for $\alpha = 5$.

Note, that for the calculations only the difference $\theta - F_0$ matters, and not the separate values F_0 and θ. This explains why the first and second set contain similar rows. We only computed the first and second moments $E(Y)$ and $E(Y^2)$ for every

F_0	λ	α	β	θ	$E(Y)$	$E(Y^2)$	$Var(Y)$	$\sigma(Y)$	E(freq)
-65	10	5	0.1	-50	9.06	108.7	26.7	5.16	110
-65	10	5	0.1	-45	48.33	3693	1357	36.8	21
-65	10	5	0.06	-45	54.4	4374	1410	37.55	18
-65	10	5	0.05	-45	55.7	4504	1404	37.47	18
-65	10	10	0.1	-45	15.4	267.5	31.9	5.65	65
-65	10	10	0.06	-45	20.6	482.6	57.6	7.59	49
-65	10	10	0.05	-45	23.2	609.1	71.5	8.46	43
-75	10	5	0.1	-65	2.28	7.51	2.30	1.52	438
-75	10	5	0.1	-60	9.06	108.7	26.7	5.16	110
-75	10	5	0.06	-60	10.6	153.5	41.7	6.46	94
-75	10	5	0.1	-55	48.33	3693	1357	36.8	21
-75	10	5	0.063	-55	53.8	4297	1404	37.47	19
-75	10	5	0.06	-55	54.4	4374	1410	37.55	18

Table 2: Results of the calculations for various values of the parameters.

choice of the parameters, but if one is interested, higher moments can be computed along the same lines, using formula (21).

3.4 Conclusions

Although it is very simple and only of a phenomenological nature, the discrete model we presented is capable of producing realistic values for the expected firing time and frequency.

By its discrete nature, the model can easily be changed without destroying the global, simple framework. If one wishes, it can for instance be adapted to fit better with clinical data or to mimic other models, by changing the basic potential $F_0(t)$, the length of the "building-up period" or the value of λ.

These are two reasons why the model may serve as an additional tool to study the relation between input and output frequencies in motor units.

References

[1] M.F. Bear, B.W. Connors and M.A. Paradiso, *Neuroscience, exploring the brain*, 2nd edition, Lippincott, Williams, Wilkins (publ.)

[2] C. Bishop, *Bayesian PCA*, Neural Information Processing Systems, vol 11, 1998, pp 382-388.

[3] J.H. Blok, J.P. van Dijk, G. Drost, M.J. Zwarts and D.F Stegeman, *A high-density multichannel surface electromyography system for the characterisation of single motor units*, Review of Scientific Instruments, vol 73 (4), 2002, pp 1887-1897.

[4] A.N. Burkitt and F.M. Clark, *Neural Computation* vol 12, 2000, pp 1789-1820.

[5] G. Golub and C. van Loan. *Matrix Computations*, 3rd edition. Johns Hopkins University Press. Baltimore, 1996.

[6] E. Henneman, *Relation between size of neurons and their susceptibility do discharge*, Science vol. 126, 1957, pp 1345-1347.

[7] P. Matthews, *Relationship of firing intervals of human motor units to the trajectory of post-spike after-hyperpolarization and synaptic noise*, Journal of Physiology vol 15, 1996, pp 597-628.

[8] T.P. Minka, *Automatic choice of dimensionality for PCA*, Neural Information Processing Systems, 2002.

[9] L.J. Myers, M. Lowery, M. O'Malley, C.L. Vaughan, C. Heneghan, A. St Clair Gibson, Y.X.R. Harley and R. Sreenivasan, *Rectification and non-linear preprocessing of EMG signals for cortico-muscular analysis*, Journal of Neuroscience Methods vol 124, 2003, pp 157-165.

[10] J.R. Norris, *Markov Chains*. Cambridge series in Statistical and Probabilistic Mathematics, 1998.

[11] M.J. Zwarts and D.F. Stegeman, *Multichannel surface EMG: Basic aspects and clinical utility*, Muscle & Nerve vol 28, 2003, pp 1-17.

