# Nenegativne matrične faktorizacije

**Jerončić, Dajana**

# Nenegativne matrične faktorizacije

**Jerončić, Dajana**

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Dajana Jerončić

# NENEGATIVNE MATRIČNE FAKTORIZACIJE

Diplomski rad

Voditelj rada:
prof. dr. sc. Luka Grubišić

Zagreb, studeni 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

# Sadržaj

# Uvod

Razvojem tehnologije, a posebice umjetne inteligencije, broj podataka koji se treba pohraniti te analizirati enormno raste. Kao jedan od načina kompresije podataka ističu se matrične aproksimacije, te nenegativne matrične faktorizacije kao jedna podvrsta istih. Upravo zbog toga, ali i zbog raznih drugih primjena osim pukog smanjenja dimenzionalnosti, popularnost nenegativnih matričnih faktorizacija zadnjih je godina u stalnom porastu. Za razliku od mnogih drugih metoda, čuva nenegativnost podataka te je upravo zbog tog svojstva te interpretabilnosti rezultata pronašla svoj put u mnoga područja gdje je nenegativnost bitno svojstvo, kao što su bioinformatika, glazba, astronomija te mnoga druga.

Glavna ideja je nenegativnu matricu aproksimirati umnoškom dviju nenegativnih matrica dimenzija manjih od početne. Dakle, govorimo o svojevrsnoj redukciji dimenzionalnosti. Stoga će za početak biti opisana upravo ovakva potreba za smanjenjem dimenzionalnosti, odnosno transformiranjem podataka iz visoko-dimenzionalnog prostora u onaj niže dimenzije te će detaljnije biti opisana linearna redukcija dimenzionalnosti kao nadskup nenegativnih matričnih faktorizacija. Također će biti dana usporedba s nekim drugim metodama redukcije dimenzionalnosti, gdje nenegativne matrične faktorizacije prednjače kod transformiranja nenegativnih podataka s obzirom da čuvaju samu prirodu podataka. Osim toga, pronalaze latentnu strukturu u podacima te omogućuju lakšu interpretabilnost rezultata od mnogih drugih metoda.

Nakon iznijete potrebe za nenegativnim matričnim faktorizacijama, slijedi formalni opis problema kao i neka njegova svojstva. Jedno od bitnijih je nejedinstvenost rješenja zbog čega će biti proučena svojstva lokalnog minimuma. S obzirom da matricu $X$ dimenzija $m \times n$ zapisujemo kao umnožak matrica $W$ i $H$ dimenzija $m \times r$ i $r \times n$ redom, ključan je odabir dimenzije $r$, odnosno reduciranog ranga. Upravo zbog toga će biti dan opis nekoliko metoda za procjenu istog, nakon čega slijedi nekoliko varijacija nenegativnih matričnih faktorizacija uz ograničenja kao što su raspršenost i ortogonalnost, a koja imaju kao cilj zadovoljiti neke specifične zahtjeve kod aproksimacije.

S obzirom da je pojam nenegativnih matričnih faktorizacija vezan za općenitu ideju aproksimacije umnoškom nenegativnih matrica, potrebni su specificirani algoritmi kojima se mogu dobiti isključivo nenegativni faktori. U poglavlju o postojećim algoritmima bit će dan opis nekih najpopularnijih, bilo to zbog toga što su nezahtjevni za implementirati

te interpretabilni (kao što je slučaj kod algoritma Hadamardovog produkta) ili jer su bili dani kao prvo rješenje problema nenegativnih matričnih faktorizacija (kao što je slučaj kod algoritma alternirajućih najmanjih kvadrata). Osim same srži algoritma, bitna je i inicijalizacija početnih matrica koje se dalje osvježavaju do konačnog rješenja, te sam kriterij zaustavljanja, što će biti opisano na kraju poglavlja.

Zatim slijedi opis nekoliko popularnih primjena nenegativnih matričnih faktorizacija. S obzirom da su neke vrste podataka isključivo nenegativne, kao što je slučaj kod piksela slike, nivoa ekspresije gena te brojanja entiteta kao što su riječi u dokumentima, primjene su raznovrsne. Neke od njih uključuju klasteriranje, filtriranje i separaciju izvora zvuka, modeliranje tema, stoga će biti opisane primjene kod tekstualne analize, obrade slike, analize glazbe te sustava za preporuku.

Jedna od novijih te zanimljivijih primjena je kod predviđanja nove veze u mreži, stoga ćemo posljednja dva poglavlja posvetiti upravo toj temi. U prvom od ta dva poglavlja bit će opisan sam problem predviđanja veze te dani primjeri postojećih rješenja pomoću nenegativnih matričnih faktorizacija. Neka od njih koriste vanjske atribute mreže, dok jedno od rješenja kao glavni adut koristi perturbaciju matrica.

Za kraj, bit će opisan konkretni algoritam za rješavanje problema predviđanja nove veze u mreži koautorstava CROSBI korištenjem nenegativnih matričnih faktorizacija uz kombiniranje perturbiranosti matrica i dodatnih atributa zasnovanih na topologiji mreže. Rezultati su uspoređeni s onima klasničnih metoda te evaluirani koristeći mjere bitne za predviđanje novih veza - preciznost i AUC.

Iz samog ovog uvoda možemo zaključiti da, iako naizgled jednostavnog naziva, nenegativne matrične faktorizacije iza svog imena kriju mnoštvo algoritama te načina i područja primjene. Zasigurno će spoznavanjem novih činjenica o ovoj metodi ta svestranost samo rasti te naći još mnoštvo novih primjena i algoritama.

# Poglavlje 1

# Redukcija dimenzionalnosti

U današnjem svijetu broj podataka raste strahovitom brzinom, od hrpe podataka koja se gomila na društvenim mrežama pa do onih nastalih pomoću umjetne inteligencije. Pretpostavlja se da je samo na internetu 2018. godine količina tih podataka iznosila 18 zetabajta[40]. Ako uzmemo u obzir da zetabajti iza gigabajta slijede tek nakon terabajta, petabajta i eksabajta, tek tada možemo pokušati percipirati tu prostranost podataka. Svi ti podaci trebaju se nekako i reprezentirati, a velik dio njih može se prikazati upravo matricama.

Uzmimo da imamo neku matricu $X \in \mathbb{R}^{m \times n}$. Tada $n$ može biti broj određenih uzoraka, korisnika, dokumenata, a stupci bi tada predstavljali $m-$dimenzionalne podatke koje odgovaraju tim jedinkama. U velikom broju slučaja stupci, a i retci, mogu se mjeriti u stotinama tisuća. Obraditi te podatke, tj. izračunati bilo kakvu statistiku ili dobiti nekakvi zaključak iz te matrice zahtijevalo bi puno više vremena nego je korisnik spreman čekati.

Redukcija dimenzionalnosti pomaže upravo u tome: transformira podatke iz visokodimenzionalnog prostora u prostor nižeg ranga pritom čuvajući neka svojstva originalnih podataka. Ovisno o korištenoj vrsti transformacije, dijeli se na linearne i nelinearne metode. Optimalna preslikavanja bi trebala biti nelinearnog karaktera, međutim, tradicionalno se koriste linearne transformacije zbog manje složenosti.

Jedna od metoda linearne redukcije dimenzionalnosti je upravo matrična aproksimacija koja nam daje mogućnost da veliku matricu $X$ zapišemo kompaktnije kao umnožak dviju manjih matrica $W \in \mathbb{R}^{m \times r}$ i $H \in \mathbb{R}^{r \times n}$ pri čemu je $r$ značajno manji od $m$ i $n$. U tom slučaju umjesto da obrađujemo $m \times n$ zapisa, dovoljno je obraditi $(m + n) \times r$ zapisa. Tako ne samo da je smanjena količina memorije potrebna za pohraniti originalnu matricu, nego je ubrzano i vrijeme izvođenja pojedinih operacija. Nadalje, često aproksimiranje može pokazati i više nego polazna matrica s obzirom da pronalazi latentnu strukturu originalne matrice. S obzirom da podaci koji nastanu kao rezultat mjerenja antenama, senzorima te drugim mjernim uređajima nisu egzaktni, često ih je vrlo bitno reprezentirati kompresirano kako bi se smanjila njihova dvoznačnost.

*Napomena*: Primijetimo da smo ovdje koristili izraz matrične aproksimacije. U izrazu *nenegativna matrična faktorizacija*, riječ faktorizacija zapravo se odnosi na aproksimaciju, tj. nije riječ o egzaktnoj matričnoj faktorizaciji kao što je egzaktna dekompozicija matrice kakvu imamo kod LU faktorizacije ili Cholesky faktorizacije. Međutim, kod nenegativnih matričnih faktorizacija izraz faktorizacije je postao uvriježen te se kao takav i koristi.

## 1.1 Linearna redukcija dimenzionalnosti

S obzirom da su nenegativne matrične faktorizacije podskup linearne redukcije dimenzionalnosti (*eng. linear dimensionality reduction*, *LDR*), u ovom odjeljku bit će opisan LDR problem, odnosno redukcija dimenzionalnosti linearnim transformacijama podataka te dano nekoliko primjera metoda koje spadaju u ovaj skup aproksimacija. Metode za LDR koriste se za kompresiju, vizualizaciju, odabir značajki te filtriranje šuma. Slijedi opis problema kako je opisano u [27].

**Definicija 1.1.1.** *Neka je dan skup uzoraka $x_j \in \mathbb{R}^m$ za $1 \leq j \leq n$ i dimenzija $r < min(m, n)$. Linearna redukcija dimenzionalnosti pronalazi skup $r$ baznih elemenata $w_k \in \mathbb{R}^m$ za $1 \leq k \leq r$ tako da linearni prostor razapet s $w_k$ aproksimira uzorke $x_j$ najbolje moguće, i da za svaki $j$ vrijedi*

$$x_j \approx \sum_{k=1}^{r} w_k h_j(k), \qquad \text{za neke težine } h_j \in \mathbb{R}^r. \tag{1.1}$$

Drugim riječima, $m$-dimenzionalni uzorci $x_j$ reprezentirani su u $r$-dimenzionalnom linearnom potprostoru razapetom elementima $w_k$ čije su koordinate dane vektorima $h_j$. LDR je ekvivalentan aproksimaciji matricom nižeg ranga, odnosno, konstruiranjem:

- matrice $X \in \mathbb{R}^{m \times n}$ tako da je svaki stupac jedan uzorak, odnosno $X(:, j) = x_j$ za $1 \leq j \leq n$,
- matrice $W \in \mathbb{R}^{m \times r}$ tako da je svaki stupac bazni element, odnosno $W(:, k) = w_k$ za $1 \leq k \leq r$, i
- matrice $H \in \mathbb{R}^{r \times n}$ tako da svaki stupac od $H$ daje koordinate uzorka $X(:, j)$ u bazi $W$, odnosno $H(:, j) = h_j$ za $1 \leq j \leq n$.

Gornji LDR model (1.1) ekvivalentan je $X \approx WH$, odnosno aproksimaciji matrice podataka $X$ matricom $WH$ gdje su matrice $W$ i $H$ nižeg ranga od matrice $X$.

## Mjera za procjenu kvalitete aproksimacije

Prvo što treba uzeti u obzir kod LDR-a je odabir mjere za procjenu kvalitete aproksimacije. Kod odabira, trebalo bi uzeti u obzir model šuma. Najčešće korištena mjera je Frobeniusova norma pogreške, odnosno

$$\|X - WH\|_F^2 = \sum_{i,j} (X - WH)_{ij}^2. \tag{1.2}$$

Razlog korištenja (1.2) je dvojak. Prvo, implicitno pretpostavlja da je šum $N$ prisutan u matrici $X = WH + N$ Gaussove distribucije, što je razumno u mnogim praktičnim primjenama. Drugo, optimalna aproksimacija može biti učinkovito dobivena pomoću skraćenog oblika dekompozicije matrice na singularne vrijednosti (*eng. truncated singular value decomposition (truncated SVD)*). Odnosno, uzmimo da za matricu $A \in \mathbb{R}^{m\times n}$ dekompozicija na singularne vrijednosti glasi $A = U \sum V^\top$, što može biti zapisano kao

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^\top \tag{1.3}$$

gdje su $\sigma_i$ singularne vrijednosti i $p$ broj singularnih vrijednosti koje su različite od nule. Nadalje, $u_i$ i $v_i$ su $i$-ti stupci od $U$ i $V$ redom. Singularne vrijednosti su nenegativne, poredane u padajućem redoslijedu.

Tada možemo aproksimirati matricu $A$ tako da uzmemo samo $k$ najvećih singularnih vrijednosti i njihove pripadne vektore, tj. imamo

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top. \tag{1.4}$$

Eckart-Young-Mirksy teorem tvrdi da je ta matrica ujedno i najbolja aproksimacija za Frobeniusovu normu (iskaz prvi put dan u [20]).

**Teorem 1.1.2.** *(Eckart-Young-Mirsky) Za svaku matricu $X \in \mathbb{R}^{m\times n}$ ranga $k$ vrijedi*

$$\|X - A\|_F^2 \geq \sigma_{k+1}^2 + \ldots + \sigma_p^2, \tag{1.5}$$

*a jednakost vrijedi kada je $X = A_k$. Nadalje, ako je $\sigma_k > \sigma_{k+1}$, $A_k$ je jedina matrica koja zadovoljava jednakost.*

## Struktura faktora $W$ i $H$

Bitno je promotriti i prirodu faktora $W$ i $H$. Skraćeni SVD i PCA ne zahtijevaju nikakvu pretpostavku nad $W$ i $H$. Različiti zahtjevi nad faktorima $W$ i $H$ vode k različitim metodama rješavanja, primjerice, ako aproksimiramo $X$ s ciljem da su stupci od $W$ nezavisni,

tada to vodi k analizi nezavisnih komponenti, zahtjev da je popunjenost matrica $W$ ili $H$ rijetka vodio bi k rijetkoj matričnoj dekompoziciji nižeg ranga, kao što je PCA za rijetko popunjene matrice [54].

Međutim, postoje problemi kod kojih su podaci isključivo nenegativni. Kako bismo sačuvali prirodu problema, izbjegli kontradikcije i olakšali intrepretaciju aproksimacije, želimo sačuvati nenegativnost faktora $W$ i $H$. Standardni pristup SVD i PCA ne može garantirati očuvanje nenegativnosti. Upravo zbog toga dolazimo do potrebe za nenegativnim matričnim faktorizacijama koje postavljaju uvjet nenegativnosti na faktore $W$ i $H$.

# Poglavlje 2

# Nenegativne matrične faktorizacije

U prošlom poglavlju iznijeli smo potrebu aproksimacije matricama nižeg ranga te spomenuli nekoliko metoda aproksimacije. Međutim, često podaci koje želimo reprezentirati su nenegativni, kao što su intenzitet piksela slike, frekvencija riječi u dokumentu, astrofizički signali, nivo ekspresije gena te mnogi drugi, o čemu će više riječi biti u poglavlju 4.

To nas dovodi do potrebe za specijaliziranijem pristupu rješavanja redukcije dimenzionalnosti. Stoga ni ne čudi da se NMF pojavio već 1994. kada su ga Paatero i Tapper prvi put uveli u [44] koristeći izraz *pozitivna matrična faktorizacija*. Ali tek nakon što su 1999. Lee i Seung objavili rad [34], NMF je dobio na popularnosti i postao lakše razumljiv. U ovom poglavlju bit će dan formalni opis problema te analizirana svojstva i rješivost NMF-a kao i odabir reduciranog ranga, a za kraj će biti dan pregled češće korištenih varijacija NMF-a uz dodatna ograničenja.

## 2.1 Opis problema

Slijedi formalni opis problema nenegativne matrične faktorizacije, kako je opisan u [4].

**Definicija 2.1.1.** *Za danu nenegativnu matricu* $X \in \mathbb{R}^{m\times n}$ *i pozitivni cijeli broj* $r < \min(m, n)$*, pronađi nenegativne matrice* $W \in \mathbb{R}^{m\times r}$ *i* $H \in \mathbb{R}^{r\times n}$ *kako bi se minimizirao izraz*

$$f(W, H) = \frac{1}{2}||X - WH||_F^2. \tag{2.1}$$

*Umnožak* WH *se naziva* nenegativna matrična faktorizacija od X*, iako* X *nije nužno jednak produktu* WH*, tj.*

$$X \approx WH. \tag{2.2}$$

**Napomena 2.1.2.** *Rang* r *iz definicije 2.1.1 nazivamo* reducirani rang.

**Napomena 2.1.3.** *Primijetimo da smo u definiciji NMF problema koristili Frobeniusovu matričnu normu, najpopularniji oblik NMF modela. Međutim, nije uvijek razumno pretpostaviti Gaussov šum za nenegativne podatke, posebno za rijetko popunjene matrice kao što su skupovi dokumenata. U praksi se koristi i velik broj drugih funkcija cilja kao što su Itakura-Saito udaljenost za analizu glazbe, $\ell_1$ norma za poboljšanje robusnosti prema odstupanjima te mnoge druge. Jedna od značajnijih mjera je Kullback-Leibler divergencija koja se često koristi u analizi teksta, a dana je izrazom*

$$D(A||B) = \sum_{ij}(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}). \tag{2.3}$$

*S obzirom da ova mjera nije simetrična u $A$ i $B$, ne naziva se udaljenosti, već divergencijom.*

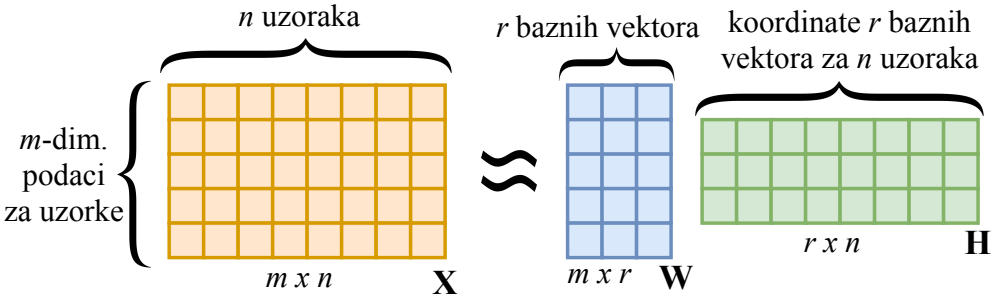
*Sada se NMF problem svodi na minimiziranje Kullback-Leibler divergencije između $X$ i $WH$:*

$$\min_{W\geq 0, H\geq 0}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(X_{ij}\log\frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij}\right). \tag{2.4}$$

*Od sada pa nadalje ćemo promatrati NMF problem uz Frobeniusovu normu, s obzirom da je to najraširenija mjera za NMF te zbog jednostavnosti dokaza.*

## Vizualizacija problema

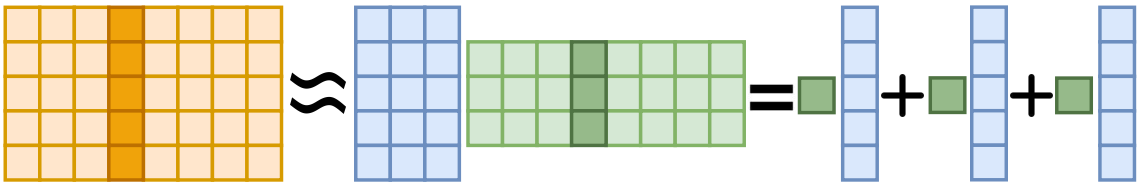
Za vizualizirati dani problem, možemo uzeti da je $X$ skup podataka gdje imamo $n$ uzoraka i za svaki gledamo $m$ karakteristika, tj. svaki stupac matrice $X$ je skup $m$-dimenzionalnih podataka koji odgovaraju nekom od $n$ uzoraka (Slika 2.1).




Slika 2.1: Vizualizacija NMF-a

Svaki stupac, odnosno uzorak matrice $X$ dobijemo kao težinsku sumu baznih vektora iz $W$, tj. koeficijenti iz matrice $H$ daju težinske vrijednosti za $r$ baznih vektora iz $W$, što je prikazano izrazom (2.5), odnosno slikom 2.2.

$$X(:, j) \approx \sum_{k=1}^{r} W(:, k)H(k, j). \tag{2.5}$$



Slika 2.2: Stupac matrice $X$ kao težinska suma baznih vektora iz $W$

Upravo je to razlog zašto se matrica $W$ naziva matricom baza, dok je $H$ matrica koeficijenata.

**Napomena 2.1.4.** *Ako retke matrice $X$ promatramo kao uzorke, tada možemo zamijeniti uloge i svaki redak $X(i, :)$ promatrati kao težinsku sumu baznih vektora, no u tom slučaju svaki redak iz matrice $H$ je bazni vektor, dok je svaki redak matrice $W$ težina baznog vektora, dakle, uloge matrice baza i matrice koeficijenata su zamijenjene.*

Geometrijski, nenegativna matrična faktorizacija može se interpretirati kao problem pronalaska *pojednostavljenog stošca* koji sadržava skup točaka i koji je sadržan u pozitivnom ortantu, kako je opisano u [37].

**Definicija 2.1.5.** *Pojednostavljeni stožac generiran s $W$ je skup*

$$C_W = \left\{ x : x = \sum_{i}^{r} h_i W_i,\ h_i \in \mathbb{R}_+ \right\}. \tag{2.6}$$

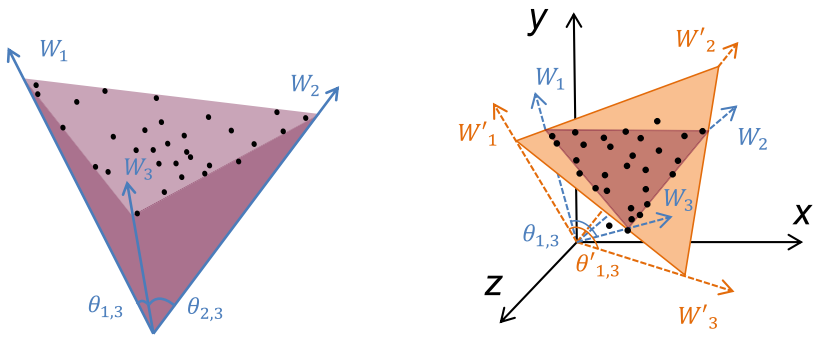
**Definicija 2.1.6.** *Ekstremna zraka pojednostavljenog stošca $C$ je zraka*

$$R_x = \{ax : a \geq 0\}, \tag{2.7}$$

*gdje se $x \in C$ ne može reprezentirati kao odgovarajuća konveksna kombinacija neke dvije točke $x_0$ i $x_1$ koje pripadaju $C$-u, ali ne i $R_x$.*

Cilj NMF-a je pronaći pojednostavljeni stožac $C_W$ koji sadržava $\mathcal{X} \subseteq \mathbb{R}^m$, kako je prikazano na slici 2.3. Primijetimo da je traženi stožac unutar nenegativnog ortanta.

Prema definiciji 2.1.6, ekstremna zraka ima beskonačan broj reprezentacija zbog skalarnog parametra $a$. Isto vrijedi i za bazne vektore za NMF, jer je $WH = WQ^{-1}QH$, gdje je




Slika 2.3: NMF problem kao pronalazak pojednostavljenog stošca koji sadrži $\mathcal{X}$ (lijevo) i usporedba stožaca razapetih s $W'$ i $W$ (desno)

$Q \in \mathbb{R}^{r\times r}$ invertibilna matrica, tako da vrijedi $WQ > 0$ i $Q^{-1}H > 0$. Možemo normalizirati bazne vektore iz $W$ odabirom

$$Q = \begin{pmatrix} \|W_1\| & & & \\ & \|W_2\| & & \\ & & \ddots & \\ & & & \|W_r\| \end{pmatrix} \tag{2.8}$$

kako bismo ograničili izbor od $W$ bez mijenjanja mogućnosti reprezentacije rješenja NMF problema.

Općenito, za dani skup podataka, bit će mnogo mogućih pojednostavljenih stožaca koji sadrže sve točke tog skupa. Ako je $C_W$ jedan stožac koji sadrži $X$ i $C_{W'}$ drugi stožac koji sadrži prvi stožac, tj. ako vrijedi $C_W \subset C_{W'}$, tada $C_{W'}$ također može služiti za reprezentaciju skupa podataka $X$.

## 2.2 Postojanje i jedinstvenost rješenja

Prvo primijetimo da $f(W,H)$ iz (2.1) nije konveksna istovremeno za $X$ i $H$. Uzmimo u obzir skalarni slučaj, tj. $m = n = 1$. Tada imamo:

$$\min_{w,h\geq 0}(x - wh)^2 = \min_{w,h\geq 0} x^2 - 2xwh + w^2h^2. \tag{2.9}$$

Tada možemo izračunati gradijent i Hessijan funkcije $\phi_x(w,h) = x^2 - 2xwh + w^2h^2$:

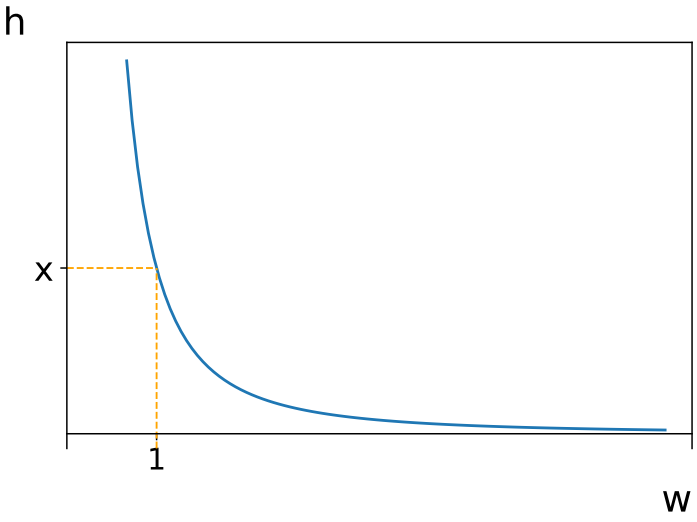
$$\nabla\phi_x(w,h) = \begin{bmatrix} 2wh^2 - 2xh \\ 2w^2h - 2xw \end{bmatrix} \tag{2.10}$$

$$\nabla^2 \phi_x(w,h) = \begin{bmatrix} 2h^2 & 4wh - 2x \\ 4wh - 2x & 2w^2 \end{bmatrix} \tag{2.11}$$

Hessijan nije pozitivno semidefinitan za sve $x, w, h \geq 0$, npr. za $(x, w, h) = (1, 2, 1)$ imamo:

$$\nabla^2 \phi_1(2,1) = \begin{bmatrix} 2 & 6 \\ 6 & 8 \end{bmatrix}, \quad \lambda_{\min}(\nabla^2 \phi_1(2,1)) = -1.7082 < 0 \tag{2.12}$$

Štoviše, za skalarni slučaj uzimamo u obzir samo egzaktne aproksimacije $(x - wh)^2 = 0$, i tada imamo beskonačan broj rješenja $x = wh$. Ako postavimo uvjet normiranosti za $w$, što bi u ovom slučaju značilo $\|w\| = 1$, tada bi imali točno jedno rješenje $(w, h) = (1, x)$ (Slika 2.4).




