

Een Bayesian random censoring model voor ontbrekende waarden in proteomica data

Auteur:
Kees van Rooijen

Begeleider:
Tjeerd Dijkstra

18 juli 2015
7.5 ECTS

In proteomica data zijn vaak veel ontbrekende waarden door beperkingen van de meetapparatuur. Het probleem van deze ontbrekende waarden is dat ze niet willekeurig voorkomen, maar vaker voorkomen bij eiwitten met lagere concentraties. Statistische methoden die hier geen rekening mee houden kunnen dus onbetrouwbare resultaten geven. In deze scriptie vergelijken we het empirical Bayesian random censoring threshold (EBRCT) model met een methode die toch aanneemt dat ontbrekende data volledig willekeurig zijn (MCAR), k-nearest neighbour imputatie (KNN), singular value thresholding imputation (SVTI), een model dat alleen de rangorde van data gebruikt (NCRI) en een fixed censoring model (FCEN).

Om modellen te vergelijken hebben we een benchmark dataset gebruikt. Deze dataset bestaat uit peptideconcentraties uit urinemonsters van 134 mannen en 134 vrouwen. Elke methode is getest op zijn vermogen om om te gaan met ontbrekende data, door de methode te combineren met een classifier en zo van proefpersonen het geslacht te voorspellen.

Voor zowel EBRCT als FCEN is het beter om op basis van de modellen te imputeren en een support vector machine te gebruiken voor classificatie, dan om een Naïve Bayes classifier te gebruiken met de parameters van het model. EBRCT en SVTI classificeren dan van alle modellen de meeste personen juist.

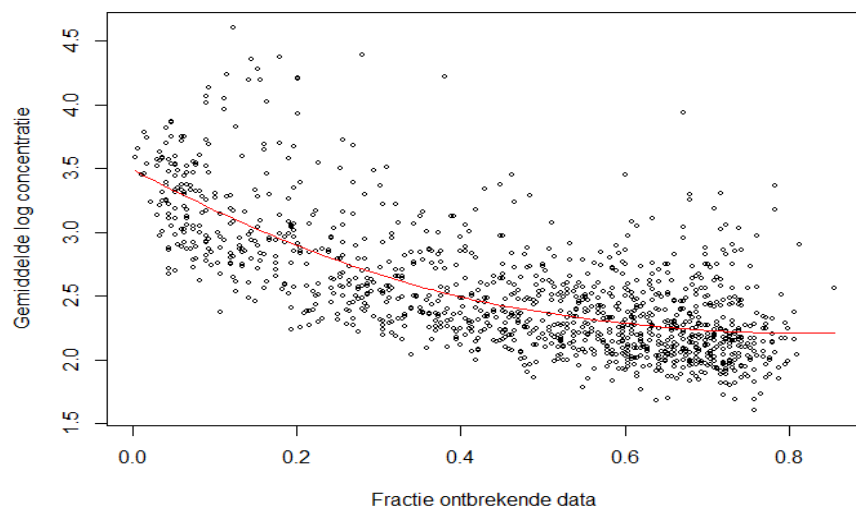
Inhoudsopgave

1	Introductie	4
2	Achtergrond	5
2.1	Bayesiaanse statistiek	5
2.1.1	Introductie	5
2.1.2	Model met een enkele parameter	7
2.1.3	Model met meerdere parameters	12
3	Methoden	16
3.1	Data en opzet	16
3.2	EBRCT	17
3.2.1	Simpel model	17
3.2.2	Sampling	18
3.2.3	Uitgebreid model	19
3.2.4	prior parameters	20
3.2.5	Classificatie	21
3.3	Andere methoden	24
3.3.1	Fixed Threshold	24
3.3.2	Imputatie en SVM	26
4	Resultaten en discussie	27
5	Conclusies en Aanbevelingen	29
6	Referenties	30
		30

1 Introductie

Biomarkers zijn essentieel voor het ontdekken van ziektes en voor het volgen van de effectiviteit van een medische behandeling. Proteomica, de studie van eiwitten, is een interessant vakgebied voor het vinden van biomarkers. Dit houdt in dat er wordt gezocht naar eiwitten die onder verschillende condities (zoals in een ziek lichaam en een gezond lichaam) in verschillende concentraties voorkomen. In dit onderzoek richten we ons niet op eiwitten maar op peptiden, die iets kleiner zijn maar een zelfde structuur hebben. Omdat het proces van identificatie veel overeenkomsten vertoont, is het verder niet nodig om hier onderscheid tussen te maken [1].

We spreken van ontbrekende data wanneer een peptide wel in de urine van de ene proefpersoon wordt waargenomen maar niet in de urine van de ander. Dit kan verschillende oorzaken hebben. Het kan zo zijn dat de waarde onder de waarnemingslimiet van de apparatuur ligt, dat de peptide echt niet voorkomt in de urine door biologische redenen, of dat deze wel voorkomt en ook boven de waarnemingslimiet ligt, maar niet wordt waargenomen door problemen met preparatie of verwerking [2]. Uit eerder onderzoek is gebleken dat peptiden die in lagere concentraties voorkomen vaker niet worden waargenomen [3, 4]. Dit is ook het geval in de dataset die wij gebruiken (figuur 1).



Figuur 1: Ieder datapunt stelt een peptide voor. Op de x-as staat de fractie ontbrekende waarden voor die peptide en op de y-as de gemiddelde log concentratie van de waarden die wel zijn waargenomen.

De onwillekeurigheid van het ontbreken van data bemoeilijkt statistische analyse en dus het ontdekken van biomarkers. Op basis van deze kennis is het EBRCT model ontworpen. In eerder onderzoek is deze methode vergeleken met andere modellen in de

toepassing van proteomica. Hier werd elke methode gebruikt om de peptiden aan te wijzen die het meest verschillen tussen verschillende condities. Het was daar mogelijk dit direct te testen, doordat van tevoren bekend was welke peptiden verschillen en welke niet. Het EBRCT model was daar in de belangrijkste dataset de beste voorspeller [4].

In dit onderzoek testen we de modellen op een benchmark dataset uit Dakna et al. [1]. In deze dataset is niet van tevoren bekend welke peptiden het meest verschillen. Daarom hebben we gekozen voor een methode waarbij classificatie van testgevallen de uiteindelijke prestatie weergeeft. De gemeten prestatie is het resultaat van een combinatie van een imputatiemethode en een methode voor classificatie.

In deze scriptie geef ik eerst de benodigde statistische achtergrond om het Bayesiaanse model te begrijpen. Daarna bekijken we op welke manier alle modellen worden gebruikt en bekijken we hoe deze presteren op de dataset.

2 Achtergrond

2.1 Bayesiaanse statistiek

Om het model dat we hebben gebruikt te begrijpen, is eerst een basiskennis van Bayesiaanse statistiek nodig. In dit hoofdstuk geef ik een introductie. Ik zal hierbij de structuur aanhouden en steeds voorbeelden gebruiken uit *Applied Bayesian Statistics* van Cowles [5].

2.1.1 Introductie

2.1.1.1 Subjectieve kansen

Bij de traditionele vorm van kansrekening wordt de kans op een gebeurtenis gezien als de relatieve frequentie van die gebeurtenis, als het experiment een groot aantal keer zou worden herhaald [6, blz. 248]. Als we zeggen dat bij het werpen van een munt de kans op kop 0.5 is, bedoelen we dat, wanneer we de munt een groot aantal keer gooien, we verwachten dat in ongeveer de helft van de gevallen kop naar boven komt.

In sommige gevallen is dit echter geen zinnige definitie. Wil je bijvoorbeeld voorspellen wat de kans is dat er een economische crisis uitbreekt, dan zou dit inhouden dat je moet kijken naar het al dan niet uitbreken van een crisis in een land met exact dezelfde economische en politieke omstandigheden. In dat geval kun je beter van subjectieve kansen spreken. Een subjectieve kans is een getal tussen 0 en 1 dat aangeeft hoe waarschijnlijk iemand het vindt dat een gebeurtenis plaatsvindt (of al heeft plaatsgevonden).

2.1.1.2 A priori kansen

Als voorbeeld bestuderen we een vrouw die wordt uitgenodigd voor een mammogram screening voor borstkanker. In dit geval zijn er twee mogelijke toestanden, die we modellen noemen:

Model	A priori kans
Borstkanker	0.0045
Geen borstkanker	0.9955

Tabel 1: de a priori kansen op borstkanker

1. Ze heeft borstkanker
2. Ze heeft geen borstkanker

Haar doel is om te bekijken wat de kans is dat ze borstkanker heeft. Omdat ze geen speciale voorgeschiedenis heeft met borstkanker, neemt ze aan dat haar kansen gelijk zijn aan de gemiddelde kans van vrouwen die deelnemen aan zo'n screening. Uit onderzoek blijkt dat van vrouwen die deelnemen aan een screening, de proportie die binnen een jaar wordt gediagnostiseerd met borstkanker gelijk is aan 0.0045 [7]. Op basis van deze gegevens bepaalt ze de waarschijnlijkheid van de twee modellen, voordat ze enige data heeft verzameld over zichzelf, de *a priori kansverdeling*, of korter *prior*. (tabel 1). Merk op dat, hoewel de deze prior is gebaseerd op objectieve informatie, het gaat om een subjectieve kans. Deze mevrouw heeft namelijk borstkanker of niet (dus de traditionele kans is ofwel 1, ofwel 0), maar de kans waar we van spreken geeft aan hoe waarschijnlijk de vrouw het zelf vindt dat ze borstkanker heeft.

2.1.1.3 Data

Op het moment dat ze de test bij de dokter laat afnemen en zo *data* verzamelt, zal ze meer te weten komen over haar situatie, en zo de kansen kunnen bijstellen. Om het eenvoudig te houden nemen we aan dat er twee mogelijke uitslagen zijn bij de mammogram: een positieve uitslag (M+) geeft aan dat er hoogstwaarschijnlijk sprake is van borstkanker, en een negatieve uitslag (M-) geeft aan dat er waarschijnlijk sprake is van geen borstkanker. De test is echter niet perfect: deze kan fout-positieve of fout-negatieve resultaten geven [7]. Zie tabel 2.

Model	P(M+)	P(M-)
Borstkanker	0.724	0.276
Geen borstkanker	0.0274	0.9766

Tabel 2: de kansverdelingen van de uitkomst van de mammogram voor beide modellen

2.1.1.4 Likelihood en a posteriori kansen

Om aan de hand van de verzamelde data de kansen te kunnen bijstellen, gebruiken we Bayes' regel. Die zegt: $P(model|data) \propto P(model) * P(data|model)$. Het teken \propto betekent *is proportioneel met*. De *likelihood* bij een bepaald model is de kans op de waargenomen data, gegeven het model ($P(data|model)$).

