Sparseness and NMF: An Overview

Hao Xu, Bingfeng Shu, Shane Deiley

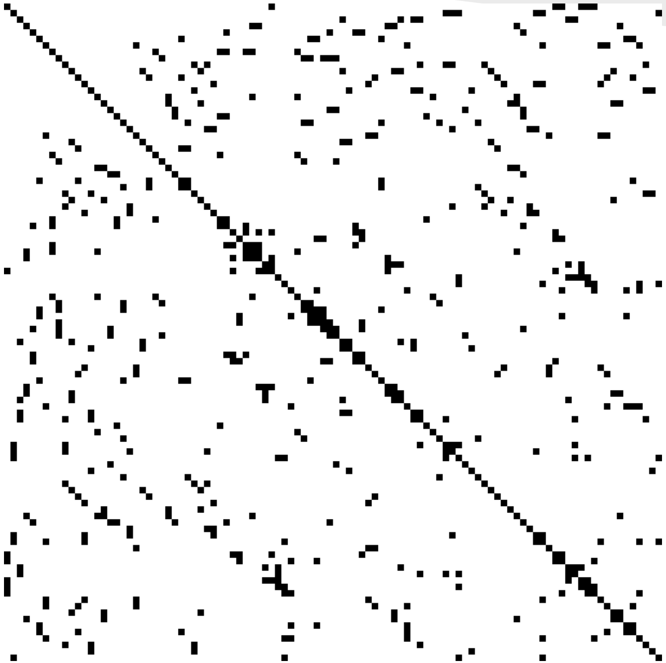
1. Sparseness
2. Insufficient Data Reduction Methods
3. Two Successful Approaches
    a. Sparseness-Constrained NMF
    b. Group Sparse Coding

Presentation Agenda

Introduction to Sparseness

Sparseness

$$\begin{pmatrix}
0 & 3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 7 \\
0 & 0 & 0 & 5 & 0 & 0 & 0 & 8 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 9 & 0 & 0 & 0 & 4 & 0 \\
0 & 1 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\
0 & 0 & 7 & 0 & 0 & 0 & 0 & 5 & 0 & 0
\end{pmatrix}$$



Naturally Sparse Data

- Text Matrices
- Disease Pattern in Patients

Natural Intrinsic Sparsity

- Data with few ‘active’ components
    - Also called latent factors
- Object or face detection
- Textual topic analysis

