

Nenegativna matrična faktorizacija

Nikola Mamić, Petar Matejčić, Jakov Tušek

3. prosinca, 2025.

Sadržaj

1	Uvod	3
2	Osnovni algoritam	3
2.1	Dokaz konvergencije	4
2.2	Analiza konvergencije	5
3	ARD-NMF	7
3.1	β -divergencija	7
3.2	Apriorne distribucije	8
3.3	Funkcija cilja	9
3.4	ARD-NMF algoritam	9
3.4.1	λ parametri	11
3.4.2	Odabir parametra ϕ	12
3.4.3	Odabir parametara a i b	12
4	Primjeri	13
4.1	Razdvajanje glazbenog zapisa u note	13
4.2	Pronalaženje latentnih komponenti u recenzijama filmova	18
4.3	Analiza slika lica	21
5	Zaključak	25

1 Uvod

Cilj ovog rada je za danu matricu $X \in \mathbb{R}^{n \times p}$ pronaći matrice s nenegativnim elementima $W \in \mathbb{R}^{n \times r}$ i $H \in \mathbb{R}^{r \times p}$ koje minimiziraju $\|X - WH\|$ tj. postići $X \approx WH$ pri čemu je WH značajno manjeg ranga od X . U primjeni stupci od X su neki dokumenti (npr. slike, audio zapisi, tekstualni dokumenti), stupci od W su komponente od kojih su dokumenti sastavljeni, a H sadrži odgovarajuće intenzitete komponenti tj. imamo:

$$X_{:,i} \approx \sum_{k=1}^r h_{k,i} W_{:,k}$$

Uvjetom nenegativnosti W i H postiže se veća interpretabilnost jer se cjelina sastavlja od dijelova, ali i veća robustnost na dodavanje komponenti te se rješenja često sastoje od dosta nula.

Prvo se bavimo algoritmom za konstrukciju W i H kada je r fiksan, zatim pokazujemo Bayesovski pristup za konstrukciju povoljnog r što je naravno ključno jer lošim izborom r gubimo informacije ili overfitamo naše podatke. Na kraju primjenjujemo algoritme na konkretnim primjerima zvučnog zapisa, recenzije filmova i slika lica.

2 Osnovni algoritam

Temeljni algoritam za NMF opisan je u radu [4]. Za pronalazak aproksimacije prvo je nužno definirati funkciju udaljenosti, tj. mjeru kvalitete aproksimacije. Prva takva je jednostavno kvadrat euklidske udaljenosti dvije matrice A i B :

$$\|A - B\|^2 = \sum_{ij} (A_{ij} - B_{ij})^2$$

Stoga razmatramo NMF kao optimizacijski problem po mjeri udaljenosti:

Problem 1. Minimizirati $\|X - WH\|^2$ s obzirom na W i H , uz $W, H \geq 0$.

U originalnom radu predlaže se multiplikativni algoritam opisan u sljedećem teoremu.

Teorem 1. Euklidska udaljenost $\|V - WH\|$ je nerastuća po sljedećim iteracijama:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \quad (1)$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(V H^T)_{i\alpha}}{(W H H^T)_{i\alpha}} \quad (2)$$

Dodatno, euklidska udaljenost je invarijantna po iteracijama ako i samo ako su W i H u stacionarnim točkama udaljenosti (diskusija nakon dokaza).

2.1 Dokaz konvergencije

Za dokaz je potrebna pomoćna funkcija:

Definicija 1. $G(h, h')$ je pomoćna funkcija za $F(h)$ ako vrijedi

$$G(h, h') \geq F(h), \quad G(h, h) = F(h)$$

Lema 1. Ako je G pomoćna funkcija, onda je F nerastuća s obzirom na ažuriranje:

$$h^{t+1} = \arg \min_h G(h, h^t) \quad (3)$$

Dokaz:

$$F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t) \quad \square$$

Lema 2: Neka je $K(h^t)$ dijagonalna matrica

$$K_{ab}(h^t) = \delta_{ab}(W^T W h^t)_a / h_a^t$$

tada je

$$G(h, h^t) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T K(h^t)(h - h^t) \quad (4)$$

pomoćna funkcija za

$$F(h) = \frac{1}{2} \sum_i (v_i - \sum_a W_{ia} h_a)^2$$

Dokaz: Budući da je $G(h, h) = F(h)$ očito iz definicije, trebamo samo pokazati da je $G(h, h^t) \geq F(h)$. Da bismo to napravili, uspoređujemo

$$F(h) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T (W^T W)(h - h^t)$$

s jednadžbom (4) i dobijemo da je $G(h, h^t) \geq F(h)$ ekvivalentno s

$$0 \leq (h - h^t)^T [K(h^t) - W^T W](h - h^t)$$

Da bismo dokazali pozitivnu semidefinitnost, promotrimo matricu:

$$M_{ab}(h^t) = h_a^t (K(h^t) - W^T W)_{ab} h_b^t$$

koja je samo dobivena skaliranjem komponenti od $K - W^T W$. Tada je $K - W^T W$ pozitivno semidefinitna ako i samo ako je M pozitivno semidefinitna, i vrijedi

$$\begin{aligned}
\nu^T M \nu &= \sum_{ab} \nu_a M_{ab} \nu_b \\
&= \sum_{ab} h_a^t (W^T W)_{ab} h_b^t \nu_a^2 - \nu_a h_a^t (W^T W)_{ab} h_b^t \nu_b \\
&= \sum_{ab} (W^T W)_{ab} h_a^t h_b^t \left[\frac{1}{2} \nu_a^2 + \frac{1}{2} \nu_b^2 - \nu_a \nu_b \right] \\
&= \frac{1}{2} \sum_{ab} (W^T W)_{ab} h_a^t h_b^t (\nu_a - \nu_b)^2 \\
&\geq 0
\end{aligned}$$

Dokaz teorema 1. Ako zamijenimo $G(h, h^t)$ u (3) s (4), dobijemo:

$$h^{t+1} = h^t - K(h^t)^{-1} \nabla F(h^t)$$

Pošto je (14) pomoćna funkcija, F je nerastuća pod ovim pravilom ažuriranja. Ako komponente zapišemo eksplisitno, dobijemo

$$h_\alpha^{t+1} = h_\alpha^t \frac{(W^T v)_\alpha}{(W^T W h^t)_\alpha}.$$

Ako zamijenimo W i H u Lemi 1 i 2, pokaže se da je F nerastuća po pravilima ažuriranja za W . \square

2.2 Analiza konvergencije

Pod opisanim iteracijama, greška konvergira. U originalnom radu [4], Lee i Seung tvrde da je konvergencija uvijek u lokalni minimum, što nije točno. Naime, dokazali smo tek da greška opada. Kako bi preciznije analizirali konvergenciju, koristit ćemo Karush-Kuhn-Tucker uvjete.

