Robust Nonnegative Matrix Factorization: Modern Dimension Reduction Procedure for Big Noisy Data Set

Yifan Xu

May 13, 2014
SAMSI – LDHD Transition Workshop

Joint with Jiayang Sun, Kenneth K. Lopiano and S. Stanley Young
**Brief preview:**

- **Non-negative Matrix Factorization (NMF):** Unsupervised dimension reduction procedure for large p data. Basis vectors spanning the subspace are likely to represent separate parts of the original data.

- **Robust NMF (rNMF):** Variation of NMF that is robust to outliers. In spirit similar to trimmed linear regressions.

**Layout:**

- PCA vs. NMF
- NMF vs. rNMF
- Examples
  - multiple corrupted images
  - single corrupted image
Large p data

Gene expressions

Face images

p genes

n samples

400 face images
4096 pixels each

Yale face database B

Spectra
Reduce Dimension

- Dimension reduction
- Feature selection
Dimension reduction through matrix factorization

\[ X \approx W^\top H \]

(standardize if necessary)

One obs. \( \approx \) k column vectors in \( W \)

**PCA:** \( W, H \) arbitrary signs

**NMF:** \( W, H \) nonnegative
Principal Component Analysis (PCA)

\[
X \approx L \times R
\]

- Orthogonal principal components
- PC’s ordered by variability explained

Teapot graphs from: http://gregemmerich.wordpress.com/2013/04/17/demystifying-big-data/
Non-negative Matrix Factorization (NMF)

- Given $X \geq 0$, a $p \times n$ data matrix, for $k < \min(k, n)$ find $W, H \geq 0$ such that:

$$
(H, W) = \arg\min_{W \geq 0, H \geq 0} \left\{ \|X - WH\|_F^2 \right\}
$$

- Each observation is expressed as a non-negative linear combination of non-negative basis vectors.

- Gives parts-based decomposition.

- No closed form solution. $H$ and $W$ are fitted iteratively.
PCA vs. NMF 1 of 3

PCA vs. NMF 2 of 3

<table>
<thead>
<tr>
<th></th>
<th>Obs 1</th>
<th>Obs 2</th>
<th>Obs 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene5</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

PCA vs. NMF 2 of 3

PCA

NMF

W

H
PCA vs. NMF 3 of 3

- 75 by 75 pixels each
- 11 images total
- Vectorize each image
- $X = 5625$ by 11 matrix
NMF in Biology: Metagene Decomposition

Each row represents one gene

Each column is one tissue sample

2 metagenes

Metagene coefficients for each of the original sample

Two distinct clusters

PCA vs. NMF Summary

• **PCA:**
  o Orthonormal basis
  o Ranked basis vectors
  o Coefficients of decomposition have arbitrary signs
  o Not intuitive to interpret

• **NMF:**
  o Parts-based reconstruction, providing a more meaningful decomposition
  o Easy to interpret
  o Mathematically more challenging than PCA
  o Reconstruction is not unique

Robust Penalized NMF (rNMF)

\[ (H, W) = \arg \min_{W \geq 0, H \geq 0} \left\{ \|X - WH\|_{F, \gamma}^2 + \beta \sum_{j=1}^{n} \|H_{.j}\|_1^2 \right\} \]

- Discovers and trims outliers: 4 variations (to be defined).
- Enforces sparsity of H through L1 norm.
rNMF – How to trim?

\[(H, W) = \arg \min_{W \geq 0, H \geq 0} \left\{ \|X - WH\|_{F, \gamma}^2 + \beta \sum_{j=1}^{n} \|H_{.j}\|_1^2 \right\}\]

rNMF objective function is not convex.

Minimizing algorithms fit \(W\) and \(H\) alternately until convergence.

◆ Solutions:
Alg-1. Trim rows (features).
Alg-2. Trim columns (observations).
Alg-3. Trim outlier cells.
Alg-4. Smooth outlier cells.

◆ Question:
How to trim (or detect) outliers?

... iteratively
rNMF

Fit

Basis
rNMF in Single Image Compression
rNMF – Tumor Image compression
rNMF – Single Face Image – 3 Types of Corruptions

Original image
192 by 168

Reconstruction
k = 15

Original Image: Yale face data base B: http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

New Challenge
A Image Cleaning Procedure by Multi-Level rNMF (MrNMF)

Semi-automatic NMF
Details omitted
Comparison

MSE Comparison - Tumor Image Simulations with 5% corruptions
PLEASE HANDLE WITH CARE

Trimmin

↑ this side up ↑

The Good

The Bad

The Ugly
Summary

- rNMF finds intuitive low dimensional representations from data set consists of large number of features and potential corruptions/outliers (images, gene expressions, documents)

- rNMF allows full automation and semi-supervised controls i.e. Can be forced to keep/drop certain features during decomposition

- Results produced by rNMF can be used for
  - direct interpretation
  - clustering in low dimension
  - feature selection from basis vectors
  - classification / regression
  - ...

Acknowledgement

- The work benefits from valuable discussions among SAMSI LDHD Dimension Reduction sub-working group. Members of the working group include:
- Naomi S. Altman (Penn State U.)
- Lingsong Zhang (Purdue U.)
- Xingye Qiao (Binghamton U.)
- Kossi Edoh (North Carolina A&T State U.)
Thank You