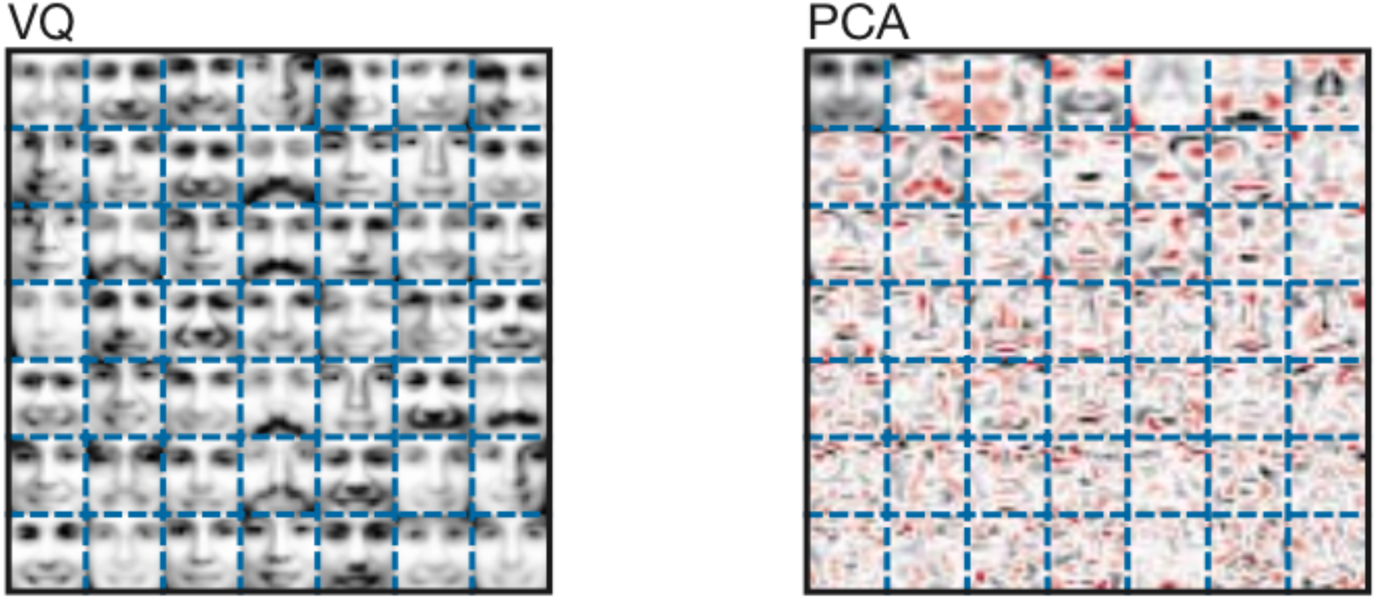
Implications of Sparse Decompositions

- Basis ~ Concise Summaries
- Coefficient ~ Low Term Linear Combination
- Parts-based representations

Data Reduction Methods and Sparsity Shortcomings

Sparsity in Data Reduction

- Holistic Representations - Not Sparse



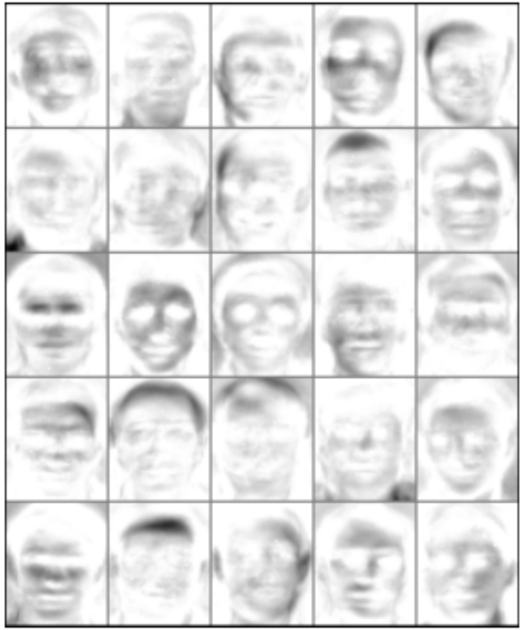
NMF

$$\min_{W \geq 0, H \geq 0} \|X - WH\|_F^2 .$$

Cannot Control Sparseness

CBCL

ORL



Non-Negative Sparse Coding

- Constraint term induces sparseness

$$\frac{1}{2}\|\mathbf{X}-\mathbf{AS}\|^2+\lambda\sum_{ij}S_{ij}$$

Low Rank Reduction (LRR)

- Low-rank naturally related to sparsity

$$\min_{Z} rank(Z), s.t., X = DZ.$$

Data Reduction Methods and Sparsity Success

Sparseness Constraints

Objective Function:

$$E(\mathbf{W},\mathbf{H}) = \|\mathbf{V} - \mathbf{WH}\|^2$$

Sparseness Constraints:

$$\begin{aligned} \text{sparseness}(\mathbf{w}_i) &= S_w, \ \forall i \\ \text{sparseness}(\mathbf{h}_i) &= S_h, \ \forall i, \end{aligned}$$

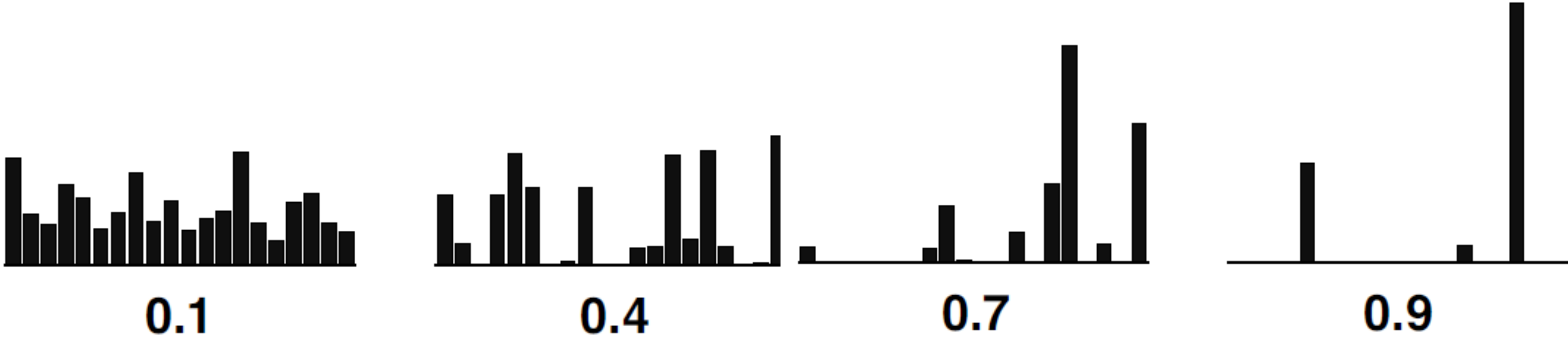
Sparseness Measure

Hoyer Measure: $$\text{sparseness}(\mathbf{x}) = \frac{\sqrt{n} - \left(\sum |x_i|\right) / \sqrt{\sum x_i^2}}{\sqrt{n} - 1}$$

