Non-negative Matrix Factorization ('Learning from Incomplete Ratings Using Non-negative Matrix Factorization'. Abril, 2006)

Jorge Schellman

Autores:
S. Zhang
W. Wang
J. Ford
F. Makedon

Introducción: NMF

En descomposición SVD admitimos que $R = U \tilde{S} V^T$ (sin restricción).

En NMF hacemos $R = P Q^T$, con todas las entradas positivas.

Al restringir la descomposición a valores no-negativos podemos reconstruir R como combinación lineal con coeficientes positivos de los vectores de la base:

$$R^{(j)} = P Q^{T(j)} = \sum_{i=1}^{k} P^{(i)} Q_i^{T(j)}$$

Introducción: NMF

$$R_{n\times m} \approx \widetilde{R}_{n\times m} = P_{n\times k} \times Q^{T}_{k\times m}$$

$$R = \begin{array}{c|cc} & i1 & i2 \\ \hline u1 & 4 & 1 \\ u2 & 3 & 2 \end{array} \qquad P = \begin{array}{c|cc} & f1 & f2 \\ \hline u1 & 1 & 2 \\ u2 & 3 & 4 \end{array} \qquad Q = \begin{array}{c|cc} & f1 & f2 \\ \hline i1 & 3 & 5 \\ i2 & 2 & 1 \end{array}$$

En paper: $A \approx X = UV$

$$\left(\widetilde{R}^{T} = QP^{T}\right)$$

$$U = \begin{array}{c|cc} & f1 & f2 \\ \hline i1 & 3 & 5 \\ i2 & 2 & 1 \end{array}$$

$$V = \begin{array}{c|cc} & u1 & u2 \\ \hline f1 & 1 & 3 \\ f2 & 2 & 4 \end{array}$$

Interpretación de Matrices

$X_{n\times m}$: matriz que se aproxima a A con rango a lo más k.

$U_{n\times k}$: matriz cuyas columnas son 'perfiles de rating representativos de cada comunidad de usuarios'.

$$U = \begin{array}{c|cc} & f1 & f2 \\ \hline i1 & 3 & 5 \\ i2 & 2 & 1 \end{array}$$

$V_{k\times m}$: matriz cuyas columnas son 'afinidades de cada usuario con todas las comunidades de usuarios'.

$$V = \begin{array}{c|cc} & u1 & u2 \\ \hline f1 & 1 & 3 \\ f2 & 2 & 4 \end{array}$$

Obtención de *X*:
Caso matriz A *fully observable*

Maximizamos el log-likelihood de la distribución de A dado X, lo que es equivalente a encontrar matrices no-negativas U y V que minimicen la norma cuadrada de Frobenius:

$$max \log Pr(A|X) \Leftrightarrow min\|A-UV\|_F^2$$

Luego se puede usar la técnica de multiplicadores de Lagrange para encontrar fórmulas de actualización para U_{ij} y V_{ij}.

Obtención de X:
Caso matriz A *sparse*

Queremos maximizar log $\Pr(A° \mid X)$.

Para esto 2 vías: **procedimiento EM** y **NMF ponderada**.

Una vez encontrado X, tomamos la predicción como X_{ij} dado que asumimos que $E(A_{ij}) = X_{ij}$ (cuando no se conoce A_{ij}) y que $\Pr(A_{ij} \mid X_{ij})$ obtiene su máximo cuando $A_{ij} = X_{ij}$

Procedimiento EM

- Expectación-Maximización:

1. Expectación: en la iteración *t* calcular el valor de expectación del log-likelihood de toda la data con respecto a los datos desconocidos (A^u) dado los observados (A^o) y la actual estimación $X^{(t-1)}$.

$$Q\left(X, X^{(t-1)}\right) = E\left[\log Pr\left(A^o, A^u \mid X\right) \mid A^o, X^{(t-1)}\right]$$

2. Maximización:

$$X^{(t)} = \underset{X}{argmax}\, Q\left(X, X^{(t-1)}\right)$$

Si A' matriz con datos observados más datos no observados reemplazados por la actual iteración, se puede demostrar que:

$$X^{(t)} = \underset{X}{argmin} \sum_{ij} \left(A'_{ij} - (UV)_{ij}\right)^2$$

WNMF

- Weighted Non-negative Matrix Factorization:

El log-likelihood de los datos observados podemos expresarlos como:

$$\log Pr(A^o|X) = \frac{-1}{2\sigma^2} \sum_{A_{ij} \in A^o} (A_{ij} - X_{ij})^2 + C$$

Maximizar lo anterior es igual a minimizar $\sum_{ij} W_{ij} (A_{ij} - (UV)_{ij})^2$

donde W_{ij} = 1 si A_{ij} es la entrada observada y 0 si no.