Definicija 2. Za opću zadaću nelinearnog programiranja

$$f(x) \rightarrow \min, \quad g_j(x) \leq 0, \quad j = 1, \dots, p,$$

gdje su funkcije f i g_j diferencijabilne, dopustiva točka x^* (dopustiva je ako ispunjava sve uvjete g_j) je **Karush-Kuhn-Tuckerova (KKT) točka** ako postoje $\lambda = (\lambda_1, \dots, \lambda_p)$ takvi da vrijedi:

- (a) $\lambda_j \geq 0, \quad j = 1, \dots, p,$
- (b) $\nabla f(x^*) + \sum_{j=1}^p \lambda_j \nabla g_j(x^*) = 0,$
- (c) $\lambda_j g_j(x^*) = 0, \quad j = 1, \dots, p.$

Svaka KKT točka je stacionarna u smislu zadaće nelinearnog programiranja, tj. za svaki dopustivi smjer pomaka d iz KKT točke x^* , $\nabla f(x^*)^T d \geq 0$.

Za NMF problem ograničenja su linearna:

$$W_{ia} \geq 0, \quad H_{aj} \geq 0,$$

što se može zapisati kao $g_{ia}(W, H) = -W_{ia} \leq 0$ i $g'_{aj}(W, H) = -H_{aj} \leq 0$. Dakle, dozvoljeni skup rješenja je konveksan skup definiram tim funkcijama. Radi jednostavnosti računanja, uzimamo

$$f(W, H) = \frac{1}{2} \|X - WH\|_F^2 = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - (WH)_{ij})^2$$

Lema 3. Limesi dobiveni iteracijama (1) i (2) koji se nalaze u interioru dozvoljenog skupa su KKT točke.

Dokaz: Neka je točka (W^*, H^*) jedna takva. Pošto je u interioru, sve nejednakosti uvjeta ispunjene su strogo: $g_{ia}(W^*, H^*) < 0$, $g'_{aj}(W^*, H^*) < 0$. Stoga su za ispunjenje KKT uvjeta (c) nužno $\lambda_{ia} = 0$, $\lambda_{aj} = 0$. Iz toga slijedi da se KKT uvjet (b) svodi na $\nabla f(W^*, H^*) = 0$. Provjerimo to.

$$\begin{aligned} H_{ij}^{(t+1)} &= H_{ij}^{(t)} \frac{[W^T X]_{ij}}{[W^T W H^{(t)}]_{ij}} \\ H_{ij}^{(t+1)} - H_{ij}^{(t)} &= H_{ij}^{(t)} \left(\frac{[W^T X]_{ij}}{[W^T W H^{(t)}]_{ij}} - 1 \right) \\ H_{ij}^{(t+1)} - H_{ij}^{(t)} &= \frac{H_{ij}^{(t)}}{[W^T W H^{(t)}]_{ij}} \left([W^T X]_{ij} - [W^T W H^{(t)}]_{ij} \right), \end{aligned} \quad (5)$$

a izraz u zagradama je točno

$$-\frac{\partial f}{\partial H_{ij}} = [W^T X]_{ij} - [W^T W H]_{ij}.$$

Pustimo limes po t u jednakosti (5) i dobijemo:

$$\begin{aligned} 0 &= \frac{H_{ij}^*}{[W^T W H^*]_{ij}} \left([W^T X]_{ij} - [W^T W H^*]_{ij} \right) \\ \Rightarrow \frac{\partial f}{\partial H_{ij}}(W^*, H^*) &= 0. \end{aligned} \quad (6)$$

Analogno se pokaže za iteracije za W , stoga su KKT uvjeti zadovoljeni za (W^*, H^*) . \square

S druge strane, ako je limes na rubu dopustivog skupa, npr. $H_{ij}^* = 0$, uvjet (6) bit će ispunjen čak i ako je $\frac{\partial f}{\partial H_{ij}}(W^*, H^*) < 0$, pa točka (W^*, H^*) nije nužno KKT točka, a ako nije KKT, onda nije ni lokalni optimum. Dodatno, u slučaju ovdje opisanom, algoritam "zapne" - pomaci bi bili u negativnom smjeru, ali pošto su iteracije multiplikativne i faktor $H_{ij}^* = 0$, ostaje u 0. Tako se održi uvjet nenegativnosti, ali se žrtvuje optimalnost rješenja.

3 ARD-NMF

Kao što smo u uvodu komentirali, jedan od ključnih problema u primjeni nenegativnih matičnih faktorizacija jest odabir broja komponenta, tj. dimenzija matrica W i H . Naime, premali broj komponenta neće moći dobro objasniti podatke, dok će preveliki broj komponenta previše prilagoditi model originalnim podacima (overfitting). Motivirani tim problemom, mi smo implementirali rješenje bazirano na Bayesovskom pristupu obrađeno u radu [1].

Ideja ARD(Automatic Relevance Determination)-NMF-a je sljedeća: stupce w_k i retke h_k "povezat" ćemo zajedničkim parametrima relevantnosti λ_k koje ćemo tijekom izvođenja algoritma ažurirati Bayesovskim pristupom. Nakon računanja W i H , izbacit ćemo one komponente za koje pripadni parametri λ_k padaju ispod određene granice, tj. one koji nam nisu relevantni.

U nastavku pretpostavljamo da je matrica $X \in \mathbb{R}^{F \times N}$

3.1 β -divergencija

Definiramo familiju funkcija za $\beta \in \mathbb{R}$:

$$d_\beta(x|y) = \begin{cases} \frac{x^\beta}{\beta(\beta-1)} + \frac{y^\beta}{\beta} - \frac{xy^{\beta-1}}{\beta-1}, & \beta \in \mathbb{R} \setminus \{0, 1\} \\ x \log \frac{x}{y} - x + y, & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1, & \beta = 0 \end{cases}$$

Funkcije $d_\beta(x|y)$ zovemo β -divergencije.

Familija β -divergencija obuhvaća široki skup funkcija cilja za probleme nenegativne matrične faktorizacije te ona uključuje Euklidsku normu ($\beta = 2$), Kullback-Leibler divergenciju ($\beta = 1$) i Itakura-Saito divergenciju ($\beta = 0$).

Dodatno, za matrice $X, Y \in \mathbb{R}^{m \times n}$ definiramo β -divergenciju:

$$D_\beta(X|Y) = \sum_{i=1}^m \sum_{j=1}^n d_\beta(x_{ij}|y_{ij})$$

Za ARD-NMF važna je veza između β -divergencija i familije Tweedie distribucija. Familija Tweedie distribucija obuhvaća distribucije s funkcijama gustoće oblika:

$$\tau(x|\mu, \phi, \beta) = h(x, \phi) \exp \left[\frac{1}{\phi} \left(\frac{1}{\beta-1} x \mu^{\beta-1} - \frac{1}{\beta} \mu^\beta \right) \right],$$

te posebno sadrži normalnu razdiobu ($\beta = 2$), Poissonovu razdiobu ($\beta = 1$) te Gama razdiobu ($\beta = 0$).