- a normalized version of the l2 / l1 measure
  - It’s some kind of weighted sum of the coefficients
  - Normalization

Sparseness Measure

- Illustration of various degrees of sparseness.



How to control?

- Fix L2 norm
- Set L1 norm to achieve desired sparseness
- For H matrix, fix the L2 norm to unity

Projected Gradient Descent

- Gradient descent Method
  - Find a local minimum of a function using gradient descent

$$\text{minimize} \qquad f(x) \qquad \text{over} \qquad x \in \mathcal{X}$$

  - Take a step in the direction of the negative gradient

$$x_{t+1} = x_t - \eta \nabla f(x_t), \qquad t \in \mathbb{N}.$$

  - Unconstrained!

Projected Gradient Descent

- points need not belong to X

$$x_{t+1} = P_{\mathcal{X}}(x_t - \eta_t \nabla f(x_t)), \qquad t \in \mathbb{N}$$

- Projection operator Px
  - Given any vector x, find the closest non-negative vector with a given constraints.

Note!

- After projecting it need not be true that $f(x_{t+1}) < f(x_t)$
- Thus we need to adjust the step-size $\mu W > 0$ and $\mu H > 0$ for convergence.

Projection operator

- L1 norm constraint

$$\Sigma|si| = L1$$

$$\text{Set } s_i := x_i + \left(L_1 - \sum x_i\right)/\mathrm{dim}(\mathbf{x}), \ \forall i$$

Projection operator

- L2 norm constraint

$$\sqrt{{s_i}^2} = L2$$

$$\mathbf{s} := \mathbf{m} + \alpha(\mathbf{s} - \mathbf{m}), \text{ where } \alpha \geq 0$$

Projection operator

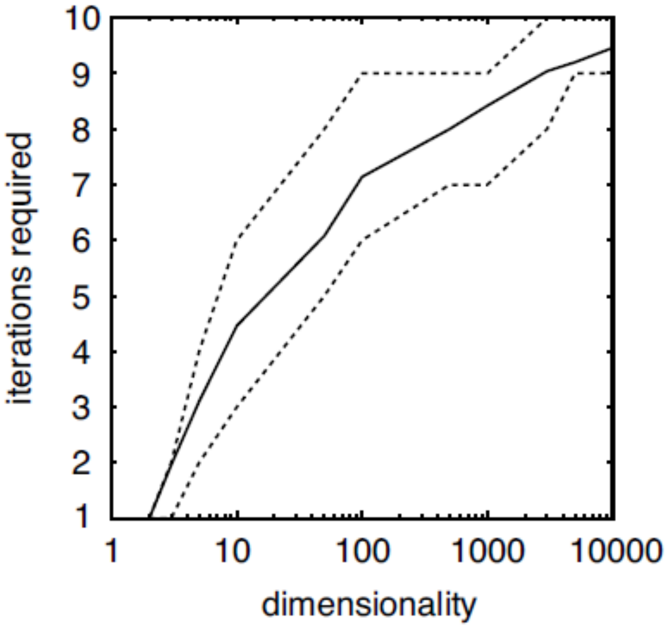
- Non-negative

$$s_i >= 0$$

Convergence of the Projection Step

- Worst-case
  - Desired sparseness 0.9, initial sparseness 0.1

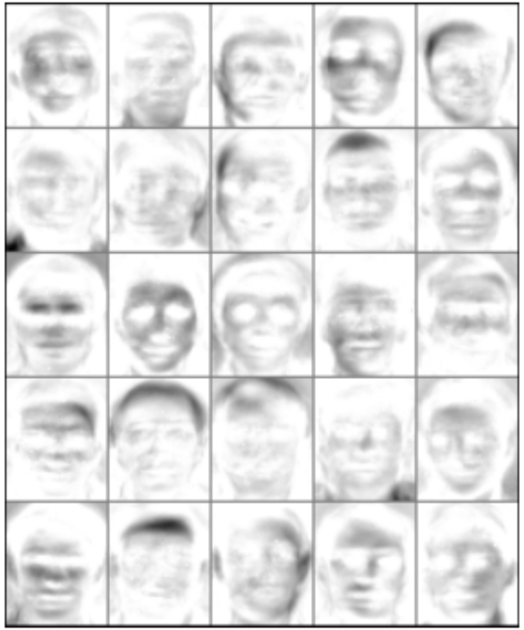
- Iterations grow slowly with dimensionality



