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

Nam H. Lee

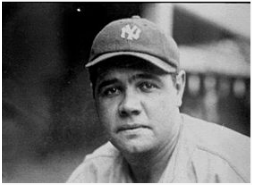
Johns Hopkins University

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

Babe Ruth



Yankees de New York - N° 3

Voltigeur, Lanceur

Frappeur gaucher **Lanceur** gaucher

Premier match

11 juillet 1914

Dernier match

30 mai 1935

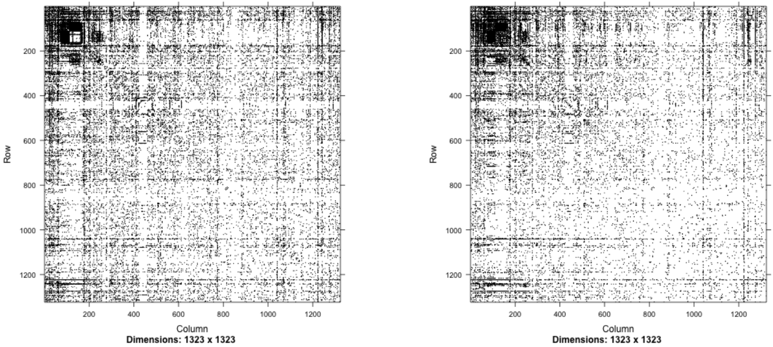
Statistiques de joueur (1914-1935)	
Parties jouées	2503
Points produits	2217
Coups de circuit	714
Moyenne au bâton	0,342
Parties lancées	163
Victoires	94

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 $\mathcal{G} = \{G_t\}_{t=1}^{T}$, what is $\widehat{\kappa}(t)$ 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

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

Then,

$$\underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 4 \\ 2 & 5 & 2 & 5 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{\overline{X}} = \underbrace{\begin{pmatrix} 0 & 0 \\ 1/3 & 4/9 \\ 2/3 & 5/9 \\ 0 & 0 \end{pmatrix}}_{\overline{W}} \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}}_{\overline{H}} \underbrace{\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}}_{\overline{\Lambda}}$$

Matrix Factorization Assumption

Noiseless Case

$$\underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 \\ 8 & 1 & 8 & 1 \\ 1 & 8 & 1 & 8 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{\overline{X}} = \underbrace{\begin{pmatrix} 0 & 0 \\ 8/9 & 1/9 \\ 1/9 & 8/9 \\ 0 & 0 \end{pmatrix}}_{\overline{W}} \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}}_{\overline{H}} \underbrace{\begin{pmatrix} 9 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}}_{\overline{\Lambda}}$$

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 $2 \to 1$ with probability $8/9$ and is for $1 \to 2$ 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 $2 \to 1$ with probability $1/9$ and is for $1 \to 2$ 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 $\overline{X} = \mathbf{E}[X]$ and (X_{ij}) are independent Poisson random variables, where

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

with $\mathbf{1}^\top \overline{W} = \mathbf{1}^\top$ and $\mathbf{1}^\top \overline{H} = \mathbf{1}^\top$.

Matrix Factorization Assumption

Wikipedia pages in two languages

1. $\overline{\Lambda}_{\ell\ell} = \overline{\Lambda}_{\ell\ell}(\tau)$, the number of edges observed for the ℓth language by time τ
2. Treat $\overline{\Lambda}$ as a nuisance parameter – two Wikigraphs might be evolving on different time scales

Inference on the inner dimension d

1. if $\overline{W}$ and $\overline{H}$ are $n^2 \times 1$ and 1×2 matrices, i.e., $(d = 1)$, then two Wikipedia graphs are noisy obs. of the "same" kind
2. if $\overline{W}$ and $\overline{H}$ are $n^2 \times 2$ and 2×2 matrices, i.e., $(d = 2)$, then two Wikipedia graphs are noisy obs. of the "different" kinds

Matrix Factorization Assumption

Multiple graphs with recurring motifs

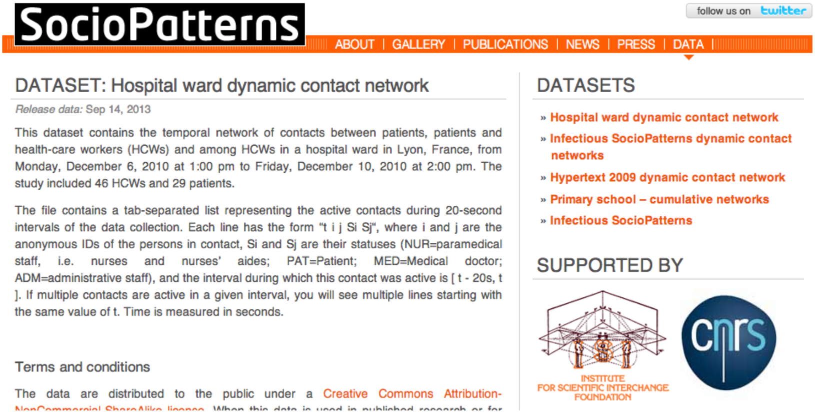
The collection $\mathcal{G}$ has d recurring motifs provided that

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

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



SocioPatterns

follow us on twitter

ABOUT | GALLERY | PUBLICATIONS | NEWS | PRESS | DATA |

DATASET: Hospital ward dynamic contact network

Release data: Sep 14, 2013

This dataset contains the temporal network of contacts between patients, patients and health-care workers (HCWs) and among HCWs in a hospital ward in Lyon, France, from Monday, December 6, 2010 at 1:00 pm to Friday, December 10, 2010 at 2:00 pm. The study included 46 HCWs and 29 patients.

