Nonnegative Matrix Factorization

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Lee and Seung, 1999

letters to nature

larvae collected randomly in the field (2° 48.12′ N, 41° 40.33′ E) by SCUBA. Between 5 and 10 juveniles were recruited successfully in each of 15, 1 l polystyrene containers ($n = 15$), the bottom of which was covered with an acetate sheet that served as substratum for sponge attachment. Containers were then randomly distributed in 3 groups, and sponges in each group were reared for 14 weeks in 3 different concentrations of $Si(OH)_4$: 0.741 ± 0.133, 30.235 ± 0.287 and 100.041 ± 0.760 μM (mean ± s.e.). All cultures were prepared using 0.22 μm polycarbonate-filtered seawater, which was collected from the sponge habitat, handled according to standard methods to prevent Si contamination[29] and enriched in dissolved silica, when treatments required, by using Na_2SiF_6. During the experiment, all sponges were fed by weekly addition of 2 ml of a bacterial culture ($40–60 \times 10^6$ bacteria ml^{-1}) to each container[30]. The seawater was replaced weekly, with regeneration of initial food and $Si(OH)_4$ levels. The concentration of $Si(OH)_4$ in cultures was determined on 3 replicates of 1 ml seawater samples per container by using a Bran-Luebbe TRAACS 2000 nutrient autoanalyser. After week 5, the accidental contamination of some culture containers by diatoms rendered subsequent estimates of Si uptake by sponges unreliable, so we discarded them for the study.

For the study of the skeleton, sponges were treated according to standard methods[30] and examined in a Hitachi S-2300 scanning electron microscope (SEM).

Received 21 April; accepted 16 August 1999.

1. Hartman, W. D., Wendt, J. W. & Wiedenmayer, F. Living and fossil sponges. Notes for a short course. *Sedimentia* 8, 1–274 (1980).
2. Ghiold, J. The sponges that spanned Europe. *New Scient.* 129, 58–62 (1991).
3. Leinfelder, R. R. Upper Jurassic reef types and controlling factors. *Profil* 5, 1–45 (1993).
4. Wiedenmayer, F. Contributions to the knowledge of post-Paleozoic neritic and archibental sponges

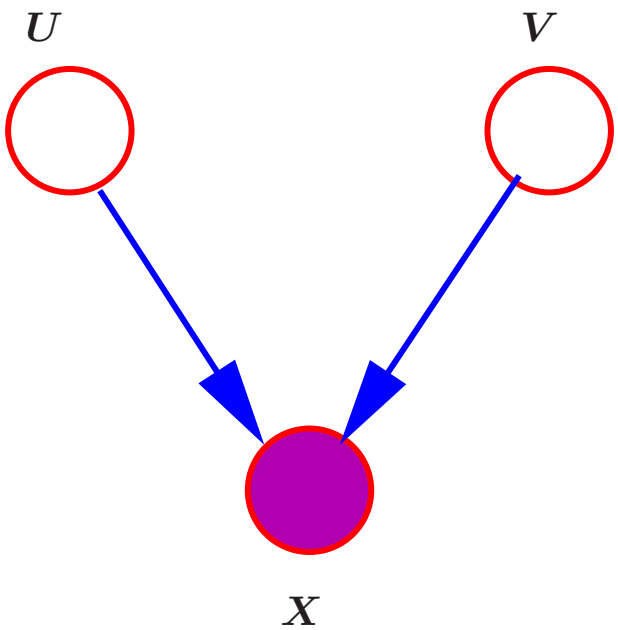
Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* **& H. Sebastian Seung*†**

** Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA*
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Is perception of the whole based on perception of its parts? There is psychological[1] and physiological[2,3] evidence for parts-based representations in the brain, and certain computational theories of object recognition rely on such representations[4,5]. But little is known about how brains or computers might learn the parts of objects. Here we demonstrate an algorithm for non-negative matrix factorization that is able to learn parts of faces and semantic features of text. This is in contrast to other methods, such as principal components analysis and vector quantization,

Nonnegative Matrix Factorization (NMF)



- Given a nonnegative target matrix $\boldsymbol{X} \in \mathbb{R}^{D \times N}$, determine a 2-factor decomposition:

