Using non-negative factorization of time series of graphs for learning from an event-actor network

Nam H. Lee

Johns Hopkins University

Youngser Park  Michael Rosen  I-Jeng Wang  Carey Priebe
Motivation – Wikipedia in multiple languages

Babe Ruth

Ruth in 1920, in New York Yankees uniform

Outfielder / Pitcher

Born: February 6, 1895
Baltimore, Maryland

Died: August 16, 1948 (aged 53)
New York City, New York

Batted: Left Threw: Left

MLB debut
July 11, 1914 for the Boston Red Sox

Last MLB appearance
May 30, 1935 for the Boston Braves

Career statistics

Batting average .342

Yankees de New York – Nº 3
Voltigeur, Lanceur
Frappeur gaucher Lanceur gaucher

Premier match
11 juillet 1914

Dernier match
30 mai 1935

Statistiques de joueur (1914-1935)

- Parties jouées 2503
- Points produits 2217
- Coups de circuit 714
- Moyenne au bâton 0,342
- Parties lancées 163
- Victoires 94

Équipes

- Red Sox de Boston (1914-1919)
- Yankees de New York (1920-1934)
- Braves de Boston (1935)
Wikipedia in French and English

Shouldn’t they be similar?

\[ G_{ij} = \begin{cases} 
1 + & \text{if topic-group } i \text{ links to topic-group } j \\
0 & \text{otherwise.}
\end{cases} \]
Generic Problem Statement

Clustering of multiple graphs

Let $(\kappa(1), G_1), \ldots , (\kappa(T), G_T)$ be an (independent) sequence of pairs of a class label $\kappa(t)$ and a (potentially weighted) graph $G_t$ on $n$ vertices. We assume that the class label $\kappa(t)$ takes values in $\{1, \ldots , K\}$ and also that given $\kappa(t) = k$, each $G_t$ is a random graph on $n$ vertices whose distribution depends only on the value of $k$. Given $G = \{G_t\}_{t=1}^T$, what is $\hat{\kappa}(t)$ for each $t = 1, \ldots , T$?

- vertex id matched across graphs (?)
- num of vertices are the same across graphs (?)
Matrix Factorization Assumption

Noiseless Case

Consider graphs $G(1)$, $G(2)$, $G(3)$ and $G(4)$ on 2 nodes such that $G(1) = G(3) = A$ and $G(2) = G(4) = B$, where

$$A = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 5 \\ 4 & 0 \end{pmatrix}$$

Then,

$$X = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 4 \\ 2 & 5 & 2 & 5 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1/3 & 4/9 \\ 2/3 & 5/9 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}$$
Matrix Factorization Assumption

Noiseless Case

\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
8/9 & 1/9 & 8/9 & 1/9 \\
1/9 & 8/9 & 0 & 0
\end{pmatrix} =
\begin{pmatrix}
0 & 0 \\
8/9 & 1/9 \\
1/9 & 8/9
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
9 & 0 & 0 & 0 \\
0 & 0 & 9 & 0 \\
0 & 0 & 0 & 9
\end{pmatrix}
\]

Interpretation

- for \( t = 1, 3 \), there were 9 interaction events, where 100 percent of them is of type \( A' \), and each of event is for \( 2 \rightarrow 1 \) with probability \( 8/9 \) and is for \( 1 \rightarrow 2 \) with probability \( 1/9 \)

- for \( t = 2, 4 \), there were 9 interaction events, where 100 percent of them is of type \( B' \), and each of event is for \( 2 \rightarrow 1 \) with probability \( 1/9 \) and is for \( 1 \rightarrow 2 \) with probability \( 8/9 \)
Matrix Factorization Assumption

Poisson Noise Case
Consider graphs $G(1), G(2), G(3)$ and $G(4)$ on 2 nodes such that $\overline{X} = E[X]$ and $(X_{ij})$ are independent Poisson random variables, where

$$
\overline{X} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} & H_{13} & H_{14} \\ H_{21} & H_{22} & H_{23} & H_{24} \end{pmatrix} \begin{pmatrix} \Lambda_{11} & 0 & 0 & 0 \\ 0 & \Lambda_{22} & 0 & 0 \\ 0 & 0 & \Lambda_{33} & 0 \\ 0 & 0 & 0 & \Lambda_{44} \end{pmatrix}
$$

with $1^T \overline{W} = 1^T$ and $1^T \overline{H} = 1^T$. 
Matrix Factorization Assumption

