MF TUTORIAL PART 4

BICLUSTERING

Sibylle Hess and Michiel Hochstenbach

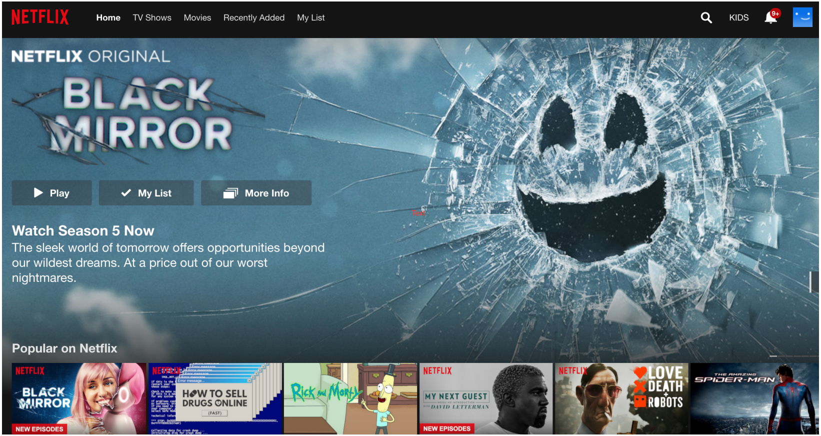
We have discussed the clustering of arbitrary shapes, but there is still one issue:

THE CURSE OF DIMENSIONALITY

In high dimensions, everything is almost equally similar

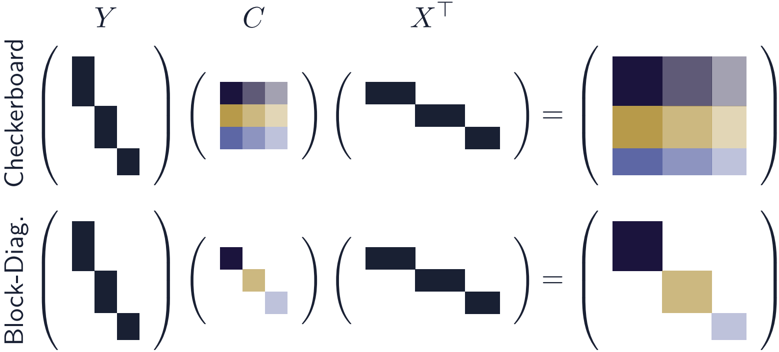
$\rightarrow$ CLUSTER FEATURES AND OBSERVATIONS

BICLUSTERING: RECOMMENDER SYSTEMS



EXCLUSIVE BICLUSTERING TRI-FACTORIZATIONS

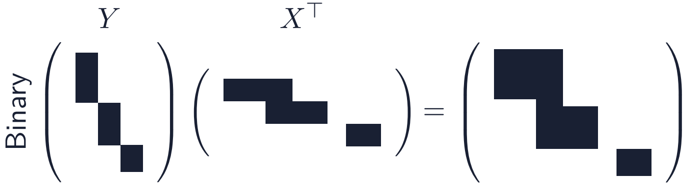
$$\min_{Y,C,X} \|D - YCX^\top\|^2 \quad \text{s.t. } Y \in \mathbb{1}^{m\times k}, C \in \mathbb{R}^{k\times k} \subseteq \mathcal{C}, X \in \mathbb{1}^{n\times k}$$



BINARY MATRIX FACTORIZATION

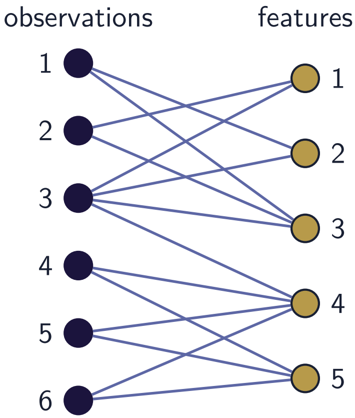
If the data matrix is binary, then a binary factorization is more interpretable:

$$\min_{Y,X} \|D - YX^\top\|^2 \quad \text{s.t. } Y \in \{0,1\}^{m\times k}, X \in \{0,1\}^{n\times k}$$



The factor matrices are here not restricted to partitions, but the objective has an inbuilt **PENALIZATION** of **OVERLAP** between clusters.