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^{\top},$$

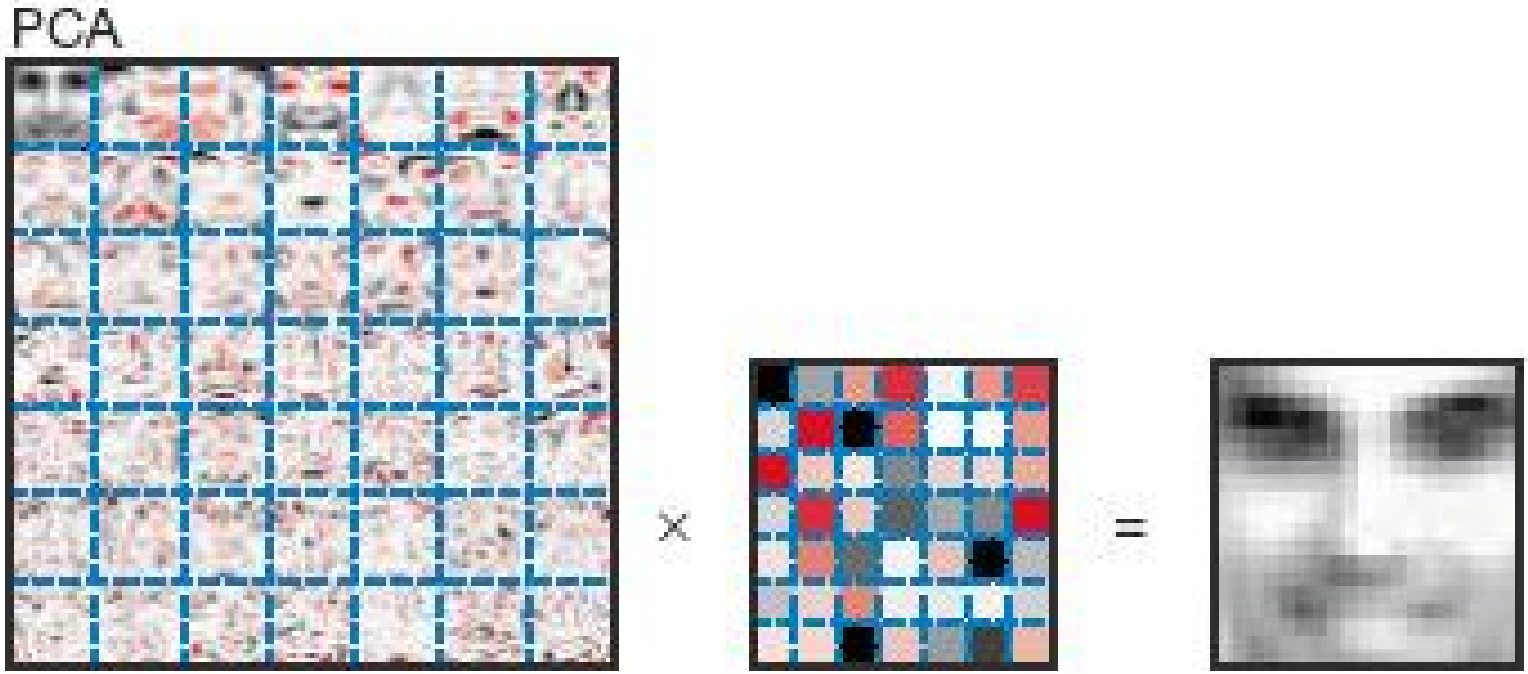
such that factor matrices $\boldsymbol{U} \in \mathbb{R}^{D \times K}$ and $\boldsymbol{V} \in \mathbb{R}^{N \times K}$ are nonnegative as well.

- Low-rank approximation of nonnegative data.
- Involves the optimization:

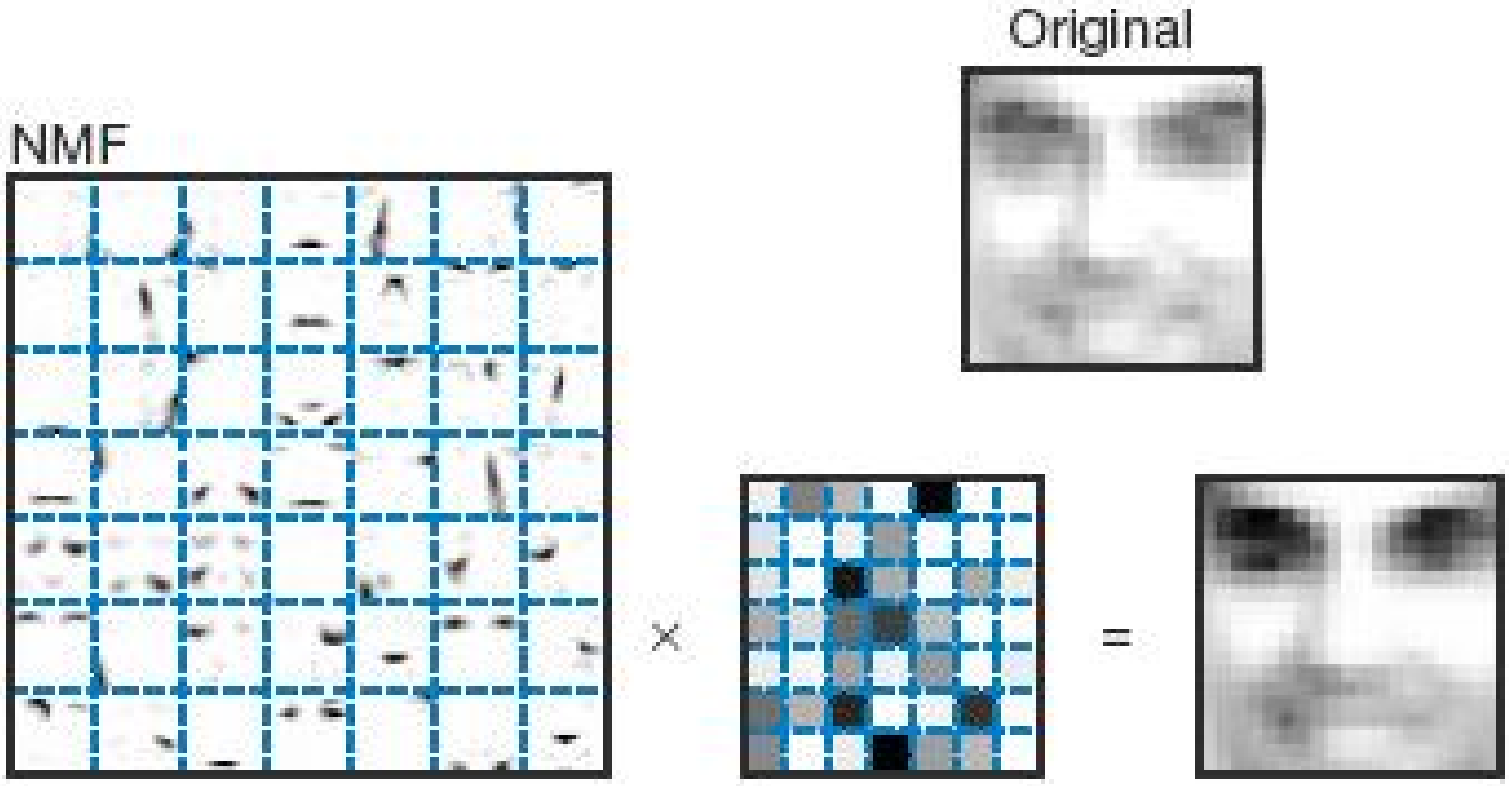
$$\min \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\top}\|^2,$$

subject to $\boldsymbol{U} > 0$ and $\boldsymbol{V} > 0$.

