Data Representation as Low Rank Matrix Factorization

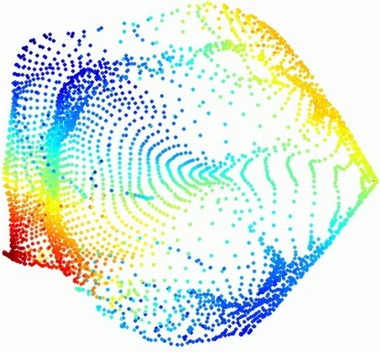
Ziv Epstein
ziv.epstein@pomona.edu

Pomona College
Advised by Blake Hunter

February 24, 2017

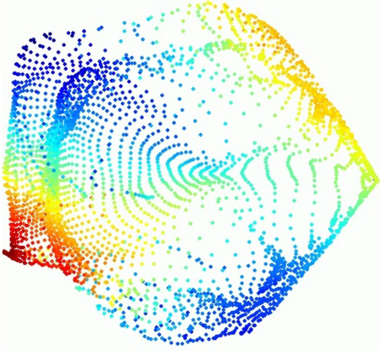
Points in $\mathbb{R}^m$

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



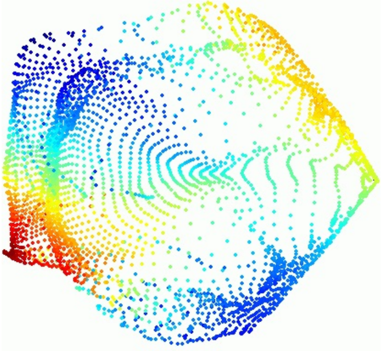
Denote these $x_i \in \mathbb{R}^m$ for $i = 1 \ldots n$.

How do we cluster x_i?

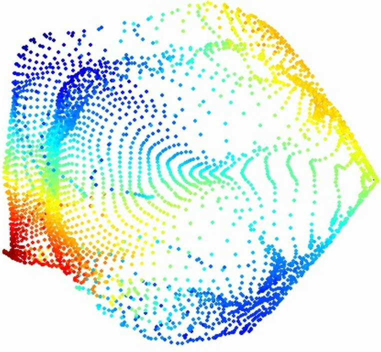


Points in $\mathbb{R}^m$

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

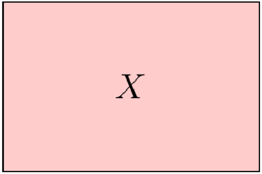


How do we cluster x_i? How do we perform dimensionality reduction? How do we visualize them?



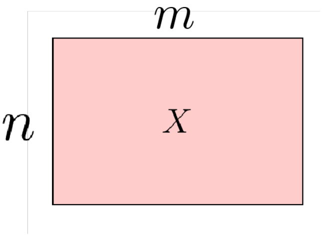
Representation and Factorization

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



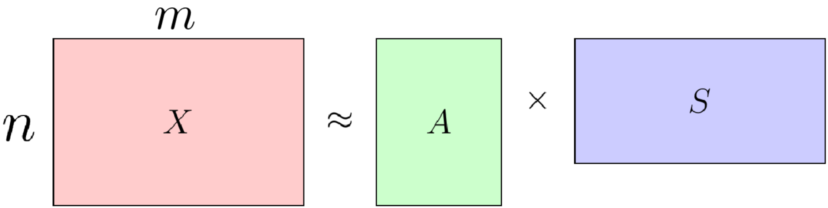
with $\text{rank}(X) = \min(m, n)$.

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



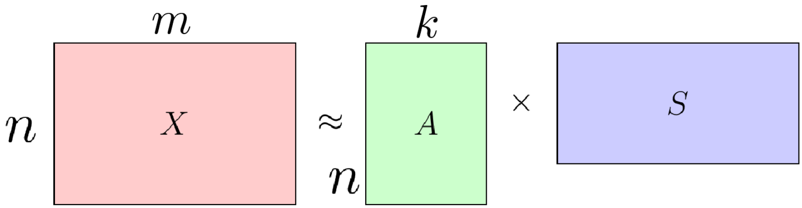
with $\text{rank}(X) = \min(m, n)$.

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



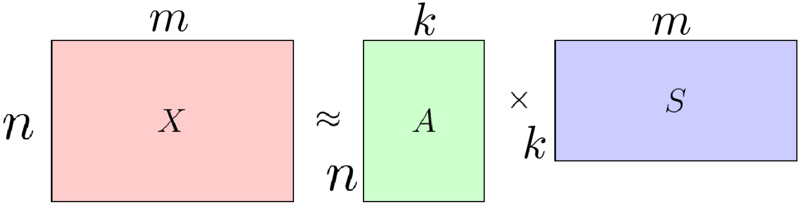
with $\text{rank}(AS) = k << \text{rank}(X)$.