Slika 2.4: Graf $x = wh$ kao skup rješenja za skalarni NMF s istaknutim normiranim rješenjem $(1, x)$

Dakle, bitno je pitanje postojanja lokalnog minimum zbog nekonveksnosti $f(W, H)$ u $W$ i $H$ istovremeno, ali i nepostojanje jedinstvenog rješenja.

Za primjer, uzmimo da je $(W, H)$ jedno rješenje. Ako uzmemo neku nenegativnu matricu $D$ reda $r$ čiji je inverz također nenegativna matrica, tada je također i $(WD, D^{-1}H)$ rješenje jer vrijedi

$$WDD^{-1}H = WH \approx X, \quad WD \geq 0, D^{-1}H \geq 0. \tag{2.13}$$

(Npr. za matricu $D$ možemo uzeti bilo koju dijagonalnu matricu s nenegativnim vrijednostima.)

## Stacionarna točka

Izvedimo sada uvjete za stacionarnu točku NMF-a kako je opisano u [31]. Napišimo prvo izraz (2.1) u obliku standardnog nelinearnog optimizacijskog problema:

$$\min_{-W\leq 0,-H\leq 0} \frac{1}{2}\|X - WH\|_F^2. \tag{2.14}$$

Tada je pripadna Lagrangeova funkcija jednaka

$$L(W, H, \lambda, \mu) = \frac{1}{2}\|X - WH\|_F^2 - \lambda \circ W - \mu \circ H, \tag{2.15}$$

gdje su $\lambda$ i $\mu$ matrice jednakih dimenzija kao i $W$ i $H$ redom, a sadrže Lagrangeove multiplikatore. Prema Karush-Kuhn-Tucker uvjetima za problem nenegativne matrične faktorizacije, ako je $(W, H)$ lokalni minimum, tada postoje $\lambda_{ij} \geq 0$ i $\mu_{ij} \geq 0$ tako da:

$$W \geq 0, \quad H \geq 0 \tag{2.16}$$

$$\nabla L_W = 0, \quad \nabla L_H = 0 \tag{2.17}$$

$$\lambda \circ W = 0, \quad \mu \circ H = 0. \tag{2.18}$$

Raspisivanjem (2.17) dobijemo:

$$XH^\top - WHH^\top - \lambda = 0, \quad W^\top X - W^\top WH - \mu = 0 \tag{2.19}$$

odnosno

$$\lambda = -(WHH^\top - XH^\top), \quad \mu = -(W^\top WH - W^\top X). \tag{2.20}$$

Kombiniranjem ovoga s $\lambda_{ij} \geq 0$ i $\mu_{ij} \geq 0$ i (2.18) dobijemo sljedeće uvjete:

$$W \geq 0, \quad H \geq 0, \tag{2.21}$$

$$\nabla F_W = WHH^\top - XH^\top \geq 0, \quad \nabla F_H = W^\top WH - W^\top X \geq 0, \tag{2.22}$$

$$W \circ (WHH^\top - XH^\top) = 0, \quad H \circ (W^\top WH - W^\top X) = 0, \tag{2.23}$$

gdje su odgovarajući Lagrangeovi multiplikatori $\lambda$ i $\mu$ za $W$ i $H$ također i gradijenti od $F$ s obzirom na $W$ i $H$ redom.

**Napomena 2.2.1.** *S obzirom da Frobeniusova mjera nije konveksna s obzirom na obje varijable* W *i* H *istovremeno, ovi uvjeti su samo nužni.*

**Definicija 2.2.2.** (W,H) *je* stacionarna točka *za NMF ako i samo ako* W *i* H *zadovoljavaju KKT uvjete* (2.21)*,* (2.22)*,* (2.23)*.*

### Ekvivalentnost aproksimacija

U poglavlju *Postojanje i jedinstvenost rješenja* u izrazu (2.13) vidjeli smo da rješenje nije jednostruko. Štoviše, za skalarni slučaj smo ograničili $w$ s $\|w\| = 1$ kako bismo osigurali jedinstveno rješenje. Za poopćiti to ograničenje u više dimenzija, možemo ograničiti matricu $W$ tako da svaki njen vektor-stupac mora biti normiran, tj. da vrijedi $\|W(:,i)\|_2 = 1$. Međutim, to sada više ne garantira jedinstvenost aproksimacije. Za daljnja promatranja definirajmo prvo kada su dvije aproksimacije ekvivalentne kako je definirano u [31].

**Definicija 2.2.3.** *Reći ćemo da su dvije aproksimacije* $(\mathrm{W}, \mathrm{H})$ *i* $(\hat{\mathrm{W}}, \hat{\mathrm{H}})$ ekvivalentne *ako i samo ako daju jednak produkt, tj.*

$$\mathrm{WH} = \hat{\mathrm{W}}\hat{\mathrm{H}}. \tag{2.24}$$

Sada se možemo pitati, ako imamo stacionarnu točku $(W, H)$, ukoliko možemo pronaći invertibilnu matricu $S$ tako da vrijedi $\hat{W} = WS \geq 0$ i $\hat{H} = S^{-1}H \geq 0$, da li smo konstruirali ekvivalentnu stacionarnu točku $(\hat{W}, \hat{H})$? Osim nenegativnosti matrica $\hat{W}$ i $\hat{H}$ trebaju biti zadovoljeni i KKT uvjeti (2.22), (2.23), odnosno treba vrijediti:

$$\begin{aligned} (S^{-1})(WHH^\top - XH^\top) &\geq 0, \\ (W^\top WH - W^\top X)S &\geq 0, \\ (WS) \circ [(WHH^\top - XH^\top)] &= 0, \\ (S^{-1}H) \circ [(W^\top WH - W^\top X)] &= 0. \end{aligned} \tag{2.25}$$

Ako uzmemo da je matrica $S$ permutacijska, tada je lako provjeriti valjanost ovih uvjeta. U tom slučaju, stupci matrica $W$ i $H$ su sačuvani u $\hat{W}$ i $\hat{H}$, samo u ispermutiranom redoslijedu. Primijetimo da matrica $S$ ne može biti *monomijalna* (tj. dobivena iz permutacijske matrice zamjenom nekih elemenata koji su jednaki 1 drugim pozitivnim brojevima), s obzirom da smo postavili ograničenje normiranosti stupaca za $W$ i $\hat{W}$. Za općeniti $S$, nije više trivijalno odrediti jedinstvenost stacionarne točke i može se odrediti samo za svaki slučaj posebno.

## 2.3 Svojstva lokalnog minimuma

Promotrimo gdje leže stacionarne točke $(W, H)$ za NMF. Za početak sumirajmo sve elemente prvog izraza uvjeta (2.23):

$$\begin{aligned} 0 &= \sum_{ij} \left( W \circ (WHH^\top - XH^\top) \right)_{ij} \\ &= \langle W, WHH^\top - XH^\top \rangle \\ &= \langle WH, WH - X \rangle \end{aligned} \tag{2.26}$$

**Teorem 2.3.1.** *Neka je* $(W, H)$ *stacionarna točka za NMF. Tada vrijedi*

$$WH \in \mathcal{B}\left(\frac{X}{2}, \frac{1}{2}\|X\|_F\right), \tag{2.27}$$

*odnosno, stacionarna točka se nalazi unutar kugle centrirane u* $\frac{X}{2}$ *radijusa* $\frac{1}{2}\|X\|_F$.

*Dokaz.* Iz (2.26) odmah slijedi

$$\left\langle \frac{X}{2} - WH, \frac{X}{2} - WH \right\rangle = \left\langle \frac{X}{2}, \frac{X}{2} \right\rangle \tag{2.28}$$

iz čega dalje slijedi

$$WH \in \mathcal{B}\left(\frac{X}{2}, \frac{1}{2}\|X\|_F\right). \tag{2.29}$$

□

**Teorem 2.3.2.** *Neka je* $(W, H)$ *stacionarna točka za NMF. Tada vrijedi*

$$\frac{1}{2}\|X - WH\|_F^2 = \frac{1}{2}\left(\|X\|_F^2 - \|WH\|_F^2\right). \tag{2.30}$$

*Dokaz.* Iz (2.26) imamo $\langle WH, X\rangle = \langle WH, WH\rangle$, stoga,

$$\begin{aligned} \frac{1}{2}\langle X - WH, X - WH\rangle &= \frac{1}{2}(\|X\|_F^2 - 2\langle WH, X\rangle + \|XH\|_F^2) \\ &= \frac{1}{2}(\|X\|_F^2 - \|WH\|_F^2). \end{aligned} \tag{2.31}$$

□

**Napomena 2.3.3.** *Primijetimo da teorem 2.3.2 također nalaže da u stacionarnoj točki* $(W, H)$ *mora vrijediti* $\|X\|_F^2 \geq \|WH\|_F^2$. *Jednakost je zadovoljena samo kod egzaktne faktorizacije, tj.* $X = WH$.

Neka je $X_r$ optimalna aproksimacija ranka $r$ nenegativne matrice $X$ koju dobijemo kako je opisano u teoremu 1.1.2. Tada lako možemo konstruirati nenegativni dio $[X_r]_+$, koji dobijemo iz $X_r$ tako da sve negativne elemente postavimo na 0. Ako promatramo kao geometrijski problem, tada $[X_r]_+$ možemo vizualizirati kao najbližu matricu matrici $X_r$ iz stošca nenegativnih matrica s obzirom na Frobeniusovu normu, odnosno, $[X_r]_+$ je projekcija matrice $X_r$ na nenegativni stožac. Sada možemo procijeniti grešku aproksimacije matrice $X$ matricom $[X_r]_+$.

**Teorem 2.3.4.** *Neka je $X_r$ najbolja aproksimacija ranka r nenegativne matrice X, i neka je $[X_r]_+$ njen nenegativni dio, tada vrijedi*

$$\|X - [X_r]_+\|_F \leq \|X - X_r\|_F. \tag{2.32}$$

*Dokaz.* Ono što trebamo pokazati je da vrijedi

$$\sum_{ij}(X - [X_r]_+)_{ij}^2 \leq \sum_{ij}(X - X_r)_{ij}^2. \tag{2.33}$$

Pokazat ćemo da vrijedi

$$|X - [X_r]_+|_{ij} \leq |X - X_r|_{ij}, \forall i, j, \tag{2.34}$$

odnosno, da je svaki pribrojnik lijeve strane manji ili jednak pribrojniku desne strane iz čega slijedi nejednakost (2.32).

- za $(X_r)_{ij} \geq 0$, po definiciji od $[X_r]_+$ slijedi da je $([X_r]_+)_{ij} = (X_r)_{ij}$, iz čega slijedi da su pribrojnik lijeve i desne strane jednaki
- za $(X_r)_{ij} < 0$, slijedi da je $|X_r|_{ij} = -(X_r)_{ij}$, a po definiciji od $[X_r]_+$ vrijedi $([X_r]_+)_{ij} = 0$ odakle imamo

$$|X - [X_r]_+|_{ij} = |X - 0|_{ij} \leq^* |X + |X_r|\,|_{ij} = |X - (X_r)|_{ij} \tag{2.35}$$

U * smo iskoristili činjenicu da su $X$ te $|X_r|$ nenegativni.

□

Pomoću ovog rezultata možemo lako usporediti dobivenu graničnu vrijednost s nenegativnim aproksimacijama.

**Korolar 2.3.5.** *Neka je $W_*H_*$ optimalna nenegativna aproksimacija ranka r i neka je $WH$ proizvoljna stacionarna točka koja zadovoljava KKT uvjete za nenegativnu aproksimaciju ranka r, tada vrijedi*

$$\|X - [X_r]_+\|_F^2 \leq \|X - X_r\|_F^2 = \sum_{i=r+1}^{n} \sigma_i^2 \leq \|X - W_*H_*\|_F^2 \leq \|X - WH\|_F^2. \tag{2.36}$$

Izraz (2.36) pokazuje da je greška aproksimacije bilo kojom stacionarnom točkom koja zadovoljava KKT uvjete uvijek veća od greške aproksimacije matrice $X$ projekcijom na stožac nenegativnih matrica.

## 2.4 Odabir reduciranog ranga

Procjena reduciranog ranga uvelike utječe na rezultat dobiven nenegativnim matričnim faktorizacijama. S obzirom da je naglasak na redukciji dimenzionalnosti, treba uzeti u obzir da rang $r$ ne smije biti prevelik, jer to ne bi doprinijelo značajnom smanjenju veličine problema; međutim, u isto vrijeme treba paziti i da odabrani rang ne bude premalen što bi vodilo k prevelikoj grešci aproksimacije.

Neki od načina procjene reduciranog ranga su sljedeći:

- **Unakrsna validacija**[42]. Za matricu $X \in \mathbb{R}^{m \times n}$ definiraju se podskupovi $\mathcal{I}_l \subset \{1, ..., m\}$ i $\mathcal{J}_l \subset \{1, ..., n\}$ za $l = 1, ..., L$ koji označavaju indekse redaka, odnosno stupaca koji se ispuštaju iz matrice $X$ te skup $\mathcal{K} \subset \{1, ..., \min(m, n)\}$ promatranih rangova. Za svaki od tih rangova računa se vrijednost $BCV(k)$ tako što se za svaki $l$ aproksimira matrica $X$ bez ispuštenih redaka i stupaca, a zatim se zbroje reziduali početne matrice i matrice dobivene kao umnožak faktora $\hat{W}$ i $\hat{H}$ gdje $\hat{W}$ i $\hat{H}$ predstavljaju matrice koje minimiziraju greške u slučaju ispuštenih redaka, odnosno stupaca. Onaj $BCV(k)$ koji je najmanji ujedno je i traženi reducirani rang $r$.

- **Kofenetska korelacija**[11]. Za svako pokretanje NMF algoritma nad matricom $X$ može se definirati matrica susjedstva $C$ gdje je $c_{ij} = 1$ ako uzorci $i$ i $j$ pripadaju istom klasteru, inače $c_{ij} = 0$ (pripadnost klasteru se odredi pomoću maksimalne vrijednosti odgovarajućeg stupca matrice koeficijenata $H$). Zatim se izračuna matrica $\bar{C}$ kao prosječna matrica svih matrica $C$ za trenutni rang. Sada je svaki element konsenzus matrice $\bar{C}$ iz intervala $[0, 1]$ te izražava vjerojatnost da su uzorci $i$ i $j$ iz istog klastera. Ako je klasteriranje stabilno, očekivano je da se matrice $C$ neće previše razlikovati, pa će vrijednosti konsenzus matrice biti blizu ili 0 ili 1. Dalje se računa mjera zasnovana na kofenetskom koeficijentu korelacije, $\rho_k(\bar{C})$, koja pokazuje disperziju matrice $\bar{C}$, a definira se kao Pearsonov koeficijent korelacije između matrice udaljenosti inducirane konsenzus matricom $\bar{C}$ i matrice udaljenosti kao rezultatom hijerarhijskog klasteriranja (za detalje pogledati [11]) . Što je vrijednost $\rho_k(\bar{C})$ bliže vrijednosti 1, to je klasteriranje stabilnije, odnosno ukazuje na bolje prepoznavanje $k$ klasa. Za reducirani rang $r$ bira se rang $k$ za koji kofenetski koeficijent korelacije kreće padati.

- **Zbroj kvadrata reziduala**[33]. Ako promatramo RSS (*eng. residual sum of squares*) krivulju, odabere se prva vrijednost ranga $k$ koja predstavlja točku infleksije; prema Hutchinsu, takva točka znači da dodatni rang ne poboljšava vrijednost greške rekonstrukcije značajno u odnosu na prošla povećanja ranga.

- **Procjena pomoću SVD-a**[46]. Qiao je sugerirao da se sumiraju singularne vrijednosti u padajućem redoslijedu te zatim uzme onaj broj singularnih vrijednosti koji

je potreban kako bi ta suma imala relativno velik udio ukupne sume. Kao graničnu vrijednost odredio je 0.9 s obzirom da sadrži dovoljno informacija od singularnih vrijednosti, a izbjegava da rang bude premalen.

## 2.5 Ograničenja na NMF

U ovom poglavlju promotrit ćemo neka najčešće korištena dodatna ograničenja na standarni NMF problem dodavanjem kojih pokušavamo zadovoljiti neke specifične zahtjeve za NMF.

Općenito, možemo koristiti sljedeću formulu za ilustrirati problem nenegativnih matričnih faktorizacija uz dodatna ograničenja:

$$f(W,H) = \frac{1}{2}\|X - WH\|_F^2 + \alpha J_1(W) + \beta J_2(H). \tag{2.37}$$

Funkcije $J_1$ i $J_2$ koriste se za neke mjere svojstava matrica $W$ i $H$ kao što su raspršenost i ortogonalnost. $\alpha$ i $\beta$ su mali regulacijski parametri koji balansiraju kompromis između pogreške aproksimacije i ograničenja.

U ovom poglavlju bit će navedena tri često korištena zahtjeva: raspršenost, ortogonalnost i glatkoća, iako uzmimo u obzir da ovisno o problemu dodatno ograničenje može biti razne prirode. Tako se može tražiti i ograničenje na udaljenost svaka dva od vektora matrice $H$, ili pak ograničenje lokaliziranosti, tj. zahtjev da korišteni bazni vektori $W(:, i)$ budu što bliži vektorima originalne matrice $X(:, j)$ kao što je slučaj u [2].

### Raspršeni NMF

Raspršenost je najčešće korišteno ograničenje za prošireni NMF iz razloga što nekad samo nekoliko karakteristika može reprezentirati cijeli skup podataka. To bi značilo da je dovoljno uzeti samo nekoliko uzoraka iz cijele populacije kako bi se efektivno reprezentirao cijeli skup podataka.

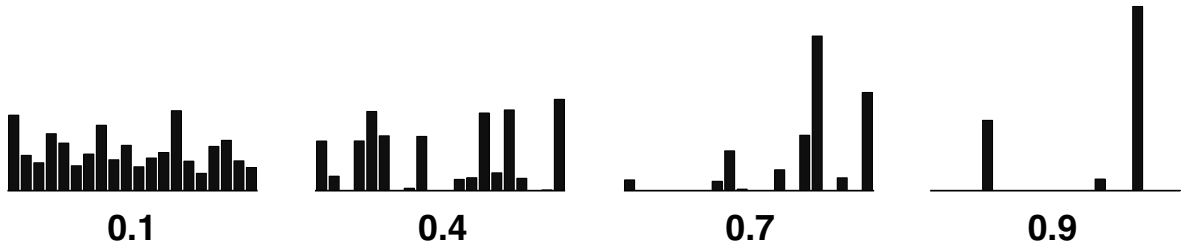
Postoji više mjera za raspršenost, tj. mapiranja iz $\mathbb{R}^n$ u $\mathbb{R}$ koja kvantificiraju koliko energije vektora je prisutno u samo nekoliko komponenti. Na normaliziranoj skali, najrjeđi vektor, onaj koji se sastoji od samo od jedne komponente koja je različita od nule, trebao bi imati raspršenost jednaku jedan, dok vektor sa svim jednakim elementima bi trebao imati raspršenost nula.

Primjer jedne takve mjere raspršenosti definirao je Hoyer u [32], a bazirana je na vezi između $L_1$ i $L_2$ norme:

$$sparseness(x) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1}, \tag{2.38}$$

gdje je $n$ dimenzionalnost od $x$. Ova funkcija evaluira izraz u vrijednost jedan ako i samo ako $x$ sadrži samo jednu komponentu različitu od nule, a jednaka je nula ako i samo ako su sve komponente jednake (do na predznak), inače interpolira glatko između te dvije vrijednosti.

Slika 2.5 prikazuje vrijednosti mjere *sparseness* za četiri različita vektora (vrijednosti vektora su prikazane u obliku histograma). Prvi vektor ima najmanju mjeru jer su njegove vrijednosti sličnih veličina, dok zadnji ima najveću jer je većina vrijednosti nula i samo nekoliko ih ima značajne vrijednosti.




Slika 2.5: Mjere raspršenosti za četiri različita vektora

## Ortogonalni NMF

Zahtjev za ortogonalnost prvi put je postavio Ding u [18]. Postavio je ograničenje ortogonalnosti na obje matrice $W$ i $H$, tj. da mora vrijediti

$$W^\top W = I, \quad H^\top H = I. \tag{2.39}$$

U tom radu dokazano je da postavljanje takvih uvjeta limitira broj rješenja NMF-a na samo jedno. Osim toga, dokazani su konvergencija lokalnom minimumu te točnost njihovog algoritma.

Jedna od bitnih primjena ortogonalnog NMF-a je u klasteriranju. U [18] dokazan je teorem da je problem za NMF uz uvjet ortogonalnosti na matricu koeficijenata $H$ ekvivalentan klasteriranju $k$-sredinama. Matrice $W$ i $H$ dobivene takvim uvjetom imaju jasnu interpretaciju: svaka baza matrice $W$ pokazuje smjer prema centru klastera u podacima, a matrica koeficijenata daje stupanj povezanosti uzorka sa svakim od klastera.

## Glatki NMF

Uvjeti na glatkoću se obično uvode u kaznenoj funkciji NMF problema kako bi se regulirao rezultat u slučaju prisustva šuma [4]. Mogu se postaviti na bilo koju od matrica $W$ i $H$

ovisno o konkretnom problemu. Npr. izraz

$$J_1(W) = \|W\|_F^2 \tag{2.40}$$

penalizira rješenja $W$ koja imaju veliku Frobeniusovu mjeru. Primijetimo da ovo implicitno penalizira stupce od $W$ s obzirom da je $\|W\|_F^2 = \sum \|w_i\|_2^2$. Općenitije, (2.40) se može napisati kao $J_1(W) = \|LW\|_F^2$, gdje je $L$ regulacijski operator.

U [4] dokazana je efikasnost ograničenja iz izraza (2.40), gdje je pokazano da se na taj način mogu dobiti karakteristike veće kvalitete nego standardnim NMF-om.

Uvjeti za glatkoću se također mogu postaviti na $H$. U [14] Chen i Cichocki su u svrhu poboljšanja analize EEG podataka za ranu detekciju Alzheimerove bolesti primorali temporalnu glatkoću stupaca matrice $H$ definirajući kaznenu funkciju za glatkoću kao:

$$J_2(H) = \frac{1}{n} \sum_i \|(I - T)h_i^\top\|_2^2 = \frac{1}{n}\|(I - T)H^\top\|_F^2, \tag{2.41}$$

gdje je $n$ ukupni broj stupaca matrice $X$, a $T$ je odgovarajuće definirani konvolucijski operator.

Spomenimo da postoje i druge definicije glatkoće, npr. koristeći B-spline kao u [47], ili pomoću $L_2$ norme kao u [45].

# Poglavlje 3

# Postojeći algoritmi

Zadnjih godina broj predloženih rješenja za problem nenegativnih matričnih faktorizacija je u stalnom porastu. U ovom poglavlju dat ćemo opis nekoliko najpopularnijih algoritama. Iako je algoritam alternirajućih nenegativnih najmanjih kvadrata prvi predložen, algoritam Hadamardovog produkta ipak je postao mnogo popularniji zbog svoje jednostavnosti i interpretabilnosti.

Osim ova dva algoritma, bit će opisane još dvije inačice algoritma alternirajućih kvadrata, od kojih hijerarhijska inačica daje ponajbolje rezultate. Također će biti dan općeniti opis algoritama zasnovanih na gradijentnom spustu te detaljnije opisan onaj zasnovan na Armijo pravilu. Za kraj, promotrit ćemo metode inicijalizacije početnih matrica te moguće kriterije zaustavljanja. Krenit ćemo od okvirnog algoritma kojeg koristi velik dio standardnih algoritama za rješavanje problema nenegativnih matričnih faktorizacija.

### Okvirni algoritam za NMF

Većina standardnih algoritama konstruiranih za rješavanje NMF-a koristi shemu koordinatnog spusta za dva bloka, tj. alternirajuće optimizira faktore $W$, odnosno $H$, dok drugog drži fiksiranog. Razlog tome je što je NMF problem konveksan kada fiksiramo jedan faktor, tj. postaje problem nenegativnih najmanjih kvadrata (*eng. nonnegative least squares*, *NNLS*). Za primjer, kada je $H$ fiksiran, moramo riješiti

$$\min_{W \geq 0} \|X - WH\|_F^2. \tag{3.1}$$

Primijetimo da ovaj problem ima određenu strukturu s obzirom da se može rastaviti u $n$ nezavisnih NNLS-a u $r$ varijabli, tj. vrijedi[27]

$$\|X - WH\|_F^2 = \sum_{i=1}^{n} \|X_{:i} - W_{:i}H\|_2^2 = \sum_{i=1}^{n} W_{:i}(HH^\top)W_{:i}^\top - 2W_{:i}(HX_{:i}^\top) + \|X_{:i}\|_2^2. \tag{3.2}$$

Za rješenje problema nenegativnih najmanjih kvadrata postoji mnogo algoritama, te se ovisno o njima i razlikuje konkretni algoritam za rješenje NMF-a pomoću koordinatnog spusta za dva bloka.

Primijetimo još da je problem simetričan u $W$ i $H$, tj. da vrijedi

$$||X - WH||_F^2 = ||X^\top - H^\top W^\top||_F^2.$$

Zbog toga možemo na isti način osvježavati oba faktora, što je slučaj u velikom broju algoritama za NMF koji se mogu prikazati okvirnim algoritmom 1 .

**Algoritam 1** Okvirni algoritam koordinatnog spusta za dva bloka

1: Inicijaliziraj početne matrice $W^0 \geq 0$ i $H^0 \geq 0$; vidi poglavlje 3.4
2: **za** $t = 1, 2, ...^\dagger$ **radi**
3: $\quad W^t = \texttt{osvježi}(X, H^{t-1}, W^{t-1})$
4: $\quad H^t = \texttt{osvježi}(X^\top, (W^t)^\top, (H^{t-1})^\top)$
5: **kraj**

† vidi poglavlje 3.5 za kriterij zaustavljanja

## 3.1 Algoritam Hadamardovog produkta

Algoritam Hadamardovog produkta (*eng. multiplicative update*, *MU*) je zasigurno najpopularniji algoritam za NMF problem. Prvi put se pojavio u [17] kao rješenje za problem nenegativnih najmanjih kvadrata. Kasnije su ga Lee i Seung uveli u [34] kao metodu rješavanja nenegativnih matričnih faktorizacija zahvaljujući kojima je NMF proširio svoj utjecaj u mnogobrojna znanstvena područja. Razlog tome je jednostavnost Hadamardovog produkta i interpretabilnost rezultata.

Da bismo formulirali pravila za iteraciju Hadamardovim produktom prvo ćemo fiksirati jedan faktor (u ovom slučaju $W$) i tada minimizirati funkciju troška s obzirom na drugi faktor ($H$). Za početak ćemo pretpostaviti da su $W$ i $H$ pozitivni na što ćemo se vratiti kasnije.

