Signal and Information Processing : The Heartbeat of a Smart Society
Shanghai, China
ICASSP·2016
The 41st IEEE International Conference on Acoustics, Speech and Signal Processing, 20-25 March 2016

GAUSS-SEIDEL BASED NON-NEGATIVE MATRIX FACTORIZATION FOR GENE EXPRESSION CLUSTERING

Qing Liao, Naiyang Guan, Qian Zhang

Dept. of Computer Science & Technology, The Hong Kong University of Science and Technology

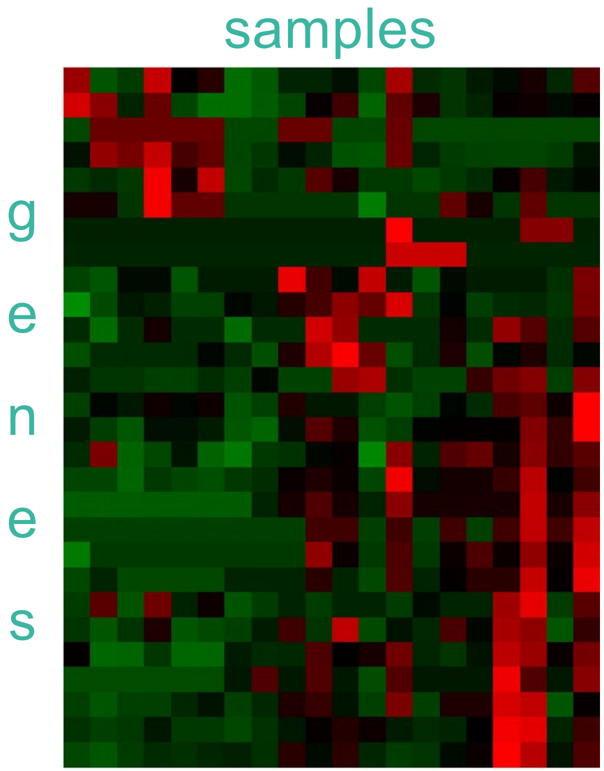
Tel: +86 18932416316

E-mail: ny.guan@gmail.com

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

INTRODUCTION

Gene Expression Clustering



Goal

Unearths similar bio-process, gene function, gene regulation, and subtypes of cells.

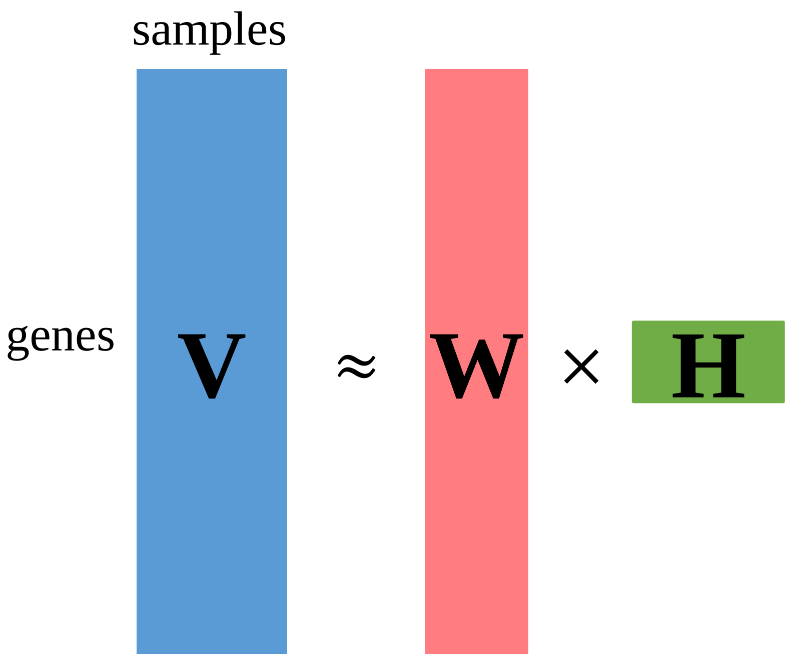
Imbalance

The number of probed genes is rather greater than the number of samples.

