

THE UNIVERSITY OF
SYDNEY

Advanced Machine Learning

(COMP 5328)

**Dictionary Learning and Non-negative Matrix
Factorisation**

Yu Yao



THE UNIVERSITY OF
SYDNEY

Announcements

- Assignment I is online now
 - Assignment I due on 9/10/2025, 11:59pm
 - Group-based (3-4 students per group). Find your teammates by yourselves.



Assignment I

Summary

The objective of this assignment is to implement Non-negative Matrix Factorisation (NMF) algorithms and analyse the robustness of NMF algorithms when the dataset is contaminated by large magnitude noise or corruption. You should implement at least two NMF algorithms and compare their robustness.



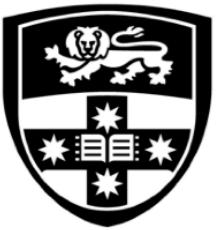
Assignment I

Data

ORL dataset: it contains 400 images of 40 distinct subjects. All images are cropped and resized to 92×112 pixels.

Extended YaleB dataset: it contains 2414 images of 38 subjects. All images are manually aligned, cropped, and then resized to 168×192 pixels.





THE UNIVERSITY OF
SYDNEY

Review

PAC learning framework

Definition:

A hypothesis class H is said to be PAC (probably approximately correct)-learnable if there exists a learning algorithm \mathcal{A} and a polynomial function $poly(\cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distribution D on $X \times Y$, the following holds for any sample of size $n > poly(1/\delta, 1/\epsilon)$ and the hypothesis h_S learned by \mathcal{A} :

$$p \left\{ R(h_S) - \min_{h \in H} R(h) \leq \epsilon \right\} \geq 1 - \delta.$$

PAC learning framework

If a hypothesis class is not learnable, it means that we cannot learn the best classifier (or classifiers close to the best) by designing a learning algorithm using a training sample with size larger than a polynomial function, e.g., $n > \text{poly}(1/\delta, 1/\epsilon)$

We need an exponentially large training sample size to learn the best classifier (or classifiers close to the best). It could be impossible to find a such a large training sample or a classical computer can handle it.

PAC-learnable checking: Empirical risk minimisation

ERM algorithm:

$$h_S = \arg \min_{h \in H} R_S(h) = \arg \min_{h \in H} \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i, h).$$

Check if the following hold when $n > \text{poly}(1/\delta, 1/\epsilon)$

$$p \left\{ R(h_S) - \min_{h \in H} R(h) \leq \epsilon \right\} \geq 1 - \delta.$$

Generalisation error

We have $R_S(h_S) \leq R_S(h^*)$, $h^* = \arg \min_{h \in H} R(h)$ then

$$\begin{aligned} & R(h_S) - \min_{h \in H} R(h) = R(h_S) - R(h^*) \\ &= R(h_S) - R_S(h_S) + R_S(h_S) - R_S(h^*) + R_S(h^*) - R(h^*) \\ &\leq R(h_S) - R_S(h_S) + R_S(h^*) - R(h^*) \\ &\leq |R(h_S) - R_S(h_S)| + |R(h^*) - R_S(h^*)| \\ &\leq \sup_{h \in H} |R(h) - R_S(h)| + \sup_{h \in H} |R(h) - R_S(h)| \\ &= 2 \sup_{h \in H} |R(h) - R_S(h)|. \end{aligned}$$

Hypothesis complexity

If A implies B , then $p\{A\} \leq p\{B\}$

$$p \left\{ \sup_{h \in H} |R(h) - R_S(h)| \geq \epsilon \right\}$$

Union bound

$$p \left\{ \bigcup_{i=1}^n A_i \right\} \leq \sum_{I=1}^n p\{A_i\}.$$

$$\leq p \left\{ \bigcup_{h \in H} |R(h) - R_S(h)| \geq \epsilon \right\}$$

$$\leq \sum_{h \in H} p \left\{ |R(h) - R_S(h)| \geq \epsilon \right\}$$

$$p \left\{ |R(h) - R_S(h)| \geq \epsilon \right\} \leq 2 \exp \left(\frac{-2n\epsilon^2}{M^2} \right). \leq 2|H| \exp \left(\frac{-2n\epsilon^2}{M^2} \right).$$

PAC learning checking

If the hypothesis class is of finite hypotheses, it is PAC learnable. Because

$$p \left\{ R(h_S) - \min_{h \in H} R(h) \leq 2 \sup_{h \in H} |R(h) - R_S(h)| \leq 2M \sqrt{\frac{\log |H| + \log(2/\delta)}{2n}} \right\} \geq 1 - \delta.$$

Since $\delta = 2|H| \exp\left(\frac{-2n\epsilon^2}{M^2}\right)$. We have

$$n = \frac{M^2}{\epsilon^2} \log\left(\frac{2|H|}{\delta}\right). \quad n > \text{poly}(1/\delta, 1/\epsilon)$$

PAC learning framework

Definition:

A hypothesis class H is said to be PAC (probably approximately correct)-learnable if there exists a learning algorithm \mathcal{A} and a polynomial function $poly(\cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distribution D on $X \times Y$, the following holds for any sample of size $n > poly(1/\delta, 1/\epsilon)$ and the hypothesis h_S learned by \mathcal{A} :

$$p \left\{ R(h_S) - \min_{h \in H} R(h) \leq \epsilon \right\} \geq 1 - \delta.$$

VC dimension

$$\begin{aligned} & p \left\{ \sup_{h \in H} |R_S(h) - R(h)| \geq \epsilon \right\} \\ & \leq 2p \left\{ \sup_{h \in H} |R_S(h) - R_{S'}(h)| \geq \epsilon/2 \right\} \\ & \leq 4p \left\{ \sup_{h \in H} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \ell(X_i, Y_i, h) \right| \geq \epsilon/4 \right\} \\ & \leq 4p \left\{ \cup_{h \in H'} \frac{1}{n} \left| \sum_{I=1}^n \sigma_i \ell(X_i, Y_i, h) \right| \geq \epsilon/4 \right\} \\ & \leq 4\Pi_H(n)p \left\{ \frac{1}{n} \left| \sum_{I=1}^n \sigma_i \ell(X_i, Y_i, h) \right| \geq \epsilon/4 \right\} \\ & \leq 8\Pi_H(n) \exp(-n\epsilon^2/32M^2). \end{aligned}$$