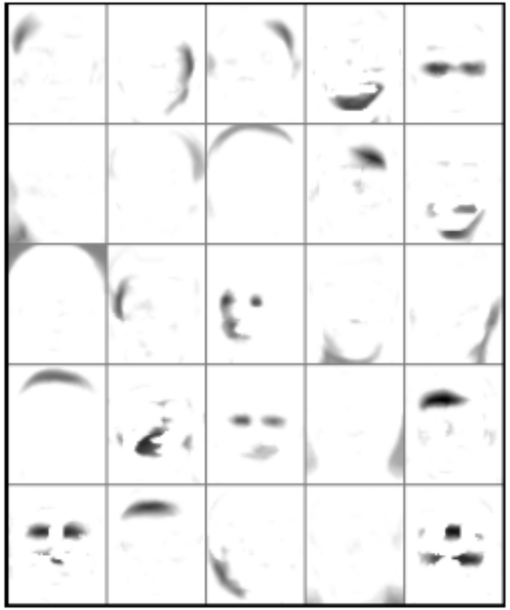
Experiment with Sparseness Constraints

- Features learned from the ORL face image database
- Sparseness level of the basis images were set 0.75.

Standard NMF



NMF with Sparseness Constraints



Using NMF in document clustering

Data matrix

news	paper	
news	paper	
news	paper	

=

Basis matrix

policy	algorithm
president	improved
peace	technical

X

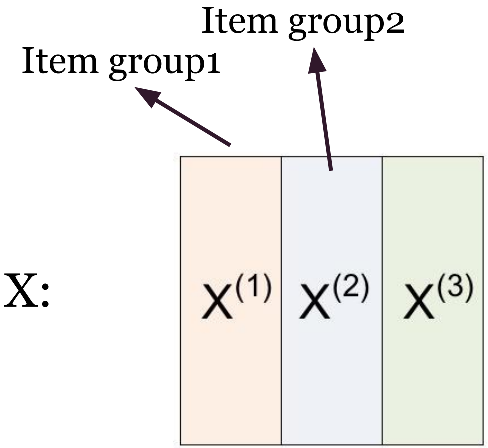
Coefficient matrix

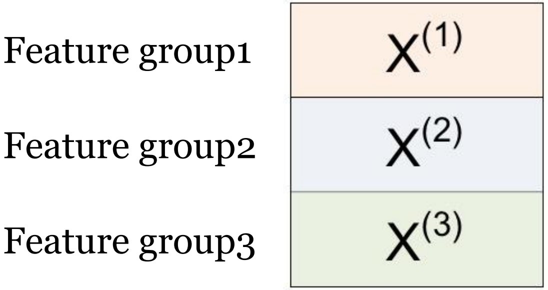
>10	0	0
2	>15	>8

A different approach
Group Sparsity in NMF

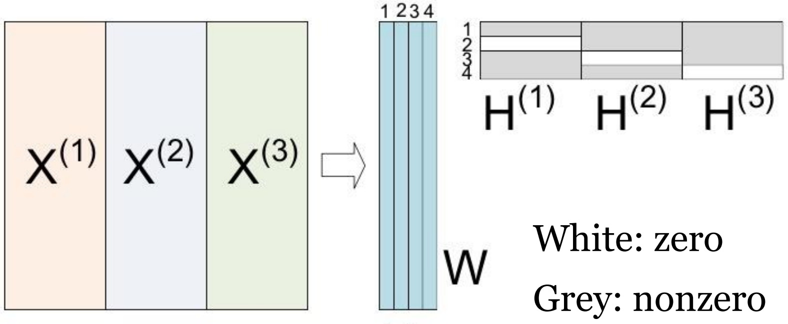
- Input data is group-structured. (Prior information)

Data matrix X:

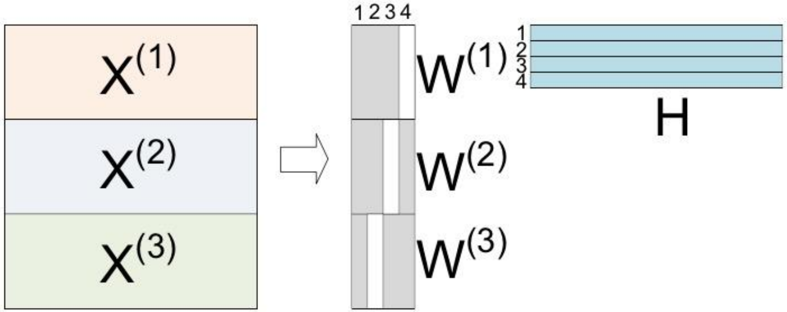




What's group sparsity?