Errata

Page	Text	Must be
8	"Tijdens cardiopulminaire bypass"	"Tijdens een cardiopulminaire bypass"
12	"hogere kerosine kosten"	"hogere kerosinekosten"
12	"het service niveau"	"het serviceniveau"
14	"is een interpolatie formule"	"is een interpolatieformule"
15	"van een kansverdeling, voor het totale"	"van een kansverdeling voor het totale"
16	onderschrift bij figuur 2.1 eindigen met een punt	
17	"toepassen van theoretische resultaten toepassen op"	"toepassen van theoretische resultaten op"
18	"model bouwen voor in de inleiding"	"model bouwen voor het in de inleiding"
20	"wordt de twee roestvrij stalen"	"wordt de tweede roestvrij stalen";
21	"Massa metrologie"	"Massametrologie"
21	"te garanderen worden de paren"	"te garanderen, worden de paren"
21	"kleinste kwadraten analyse"	"kleinste-kwadratenanalyse"
21	"STS procedure" "STS metingen"	"STS-procedure" "STS-metingen"
24	"STS serie"	"STS-serie"
24	tweemaal: "massa metrologie"	"massametrologie"
25	"standaard gewichten"	"standaardgewichten"
26	"het softwaresystemen"	waarschijnlijk "het softwaresysteem" of anders "de softwaresystemen"
33	"topassen"	"toepassen"
49	"Maybe also include explanation of Mark here..."	Should have been omitted
55	"different approach"	"different approaches"
58	"These probability"	"These probabilities"
65	"mean and variance of water usage has been"	"mean and variance of water usage have been"
65	"waterusage"	"water usage"
75	"the uspect"	"the suspect"
75	"Hence, The"	"Hence, the"
76	"orestimating"	"or estimating"
76	"shouldbe"	"should be"
77	"Further, we write"	"Furthermore, we write"
92	"weight one the scale"	"weight on the scale"
95 bottom	`rudi@win.tue.nl`	`{rudi,averhoeven,mroeger}@win.tue.nl`
98	"ILP"	"ILP problem (3 times)"
105	The colours red and blue are not visible, because this syllabus is in b-w. See `http://www.cwi.nl/publications/Abstracts_syll/syllabus55.pdf` for the coloured page 105.	
110	"a muscles has."	"a muscle has."
112	"muscles.The data"	"muscles. The data"
112, 117, 121	"ie."	"i.e.,"

Page	Text	Must be
120	"This is "simulated" from one of our industrial collaborators."	"This is simulated data from one of our industrial collaborators."
121	"We leave this is a direction"	"We leave this as a direction"
122	"pulse-form"	"pulse form"
	"neuron-system"	"neuron system"
125	"msec."	"ms" (five times)