Stel we krijgen te horen dat het resultaat van de test positief is (M+). We kunnen nu de verwachtingen bijstellen (tabel 3). Omdat Bayes' regel slechts iets zegt over proportionaliteit, moeten we de resultaten normaliseren, zodat de totale kans uitkomt op 1. De genormaliseerde uitkomst ($P(model|data)$) heet dan de *a posteriori kans*.

Model	Prior	Likelihood M+	Prior * Likelihood	Posterior
Borstkanker	0.0045	0.724	0.0033	0.107
Geen borstkanker	0.9955	0.0274	0.0273	0.893

Tabel 3: de berekening van de a posteriori kansen voor beide modellen

Zoals te zien is in de laatste kolom, is de kans dat er inderdaad sprake is van borstkanker, ondanks het positieve resultaat, nog steeds vrij klein. Als er nu besloten wordt verder onderzoek uit te voeren, kunnen de resultaten van de tests weer worden gebruikt om de verwachtingen bij te stellen. De posterior die we net hebben berekend wordt dan de prior voor de volgende stap.

In dit voorbeeld werd duidelijk dat Bayes' regel een goede manier geeft om aan de hand van verkregen data je subjectieve kansen bij te stellen.

2.1.2 Model met een enkele parameter

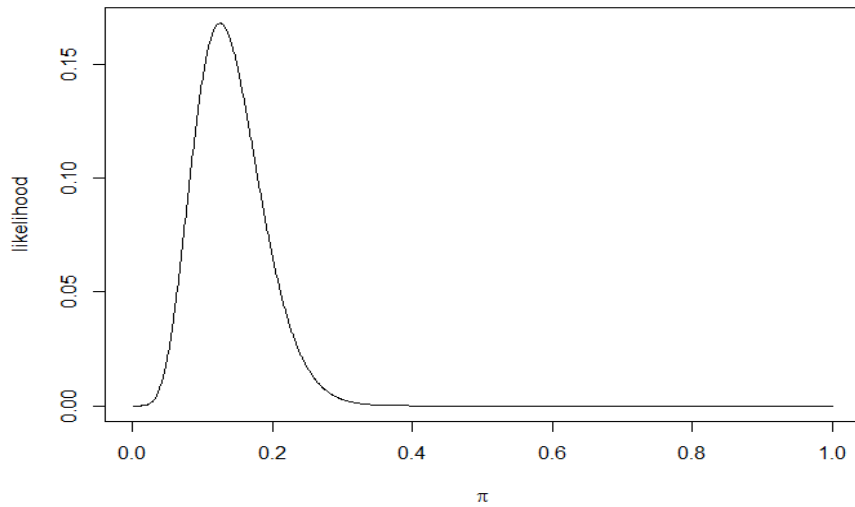
In dit hoofdstuk volgen we een ander voorbeeld, weer uit *Applied Bayesian Statistics* van Cowles [5]. In 2002 werd er op de University of Iowa onderzoek gedaan wat het effect zou zijn van een verhoging van het collegegeld met 19%. De onbekende parameter π is de proportie van de studenten die als gevolg van deze verhoging de universiteit zou verlaten. Omdat het te veel werk is om alle studenten te ondervragen, wordt slechts een kleine steekproef ondervraagd en op basis daarvan wordt de verwachting voor π bepaald.

2.1.2.1 Kansverdeling

Allereerst is het belangrijk om een kansverdeling te kiezen die de situatie goed omschrijft. Op de vraag of een student de school zal verlaten, kan deze "ja" of "nee" antwoorden. Zoals eerder genoemd, is de proportie van studenten die positief antwoordt gelijk aan π . Omdat we geen verdere informatie hebben over de studenten, kunnen we het antwoord van een student behandelen als een random variabele, met een succeskans (de kans op "ja") van π .

Omdat de steekproef willekeurig wordt gekozen uit alle studenten, kunnen we aannemen dat de antwoorden onafhankelijk van elkaar zijn. Het aantal successen bij een ondervraagde groep is dan binomiaal verdeeld, met kans π . De kans op y successen bij een steekproef van n studenten kan dan als volgt berekend worden:

$$p(y|\pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad (1)$$



Figuur 2: de binomiale likelikhoo verdeling voor $n = 50$ en $y = 7$

Er zijn namelijk y successen met succeskans π (dus een totale kans van π^y), $n - y$ mislukkingen met kans $(1 - \pi)$ (totale kans van $(1 - \pi)^{n-y}$). De successen kunnen echter op elke plek voorkomen, en het aantal verschillende verdelingen is te vinden als $\binom{n}{y}$.

2.1.2.2 Likelihood functie

Nadat de steekproef is gehouden en de studenten zijn ondervraagd, is het aantal successen y bekend. In het vorige hoofdstuk werd de likelihood van een model gedefiniëerd als de kans op de waargenomen data, als dat model waar zou zijn. In dit geval kunnen we dus voor elke π de likelihood berekenen.

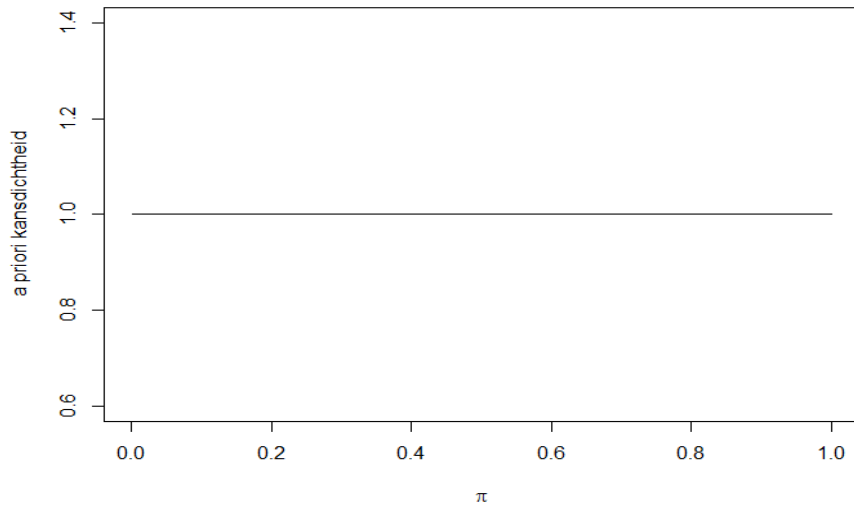
$$L(\pi; y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad (2)$$

Het eerste deel van die term is echter voor elke π gelijk, dus kunnen we ook schrijven als in vergelijking 3. Een voorbeeld van hoe deze functie er uit ziet is te zien in figuur 2.

$$L(\pi; y) \propto \pi^y (1 - \pi)^{n-y} \quad (3)$$

2.1.2.3 Prior

Net als in het vorige geval, moeten we ook hier op een bepaalde manier onze verwachting voor de hoogte van π aangeven. Als we bijvoorbeeld nog geen enkele verwachting hebben



Figuur 3: non-informatieve a priori kansverdeling

van de echte waarde van π , gebruiken we een non-informatieve prior (figuur 3). Het is echter mogelijk om elke verwachting of kennis over π in een prior te verwerken. Zouden we bijvoorbeeld op magische wijze weten dat π niet groter is dan 0.5, dan zou de prior er uit kunnen zien als in figuur 4. Straks zullen we een gebruikelijke manier bekijken om de prior te kiezen.

2.1.2.4 A posteriori verdeling

Nadat we een prior kiezen, kunnen we de resultaten van het onderzoek gebruiken om onze verwachtingen bij te stellen, volgens dezelfde regels als in het simpele geval. De a posteriori kansverdeling wordt gegeven door de a priori kansverdeling voor π te vermenigvuldigen met de likelihood $L(\pi; data)$.

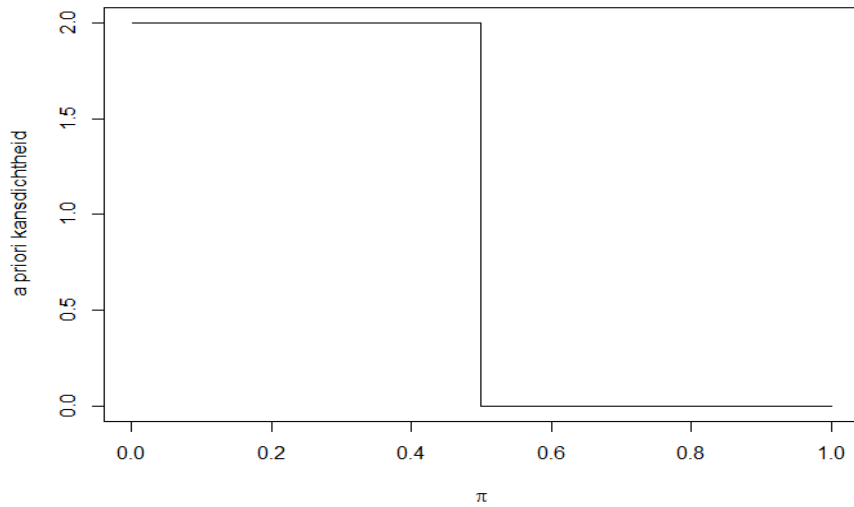
$$p(\pi|data) \propto p(\pi) * L(\pi; data) \quad (4)$$

In dit voorbeeld kunnen we dat schrijven als:

$$p(\pi|y) \propto p(\pi) * \pi^y (1 - \pi)^{n-y}, \quad 0 < \pi < 1 \quad (5)$$

We kiezen ervoor de noninformatie prior te gebruiken, dus $p(\pi) = 1$ voor $0 < \pi < 1$. Stel dat we bij een steekproef van 50 studenten 7 successen hebben waargenomen, wordt de a posteriori kansverdeling gegeven door vergelijking 6. In figuur 5 is deze gevisualiseerd.

$$p(\pi|y) \propto 1 * \pi^7 (1 - \pi)^{43}, \quad 0 < \pi < 1 \quad (6)$$



Figuur 4: a priori kansverdeling als π niet groter kan zijn dan 0.5

Dit geeft dus weer wat de verwachting is over het totale percentage schoolverlaters. De verwachtingswaarde voor π , en dus het percentage schoolverlaters, is nu $\frac{7}{50} = 0.14$. Als we een steekproef hadden gehouden bij 100 mensen met 14 positieve antwoorden, zou de verwachtingswaarde hetzelfde zijn. De verdeling voor π zou er dan wel anders uitzien, omdat we dan zekerder zouden zijn van onze verwachting.

2.1.2.5 Conjugate Priors

Bij het kiezen van een prior is het vaak wenselijk om rekening te houden met het gemak van de berekening van de a posteriori kansverdeling. Door een prior te kiezen die uit dezelfde familie van functies komt als de likelihood functie (dit noemen we een conjugate prior), kan de berekening erg snel verlopen.