(a)



(b)

Group sparsity:

Share same sparsity pattern in factor matrix.

Ex:

Reconstruct X(1): only 1st, 3rd, and 4th basis components are used.

Benefit:

Better understanding.

More intuitive.

How to promote group sparsity?

Sparse coding

- Add constraint
- Penalty term induces sparsity

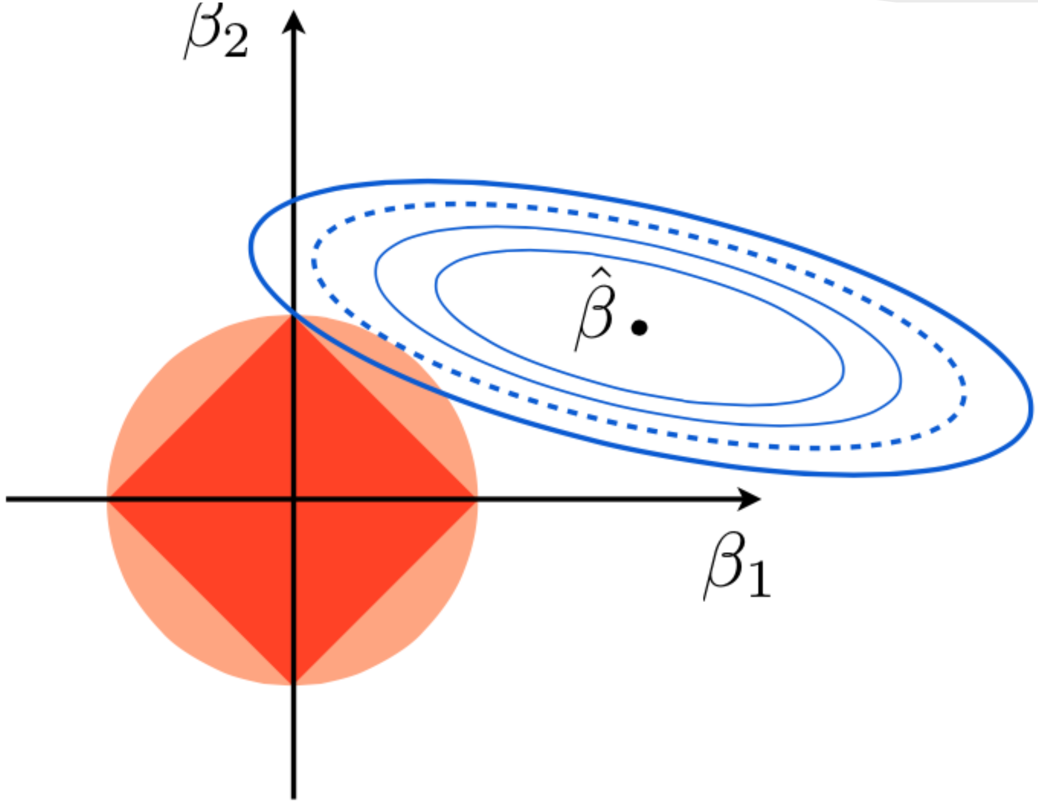
$$\frac{1}{2}\|\mathbf{X}-\mathbf{AS}\|^2+\lambda\sum_{ij}S_{ij}$$

Lq-norm $$\|x\|_p=\left(\sum_{i\in\mathbb{N}}|x_i|^p\right)^{1/p}$$

L1-norm $$\|\boldsymbol{x}\|_1:=\sum_{i=1}^{n}|x_i|.$$

L1-norm constraint $$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} f(\mathbf{W},\mathbf{H})\ +\alpha\sum_{j=1}^{n}\|\mathbf{h}_{\cdot j}\|$$

How it works?