NMF clustering (Brunet *et al.* 2004)

Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

Nonnegative Matrix Factorization (NMF)



- Model:

$$\min_{W \geq 0, H \geq 0} \| X - WH \|_F^2$$

- Optimization:
  - Multiplicative Update Rule (MUR) (Lee *et al.* 2001)
  - NeNMF (Guan *et al.* 2012)

Daniel D Lee and H Sebastian Seung, “Algorithms for nonnegative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2001, pp. 556-562.
Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, “NeNMF: an optimal gradient method for nonnegative matrix factorization,” *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882-2898, 2012.

PROBLEM FORMULATION

Gauss-Seidel Based Nonnegative Matrix Factorization (GSNMF)

- Gauss-Seidel Method - An Example

Problem:

Given $V = \begin{bmatrix} 8 & 9 \\ 9 & 6 \\ 1 & 1 \end{bmatrix}$ and $A = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 2 \end{bmatrix}$, find

$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ h_{31} & h_{32} \end{bmatrix}$ such that $V = AH$.

PROBLEM FORMULATION

Gauss-seidel Based Nonnegative Matrix Factorization (GSNMF)

- Gauss-Seidel Method - An Example

1. Initialize: $\mathrm{H}^1 = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 5 & 3 \end{bmatrix}.$

2. Decompose: $A = U + D + U^T = \begin{bmatrix} 0 & 2 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 2 & 2 \end{bmatrix}.$

3. Solve the linear system:

$$V = AH = UH + DH + U^T H \xrightarrow{H^k} V - UH^k = DH^{k+1} + U^T H^{k+1}$$

$$\begin{bmatrix} 8 & 9 \\ 9 & 6 \\ 1 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 2 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 5 & 3 \end{bmatrix} = \begin{bmatrix} -6 & -7 \\ -1 & 0 \\ 1 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 2 & 0 \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ h_{31} & h_{32} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ h_{31} & h_{32} \end{bmatrix} \Longrightarrow H^2 = \begin{bmatrix} -6 & -7 \\ 11 & 14 \\ -9 & -13 \end{bmatrix}$$

PROBLEM FORMULATION

Gauss-seidel Based Nonnegative Matrix Factorization (GSNMF)

- Gauss-Seidel Method – Metric

1. Avoid the inverse operation
2. Fast convergence rate

- Gauss-Seidel Method – Problem

GS method: $\|X - AH\|_F^2$, where $A = A^T$

NMF: $\|X - WH\|_F^2$, where $W \neq W^T$

How to adopt the GS method to solve NMF?

PROBLEM FORMULATION

Gauss-seidel Based Nonnegative Matrix Factorization (GSNMF)

- Gauss-Seidel Method – An approximation model

$$\min_{H\geq 0}\left\|W^{T}X-W^{T}WH\right\|_{F}^{2}$$

Is the approximation reasonable ?

$$\min_{W\geq 0,H\geq 0}\left\|W^{T}X-W^{T}WH\right\|_{F}^{2}$$

$$\left\|W^{T}X-W^{T}WH\right\|_{F}^{2}\leq\left\|W\right\|_{F}^{2}\left\|X-WH\right\|_{F}^{2}$$