We bekijken wat dit zou betekenen in het geval van een binomiale verdeling, zoals in het vorige voorbeeld. De likelihood functie zag er uit als:

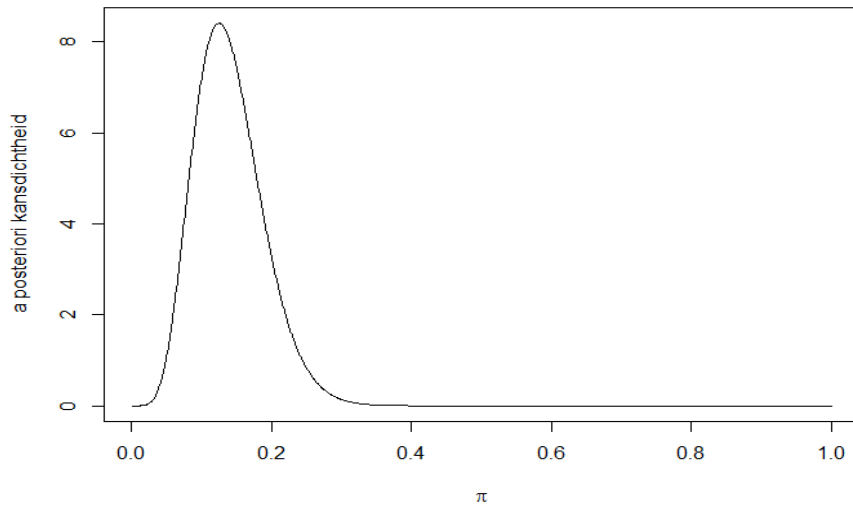
$$L(\pi; y) \propto \pi^y (1 - \pi)^{n-y} \quad (7)$$

Merk op dat de likelihood functie gaat over het interval $(0,1)$, en dat zowel π als $(1 - \pi)$ voorkomen en worden verheven tot een nonnegatieve macht. Er bestaat een kansverdeling met dezelfde eigenschappen, namelijk de *beta* kansverdeling. De kansverdeling met parameters α en β en random variabele π wordt geschreven als

$$\pi \sim \mathcal{B}(\alpha, \beta) \quad (8)$$

en is gedefiniëerd door

$$p(\pi) \propto \pi^{\alpha-1} (1 - \pi)^{\beta-1}, \quad 0 < \pi < 1 \quad (9)$$



Figuur 5: a posteriori kansverdeling

Het is nu erg makkelijk om de a posteriori kansverdeling te berekenen. Met een beta-prior en een binomiale likelihood functie, geldt dan namelijk:

$$\begin{aligned}
 p(\pi|y) &\propto p(\pi) * L(\pi; y) \\
 &\propto \pi^{\alpha-1}(1-\pi)^{\beta-1} * \pi^y(1-\pi)^{n-y} \\
 &= \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1} \\
 p(\pi|y) &= \mathcal{B}(y+\alpha, n-y+\beta)
 \end{aligned} \tag{10}$$

Het enige resterende probleem is dan het kiezen van parameters α en β zodat de prior een goede representatie is van de voorkennis over de parameter π . In het vorige voorbeeld zagen we dat het gebruik van een non-informatieve prior de volgende a posteriori kansverdeling opleverde:

$$\begin{aligned}
 p(\pi|y) &\propto \pi^y(1-\pi)^{n-y} \\
 p(\pi|y) &= \mathcal{B}(y+1, n-y+1)
 \end{aligned} \tag{11}$$

Dit betekent dus dat de informatie van $\mathcal{B}(\alpha, \beta)$ gelijk staat aan de informatie na het hebben gedaan van een steekproef van grootte $\alpha + \beta - 2$, met $\alpha - 1$ successen, want $\alpha = y + 1$ en $\beta = n - y + 1$. Dit kan worden gebruikt om de parameters van de beta-prior te kiezen wanneer er al een eerdere steekproef is gedaan, maar ook wanneer de kennis vager is. Een vaag vermoeden dat ongeveer de helft van de studenten de universiteit zou verlaten kan worden aangegeven met $\mathcal{B}(4, 4)$. Een sterk vermoeden dat slechts een kwart de school zou verlaten zou kunnen worden aangegeven met $\mathcal{B}(19, 59)$.

Stel dat we dit laatste vermoeden zouden hebben (dus de prior is $p(\pi) = \mathcal{B}(19, 59)$), en we vervolgens in een steekproef van 30 studenten 5 successen hebben, dan is de a posteriori kansverdeling dus snel te berekenen en gelijk aan $p(\pi|y) = \mathcal{B}(24, 84)$.

2.1.3 Model met meerdere parameters

Vaak is er niet slechts één, maar enkele onbekende parameters die onderzocht moeten worden. In het geval van normaal verdeelde data is het vaak zo dat zowel het gemiddelde als de variantie onbekend zijn. Eerst bekijken we de situaties waarin slechts één van beiden onbekend is, en daarna combineren we de priors voor een model waarin beide parameters onbekend zijn.

2.1.3.1 Onbekend gemiddelde, bekende variantie

Zoals gezegd nemen we aan dat de data een normale verdeling kent.

$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad (12)$$

Voor een trekking uit die data ziet de kansverdeling er dan als volgt uit:

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) \quad (13)$$

We noteren de n waarnemingen uit de steekproef y_1, y_2, \dots, y_n . Omdat we hier ook weer aannemen dat de waarnemingen onafhankelijk zijn, kan de kansverdeling van de waarnemingen worden berekend als

$$\begin{aligned} p(y_1, y_2, \dots, y_n|\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^{2n}}} * \exp\left(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^{2n}}} * \exp\left(-\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{2\sigma^2}\right) * \exp\left(-\frac{\sum_{i=1}^n (\bar{y} - \mu)^2}{2\sigma^2}\right) \end{aligned} \quad (14)$$

Waar \bar{y} het gemiddelde van de steekproef is. Omdat we in dit geval aannemen dat de variantie bekend en constant is, kunnen we de likelihood berekenen voor alleen het onbekende gemiddelde. Alle delen uit deze vergelijking die gelijk zijn voor alle μ kunnen we dan wegwerken zoals we dat deden in vergelijking 3.

$$\begin{aligned} L(\mu|y) &\propto \exp\left(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{n(\mu - \bar{y})^2}{2\sigma^2}\right) \end{aligned} \quad (15)$$

Om nu een conjugate prior voor de normale verdeling te vinden, hebben we een verdeling nodig waarin de random variabele op dezelfde plaats voorkomt als μ in de vergelijking

voor de likelihood. Merk op dat de normale verdeling zelf deze vorm heeft, zoals te zien is in vergelijking 13. We kunnen nu dus een normaal verdeelde prior voor μ gebruiken.

$$\begin{aligned}\mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ p(\mu) &\propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)\end{aligned}\tag{16}$$

Om te laten zien dat we inderdaad te maken hebben met een conjugate prior, moet ook de a posteriori kansverdeling een normale verdeling zijn. De berekening van die verdeling gaat als volgt:

$$\begin{aligned}p(\mu|y) &\propto p(\mu) * L(\mu|y) \\ &\propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) * \exp\left(-\frac{n(\mu - \bar{y})^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \frac{n(\mu - \bar{y})^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\sigma^2(\mu - \mu_0)^2 + n\sigma_0^2(\mu - \bar{y})^2}{2\sigma_0^2\sigma^2}\right) \\ &= \exp\left(-\frac{\sigma^2(\mu^2 - 2\mu\mu_0 + \mu_0^2) + n\sigma_0^2(\mu^2 - 2\mu\bar{y} + \bar{y}^2)}{2\sigma_0^2\sigma^2}\right) \\ &= \exp\left(-\frac{\sigma^2\mu^2 - 2\sigma^2\mu\mu_0 + \sigma^2\mu_0^2 + n\sigma_0^2\mu^2 - 2n\sigma_0^2\mu\bar{y} + n\sigma_0^2\bar{y}^2}{2\sigma_0^2\sigma^2}\right) \\ &\propto \exp\left(-\frac{\sigma^2\mu^2 - 2\sigma^2\mu\mu_0 + n\sigma_0^2\mu^2 - 2n\sigma_0^2\mu\bar{y}}{2\sigma_0^2\sigma^2}\right) \\ &= \exp\left(-\frac{(\sigma^2 + n\sigma_0^2)\left(\mu^2 - \frac{2\sigma^2\mu_0 + 2n\sigma_0^2\mu\bar{y}}{\sigma^2 + n\sigma_0^2}\right)}{2\sigma_0^2\sigma^2}\right) \\ &= \exp\left(-\frac{(\sigma^2 + n\sigma_0^2)\left(\mu^2 - 2\mu\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{y}}{\sigma^2 + n\sigma_0^2}\right)}{2\sigma_0^2\sigma^2}\right) \\ &\propto \exp\left(-\frac{(\sigma^2 + n\sigma_0^2)\left(\mu - \frac{\sigma^2\mu_0 + n\sigma_0^2\bar{y}}{\sigma^2 + n\sigma_0^2}\right)^2}{2\sigma_0^2\sigma^2}\right) \\ &= \mathcal{N}\left(\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{y}}{\sigma^2 + n\sigma_0^2}, \frac{\sigma_0^2\sigma^2}{\sigma^2 + n\sigma_0^2}\right)\end{aligned}\tag{17}$$

De a posteriori kansverdeling is dus inderdaad een normale verdeling. De nieuwe parameters van die normale verdeling zijn dan $\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{y}}{\sigma^2 + n\sigma_0^2}$ en $\frac{\sigma_0^2\sigma^2}{\sigma^2 + n\sigma_0^2}$.

2.1.3.2 Onbekende variantie, bekend gemiddelde

Voor het andere geval, waar juist het gemiddelde als bekend wordt verondersteld en de variantie moet worden onderzocht, kan een soortgelijke afleiding worden gemaakt.