Naime, β -divergencija je proporcionalna razlici log-vjerodostojnosti saturiranog i generalnog modela podataka, tj.

$$\log \frac{\tau(x|\mu = x, \phi, \beta)}{\tau(x|\mu, \phi, \beta)} = \frac{1}{\phi} d_\beta(x|\mu).$$

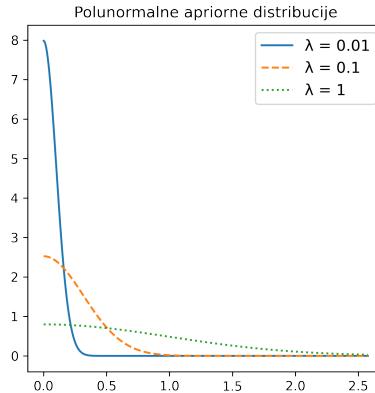
Zbog te činjenice, problem formuliramo na slijedeći način: pretpostavimo da zadana matrica X dolazi iz Tweedie distribucije $\tau(WH, \phi, \beta)$ (po elementima), gdje nam je parametar β zadan. Tada želimo odrediti najvjerojatnije matrice W i H s obzirom na X i β .

3.2 Apriorne distribucije

Kako bi odredili najvjerojatnije matrice W i H te parametre relevantnosti $\lambda = (\lambda_1, \dots, \lambda_K)$, trebamo definirati pripadne apriorne distribucije.

Apriorne distribucije za W i H moraju zadovoljavati dva uvjeta: (1) elementi matrica moraju biti nenegativni i (2) stupci w_i i retci h_i moraju biti povezani parametrima relevantnosti λ_i . U našem radu, mi ćemo koristiti polunormalne distribucije:

$$p(w_{ik}|\lambda_k) = \left(\frac{2}{\pi\lambda_k}\right)^{\frac{1}{2}} \exp\left(-\frac{w_{ik}^2}{2\lambda_k}\right), \quad p(h_{kj}|\lambda_k) = \left(\frac{2}{\pi\lambda_k}\right)^{\frac{1}{2}} \exp\left(-\frac{h_{kj}^2}{2\lambda_k}\right)$$



Slika 1: Na grafu su nacrtane apriorne polunormalne razdiobe s različitim vrijednostima parametra λ . Vidimo da manja vrijednost parametra λ stavlja veću težinu oko 0. Drugim riječima, mala vrijednost parametra λ_k implicira da je w_k nul-stupac,a h_k nul-redak, te oni ne pridonose konačnoj faktorizaciji.

Za apriorne distribucije parametara λ_k uzimam inverzne Gama distribucije:

$$p(\lambda_k : a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\left(-\frac{b}{\lambda_k}\right),$$

gdje su parametri a i b jednaki za sve λ_k i dodatno tražimo da su parametri međusobno nezavisno, tj. $p(\boldsymbol{\lambda} : a, b) = \prod_{k=1}^K p(\lambda_k : a, b)$.

3.3 Funkcija cilja

Želimo pronaći W, H i λ maksimalne vjerodostojnosti za danu matricu X , pa funkciju cilja (koju želimo minimizirati) definiramo:

$$\begin{aligned}
C(W, H, \lambda) &:= -\log p(W, H, \lambda | X) \\
&= -\log \left(\frac{p(X|W, H, \lambda)p(W, H, \lambda)}{p(X)} \right) \\
&= -\log p(X|W, H, \lambda) - \log p(W, H, \lambda) + \log p(X) \\
&= -\log p(X|W, H) - \log p(W, H|\lambda) - \log p(\lambda) + const. \\
&= \frac{1}{\phi} D_\beta(X|WH) + \sum_{k=1}^r \left[\frac{1}{\lambda_i} \left(\frac{1}{2} \|W_k\|_2^2 + \frac{1}{2} \|H_k\|_2^2 + b \right) + c \log \lambda_k \right] + const.,
\end{aligned}$$

pri čemu prva jednakost slijedi iz Bayesovog teorema, treća primjenom $p(W, H, \lambda) = p(W, H|\lambda)p(\lambda)$ i posljednja iz svojstva β -divergencije te pretpostavke da su W i H uvjetno nezavisni uz dano λ i $c := \frac{F+N}{2} + a + 1$.

3.4 ARD-NMF algoritam

Kako i kod MU-NMF-a, algoritmi za računanje W i H su simetrični, pa neka je W fiksani i želimo ažurirati H .

Kao i kod izvoda osnovnog algoritma, koristit ćemo pomoćne funkcije koje odozgo ograničavaju funkciju cilja. U [3] autori su pokazali da su funkcije $G(H|\tilde{H})$ raspisane u tablici 1 pomoćne funkcije za nepenaliziranu funkciju cilja $\frac{1}{\phi} D_\beta(X|WH)$, gdje su p_{kn} i q_{kn} (uz označku $\hat{x}_{ij} := [WH]_{ij}$) definirani:

$$p_{kn} = \sum_f w_{fk} x_{fn} \hat{x}_{fn}^{\beta-2}, \quad q_{kn} = \sum_f w_{fk} \hat{x}_{fn}^{\beta-1}$$

Definiramo funkciju $R_2(H)$:

$$R_2(H) := \sum_{k=1}^r \frac{1}{2\lambda_k} \|h_k\|_2^2,$$

pa $C(W, H, \lambda)$ možemo zapisati na slijedeći način (W nam je fiksani pa $\|w_k\|_2^2$ ulaze u konstantu):

$$C(W, H, \lambda) = \frac{1}{\phi} D_\beta(X|WH) + R_2(H) + const.$$

Neka je $G(H, \tilde{H})$ pomoćna funkcija iz tablice 1. Tada se može pokazati [6] da je $F(H|\tilde{H}) = \phi^{-1}G(H|\tilde{H}) + R_2(H)$ pomoćna funkcija za penaliziranu funkciju cilja $C(W, H, \lambda)$.

Jednadžba $\nabla_H F(H|\tilde{H}) = 0$ ne može se riješiti analitički za sve $\beta \in \mathbb{R}$, pa nam je cilj pronaći neku funkciju $J(H|\tilde{H})$ koja će biti pomoćna funkcija za F i za

koju možemo lako riješiti jednadžbu $\nabla_H J(H|\tilde{H}) = 0$. To radimo za dva slučaja: $\beta > 2$ i $\beta \leq 2$.

U određivanju J koristimo se sljedećom lemom:

Lema 4. Za svaki $\nu > 0$, funkcija $g_\nu(t) = \frac{1}{t}(\nu^t - 1)$ je monotona i nerastuća po $t \in \mathbb{R}$. Posebno, za $\nu \neq 1$, $g_\nu(t)$ je monotono rastuća.

