Les 13 Gaussian mixture modellen

We zullen in deze les op een paar vragen ingaan die bij de beschrijving van waarnemingen of processen door probabilistische modellen een belangrijke rol spelen. Hierbij zullen we vooral naar de Gauss-verdeling (vaak normaalverdeling genoemd) kijken. Hier zijn (minstens) twee redenen waarom de Gauss-verdeling zo'n vooraanstaande functie inneemt:

(1) De centrale limietstelling zegt (in het grof) dat de som van bijna willekeurige stochasten tegen een Gauss-verdeling convergeert.

(2) De Gauss-functie waarmee we de Gauss-verdeling beschrijven heeft goede eigenschappen die een analytische behandeling van verschillende vragen mogelijk maken.

De aspecten waarop we in deze les vooral zullen ingaan zijn:

- Het uitbreiden van de 1-dimensionale Gauss-verdeling tot Gauss-verdelingen van n-dimensionale vectoren.
- Het combineren van verschillende Gauss-verdelingen in een zogeheten *Gaussian mixture model*.
- Het schatten van parameters, vooral voor Gaussian mixture modellen.

13.1 Meerdimensionale Gauss-verdelingen

Continue kansverdelingen beschrijven we meestal door een dichtheidsfunctie $f(x)$, waarbij we dan voor een stochast X met deze verdeling de kans $p(X \leq x)$ berekenen door $p(X \leq x) = \int_{-\infty}^{x} f(u)\, du$.

Voor een normaalverdeelde stochast X hebben we de dichtheidsfunctie $f(x)$ al eerder bekeken, dit is de Gauss functie

$$f(x) := \frac{1}{\sqrt{2\pi}\,\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

waarbij $\mu = E[X]$ de verwachtingswaarde en $\sigma^2 = E[(X-\mu)^2]$ de variantie van X is. Een Gauss-verdeling met parameters μ en σ noteren we ook als $\mathcal{N}(\mu, \sigma)$.

Vaak hebben we het echter met waarnemingen te maken die niet door een enkele waarde maar door een vector van waarden beschreven wordt. We kunnen hier bijvoorbeeld aan de gemiddelde intensiteiten voor verschillende intervallen van frequenties denken die in de spraakherkenning als kenmerken voor de verschillende klinkers en medeklinkers gebruikt worden. Maar ook algemeen is het handig, waarnemingen door vectoren met verschillende kenmerken te beschrijven en het is dus voor de hand liggend, naar kansverdelingen voor vectoren te kijken.

Als de componenten van de vectoren onafhankelijke stochasten zijn, is dit makkelijk, want dan is de gemeenschappelijke verdeling van de verschillende componenten gewoon het product van de aparte verdelingen:

Stel we hebben n onafhankelijke stochasten $X_1, \ldots, X_n$ die we door de vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ beschrijven, waarbij X_i normaalverdeeld met kansverdeling $\mathcal{N}(\mu_i, \sigma_i)$ is. Dan heeft de kansverdeling $p(\boldsymbol{X} = \boldsymbol{x}) = p(\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix})$ van de stochast $\boldsymbol{X}$ de dichtheidsfunctie

$$\begin{aligned} f(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma_i} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right). \end{aligned}$$

De vraag is nu, wat er gebeurt als de componenten niet meer onafhankelijk zijn. Voor de verwachtingswaarde maakt dat niets uit, want die kunnen we nog steeds componentsgewijs berekenen:

Stel we hebben een stochast $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ en we definiëren de stochasten Y en Z door $Y = \begin{pmatrix} X_1 \\ 0 \end{pmatrix}$ en $Z = \begin{pmatrix} 0 \\ X_2 \end{pmatrix}$, dan is $\boldsymbol{X} = Y + Z$ en we hebben $E[Y] = \begin{pmatrix} E[X_1] \\ 0 \end{pmatrix}$ en $E[Z] = \begin{pmatrix} 0 \\ E[X_2] \end{pmatrix}$. Omdat we verwachtingswaarden bij elkaar kunnen optellen, geldt

$$E[\boldsymbol{X}] = E[Y + Z] = E[Y] + E[Z] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix}.$$

Het probleem ligt echter in de variantie, want die mogen we niet zo maar als som van de enkele componenten berekenen. Voor twee stochasten Y en Z geldt namelijk dat $Var(Y + Z) = Var(Y) + Var(Z) + 2Cov(Y, Z)$, waarbij de *covariantie* $Cov(Y, Z)$ gedefinieerd is door

$$Cov(Y, Z) = E[(Y - E[Y]) \cdot (Z - E[Z])].$$

Als Y en Z onafhankelijke stochasten zijn, geldt $Cov(Y, Z) = 0$ en dan is inderdaad $Var(Y + Z) = Var(Y) + Var(Z)$, maar in het algemeen is dat niet zo. Merk op dat twee stochasten wel covariantie 0 kunnen hebben, zonder onafhankelijk te zijn.

Voor een kansverdeling van een n-dimensionale stochast $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ met verwachtingswaarde $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$ definiëren we nu de *covariantie matrix* Σ als

matrix van de covarianties van de enkele componenten, dus:

$$\Sigma := (\sigma_{ij}) \text{ met } \sigma_{ij} = E[(X_i - \mu_i) \cdot (X_j - \mu_j)].$$

Merk op dat Σ een symmetrische matrix is, dus dat $\Sigma_{ij} = \Sigma_{ji}$ en dat de diagonaalelementen van Σ juist de varianties van de enkele componenten X_i zijn, want $\sigma_{ii} = E[(X_i - \mu_i)^2] = Var(X_i) = \sigma_i^2$.

De grap is nu dat we met behulp van de covariantie matrix ook de variantie van een willekeurige lineaire combinatie van de componenten X_i uit kunnen rekenen, namelijk als een soort inproduct. Stel we hebben de stochast $Y := a_1X_1 + \ldots + a_nX_n$, dan heeft Y verwachtingswaarde

$$E[Y] = a_1E[X_1] + \ldots + a_nE[X_n] = a_1\mu_1 + \ldots + a_n\mu_n.$$

Voor de variantie van Y geldt dan:

$$\begin{aligned}
Var(Y) &= E[(Y - E[Y])^2] \\
&= E[((a_1X_1 + \ldots + a_nX_n) - (a_1\mu_1 + \ldots + a_n\mu_n))^2] \\
&= E[((a_1(X_1 - \mu_1) + \ldots + a_n(X_n - \mu_n))^2] \\
&= E[\sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j(X_i - \mu_i)(X_j - \mu_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_jE[(X_i - \mu_i)(X_j - \mu_j)] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\Sigma_{ij} \\
&= \begin{pmatrix} a_1 & \ldots & a_n \end{pmatrix} \cdot \Sigma \cdot \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}.
\end{aligned}$$

We berekenen dus de variantie van een lineaire combinatie van de stochasten $X_1, \ldots, X_n$ met coördinatenvector $(a_1, \ldots, a_n)$ als inproduct van de coördinatenvector met zich zelf, waarbij we de covariantie matrix Σ als Gram matrix van het inproduct beschouwen.

