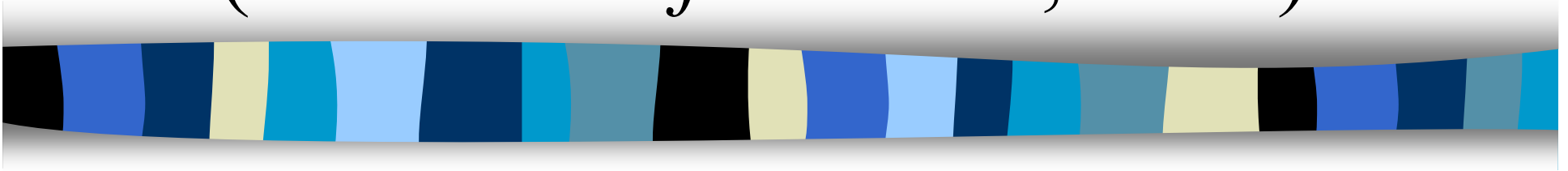


# Multi-Dimensional Scaling for Large Data Sets (*BMC Bioinformatics*, 2008)



Jengnan Tzeng<sup>1</sup>,  
Henry Horng-Shing Lu<sup>2</sup> and  
Wen-Hsiung Li<sup>3</sup>

[1.jengnan@gmail.com](mailto:1.jengnan@gmail.com)

[2.hslu@stat.nctu.edu.tw](mailto:2.hslu@stat.nctu.edu.tw)

[3.wli@uchicago.edu](mailto:3.wli@uchicago.edu)



# Outline

- Introduction to Multi-dimensional Scaling (MDS)
- Overview of recent techniques
- Methods to improve MDS
- Simulation results
- Applications to Microarray data
- Conclusion and Discussion



# Introduction to MDS

- Multi-dimensional Scaling (MDS) is a set of techniques for representing high dimensional data into low dimensional space
- The dissimilarities between pairs of data points in low dimensional space are similar to those of the dissimilarities in high dimensional space

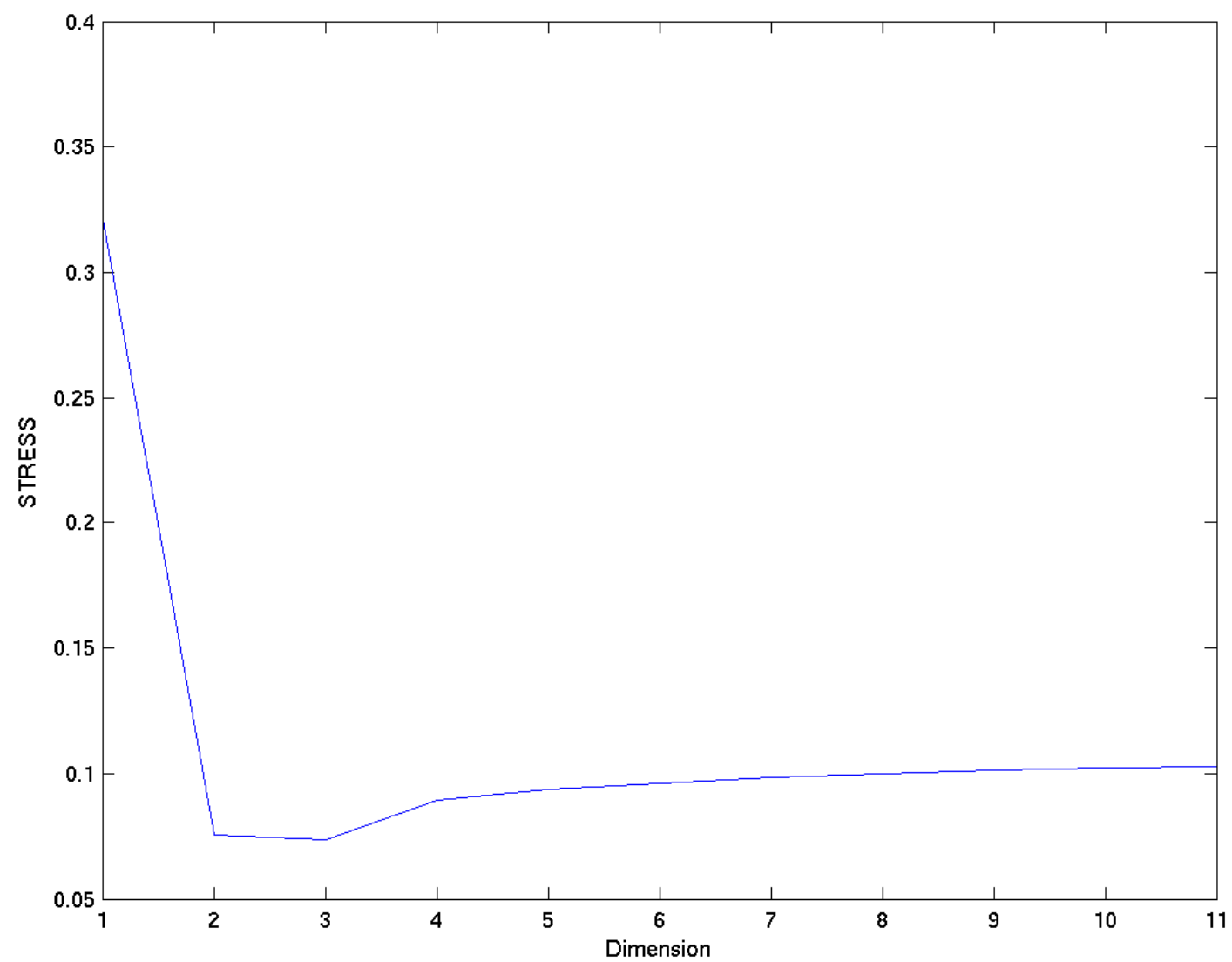
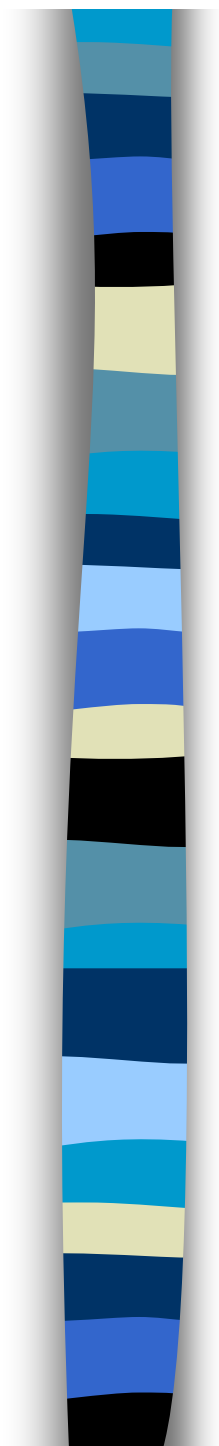


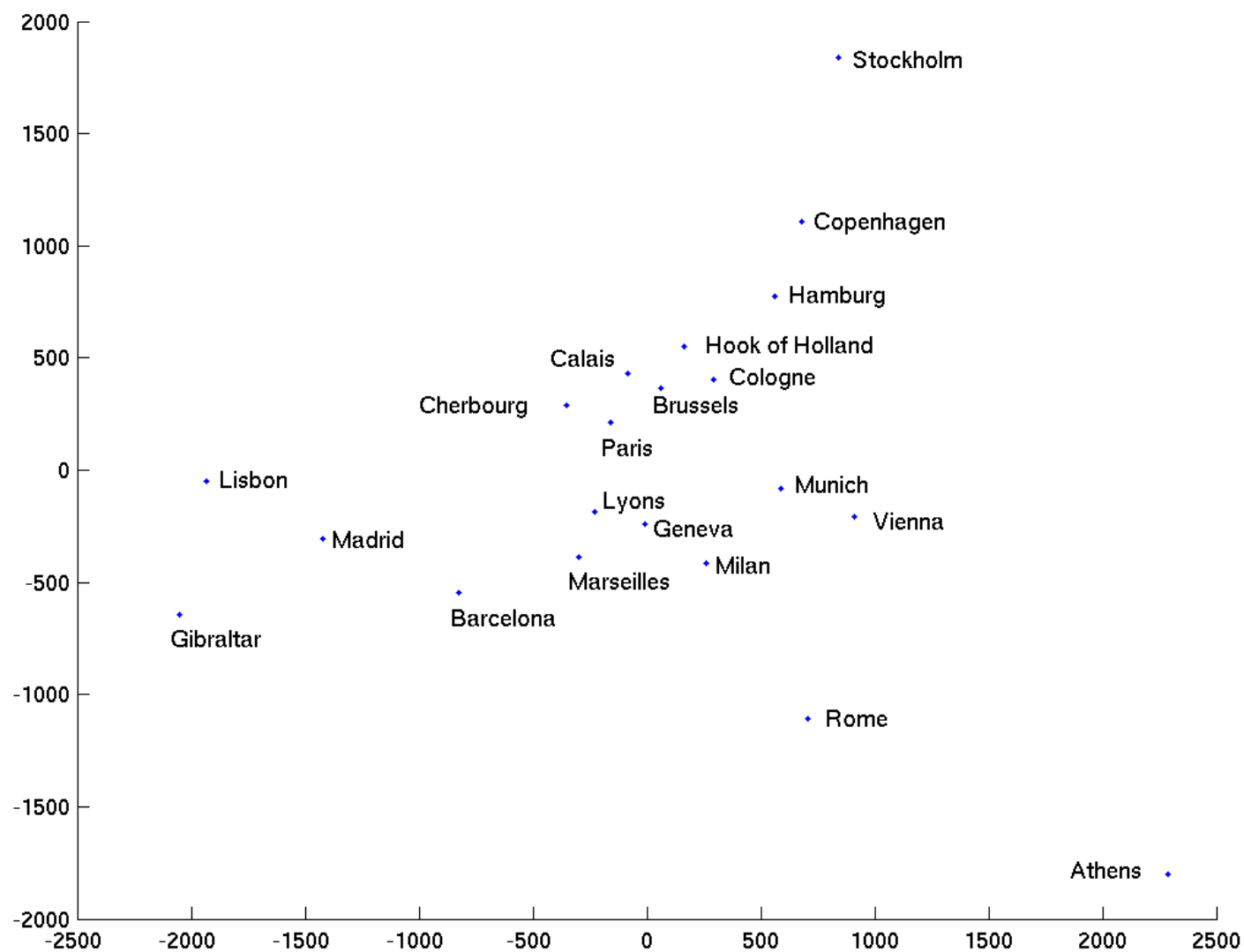
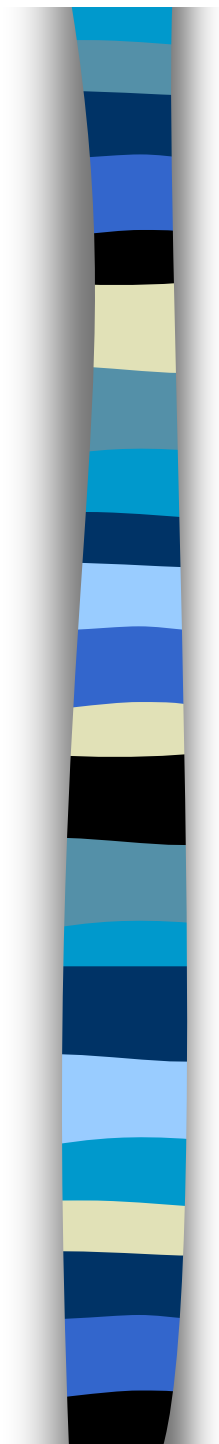
# The Applications of MDS