Proučimo sada slučaj kada je $\beta > 2$: Ključna ideja je da, koristeći prethodnu lemu, ograničimo $R_2(H)$ po elementima. Uočimo da za $\nu = h_{kn}/\tilde{h}_{kn}$, zbog $\beta > 2$, imamo:

$$\frac{1}{2} \left[\left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^2 - 1 \right] \leq \frac{1}{\beta} \left[\left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^\beta - 1 \right]$$

Raspisivanjem dobijemo:

$$\frac{1}{2\lambda_k} h_{kn}^2 \leq \frac{1}{\lambda_k \beta} \tilde{h}_{kn}^2 \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^\beta + const.,$$

iz čega lako vidimo:

$$R_2(H) \leq \sum_k \sum_n \frac{1}{\lambda_k \beta} \tilde{h}_{kn}^2 \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^\beta.$$

Konačno, definiramo funkciju J :

$$J(H|\tilde{H}) = \frac{1}{\phi} G(H|\tilde{H}) + \sum_k \sum_n \frac{1}{\lambda_k \beta} \tilde{h}_{kn}^2 \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^\beta$$

Lako vidimo da vrijedi $F(H|\tilde{H}) \leq J(H|\tilde{H})$, te posebno jednakost vrijedi ako i samo $\nu = h_{kn}/\tilde{h}_{kn} = 1$, tj. $H = \tilde{H}$. Stoga, J je pomoćna funkcija za F .

Korištenjem $G(H|\tilde{H})$ iz tablice 1, deriviranjem J s obzirom na element h_{kn} i izjednačavanjem rezultata s nula, dobijemo pravilo ažuriranja:

$$h_{kn} = \tilde{h}_{kn} \left(\frac{p_{kn}}{q_{kn} + (\phi/\lambda_k)\tilde{h}_{kn}} \right)^{1/(\beta-1)}$$

Proučimo sada slučaj kada je $\beta \leq 2$:

Za razliku od prethodnog slučaja gdje smo odozgo ograničavali $R_2(H)$, sada ćemo ograničiti $G(H|\tilde{H})$.

Analogno prethodnom slučaju, prema lemi vrijedi:

$$\frac{1}{\beta} \left[\left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^\beta - 1 \right] \leq \frac{1}{2} \left[\left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^2 - 1 \right],$$

iz čeka raspisivanjem dobijemo:

$$\frac{1}{\beta} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^\beta \leq \frac{1}{2} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^2 + const.$$

β	$G(H \tilde{H})$
$\beta < 1$	$\sum_{kn} q_{kn} h_{kn} - \frac{1}{\beta-1} p_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta-1} + const.$
$\beta = 1$	$\sum_{kn} q_{kn} h_{kn} - p_{kn} \tilde{h}_{kn} \log\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right) + const.$
$\beta \in (1, 2]$	$\sum_{kn} \frac{1}{\beta} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta} - \frac{1}{\beta-1} p_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta-1} + const.$
$\beta > 2$	$\sum_{kn} \frac{1}{\beta} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta} - p_{kn} \tilde{h}_{kn} + const.$

Tablica 1: Pomoćne funkcije korištene za MM za β -NMF. Izvod se može vidjeti u $\| \cdot \|$

Stoga, J definiramo s:

$$J(H|\tilde{H}) = \sum_k \sum_n \frac{q_{kn} \tilde{h}_{kn}}{2\phi} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^2 - \frac{q_{kn} \tilde{h}_{kn}}{\phi(\beta-1)} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta-1} + \frac{h_{kn}^2}{2\lambda_k}$$

i lako vidimo da je to pomoćna funkcija za F .

Izjednačavanjem derivacije obzirom na element h_{kn} s 0, dobijemo pravilo ažuriranja:

$$h_{kn} = \tilde{h}_{kn} \left(\frac{p_{kn}}{q_{kn} + (\phi/\lambda_k) \tilde{h}_{kn}} \right)^{1/(3-\beta)}$$

Konačno, uz oznaku:

$$\xi(\beta) = \begin{cases} 1/(3-\beta), & \beta \leq 2 \\ 1/(\beta-1), & \beta > 2 \end{cases}$$

imamo matrični zapis pravila ažuriranja (u MATLAB notaciji):

$$\begin{aligned} H &\leftarrow H.*\left(\frac{W^T((WH)\cdot^2.*X)}{W^T(WH)\cdot^{(\beta-1)} + \phi H./repmat(\lambda, 1, N)}\right)^{\cdot\xi(\beta)} \\ W &\leftarrow W.*\left(\frac{((WH)\cdot^{(\beta-2)}.*X)H^T}{(WH)\cdot^{(\beta-1)}H^T + \phi W./repmat(\lambda, F, 1)}\right)^{\cdot\xi(\beta)} \end{aligned}$$

3.4.1 λ parametri

Prema [6] nakon ažuriranja W i H , novi λ_k dobijemo izjednačavanjem parcijalne derivacije $C(W, H, \lambda)$ po λ_k s nulom, te dobijemo:

$$\lambda_k = \frac{\frac{1}{2}\|w_k\|_2^2 + \frac{1}{2}\|h_k\|_2^2 + b}{c}.$$

No, možemo pokazati da je novi λ_k zapravo Bayesovski procjenitelj koji maksimizira aposteriornu distribuciju $p(\lambda_k|w_k, h_k)$.

$$\begin{aligned}
p(\lambda_k|w_k, h_k) &\propto p(w_k, h_k|\lambda_k)p(\lambda_k) \\
&= p(w_k|\lambda_k)p(h_k|\lambda_k)p(\lambda_k) \\
&= \prod_{i=1}^F \left[\left(\frac{2}{\pi\lambda_k} \right)^{1/2} \exp \left(-\frac{w_{ik}^2}{2\lambda_k} \right) \right] \prod_{j=1}^N \left[\left(\frac{2}{\pi\lambda_k} \right)^{1/2} \exp \left(-\frac{h_{kj}^2}{2\lambda_k} \right) \right] p(\lambda_k) \\
&\propto \left(\frac{2}{\pi\lambda_k} \right)^{(F+N)/2} \exp \left(-\frac{\|w_k\|_2^2 + \|h_k\|_2^2}{2\lambda_k} \right) \lambda_k^{-(a+1)} \exp \left(-\frac{b}{\lambda_k} \right) \\
&\propto \lambda_k^{-\frac{F+N}{2}-a-1} \exp \left(-\frac{1}{\lambda_k} (\frac{1}{2}\|w_k\|_2^2 + \frac{1}{2}\|h_k\|_2^2 + b) \right)
\end{aligned}$$

Iz prethodnog računa vidimo da je aposteriorna distribucija λ_k , s danim w_k i h_k , opet inverzna gama distribucija pa slijedi da je procjenitelj koji maksimizira aposteriornu distribuciju (mod funkcije gustoće) jednak:

$$\frac{\frac{1}{2}\|w_k\|_2^2 + \frac{1}{2}\|h_k\|_2^2 + b}{c}$$

3.4.2 Odabir parametra ϕ

Odabir parametra ϕ vezan je uz sami praktični problem koji se rješava odnosno uz distribuciju šuma u modelu. Ukoliko je šum normalno distribuiran tada je ϕ pripadna varijanca i može se procjeniti kao uzoračka varijanca ulazne matrice. Ako šum ima Poissonovu distribuciju tada je prikladan $\phi = 1$. Postoje pristupi u kojima se i tom parametru dodjeli prior pa se optimizacija provodi i po ϕ , ali to dodatno komplificira model. S obzirom da je iz samog oblika cost-funkcije koju promatramo jasno da manji ϕ jače penalizira samu udaljenost između X i WH , a veći broj komponenti, to treba imati na umu ako smo kao rezultat dobili premali broj komponenti ili lošu aproksimaciju polazne matrice.

