Weighted Low Rank Approximations

Nathan Srebro and Tommi Jaakkola

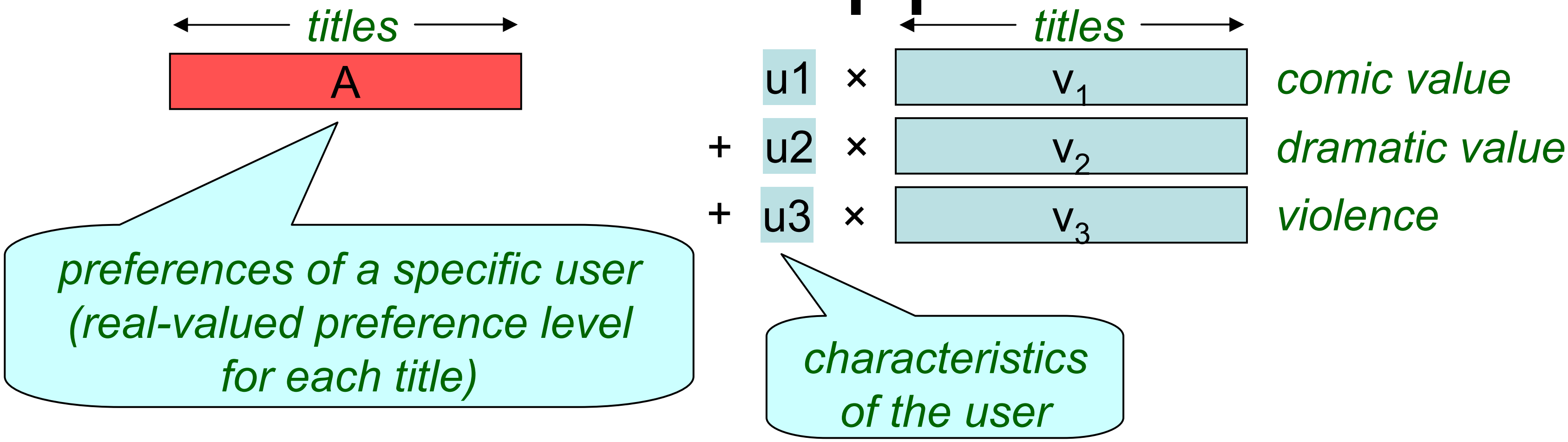
Computer **S**cience and **A**rtificial **I**ntelligence **L**aboratory

Massachusetts Institute of Technology

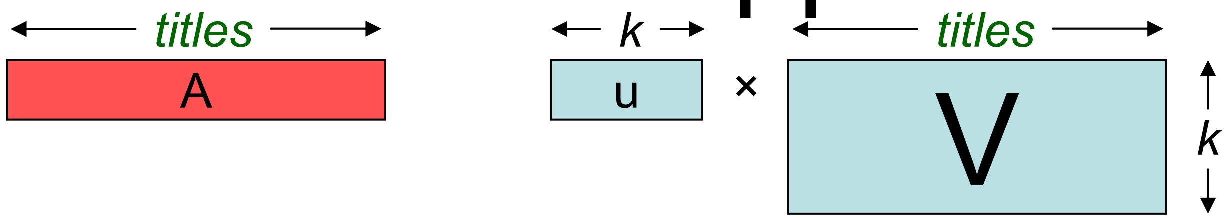
Weighted Low Rank Approximations

- What is a ‘weighted low rank approximation’?
  - What is a ‘low rank approximation’?
- Why weighted low rank approximations?
- How do we find a weighted low rank approximation?
- What can we do with weighted low rank approximations?