We beginnen weer bij de kansverdeling voor de waarnemingen. In plaats van variantie σ^2 gebruiken we in dit geval de precisie $\tau^2 = 1/\sigma^2$. Vergelijking 18 gebruikt dezelfde definitie voor de normale verdeling als vergelijking 13, alleen met andere notatie.

$$\begin{aligned} p(y_1, y_2, \dots, y_n | \mu, \tau^2) &= \prod_{i=1}^n \frac{\sqrt{\tau^2}}{\sqrt{2\pi}} * \exp\left(-\frac{\tau^2(y_i - \mu)^2}{2}\right) \\ &\propto (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2} \tau^2\right) \end{aligned} \quad (18)$$

Schrijf nu $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ voor de variantie van de steekproef. Omdat we aannemen dat het gemiddelde constant is, kunnen we nu de likelihood voor alleen de onbekende precisie berekenen.

$$L(\tau^2 | y) \propto (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{(n-1)s^2}{2} \tau^2\right) \quad (19)$$

De conjugate prior die hier bij hoort is de Gamma functie. Deze is gegeven door

$$\begin{aligned} \tau^2 &\sim \mathcal{G}(\alpha, \beta) \\ p(\tau^2 | \alpha, \beta) &\propto (\tau^2)^{\alpha-1} * \exp(-\beta \tau^2) \end{aligned} \quad (20)$$

Nu kunnen we de a posteriori kansverdeling weer op een soortgelijke manier afleiden

$$\begin{aligned} p(\tau^2 | y) &= p(\tau^2) * L(\tau^2 | y) \\ &\propto (\tau^2)^{\alpha-1} * \exp(-\beta \tau^2) * (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{(n-1)s^2}{2} \tau^2\right) \\ &= (\tau^2)^{\alpha + \frac{n}{2} - 1} * \exp\left(-\left(\beta + \frac{(n-1)s^2}{2}\right) \tau^2\right) \\ &= \mathcal{G}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2}\right) \end{aligned} \quad (21)$$

De a posteriori is weer een Gamma verdeling, met nieuwe parameters $\alpha + \frac{n}{2}$ en $\beta + \frac{(n-1)s^2}{2}$.

2.1.3.3 Gemiddelde en variantie allebei onbekend

In de situatie waar we het model willen gebruiken is er echter sprake van een situatie waarin zowel het gemiddelde als de variantie onbekend is. We gebruiken hier voor een combinatie van de vorige twee priors, namelijk een Normal-Gamma prior:

$$\begin{aligned} p(\mu, \tau^2) &= p(\tau^2) * p(\mu | \tau^2) \\ &= \mathcal{G}(\alpha, \beta) * \mathcal{N}\left(\mu_0, \frac{1}{\kappa \tau^2}\right) \\ &\propto (\tau^2)^{\alpha-1} * \exp(-\beta \tau^2) * \sqrt{\tau^2} * \exp\left(-\frac{(\mu - \mu_0)^2}{2} \kappa \tau^2\right) \end{aligned} \quad (22)$$

Bij het berekenen van de likelihood kunnen we de variantie van de steekproef $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ gebruiken, waar \bar{y} weer het gemiddelde van de steekproef is. Dan

$$\begin{aligned}
L(\mu, \tau^2 | y) &\propto (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2} \tau^2\right) \\
&= (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2}{2} \tau^2\right) \\
&= (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{\tau^2((n-1)s^2 + n(\bar{y} - \mu)^2)}{2}\right) \\
&= (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{\tau^2 s^2 (n-1)}{2}\right) * \exp\left(-\frac{\tau^2 n (\bar{y} - \mu)^2}{2}\right)
\end{aligned} \tag{23}$$

Ten slotte kan de a posteriori kansverdeling weer worden gevonden:

$$\begin{aligned}
p(\mu, \tau^2 | y) &= p(\mu, \tau^2) * L(\mu, \tau^2 | y) \\
&\propto (\tau^2)^{\alpha-1} * \exp(-\beta \tau^2) * \sqrt{\tau^2} * \exp\left(-\frac{(\mu - \mu_0)^2}{2} \kappa \tau^2\right) \\
&\quad * (\tau^2)^{\frac{n}{2}} * \exp\left(-\frac{\tau^2 s^2 (n-1)}{2}\right) * \exp\left(-\frac{\tau^2 n (\bar{y} - \mu)^2}{2}\right) \\
&= (\tau^2)^{\alpha-1+\frac{n}{2}} * \exp\left(-\left(\beta + \frac{s^2(n-1)}{2}\right) \tau^2\right) \\
&\quad * \sqrt{\tau^2} * \exp\left(-\frac{\kappa(\mu - \mu_0)^2 + n(\bar{y} - \mu)^2}{2} \tau^2\right)
\end{aligned} \tag{24}$$

De noemer van de breuk in de laatste exponent moet van de vorm $k_1(\mu - \mu_1)^2$ zijn, zodat het resultaat weer een Normal-Gamma verdeling is. Merk op dat alle restproducten die geen μ bevatten weggewerkt kunnen worden naar de eerste exponent.

$$\begin{aligned}
\kappa(\mu - \mu_0)^2 + n(\bar{y} - \mu)^2 &= \kappa(\mu^2 - 2\mu\mu_0 + \mu_0^2) + n(\bar{y}^2 - 2\bar{y}\mu + \mu^2) \\
&= \kappa\mu^2 - 2\kappa\mu\mu_0 + \kappa\mu_0^2 + n\bar{y}^2 - 2n\bar{y}\mu + n\mu^2 \\
&= (\kappa + n)\left(\mu_0^2 - \frac{2\kappa\mu\mu_0}{\kappa + n} - \frac{2n\bar{y}\mu}{\kappa + n}\right) + \kappa\mu_0^2 + n\bar{y}^2 \\
&= (\kappa + n)\left(\mu^2 - \frac{2\mu(\kappa\mu_0 + n\bar{y})}{\kappa + n}\right) + \kappa\mu_0^2 + n\bar{y}^2 \\
&= (\kappa + n)\left(\left(\mu - \frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}\right)^2 - \left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}\right)^2\right) + \kappa\mu_0^2 + n\bar{y}^2 \\
&= (\kappa + n)\left(\mu - \frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}\right)^2 - \frac{(\kappa\mu_0 + n\bar{y})^2}{\kappa + n} + \kappa\mu_0^2 + n\bar{y}^2
\end{aligned} \tag{25}$$

Het eerste gedeelte heeft nu de gewenste vorm, het tweede gedeelte bevat geen μ en kan uiteindelijk dus verplaatst worden naar de eerste exponent (in vergelijking 24). Eerst

kan dit deel echter nog wat vereenvoudigd worden:

$$\begin{aligned}
-\frac{(\kappa\mu_0 + n\bar{y})^2}{\kappa + n} + \kappa\mu_0^2 + n\bar{y}^2 &= \frac{(\kappa + n)(\kappa\mu_0^2 + n\bar{y}^2) - (\kappa\mu_0 + n\bar{y})^2}{\kappa + n} \\
&= \frac{(\kappa^2\mu_0^2 + \kappa n\mu_0^2 + \kappa n\bar{y}^2 + n^2\bar{y}^2) - (\kappa^2\mu_0^2 + 2\kappa n\mu_0\bar{y} + n^2\bar{y}^2)}{\kappa + n} \\
&= \frac{\kappa^2\mu_0^2 + \kappa n\mu_0^2 + \kappa n\bar{y}^2 + n^2\bar{y}^2 - \kappa^2\mu_0^2 - 2\kappa n\mu_0\bar{y} - n^2\bar{y}^2}{\kappa + n} \\
&= \frac{\kappa n\mu_0^2 + \kappa n\bar{y}^2 - 2\kappa n\mu_0\bar{y}}{\kappa + n} \\
&= \frac{\kappa n(\mu_0^2 + \bar{y}^2 - 2\mu_0\bar{y})}{\kappa + n} \\
&= \frac{\kappa n(\mu_0 - \bar{y})^2}{\kappa + n}
\end{aligned} \tag{26}$$

Nu de noemer de gewenste vorm heeft en het restproduct is uitgewerkt, keren we voor de laatste keer terug naar de berekening van de a posteriori verdeling.

$$\begin{aligned}
p(\mu, \tau^2 | y) &\propto (\tau^2)^{\alpha-1+\frac{n}{2}} * \exp\left(-\left(\beta + \frac{s^2(n-1)}{2}\right)\tau^2\right) \\
&\quad * \sqrt{\tau^2} * \exp\left(-\frac{(\kappa + n)\left(\mu - \frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}\right)^2 + \frac{\kappa n(\mu_0 - \bar{y})^2}{\kappa + n}}{2}\tau^2\right) \\
&= (\tau^2)^{\alpha-1+\frac{n}{2}} * \exp\left(-\left(\beta + \frac{s^2(n-1)}{2} + \frac{\kappa n(\mu_0 - \bar{y})^2}{2(\kappa + n)}\right)\tau^2\right) \\
&\quad * \sqrt{\tau^2} * \exp\left(-\frac{(\kappa + n)\left(\mu - \frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}\right)^2}{2}\tau^2\right) \\
&= \mathcal{G}\left(\alpha + \frac{n}{2}, \beta + \frac{s^2(n-1)}{2} + \frac{\kappa n(\mu_0 - \bar{y})^2}{2(\kappa + n)}\right) \\
&\quad * \mathcal{N}\left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}, \frac{1}{(\kappa + n)\tau^2}\right)
\end{aligned} \tag{27}$$

En dit is weer een Normal-Gamma verdeling. Het feit dat er een conjugate prior is voor normaal verdeelde data is voor het model dat we gaan bestuderen erg belangrijk, omdat dit het proces van de berekening zeer snel maakt. Met kennis van de data, kan met de informatie (n, \bar{y}, s^2) direct de kennis over het gemiddelde en de variantie worden bijgesteld.

3 Methoden

3.1 Data en opzet

Voor het vergelijken van de methoden hebben we een benchmark proteomica dataset gebruikt uit Dakna et al. [1]. Voor deze dataset is van 134 mannen en 134 vrouwen een

urinemonster afgenomen. Door middel van capillaire elektroforese–massaspectrometrie (CE-MS) zijn peptideconcentraties gemeten. We hebben alleen deelnemers geselecteerd waarbij ten minste 25% van de peptiden gemeten waren. Na deze selectie bleven 130 mannen en 132 vrouwen over. Ook zijn alleen de peptiden geselecteerd die bij ten minste 25% van de deelnemers gemeten zijn. Dit leverde een selectie van 1156 peptiden op. In de resterende dataset ontbreekt 44.2% van de data.

Om de verschillende methoden te testen op hun vermogen om te gaan met ontbrekende data hebben we 2-fold cross-validatie gebruikt. De te onderzoeken conditie is in het geval van deze benchmark dataset het geslacht. De dataset is opgesplitst in een deel voor training en een deel voor testen. Elke methode is daarna gebruikt om, op basis van de kennis uit de training set, het geslacht van elke persoon uit de test set te bepalen. Hoe dat precies gebeurt verschilt per methode en wordt in de rest van dit hoofdstuk uitgelegd. Om de resultaten te vergelijken berekenen we steeds de receiver operating characteristic (ROC), waar mannen de positieven zijn en vrouwen de negatieven.