Tegelijkertijd zien we zo ook in dat de covariantie matrix Σ positief definiet is, dus dat $v^{tr} \cdot \Sigma \cdot v \geq 0$ voor elke vector v, want varianties zijn als verwachtingswaarden van de kwadraten $(X - \mu)^2$ altijd positief.

Met behulp van de covariantie matrix Σ kunnen we nu ook de algemene vorm van de n-dimensionale Gauss-verdeling aangeven, dus ook voor het geval dat de componenten X_i van de stochast $\boldsymbol{X}$ niet onafhankelijk zijn. De dichtheidsfunctie is gegeven door

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\ \det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{tr}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right).$$

Voor onafhankelijke componenten X_i is deze formule precies hetzelfde wat we eerder al hadden, want voor de covariantie matrix bij onafhankelijke com-

ponenten geldt

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots \\ 0 & \ddots & \\ \vdots & & \sigma_n^2 \end{pmatrix}, \quad \det(\Sigma)^{\frac{1}{2}} = \prod_{i=1}^{n} \sigma_i, \quad \Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots \\ 0 & \ddots & \\ \vdots & & \frac{1}{\sigma_n^2} \end{pmatrix}$$

en dus $-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{tr}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = -\sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{2\sigma_i^2}$.

Om te zien dat de formule voor de Gauss-verdeling ook voor een algemene covariantie matrix geldt, hebben we een stelling uit de Lineaire Algebra nodig:

Stelling: Als Σ een symmetrische, positief definiete matrix is, dan bestaat er een orthogonale basis transformatie T (d.w.z. een matrix T met $T^{tr} = T^{-1}$) zo dat $T^{tr}\Sigma T$ een diagonaalmatrix is.

We kunnen de kolommen van T als nieuwe stochasten X_i' zien, namelijk de i-de kolom als de stochast $X_i' := T_{1i}X_1 + T_{2i}X_2 + \ldots + T_{ni}X_n$, dan geeft de matrix $T^{tr}\Sigma T$ juist de covarianties $Cov(X_i', X_j')$ van de nieuwe stochasten. Maar omdat $T^{tr}\Sigma T = D$ een diagonaalmatrix is, betekent dit dat de nieuwe stochasten paarsgewijs covariantie 0 hebben.

Maar we kunnen ook een vector $\boldsymbol{x}$ met betrekking tot de nieuwe basis van stochasten uitdrukken, namelijk als $\boldsymbol{x}' = T^{-1}\boldsymbol{x}$. Net zo moeten we ook de verwachtingswaarde $\boldsymbol{\mu}$ op de nieuwe basis transformeren, als $\boldsymbol{\mu}' = T^{-1}\boldsymbol{\mu}$. Dan geldt wegens $T^{-1} = T^{tr}$:

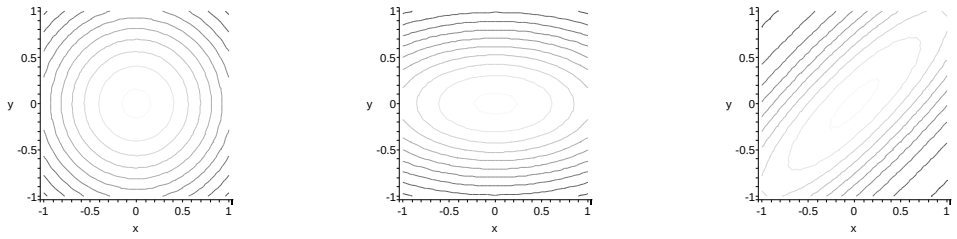
$$\begin{aligned}(\boldsymbol{x}'-\boldsymbol{\mu}')^{tr}D^{-1}(\boldsymbol{x}'-\boldsymbol{\mu}') &= ((\boldsymbol{x}-\boldsymbol{\mu})^{tr}T^{-tr})(T^{-1}\Sigma^{-1}T^{-tr})(T^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^{tr}(T^{-tr}T^{-1})\Sigma^{-1}(T^{-tr}T^{-1})(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^{tr}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}).\end{aligned}$$

Omdat we weten dat onze formule voor de Gauss-verdeling voor het geval van de getransformeerde coördinaten $\boldsymbol{x}'$ met diagonale covariantie matrix D geldt, volgt hieruit dat de formule ook voor de algemene covariantie matrix Σ geldt.

We zouden de transformatie T op nieuwe stochasten met onderling covariantie 0 in principe zo kunnen interpreteren, dat we maar per ongeluk de verkeerde stochasten hebben gekozen die niet onafhankelijk zijn, maar dat we dit door een orthogonale transformatie recht kunnen zetten. Maar we hebben alleen maar ervoor gezorgd dat de nieuwe stochasten covariantie 0 hebben, ze hoeven nog steeds niet onafhankelijk te zijn (en zijn dit in de praktijk ook vaak niet).

Het effect van de transformatie T is in een 2-dimensionaal voorbeeld makkelijk te zien, de orthogonale matrix T is in dit geval gewoon een draaiing van het coördinaat-stelsel. De drie plaatjes in Figuur II.8 geven krommen van constante dichtheid voor de Gauss-verdelingen met verwachtingswaarde $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ en covariantie matrices

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 2 & \sqrt{3} \\ \sqrt{3} & 2 \end{pmatrix}.$$



Figuur II.8: Normaalverdelingen in dimensie 2 met verschillende covariantie matrices.

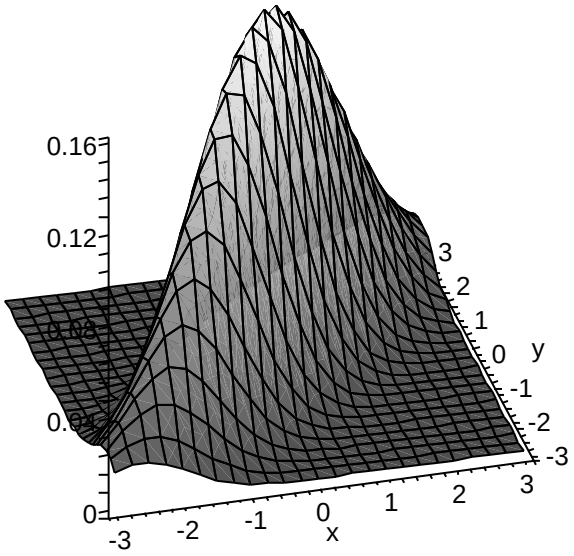
In het eerste geval is de variantie voor de twee componenten hetzelfde, daarom zijn de krommen van constante dichtheid cirkels. In het tweede geval heeft de stochast op de x-as variantie 2 en de stochast op de y-as variantie $\frac{1}{2}$, en we zien ook dat de spreiding in de richting van de x-as groter is dan in de richting van de y-as. In dit geval zijn de krommen van constante dichtheid ellipsen. In het derde plaatje is de covariantie matrix niet meer diagonaal, maar we zien dat we met een rotatie om 45° weer in dezelfde situatie als bij het tweede plaatje terecht komen. De transformatie die bij de rotatie om 45° hoort is de matrix $T = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ en we hebben

$$T^{tr}\Sigma_3 T = \begin{pmatrix} 2+\sqrt{3} & 0 \\ 0 & 2-\sqrt{3} \end{pmatrix}.$$

In de richting van de vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ heeft deze kansverdeling dus een variantie van $2+\sqrt{3} \approx 3.73$ en in de richting van de vector $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ is de variantie $2-\sqrt{3} \approx 0.27$.