$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W}, \mathbf{H}) \; + \alpha \sum_{j=1}^{n} \|\mathbf{h}_{\cdot j}\|$$

$L_{1,q}$ norm

Definition:

$$\|\mathbf{Y}\|_{1,q} = \sum_{j=1}^{a} \|\mathbf{y}_{j\cdot}\|_q = \|\mathbf{y}_{1\cdot}\|_q + \cdots + \|\mathbf{y}_{a\cdot}\|_q \,.$$

L1,q - norm of a matrix is the sum of Lq -norms of its rows.

L1,q -norm promotes as many number of zero rows as possible to appear in Y

$L_{1,q}$ mixed norm constraints

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2}\sum_{b=1}^{B}\left\|\mathbf{X}^{(b)} - \mathbf{W}\mathbf{H}^{(b)}\right\|_F^2.$$

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} f(\mathbf{W},\mathbf{H}) + \alpha\|\mathbf{W}\|_F^2 + \beta\sum_{b=1}^{B}\left\|\mathbf{H}^{(b)}\right\|_{1,q}.$$

Data matrix are divided into submatrix.

$\|\mathbf{W}\|_F^2$ is used to prevent the elements of W from growing arbitrarily large.

β: control the strength of constraint

Demonstration

- Comparison among 4 kinds of constraints.
- Factorize with group sparsity and see what W and H we get

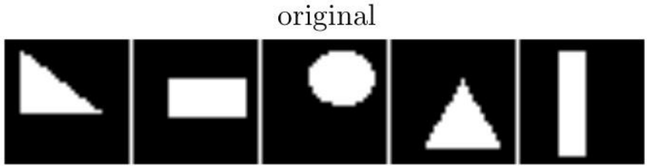
$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W}, \mathbf{H}) + \alpha \|\mathbf{W}\|_F^2 + \beta \|\mathbf{H}\|_F^2 ,$$ Frobenius norm

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W}, \mathbf{H}) + \alpha \|\mathbf{W}\|_F^2 + \beta \sum_{j=1}^{n} \|\mathbf{h}_{\cdot j}\|_1^2 .$$ L1 norm

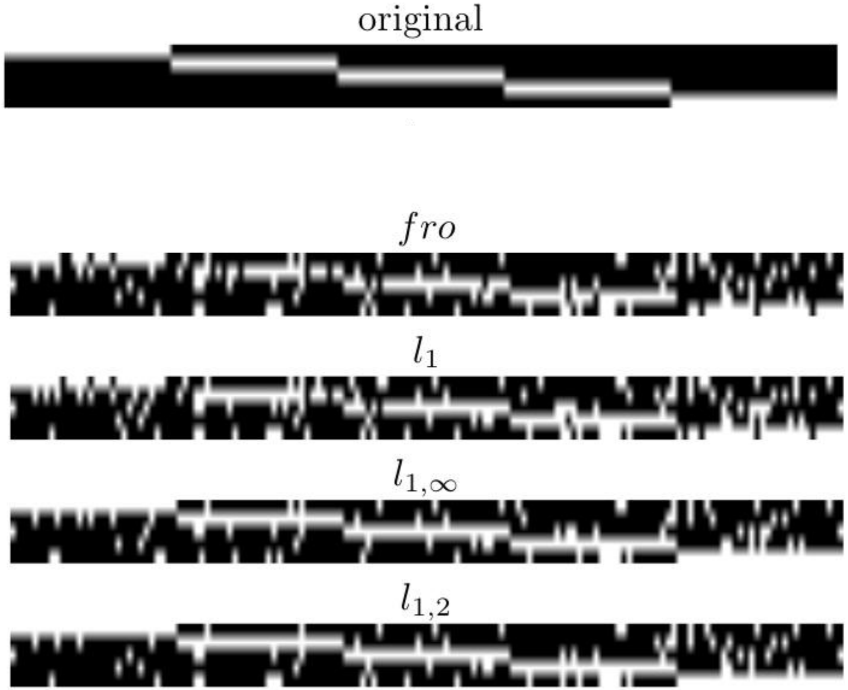
$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W}, \mathbf{H}) + \alpha \|\mathbf{W}\|_F^2 + \beta \sum_{b=1}^{B} \left\| \mathbf{H}^{(b)} \right\|_{1\infty}$$ Mixed norm with q=∞

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W}, \mathbf{H}) + \alpha \|\mathbf{W}\|_F^2 + \beta \sum_{b=1}^{B} \left\| \mathbf{H}^{(b)} \right\|_{1 2}$$ Mixed norm with q=2

Demonstration (2)

