A.F. Schilperoort

Polya tree priors

Bachelorscriptie

Scriptiebegeleider: prof.dr. A.W. van der Vaart

Datum Bachelorexamen: 9 Juni 2015

Mathematisch Instituut, Universiteit Leiden

Inhoudsopgave

1 Inleiding

In dit verslag wordt gekeken naar Polya tree priors. De theorie achter Polya tree priors is uitbundig bestudeerd door Ferguson in 1974, Mauldin in 1992 en Lavine in 1994, maar een aantal essentiële vragen staan nog steeds open. Polya tree priors zijn a priori verdelingen in de Bayesiaanse statistiek. In het eerste deel van dit verslag wordt het verschil tussen de klassieke en de bayesiaanse statistiek uitgelegd. In het tweede deel wordt de Polya tree definieerd en daarna zullen we aan de hand van simulaties in R gaan kijken wat voor posteriori dichtheden volgen voor verschillende parameters.
Polya tree priors kunnen worden geconstrueerd met een bepaald algoritme die we in dit verslag zullen gaan zien. Na waarnemingen uit een bepaalde verdeling zal de a posteriori dichtheid van de Polya tree prior zich vormen tot de echte dichtheid van deze verdeling. In dit verslag gaan we ook bekijken wat het effect is van de keuze van de parameters van de Polya tree. De vorm van de posteriori dichtheid is namelijk afhankelijk van de keuze van de parameters. De parameters bepalen namelijk of de posteriori dichtheid discreet of continu is.
Polya tree priors zijn heel aantrekkelijk voor het modelleren van, in principe, elke verdeling. Als de verdeling van een aantal waarnemingen onbekend is, zal je met behulp van de Polya tree toch een redelijke schatting kunnen maken van hoe de verdeling eruit ziet.
Een nadeel van Polya tree verdelingen is de afhankelijkheid van het model dat gebruikt wordt. De partities waarin de Polya tree wordt opgedeeld bepalen grotendeels hoe de posteriori dichtheid eruit ziet. De posteriori dichtheid heeft namelijk discontinuiteiten, met kans één, bij de eindpunten van de partities. Nou is dit in het veld van statistisch modeleren niet echt een uitzondering, maar het is wel iets om rekening mee te houden.
Aan het einde van het verslag zal ik uiteenzetten wat nog handig is om te bekijken qua verbeteringen en wat verder nog het bestuderen waard is.

2 Bayesiaanse Statistiek

Zij X een discrete random variabele, met verdelingsfunctie F, afhankelijk van één parameter θ. In de statistiek weten we niet wat de waarde van θ is, maar is het onze taak om informatie over de onbekende parameter te vinden. De manier om wat over deze θ te weten te komen is om veel realisaties van X te analyseren, oftewel we gaan kijken naar een random dataset $X = (X_1, X_2, \ldots, X_n)$, waarbij elke X_i onafhankelijk verdeeld is volgens F. De echte uitkomst die we waarnemen, noteer ik met $x = (x_1, x_2, \ldots, x_n)$. Verschillende waarden van θ leiden tot verschillende kansen voor de uitkomst, oftewel $P(X = x|\theta)$ varieert met θ. In de klassieke statistiek wordt deze kans ook wel de likelihood van θ, $L(\theta|x)$, genoemd. Voor de echte waarde van θ zal de kans $P(X = x|\theta)$ waarschijnlijk groter zijn.

In de Bayesiaanse statistiek wordt het concept van likelihood vervangen door een concept van echte kans. Om dat te doen beschouwen we θ als een onbekende variabele in plaats van een onbekende constante. Aangezien θ wordt beschouwd als een random variabele geven we dit voor de duidelijkheid even aan met Θ. Deze random variabele Θ heeft natuurlijk zelf ook een kansverdeling. Deze geven we aan met $\pi_\Theta(\theta) = P(\Theta = \theta)$, en wordt de a priori verdeling genoemd. De a priori verdeling kunnen we kiezen aan de hand van de kennis die we al hebben over θ. Bayesiaanse statistici berekenen dan na een waarneming de a posteriori verdeling voor Θ uit met behulp van de regel van Bayes.

Stelling 2.1 (Regel van Bayes)**.** *Voor evenementen A en B gegeven $P(B) \neq 0$ geldt*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Als we aannemen dat Θ een random variabele is, is de kans $P(X = x|\theta)$ een conditionele kans en wordt geschreven als $P(X = x|\Theta = \theta)$. Gebruiken van de regel van Bayes geeft

$$P(\Theta = \theta|X = x) = \frac{P(X = x|\Theta = \theta)P(\Theta = \theta)}{P(X = x)}.$$

Deze conditionele kans $P(\Theta = \theta|X = x)$, afgekort $P(\theta|x)$, is de al eerder genoemde a posteriori verdeling van Θ. Wanneer we verschillende verdelingen voor X beschouwen, zal de random variabele X discreet of continu zijn, maar onze parameters zijn continu. Dat betekent dat we voor Θ met kansverdelingen werken in plaats van kansen. Een kleine aanpassing van de regel van Bayes geeft ons dan

$$f(\theta|x) = \frac{P(x|\theta)\pi(\theta)}{P(x)},$$

in het geval dat X discreet is en

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{P(x)},$$

in het geval dat X continu is, waarbij $\pi(\theta)$ de a priori verdeling op Θ is, $f(\theta|x)$ de a posteriori verdeling op Θ, en $f(x|\theta)$ de conditionele dichtheid. In beide

gevallen is de a posteriori verdeling afhankelijk van $P(X = x)$, deze is constant, dus we kunnen ook schrijven dat de a posteriori proportioneel is tot, $P(x|\theta)\pi(\theta)$ of $f(x|\theta)\pi(\theta)$, in resp. het discrete of continue geval. We geven dit aan met het standaard symbool $\propto$, dus

$$f(\theta|x) \propto \begin{cases} P(x|\theta)\pi(\theta) & \text{als } X \text{ discreet} \\ f(x|\theta)\pi(\theta) & \text{als } X \text{ continu} \end{cases}.$$

$P(x)$, wordt gezien als een normeringsconstante die van de a posteriori, $f(\theta|x)$, een dichtheid maakt.