3.4.3 Odabir parametara a i b

Prvo diskutirajmo odabir b za fiksan a . Iz zakona velikih brojeva i činjenice da je $\mathbb{E}[x_{fn} | \hat{x}_{fn}] = \hat{x}_{fn}$ (Pri čemu s x_{fn} označavamo fn -ti element matrice X dimenzije $F \times N$, a s \hat{x}_{fn} fn -ti element matrice WH):

$$\hat{\mu}_X \triangleq \frac{1}{FN} \sum_{fn} x_{fn} \approx \mathbb{E}[x_{fn}] = \mathbb{E}[\hat{x}_{fn}] = \sum_k \mathbb{E}[w_{fk}h_{kn}]$$

Koristimo da w_{fk} and h_{kn} uvjetno nezavisni uz dani λ_k :

$$\mathbb{E}[w_{fk}h_{kn}] = \mathbb{E}[\mathbb{E}[w_{fk}h_{kn} | \lambda_k]] = \mathbb{E}[\mathbb{E}[w_{fk} | \lambda_k] \mathbb{E}[h_{kn} | \lambda_k]].$$

Jer w_{fk} i h_{kn} imaju HF distribuciju s parametrom λ_k imamo:

$$\mathbb{E}[w_{fk} \mid \lambda_k] = \sqrt{\frac{2\lambda_k}{\pi}}, \quad \mathbb{E}[h_{kn} \mid \lambda_k] = \sqrt{\frac{2\lambda_k}{\pi}}.$$

Dakle

$$\mathbb{E}[w_{fk}h_{kn}] = \mathbb{E}\left[\frac{2\lambda_k}{\pi}\right] = \frac{2}{\pi} \mathbb{E}[\lambda_k].$$

Jer $\lambda_k \sim \text{Inv-Gamma}(a, b)$ slijedi

$$\mathbb{E}[\lambda_k] = \frac{b}{a-1}.$$

Dakle

$$\mathbb{E}[w_{fk}h_{kn}] = \frac{2}{\pi} \frac{b}{a-1}.$$

Vratimo u početnu jednakost:

$$\hat{\mu}_X \approx \sum_{k=1}^K \mathbb{E}[w_{fk}h_{kn}] = \sum_{k=1}^K \frac{2}{\pi} \frac{b}{a-1} = \frac{2Kb}{\pi(a-1)}.$$

Pa imamo izraz za b :

$$b = \hat{\mu}_X \frac{\pi(a-1)}{2K}.$$

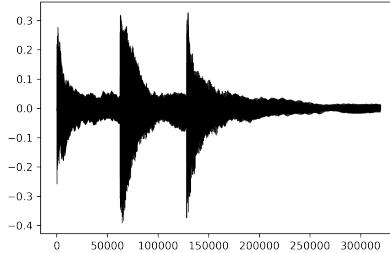
Odabir a s druge strane je dosta kompleksniji problem koji nije do kraja razriješen. Postoji pristup metodom momenata kao gore koji računa i drugi moment što daje sustav iz kojeg se može izračunati a , ali je problem da je izračunati a često izvan parametarskog prostora. Srećom je problem dosta stabilan s obzirom na izbor a jer a utječe kroz $c = \frac{F+N}{2} + a + 1$ i ako je a malen u odnosu na $F+N$ njegov doprinos je malen što ćemo kasnije vidjeti na primjerima.

4 Primjeri

4.1 Razdvajanje glazbenog zapisa u note

U ovom dijelu opisat ćemo primjenu NMF algoritma na audio zapise klavira s ciljem da dobijemo informacije o prisutnim notama i njihovim vremenima aktivacije.

Neka je x audiozapis spremlijen kako vektor duljine n . Kako bi dobili nenegativnu matricu X , nad x radimo ShortTime Fourierove transformacije (STFT). U STFT algoritmu x dijelimo na segmente duljine $L \ll n$ koje rastavljam diskretnim Fourierovim transformacijama, te rezultate tih transformacija spremamo u retke matrice Z . Kako je matrica Z kompleksna, nenegativnu matricu X dobivamo formulom: $X = |Z|^2$.



Slika 2: Originalni audiozapis s tri uzastopne note

Za funkciju cilja koristit ćemo Itakura-Saito divergenciju što odgovara $\beta = 0$ u ARD-NMF-u. Razlog tome je što je IS-divergencija invarijantna na skaliranje u smislu da vrijedi: $d_0(ax|ay) = d_0(x|y)$ za sve $a \in \mathbb{R}$. Drugim riječima, funkcija cilja će davati jednake relativne težine malim i velikim komponentama maticе X , što nam je važno zbog velikih raspona vrijednosti u zadanoj matici.

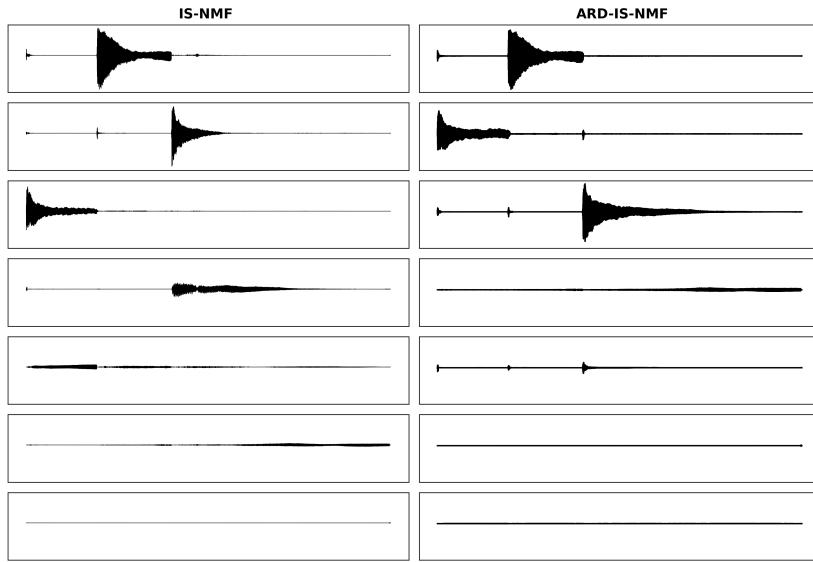
Za usporedbu, na svim audiozapisima je izvršen IS-NMF algoritam (bez detekcije relevantnih komponenti) dan u [2] i ARD-NMF algoritam obrađen u prethodnom poglavlju za $\beta = 0$. Svi algoritmi su izvršeni 10 puta uz iste početne matrice te je uzeto rješenje s najmanjom realizacijom pripadne funkcije cilja.

