Group Meeting

2014.09.29

1. Sparse Subspace Clustering
2. Partial Least Squares
3. TCGA Data Analysis

Sparse Subspace Clustering

PCA in a Geometric View:

Given a set of points $\{x_j\}_{j=1}^N$ in R^D, try to find an (affine) subspace $S \subset R^D$ of dimension d, $\dim(S) = d$ that best fits these points.

$$x_j = \mu + U_d y_j + \epsilon_j \; j = 1, \cdots, N$$

where $U_d \in R^{D \times d}$ whose columns form a basis for the subspace and y_j is the vector of new coordinates of x_j in the subspace.

$$\min_{\mu, U_d, \{y_j\}} \quad \sum_{j=1}^{N} ||x_j - \mu - U_d y_j||^2$$

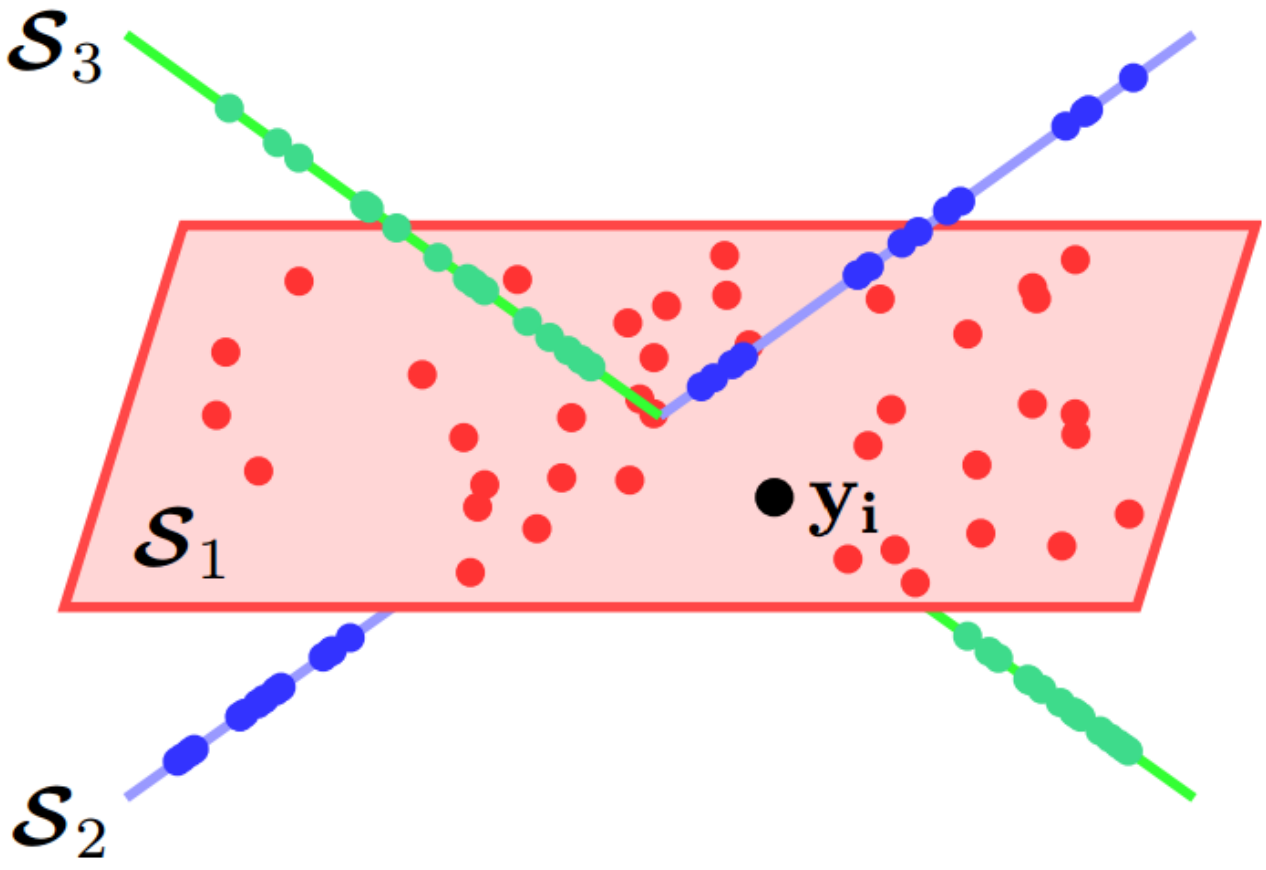
$$s.t. \qquad U_d^T U_d = I_d, \qquad \sum_{j=1}^{N} y_j = 0$$

Subspace Clustering Formulation:

Given a set of sample points $X = \{x_j \in R^D\}_{j=1}^N$ drawn from $n \geq 1$ distinct linear subspaces $S_i \subset R^D$ of dimensions $d_i < D$, $j = 1, \cdots, n$.