- Visualization of high dimensional data
- Feature extraction
- Pattern recognition
- Mapmaking
- Protein folding
- Mapping of disease genes

# Distances Between European Cities

	Athens	Barcelona	Brusse	Calais	Cherbourg	Cologne	Copenhagen	Geneva	Gibraltar	Hamburg	Hook of	Lisbon	Lyons	Madrid	Marseilles	Mila
Athens	0	3313	2963	3175	3339	2762	3276	2610	4485	2977	3030	4532	2753	3949	2865	22
Barcelona	3313	0	1318	1326	1294	1498	2218	803	1172	2018	1490	1305	645	636	521	10
Brussels	2963	1318	0	204	583	206	966	677	2256	597	172	2084	690	1558	1011	9
Calais	3175	1326	204	0	460	409	1136	747	2224	714	330	2052	739	1550	1059	10
Cherbourg	3339	1294	583	460	0	785	1545	853	2047	1115	731	1827	789	1347	1101	12
Cologne	2762	1498	206	409	785	0	760	1662	2436	460	269	2290	714	1764	1035	9
Copenhagen	3276	2218	966	1136	1545	760	0	1418	3196	460	269	2971	1458	2498	1778	15
Geneva	2610	803	677	747	853	1662	1418	0	1975	1118	895	1936	158	1439	425	3
Gibraltar	4485	1172	2256	2224	2047	2436	3196	1975	0	2897	2428	676	1817	698	1693	21
Hamburg	2977	2018	597	714	1115	460	460	1118	2897	0	550	2671	1159	2198	1479	12
Hook of Holland	3030	1490	172	330	731	269	269	895	2428	550	0	2280	863	1730	1183	10
Lisbon	4532	1305	2084	2052	1827	2290	2971	1936	676	2671	2280	0	1178	668	1762	22
Lyons	2753	645	690	739	789	714	1458	158	1817	1159	863	1178	0	1281	320	3
Madrid	3949	636	1558	1550	1347	1764	2498	1439	698	2198	1730	668	1281	0	1157	17
Marseilles	2865	521	1011	1059	1101	1035	1778	425	1693	1479	1183	1762	320	1157	0	6
Milan	2282	1014	925	1077	1209	911	1537	328	2185	1238	1098	2250	328	1724	618	
Munich	2179	1365	747	977	1160	583	1104	591	2565	805	851	2507	724	2010	1109	3
Paris	3000	1033	285	280	340	465	1176	513	1971	877	457	1799	471	1273	792	8
Rome	817	1460	1511	1662	1794	1497	2050	995	2631	1751	1683	2700	1048	2097	1011	5
Stockholm	3927	2868	1616	1786	2196	1403	650	2068	3886	949	1500	3231	2108	3188	2428	21
Vienna	1991	1802	1175	1381	1588	937	1455	1019	2974	1155	1205	2937	1157	2409	1363	8











# Variants of MDS

- Metric or Non-metric
- Weighted or Non-weighted
- Single Matrix or Multiple Matrices
- Deterministic or Probabilistic



# Overviews of MDS Techniques

- Torgerson's Classical MDS (CMDS)
- Spring Model
- Chalmers' Linear Iteration Algorithm
- Anchor Point Method



# Torgerson's Classical MDS

- Proposed by Torgerson (1952)
- Given symmetric distance/correlation matrix  $D$
- Using “Double Centering” to transform  $D$  to  $B$
- Process singular value decomposition (SVD) on  $B$  to solve the configuration matrix  $X$

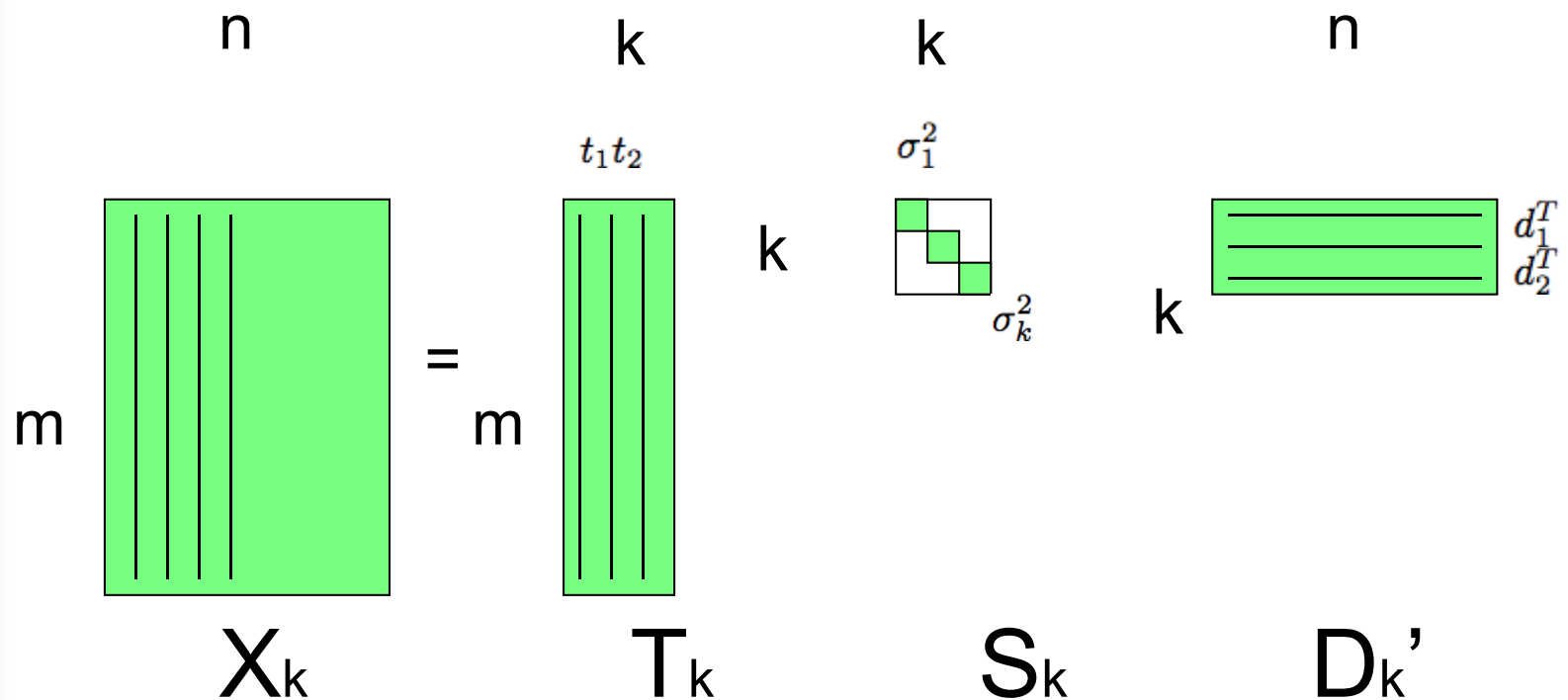


# Double Centering

$$X - \frac{1}{N} X \mathbf{i} \mathbf{i}^T \quad \text{Shift } X \text{ to zero mean}$$

$$\begin{aligned} B &= \left( X - \frac{1}{N} X \mathbf{i} \mathbf{i}^T \right)^T \left( X - \frac{1}{N} X \mathbf{i} \mathbf{i}^T \right) \\ &= X^T X - \frac{1}{N} X^T X \mathbf{i} \mathbf{i}^T - \frac{1}{N} \mathbf{i} \mathbf{i}^T X^T X + \frac{1}{N^2} \mathbf{i} \mathbf{i}^T X^T X \mathbf{i} \mathbf{i}^T \\ &= D - \frac{1}{N} D \mathbf{i} \mathbf{i}^T - \frac{1}{N} \mathbf{i} \mathbf{i}^T D + \frac{1}{N^2} \mathbf{i} \mathbf{i}^T D \mathbf{i} \mathbf{i}^T \\ &= D - \bar{D}_r - \bar{D}_c + \bar{D}_g, \end{aligned}$$

# Truncated-SVD



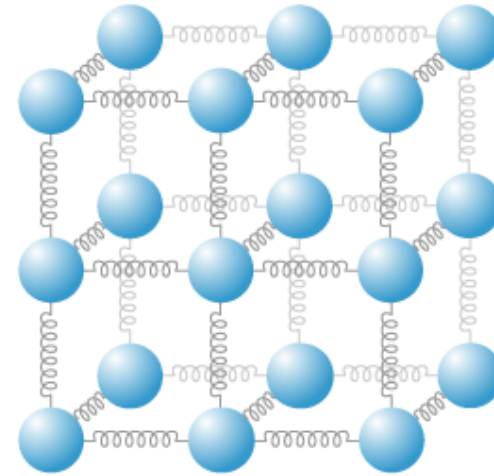
$X_k$  is truncated configuration  
 $X - X_k$  can be consider as the noise



# Properties of CMDS

- Deterministic
- No missing value
- PCA representation
- Denoise automatically
- $O(N^3)$  computation cost

# Spring Model



$$\overrightarrow{E}_i = \frac{1}{N-1} \sum_{j \neq i} \left( \frac{\delta_{i,j} - d_{i,j}}{\delta_{i,j}} \right) (x_j - x_i)$$

$d_{i,j}$  the given distance

$\delta_{i,j}$  the layout distance



# Properties of Springs Model

- Proposed by Eades (1984)
- For each iteration and for each node, it search  $N-1$  nodes to compute the spring force
- Solution is not unique
- Probably converge to local solution
- For each iteration, the computational cost is  $O(N^3)$





# Chalmers' Linear Iteration Algorithm

- Proposed by M. Chalmers (1996)
- Improvement of Springs model
- For each iteration and for each node, it searches only two sets of constant nodes
  - Neighborhood region
  - Randomly choose constant nodes from afar
- Process only constant iterations



# Properties of Chalmers' Linear Algorithm

- Computational cost is  $O(N)$
- Good for updating new node
- The result is not deterministic
- Need good initial guess



# Anchor Point Method

- Proposed by Buja *et al.* 1998
- The distances between points in different clusters are less meaningful
- $K$  points in the same cluster are chosen as anchors, and others ( $N-K$ ) are considered as floaters



# Properties of Anchor Point Method

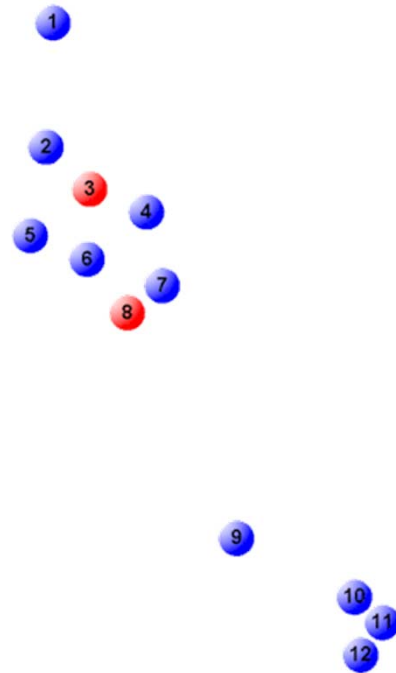
- Compute only  $N \times K$  matrix
- Need pre-knowledge of grouping structure
- The number of anchors could not be smaller than the dimension  $p$  of the given data