Svi kodovi za ovaj primjer nalaze se u datoteci *Usporedba algoritama za zvuk.ipynb*.

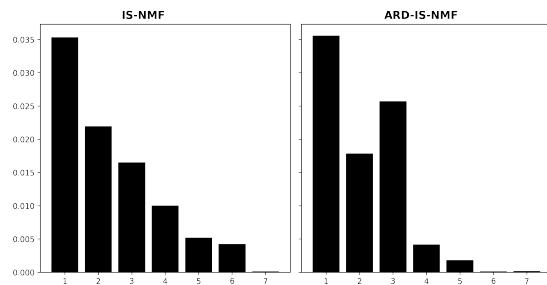
Tri uzastopne note

Za početak smo uzeli najjednostavniji audiozapis s tri različite, uzastopne note. Za računanje matrice X korišten je STFT algoritam s $L = 1024$, uz pomake $hop = 512$, dok su parametri ARD-NMF-a: $a = 250$, $\tau = 1e-3$ i $pruning_threshold = 10$ (dobiveni eksperimentom). Početni broj komponenta bio je 7.

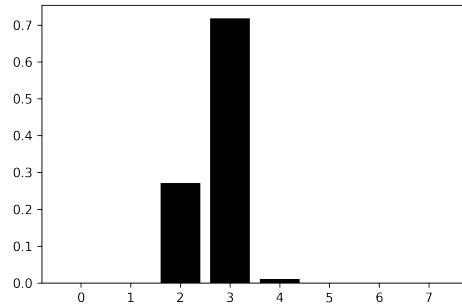
Na slici 3 prikazane su rekonstruirane komponente sortirane silazno po standardnoj devijaciji za IS-NMF i parametrima relevantnosti za ARD-NMF. U *Jupyter Notebook* datoteci nalaze se i audiozapisi rekonstruiranih komponenta.



Slika 3: Rekonstruirane komponente dobivene s IS-NMF i ARD-NMF algoritmom.



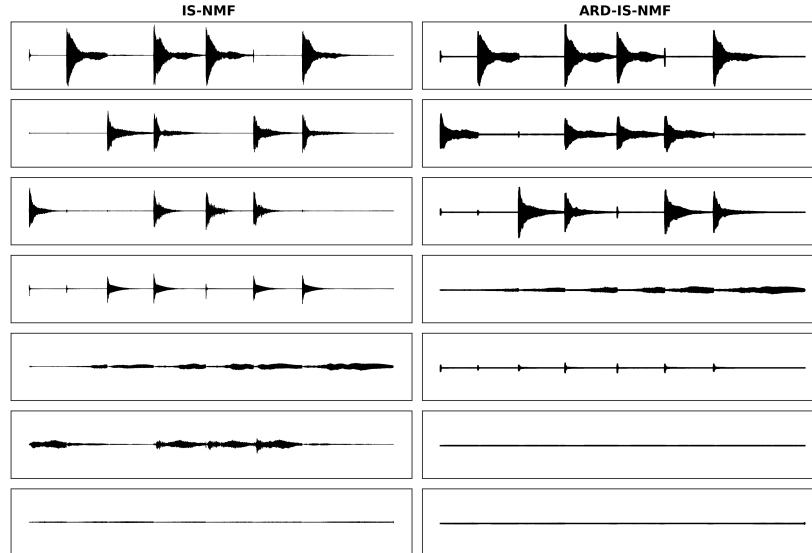
Slika 4: Standardne devijacije rekonstruiranih komponenti. Male standardne devijacije komponenti impliciraju da su one konstante pa ne pridonose rekonstrukciji signala. Ovdje vidimo kako standardne devijacije komponenta dobivenih s ARD-NMF-om padaju nakon 3. komponente.



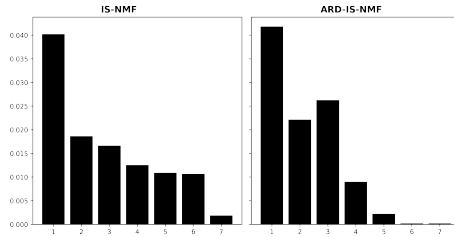
Slika 5: Izveli smo ARD-NMF algoritam 1000 puta, te smo na grafu prikazali histogram broja relevantnih komponenti. Vidimo kako u 70% slučajeva ARD-NMF završi s točnim brojem komponenta.

Akord

Ovaj audiozapis sadrži tri note kao i prethodni, ali uz pojedinačne aktivacije, note se pojavljuju i u svim parovima te kao akrod. Parametri su bili isti kao za prethodni primjer, te su rezultati prikazani na slikama 6 i 7.



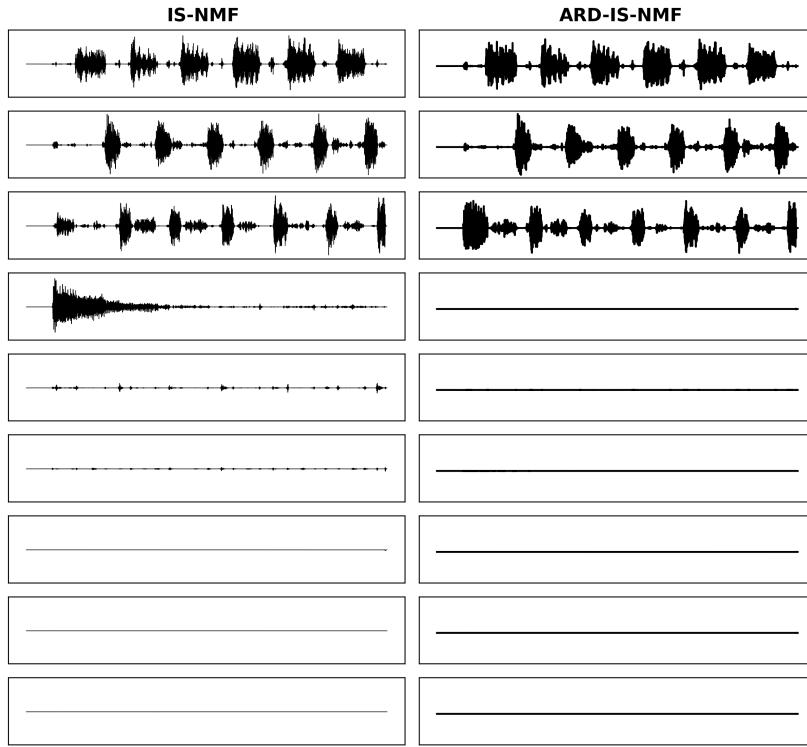
Slika 6: Rekonstrukcije akorda dobivene s IS-NMF i ARD-NMF algoritmima.