VC dimension

Let H be a hypothesis set with $\text{VC dimension}(H) = d$ then for all $n \geq d$

$$\Pi_H(n) \leq \left(\frac{en}{d}\right)^d.$$

The proof is in Chapter 3 of the book “Foundations of ML”

VC dimension

Definition:

VC dimension

The VC dimension of a hypothesis class H is the size of the largest set that can be fully shattered by H :

$$\text{VC dimension}(H) = \max_n \{n : \Pi_H(n) = 2^n\}.$$

Note that VC dimension is designed for binary classification problem. Why?

VC dimension

Definition:

Shattering

The data points $\{X_1, \dots, X_n\}$ is said to be shattered by a hypothesis class H when H realises all possible binary predictions. That is $\Pi_H(n) = 2^n$.

PAC learning checking

If the hypothesis class is of finite VC dimension, it is PAC learnable. Because

$$p \left\{ R(h_S) - \min_{h \in H} R(h) \leq 2 \sup_{h \in H} |R(h) - R_S(h)| \leq 2M \sqrt{\frac{32(d \log(en/d) + \log(8/\delta))}{n}} \right\} \geq 1 - \delta.$$

Since $\delta = 8 \left(\frac{en}{d}\right)^d \exp(-n\epsilon^2/32M^2)$, we have

$$n = \frac{32M^2}{\epsilon^2} (d \log(en/d) + \log(8/\delta)).$$

$$n > \text{poly}(1/\delta, 1/\epsilon)$$

Generalisation error

We have proven that

$$R(h_S) \leq R(h^*) + 2 \sup_{h \in H} |R(h) - R_S(h)|.$$

Some discussion on the differences between training error and generalisation error.

Generalisation bound

Why the bound is useful?

To prove PAC learnable.

To analyse the test error.

To improve learning algorithm, e.g., if we can prove

$$R(h_S) \leq R(h^*) + 2 \sup_{h \in H} |R(h) - R_S(h)|$$

$$\triangleq UB(h_S).$$

We could propose a new algorithm,

$$h_S = \arg \min_{h \in H} \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i, h) + UB(h),$$

which is call the regularised ERM algorithm.

