

Nonnegative Matrix Factorization Model and its Applications

-Junior Math Seminar-

Taehyeong Kim

Mathematics, Pusan National University

October 5, 2021

Content

■ Introduction

- Who am I?
 - Nonnegative Matrix Factorization

■ Error measures

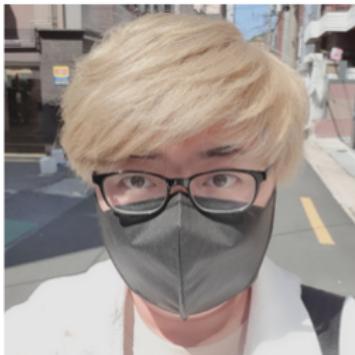
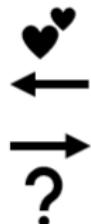
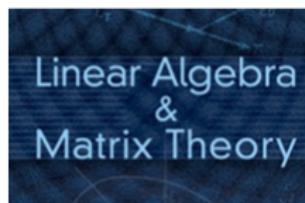
- Statistical model and maximum likelihood
 - β -divergence
 - Choice of the error measure

Applications of NMF Models

- Nonnegative Matrix Factorization
 - Feature extraction in a set of images
 - Blind hyperspectral unmixing
 - Audio source separation
 - Symmetric Nonnegative Matrix Factorization
 - Symmetric Nonnegative Matrix Trifactorization
 - Text mining: topic recovery and document classification
 - Topic modeling
 - Probabilistic Latent Semantic Analysis and Indexing

Conclusion

Who am I?



I am interested in...

- Numerical linear algebra
 - Matrix analysis
 - Applied mathematics

Nonnegative Matrix Factorization



“All models are wrong, but some are useful” - George E. P. Box

Nonnegative Matrix Factorization

There are many matrix factorizations(decompositions).

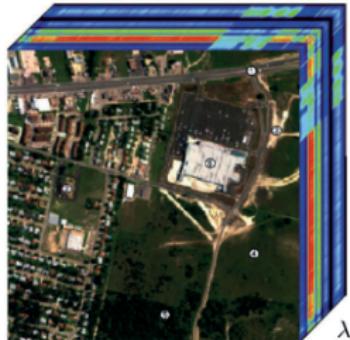
- LU decomposition : $A = LU$
 - Cholesky decomposition : $A = UU^*$
 - QR decomposition : $A = QR$
 - Eigenvalue decomposition : $A = VDV^{-1}$
 - Singular value decomposition : $A = U\Sigma V^*$
 - CUR decomposition : $A = CUR$

Each of the factorizations is used in several applications, such as solving matrix equations, clustering.

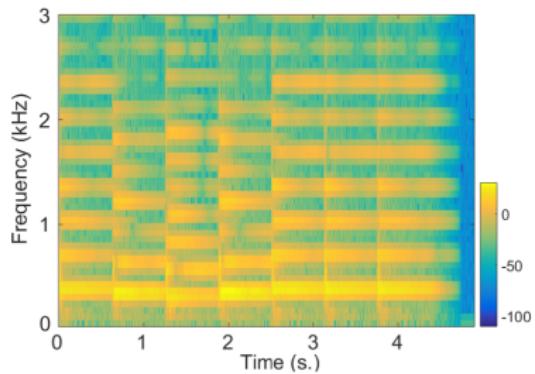
Nonnegative Matrix Factorization



Image data



Hyperspectral data

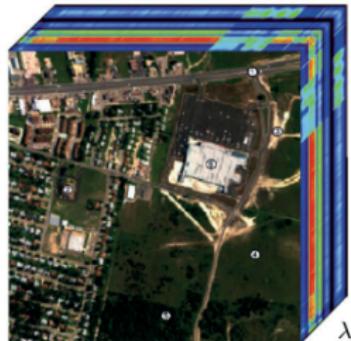


Signal data

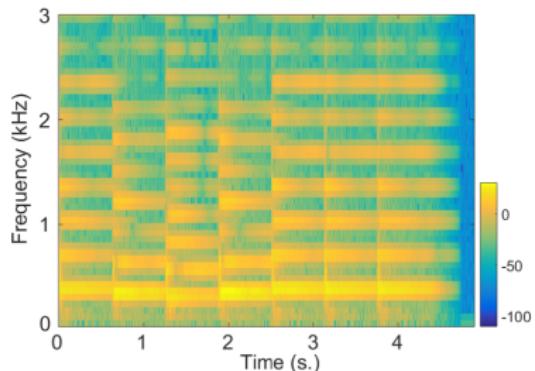
Nonnegative Matrix Factorization



Image data



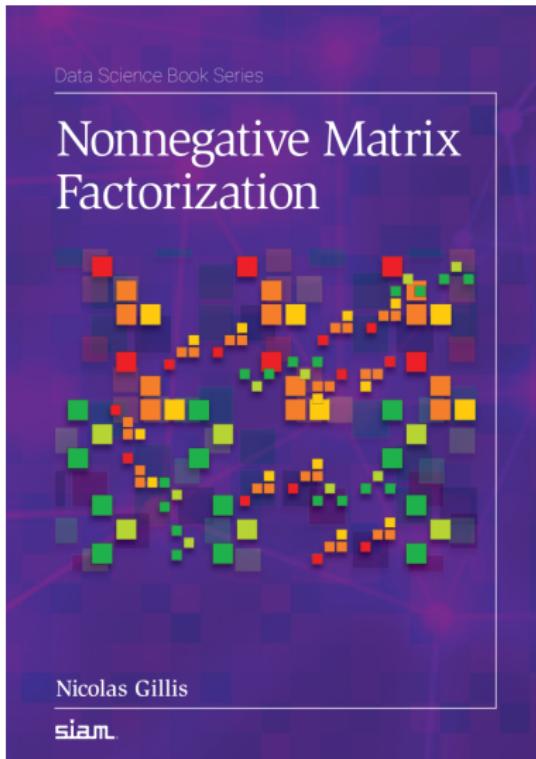
Hyperspectral data



Signal data

Most of data can be expressed as nonnegative matrix!