The file contains a tab-separated list representing the active contacts during 20-second intervals of the data collection. Each line has the form "t i j Si Sj", where i and j are the anonymous IDs of the persons in contact, Si and Sj are their statuses (NUR=paramedical staff, i.e. nurses and nurses' aides; PAT=Patient; MED=Medical doctor; ADM=administrative staff), and the interval during which this contact was active is [t - 20s, t]. If multiple contacts are active in a given interval, you will see multiple lines starting with the same value of t. Time is measured in seconds.

Terms and conditions

The data are distributed to the public under a Creative Commons Attribution-

DATASETS

- » Hospital ward dynamic contact network
- » Infectious SocioPatterns dynamic contact networks
- » Hypertext 2009 dynamic contact network
- » Primary school – cumulative networks
- » Infectious SocioPatterns

SUPPORTED BY

INSTITUTE FOR SCIENTIFIC INTERCHANGE FOUNDATION

cnrs

- 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 $\widehat{d} = 1, \ldots, T$,

$$(\widehat{W}, \widehat{H}) := \underset{W \geq 0, H \geq 0}{\arg\min}\ D(\widehat{P}; W, H), \tag{1}$$

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

Penalized-Loss Minimization

- AICc

Penalized-Loss Minimization

AICc = Loss + Penalty

$$\text{AICc} := \underbrace{-2\sum_{ij,t}\widehat{P}_{ij,t}\log(\widehat{P}_{ij,t})}_{Loss} + \underbrace{2\sum_{k=1}^{\hat{d}}\frac{\widehat{C}_k - 1}{\widehat{N}_k}}_{Penalty} \tag{2}$$

where $\widehat{N}_k = \sum_t \widehat{H}_{kt} N_t$, and $\widehat{C}_k = \sum_{ij} \mathbf{1}\{\widehat{W}_{ij,k} > 0\}$.

Intuition

- An empirical estimate of entropy – $\frac{1}{N_t}\sum_{ij} X_{ij,t}\log(\widehat{P}_{ij,t}) = \frac{1}{N_t}\sum_{\ell=1}^{N_t}\sum_{ij}\mathbf{1}_{B_{ij}}(\xi_\ell(t))\log(\widehat{P}_{ij}) \approx \mathbf{E}[\sum_{ij}\mathbf{1}_{B_{ij}}(\xi_0(t))\log(P_{ij})] = \mathbf{E}[\log(p_t(\xi_0(t)))]$, where each $\xi_\ell(t) \sim p_t$ independently
- Non-redundancy – $\widehat{C}_k$ penalizes the models with $(\widehat{W}_{ij,1}, \ldots, \widehat{W}_{ij,\widehat{d}})$ having too many non-zero terms for too many ij

Penalized-Loss Minimization

Unbiased in the limit

Under some simplifying asymptotic condition,

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

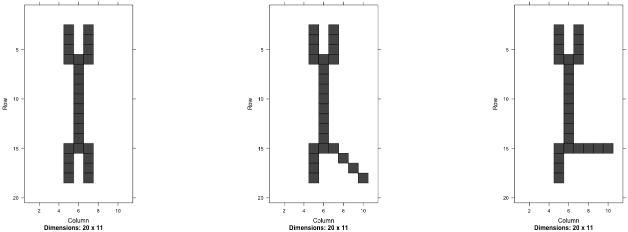
where the expectation is taken w.r.t. the parameter $(\overline{W}, \overline{H}, \mathbf{N})$,

$$\varphi(W, H) := \mathbf{E} \left[\sum_{ij,t} (X_{ij,t}/N_t) \log((WH)_{ij,t}) \right],$$

$$\overline{C}_k = \sum_{ij} \mathbf{1}\{\overline{W}_{ij,k} > 0\},$$

$$\overline{n}_k \overline{\lambda}_k \approx N_k(t)/\ell.$$

AICc on Swimmer

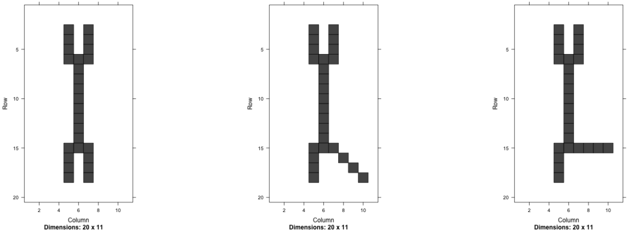


Swimmer Data Set[1]

The swimmer data set is a frequently-tested data set for bench-marking NMF algorithms. In our present notation, each column of 220×256 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.

