Sparse Matrix Factorization for Gene Expression Analysis
(Work in Progress)

Nati Srebro and Tommi Jaakkola
MIT EECS
Genes within cluster follow same expression pattern – deviation from cluster consensus is noise
Genes within cluster follow same expression pattern – deviation from cluster consensus is noise
• Transcriptional factors
• Regulatory cascades
• Responses / stimuli
• Processes
  • Protein complexes
  • Pathways
  • Cell activities
Modeling Data as Combinations of Factors

• Instead of being assigned to a cluster, each data vector is a linear combination of ‘factors’.
• ‘Factors’ represent basic structural components that are combined to get the data vectors.
Modeling Data as \text{Linear}
Combinations of Factors

\[ C \times F \approx A \]
Limitations of hierarchical clustering
SVD Analysis of Gene Expression Patterns

- Alter, Brown, Botstein: PNAS 2000
  - Yeast cell-cycle
- Raychaudhuri, Stuart, Altman: PSB 2000
  - Yeast cell-cycle and sporulation; serum-treated human fibroblast
  - Yeast cell-cycle
Expression of cell-cycle genes projected to leading two eigenfactors:

### True (Planted) Factors

<table>
<thead>
<tr>
<th>Eigenfactors</th>
<th>-0.6431</th>
<th>-0.9408</th>
<th>-0.9932</th>
<th>-0.0034</th>
<th>0.6628</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.4274</td>
<td>0.2364</td>
<td>-0.0917</td>
<td>0.9128</td>
<td>-0.0681</td>
</tr>
<tr>
<td></td>
<td>0.6019</td>
<td>-0.0967</td>
<td>-0.0332</td>
<td>0.3965</td>
<td>0.5372</td>
</tr>
<tr>
<td></td>
<td>-0.2036</td>
<td>0.2226</td>
<td>0.0622</td>
<td>-0.0977</td>
<td>0.5172</td>
</tr>
<tr>
<td></td>
<td>0.0067</td>
<td>0.0144</td>
<td>-0.0130</td>
<td>-0.0069</td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td>-0.0012</td>
<td>-0.0025</td>
<td>0.0003</td>
<td>0.0006</td>
<td>-0.0037</td>
</tr>
<tr>
<td></td>
<td>-0.0033</td>
<td>0.0019</td>
<td>0.0003</td>
<td>-0.0001</td>
<td>0.0048</td>
</tr>
</tbody>
</table>
SVD recovers subspaces – eigenfactors describe them

Are eigenfactors interpretable?

- Degrees of freedom in choosing factors
- Is orthogonality desired?
- Can only reconstruct a few factors (<<dimension)
- Additional eigenfactors used to refine non-linear interactions, instead of corresponding to new factors

Does each data vector really depend on all factors?
Sparse Matrix Factorization:
combinations of $m$ factors, from a pool of $k$

At most $m$ non-zero entries in row
### True (Planted) Factors

<table>
<thead>
<tr>
<th>Factors</th>
<th>0.6431</th>
<th>-0.9408</th>
<th>-0.9932</th>
<th>-0.0034</th>
<th>0.6628</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4274</td>
<td>0.2364</td>
<td>-0.0917</td>
<td>0.9128</td>
<td>-0.0681</td>
<td></td>
</tr>
<tr>
<td>0.6019</td>
<td>-0.0967</td>
<td>-0.0332</td>
<td>0.3965</td>
<td>0.5372</td>
<td></td>
</tr>
<tr>
<td>-0.2036</td>
<td>0.2226</td>
<td>0.0622</td>
<td>-0.0977</td>
<td>0.5172</td>
<td></td>
</tr>
<tr>
<td>0.0067</td>
<td>0.0144</td>
<td>-0.0130</td>
<td>-0.0069</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>-0.0012</td>
<td>-0.0025</td>
<td>0.0003</td>
<td>0.0006</td>
<td>-0.0037</td>
<td></td>
</tr>
<tr>
<td>-0.0033</td>
<td>0.0019</td>
<td>0.0003</td>
<td>-0.0001</td>
<td>0.0048</td>
<td></td>
</tr>
</tbody>
</table>

### Eigenfactors

<table>
<thead>
<tr>
<th>Factors</th>
<th>0.4006</th>
<th>1.0000</th>
<th>0.9303</th>
<th>0.1553</th>
<th>-0.5772</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>0.3976</td>
<td>0.6425</td>
<td>-0.1286</td>
<td>-0.1773</td>
<td></td>
</tr>
<tr>
<td>0.1814</td>
<td>0.5768</td>
<td>0.6388</td>
<td>-0.0995</td>
<td>-0.9999</td>
<td></td>
</tr>
<tr>
<td>-0.1335</td>
<td>0.1487</td>
<td>-0.1094</td>
<td>0.9999</td>
<td>0.1056</td>
<td></td>
</tr>
<tr>
<td>-0.6435</td>
<td>-0.9306</td>
<td>-1.0000</td>
<td>0.0973</td>
<td>0.6378</td>
<td></td>
</tr>
</tbody>
</table>
Sparse Matrix Factorization

\[ \text{sparse} \quad \text{Sparsity (} m \text{)} \quad \text{dense} \]

\[ m=k: \text{low rank approximation} \]
If $m=1$, and coefficients are 0/1, matrix decomposition is equivalent to k-means clustering.