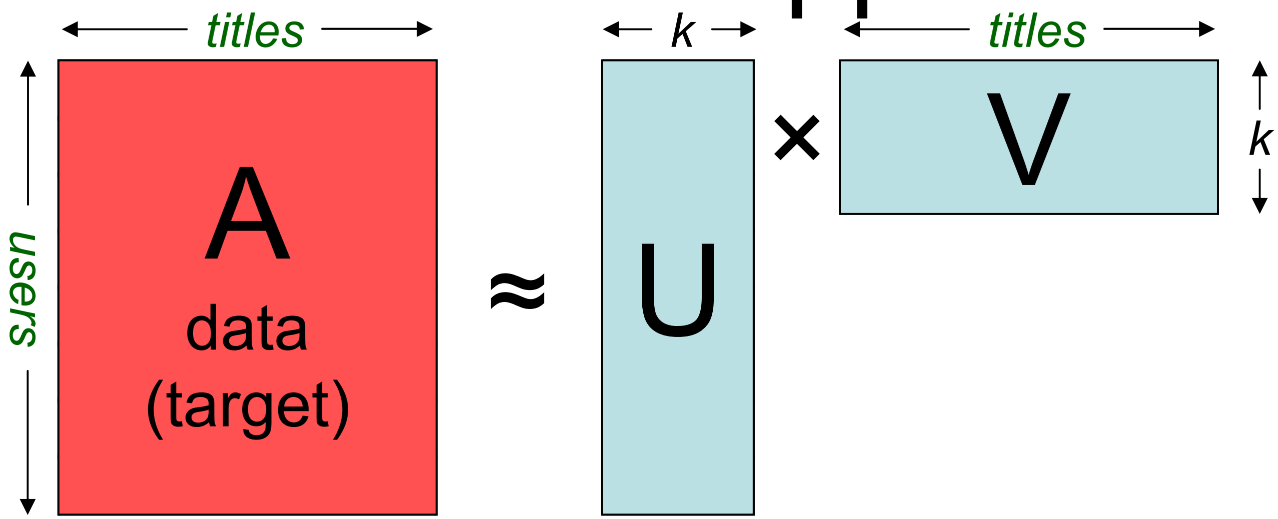
Low Rank Approximation



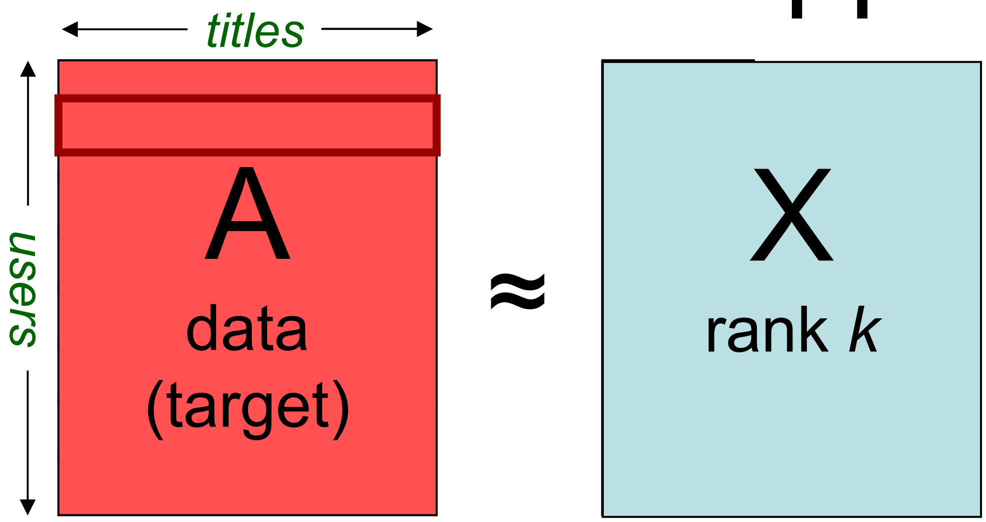
Low Rank Approximation

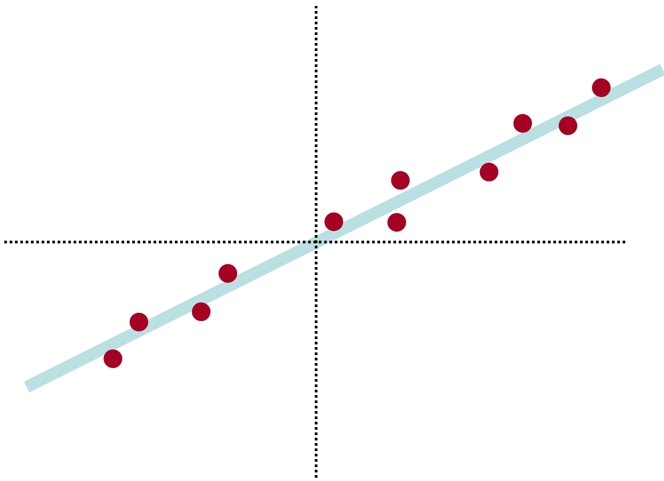


Low Rank Approximation

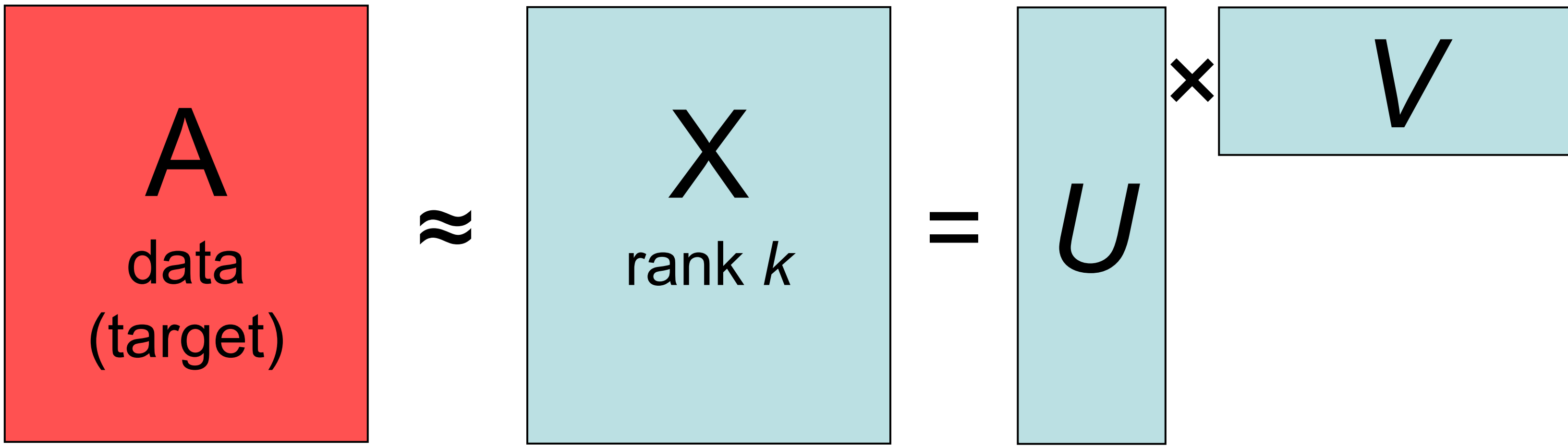


Low Rank Approximation



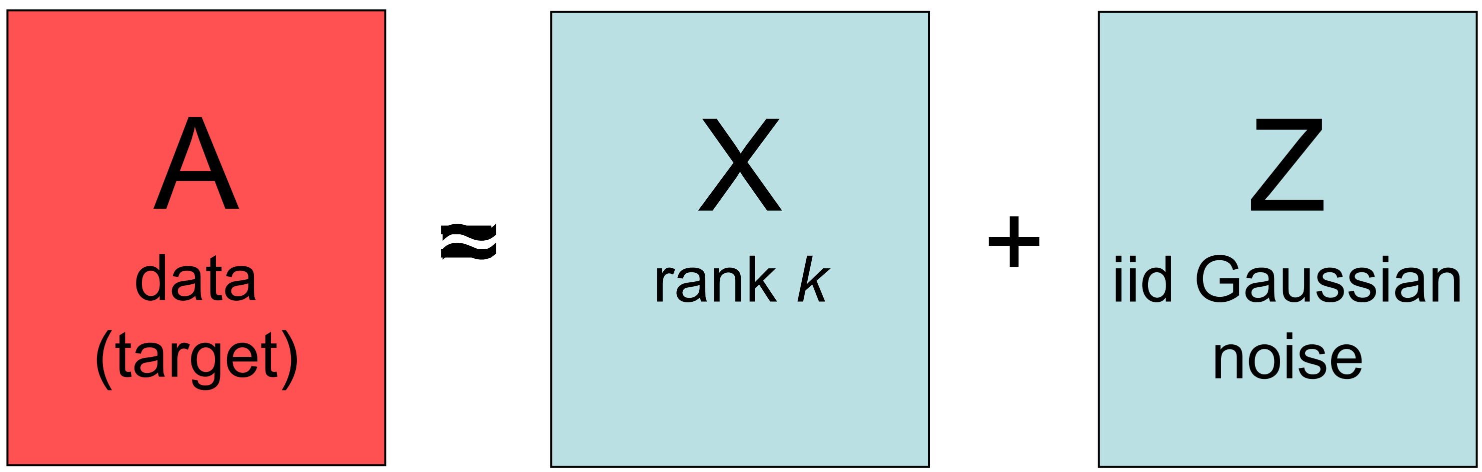


Low Rank Approximation



- Compression (mostly to reduce processing time)
- Prediction: collaborative filtering
- Reconstructing latent signal
  - biological processes through gene expression
- Capturing structure in a corpus
  - documents, images, etc
- Basic building block, e.g. for non-linear dimensionality reduction

Low Rank Approximation



$$\log L(X;A) = \sum_{ij} \log P(A_{ij} \mid X_{ij}) = \frac{-1}{2\sigma^2} \sum_{ij} (A_{ij} - X_{ij})^2 + const$$

Max likelihood low-rank matrix given iid Gaussian noise

⟹ low-rank *X* minimizing sum-squared error

given explicitly in terms of SVD

Weighted Low Rank Approximation

A data (target) = **X** rank k + **Z** Gaussian noise

$$Z_{ij} \sim N(0, \sigma_{ij}^2)$$

$$\log L(X; A) = \sum_{ij} \log P(A_{ij} \mid X_{ij}) = -\sum_{ij} w_{ij} (A_{ij} - X_{ij})^2 + const$$

$$w_{ij} = 1/\sigma_{ij}^2$$

Find low-rank matrix minimizing weighted sum-squared-error

[Young 1940]

Weighted Low Rank Approximation

low-rank matrix minimizing weighted sum-squared error

- External information about noise variance for each measurement

  e.g. in gene expression analysis

- Missing data (0/1 weights)

  collaborative filtering

- Different number of samples

  e.g. separating style and content [Tenenbaum Freeman 00]