Constraint: $\left\|W\right\|_{F}^{2}\leq 1$

$$\left\|W^{T}X-W^{T}WH\right\|_{F}^{2}\leq\left\|X-WH\right\|_{F}^{2}$$

OPTIMIZATION

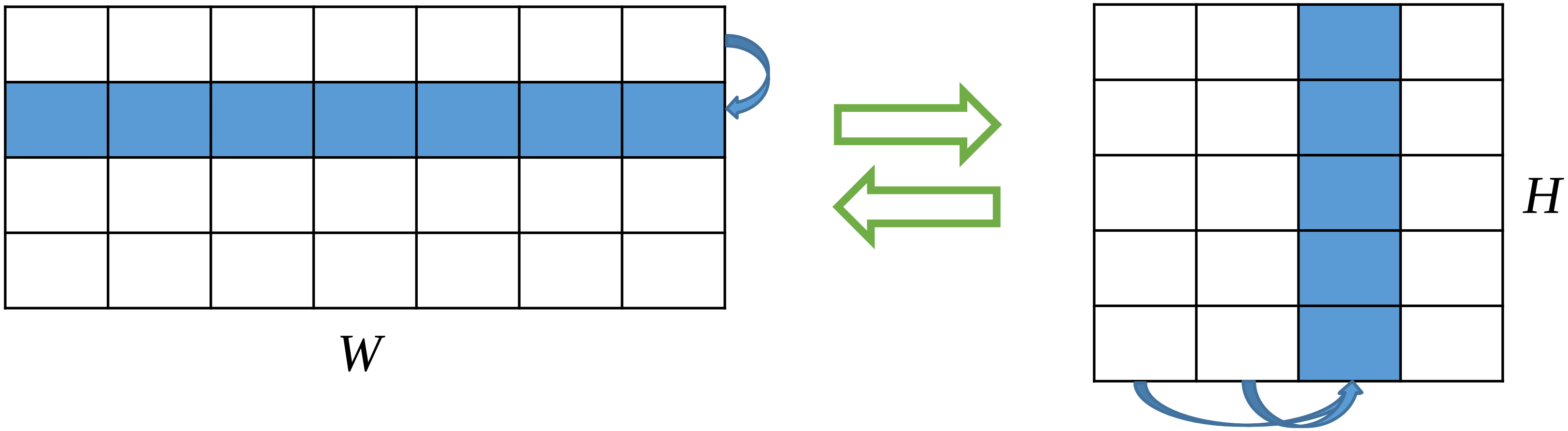
Gauss-seidel Based Nonnegative Matrix Factorization (GSNMF)

- Divide: $W^T W = U^T + D + U$
- Transform: $W^T V = W^T W H \Longrightarrow W^T V = U^T H + DH + UH$

$$\Longrightarrow W^T V - UH = U^T H + DH$$

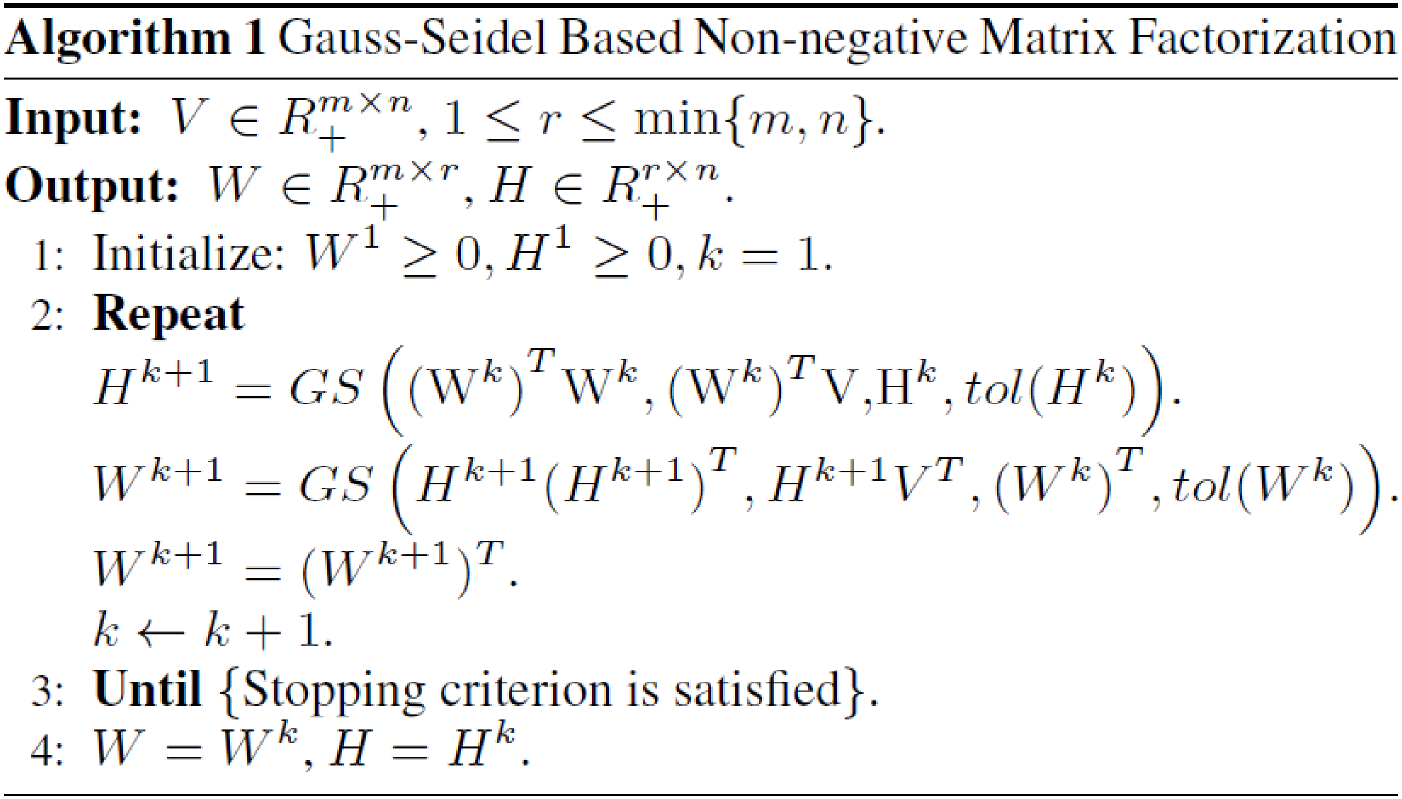
- Update H row by row:

$$H_{i\bullet}^{k+1} = \frac{1}{(W^{kT} W^k)_{ii}} \prod_+ \left(\left(UH^k\right)_{i\bullet} - \sum_{j>i} \left(W^{kT} W^k\right)_{ij} H_{j\bullet}^k - \sum_{j<i} \left(W^{kT} W^k\right)_{ji} H_{j\bullet}^{k+1} \right)$$



OPTIMIZATION

Gauss-seidel Based Nonnegative Matrix Factorization (GSNMF)



Algorithm 1 Gauss-Seidel Based Non-negative Matrix Factorization

```
Input: $V \in R_{+}^{m\times n}, 1 \leq r \leq \min\{m, n\}$.
Output: $W \in R_{+}^{m\times r}, H \in R_{+}^{r\times n}$.
Initialize: $W^1 \geq 0, H^1 \geq 0, k = 1$.
Repeat
    $H^{k+1} = GS\left(({\rm W}^k)^T{\rm W}^k, ({\rm W}^k)^T{\rm V},{\rm H}^k, tol(H^k)\right)$.
    $W^{k+1} = GS\left(H^{k+1}(H^{k+1})^T, H^{k+1}V^T, (W^k)^T, tol(W^k)\right)$.
    $W^{k+1} = (W^{k+1})^T$.
    $k \leftarrow k + 1$.
Until {Stopping criterion is satisfied}.
$W = W^k, H = H^k$.
```

Algorithms	Time complexities of one iteration round
NMF	$O(mnr + mr^2 + nr^2)$
NeNMF	$O(mnr + mr^2 + nr^2) + K \times O(mr^2 + nr^2)$
GSNMF	$O(mnr + mr^2 + nr^2)$

Datasets

Table 1. Summarization of six cancer gene expression datasets.