Funkcija troška (2.1) iz definicije nenegativne matrične faktorizacije može se napisati kao

$$\frac{1}{2}||X - WH||_F^2 = \frac{1}{2}\sum_{i=1}^{n} ||X_{:i} - WH_{:i}||_2^2, \tag{3.3}$$

odnosno, može se rastaviti na $n$ nezavisnih problema gdje minimiziramo svaki stupac $h$ od $H$ posebno. Tada imamo niz kvadratnih problema koje možemo izraziti kao

$$\min_{h \geq 0} F(h), \quad \text{gdje je } F(h) = \frac{1}{2}||x - Wh||_2^2. \tag{3.4}$$

Pretpostavimo da je $\overline{h} \geq 0$ neka trenutna aproksimacija NMF problema. Tada možemo formulirati sljedeći problem:

$$\min_{h\geq 0} \bar{F}(h) = \min_{h\geq 0} \frac{1}{2}\left[\|x - Wh\|_2^2 + (h - \bar{h})^\top V_{\bar{h}}(h - \bar{h})\right] \tag{3.5}$$

gdje je $V_{\bar{h}} = D_y - W^\top W$ pri čemu je $y = \frac{[W^\top W\bar{h}]}{[\bar{h}]}$. Može se pokazati pozitivna semidefinitnost matrice $V_{\bar{h}}$, stoga je drugi pribrojnik $(h-\bar{h})^\top V_{\bar{h}}(h-\bar{h}) \geq 0$, iz čega slijedi da je $\bar{F}(h) \geq F(h)$ za svaki $h$, a posebno vrijedi $\bar{F}(\bar{h}) = F(\bar{h})$. Raspisivanjem se dobije da je gradijent funkcije $\bar{F}(h)$ jednak

$$\nabla_h \bar{F} = W^\top Wh - W^\top x + V_{\bar{h}}(h - \bar{h}). \tag{3.6}$$

Izjednačavanjem $\nabla_h \bar{F} = 0$ kako bismo dobili minimum $h^*$ dobijemo

$$(W^\top W + V_{\bar{h}})h^* = W^\top x - V_{\bar{h}}\bar{h}. \tag{3.7}$$

Kako vrijedi $W^\top W + V_{\bar{h}} = D_{W^\top W\bar{h}} D_{\bar{h}}^{-1}$ i $V_{\bar{h}}\bar{h} = 0$, konačno imamo:

$$v^* = \bar{v} \circ \frac{[W^\top x]}{\left[W^\top W\bar{h}\right]}, \tag{3.8}$$

gdje $\frac{[]}{[]}$ označava dijeljenje matrica po komponentama.

S obzirom da je $h^*$ globalni minimum za $\bar{F}(h)$, vrijedi $\bar{F}(h^*) \leq \bar{F}(\bar{h})$. Nadalje, već smo vidjeli da vrijedi $\bar{F}(h) \geq F(h)$ za svaki $h$. Iz svega toga slijedi da vrijedi

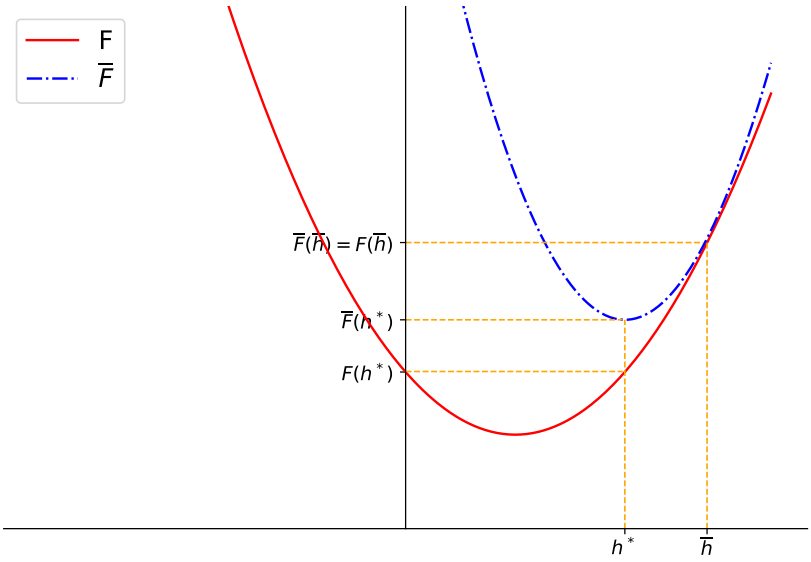
$$F(h^*) \leq \bar{F}(h^*) \leq \bar{F}(\bar{h}) = F(\bar{h}), \tag{3.9}$$

odnosno, imamo spust u funkciji troška. Slika 3.1 opisuje ovu nejednakost.

Ovdje smo izveli pravilo osvježenja za jedan stupac matrice $H$. Ako isti postupak ponovimo za svaki stupac matrice, tada dobijemo pravilo osvježenja za cijelu matricu $H$. Na sličan način možemo izvesti i pravilo za osvježenje matrice $W$ te sada imamo sve korake za rješenje NMF-a pomoću Hadamardovog produkta što je prikazano u algoritmu 2.

Primijetimo sada da u izrazu (3.5) drugi pribrojnik se može interpretirati kao kaznena funkcija kako bi se spriječilo da rješenje optimizacijskog problema bude nula. Nadalje, matrica $D_y - W^\top W$ gdje je $y = \frac{[W^\top W\bar{h}]}{[\bar{h}]}$ se može promatrati kao aproksimacija Hesseove matrice $W^\top W$. Gornji razvoj možemo sažeti teoremom 3.1.1.

**Teorem 3.1.1.** *Euklidska udaljenost $\|X - WH\|_F^2$ je nerastuća kod korištenja pravila za iteraciju kao u algoritmu 2.*




Slika 3.1: Vizualizacija nerastuće ciljne funkcije kod algoritma Hadamardovog produkta

**Algoritam 2** Algoritam Hadamardovog produkta (MU)

1: Inicijaliziraj neke početne matrice $W^0 \geq 0$ i $H^0 \geq 0$ i $k = 0$
2: **ponavljaj**
3: $\quad W^{k+1} = W^k \circ \dfrac{\left[X(H^k)^\top\right]}{\left[W^k H^k (H^k)^\top\right]}$
4: $\quad H^{k+1} = H^k \circ \dfrac{\left[(W^{k+1})^\top X\right]}{\left[(W^{k+1})^\top (W^{k+1}) H^k\right]}$
5: $\quad k = k + 1$
6: **dok** nije zadovoljen kriterij zaustavljanja

Ostaje nam problem očuvanja nenegativnosti matrica $W^{k+1}$ i $H^{k+1}$ ukoliko znamo da su $W^k$ i $H^k$ nenegativne. Dakle, ono što želimo izbjeći je da zapnemo u nekoj graničnoj točki koja nije stacionarna, tj. nije zadovoljen KKT uvjet (2.22), nego vrijedi:

$$\nabla_{H_{ij}} F < 0. \tag{3.10}$$

Ako pretpostavimo da su početni $W$ i $H$ pozitivni, tada vrijedi

$$\nabla_{H_{ij}^k} F < 0, \quad \text{odnosno} \quad [(W^k)^\top W^k H^k]_{ij} - [(W^k)^\top X]_{ij} < 0, \tag{3.11}$$

pa iz pravila za iteraciju Hadamardovim produktom imamo

$$H_{ij}^{k+1} = H_{ij}^{k} \frac{[(W^k)^\top X]_{ij}}{[(W^k)^\top W^k H^k]_{ij}} > H_{ij}^{k} > 0. \tag{3.12}$$

Ovaj rezultat je iskazan u sljedećoj lemi.

**Lema 3.1.2.** *Ako su matrice $W^0$ i $H^0$ pozitivne, a matrica $X$ nema nul-redaka ni nul-stupaca, tada su $W^k$ i $H^k$ također pozitivne.*

U praksi, zbog konačne preciznosti računala, brojevi čija je apsolutna vrijednost manja od neke konstante $\epsilon_M$ ne mogu se prikazati u računalu dovoljno precizno zbog čega budu zaokruženi te reprezentirani kao nula. Zbog toga, jednom kada $H_{ij}^k$ postane nula, pravila za iteraciju Hadamardovim produktom ga ne mogu ponovno učiniti pozitivnim te je tada velika vjerojatnost da zapnemo u nestacionarnoj točki. Tada možemo umjesto $H_{ij}^k = 0$ postaviti $H_{ij}^k$ na neki $\epsilon > \epsilon_M$ kako je sugerirano u [35].

Za kraj, primijetimo još da nazivnik $\left[(W^k)^\top (W^k) H^k\right]$ iz pravila za iteraciju Hadamardovim produktom za neki element $(i, j)$ može biti jednak nuli, što bi u tom slučaju rezultiralo iznimkom zbog dijeljenja s nulom. Tada imamo sljedeće dvije situacije:

- $H_{ij}^k > 0$, tada $\left[(W^k)^\top (W^k) H^k\right]_{ij} = 0$ daje

$$0 = \sum_{lt} W_{ti}^k W_{tl}^k H_{lj}^k \geq \sum_{t} W_{ti}^k W_{ti}^k H_{ij}^k = H_{ij}^k \sum_{t} W_{ti}^k W_{ti}^k. \tag{3.13}$$

  Međutim, ovaj izraz je jednak nuli samo kada je $W_{:i} = 0$, do čega dolazi zbog aproksimacije nedovoljnog ranka. Ovakav slučaj se riješi generiranjem zamjenskog vektora za stupac $W_{:i}$.

- $H_{ij}^k = 0$, tada imamo neodređeni izraz $0/0$ u kojem slučaju vrijedi

$$\nabla_{H_{ij}} F = [W^\top X]_{ij} \leq 0. \tag{3.14}$$

  Tada, umjesto da zamijenimo $\left[(W^k)^\top (W^1) H^k\right]_{ij}$ s $\epsilon$ (što bi dalje propagiralo $H_{ij}^{k+1} = 0$), treba postaviti $H_{ij}^{k+1} = \epsilon > 0$.

Jedna od mana algoritma Hadamardovog produkta je ta što konvergira relativno sporo (teoretska analiza provedena u [30]). Primijetimo da standardni MU algoritam osvježava $W$ samo jednom prije nego osvježi $H$. Algoritam se može znatno ubrzati koristeći strategiju iz [28]. Ideja je ta da se osvježi $W$ nekoliko puta prije nego osvježimo $H$ jer se tada produkti $HH^\top$ i $XH^\top$ ne trebaju ponovno računati.

## 3.2 Algoritmi alternirajućih najmanjih kvadrata

**Algoritam alternirajućih nenegativnih najmanjih kvadrata** (*eng. alternating nonnegative least squares*, *ANLS*) prvi je algoritam koji su predložili Paatero i Tapper u [44] kada su definirali problem nenegativne matrične faktorizacije. Motivacija za to je konveksnost funkcije $F$ za NMF iz (2.1) ukoliko fiksiramo jedan od faktora $W$ ili $H$. Npr., ako fiksiramo $W$, tada je funkcija $F$ kompozicija Frobeniusove mjere i linearne transformacije od $H$. Slično vrijedi i za slučaj kada fiksiramo $H$.

Fiksiranjem neke od te dvije matrice problem NMF-a se preoblikuje u problem nenegativnih najmanjih kvadrata. Dakle, u svakom koraku osvježavanja faktora $W$ i $H$ rješavamo dva odvojena potproblema najmanjih kvadrata kao što je prikazano u algoritmu 3. Nadalje, svaki od ta dva odvojena problema može se rastaviti na $n$ nezavisnih problema koji odgovaraju stupcima matrice $X$. Svaki taj manji osnovni problem nenegativnih najmanjih kvadrata (*eng. nonnegative least squares*, *NNLS*) može se riješiti nekom od metoda iz [36][24][10].

**Algoritam 3** Algoritam alternirajućih nenegativnih najmanjih kvadrata (ANLS)

1: Inicijaliziraj početne matrice $W^0 \geq 0$ i $H^0 \geq 0$
2: **ponavljaj**
3: Riješi: $W^{k+1} = \text{argmin}_{W \geq 0} \|X - WH^k\|_F^2$
4: Riješi: $H^{k+1} = \text{argmin}_{H \geq 0} \|X - W^{k+1}H\|_F^2$
5: $k = k + 1$
6: **dok** nije zadovoljen kriterij zaustavljanja

Pomoću teorema o konvergenciji metode koordinatnog spusta može se dokazati da ako su potproblemi iz algoritma 3 egzaktno i jedinstveno rješivi, tada je svaka granična točka algoritma ujedno i stacionarna točka problema nenegativne matrične faktorizacije.

S obzirom da svaka iteracija ovog algoritma računa optimalno rješenje za svaki NNLS, to znači da svaka iteracija smanjuje grešku više od ostalih algoritama koji imaju oblik okvirnog algoritma 1 za koordinatni spust za dva bloka. Međutim, svaka iteracija je zbog toga spora, te teška za implementirati. Zbog toga je često bolje rješavati NNLS probleme egzaktno u kasnijim koracima nekog računski manje zahtjevnog algoritma kao što je već spomenuti algoritam Hadamardovog produkta, ili pak algoritam alternirajućih najmanjih kvadrata čiji opis slijedi.

**Algoritam alternirajućih najmanjih kvadrata** (*eng. alternating least squares*, *ALS*) neegzaktna je verzija algoritma nenegativnih alternirajućih najmanjih kvadrata. Razlika je ta što kod rješavanja potproblema nema ograničenja nenegativnosti. Egzaktno rješenje problema NNLS zamijenjeno je projekcijom rješenja neograničenog problema najmanjih kvadrata u nenegativni ortant kako je prikazano u algoritmu 4. Ovo ubrzava algoritam, međutim svojstvo konvergencije više nije zadovoljeno.

**Algoritam 4** Algoritam alternirajućih najmanjih kvadrata (ALS)

1: Inicijaliziraj početne matrice $W^0$ i $H^0$
2: **ponavljaj**
3: Riješi: $W^{k+1} = \text{argmin} \|X - WH^k\|_F^2$
4: $W^{k+1} = [W^{k+1}]_+$
5: Riješi: $H^{k+1} = \text{argmin} \|X - W^{k+1}H\|_F^2$
6: $H^{k+1} = [H^{k+1}]_+$
7: $k = k + 1$
8: **dok** nije zadovoljen kriterij zaustavljanja

Primijetimo da u algoritmu 4 problem nenegativnosti je riješen na najjednostavniji način, zamjenom negativnih elemenata nulom. Ovom tehnikom se potpomaže rijetka popunjenost matrice, što pruža fleksibilnost algoritmu. Za razliku od algoritama baziranih na Hadamardovom produktu kao što je MU, gdje jednom kada element matrice $W$ ili $H$ dođe u nulu, tada i ostaje u nuli, ovdje se kroz iteracije element može ponovno odmaknuti od nule.

Ako usporedimo ovaj algoritam s algoritmom alternirajućih nenegativnih najmanjih kvadrata, vidimo da ANLS ima bolju aproksimacijsku grešku te uvijek dolazi do spusta u funkciji $F$. Međutim, ANLS za isti broj iteracija treba značajno više vremena.

U praksi, ANLS se rijetko koristi zbog svoje neefikasnosti. Međutim, ALS ima manu što općenito ne konvergira u stacionarnu točku. Zbog toga se preporučuje da se ALS koristi u početnim koracima hibridnih algoritama, što je posebno korisno za rijetko popunjenje matrice kako je sugerirano u [16].

**Algoritam hijerarhijskih alternirajućih najmanjih kvadrata** (*eng. hierarchical alternating least squares*, *HALS*) rješava NNLS potprobleme koristeći metodu egzaktnog koordinatnog spusta, svakog puta osvježavajući jedan stupac matrice $W$, odnosno redak matrice $H$. Dakle, ideja je fiksirati $r - 1$ stupaca (redaka), te tada minimizirati $j$-ti stupac (redak) kao u [21], odnosno

$$\min J_j(W_{:j}, H_{j:}) = \|R^{(j)} - W_{:j}H_{j:}\|_F^2, \tag{3.15}$$

gdje je $R^{(j)}$ $j$-ti rezidual

$$R^{(j)} \coloneqq X - \sum_{i \neq j}^{r} W_{:i}H_{i:}. \tag{3.16}$$

Sada možemo naći stacionarne točke računajući gradijent u $W_{:j}$

$$0 = \frac{\partial J_j}{\partial W_{:j}} = W_{:j}H_{j:}H_{j:}^\top - R^{(j)}H_{j:}^\top, \tag{3.17}$$

i gradijent u $H_{j:}$

$$0 = \frac{\partial J_j}{\partial H_{j:}} = H_{j:}^\top W_{:j}^\top W_{:j} - (R^{(j)})^\top W_{:j}. \tag{3.18}$$

Slijedi da su pravila osvježenja $j$-te komponente od $W$ i $H$ jednaka

$$W_{:j} \leftarrow \left[\frac{R^{(j)} H_{j:}^\top}{H_{j:} H_{j:}^\top}\right]_+ \tag{3.19}$$

$$H_{j:} \leftarrow \left[\frac{(R^{(j)})^\top W_{:j}}{W_{:j}^\top W_{:j}}\right]_+ . \tag{3.20}$$

Primijetimo sada da rezidual (3.16) možemo raspisati kao:

$$R^{(j)} := X - \sum_{i \neq j}^{r} W_{:i} H_{i:} = X - WH + W_{:j} H_{j:}. \tag{3.21}$$

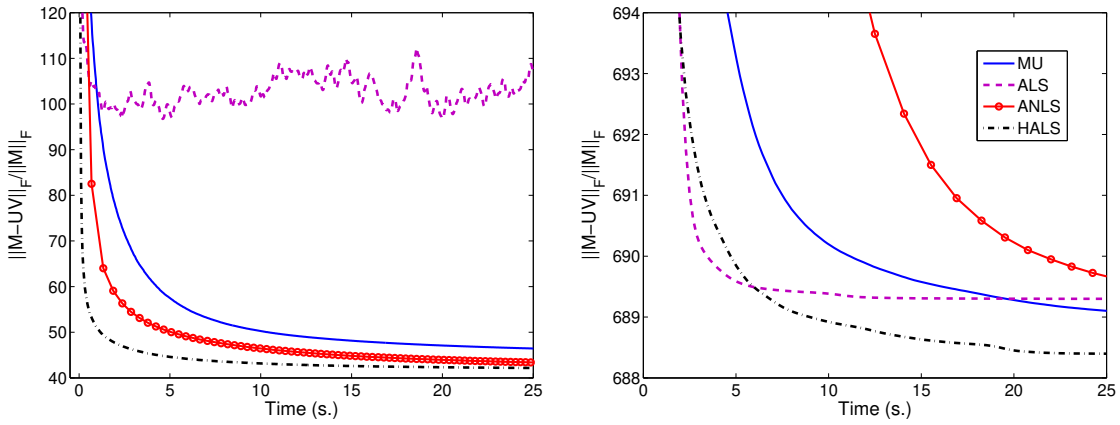
Kako bismo izbjegli eksplicitno računanje reziduala, možemo ovaj izraz supstituirati u (3.19) i (3.20) pa konačno imamo pravila prikazana algoritmom 5.

**Algoritam 5** Algoritam hijerarhijskih alternirajućih najmanjih kvadrata (HALS)

1: Inicijaliziraj početne matrice $W^0 \geq 0$ i $H^0 \geq 0$
2: **ponavljaj**
3: **za** $j = 1, .., r$ **radi**
4: $$W_{:j}^{(k+1)} = \left[W_{:j} + \frac{[XH]_{:j}^\top - W\left[HH^\top\right]_{:j}}{[HH^\top]_{jj}}\right]_+$$
5: $$H_{j:}^{(k+1)} = \left[H_{j:} + \frac{[X^\top W]_{:j} - H^\top\left[W^\top W\right]_{:j}}{[W^\top W]_{jj}}\right]_+$$
6: **kraj**
7: $k = k + 1$
8: **dok** nije zadovoljen kriterij zaustavljanja

HALS konvergira mnogo brže nego MU ([26]), dok je kompleksnost izvođenja skoro jednaka. Nadalje, HALS pod blagim pretpostavkama konvergira u stacionarnu točku (vidi [28]).

Sada slijedi usporedba navedenih algoritama alternirajućih kvadrata s najčešće korištenim algoritmom MU.




Slika 3.2: Usporedba MU, ALS, ANLS i HALS algoritama na NMF problemu za ”gusti” (lijevo) i ”raspršeni” (desno) skup podataka

**Usporedba algoritama alternirajućih najmanjih kvadrata s algoritmom MU**

Usporedba iz [27] uspoređuje ALS algoritme s MU za dva skupa podataka: prvi dataset je CBCL baza slika lica, dakle ”gusti” skup podataka, dok je drugi dataset klasična baza dokumenata, dakle ”raspršeni” skup podataka.

Slika 3.2 prikazuje iznos funkcije troška $F$ problema NMF kroz vrijeme. Na lijevom grafu su prikazani rezultati za CBCL, a na desnom za dokumente. Možemo primijetiti da:

- MU konvergira jako sporo
- ALS oscilira za gustu matricu te ima vidno najveću grešku od svih algoritama, dok za rijetko popunjenu matricu u početku konvergira jako brzo, međutim onda se stabilizira i ne uspijeva doći do rješenja koje bi imalo nisku funkciju troška
- ANLS je nakon HALS-a drugi najbolji za gustu matricu te daje dobre rezultate, međutim za rijetko popunjenu matricu daje jako loše rezultate
- HALS u oba slučaja daje najbolje rješenje u promatranom vremenu.

## 3.3 Projicirani gradijentni spust

Jedan od načina za promotriti NMF je kao nelinearni optimizacijski problem na konveksnom setu, tj. u nenegativnom ortantu. Postavljanjem negativnih elemenata matrice na

nulu lako možemo napraviti projekciju na nenegativni ortant. Zbog toga se može koristiti projicirani gradijentni spust (*eng. projected gradient descent, PGD*) koji se sastoji od računanja gradijenta $\nabla F(h^k)$, odabira veličine koraka $\alpha^k$ te projiciranja osvježenog rješenja na nenegativni ortant $\mathbb{R}^n_+$ na sljedeći način:

$$h^{k+1} = [h^k - \alpha_k \nabla F(h^k)]_+. \tag{3.22}$$

## PGD koristeći Armijo pravilo

Kako bi se poboljšala brzina konvergencije PGD-a, često se umjesto korištenja fiksne veličine koraka koristi neka od metoda za odabir $\alpha^k$ koja osigurava varijabilnost veličine koraka. Jedan od načina odabira $\alpha^k$ je pomoću Armijo pravila [5][6]. Algoritam 6 prikazuje rješenje problema NMF koristeći PGD uz Armijo pravilo kako je prikazano u [50].

**Algoritam 6** Projicirani gradijentni spust koristeći Armijo pravilo (PGD-Armijo)

1: Inicijaliziraj početne matrice $W^0 \geq 0$ i $H^0 \geq 0$, te $0 < \beta < 1$, $0 < \sigma < 1$ i $k = 1$
2: **ponavljaj**
3: $W^{k+1} = [W^k - \alpha_k \nabla_W F(W^k, H^k)]_+$
($\alpha_k = \beta^{t_k}$, gdje je $t_k$ najmanji $t \in \mathbb{N}$ t.d. vrijedi (3.23))
4: $H^{k+1} = [H^k - \alpha_k \nabla_H F(W^{k+1}, H^k)]_+$
($\alpha_k = \beta^{t_k}$, gdje je $t_k$ najmanji $t \in \mathbb{N}$ t.d. vrijedi (3.24))
5: $k = k + 1$
6: **dok** nije zadovoljen kriterij zaustavljanja

U koracima 3 i 4 algoritma 6, Armijo pravilo koristi izraze (3.23) i (3.24) kako bi se osigurao dovoljan spust u svakoj iteraciji, zbog čega daje bolju konvergenciju od fiksnog koraka $\alpha$.

$$\begin{aligned}(1-\sigma)\left\langle \nabla_W F(W^k, H^k), W^{k+1} - W^k \right\rangle + \\ +\frac{1}{2}\left\langle W^{k+1} - W^k, (W^{k+1} - W^k)((H^k)(H^k)^\top) \right\rangle \leq 0\end{aligned} \tag{3.23}$$

$$\begin{aligned}(1-\sigma)\left\langle \nabla_H F(W^{k+1}, H^k), H^{k+1} - H^k \right\rangle + \\ +\frac{1}{2}\left\langle H^{k+1} - H^k, ((W^{k+1})^\top W^{k+1})(H^{k+1} - H^k) \right\rangle \leq 0\end{aligned} \tag{3.24}$$

Daljnje poboljšanje konvergencije može se dobiti primjenom pravila kojeg je Lin opisao u [35]. Sugerirao je da se $\alpha_{k-1}$ koristi kao inicijalni pokušaj za vrijednost $\alpha_k$, zbog čega je potrebno manje koraka za naći $\alpha_k$. Algoritam za PGD uz Linovo pravilo može se naći u [50]. Sličan je PGD-u uz Armijo pravilo, a glavna razlika je u odabiru koraka $\alpha_k$.

## 3.4 Inicijalizacija početnih matrica

Već smo vidjeli da problem nenegativnih matričnih faktorizacija nije konveksan, zbog čega očekujemo da ima lokalnih minimuma. Jedna od najjednostavnijih metoda inicijalizacije početnih matrica bila bi nasumičnim generiranjem matrica, međutim, na taj način možemo pokrenuti algoritam daleko od stacionarne točke, zbog čega bi bilo potrebno mnogo iteracija kako bismo se približili nekoj od stacionarnih točki.

Dobra inicijalna aproksimacija ne samo da bi vodila ka konvergenciji dobrom lokalnom minimumu, već bi i smanjila broj iteracija potrebnih za doći do stacionarne točke. Većina strategija za inicijalizaciju nema garancije o graničnoj vrijednosti udaljenosti inicijalne točke od one koja bi bila optimalna. Ovdje su navedene neke od korištenih metoda inicijalizacije[27][31].

- **Poboljšana metoda nasumične inicijalizacije:** Umjesto direktnog korištenja nasumično generirane početne točke, dodatnim korakom može se znatno poboljšati inicijalna točka. Npr., za algoritme koji su osjetljivi na skaliranja (kao što su MU i PGD) možemo naći faktor za skaliranje

$$\alpha := \frac{\langle X, W_0H_0\rangle}{\langle W_0H_0, W_0H_0\rangle}, \quad W_0 = W_0\sqrt{\alpha}, \quad H_0 = H_0\sqrt{\alpha}, \tag{3.25}$$

  gdje je $\alpha$ zapravo optimalno rješenje problema

$$\min_{\alpha} \|X - \alpha W_0H_0\|_F^2. \tag{3.26}$$

- **Metode za klasteriranje:** Korištenjem centroida izračunatim pomoću metoda za klasteriranje, npr. klasteriranjem $k$-sredinama ili sfernim $k$-sredinama, kako bi se inicijalizirali stupci od $W$, i zatim inicijalizirao $H$ kao odgovarajuće skaliranje matrice indikatora (tj. $H_{kj} \neq 0 \iff X(:, j)$ pripada $k$-tom klasteru).