Experimentos

- 2 datasets:

  (1) 1426x2945 (items x users) rating matrix de MovieLens.

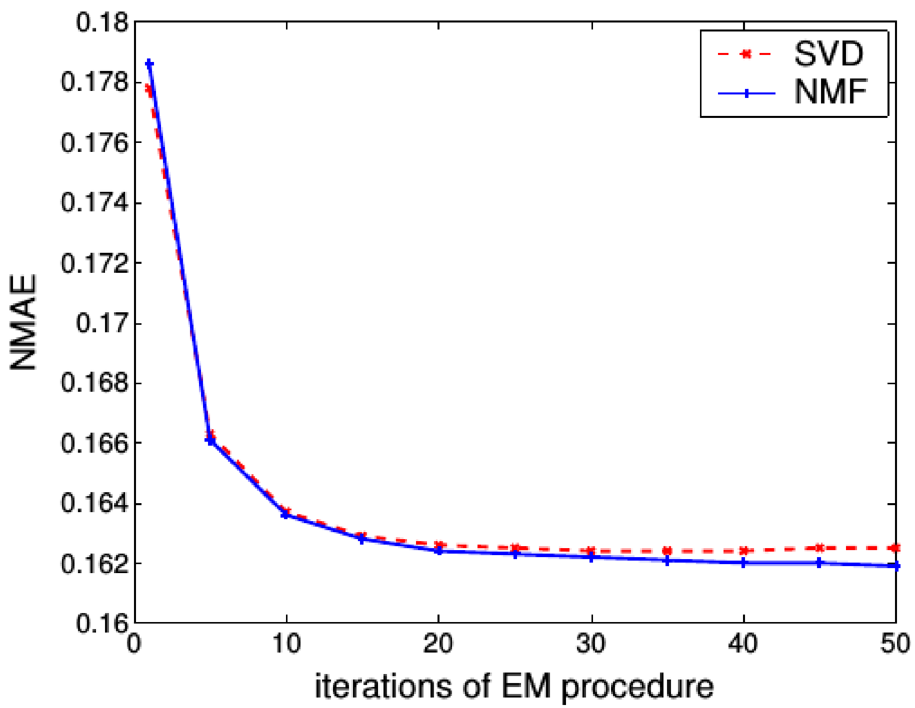
  (2) 100 x 3000 de Jester

- 5-fold cross-validation

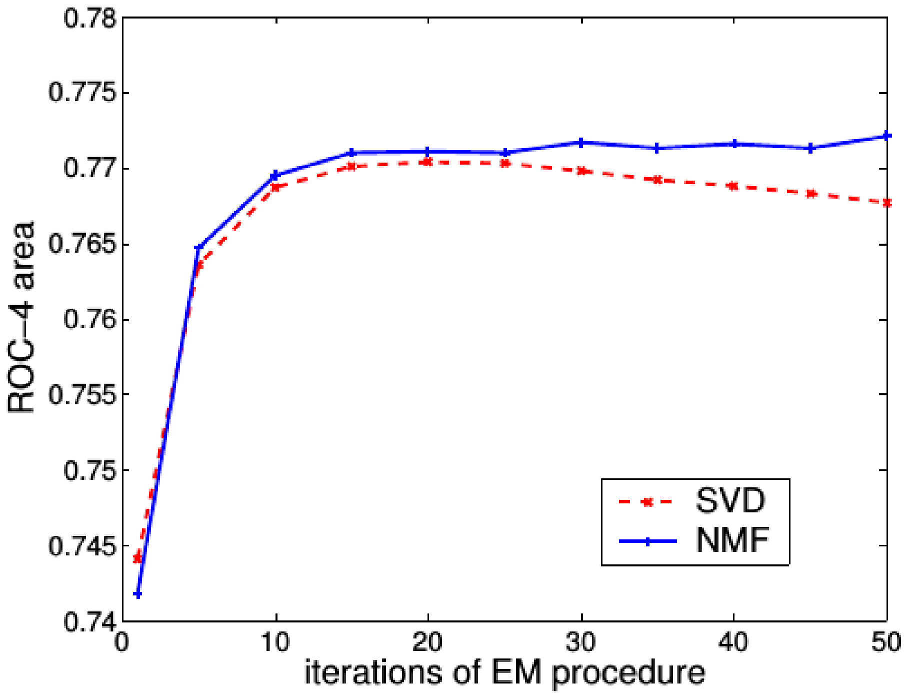
- Métricas: NMAE + área ROC-4

Resultados:
EM proc. vs WNMF

- SVD-based vs NMF-based EM approach:
  - En SVD, matriz de ratings normalizada por z-scores.
  - k = 10 en SVD y 20 en NMF.