CWI SYLLABI

1 Vakantiecursus 1984: *Hewet - plus wiskunde.* 1984.
2 E.M. de Jager, H.G.J. Pijls (eds.). *Proceedings Seminar 1981–1982. Mathematical structures in field theories.* 1984.
3 W.C.M. Kallenberg, et al. *Testing statistical hypotheses: worked solutions.* 1984.
4 J.G. Verwer (ed.). *Colloquium topics in applied numerical analysis, volume 1.* 1984.
5 J.G. Verwer (ed.). *Colloquium topics in applied numerical analysis, volume 2.* 1984.
6 P.J.M. Bongaarts, J.N. Buur, E.A. de Kerf, R. Martini, H.G.J. Pijls, J.W. de Roever. *Proceedings Seminar 1982–1983. Mathematical structures in field theories.* 1985.
7 Vacantiecurus 1985: *Variatierekening.* 1985.
8 G.M. Tuynman. *Proceedings Seminar 1983–1985. Mathematical structures in field theories, Vol.1 Geometric quantization.* 1985.
9 J. van Leeuwen, J.K. Lenstra (eds.). *Parallel computers and computations.* 1985.
10 Vakantiecursus 1986: *Matrices.* 1986.
11 P.W.H. Lemmens. *Discrete wiskunde: tellen, grafen, spelen en codes.* 1986.
12 J. van de Lune. *An introduction to Tauberian theory: from Tauber to Wiener.* 1986.
13 G.M. Tuynman, M.J. Bergvelt, A.P.E. ten Kroode. *Proceedings Seminar 1983–1985. Mathematical structures in field theories, Vol.2.* 1987.
14 Vakantiecursus 1987: *De personal computer en de wiskunde op school.* 1987.
15 Vakantiecursus 1983: *Complexe getallen.* 1987.
16 P.J.M. Bongaarts, E.A. de Kerf, P.H.M. Kersten. *Proceedings Seminar 1984–1986. Mathematical structures in field theories, Vol.1.* 1988.
17 F. den Hollander, H. Maassen (eds.). *Mark Kac seminar on probability and physics. Syllabus 1985–1987.* 1988.
18 Vakantiecursus 1988. *Differentierekening.* 1988.
19 R. de Bruin, C.G. van der Laan, J. Luyten, H.F. Vogt. *Publiceren met LATEX.* 1988.
20 R. van der Horst, R.D. Gill (eds.). *STATAL: statistical procedures in Algol 60, part 1.* 1988.
21 R. van der Horst, R.D. Gill (eds.). *STATAL: statistical procedures in Algol 60, part 2.* 1988.
22 R. van der Horst, R.D. Gill (eds.). *STATAL: statistical procedures in Algol 60, part 3.* 1988.
23 J. van Mill, G.Y. Nieuwland (eds.). *Proceedings van het symposium wiskunde en de computer.* 1989.
24 P.W.H. Lemmens (red.). *Bewijzen in de wiskunde.* 1989.
25 Vakantiecursus 1989: *Wiskunde in de Gouden Eeuw.* 1989.
26 G.G.A. Bäuerle et al. *Proceedings Seminar 1986–1987. Mathematical structures in field theories.* 1990.
27 Vakantiecursus 1990: *Getallentheorie en haar toepassingen.* 1990.
28 Vakantiecursus 1991: *Meetkundige structuren.* 1991.
29 A.G. van Asch, F. van der Blij. *Hoeken en hun Maat.* 1992.
30 M.J. Bergvelt, A.P.E. ten Kroode. *Proceedings seminar 1986–1987. Lectures on Kac-Moody algebras.* 1992.
31 Vakantiecursus 1992: *Systeemtheorie.* 1992.
32 F. den Hollander, H. Maassen (eds.). *Mark Kac seminar on probability and physics. Syllabus 1987–1992.* 1992.
33 P.W.H. Lemmens (ed.). *Meetkunde van kunst tot kunde, vroeger en nu.* 1993.
34 J.H. Kruizinga. *Toegepaste wiskunde op een PC.* 1992.
35 Vakantiecursus 1993: *Het reële getal.* 1993.
36 Vakantiecursus 1994: *Computeralgebra.* 1994.
37 G. Alberts. *Wiskunde en praktijk in historisch perspectief. Syllabus.* 1994.
38 G. Alberts, J. Schut (eds.). *Wiskunde en praktijk in historisch perspectief. Reader.* 1994.
39 E.A. de Kerf, H.G.J. Pijls (eds.). *Proceedings Seminar 1989–1990. Mathematical structures in field theory.* 1996.
40 Vakantiecursus 1995: *Kegelsneden en kwadratische vormen.* 1995.
41 Vakantiecursus 1996: *Chaos.* 1996.
42 H.C. Doets. *Wijzer in Wiskunde.* 1996.
43 Vakantiecursus 1997: *Rekenen op het Toeval.* 1997.
44 Vakantiecursus 1998: *Meetkunde, Oud en Nieuw.* 1998.
45 Vakantiecursus 1999: *Onbewezen Vermoedens.* 1999.
46 P.W. Hemker, B.W. van de Fliert (eds.). *Proceedings of the 33^{rd} European Study Group with Industry.* 1999.
47 K.O. Dzhaparidze. *Introduction to Option Pricing in a Securities Market.* 2000.
48 Vakantiecursus 2000: *Is wiskunde nog wel mensenwerk?* 2000.
49 Vakantiecursus 2001: *Experimentele wiskunde.* 2001.
50 Vakantiecursus 2002: *Wiskunde en gezondheid.* 2002.
51 G.M. Hek (ed.). *Proceedings of the 42^{nd} European Study Group with Industry.* 2002.
52 Vakantiecursus 2003: *Wiskunde in het dagelijks leven.* 2003.
53 Vakantiecursus 2004: *Structuur in schoonheid.* 2004.
54 Vakantiecursus 2005: *De schijf van vijf – meetkunde, algebra, analyse, discrete wiskunde, stochastiek.* 2005.
55 J. Hulshof (ed.). *Proceedings of the 52^{nd} European Study Group with Industry.* 2006.