Nonnegative Matrix Factorization



Nonnegative Matrix Factorization

Nonnegative Matrix Factorization

Given a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$, a factorization rank r , and a distance measure $D(\cdot, \cdot)$ between two matrices, compute two nonnegative matrices $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$ such that $D(X, WH)$ is minimized, that is solve

$$\min_{W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}} D(X, WH). \quad (1)$$

We call an NMF model is an optimization model that requires the choice of

- the variables (in the standard NMF model, the factors W and H),
 - the objective function (such as the standard least squares error $\|X - WH\|_2^F$) with or without regularizers (such as $\|H\|_1$ to induce sparse solutions),
 - constraints on the variables (such as nonnegativity of W and H in the standard NMF model, and orthogonality with $HH^\top = I$ in the ONMF model).

Statistical model and maximum likelihood

Error measure

Error measure used to evaluate the quality of the approximation, WH of X , denoted as $D(X, WH)$.

Suppose that the entry at position (i, j) of matrix X contains the observations of a random variable, \tilde{X} , defined by the parameter $(\hat{W}\hat{H})_{ij}$

Example

Consider $\tilde{X} = \hat{W}\hat{H} + \tilde{N}$, where the factor $\hat{W} \geq 0$ and $\hat{H} \geq 0$ are deterministic, and the noise is i.i.d. Gaussian with mean 0 and standard deviation σ .

$$\tilde{X}_{ij} \sim \mathcal{N} \left((\hat{W}\hat{H})_{ij}, \sigma \right) \quad \text{for all } i, j \text{ and some } \sigma > 0.$$

Thus the probability density function of \tilde{X}_{ij} is

$$p\left(\tilde{X}_{ij}; (\hat{W}\hat{H})_{ij}, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2} \left(\tilde{X}_{ij} - (\hat{W}\hat{H})_{ij}\right)^2}$$

Statistical model and maximum likelihood

Example

Since the noise is assumed to be i.i.d., the likelihood of the sample X with respect to $(\hat{W}\hat{H})_{ij}$ and σ is

$$\ell(X; \hat{W}\hat{H}, \sigma) = \prod_{i,j} p\left(X_{ij}; (\hat{W}\hat{H})_{ij}, \sigma\right). \quad (2)$$

Given a sample X , the unknown parameters, \hat{W} , \hat{H} , and σ , can be estimated by solving the optimization problem

$$\max_{W \geq 0, H \geq 0, \sigma} \ell(X; \hat{W}\hat{H}, \sigma).$$

We can modify this optimization problem as

$$\min_{W \geq 0, H \geq 0} D(X, WH) \text{ where } D(X, WH) = \sum_{i,j} (X - WH)_{ij}^2 = \|X - WH\|_F^2.$$

which is obtained by taking the logarithm of (2).

Statistical model and maximum likelihood

Acronym	$D(X, WH)$	Distribution [†]
ℓ_2 -NMF [303]	$\ X - WH\ _F^2 = \sum_{i,j} (X - WH)_{ij}^2$	Gaussian
Weighted NMF [179]	$\sum_{i,j} P_{ij} (X - WH)_{ij}^2$	independently distributed entries, Gaussian
ℓ_1 -NMF [273]	$\ X - WH\ _1 = \sum_{i,j} X - WH _{ij}$	Laplace
ℓ_∞ -NMF [209]	$\ X - WH\ _\infty = \max_{i,j} X - WH _{ij}$	Uniform
KL-NMF [303]	$D_1(X, WH)$	Poisson
IS-NMF [158]	$D_0(X, WH)$	multiplicative Gamma
β -NMF [160]	$D_\beta(X, WH)$	Tweedie distributions

[†]If not specified, the noise is i.i.d.

Table 1: Several error measures for NMF and the corresponding distribution.

β -divergence

An important class of estimators is based on the β -divergences. Given two nonnegative scalars z and y , the β -divergence between z and y is defined as follows:

$$d_\beta(z, y) = \begin{cases} \frac{z}{y} - \log \frac{z}{y} - 1 & \text{for } \beta = 0 \\ z \log \frac{z}{y} - z + y & \text{for } \beta = 1 \\ \frac{1}{\beta(\beta-1)} \left(z^\beta + (\beta-1)y^\beta - \beta z y^{\beta-1} \right) & \text{for } \beta \neq 0, 1 \end{cases} \quad (3)$$

And the β -divergence between two matrices A and B is

$$D_\beta(A, B) = \sum_{i,j} d_\beta(A_{ij}, B_{ij}).$$

β -divergence

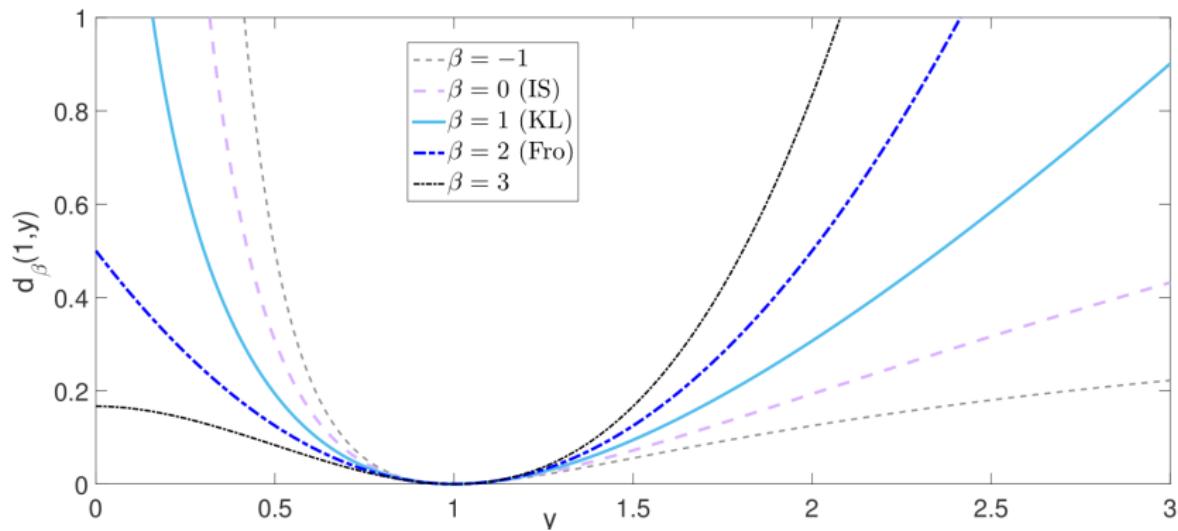


Figure 1: Illustration of the β -divergences $d_\beta(1, y)$ for $\beta = -1, 0, 1, 2, 3$.