- **Dekompozicija na singularne vrijednosti:** Neka je $\sum_{k=1}^{r} u_k v_k^\top$ najbolja aproksimacija ranga $r$ od $X$. Svaki faktor ranka 1 $u_k v_k^\top$ može sadržavati pozitivne i negativne elemente (osim prvoga, prema Perron-Frobeniusovu teoremu[7]). Ako označimo $[x]_+ = max(x, 0)$, imamo

$$u_k v_k^\top = [u_k]_+[v_k^\top]_+ + [-u_k]_+[-v_k^\top]_+ - [-u_k]_+[v_k^\top]_+ - [u_k]_+[-v_k^\top]_+, \tag{3.27}$$

  gjde su prva dva faktora ranga 1 dekompozicije nenegativna. U [9] je predloženo da se svaki faktor ranka 1 u $\sum_{k=1}^{r} u_k v_k^\top$ zamijeni s $[u_k]_+[v_k^\top]_+$ ili $[-u_k]_+[-v_k^\top]_+$, odabirući onoga s većom normom i zatim skaliranjem.

- **Odabir podskupa stupaca**: Moguće je inicijalizirati stupce od $W$ korištenjem uzoraka, odnosno, inicijaliziranjem $W = X(:, \mathcal{K})$, gdje je $\mathcal{K}$ neki skup kardinalnosti $r$.

U praksi, moguće je koristiti više različitih inicijalizacija, a zatim zadržati najbolje rješenje.

## 3.5 Kriterij zaustavljanja

Postoji nekoliko vrsta kriterija zaustavljanja, mogu se bazirati na razvoju funkcije troška, uvjetima optimalnosti ili razlici funkcije troška između dvije iteracije. Ovi kriteriji se uglavnom kombiniraju s maksimalnim brojem iteracija ili vremenskim ograničenjem kako bi se osigurao završetak algoritma.

Kada je kriterij zaustavljanja razlika funkcije troška između dvije iteracije, tada stajemo kada ne uspijemo funkciju troška ili njenu skaliranu vrijednost smanjiti za neki određeni $\epsilon > 0$, tj.

$$F(W^{k+1}, H^{k+1}) - F(W^k, H^k) < \epsilon \quad \text{ili} \quad \frac{F(W^{k+1}, H^{k+1}) - F(W^k, H^k)}{F(W^k, H^k)} < \epsilon. \tag{3.28}$$

Ovaj kriterij nije dobar izbor za sve slučajeve s obzirom da algoritam može stati u točki koja je daleko od stacionarne. Bolji izbor bio bi kada bismo ograničili normu projiciranog gradijenta kako je sugerirano u [35]. Za NMF definiramo

$$[\nabla^P F_Y]_{ij} = \begin{cases} [\nabla F_Y]_{ij}, & \text{ako je } Y_{ij} > 0, \\ \min(0, [\nabla F_Y]_{ij}), & \text{ako je } Y_{ij} = 0, \end{cases} \tag{3.29}$$

gdje je $Y$ umjesto $W$ ili $H$. Tada predloženi uvjet glasi

$$\left\| \begin{bmatrix} \nabla^P F_{W^k} \\ \nabla^P F_{H^k} \end{bmatrix} \right\|_F \leq \epsilon \left\| \begin{bmatrix} \nabla F_{W^1} \\ \nabla F_{H^1} \end{bmatrix} \right\|_F \tag{3.30}$$

Izrazom (3.30) možemo provjeriti da li je gradijent trenutne točke $(W^k, H^k)$ bliži nuli za određen broj puta u odnosu na gradijent prve dobivene aproksimacije. Ukoliko je, znači da smo blizu nule, te ujedno i blizu stacionarne točke.

Primijetimo da ako imamo dvije jednake aproksimacije $\tilde{W}\tilde{H} = WH$ gornja gradijentna norma nije jednaka. Npr. postavljanjem $\tilde{W} = \gamma W$ i $\tilde{H} = \frac{1}{\gamma} H$ ne mijenja se aproksimacija, ali je sada gradijentna norma različita, tj. vrijedi

$$\begin{aligned} \left\| \begin{bmatrix} \nabla^P F_{\tilde{W}} \\ \nabla^P F_{\tilde{H}} \end{bmatrix} \right\|_F^2 &= \left\| \nabla^P F_{\tilde{W}} \right\|_F^2 + \left\| \nabla^P F_{\tilde{H}} \right\|_F^2 = \frac{1}{\gamma^2} \left\| \nabla^P F_W \right\|_F^2 + \gamma^2 \left\| \nabla^P F_H \right\|_F^2 \\ &\neq \left\| \begin{bmatrix} \nabla^P F_W \\ \nabla^P F_H \end{bmatrix} \right\|_F^2. \end{aligned} \tag{3.31}$$

Dvije faktorizacije koje rezultiraju jednakom aproksimacijom trebale bi se smatrati jednakima što se tiče preciznosti. Mogli bi npr. definirati $\gamma = \frac{\|\nabla^P F_W\|_F}{\|\nabla^P F_H\|_F}$, što bi minimiziralo izraz (3.31) i iz čega bi slijedilo $\|\nabla^P F_{\tilde{W}}\|_F = \|\nabla^P F_{\tilde{H}}\|_F$ međutim to nije dobro rješenje kada je samo jedan od gradijenata $\|\nabla^P F_{\tilde{W}}\|_F$ ili $\|\nabla^P F_{\tilde{H}}\|_F$ blizu nule.

Dakle, skaliranje utječe na gradijent $\begin{bmatrix} \nabla^P F_W \\ \nabla^P F_H \end{bmatrix}$ NMF problema i svaki kriterij zaustavljanja koji koristi informacije o gradijentu treba uzeti u obzir skaliranje. Kako bi se ograničio taj utjecaj, u [31] se predlaže sljedeća metoda skaliranja nakon svake iteracije:

$$\tilde{W}^k \leftarrow W^k D^k \quad \text{i} \quad \tilde{H}^k \leftarrow (D^k)^{-1} H^k \tag{3.32}$$

gdje je $D^k$ pozitivna dijagonalna matrica kojoj su elementi jednaki

$$(D^k)_{ii} = \sqrt{\frac{\left\|(H^k)_{i:}\right\|_2}{\left\|(W^k)_{:i}\right\|_2}}. \tag{3.33}$$

Ovo skaliranje osigurava da vrijedi $\|\tilde{W}_{:i}\|_F^2 = \|\tilde{H}_{i:}\|_F^2$ te pomaže smanjiti razliku između $\|\nabla^P F_{\tilde{W}}\|_F$ i $\|\nabla^P F_{\tilde{H}}\|_F$. Na ovaj način doprinosi se i smanjenju numeričke nestabilnosti. Napomenimo samo da kod korištenja (3.30) kao kriterija zaustavljanja potrebno je primijeniti jednako skaliranje na inicijalnu točku kao i na $(W^1, H^1)$.

# Poglavlje 4

# Primjeri primjene

Nenegativne matrične faktorizacije sve više dobivaju na važnosti u mnogim znanstvenim područjima gdje su ulazni podaci nenegativni. S obzirom da se mogu koristiti za modeliranje tema, učenje karakteristika, klasteriranje, vremensku segmentaciju, filtriranje i separaciju izvora te kodiranje, mogućnosti za primjenu su mnogobrojne.

Ovdje ćemo spomenuti neke od češćih primjena. S obzirom da je primjena nenegativnih matričnih faktorizacija u tekstualnoj analizi dosta slikovita, a ujedno i popularna, bit će dana detaljnija analiza primjene NMF-a na problemu modeliranja tema. Zatim slijedi kraći pregled primjene u obradi slika, sustavima za preporuku te glazbenoj analizi gdje je naglasak na vizualizaciji primjene, a ne na tehničkoj prirodi problema.

Osim navedenih primjena, NMF je našao i svoju primjenu u medicini, npr. za analizu EEG zapisa [15], bioinformatici za analizu ekspresije gena [11], astronomiji [8] te mnogim drugim područjima.

Jedna od novijih primjena nenegativnih matričnih faktorizacija je kod predviđanja nove veze u mrežama što će biti opisano u poglavlju *NMF na problemu predviđanja veze u mreži* s obzirom da će u poglavlju *Primjena NMF-a na predviđanju veze u društvenoj mreži* bit dan primjer konkretnog algoritma zasnovanog na NMF-u za rješenje tog problema.

## 4.1 Tekstualna analiza

Uzmimo da imamo neki skup od $n$ dokumenata. Tada možemo konstruirati matricu *pojam-dokument* dimenzije $m \times n$ gdje je $m$ broj pojmova koji se pojavljuju u cijelom skupu dokumenata, odnosno svojevrsnom rječniku baze dokumenata. $X(i, j)$ bi tada mogao označavati broj puta koji se $i$-ta riječ rječnika pojavljuje u $j$-tom dokumentu; u tom slučaju je svaki stupac od $X$ vektor koji je brojač pojmova u dokumentu.

U praksi, vrlo često se koristi *TFIDF* vrijednost (*eng. term frequency - inverse document frequency, tf-idf*). Tada vrijednost $X(i, j)$ uzima u obzir ne samo koliko se puta

određena riječ pojavljuje u dokumentu, već i broj puta koliko se pojavljuje u ostalim dokumentima. Na taj način mogu se profiltrirati zaustavne riječi te dati manja važnost riječima koje se pojavljuju u mnogo dokumenata pa samim time nisu specifične za neki dokument.

U ovom slučaju, rezultat faktorizacije matrice $X$ su matrice $W \in \mathbb{R}^{m \times r}$ i $H \in \mathbb{R}^{r \times n}$, odnosno matrice *pojam-tema* te *tema-dokument*. Dakle, svaki stupac matrice $W$ predstavlja jednu od $r$ tema, tj. grupira riječi rječnika prema tome kako se istovremeno mogu naći u nekoliko različitih dokumenata. Svaki stupac matrice $H$ pokazuje važnost tema za određeni dokument, tj. $H(k, j)$ prikazuje važnost $k$-te teme za $j$-ti dokument.

Ovakvu analizu ćemo sada provesti na bazi prvih paragrafa članaka dobivenih scrapeanjem stranice vecernji.hr po uzoru na [48]. Članci su podijeljeni u pet kategorija: politika (635), sport (298), hrana (948), životinje(337), obrazovanje(679), gdje je broj članaka prikazan u zagradama te se razlikuje za svaku kategoriju.

Predprocesiranje podataka sastoji se od nekoliko koraka: izbacivanje zaustavnih riječi te interpukcijskih znakova, transformacija riječi u osnovni oblik (npr. da ne bi rod ili padež utjecali na razlikovanje riječi) koristeći *stemmer* [38] te određivanje korpusa koji ćemo koristiti za određivanje koherencije.

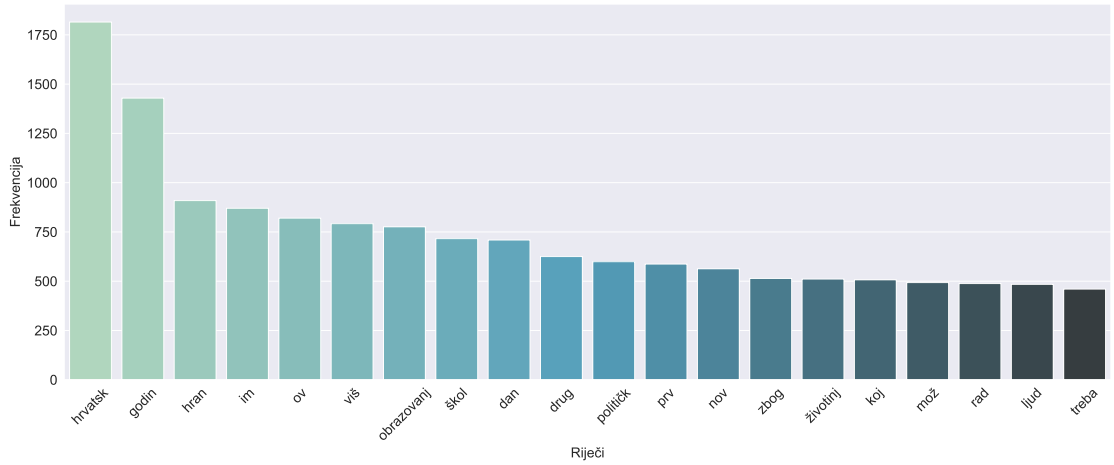
| title | text | word_count | processed_text |
|---|---|---|---|
| 'Nije normalno da grad ima 218 mjesnih odbora ... | Darko Klasić inženjer je radiologije, a politi... | 97 | [dark, klasić, inženjer, radiologij, politik, ... |
| Atena pokazuje mišiće Turskoj: Macron prodaje ... | Grčki premijer Kyriakos Mitsotakis najavio je ... | 80 | [grčk, premijer, kyriakos, mitsotakis, najavi,... |
| Domoljubi nose zaštitnu masku, kaže sad Trump ... | Negatori koronavirusa odjednom su počeli govor... | 108 | [negator, koronavirus, odjedn, počel, govori, ... |
| Što ili tko stoji iza novog krvoprolića na uža... | Oružani sukobi na granici između Armenije i Az... | 118 | [oružan, sukob, granic, između, armen, azerbajd... |
| Hoće li proračunati potezi iz SAD-a uvući Iran... | Nešto se zagonetno događa u Iranu. Već mjesec ... | 217 | [zagonetn, događ, iran, mjesec, dan, zemlj, po... |

Slika 4.1: Primjeri procesiranog teksta za nekoliko članaka

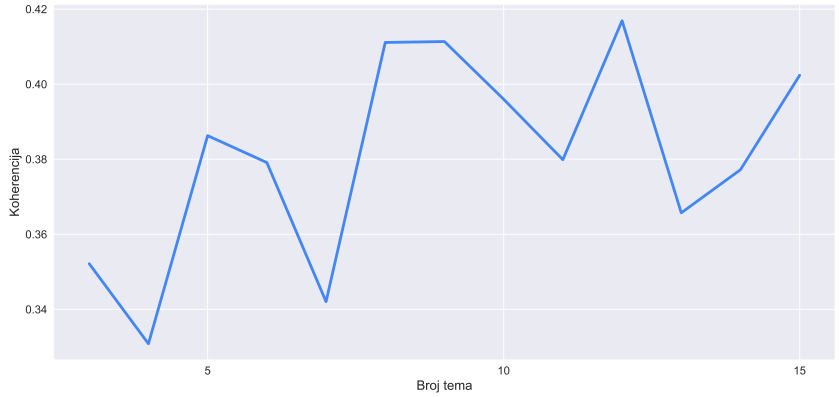
Na slici 4.1 može se vidjeti kako izgleda procesirani tekst za nekoliko prvih članaka. Vidimo da se sastoji od niza riječi koje su u nekom osnovnom obliku. Na slici 4.2 vidimo koje su najčešće korištene riječi u cijelom skupu članaka. Vidimo da su česte pojave nekih riječi koje bi mogle sugerirati neke teme na koje će se podijeliti riječi, kao što su *političk* i *životinj*.

Za odabrati broj tema $r$, odnosno broj stupaca matrice $W$, koristit ćemo koherenciju koja se zasniva na relativnoj udaljenosti između riječi unutar teme. Provjerit ćemo vrijednost koherencije za $3 \leq k \leq 15$ te izabrati najbolji rezultat za odrediti NMF. Slika 4.3 prikazuje dobivene vrijednosti koherencije.

Možemo primijetiti da je za $k = 12$ dobiven najbolji rezultat, što se razlikuje od broja različitih kategorija iz kojih su preuzeti članci. Sada ćemo provesti NMF za $r = k_{opt} = 12$ koristeći sklearn nmf model, inicijalizirajući početne matrice pomoću NNDSVD-a (eng. nonnegative double singular value decomposition) i uzimajući u obzir *TFIDF* vrijednosti.




Slika 4.2: Najčešće korištene riječi u cijelom skupu članaka




Slika 4.3: Vrijednosti koherencije za broj tema $3 \leq k \leq 15$

U ovom konkretnom slučaju, *TFIDF* vrijednost je izračunata pomoću TfidfVectorizer-a iz scikit-learna koji je implementiran na sljedeći način:

$$\text{tfidf}(t, d, D) = f_{td}\left(\log \frac{|D| + 1}{|\{d \in D : t \in d\}| + 1} + 1\right), \tag{4.1}$$

gdje su $t, d, D$ pojam, dokument i skup dokumenata redom, a $f_{td}$ frekvencija pojma $t$ u dokumentu $d$.

Provjerimo na slici 4.4 kakve se različite teme dobiju pomoću takvog modela te da li na ikakvi način oslikavaju početne kategorije iz kojih su preuzeti tekstovi.

| topic_num | topics |
|---|---|
| 0 | im ov dan zna ljud restoran svak mog |
| 1 | škol učenik osnovn razred srednj školsk |
| 2 | strank hdz političk izbor sdp predsjednik vlad... |
| 3 | životinj udrug pas zaštit ljubimc skloništ nap... |
| 4 | obrazovanj reform znanost ministric divjak kur... |
| 5 | hran proizvod namirnic sigurnost europsk potro... |
| 6 | sport sportsk sportaš olimpijsk aktivnost jani... |
| 7 | sol ulj jaj žlic luk mes priprem sastojc |
| 8 | hrvatsk držav politik europsk političk društv ... |
| 9 | post kun milijun godin viš cijen milijard |
| 10 | zoološk vrt posjetitelj zagrebačk grad zoo |
| 11 | kn champion cijen tenisic umbr majic sportsk mond |

Slika 4.4: Najčešće korištene riječi za $r = 12$ tema

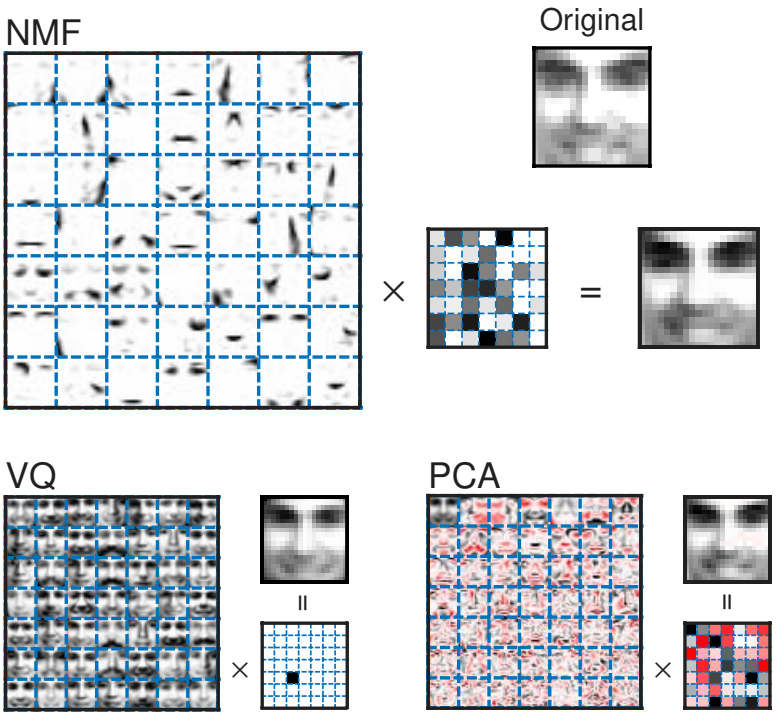
Sada možemo primijetiti zašto je broj tema veći nego broj početnih kategorija kojih je bilo pet. Npr. za kategoriju sport, vidimo da su nastale dvije slične teme, tema 6 koja govori o sportašima i sportskim aktivnostima te tema 11 koja govori o sportskoj modi. Nadalje, teme 1 i 4 vezane su uz obrazovanje, međutim tema 1 govori o učenicima i školi, dok tema 4 govori o reformama u obrazovanju. Sada možemo zaključiti da je NMF pronašao neke podkategorije u početnim kategorijama te je zato broj tema veći.

## 4.2 Obrada slika

Kako bismo neku sliku prikazali vektorizirano, treba svaki piksel reprezentirati njegovim intenzitetom kod sive verzije te slike. Tada za sliku dimenzija $d \times d$ dobijemo $d^2$ vrijednosti koje ju reprezentiraju. Na taj način možemo za $n$ slika konstruirati matricu $X \in \mathbb{R}^{m \times n}$ gdje je $m = d^2$ te koja ima sve elemente nenegativne. Tada svaki stupac matrice odgovara jednoj slici iz baze.

U ovom slučaju, NMF generira nenegativne matrice $W \in \mathbb{R}^{m \times r}$ i $H \in \mathbb{R}^{r \times n}$ gdje je svaka slika $X(:, j)$ aproksimirana linearnom kombinacijom vektora iz $W$. Upravo zbog toga stupce matrice $W$ nazivamo *baznim slikama*. S obzirom da su obje matrice nenegativne, NMF dopušta samo aditivnost, zbog čega omogućuje da dobijemo reprezentaciju koja se bazira na dijelovima.

Pogledajmo NMF na primjeru baze lica koja je analizirana u [34]. Lee i Seung su usporedili tri metode za analizu slika: nenegativne matrične faktorizacije (NMF), analizu glavnih komponenata (PCA) te vektorsku kvantizaciju (VQ). Za razliku od NMF-a koji zbog aditivnosti za reprezentaciju slike koristi dijelove, PCA i VQ uzimaju u obzir cjelinu.




Slika 4.5: Usporedba rekonstrukcija lica kod algoritama NMF, PCA i VQ

Konkretno, to bi značilo da su bazne slike kod NMF-a dijelovi lica kao što su oči, nos, usta, brkovi, dok su to kod PCA i VQ cijela lica, bila to prototipovi lica kao što je slučaj kod VQ-a ili nešto nalik na iskrivljene slike lica kao kod PCA, što se može primijetiti na slici 4.5.

Vrijednosti u matricama su prikazane odgovarajućom sivom bojom koja odgovara vrijednostima intenziteta, dok su nenegativne vrijednosti kod PCA prikazane crvenom bojom. U ovom primjeru, slike imaju $19 \times 19$ piksela, dok je pripradni broj baznih slika jednak 49, gdje se svaka bazna slika ponovno sastoji od $19 \times 19$ piksela.

Primijetimo da kod NMF-a, matrica baza te matrica koeficijenata su obje rijetko popunjene, s obzirom da nisu globalne te se sastoje od nekoliko verzija usta, očiju i drugih dijelova lica. Varijabilnost lica se dobije kombiniranjem različitih dijelova. Iako se svi

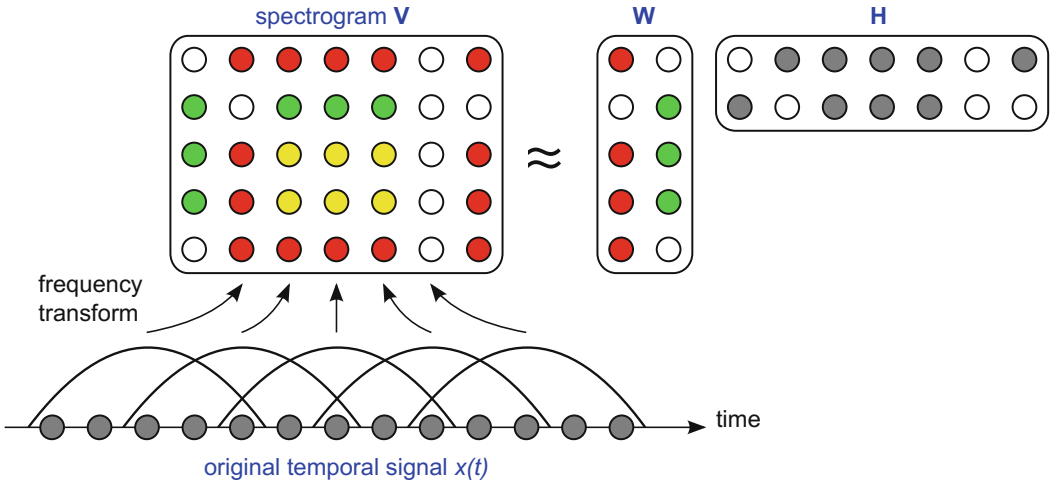
dijelovi lica (bazne slike) koriste u barem jednom licu, svako lice ne koristi sve dijelove. To rezultira rijetkim kodiranjem, za razliku od jediničnog kodiranja kod VQ-a, te potpuno distribuiranog kod PCA.

Još jedna od potencijalnih primjena NMF-a je kod prepoznavanja lica. U [29] dana je usporedba PCA i NMF-a kod prepoznavanja lica kod kojih su neki dijelovi prekriveni pomoću sunčanih naočala ili šala. U tim situacijama, NMF je pokazao bolje rezultate od PCA s obzirom da uzima u obzir dijelove pa je i dalje mogao dobro aproksimirati neskrivene diijelove lica kao što su oči i nos.

## 4.3 Analiza glazbe

Jedan od načina primjene NMF-a u analizi zvuka je kod transkripcije glazbe, odnosno prepoznavanja nota. Pogledat ćemo primjer iz [22] gdje su autori separirali individualne note kod zvučnog zapisa klavira.

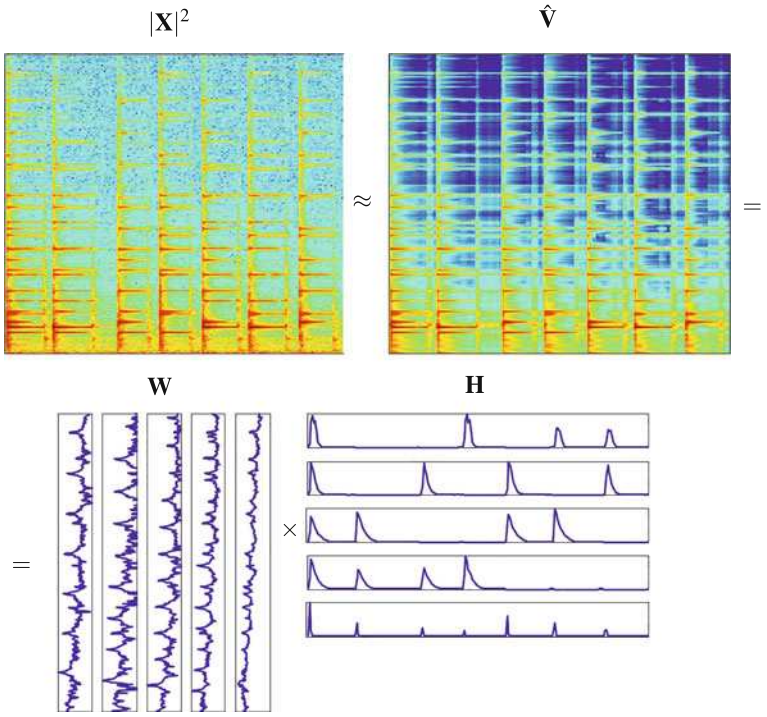
Slika[23] 4.6 prikazuje problem aproksimacije vremenskog signala $x(t)$ nenegativnim matricama $W$ i $H$. Ideja je neki duži signal podijeliti na kraće segmente i tada izračunati Fourierovu transformaciju odvojeno na svakom kraćem segmentu (*eng. short-time Fourier transform, STFT*). Dobivena matrica se zatim aproksimira dvjema nenegativnim matricama $W$ i $H$ gdje stupci od $W$ predstavljaju određene spektralne uzorke, a vrijednosti u $H$ vremenske aktivacije.




