

# Data Representation as Low Rank Matrix Factorization

Ziv Epstein

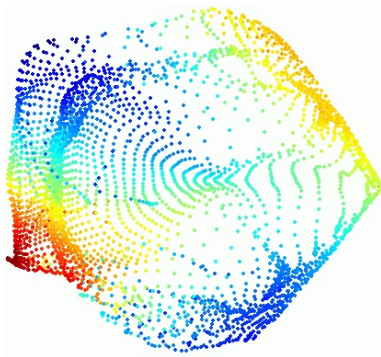
`ziv.epstein@pomona.edu`

Pomona College

Advised by Blake Hunter

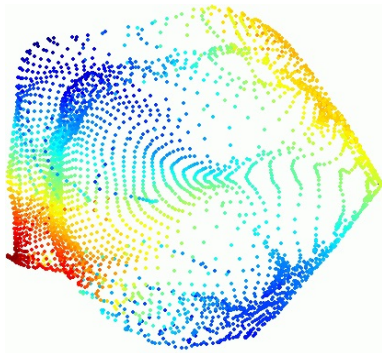
February 21, 2017

In many contexts in data science and linear algebra, we have lots of points in a high-dimensional space.

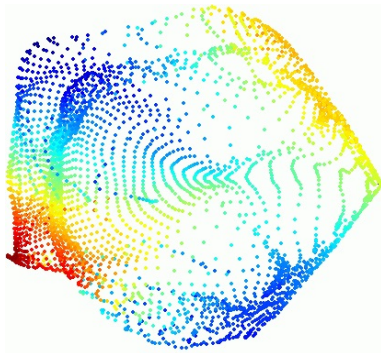


Denote these  $x_i \in \mathbb{R}^m$  for  $i = 1 \dots n$ .

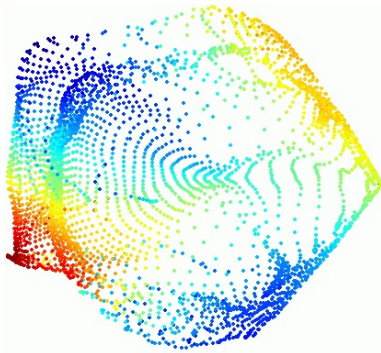
How do we cluster  $x_i$ ?



How do we cluster  $x_i$ ?    How do we perform dimensionality reduction?

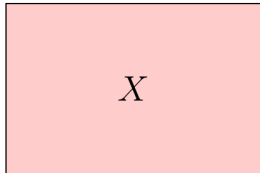


How do we cluster  $x_i$ ? How do we perform dimensionality reduction? How do we visualize them?



# Representation and Factorization

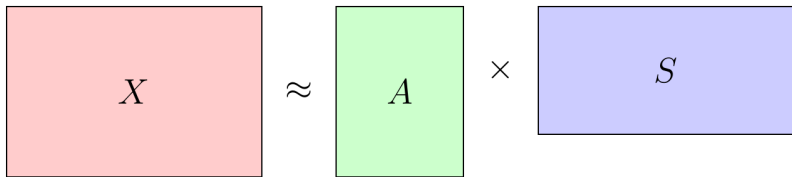
We can *represent*  $X \in \mathbb{R}^{n \times m}$ .



with  $\text{rank}(X) = \min(m, n)$ .

# Representation and Factorization

We can *approximate*  $X \in \mathbb{R}^{n \times m}$  as  $AS$ .



with  $\text{rank}(AS) = k \ll \min(m, n)$ .

# Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of “basis elements’.

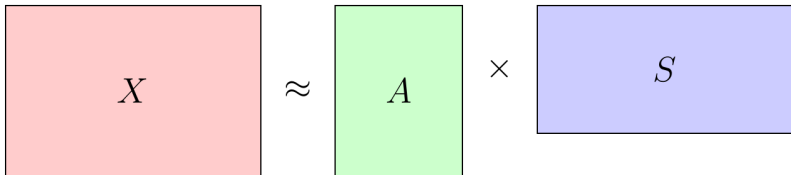
$$x_i = \sum_{j=1}^k A_{i,j} S_{j,-}$$



# Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of “basis elements”.