In Figuur II.9 is een 3-dimensionaal plaatje van de dichtheidsfunctie met covariantie matrix Σ_3 te zien.

> Bij de omgekeerde taak, dat we bij een stelsel waarnemingen de parameters van een verdeling moeten schatten, is het natuurlijk voor de hand liggend het aantal parameters te beperken als er hiervoor plausibele redenen zijn. De verwachtingswaarde van een n-dimensionale Gauss-verdeling geeft n parameters en de covariantie matrix nog eens $\frac{n(n+1)}{2}$ (omdat de matrix symmetrisch is), en dit is vaak redelijk veel. Men veronderstelt daarom vaak gewoon dat de componenten onafhankelijk zijn, dus dat de covariantie matrix diagonaal is, dit reduceert het aantal parameters tot $2n$. Soms wordt ook nog de variantie van de componenten gelijk gekozen, dan is de covariantie matrix een veelvoud van de eenheidsmatrix.



Figuur II.9: Normaalverdeling in dimensie 2 met covariantie matrix Σ_3.

13.2 Mixture modellen

Een situatie die we vaak tegenkomen is het volgende: We hebben een aantal waarnemingen $o_1, \ldots, o_N$ die volgens een (onbekend) proces geproduceerd zijn. Aan de hand van deze waarnemingen willen we een probabilistisch model bepalen dat de waarnemingen zo goed mogelijk beschrijft.

In de meeste gevallen kunnen we om theoretische redenen ervan uitgaan dat we het met een zeker type van kansverdelingen te maken hebben, bijvoorbeeld met een Gauss-verdeling, een Poisson verdeling of een binomiale verdeling. We zullen ons hier min of meer tot Gauss-verdelingen beperken, omdat dit aan de ene kant de belangrijkste verdeling is, en omdat de methoden analoog op andere verdelingen toegepast kunnen worden.

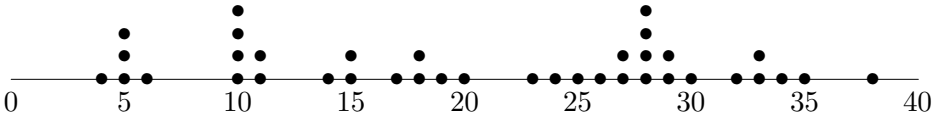
Een voorbeeld

Als voorbeeld kijken we naar de verdeling van cijfers bij een tentamen. Vaak is het zo dat de uitslagen van een tentamen ongeveer normaalverdeeld zijn, de meeste mensen halen een 6, 7 of 8, al minder een 5 of 9 en nog minder hoogstens een 4 of een 10. Als het gemiddelde bij 8 ligt, acht men het tentamen als (te) makkelijk, als het bij 6 of lager ligt, was het tentamen misschien echt te moeilijk.

Maar soms zijn er ook uitslagen als de volgende, waarbij voor elke student het aantal behaalde punten aangegeven is (het maximaal aantal punten was

40): 4, 5, 5, 5, 6, 10, 10, 10, 10, 11, 11, 14, 15, 15, 17, 18, 18, 19, 20, 23, 24, 25, 26, 27, 27, 28, 28, 28, 28, 29, 29, 30, 32, 33, 33, 34, 35, 38.

Als we dit als enkele waarden op een as laten zien, hebben we



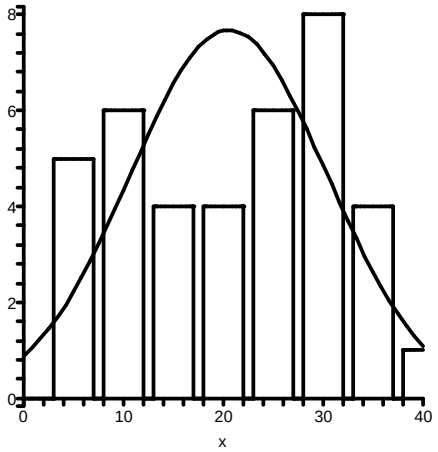
We hebben in Wiskunde 1 gezien dat we met de *maximum likelihood schatting* voor een Gauss-verdeling de verwachtingswaarde μ schatten door het gemiddelde van de waarnemingen, in het voorbeeld dus

$$\mu = \frac{1}{N}\sum_{i=1}^{N} o_i = \frac{780}{38} \approx 20.53.$$

De variantie σ^2 schatten we als gemiddelde van de kwadratische afwijkingen van de schatting voor μ, in het voorbeeld geeft dit

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(o_i - \mu)^2 \Rightarrow \sigma \approx 9.88.$$

Het plaatje in Figuur II.10 geeft een vergelijk van de geschatte Gauss-verdeling met de daadwerkelijke verdeling, waarbij we de uitslagen door een histogram beschrijven dat de uitslagen binnen een interval van 5 punten samenvat. Het is duidelijk dat de Gauss-verdeling de waarnemingen erg slecht beschrijft, bijvoorbeeld is er rond de verwachtingswaarde eigenlijk een dip en wordt de hoge dichtheid rond 30 punten niet adequaat weergegeven.



Figuur II.10: Beschrijving van tentamen uitslagen door een Gauss-verdeling.

In dit voorbeeld krijgt men het idee dat een Gauss-verdeling niet flexibel genoeg is, om dit soort waarnemingen te beschrijven, en dat een combinatie van twee Gauss-verdelingen misschien beter zou passen.