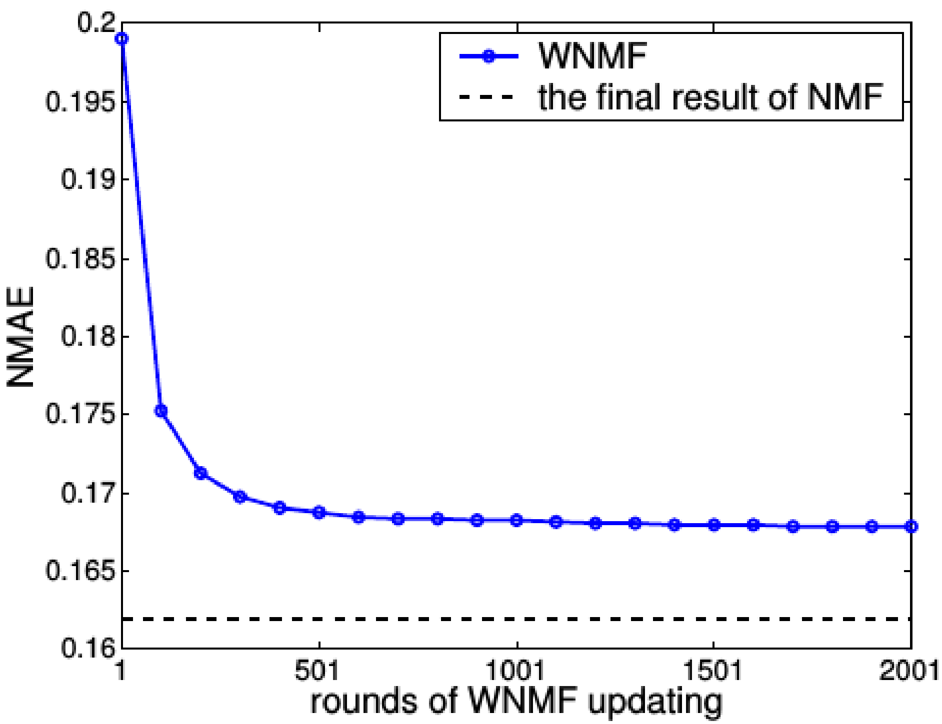
(a) NMAE



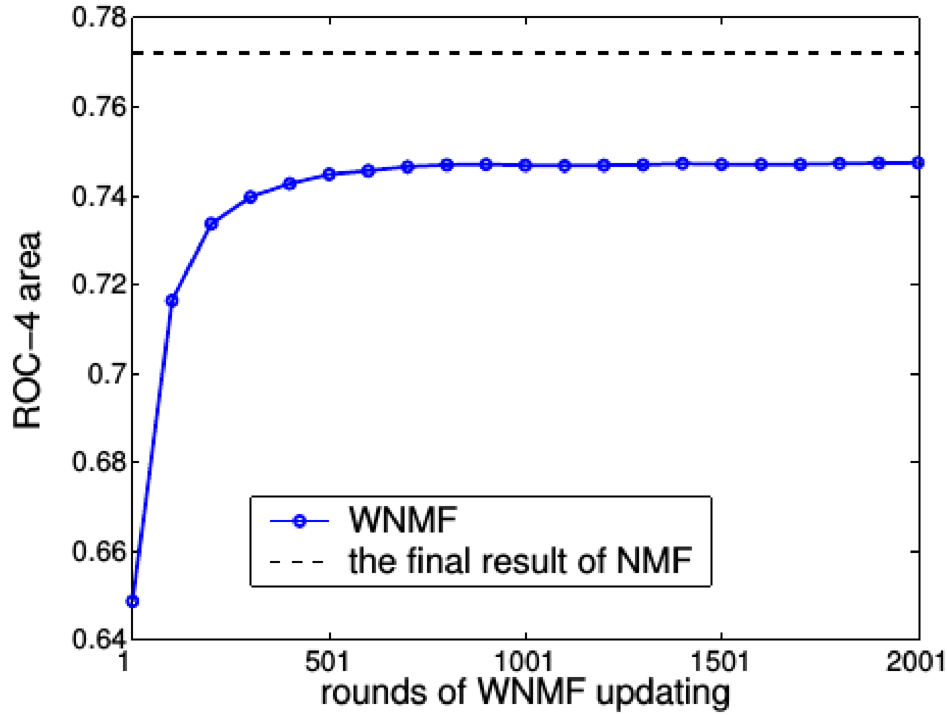
(b) ROC-4 area

Resultados: EM proc. vs WNMF

- WNMF (en NFM) en donde valores iniciales de *U* y *V* son aleatorios.



