THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Week 6 Tutorial:

Dictionary Learning and Non-negative Matrix Factorisation

Anjin Liu

anjin.liu@sydney.edu.au

Tutorial Contents

- Review (20min):
  - Lecture 4: Dictionary Learning and Non-negative Matrix Factorisation
- Tutorial exercise & QA (40min):

Key points

- Dictionary Learning
- Non-negative Matrix Factorisation

Announcements

- Assignment 1 is online now
  - Assignment 1 due on 9/10/2025, 11:59pm
  - Group-based (3-4 students per group). Find you teammates by yourselves.
  - Put your team member names in the report

Dictionary learning

What is a dictionary in machine learning?

A dictionary is a collection of words in one specific languages.

Can we find some common "words" (elements) to express data?

Dictionary Learning

Step 1. Data with Labels

Training Samples:

"Stocks fell as interest rates rose in the US." → **Finance**
"The central bank plans to increase interest rates again." → **Finance**
"The team won the championship after a thrilling final." → **Sports**
"The coach praised the players for their defense." → **Sports**
"New smartphone released with advanced AI camera features." → **Technology**
"Tech companies compete to release faster chips." → **Technology**

Testing Samples:
"Investors are worried about inflation and market volatility." → ???
"Oil prices climbed after new trade restrictions." → ???
"The new season starts next month with tough rivalries." → ???

Dictionary Learning

THE UNIVERSITY OF SYDNEY

Step 2. Dictionary Learning Outcome

The algorithm learns latent "atoms" that roughly align with our labels:

- **Atom 1 (Finance)** ≈ words about **stocks, interest rates, bank, investors**
- **Atom 2 (Sports)** ≈ words about **team, coach, championship**
- **Atom 3 (Technology)** ≈ words about **smartphone, AI, chips, tech**

Dictionary Learning

Step 3. Sparse Representation of New Data

When a new article comes in, dictionary learning **represents it as a mixture of atoms** rather than a single cluster.

Example:
*"**Investors** worry as <u>**tech**</u> **stocks** plunge after poor earnings reports with the new <u>**AI chips**</u>."*
= 60% **Finance Atom** + 40% <u>**Technology Atom**</u> → **Finance-Tech hybrid label**

Dictionary Learning

house

plane

$$
\begin{array}{c|cccc}
 & \text{Small house} & \text{big house} & \text{Small plane} & \text{big plane} \\
\hline
\text{Pack A} & & & & \\
\text{Pack B} & & \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix} & & \\
\text{Pack C} & & & & \\
\text{Pack D} & & & &
\end{array}
$$

Column	Model	Interpretation (feature counts)
1	Small house	1 pack A, 1 pack B, 1 pack C, 1 pack D
2	Big house	2 pack A, 2 pack B, 2 pack C, 2 pack D
3	Small airplane	2 pack A, 1 pack B, 2 pack C, 1 pack D
4	Big airplane	4 pack A, 2 pack B, 4 pack C, 2 pack D

Dictionary learning

What is a dictionary in machine learning?

$$\begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 2 & 2 & 4\\1 & 2 & 1 & 2\\1 & 2 & 2 & 4\\1 & 2 & 1 & 2\end{bmatrix}=\begin{bmatrix}1 & 2\\1 & 1\\1 & 2\\1 & 1\end{bmatrix}\begin{bmatrix}1 & 2 & 0 & 0\\0 & 0 & 1 & 2\end{bmatrix}$$

Lego Set 1={A,B,C,D}

Lego Set 2={A,A,B,C,C,D}

$$\begin{bmatrix}1 & 2\\1 & 1\\1 & 2\\1 & 1\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix}=a\times\begin{bmatrix}1\\1\\1\\1\end{bmatrix}+b\times\begin{bmatrix}2\\1\\2\\1\end{bmatrix}$$

Dictionary learning

What is a dictionary in machine learning?

Let $x \in \mathbb{R}^d$, $D \in \mathbb{R}^{d \times k}$

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$

Note that $\|x\| = \sqrt{x^\top x}$ is the ell 2 norm.