- Varying importance of different entries

  e.g. 2D filters [Shpak 90, Lu et al 97]

- Subroutine for further generalizations

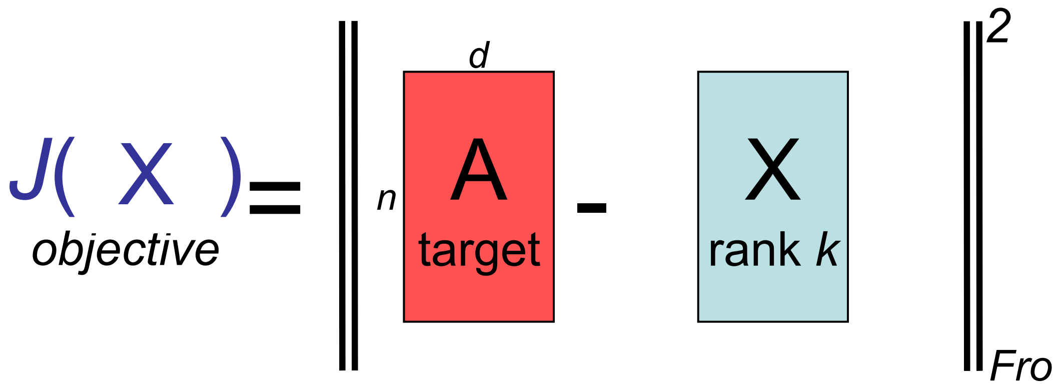
How?

Given A and W, find rank *k* matrix X minimizing the weighted sum-square difference

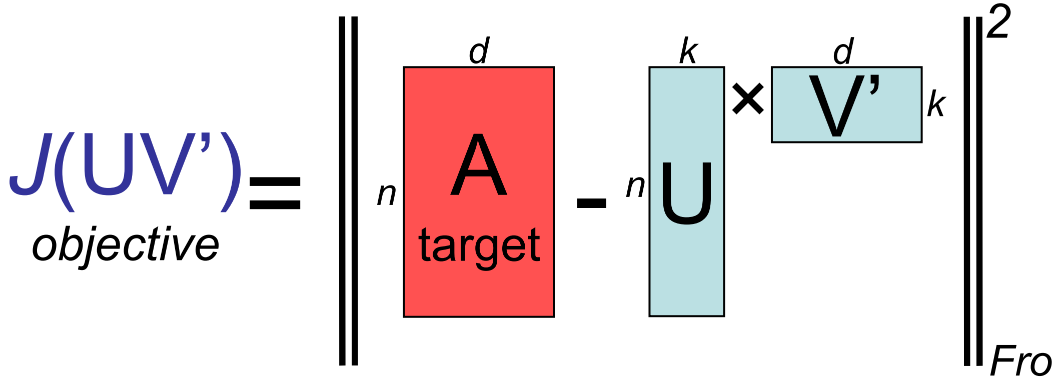
$$\sum_{ij} W_{ij}(A_{ij} - X_{ij})^2$$

(Unweighted) Low Rank Approximation

(Unweighted) Low Rank Approximation



(Unweighted) Low Rank Approximation



$$\frac{\partial J}{\partial U} = 2(UV'-A)V = 0$$

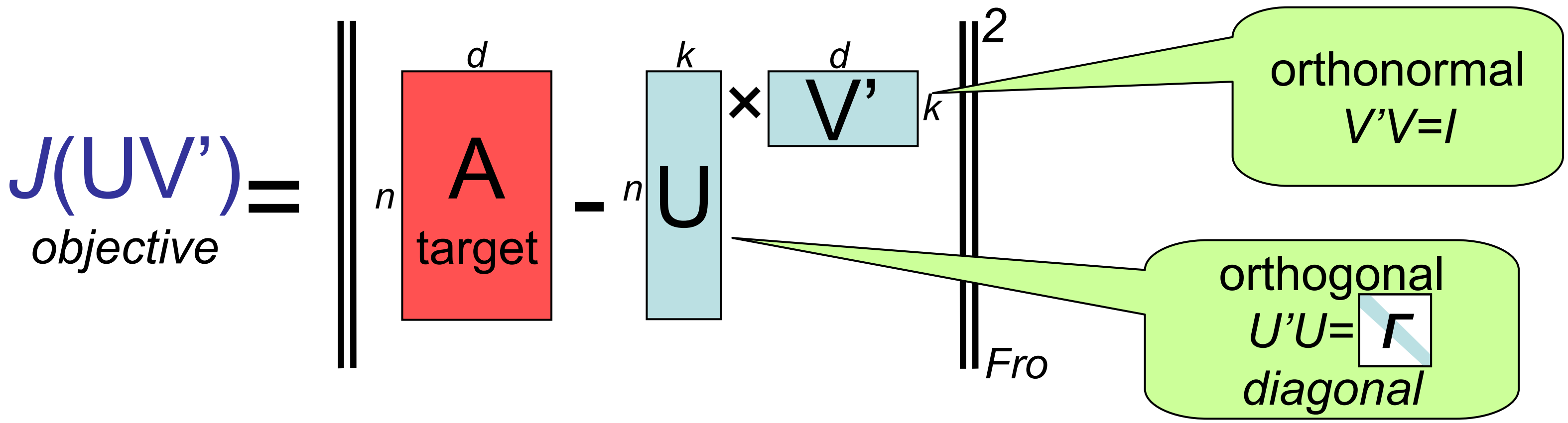
$$\frac{\partial J}{\partial V} = 2(VU'-A')U = 0$$

UV’V=AV

V =A’U

(Unweighted) Low Rank Approximation

$$\frac{\partial J(UV')}{\partial U,V} = 0 \iff$$ *U, V* are correspondingly spanned by eigenvectors of *AA'* and *A'A*



$$\frac{\partial J}{\partial U} = 2(UV' - A)V = 0$$

$$\frac{\partial J}{\partial V} = 2(VU' - A')U = 0$$

$$UV'V = AV$$

$$V\Gamma = A'AV$$

(Unweighted) Low Rank Approximation

$\frac{\partial J(UV')}{\partial U,V} = 0$ $\Leftrightarrow$ U,V are correspondingly spanned by eigenvectors of AA' and $A'A$

Global Minimum $\Leftrightarrow$ U,V are correspondingly spanned by leading eigenvectors

U,V spanned by eigenvectors of AA' and AA'

$$\Longrightarrow \quad J(UV') = \|A - UV'\|^2 = \sum_{\substack{\text{unselected} \\ \text{eigenvalues}}}$$

(Unweighted) Low Rank Approximation

$\frac{\partial J(UV')}{\partial U,V} = 0$ $\Leftrightarrow$ *U,V* are correspondingly spanned by eigenvectors of *AA'* and *A'A*

Global Minimum $\Leftrightarrow$ *U,V* are correspondingly spanned by leading eigenvectors

All other fixed points are saddle points

(no non-global local minima)