3.2 EBRCT

3.2.1 Simpel model

Allereerst bekijken we het EBRCT model [4]. Voor een gesimplificeerde versie van het EBRCT model bekijken we van K proefpersonen de concentratie van een bepaalde peptide. De waargenomen concentraties (inclusief missende waarden) noteren we als y_1, \dots, y_K . De werkelijke concentraties zonder missende data noteren we als x_1, \dots, x_K . Ook nemen we aan dat er thresholds c_1, \dots, c_K bestaan. Een waarde wordt alleen waargenomen wanneer deze boven de threshold ligt. We nemen aan dat zowel de werkelijke concentraties als de thresholds uit een normale verdeling komen. Met NA noteren we een missende waarde.

$$\begin{aligned} x_k &\sim \mathcal{N}(\mu_x, \sigma_x^2) \\ c_k &\sim \mathcal{N}(\mu_c, \sigma_c^2) \\ y_k &= \begin{cases} x_k & \text{als } x_k > c_k \\ \text{NA} & \text{als } x_k < c_k \end{cases} \end{aligned} \tag{28}$$

μ_x, σ_x^2, μ_c en σ_c^2 zijn de onbekende parameters. Om de informatie uit de data te kunnen verwerken, gebruiken we weer de Normal-Gamma verdeling voor deze parameters. Merk op dat waar in de afleiding van de Normal-Gamma a posteriori kansverdeling de precisie $\tau^2 = \frac{1}{\sigma^2}$ werd gebruikt om de berekening overzichtelijk te houden, we nu weer de variantie

σ^2 gebruiken in de notatie.

$$\begin{aligned}
\mu_x &\sim \mathcal{N}(m_{x0}, \frac{\sigma_x^2}{k_{x0}}) \\
\frac{1}{\sigma_x} &\sim \mathcal{G}(a_{x0}, b_{x0}) \\
\mu_c &\sim \mathcal{N}(m_{c0}, \frac{\sigma_c^2}{k_{c0}}) \\
\frac{1}{\sigma_c} &\sim \mathcal{G}(a_{c0}, b_{c0})
\end{aligned} \tag{29}$$

In de situatie waarin x_1, \dots, x_K bekend zijn, is de a posteriori kansverdeling weer een Normal-Gamma verdeling met nieuwe parameters, zoals in 27:

$$\begin{aligned}
m_x &= \frac{k_{x0}m_{x0} + K\bar{x}}{k_{x0} + K} \\
k_x &= k_{x0} + K \\
a_x &= a_{x0} + \frac{K}{2} \\
b_x &= b_{x0} + \frac{s_x^2(K-1)}{2} + \frac{k_{x0}K(m_{x0} - \bar{x})^2}{2(k_{x0} + K)}
\end{aligned} \tag{30}$$

En op soortgelijke manier voor de parameters zijn m_c, k_c, a_c en b_c te vinden wanneer c_1, \dots, c_K bekend zijn.

3.2.2 Sampling

In werkelijkheid zijn zowel x_1, \dots, x_K als c_1, \dots, c_K onbekend. Sommige x_k kunnen direct uit de data worden afgeleid, wanneer y_k is waargenomen. Om waarden voor de andere gevallen te vinden, combineren we de waarnemingen met de huidige kennis van de parameters in vergelijking 29. Door uit de juiste verdelingen te samplen verkrijgen we dan waarden voor x_k en c_k .

1. Sample uit de verdelingen uit vergelijking 29 om op basis van de hyperparameters $m_{x0}, k_{x0}, a_{x0}, b_{x0}, m_{c0}, k_{c0}, a_{c0}$ en b_{c0} waarden te krijgen voor $\sigma_x, \mu_x, \sigma_c, \mu_c$.
2. Combineer voor elke k de waargenomen waarde y_k met de kennis uit vergelijking 28 om waarden voor x_k en c_k te vinden. Er zijn hier twee gevallen te onderscheiden: de concentratie is waargenomen of niet waargenomen.

Waargenomen

Uit het feit dat de concentratie waargenomen is blijkt dat er sprake is van het eerste geval in vergelijking 28. Dus volgt dat $y_k = x_k$ en $x_k > c_k$. Dit betekent dat we een waarde voor c_k kunnen vinden door te samplen uit

$$c_k \sim \mathcal{N}_{c_k < x_k}(\mu_c, \sigma_c^2) \tag{31}$$

Niet waargenomen

Uit $y_k = \text{NA}$ en vergelijking 28 volgt dat $x_k < c_k$. In dit geval zijn x_k en c_k allebei onbekend, en moeten we voor beide waarden samplen uit de bijpassende normale verdeling, met de extra eigenschap dat $x_k < c_k$.

$$\begin{pmatrix} c_k \\ x_k \end{pmatrix} \sim \mathcal{N}_{x_k < c_k} \left(\begin{pmatrix} \mu_c \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_c & 0 \\ 0 & \sigma_x \end{pmatrix} \right) \quad (32)$$

3. Gebruik de gevonden waarden voor x_1, \dots, x_K en c_1, \dots, c_K om volgens vergelijking 30 de hyperparameters bij te stellen.

De bijgestelde hyperparameters uit stap 3 worden dan gebruikt om de stappen opnieuw door te lopen, net zolang totdat het model voldoende convergentie vertoont. De distributies van de parameters μ_x , σ_x , μ_c en σ_c kunnen dan worden gebruikt om conclusies te trekken op basis van het model.

3.2.3 Uitgebreid model

In het vorige geval keken we naar slechts één conditie en één peptidesoort. Voor het werkelijke geval moet het model worden uitgebreid. In het uitgebreide model bekijken we van alle gemeten peptiden de concentratie, met een onderscheid tussen de verschillende condities.

In een algemeen geval schrijven we voor de gemeten concentratie $y_{jk}^{(l)}$ met drie indices, waar $j \in (1 \dots J)$ staat voor de soort peptide, $l \in (1 \dots L)$ voor de conditie en $k \in (1 \dots K^{(l)})$ voor de proefpersoon. In het uitgebreide model geldt verder:

- de gemiddelde peptideconcentratie $\mu_{x,j}^{(l)}$ uniek is voor elke conditie en peptide
- de variantie van de peptideconcentratie $\sigma_{x,j}^2$ en de gemiddelde thresholdconcentratie $\mu_{c,j}$ worden gedeeld door de verschillende condities maar zijn verschillend voor elke peptide
- de variantie van de thresholdconcentratie σ_c^2 gelijk is voor alle condities en peptiden

Verder verlopen de berekeningen aan het model op een zelfde wijze als die bij het eenvoudige model:

$$\begin{aligned} x_{jk}^{(l)} &\sim \mathcal{N}(\mu_{x,j}^{(l)}, \sigma_{x,j}^2) \\ c_{jk}^{(l)} &\sim \mathcal{N}(\mu_{c,j}, \sigma_c^2) \\ y_{jk}^{(l)} &= \begin{cases} x_{jk}^{(l)} & \text{als } x_{jk}^{(l)} > c_{jk}^{(l)} \\ \text{NA} & \text{als } x_{jk}^{(l)} < c_{jk}^{(l)} \end{cases} \end{aligned} \quad (33)$$

De berekening van die parameters gaat dan weer als in vergelijking 29.

$$\begin{aligned}
\mu_{x,j}^{(l)} &\sim \mathcal{N}(m_{x,j}^{(l)}, \frac{\sigma_{x,j}^2}{k_x^{(l)}}) \\
\frac{1}{\sigma_{x,j}^2} &\sim \mathcal{G}(a_x, b_{x,j}) \\
\mu_{c,j} &\sim \mathcal{N}(m_{c,j}, \frac{\sigma_c^2}{k_c}) \\
\frac{1}{\sigma_c} &\sim \mathcal{G}(a_c, b_c)
\end{aligned} \tag{34}$$

Het updaten van de hyperparameters gaat dan als volgt.

$$\begin{aligned}
m_{x,j}^{(l)} &= \frac{k_{x0}^{(l)} m_{x0,j}^{(l)} + K^{(l)} \bar{x}_j^{(l)}}{k_{x0}^{(l)} + K^{(l)}} \\
k_x^{(l)} &= k_{x0}^{(l)} + K^{(l)} \\
a_x &= a_{x0} + \frac{K}{2} \\
b_{x,j} &= b_{x0,j} + \frac{s_{x,j}^2 (K-1)}{2} + \frac{k_{x0} K (m_{x0,j} - \bar{x}_j)^2}{2(k_{x0} + K)} \\
m_{c,j} &= \frac{k_{c0} m_{c0,j} + K \bar{c}_j}{k_{c0} + K} \\
k_c &= k_{c0} + K \\
a_c &= a_{c0} + \frac{K}{2} \\
b_c &= b_{c0} + \frac{s_c^2 (K-1)}{2} + \frac{k_{c0} K (m_{c0} - \bar{c})^2}{2(k_{c0} + K)}
\end{aligned} \tag{35}$$

Hier is $\bar{x}_j^{(l)}$ het gemiddelde van de werkelijke concentraties van een bepaalde peptide bij een bepaalde conditie, K de som van alle $K^{(l)}$, $s_{x,j}^2$ en \bar{x}_j de variantie en het gemiddelde van een peptideconcentratie over alle condities, \bar{c}_j de gemiddelde threshold waarde over alle condities en ten slotte s_c^2 en \bar{c} de variantie en het gemiddelde van de threshold waarden over alle condities en peptiden.

Het model bestaat nu uit samplen uit vergelijking 34 voor parameters, met die parameters en de data samplen zoals in vergelijking 31 en vergelijking 32, en vervolgens de hyperparameters updaten zoals in vergelijking 35.

3.2.4 prior parameters

Om het model compleet te maken zijn nu alleen nog de parameters voor de a priori kansverdeling nodig. Het EBRCT model doet dat op basis van de data. Door dit hergebruik van de data heet het een *empirisch* Bayesiaans model in plaats van een volledig

Bayesiaans model. Bij het opstellen van de prior wordt geen onderscheid gemaakt tussen verschillende peptiden en condities, dit gebeurt pas bij het updaten. Zo worden m_{x0} en σ_{x0}^2 berekend als het gemiddelde en de variantie van alle geobserveerde concentraties. Voor m_{c0} en σ_{c0}^2 wordt van elke peptidesoort alleen de laagste geobserveerde waarde bekeken, en van die minima worden gemiddelde en variantie berekend.

Hyperparameters k_{x0} , a_{x0} , k_{c0} en a_{c0} worden berekend als fracties f_{μ_x} , f_{σ_x} , f_{μ_c} en f_{σ_c} van de gemiddelde hoeveelheid relevante data $\frac{1}{I} \sum_{l=1}^L K^{(l)}$, $\sum_{l=1}^L K^{(l)}$, $\sum_{l=1}^L K^{(l)}$ en $J \sum_{l=1}^L K^{(l)}$. De fracties geven de weging van de prior aan. Omdat de berekening van de prior biased is en deze voornamelijk wordt gebruikt om het model ook stabiel te laten werken in het geval van veel ontbrekende data, kunnen de fracties relatief klein zijn. In Koopmans et al. [4] wordt voor de fracties 0.01, 2, 0.2 en 0.01 gebruikt. De data die we hier gebruiken heeft per conditie echter 10 keer zo veel monsters, dus vooral de fractie f_{σ_x} hoeft niet zo groot te zijn. We gebruiken daarom 0.01 voor alle fracties.