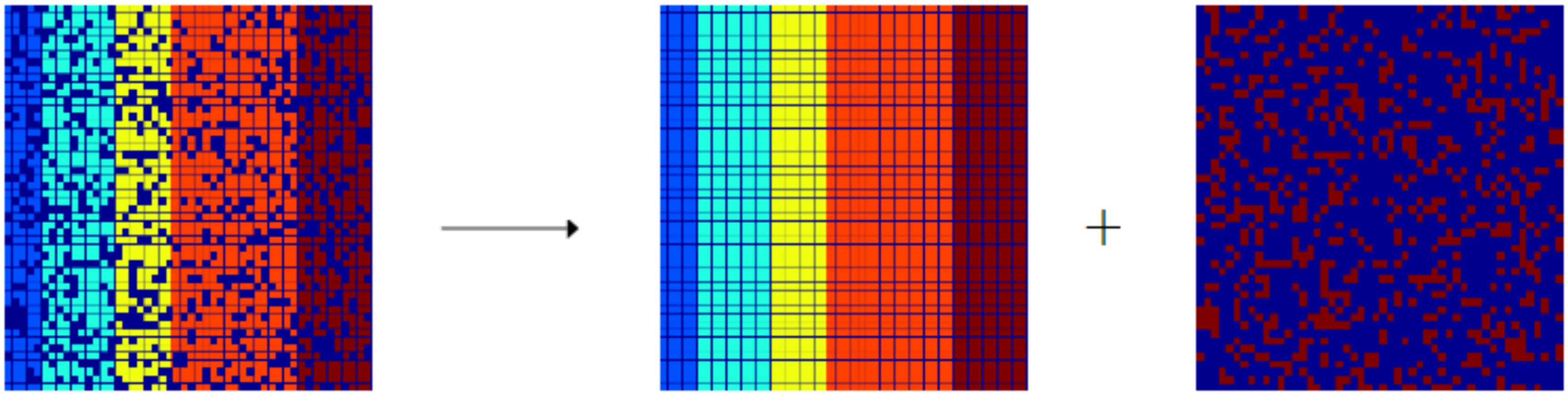


- Frobenius norm and L1 norm cannot successfully recover the group structure.
- Misinterpretation about the role of latent components.
- Sparsity destroyed.



Recovered coefficient matrices

Low-rank Data Recovery



Matrix of corrupted observations *Underlying low-rank matrix* *Sparse error matrix*

Example of Group Sparse Error

Non-negative Low-Rank and Group Sparse Matrix Factorization

- Approximate rank with Nuclear Norm
- Assume Error Group Sparse

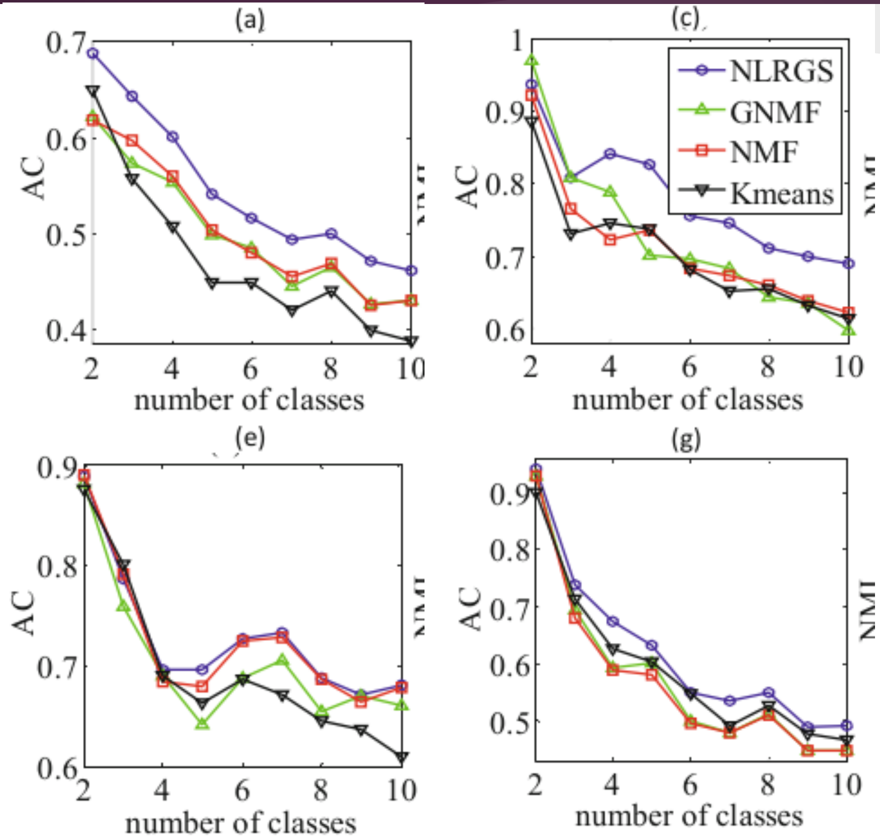
$$\min_{Z\geq 0, D\geq 0} \|Z\|_* + \lambda \|E\|_g, s.t., X = DZ + E.$$