We kunnen bijvoorbeeld de punten in twee delen onderverdelen, waarbij de punten beneden het gemiddelde in de ene helft en de punten boven het gemiddelde in de andere helft belanden. In het voorbeeld komt het er toevallig zo uit, dat dan in beide helften even veel punten zitten, namelijk 19. Voor de twee helften schatten we nu aparte Gauss-verdelingen $\mathcal{N}(\mu_i, \sigma_i)$ met $i = 1, 2$. Voor de helft met de lagere uitslagen geeft dit

$$\mu_1 = \frac{223}{19} \approx 11.74 \text{ en } \sigma_1 \approx 5.11$$

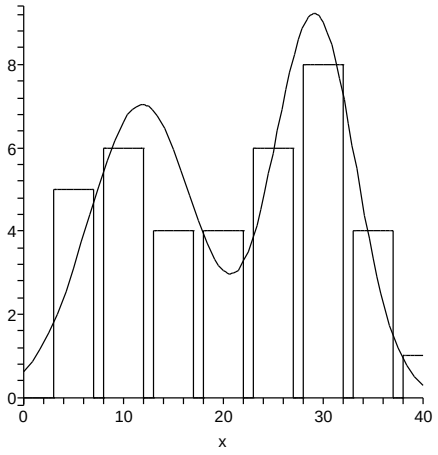
voor de helft met de hogere uitslagen krijgen we

$$\mu_2 = \frac{557}{19} \approx 29.32 \text{ en } \sigma_2 \approx 3.84.$$

De twee Gauss-verdelingen worden nu gecombineerd door ze bij elkaar op te tellen, waarbij we gewichten aan de twee componenten moeten toekennen, zo dat de totale kansmassa van de gecombineerde verdeling weer 1 is. Omdat in ons voorbeeld even veel punten in de twee delen zitten, is het voor de hand liggend voor beide componenten gewicht $\frac{1}{2}$ te nemen, dit geeft dan de dichtheidsfunctie:

$$f(x) = \frac{1}{2}\,\frac{1}{\sqrt{2\pi}\,\sigma_1}\,e^{-\frac{-(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1}{2}\,\frac{1}{\sqrt{2\pi}\,\sigma_2}\,e^{-\frac{-(x-\mu_2)^2}{2\sigma_2^2}}.$$

In Figuur II.11 zien we dat dit al een duidelijk betere beschrijving van de uitslag van het tentamen is. Uit het feit dat een gecombineerde verdeling met twee componenten een goede beschrijving van de waarnemingen geeft, kan men omgekeerd ook concluderen, dat de studenten in twee klassen ingedeeld kunnen worden: degene die veel van de stof hebben begrepen (en misschien al veel ervan kenden) en degene die moeite met de stof hebben.



Figuur II.11: Beschrijving van tentamen uitslagen door een mixture van twee Gauss-verdelingen.

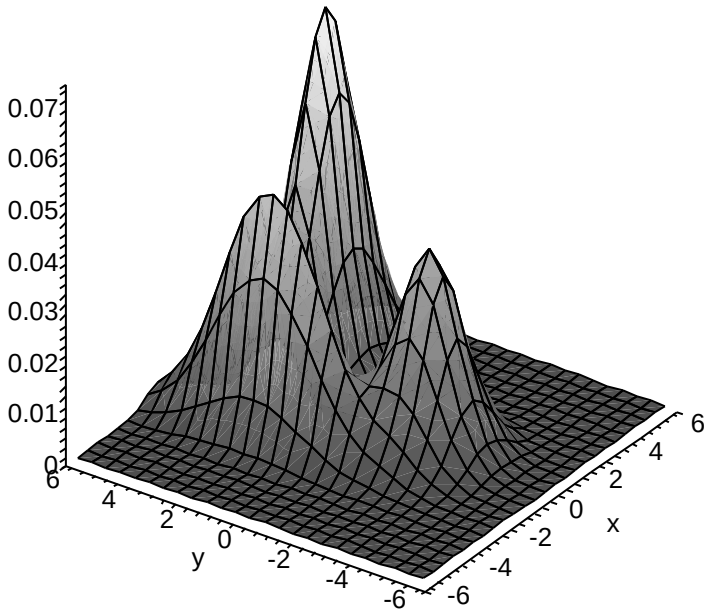
Gaussian mixture modellen

Algemeen noemt men een lineaire combinatie van verschillende kansverdelingen een *mixture model*, waarbij meestal stiekem verondersteld wordt dat de enkele componenten eenvoudige kansverdelingen zijn. In het belangrijke geval dat alle componenten Gauss-verdelingen zijn, spreekt men van een *Gaussian mixture model*.

Een n-dimensionaal Gaussian mixture model met r componenten krijgt men als volgt: Elk van de r componenten is een n-dimensionale Gauss-verdeling, gegeven door de verwachtingswaarde(-vector) $\boldsymbol{\mu}_i$ en de covariantie matrix Σ_i. Verder hebben we gewichten $w_1, \ldots, w_r$ met $\sum_{i=1}^{r} w_i = 1$. Dan heeft het Gaussian mixture model met de Gauss-verdelingen $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ en gewichten $w_1, \ldots, w_r$ de dichtheidsfunctie

$$f(\boldsymbol{x}) = \sum_{i=1}^{r} w_i \, \frac{1}{(2\pi)^{\frac{n}{2}} \, \det(\Sigma_i)^{\frac{1}{2}}} \, \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^{tr} \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right).$$

In Figuur II.12 is een 2-dimensionaal Gaussian mixture model met drie componenten te zien, waarbij de componenten verschillende gewichten hebben.



Figuur II.12: Gaussian mixture model met drie componenten.

13.3 Schatten van parameters

We hebben in het voorbeeld van de tentamensuitslagen de parameters zo bepaald dat de kans op de waarnemingen, gegeven het model, maximaal wordt. Zo'n schatting van de parameters (die we ook in Wiskunde 1 al eens hebben bekeken) noemt men een *maximum likelihood schatting*. Dit is echter niet de enige mogelijkheid om te definiëren welke parameters optimaal zijn voor het beschrijven van een rij waarnemingen.

De meest gebruikte definities voor optimale zullen we nu kort bespreken. Merk op dat geen van de verschillende aanpakken per se beter is dan de anderen. Ook geldt voor elke definitie dat de optimale parameters bijna in elk geval alleen maar numeriek benadert kunnen worden, dus er is ook algoritmisch geen duidelijke voorkeur aan een van de toegangen te geven.

> Aan de andere kant laat zich aantonen dat alle methoden naar hetzelfde model convergeren, als de lengte N van de rij waarnemingen tegen oneindig gaat. De verschillen liggen in de schattingen voor kortere rijen $O = o_1, \ldots, o_N$ van waarnemingen, maar dit is natuurlijk in de praktijk het belangrijke geval, want er is nooit voldoende training materiaal: *There is no data like more data.*

Maximum likelihood schatting

De maximum likelihood methode zijn we al een paar keer tegen gekomen. Het idee is, een aantal waarnemingen $O = o_1, \ldots, o_N$ te bekijken en de voorwaardelijke kans $p(O \mid \lambda(\theta))$ te berekenen, waarbij het model λ van parameters θ afhangt. Degene parameters $\hat{\theta}$ waarvoor $p(O \mid \lambda(\theta))$ zijn maximum aanneemt, zijn de maximum likelihood schatting. Dit schrijft men vaak als

$$\hat{\theta} := \operatorname*{argmax}_{\theta} p(O \mid \lambda(\theta))$$

waarmee uitgedrukt wordt dat $\hat{\theta}$ niet de maximale waarde van $p(O \mid \lambda(\theta))$ is, maar het argument waar het maximum aangenomen wordt.