Holistic Representation

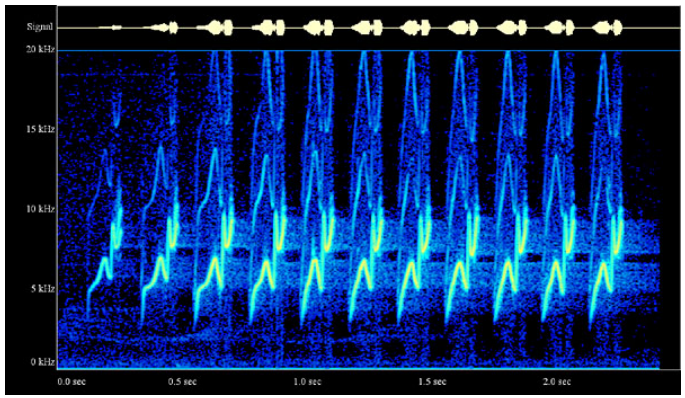


Parts-Based Representation

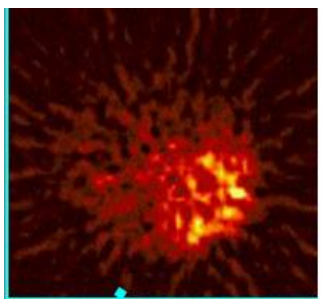


Nonnegative Data

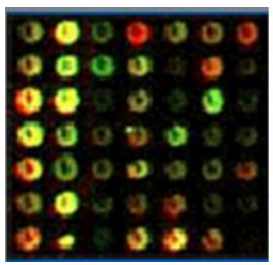
(a) Document



(b) Spectrogram

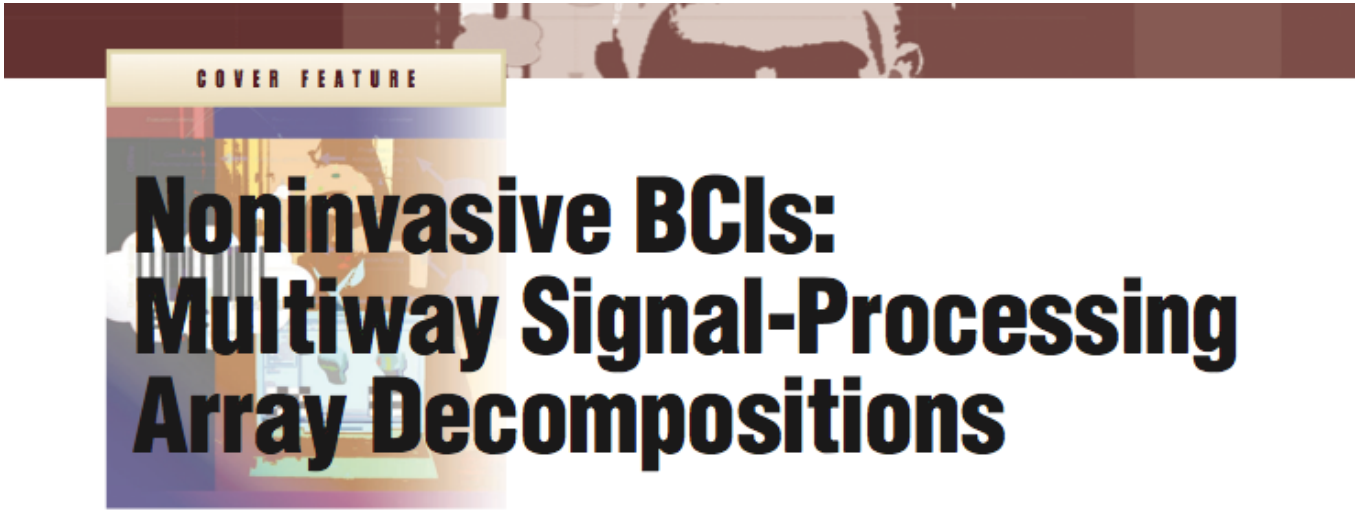


(c) Image



(d) Gene Expression

Matrix Factorization for Brain Computer Interface, 2008



COVER FEATURE

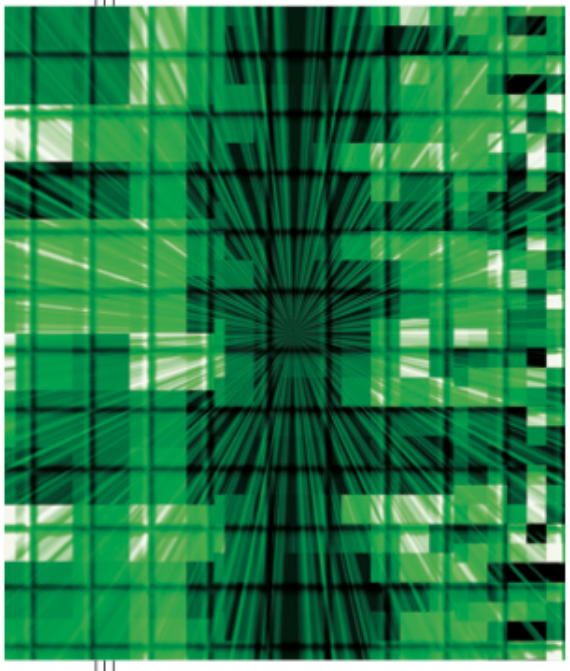
Noninvasive BCIs: Multiway Signal-Processing Array Decompositions

Andrzej Cichocki, Yoshikazu Washizawa, Tomasz Rutkowski, Hovagim Bakardjian, and Anh-Huy Phan, RIKEN Brain Science Institute, Japan
Seungjin Choi and Hyekyoung Lee, Pohang University of Science and Technology, Korea
Qibin Zhao and Liqing Zhang, Shanghai Jiao Tong University, China
Yuanqing Li, South China University of Technology, China

In addition to helping better understand how the human brain works, the brain-computer interface neuroscience paradigm allows researchers to develop a new class of bioengineering control devices and robots, offering promise for rehabilitation and other medical applications as well as exploring possibilities for advanced human-computer interfaces.