Voorbeeld 2.2. Zij $X = \{X_1, X_2, \ldots X_n\}$ i.i.d. Bernoulli met $P_\theta(X = 1) = \theta$ en $\pi(\theta)$ een a priori dichtheid, continu en positief op (0,1). Een standaard kansverdeling op [0,1] is de Beta(α, β) verdeling, deze is dus geschikt voor onze prior. Zij

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

de a priori dichtheid voor θ. Dan is de a posteriori dichtheid, gegeven X,

$$\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+r)\Gamma(\beta+(n-r))}\theta^{\alpha+r-1}(1-\theta)^{\beta+(n-r)-1},$$

waarbij $r = \sum_{i=1}^{n} X_i$, het aantal waarnemingen gelijk aan 1. De a posteriori is weer een beta verdeling, Beta$(\alpha + r, \beta + n - r)$. Een mooie schatter $\hat{\theta}$ voor θ is de verwachting van de a posteriori. Deze is bekend voor een beta verdeling, dus we krijgen

$$\hat{\theta} = E(\theta|X_1, X_2, \ldots, X_n) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\frac{r}{n}$$

Bovenstaand voorbeeld maakt de werkwijze van de bayesiaanse statistiek wat duidelijker. In het klassieke geval is namelijk een schatter als $\hat{\theta} = \frac{r}{n}$, toepasselijk. In gevallen waar we maar beschikking hebben tot een kleine steekproef is de Bayesiaanse aanpak soms handiger.

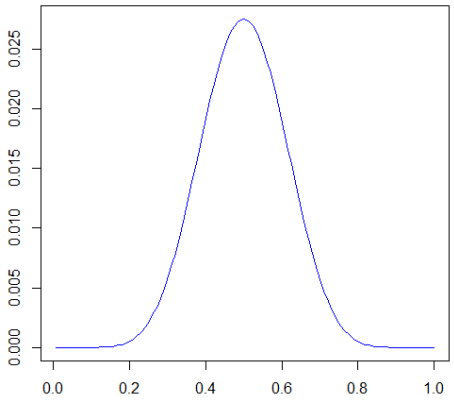
Voorbeeld 2.3. Stel we gooien een muntstuk op, waarbij we waarnemen: $X_i = 1$ als kop en $X_i = 0$ als munt. Dan $X_1, X_2, \ldots X_n$ i.i.d. Bernoulli met $P_\theta(X = 1) = \theta$.
Stel we hebben een kleine steekproef, bijvoorbeeld $n = 3$, waarbij $X_1 = X_2 = X_3 = 1$. In de klassieke statistiek: $\hat{\theta} = \frac{r}{n} = \frac{3}{3} = 1$.
In de bayesiaanse statistiek: We gooien met een muntstuk, dus met enige zekerheid kunnen we zeggen dat θ in de buurt van $\frac{1}{2}$ zal liggen. Dus a priori geven we dit ook mee, bijvoorbeeld met $\pi(\theta) = \text{Beta}(10, 10)$.

De a posteriori verdeling wordt dan gegeven door

$$f(\theta|x) = \text{Beta}(\alpha + r, \beta + n - r) = \text{Beta}(13, 10),$$

en

$$\hat{\theta} = E(\theta|X_i) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\frac{r}{n} = \frac{39}{69} = 0.57.$$



Figuur 1: Beta(10,10) dichtheid

In dit geval levert de Bayesiaanse statistiek een veel betere schatter op dan de klassieke statistiek. Om dit wiskundig hard te maken zou men kunnen kijken naar de Mean Square Error van beide methodes. De Bayesiaanse methode geeft natuurlijk niet altijd betere statistische resultaten, alleen het voorbeeld hierboven illustreert het idee.

3 Polya tree priors

3.1 De Polya tree

Zoals de naam al zegt zijn Polya tree priors, a priori verdelingen in de Bayesiaanse statistiek. Zowel de a priori als de a posteriori verdeling zijn kansverdelingen op de de verzameling van verdelingsfuncties. In het algemeen is dit vrij ingewikkeld. Polya tree priors zijn een speciaal geval, omdat ze geconstrueerd kunnen worden met een algoritme. We definiëren de Polya tree verdeling in stappen.
Zij $\{B_0, B_1\}$ een partitie van $\mathbb{R}$. Zij $\{B_{00}, B_{00}\}$ weer een partitie van B_0 en $\{B_{10}, B_{11}\}$ een partitie van B_1. We gaan verder op dezelfde manier. Zij

$$E^m = \{\epsilon = \epsilon_1 \dots \epsilon_m : \epsilon_k \in \{0,1\}\},\ m = 1, 2, \dots \text{ met } E^0 = \{\emptyset\},$$

en zij

$$E^* = \bigcup_{m=0}^{\infty} E^m.$$

Op het m^{e} niveau, wordt B_ϵ voor $\epsilon \in E^m$ verdeeld in $\{B_{\epsilon 0}, B_{\epsilon 1}\}$. Het is mogelijk dat $\{B_{\epsilon 0} = \emptyset, B_{\epsilon 1} = B_\epsilon\}$. Zij $\Pi = \{B_0, B_1, B_{00}, B_{01}, \dots\}$.
Definieer een rij van random variabelen $Y = \{Y_0, Y_{00}, Y_{10}, \dots\}$, en een rij van niet negatieve reeele parameters $A = \{\alpha_0, \alpha_1, \alpha_{00}, \alpha_{01}, \dots\}$, met $Y_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ voor $\epsilon \in E^*$. De random variabelen in Y bepalen de conditionele kansen in P als $Y_{\epsilon 0} = P(B_{\epsilon 0}|B_\epsilon)$.
De Polya tree verdeling wordt bepaald door de partities in Π en de Beta parameters in A.

Definitie 3.1. Een stochastische verdelingsfunctie F op $\mathbb{R}$ heeft een Polya tree verdeling met parameters (Π, A) als er random variabelen $Y = \{Y_0, Y_{00}, Y_{10}, \dots\}$ bestaan zdd.