Weighted Low Rank Approximation

$$J(UV') = \sum_{ij} W_{ij} (A - UV')_{ij}^2$$

$$\frac{\partial J}{\partial U} = 2((UV' - A) \cdot W)V = 0$$

$$U_i = A_i \underline{W_i} V (V' \underline{W_i} V)^{-1}$$

If rank(W)=1, $(V'\underline{W_i}V)$ can be simultaneously diagonalized
⇒ eigen-methods apply [IraniAnandan 2000]

Otherwise, eigen-methods cannot be used,
solutions are not incremental

WLRA: Optimization

$$J(UV') = \sum_{ij} W_{ij}(A - UV')_{ij}^2$$

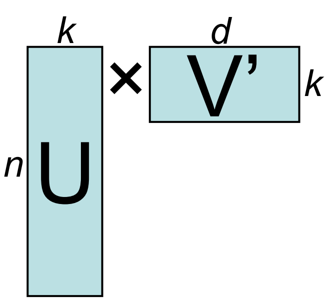
For fixed *V*, find optimal *U*

For fixed *U*, find optimal V

$$J^*(V) = \min_U J(UV')$$

$$\tfrac{\partial}{\partial V} J^*(V) = 2U^{*\prime}((U^* V' - Y) \otimes W)$$

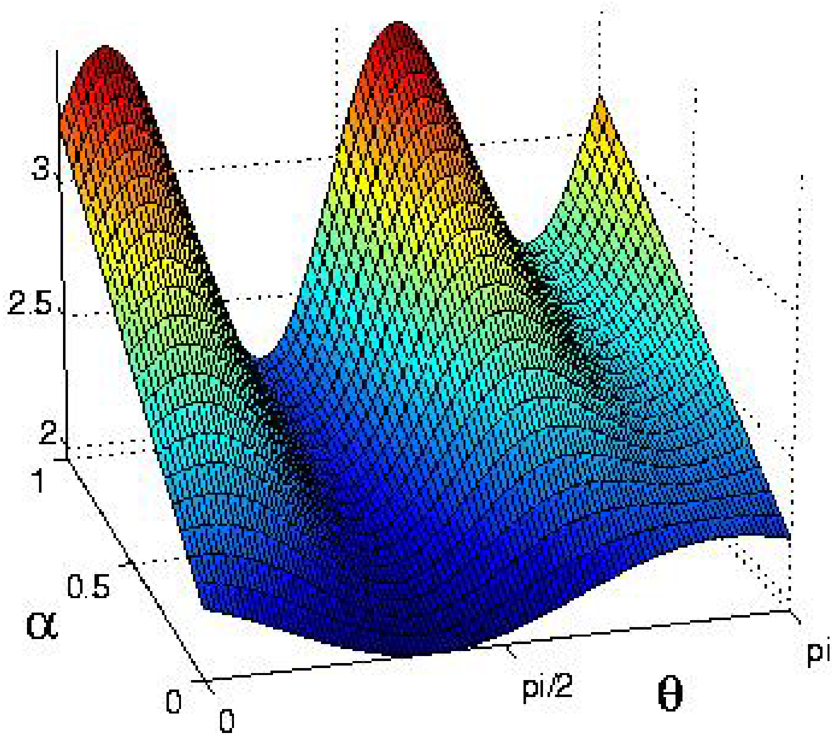
Conjugate gradient descent on *J**

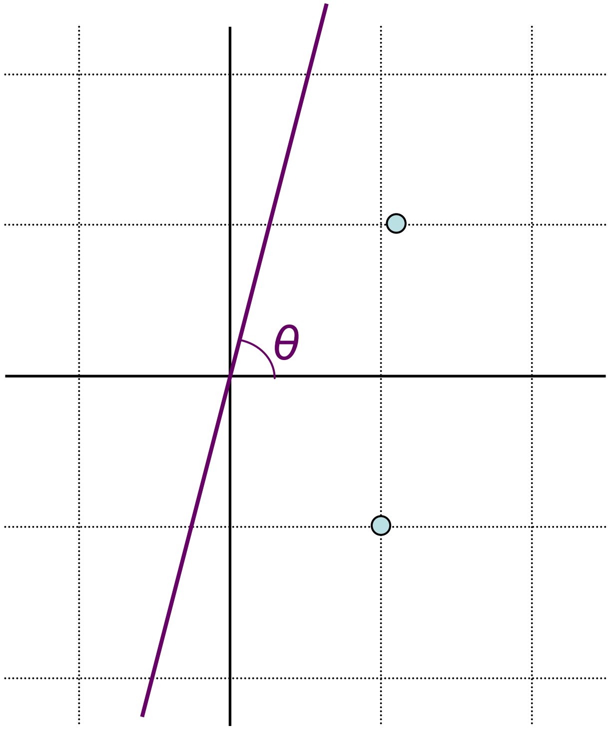


Optimize *kd* parameters instead of *k(d+n)*

Local Minima in WLRA

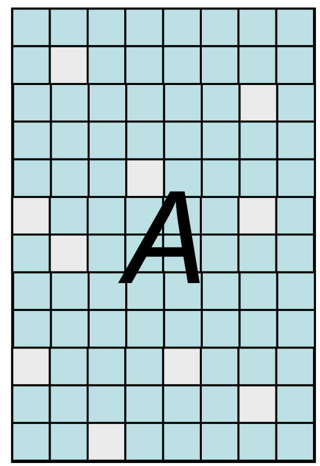
$$A = \begin{bmatrix} 1 & 1.1 \\ 1 & -1 \end{bmatrix} \quad W = \begin{bmatrix} 1+\alpha & 1 \\ 1 & 1+\alpha \end{bmatrix}$$





WLRA: An EM Approach

0/1 weights:



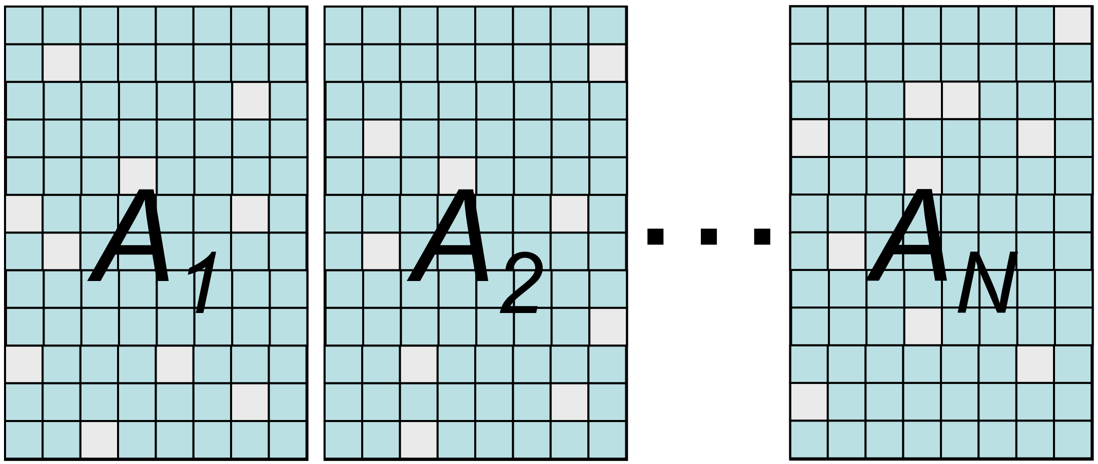
$$A = X + Z$$

observations

parameters

N(0,1) noise

WLRA: An EM Approach



$A_r = X + Z_r$

same low-rank X for all targets A_r

independent noise Z_r for each target A_r

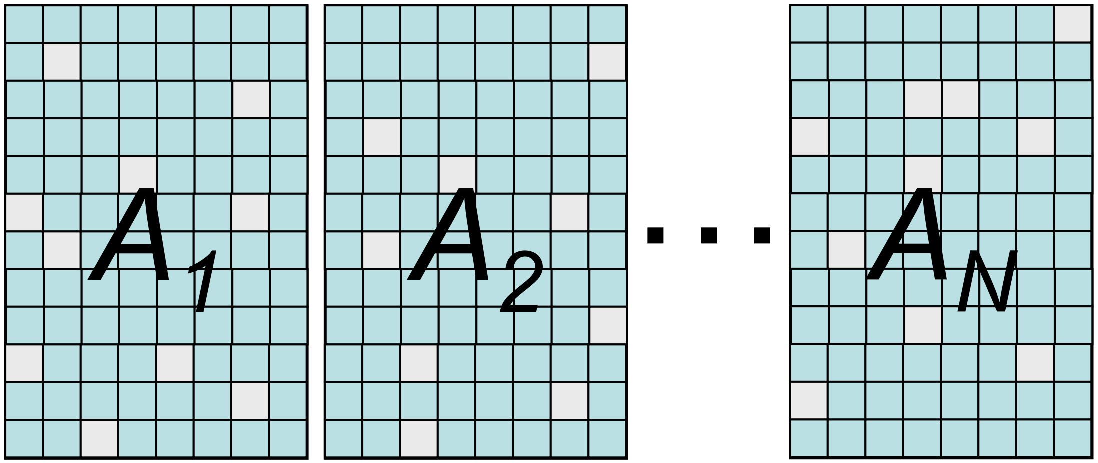
Expectation Step:

missing $A_r[i,j] \leftarrow X[i,j]$

Maximization Step:

$$X \leftarrow \mathrm{LRA}\left(\frac{1}{N} \sum_r A_r \right)$$

WLRA: An EM Approach



$A_r = X + Z_r$

Expectation Step:

missing $A_r[i,j] \leftarrow X[i,j]$

Maximization Step:

$$X \leftarrow \mathrm{LRA}\left(\frac{1}{N}\sum_r A_r\right)$$

integer

WLRA(A,W) with $W[i,j]=w[i,j]/N$

$\Rightarrow$ $A_r[i,j]=A[i,j]$, *or missing if* $w[i,j]<r$

$$X \leftarrow \mathrm{LRA}\big(W \otimes A + (1-W) \otimes X\big)$$

WLRA: Optimization

$$J(UV') = \sum_{ij} W_{ij}(A - UV')^2_{ij}$$

For fixed V, find optimal U

For fixed U, find optimal V

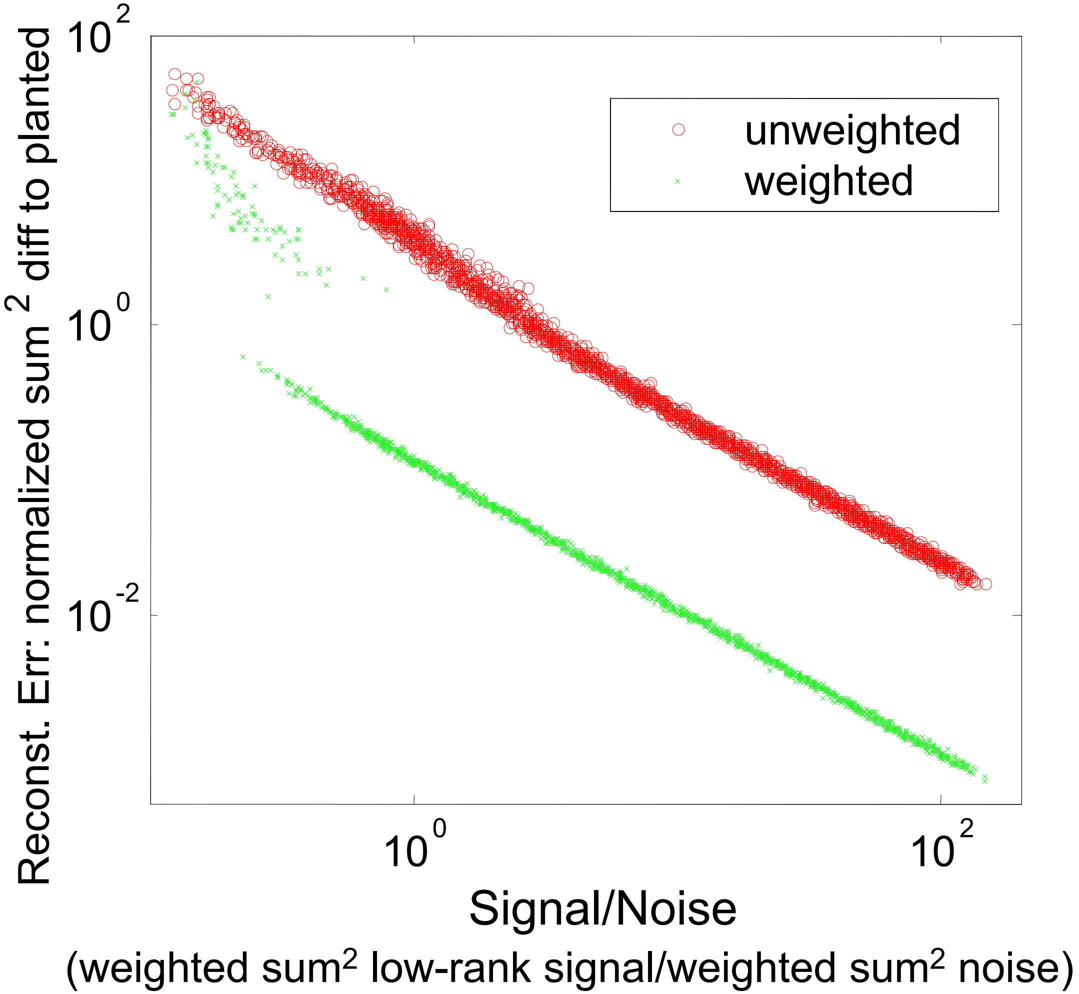
$$J^*(V) = \min_U J(UV')$$

$$\tfrac{\partial}{\partial V} J^*(V') = 2U^{*\prime}((U^* V' - A) \otimes W)$$

Conjugate gradient descent on J^*

$X \leftarrow \mathrm{LRA}\big(W \otimes A + (1 - W) \otimes X\big)$

Estimations of a "planted" X with Non-Identical Noise Variances



Collaborative Filtering of Joke Preferences (“Jester”)