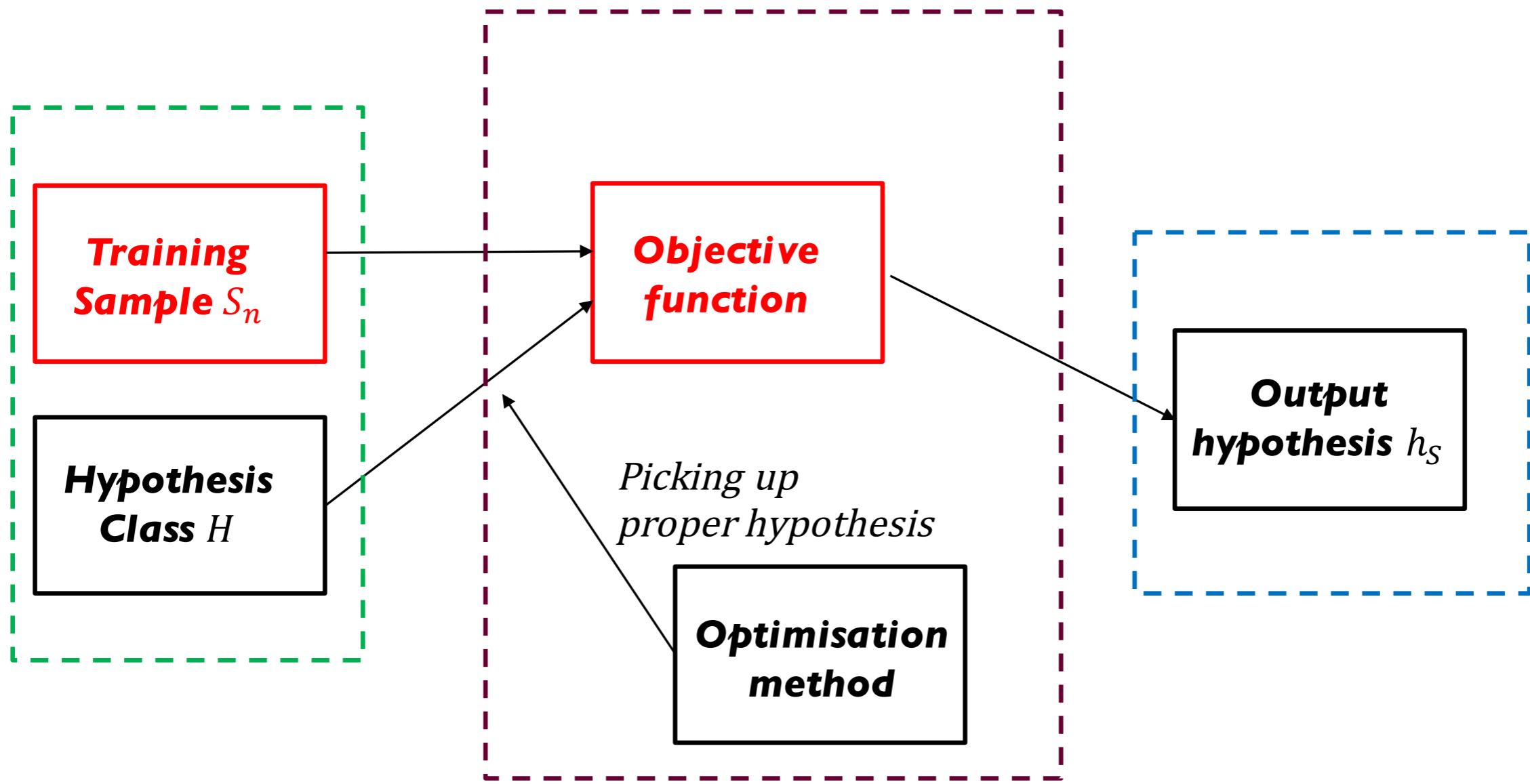
PAC learning framework

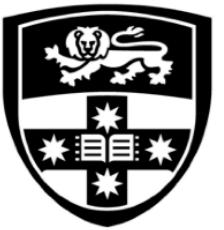
For a given learning algorithm, the PAC learning framework can be used to calculate how many examples we will need such that for a given confidence level, e.g., with probability at least $1 - \delta$, and a given approximation error ϵ , the learned hypothesis h_S satisfies $R(h_S) - R(h^*) \leq \epsilon$.

Note that the claim holds for all distributions of the training data, including the worst case. This means the bound is loose for some applications. We may need a smaller training sample size.



Machine Learning Algorithms





THE UNIVERSITY OF
SYDNEY

Dictionary learning

Dictionary learning

What is a dictionary in machine learning?

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

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

Dictionary learning

What is a dictionary in machine learning?

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

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

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

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

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

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

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

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.

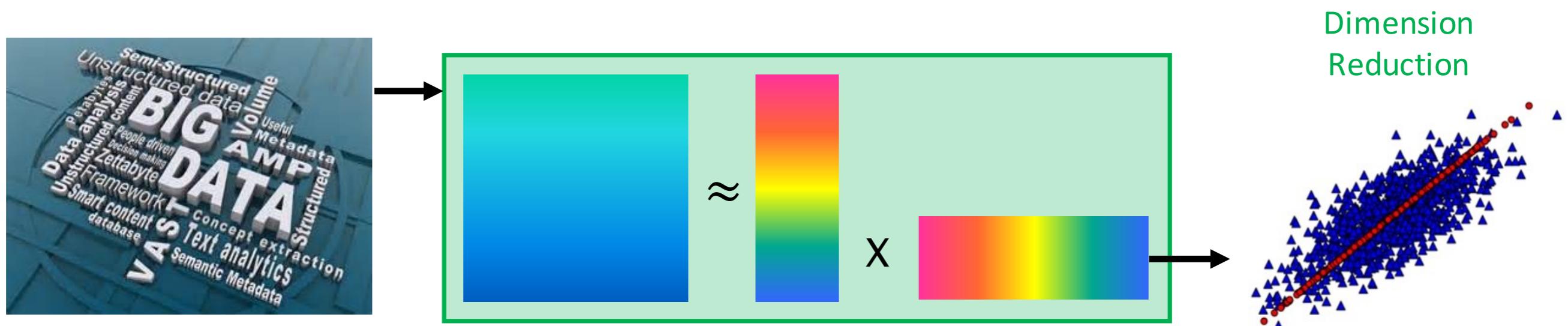
Suppose D^* and R^* are the local minimisers for the objective, we have

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

Normalisation (optional):

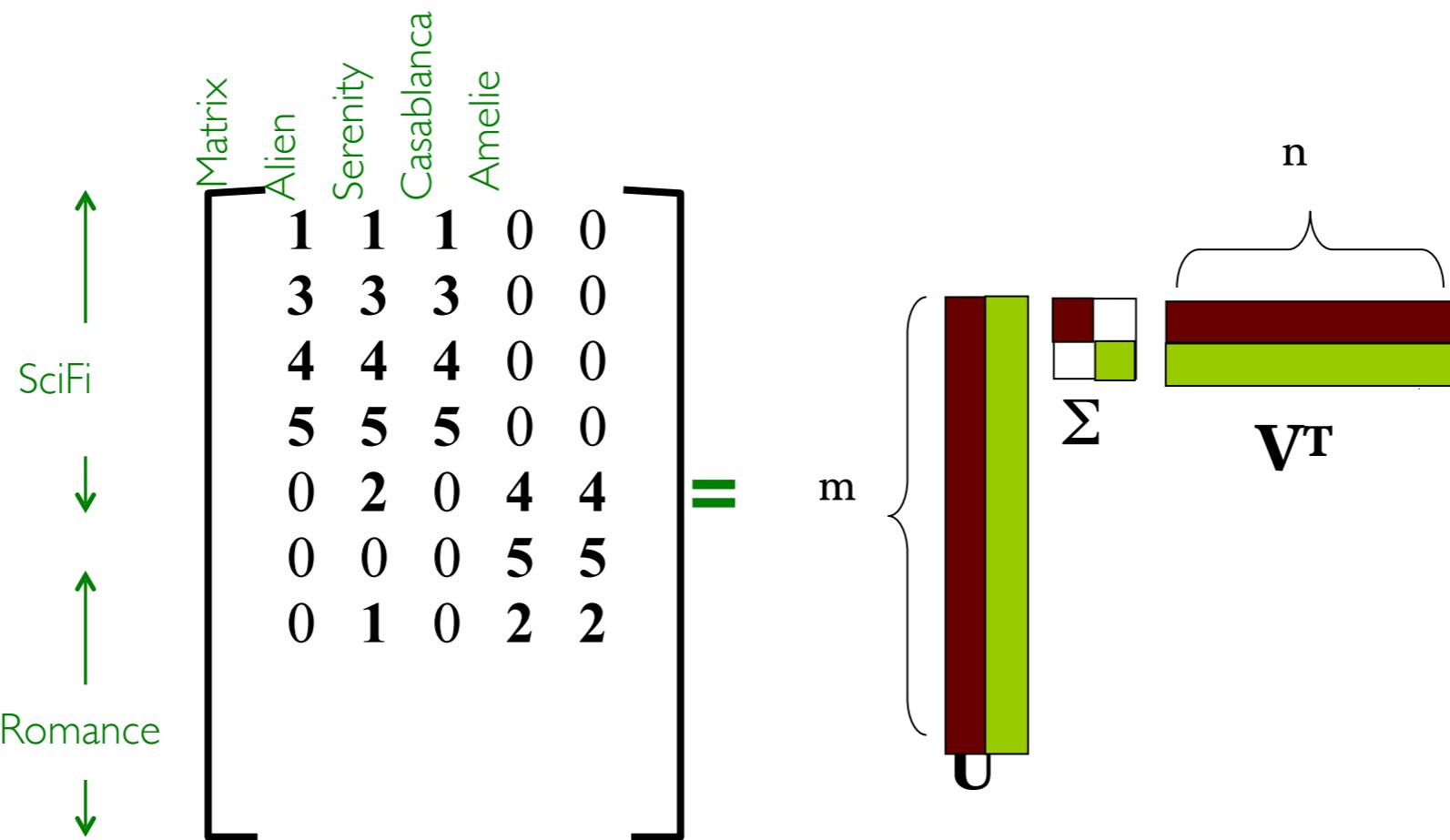
$$D_{:,i} \leftarrow D_{:,i} / \|D_{:,i}\|$$