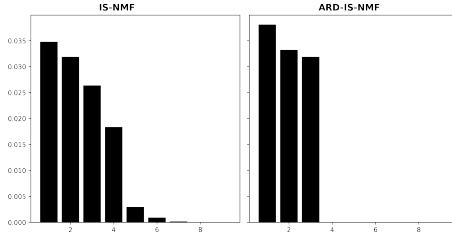
Slika 7: Standardne devijacije komponenata dobivenih za snimku akorda. Vidimo da standardne devijacije kod ARD-NMF-a padaju nakon 3. komponente, tj. one kasnije komponente postaju irelevantne.

The Ecstasy of Gold

Konačno, uzeli smo prvih 5 sekundi audiozapisa na klaviru izvedene skladbe "The Ecstasy of Gold" preuzete s YouTube-a. Početni broj komponenta bio je 10, dok su parametri za ARD-NMF bili: $a = 2$, $\tau = 1e-6$, $pruning_threshold = 10$ ponovno dobiveni eksperimentom. Rezultati su prikazani na slici 8.



Slika 8: Prikazane su rekonstruirane komponente "The Ecstasy of Gold" dobivene IS-NMF i ARD-NMF algoritmima.



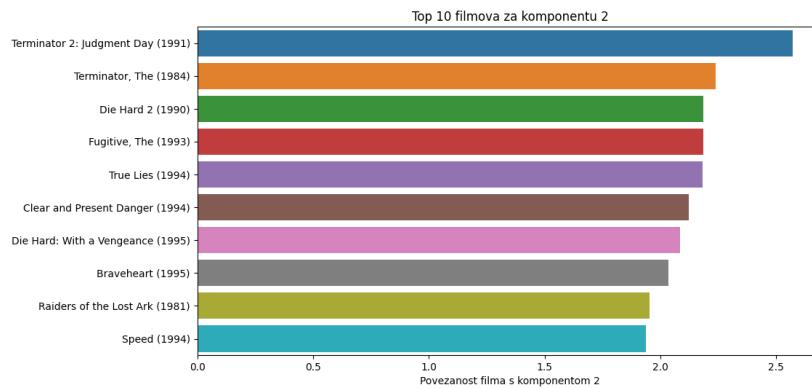
Slika 9: Standardne devijacije rekonstruiranih komponenta "The Ecstasy of Gold"

4.2 Pronalaženje latentnih komponenti u recenzijama filmova

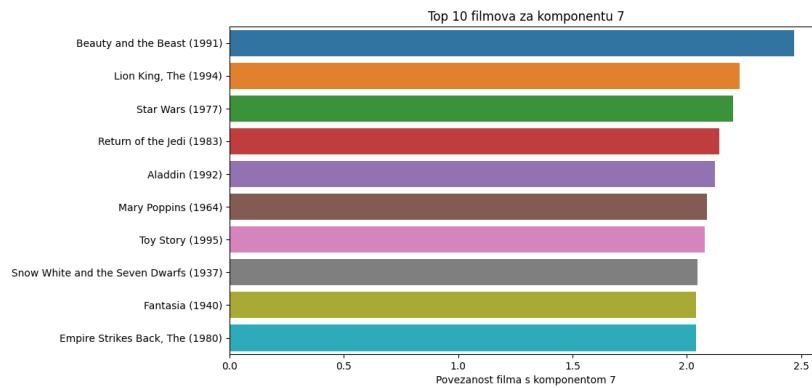
Primjenjujemo NMF algoritam na skup recenzija filmova korisnika stranice MovieLens. Podaci su preuzeti s Kaggle-a, radi se o recenzijama 1682 filma od strane 943 korisnika, od kojih je svaki korisnik ocijenio barem 20 filmova ocjenom 1-5. Matrica podataka X je tipa 943×1682 , gdje su retci pojedini korisnici, stupci filmovi, a X_{ij} je ocjena kojom je i -ti korisnik ocijenio j -ti film. 0 označava da korisnik nije ocijenio film. Korišten je osnovni ARD-NMF algoritam s $\beta = 2$, $a = 2$, što se testiranjem pokazalo najboljim. Početni broj komponenti bio je 30.

Prikaz pronađenih komponenti

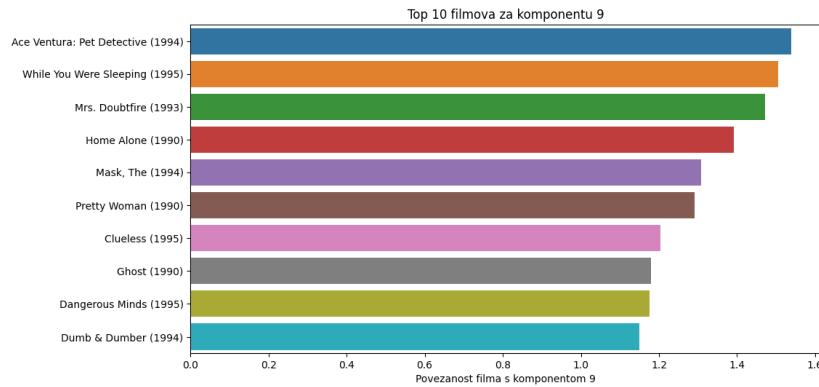
Nakon izvođenja faktorizacije, retci u matrici H , koja je tipa 17×1682 su komponente dobivene NMF-om. Element H_{ij} govori koliko j -ti film pripada i -toj komponenti. Gledajući deset najznačajnijih filmova za razne komponente, možemo jasno vidjeti da je algoritam uspio pronaći veze između njih, npr. akcijske filmove, animirane filmove, komedije, itd. stavio je u zasebne kategorije.



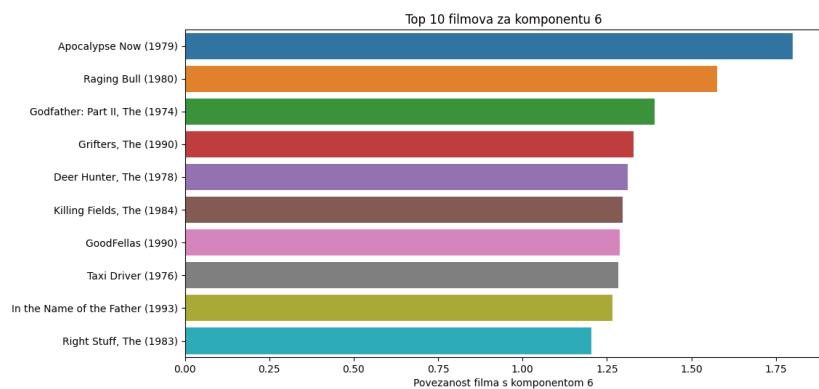
Slika 10: Ova komponenta jasno opisuje akcijske filmove