For general coefficients with $m=1$, matrix decomposition is equivalent to clustering with a correlation distance measure.
Sparse Matrix Factorization

Sparsity ($m$)

$m=1$: clustering

$m=2,3,4$}

$m=k$: low rank approximation
Sparse Matrix Factorization
(m>1, but small)

• Model limited interactions
• Recovery even with large number of factors (beyond dimensionality of data / width of data matrix).
• No* degrees of freedom in recovery. *except scaling and permutation

• More interpretable factors?
An Encoding of the Data

\[ A \approx C \times F \]

(encoding) \quad \times \quad (the code)

Constraints reduce description length
Constrained Matrix Factorization
Lee & Seung, NIPS 97, Nature 99, NIPS 00

• Conic (non-negative coefficients)
• Convex (stochastic coefficients)
• Non-negative coefficients AND factors

Non-negativity appropriate for gene expression?
Viewed as PRMs

Expression = $G(F) + \text{error}$
Reconstructing a SMF from (noisy) Data: An Optimization Problem
Finding SMFs

Given $A$, find $C,F$ that minimize

$$\| A - C F \|_2$$

Subject to: at most $m$ non-zero entries in each row of $C$
Iterative Alternate Optimization

Optimize $F$ given $C$, and $C$ given $F$

Generalization of k-means clustering
Iterative optimization

• For fixed $C$, finding optimal $F$ is easy:

\[ A \approx CF \Rightarrow F = pinv(C)A \]

• For fixed $F$: each row of $A$ should be projected to a subspace spanned by $m$ of the rows of $F$
Optimizing $C$ for fixed $F$

(decoding)
Optimizing $C$ for fixed $F$ 
*(decoding)*

- For each row, find best projection to subspace spanned by $m$ of the rows of $F$.
  - need $\binom{k}{m}n$ projections
  - Perhaps with geometric data structure $\binom{k}{m}+n$

- Heuristic approach: change one coefficient at a time
  - With other coefficients fixed (simple projection)
  - With only coefficient *mask* fixed
Optimizing $C,F$ for fixed mask

$A \approx F \times$
Initializing the Factors $F$

- Where do we start our alternate-maximization search?
- In k-means: start with random rows of $A$
  - Problematic for SMF: too close to local minima with factors resembling cluster centers.
Jumping out of local minima

• Instead of restarting from scratch, keep the useful factors, replace the less-used factors.
• Can measure the effect of each factor on reducing the error.
• Back to a familiar problem: how do we pick new factors to replace those removed?
• Regularization penalty promoting sparseness (instead of hard constraint)

• EM instead of MM:
  – Search for distribution over $C$
  – Optimize $F$ for $C=\mathbb{E}[C|\mathcal{F}]$
Maximum Entropy Setting

\[ \min D(Q \| P_0) \quad \text{s.t.} \quad \|A - E_Q[(C.M)]F\|_2 < R \]

\[ P_0(M_{i,j}) \sim \text{Bernoulli}(q) \quad P_0(C_{i,j}) \sim N(0, \sigma^2) \]

\[ P_0(C, M) = P_0(C)P_0(M) \]
SMF with partially known $C,F$

- Some factors are known:
  - How well can they combine to explain data?
  - Find additional factors beyond known ones

- Combined with factor localization data: partial knowledge about coefficients
Reconstructing a SMF from (noisy) Data: A Statistical Problem

For \( A = C \times F + E \), up to what level of noise is \( C \times F \) the optimal factorization?

Measure: correlation of reconstructed \( F \) to true \( F \), as a function of \( \text{Var}(E)/\text{Var}(C \times F) \)
Reconstruction in the Presence of Noise

% Factors within >0.95

noise/signal

n=800
n=200
n=3200

k=10, m=2
d=20
Reconstruction in the Presence of Noise – low dimension

![Graph](image)

- k=10, m=2
- n=800

% Factors within >0.95 vs. noise/signal (logarithmic)
Current directions

• Better optimization methods
• Investigating the SMF of expression data (cell cycle, stress response)
• Model selection: choosing $k, m$