Dictionary learning application



Dictionary learning

$$\text{SVD: } X = U\Sigma V^T$$



Dictionary learning

$$\text{PCA: } A = U\Lambda U^T$$

$$A = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_k \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{bmatrix} \begin{bmatrix} | & \mathbf{u}_1^T & | \\ | & \mathbf{u}_2^T & | \\ \vdots & & \vdots \\ | & \mathbf{u}_k^T & | \end{bmatrix}$$

D R

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

Special requirement: columns of D are orthonormal to each other.

Dictionary learning application



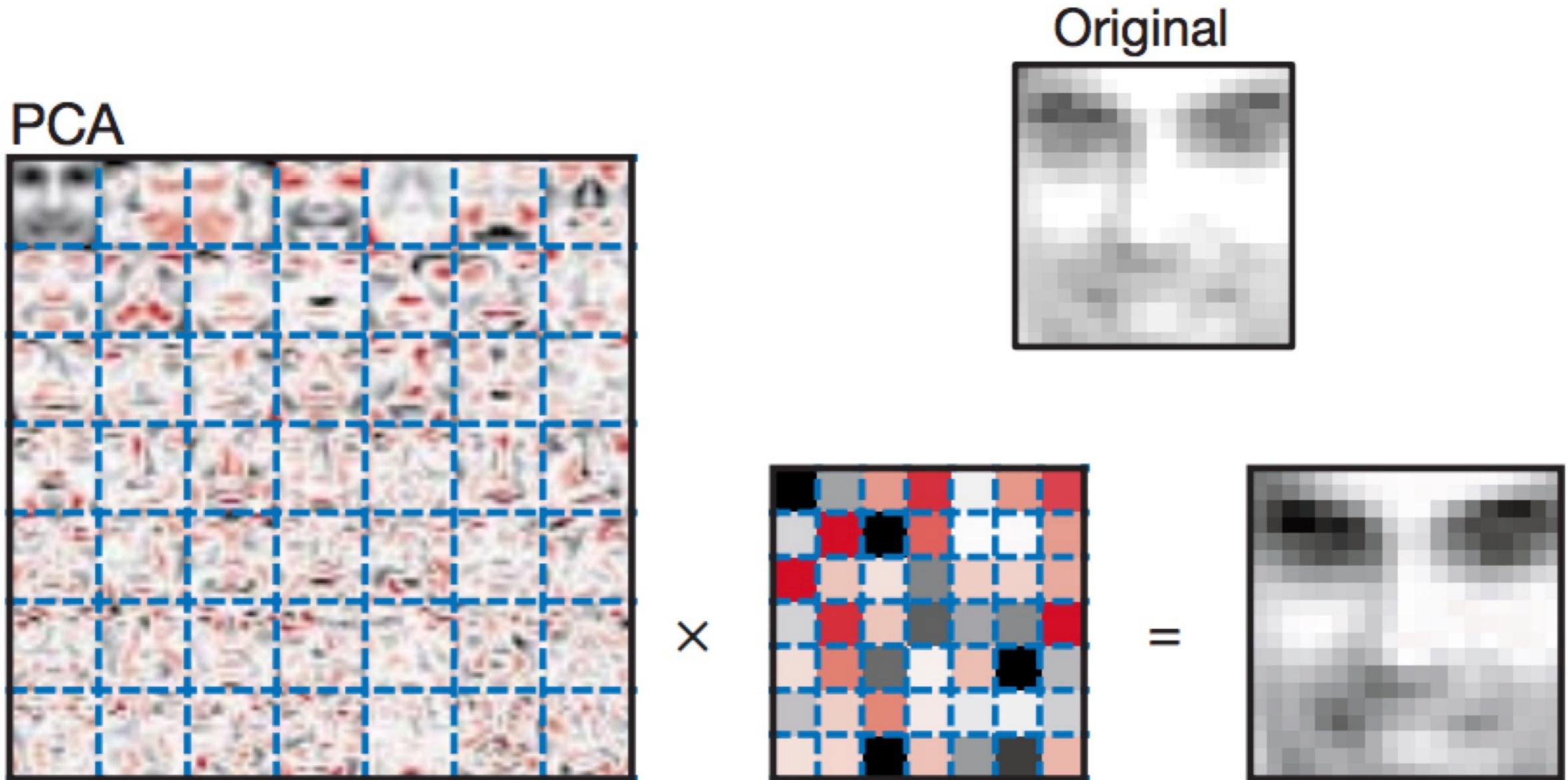
Question: what are the common bases D of human face?

