Nonnegative Matrix Factorization
NMF model and Applications

Taehyeong Kim

Mathematics, Pusan National University

July 10, 2021

Content

- Introduction
- Error measures
  - Statistical model and maximum likelihood
  - β-divergence
  - Choice of the error measure
- Application of NMF Model
  - 5.4.7. Symmetric Nonnegative Matrix Factorization
  - 5.4.9. Symmetric Nonnegative Matrix Trifactorization
  - 1.3.3. Text mining: topic recovery and document classification
  - 5.4.9.1. Topic modeling
  - 5.5.4. Probabilistic Latent Semantic Analysis and Indexing
- Summary

Introduction

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).

Introduction

In some applications, the input matrix is not close to a low-rank matrix. Typical examples is word count data sets used in text mining (Sections 1.3.3, 5.5.4, and 5.4.9.1).

Related part in Book

- 1.3.3. Text mining: topic recovery and document classification
- 5.4.9. Symmetric nonnegative matrix trifactorization
- 5.4.9.1. Topic modeling
- 5.5.4. Probabilistic latent semantic analysis and indexing

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

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

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

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

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

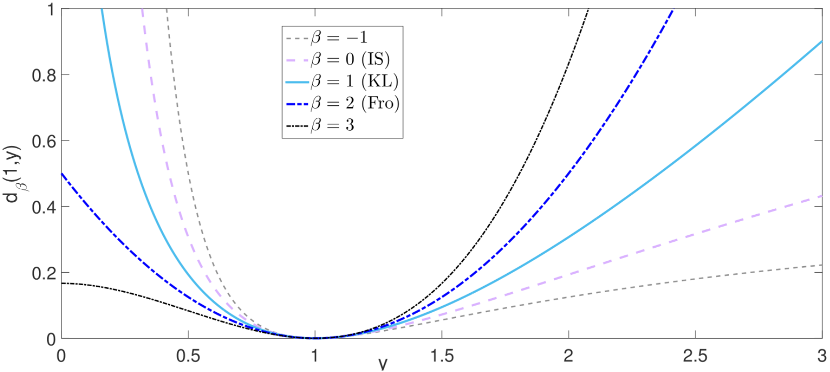


Figure 1: Illustration of the β-divergences $d_\beta(1, y)$ for $\beta = -1, 0, 1, 2, 3$.

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 $\texttt{X} = \texttt{sprand}(100, 100, 0.5)$ and compute a β-NMF (W, 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.

Application of NMF Model

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	$W = X(:, \mathcal{K})$ with $\lvert\mathcal{K}\rvert = r, H \geq 0$
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	$X \approx \sum_{\ell=1}^{r} \sum_{k=1}^{p} W_\ell(:, k) \left[0_{1 \times (k-1)}\ H(\ell, 1 : n-k+1)\right]$, $W_\ell \in \mathbb{R}_+^{m \times p}\ (1 \leq \ell \leq r), H \in \mathbb{R}_+^{r \times n}$
symNMF	$W = H^\top \geq 0$
tri-NMF	$X \approx WSH, W \in \mathbb{R}_+^{m \times r_1}, S \in \mathbb{R}_+^{r_1 \times r_2}, H \in \mathbb{R}_+^{r_2 \times n}$
tri-ONMF	tri-NMF & $W^\top W = I_{r_1}, HH^\top = I_{r_2}$
tri-symNMF	tri-NMF & $W = H^\top, S = S^\top$
deep NMF	$X \approx WH_1H_2 \ldots H_t, W \geq 0, H_i \geq 0$ for all i
binary NMF	$W \in \{0,1\}^{m \times r}, H \in \{0,1\}^{r \times n}$
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	$W \geq 0, H \geq 0, H^\circ \geq 0$, $X(:,j) \approx WH(:,j) + \sum_{k<\ell} \left(W(:,k) \circ W(:,\ell)\right) H^\circ(k,\ell,j)$

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

5.4.7. Symmetric Nonnegative Matrix Factorization

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.

And there are several applications of symNMF.

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

5.4.7. 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[3].

Zachary's karate club[3]

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 $i = 1, 2, \ldots, n$.