β -divergence

There are two important properties of the β -divergences:

■ Convexity

The function $d_\beta(z, y)$ is convex in the second argument, y , for $\beta \in [1, 2]$.

This implies that $D_\beta(X, WH)$ is convex in H for W fixed and vice versa.

■ Scaling

$$d_\beta(\gamma z, \gamma y) = \gamma^\beta d_\beta(z, y)$$

This implies that the larger the β , the more sensitive is the β -divergence to large values of z , and vice versa.

The NMF problem using the β -divergence, which we refer to as β -NMF, is the following: Given $X \in \mathbb{R}_+^{m \times n}$ and r , solve

$$\min_{W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}} D_\beta(X, WH)$$

β -divergence

Example (Over-/underapproximations)

Let $\mathbf{X} = \text{sprand}(100, 100, 0.5)$ and compute a β -NMF (\mathbf{W}, \mathbf{H}) for $r = 10$ via 100 iterations of the multiplicative update(MU) technique.

For $\beta = 0$ (IS-NMF),

$$\frac{\|\max(0, WH - X)\|_F}{\|X - WH\|_F} \geq 100.00\% \text{ while } \frac{\|\max(0, X - WH)\|_F}{\|X - WH\|_F} \leq 0.33\%$$

so that WH over-approximates X in all cases as most entries of WH are larger than X .

And for $\beta = 2$ (ℓ_2 -NMF),

$$\frac{\|\max(0, WH - X)\|_F}{\|X - WH\|_F} \leq 59.84\% \text{ while } \frac{\|\max(0, X - WH)\|_F}{\|X - WH\|_F} \geq 80.12\%$$

so that WH is more balanced around X although it tends to under-approximate it.

Choice of the error measure

Choosing the right objective function for your NMF model can be crucial.

- Empirical choice
 - Cross validation
 - for music transcription based on NMF, the β -divergence with $\beta = 0.5$ performs best.
 - for hyperspectral images, the β -divergence with $\beta \approx 1.5$ performs best.
 - Statistical approaches
 - score matching minimizes the expected squared Euclidean distance between the scores of the true distribution and the model.
 - A maximum likelihood approach can also be used to assess whether the observed data is more likely to follow a given distribution.
 - Distributional robustness
 - More recently, a distributionally robust NMF (DR-NMF) model was proposed.

$$\min_{W \geq 0, H \geq 0} \max_{\beta \in \Omega} D_\beta(X, WH),$$

where Ω is a subset of β 's interest.

- for audio signals where both KL and IS divergences are often used, using DR-NMF with $\Omega = \{0, 1\}$ leads to a low reconstruction error for both IS and KL divergences.

Applications of NMF Models

Name	Model
NMF	$W \geq 0, H \geq 0$
ONMF	$W \geq 0, H \geq 0, HH^\top = I_r$
projective NMF	$W = XH^\top, H \geq 0$
convex NMF	$W = XC, C \geq 0, H \geq 0$
separable NMF	$W = X(\cdot, \mathcal{K})$ with $ \mathcal{K} = r, H \geq 0$
dictionary NMF	$W = DC \geq 0, D$ dictionary, $H \geq 0$
semi-NMF	$H \geq 0$
sparse NMF	$W \geq 0, H \geq 0, W$ and/or H sparse
affine NMF	$X \approx WH + we^\top, W \geq 0, H \geq 0, w \geq 0$
NMU	$WH \leq X, W \geq 0, H \geq 0$
convolutive NMF	$X \approx \sum_{\ell=1}^r \sum_{k=1}^p W_\ell(\cdot, k) [0_{1 \times (k-1)} H(\ell, 1 : n - k + 1)]$ $W_\ell \in \mathbb{R}_+^{m \times p} \quad (1 \leq \ell \leq r), H \in \mathbb{R}_+^{r \times n}$ $W = H^\top \geq 0$
symNMF	$X \approx WSH, W \in \mathbb{R}_+^{m \times r_1}, S \in \mathbb{R}_+^{r_1 \times r_2}, H \in \mathbb{R}_+^{r_2 \times n}$
tri-NMF	$tri\text{-}NMF \& W^\top W = I_{r_1}, HH^\top = I_{r_2}$
tri-ONMF	$tri\text{-}NMF \& W = H^\top, S = S^\top$
tri-symNMF	$X \approx WH_1 H_2 \dots H_t, W \geq 0, H_i \geq 0 \text{ for all } i$
deep NMF	$W \in \{0, 1\}^{m \times r}, H \in \{0, 1\}^{r \times n}$
binary NMF	
Boolean NMF	$X \approx \min(WH, 1), W \in \{0, 1\}^{m \times r}, H \in \{0, 1\}^{r \times n}$
interval-valued NMF	$(WH)_{i,j} \in X(i, j) = [a(i, j), b(i, j)]$
kernel NMF	$\Phi(X) \approx WH, W \geq 0, H \geq 0$
bilinear NMF	$W \geq 0, H \geq 0, H^\circ \geq 0$
	$X(\cdot, j) \approx WH(\cdot, j) + \sum_{k < \ell} (W(\cdot, k) \circ W(\cdot, \ell)) H^\circ(k, \ell, j)$

Table 2: NMF variants for a given data matrix X .

Nonnegative Matrix Factorization

NMF in data analysis

- Feature extraction in a set of images
 - Blind hyperspectral unmixing
 - Audio source separation

Recall : NMF

Given a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$, find nonnegative matrices W and H such that

$$X \cong WH$$

Feature extraction in a set of images

Feature extraction in a set of images

We applied NMF on the CBCL face data set[4].