Given $x_1, \ldots, x_n \in \mathbb{R}^d$

$$\{D^*, \alpha_1^*, \ldots, \alpha_n^*\} = \arg \min_{D \in \mathbb{R}^{d \times k}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \|x_i - D\alpha_i\|^2.$$

$$D = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix}$$

What is the reconstruction error of x_1 $\|x_1 - Da_1\|^2$?

$$D = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix}$$

What is the reconstruction error of x_1 $\|x_1 - Da_1\|^2$?

$$\sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2} = 0$$

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

What is the reconstruction error of x_1 $\|x_1 - Da_1\|^2$?

$$D = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix}$$

What is the reconstruction error of x_1 $\|x_1 - Da_1\|^2$?

$$\sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-2)^2} = 1$$

Dictionary learning

Note that

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i - D\alpha_i\|^2 = \frac{1}{n}\|X - DR\|_F^2,$$

where $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$,

$R = [\alpha_1, \alpha_2, \ldots, \alpha_n] \in \mathbb{R}^{k\times n}$,

$\|X\|_F = \sqrt{\mathrm{trace}(X^\top X)} = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{n} X_{i,j}^2}$ is the Frobenius norm of X.

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$E = X - DA = 0$$

$$\|E\|_F = 0$$

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix} \qquad D = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 2 & 2 & 4 \end{bmatrix}$$

$$X' = DA = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 2 & 4 \end{bmatrix}$$

$$E = X - X' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -2 \end{bmatrix}$$

$$\|E\|_F = \sqrt{0^2 + \cdots + (-2)^2 + (-1)^2 + \dots} = \sqrt{10} \approx 3.16$$

$$\|X\|_F \approx 8.37$$

Dictionary learning

Note that

$$\arg \min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2,$$

where $\mathcal{D}$ and $\mathcal{R}$ are some specific domains for D and R.

Optimisation

Objective:

$$\min_{D\in\mathcal{D},R\in\mathcal{R}} \|X - DR\|_F^2$$

The objective is convex with respect to either R or D but not to both.

Fix R, solve for D

$$\min_{D\in\mathcal{D}} \|X - DR\|_F^2$$

Fix D, solve for R

$$\min_{R\in\mathcal{R}} \|X - DR\|_F^2$$

Engan, Kjersti, Sven Ole Aase, and J. Hakon Husoy. "Method of optimal directions for frame design." Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on. Vol. 5. IEEE, 1999.

Optimisation

Objective:

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

The objective is convex with respect to either D or R but not to both.

Suppose D^* and R^* are the local minimisers for the objective, we have

$$X \approxeq D^* R^* = (D^* A)(A^{-1} R^*).$$

Normalisation (optional):

$$D_{:,i} \leftarrow D_{:,i} / \|D_{:,i}\|$$

Scaling Ambiguity

$$Q = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}, \qquad Q^{-1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}.$$

$$D' = DQ, \qquad A' = Q^{-1}A.$$

$$D' = \begin{bmatrix} 2 & 1 \\ 2 & 0.5 \\ 2 & 1 \\ 2 & 0.5 \end{bmatrix}, \qquad A' = \begin{bmatrix} 0.5 & 1 & 0 & 0 \\ 0 & 0 & 2 & 4 \end{bmatrix}.$$