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



with $\text{rank}(AS) = k << \text{rank}(X)$.

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



with $\text{rank}(AS) = k << \text{rank}(X)$.

Our original high-dimensional points are linear combinations of "basis elements'.'

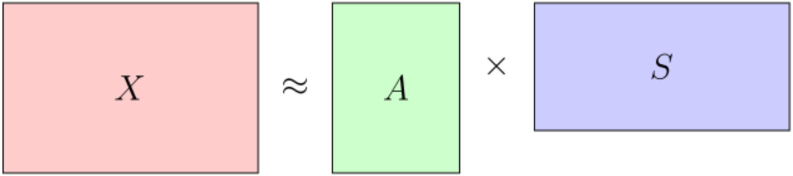
$$x_i = \sum_{j=1}^{k} A_{i,j} S_{j,\_}$$

.

Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of "basis elements'.'

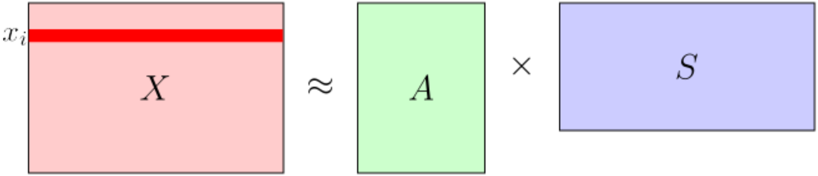
$$x_i = \sum_{j=1}^{k} A_{i,j} S_{j,\_}$$



Low Dimensional Interpretation

Our original high-dimensional points are linear combinations of "basis elements'.'

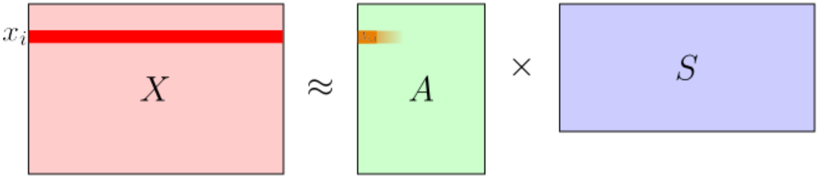
$$x_i = \sum_{j=1}^{k} A_{i,j} S_{j,-}$$



Low Dimensional Interpretation

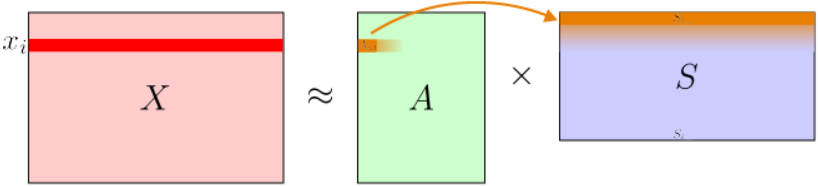
Our original high-dimensional points are linear combinations of "basis elements'.'

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



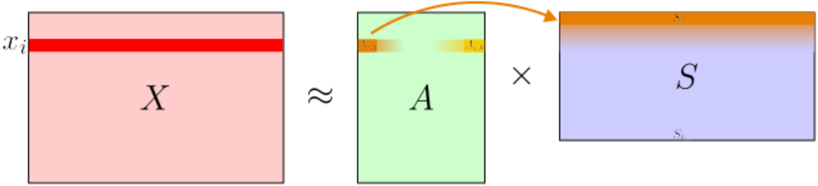
Our original high-dimensional points are linear combinations of "basis elements'.'

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



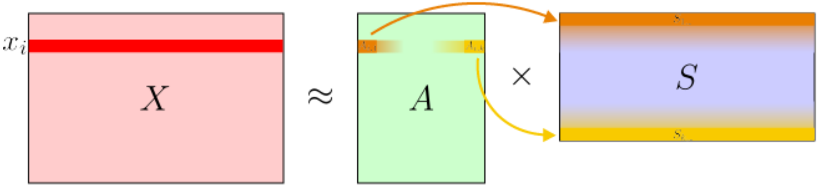
Our original high-dimensional points are linear combinations of "basis elements'.'

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



Our original high-dimensional points are linear combinations of "basis elements'.'