CBCL face data set is

- greyscale
 - $m = 2429$ images
 - each images is $n = 19 \times 19$ pixels

We create $V \in \mathbb{R}^{n \times m}$ and find W and H by NMF.

$$\begin{array}{c}
 \text{Image} \approx X(:, j) * H(:, j) = WH(:, j) \\
 \text{Image} \\
 \approx \\
 X(:, j) \\
 * \\
 H(:, j) \\
 = \\
 WH(:, j)
 \end{array}$$

Feature extraction in a set of images

Q. What are the advantages of NMF?

A. NMF learns to represent faces with a set of basis images resembling parts of faces.

Feature extraction in a set of images

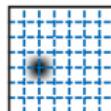
Q. What are the advantages of NMF?

A. NMF learns to represent faces with a set of basis images resembling parts of faces.

VG



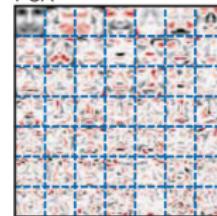
×



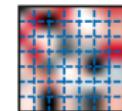
2



PCA



1



2



Feature extraction in a set of images

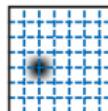
Q. What are the advantages of NMF?

A. NMF learns to represent faces with a set of basis images resembling parts of faces.

VC



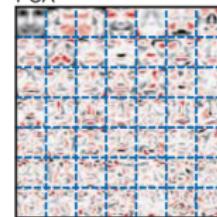
×



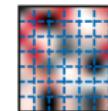
2



PGM



10

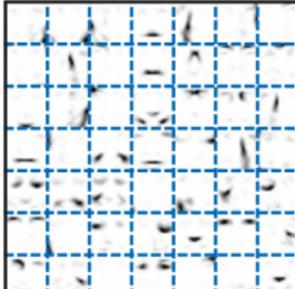


2

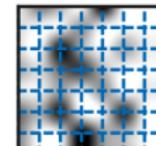


Original

NMF



X



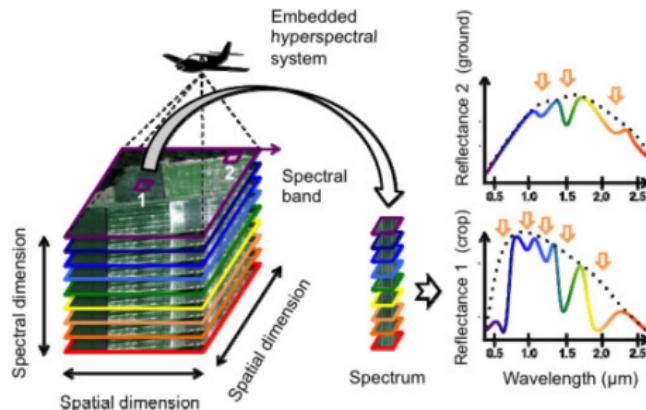
6



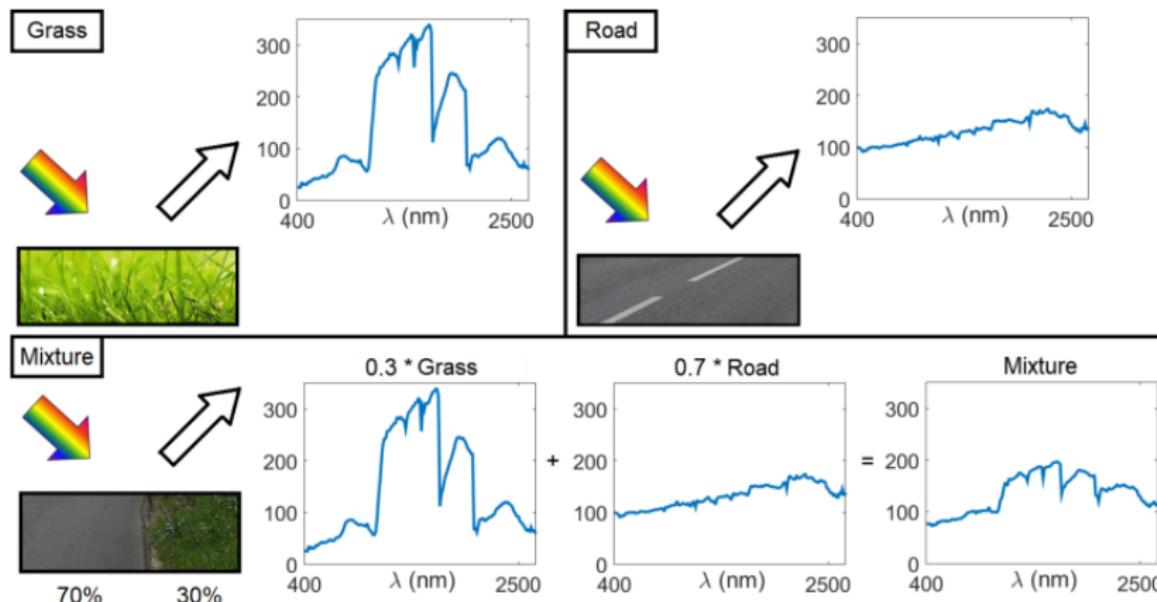
Blind hyperspectral unmixing

Blind hyperspectral unmixing

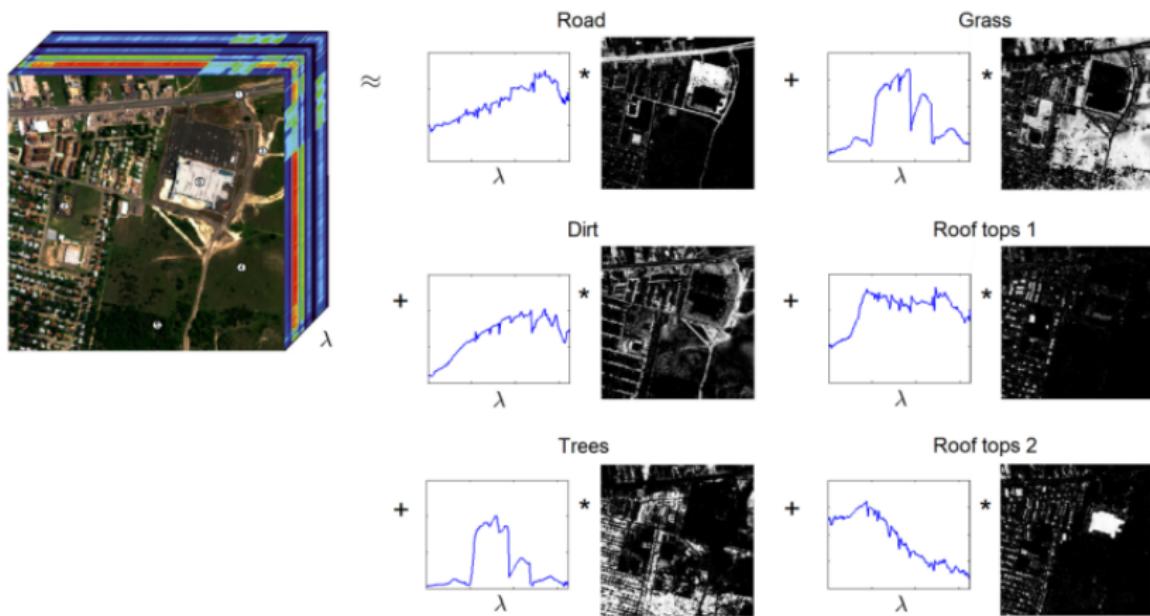
A hyperspectral image measures the intensity of the light within a scene for many different wavelengths. Hence, for each pixel, a vector of intensities is recorded that is equal to the fraction of light reflected by that pixel depending on the wavelength; this is referred to as the spectral signature of the pixel. Given a hyperspectral image, the goal of blind hyperspectral unmixing (blind HU) is to recover the materials present in an image, referred to as the endmembers, and their proportions in each pixel, referred to as abundances.



Blind hyperspectral unmixing



Blind hyperspectral unmixing



Audio source separation

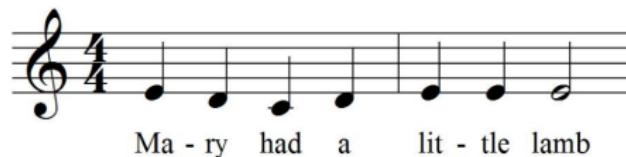
Audio source separation

Given an audio signal recorded from a single microphone, we can define matrix X following as. The entry of X at position (i, j) is the magnitude of the Fourier coefficient for the j th time frame at the i th frequency.