Wikipedia pages in two languages

1. \( \bar{\Lambda}_{\ell\ell} = \bar{\Lambda}_{\ell\ell}(\tau) \), the number of edges observed for the \( \ell \)th language by time \( \tau \)

2. Treat \( \bar{\Lambda} \) as a nuisance parameter – two Wikigraphs might be evolving on different time scales

Inference on the inner dimension \( d \)

1. if \( \bar{W} \) and \( \bar{H} \) are \( n^2 \times 1 \) and \( 1 \times 2 \) matrices, i.e., \( (d = 1) \), then two Wikipedia graphs are noisy obs. of the “same” kind

2. if \( \bar{W} \) and \( \bar{H} \) are \( n^2 \times 2 \) and \( 2 \times 2 \) matrices, i.e., \( (d = 2) \), then two Wikipedia graphs are noisy obs. of the “different” kinds
Matrix Factorization Assumption

Multiple graphs with recurring motifs

The collection $G$ has $d$ recurring motifs provided that

$$X = \overline{WH\Lambda},$$

where $\overline{W}$, $\overline{H}$ and $\overline{\Lambda}$ are $n^2 \times d$, $d \times T$ and $T \times T$ full “positive-rank” non-negative matrices such that $\mathbf{1}^\top \overline{W} = \mathbf{1}^\top$, $\mathbf{1}^\top \overline{H} = \mathbf{1}^\top$ and $\overline{\Lambda}$ is diagonal.

1. $\overline{X}_{\ell,t} = \mathbf{E}[G_{ij}(t)]$, where $\ell = i + (j - 1)n$

2. $\sum_{ij} \mathbf{E}[G_{ij}(t)] = \mathbf{E}[\mathbf{1}^\top G(t)\mathbf{1}] = \overline{\Lambda}_{tt}$
Related Works

- Elbow finding method (Zhu & Godshi)
- Core-consistency for PARAFAC tensor models (Bro & Kieffer)
- Two sample hypothesis testing procedure (Tang et al.)
- Clustering in a mixture distribution (Schiebinger et al.)
Model Selection – estimating $d$, the inner dimension

NMF
Given $\hat{d} = 1, \ldots, T$,

$$\left(\hat{W}, \hat{H}\right) := \arg\min_{W \geq 0, H \geq 0} D(\hat{P}; W, H),$$

where $\hat{P}_{ij,t} := X_{ij,t}/N_t$ and $W$ has $\hat{d}$ columns and $H$ has $\hat{d}$ rows.

Penalized-Loss Minimization

- AICc
Penalized-Loss Minimization

\[ \text{AICc} = \text{Loss} + \text{Penalty} \]

\[ \text{AICc} := -2 \sum_{ij,t} \hat{P}_{ij,t} \log(\hat{P}_{ij,t}) + 2 \sum_{k=1}^{\hat{d}} \frac{\hat{C}_k - 1}{\hat{N}_k} \]

where \( \hat{N}_k = \sum_t \hat{H}_{kt} N_t \), and \( \hat{C}_k = \sum_{ij} 1\{\hat{W}_{ij,k} > 0\} \).

Intuition

- An empirical estimate of entropy –
  \[ \frac{1}{N_t} \sum_{ij} X_{ij,t} \log(\hat{P}_{ij,t}) = \frac{1}{N_t} \sum_{\ell=1}^{N_t} \sum_{ij} 1_{B_{ij}}(\xi_{\ell}(t)) \log(\hat{P}_{ij}) \approx E[\sum_{ij} 1_{B_{ij}}(\xi_0(t)) \log(P_{ij})] = E[\log(p_t(\xi_0(t))))], \]
  where each \( \xi_{\ell}(t) \sim p_t \) independently

- Non-redundancy – \( \hat{C}_k \) penalizes the models with \((\hat{W}_{ij,1}, \ldots, \hat{W}_{ij,\hat{d}})\) having too many non-zero terms for too many \( ij \)
Penalized-Loss Minimization