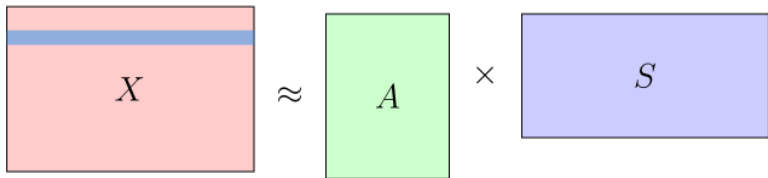
$$x_i = \sum_{j=1}^k A_{i,j} S_{j,-}$$



# Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of “basis elements”.

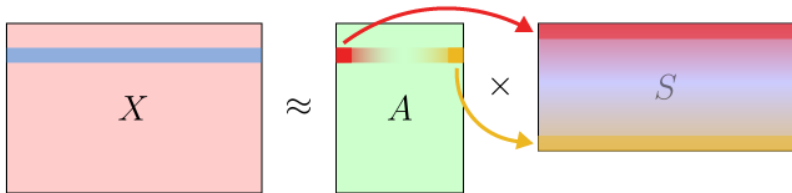
$$x_i = \sum_{j=1}^k A_{i,j} S_{j,-}$$



# Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of “basis elements”.

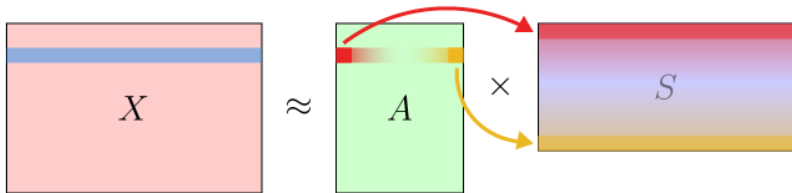
$$x_i = \sum_{j=1}^k A_{i,j} S_{j,-}$$



# Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of “basis elements”.

$$x_i = \sum_{j=1}^k A_{i,j} S_{j,-}$$



Find factorization by solving optimization problem

$$\min_{A,S} ||X - AS||$$

## Example (Lee and Seung 1999)

## Example (Lee and Seung 1999)

A 19x19 pixel greyscale image of a face is a data point ( $x_i \in \mathbb{R}^{361}$  with values between 0 and 256)



## Example (Lee and Seung 1999)

A 19x19 pixel greyscale image of a face is a data point ( $x_i \in \mathbb{R}^{361}$  with values between 0 and 256)



## Example (Lee and Seung 1999)

A 19×19 pixel greyscale image of a face is a data point ( $x_i \in \mathbb{R}^{361}$  with values between 0 and 256)

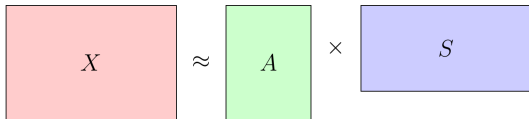


Take a database of 2,429 faces to get  $X \in \mathbb{R}^{2429 \times 361}$



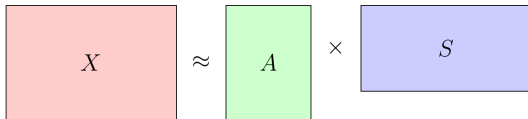
## Example (Lee and Seung 1999)

**Before:** Find  $A$  and  $S$  such that  $\|X - AS\|$  is minimized.



## Example (Lee and Seung 1999)

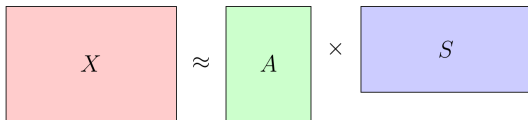
**Before:** Find  $A$  and  $S$  such that  $\|X - AS\|$  is minimized.