BIPARTITE GRAPH CUT INTERPRETATION OF BINARY AND BLOCK-DIAGONAL MF



Interpret D as adjacency matrix of bipartite graph, the similarity matrix is defined as

$$W = \begin{pmatrix} \mathbf{0} & D \\ D^\top & \mathbf{0} \end{pmatrix}$$

The **MINIMUM CUT** objective can then be used for the identification of biclusters.

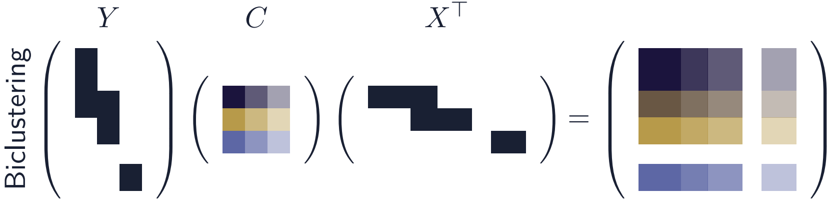
Most cluster models assume **EXCLUSIVITY OF CLUSTERS** to use optimization tricks from k-means

However, especially for biclustering applications, the **EXCLUSIVITY ASSUMPTION** is often **INEPT**.

A movie does not only belong to one cluster (e.g. genre)

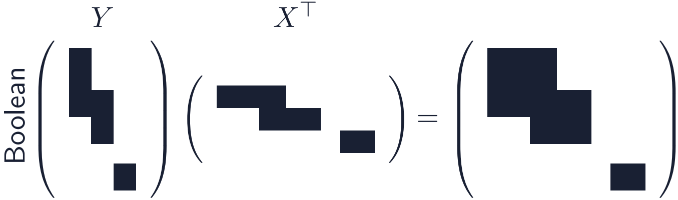
Biclustering with Overlap and Outliers

$$\min_{Y,C,X} \|D - YCX^\top\|^2 \quad \text{s.t. } Y \in \{0,1\}^{m\times k}, C \in \mathbb{R}^{k\times k}, X \in \{0,1\}^{n\times k}$$



BOOLEAN MATRIX FACTORIZATION

$$\min_{Y,X} \|D - Y \otimes X^\top\|^2 \quad \text{s.t. } Y \in \{0,1\}^{m \times k}, X \in \{0,1\}^{n \times k}$$



The **Boolean** matrix product **allows for overlap** :

$$Y \otimes X^\top = Y_{\cdot 1} X_{\cdot 1}^\top \oplus \ldots \oplus Y_{\cdot k} X_{\cdot k}^\top, \quad \text{where } 1 \oplus 1 = 1$$

Miettinen et al. 2008

Optimization

Alternating Minimization

$$X_{t+1} = \arg\min_X \|D - Y_t C_t X^\top\|^2 \qquad \text{s.t. } X \in \mathbb{1}^{n\times k}$$

$$C_{t+1} = \arg\min_C \|D - Y_t C X_{t+1}^\top\|^2 \qquad \text{s.t. } C \in \mathbb{R}^{k\times k}$$

$$Y_{t+1} = \arg\min_Y \|D - Y C_{t+1} X_{t+1}^\top\|^2 \qquad \text{s.t. } Y \in \mathbb{1}^{m\times k}$$

The **EXCLUSIVITY ASSUMPTION** makes the analytical computation of optimal X_{t+1}, C_{t+1} and Y_{t+1} possible.

Pros:

- fast convergence
- no hyperparameters except for rank

Cons:

- sensitive to initialization
- only applicable to partitioning clusters

Gaul andSchader 1996, Mirkin et al. 1995

Alternating Minimization in the Literature

- **Checkerboard Biclustering**

  Vichi 2001, Cho et al. 2004, Wang et al. 2011
- **Diagonal Biclustering**

  Han et al. 2017; Song et al. 2020
- **Binary Matrix Factorization**
  (one factor satisfies the exclusivity assumption)

  Koyotürk and Grama 2003, Li 2005

Bipartite Graph Cuts

Minimize the graph cut of

$$W = \begin{pmatrix} \mathbf{0} & D \\ D^\top & \mathbf{0} \end{pmatrix}$$

Pros:

- Fast, TSVD of D is sufficient to get the eigendecomposition of W

Cons:

- Sensitive to similarity measure and noise

Literature

- **Block-Diag. MF** Zha et al. 2001, Dhillon 2001
  Learn suitable graph representation: Nie et al. 2017
- **Binary MF** Neumann 2018

Nonnegative (soft)-orthogonal relaxation

$$\min_{\substack{Y \in \mathbb{R}_+^{m\times k}, \\ C \in \mathbb{R}_+^{k\times k},\ X \in \mathbb{R}_+^{n\times k}}} \left\| D - YCX^\top \right\|^2, \text{ s.t. } Y^\top Y = I,\ X^\top X = I.$$

Orthogonal NMF is NP-hard, in practice only soft-orthogonality of the matrices is achieved **Asteris et al. 2015**

Pros:

- NMF optimization methods can be applied
- Soft orthogonality allows for overlap and identification of outliers

Cons:

- Discretization is necessary $\rightarrow$ quality of binary end-result unclear

Ding et al. 2006, Yoo and Choi 2010, Del Buono and Pio 2015

Nonbinary Penalization of NMF

$$\min_{\substack{Y\in\mathbb{R}^{m\times k},\\ C\in\mathbb{R}_+^{k\times k},\ X\in\mathbb{R}^{n\times k}}} \|D - YCX^\top\|^2 + \phi(X) + \phi(Y)$$

ϕ penalizes nonbinary elements in the factor matrices X and Y.

Pros:

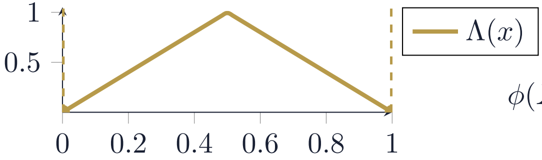
- NMF optimization methods can be applied
- No exclusivity, models overlap and outliers
- Results are very close to or equal to binary matrices

Cons:

- Discretization might be still necessary
- Scaling of matrices during optimization is an issue

Zhang et al. 2007/2010/2013, Hess et al. 2017/2021

PROX-OPERATOR FOR NONBINARY PENALIZATION



$$\phi(X) = \sum_{i,c} \Lambda(X_{ic})$$

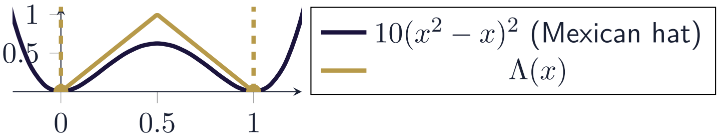
The Prox-Operator for ϕ is given by

$$\left(\operatorname{prox}_{\alpha\phi}(X)\right)_{ic} = \operatorname{prox}_{\alpha\Lambda}(X_{ic}) \text{with}$$

$$\operatorname{prox}_{\alpha\Lambda}(x) = \begin{cases} \max\{0, x - \alpha\} & x \leq 0.5 \\ \min\{1, x + \alpha\} & x > 0.5. \end{cases}$$

Hess et al. 2017

Nonbinary Penalization in the Literature



- **Checkerboard Biclustering**
  Stochastic PALM with Λ-Penalization: Hess et al. 2021
- **Binary Matrix Factorization**
  Multiplicative Updates with Mexican hat penalization
  Zhang et al. 2007/2010/2013
- **Boolean Matrix Factorization**
  PALM with Λ-Penalization: Hess et al. 2017

Conclusions

DISCUSSION

- Many of the one-sided clustering tricks are taken over for biclustering, but having two binary factor matrices poses additional challenges
- Theory for discretization steps is needed
- Different penalization functions could be tried (inspiration from regularized regression)
- Good biclustering optimizations are also relevant for tensor factorizations **Balažević et al. 2019**
- Modeling overlap and outliers in biclustering is still an active field of research

SOME REFERENCES PART IV

Miettinen & Neumann 2020	Survey Boolean MF
Hess et al. 2021	Penalized optimization of biclustering with overlap and outliers
Del Buono & Pio. 2015	Analysis of nonnegative matrix tri-factorization for biclustering