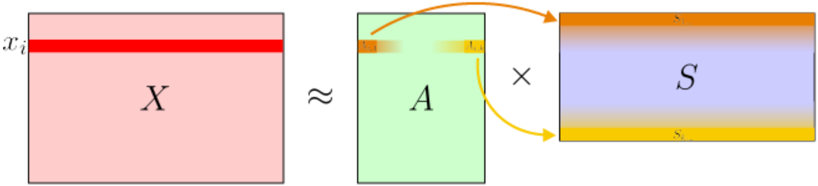
$$x_i = \sum_{j=1}^{k} A_{i,j} S_{j,-}$$



.

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

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

.

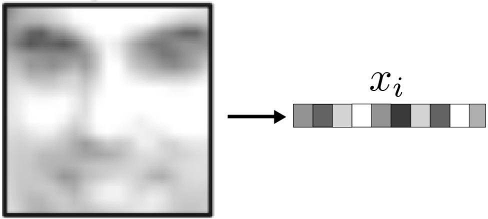
Example (Lee and Seung 1999)

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)

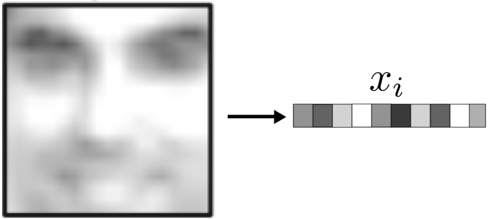
Example (Lee and Seung 1999)

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)

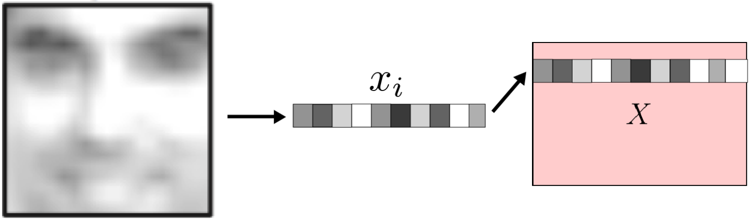
A 19x19 pixel greyscale image of a face is a data point ($x_i \in \mathbb{R}^{361}$ with values between 0 and 256)



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

Example (Lee and Seung 1999)

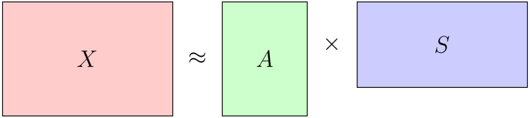
A 19x19 pixel greyscale image of a face is a data point ($x_i \in \mathbb{R}^{361}$ with values between 0 and 256)



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.



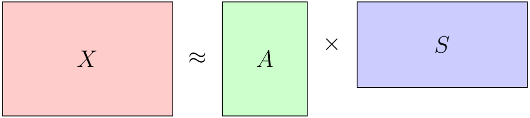
Example (Lee and Seung 1999)

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

$$X \approx A \times S$$

Now: We add some constraints

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)

Example (Lee and Seung 1999)

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

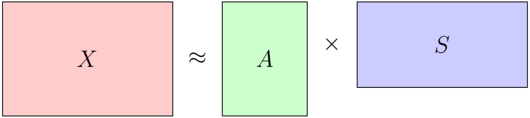
$$X \approx A \times S$$

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 (NMF)

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

$$||X - AS||$$

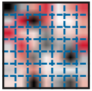
Example (Lee and Seung 1999)

1: PCA

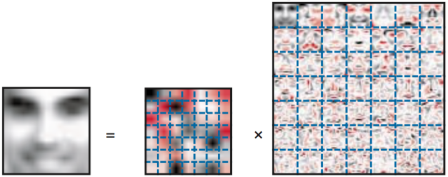
Example (Lee and Seung 1999)

1: PCA  =

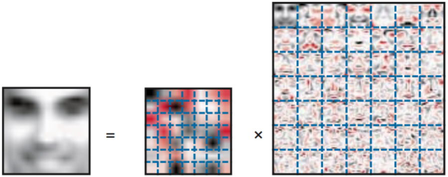
Example (Lee and Seung 1999)

1: PCA  = 

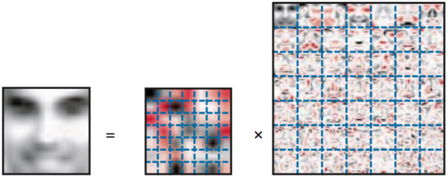
Example (Lee and Seung 1999)

1: PCA 

Example (Lee and Seung 1999)

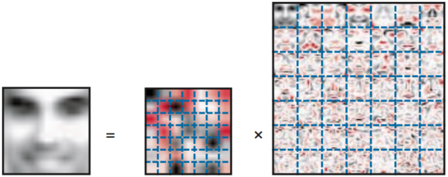
1: PCA  "eigenfaces"

Example (Lee and Seung 1999)