Slika 4.6: Prikaz aproksimacije vremenskog signala $x(t)$ nenegativnim matricama $W$ i $H$

U primjeru s klavirom korištena je glazbena sekvenca koja se sastoji od četiri različite note, u prvoj mjeri su sve note odsvirane odjednom te nakon toga u parovima gdje su pokrivene sve moguće kombinacije.

Autori su uspjeli za $r = 5$ dobiti potpunu separaciju individualnih nota. Svaka komponenta matrice $W$ odgovara po jednoj odsviranoj noti, dok zadnja komponenta odgovara zvukovima proizvedenim prijelazima kao što su otpuštanje pedale i pritiskanje određene tipke.




Slika 4.7: Aproksimacija spektrograma zvuka klavira komponiranog od četiri note za $r = 5$

Na slici[23] 4.7 možemo primijetiti uzorke u stupcima od $W$ koji predstavljaju bazne signale za početni spektrogram.

Neke od ostalih primjena nenegativnih matričnih faktorizacija u analizi glazbe i zvuka su redukcija šuma [53], kompresija zvuka [43] te separacija izvora zvuka [49].

## 4.4 Sustavi za preporuku

Sustavi za preporuku (*eng. recommender systems*) zajednički je naziv za skup algoritama kojima je cilj predvidjeti ocjenu odnosno preferencu korisnika na temelju informacija o korisnicima i proizvodima. Jedan od načina preporuke je tzv. kolaborativno filtriranje (*eng. collaborative filtering, CF*) koje stvara preporuku na osnovu drugih korisnika uzimajući u obzir sličnosti u ukusu.

| movieId | title | genres |
|---|---|---|
| 260 | Star Wars: Episode IV - A New Hope (1977) | Action\|Adventure\|Sci-Fi |
| 919 | Wizard of Oz The (1939) | Adventure\|Children\|Fantasy\|Musical |
| 1270 | Back to the Future (1985) | Adventure\|Comedy\|Sci-Fi |
| 1380 | Grease (1978) | Comedy\|Musical\|Romance |
| 163 | Desperado (1995) | Action\|Romance\|Western |
| 1288 | This Is Spinal Tap (1984) | Comedy |
| 2134 | Weird Science (1985) | Comedy\|Fantasy\|Sci-Fi |

Slika 4.8: Top 7 najbolje ocijenjenih filmova za jednog korisnika

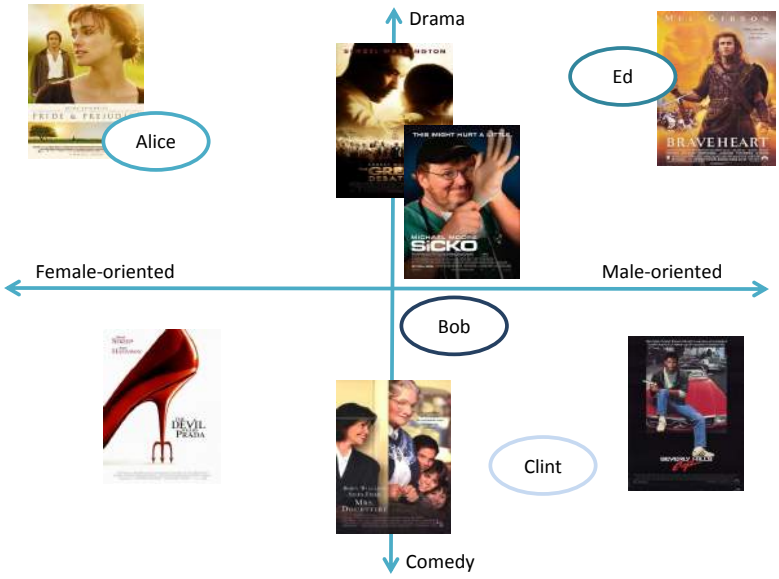
Jedan jednostavni sustav za preporuku filmova pomoću NMF-a dan je u [41]. Cilj je konstruirati matricu $X \in \mathbb{R}^{m \times n}$ gdje je $n$ broj korisnika, $m$ broj filmova te $X(i, j)$ označava ocjenu $i$-tog filma za $j$-tog korisnika. Pomoću dobivenih nenegativnih matrica $W \in \mathbb{R}^{m \times r}$ i $H \in \mathbb{R}^{r \times n}$, aproksimira se matrica $X$ i tako dobiju vrijednosti koje nedostaju. Za svakog korisnika preporuka je onaj film koji ima najveću vrijednost u tako rekonstruiranoj matrici, a još ga nije ocijenio.

| movieId | title | genres |
|---|---|---|
| 260 | Killer's Kiss (1955) | Crime\|Film-Noir\|Romance |
| 919 | My Life as a Dog (Mitt liv som hund) (1985) | Comedy\|Drama |
| 1270 | 9 (2009) | Adventure\|Animation\|Sci-Fi |
| 1380 | White Sound  The (Das weiÌŬe Rauschen) (2001) | Comedy\|Musical\|Romance |
| 163 | Scout  The (1994) | Comedy\|Drama |
| 1288 | Spellbound (1945) | Mystery\|Romance\|Thriller |
| 2134 | Bride of the Monster (1955) | Comedy\|Fantasy\|Sci-Fi |

Slika 4.9: Preporuka 7 filmova za korisnika dobivena pomoću NMF-a

Na primjeru jednog određenog korisnika, možemo vidjeti 7 filmova kojima je dao najveću ocjenu (Slika 4.8), a zatim 7 filmova koje bi mu preporučio opisani NMF za preporuku (Slika 4.9).

U prikazanim tablicama se mogu primijetiti neke preference korisnika, kao da preferira filmove koji nisu trileri, već komedije, avanturističke, znanstveno-fantastične te romantične.




Slika 4.10: Prikaz pozicioniranja korisnika u koordinatnom sustavu latentnih faktora

Ako uzmemo za primjer sliku 4.10 iz [39] gdje je Louppe prikazao jedan jednostavni prikaz pozicioniranja korisnika s obzirom na neke latentne faktore kao što su spol ili ozbiljnost filma, tada možemo primijetiti da bi se naš korisnik nalazio u donjem srednjem dijelu koordinatnog sustava.

# Poglavlje 5

# NMF na problemu predviđanja veze u mreži

Mnogi društveni, biološki, informacijski te tehnološki sustavi mogu se prikazati u obliku mreže gdje čvorovi označavaju entitete, a bridovi poveznicu, odnosno interakciju između čvorova. U mnogim slučajevima, prikupljeni podaci o nekoj mreži nisu potpuni te zbog toga želimo saznati veze koje nedostaju ili pak predvidjeti neke veze koje bi mogle nastati u budućnosti.

Primjerice, kod kompleksnih bioloških mreža korisno je saznati postoji li veza između dva čvora s obzirom da laboratorijski eksperimenti često predstavljaju visok trošak [12]. Nadalje, kod društvenih mreža, predviđanje veza može pomoći kod preporuke novih prijatelja ili entiteta koji su korisniku od interesa kao što je preporuka restorana, proizvoda ili turističkih odredišta. Jedna od korisnijih te zasigurno zanimljivijih primjena je kod prevencije kriminala i terorističkih aktivnosti gdje predviđanje veza može otkriti skrivenu povezanost između kriminalaca [3].

Većina postojećih metoda koje rješavaju problem predviđanja veze u mreži uzima u obzir samo topologiju mreže te zanemaruje latentne značajke čvorova koji se ne mogu direktno iščitati iz mreže. Zbog toga je zadnjih godina proučavana primjena nenegativnih matričnih faktorizacija na ovom problemu gdje se pokazalo da ovakav model može pronaći potencijalnu strukturu povezanosti entiteta mreže te ostavlja mogućnost proširenja modela eksternim atributima.

U ovom poglavlju bit će dan opis predviđanja nove veze u mreži te zatim pregled nekoliko postojećih rješenja primjenom nenegativnih matričnih faktorizacija.

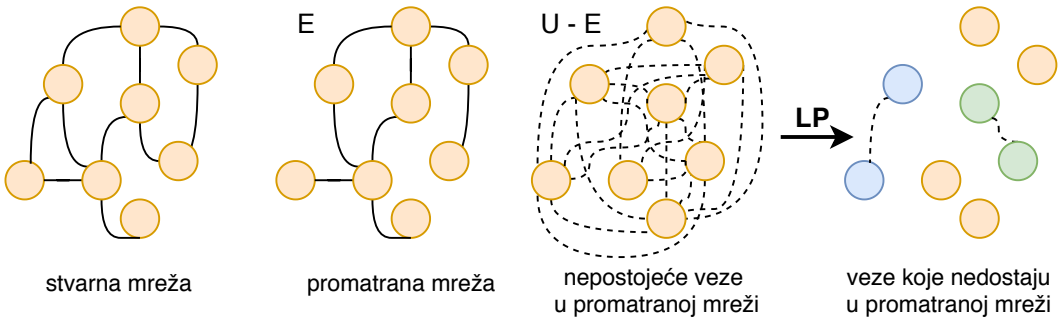
Spomenimo samo da kod kompleksnih mreža zbog velikog broja podataka, nepouzdanosti mjernih instrumenata te ograničene reprezentacije, osim što može doći do nedostatka postojećih veza, velika je mogućnost pojave lažnih veza. Stoga osim predviđanja novih veza postoji i problem identifikacije lažnih veza (*eng. spurious links*).

## 5.1 Problem predviđanja veze u mreži

Neka je dan neusmjereni, netežinski graf $G = (V, E)$ gdje $V$ i $E$ reprezentiraju skup čvorova, odnosno skup bridova. Tada su $N = |V|$ i $M = |E|$ brojevi čvorova i brojevi bridova redom.

Sada mreža može biti prikazana pomoću matrice susjedstva $A \in \{0, 1\}^{N \times N}$ gdje je $A_{ij} = A_{ji} = 1$ ako postoji veza između čvorova $i$ i $j$, inače vrijedi $A_{ij} = A_{ji} = 0$.

Ako s $U$ označimo univerzalni skup svih mogućih bridova u grafu $G$, tada $U - E$ predstavlja nepostojeće veze u promatranom grafu. Cilj predviđanja veze u mreži (*eng. link prediction, LP*) je pronaći vezu iz skupa $U - E$ koja nedostaje u trenutnom grafu što je prikazano slikom 5.1, odnosno predvidjeti novu vezu u budućnosti.




Slika 5.1: Problem predviđanja veze iz skupa nepostojećih veza $U - E$

Taj problem se svodi na pronalazak vjerojatnosti da postoji veza između neka dva čvora, odnosno vrijednosti *Similarity(u,v)* koja pokazuje sličnost čvorova $u$ i $v$. Što je vrijednost $Similarity(u, v)$ veća, to su $u$ i $v$ sličniji. Nepostojeće veze u grafu moguse poredati silazno prema vrijednosti $Similarity(u, v)$. Tada su veze, odnosno bridovi na početku poretka oni koji imaju najveću vjerojatnost da postoje u grafu ili da će postojati u budućnosti. Sada ćemo navesti nekoliko načina računanja sličnosti prema [19] [25] .

**Zajednički susjedi** Pretpostavimo da je čvor $v \in V$; tada skup susjednih čvorova od $v$ možemo označiti kao $\Gamma(v) := \{t \,|\, (t, v) \in E \vee (v, t) \in E \wedge t \neq v\}$. Zajednički susjedi od $u$ i $v$ odnose se na susjede koji se nalaze u skupu susjeda od $u$ i u skupu susjeda od $v$. Sada se sličnost čvorova može izračunati kao

$$Similarity(u, v) = |\Gamma(u) \cap \Gamma(v)|\,. \tag{5.1}$$

Ova metoda zasniva se na pretpostavci da čvorovi koji imaju više zajedničkih susjeda imaju veću vjerojatnost da će se nekada spojiti u budućnosti.

**Preferencijalna privrženost** Uz metodu zajedničkih susjeda, jedna je od osnovnih metoda računanja sličnosti. Svodi se na pretpostavku da čvorovi koji imaju veći stupanj imaju veću vjerojatnost da dobiju novu vezu. Kod pronalaska sličnosti, treba uzeti u obzir stupanj oba čvora za koje se gleda možebitna veza. Tada je vjerojatnost generiranja veze između čvorova $u$ i $v$ direktno proporcionalna umnošku stupnjeva čvorova, odnosno

$$Similarity(u, v) = |\Gamma(u)| * |\Gamma(v)| . \tag{5.2}$$

Primijetimo da indeks sličnosti ne zahtijeva nikakve informacije o susjedima, stoga metoda preferencijalne privrženosti ima najniži trošak izračuna.

**Jaccardov koeficijent** Poznat i kao Jaccardov indeks ili Jaccardov koeficijent sličnosti, statistička je mjera za sličnost koja se koristi kod usporedbe dva skupa. Kod predviđanja veze, za svaki par čvorova računa se kvocijent broja zajedničkih susjeda čvorova s brojem ukupnog broja susjeda, odnosno

$$Similarity(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}. \tag{5.3}$$

**Adamic-Adar indeks** Za razliku od preferencijalne privrženosti, daje veći indeks sličnosti onim čvorovima manjeg stupnja. Inicijalno je osmišljen za mjerenje sličnosti između osobnih stranica. Ako uzmemo u obzir zajedničke susjede neka dva čvora, tada će oni zajednički susjedi koji imaju manje svojih susjeda davati veći doprinos Adamic-Adar indeksu od onih zajedničkih susjeda s velikim brojem svojih susjeda, tj.

$$Similarity(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|}. \tag{5.4}$$

U stvarnosti, ovaj indeks može se interpretirati u slučaju društvenih mreža: ako zajednički poznanik dvije osobe ima mnogo prijatelja, tada je manja vjerojatnost da će međusobno upoznati dvoje svojih prijatelja nego u slučaju da ima manji broj poznanika. Ovakva strategija pokazuje dobre rezultate kod predviđanja prijateljstva, međutim kod koautorstva je pokazalo loše rezultate [1].

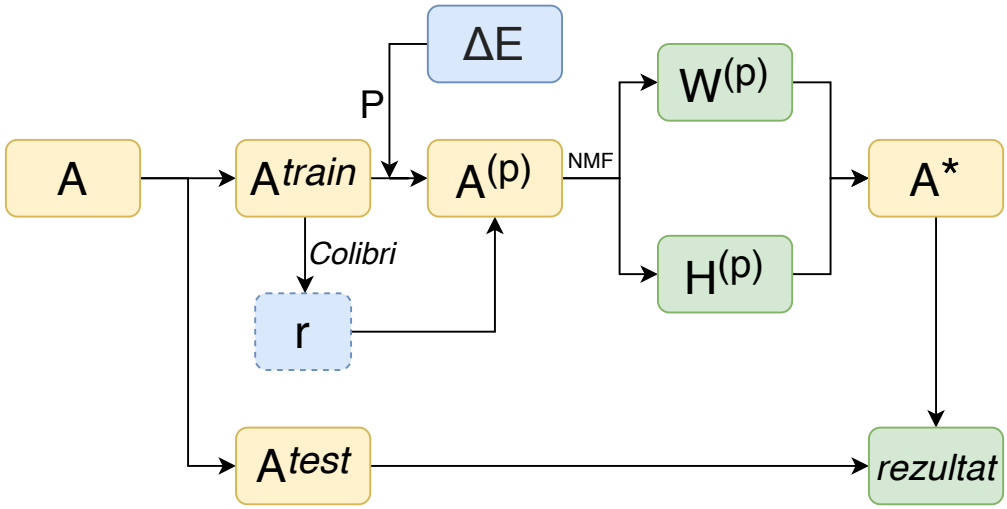
**Katz metoda** Glavna ideja ove metode je da što više staza postoji između neka dva čvora, to je veća sličnost čvorova. Osim broja staza, uzima i obzir njihovu dužinu tako da veći doprinos indeksu sličnosti daju one staze koje su kraće. Izraz se može zapisati kao

$$Similarity(u, v) = \sum_{l=1}^{\infty} \beta^l \cdot |path_{uv}^{<l>}| \tag{5.5}$$

gdje je $|path_{uv}^{<l>}|$ broj staza između čvorova $u$ i $v$ duljine $l$, a parametar $\beta$ za koji vrijedi $0 < \beta < 1$ služi kako bi kontrolirao doprinos staze u ovisnosti o duljini. S obzirom da Katz metoda uzima u obzir topologiju cijele mreže, složenija je od prijašnjih metoda. Računska složenost ove metode je $O(N^3)$, stoga nije pogodna za velike mreže. Međutim, ovo je jedna od metoda koja pokazuje veliku točnost kod predviđanja.

## 5.2 Perturbacija matrice susjedstva

U [51] 2016. predložena je metoda zasnovana na perturbaciji matrice susjedstva. Glavna motivacije ove metode je prisustvo svojevrsnog šuma kod društvenih mreža te nepredviđenih, ali stvarnih veza. Primjerice, dvoje ljudi koji nemaju zajedničke prijatelje mogu se igrom slučaja sprijateljiti, što se ne može objasniti nekim općenitim modelom predviđanja veza.




Slika 5.2: Metoda zasnovana na perturbaciji za rješenje problema predviđanja linka

Procedura se sastoji od sljedećih nekoliko koraka prikazanih slikom 5.2:

- mreža, odnosno skup bridova $E$ nasumično se podijeli u skup za učenje $E^{train}$ te testni skup $E^{test}$, nakon čega se kreiraju matrice susjedstva $A^{train}$ i $A^{test}$ koje odgovaraju skupovima bridova $E^{train}$ i $E^{test}$ redom
- broj skrivenih značajki $r$ automatski se optimizira pomoću *Colibri* metode za matricu susjedstva $A^{train}$
- konstruira se perturbacijski skup $\Delta E$ kako bi se $P$ puta perturbirao skup $E^{train}$ i na taj način dobio niz novih perturbiranih matrica $A^{(p)}$, $p = 1, ..., P$

- pomoću perturbacijskih matrica $A^{(p)}$ i reduciranog ranga $r$ dobije se niz baznih matrica $W^{(p)}$ i matrica koeficijenata $H^{(p)}$ koristeći za ciljnu funkciju Frobeniusovu mjeru te Kullback-Leibler divergenciju

- konačno se izračuna matrica sličnosti pomoću baznih matrica i matrica sličnosti kao

$$A^* = \frac{1}{P}\sum_{p=1}^{P} W^{(p)}H^{(p)}. \tag{5.6}$$

Za konstruirati perturbacijski skup $\Delta E$ prvo se odabere parametar $\eta$ koji pokazuje količinu perturbiranosti skupa $E^{train}$. Postoje dvije vrste perturbiranosti; kod prve se nasumično odabire $\eta(M-L)$ bridova koji se brišu iz $E^{train}$, čime se pokušava riješiti problem šuma u mreži, dok se kod druge nasumično odabire $\eta(M-L)$ bridova iz $U - E^{train}$ te dodaje u $E^{train}$ s ciljem da se riješi problem stvarnih, neočekivanih veza.

Četiri predložene metode su sljedeće: $NMF-D1$, $NMF-A1$, $NMF-D2$ i $NMF-A2$, gdje $D$ označava metodu kod koje se koristi perturbacija za brisanje bridova, $A$ označava perturbaciju za dodavanje bridova, te brojevi 1 i 2 označavaju funkciju cilja koja se koristi: Frobeniusovu mjeru i Kullback-Leiblerovu divergenciju redom.

Autori članka su proveli navedene metode na primjeru 15 mreža te dobivene rezultate usporedili s nekim klasičnim metodama baziranim na indeksu sličnosti. Najbolje rezultate među ostalim dala je metoda $NMF-A2$, dok ostale tri imaju bolje rezultate od prosjeka. Nadalje, metode koje koriste KL divergenciju pokazale su se uspješnijim od onih s Frobeniusovom mjerom.

Kada promatramo korelaciju najuspješnije metode $NMF-A2$ sa statistikama korištenih mreža, možemo primijetiti da ima pozitivnu korelaciju s prosječnim stupnjem i koeficijentom klasteriranja, dok je korelacija negativna za broj čvorova grafa.

## 5.3 Sličnost vektora težinske matrice

Autori članka [13] osim topologije mreže, uzeli su u obzir i atribute čvorova. Tako su faktorizirali matricu susjedstva $A \in \{0,1\}^{n\times n}$ i matricu atributa $B \in \mathbb{R}^{n\times m}$ gdje je $m$ broj atributa. Vrijednosti matrice $B$ mogu se definirati unutar intervala $[0,1]$ normaliziranjem redaka.

Faktoriziranje nenegativnih matrica $A$ i $B$ može se promatrati kao sljedeća dva NMF problema

$$\min_{W,H\geq 0} \|A - WH\|_F^2, \tag{5.7}$$

$$\min_{W^B,H^B\geq 0} \|A - W^BH^B\|_F^2, \tag{5.8}$$

gdje je $r$ reducirani rang u oba slučaja, tj. dimenzija matrica su $W \in \mathbb{R}^{n\times r}$, $H \in \mathbb{R}^{r\times n}$, $W^B \in \mathbb{R}^{n\times r}$ i $H^B \in \mathbb{R}^{r\times m}$. Sljedeći korak je preslikati matrice $A$ i $B$ u isti latentni prostor, međutim istovremenim faktoriziranjem ne može se osigurati identičnost matrica $H$ i $H^B$ . Zbog toga se uvodi matrica $H^*$ kako bi se minimizirala udaljenost između matrica $H$ i $H^B$ pa je izraz koji se treba minimizirati

$$\min_{W,H,W^B,H^B\geq 0} ||A - WH||_F^2 + ||A - W^B H^B||_F^2 + \lambda||H - H^*||_F^2 + \mu||H^B - H^*||_F^2 \tag{5.9}$$

$$||W||_1 = 1,\ ||W^B||_1 = 1. \tag{5.10}$$

Kako bi se pomakao uvjet (5.10), definiraju se pomoćne dijagonalne matrice

$$\begin{aligned} Q &= diag(\textstyle\sum_i w_{i1}, \sum_i w_{i2}, \cdots, \sum_i w_{ir}),\\ Q^B &= diag(\textstyle\sum_i w^B_{i1}, \sum_i w^B_{i2}, \cdots, \sum_i w^B_{ir}). \end{aligned} \tag{5.11}$$

Sada se matrice $W$ i $W^B$ mogu normalizirati kao $WQ^{-1}$ i $W^BQ^{B^{-1}}$. S obzirom da vrijedi $WH = (WQ^{-1})(QH)$ i $W^BH^B = (W^BQ^{B^{-1}})(Q^BH^B)$, matrice $H$ i $H^B$ mogu se normalizirati kao $HQ$ i $H^BQ^B$ redom. Sada je problem iz (5.9) i (5.10) ekvivalentan s

$$\min_{w,h,w^B,h^B} J(w, w^B, h, h^B, h^*),\ \text{t.d vrijedi } W, H, W^B, H^B \geq 0, \tag{5.12}$$

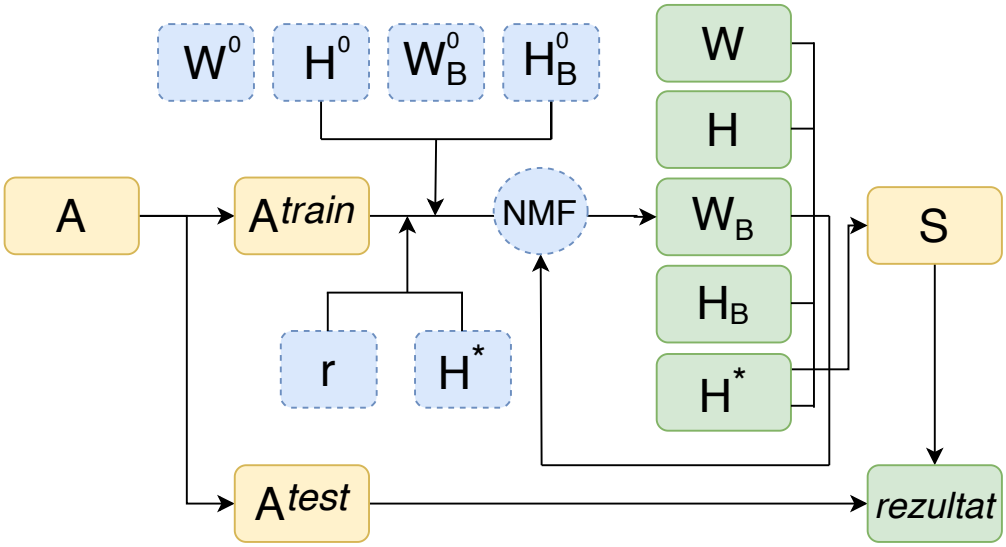
gdje je

$$\begin{aligned} J(w, w^B, h, h^B, h^*) = ||A - WH||_F^2 + ||B - W^BH^B||_F^2 +\\ + \lambda||QH - H^*||_F^2 + \mu||Q^BH^B - H^*||_F^2. \end{aligned} \tag{5.13}$$

Za rješenje problema (5.12) koristi se iterativna metoda. Fiksiranjem četiri od pet varijabli $W, H, W^B, H^B, H^*$ može se minimizirati funkcija $J$ s obzirom na petu varijablu. U [13] mogu se naći pravila osvježenja za ovih pet varijabli.

Sada se algoritam sastoji od sljedećih koraka, kako je prikazano na slici 5.3:

- matrica susjedstva $A$ se podijeli u skup za učenje $A^{train}$ i skup za testiranje $A^{test}$
- inicijaliziraju se parametri $r, \lambda, \mu$ te početne matrice $W, H, W^B, H^B, H^*$
- dok nije zadovoljen uvjet konvergencije
  - osvježi se $W$ prema pravilu osvježenja
  - izračunaju se elementi dijagonale od $Q$ s obzirom na $W$
  - osvježe se matrice $W^B, H, H^B, H^*$
- rezultat je matrica sličnosti $S$ koja se dobije kao sličnost stupaca matrice $H^*$




Slika 5.3: Metoda zasnovana na sličnosti težinske matrice $H^*$

Autori su usporedili navedeni algoritam na 13 mreža sa 7 drugih metoda koje su bazirane na indeksu sličnosti. Ova metoda se pokazala uspješnijom na gotovo svim mrežama. U sljedećem odjeljku *Uparivanje multivarijantnih informacija* dana je usporedba metode bazirane na uparivanju multivarijantnih informacija s ovom metodom, gdje je druga metoda ipak bila uspješnija.