In de praktijk werkt men meestal met de logaritme van de kans, de zogeheten loglikelihood. In principe zou men hiervan een maximum kunnen vinden door de partiële afgeleiden naar de parameters θ gelijk aan 0 te zetten, maar meestal lukt dit niet meer analytisch.

Een speciaal geval is de n-dimensionale Gauss-verdeling, hier gaat men na dat analoog met de gewone Gauss-verdeling de maximum likelihood schatting voor de verwachtingswaarde en de covariantie matrix er zo uit ziet:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} o_i \quad \text{en} \quad \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(o_i - \hat{\boldsymbol{\mu}})(o_i - \hat{\boldsymbol{\mu}})^{tr}.$$

Merk op dat $(o_i - \hat{\boldsymbol{\mu}})$ een kolom vector is, en een kolom vector maal een rij vector geeft inderdaad een matrix.

In het algemeen moet men bij de maximum likelihood schatting het maximum benaderen, dit geldt met name voor Gaussian mixture modellen waarbij

naast de verwachtingswaarden $\boldsymbol{\mu}_i$ en de covariantie matrices Σ_i ook de gewichten w_i geschat moeten worden.

Maximum a posteriori schatting

Bij de maximum likelihood schatting hebben we de parameters θ van het model λ als variabelen geïnterpreteerd. Maar we kunnen hier ook anders na kijken, namelijk de parameters θ als toevalsvariabelen zien onder de voorwaarde dat de waarnemingen $O = o_1, \ldots, o_N$ zijn gebeurd. Dit betekent dat we $p(\theta \mid O)$ willen maximaliseren, dus bepalen we de optimale parameters $\hat{\theta}$ door

$$\hat{\theta} := \operatorname*{argmax}_{\theta} p(\theta \mid O)$$

en we noemen dit de *maximum a posteriori* schatting voor θ. Volgens de regel van Bayes geldt

$$p(\theta \mid O) = \frac{p(O \mid \theta)p(\theta)}{p(O)}$$

en omdat de noemer bij het bepalen van het maximum geen rol speelt, vinden we het maximum van $p(\theta \mid O)$ door $p(O \mid \theta)p(\theta)$ te maximaliseren. Het verschil tegenover de maximum likelihood aanpak ligt dus in de *a priori kans* $p(\theta)$, die rekening ermee houdt, dat misschien niet alle stelsels van parameters dezelfde kans hebben. Als $p(\theta)$ een bijna uniforme verdeling is, zullen de twee aanpakken niet veel verschillen, maar als we al zekere informatie over het model hebben, kan dit er heel anders uitzien.

> De naam *a posteriori schatting* benadrukt het feit, dat we *voor* het waarnemen van O al een kansverdeling $p(\theta)$ voor θ hadden, maar dat we deze *na* het waarnemen van O tot $p(\theta \mid O)$ aanpassen.

Bayesian Learning

Bij de maximum likelihood en de maximum a posteriori schatting proberen we, parameters van de kansverdeling te schatten, waarmee we de kansen $p(X = x)$ berekenen. Maar in feite zijn we vooral aan de kansverdeling $p(X = x)$ geïnteresseerd. Omdat we de waarnemingen O als informatie bron hebben, kunnen we ook eens kijken wat er over de voorwaardelijke kans $p(x \mid O)$ te zeggen valt.

Natuurlijk veronderstellen we nu ook weer een kansverdeling die van parameters θ afhangt, en door over de mogelijke parameterwaarden te integreren krijgen we een schatting voor de kansverdeling als volgt:

$$p(x \mid O) = \int_{\theta} p(x, \theta \mid O)\, d\theta,$$

waarbij de parameters θ over alle mogelijke waarden lopen. Met behulp van de Bayes regel zien we dat $p(x, \theta \mid O) = p(x \mid \theta, O)\, p(\theta \mid O)$, maar er geldt

$p(x \mid \theta, O) = p(x \mid \theta)$, omdat de waarde x niet van de waarnemingen o_i afhangt. We hebben dus

$$p(x \mid O) = \int_\theta p(x \mid \theta)\, p(\theta \mid O)\, d\theta.$$

Als $p(\theta \mid O)$ nu bijvoorbeeld een scherpe peak rond zekere (de optimale) parameters $\hat{\theta}$ heeft, dus bijna een Dirac δ-functie is, dan geeft de integraal de benadering $p(x \mid O) \approx p(x \mid \hat{\theta})$, d.w.z. we vullen de optimale parameters in. Op die manier zouden we dezelfde kansverdeling krijgen als bij een maximum a posteriori schatting van de optimale parameters $\hat{\theta}$. Het feit dat we onzeker erover zijn welke parameters optimaal zijn, geeft aanleiding over de verschillende mogelijkheden te middelen en dit noemt men *Bayesian learning*.

Het probleem bij de integratie over $p(x \mid \theta)\, p(\theta \mid O)$ is de kansverdeling $p(\theta \mid O)$, die we ook bij de maximum a posteriori aanpak tegen zijn gekomen. Toen hoefden we alleen maar de parameters $\hat{\theta}$ te bepalen waar de kansverdeling maximaal wordt, maar nu hebben we de volledige verdeling nodig. Dit laat zien dat deze aanpak in het algemeen een zware rekenlast vergt.

Maar gelukkig is het onder zekere voorwaarden wel mogelijk de kansverdeling $p(\theta \mid O)$ te bepalen. We kijken naar het speciaal geval van een 1-dimensionale Gauss-verdeling $p(x) = \mathcal{N}(\mu, \sigma)$ met gegeven variantie σ^2 maar onbekende verwachtingswaarde μ. We veronderstellen dat de verwachtingswaarde μ zelf ook normaalverdeeld is, namelijk $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, een informatie die we bijvoorbeeld door eerdere waarnemingen hebben geschat. In dit geval kunnen we niet alleen maar de maximum a posteriori schatting van $p(\mu \mid O)$, maar ook de kansverdeling zelfs expliciet uitrekenen, en als resultaat krijgen we een Gauss-verdeling $\mathcal{N}(\hat{\mu}, \hat{\sigma})$ met parameters:

$$\hat{\mu} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\, \overline{o} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\, \mu_0 \quad \text{en} \quad \hat{\sigma} = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2},$$

waarbij $\overline{o} := \frac{1}{N} \sum_{i=1}^{N} o_i$. We zien dat $\hat{\mu}$ een gewogen gemiddelde tussen het gemiddelde $\overline{o}$ van de waarnemingen en de a priori schatting μ_0 van de verwachtingswaarde is, waarbij het gewicht voor $\overline{o}$ met groeiend aantal N van waarnemingen toeneemt.