**Now:** We add some constraints

## Example (Lee and Seung 1999)

**Before:** Find  $A$  and  $S$  such that  $\|X - AS\|$  is minimized.

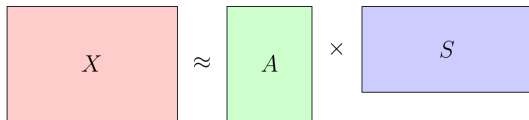

$$X \approx A \times S$$

**Now:** We add some constraints

- ① Columns of  $A$  to be orthonormal; Rows of  $S$  to be orthogonal (PCA)

## Example (Lee and Seung 1999)

**Before:** Find  $A$  and  $S$  such that  $\|X - AS\|$  is minimized.

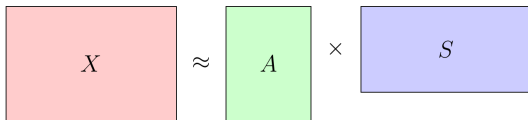

$$X \approx A \times S$$

**Now:** We add some constraints

- ① Columns of  $A$  to be orthonormal; Rows of  $S$  to be orthogonal (PCA)
- ② All entries of  $A$  and  $S$  are non-negative (NNMF)

## Example (Lee and Seung 1999)

**Before:** Find  $A$  and  $S$  such that  $\|X - AS\|$  is minimized.



**Now:** We add some constraints

- ① Columns of  $A$  to be orthonormal; Rows of  $S$  to be orthogonal (PCA)
- ② All entries of  $A$  and  $S$  are non-negative (NMF)

For each, find  $A$  and  $S$  subject to constraints that minimize

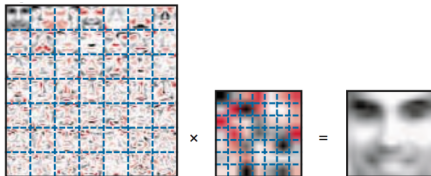
$$\|X - AS\|$$

## Example (Lee and Seung 1999)

1: PCA

# Example (Lee and Seung 1999)

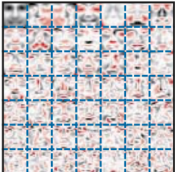


1: PCA





# Example (Lee and Seung 1999)

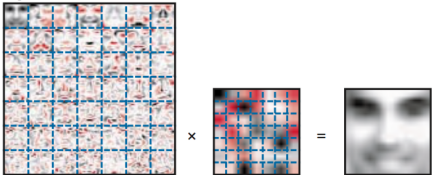
1: PCA

 $\times$  $=$ 

“eigenfaces”

# Example (Lee and Seung 1999)

1: PCA



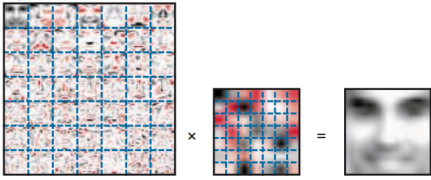
The diagram illustrates the PCA process for face images. On the left is a large grid of 25 small face images, each with a blue dashed border. This grid is multiplied (indicated by a small 'x') by a smaller grid of 25 color patches, also with blue dashed borders. This product is equal (indicated by an '=' sign) to a single grayscale face image on the right, which is labeled "eigenfaces".

$\times$   $=$  "eigenfaces"

2: NMF

# Example (Lee and Seung 1999)

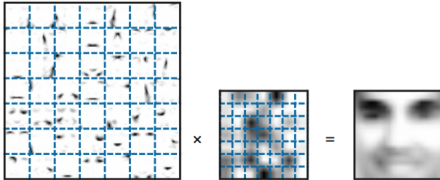
1: PCA



The diagram illustrates the PCA process. On the left is a 10x10 grid of small face images, each with a blue dashed border. To its right is a small 10x10 grid of feature vectors, also with blue dashed borders. A multiplication sign 'x' is between them. To the right of the feature vectors is an equals sign '=', followed by a single grayscale face image. The text "eigenfaces" is written to the right of the final image.

$\times$  = "eigenfaces"

2: NMF

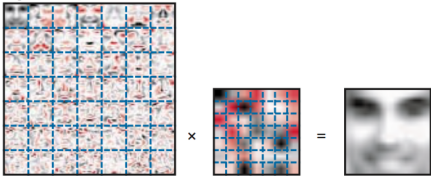


The diagram illustrates the NMF process. On the left is a 10x10 grid of small face images, each with a blue dashed border. To its right is a small 10x10 grid of feature vectors, also with blue dashed borders. A multiplication sign 'x' is between them. To the right of the feature vectors is an equals sign '=', followed by a single grayscale face image.