Matrix Factorization for Collaborative Filtering, 2009

COVER FEATURE



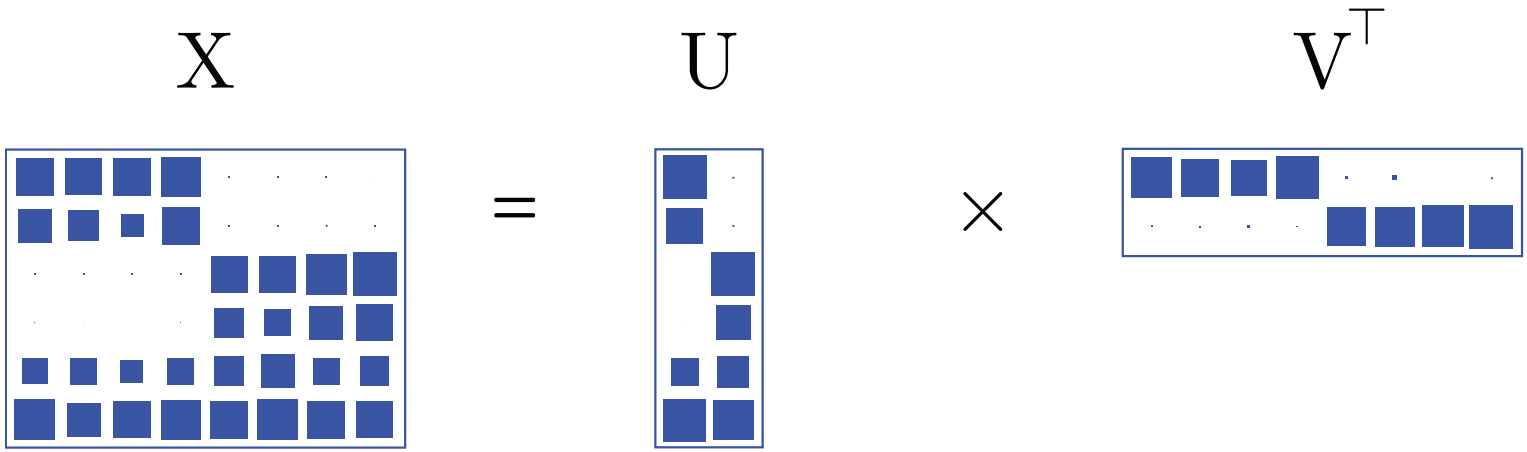
MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS

Yehuda Koren, *Yahoo Research*
Robert Bell and Chris Volinsky, *AT&T Labs—Research*

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest-neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

Such systems are particularly useful for entertainment products such as movies, music, and TV shows. Many customers will view the same movie, and each customer is likely to view numerous different movies. Customers have proven willing to indicate their level of satisfaction with particular movies, so a huge volume of data is available about which movies appeal to which customers. Companies can analyze this data to recommend movies to particular customers.

NMF for Clustering



Nonnegative Matrix Factorization (NMF)

Given a nonnegative matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ (with $X_{ij} \geq 0$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$), NMF seeks a decomposition of $\boldsymbol{X}$ that is of the form:

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^{\top},$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times k}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times k}$ are restricted to be nonnegative matrices as well.

- When columns in $\boldsymbol{X}$ are treated as data points in m-dimensional space, columns in $\boldsymbol{U}$ are considered as basis vectors (or factor loadings) and each row in $\boldsymbol{V}$ is encoding that represents the extent to which each basis vector is used to reconstruct each data vector.
- Alternatively, when rows in $\boldsymbol{X}$ are data points in n-dimensional space, columns in $\boldsymbol{V}$ correspond to basis vectors and each row in $\boldsymbol{U}$ represents encoding.

NMF: Least Squares

The LS error function is given by

$$\begin{aligned} \mathcal{J}_{LS} &= \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top\|^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - [\boldsymbol{U}\boldsymbol{V}^\top]_{ij}\right)^2. \end{aligned}$$

Then, NMF involves the following optimization:

$$\underset{\boldsymbol{U}\geq 0, \boldsymbol{V}\geq 0}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top\|^2.$$

Gradient descent yields

$$\begin{aligned} U_{ij} &\leftarrow U_{ij} + \eta_{ij}^U \left([\boldsymbol{X}\boldsymbol{V}]_{ij} - [\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{V}]_{ij}\right), \\ V_{ij} &\leftarrow V_{ij} + \eta_{ij}^V \left([\boldsymbol{X}^\top\boldsymbol{U}]_{ij} - [\boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{U}]_{ij}\right). \end{aligned}$$