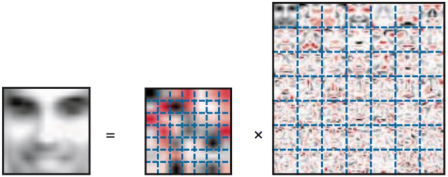
1: PCA  “eigenfaces”

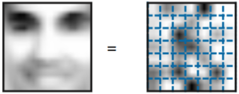
1: NMF

Example (Lee and Seung 1999)

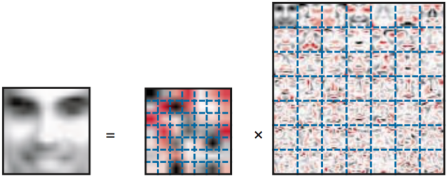
1: PCA  "eigenfaces"

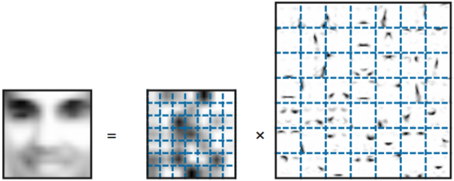
1: NMF  =

1: PCA  “eigenfaces”

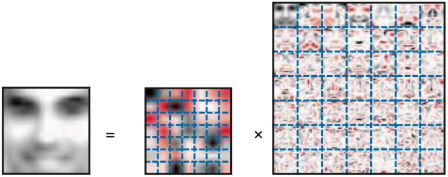
1: NMF 

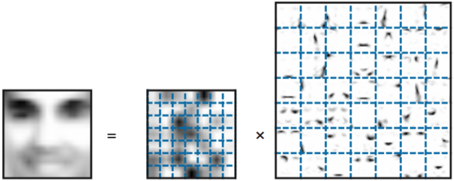
Example (Lee and Seung 1999)

1: PCA  "eigenfaces"

1: NMF 

Example (Lee and Seung 1999)

1: PCA  “eigenfaces”

1: NMF  parts based

NMF learns an additive, parts based model

NMF learns an additive, parts based model

which is what our brain does when recognizing images!

Recognition-by-parts

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. Minsky, Marvin. "A framework for representing knowledge." (1974).

Recognition-by-parts

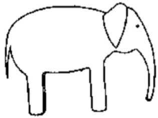
There is a lot of evidence in cognitive science that we understand high level concepts using a parts-based representation

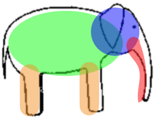
1. Minsky, Marvin. "A framework for representing knowledge." (1974).
2. Palmer, Stephen E. "Hierarchical structure in perceptual representation." Cognitive psychology 9.4 (1977): 441-474.

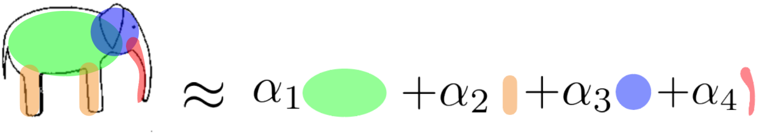
Recognition-by-parts

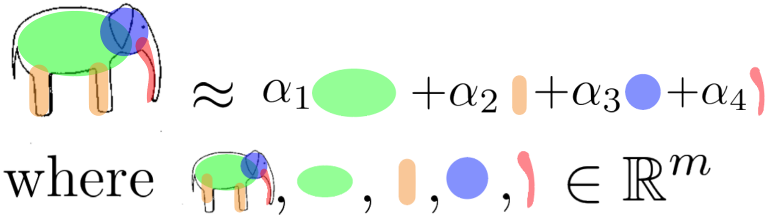
There is a lot of evidence in cognitive science that we understand high level concepts using a parts-based representation

