**Main Contribution:** We propose a novel method called Non-negative Multiple Matrix Factorization (NM2F), which integrates the information of multiple matrices and extracts common factors from the matrices.

- Generalize Non-negative Matrix Factorization (NMF) to decompose multiple matrices
- Improve generalization performance on factorizing a highly sparse target matrix
- Extract common factors of the target and auxiliary matrices simultaneously

**Problem:** Extract non-negative factors from a sparse matrix

NMF extracts base and coefficient factors from a target matrix \( X \)

Ex: Music listening data set

<table>
<thead>
<tr>
<th>Artist</th>
<th>Target Matrix ( X )</th>
<th>Matrix Factorization ( X \sim WH )</th>
</tr>
</thead>
<tbody>
<tr>
<td>User /</td>
<td>Artist / Topic k</td>
<td>Base ( W ) Coefficient ( H ) Artist /</td>
</tr>
</tbody>
</table>

\( x_{ij} \) is the listening count of user / on artist

**Solution:** Utilize complementary data as auxiliary matrices

Auxiliary matrices share the row or column indices with the target matrix

**Intuitive Explanation**

- **Row-wise information (Multi-Task like)**
- **Column-wise information (Multi-View like)**

- Auxiliary matrices mitigate the sparseness
- High generalization performance

**New Pattern**

1. Enable to factorize the highly sparse target matrix
2. Extract common factors among the target and auxiliary matrices

**Experiments**

**Synthetic data experiment:**

- Auxiliary matrices \( Y \) and \( Z \) improved average test log-likelihoods on \( X \) in the highly sparse situations

<table>
<thead>
<tr>
<th>Sparse</th>
<th>Dense</th>
<th>Near Dense</th>
<th>Dense</th>
<th>NM2F</th>
<th>VBMF</th>
<th>PMF</th>
<th>PAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.9%</td>
<td>-1.06 - 0.09</td>
<td>-1.24 ± 0.23</td>
<td>-2.90 ± 1.00</td>
<td>-7.13 ± 3.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99.9%</td>
<td>-1.07 ± 0.25</td>
<td>-42.45 ± 6.30</td>
<td>-14.55 ± 1.40</td>
<td>-16.25 ± 6.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99.9%</td>
<td>-0.86 ± 0.55</td>
<td>-41.35 ± 3.34</td>
<td>-15.30 ± 6.30</td>
<td>-12.85 ± 11.20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Real-world large data experiment:**

1. Social Curation (Togetter) data set:
   - NM2F improved the performance (\( \alpha=0.1, \beta=0.01 \))

<table>
<thead>
<tr>
<th>Data Set</th>
<th>NM2F</th>
<th>VBMF</th>
<th>PMF</th>
<th>PAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Togetter</td>
<td>12.97 ± 0.48</td>
<td>-21.27 ± 0.23</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

1.8 Million users x 235,000 stories matrix \( X \) is sparse (99.98%)

**NM2F**

- Relative artists and tags are simultaneously extracted

- Top 10 highest valued artists and tags
- Popular women singers and its related tags
- Classical UK rock
- Industrial rap and hip-hop

**Ex: A story of the Japan disaster**

Mass media, Japan and US government officials who post emergency information on the Japan disaster

**Mathematical Details:**

Objective function: minimizing reconstruction error

\[
\min_{W,H,A,B} D(X,Y,Z,W,H,A,B,A) = \alpha D(Y,A) + \beta D(Z,W,B)
\]

\[
D(X,Y,Z,W,H,A,B,A) = \sum_{i,j} d_{i,j}^{(X)} + \sum_{i,j} d_{i,j}^{(Y)} + \sum_{i,j} d_{i,j}^{(Z)}
\]

Where \( d_{i,j}^{(X)} \) is the squared difference between the \( i,j \)th element of \( X \) and \( \hat{X} \),

\[
\hat{X}_{i,j} = \sum_{k} w_{i,k} h_{k,j} - \sum_{k} a_{i,k} h_{k,j} - \sum_{k} w_{i,k} a_{j,k} + \sum_{k} w_{i,k} a_{j,k}
\]

\( \alpha \) and \( \beta \) are scaling parameters,

\( d \) is the generalized Kullback-Leibler divergence

**Existing methods fail when matrix X is sparse**

**Dense**

<table>
<thead>
<tr>
<th>Dense</th>
<th>Sparse</th>
<th>Poor generalization performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>NM2F</td>
<td></td>
</tr>
</tbody>
</table>

**Mathematical Explanation of NM2F**

**NM2F** is a generalization of NMF, for "Large" matrix including undefined block missing region

**NM2F**

- Auxiliary data

**Mathematical Details:**

Objective function: minimizing reconstruction error

\[
\min_{W,H,A,B} D(X,Y,Z,W,H,A,B,A) = \alpha D(Y,A) + \beta D(Z,W,B)
\]

\[
D(X,Y,Z,W,H,A,B,A) = \sum_{i,j} d_{i,j}^{(X)} + \sum_{i,j} d_{i,j}^{(Y)} + \sum_{i,j} d_{i,j}^{(Z)}
\]

Where \( d_{i,j}^{(X)} \) is the squared difference between the \( i,j \)th element of \( X \) and \( \hat{X} \),

\[
\hat{X}_{i,j} = \sum_{k} w_{i,k} h_{k,j} - \sum_{k} a_{i,k} h_{k,j} - \sum_{k} w_{i,k} a_{j,k} + \sum_{k} w_{i,k} a_{j,k}
\]

\( \alpha \) and \( \beta \) are scaling parameters,

\( d \) is the generalized Kullback-Leibler divergence

**Ex: Togetter story**

Mass media, Japan and US government officials who post emergency information on the Japan disaster