Multiplicative Updates

Gradient descent algorithms do not preserve that elements of $\boldsymbol{U}$ and $\boldsymbol{V}$ are nonnegative.
Choose learning rates

$$\begin{aligned} \eta_{ij}^{U} &= \frac{U_{ij}}{[\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{V}]_{ij}}, \\ \eta_{ij}^{V} &= \frac{V_{ij}}{[\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}]_{ij}}. \end{aligned}$$

Then we have multiplicative updates:

$$\begin{aligned} U_{ij} &\leftarrow U_{ij} \frac{[\boldsymbol{X}\boldsymbol{V}]_{ij}}{[\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{V}]_{ij}}, \\ V_{ij} &\leftarrow V_{ij} \frac{[\boldsymbol{X}^\top \boldsymbol{U}]_{ij}}{[\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}]_{ij}}. \end{aligned}$$

Multiplicative Updates for NMF: Alternative Derivation

- Suppose that the gradient of an error function has a decomposition that is of the form

$$\nabla \mathcal{J} = [\nabla \mathcal{J}]^+ - [\nabla \mathcal{J}]^- ,$$

  where $[\nabla \mathcal{J}]^+ > 0$ and $[\nabla \mathcal{J}]^- > 0$.
- Then multiplicative update for parameters Θ has the form

$$\boxed{\Theta \leftarrow \Theta \odot \left(\frac{[\nabla \mathcal{J}]^-}{[\nabla \mathcal{J}]^+} \right)^{.\eta}.}$$

- Compute derivatives:

$$\begin{aligned} \nabla_U \mathcal{J} &= [\nabla_U \mathcal{J}]^+ - [\nabla_U \mathcal{J}]^- = \boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{V} - \boldsymbol{X}\boldsymbol{V}, \\ \nabla_V \mathcal{J} &= [\nabla_V \mathcal{J}]^+ - [\nabla_V \mathcal{J}]^- = \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{X}^\top \boldsymbol{U}. \end{aligned}$$

- Choosing $\eta = 1$ yields

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \odot \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{V}}, \qquad \boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^\top \boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}}.$$

Term-Document Matrix

A term-document matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is a collection of vector space representations of documents, where rows are terms (words) and columns are documents

$$X_{ij} = t_{ij} \log \left(\frac{n}{idf_i} \right),$$

where t_{ij} is the term frequency of word i in document j and idf_i is the number of documents containing word i.

Clustering by Factorization

NMF yields a factorization $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$:

- U_{ij}: the degree to which term i belongs to cluster j.
- V_{ij}: the degree document i is associated with cluster j.

Document clustering is based on column vectors of $\boldsymbol{V}^\top$.

Assign document i to cluster j^* if

$$j^* = \arg\max_j V_{ij}.$$

Document Clustering by NMF

1. Construct a term-document matrix $\boldsymbol{X}$.
2. Perform NMF on $\boldsymbol{X}$, yielding $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$.
3. Normalization

$$U_{ij} \leftarrow \frac{U_{ij}}{\sqrt{\sum_i U_{ij}^2}}, \quad V_{ij} \leftarrow V_{ij} \left(\sqrt{\sum_i U_{ij}^2} \right).$$

4. Use $\boldsymbol{V}$ to determine the cluster label for each document. Assign document d_i to cluster j^* if

$$j^* = \arg\max_j V_{ij}.$$

k-means: Matrix Factorization Perspective

Recall the objective function for k-means clustering

$$\mathcal{J} = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{K} V_{ij} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2,$$

where

$$V_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases}$$

This can be written as

$$\mathcal{J} = \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top\|^2,$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and $\boldsymbol{U} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]$ contains centers (prototype vectors) in columns and $\boldsymbol{V}$ is the indicator matrix.

NMF: I-Divergence

- Consider the *I*-divergence between the data and the model:

$$\mathcal{J}_I = \sum_i \sum_j \left\{ X_{ij} \log \frac{X_{ij}}{\left[\boldsymbol{U}\boldsymbol{V}^\top\right]_{ij}} - X_{ij} + \left[\boldsymbol{U}\boldsymbol{V}^\top\right]_{ij} \right\}.$$

  Note that I-divergence is identical to Kullback-Leibler divergence when $\sum_i \sum_j X_{ij} = \sum_i \sum_j \left[\boldsymbol{U}\boldsymbol{V}^\top\right]_{ij} = 1$.