# Challenges of MDS

- Dimension Estimation
- Missing data
- Inconsistent data
- Computational Cost

# Distance matrix has many redundancies



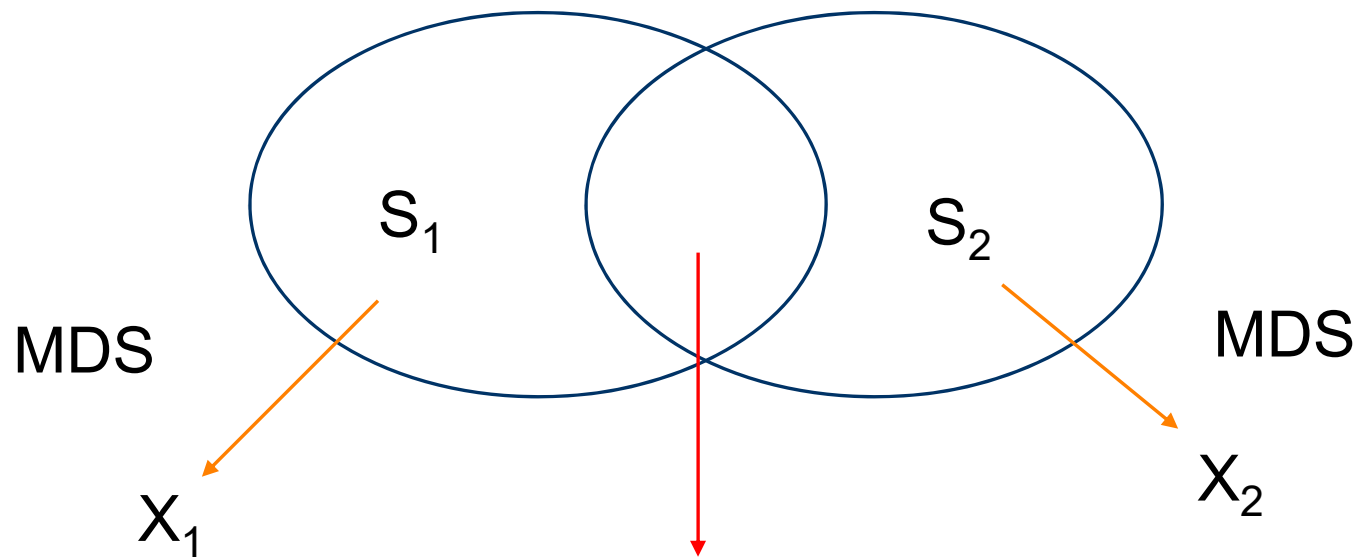
- $p+1$  points can determinate the configurations of points in  $p$  dimension
- For each node, there are  $N-p-1$  redundant information



## Split-and-combine MDS (SCMDS)

- We group  $N$  data into  $K$  overlapping groups
- $N_g$  is the maximal number of points in each group
- The minimal number of points in each intersection region is  $N_i > p$
- Apply MDS to each group
- Using information of each intersection points to recombine the points into one group

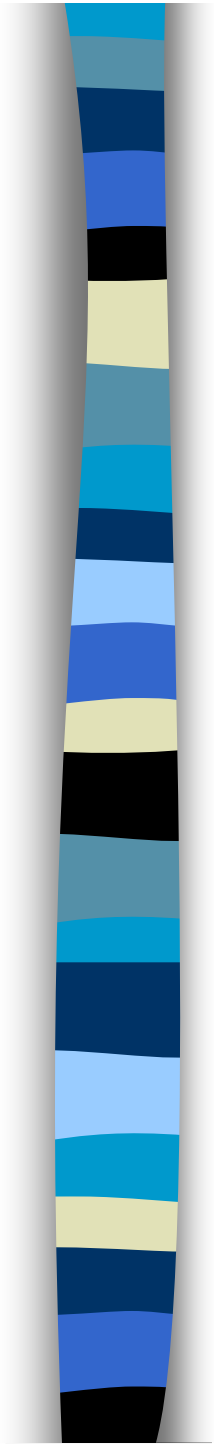
# Example: Two groups



$$x_{k,1} = Ux_{j,2} + b$$

$U$  is a volume-preserving operator





Apply QR transform to  $X_1 - \bar{X}_1 \mathbf{i}^T$  and  $X_2 - \bar{X}_2 \mathbf{i}^T$

$$X_1 - \bar{X}_1 \mathbf{i}^T = Q_1 R_1$$

$$X_2 - \bar{X}_2 \mathbf{i}^T = Q_2 R_2$$

$R_1 = R_2$  after sign modification of  $Q_2$

$$Q_1^T (X_1 - \bar{X}_1 \mathbf{i}^T) = Q_2^T (X_2 - \bar{X}_2 \mathbf{i}^T)$$

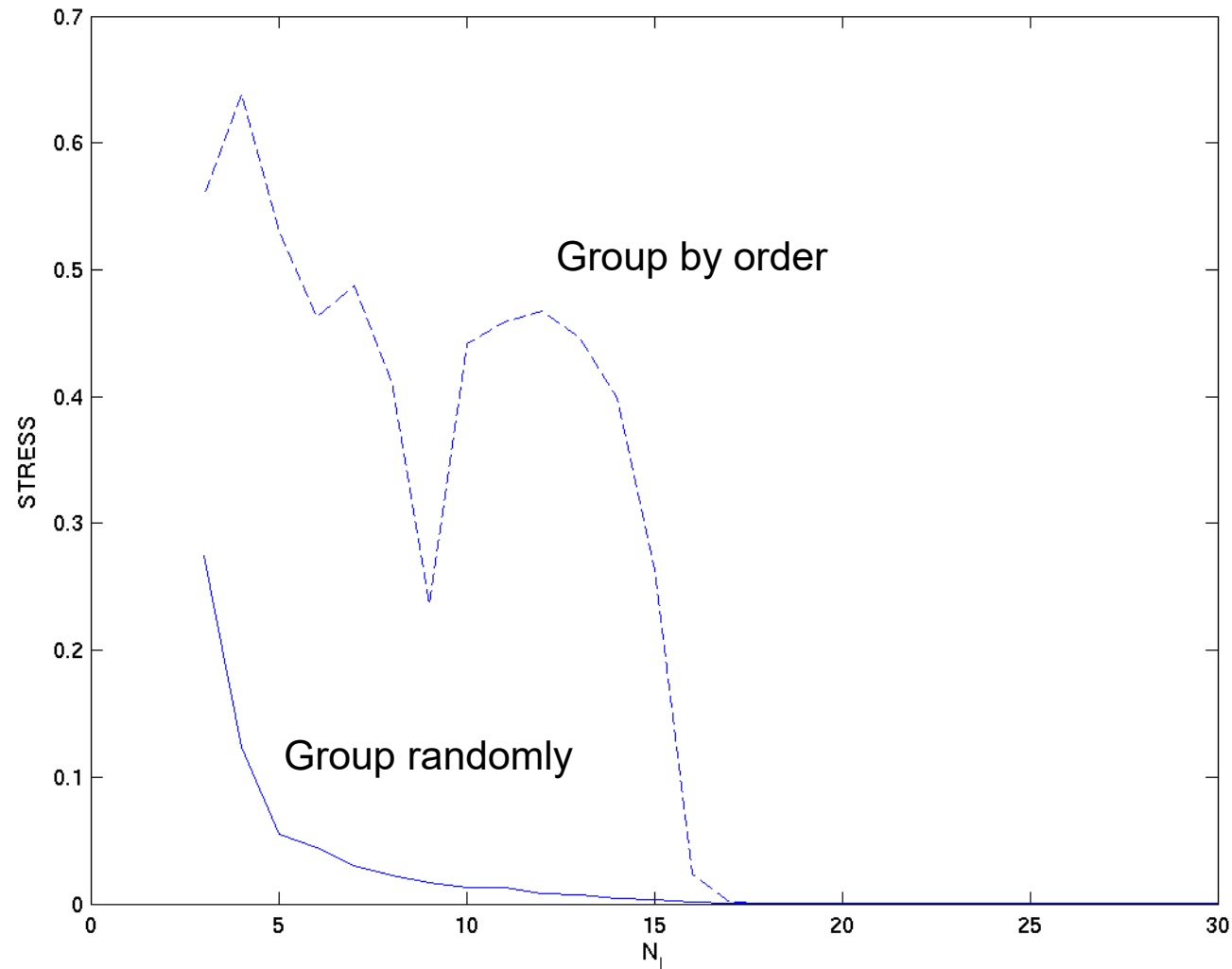
$$X_1 = Q_1 Q_2^T X_2 - Q_1 Q_2^T (\bar{X}_2 \mathbf{i}^T) + \bar{X}_1 \mathbf{i}^T.$$



# Split Criteria

- The minimal number  $N_i$  of points in the intersection region should be larger than the original dimension  $p$
- The points cluster in the same group should contain both neighboring and afar

# Compare errors of grouping method





# Definition of STRESS

$$\textit{Stress} = \sqrt{\frac{\sum_{i,j} (d_{i,j} - \hat{d}_{i,j})^2}{\sum_{i,j} d_{i,j}^2}}$$



## $O(N^3)$ to $O(N)$

$N_g$  is the number of each group

$N_I$  is the number of points in the intersection

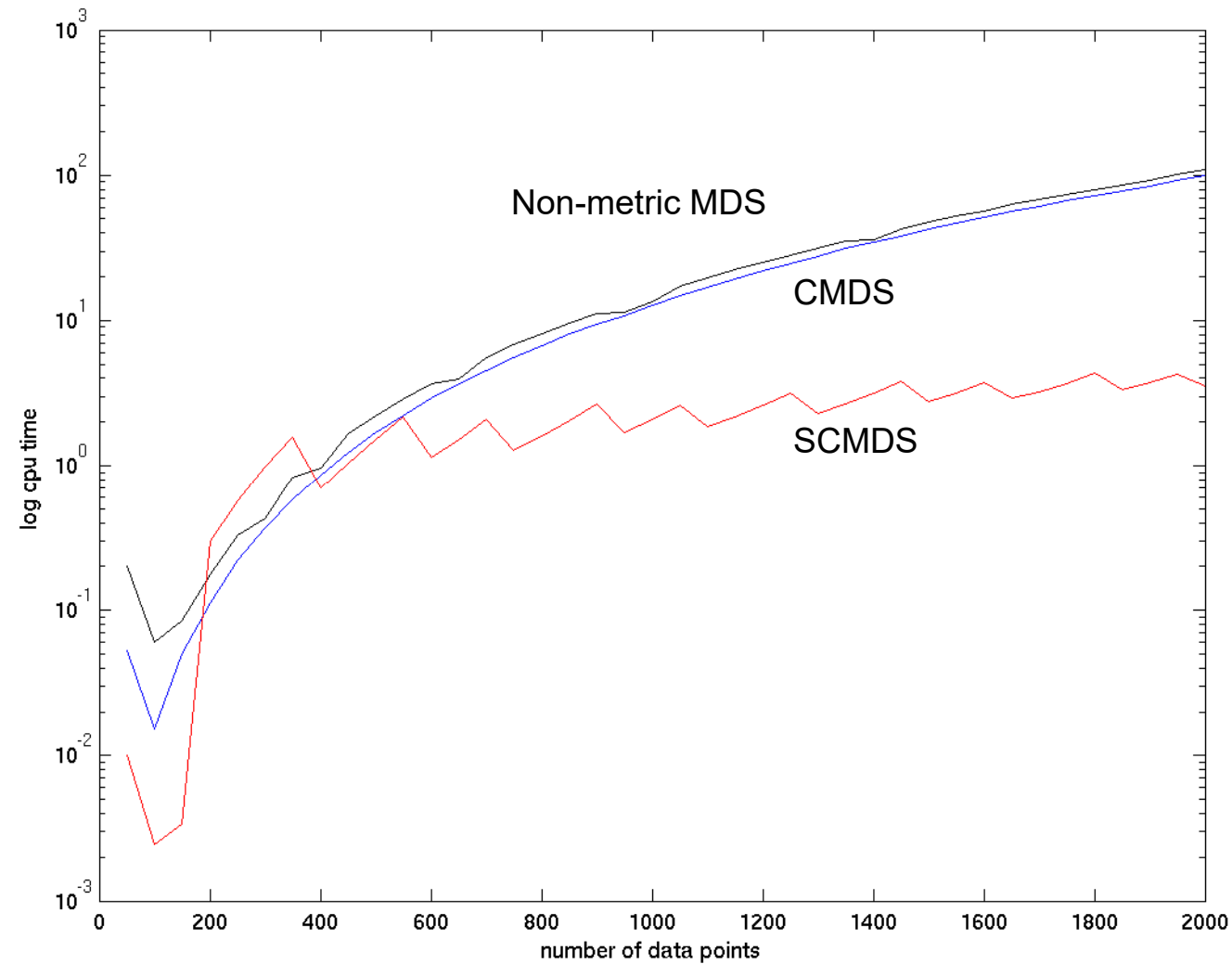
$K$  is the group number  $K = \frac{(N - N_I)}{(N_g - N_I)} \sim O(N)$