Goal : to blindly separate the sources that compose the signal.

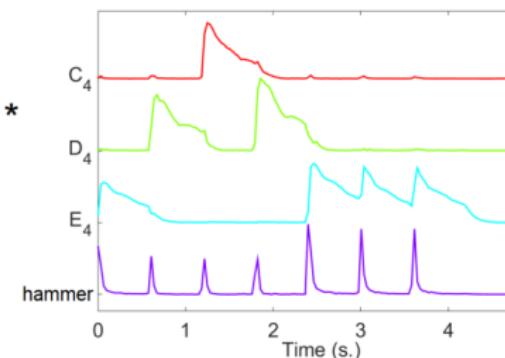
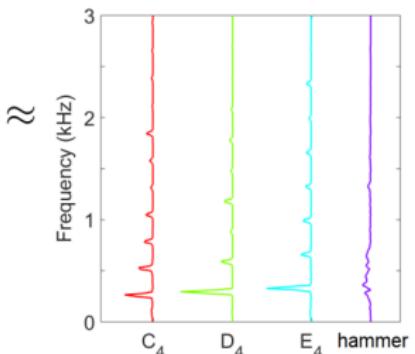
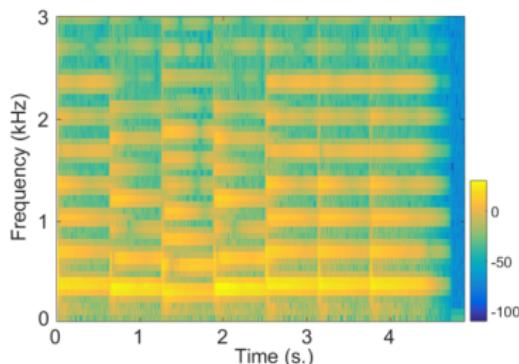
Example : separate the voice and the instruments in a song.

Let us use a simple monophonic signal for illustrative purposes, namely a piano recording of "Mary Had a Little Lamb," whose musical score is shown below.



The sequence is composed of three notes, C_4 , D_4 , and E_4 , that activate as follows: E_4 , D_4 , C_4 , D_4 , E_4 , E_4 , E_4 .

Audio source separation



Symmetric Nonnegative Matrix Factorization

Symmetric nonnegative matrix factorization(SymNMF)

SymNMF requires $W = H^\top$, that is, $X \approx WW^\top$. SymNMF allows us to perform such a task. SymNMF decomposes X as follows:

$$X \approx WW^\top = \sum_{k=1}^r W(:, k)W(:, k)^\top.$$

SymNMF can be applied to graph theory. In the exact case, when $X = WW^\top$, X is decomposed into r cliques. In summary, each rank-one matrix $W(:, k)W(:, k)^\top$ in a symNMF of X corresponds to a subset of nodes that are highly connected.

- Pixel clustering
If $X(i, j)$ is the similarity between pixels in an image, a symNMF of X provides a soft clustering of the pixels into homogeneous regions.
 - Document clustering
If $X(i, j)$ is the similarity between documents, symNMF classifies these documents into similar topics.

Symmetric Nonnegative Matrix Factorization

Let us illustrate the capacity of symNMF to split the nodes of a graph into different communities on a simple example using the Zachary's karate club data set[6].

Zachary's karate club[6]

Zachary is a researcher who studied the relationships between the members of a karate club. Each edge in the graph represents the friendship between two members of the club. There are 34 members and 78 friendship links.

During his study, Zachary observed a dispute between the administrator and the instructor of the club, which resulted in the instructor leaving the club to start a new one, taking about half of the original club's members with him.

Applying symNMF with $r = 2$ to the symmetric adjacency matrix of this graph, $X \in \mathbb{R}_+^{34 \times 34}$, allows two communities to be identified, where each column of W represents a community. Note that $X(i, j)$ represents the affinity between i and j , and hence we set $X(i, i) = 1$ for $i = 1, 2, \dots, n$.