Ten slotte moeten b_{x0} en b_{c0} nog berekend worden. Dit gaat volgens vergelijking 36.

$$\begin{aligned} b_{x0} &= \frac{a_{x0}}{2} \frac{1}{\sigma_{x0}^2} \\ b_{c0} &= \frac{a_{c0}}{2} \frac{1}{\sigma_{c0}^2} \end{aligned} \tag{36}$$

3.2.5 Classificatie

Na het uitvoeren van de Gibbs sampler vinden we per peptide de gemiddelde waarden voor $\mu_x^{(1)}$, $\mu_x^{(2)}$, σ_x en μ_c , en voor alle peptiden samen σ_c . Het doel is nu om voor een nieuwe proefpersoon met (al dan niet waargenomen) peptideconcentraties y_1, \dots, y_L zo goed mogelijk te bepalen of deze mannelijk of vrouwelijk is. Met andere woorden, als de geslachten worden weergegeven met C_1 en C_2 , wat is dan $p(C_1|y)$?

3.2.5.1 Naive Bayes

De Naive Bayes classifier gebruikt Bayes' regel in combinatie met de aanname dat verschillende peptiden onafhankelijk zijn. Voor het opstellen van die classifier volgen we Bishop [8]. De kans op geslacht C_1 op basis van de data y is volgens Bayes' regel te vinden als:

$$\begin{aligned} p(C_1|y) &= \frac{p(y|C_1)p(C_1)}{p(y)} \\ &= \frac{p(y|C_1)p(C_1)}{p(y|C_1)p(C_1) + p(y|C_2)p(C_2)} \\ &= \frac{1}{1 + \exp(-a)} \\ &= \sigma(a) \end{aligned} \tag{37}$$

Waar $p(C)$ de a priori kans op de conditie is, σ de sigmoid functie is en

$$\begin{aligned}
a &= \log \frac{p(y|C_1)p(C_1)}{p(y|C_2)p(C_2)} \\
&= \log \frac{p(y|C_1)}{p(y|C_2)} \\
&= \log p(y|C_1) - \log p(y|C_2)
\end{aligned} \tag{38}$$

wanneer we aannemen dat $p(C_1) = p(C_2)$. Omdat we nog geen kennis hebben op het moment dat de data wordt bekeken is dit een redelijke aanname. Nu gebruiken we de aanname van onafhankelijkheid tussen verschillende peptiden:

$$\begin{aligned}
a &= \log p(y_1, \dots, y_J|C_1) - \log p(y_1, \dots, y_J|C_2) \\
&= \log \prod_{j=1}^J p(y_j|C_1) - \log \prod_{j=1}^J p(y_j|C_2) \\
&= \sum_{j=1}^J \log p(y_j|C_1) - \sum_{j=1}^J \log p(y_j|C_2) \\
&= \sum_{j=1}^J [\log p(y_j|C_1) - \log p(y_j|C_2)]
\end{aligned} \tag{39}$$

Om $p(y_j|C_i)$ te berekenen moet onderscheid gemaakt worden tussen de situatie waarin y_j is waargenomen en de situatie waarin hij niet is waargenomen:

$$\begin{aligned}
p(y_j \in \mathbb{R}|C^{(l)}) &= \mathcal{N}(y_j|\mu_{x,j}^{(l)}, \sigma_{x,j}^2) \Phi\left(\frac{y_j - \mu_{c,j}}{\sigma_c}\right) \\
p(y_j = \text{NA}|C^{(l)}) &= \Phi\left(\frac{\mu_{c,j} - \mu_{x,j}^{(l)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}}\right)
\end{aligned} \tag{40}$$

Hier is Φ de cumulatieve verdelingsfunctie. Als we alles dan samenvoegen krijgen we:

$$\begin{aligned}
a &= \sum_{\forall j: y_j \in \mathbb{R}} \left[\log \left(\mathcal{N}(y_j|\mu_{x,j}^{(1)}, \sigma_{x,j}^2) \Phi\left(\frac{y_j - \mu_{c,j}}{\sigma_c}\right) \right) - \log \left(\mathcal{N}(y_j|\mu_{x,j}^{(2)}, \sigma_{x,j}^2) \Phi\left(\frac{y_j - \mu_{c,j}}{\sigma_c}\right) \right) \right] \\
&\quad + \sum_{\forall j: y_j = \text{NA}} \left[\log \Phi\left(\frac{\mu_{c,j} - \mu_{x,j}^{(1)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}}\right) - \log \Phi\left(\frac{\mu_{c,j} - \mu_{x,j}^{(2)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}}\right) \right]
\end{aligned} \tag{41}$$

Het stuk in de eerste som kan veel vereenvoudigd worden:

$$\begin{aligned}
& \log \left(\mathcal{N}(y_j | \mu_{x,j}^{(1)}, \sigma_{x,j}^2) \Phi \left(\frac{y_j - \mu_{c,j}}{\sigma_c} \right) \right) - \log \left(\mathcal{N}(y_j | \mu_{x,j}^{(2)}, \sigma_{x,j}^2) \Phi \left(\frac{y_j - \mu_{c,j}}{\sigma_c} \right) \right) \\
&= \log \left(\mathcal{N}(y_j | \mu_{x,j}^{(1)}, \sigma_{x,j}^2) \right) + \log \left(\Phi \left(\frac{y_j - \mu_{c,j}}{\sigma_c} \right) \right) - \log \left(\mathcal{N}(y_j | \mu_{x,j}^{(2)}, \sigma_{x,j}^2) \right) - \log \left(\Phi \left(\frac{y_j - \mu_{c,j}}{\sigma_c} \right) \right) \\
&= \log \left(\mathcal{N}(y_j | \mu_{x,j}^{(1)}, \sigma_{x,j}^2) \right) - \log \left(\mathcal{N}(y_j | \mu_{x,j}^{(2)}, \sigma_{x,j}^2) \right) \\
&= \log \left(\frac{1}{\sqrt{2\pi}\sigma_{x,j}} \exp \left(-\frac{(y_j - \mu_{x,j}^{(1)})^2}{2\sigma_{x,j}^2} \right) \right) - \log \left(\frac{1}{\sqrt{2\pi}\sigma_{x,j}} \exp \left(-\frac{(y_j - \mu_{x,j}^{(2)})^2}{2\sigma_{x,j}^2} \right) \right) \\
&= \log \left(\frac{1}{\sqrt{2\pi}\sigma_{x,j}} \right) + \log \left(\exp \left(-\frac{(y_j - \mu_{x,j}^{(1)})^2}{2\sigma_{x,j}^2} \right) \right) - \log \left(\frac{1}{\sqrt{2\pi}\sigma_{x,j}} \right) - \log \left(\exp \left(-\frac{(y_j - \mu_{x,j}^{(2)})^2}{2\sigma_{x,j}^2} \right) \right) \\
&= -\frac{(y_j - \mu_{x,j}^{(1)})^2}{2\sigma_{x,j}^2} + \frac{(y_j - \mu_{x,j}^{(2)})^2}{2\sigma_{x,j}^2} \\
&= -\frac{y_j^2 - 2y_j\mu_{x,j}^{(1)} + \mu_{x,j}^{(1)2}}{2\sigma_{x,j}^2} + \frac{y_j^2 - 2y_j\mu_{x,j}^{(2)} + \mu_{x,j}^{(2)2}}{2\sigma_{x,j}^2} \\
&= \frac{-\mu_{x,j}^{(1)2} + 2y_j\mu_{x,j}^{(1)} - 2y_j\mu_{x,j}^{(2)} + \mu_{x,j}^{(2)2}}{2\sigma_{x,j}^2} \\
&= \frac{(\mu_{x,j}^{(1)} - \mu_{x,j}^{(2)})(y_j - \frac{1}{2}(\mu_{x,j}^{(1)} + \mu_{x,j}^{(2)}))}{\sigma_{x,j}^2}
\end{aligned} \tag{42}$$

Dit resultaat invullen geeft:

$$\begin{aligned}
a = & \sum_{\forall j: y_j \in \mathbb{R}} \frac{(\mu_{x,j}^{(1)} - \mu_{x,j}^{(2)})(y_j - \frac{1}{2}(\mu_{x,j}^{(1)} + \mu_{x,j}^{(2)}))}{\sigma_{x,j}^2} \\
& + \sum_{\forall j: y_j = \text{NA}} \left[\log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(1)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) - \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(2)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) \right]
\end{aligned} \tag{43}$$

Door met vergelijking (43) de waarde voor a te berekenen, kan nu volgens $p(C_1|y) = \frac{1}{1+\exp(-a)}$ de gezochte kans berekend worden. In het geval van het berekenen van een ROC curve is het niet nodig om deze kans te berekenen. Omdat de sigmoid functie strikt monotoon stijgend is kunnen personen aan de hand van de waardes van a gerangschikt worden.

3.2.5.2 Support Vector Machine

Een andere mogelijkheid voor classificatie is een Support Vector Machine (SVM). Hiervoor gebruiken we de R functie "svm" uit het package "e1071" versie 1.6-4 met default

parameters. Het is nu wel nodig om de ontbrekende waarden te imputeren met een enkele waarde. Weer worden eerst op basis van de trainingsdata per peptide waarden gevonden voor $\mu_x^{(1)}$, $\mu_x^{(2)}$, σ_x en μ_c , en voor alle peptiden samen σ_c . Daarna moeten zowel in de trainingsdata als in de testdata de ontbrekende waarden worden geïmputeerd.

Wanneer het geslacht bekend is, is μ_x ook bekend. Omdat de concentratie niet is waargenomen, weten we dat er voor $(c, x)^T$ wordt gesampled uit

$$\mathcal{N}_{x < c} \left(\begin{pmatrix} \mu_c \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_c & 0 \\ 0 & \sigma_x \end{pmatrix} \right) \quad (44)$$

De gemiddelde waarde die x dan krijgt is de waarde die we imputeren.