Unbiased in the limit
Under some simplifying asymptotic condition,

$\lim_{\ell \to \infty} \ell \left( E[\varphi(\hat{W}, \hat{H})] - \varphi(W, H) \right) = \sum_{k=1}^{d} \frac{\overline{C}_k - 1}{\overline{n}_k \overline{\lambda}_k}, \quad (3)$

where the expectation is taken w.r.t. the parameter $(W, H, N)$,

$\varphi(W, H) := E \left[ \sum_{ij,t} (X_{ij,t}/N_t) \log((WH)_{ij,t}) \right],
\overline{C}_k = \sum_{ij} 1\{W_{ij,k} > 0\},
\overline{n}_k \overline{\lambda}_k \approx N_k(t)/\ell.$
Swimmer Data Set\textsuperscript{1}

The swimmer data set is a frequently-tested data set for benchmarking NMF algorithms. In our present notation, each column of $220 \times 256$ data matrix $X$ is a vectorization of a binary image, and each row corresponds to a particular pixel. Each image is a binary images (20-by-11 pixels) of a body with four limbs which can be each in four different positions. It is known that the matrix $X$ is 16-separable while the rank of $X$ is 13.

\textsuperscript{1}D. Donoho and V. Stodden. “When Does Non-Negative Matrix Factorization Give Correct Decomposition into Parts?” In: 2003.
Swimmer Data Set

Application of our AICc criteria using \texttt{nmf} with option \texttt{pe-nmf} with $\alpha = 0$ and $\beta = 1$ yields the estimated $\hat{d}$ as 16 while using \texttt{nmf} with option \texttt{lee} yields $\hat{d} = 18$. Application of our FIC criteria using \texttt{FastConicalHull} and \texttt{FastSepNMF} yields the estimated $\hat{d}$ as 13 while using \texttt{nnmf} yields $\hat{d} = 1$. 
Table 1: The baseline procedure “dimSelect \( \circ \) svd” is compared against the NMF procedure “getAICc \( \circ \) gclust” for choosing \( \hat{d} \) for each of 100 Monte Carlo simulation experiments. The true rank \( r \) is 3. For \( \kappa = 0.5 \), the baseline procedure performs poorly.

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>( \hat{d} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^1</td>
<td>48</td>
<td>39</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10^2</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10^3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10^4</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

(a) “dimSelect \( \circ \) svd”

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>( \hat{d} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^1</td>
<td>0</td>
<td>19</td>
<td>18</td>
<td>28</td>
<td>29</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>10^2</td>
<td>0</td>
<td>14</td>
<td>38</td>
<td>28</td>
<td>19</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10^3</td>
<td>0</td>
<td>16</td>
<td>75</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10^4</td>
<td>0</td>
<td>17</td>
<td>83</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

(b) “getAICc \( \circ \) gclust”
AICc vs. Others – Biologically-motivated simulation data

Figure 1: Comparison of three approaches through ARI for the model selection performance. The different symbols distinguish the different levels of intensity. The different line types distinguish the different algorithms. In all cases, our procedure either outperforms or nearly on par with the two baseline algorithms.
AICc on Wikigraphs

Wikigraphs

Table 3: Do English and French Wiki-graphs represent the same connectivity structure?

<table>
<thead>
<tr>
<th>$\hat{d}$</th>
<th>Neg. Log Likelihood</th>
<th>Penalty</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.05</td>
<td>1.52</td>
<td>44.57</td>
</tr>
<tr>
<td>2</td>
<td>41.65</td>
<td>4</td>
<td>45.65</td>
</tr>
</tbody>
</table>
AIC\textsubscript{c} with repeated SVT

Figure 2: More curves for “implied clustering” are minimized at \( \hat{d} = 2 \), i.e., 3 and 7 curves out of 9 are marked red resp. for apparent (Left) and implied (Right) clustering. As moving from the bottom curve to the top curve, \( \varepsilon \) assumes the different values.
AICc with Seeded Graph Matching

Figure 3: The number $m_0$ of seeds can be at most the number $m$ of vertices in the graph. A noticeable trend is that for each $m$, the bigger $m_0$ is, the larger ARI value becomes. Another notable trend is that for the $m_0 = m$ cases, as $m$ gets larger, ARI becomes larger as well.
Summary

Open Issues

Contact
  ▶ nhlee@jhu.edu