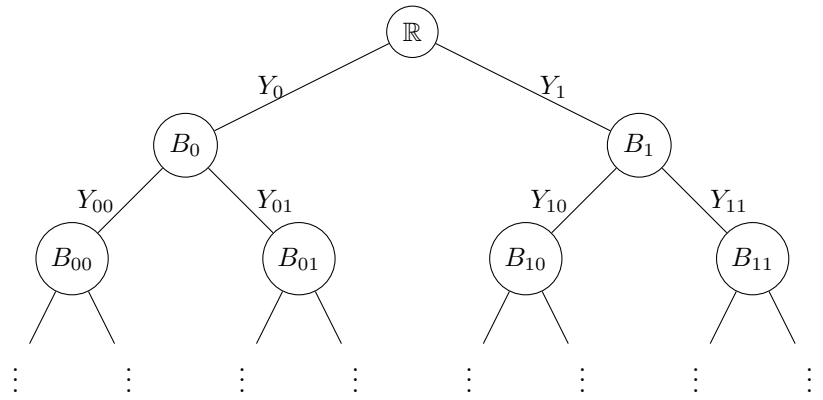
- (i) De random variabelen in Y zijn onafhankelijk;
- (ii) $\forall \epsilon \in E^*$, $Y_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon_0}, \alpha_{\epsilon_1})$;
- (iii) $\forall m = 1, 2, \dots$, en $\forall \epsilon \in E^m$,

$$F(B_\epsilon) = \left[\prod_{\substack{j=1 \\ \{\epsilon_j = 0\}}}^{m} Y_{\epsilon_1 \dots \epsilon_{j-1} 0}\right] \cdot \left[\prod_{\substack{j=1 \\ \{\epsilon_j = 1\}}}^{m} \left(1 - Y_{\epsilon_1 \dots \epsilon_{j-1} 0}\right)\right].$$

De andere manier om dit weer te geven is

$$F(B_\epsilon) = \prod_{j=1}^{m} (Y_{\epsilon_1 \dots \epsilon_j})^{(1-\epsilon_j)} (1 - Y_{\epsilon_1 \dots \epsilon_j})^{\epsilon_j}.$$

En we gebruiken de notatie $F \sim PT(\Pi, A)$.



Figuur 2: Polya tree verdeling

Het gewicht van de informatie op $F(B_{\epsilon 0}|B_\epsilon)$ wordt gegeven door $\alpha_{\epsilon 0} + \alpha_{\epsilon 1}$.
Zij $W_1, \ldots, W_n$ random variabelen van steekproefgrootte n uit F. De verdeling van observatie W_i is gegeven voor $\epsilon \in E^*$, als

$$\begin{aligned} P[W_i \in B_\epsilon] &= E[P[W_i \in B_\epsilon|F]] \\ &= E[F(B_\epsilon)] \\ &= E[F(B_{\epsilon_1})F(B_{\epsilon_1\epsilon_2}|B_{\epsilon_1}) \ldots F(B_\epsilon|B_{\epsilon_1 \ldots \epsilon_{m-1}})] \\ &= E[Y_{\epsilon_1} \cdot Y_{\epsilon_1\epsilon_2} \cdot \ldots \cdot Y_\epsilon] \\ &= \frac{\alpha_{\epsilon_1}}{\alpha_0 + \alpha_1} \cdots \frac{\alpha_{\epsilon_1 \ldots \epsilon_m}}{\alpha_{\epsilon_1 \ldots \epsilon_{m-1}0} + \alpha_{\epsilon_1 \ldots \epsilon_{m-1}1}}, \end{aligned} \tag{1}$$

waarbij de laatste gelijkheid volgt uit het feit dat alle Y_i onafhankelijk van elkaar zijn. We kunnen dit zien als een a priori schatter $\hat{F}_0(B_\epsilon)$ van $F(B_\epsilon)$.

3.2 De 'a posteriori' dichtheid

In deze paragraaf zullen we de handigheid van Polya tree priors gaan zien Het is namelijk zo dat de 'a posteriori' verdeling van een 'a priori' Polya tree verdeling, weer een Polya tree oplevert. Om dit inzichtelijk te maken is het handig om eerst het begrip Tailfree process uit te leggen.
Zij $\{\pi_m; m = 1, 2, \ldots\}$ een boom van meetbare partities van $\mathbb{R}$; oftewel, zij $\pi_1, \pi_2, \ldots$ een rij van meetbare partities zodanig dat π_{m+1} een verfijning is van π_m voor elke $m = 1, 2, \ldots$. Dit leidt tot de volgende definitie,

Definitie 3.2. De verdeling van een kansmaat P op $\mathbb{R}$ is tailfree ten opzichte van $\{\pi_m\}$ als er een niet negatieve random variabelen $\{Y_{m,B}; m = 1, 2 \ldots, B \in \pi_m\}$ bestaan zodanig dat

(1) De families $\{Y_{1,B}; B \in \pi_1\}, \{Y_{2,B}; B \in \pi_2\}, \ldots$ zijn onafhankelijk

(2) Voor elke $m = 1, 2, \ldots$, als $B_j \in \pi_j, j = 1, 2, \ldots, m$ en $B_1 \supset B_2 \supset \ldots \supset B_m$, dan $P(B_m) = \prod_{j=1}^{m} Y_{j,B_j}$.

Volgens Theorem 3.14 (Conjugacy) in [1] geldt

Stelling 3.3. *Als de verdeling van P tailfree is ten opzichte van $\{\pi_m\}$ en als $X_1, X_2, \ldots X_n$ een steekproef is uit P, dan is de posteriori verdeling van P gegeven $X_1, X_2, \ldots, X_n$ weer tailfree ten opzichte van $\{\pi_m\}$.*

Echter een belangrijk gevolg van deze stelling is.

Stelling 3.4. *De posteriori verdeling corresponderend met een observatie $X_1, \ldots, X_n$ van een verdeling P die a priori is opgesteld via een Polya tree process $PT(\Pi_m, \alpha_\epsilon)$ is een Polya tree process $PT(\Pi_m, \alpha^*_\epsilon)$, waarbij $\alpha^*_\epsilon := \alpha_\epsilon + \sum_{i=1}^{n} \#(X_i \in B_\epsilon)$, voor elke $\epsilon \in E^*$.*

Bewijs. Omdat de a posteriori tail free is volgens Stelling 3.3, is het genoeg om aan te tonen dat onder de posteriori verdeling de variabelen $Y_{\epsilon 0} = P(A_{\epsilon 0}|A_\epsilon)$ binnen elk gegeven niveau onafhankelijk beta variabelen met parameters $(\alpha^*_{\epsilon 0}, \alpha^*_{\epsilon 1})$ zijn. Zij $m \in \mathbb{N}$ en zij $Y = (Y_\epsilon : \epsilon \in \cup_{k \leq m} E^k)$, waarbij $Y_{\epsilon 1} = 1 - Y_{\epsilon 0}$. Volgens Theorem 3.13 in [1] geldt dat de verdeling van Y gegeven $X_1, \ldots, X_n$ hetzelfde is als de conditionele verdeling gegeven de vector die alle cellen telt, $N = (N_\epsilon : \epsilon \in E^m)$. De marginale likelihood van $Y_{\epsilon 0}$ is proportioneel tot $Y_{\epsilon 0}^{\alpha_{\epsilon 0}}(1-Y_{\epsilon 0})^{\alpha_{\epsilon 1}} = Y_{\epsilon 0}^{\alpha_{\epsilon 0}} Y_{\epsilon 1}^{\alpha_{\epsilon 1}}$, en deze variabelen zijn onafhankelijk. De conditionele likelihood van N gegeven Y is multinomiaal, met de kansen $P(A_\epsilon)$, gegeven door het product van alle $Y_{\epsilon_1} Y_{\epsilon_1 \epsilon_2} \ldots Y_{\epsilon_1 \ldots \epsilon_m}$. Dus de gezamelijke likelihood van (N, Y) is proportioneel tot