Scaling Ambiguity

$$X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{bmatrix}$$

Solution 1

$$D = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

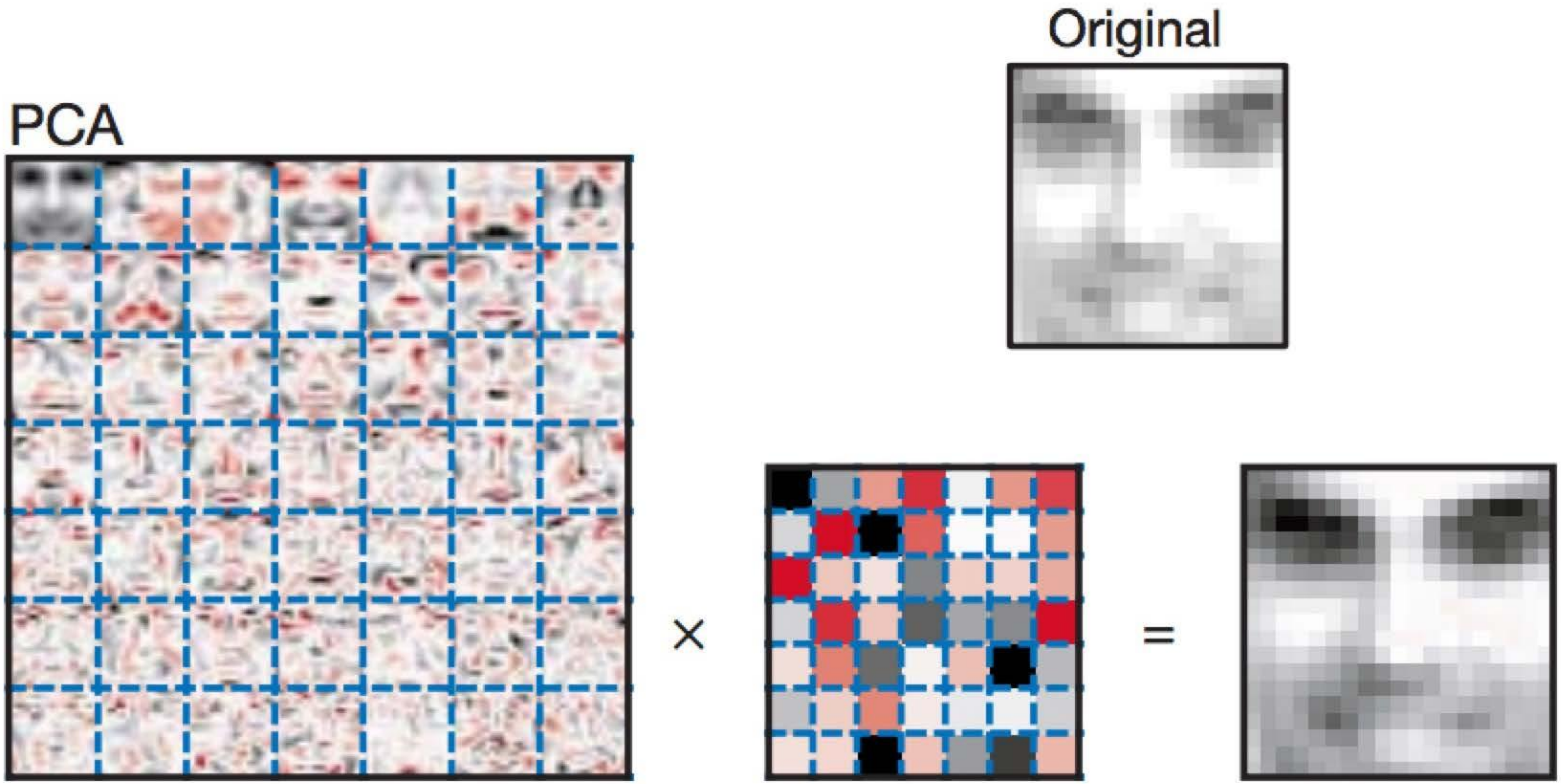
Solution 2

$$D' = \begin{bmatrix} 2 & 1 \\ 2 & 0.5 \\ 2 & 1 \\ 2 & 0.5 \end{bmatrix}, \qquad A' = \begin{bmatrix} 0.5 & 1 & 0 & 0 \\ 0 & 0 & 2 & 4 \end{bmatrix}$$