The low bound of  $N_I$  is  $p+1$ , assume  $N_g = \alpha p$

The computational complexity is

$$\frac{N - p}{(\alpha - 1)p} O(\alpha^3 p^3) + \frac{N - \alpha p}{(\alpha - 1)p} O(p^3) \sim O(p^2 N).$$

# Speed comparison for MDS





# Simulation experiment - Spiral data

$$X = [q_1, q_2, n_3, \dots, n_p]$$

$$q_{j,i} = \alpha 2\theta_{\lfloor i/k \rfloor} \cos \theta_{\lfloor i/k \rfloor} + n_{j,i}$$

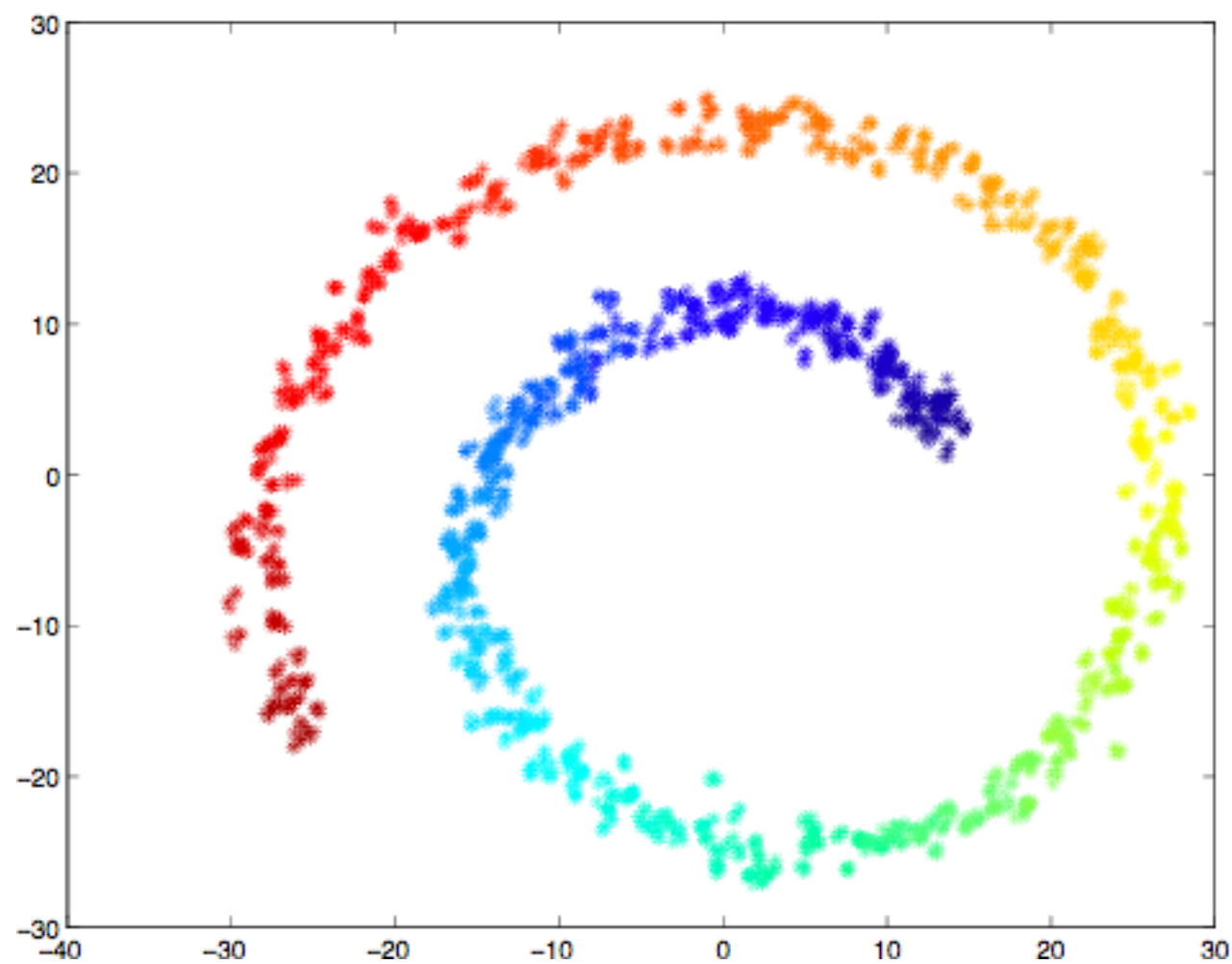
$$n_i \sim N(0,1)$$

$$p = 17$$

$$j = 1, 2$$

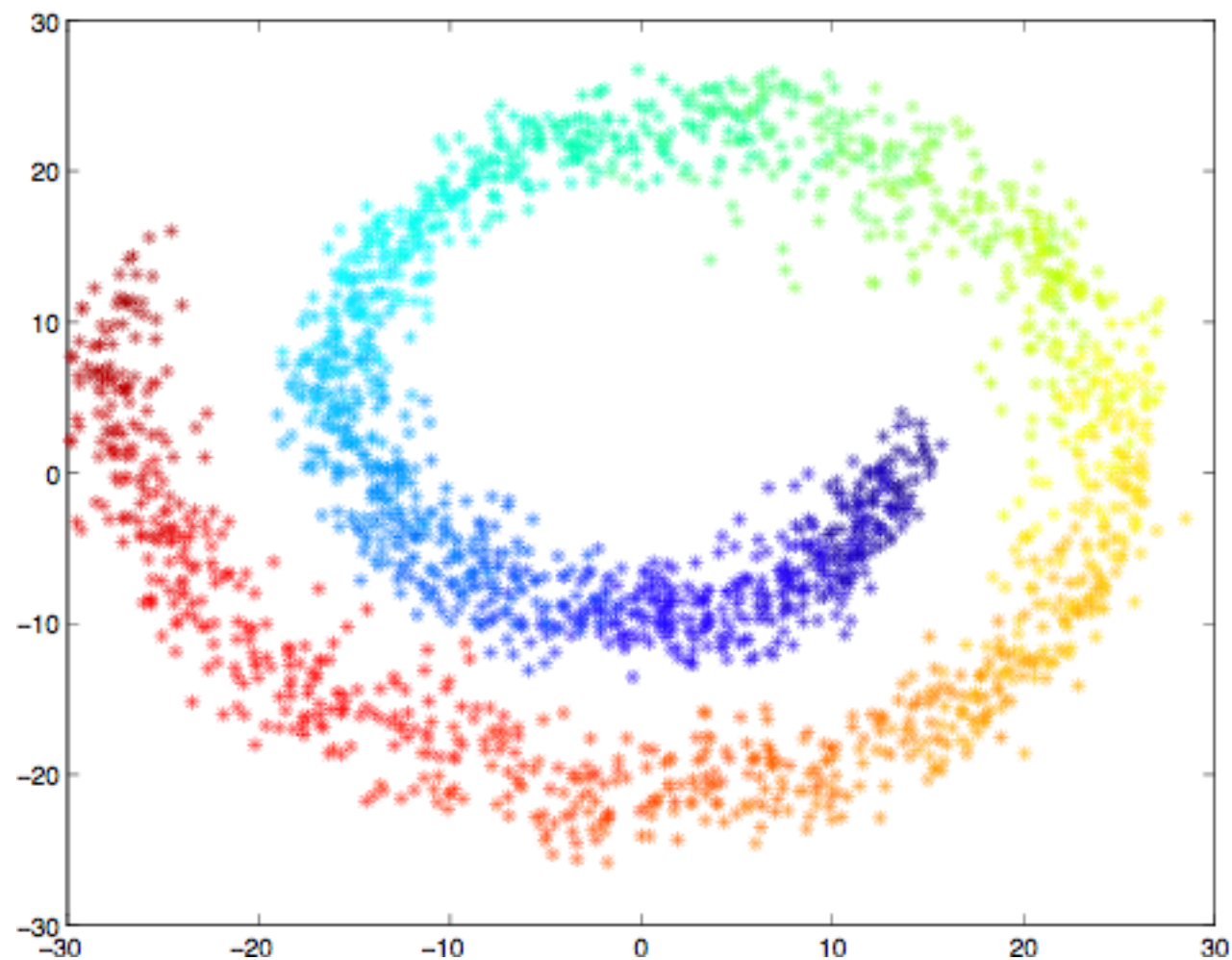
$$i = 1, \dots, kN$$