Met behulp van de verdeling $p(\mu \mid O)$ kunnen we nu ook de kansverdeling $p(x \mid 0)$ zelf herschatten, dit wordt in dit geval een Gauss-verdeling met verwachtingswaarde $\hat{\mu}$ en variantie $\sigma^2 + \hat{\sigma}^2$.

Het voorbeeld laat zien hoe we de eerdere informatie over de verwachtingswaarde μ met de nieuwe informatie uit de waarnemingen O combineren. Dit geeft de mogelijkheid om parameters stapsgewijs met betrekking tot nieuwe waarnemingen aan te passen. Een toepassing van deze methode ligt bijvoorbeeld in de spraakherkenning: Als uitgangspunt (analoog met de verdeling $\mathcal{N}(\mu_0, \sigma_0^2)$) neemt men een model dat op training materiaal van verschillende sprekers baseert en daarom enigszins onafhankelijk is van de sprekers. Vervolgens wordt dit model door waarnemingen van een individuele spreker volgens de Bayesian learning methode aan deze spreker aangepast. Voor deze aanpassing

is veel minder training materiaal (typisch 0.5 tot 2 uur) van de spreker nodig dan voor het volledige trainen van de modellen (waar vaak honderden van uren materiaal gebruikt worden).

Maximum entropie schatting

Een iets andere aanpak voor het schatten van de parameters van een probabilistisch model gebruikt de entropie. Het idee achter deze methode is, de meest algemene kansverdeling te bepalen die aan gegeven randvoorwaarden voldoet.

Voor een discrete kansverdeling $P = (p_1, \ldots, p_N)$ hadden we de entropie gedefinieerd als $H(X) = -\sum_{i=1}^{N} p_i \, {}^2\log(p_i)$ en voor een continue kansverdeling met dichtheidsfunctie $f(x)$ als $H(X) = -\int_{-\infty}^{\infty} f(x) \, {}^2\log(f(x)) \, dx$. Als de kansverdeling van parameters θ afhangt dan geldt hetzelfde natuurlijk ook voor de entropie $H(X)$. De maximum entropie methode bepaald de parameters van de kansverdeling zo dat de entropie maximaal wordt, onder mogelijke gegeven randvoorwaarden. Hiermee wordt uitgedrukt dat we onzeker over de kansverdeling zijn, behalve over de dingen die we in de randvoorwaarden hebben geformuleerd. De kunst ligt hierbij in het opzetten van de randvoorwaarden die meestal uit waarnemingen afgeleid worden.

Als men bijvoorbeeld alleen maar verondersteld wordt, dat $f(x) = 0$ voor x buiten het interval $[a, b]$, dan gaat men na, dat de entropie maximaal wordt voor de uniforme verdeling op het interval $[a, b]$. Maar als men (op grond van zekere waarnemingen) ook nog de verwachtingswaarde en de variantie als randvoorwaarde vastlegt, krijgt men een normaalverdeling als maximum entropie schatting.

13.4 Hidden Markov modellen met continue emissie kansen

We hebben in de vorige les verondersteld dat de states van een HMM alleen maar eindig veel mogelijke waarnemingen kunnen produceren. In de praktijk kan men dit natuurlijk altijd bereiken door verschillende waarnemingen tot een klasse samen te vatten (bijvoorbeeld door vector quantisering), maar vaak is dat een te grove benadering.

Een oplossing hiervoor bestaat erin, de emissiekansen $b_i(o_t)$ niet meer door discrete kansverdelingen maar door dichtheidsfuncties van continue kansverdelingen te beschrijven. Hierbij neemt men bijna altijd kansverdelingen die door parameters zijn beschreven, en de meest gebruikte vorm van deze kansverdelingen zijn Gaussian mixture modellen. Als de k-de component van de (n-dimensionale) Gaussian mixture voor state S_i de Gauss-verdeling $\mathcal{N}(\mu_{ik}, \Sigma_{ik})$ en deze gewicht w_{ik} heeft, dan is de emissiekans voor een waarneming $\boldsymbol{x}$ vanuit state S_i gegeven door:

$$b_i(\boldsymbol{x}) = \sum_{k=1}^{K} w_{ik} \, \frac{1}{(2\pi)^{\frac{n}{2}} \det(\Sigma_{ik})^{\frac{1}{2}}} \, \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_{ik})^{tr} \Sigma_{ik}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{ik}) \right).$$

Merk op dat hierbij de gewichten w_{ik} voor één state S_i bij elkaar opgeteld 1 moeten geven, dus $\sum_{k=1}^{K} w_{ik} = 1$.

Een Hidden Markov model met continue emissiekansen wordt dus beschreven door de volgende parameters:

- de beginverdeling π van de states,
- de overgangskansen a_{ij},
- de emissiekansen $b_i(\boldsymbol{x})$, gegeven door de gewichten w_{ik}, de verwachtingswaarden $\boldsymbol{\mu}_{ik}$ en de covariantie matrices Σ_{ik}.

Als deze parameters van een HMM λ bekend zijn, dan hebben de eerste twee fundamentele problemen voor HMMs, namelijk het berekenen van de kans $p(O \mid \lambda)$ van een rij O van waarnemingen en het vinden van de optimale rij states $q = \text{argmax}_q\, p(O, q \mid \lambda)$ voor een rij waarnemingen precies dezelfde oplossing als in het geval van discrete emissiekansen. In het forward algoritme en in het Viterbi algoritme zijn namelijk van de kansverdelingen $b_i(\boldsymbol{x})$ alleen maar hun waarden op de waarnemingen $o_1, \ldots, o_T$ nodig, en deze kunnen we natuurlijk uitrekenen. Aan deze algoritmen verandert dus helemaal niets, behalve van de manier hoe de emissiekansen $b_i(o_t)$ uitgerekend worden.

Anders zit het met het derde fundamentele probleem, het bepalen van de parameters. Hier moeten namelijk behalve van de verwachtingswaarden $\boldsymbol{\mu}_{ik}$ en de covariantie matrices Σ_{ik} ook nog de gewichten w_{ik} geschat worden. Hiervoor wordt meestal het *expectation maximization algoritme* toegepast, een veralgemening van het Baum-Welch algoritme dat we bij de discrete emissiekansen gezien hebben.

Expectation maximization algoritme

Het *expectation maximization* algoritme, kort *EM-algoritme* (soms ook *expectation modification* algoritme geheten) is een algemene methode waarbij men de likelihood van een probabilistisch model verbetert door een zekere hulpfunctie te optimaliseren. Het wordt toegepast in situaties waar een model verborgen parameters bevat die niet geobserveerd kunnen worden, zo als de states van een HMM.