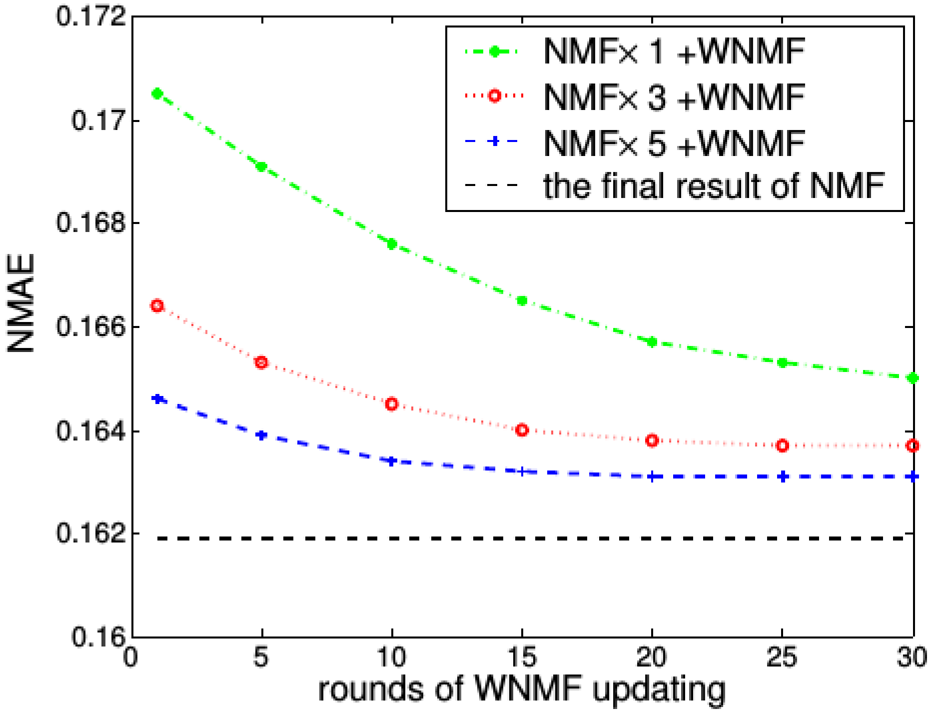
(a) NMAE



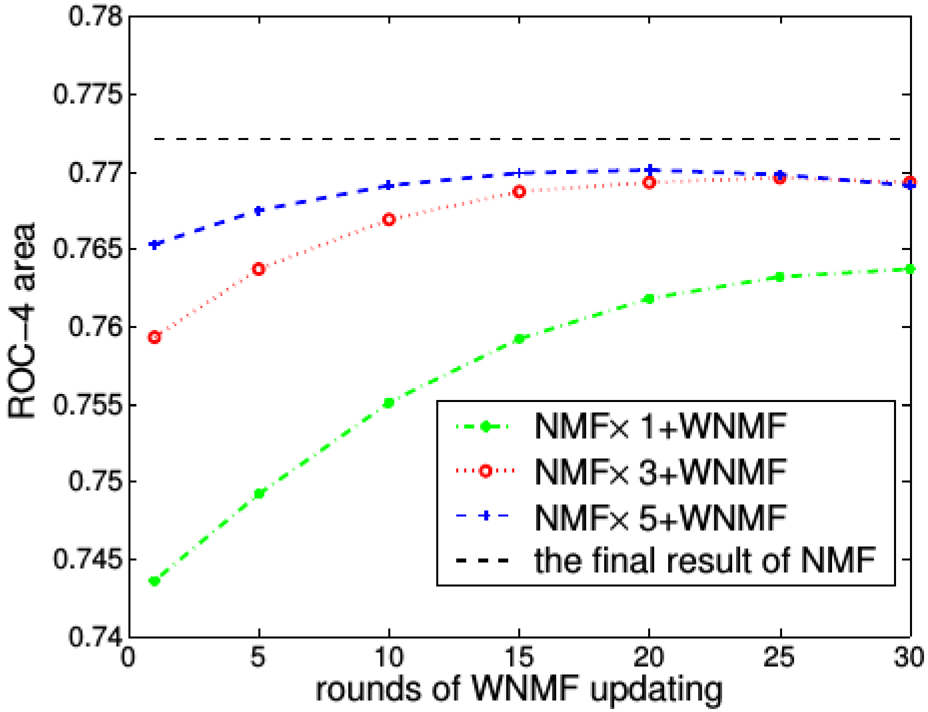
(b) ROC-4 area

Resultados:
EM proc. vs WNMF

- Cada iteración del EM proc.: 303.3s. Cada ronda de actualización en WNMF: 1.27s
- Hibridación EM/WNMF: aplicación procedimiento EM por varias iteraciones → obtención de un buen par *U*, *V* → aplicación WNMF.



(a) NMAE



(b) ROC-4 area

Resultados:
Algoritmos basados en NMF

- En NMAE, NMF+EM y SVD+EM obtienen mejores resultados.
- En área ROC-4, NMF+EM.

Table 1: Performance of CF algorithms on MovieLens.

	Pearson	SVD EM	NMF EM	Hybrid NMF
NMAE	0.1707	0.1629	0.1623	0.1634
ROC-4	0.7471	0.7682	0.7723	0.7691

Table 2: Performance of CF algorithms on Jester.

	Pearson	SVD EM	NMF EM	Hybrid NMF
NMAE	0.1634	0.1605	0.1599	0.1599
ROC-4	0.7539	0.7588	0.7612	0.7608

Discusión

Ventaja de NMF: coordenadas obtenidas reflejan directamente las características de las *comunidades de usuarios.*

Al ver las columnas de *U* podemos obtener una lista rankeada de las preferencias que tiene cada *grupo* sobre todos los ítems.