- Multiplicative updates for $\boldsymbol{U}$ and $\boldsymbol{V}$ are determined by minimizing $\mathcal{J}_I$ with nonnegativity constraints $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$ satisfied:

$$\begin{aligned} U_{ij} &\leftarrow U_{ij} \frac{\sum_k (X_{ij}/[\boldsymbol{U}\boldsymbol{V}^\top]_{ik}) V_{kj}}{\sum_k V_{kj}}, \\ V_{ij} &\leftarrow V_{ij} \frac{\sum_k (X_{ki}/[\boldsymbol{U}\boldsymbol{V}^\top]_{ki}) U_{kj}}{\sum_k U_{kj}}. \end{aligned}$$

- Equivalence between NMF and PLSA was shown by Gaussier and Goutte (SIGIR-2005).

Weighted NMF

- In practice, the data matrix is often incomplete, i.e., some of entries are missing or unobserved.
- In order to handle missing entries in the decomposition, we consider an objective function that is a sum of weighted residuals:

$$\mathcal{J} = \sum_{i,j} W_{ij}(X_{ij} - [\boldsymbol{U}\boldsymbol{V}^\top]_{ij})^2 = \|\boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top)\|^2,$$

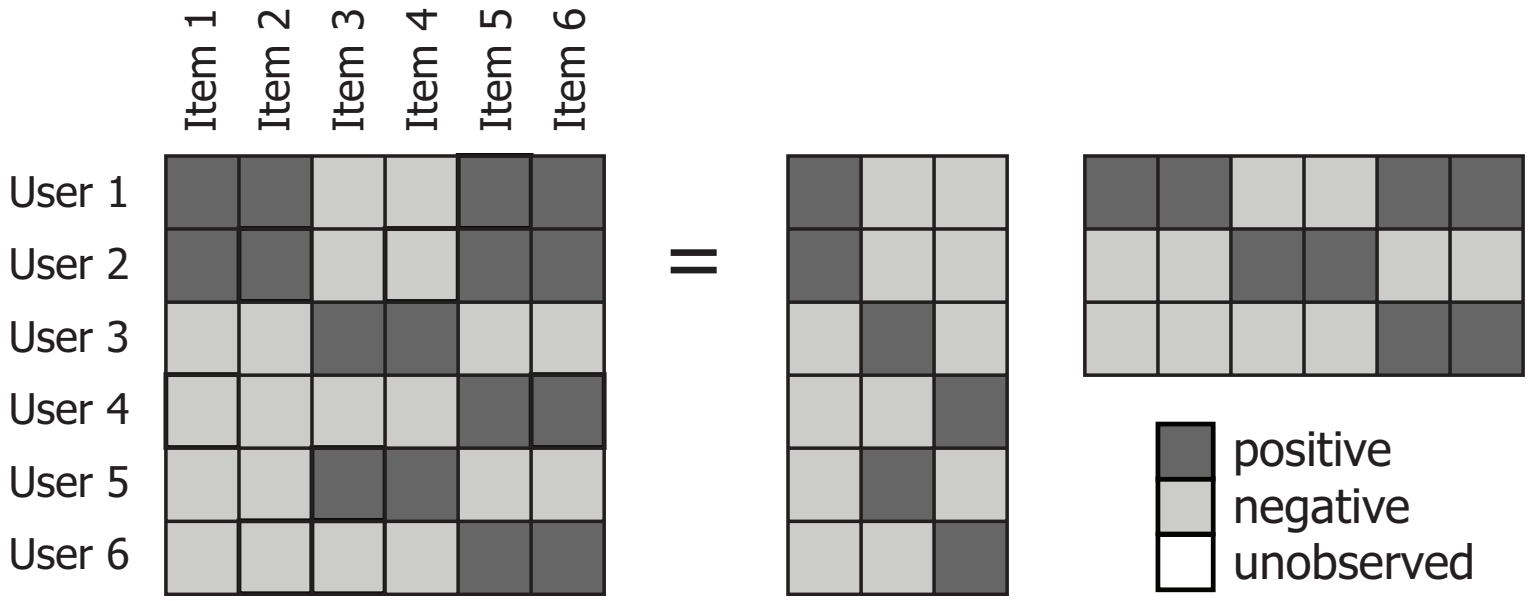
  where W_{ij} are binary weights, i.e.,

$$W_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{ is observed} \\ 0 & \text{if } X_{ij} \text{ is unobserved.} \end{cases}$$