$$\prod_{|\epsilon|=m} (Y_{\epsilon_1} Y_{\epsilon_2 \epsilon_2} \ldots Y_{\epsilon_1 \ldots \epsilon_m})^{N_\epsilon} \prod_{|\epsilon| \leq m} Y_\epsilon^{\alpha_\epsilon} = \prod_{|\epsilon| \leq m} Y_\epsilon^{N_\epsilon + \alpha_\epsilon},$$

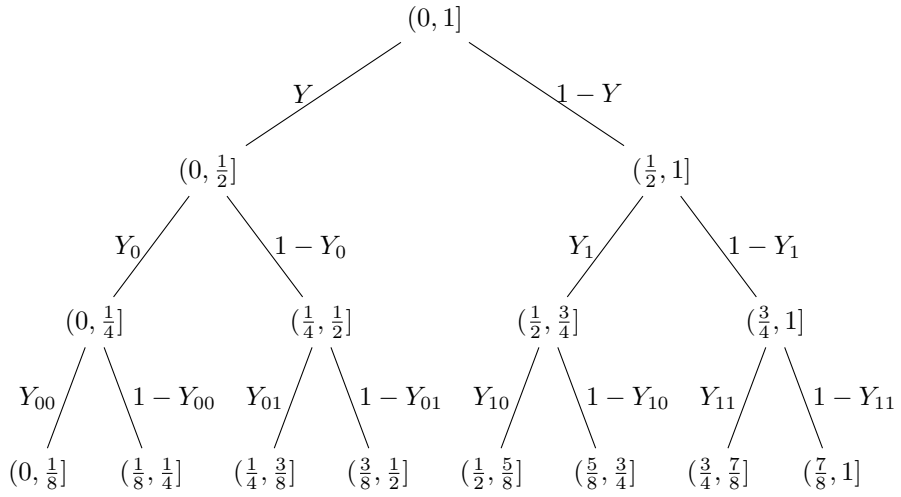
waarbij $N_\epsilon = \sum_{\delta \in E^{m-|\epsilon|}} N_{\epsilon\delta}$ het nummer van observaties aangeeft die vallen in A_ϵ. We zien dat het rechterlid een product is van beta-likelihoods met parameters α^*_ϵ. □

Uit Definitie 3.1 volgt dat voor een Polya tree verdeling op $\mathbb{R}$ we niet negatieve onafhankelijke random variabele Y_ϵ hebben. Ook geldt voor willekeurige m en als $B_{\epsilon_1} \supset B_{\epsilon_1 \epsilon_2} \supset \ldots \supset B_{\epsilon_1 \epsilon_2 \ldots \epsilon_m}$, dan

$$P(B_{\epsilon_1 \ldots \epsilon_m}) = Y_{\epsilon_1} \cdot Y_{\epsilon_1 \epsilon_2} \cdot \ldots \cdot Y_{\epsilon_1 \epsilon_2 \ldots \epsilon_m} = \prod_{j=1}^{m} Y_{\epsilon_1 \ldots \epsilon_j}.$$

Oftewel Polya tree verdelingen zijn tail free. Uit Stelling 3.3 volgt dan dat de posteriori weer een Polya tree is. Hierdoor kunnen we a posteriori verdelingen voor Polya tree verdelingen gemakkelijk bepalen. Een voorbeeld verduidelijkt dit idee.

Voorbeeld 3.5. We construeren een tailfree process op het interval (0,1] ten opzichte van $\{\pi_m\}$, waarbij π_m de verzamling is van intervallen van lengte $1/2^m$, $\pi_m = \{((i-1)/2^m, i/2^m];\ i = 1, \ldots, 2^m\}$.



Figuur 3: Een Polya tree voor het interval (0,1] met dyadische interallen

We gebruiken hier een makkelijkere notatie voor de variabelen $Y_{m,B}$ uit Definitie 3.2. Zij $.\epsilon_1\epsilon_2\ldots\epsilon_m$ de binaire code van een dyadische breuk $\sum_1^m \epsilon_j 2^{-j}$, waarbij elke ϵ_j gelijk aan nul of één is.
Als $B \in \pi_m$ van de vorm $(.\epsilon_1\ldots\epsilon_m, .\epsilon_1\ldots\epsilon_m + 2^{-m}]$ is, dan gebruiken we $Y_{\epsilon_1\ldots\epsilon_m}$ als $\epsilon_m = 0$, en $1 - Y_{\epsilon_1\ldots\epsilon_m}$ als $\epsilon_m = 1$, om $Y_{m,B}$ aan te geven.
Dan geldt dat $P(B)$ het product is van alle variabelen die horen bij het pad in de boom van $(0,1]$ tot B, zodat

$$P(B) = \left[\prod_{\substack{j=1\\ \{\epsilon_j=0\}}}^{m} Y_{\epsilon_1\ldots\epsilon_{j-1}}\right] \cdot \left[\prod_{\substack{j=1\\ \{\epsilon_j=1\}}}^{m} \left(1 - Y_{\epsilon_1\ldots\epsilon_{j-1}}\right)\right].$$