Interest Group 5	Interest Group 8	Interest Group 9
Austin Powers: International Man of My	Terminator 2: Judgment Day	Dumb & Dumber
(Comedy)	(Action|Sci-Fi|Thriller)	(Comedy)
Austin Powers: The Spy Who Shagged	Die Hard	Ace Ventura: When Nature Calls
(Comedy)	(Action|Thriller)	(Comedy)
Clerks	Independence Day (ID4)	Ace Ventura: Pet Detective
(Comedy)	(Action|Sci-Fi|War)	(Comedy)
Big Lebowski	Matrix	Home Alone 2: Lost in New York
(Comedy|Crime|Mystery|Thrill	(Action|Sci-Fi|Thriller)	(children's|Comedy)
Happy Gilmore	Speed	Nutty Professor
(Comedy)	(Action|Romance|Thriller)	(Comedy|Fantasy|Romance|Sci-I

Interest Group 17	Interest Group 19
Pretty Woman	Sound of Music
(Comedy|Romance)	(Musical)
Notting Hill	Grease
(Comedy|Romance)	(Comedy|Musical|Romance)
Steel Magnolias	Little Mermaid
(Drama)	(Animation|Children's|Comedy
Erin Brockovich	Wizard of Oz
(Drama)	(Adventure|Children's|Drama|I
Sleepless in Seattle	Cinderella
(Comedy|Romance)	(Animation|Children's|Musica