## 5.4 Uparivanje multivarijantnih informacija

U [52] autori su pokušali u potpunosti integrirati pomoćne informacije o mreži kako bi poboljšali učinkovitost predviđanja unatoč mreži čija je topologija nepotpuna. Zbog toga su odlučili faktorizirati dvije matrice: matricu susjedstva $A \in \{0,1\}^{N\times N}$ i matricu sličnosti za pomoćne atribute $S \in \mathbb{R}^{N\times N}$ kao $A = W_1H_1$ i $S = W_2H_2$. Zatim se preslikaju dobivene informacije u dva prostora manjeg ranga u kojima su $W_1$ i $W_2$ baze latentnih prostora, odnosno imamo

$$\min_{W_1,H_1\geq 0} \|A - W_1H_1\|_F^2 \tag{5.14}$$

$$\min_{W_2,H_2\geq 0} \|S - W_2H_2\|_F^2. \tag{5.15}$$

Cilj je sada upariti multivarijantne informacije kako bi se poboljšala točnost predviđanja veze, stoga možemo kombinirano zapisati formule (5.14) i (5.15) kao

$$Q = \min_{W_1,H_1\geq 0} \|A - W_1H_1\|_F^2 + \min_{W_2,H_2\geq 0} \|S - W_2H_2\|_F^2. \tag{5.16}$$

U izrazu (5.16) kombinirane su topološka struktura i pomoćni atributi, međutim nisu integrirani u isti prostor značajki. Zbog toga je potrebno pronaći zajedničku matricu $W$, tj. $W = W_1 = W_2$. Nadalje, autori su dodali dodatne parametre kako bi izbjegli *overfitting* i izbalansirali utjecaj pomoćnih atributa u odnosu na topologiju mreže te je konačna funkcija cilja sljedeća:

$$Q = \min_{W,H_1,H_2\geq 0} \left(\|A - WH_1\|_F^2 + \alpha\|S - WH_2\|_F^2 + \beta\left(\|H_1\|_F^2 + \|H_2\|_F^2\right)\right). \tag{5.17}$$

Uvođenjem Lagrangeovih multiplikatora te pomoću parcijalnih derivacija od Lagrangeove funkcije s obzirom na $W, H_1$ i $H_2$ dobiju se izrazi za sljedeća pravila osvježenja

$$\begin{aligned} W &\longleftarrow W \circ \frac{\left[AH_1^\top + \alpha S H_2^\top\right]}{\left[WH_1H_1^\top + \alpha WH_2H_2^\top\right]} \\ H_1 &\longleftarrow H_1 \circ \frac{[W^\top A]}{[W^\top WH_1 + \beta H_1]} \\ H_2 &\longleftarrow H_2 \circ \frac{[\alpha W^\top S]}{[\alpha W^\top WH_2 + \beta H_2]} \end{aligned} \tag{5.18}$$

Još ostaje pitanje određivanja matrice sličnosti za pomoćne atribute. Postoje dvije vrste pomoćnih atributa: oni koji su dobiveni iz strukture mreže, zbog čega se nazivaju internima te oni koji se odnose na atribute čvorova, odnosno eksterni.

Eksterne atribute potrebno je predprocesirati, pa tako one informacije koje imaju vrijednost (kao što su godine, visina, broj članaka...) mogu se direktno preslikati u odgovarajuće vrijednosti, dok je onim informacijama koje nemaju brojčanu vrijednost dodijeljena vrijednost unutar nekog intervala. Nadalje, za one atribute koji imaju dva moguća statusa koriste se brojevi 0 i 1 kako bi se odredilo kojoj kategoriji entitet pripada.

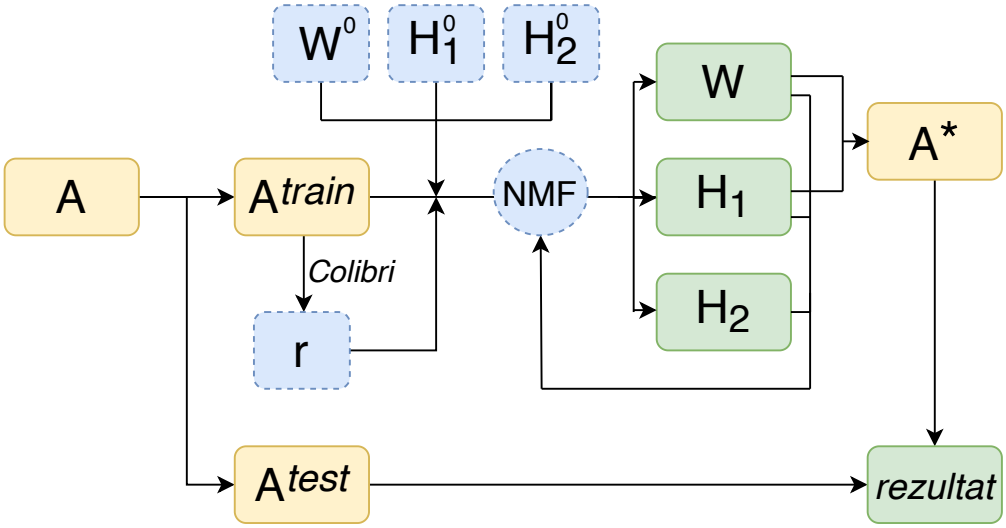
Sada se atributi svakog pojedinog čvora mogu predstaviti vektorom $Z_m$ gdje je $m$ broj atributa. Matrica $Z \in \mathbb{R}^{n\times m}$ sadrži niz redaka gdje svaki redak predstavlja vektor atributa koji pripada čvoru, odnosno element $Z_{ij}$ predstavlja vrijednost $j$-tog atributa za $i$-ti čvor. Prije sljedećeg koraka potrebno je još normalizirati vrijednosti atributa s obzirom na stupce.

Sljedeći korak je izračunati sličnosti između vektora $Z_m$ za svaka dva čvora te tako formirati matricu sličnosti atributa $S$. Autori su odabrali mjeru sličnosti baziranu na kosinusu, tj.

$$S_{ij} = \frac{\sum_{l=1}^{m} Z_{il}Z_{jl}}{\sqrt{\sum_{l=1}^{m} Z_{il}^2 \sum_{l=1}^{m} Z_{jl}^2}} \tag{5.19}$$

Ukratko, metoda se može opisati pomoću sljedećih koraka što je prikazano na slici 5.4:

- matrica susjedstva $A$ se podijeli u skup za učenje $A^{train}$ i skup za testiranje $A^{test}$
- pomoću *Colibri* metode dobije se broj latentnih značajki $r$
- inicijaliziraju se $W, H_1$ i $H_2$
- dok nije zadovoljen uvjet konvergencije, osvježavaju se $W, H_1$ i $H_2$ pomoću izraza (5.18)
- konačni rezultat je umnožak $A^* = WH_1$




Slika 5.4: Metoda zasnovana na uparivanju multivarijantnih informacija

Autori su testirali navedeni algoritam na 13 mreža te ga usporedili s 11 postojećih metoda. U globalu predložene metode pokazale su se uspješnijima od ostalih metoda baziranih na indeksu sličnosti te modernim metodama kao što je NMF baziran na sličnosti vektora težinske matrice koji je opisan u prošlom odjeljku *Sličnost vektora težinske matrice*. Jedina metoda koja je pokazala bolje rezultate od navedenih je ona bazirana na perturbacijama, a koja je opisana u odjeljku *Perturbacija matrice susjedstva*.

# Poglavlje 6

# Primjena NMF-a na predviđanju veze u društvenoj mreži

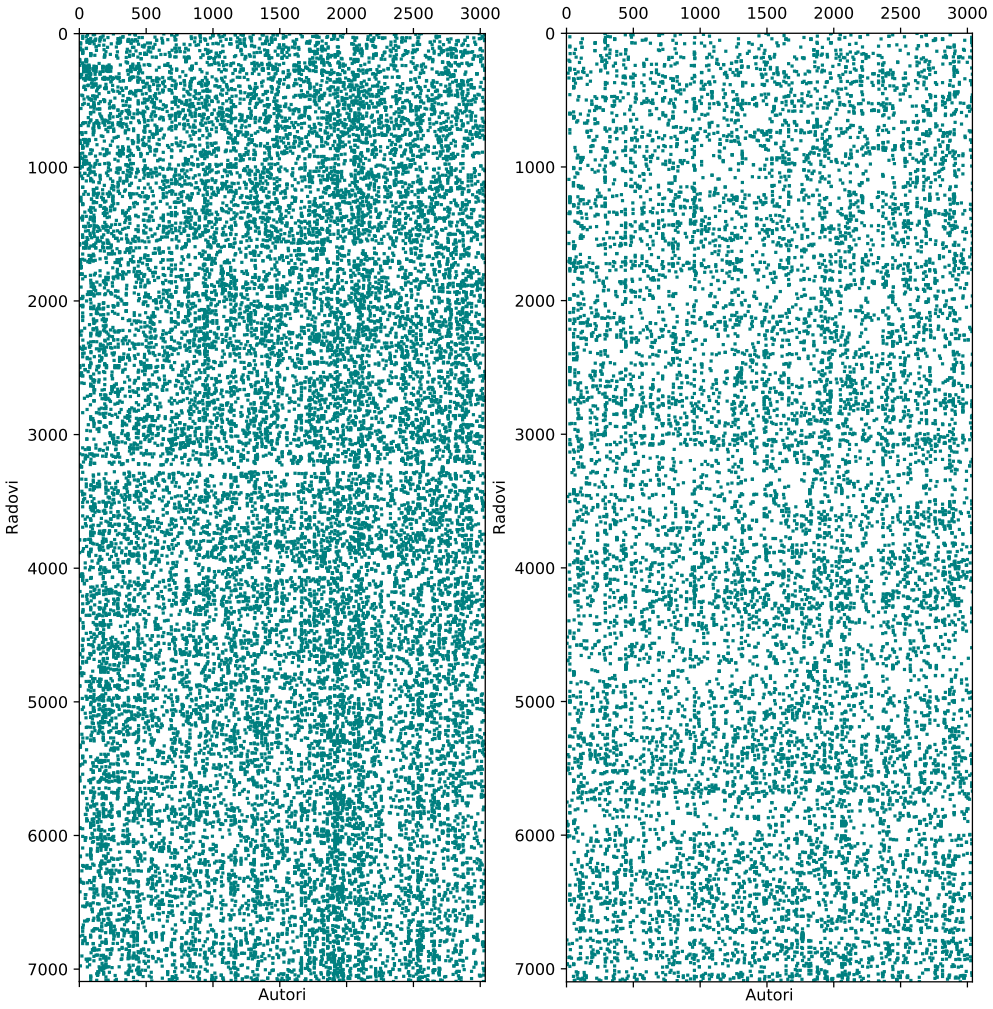
U ovom poglavlju bit će dan primjer primjene nenegativnih matričnih faktorizacija na problemu predviđanja nove veze u društvenoj mreži, konkretnije, mreži koautorstava članaka. Korišteni skup podataka su članci navedeni u hrvatskoj znanstvenoj bibliografiji CROSBI.

Prvo će biti opisan korišteni skup podataka i njegovo predprocesiranje, a zatim dan algoritam za predviđanje veze u mreži koautorstava koji se sastoji od nekoliko koraka: određivanje dodatnih značajki, procjena reduciranog ranga, perturbacija matrice susjedstva te računanje matrice sličnosti koristeći NMF. Nakon toga slijedi evaluacija rezultata dobivenih navedenim algoritmom te, konačno, zaključak primjene.

## 6.1 Skup podataka i predprocesiranje

Polazna baza podataka su sva koautorstva iz bibliografije CROSBI označena kao da pripadaju Prirodoslovno-matematičkom fakultetu u Zagrebu, odnosno svi radovi gdje je barem jedan od autora s PMF-a. Za razmatranje ćemo uzeti sve radove objavljene između 2005. i 2020. godine. S obzirom da su u skupu podataka mnogi koautori strani znanstvenici, a i navedeni su neki diplomski radovi, postoji mnogo autora koji su u cijelom skupu podataka napisali samo jedan ili dva rada pa ih ne smatramo aktivnim znanstvenicima PMF-a. Zbog toga ćemo uzeti sve one autore koji su u navedenom intervalu napisali barem tri rada.

Sada dobiveni skup podataka možemo podijeliti na dva dijela: skup za učenje i skup za testiranje. Kao skup za učenje uzet ćemo koautorstva ostvarena između 2005. i 2014. godine, a zatim ćemo algoritam testirati na onim koautorstvima ostvarenim između 2015. i 2020. godine. Dakle, skup autora koje promatramo su oni autori koji se pojavljuju u skupu godina za učenje, te za njih pokušavamo predvidjeti nove suradnje u godinama za testiranje.

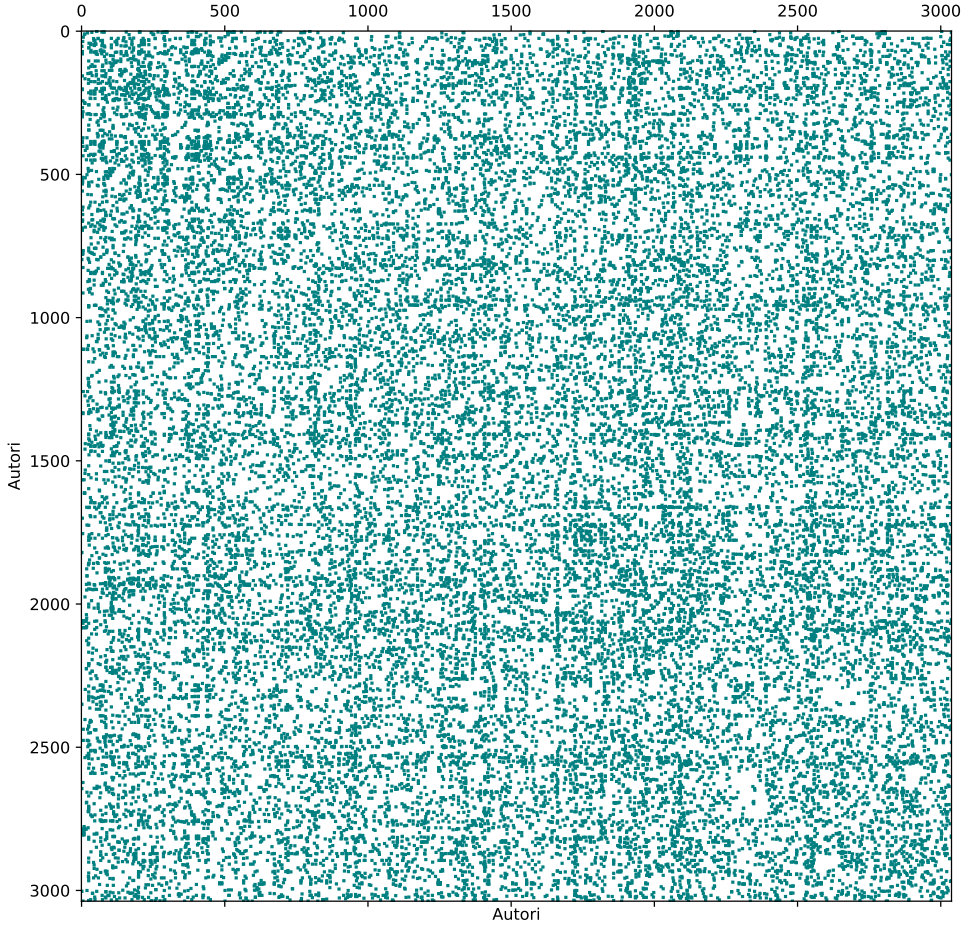



Slika 6.1: Prikaz matrica *članak-autor* za skup za učenje (lijevo) i skup za testiranje (desno)

Slika 6.1 prikazuje matrice *CON_train* i *CON_test* gdje je $CON_{ij} = 1$ ako je u $i$-tom radu sudjelovao $j$-ti autor u godinama za učenje, odnosno godinama za testiranje, inače $CON_{ij} = 0$ . Iz matrice *CON_train* smo izbacili sve članke gdje je sudjelovao jedan autor s obzirom da nam ne govore ništa o povezanosti s drugim autorima, te sve autore koji nisu surađivali s nijednim od ostalih autora. Na ovaj način dobijemo konačan broj autora 3038, broj članaka razmatranih u skupu za učenje 7093 te broj članaka razmatranih u skupu za testiranje 7098.

Dalje je potrebno stvoriti matricu susjedstva *A_train* gdje vrijedi da je $A\_train_{ij} = 1$

ukoliko su autori $i$ i $j$ ostvarili suradnju u godinama za učenje, inače $A\_train_{ij} = 0$. Tako dobivena matrica je simetrična s obzirom da je koautorstvo neusmjereno svojstvo. Slika 6.1 prikazuje navedenu matricu.




Slika 6.2: Matrica susjedstva *autor-autor* za skup za učenje

Ako promatramo graf $G$ induciran matricom $A\_train$, tj. graf kojemu su čvorovi autori, a brid postoji između čvorova ako su autori surađivali, tada su neke od statistika za $G$ prikazane na slici 6.3.

Kako bismo dobili skup za testiranje, uzimamo skup svih parova autora koji nisu surađivali u godinama za učenje (4597223 para autora). Iz tog skupa je potrebno razlučiti parove autora koji će biti ili pozitivni ili negativni primjeri za testiranje. Svaki par autora $(i, j)$ koji je ostvario suradnju u godinama za testiranje je pozitivni primjerak (4282 para

| statistike | |
|---|---|
| broj čvorova | 3038 |
| broj bridova | 15980 |
| prosječni stupanj čvora | 0.1901 |
| prosječni koeficijent klasteriranja | 0.672975 |
| broj povezanih komponenti | 21 |

Slika 6.3: Statistike grafa induciranog matricom susjedstva skupa za učenje

autora), dok svaki onaj par koji nije ostvario suradnju je negativni primjerak (4592941 par autora).

## 6.2 Algoritam

Jedini ulazni podatak koji koristimo za algoritam je matrica susjedstva *A_train*. Pomoću nje ćemo prvo procijeniti reducirani rang, a zatim ćemo je koristiti za ekstrakciju dodatnih atributa koji se zasnivaju na topologiji mreže te perturbaciju kojom ćemo simulirati neke iznenadne veze u grafu ili neke veze koje su možda nastale prije promatranog intervala godina. Nakon perturbacije slijedi samo računanje matrice sličnosti pomoću nekoliko varijacija nenegativnih matričnih faktorizacija koje se razlikuju u korištenom algoritmu NMF-a i korištenim dodatnim atributima.

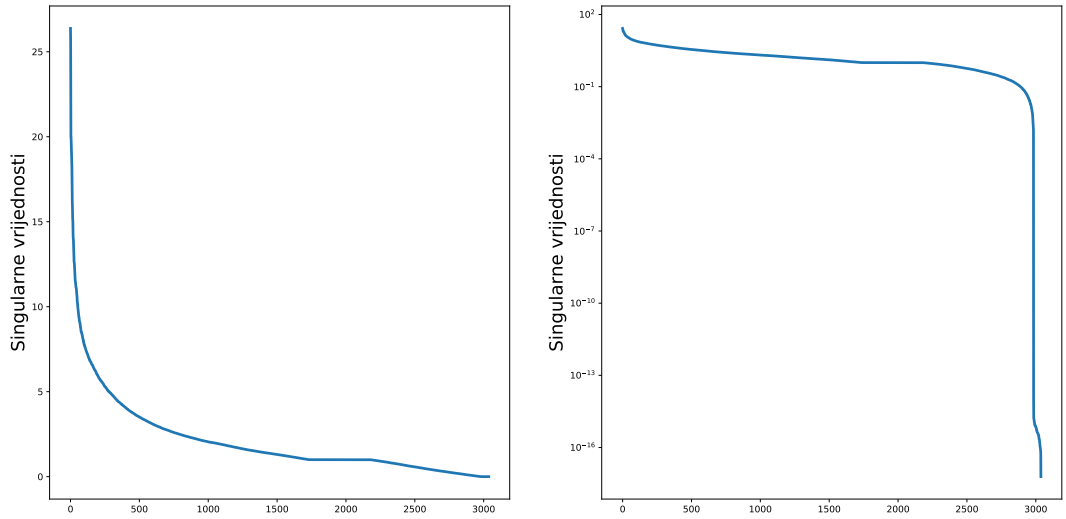
### Procjena reduciranog ranga

S obzirom da je računski najmanje zahtijevno ocijeniti reducirani rang pomoću dekompozicije na singularne vrijednosti, možemo prvo promotriti razdiobu singularnih vrijednosti te pokušati pomoću njihovog zbroja odrediti reducirani rang $r$.

Provođenjem SVD-a nad matricom *A_train* dobijemo dijagonalnu matricu singularnih vrijednosti poredanih u padajućem redoslijedu. Njihova razdioba prikazana je slikom 6.4 na linearnoj i logaritamskoj skali.

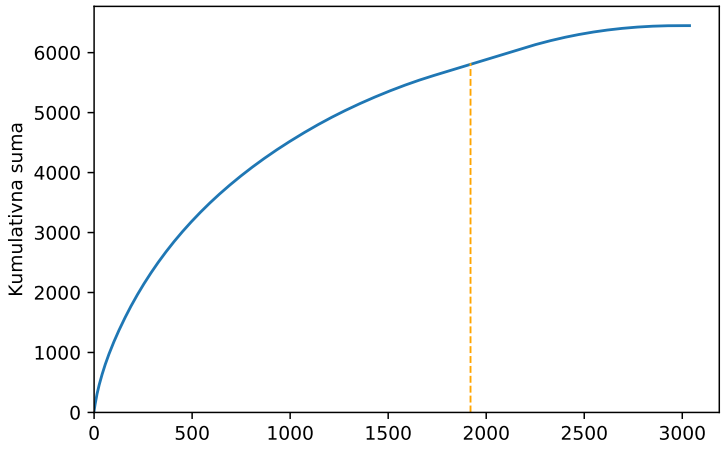
Prema Qiao, ako za reducirani rang $r$ uzmemo broj singularnih vrijednosti čiji zbroj čini 90% ukupnog zbroja singularnih vrijednosti, tada bi $r$ trebao biti dovoljno malen da smanji dimenziju problema, a dovoljno velik da sačuva zadovoljavajuć broj podataka.

U ovom slučaju, reducirani rang bio bi 1920, što ne smanjuje veličinu problema. Tada bi matrice $W$ i $H$ bile dimenzija $3028 \times 1920$ i $1920 \times 3038$ redom, što zajedno daje




Slika 6.4: Singularne vrijednosti u padajućem poretku na linearnoj skali (lijevo) i logaritamskoj skali (desno)

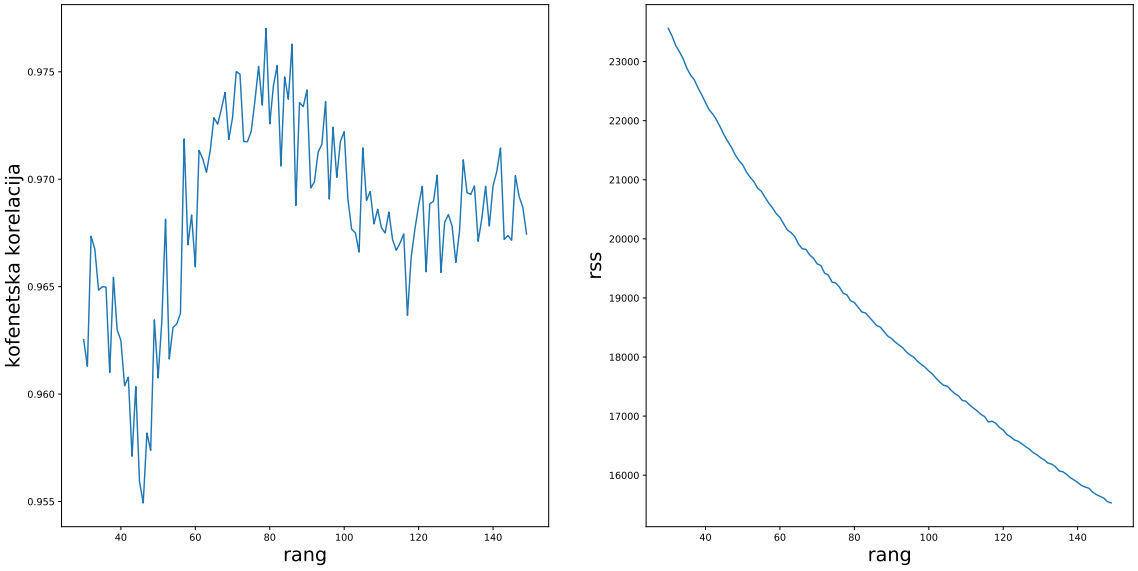
11665920 zapisa, dok originalni problem ima 9229444 zapisa, odnosno imali bismo povećanje od čak 26.34%. Slika 6.5 prikazuje graf za kumulativni zbroj singularnih vrijednosti s istaknutom vrijednosti 1920 za izbor reduciranog ranga prema ovoj metodi.




Slika 6.5: Kumulativni zbroj singularnih vrijednosti

Druga metoda procjene reduciranog ranga je pomoću zbroja kvadrata reziduala i kofenetskog koeficijenta korelacije. Promotrit ćemo navedene mjere za vrijednosti ranga između 30 i 150 s obzirom da želimo smanjiti veličinu problema.

Za svaku vrijednost ranga iz navedenog intervala pokreće se NMF algoritam 30 puta te se računa prosječna rss vrijednost i kofenetski koeficijent korelacije. Slika 6.6 prikazuje dobivene krivulje za kofenetski koeficijent te rss krivulju.

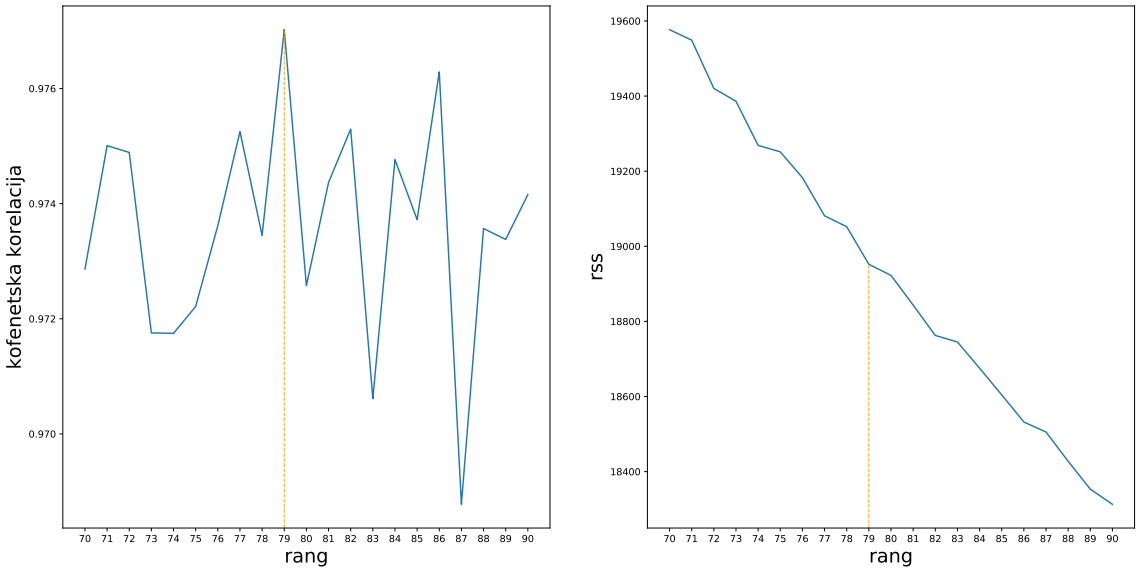



Slika 6.6: Kofenetski koeficijent korelacije (lijevo) i rss (desno) za vrijednosti ranga između 30 i 150

Vidimo da kofenetski koeficijent kreće padati negdje za vrijednosti blizu 80 pa možemo pobliže promotriti krivulje za vrijednosti ranga između 70 i 90. Slika 6.7 prikazuje navedene krivulje s istaknutom vrijednosti za koju je kofenetski koeficijent najviši, tj. $k = 79$ te nakon toga vrijednost kreće padati. S obzirom da rss krivulja nema nekih istaknutih točaka, kao reducirani rank ćemo odabrati vrijednost određenu pomoću kofenetskog koeficijenta.