Datasets	Samples	Genes	Classes
gastricGSE2685 2razreda[1]	30	4522	2
gastricGSE2685[1]	30	4522	3
LL GSE1577[1]	29	15434	3
LL GSE1577 2razreda[1]	19	15434	2
AE GSE5060 GPL96[2]	22	18651	4
HSTS GSE2719[2]	39	18753	8

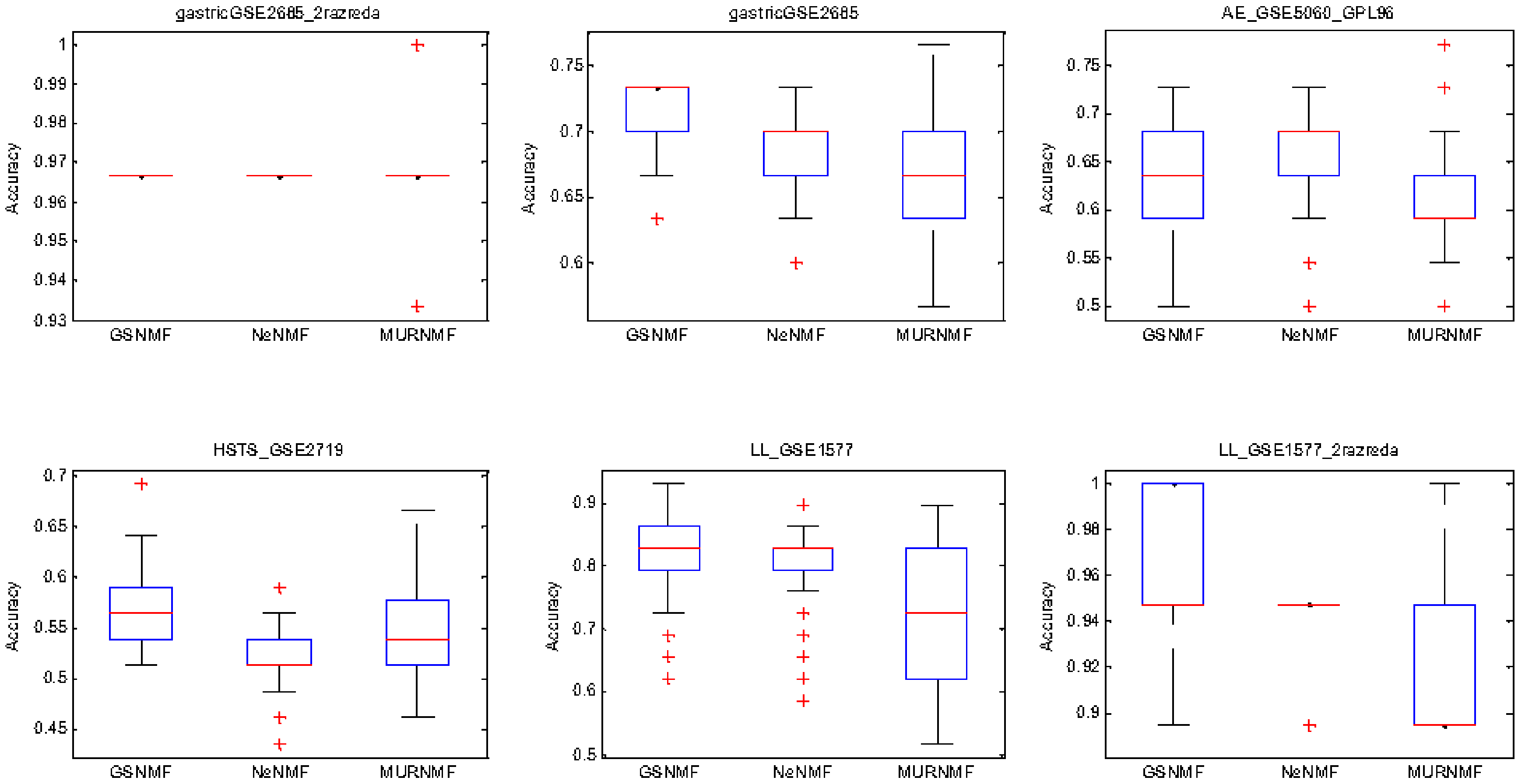
[1] The orange datasets: http://www.biolab.si/supp/bi-cancer/projections/.

[2] The GSE data : http://www.ncbi.nlm.nih.gov/gds.