MC SYLLABI

1.1 F. Göbel, J. van de Lune. Leergang besliskunde, deel l: wiskundige basiskennis. 1965.
1.2 J. Hemelrijk, J. Kriens. Leergang besliskunde, deel 2: kansberekening. 1965.
1.3 J. Hemelrijk, J. Kriens. Leergang besliskunde, deel 3: statistiek. 1966.
1.4 G. de Leve, W. Molenaar. Leergang besliskunde, deel 4: Markovketens en wachttijden. 1966
1.5 J. Kriens, G. de Leve. Leergang besliskunde, deel 5: inleiding tot de mathematische besliskunde. 1966.
1.6a B. Dorhout, J. Kriens. Leergang besliskunde, deel 6a: wiskundige programmering. 1967.
1.6b B. Dorhout, J. Kriens, J.Th. van Lieshout. Leergang besliskunde deel 6b: wiskundige programmering. 1967
1.7a G. de Leve. Leergang besliskunde, deel 7a: dynamische programmering 1. 1969
1.7b G. de Leve, H.C. Tijms. Leergang besliskunde, deel 7b: dynamische programmering 2. 1970.
1.7c G. de Leve, H.C. Tijms. Leergang besliskunde deel 7e: dynamische programmering 3. 1971.
1.8 J. Kriens, F. Göbel, W. Molenaar. Leergang besliskunde, deel 8: minimaxmethode, netwerkplanning, simulatie. 1968.
2.1 G.J.R. Förch, P.J. van der Houwen, RP. van de Riet. Colloquium stabiliteit van differentieschema's deel 1. 1967.
2.2 L. Dekker, T.J. Dekker, P.J. van der Houwen, M.N. Spijker. Colloquium stabiliteit van differentieschema's deel 2. 1968.
3.1 H.A. Lauwerier. Randwaardeproblemen, deel 1. 1967.
3.2 H.A. Lauwerier. Randwaardeproblemen, deel 2. 1968.
3.3 H.A. Lauwerier. Randwaardeproblemen, deel 3. 1968.
4 H.A. Lauwerier. Representaties van groepen. 1968.
5 J.H. van Lint, J.J. Seidel, P.C. Baayen. Colloquium discrete wiskunde. 1968.
6 K.K. Koksma. Cursus ALGOL 60. 1969.
7.1 Colloquium moderne rekenmachines, deel 1. 1969.
7.2 Colloquium moderne rekenmachines, deel 2. 1969.
8 H. Bavinck, J. Grasman. Relaxatietrillingen. 1969.
9.1 T.M.T. Coolen, G.J.R. Förch, E.M. de Jager, H.G.J. Pijls. Colloquium elliptische differentiaalvergelijkingen, deel 1. 1970.
9.2 W.P. van den Brink, T.M.T. Coolen, B. Dijkhuis, P.P.N. de Groen, P.J. van der Houwen, EM. de Jager, N.M. Temme, R.J. de Vogelaere. Colloquium elliptische differentiaalvergelijkingen, deel 2. 1970.
10 l. J. Fabius, W.R. van Zwet. Grondbegrippen van de waarschijnlijkheidsrekening. 1970.
11 H. Bart, M.A. Kaashoek, H.G.J. Pijls, W.J. de Schipper, J. de Vries. Colloquium halfalgebra's en positieve operatoren. 1971.
12 T.J. Dekker. Numerieke algebra. 1971.
13 F.E.J. Kruseman Aretz. Programmeren voor rekenautomaten; de MC ALGOL 60 vertaler voor de EL X8. 1971.
14 H. Bavinck, W. Gautschi, G.M. Willems. Colloquium approximatietheorie. 1971.
15.1 T.J. Dekker, P. W. Hemker, P.J. van der Houwen. Colloquium stijve differentiaalvergelijkingen, deel 1. 1972.
15.2 P.A. Beentjes, K. Dekker, H.C. Hemker, S.P.N. van Kampen, G.M. Willems. Colloquium stijve differentiaalvergelijkingen, deel 2. 1973.
15.3 P.A. Beentjes, K. Dekker, P.W. Hemker, M. van Veldhuizen. Colloquium stijve differentiaalvergelijkingen, deel 3. 1975.
16.1 L. Geurts. Cursus programmeren, deel 1: de elementen van het programmeren. 1973.
16.2 L. Geurts. Cursus programmeren, deel 2: de programmeertaal ALGOL 60. 1973.
17.1 P.S. Stobbe. Lineaire algebra, deel 1. 1973.
17.2 P.S. Stobbe. Lineaire algebra, deel 2. 1973.
17.3 N.M. Temme. Lineaire algebra, deel 3. 1976.
18 F. van der Blij, H. Freudenthal, J.J. de Iongh, J.J. Seidel, A. van Wijngaarden. Een kwart eeuw wiskunde 1946-1971, syllabus van de vakantiecursus 1971. 1973.
19 A. Hordijk, R. Potharst, J.Th. Runnenburg. Optimaal stoppen van Markovketens. 1973.
20 T.M.T. Coolen, P.W. Hemker, P.J. van der Houwen, E Slagt. ALGOL 60 procedures voor begin- en randwaardeproblemen. 1976.
21 J.W. de Bakker (red.). Colloquium programma-correctheid. 1975.