Voor twee modellen λ en λ' en een waarneming O is deze hulpfunctie (ook Q-functie geheten) gedefinieerd door:

$$Q(\lambda, \lambda') := \sum_q p(O, q \mid \lambda) \cdot \log(p(O, q \mid \lambda'))$$

waarbij de som over alle rijen $q = q_1 q_2 \ldots q_T$ van states loopt. (Het is overigens geen toeval dat de Q-functie sterk op de uitdrukkingen lijkt die we in het kader van de entropie hebben gezien.) De cruciale eigenschap van de Q-functie is:

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \Rightarrow p(O \mid \lambda') \geq p(O \mid \lambda),$$

d.w.z. men vindt een beter model voor de beschrijving van O, door de functie $Q(\lambda, \lambda')$ over λ' te maximaliseren.

In het verband met HMMs geldt:

$$p(O, q \mid \lambda') = \pi'(q_1) b'_{q_1}(o_1) \prod_{t=1}^{T-1} a'_{q_t q_{t+1}} b'_{q_{t+1}}(o_{t+1}),$$

dus is

$$\log(p(O, q \mid \lambda')) = \log(\pi'(q_1)) + \sum_{t=1}^{T-1} \log(a'_{q_t q_{t+1}}) + \sum_{t=1}^{T} \log(b'_{q_t}(o_t)).$$

We kunnen daarom $Q(\lambda, \lambda')$ schrijven als

$$Q(\lambda, \lambda') = Q_{\pi'}(\lambda, \pi') + \sum_{i=1}^{N} Q_{a'_i}(\lambda, a'_i) + \sum_{i=1}^{N} Q_{b'_i}(\lambda, b'_i)$$

waarbij

$$\begin{aligned}
Q_{\pi'}(\lambda, \pi') &= \sum_{i=1}^{N} p(O, q_1 = S_i \mid \lambda) \log(\pi'_i), \\
Q_{a'_i}(\lambda, a'_i) &= \sum_{j=1}^{N} \sum_{t=1}^{T-1} p(O, q_t = S_i, q_{t+1} = S_j \mid \lambda) \log(a'_{ij}), \\
Q_{b'_i}(\lambda, b'_i) &= \sum_{t=1}^{T} p(O, q_t = S_i \mid \lambda) \log(b'_i(o_t)).
\end{aligned}$$

We kunnen deze functies apart maximaliseren, omdat $Q_{\pi'}$ alleen maar van de parameters $\pi'(i)$, $Q_{a'_i}$ alleen maar van de parameters a'_{ij} en $Q_{b'_i}$ alleen maar van de parameters van b'_i afhangt. Hierbij moeten we wel opletten, dat de parameters aan de randvoorwaarden voor kansverdelingen moeten voldoen, dus dat alle parameters ≥ 0 zijn en dat:

$$\sum_{i=1}^{N} \pi'(i) = 1, \quad \sum_{j=1}^{N} a'_{ij} = 1 \text{ voor } 1 \leq i \leq N, \quad \int b'_i(\boldsymbol{x})\, d\boldsymbol{x} = 1 \text{ voor } 1 \leq i \leq N.$$

We zien dat de functies voor de beginverdeling en de overgangskansen van de vorm $Q(y_1, \ldots, y_M) = \sum_{j=1}^{M} w_j \log(y_j)$ met de randvoorwaarde $\sum_{j=1}^{M} y_j = 1$ zijn. Als we met behulp van deze vergelijking y_M vervangen door $y_M = 1 - \sum_{j=1}^{M-1} y_j$ en vervolgens partieel naar y_k afleiden, krijgen we

$$\begin{aligned}
\frac{\partial}{\partial y_k} Q(y_1, \ldots, y_M) &= \frac{\partial}{\partial y_k} \Big(\sum_{j=1}^{M-1} w_j \log(y_j) + w_M \log(1 - y_1 - \ldots - y_{M-1}) \Big) \\
&= \frac{w_k}{y_k} - \frac{w_M}{1 - y_1 - \ldots - y_{M-1}} = \frac{w_k}{y_k} - \frac{w_M}{y_M}.
\end{aligned}$$

In een lokaal maximum moeten alle partiële afgeleiden 0 zijn, hieruit volgt dat $\frac{w_k}{y_k} = \frac{w_M}{y_M}$ voor alle k, d.w.z. $y_k = c w_k$ voor een vaste constante c. Maar omdat $\sum_{j=1}^{M} y_j = 1$, is deze constante noodzakelijk $c = \frac{1}{\sum_{j=1}^{M} w_j}$, dus hebben we

$$y_k = \frac{w_k}{\sum_{j=1}^{M} w_j}.$$

Als we dit weer voor de functies $Q_{\pi'}(\lambda, \pi')$ en $Q_{a'_i}(\lambda, a'_i)$ invullen, krijgen we precies de vergelijkingen die we al in de vorige les hebben gevonden, namelijk

$$\pi'(i) = \frac{\alpha_1(i)\ \beta_1(i)}{\sum_{i=1}^{N} \alpha_t(i)\ \beta_t(i)},$$
$$a'_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\ a_{ij}\ b_j(o_{t+1})\ \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)}.$$

Schatten van de Gaussian mixture modellen

We moeten nu nog de functies $Q_{b'_i}(\lambda, b'_i)$ maximaliseren, waarbij $b'_i(\boldsymbol{x})$ een Gaussian mixture model is. De expliciete oplossing van dit probleem zullen we hier niet behandelen, dit vraagt vervelend veel rekenwerk. Maar we zullen wel het resultaat aangeven en nagaan dat dit plausibel lijkt.

We veronderstellen eerst eens, dat de kansverdelingen $b_i(\boldsymbol{x})$ van de modellen λ en λ' gewone n-dimensionale Gauss-verdelingen met maar één component zijn, gegeven door de verwachtingswaarden $\boldsymbol{\mu}_i$ en de covariantie matrices Σ_i.

Net als in de vorige les noteren we met

$$\gamma_t(i) := p(q_t = S_i \mid O, \lambda)$$

de kans dat de waarneming op tijdstip t door de state S_i geproduceerd is. Deze kans konden we met behulp van de vooruitkansen $\alpha_t(i)$ en achteruitkansen $\beta_t(i)$ makkelijk uitrekenen, namelijk als

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}.$$

Als verbeterde schatting voor de parameters van de Gauss-verdelingen krijgen we nu:

$$\boldsymbol{\mu}'_i = \frac{1}{\sum_{t=1}^{T} \gamma_t(i)} \sum_{t=1}^{T} \gamma_t(i)\, o_t,$$
$$\Sigma'_i = \frac{1}{\sum_{t=1}^{T} \gamma_t(i)} \sum_{t=1}^{T} \gamma_t(i)\ (o_t - \boldsymbol{\mu}'_i)(o_t - \boldsymbol{\mu}'_i)^{tr}$$
$$= \left(\frac{1}{\sum_{t=1}^{T} \gamma_t(i)} \sum_{t=1}^{T} \gamma_t(i)\ o_t o_t^{tr} \right) - \boldsymbol{\mu}'_i \boldsymbol{\mu}_i'^{tr}$$