Dictionary learning

PCA: $A = U\Lambda U^T$

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$

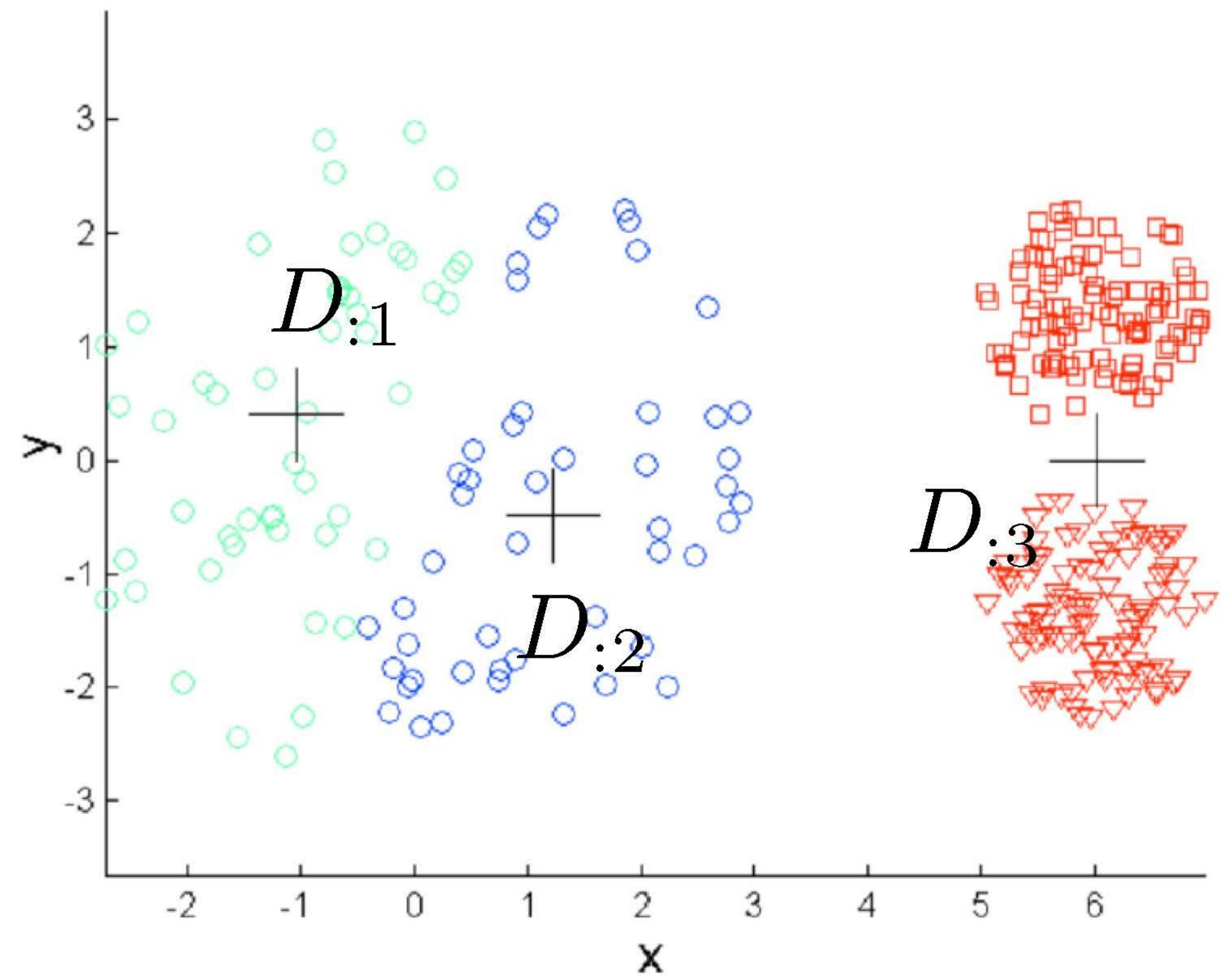


Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

Dictionary learning

K-means clustering:

$$\min_{D\in\mathcal{D},R\in\mathcal{R}} \|X-DR\|_F^2$$

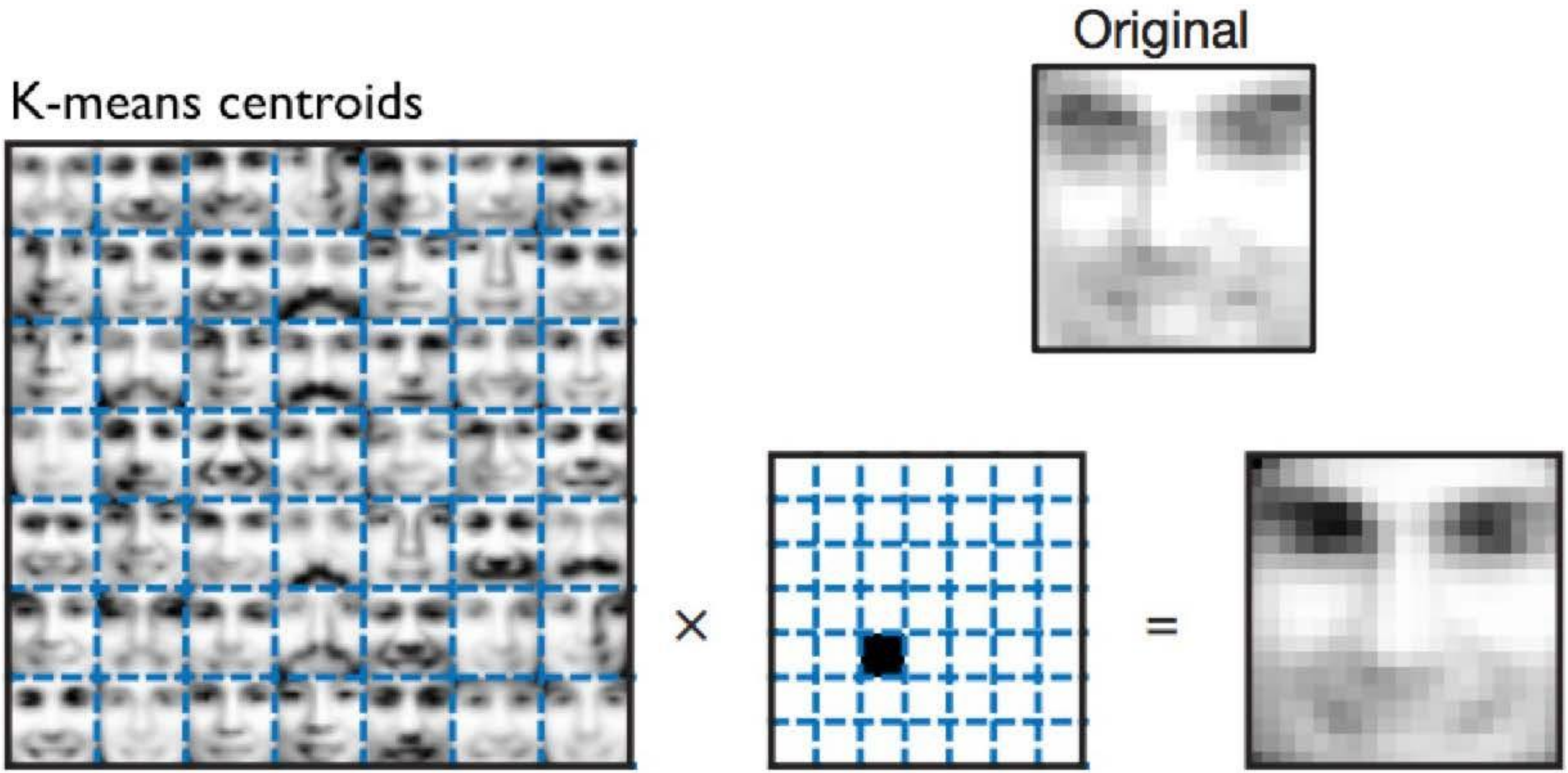


Special requirement: each column of R only one have entry equals to one, the other entries are all zeros.

Dictionary learning

K-means clustering:

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$



Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

Dictionary Matrix or Code Matrix
may have negative values
which may not to explain in real-world applications
such as *use Lego pack N but remove pack M*

$$DA = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 & 8 \\ 0 & 0 & 0 & 0 \\ 2 & 4 & 6 & 8 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Non-negative matrix factorisation

- **Why non-negativity of data?**

Data is often nonnegative by nature
Image intensities
Movie ratings
Document-term counts
Microarray data
Stock market values