22 R. Helmers, J. Oosterhoff, F.H. Ruymgaart, M.C.A. van Zuylen. Asymptotische methoden in de toe-tsingstheorie; toepassingen van naburigheid. 1976.
23.1 J.W. de Roever (red.). Colloquium onderwerpen uit de biomathematica, deel l. 1976.
23.2 J.W. de Roever (red.). Colloquium onderwerpen uit de biomathematica. deel 2. 1977.
24.1 P..J. van der Houwen. Numerieke integratie van differentiaalvergelijkingen . deel 1: eenstapsmethaden. 1974.
25 Colloquium structuur van programmeertalen. 1976.
26.1 N.M. Temme (ed.). Nonlinear analysis, volume 1. 1976.
26.2 N.M. Temme (ed.). Nonlinear analysis, volume 2. 1976.
27. M. Bakker, P.W. Hemker, P..J. van der Houwen, SJ. Polak, M, van Veldhuizen. Colloquium discretiseringsmethoden. 1976.
28 O. Diekmann, N.M. Temme (eds.). Nonlinear diffusion problems. 1976.
29.1 J.C.P. Bus (red.). Colloquium numerieke programmatuur, deel 1A, deel 1 B. 1976.
29.2 H.J.J. te Riele (red.). Colloquium numerieke programmatuur, deel 2. 1977.
30 J. Heering, P. Klint (red.). Colloquium programmeeromgevingen. 1983.
31 J.H. van Lint (red.). Inleiding in de coderingstheorie. 1976.
32 L. Geurts (red.). Colloquium bedrijfsystemen. 1976.
33 P..J. van der Houwen. Berekening van waerstanden in zeeën en rivieren. 1977.
34 J. Hemelrijk. Oriënterende cursus mathematische statistiek. 1977.
35 P.J.W. ten Hagen (red.). Colloquium, computer graphics. 1978.
36 J.M. Aarts, J. de Vries. Colloquium topologische dynamische systemen. 1977.
37 J.C. van Vliet (red.). Colloquium capita datastructuren. 1978.
38.1 T.H. Koornwinder (ed.). Representations of locally compact groups with applications, part I. 1979.
38.2 T.H. Koornwinder (ed.). Representations of locally compact groups with applications, part II. 1979.
39 O.J. Vrieze, G.L. Wanrooy. Colloquium stochastische spelen. 1978.
40 J. van Tiel. Convexe analyse. 1979.
41 H.J.J. te Riele (ed.) Colloquium numerical treatment of integral equations. 1979.
42 J.C. van Vliet (red.). Colloquium capita implementatie van programmeertalen. 1980.
43 A.M. Cohen, H.A. Wilbrink. Eindige groepen (een inleidende cursus). 1980.
44 J.G. Verwer (ed.) . Colloquium numerical solution of partial differential equations. 1980.
45 P. Klint (red.). Colloquium; hogere programmeertalen en computerarchitectuur. 1980.
46.1 P.M.G. Apers (red.). Colloquium databankorganisatie, deel l. 1981.
46.2 P.G.M. Apers (red.). Colloquium databankorganisatie, deel 2. 1981. .
47.1 P. W. Hemker (ed.). NUMAL, numerical procedures in ALGOL 60: general information and indices. 1981.
47.2 P.W. Hemker (ed.). NUMAL, numerical procedures procedures in ALGOL 60, vol. l: elementary procedures; vol. 2: algebraic evaluatons. 1981.
47.3 P.W. Hemker (ed.). NUMAL, numerical procedures in ALGOL 60, vol. 3A: linear algebra part I. 1981.
47.4 P.W. Hemker (ed.). NUMAL, numerical procedures in ALGOL 60, vol 3B: linear algebra, part II. 1981.
47.5 P.W. Hemker (ed.). NUMAL, procedures in ALGOL 60, vol. 4: analytical evaluations; vol. 5A: analytical problems. part I. 1981
47.6 P.W. Hemker (ed.). NUMAL, procedures in ALGOL 60, vol 5B: analytical problems, part II. 1981
47.7 P.W. Hemker (ed.). NUMAL, procedures in ALGOL 60, vol. 6: special functions and constants; vol. 7: interpolation and approximation. 1981
48.1 P.M.B. Vitányi, J. van Leeuwen, P. van Emde Boas (red.). Colloquium complexiteit en algoritmen, deel 1. 1982.
48.2 P.M.B. Vitányi, J. van Leeuwen, P. van Emde Boas (red.). Colloquium complexiteit en algoritmen, deel II. 1982.
49 T.H. Koomwinder (ed.) The structure of real semisimple Lie groups. 1982
50 H. Nijmeijer. Inleiding systeemtheorie. 1982.
51 P..J. Hoogendoorn (red.). Cursus cryptografie. 1983.