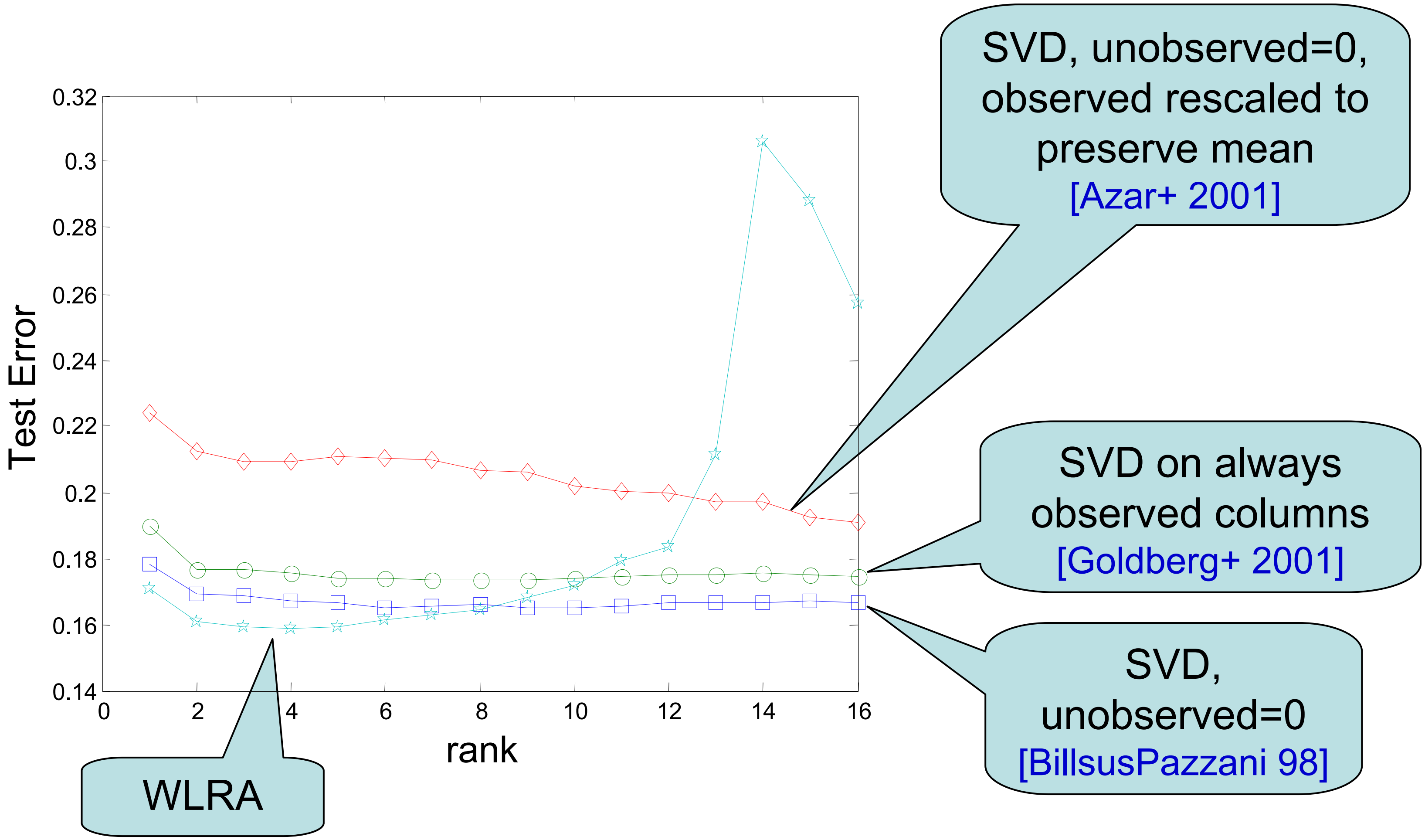
jokes

users

+1	-2	+3			+6			+5			-2		
+6	+4	-7				-2		+4		-3			
-5	+7	-2	+6	+2				-1	-7				
-1	+2	+4		+1			+3		-3			+7	
+2	-3	-6			+4								
0	-2	-1				+5	-2		+6				+1
+5	+2	-7		+3					-2	-3	-1		
+7	-3	-2	-6		+2	+2		-4		-7			+3
-4	+2	+2				+1					+2	-5	
-3	-6	-1							+6	+5			-7
+1	-6	+7	-3	-2			+6	+5					
+9	+4	+5								-2		-6	
+7	-5	+2			+4		-2			-3			
+1	+6	-2						+2	+4		-2		+3
-3	+3	-6		-4	-9		+7			0		+2	
-4	+3	+2	-2			0					-4		

[Goldberg 2001]

Collaborative Filtering of Joke Preferences (“Jester”)



WLRA as a Subroutine for Other Cost Functions

$$\sum_{ij} (A_{ij} - X_{ij})^2$$

$\Rightarrow$ Low Rank Approximation (PCA)

$$\sum_{ij} W_{ij} (A_{ij} - X_{ij})^2$$

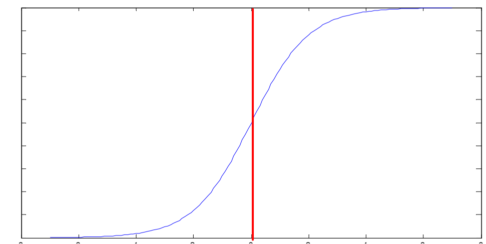
$\Rightarrow$ Weighted Low Rank Approximation

$$\sum_{ij} Loss(A_{ij}, X_{ij})$$

Logistic Low Rank Regression

[Collins+ 2002, Gordon 2003, Schein+ 2003]

$$\Pr\left(\underset{\textit{observed}}{Y} \;\middle|\; \underset{\text{rank } k}{X} \right) = g\left(\underset{\text{rank } k}{X} \right)$$



$$g(x) = 1 \big/ 1 + e^{-x}$$

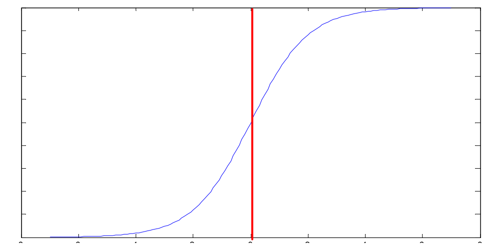
$$Loss(Y, X) = \sum_{ij} \log g(Y_{ij}, X_{ij})$$

$$\approx -\sum_{ij} \frac{g(Y_{ij} X_{ij}^{(t-1)}) g(-Y_{ij} X_{ij}^{(t-1)})}{2} \left(X_{ij}^{(t)} - \left(X_{ij}^{(t-1)} + \frac{Y_{ij}}{g(Y_{ij} X_{ij}^{(t-1)})} \right) \right)^2 + Const$$