$\times$  =

# Example (Lee and Seung 1999)

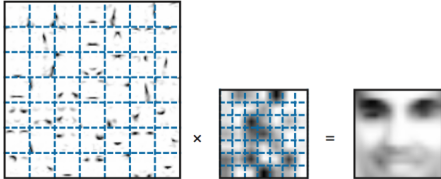
1: PCA



The diagram illustrates the PCA process. On the left is a 10x10 grid of small face images, each with a blue dashed bounding box. This grid is multiplied (indicated by a small 'x') by a single 10x10 grid of small images, which are the principal components. This is followed by an equals sign (=) and a single larger grayscale face image, labeled "eigenfaces".

$\times$  = "eigenfaces"

2: NMF



The diagram illustrates the NMF process. On the left is a 10x10 grid of small face images, each with a blue dashed bounding box. This grid is multiplied (indicated by a small 'x') by a single 10x10 grid of small images, which are the non-negative matrix factors. This is followed by an equals sign (=) and a single larger grayscale face image, labeled "parts-based".

$\times$  = parts-based

NNMF learns an additive, parts based model



NNMF learns an additive, parts based model



which is what our brain does when recognizing images!

There is a lot of evidence in cognitive science that we understand high level concepts using a parts-based representation

There is a lot of evidence in cognitive science that we understand high level concepts using a parts-based representation

- ① Biederman, Irving. "Recognition-by-components: a theory of human image understanding." Psychological review 94.2 (1987): 115.



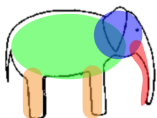
There is a lot of evidence in cognitive science that we understand high level concepts using a parts-based representation

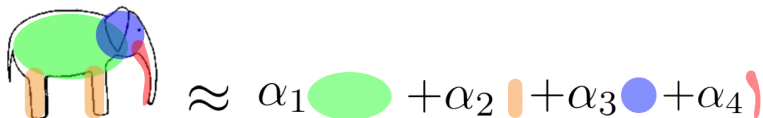
- ① Biederman, Irving. "Recognition-by-components: a theory of human image understanding." Psychological review 94.2 (1987): 115.
- ② Minsky, Marvin. "A framework for representing knowledge." (1974).

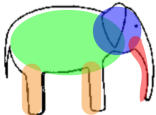
There is a lot of evidence in cognitive science that we understand high level concepts using a parts-based representation

- ① Biederman, Irving. "Recognition-by-components: a theory of human image understanding." *Psychological review* 94.2 (1987): 115.
- ② Minsky, Marvin. "A framework for representing knowledge." (1974).
- ③ Palmer, Stephen E. "Hierarchical structure in perceptual representation." *Cognitive psychology* 9.4 (1977): 441-474.







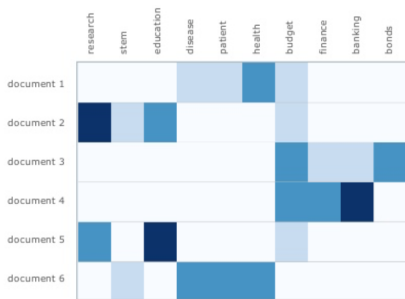


$$\approx \alpha_1 \text{ (green body)} + \alpha_2 \text{ (orange leg)} + \alpha_3 \text{ (blue head)} + \alpha_4 \text{ (red trunk)}$$