Complexity

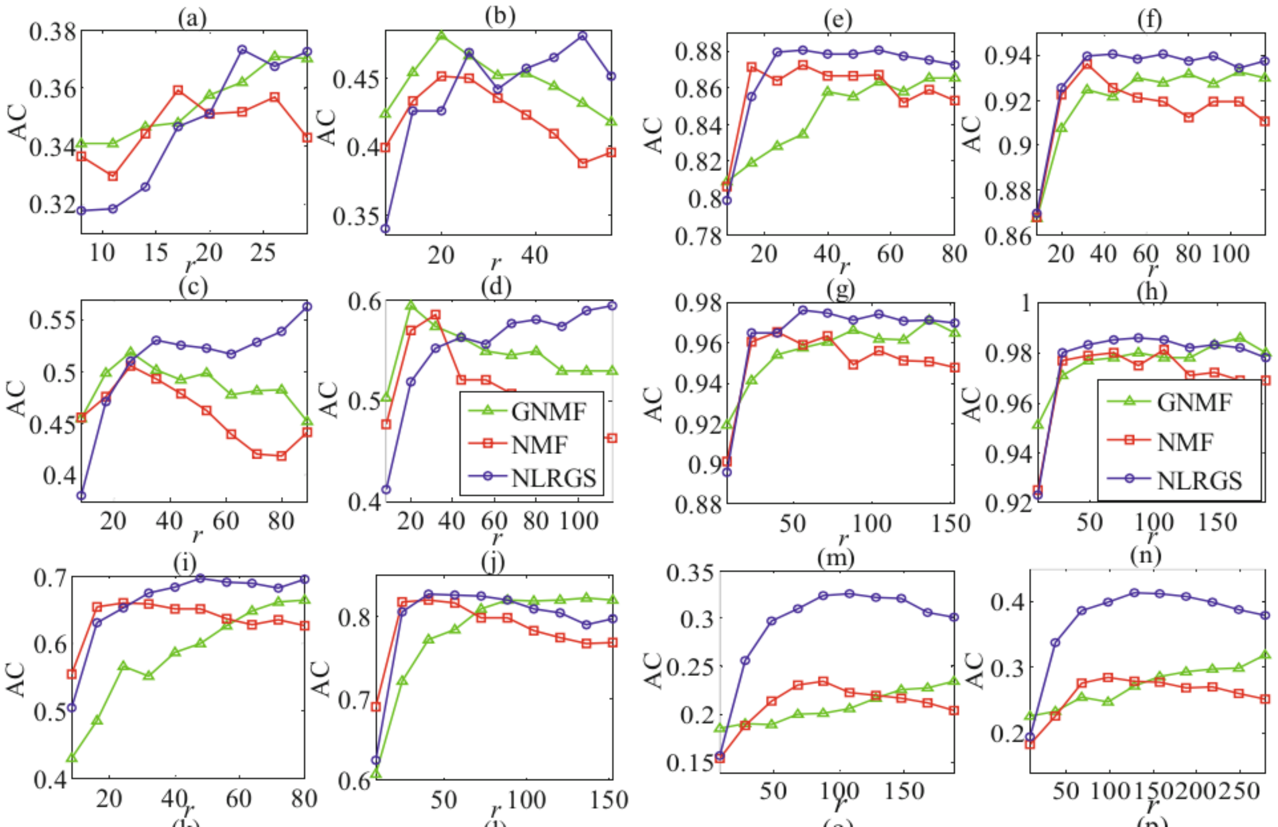
$$O(mnr + nr^2 + mnr^2 + 2n^2r + mn) + K \times O(mr^2 + m^2r).$$

- Algorithm solved with gradient descent
  - K is the number of iterations
- $K < r$ & $r << n,m$

Experiment: Face Clustering



Experiment: Face Recognition



1. Sparsity is desirable in data reduction
2. Various methods fail
3. Two Improve Methods Presented
    a. Constraining the decomposition
    b. Group Sparse Coding

Conclusion