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

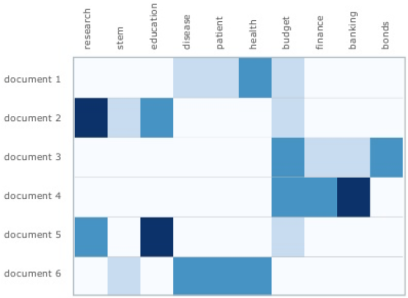


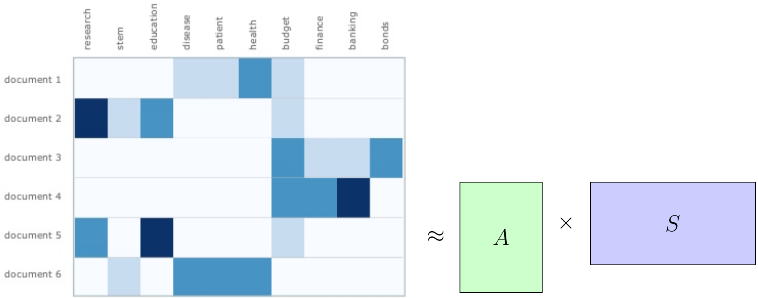



≈ α1 +α2 +α3 +α4

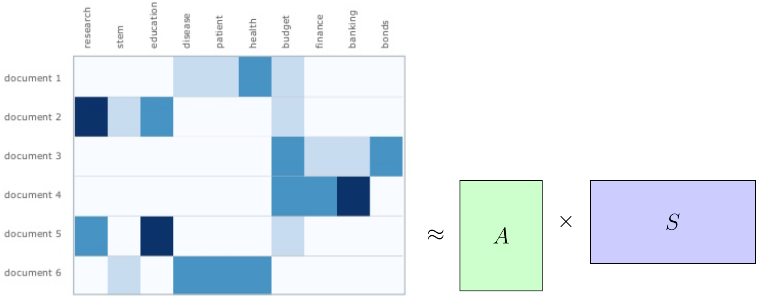

$\approx \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$
where , , , , $\in \mathbb{R}^m$

Topic Modelling





research
stem
education
disease
patient
health
budget
finance
banking
bonds
document 1
document 2
document 3
document 4
document 5
document 6
≈
A
×
S



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

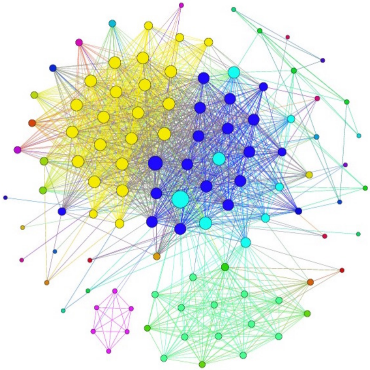
Topic Modelling

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	administratio	sme	government	fell	likely	hear	forces

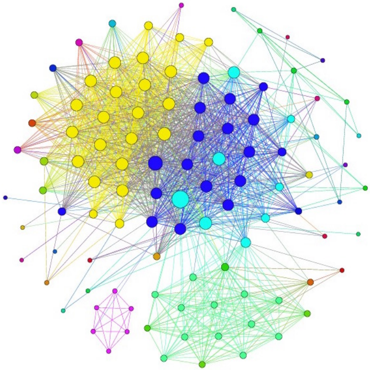
One final application

One final application

Graph clustering!



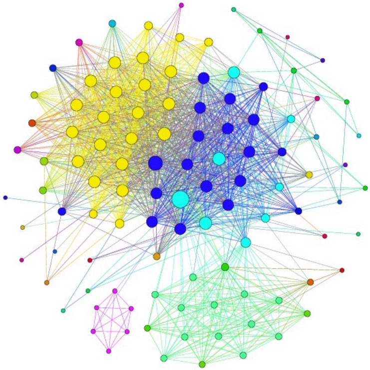
Graph clustering!



Definition

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

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

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} 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} if $i \notin C_r$.

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

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} if $i \notin C_r$.

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

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} if $i \notin C_r$.

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

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

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

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} if $i \notin C_r$.

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

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

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

Graph Clustering

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} if $i \notin C_r$.

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

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

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

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

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} if $i \notin C_r$.

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

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

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

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

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

Conclusion

- Many important problems are actually NMF in disguise!

Conclusion

- Many important problems are actually NMF in disguise!
- Powerful, visual and intuitive technique for any data set that can be represented as a matrix

Conclusion

- Many important problems are actually NMF in disguise!
- Powerful, visual and intuitive technique for any data set that can be represented as a matrix
- My thesis also explores:

Conclusion

- Many important problems are actually NMF in disguise!
- Powerful, visual and intuitive technique for any data set that can be represented as a matrix
- My thesis also explores:
  - more types of NMF and relation to cognition: supervised, hierarchical

Conclusion

- Many important problems are actually NMF in disguise!
- Powerful, visual and intuitive technique for any data set that can be represented as a matrix
- My thesis also explores:
  - more types of NMF and relation to cognition: supervised, hierarchical
  - neural networks, they can even be thought of as recursive NMF!

Conclusion

- Many important problems are actually NMF in disguise!
- Powerful, visual and intuitive technique for any data set that can be represented as a matrix
- My thesis also explores:
    - more types of NMF and relation to cognition: supervised, hierarchical
    - neural networks, they can even be thought of as recursive NMF!
    - Visualization techniques and implementations

Thanks!