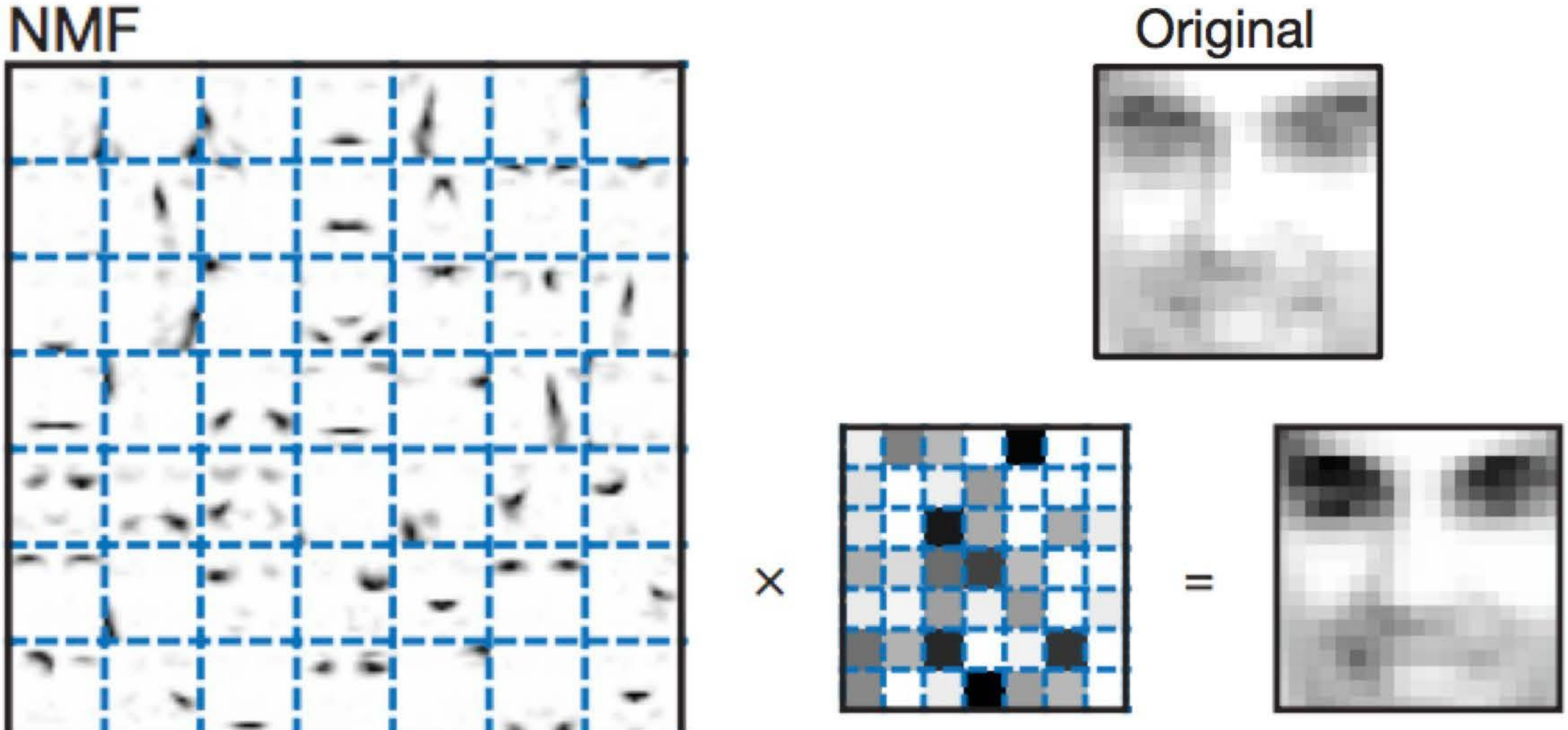
Non-negative matrix factorisation

$$\min_{D\in\mathcal{D},R\in\mathcal{R}} \|X - DR\|_F^2$$

Special requirement: $\mathcal{D} = \mathbb{R}_+^{d\times k}, \quad \mathcal{R} = \mathbb{R}_+^{k\times n}.$

Non-negative matrix factorisation

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$



Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

NMF optimisation

MUR (Multiplicative Update Rules):

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

Fix D, solve for R

$$\frac{\partial \|X - DR\|_F^2}{\partial R} = -2D^\top X + 2D^\top DR$$

The Matrix Cookbook: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
Online helping tool: http://www.matrixcalculus.org/

$$\|X - DR\|_F^2 = \mathrm{trace}((X - DR)^\top (X - DR))$$

Dictionary Learning

- Dictionary Learning
- Non-negative Matrix Factorisation