Identify each subspace S_i without knowing which sample point belong to which subspace.

1. Identifying the number of subspaces n and their dimensions $d_i = \dim(S_i)$
2. Identifying the orthonormal basis for each subspace S_i.(or equivalently a basis for its orthogonal complement $S_i^{\perp}$)
3. Clustering the N points into the subspaces to which they belong


$\mathcal{S}_3$
$\mathcal{S}_1$
$\mathbf{y_i}$
$\mathcal{S}_2$

How to define a good similarity matrix?

Characterize the local or global subspace structures around the points of interest

- Two points could be very close, but lie in different subspaces(e.g. near the intersection of two subspaces)
- Two points could be far from each other, but belong to the same subspace

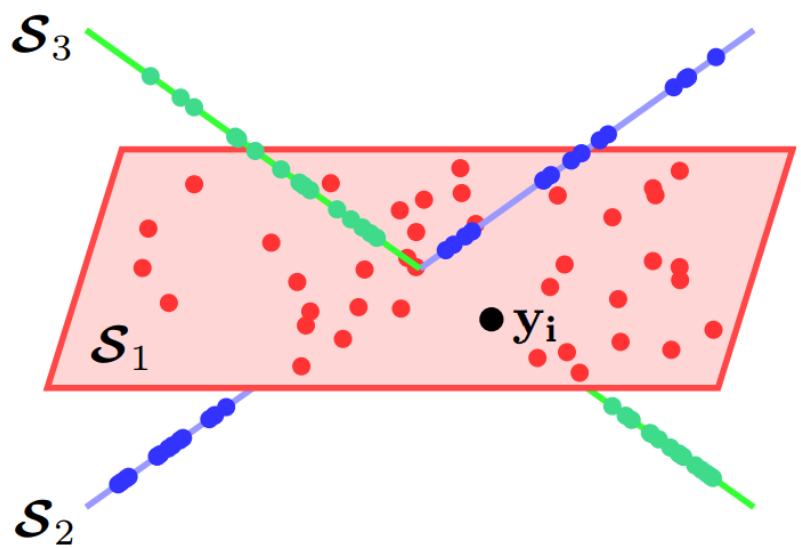
Sparse Subspace Clustering: Key observations and conclusions

Sparse Subspace Clustering: Key observations and conclusions

Self-expressiveness

Each data point in a union of subspaces can be expressed as a (sparse) linear or affine combination of all other points in the dataset.

$$x_j = \sum_{i \neq j} c_{ij} x_i, \qquad \text{or} \qquad X = XC, \text{diag}(C) = 0$$



Sparse Representation Theory

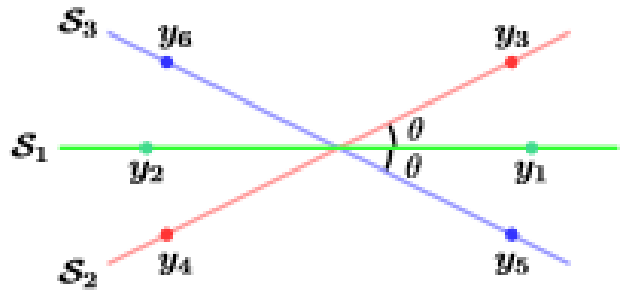
Given the following conditions:

- The subspaces are sufficiently seperated
- The data within the subspaces are well distributed