## Određivanje dodatnih atributa

Neki od češće korištenih dodatnih atributa za predviđanje veze u društvenim mrežama su sličnost ključnih riječi, zbroj susjeda, zbroj radova, duljina najkraćeg puta te mnogi drugi. Ovdje ćemo uzeti u obzir dva topološka atributa, a to su zbroj susjeda i duljina najkraćeg puta. Promotrit ćemo razdiobu navedenih atributa na dva skupa podataka: pozitivnim parovima i negativnim parovima skupa za testiranje.




Slika 6.7: Kofenetski koeficijent korelacije (lijevo) i rss (desno) za vrijednosti ranga između 70 i 90

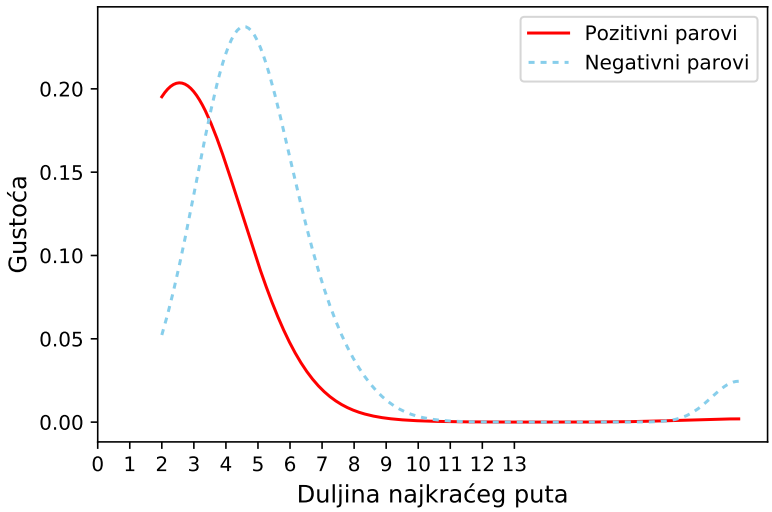
### Duljina najkraćeg puta

Za svaki par autora $(i, j)$ računamo najkraći put između njih, tj. broj bridova u grafu $G$ da bismo došli od čvora $i$ do čvora $j$. Ako su autori surađivali tada je ta duljina 1, ako nisu surađivali, ali imaju zajedničke susjede tada iznosi 2. Ako nemaju zajedničke susjede, ali su njihovi susjedi međusobno surađivali tada je ta duljina 3 te analogno za veće iznose duljine najkraćeg puta.

Polazimo s pretpostavkom da što je kraća udaljenost između dva autora, to je veća vjerojatnost da će oni surađivati. Naime, možemo pretpostaviti da ako su im susjedi surađivali, tada je njihovo područje interesa blisko te će možda i oni međusobno ostvariti koautorstvo.

Slika 6.8 prikazuje razdiobu duljine najkraćeg puta za skup pozitivnih parova i skup negativnih parova. Možemo primijetiti da skup pozitivnih parova ima mnogo veću vjerojatnost da mu duljina najkraćeg puta bude 2 ili 3, nego što je to slučaj kod skupa negativnih parova. To je razumljivo, s obzirom da ako je prevelika duljina najkraćeg puta između dva autora, tada nije ni velika vjerojatnost da će ih susjedi susjeda upoznati.

Kako bismo sada stvorili matricu $A\_shp$ koja prikazuje duljinu najkraćeg puta kao atribut koji možemo koristiti za NMF, trebamo takvu funkciju koja će preslikati duljine najkraćeg puta na sljedeći način: za one autore koji imaju manju međusobnu udaljenost, vrijednost atributa $a\_shp$ će biti najveća; kako duljina puta raste, tako vrijednost atributa




Slika 6.8: Razdioba duljine najkraćeg puta za skup pozitivnih parova i skup negativnih parova

$a\_shp$ pada, s tim da duljine 2 i 3 imaju značajnije vrijednosti od svih sljedećih duljina puta.

Jedna takva funkcija je sljedeća: ako je $shp$ vrijednost najkraćeg puta, a $ashp$ pripadni atribut, tada možemo uzeti $ashp = \frac{1}{(shp-1)^2}$. Slika 6.9 prikazuje navedenu funkciju.

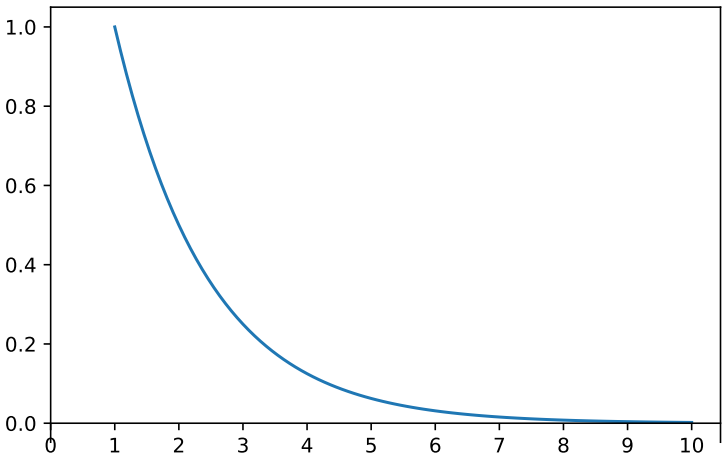
Primijetimo da je vrijednost jednaka 0.5 kada je duljina jednaka 2, odnosno kada su surađivali susjedi autora. To je ujedno i najveća vrijednost u skupu za testiranje.

Sada ćemo promotriti razdiobu za zbroj susjeda te stvoriti pripadnu matricu atributa gdje će također najveća vrijednost u skupu za testiranje iznositi 0.5 tako da zajedno u zbroju s atributom $ashp$ daje maksimalnu vrijednost 1.

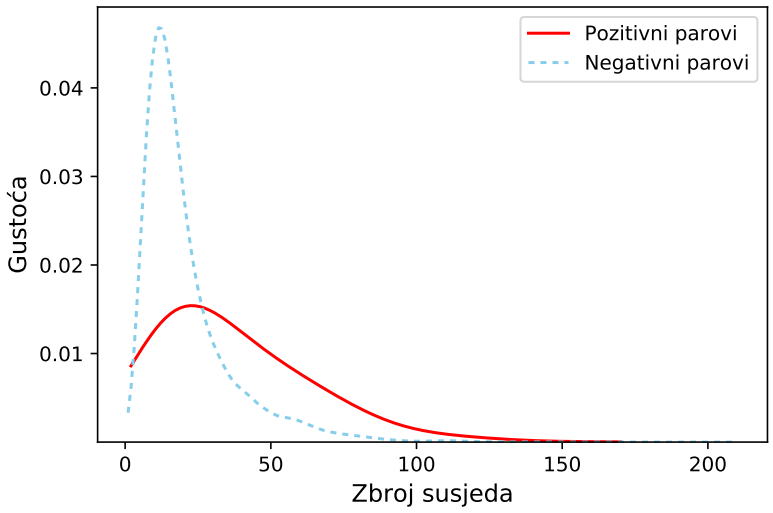
Za dva autora zbroj susjeda računamo tako da zbrojimo broj susjeda svakog od autora uz to da zajedničke susjede računamo samo jednom. Ovdje polazimo od pretpostavke da ako autori surađuju s velikim brojem autora, tada je i veća vjerojatnost da će oni međusobno ostvariti suradnju.

Promotrimo na slici 6.10 razdiobu zbroja susjeda za skup pozitivnih parova i negativnih parova skupa za testiranje. Primijetimo da je za skupu pozitivnih parova veća vjerojatnost za veći zbroj nego u skupu negativnih parova.

Ako je $son$ vrijednost zbroja atributa, a $ason$ pripadni atribut, tada želimo da niske $son$ vrijednosti dobiju zanemarivu vrijednost atributa $ason$, a zatim naglo porastu. Kao granicu $q$ naglog porasta odredit ćemo da zadnjih 20% parova autora s najvećom vrijednosti zbroja susjeda dobije značajnije vrijednosti atributa $ason$.

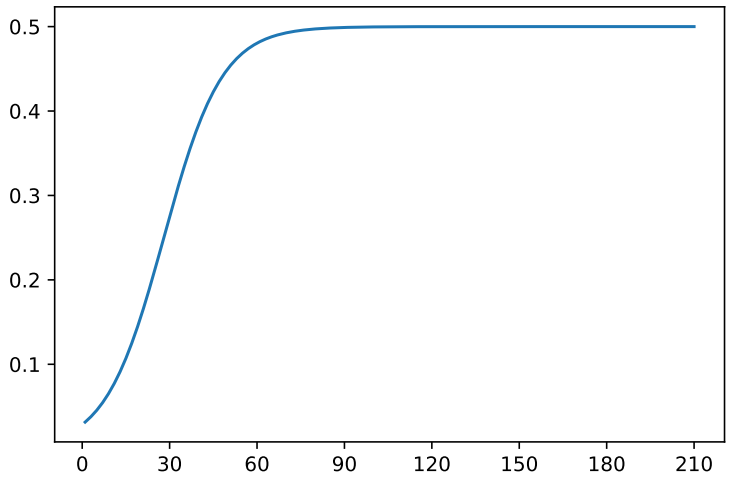



Slika 6.9: Funkcija $ashp = \frac{1}{(shp-1)^2}$ koja preslikava duljine najkraćeg puta u pripadni atribut




Slika 6.10: Razdioba zbroja susjeda za skup pozitivnih parova i skup negativnih parova

Jedna takva funkcija opisana je logističkom regresijom, s obzirom da iz jako niskih vrijednosti naglo poraste u jako visoke. Korištena funkcija $ason = \frac{1}{2}\frac{1}{1+e^{-0.1-0.1(son-q)}}$ prikazana je na slici 6.11. Primijetimo da je ovdje također najviša vrijednost 0.5.




Slika 6.11: Funkcija $ason = \frac{1}{2}\frac{1}{1+e^{-0.1-0.1(son-q)}}$ koja preslikava zbroj susjeda u pripadni atribut

## Perturbacija matrice susjedstva

Kako bismo simulirali ona koautorstva koja nastaju slučajno, a ujedno i ona koja možda nisu u matrici susjedstva $A\_train$ jer su nastala prije prije promatranog intervala godina, poslužit ćemo se perturbacijom matrice susjedstva.

Za par autora $(i, j)$ koji nije surađivao u skupu godina za učenje, tj. $A\_train_{ij} = 0$, simulirat ćemo koautorstvo tako što ćemo postaviti $A\_train_{ij} = 1$. Kao koeficijent perturbiranosti uzet ćemo $\eta = 0.07$, odnosno za 7% parova koji nisu koautori simulirat ćemo koautorstvo. Na ovaj način dobijemo matricu $A\_p$.

Kako bi vjerojatnost da neka veza ne bude perturbirana bila relativno mala, postupak ćemo ponoviti 30 puta te tako dobiti niz perturbiranih matrica susjedstva $A\_p$. Sada je vjerojatnost da neka veza u bilo kojoj od tih 30 matrica ostane neperturbirana $(1-0.07)^{30} = 0.113367$.

## Računanje matrice sličnosti

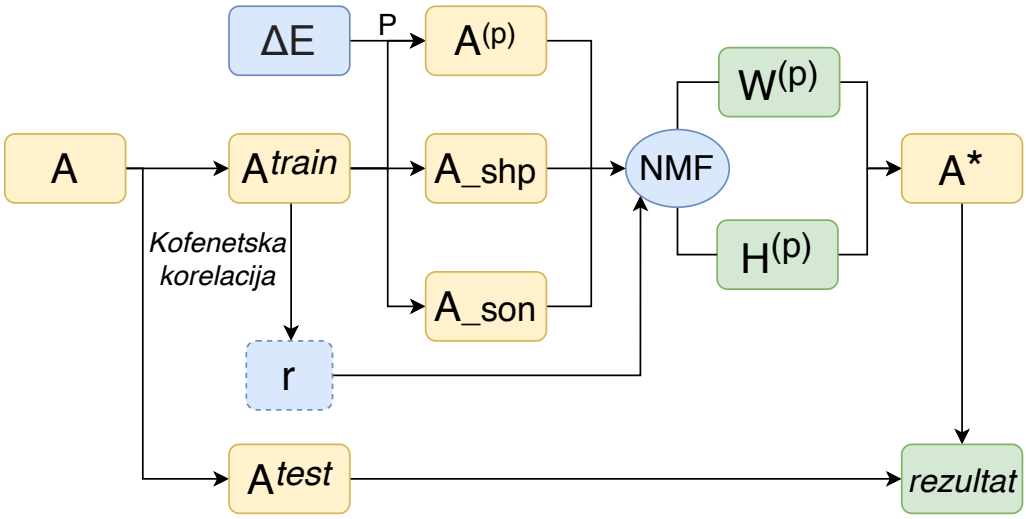
Matricu sličnosti ćemo izračunati koristeći nenegativne matrične faktorizacije uz perturbaciju matrice susjedstva te dodatne atribute. Za računanje NMF-a korištene su implementacije koje se nalaze u python paketima *scikit-learn* i *nimfa*.

Za svaku od korištenih metoda, pripadni NMF algoritam je proveden uz perturbaciju matrice susjedstva, odnosno NMF algoritam je proveden na svakoj od matrica koja nastane kao zbroj perturbiranih matrica $A\_p$ i korištenih dodatnih atributa te je zatim uzeta arit-

metička sredina dobivenih rezultata kako bi se dobila matrica sličnosti. Korišteni algoritmi su sljedeći:

- *MU+FR+SON+SHP* - korištena implementacija je iz *scikit-learn* paketa, funkcija *sklearn.decomposition.non_negative_factorization*; korišten je algoritam Hadamardovog produkta uz Frobeniusovu mjeru te inicijalizacija *nndsvda*, verzijom *nndsvd*-a gdje se nule u dobivenim inicijalnim matricama zamijene prosječnom vrijednosti matrice $X$; korištene su obje matrice atributa $A\_shp$ i $A\_son$
- *MU+KL+SON+SHP* - korištena je implementacija algoritma Hadamardovog produkta iz *nimfa* paketa, funkcija $Nmf$ uz *Kullback-Leibler* funkciju cilja te odgovarajući korak osvježenja baziran na divergenciji; inicijalizacija je pomoću *random_vcol* metode gdje se svaki stupac matrice $W$ inicijalizira prosjekom $p$ nasumično odabranih stupaca matrice $X$, i slično tome, svaki redak matrice $H$ prosjekom $p$ nasumično odabranih redaka; korištene su obje matrice atributa $A\_shp$ i $A\_son$
- *LS+SON+SHP* - korištena implementacija je iz *nimfa* paketa, funkcija $Lsnmf$ koja implementira algoritam alternirajućih nenegativnih najmanjih kvadrata gdje je svaki potproblem riješen koristeći projicirani gradijent, a korak $\alpha$ se mijenja po iteracijama uz Linovo pravilo; kao i kod prošlog algoritma, inicijalizacija je pomoću metode *random_vcol*; korištene su obje matrice atributa $A\_shp$ i $A\_son$
- *CD+SON+SHP* - korištena implementacija je iz *scikit-learn* paketa, funkcija *sklearn.decomposition.non_negative_factorization*; korišten je algoritam koordinatnog spusta koji koristi *Fast HALS*, odnosno ubrzanu verziju hijerarhijskih alternirajućih najmanjih kvadrata; inicijalizacija je pomoću *nndsvd*-a, a funkcija cilja frobeniusova; korištene su obje matrice atributa $A\_shp$ i $A\_son$
- *MU+KL* - korišteni algoritam je jednak kao kod *MU+KL+SON+SHP* , osim što nisu korištene matrice atributa $A\_shp$ i $A\_son$
- *MU+KL+SHP* - jednak algoritam kao *MU+KL* uz korištenje samo matrice atributa $A\_shp$
- *MU+KL+SON* - jednak algoritam kao *MU+KL* uz korištenje samo matrice atributa $A\_son$

Svaki od algoritama pokrenut je 10 puta uz prethodno određeni reducirani rang $r$ koji iznosi 79. Koeficijent perturbacije je 0.07 te broj perturbiranih matrica 30. Opisani algoritmi prikazani su na slici 6.12 gdje je $NMF \in \{MU+FR+SON+SHP, MU+KL+SON+SHP, LS+SON+SHP, CD+SON+SHP, MU+KL, MU+KL+SHP, MU+KL+SON\}$. Uzmimo u obzir da je na slici prikazan okvirni algoritam u kojem su prikazane obje matrice atributa, međutim u nekim algoritmima se ne koriste ili se koristi samo jedna.




Slika 6.12: NMF algoritam uz perturbacije matrice susjedstva i matrice atributa

## 6.3 Evaluacija rezultata

U ovom odjeljku bit će dana usporedba navedenih algoritama zasnovanih na NMF-u s nekim klasičnim metodama baziranim na indeksima sličnosti. Svaki od algoritama pokrenut je 10 puta te je kao konačna mjera uspješnosti uzet prosjek dobivenih mjera za svako pokretanje.

### Klasične metode

Metode za usporedbu su sljedeće: zajednički susjedi (*eng. Commong Neighbors, CN*), Jaccard, Adamic-Adar (AA), alokacija resursa ( *eng. Resource Allocation, RA*) te Salton. Prva tri indeksa opisana su u poglavlju *Problem predviđanja veze u mreži*. Slijede izrazi za izračun posljednja dva indeksa.

$$Similarity_{RA}(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|} \tag{6.1}$$

$$Similarity_{Salton}(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)| \times |\Gamma(v)|}}. \tag{6.2}$$

## Mjere za evaluaciju

Često korištene mjere za evaluaciju kod predviđanja nove veze su preciznost (*eng. precision*) i AUC (*eng. area under the ROC curve*). AUC mjera, osim što predstavlja površinu ispod ROC krivulje, može se interpretirati i kao vjerojatnost da dobiveni model nasumično odabranom pozitivnom primjeru dodijeli viši rang nego nekom nasumično odabranom negativnom primjeru. U slučaju predviđanja nove veze, to bi značilo da ako promatramo skup svih autora koji nisu surađivali u skupu za učenje, AUC računa vjerojatnost da naš model pridijeli veću vjerojatnost nove veze između dva nasumično odabrana autora koji su surađivali u skupu za testiranje, nego nekom nasumično odabranom paru autora koji nije surađivao u skupu za testiranje.

Konkretno, $n$ puta nasumično uzmemo jedan pozitivni par autora $(i_P, j_P)$ i jedan negativni par autora $(i_N, j_N)$ te promotrimo pripadajuće indekse sličnosti $s_P$ i $s_N$ dobivene nekim modelom. Ako je $n'$ broj puta koliko je vrijednost $s_P$ bila veća od $s_N$, a $n''$ broj puta koliko su vrijednosti bile jednake, tada je formula za izračun AUC vrijednosti sljedeća:

$$AUC = \frac{n' + 0.5n''}{n}. \tag{6.3}$$

Za onaj model koji nasumično određuje nove veze, vrijednost AUC trebala bi iznositi 0.5. Dakle, sve one vrijednosti više od 0.5 pokazuju koliko je neki algoritam bolji od nasumičnog nagađanja. U svim evaluacijama rezultata usporedba parova izvršena je 10000 puta.

S obzirom da nam često kod predviđanja novih veza nije bitno koliko dobro algoritam predviđa općenito sve veze, već nam je samo bitno da nekada odredimo manji broj mogućih veza, druga mjera koja se koristi je preciznost.

Mjera za preciznost je intuitivna, a računa se tako da poredamo indekse sličnosti pridijeljene algoritmom u padajućem redoslijedu te za prvih $top\_L$ parova vidimo koliko ih je istinito pozitivnih. Tada je preciznost omjer točno predviđenih $l$ veza u odnosu na $top\_L$ veza koje smo uzeli u razmatranje, odnosno:

$$Precision = \frac{l}{top\_L}. \tag{6.4}$$

Primijetimo da za drugačiji odabir prvih $top\_L$ parova vrijednosti za preciznost mogu biti drugačije (gotovo uvijek i jesu). Zbog toga je za svaku od metoda izračunata preciznost za $top\_L$ od 100 do 1000 s korakom od 100 te uzeta prosječna preciznost dobivenih vrijednosti.

S obzirom da je svaka od metoda pokrenuta 10 puta, AUC i preciznost su izračunati za svaki od dobivenih rezultata i zatim je uzeta prosječna vrijednost.

## Rezultati

Na slici 6.13 vidimo vrijednosti za preciznost i AUC dobivene za navedene metode zasnovane na NMF-u te za klasične metode. Za svaku od ove dvije kategorije podebljani su najbolji rezultati za svaku od mjera.

| metoda | preciznost - top_L | | | | | | | | | | prosječna preciznost | AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | | |
| mu+fr+son+shp | 0.1140 | 0.1195 | 0.1217 | 0.1203 | 0.1168 | 0.1162 | 0.1141 | 0.1113 | 0.1078 | 0.1047 | 0.1146 | 0.8641 |
| mu+kl+son+shp | **0.1600** | **0.1455** | **0.1453** | **0.1388** | **0.1338** | **0.1278** | **0.1243** | **0.1195** | **0.1148** | **0.1126** | **0.1322** | 0.8695 |
| ls+son+shp | 0.1320 | 0.1115 | 0.0950 | 0.0933 | 0.0940 | 0.0945 | 0.0926 | 0.0916 | 0.0902 | 0.0881 | 0.0983 | 0.8724 |
| cd+son+shp | 0.1190 | 0.0845 | 0.0863 | 0.0880 | 0.0910 | 0.0923 | 0.0924 | 0.0920 | 0.0903 | 0.0885 | 0.0924 | 0.8734 |
| mu+kl | 0.1130 | 0.1125 | 0.1050 | 0.1035 | 0.1044 | 0.1008 | 0.0971 | 0.0936 | 0.0912 | 0.0890 | 0.1010 | 0.7075 |
| mu+kl+shp | 0.0760 | 0.1015 | 0.0980 | 0.0948 | 0.0894 | 0.0892 | 0.0900 | 0.0866 | 0.0838 | 0.0843 | 0.0894 | **0.9039** |
| mu+kl+son | 0.0300 | 0.0300 | 0.0410 | 0.0365 | 0.0328 | 0.0303 | 0.0301 | 0.0289 | 0.0286 | 0.0276 | 0.0316 | 0.7169 |
| cn | **0.1000** | **0.1100** | **0.1000** | 0.1000 | 0.1080 | 0.0950 | **0.0929** | **0.0875** | 0.8333 | 0.0810 | **0.0958** | **0.7636** |
| jaccard | 0.0400 | 0.0550 | 0.0500 | 0.0550 | 0.0520 | 0.0533 | 0.0457 | 0.0400 | 0.0411 | 0.0450 | 0.0477 | 0.7611 |
| salton | 0.0400 | 0.0550 | 0.0567 | 0.0500 | 0.0460 | 0.0433 | 0.0457 | 0.0413 | 0.0444 | 0.0450 | 0.0467 | 0.7598 |
| ra | 0.0300 | 0.0750 | 0.0633 | 0.0750 | 0.0680 | 0.0667 | 0.0700 | **0.0875** | **0.0889** | **0.0860** | 0.0710 | 0.7614 |
| aa | 0.0600 | 0.0900 | 0.0900 | **0.1075** | **0.1100** | **0.1000** | 0.0900 | **0.0875** | 0.0833 | 0.0810 | 0.0899 | 0.7625 |
| random | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.5000 |

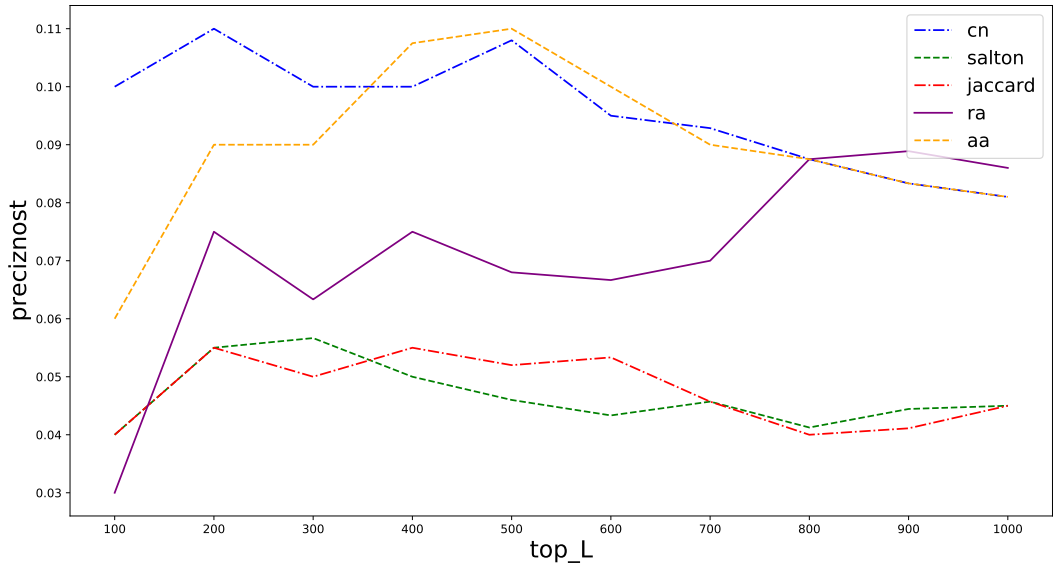
Slika 6.13: Preciznost i AUC za NMF algoritme i klasične metode

Možemo primijetiti da uglavnom kod svih metoda pada preciznost povećanjem $top\_L$ varijable. Također, što se tiče preciznosti, metoda *MU+KL+SON+SHP* ima najbolje rezultate u svim segmentima. Druga najbolja metoda po preciznosti je *MU+FR+SON+SHP*. Zanimljivo je uočiti da su metode *MU+KL* uz *SON* i *SHP* atribute pojedinačno, značajno lošije od metode koja kombinira oba atributa istovremeno.