Bijvoorbeeld, $P((\frac{3}{8},\frac{1}{2}]) = Y(1-Y_0)(1-Y_{01})$. De onafhankelijkheid uit Definitie 3.2 voor de Y-variabelen is in principe alleen nodig tussen de rijen in Figuur 3.
Als we alle Y-variabelen onafhankelijk kiezen en $Y_\epsilon \in \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$, dan geldt voor de posteriori verdelingen van de variabelen Y_ϵ, gegeven een dataset $X = \{X_1, X_2, \ldots, X_n\}$ uit P, dat deze weer dezelfde vorm hebben, namelijk onafhankelijke beta verdelingen.
De posteriori verdeling van $Y_{\epsilon_1\ldots\epsilon_{m-1}}$ gegeven $X_1, X_2, \ldots, X_n$ is $\text{Beta}(\alpha_{\epsilon_1\ldots\epsilon_{m-1}0} + M, \alpha_{\epsilon_1\ldots\epsilon_{m-1}1} + N)$, waarbij M het aan X_i is dat tussen $(.\epsilon_1\ldots\epsilon_{m-1}0, .\epsilon_1\ldots\epsilon_{m-1}1]$ valt, en N is het aantal X_i dat in $(.\epsilon_1\ldots\epsilon_{m-1}1, .\epsilon_1\ldots\epsilon_{m-1}1 + 2^{-m}]$ valt.
Bij Polya tree verdelingen zoals in Definitie 3.1 gedefinieerd geldt dus dat de a posteriori van dezelfde vorm is als de a priori verdeling. In (1) hebben we al een schatter gevonden voor de a priori verdeling. De a posteriori verdeling, na

een waarneming $X = \{X_1, X_2, \ldots, X_n\}$ wordt verkregen door

$$F|\{X_i\} \sim PT(\Pi, A(X_1, \ldots, X_n)),$$

waarbij

$$\alpha_\epsilon(X_1, \ldots, X_n) = \alpha_\epsilon + n_\epsilon, \text{ en } \quad n_\epsilon = \#\{X_i \in B_\epsilon\}.$$

Iedere waarneming X_i zorgt er dus voor dat alle parameters die corresponderen met de verzamelingen $X \subseteq B_\epsilon$ worden opgehoogd met waarde 1. Een a posteriori schatter voor de kans op B_ϵ, $\epsilon \in E^m$, is dus

$$\begin{aligned} \hat{P}(B_\epsilon) &= E[P(B_\epsilon)|X_1, \ldots, X_n] \\ &= \frac{\alpha_{\epsilon_1}+n_{\epsilon_1}}{\alpha_0+\alpha_1+n} \cdots \frac{\alpha_{\epsilon_1\ldots\epsilon_m}+n_{\epsilon_1\ldots\epsilon_m}}{\alpha_{\epsilon_1\ldots\epsilon_{m-1}0}+\alpha_{\epsilon_1\ldots\epsilon_{m-1}1}+n_{\epsilon_1\ldots\epsilon_{m-1}}}. \end{aligned}$$

Voorbeeld 3.6. Zij $m = 2$ en $\Pi = \{B_0, B_1, B_{00}, B_{01}, B_{10}, B_{11}\}$ en stel dat we de Polya tree parametriseren met $\alpha_0 = 1, \alpha_1 = 3, \alpha_{00} = 1, \alpha_{01} = 1$, $\alpha_{10} = 3, \alpha_{11} = 1$. De 'a priori' schatter van $P(B_0)$ wordt gegegeven door $\frac{1}{4}$ en de a priori schatters van $P(B_{00}|B_0)$ en $P(B_{10}|B_1)$ zijn $\frac{1}{2}$ en $\frac{3}{4}$ respectievelijk. Dus de kansvector $[\hat{P}_0(B_{00}), \hat{P}_0(B_{01}), \hat{P}_0(B_{10}), \hat{P}_0(B_{11})]$ wordt a priori gegeven door

$$[(\frac{1}{4})(\frac{1}{2}), (\frac{1}{4})(\frac{1}{2}), (\frac{3}{4})(\frac{3}{4}), (\frac{3}{4})(\frac{1}{4})] = [.125, .125, .5625, .1875].$$

Stel we hebben twee observaties $W_1 \in B_{01}$ en $W_2 \in B_{11}$. De nieuwe parameters worden dan gegeven door $\alpha_{01} = 2$ en $\alpha_{11} = 2$ en de schatter voor de nieuwe kansvector wordt dan gegeven door

$$\begin{aligned} [\hat{P}(B_{00}), \hat{P}(B_{01}), \hat{P}(B_{10}), \hat{P}(B_{11})] &= [(\tfrac{1}{3})(\tfrac{1}{3}), (\tfrac{1}{3})(\tfrac{2}{3}), (\tfrac{2}{3})(\tfrac{3}{5}), (\tfrac{2}{3})(\tfrac{2}{5})] \\ &= [.111, .222, .4, .266] \end{aligned}$$

4 Simulaties

Het is handig om inzicht te verkrijgen in het gebruik van Polya tree priors doormiddel van R. Om in R Polya tree priors te gebruiken, gebruiken we de volgende functie.

```
1  R<-function(bin){
2    l<-length(bin)
3    rn=0
4    for (i in 1:l) {rn<-rn+2^i}
5    r<-(rn-2^l+1):rn
6    lr<-length(r)
7    for (j in 1:l) {
8      if (bin[j]==1) {
9        r<-r[-c(1:(lr/2))]
10     }
11     else {
12       r<-r[-c((lr/2+1):lr)]
13     }
14     lr<-length(r)
15   }
16   return(r)
17 }
```

