Weighted Low Rank Approximations

Nathan Srebro and Tommi Jaakkola

Computer Science and Artificial Intelligence Laboratory
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

preferences of a specific user (real-valued preference level for each title)

A

u1 \times v_1
+ u2 \times v_2
+ u3 \times v_3

titles

comic value
dramatic value
violence

characteristics of the user
Low Rank Approximation

\[ A \approx uv^T \]
Low Rank Approximation

\[ \text{data (target)} \approx U \times V \]
Low Rank Approximation

\[ \mathbf{A} \approx \mathbf{X} \text{ rank } k \]
Low Rank Approximation

\[ A \approx X \times V \]

- 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

\[ \sum_{ij} \log P(A_{ij} \mid X_{ij}) = \frac{-1}{2\sigma^2} \sum_{ij} (A_{ij} - X_{ij})^2 + \text{const} \]

Max likelihood low-rank matrix given iid Gaussian noise

\[ \text{low-rank } X \text{ minimizing sum-squared error} \]

given explicitly in terms of SVD
Weighted Low Rank Approximation

\[ A = X_{\text{rank } k} + Z \]

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

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

\[ w_{ij} = 1 / \sigma_{ij}^2 \]

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

\[
J(X) = \| X - \text{target} - \text{rank } k \|_F^2
\]
(Unweighted) Low Rank Approximation

\[ J(UV') = \| \text{target} - UV' \|_{Fro}^2 \]

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

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

\[ UV'V = AV \]

\[ V = A'U \]
(Unweighted) Low Rank Approximation

\[ \frac{\partial J(UV')}{\partial U,V} = 0 \quad \leftrightarrow \quad U,V \text{ are correspondingly spanned by eigenvectors of } AA' \text{ and } A'A \]

**Objective**

\[ J(UV') = \| A - UV' \|_F^2 \]

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

\[ UV'V = AV \]

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

\[ V \Gamma = A'AV \]

**Orthogonal**

\[ U'U = \Gamma \]

**Orthonormal**

\[ V'V = I \]
(Unweighted) Low Rank Approximation

\[ \frac{\partial J(UV')}{\partial U,V} = 0 \quad \leftrightarrow \quad U,V \text{ are correspondingly spanned by eigenvectors of } AA' \text{ and } A'A \]

Global Minimum \quad \leftrightarrow \quad U,V \text{ are correspondingly spanned by leading eigenvectors}

\[ U,V \text{ spanned by eigenvectors of } AA' \text{ and } A'A' \]

\[ J(UV') = \| A - UV' \|^2 = \sum \text{unselected eigenvalues} \]
(Unweighted) Low Rank Approximation

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

Global Minimum \iff \quad U,V \text{ are correspondingly spanned by leading eigenvectors}

All other fixed points are saddle points

(no non-global local minima)
Weighted Low Rank Approximation

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

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

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

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

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

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

For fixed \( V \), find optimal \( U \)
For fixed \( U \), find optimal \( V \)

\[ J^*(V) = \min_U J(UV') \]
\[ \frac{\partial}{\partial V} J^*(V) = 2U'^*((U^*V' - Y) \otimes W) \]

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

\[ A = X + Z \]

0/1 weights:

- Observations
- Parameters
- \( N(0,1) \) noise
WLRA: An EM Approach

\textbf{Expectation Step:}
missing $A_r[i,j] \leftarrow X[i,j]$

\textbf{Maximization Step:}

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

$A_r = X + Z_r$

- same low-rank $X$ for all targets $A_r$
- independent noise $Z_r$ for each target $A_r$
WLRA: An EM Approach

**Expectation Step:**
missing $A_r[i,j] \leftarrow X[i,j]$

**Maximization Step:**

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

$$A_r = X + Z_r$$

$W[1,2,\ldots,N]$ with $W[i,j] = w[i,j]/N$

⇒ $A_r[i,j] = A[i,j]$, or missing if $w[i,j] < r$

$$X \leftarrow LRA(W \otimes A + (1-W) \otimes X)$$
WLRA: Optimization

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

For fixed \( V \), find optimal \( U \)
For fixed \( U \), find optimal \( V \)

\[ J^*(V) = \min_{U} J(UV') \]

\[ \frac{\partial}{\partial V} J^*(V') = 2U^*((U^*V' - A) \otimes W) \]

Conjugate gradient descent on \( J^* \)

\[ X \leftarrow \text{LRA}(W \otimes A + (1-W) \otimes X) \]
Estimations of a “planted” $X$ with Non-Identical Noise Variances