In de testdata is het geslacht juist nog niet bekend. Hierdoor weten we niet of we gebruik moeten maken van $\mu_x^{(1)}$ of van $\mu_x^{(2)}$. Merk op dat we voor een enkele peptide j , de kans op een bepaald geslacht berekenen als

$$p(C_1|y_j) = \sigma(a) \quad (45)$$

$$a = \begin{cases} \frac{(\mu_{x,j}^{(1)} - \mu_{x,j}^{(2)})(y_j - \frac{1}{2}(\mu_{x,j}^{(1)} + \mu_{x,j}^{(2)}))}{\sigma_{x,j}^2} & \text{als } y_j \in \mathbb{R} \\ \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(1)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) - \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(2)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) & \text{als } y_j = \text{NA} \end{cases}$$

Door deze twee mogelijkheden om a te berekenen aan elkaar gelijk te stellen, kunnen we, wanneer $y_j = \text{NA}$ de waarde voor y_j vinden waarmee a , en dus $p(C_1|y_j)$ dezelfde waarde had gekregen. Deze waarde kunnen we dan gebruiken om te imputeren.

$$\frac{(\mu_{x,j}^{(1)} - \mu_{x,j}^{(2)})(y_j - \frac{1}{2}(\mu_{x,j}^{(1)} + \mu_{x,j}^{(2)}))}{\sigma_{x,j}^2} = \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(1)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) - \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(2)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right)$$

$$(y_j - \frac{1}{2}(\mu_{x,j}^{(1)} + \mu_{x,j}^{(2)})) = \left[\log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(1)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) - \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(2)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) \right] \frac{\sigma_{x,j}^2}{(\mu_{x,j}^{(1)} - \mu_{x,j}^{(2)})}$$

$$y_j = \left[\log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(1)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) - \log \Phi \left(\frac{\mu_{c,j} - \mu_{x,j}^{(2)}}{\sqrt{\sigma_c^2 + \sigma_{x,j}^2}} \right) \right] \frac{\sigma_{x,j}^2}{(\mu_{x,j}^{(1)} - \mu_{x,j}^{(2)})} + \frac{1}{2}(\mu_{x,j}^{(1)} + \mu_{x,j}^{(2)}) \quad (46)$$

Nu kunnen we een SVM trainen op de geïmputeerde trainingsdata, en die vervolgens gebruiken om geïmputeerde testdata te classificeren.

3.3 Andere methoden

3.3.1 Fixed Threshold

Het Fixed Censoring (FCEN) [2, 3] model gebruikt ook de kennis over de thresholds, alleen gebruikt het een vaste threshold per peptide in plaats van een random threshold.

Dit vaste threshold wordt gekozen als de laagste gemeten waarde van een peptideconcentratie over beide geslachten min 10^{-6} . Per peptide kent dit model $\mu_x^{(1)}$, $\mu_x^{(2)}$, $\sigma_x^{(1)}$, $\sigma_x^{(2)}$ en c . Deze parameters worden benaderd door middel van maximum Likelihood estimation. Hiervoor gebruiken we het BFGS algoritme uit R functie "optim".

3.3.1.1 Naive Bayes

Classificatie kan net als bij EBRCT met een Naive Bayes classifier. Omdat dit model met een aparte σ_x werkt voor beide condities, is de classifier minder ver te vereenvoudigen. De berekening van a gaat dan als volgt:

$$a = \sum_{\forall j: y_j \in \mathbb{R}} \left[\log \mathcal{N}(y_j | \mu_{x,j}^{(1)}, \sigma_{x,j}^{(1)2}) - \log \mathcal{N}(y_j | \mu_{x,j}^{(2)}, \sigma_{x,j}^{(2)2}) \right] + \sum_{\forall j: y_j = \text{NA}} \left[\log \Phi \left(\frac{c_j - \mu_{x,j}^{(1)}}{\sigma_{x,j}^{(1)2}} \right) - \log \Phi \left(\frac{c_j - \mu_{x,j}^{(2)}}{\sigma_{x,j}^{(2)2}} \right) \right] \quad (47)$$

3.3.1.2 Support Vector Machine

Ook hier kunnen we de methode omzetten naar een methode om mee te imputeren. Omdat het gaat om het imputeren van ontbrekende waarden, weten we dat $y = \text{NA}$. Wederom moet er onderscheid gemaakt worden tussen bekend en onbekend geslacht. Wanneer het geslacht bekend is, kunnen we voor een waarde van x samplen uit

$$\mathcal{N}_{x < c}(\mu_x^{(l)}, \sigma_x^{(l)}) \quad (48)$$

De gemiddelde waarde die x krijgt is dan de waarde om te imputeren.

Het geval waarin geslacht onbekend is, is wederom wat lastiger, omdat hier met verschillende varianties voor verschillende condities wordt gewerkt. De waarde van y wordt weer verkregen door de formules voor $p(C_1 | y_j)$ in het geval dat $y_j \in \mathbb{R}$ en in het geval dat $y = \text{NA}$ aan elkaar gelijk te stellen, en op te lossen voor y . Eerst herschrijven we het eerste deel, daarna stellen we ze aan elkaar gelijk.

$$\begin{aligned} & \log \mathcal{N}(y | \mu_x^{(1)}, \sigma_x^{(1)2}) - \log \mathcal{N}(y | \mu_x^{(2)}, \sigma_x^{(2)2}) \\ &= \log \left(\frac{1}{\sqrt{2\pi}\sigma_x^{(1)}} \exp \left(-\frac{(y - \mu_x^{(1)})^2}{2\sigma_x^{(1)2}} \right) \right) - \log \left(\frac{1}{\sqrt{2\pi}\sigma_x^{(2)}} \exp \left(-\frac{(y - \mu_x^{(2)})^2}{2\sigma_x^{(2)2}} \right) \right) \\ &= -\log(\sigma_x^{(1)}) - \frac{(y - \mu_x^{(1)})^2}{2\sigma_x^{(1)2}} + \log(\sigma_x^{(2)}) + \frac{(y - \mu_x^{(2)})^2}{2\sigma_x^{(2)2}} \\ &= -\log(\sigma_x^{(1)}) - \frac{y^2 - 2y\mu_x^{(1)} + \mu_x^{(1)2}}{2\sigma_x^{(1)2}} + \log(\sigma_x^{(2)}) + \frac{y^2 - 2y\mu_x^{(2)} + \mu_x^{(2)2}}{2\sigma_x^{(2)2}} \\ &= \left(\frac{1}{2\sigma_x^{(2)2}} - \frac{1}{2\sigma_x^{(1)2}} \right) y^2 + \left(\frac{\mu_x^{(1)}}{\sigma_x^{(1)2}} - \frac{\mu_x^{(2)}}{\sigma_x^{(2)2}} \right) y + \frac{\mu_x^{(2)2}}{2\sigma_x^{(2)2}} - \frac{\mu_x^{(1)2}}{2\sigma_x^{(1)2}} + \log(\sigma_x^{(2)}) - \log(\sigma_x^{(1)}) \end{aligned} \quad (49)$$

Nu stellen we beide delen aan elkaar gelijk

$$\log \mathcal{N}(y|\mu_x^{(1)}, \sigma_x^{(1)2}) - \log \mathcal{N}(y|\mu_x^{(2)}, \sigma_x^{(2)2}) = \log \Phi \left(\frac{c_j - \mu_{x,j}^{(1)}}{\sigma_{x,j}^{(1)2}} \right) - \log \Phi \left(\frac{c_j - \mu_{x,j}^{(2)}}{\sigma_{x,j}^{(2)2}} \right) \quad (50)$$

En herschrijven we dit als

$$\begin{aligned} 0 &= ay^2 + by + c \\ a &= \frac{1}{2\sigma_x^{(2)2}} - \frac{1}{2\sigma_x^{(1)2}} \\ b &= \frac{\mu_x^{(1)}}{\sigma_x^{(1)2}} - \frac{\mu_x^{(2)}}{\sigma_x^{(2)2}} \\ c &= \frac{\mu_x^{(2)2}}{2\sigma_x^{(2)2}} - \frac{\mu_x^{(1)2}}{2\sigma_x^{(1)2}} + \log(\sigma_x^{(2)}) - \log(\sigma_x^{(1)}) - \log \Phi \left(\frac{c_j - \mu_{x,j}^{(1)}}{\sigma_{x,j}^{(1)2}} \right) + \log \Phi \left(\frac{c_j - \mu_{x,j}^{(2)}}{\sigma_{x,j}^{(2)2}} \right) \end{aligned} \quad (51)$$

Tot slot kunnen we y vinden als

$$y = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \text{ of } y = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad (52)$$

y is hier dus de waarde waarvoor $p(C_1|y_j = y) = p(C_1|y_j = \text{NA})$. Omdat we al weten dat $y_j = \text{NA}$, kiezen we van de twee mogelijke uitkomsten voor y de kleinste positieve waarde. Merk op dat voor het speciale geval waar $\sigma_x^{(1)} = \sigma_x^{(2)}$ geldt dat $y = -\frac{c}{b}$.

Net als bij EBRCT imputeren we nu de ontbrekende data waarbij geslacht bekend is en waarbij het geslacht onbekend is. Een SVM wordt getraind op de geïmputeerde trainingsdata en getest op de geïmputeerde test data.

3.3.2 Imputatie en SVM

Alle andere modellen die we gebruiken zijn sowieso al gericht op het imputeren van een enkele waarde. Daardoor kunnen ze direct in combinatie met een SVM worden gebruikt.

Bij elke methode wordt eerst de volledige training set geïmputeerd. Vervolgens wordt op die geïmputeerde data een SVM getraind. Vervolgens wordt elk afzonderlijk testgeval samengenomen met alle trainings data en geïmputeerd. Het geïmputeerde testgeval kan dan worden geclassificeerd met de getrainde SVM.

3.3.2.1 Mean imputation

Een methode die aanneemt dat ontbreken van de data volledig willekeurig is (MCAR) is mean imputation. Hierbij worden ontbrekende waarden met het gemiddelde van alle wél waargenomen waarden voor die peptide en dat geslacht geïmputeerd. Hierdoor verandert het gemiddelde van een peptide bij een bepaald geslacht niet. Omdat elke peptide afzonderlijk wordt bekeken wordt hier net als bij het EBRCT model geen rekening gehouden met eventuele afhankelijkheid tussen verschillende peptiden.

3.3.2.2 Rank imputation

In Dakna et al. [1] wordt de Wilcoxon rank sum test [9] gebruikt om te testen op een verschil in de peptideconcentratie tussen beide geslachten. Hierbij wordt niet gekeken naar de concentratie zelf van een peptide, maar naar de rang van die concentratie. Concentraties van beide geslachten worden samen genomen en gesorteerd, en de positie in die gesorteerde rij bepaalt de rang. Alle niet geobserveerde waarden krijgen dezelfde laagste rang. Om dit resultaat te kunnen gebruiken in combinatie met een SVM, vervangen we simpelweg de waarden door hun rang. Dit is dus een Non-parametric Censoring Rank Imputatie (NCRI).

3.3.2.3 KNN

De k-nearest neighbours methode (KNN) [10] bekijkt peptiden met soortgelijke waarden om een waarde te imputeren. Voor een peptide met een ontbrekende concentratie zoekt deze methode de k peptiden die op de wel waargenomen waarden het meest met die peptide overeenkomt. Om de ontbrekende waarden nu in te vullen wordt een gewogen gemiddelde genomen van de waargenomen concentraties bij die k peptiden, waarbij peptiden die meer overeenkomen zwaarder meewegen. We gebruiken hiervoor de R functie "impute.knn" uit het "impute" package versie 1.40.0, met $k = 75$.