The Yale Face Database B

Dictionary learning

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

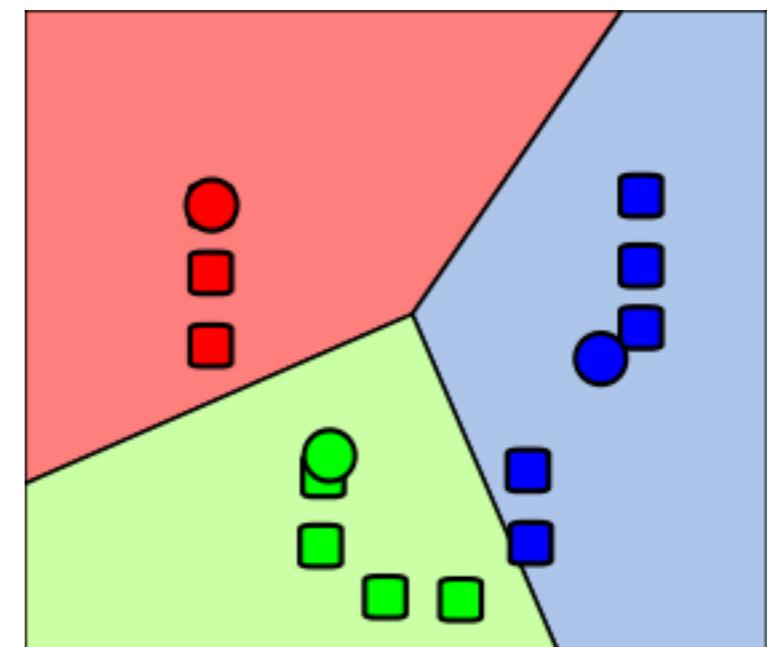
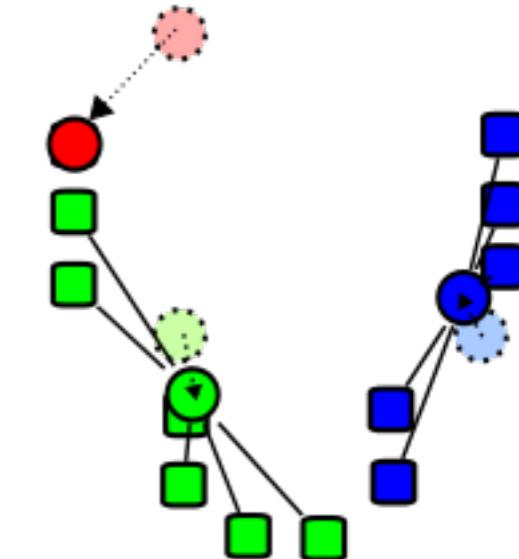
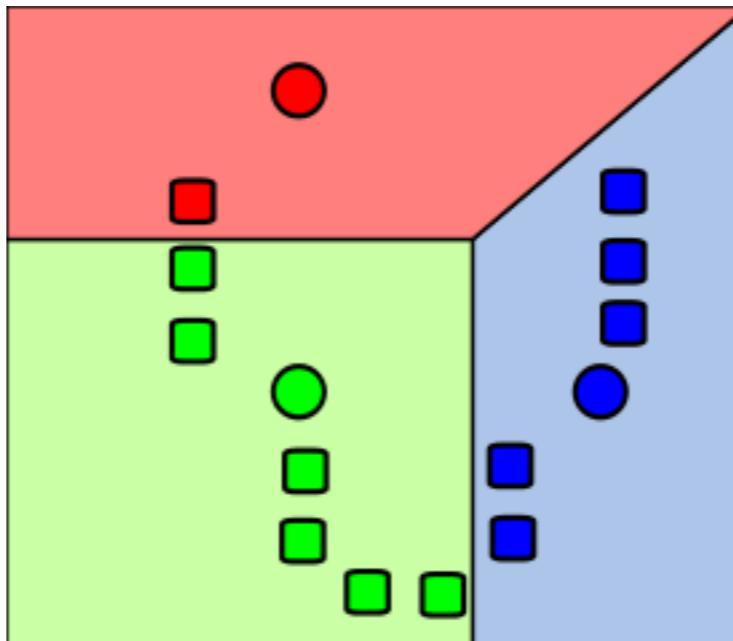
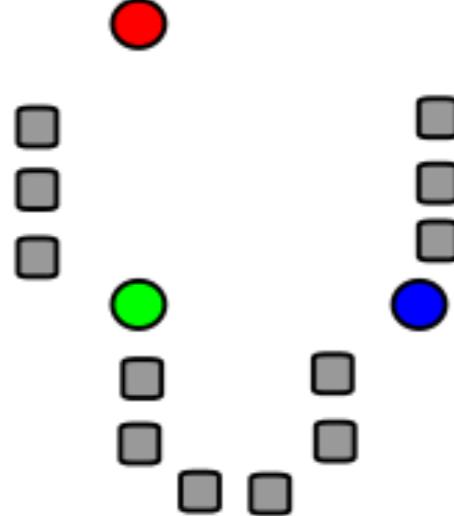
$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$



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:

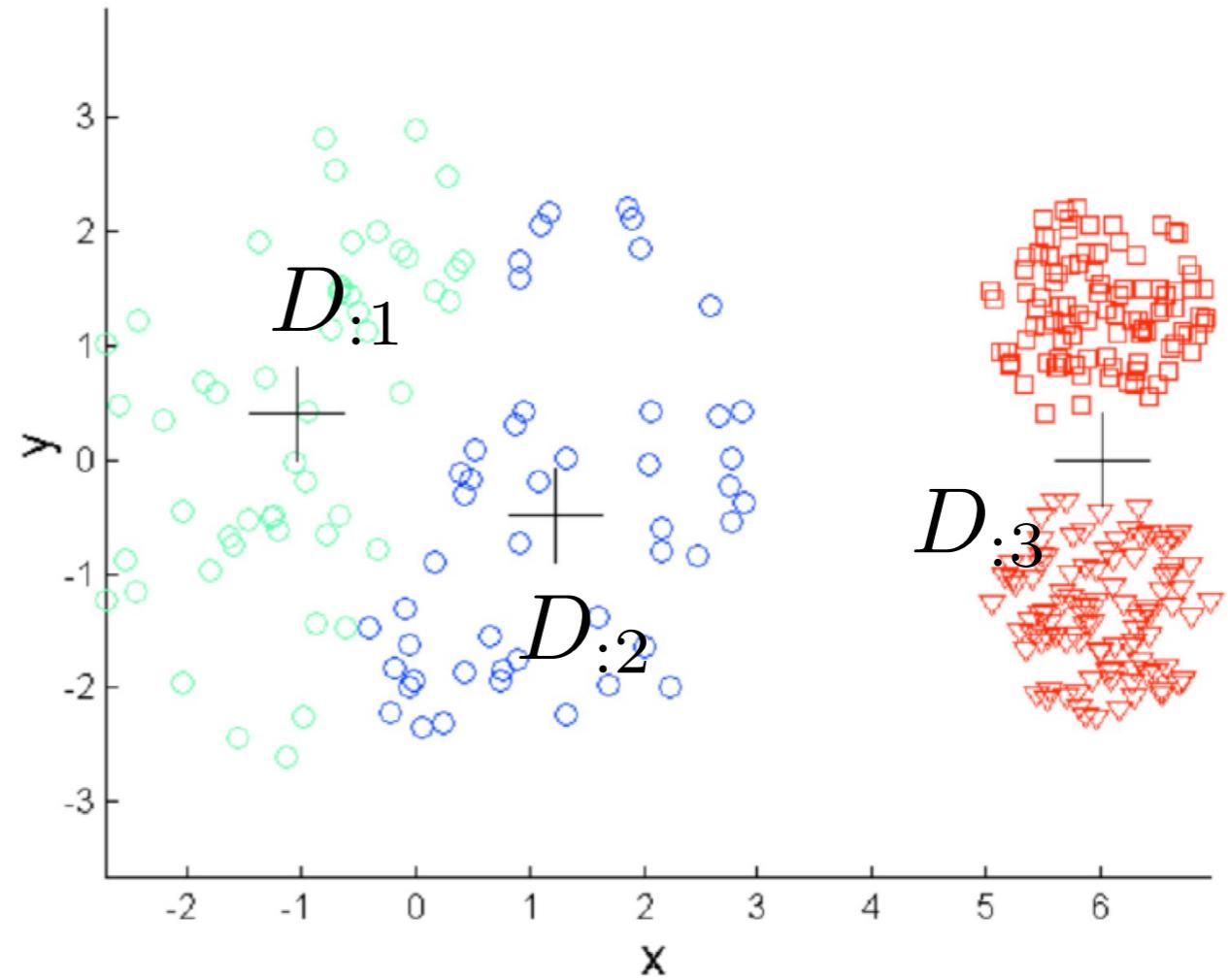


An updating rule to find the k centroids.

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

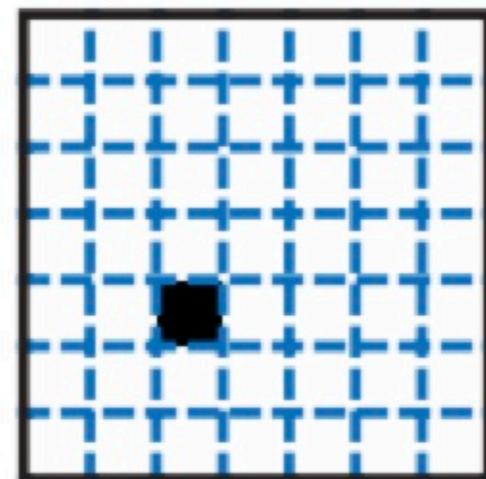
K-means clustering:

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

K-means centroids



\times



$=$



Original



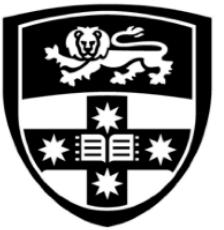
Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

Revisit human faces



Question: is that possible to extract parts-based structures of human faces? E.g., eye, nose, mouth, ear, forehead, chin...