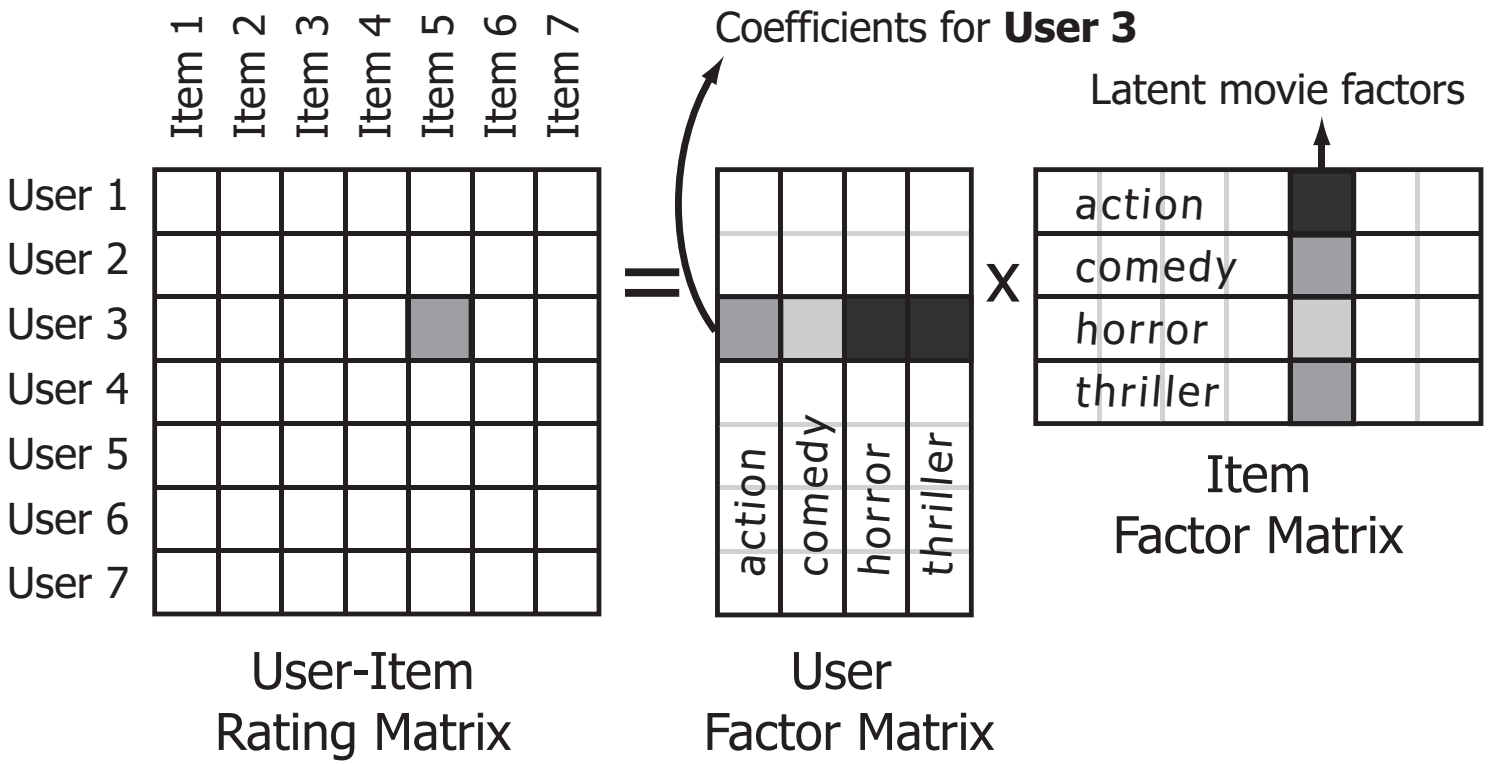
- Multiplicative updates for WNMF are given as:

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \odot \frac{[\boldsymbol{W} \odot \boldsymbol{X}]\boldsymbol{V}}{[\boldsymbol{W} \odot \boldsymbol{U}\boldsymbol{V}^\top]\boldsymbol{V}}, \qquad \boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{[\boldsymbol{W} \odot \boldsymbol{X}]^\top \boldsymbol{U}}{[\boldsymbol{W} \odot \boldsymbol{V}\boldsymbol{U}^\top]\boldsymbol{U}}.$$

Weighted NMF for Collaborative Filtering



Weighted NMF for Collaborative Filtering



$$X_{ij} \approx \boldsymbol{u}_i^\top \boldsymbol{v}_j$$