3.3.2.4 SVTI

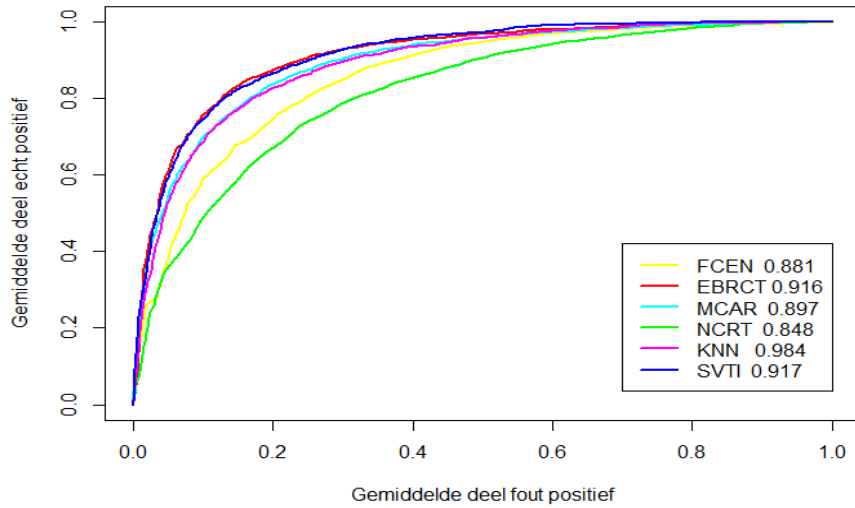
Ten slotte gebruiken we Singular Value Threshold Imputation (SVTI) [11]. Dit algoritme imputeert waarden op basis van de gehele matrix. Hiervoor gebruiken we R functie "SVTImpute" uit het "imputation" package versie 2.0.1, met $\lambda = 1000$ en $\text{threshold} = 10^{-5}$.

4 Resultaten en discussie

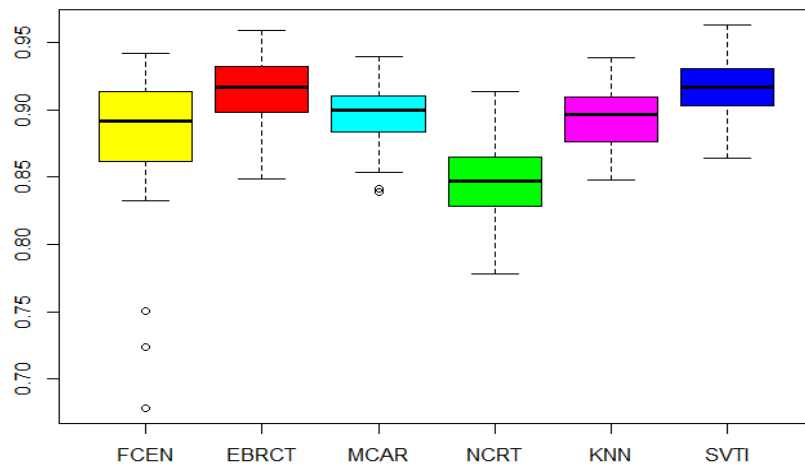
Elke methode is, zoals in het vorige hoofdstuk uitgelegd, gebruikt om voor elke proefpersoon een voorspelling van het geslacht te doen. Deze procedure is 50 keer herhaald, en ter vergelijking bekijken we de ROC curve van de voorspellingen, gemiddeld over die 50 herhalingen (Figuur 6).

Voor een beschrijving van de resultaten van alle 50 herhalingen bekijken we ook een boxplot van de Area Under the Curve (AUC) voor elke methode (Figuur 7). Dit is dus de oppervlakte onder de curve in Figuur 6. Merk op dat een willekeurige voorspeller hier een gemiddelde score van 0.5 zou halen, en de perfecte voorspeller altijd een score van 1.0.

Bij zowel EBRCT als FCEN presteerde de variant die gebruik maakt van een SVM beter dan de Naive Bayes classifier. Bij EBRCT was de gemiddelde AUC score bij de SVM 0.92 en bij de Naive Bayes 0.85. Bij FCEN behaalde de SVM een score van 0.88 en de Naive Bayes classifier 0.75. In de vergelijkingen hebben we van beide modellen alleen de SVM variant meegenomen.



Figuur 6: ROC curve voor de voorspelling van het geslacht in de proefpersonen in de testgroep. De curve ontstaat door de threshold voor classificatie te variëren. Ook is de gemiddelde AUC weergegeven.



Figuur 7: Boxplot van de AUC statistiek van de ROC curve over 50 iteraties van de 2-fold cross validatie voor elke methode.

De resultaten komen deels overeen met eerdere vindingen [4]. SVTI werkte in dat onderzoek het best wanneer slechts naar het gemiddelde van geïmputeerde waarden werd gekeken. EBRCT werkte beter wanneer ook de andere informatie van het model gebruikt werd. In dit onderzoek gebeurt er mogelijk iets soortgelijks. Wanneer bij het EBRCT model door middel van een SVM wordt geclassificeerd, moet het model gebruikt worden om een enkele waarde te imputeren, waar informatie over het gemiddelde en de variantie in wordt samengenomen. Omdat met een andere variantie en andere gemiddeldes soms dezelfde imputatiewaarde wordt gevonden, gaat in die stap informatie verloren. Het SVTI algoritme imputeert sowieso al een enkele waarde, en verliest in deze stap dus geen informatie. Dit verklaart mogelijk waarom SVTI net iets beter presteert dan EBRCT, net als het eerste geval uit Koopmans et al. [4].

FCEN geeft in dit onderzoek, wanneer de methode wordt gebruikt om te imputeren en wordt gecombineerd met een SVM, betere resultaten in verhouding tot de andere modellen dan in Koopmans et al. [4]. Een mogelijke oorzaak hiervan is de overfitting die daar plaats kan hebben gevonden. Het FCEN model schat 5 parameters per peptide, waar per peptide slechts tussen de 2 en 12 concentraties zijn waargenomen. In de dataset die wij hebben gebruikt zijn er tussen de 66 en 259 waargenomen concentraties.

KNN presteert minder goed dan MCAR. In Koopmans et al. [4] is dit ook het geval wanneer er sprake is van een vergelijkbaar percentage ontbrekende data.

NCRI geeft de minst goede classificatie. Deze techniek is gebaseerd op de Wilcoxon rank sum test. Uit Dakna et al. [1] blijkt dat deze test niet veel beter presteert dan een standaard t-test die de ontbrekende waarden negeert. De manier waarop de SVM met de waarden omgaat moet de slechtere prestatie van deze techniek verklaren.

5 Conclusies en Aanbevelingen

We hebben verschillende imputatietechnieken getest op een benchmark dataset. Het empirical Bayesian random censoring threshold (EBRCT) model, imputatie van het gemiddelde (MCAR), k-nearest neighbour imputatie (KNN), singular value thresholding imputation (SVTI), een model dat gebruik maakt van de rangorde van de data (NCRI) en een fixed censoring model (FCEN) werden beoordeeld op hun vermogen om het geslacht van proefpersonen te voorspellen. SVTI en EBRCT waren de meest geschikte technieken wanneer ze werden gecombineerd met een SVM.

Voor zowel EBRCT als FCEN is het mogelijk om enkele waarden te imputeren op basis van de waarden van de parameters die door het model berekend worden. Deze methode van imputeren gaf, wanneer deze werd gecombineerd met een SVM, voor beide modellen een beter resultaat dan het oorspronkelijke model gecombineerd met de Naïve Bayes classifier.

De slechte prestatie van de Naïve Bayes classifier doet vermoeden dat de aanname van onafhankelijkheid die wordt gedaan niet juist is. Deze aanname wordt echter ook in EBRCT en FCEN zelf gedaan. Een mogelijke verbetering aan die methoden zou zijn om die onafhankelijkheid niet aan te nemen. Het is echter de vraag in hoeverre de methode

dan te complex wordt.

Een andere mogelijke verbetering in het geval van een dataset waarbij het aantal proefpersonen relatief groot is, is het toevoegen van aparte parameters voor de variantie van de peptideconcentratie voor beide condities. Ook kan er per peptide een aparte variantie voor het threshold worden berekend. De versie van het model die we nu gebruiken hebben is ontwikkeld voor een dataset waarbij per conditie slechts peptiden van zes gevallen waren gemeten, en bij meerdere parameters zou het model te weinig data kunnen hebben om deze goed te kunnen berekenen. In deze dataset zijn er 130 proefpersonen per conditie, dus zouden meer parameters kunnen worden toegevoegd om het model meer mogelijkheden te geven verschil tussen condities te herkennen.

Zoals beschreven in de discussie leidt de imputatie van slechts een enkele waarde bij het EBRCT model en het FCEN model tot verlies van informatie van het model, en eventueel tot slechte resultaten. Als alternatief zou kunnen worden getest wat het effect is van multiële imputatie [12] op de testresultaten.

6 Referenties

- [1] Mohammed Dakna, Keith Harris, Alexandros Kalousis, Sebastien Carpentier, Walter Kolch, Joost P Schanstra, Marion Haubitz, Antonia Vlahou, Harald Mischak, and Mark Girolami. Addressing the challenge of defining valid proteomic biomarkers and classifiers. *BMC bioinformatics*, 11(1):594, 2010.
- [2] Sandra L Taylor, Gary S Leiserowitz, and Kyoungmi Kim. Accounting for undetected compounds in statistical analyses of mass spectrometry ‘omic studies. *Statistical applications in genetics and molecular biology*, 12(6):703–722, 2013.
- [3] Yuliya Karpievitch, Jeff Stanley, Thomas Taverner, Jianhua Huang, Joshua N Adkins, Charles Ansong, Fred Heffron, Thomas O Metz, Wei-Jun Qian, Hyunjin Yoon, et al. A statistical framework for protein quantitation in bottom-up ms-based proteomics. *Bioinformatics*, 25(16):2028–2034, 2009.
- [4] Frank Koopmans, L Niels Cornelisse, Tom Heskes, and Tjeerd MH Dijkstra. Empirical bayesian random censoring threshold model improves detection of differentially abundant proteins. *Journal of proteome research*, 13(9):3871–3880, 2014.
- [5] Mary Kathryn Cowles. *Applied Bayesian statistics: with R and OpenBUGS examples*, volume 98. Springer Science & Business Media, 2013.
- [6] David S Moore. *The basic practice of statistics*. Palgrave Macmillan, 2010.
- [7] Steven P Poplack, Anna N Tosteson, Margaret R Grove, Wendy A Wells, and Patricia A Carney. Mammography in 53,803 women from the new hampshire mammography network 1. *Radiology*, 217(3):832–840, 2000.
- [8] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

- [9] Dennis R Helsel et al. *Nondetects and data analysis. Statistics for censored environmental data.* Wiley-Interscience, 2005.
- [10] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
- [11] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- [12] Joseph L Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15, 1999.