# Simulation results

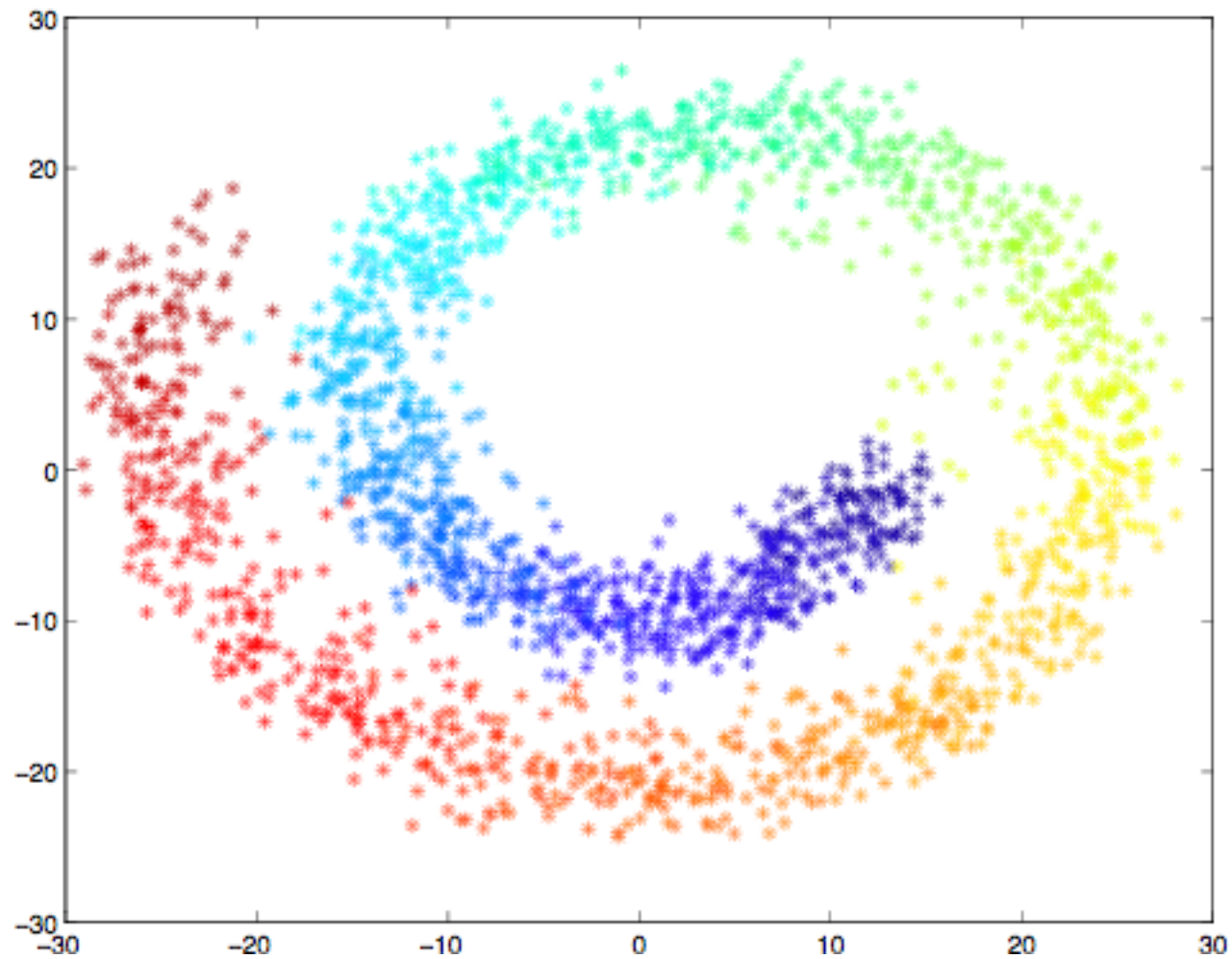




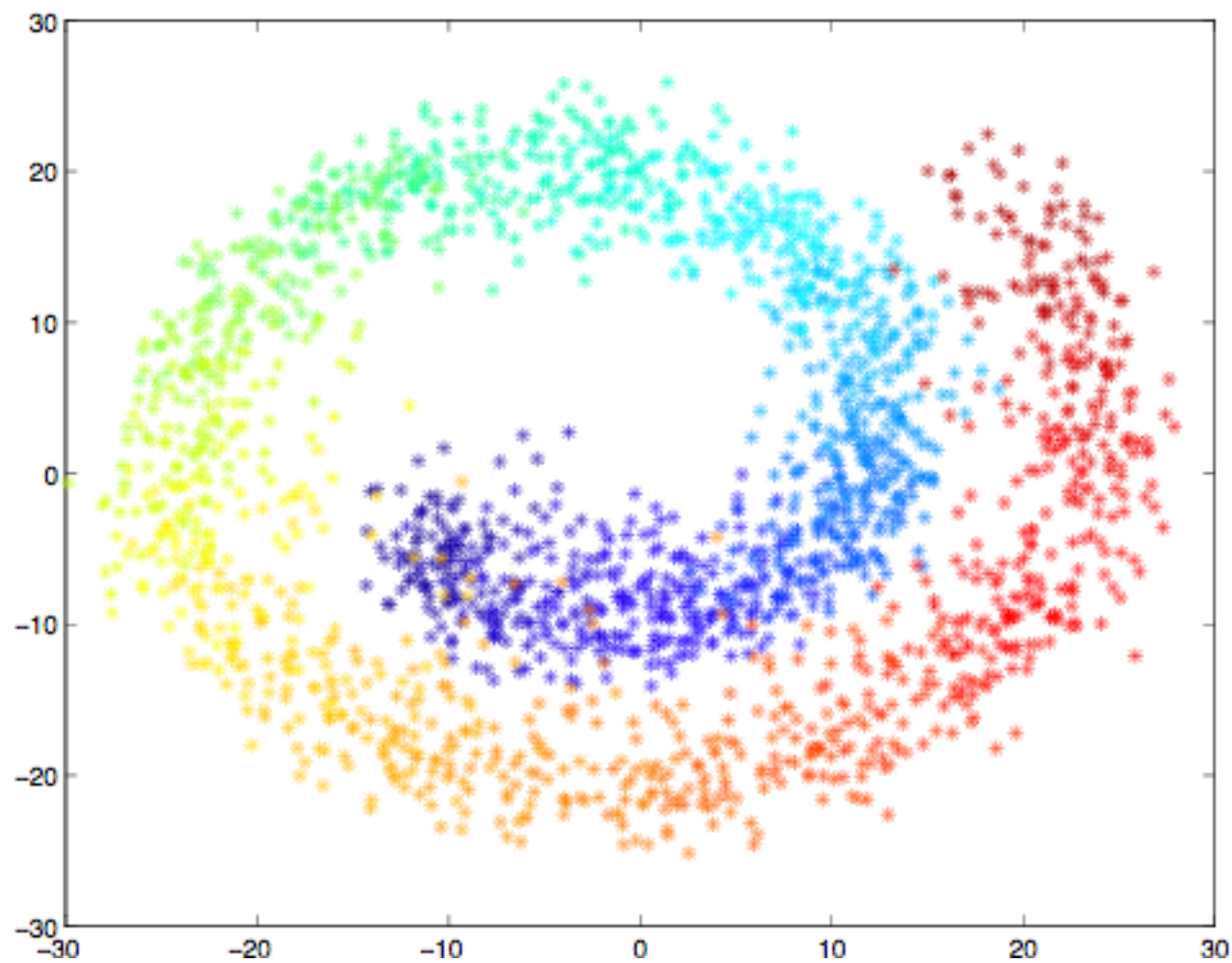
# One Dimension Loss



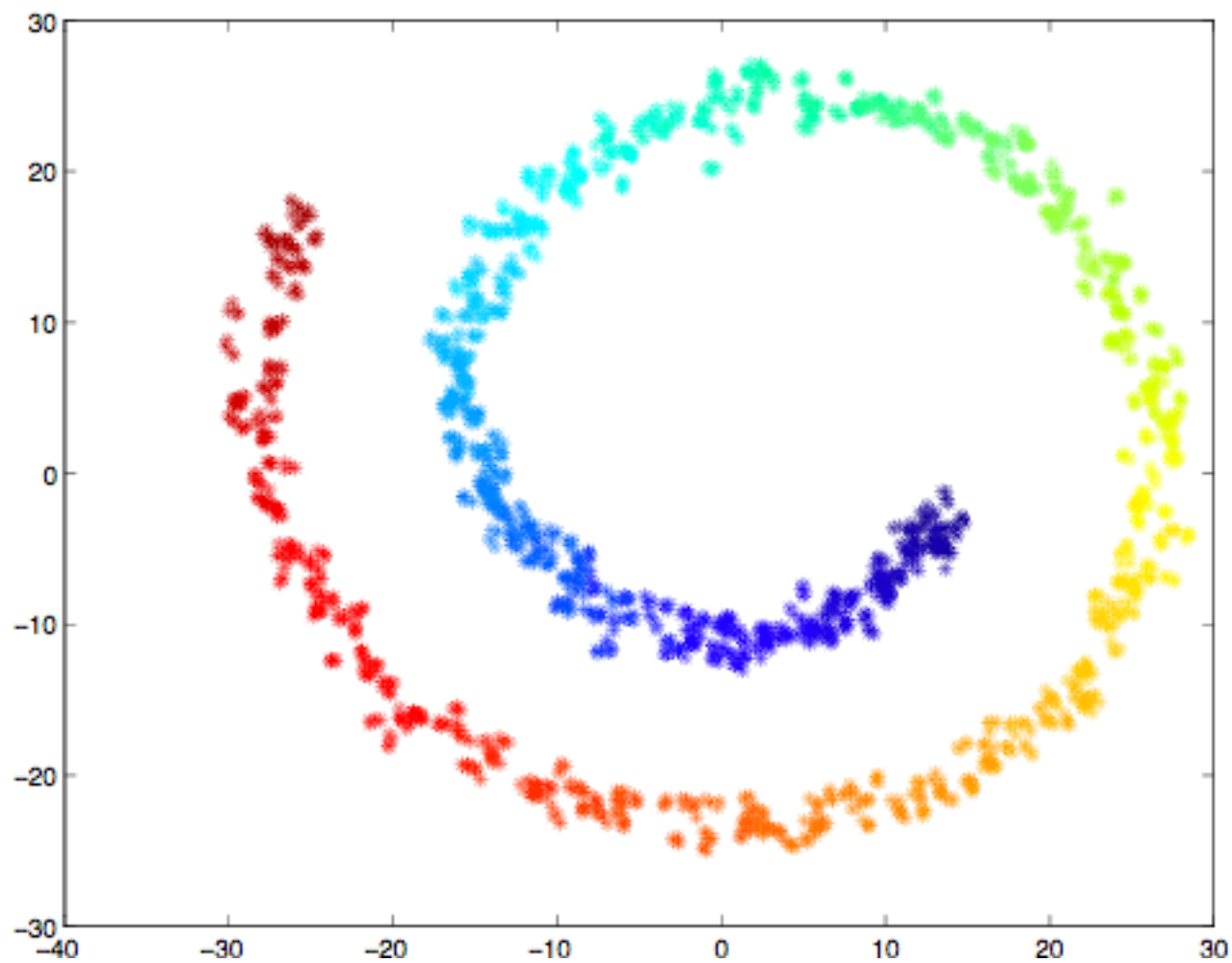
# Two Dimensions Loss



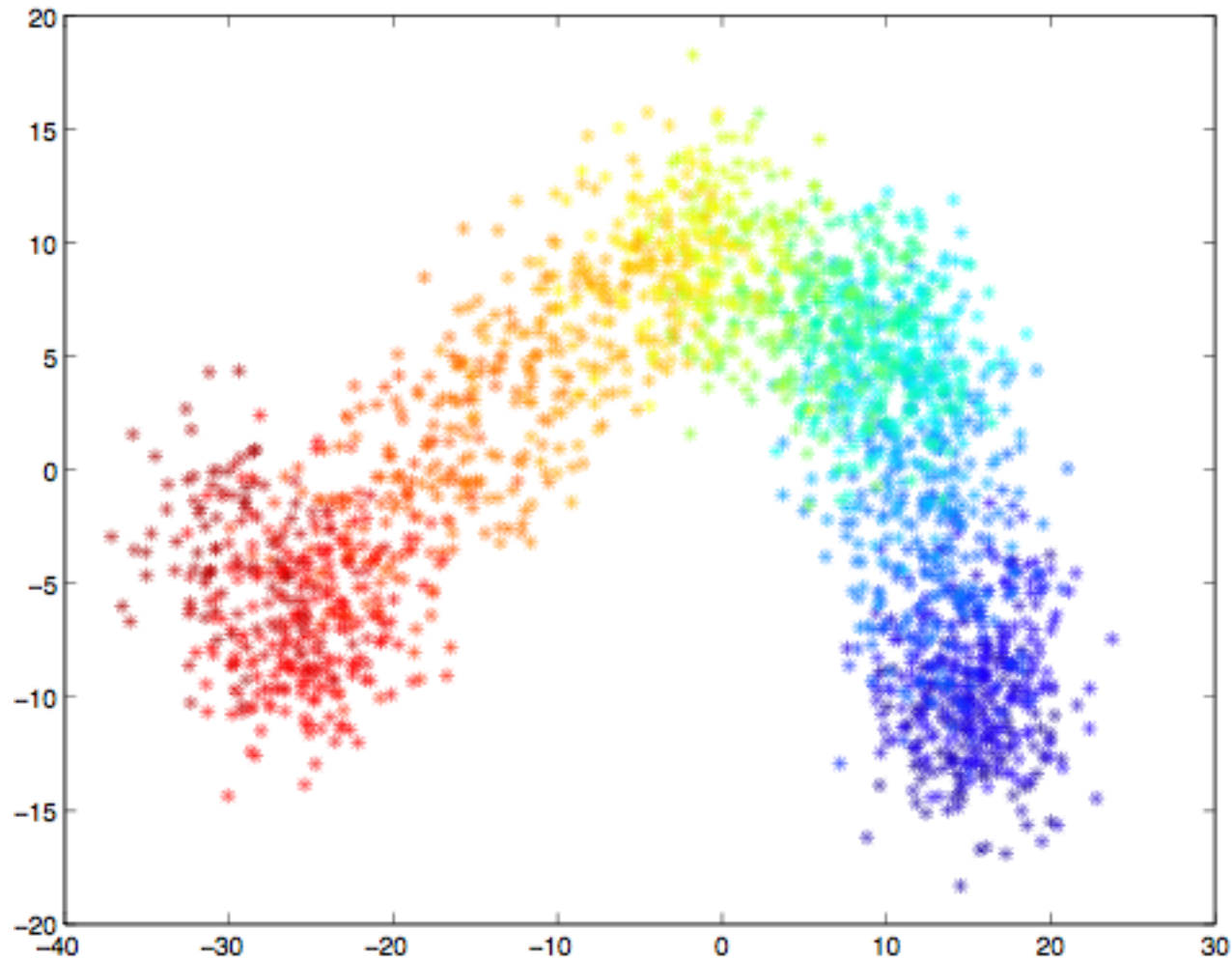
# Three Dimensions Loss



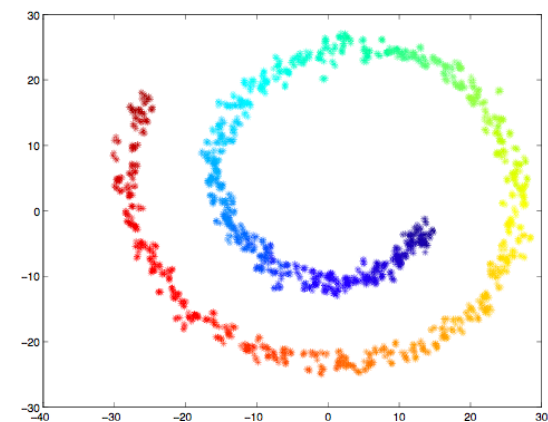
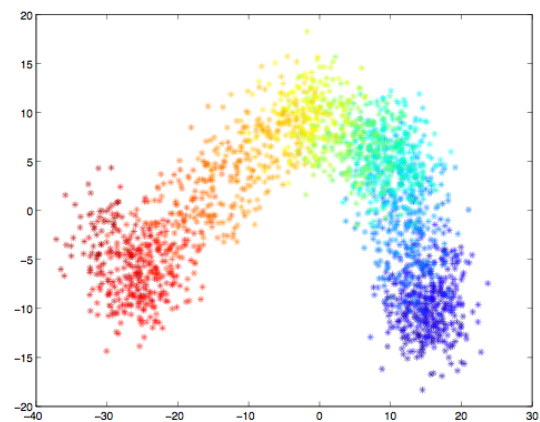
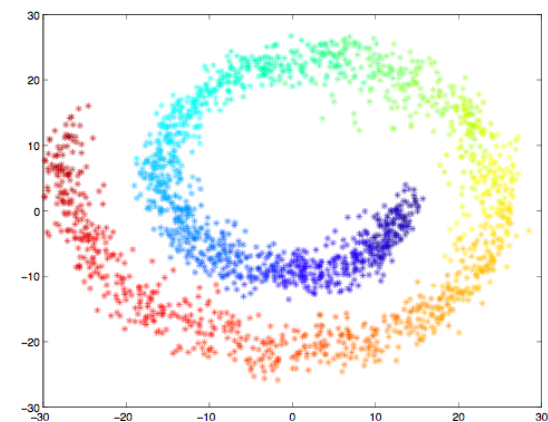
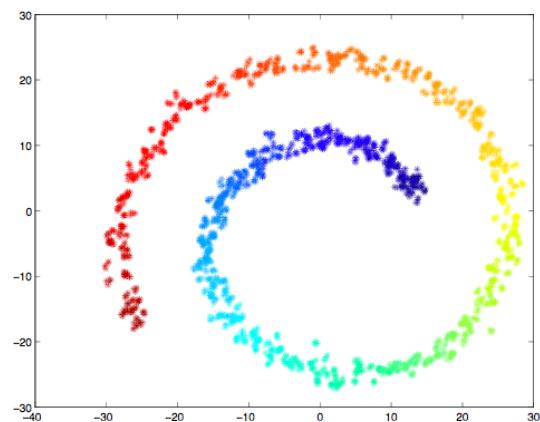
# No Dimension Loss



# No Afar information & One Dimension Loss



# Split Condition





# Properties of SCMDS

- Computational complexity is  $O(N)$