EXPERIMENTS

Results

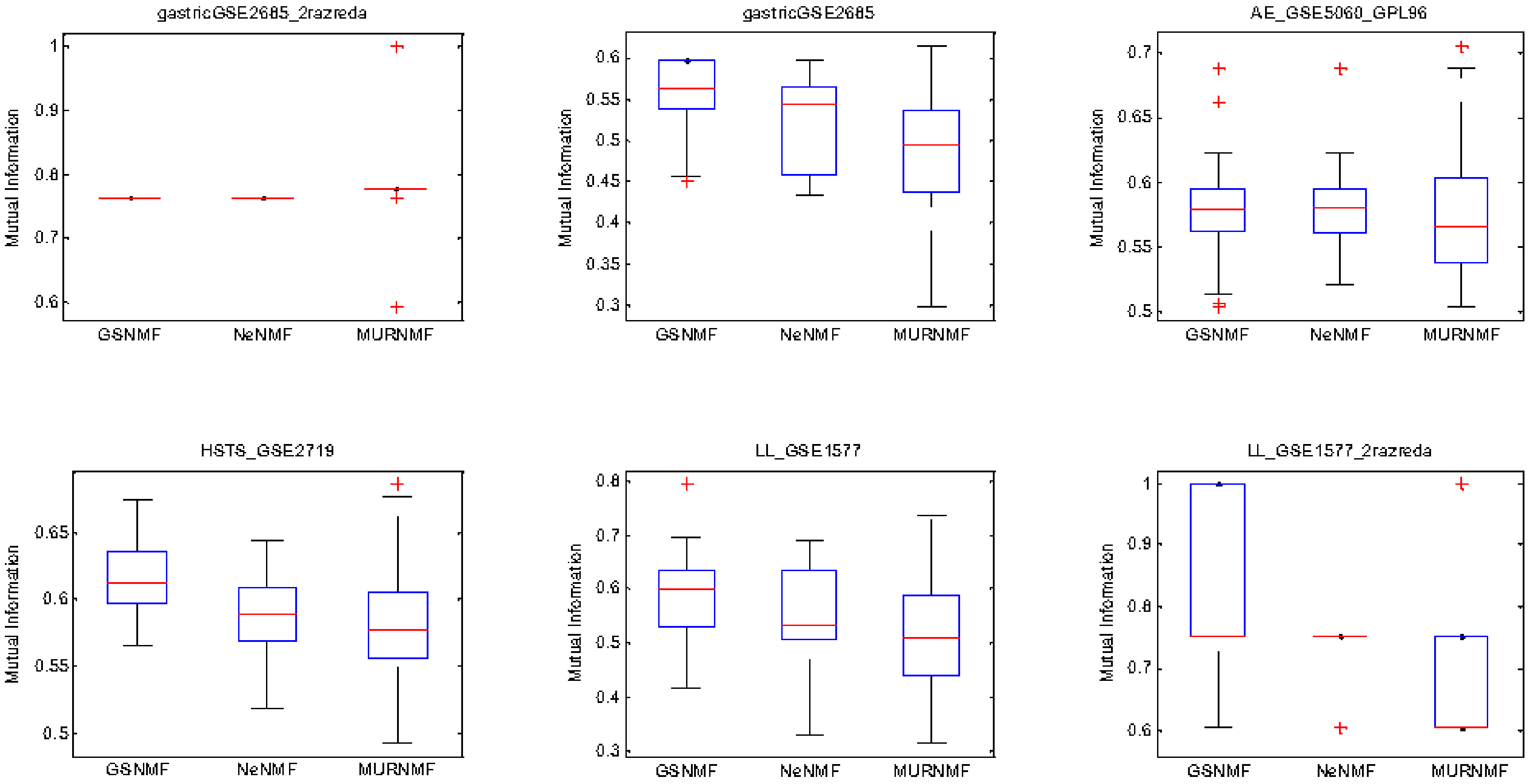
Fig. 1. The box plots of the clustering accuracies for the three NMF algorithms over 100 runs on all the six gene expression datasets.