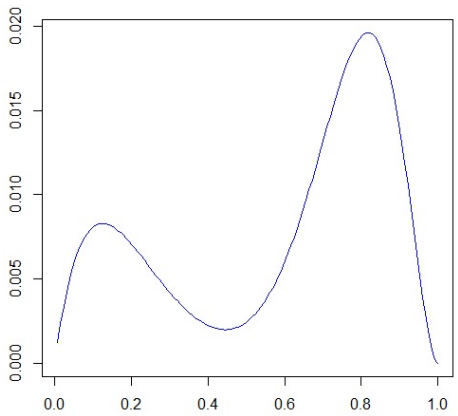
Deze functie geeft voor een binair getal de juiste positie in de vector Y in onderstaand programma. Onderstaande code genereert een vector Pb in regel 65-71. Dit is de posteriori mean van de Polya tree verdeling, onze Bayes schatter $\hat{P}$. De vector is gedefinieerd op de verzameling van intervallen van lengte $1/2^m$, $\{B_\epsilon | \epsilon \in E^m\}$. Pb is de vector bestaande uit de waarden $\hat{P}(B_\epsilon) = E[P(B_\epsilon)|X_1, \ldots, X_n]$, waarbij Pb[i] $= \hat{P}(((i-1)/2^m, i/2^m])$. Verder staat in de code als commentaar erbij geschreven welke stap we van het algoritme we doen, voor het simuleren van de Polya tree.

```
setwd("/Users/Aschwin/Desktop/School/3e Jaar/Bachelor Seminarium/R")
source("R.R")

m<-7
B<-seq(from = 0, to = 1, by = (1/(2^m))) ##diadische intervallen op [0,1]
u<-(1:2^m)/2^m

##Data
v<-rbinom(10000,rep(1,20),0.3)
data<-v*rbeta(10000,2,8)+(1-v)*rbeta(10000,10,3)

##Parameters
A<-vector('numeric',254)+1

##Makegrid matrix met alle binaire getallen van grootte m
M<-matrix(data = 0, nrow = 2^m, ncol = m)
for (i in 1:m){
  grote<-2^(i-1)
  stap<-2^i
  teller<-1
  while (M[2^m,m+1-i]!=1){
    M[(teller+grote):(teller+grote+grote-1),m+1-i]<-1
    teller<-teller+stap
  }
}

##--------------code---------------

Y<-vector('numeric',length(A))            ##splitting variables
##FillY:
for (i in 1:length(A)) {
  if (i%%2==0) {
    Y[i-1]<-mean(rbeta(9999,A[i-1],A[i]))
    Y[i]<-1-Y[i-1]
  }
}

##fillPb
Pb<-1+vector('numeric',2^m)               ##Kans op B_i
for (i in 1:(2^m)){
  for (j in 0:(m-1)){
    Pb[i]<-Pb[i]*Y[R(M[i,1:(m-j)])]
  }
}

##Update parameters
for (i in 1:length(data)) {
  for (j in 1:(length(B)-1)) {
    if (B[j]<data[i] & data[i]<B[j+1]) {
      for (t in 0:(m-1)) {
        A[R(M[j,1:(m-t)])]<-A[R(M[j,1:(m-t)])]+1
      }
    }
  }
}

##Posteriori----------------------
##FillY:
for (i in 1:length(A)) {
  if (i%%2==0) {
    Y[i-1]<-mean(rbeta(9999,A[i-1],A[i]))
    Y[i]<-1-Y[i-1]
  }
}
##fillPb
Pb<-1+vector('numeric',2^m)
for (i in 1:(2^m)){
  for (j in 0:(m-1)){
    Pb[i]<-Pb[i]*Y[R(M[i,1:(m-j)])]
  }
}
```

We kunnen in deze code eenvoudig veranderingen aanbrengen van bijvoorbeeld m, de data of de parameters. In bovenstaand geval hebben we $m = 7$ gekozen en kiezen we voor alle parameters α de waarde 1. We geven a priori dus geen informatie mee. Als data gebruiken we 10.000 trekkingen uit een gemengde beta verdeling, waarvoor we de kans 0.3 kiezen voor een trekking uit beta(2,8) en kans 0.7 voor een trekking uit beta(10,3). De werkelijke verdeling van deze functie ziet er als volgt uit. R geeft voor

```
plot(u,(0.3*dbeta(u,2,8)+0.7*dbeta(u,10,3))/length(u),type="l",col='red',xlab
    ='',ylab='')
```

het volgende plaatje

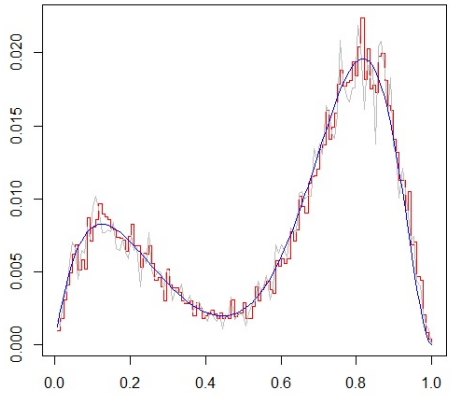


Figuur 4: 0.3*Beta(2,8)+0.7*Beta(10,3) dichtheid

Als we vervolgens onze code runnen, dan zien we dat de a posteriori zich naar de werkelijke verdeling vormt. Na

```
plot(u,Pb,type='s',col='red',xlab='',ylab='')
lines(u,(0.3*dbeta(u,2,8)+0.7*dbeta(u,10,3))/length(u),type='l',col='blue')
lines(u,Pb1,type='l',col='grey')
```

krijgen we: (zie volgende pagina)

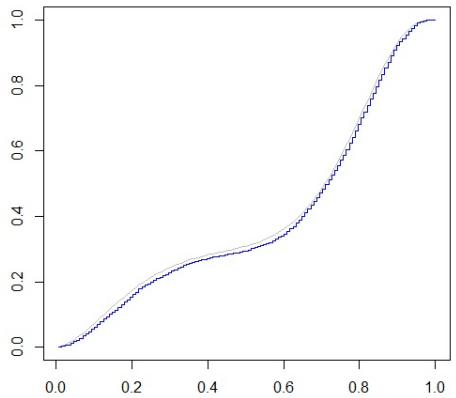


Figuur 5: De a posteriori mean dichtheid

We zien hier dat de posteriori mean (de rode lijn) zich redelijk naar de werkelijke verdeling (de blauwe lijn) vormt. De grijze lijn stelt één trekking uit de posteriori verdeling voor. Om de kansverdeling van de a posteriori mean te bekijken, runnen we de volgende code:

```
1 plot(u,(cumsum(0.3*dbeta(u,2,8)+0.7*dbeta(u,10,3)))/length(u),type='l',col='
      grey',xlab='',ylab='')
2 lines(u,cumsum(Pb),type='s',col='blue')
```

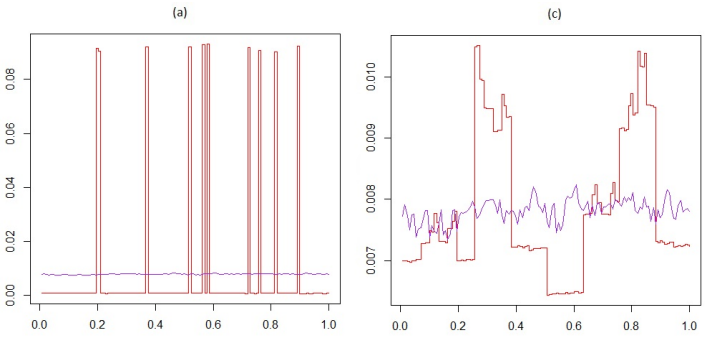
en krijgen we het volgende plaatje.