- Allow missing data
- Good for local updating
- Deterministic
- Allow parallel computing



# Drawbacks of SCMDS

- Possible dimension loss when splitting criteria are not satisfied
- Not suitable for  $p$  is large,  $p > \sqrt{N}$





# Applications to Microarray Data

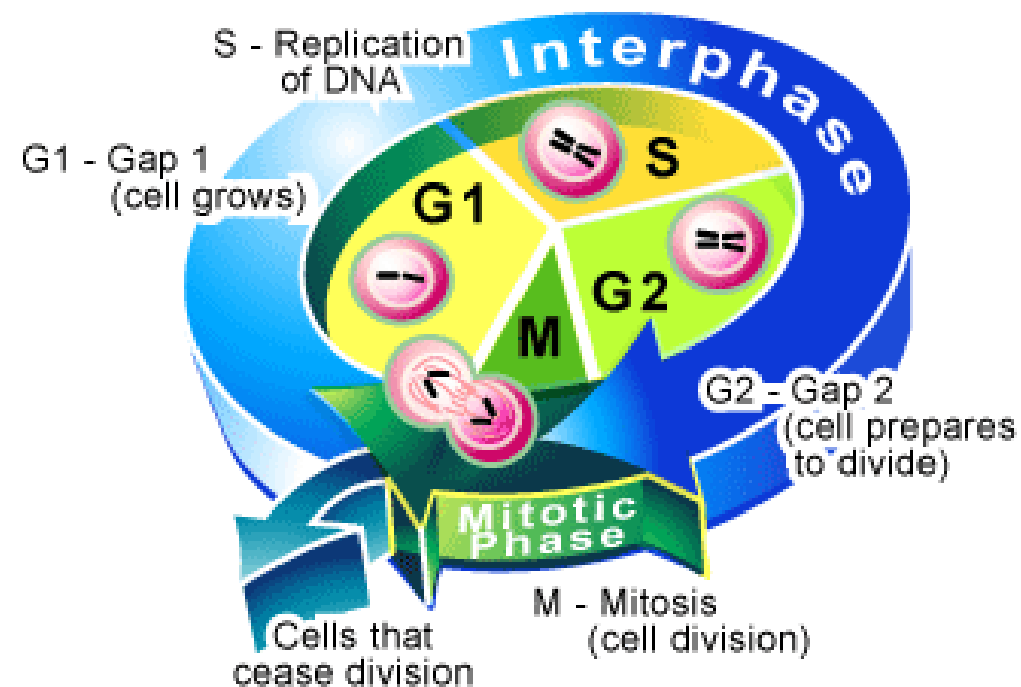
- $N$ : Large number of Genes
- $p$ : Small number of time experiment
- With missing value
- With noise



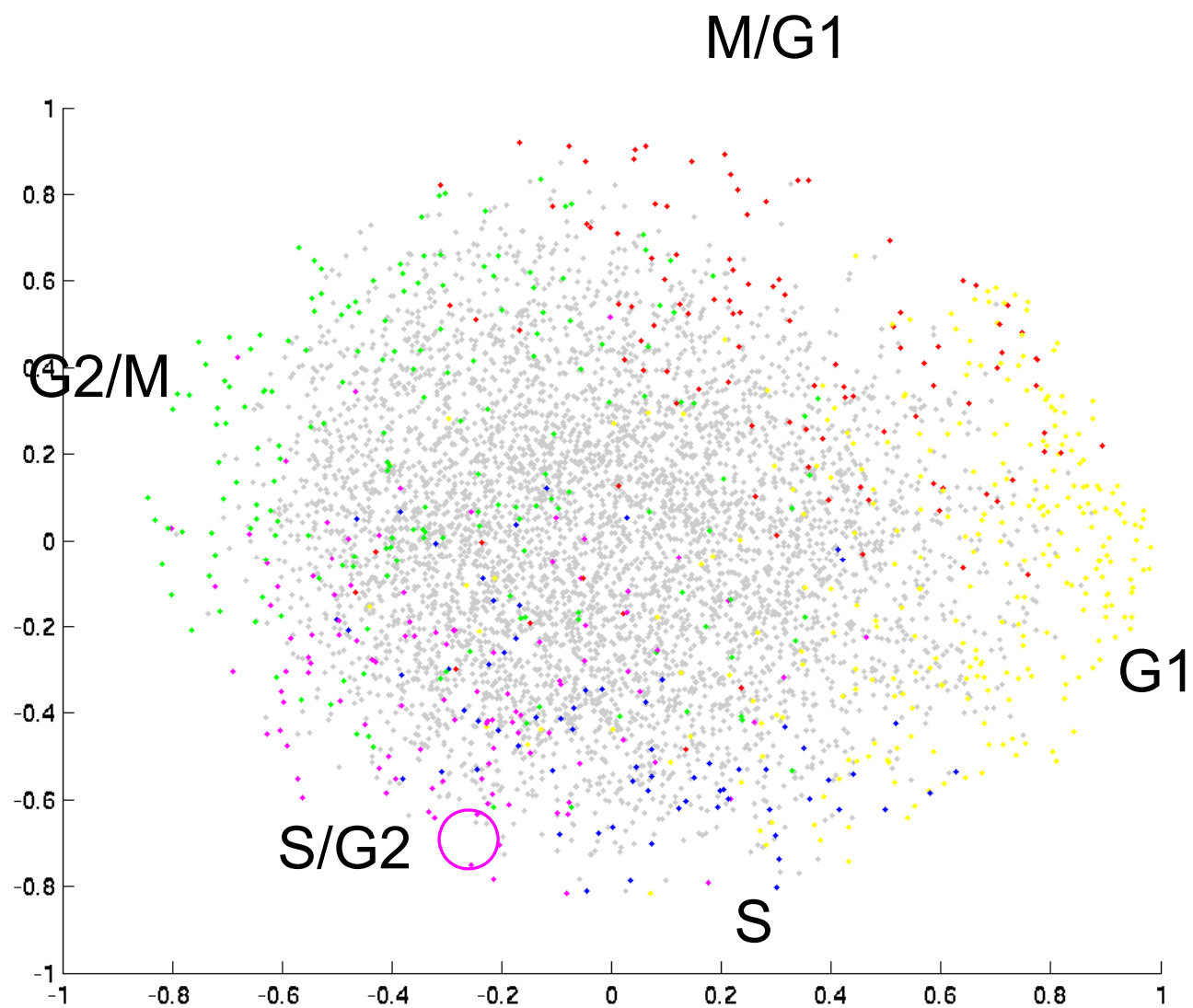
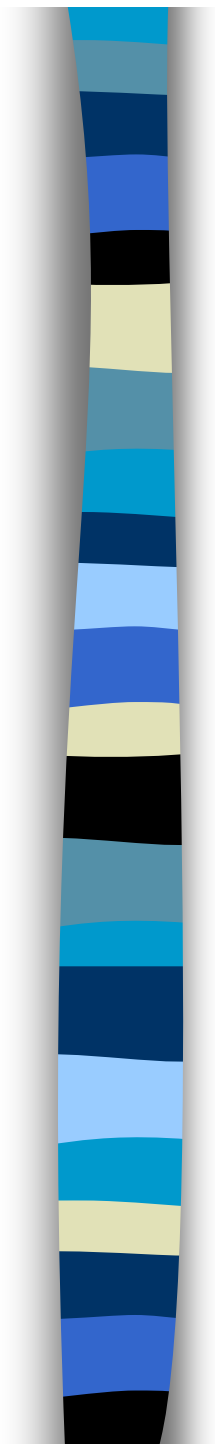
# Yeast Cell Cycle Data (Spellman et al.)