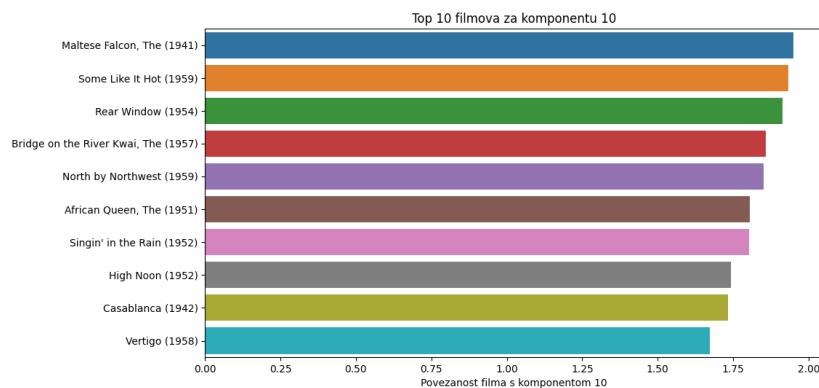
Slika 11: Ova komponenta jasno opisuje animirane/dječje filmove



Slika 12: Ova komponenta jasno opisuje komedije



Slika 13: Ova komponenta jasno opisuje trilere



Slika 14: Ova komponenta jasno opisuje klasične filmove

Predlaganje filmova korisniku

U matrici W je opisano koliko je dani korisnik povezan s nekom komponentom. Stoga je intuitivna predlaganje filmova s obzirom na to - red rekonstruirana matrica $(WH)_u$ su predviđene ocjene korisnika indeksa u - za sve filmove, pa i one koje nije ocijenio! Stoga na skupu filmova koje još nije ocijenio možemo pronaći one za koje model misli da će mu se najviše svidjeti:

- Top 10 ocijenjenih filmova:
1. Air Force One (1997)
 2. Unforgettable (1996)
 3. Murder at 1600 (1997)
 4. Just Cause (1995)
 5. Clear and Present Danger (1994)
 6. Net, The (1995)
 7. Client, The (1994)
 8. Tombstone (1993)
 9. Firm, The (1993)
 10. Murder in the First (1995)

Slika 15: Top 10 filmova koje je korisnik ocijenio

- Top 10 predloženih dosad neocijenjenih filmova:
1. Terminator 2: Judgment Day (1991)
 2. Eraser (1996)
 3. Executive Decision (1996)
 4. Titanic (1997)
 5. Terminator, The (1984)
 6. Time to Kill, A (1996)
 7. Phenomenon (1996)
 8. Braveheart (1995)
 9. Saint, The (1997)
 10. Courage Under Fire (1996)

Slika 16: Top 10 predloženih filmova

4.3 Analiza slika lica

U ovom poglavlju primjenjujemo ARD-NMF algoritam na 400 slika lica iz Pythonove biblioteke `sklearn.datasets`. Algoritam je izведен u 500 iteracija i dobivene su sljedeće komponente:



Slika 17: Komponente lica

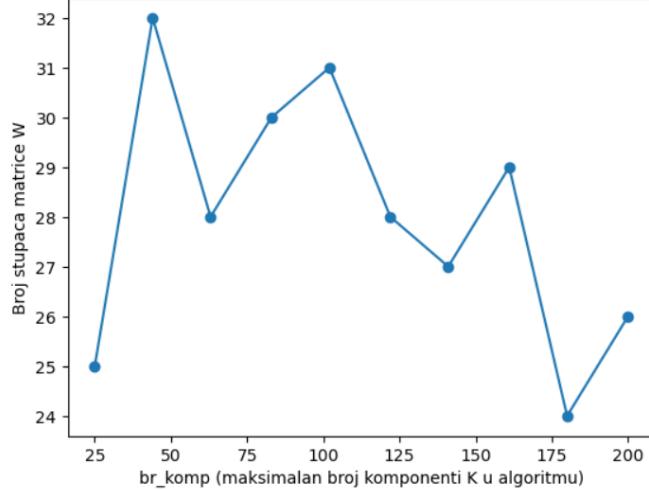
Sada smo za nekoliko slika rekonstrurali originalnu sliku iz aproksimacije:



Slika 18: Originalna i rekonstruirana lica

Vidimo da dosta dobro rekonstruiramo konture lica,a najveća mane su dosta veliki gubitak rezolucije i činjenica da se obrisi naočala pojavljuju na slikama bez naočala. Dalje nas zanima uspijeva li algoritam detektirati povoljan broj komponenti tj. r bez obzira koliko mu damo gornje ograničenje na r pa smo za deset različitih vrijednosti maksimalnog broj komponenti između 25 i 200 proveli algoritam i zabilježili izračunati r.

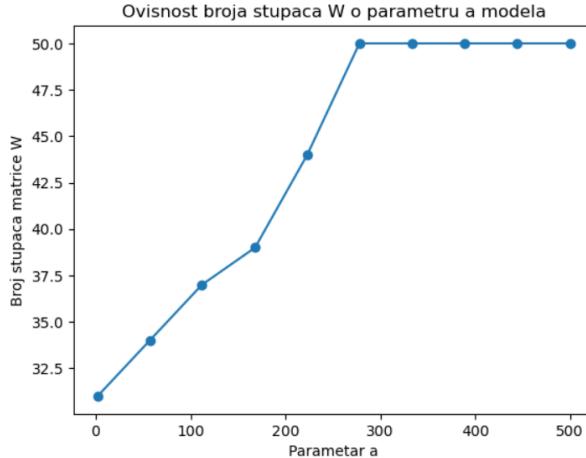
Ovisnost izračunatog broja komponenti (stupaca od W) o maksimalnom broju komponenti



Slika 19

Vidimo da bez obzira na gornje ograničenje na kraju završimo u relativno uskom intervalu od 24 do 32 komponente.

Nakon toga smo promatrali stabilnost s obzirom na promjenu parametra a za kojeg smo već ranije komentirali da nemamo razvijenih dobrih metoda za procjenu



Slika 20

Naš rezultat je konzistentan s rezultatima iz [2] da utjecaj a nije velik dok je a relativno malen u odnosu na dimenzije matrice. Kada a previše povećamo on

voli "povući" rang prema našem zadanim maksimumu što naravno nije dobro (ovdje je maksimalan broj komponenti bio 50):

5 Zaključak

U ovom radu smo izveli i testirali dva algoritma za računanje nenegativnih matričnih faktorizacija. Na primjerima audiozapisa, testirali smo točnost odabira ranga algoritmom ARD-NMF te smo u tom konkretnom primjeru dobili dobru preciznost. Na primjeru slike lica smo testirali utjecaj parametra a za kojeg nismo pronašli dobre teorijska procjene, te smo dobili da je problem relativno stabilan dok je a "malen" u odnosu na dimenzije matrice X . Također, na našem primjeru je određeni rang bio nezavisan od početne vrijednosti ranga u algoritmu. U primjeru s rezenzijama filmova, prikazali smo snagu NMF-a u klasteriranju filmova.

Literatura

- [1] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and applications for approximate non-negative matrix factorization. *Computational Statistics Data Analysis*, 52(1):155–173, 2007.
- [2] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.
- [3] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β -divergence, 2011.
- [4] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.
- [5] Vincent Y. F. Tan and Cédric Févotte. Supplementary material to “ automatic relevance determination in nonnegative matrix factorization with the β -divergence ”, 2012.
- [6] Vincent Y. F. Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization with the β -divergence, 2012.