Logistic Low Rank Regression

[Collins+ 2002, Gordon 2003, Schein+ 2003]

$$\Pr\left(\underset{\text{observed}}{Y} \,\middle|\, \underset{\text{rank } k}{X}\right) = g\left(\underset{\text{rank } k}{X}\right)$$



$$g(x) = 1/(1+e^{-x})$$

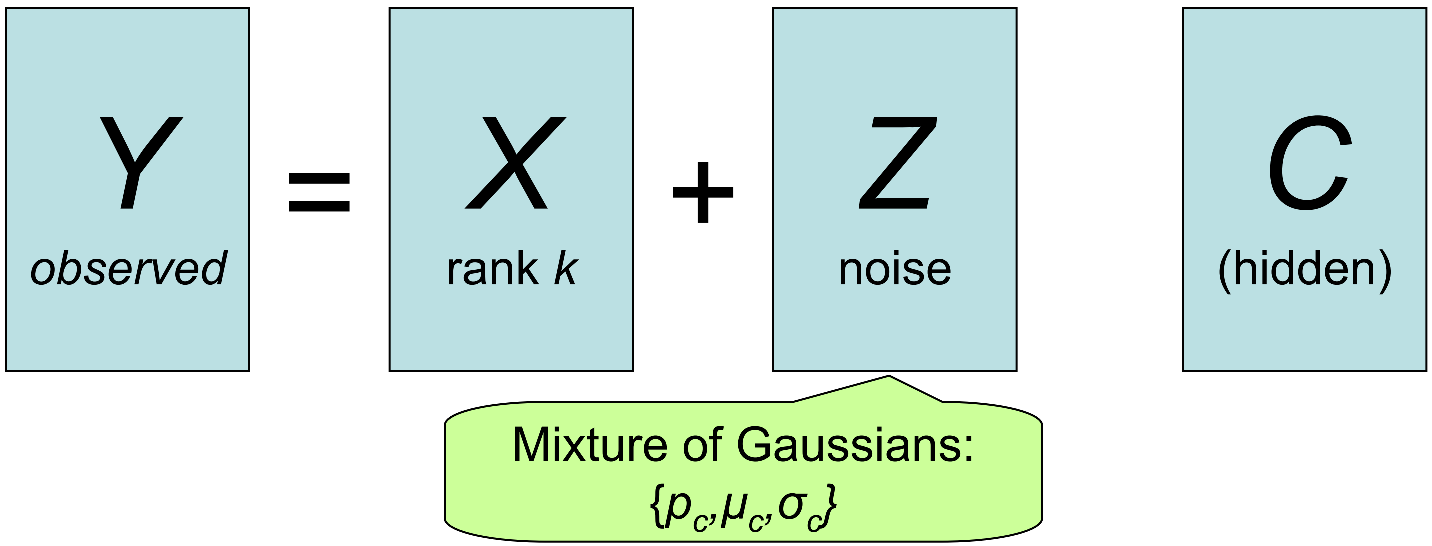
$$Loss(Y, X) = \sum_{ij} \log g(Y_{ij}, X_{ij})$$

$$\approx -\sum_{ij} W_{ij}^{(t)} \left(X_{ij}^{(t)} - \left(A_{ij}^{(t)} \right) \right)^2 + Const$$

$$W_{ij}^{(t)} = \frac{g(Y_{ij} X_{ij}^{(t-1)})\, g(-Y_{ij} X_{ij}^{(t-1)})}{2}$$

$$A_{ij}^{(t)} = X_{ij}^{(t-1)} + \frac{Y_{ij}}{g(Y_{ij} X_{ij}^{(t-1)})}$$

Maximum Likelihood Estimation with Gaussian Mixture Noise



E step: calculate posteriors of C

M step: WLRA with

$$W_{ij} = \sum_c \frac{\Pr(C_{ij} = c)}{\sigma_C^2} \qquad A_{ij} = Y_{ij} + \sum_c \frac{\Pr(C_{ij} = c)\mu_C}{\sigma_C^2} \Big/ W_{ij}$$

Weighted Low Rank Approximations

Nathan Srebro   Tommi Jaakkola

- Weights often appropriate in low rank approximation
- WLRA more complicated than unweighted LRA
  - Eigenmethods (SVD) do not apply
  - Non-incremental
  - Local minima
- Optimization approaches:
  - Alternate optimization
  - Gradient methods on $J^*(V)$
  - EM method: $X \leftarrow \mathrm{LRA}(W \otimes A + (1-W) \otimes X)$
- WLRA useful as subroutine for more general loss functions

www.csail.mit.edu/~nati/LowRank/

END

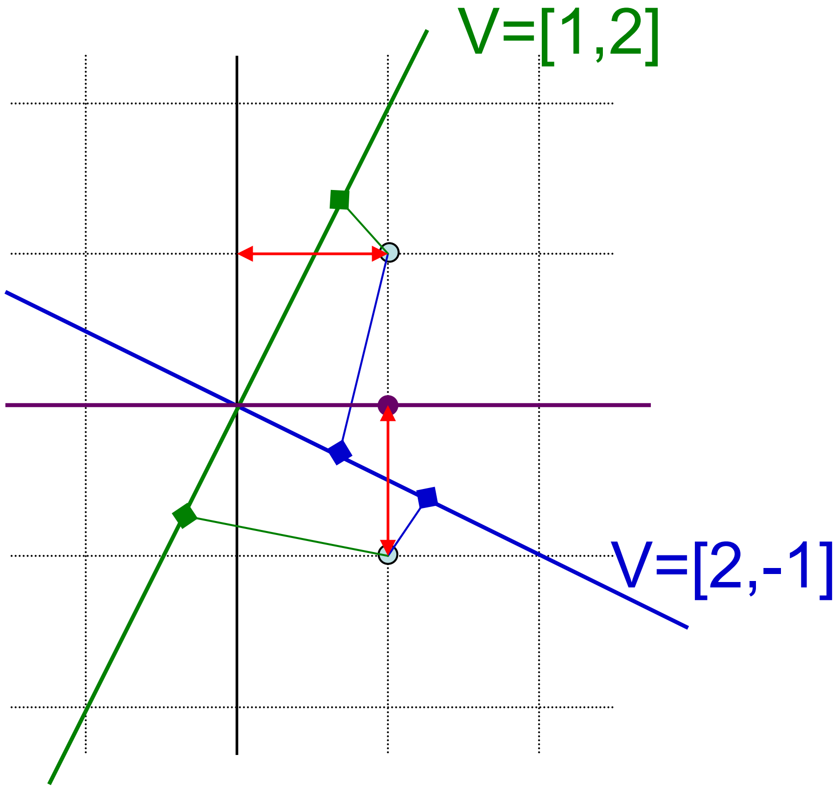
Local Minima in WLRA

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad W = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$J=3$

$$X = \begin{bmatrix} 2/3 & -1/3 \\ 1\,1/3 & -2/3 \end{bmatrix} \quad X = \begin{bmatrix} 2/3 & 1\,1/3 \\ -1/3 & -2/3 \end{bmatrix}$$

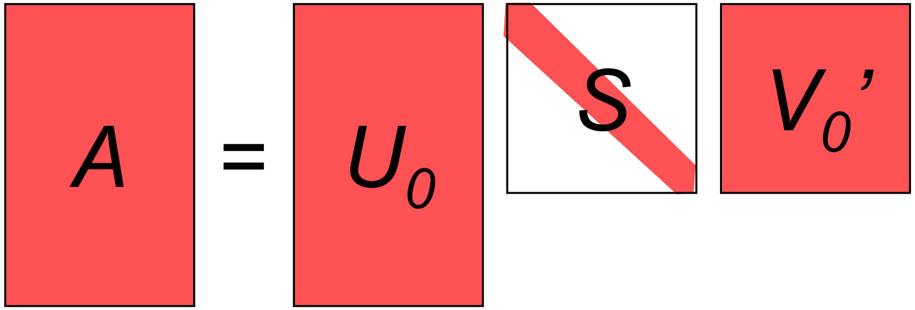
$J=2\tfrac{1}{3}$ $\qquad$ $J=2\tfrac{1}{3}$