- Data contains 6178 genes
- 5672 genes with less than two missing values
- 753 genes are labeled
- There are 18 time samplings for each gene
- Transform time course data to frequency base by DCT
- $N_g=60$  &  $N_l=30$ , CPU time is 9 sec

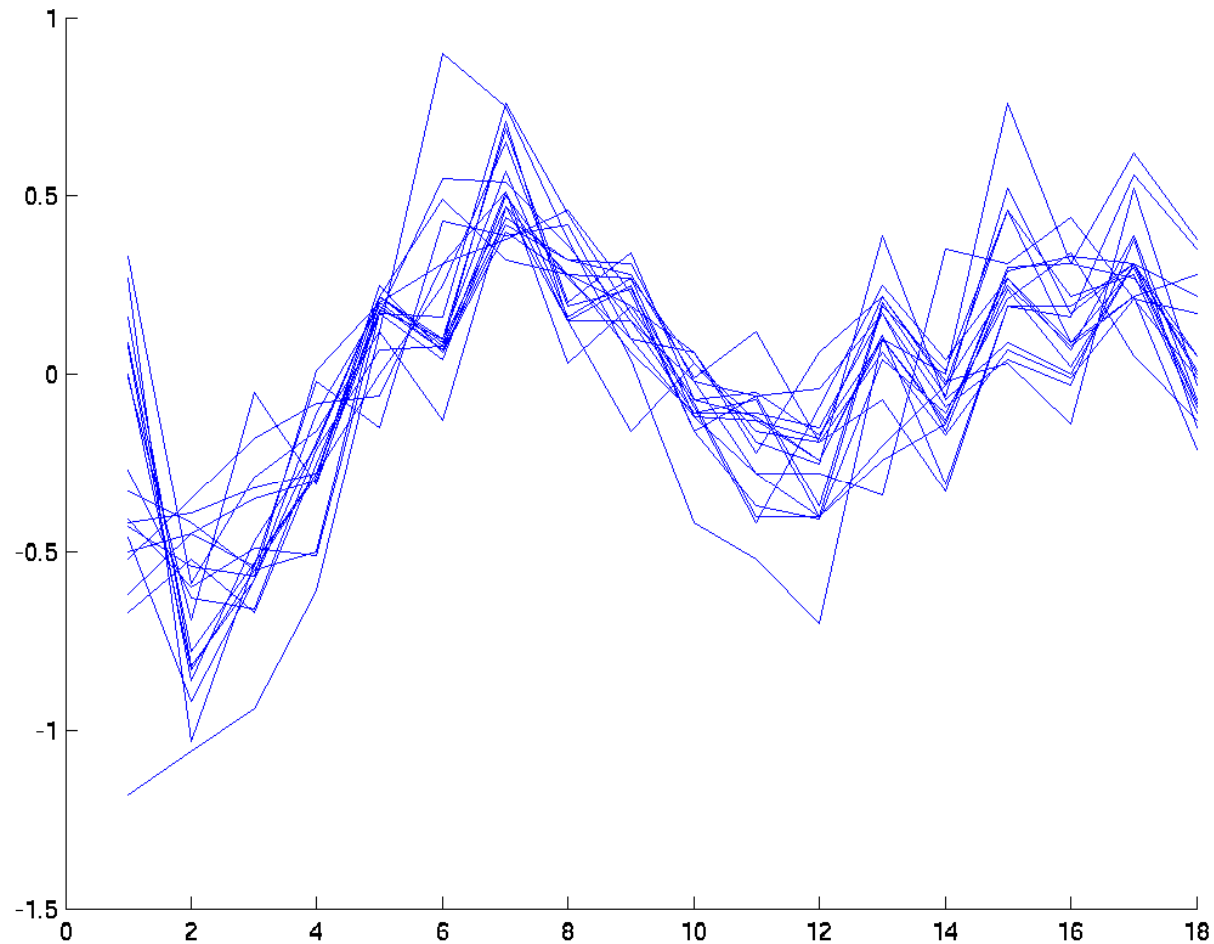
# Cell Cycle



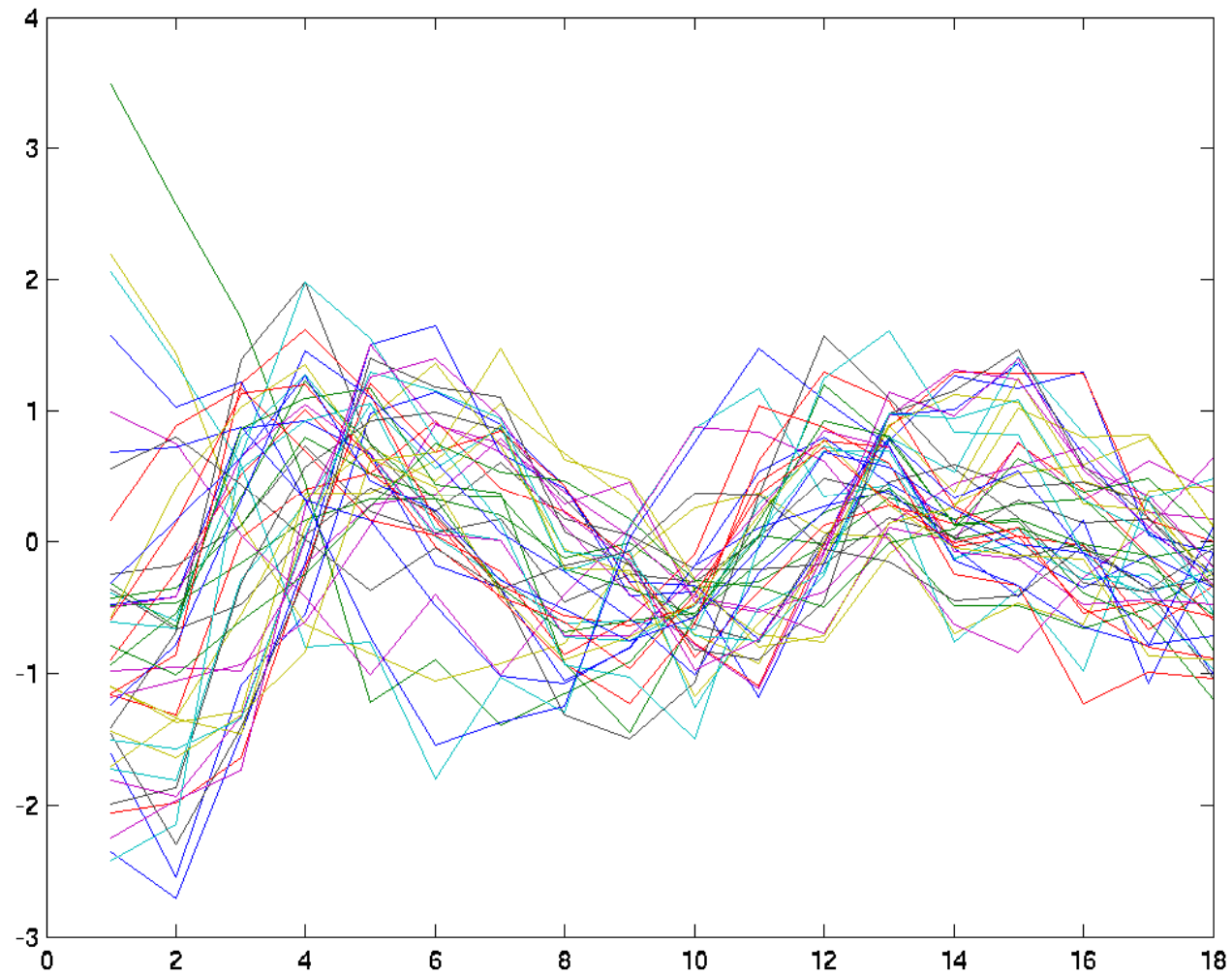
<http://www.bioteach.ubc.ca/CellBiology/TheCellCycle/>