The Yale Face Database B



THE UNIVERSITY OF
SYDNEY

Non-negative matrix factorisation

Non-negative matrix factorisation



THE UNIVERSITY OF
SYDNEY

- **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



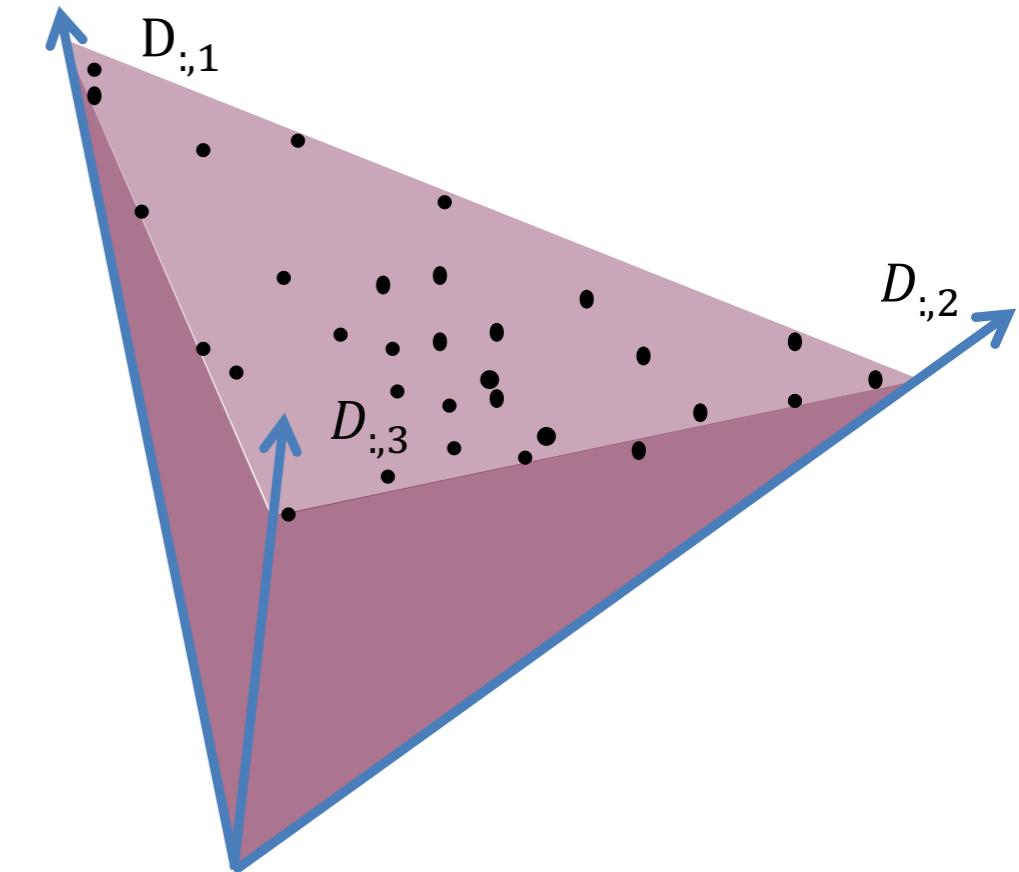
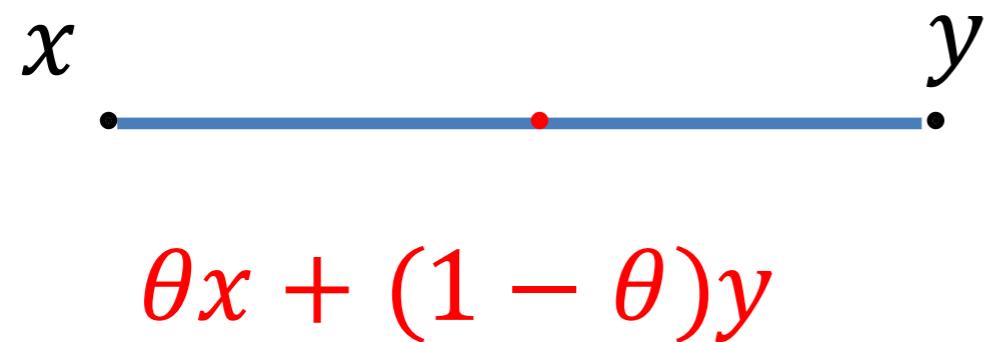
THE UNIVERSITY OF
SYDNEY

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

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



Geometry interpretation of non-negativity constraints



$$\theta \in [0, 1]$$

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

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



Geometry interpretation of non-negativity constraints

THE UNIVERSITY OF
SYDNEY

Question: what will the bases (columns of D , e.g., $D_{:1}$) be look like?



THE UNIVERSITY OF
SYDNEY

Geometry interpretation of non-negativity constraints

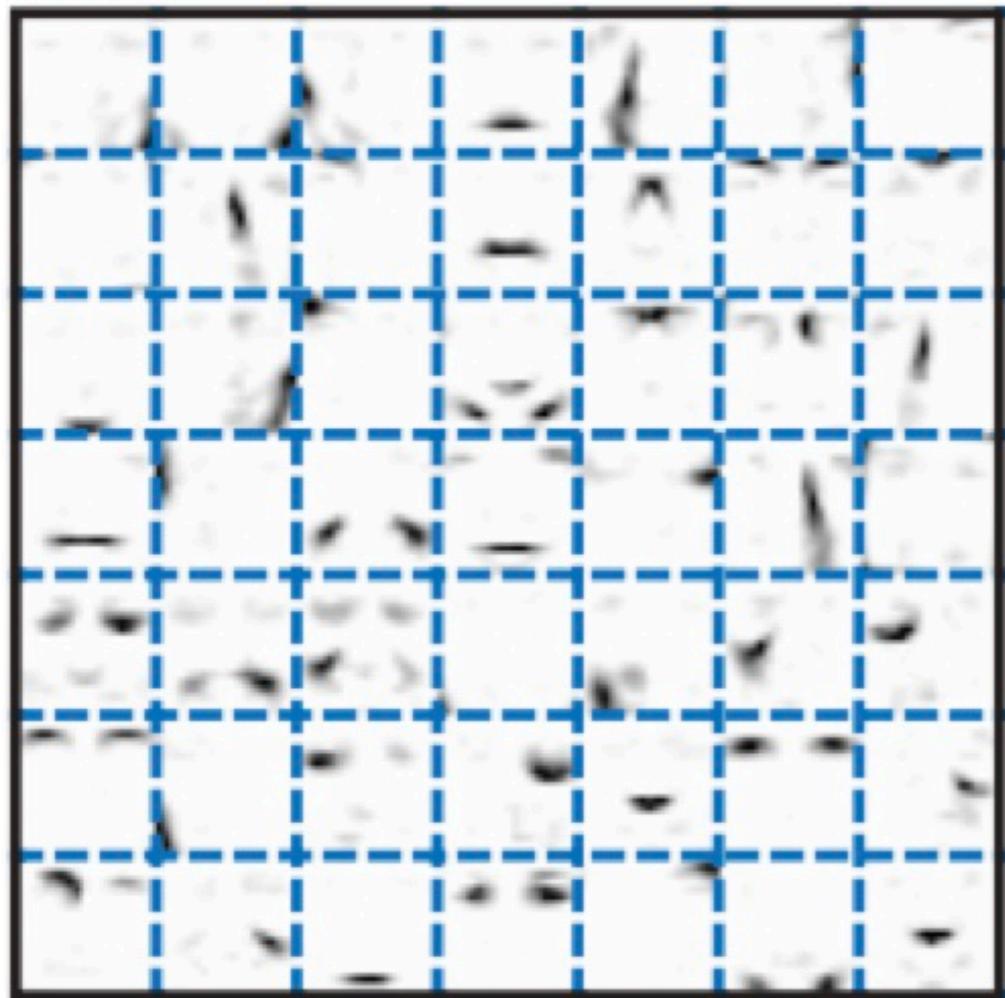
These constraints lead to a **parts-based representation** because they allow **only additive**, not subtractive, **combinations**.