[1] D. Donoho and V. Stodden. "When Does Non-Negative Matrix Factorization Give Correct Decomposition into Parts?" In: 2003.

AICc on Swimmer



Swimmer Data Set

Application of our AICc criteria using `nmf` with option `pe-nmf` with $\alpha = 0$ and $\beta = 1$ yields the estimated $\widehat{d}$ as 16 while using `nmf` with option `lee` yields $\widehat{d} = 18$. Application of our FIC criteria using `FastConicalHull` and `FastSepNMF` yields the estimated $\widehat{d}$ as 13 while using `nnmf` yields $\widehat{d} = 1$.

AICc vs. Others – Biologically-motivated simulation data

Table 1 : The baseline procedure "dimSelect ∘ svd" is compared against the NMF procedure "getAICc ∘ gclust" for choosing $\widehat{d}$ for each of 100 Monte Carlo simulation experiments. The true rank r is 3. For $\kappa = 0.5$, the baseline procedure performs poorly.

	$\widehat{d}$					
λ	1	2	**3**	4	5	6
10^1	48	39	**12**	1	0	0
10^2	100	0	**0**	0	0	0
10^3	100	0	**0**	0	0	0
10^4	100	0	**0**	0	0	0

(a) "dimSelect ∘ svd"

	$\widehat{d}$					
λ	1	2	**3**	4	5	6
10^1	0	19	**18**	28	29	6
10^2	0	14	**38**	28	19	1
10^3	0	16	**75**	9	0	0
10^4	0	17	**83**	0	0	0

(b) "getAICc ∘ gclust"

AICc vs. Others – Biologically-motivated simulation data

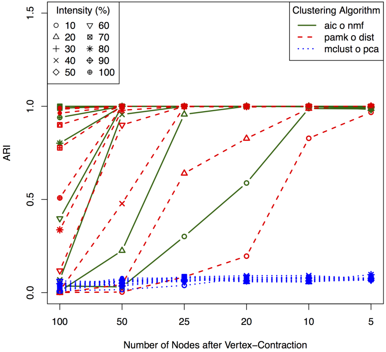


Figure 1 : Comparison of three approaches through ARI for the model selection performance. The different symbols distinguish the different levels of intensity. The different line types distinguish the different algorithms. In all cases, our procedure either outperforms or nearly on par with the two baseline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

AICc on Wikigraphs

Wikigraphs

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

$\widehat{d}$	Neg. Log Likelihood	Penalty	AICc
1	**43.05**	**1.52**	**44.57**
2	41.65	4	45.65

AICc with repeated SVT

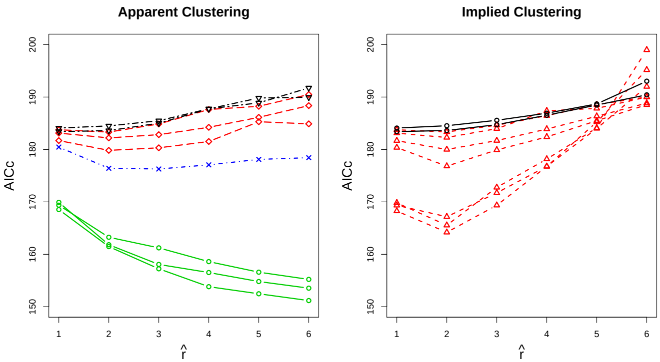


Figure 2 : More curves for “implied clustering ” are minimized at $\widehat{d} = 2$, i.e., 3 and 7 curves out of 9 are marked red resp. for apparent (Left) and implied (Right) clustering. As moving from the bottom curve to the top curve, ε assumes the different values

AICc with Seeded Graph Matching

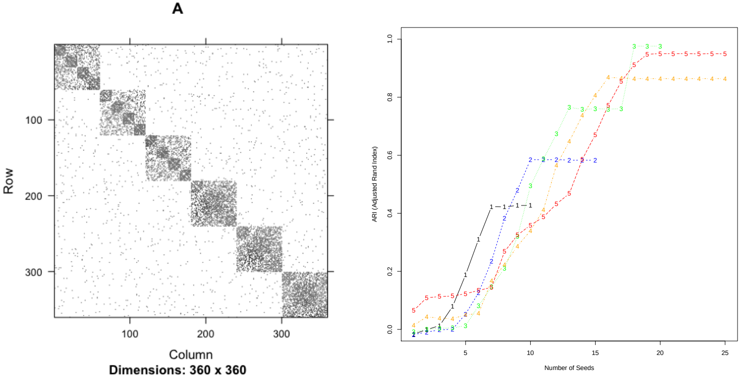


Figure 3 : The number m_0 of seeds can be at most the number m of vertices in the graph. A noticeable trend is that for each m, the bigger m_0 is, the larger ARI value becomes. Another notable trend is that for the $m_0 = m$ cases, as m gets larger, ARI becomes larger as well.

Summary

Open Issues

Contact

- nhlee@jhu.edu