Symmetric Nonnegative Matrix Factorization

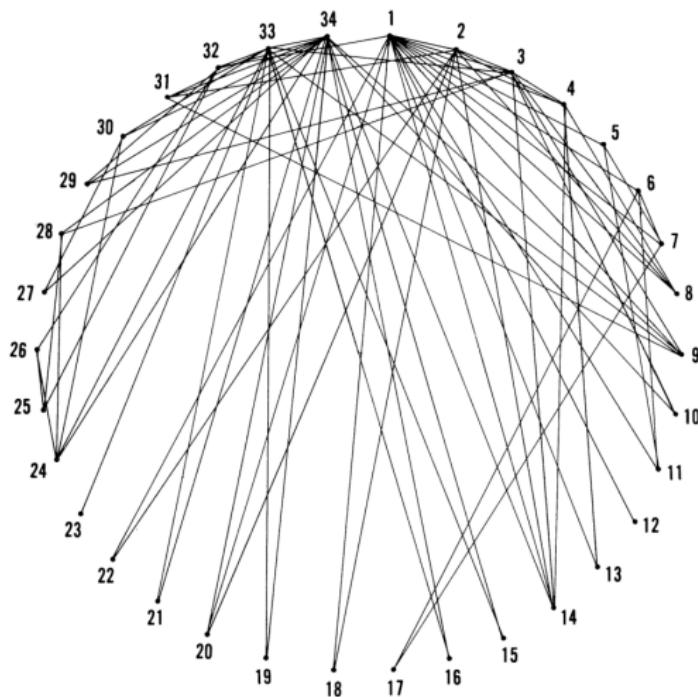
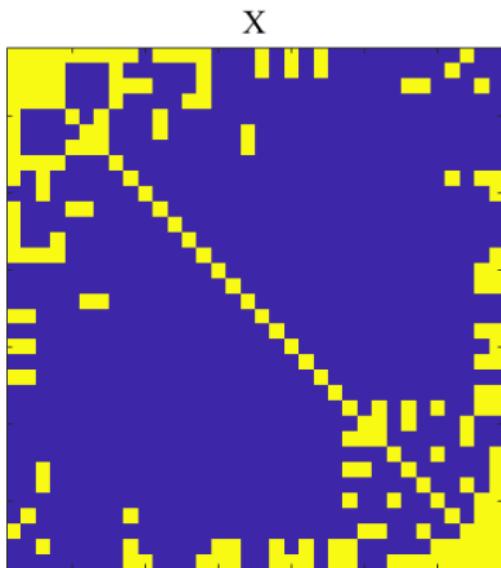
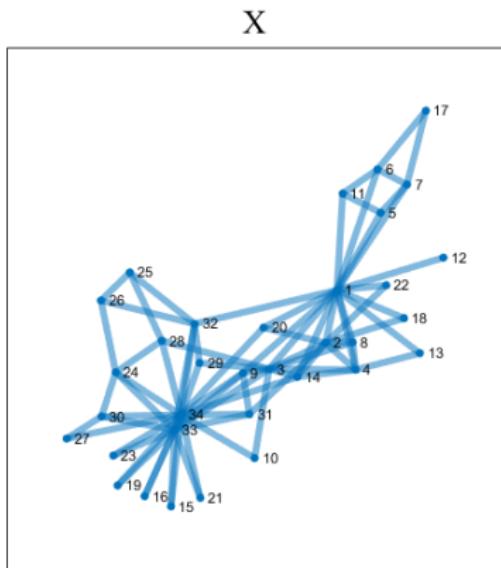


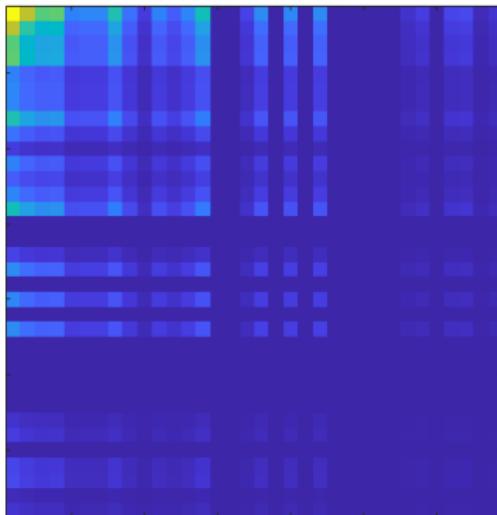
Figure 2: Social Network Model of Relationships in the Karate Club[6]