where , , , ,   $\in \mathbb{R}^m$

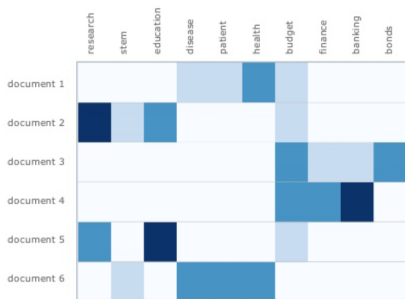


# Topic Modelling



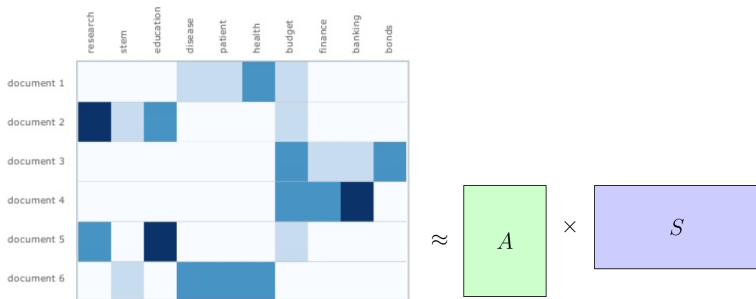


# Topic Modelling



$$\approx \begin{matrix} \boxed{A} \\ \times \\ \boxed{S} \end{matrix}$$

# Topic Modelling



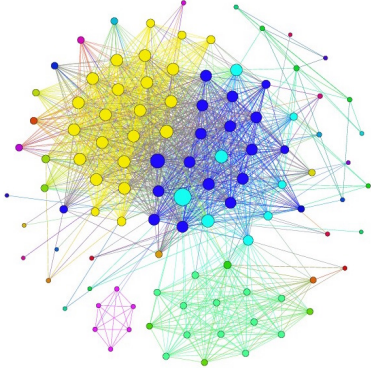
Perform with corpus of news documents with  $k = 10$

# Topic Modelling

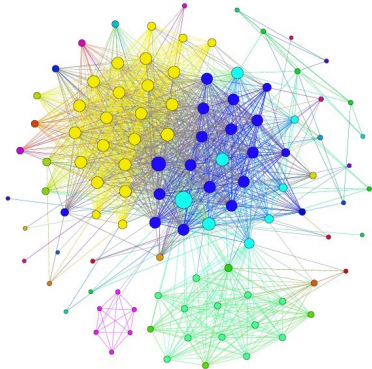
topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
reds	twins	jays	report	banks	china	percent	watt	truth	invasion
cincinnati	minnesota	blue	hhs	loans	economic	revenue	freddie	opinions	allied
pirates	runs	toronto	exchanges	collateral	reforms	quarter	mae	reason	troops
hit	game	hit	consumers	abs	growth	billion	fannie	certain	normandy
season	said	game	plans	ecb	beijing	million	mac	nature	german
cueto	innings	said	cost	bank	economy	company	senate	objects	british
latos	inning	gibbons	health	lending	said	share	republican	TRUE	germans
pittsburgh	indians	run	costs	small	li	cents	nomination	men	landing
game	run	bautista	healthcare	european	year	year	panel	god	0
bruce	plouffe	rockies	premiums	assets	urged	sales	committee	thought	beaches
run	hit	davis	insurance	loan	fiscal	shares	obama	mind	france
left	sox	right	individual	businesses	ministry	rose	sec	order	eisenhower
inning	left	runs	reform	euro	speed	said	fhfa	knowledge	divisions
arizona	white	single	states	funds	spending	earnings	nominee	thoughts	beach
innings	tigers	reyes	lower	backed	premier	trading	government	doubt	allies
games	dozier	rays	affordable	credit	local	ebay	democrat	sciences	operation
got	cleveland	johnson	month	smes	policy	sandisk	carolina	ought	june
said	hits	left	silver	firms	sources	profit	demarco	life	1944
second	detroit	kawasaki	lowest	rbs	banks	stock	housing	principles	day
homer	season	walked	administrative	sme	government	fell	likely	hear	forces



Graph clustering!



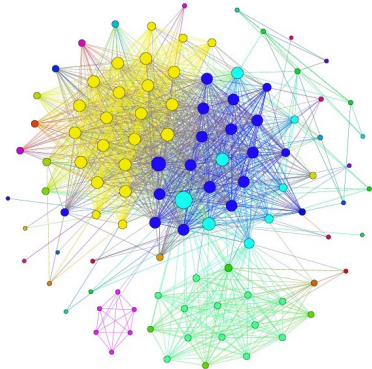
## Graph clustering!



### Definition

A **clustering**  $C = \{C_1, \dots, C_k\}$  of a graph  $G = (E, V)$  is a partition of the vertices. A **good clustering** optimizes some measure of community.

## Graph clustering!



### Definition

A **clustering**  $C = \{C_1, \dots, C_k\}$  of a graph  $G = (E, V)$  is a partition of the vertices. A **good clustering** optimizes some measure of community.

$$\text{connectivity}(G, C) = \frac{\#\{(i, j) \in E \mid (i, j) \in C\} + \#\{(i, j) \notin E \mid (i, j) \notin C\}}{|V|^2}$$

The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .



The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .

$$\text{connectivity}(G, C) = 1 - \frac{\sum_{i,j} \left( A_{ij} - \sum_{r=1}^k M_{ir} M_{jr} \right)^2}{|V|^2}$$

The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .

$$\begin{aligned} \text{connectivity}(G, C) &= 1 - \frac{\sum_{i,j} \left( A_{ij} - \sum_{r=1}^k M_{ir} M_{jr} \right)^2}{|V|^2} \\ &= 1 - \frac{\|A - MM^T\|_F^2}{|V|^2} \end{aligned}$$

The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .

$$\begin{aligned} \text{connectivity}(G, C) &= 1 - \frac{\sum_{i,j} \left( A_{ij} - \sum_{r=1}^k M_{ir} M_{jr} \right)^2}{|V|^2} \\ &= 1 - \frac{\|A - MM^T\|_F^2}{|V|^2} \end{aligned}$$

Finding a clustering  $C$  that maximizes the connectivity is equivalent to the minimization problem

$$\min_{M \in \{0,1\}^{n \times k}} \|A - MM^T\|_F^2$$

The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .

$$\begin{aligned} \text{connectivity}(G, C) &= 1 - \frac{\sum_{i,j} \left( A_{ij} - \sum_{r=1}^k M_{ir} M_{jr} \right)^2}{|V|^2} \\ &= 1 - \frac{\|A - MM^T\|_F^2}{|V|^2} \end{aligned}$$

Finding a clustering  $C$  that maximizes the connectivity is equivalent to the minimization problem

~~$$\min_{M \in \{0,1\}^{n \times k}} \|A - MM^T\|_F^2$$~~

The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .

$$\begin{aligned} \text{connectivity}(G, C) &= 1 - \frac{\sum_{i,j} \left( A_{ij} - \sum_{r=1}^k M_{ir} M_{jr} \right)^2}{|V|^2} \\ &= 1 - \frac{\|A - MM^T\|_F^2}{|V|^2} \end{aligned}$$

Finding a clustering  $C$  that maximizes the connectivity is equivalent to the minimization problem

~~$$\min_{M \in \{0,1\}^{n \times k}} \|A - MM^T\|_F^2$$~~

$$\min_{M \in \mathbb{R}_+^{n \times k}} \|A - MM^T\|_F^2$$

The clustering  $C$  can be formulated as a membership matrix  $M \in \{0, 1\}^{n \times k}$  with  $M_{ir} = 1$  if  $i \in C_r$  and  $M_{ir} = 0$  if  $i \notin C_r$ .

$$\begin{aligned} \text{connectivity}(G, C) &= 1 - \frac{\sum_{i,j} \left( A_{ij} - \sum_{r=1}^k M_{ir} M_{jr} \right)^2}{|V|^2} \\ &= 1 - \frac{\|A - MM^T\|_F^2}{|V|^2} \end{aligned}$$

Finding a clustering  $C$  that maximizes the connectivity is equivalent to the minimization problem

~~$$\min_{M \in \{0,1\}^{n \times k}} \|A - MM^T\|_F^2$$~~

$$\min_{M \in \mathbb{R}_+^{n \times k}} \|A - MM^T\|_F^2$$

where  $M$  is a **weak membership matrix**. That is, the larger  $M_{i,j}$ , the stronger membership of vertex  $i$  in cluster  $j$ .



- Turns out graph clustering is NNMF in disguise!





- Turns out graph clustering is NMF in disguise!
- Powerful, visual and intuitive technique for any data that can be represented as a matrix



- Turns out graph clustering is NNMF in disguise!
- Powerful, visual and intuitive technique for any data that can be represented as a matrix
- My thesis:



- Turns out graph clustering is NNMF in disguise!
- Powerful, visual and intuitive technique for any data that can be represented as a matrix
- My thesis:
  - Explore more types of NNMF and relation to cognition: supervised, hierarchical



- Turns out graph clustering is NNMF in disguise!
- Powerful, visual and intuitive technique for any data that can be represented as a matrix
- My thesis:
  - Explore more types of NNMF and relation to cognition: supervised, hierarchical
  - Neural Networks can even be thought of as recursive NNMF!
  - viz