Non-negative matrix factorisation

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

NMF



Original



$$\text{Original} \times \begin{matrix} \text{Matrix A} \\ \times \\ \text{Matrix B} \end{matrix} = \text{NMF Result}$$

The diagram illustrates the Non-Negative Matrix Factorization (NMF) process. It shows the original input image being multiplied by two non-negative matrices (A and B) to produce the NMF result. Matrix A is a sparse matrix with a checkered pattern of black and white pixels, while Matrix B is a denser matrix with various gray tones.

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/>

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

NMF optimisation

MUR (Multiplicative Update Rules):

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

Fix D^k , solve for R^{k+1}

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

$$R_{i,j}^{k+1} = R_{i,j}^k + \frac{\eta_{i,j}}{2} (2D^{k^\top} X - 2D^{k^\top} D^k R^k)_{i,j}$$

$$\eta_{i,j} = \frac{R_{i,j}^k}{(D^{k^\top} D^k R^k)_{i,j}}$$

NMF optimisation

MUR (Multiplicative Update Rules):

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

Fix D^k , solve for R^{k+1}

$$R_{i,j}^{k+1} = R_{i,j}^k \frac{(D^{k^\top} X)_{i,j}}{(D^{k^\top} D^k R^k)_{i,j}}$$

Fix R^{k+1} , solve for D^{k+1}

$$D_{i,j}^{k+1} = D_{i,j}^k \frac{(X R^{k+1})_{i,j}^\top}{(D^k R^{k+1} R^{k+1})_{i,j}^\top}$$

NMF optimisation

MUR (Multiplicative Update Rules):

$$R_{i,j}^{k+1} = R_{i,j}^k \frac{(D^{k^\top} X)_{i,j}}{(D^{k^\top} D^k R^k)_{i,j}}$$

$$D_{i,j}^{k+1} = D_{i,j}^k \frac{(X R^{k+1})_{i,j}^\top}{(D^k R^{k+1} R^{k+1})_{i,j}^\top}$$

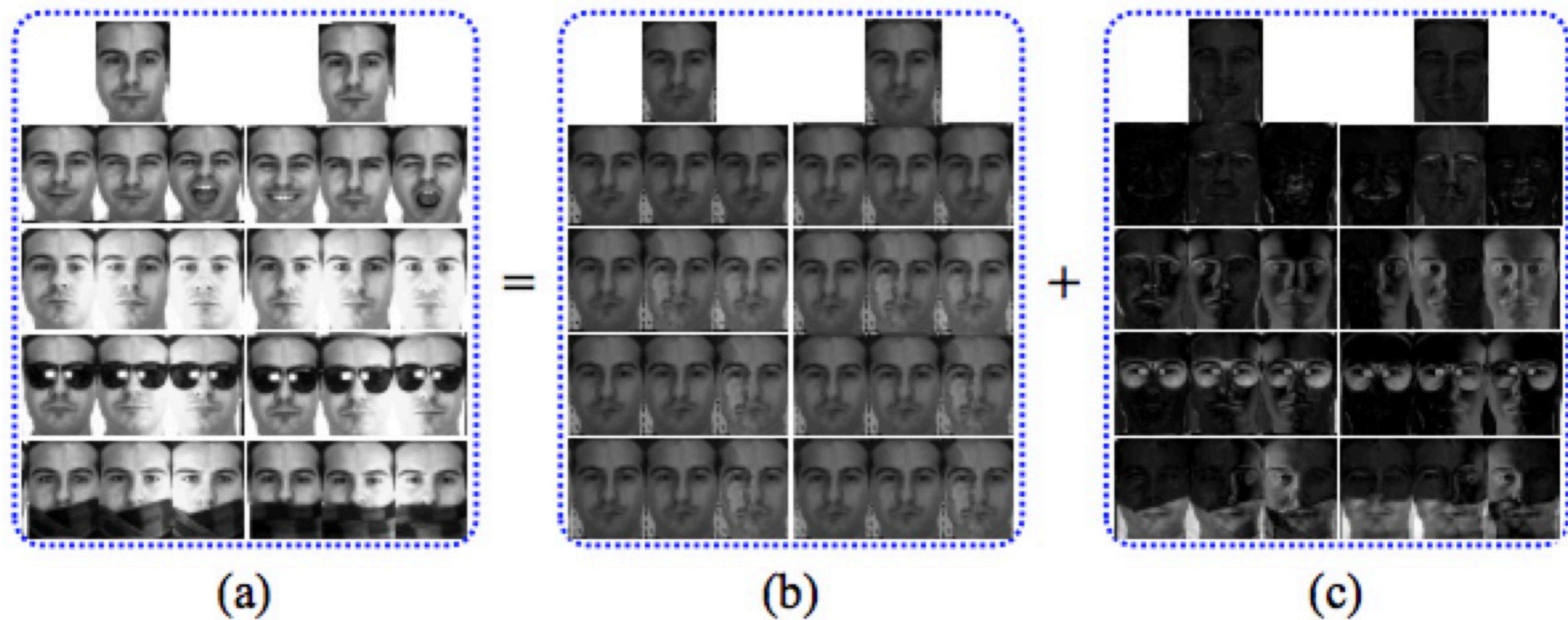
Theorem

The above update rule will make the objective $\|X - DR\|_F^2$ non-increasing. The objective is invariant if and only if D^{k+1} and R^{k+1} are at a stationary point of the distance.

Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." Advances in neural information processing systems. 2001.

NMF applications

Face processing



Guan, Naiyang, et al. "Truncated Cauchy Non-negative Matrix Factorization for Robust Subspace Learning." IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).



Pros and cons

- **Pros:**
Great interpretability
- **Cons:**
Factorisation is not unique
Feature redundancy among the bases



NMF variants

- Model developments with **various loss functions** in addition to squared loss
- to compensate for feature redundancy in the columns
- to improve generalisation ability