(Unweighted) Low Rank Approximation

$$\frac{\partial J(UV')}{\partial U,V} = 0 \quad \Leftrightarrow$$

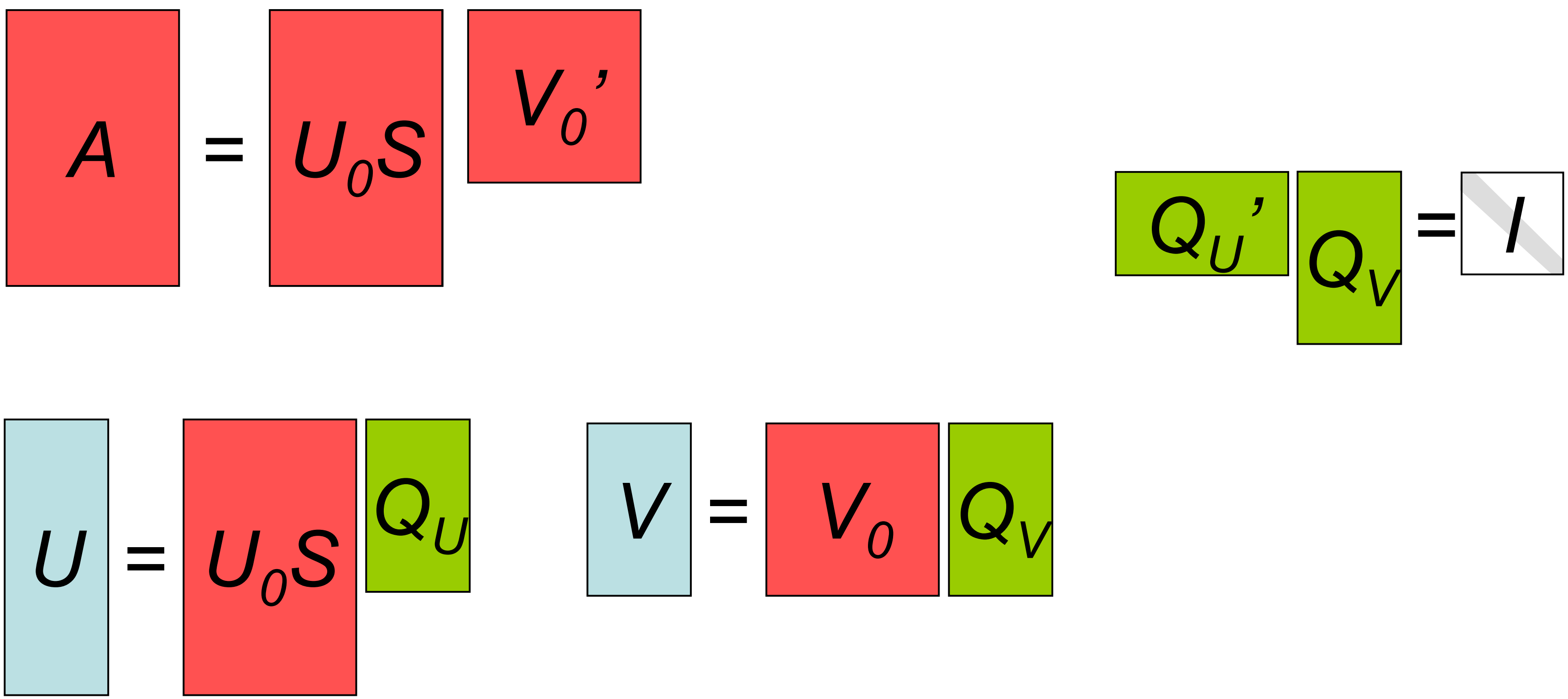
U,V are correspondingly spanned by eigenvectors of *AA'* and *A'A*



(Unweighted) Low Rank Approximation

$$\frac{\partial J(UV')}{\partial U,V} = 0 \quad \Leftrightarrow$$

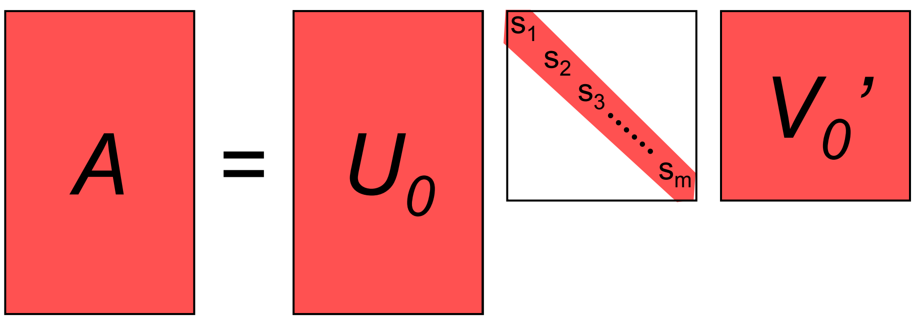
U,V are correspondingly spanned by eigenvectors of *AA'* and *A'A*



(Unweighted) Low Rank Approximation

$$\frac{\partial J(UV')}{\partial U,V} = 0 \quad \Leftrightarrow$$

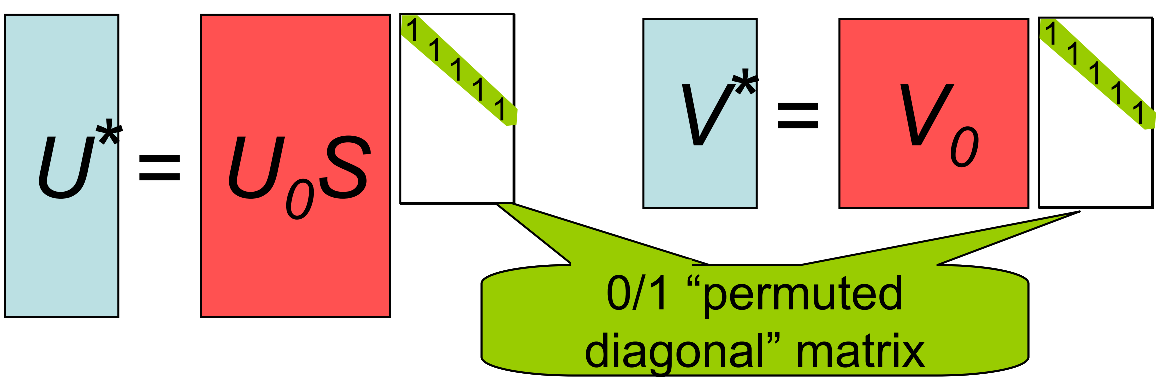
U,V are correspondingly spanned by eigenvectors of *AA'* and *A'A*



$$\|A\|^2 = \sum_{\alpha} s_\alpha^2$$

$$\|UV'\|^2 = \sum_{\text{selected } \alpha} s_\alpha^2$$

$$\|A - UV'\|^2 = \sum_{\text{unselected } \alpha} s_\alpha^2$$



(Unweighted) Low Rank Approximation

$$\frac{\partial J(UV')}{\partial U,V} = 0 \iff$$ *U,V* are correspondingly spanned by eigenvectors of *AA'* and *A'A*

Global Minimum $\iff$ *U,V* are correspondingly spanned by leading eigenvectors