(Weighted sum$^2$ low-rank signal/weighted sum$^2$ noise)
Collaborative Filtering of Joke Preferences ("Jester")

![Matrix Representation of Joke Preferences]

[Goldberg 2001]
Collaborative Filtering of Joke Preferences (“Jester”)
WLRA as a Subroutine for Other Cost Functions

\[
\sum_{ij} (A_{ij} - X_{ij})^2 \quad \Rightarrow \text{Low Rank Approximation (PCA)}
\]

\[
\sum_{ij} W_{ij} (A_{ij} - X_{ij})^2 \quad \Rightarrow \text{Weighted Low Rank Approximation}
\]

\[
\sum_{ij} \text{Loss}(A_{ij}, X_{ij})
\]
**Logistic Low Rank Regression**

\[ \text{Pr} \begin{pmatrix} Y_{\text{observed}} \\ X_{\text{rank }k} \end{pmatrix} = g \begin{pmatrix} X_{\text{rank }k} \end{pmatrix} \]

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

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

\[ g(x) = \frac{1}{1+e^{-x}} \]
Logistic Low Rank Regression

\[ \text{Pr}\left( \begin{array}{c} Y \\ \text{observed} \end{array} \right) \left| \begin{array}{c} X \\ \text{rank } k \end{array} \right) = g\left( \begin{array}{c} X \\ \text{rank } k \end{array} \right) \]

\[ g(x) = \frac{1}{1 + e^{-x}} \]

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

\[ \approx -\sum_{ij} \frac{W_{ij}^{(t)}}{2} \left( X_{ij}^{(t)} - A_{ij}^{(t)} \right)^2 + \text{Const} \]

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

\[ A_{ij}^{(t)} = X_{ij}^{(t-1)} + \frac{Y_{ij}}{g(Y_{ij}X_{ij}^{(t-1)})} \]
Maximum Likelihood Estimation with Gaussian Mixture Noise

\[ Y_{\text{observed}} = X_{\text{rank } k} + Z_{\text{noise}} + C_{(\text{hidden})} \]

Mixture of Gaussians: \( \{p_c, \mu_c, \sigma_c\} \)

**E step:** calculate posteriors of \( C \)

\[ W_{ij} = \sum_c \frac{\Pr(C_{ij} = c)}{\sigma_c^2} \]

\[ A_{ij} = Y_{ij} + \sum_c \frac{\Pr(C_{ij} = c)\mu_c}{\sigma_c^2} \]

**M step:** WLRA with
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: $X \leftarrow \text{LRA}(W \otimes A + (1-W) \otimes X)$
• WLRA useful as subroutine for more general loss functions

www.csail.mit.edu/~nati/LowRank/
END
Local Minima in WLRA

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

\[ X = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ 1\frac{1}{3} & -\frac{2}{3} \end{bmatrix} \quad \text{J=2}\frac{1}{3} \]

\[ V=[1,2] \]

\[ V=[2,-1] \]
(Unweighted) Low Rank Approximation

\[ \frac{\partial J(UV')}{\partial U,V} = 0 \quad \leftrightarrow \quad U, V \text{ are correspondingly spanned by eigenvectors of } AA' \text{ and } A'A \]

\[ A = U_0 S V_0' \]
(Unweighted) Low Rank Approximation

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

\[
A = U_0S \quad V' = V_0'
\]

\[
U = U_0S \quad Q_U \quad V = V_0 \quad Q_V
\]

\[
Q_{U'} \quad Q_V = \begin{bmatrix} I \end{bmatrix}
\]
(Unweighted) Low Rank Approximation

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

\[
A = U_0 V_0' \\
U^* = U_0 S \\
V^* = V_0
\]

\[
\|A\|^2 = \sum_{\alpha} s^2_{\alpha} \\
\|UV'\|^2 = \sum_{\text{selected } \alpha} s^2_{\alpha} \\
\|A - UV'\|^2 = \sum_{\text{unselected } \alpha} s^2_{\alpha}
\]

0/1 “permuted diagonal” matrix
(Unweighted) Low Rank Approximation

\[
\frac{\partial J(UV')}{\partial U,V} = 0 \quad \leftrightarrow \quad U, V \text{ are correspondingly spanned by eigenvectors of } AA' \text{ and } A'A
\]

Global Minimum \quad \leftrightarrow \quad U, V \text{ are correspondingly spanned by leading eigenvectors}

\[
U = U_0 S \quad \quad V = V_0
\]