EXPERIMENTS

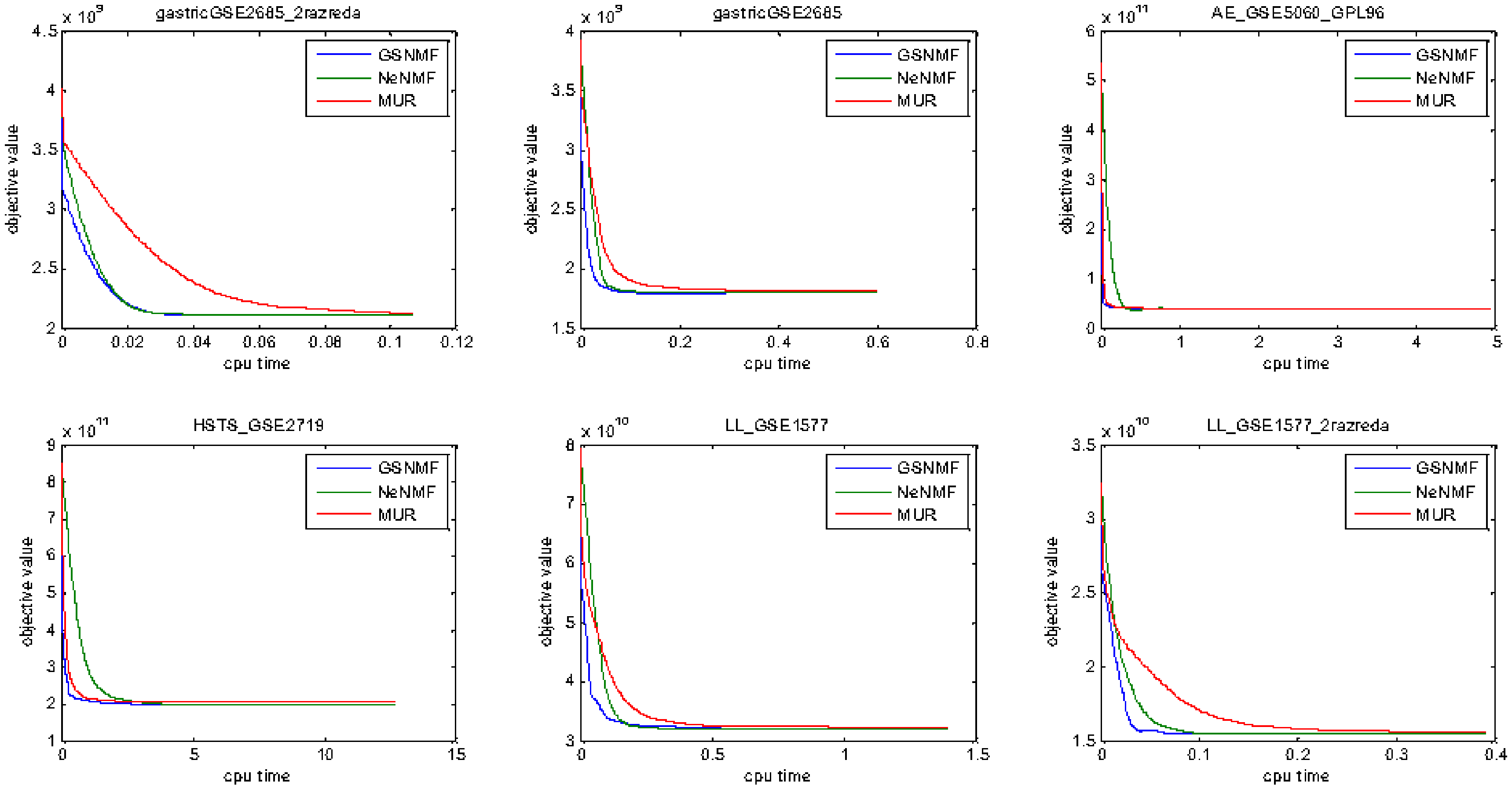
Results

Fig. 2. The box plots of the clustering mutual information for the three NMF algorithms over 100 runs on all the six gene expression datasets.



Results

Fig. 3. The average time consumption plots for the three NMF algorithms over 100 runs on all six gene expression datasets.



Thanks!

Questions?