Figuur 6: 0.3*Beta(2,8)+0.7*Beta(10,3) kansverdeling en a posteriori kansverdeling

4.1 Keuze van de α_ϵ's

Bij het gebruik van Polya tree priors kunnen we dus van te voren informatie meegeven van wat we al weten over de parameter die we gaan schatten. Die informatie geef je mee aan de hand van de parameters $A = \{\alpha_0, \alpha_1, \ldots, \}$. De vraag is natuurlijk welke waarde geef je deze α_ϵ mee? Volgens Lavine zijn er drie dingen waarmee we rekening moeten houden bij het kiezen van de α_ϵ's. Als eerste zorgen de α_ϵ ervoor hoe snel de a posteriori verdeling verandert vanaf de a priori verdeling. Als de α_ϵ's groot worden gekozen dan zal de verdeling van de a posteriori verdeling van een onbekende parameter Θ, $f(\Theta|X_1, \ldots, X_n)$, na de waarnemening X, dichterbij de a priori verdeling van deze parameter liggen. Daarentegen geldt dat als de α_ϵ's klein worden gekozen dat dan de a posteriori verdeling dichterbij de verdeling van de waarneming ligt. Dit is goed te zien in Figuur 7.



Figuur 7: De priori mean (paars) en de posteriori mean (rood)
voor geval (a) $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$ en geval (c) $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$. Dataset van grootte 10.

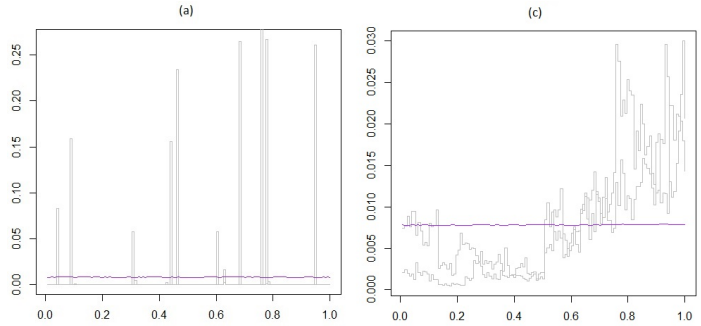
Het tweede wat overwegen moet worden bij het kiezen van de α_ϵ's is dat de α_ϵ's bepalen hoe glad $F \sim PT(\Pi, A)$ is. Als we $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$ groot kiezen, dan heeft $F(B_{\epsilon 0}|B_\epsilon)$ een verdeling geconcentreerd rondom de waarde $\frac{1}{2}$. Dit zorgt ervoor zorgt dat F continu is, omdat met grote kans geldt dat $F(B_{\epsilon 0})$ en $F(B_{\epsilon 1})$ ongeveer aan elkaar gelijk zijn. Volgens Ferguson is het nuttig om drie verschillende gevallen te beschouwen:

(a) $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$. Dit geeft een Dirichlet proces, want $\alpha_{\epsilon 0} = \alpha_{\epsilon 1} \quad \forall \epsilon \in E^*$. Met deze voorwaarde voor de parameters α is F discreet met kans 1. [5]

(b) $\alpha_{\epsilon_1 \ldots \epsilon_m} = 1$, F is continu singulier met kans 1.

(c) $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$, F is absoluut continu met kans 1.

Een Beta(α_0, α_1) priori heeft verwachting $\frac{\alpha_0}{\alpha_0 + \alpha_1}$ en variantie $\frac{1}{\alpha_0 + \alpha_1}$. Als $\alpha_0 + \alpha_1$ groot is dan is de variantie kleiner en is de prior dus meer geconcentreerd rond zijn verwachting. Als $\alpha_0 + \alpha_1$ klein is, dan is de variantie dus groter en zal één trekking uit de priori verdeling minder concentreerd zijn rond de prior mean. Dit is goed te zien in Figuur 8, waarbij we de prior mean samen met twee trekkingen uit de priori verdeling, voor de twee verschillende gevallen (a) α_ϵ klein, en (c) α_ϵ groot, zien. Updaten met de data helpt wel, maar minder zoals je goed kunt zien in Figuur 7.



Figuur 8: De priori mean (paars) met twee trekkingen uit de priori verdeling (grijs) voor twee verschillende keuze in parameters.

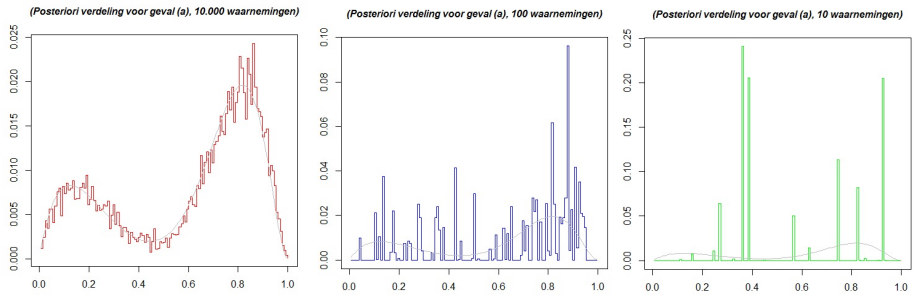
Laten we eens kijken wat er gebeurd als we in de bovenstaande code de parameters veranderen. We beginnen weer met een dataset van 10.000 uit de gemengde beta verdeling.

```
##Data
v<-rbinom(10000,rep(1,20),0.3)
data<-v*rbeta(10000,2,8)+(1-v)*rbeta(10000,10,3)
```

Voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$, definieren we de parameters als volgt:

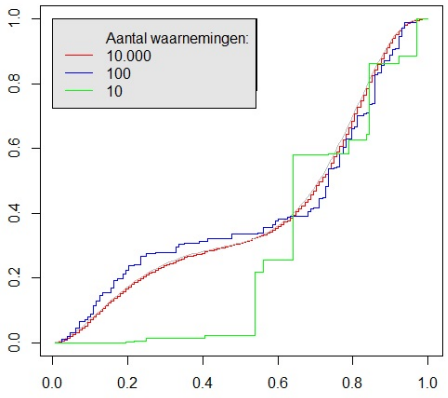
```
##Parameters
lengtha<-0
for (i in 1:m){
  lengtha<-lengtha+2^i
}
A<-vector('numeric',lengtha)
n<-1
t<-2
for (i in 1:lengtha){
  A[i]<-1/2^n
  if (i>=t){
    n<-n+1
    t<-t+2^n
  }
}
```

Als we de posteriori dichtheden plotten met als parameters $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$, krijgen we drie verschillende dichtheden voor drie verschillende groottes van de datasets.