Clustering con NMF

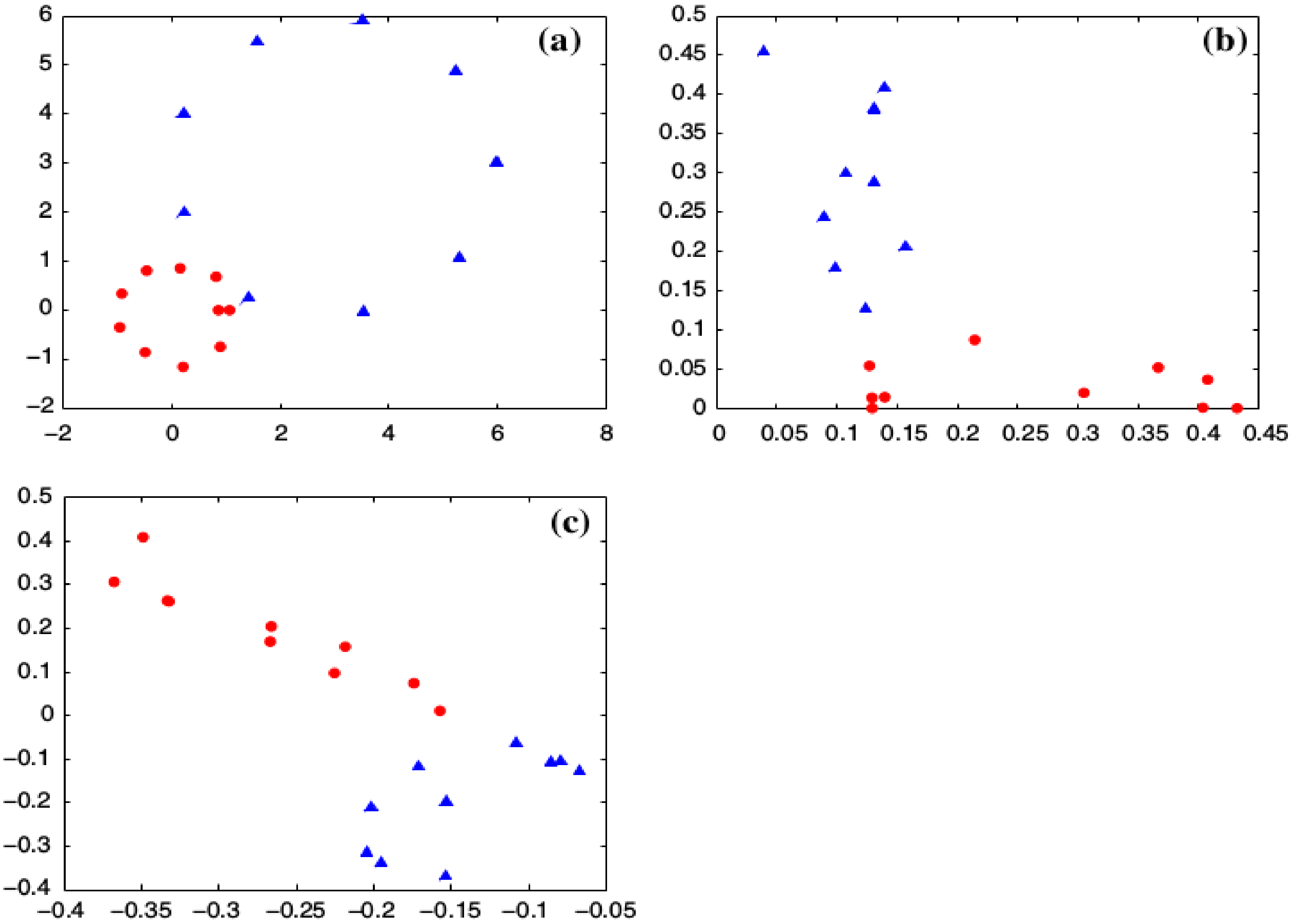


Fig. 1 **a** An artificial toy dataset consisting of two natural clusters. **b** Data distribution in the SS-NMF subspace of the two column vectors of **G**. The data points from the two clusters get distributed along the two axes. **c** Data distribution in the SS-SNC subspace of the first two singular vectors. There is no relationship between the axes and the clusters

Referencias

- S. Zhang, W. Wang, J. Ford, and F. Makedon, “Learning from Incomplete Ratings Using Non-Negative Matrix Factorization,” Proc. Sixth SIAM Int’l Conf. Data Mining (SDM), pp. 549-553, 2006.
- Y. Chen, M. Rege, M. Dong, and J. Hua, “Non-negative matrix factorization for semi-supervised data clustering,“ Knowl. Inf. Syst., 17, pp. 355–379, 2008.
- W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,“ Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, pp. 267–273, 2003.
- D. Lee and H.S. Seung, “Learning parts of objects by non-negative matrix factorization,“ Nature 401, pp. 788-791, 1999.