Dit resultaat wordt begrijpelijk als we het met de maximum likelihood schatting van een gewone Gauss-verdeling vergelijken, die gegeven is door

$$\boldsymbol{\mu} = \frac{1}{T} \sum_{t=1}^{T} o_t \quad \text{en} \quad \Sigma = \frac{1}{T} \sum_{t=1}^{T} (o_t - \boldsymbol{\mu})(o_t - \boldsymbol{\mu})^{tr}.$$

In plaats van het gewone gemiddelde is $\boldsymbol{\mu}'_i$ het gewogen gemiddelde van de waarnemingen o_t, waarbij de gewichten $\gamma_t(i)$ aangeven, met welke kans we het

systeem op tijdstip t in state S_i verwachten. Hierdoor houden we rekening ermee, dat waarnemingen die bijna zeker door de state S_i geproduceerd zijn een grote rol bij het bepalen van de verwachtingswaarde in state S_i spelen. Omgekeerd spelen waarnemingen op tijdstippen waar het systeem zeker niet in state S_i is op deze manier ook geen rol voor de verwachtingswaarde in state S_i.

Dezelfde redenering geldt ook voor de covariantie matrices Σ_i', want ook hierbij vervangen we het gewone gemiddelde van de covariantie matrices $(o_t - \boldsymbol{\mu})(o_t - \boldsymbol{\mu})^{tr}$ door het gewogen gemiddelde met gewichten $\gamma_t(i)$.

> We kunnen de schatting van de parameters van een HMM met k states ook interpreteren als schatting voor de parameters van een Gaussian mixture model, waarmee we een rij $O = o_1, \ldots, o_T$ van waarnemingen willen beschrijven (die we niet noodzakelijk als waarnemingen op verschillende tijdstippen hoeven te zien). De states van het HMM worden dan de componenten van het mixture model. De verwachtingswaarden en covariantie matrices worden dan precies met de boven aangegeven formules bepaald, maar in plaats van de overgangskansen tussen de states hebben we nu de relatieve gewichten van de enkele componenten nodig. Maar $\gamma_t(i)$ geeft de kans aan, dat de waarneming o_t door de i-de component is geproduceerd, daarom geeft het gemiddelde over de $\gamma_t(i)$ de kans aan, dat een waarneming überhaupt door de i-de component geproduceerd is, en we krijgen als herschatting w_i' voor het gewicht w_i van de i-de component:
>
> $$w_i' = \frac{1}{T} \sum_{t=1}^{T} \gamma_t(i).$$

We kijken nu naar het algemenere geval dat de verdeling $b_i(\boldsymbol{x})$ van de emissiekansen vanuit state S_i een Gaussian mixture model is, d.w.z. we hebben

$$b_i(x) = \sum_{k=1}^{K} w_{ik}\, b_{ik}(\boldsymbol{x}),$$

waarbij $b_{ik}(\boldsymbol{x})$ een Gauss-verdeling $\mathcal{N}(\boldsymbol{\mu}_{ik}, \Sigma_{ik})$ met verwachtingswaarde $\boldsymbol{\mu}_{ik}$ en covariantie matrix Σ_{ik} is en natuurlijk $\sum_{k=1}^{K} w_{ik} = 1$ voor alle i.

We kunnen dit geval terug brengen naar het net behandelde geval van eenvoudige Gauss-verdelingen door de componenten van de mixture modellen als tweede level van verborgen states te beschouwen. In plaats van alleen maar de states $S_1, \ldots, S_N$ hebben we zo paren (S_i, k) van verborgen states, waarbij we met $q_t = (S_i, k)$ uitdrukken dat de waarneming op tijdstip t door de k-de component van state S_i geproduceerd is. Met k_t noteren we de component die op tijdstip t vuurt.

Analoog met de kansen $\gamma_t(i) = p(q_t = S_i \mid O, \lambda)$ definiëren we nu de kans, dat het systeem op tijdstip t een waarneming vanuit de k-de component van state S_i produceert en noemen deze kans $\zeta_t(i,k)$. We hebben dus (onder gebruik van de regel van Bayes):

$$\zeta_t(i,k) := p(q_t = i, k_t = k \mid O, \lambda) = \frac{p(O, q_t = i, k_t = k \mid \lambda)}{p(O \mid \lambda)}.$$

Net als de kansen $\gamma_t(i)$ kunnen we ook de kansen $\zeta_t(i,k)$ met behulp van de vooruitkansen $\alpha_t(i)$ en de achteruitkansen $\beta_t(i)$ makkelijk uitrekenen, er geldt:

$$\zeta_1(i,k) = \frac{\pi(i)\, w_{ik}\, b_{ik}(o_1)\, \beta_1(i)}{p(O \mid \lambda)}$$
$$\zeta_t(i,k) = \frac{\sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji}\, w_{ik}\, b_{ik}(o_t)\, \beta_t(i)}{p(O \mid \lambda)}$$

Analoog met de formules voor de $b'_i(\boldsymbol{x})$ met eenvoudige Gauss-verdelingen krijgen we zo de schattingen voor de mixture modellen $b'_i(\boldsymbol{x}) = \sum_{k=1}^{K} w'_{ik}\, b'_{ik}(\boldsymbol{x})$ als volgt:

$$w'_{ik} = \frac{1}{\sum_{t=1}^{T} \gamma_t(i)} \sum_{t=1}^{T} \zeta_t(i,k),$$
$$\boldsymbol{\mu}'_{ik} = \frac{1}{\sum_{t=1}^{T} \zeta_t(i,k)} \sum_{t=1}^{T} \zeta_t(i,k)\; o_t,$$
$$\Sigma'_{ik} = \left(\frac{1}{\sum_{t=1}^{T} \zeta_t(i,k)} \sum_{t=1}^{T} \zeta_t(i,k)\; o_t o_t^{tr} \right) - \boldsymbol{\mu}'_{ij} \boldsymbol{\mu}''^{tr}_{ik}.$$

Ook hier blijken de nieuwe parameters met onze intuïtie te kloppen. De nieuwe gewichten w'_{ik} zijn gewoon de relatieve frequenties waarmee we in state S_i de k-de component kiezen om een waarneming te produceren. De verwachtingswaarden $\boldsymbol{\mu}'_{ik}$ en de covariantie matrices Σ'_{ik} zijn de gewogen gemiddelden van de maximum likelihood schattingen voor gewone Gauss-verdelingen, waarbij ook hier de gewichten de kans aangeven, waarmee we in state S_i een door de k-de component geproduceerde waarneming verwachtende.

Belangrijke begrippen in deze les

- n-dimensionale Gauss verdeling
- covariantie matrix
- (Gaussian) mixture modellen
- maximum likelihood schatting
- maximum a posteriori schatting
- Hidden Markov modellen met continue emissie kansen
- expectation maximization algoritme