Figuur 9: De a posteriori dichtheden voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$

De kansverdeling van de posteriori dichtheid ziet er als volgt uit:

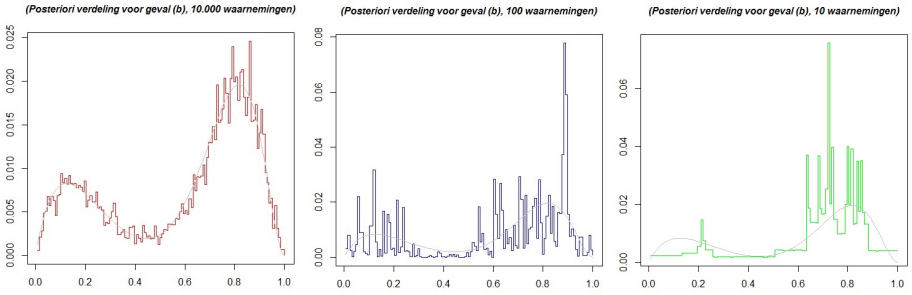


Figuur 10: De kansverdelingen voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$

Voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = 1$ definieren we de parameters als:

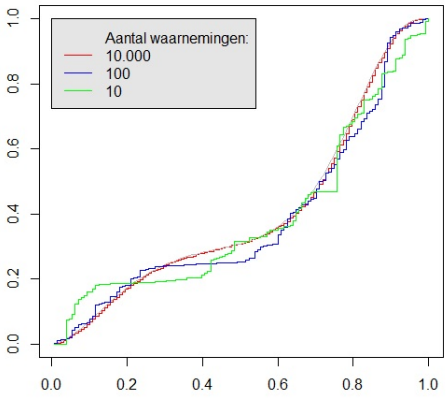
```
##Parameters
lengtha<-0
for (i in 1:m){
  lengtha<-lengtha+2^i
}
A<-vector('numeric',lengtha)+1
```

We krijgen dan de volgende drie dichtheden voor de verschillende groottes van de datasets.



Figuur 11: De a posteriori dichtheden voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = 1$

De kansverdeling van de posteriori dichtheid ziet er als volgt uit:

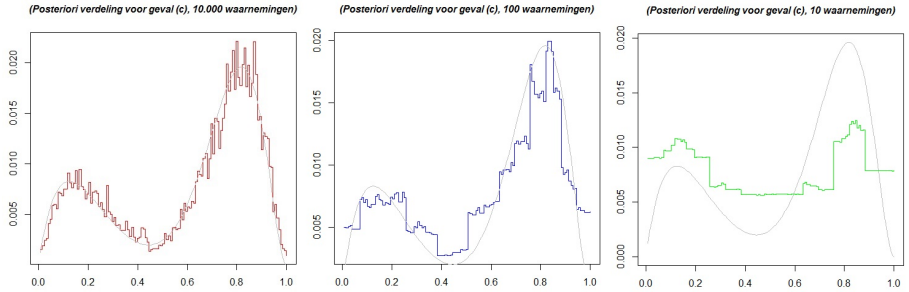


Figuur 12: De kansverdelingen voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = 1$

Voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$ definieren we de parameters als:

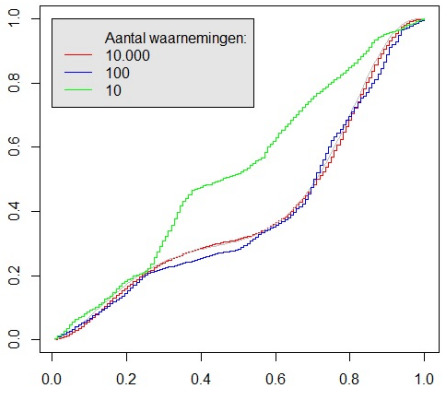
```
##Parameters
lengtha<-0
for (i in 1:m){
  lengtha<-lengtha+2^i
}
n<-1
t<-2
A<-vector('numeric',lengtha)
for (i in 1:lengtha){
  A[i]<-n^2
  if (i>=t){
    n<-n+1
    t<-t+2^n
  }
}
```

en krijgen we weer de volgende drie dichtheden



Figuur 13: De a posteriori dichtheden voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$

De a posteriori verdelingen voor de verschillende groottes van de datasets zijn:



Figuur 14: De kansverdelingen voor $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$

Uit Figuur 9, 11 en 13 kunnen we opmaken, dat wat Ferguson al bewezen heeft, wel degelijk klopt. Voor grote datasets maakt het niet veel uit hoe je de parameters kiest, omdat de invloed van de dataset voor de verschillende intervallen de invloed van de parameters overheerst. Voor een kleinere dataset van bijvoorbeeld 10, zien we dat we in geval (a), $\alpha_{\epsilon_1 \ldots \epsilon_m} = 2^{-m}$, een discrete dichtheid krijgen, terwijl dit voor geval (c), $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$, niet zo is.
Het is dus belangrijk om van tevoren goed te kiezen wat soort parameters je voor je Polya tree wil gebruiken. Heb je veel vertrouwen in je experts bij het modelleren van een probleem en is een continue dichtheid gewenst, kies dan de parameters groot genoeg. Heb je van tevoren weinig informatie over de vorm van je verdeling en maakt het je weinig uit dat je een discrete dichtheid krijgt, kies dan de parameters laag, zodat de posteriori dichtheid dichter bij de echte dichtheid ligt die je probeert te schatten.

5 Tot slot

Wat verder nog onderzocht kan worden is of de splits met andere dan Beta variabelen betere statistische resultaten oplevert. Ook zouden simulaties op een ander interval dan $(0, 1)$ meer inzicht kunnen geven op het kiezen van de parameters.
Het bekijken van de fout van de a posteriori ten opzichte van de werkelijke verdeling zou ook interresant kunnen zijn. Je zou kunnen proberen doormiddel van het varieren van de parameters of andere splits variabelen gebruiken deze fout te kunnen minimaliseren. Deze fout zou je weer kunnen vergelijken met andere methoden voor het schatten van een onbekende verdeling om te kijken welke methode het beste is.

Referenties

[1] Subhashis Ghosal, Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Wiskunde, Leiden.

[2] Bas Kleijn, Aad van der Vaart, Harry van Zanten. *Lectures on Nonparametric Bayesian Statistics*. Wiskunde, Leiden, 9-3-2013.

[3] Thomas S. Ferguson. *Prior distributions on spaces of probability measures*. University of California, 1974

[4] Michael Lavine. *Some aspects of Polya tree distributions for statistical modelling*. Duke University, 1992

[5] David Blackwell. *Discreteness of Ferguson selections*. University of California, Berkeley, 1973