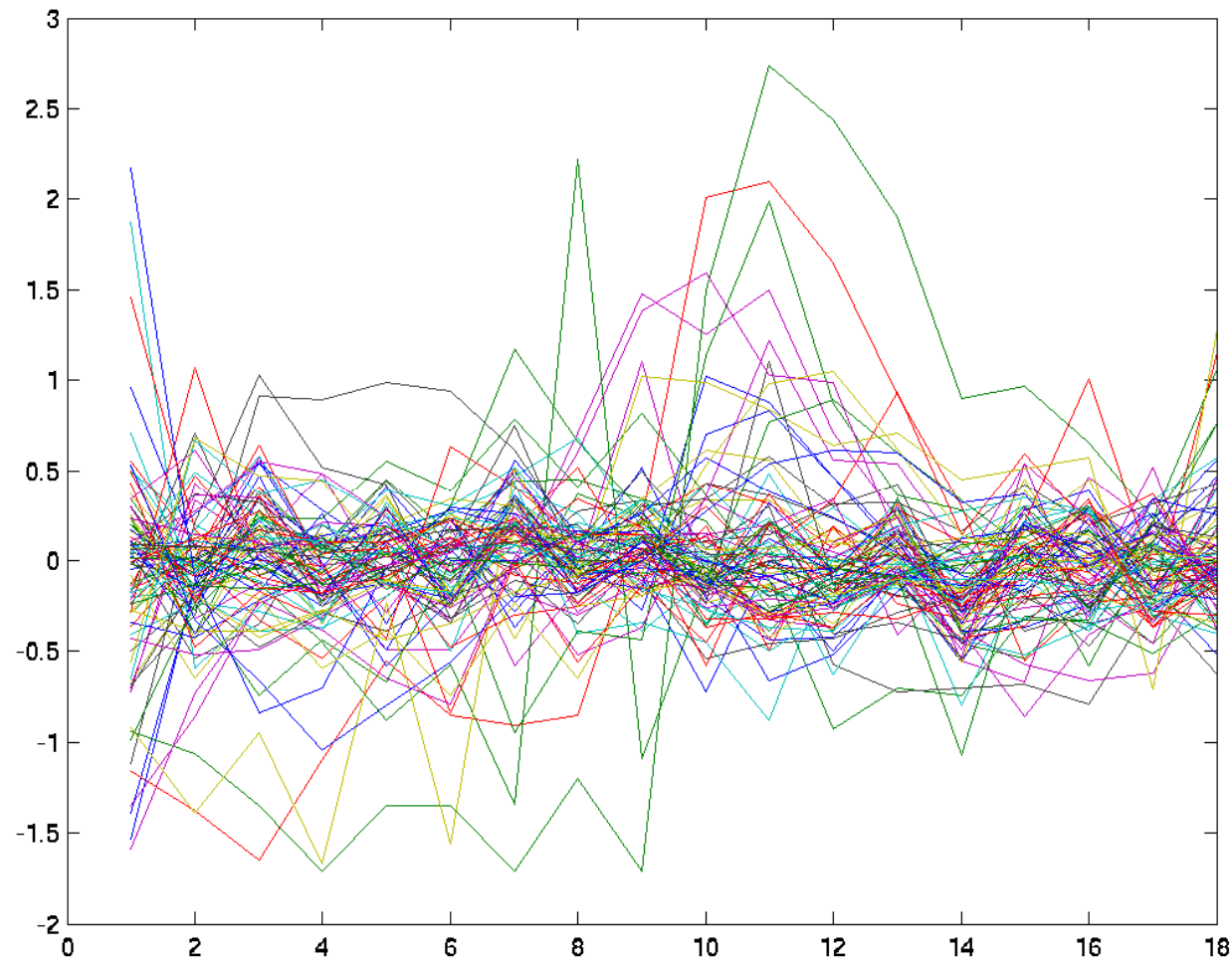
# Expression profile of genes within pink circle

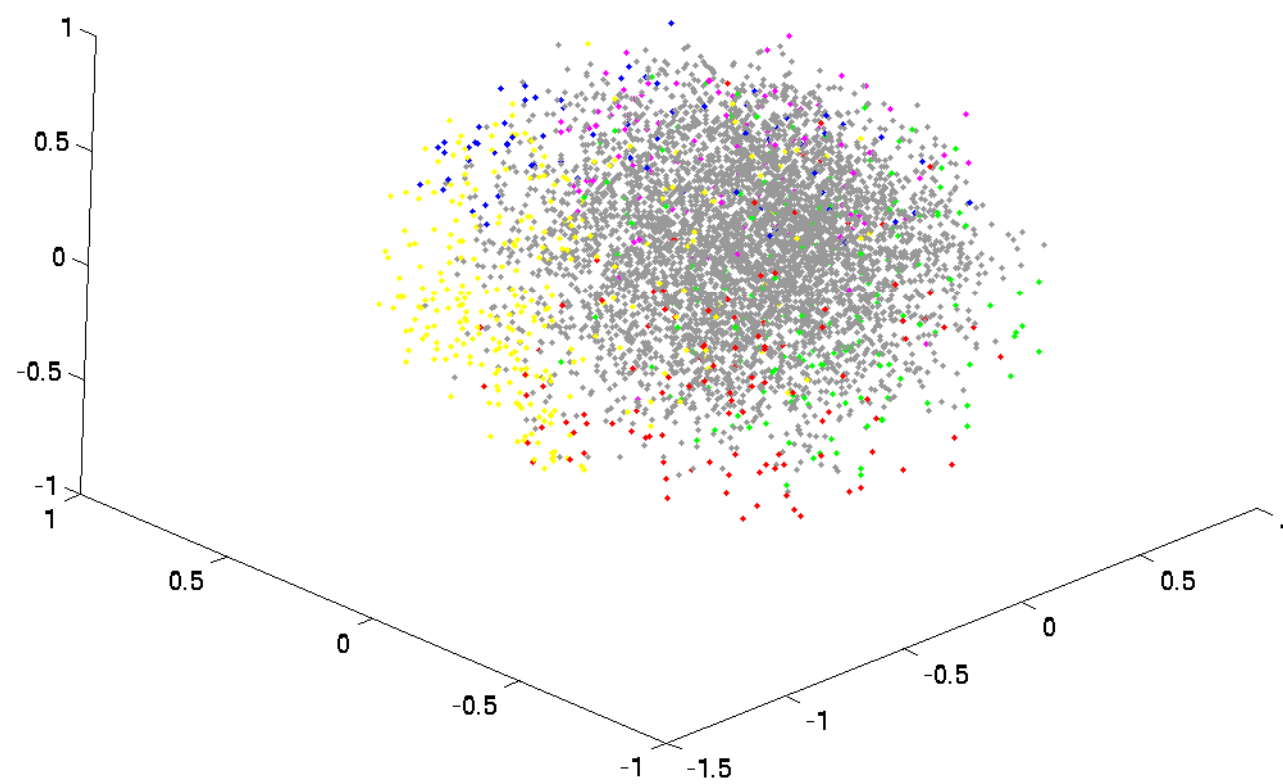
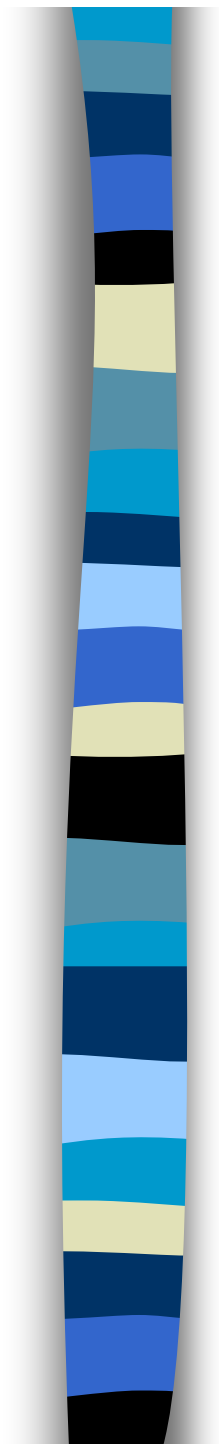


# Expression level in the outer part



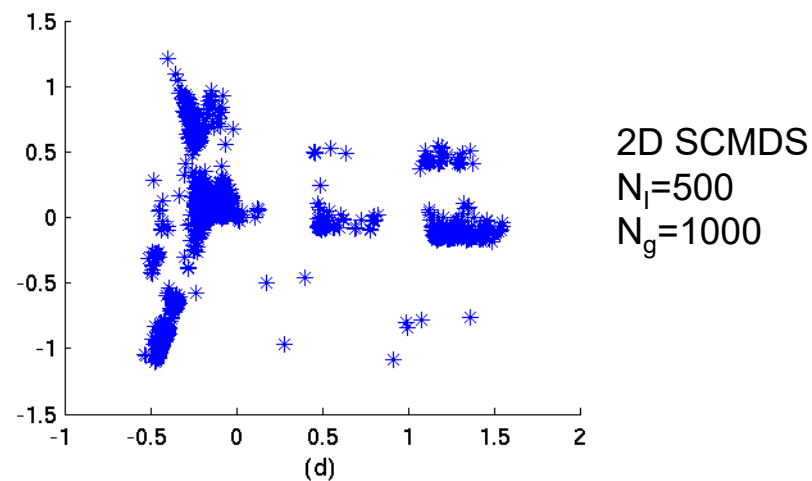
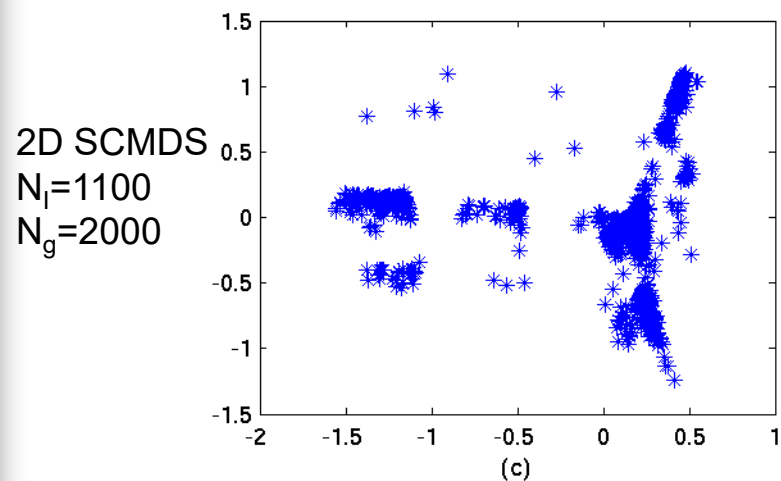
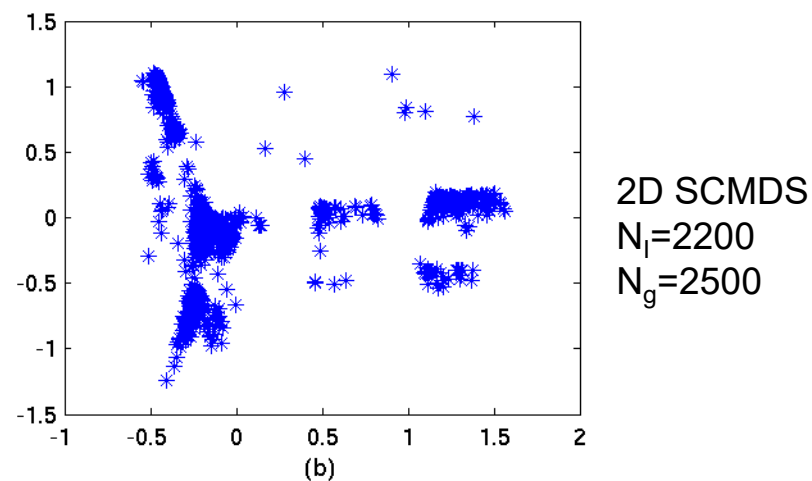
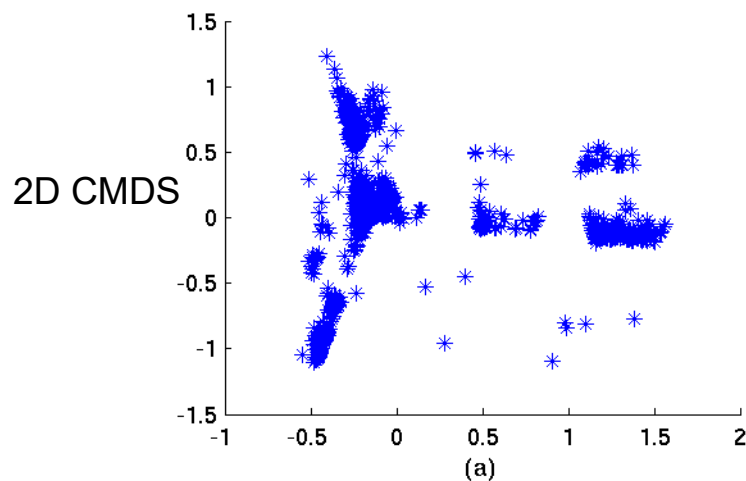
# Expression level in the inner part







# GO Correlation Map





# Conclusion and Discussion

- The proposed SCMDS method is good for  $p \ll N$
- This method is  $O(N)$  complexity
- Grouping method and dimension estimate will affect the outcome
- Grouping method is more crucial than dimension estimation