Symmetric Nonnegative Matrix Factorization

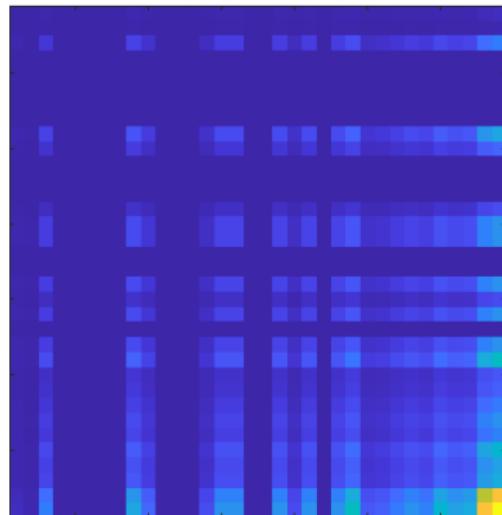


Symmetric Nonnegative Matrix Factorization

$$W(:,1) \times W(:,1)^T$$

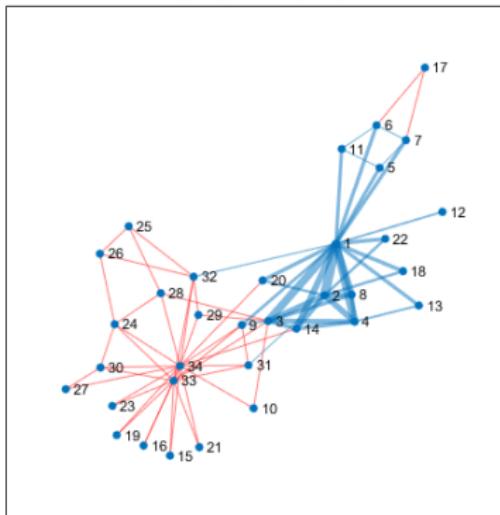


$$W(:,2) \times W(:,2)^T$$

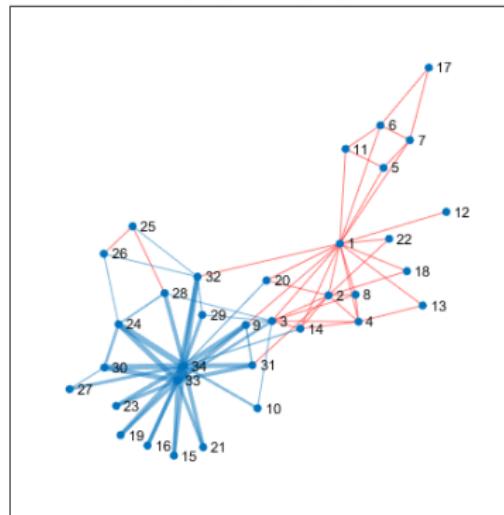


Symmetric Nonnegative Matrix Factorization

$$W(:,1) \times W(:,1)^T$$



$$W(:,2) \times W(:,2)^T$$



Symmetric Nonnegative Matrix Factorization

Recall

Recall that symNMF leads to a soft clustering: some vertices belong to the two communities with different intensities. For example, node 9 is rather central in the graph and is shared among the two communities, with $W(9, 1) = 0.32$ and $W(9, 2) = 0.54$. This node is actually the only one “misclassified” by symNMF in the sense that the person represented by node 9 left the club with the instructor (node 1), not with the administrator (node 34).

Symmetric Nonnegative Matrix Trifactorization

Nonnegative matrix trifactorization(tri-NMF)

The NMF model with three factor matrices, referred to as nonnegative matrix trifactorization(tri-NMF), is the following: Given $X \in \mathbb{R}_+^{m \times n}$, r_1 and r_2 , find $W \in \mathbb{R}_+^{m \times r_1}$, $S \in \mathbb{R}_+^{r_1 \times r_2}$, and $H \in \mathbb{R}_+^{r_2 \times n}$ such that

$$X \approx WSH$$

Symmetric nonnegative matrix trifactorization(tri-symNMF)

Given a symmetric nonnegative matrix $X \in \mathbb{R}_+^{m \times m}$ and a factorization rank r , it looks for a nonnegative matrix $R \in \mathbb{R}_+^{m \times r}$ and a symmetric nonnegative matrix $S \in \mathbb{R}_+^{r \times r}$ such that

$$X \approx WSW^\top$$

i.e., tri-NMF & $W = H^\top, S = S^\top$

Symmetric Nonnegative Matrix Trifactorization

Interpretations of symNMF and tri-NMF

As for tri-NMF, tri-symNMF allows these communities to interact via the factor S . The entry $W(j, k)$ can be interpreted as the membership indicator of item j for community k . The entry $S(k, l)$ is the strength of the connection between communities k and l .

So,

$$X(i, j) \approx W(i, :) S W(j, :)^\top = \sum_{k=1}^r \sum_{l=1}^r W(i, k) S(k, l) W(j, l)$$

The value $X(i, j)$ reflects the memberships of items i and j in the different communities and how these communities interact together.

Text mining: topic recovery and document classification

Let each column of the matrix X correspond to a document, that is, a nonnegative vector of word counts. For example, the entry of X at position (i, j) can be the number of times word i appears in document j .

Term-Document Matrix(TDM)

- D1 = "I like databases"
 - D2 = "I dislike databases"

then the document-term matrix would be:

	I	like	dislike	databases
D1	1	1	0	1
D2	1	0	1	1

Text mining: topic recovery and document classification

The matrix X can also be constructed in different, more sophisticated ways, for example, with the term frequency-inverse document frequency (tf-idf)[5].

Term Frequency times Inverse Document Frequency(TF-IDF)

Suppose we have a collection of N documents. Define f_{ij} to be the frequency (number of occurrences) of term (word) i in document j . And suppose term i appears in n_i of the N documents in the collection.

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}} \quad \text{and} \quad IDF_i = \log_2(N/ni)$$

Finally, The TF-IDF score for term i in document j is then defined to be

$$TF-IDF_{ij} = TF_{ij} \times IDF_i$$

The terms with the highest TF-IDF score are often the terms that best characterize the topic of the document.

Text mining: topic recovery and document classification

This is the so-called bag of words model where the positions of the words in a document are not taken into account. The NMF of X provides the model

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

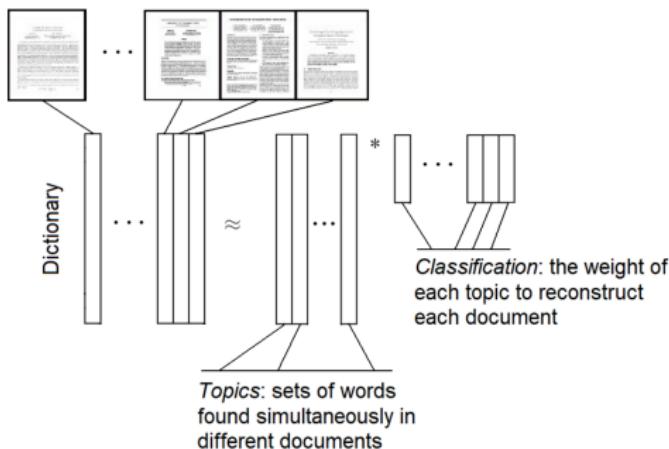


Figure 3: Illustration of NMF for text mining: extraction of topics, and classification of each document with respect to these topics.

Topic modeling

Since the word-by-document matrix X is usually full rank, X is typically far from a low-rank matrix, and it does not follow the NMF model $X \approx WH$ very closely. The vector $X(:, j)$ is a sample of a random variable $\tilde{x}_j \in \mathbb{R}^m$. The distribution of \tilde{x}_j is such that $\mathbb{E}(\tilde{x}_j) = \hat{W}\hat{H}(:, j)$ where (\hat{W}, \hat{H}) are deterministic but unknown parameters to be estimated. In the context of topic modeling, these parameters can be interpreted as follows

- The columns of \hat{W} correspond to topics.
 - $\sum \hat{W}(:, k) = 1$ for all k
 - $\hat{W}(i, k)$ is the probability of picking the word i when discussing the topic k .
- The vector $\frac{\hat{H}(:, j)}{\|\hat{H}(:, j)\|_1}$ indicates the proportion of each topic discussed in the j th document, while $\|\hat{H}(:, j)\|_1$ equals the number of words present in the document.