5.4.7. Symmetric Nonnegative Matrix Factorization

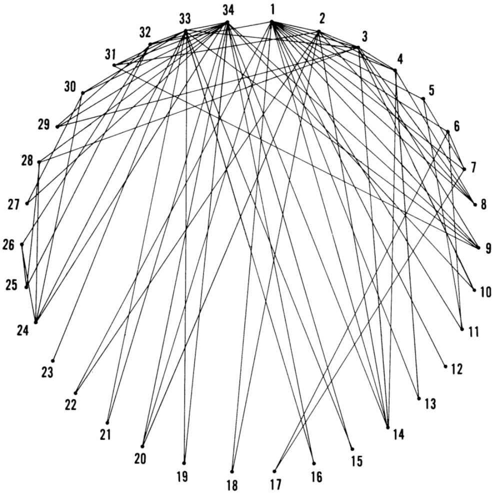
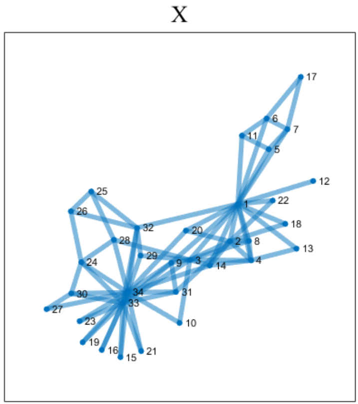
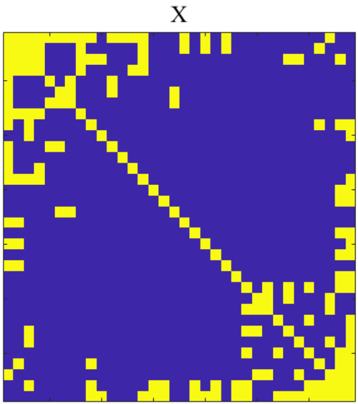


Figure 2: Social Network Model of Relationships in the Karate Club[3]

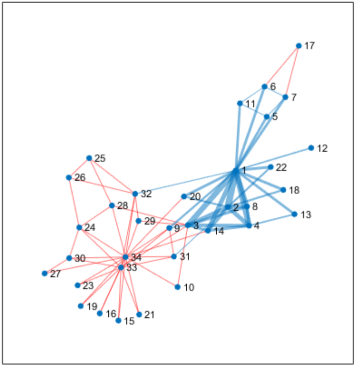
5.4.7. Symmetric Nonnegative Matrix Factorization



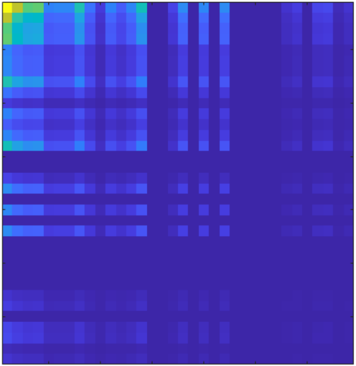


5.4.7. Symmetric Nonnegative Matrix Factorization

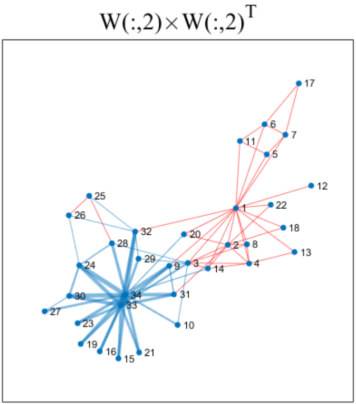
$W(:,1) \times W(:,1)^T$

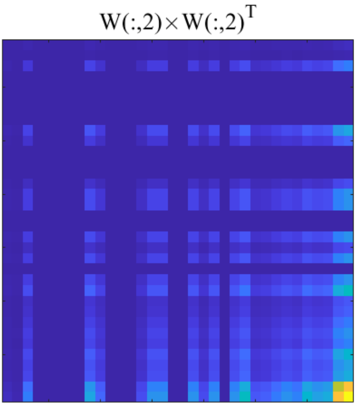


$W(:,1) \times W(:,1)^T$



5.4.7. Symmetric Nonnegative Matrix Factorization





5.4.7. Symmetric Nonnegative Matrix Factorization

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).

5.4.9. 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 & $W = H^\top, S = S^\top$

5.4.9. 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,

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

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

1.3.3. 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

1.3.3. 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)[2].