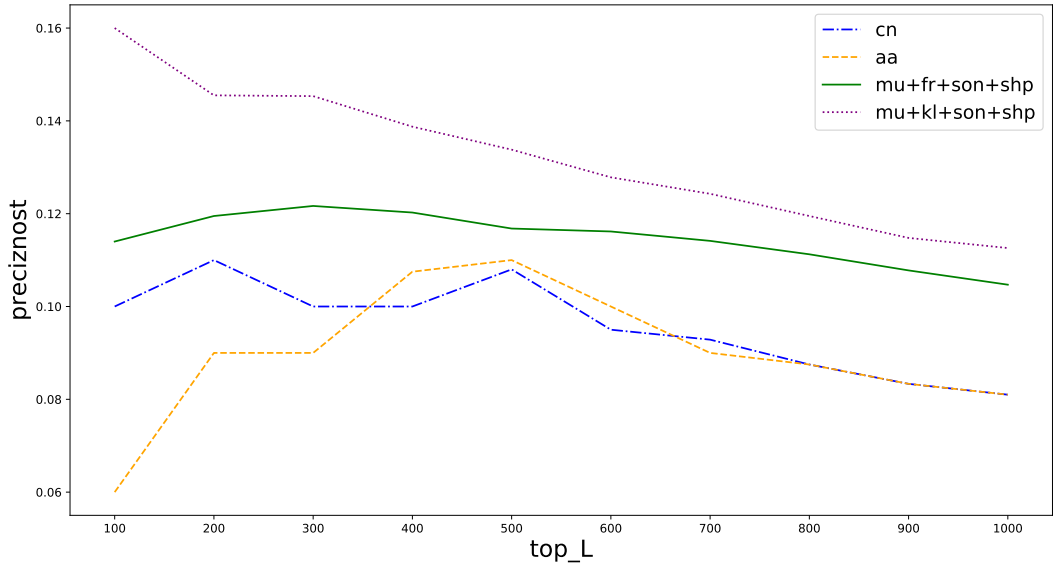
Što se tiče AUC mjere, tu je najbolja *MU+KL+SHP* metoda, što bi dalo sugerirati da bi za neke veće izbore varijable $top\_L$ ta metoda počela davati bolje postotke preciznosti u odnosu na druge metode. Također možemo primijetiti da sve klasične metode imaju podjednaku AUC vrijednost. Nadalje, NMF metode, osim *MU+KL+SON* i *MU+KL* imaju značajno više AUC vrijednosti od klasičnih metoda.

Slika 6.14 prikazuje kako se mijenja preciznost pri promjeni $top\_L$ za klasične metode. Vidimo da se najboljom pokazala *CN*, a zatim *AA*, dok je *RA* podbacila za niže vrijednosti $top\_L$ te pri većim vrijednostima sustigla prve dvije metode.

Na slici 6.15 prikazana je usporedba preciznosti za po dvije najbolje metode iz oba skupa NMF i klasičnih metoda: *MU+KL+SHP+SON*, *MU+FR+SHP+SON*, *CN* i *AA*. Vidimo da su preciznosti NMF metoda veće od preciznosti klasičnih metoda za sve vrijednosti $top\_L$.




Slika 6.14: Graf preciznosti klasičnih metoda za različite vrijednosti *top_L*




Slika 6.15: Usporedba preciznosti pri porastu *top_L* za po dvije najbolje metode iz skupa NMF algoritama i klasičnih metoda

Vidjeli smo da se broj zajedničkih susjeda pokazao kao najbolji među klasičnim indeksima, stoga ga možemo probati kombinirati s metodama zasnovanim na nenegativnim matričnim faktorizacijama te vidjeti da li poboljšava rezultate. Kao NMF metodu na kojoj ćemo provesti analizu uzet ćemo *MU+FR+SHP+SON* s obzirom da je dala najbolje rezultate odmah iza *MU+KL+SHP+SON*, a izvodi se mnogo brže. Osim dodavanja dodatnog atributa, proučit ćemo i utjecaj promjene reduciranog ranga te perturbacijskih parametara.

Možemo primijetiti da su duljina najkraćeg puta i broj zajedničkih susjeda povezani. Ukoliko zajednički susjedi postoje, a autori nisu surađivali, tada duljina najkraćeg puta iznosi 2. Dakle, broj zajedničkih susjeda možemo iskoristiti za distinkciju onih autora koji imaju jednak, odnosno maksimalni iznos *shp* atributa. Za te autore vrijednost $ashp$ iznosi 0.5, a sljedeća dodijeljena vrijednost iznosi 0.25 za duljinu puta 3, pa ćemo broj zajedničkih susjeda preslikati u vrijednosti između 0 i 0.25 kako bismo napravili distinkciju autora, ali opet dali tom atributu manju važnost nego ostalim dvama atributima.

S obzirom da mnogo parova autora ima manje od 5 zajedničkih susjeda *cn*, svima njima ćemo pridijeliti vrijednost atributa *acn* u iznosu 0. Također ćemo i ograničiti broj zajedničkih susjeda kako bi preslikane vrijednosti bile što bliže jedne drugima. Na kraju ćemo skalirati dobivene iznose kako bi na kraju bili u intervalu $[0, 0.25]$. Konačni izraz kojim broj autora *cn* preslikavamo u atribut *acn* je sljedeći:

$$acn = \frac{1}{4}\frac{\min(\max(cn - 5, 0), 20)}{20} \tag{6.5}$$

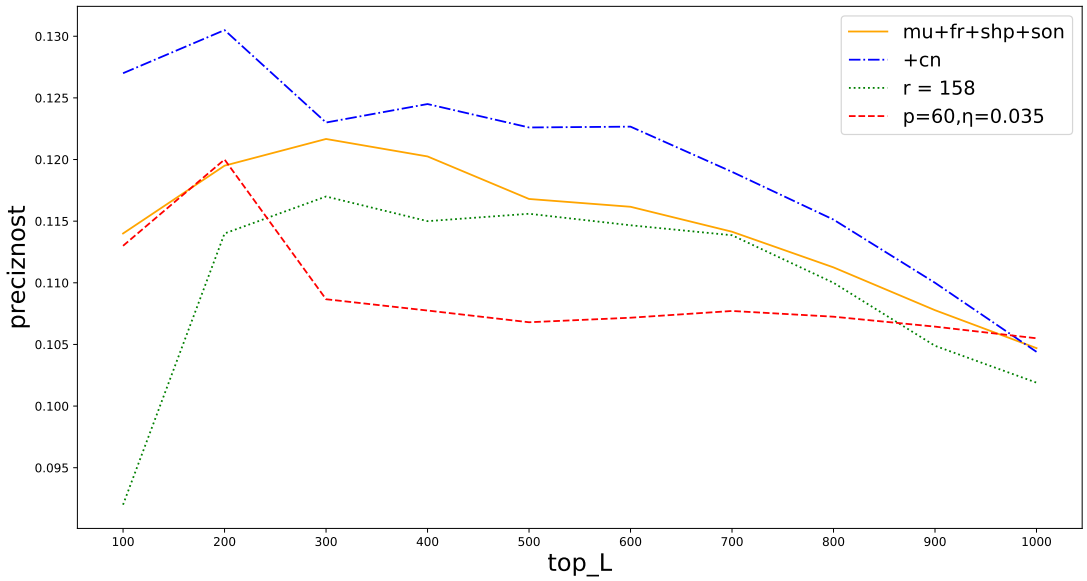
U daljnjoj analizi možemo vidjeti kako promjena reduciranog ranga utječe na rezultat. S obzirom da se porastom reduciranog ranga smanjuje greška u odnosu na polaznu matricu, možda bi se dalo pretpostaviti da bi rezultat trebao biti bolji. Međutim, prema kofenetskom koeficijentu, za veći rang prepoznavanje klasa pomoću NMF-a trebalo bi biti nestabilnije, stoga možemo očekivati lošiji rezultat. Kao drugi rang za koji ćemo analizirati rezultate dobivene metodom *MU+FR+SHP+SON* dvostruko je veći od korištenog ranga (79) te iznosi 158.

Posljednju promjenu koju ćemo napraviti je mijenjanje perturbacijskih parametara. Kod algoritama u tablici 6.13 korišteni su $p = 30$ za broj puta koliko je matrica $A\_train$ perturbirana, te $\eta = 0.07$ kao koeficijent perturbacije, odnosno udio elemenata koje perturbiramo u polaznoj matrici susjedstva. Sada ćemo pokušati "profiniti" perturbiranost, tj. uzeti manji postotak veza koje ćemo mijenjati, te tako pokušati suptilnije utjecati na rezultat NMF metode. Za $\eta$ ćemo uzeti dvostruko manji iznos, odnosno 0.035, zbog čega moramo povećati $p$ kako bi vjerojatnost da neka veza ostane neperturbirana ostala i dalje mala, pa ćemo uzeti dvostruko veći broj za $p$ , odnosno 60, pa sada ta vjerojatnost iznosi $(1 - 0.035)^{60} = 0.1179$.

Rezultati za navedene promjene prikazani su na slici 6.16 gdje su podebljani najbolji rezultati za svaki od $top\_L$ te AUC dok je na slici 6.17 pripadajući graf za preciznosti.

| metoda | preciznost - top_L | | | | | | | | | | prosječna preciznost | AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | | |
| mu+fr+son+shp | 0.1140 | 0.1195 | 0.1217 | 0.1203 | 0.1168 | 0.1162 | 0.1141 | 0.1113 | 0.1078 | 0.1047 | 0.1146 | 0.8641 |
| cn | **0.1270** | **0.1305** | **0.1230** | **0.1245** | **0.1226** | **0.1227** | **0.1190** | **0.1151** | **0.1100** | 0.1044 | **0.1199** | 0.8604 |
| r = 158 | 0.0920 | 0.1140 | 0.1170 | 0.1150 | 0.1156 | 0.1147 | 0.1139 | 0.1100 | 0.1049 | 0.1019 | 0.1099 | 0.8616 |
| p=60, η=0.035 | 0.1130 | 0.1200 | 0.1087 | 0.1078 | 0.1068 | 0.1072 | 0.1077 | 0.1073 | 0.1065 | **0.1055** | 0.1090 | **0.8679** |

Slika 6.16: Rezultati pri promjenama parametara te dodatka matrice atributa




Slika 6.17: Usporedba preciznosti pri promjenama parametara te dodatka matrice atributa

Možemo primijetiti da su rezultati što se tiče preciznosti najbolji kod dodatka matrice atributa *CN*, a zatim kod izvornog algoritma *MU+FR+SHP+SON*. Nadalje, povećanjem reduciranog ranga *r* rezultati su samo lošiji, stoga vidimo da je početni izbor reduciranog ranga pomoću kofenetske korelacije bio opravdan. Ipak, porastom varijable *top_L*, vrijednosti preciznosti za sve metoda se približavaju jedna drugoj što možda objašnjava slične vrijednosti za AUC.

Iako ne značajno, AUC vrijednost je najviša kod profinjene perturbiranosti. Preciznost je, međutim, najniža, no pri većim izborima za *top_L* prestigla je ostale vrijednosti, te bi stoga za više vrijednosti *top_L* možda ipak imala bolju prosječnu vrijednost za preciznost u odnosu na ostale metode.

Iz ove analize možemo zaključiti da nenegativne matrične faktorizacije implementirane algoritmom Hadamardovog produkta uz adekvatne atribute polučuju bolje rezultate od klasičnih metoda. Također, viši reducirani rang ne znači bolji rezultat, već je broj klasa koje algoritam može prepoznati ključan u odabiru. Možda bi se dalje rezultati mogli poboljšati dodavanjem još nekih atributa koji bi bili specifični za prirodu neke mreže, odnosno uzimajući u obzir što mreža predstavlja te u kojim slučajevima su veze među entitetima najvjerojatnije. Ovakav pristup svakako ostavlja mnogo prostora za nadogradnju, bilo eksperimentiranjem s dodatnim atributima ili drugačijim implementacijama samog NMF algoritma.

# Bibliografija


[1] L. A. Adamic i E. Adar, *Friends and neighbors on the Web*, Social Networks **25** (2003), br. 3, 211–230, ISSN 03788733, https://linkinghub.elsevier.com/retrieve/pii/S0378873303000091.

[2] R. Aihara, T. Takiguchi i Y. Ariki, *Individuality-Preserving Voice Conversion for Articulation Disorders Using Locality-Constrained NMF*, Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies (Grenoble, France), Association for Computational Linguistics, kolovoz 2013, str. 3–8, https://www.aclweb.org/anthology/W13-3902.

[3] G. Berlusconi, F. Calderoni, N. Parolini, M. Verani i C. Piccardi, *Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis*, PLOS ONE **11** (2016), br. 4, 1–21, https://doi.org/10.1371/journal.pone.0154244.

[4] M. W. Berry, M. Browne, A. N. Langville, V.P Pauca i R. J. Plemmons, *Algorithms and Applications for Approximate Nonnegative Matrix Factorization*, Teh. izv.

[5] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Transactions on Automatic Control **21** (1976), br. 2, 174–184.

[6] D. Bertsekas, *Nonlinear Programming*, SIAM AMS Proc (Belmont Massachusetts), sv. 9, Athena Scientific, 1999, ISBN 9781886529007.

[7] J. M. Bioucas-Dias i J. M. P. Nascimento, *Estimation of signal subspace on hyperspectral data*, Image and Signal Processing for Remote Sensing XI (Lorenzo Bruzzone, ur.), sv. 5982, International Society for Optics and Photonics, SPIE, 2005, str. 191 – 198, https://doi.org/10.1117/12.620061.

[8] M. R. Blanton i S. Roweis, *K-Corrections and Filter Transformations in the Ultraviolet, Optical, and Near-Infrared*, The Astronomical Journal **133** (2007), br. 2, 734–754, https://doi.org/10.1086%2F510127.

[9] C. Boutsidis i E. Gallopoulos, *SVD based initialization: A head start for nonnegative matrix factorization*, Pattern Recognition **41** (2008), br. 4, 1350–1362, ISSN 00313203.

[10] R. Bro i S. De Jong, *A fast non-negativity-constrained least squares algorithm*, Journal of Chemometrics **11** (1997), br. 5, 393–401, ISSN 08869383.

[11] J. P. Brunet, P. Tamayo, T. R. Golub i J. P. Mesirov, *Metagenes and molecular pattern discovery using matrix factorization*, Proceedings of the National Academy of Sciences of the United States of America **101** (2004), br. 12, 4164–4169, ISSN 00278424, www.pnas.orgcgidoi10.1073pnas.0308531101.

[12] C. V. Cannistraci, G. Alanis-Lobato i T. Ravasi, *From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks*, Scientific Reports **3** (2013), ISSN 20452322, https://pubmed.ncbi.nlm.nih.gov/23563395/.

[13] B. Chen, F. Li, S. Chen, R. Hu i L. Chen, *Link prediction based on non-negative matrix factorization*, PLoS ONE **12** (2017), br. 8, e0182968, ISSN 19326203, https://doi.org/10.1371/journal.pone.0182968.

[14] Z. Chen i A. Cichocki, *Nonnegative Matrix Factorization with Temporal Smoothness and / or Spatial Decorrelation Constraints*, Signal Processing (2005).

[15] Z. Chen, A. Cichocki i T. M. Rutkowski, *Constrained non-negative matrix factorization method for EEG analysis in early detection of alzheimer disease*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, sv. 5, 2006, str. 893—-896, ISBN 142440469X, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.78.8613.

[16] A. Cichocki, R. Zdunek, A. H. Phan i S. Amari, *Nonnegative Matrix and Tensor Factorizations*, 2009.

[17] M. E. Daube-Witherspoon i G. Muehllehner, *An Iterative Image Space Reconstruction Algorithm Suitable for Volume ECT*, IEEE Transactions on Medical Imaging **5** (1986), br. 2, 61–66, ISSN 1558254X, https://pubmed.ncbi.nlm.nih.gov/18243988/.

[18] C. Ding, T. Li, W. Peng i H. Park, *Orthogonal nonnegative matrix tri-factorizations for clustering*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, sv. 2006, 2006, str. 126–135, ISBN 1595933395.

[19] L. Dong, Y. Li, H. Yin, H. Le i M. Rui, *The algorithm of link prediction on social network*, Mathematical Problems in Engineering **2013** (2013), ISSN 1024123X, http://dx.

[20] C. Eckart i G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika **1** (1936), br. 3, 211–218, ISSN 00333123, https://link.springer.com/article/10.1007/BF02288367.

[21] N. B. Erichson, A. Mendible, S. Wihlborn i J. N. Kutz, *Randomized nonnegative matrix factorization*, Pattern Recognition Letters **104** (2018), 1–7, ISSN 01678655, https://github.com/erichson/ristretto.

[22] C. Févotte, N. Bertin i J. L. Durrieu, *Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis*, mar 2009, ISSN 08997667, str. 793–830.

[23] C. Févotte, E. Vincent i A. Ozerov, *Single-channel audio source separation with NMF: Divergences, constraints and algorithms*, Signals and Communication Technology, 2018, str. 1–24, https://hal.inria.fr/hal-01631185.

[24] V. Franc, V. Hlaváč i M. Navara, *Sequential coordinate-wise algorithm for the nonnegative least squares problem*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), sv. 3691 LNCS, Springer, Berlin, Heidelberg, 2005, str. 407–414, ISBN 3540289690, http://link.springer.com/10.1007/11556121{_}50.

[25] F. Gao, K. Musial, C. Cooper i S. Tsoka, *Link prediction methods and their accuracy for different social networks and network metrics*, Scientific Programming **2015** (2015), ISSN 10589244, http://dx.doi.org/10.1155/2015/172879.

[26] N. Gillis, *Nonnegative Matrix Factorization Complexity , Algorithms and Applications*, PhD Thesis (2011), br. February.

[27] ———, *The Why and How of Nonnegative Matrix Factorization*, (2014), 1–25, http://arxiv.org/abs/1401.5226.

[28] N. Gillis i F. Glineur, *Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization*, jul 2012, ISSN 08997667, http://arxiv.org/abs/1107.5194http://dx.doi.org/10.1162/NECO{_}a{_}00256, str. 1085–1105.

[29] D. Guillamet i J. Vitrià, *Non-negative matrix factorization for face recognition*, Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), sv. 2504, 2002, str. 336–344, ISBN 3540000119.

[30] J. Han, L. Han, M. Neumann i U. Prasad, *On the rate of convergence of the image space reconstruction algorithm*, Operators and Matrices **3** (2009), br. 1, 41–58, ISSN 18463886.

[31] N. Ho, *Nonnegative Matrix Factorization Algorithms and Applications*, Disertacija, Université Catholique de Louvain, 2008, ISBN 9781450334358.

[32] P.O. Hoyer, *Non-negative matrix factorization with sparseness constraints*, Journal of Machine Learning Research **5** (2004), 1457–1469, ISSN 15337928, http://cbcl.mit.edu/cbcl/software-datasets/FaceData2.html.

[33] L. N. Hutchins, S. M. Murphy, P. Singh i J. H. Graber, *Position-dependent motif characterization using non-negative matrix factorization*, Bioinformatics **24** (2008), br. 23, 2684–2690, ISSN 13674803, https://pubmed.ncbi.nlm.nih.gov/18852176/.

[34] D. D. Lee i H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature **401** (1999), br. 6755, 788–791, ISSN 00280836, https://www.nature.com/articles/44565.

[35] C. J. Lin, *On the convergence of multiplicative update algorithms for nonnegative matrix factorization*, IEEE Transactions on Neural Networks **18** (2007), br. 6, 1589–1596, ISSN 10459227.

[36] R. F. Ling, C. L. Lawson i R. J. Hanson, *Solving Least Squares Problems.*, Journal of the American Statistical Association **72** (1977), br. 360, 930, ISSN 01621459.

[37] T. Liu, M. Gong i D. Tao, *Large-Cone Nonnegative Matrix Factorization*, IEEE Transactions on Neural Networks and Learning Systems **28** (2017), br. 9, 2129–2142, ISSN 21622388, http://www.ieee.org/publications{_}standards/publications/rights/index.html.

[38] N. Ljubešić i I. Pandžić, *Stemmer for Croatian — Natural Language Processing group*, http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/, posjećena 2020-09-24.

[39] G. Louppe, *Collaborative filtering: Scalable approaches using restricted Boltzmann machines*, (2010), https://www.researchgate.net/publication/264046872http://orbi.ulg.ac.be/handle/2268/74400.

[40] B. Marr, *How Much Data Is There in the World?*, 2019, https://www.bernardmarr.com/default.asp?contentID=1846, posjećena 2020-09-02.

[41] V. Metpally, *GitHub - metpallyv/MovieRecommendation*, https://github.com/metpallyv/MovieRecommendation, posjećena 2020-09-23.

[42] A. B. Owen i P. O. Perry, *Bi-cross-validation of the SVD and the nonnegative matrix factorization*, Annals of Applied Statistics **3** (2009), br. 2, 564–594, ISSN 19326157.

[43] A. Ozerov, *Contributions in Audio Modeling for Solving Inverse Problems: Source Separation, Compression and Inpainting*, Habilitation à diriger des recherches, Université Rennes 1, studeni 2019, https://hal.archives-ouvertes.fr/tel-02370669.

[44] P. Paatero i U. Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics **5** (1994), br. 2, 111–126, ISSN 1099095X.

[45] V. P. Pauca, J. Piper i T. J. Plemmons, *Nonnegative matrix factorization for spectral data analysis*, Linear Algebra and Its Applications **416** (2006), br. 1, 29–47, ISSN 00243795.

[46] H. Qiao, *New SVD based initialization strategy for non-negative matrix factorization*, Pattern Recognition Letters **63** (2015), 71–77, ISSN 01678655, http://arxiv.org/abs/1410.2786.

[47] T. Sadowski i R. Zdunek, *Image completion with smooth nonnegative matrix factorization*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), sv. 10842 LNAI, Springer Verlag, 2018, str. 62–72, ISBN 9783319912615.

[48] R. Salgado, *Topic Modeling Articles with NMF. Extracting topics is a good... — Towards Data Science*, https://towardsdatascience.com/topic-modeling-articles-with-nmf-8c6b2a227a45, posjećena 2020-09-24.

[49] T. Virtanen, *Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria*, IEEE Transactions on Audio, Speech and Language Processing **15** (2007), br. 3, 1066–1074, ISSN 15587916.

[50] W. Wang, *Non-Negative Matrix Factorization Based on Projected Nonlinear Conjugate Gradient Algorithm*, ICA Research Network International Workshop (2008), br. 1, 2–5.

[51] W. Wang, F. Cai, Pengfei J. i L. Pan, *A perturbation-based framework for link prediction via non-negative matrix factorization*, Scientific Reports **6** (2016), br. November, 1–11, ISSN 20452322, http://dx.doi.org/10.1038/srep38938.

[52] W. Wang, M. Tang i P. Jiao, *A unified framework for link prediction based on non-negative matrix factorization with coupling multivariate information*, PLoS ONE **13** (2018), br. 11, 1–22, ISSN 19326203.

[53] K. W. Wilson, B. Raj i P. Smaragdis, *Regularized non-negative matrix factorization with temporal dependencies for speech denoising*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, str. 411–414.

[54] H. Zou, T. Hastie i R. Tibshirani, *Sparse Principal Component Analysis*.

# Sažetak

Nenegativne matrične faktorizacije vrsta su linearne redukcije dimenzionalnosti gdje je glavni cilj aproksimirati nenegativnu matricu umnoškom dviju nenegativnih matrica manjih dimenzija od početne. Ovakvom transformacijom podataka pronalaze se latentne značajke te se čuva sama priroda podataka čime se olakšava interpretabilnost. Najčešće korištena funkcija troška zasniva se na Frobeniusovoj mjeri, dok je Kullback-Leibler divergencija pokazala dobre rezultate za rijetko popunjene matrice. S obzirom da rješenje problema nije jedinstveno, a kako bi se smanjila greška aproksimacije, potrebno je procijeniti reducirani rang, odnosno nižu dimenziju u koju preslikavamo početne podatke. Pri tome je bitno voditi računa o broju klasa koje algoritam prepoznaje. Za određivanje reduciranog ranga, a dalje i samih faktora, tradicionalno se koriste algoritmi alternirajućih najmanjih kvadrata te najčešće algoritam Hadamardovog produkta zbog svoje jednostavnosti. Zahvaljujući svojoj svestranoj primjeni kod modeliranja tema, separacije izvora zvuka, klasteriranja te vremenske segmentacije, nenegativne matrične faktorizacije našle su svoj put u mnoga područja gdje su podaci nenegativni, kao što je bioinformatika, astronomija, glazba, tekstualna analiza te mnoga druga. Jedna od novijih primjena je kod predviđanja nove veze u mreži, gdje se, uz perturbacije ili dodatne matrice atributa, uspješno mogu predvidjeti nova prijateljstva, koautorstva ili pak neuronske veze. Ovdje je pokazano na mreži koautorstava CROSBI da se nenegativnim matričnim faktorizacijama u kombinaciji s perturbacijama te matricama atributa dobivenih iz topologije mreže, kao što je duljina najkraćeg puta te zbroj susjeda, mogu dobiti znatno bolji rezultati od onih koristeći klasične metode.

# Summary

Non-negative matrix factorization belongs to the group of linear dimensionality reduction methods and its main goal is to approximate non-negative matrix with the product of two low-rank non-negative matrices. This kind of transformation identifies latent features preserving non-negative structure of the original data which leads to easier interpretability. The most widely used cost function is based on Frobenius norm, while Kullback-Leibler divergence has shown to be effective for sparseness. Taking into account that the solution to this problem is not unique, and in order to decrease approximation error, it is essential to estimate reduced rank, i.e. lower dimension into which the original data is being transformed. One of the key factors here is the number of classes recognized by the algorithm. Both reduced rank estimation and approximation factors are typically obtained using algorithms based on alternating least squares, and, more often, multiplicative update thanks to its simplicity. Due to its versatile applications such as topic modeling, audio source separation, clustering and temporal segmentation, non-negative matrix factorization found its way into various fields characterized by non-negative data, such as bioinformatics, astronomy, music, textual analysis, etc. One of the recent applications is regarding link prediction in networks, where new friendships, coauthorships and even neural connections can be successfully obtained using perturbations or attribute matrices. Using the coauthorship network CROSBI as an example, in this work it was shown that non-negative matrix factorization in combination with perturbations and attribute matrices based on the network topology, such as shortest path distance and sum of neighbors, outperforms results obtained by classical link prediction methods.

# Životopis

Rođena sam 20. siječnja 1995. godine u Splitu. Osnovnu školu "Josip Pupačić" pohađala sam u Omišu od 2001. do 2009. nakon čega sam upisala Prirodoslovno-matematičku gimnaziju u Splitu. Svoje srednjoškolsko školovanje završavam 2013. odličnim uspjehom kako u školi, tako i na maturi, gdje sam između ostaloga imala 100%-tnu riješenost državne mature iz matematike.

Iste godine upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu, a isti završavam 2016. postavši sveučilišna prvostupnica matematike. Godinu nakon upisujem diplomski studij Računarstvo i matematika na istom fakultetu. U lipnju 2018. pohađam Ruby on Rails tečaj Infinuma, gdje i ostajem raditi od rujna do prosinca iste godine na poziciji backend engineer.