Probabilistic Latent Semantic Analysis and Indexing

In PLSA, the number of documents, n , is assumed to be fixed, while the dictionary contains m words. The observation is a matrix of word counts, $X \in \mathbb{Z}_+^{m \times n}$, where $X(i, j)$ is the number of times word i appears in document j .

$$\ell = e^\top X e$$

is length of a set of documents.

Let us define

- the vector $\hat{s} \in \mathbb{R}_+^r$ where $\hat{s}(k)$ is the probability of a word sampled randomly to be associated to with the k th topic for $k = 1, 2, \dots, r$ with $\hat{s}^\top e = 1$.
- the matrix $\hat{A} \in \mathbb{R}_+^{m \times r}$ where $\hat{A}(i, k)$ is the probability of using the i th word in the dictionary assuming we are discussing the k th topic, for $i = 1, 2, \dots, m$ and $k = 1, 2, \dots, r$ with $\hat{A}^\top e = e$ and
- the matrix $\hat{B} \in \mathbb{R}_+^{r \times n}$ where $\hat{B}(k, j)$ is the probability of using the j th document assuming we are discussing the k th topic, for $k = 1, 2, \dots, r$ and $j = 1, 2, \dots, n$ with $\hat{B}e = e$.

Probabilistic Latent Semantic Analysis and Indexing

Then, PLSA assumes the word co-occurrence matrix X of length ℓ is a sample of a random variable \tilde{X} and is generated by sampling ℓ words as follows:

- 0 Set $X(i, j) = 0$ for $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$.
 - 1 For $p = 1, 2, \dots, \ell$,
 - 1.1 Pick a topic $k \in \{1, 2, \dots, r\}$ with probability given by \hat{s} .
 - 1.2 Pick a word $i \in \{1, 2, \dots, n\}$ with probability given by $\hat{A}(:, k)$.
 - 1.3 Pick a document $j \in \{1, 2, \dots, m\}$ with probability given by $\hat{B}(k, :)$.
 - 1.4 $X(i, j) = X(i, j) + 1$.

Probabilistic Latent Semantic Analysis and Indexing

Example of PLSA

- D1 = "I like databases"
 - D2 = "I love databases"
 - D3 = "Math is my life"
 - D4 = "Math is funny"

The TDM of these documents as follow:

	I	like	love	databases	Math	is	my	life	funny
D1	1	1	0	1	0	0	0	0	0
D2	1	0	1	1	0	0	0	0	0
D3	0	0	0	0	1	1	1	1	0
D4	0	0	0	0	1	1	0	0	1

Probabilistic Latent Semantic Analysis and Indexing

Example of PLSA

Let the topic is “databases” and “math” then the number of topic is 2, so example of \hat{s} is

$$\hat{s}^T = [0.5, 0.5]$$

And example of \hat{A} and \hat{B} is as follows:

$$\hat{A} = \begin{bmatrix} 0.333 & 0 \\ 0.167 & 0 \\ 0.167 & 0 \\ 0.333 & 0 \\ 0 & 0.286 \\ 0 & 0.286 \\ 0 & 0.143 \\ 0 & 0.143 \\ 0 & 0.143 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 0.5 & 0.5 \\ 0.571 & 0.429 \end{bmatrix}$$

And the length of all document $\ell = 13$.

Probabilistic Latent Semantic Analysis and Indexing

PLSA assumes that each word sampled in the data set is generated so that the words and documents are conditionally independent given the hidden topic. The above model implies that

$$\frac{1}{\ell} \mathbb{E} \left(\tilde{X} \right) = \hat{A} \operatorname{diag}(\hat{s}) \hat{B}$$

since $\frac{1}{\ell} \mathbb{E} \left(\tilde{X}_{ij} \right) = \sum_{k=1}^r \hat{s}(k) \hat{A}(i, k) \hat{B}(k, j)$.

Moreover, if ℓ is sufficiently large, $\frac{1}{\ell}X$ get closer to $\frac{1}{\ell}\mathbb{E}(\tilde{X})$.

Finally, we have

$$X \approx \ell \hat{A} \operatorname{diag}(\hat{s}) \hat{B}$$

Probabilistic Latent Semantic Analysis and Indexing

Now our goal of PLSA is to estimate \hat{s} , \hat{A} , and \hat{B} for given X and r .

We assume that $\tilde{X}(i, j)$ follows Poisson distribution of parameter $(\hat{A} \text{ diag}(\hat{s}) \hat{B})_{i, j}$ for PLSA, i.e., a probability mass function given by

$$\Pr(\tilde{X}(i, j) = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where } \lambda = (\hat{A} \text{ diag}(\hat{s}) \hat{B})_{i,j}$$

It then uses the maximum likelihood estimator for $(\hat{A}, \hat{s}, \hat{B})$ which is obtained by solving

$$\max_{(A,s,B) \geq 0} \sum_{i,j,k} X_{i,j} \log(A \operatorname{diag}(s)B)_{i,j} \text{ such that } s^\top e = 1, A^\top e = e \text{ and } B^\top e = e. \quad (4)$$

A solution (A, s, B) can be used to construct an NMF (W, H) of X , by choosing $W = A$ and $H = \ell \operatorname{diag}(s)B$ so that

$$X \approx \ell A \operatorname{diag}(s) B = W H.$$

Conclusion



“All models are wrong, but some are useful” - George E. P. Box

Reference

- [1] Sanjeev Arora et al. "A practical algorithm for topic modeling with provable guarantees". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 280–288.
- [2] Nicolas Gillis. *Nonnegative Matrix Factorization*. SIAM, 2020.
- [3] Majid Janzamin et al. "Spectral learning on matrices and tensors". In: *arXiv preprint arXiv:2004.07984* (2020).
- [4] Daniel D Lee and H Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.
- [5] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.
- [6] Wayne W Zachary. "An information flow model for conflict and fission in small groups". In: *Journal of anthropological research* 33.4 (1977), pp. 452–473.

Thank you!