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

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

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

1.3.3. 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

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

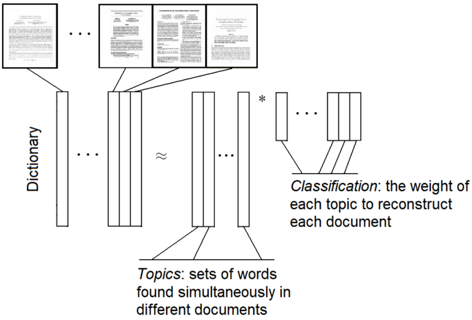


Figure 3: Illustration of NMF for text mining: extraction of topics, and classification of each document with respect to these topics.

5.4.9.1. 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.
  - $\sum \hat{W}(:, k) = 0$ 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 jth document, while $\|\hat{H}(:, j)\|_1$ equals the number of words present in the document.

5.4.9.1. Topic modeling

Consider

$$XX^\top$$

The entry $(XX^\top)_{i,j}$ is equal to the number of different combinations of the words i and j appearing in the same document. The symmetric matrix $XX^\top$ can be interpreted as the weighted adjacency matrix of a graph connecting nodes corresponding to the words in the dictionary.
Let the matrix $\hat{W} \in \mathbb{R}_+^{r \times n}$ be following as.

- deterministic but unknown
- word-by-topic matrix whose entry at position (i, k) contains the probability for word i to be used in topic k

And let vector $\tilde{h}$ be a random variable corresponding to the proportions of the topics discussed within a document. Then the columns of X are assumed to be generated as follows.

5.4.9.1. Topic modeling

For $j = 1, 2 \dots, n$,

1. let the vector $H(:, j) \in \Delta^r$ be a sample of the random variable $\tilde{h}$
2. $X(:, j)$ is the sample of a multinomial distribution of parameters $\hat{W}H(:, j)$ the probability to pick the ith word in the dictionary is $(\hat{W}H(:, j))_i$.

There are two key differences of the above model with NMF:

- The columns of X are sampled from the same distribution with the same parameters.
  In NMF, the columns of X are sampled from the same distributions but with different parameters, namely with parameters $\hat{W}\hat{H}(:, j)$ for the jth column of X.
- When the number of words sampled in the j document, $e^\top X(:, j)$, is not sufficiently large, we will not have

$$\frac{X(:, j)}{e^\top X(:, j)} \approx \hat{W}H(:, j).$$

5.4.9.1. Topic modeling

Finally, as the number n of sampled documents goes to infinity, we have

$$\lim_{n\to\infty} \frac{XX^\top}{e^\top XX^\top e} = \mathbb{E}\left(\hat{W}\tilde{h}\tilde{h}^\top \hat{W}^\top\right) = \hat{W}\underbrace{\mathbb{E}\left(\tilde{h}\tilde{h}^\top\right)}_{=S}\hat{W}^\top,$$

where $S \in \mathbb{R}^{r\times r}$ is the topic-by-topic matrix, which is the second-order moment of $\tilde{h}$. If the number of documents observed is sufficiently large, the use of the tri-symNMF,

$$\frac{XX^\top}{e^\top XX^\top e} \approx \hat{W}S\hat{W}^\top$$

is justified by the probabilistic topic models as described before. For more details, see [1].

5.5.4. 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 kth topic for $k = 1, 2, \ldots, r$ 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 ith word in the dictionary assuming we are discussing the kth topic, for $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, r$ 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 jth document assuming we are discussing the kth topic, for $k = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, n$ with $\hat{B}e = e$.

5.5.4. 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 $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

1 For $p = 1, 2, \ldots, \ell$,

  1.1 Pick a topic $k \in \{1, 2, ..., r\}$ with probability given by $\hat{s}$.

  1.2 Pick a word $i \in \{1, 2, ..., n\}$ with probability given by $\hat{A}(:, k)$.

  1.3 Pick a document $j \in \{1, 2, ..., m\}$ with probability given by $\hat{B}(k, :)$.

  1.4 $X(i, j) = X(i, j) + 1$.

5.5.4. 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

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

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 $\frac{1}{\ell}\mathbb{E}\left(\tilde{X}\right)$.

Finally, we have

$$X \approx \ell\hat{A}\,\mathrm{diag}\left(\hat{s}\right)\hat{B}$$

5.5.4. 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 $(\hat{A}\,\mathrm{diag}(\hat{s})\hat{B})_{i,j}$ for PLSA, i.e., a probability mass function given by

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

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

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

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

$$X\approx \ell A\,\mathrm{diag}(s)B=WH.$$

Summary

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] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[3] 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!