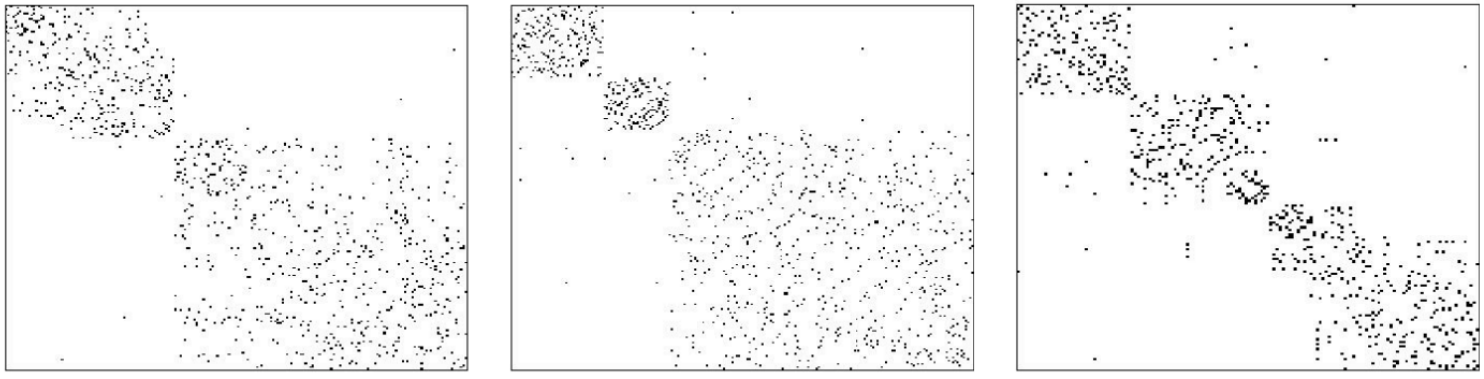
we can rescover the sparse subspace representation by the optimization problem:

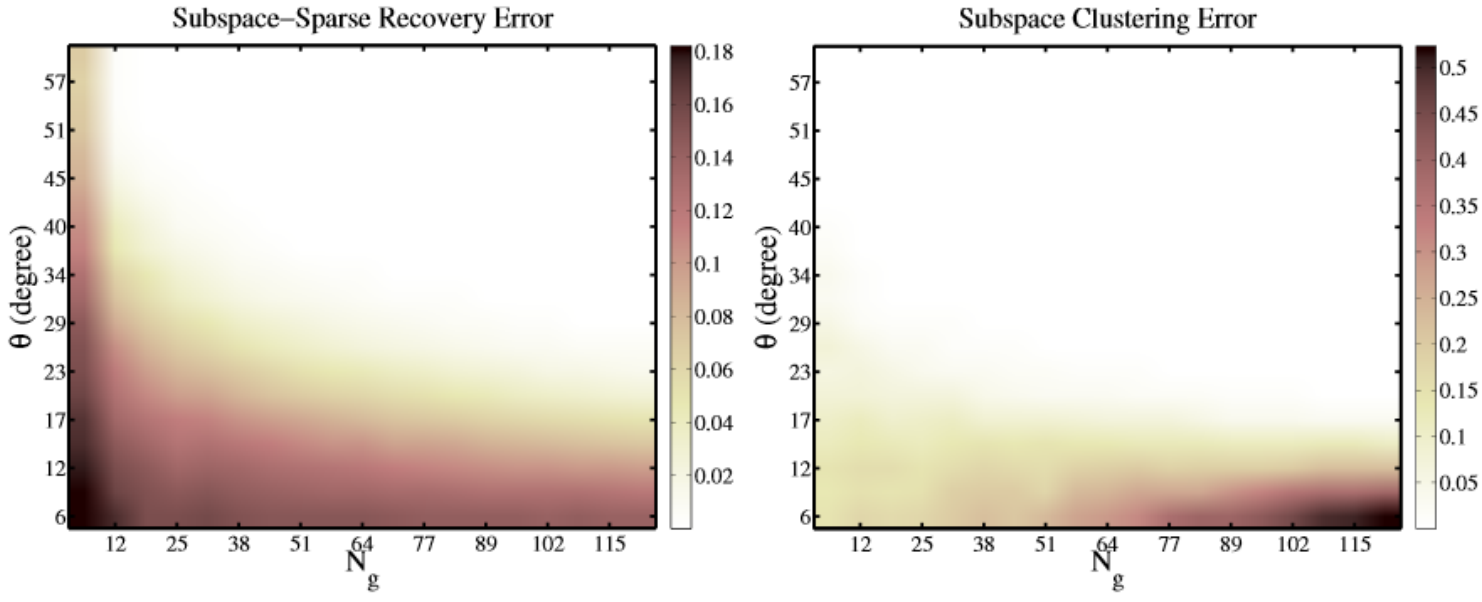
$$\min_{C} ||C||_1 \qquad \text{s.t. } X = XC, \text{diag}(C) = 0$$

where $||C||_1 = \sum_{i,j=1}^{N} |c_{ij}|$



$$C_1 = \left(\begin{array}{cc|cc|cc} 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right)$$




Subspace–Sparse Recovery Error
θ (degree)
57
51
45
40
34
29
23
17
12
6
12 25 38 51 64 77 89 102 115
N_g
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
Subspace Clustering Error
θ (degree)
57
51
45
40
34
29
23
17
12
6
12 25 38 51 64 77 89 102 115
N_g
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05

Clustering Using Sparse Coefficients

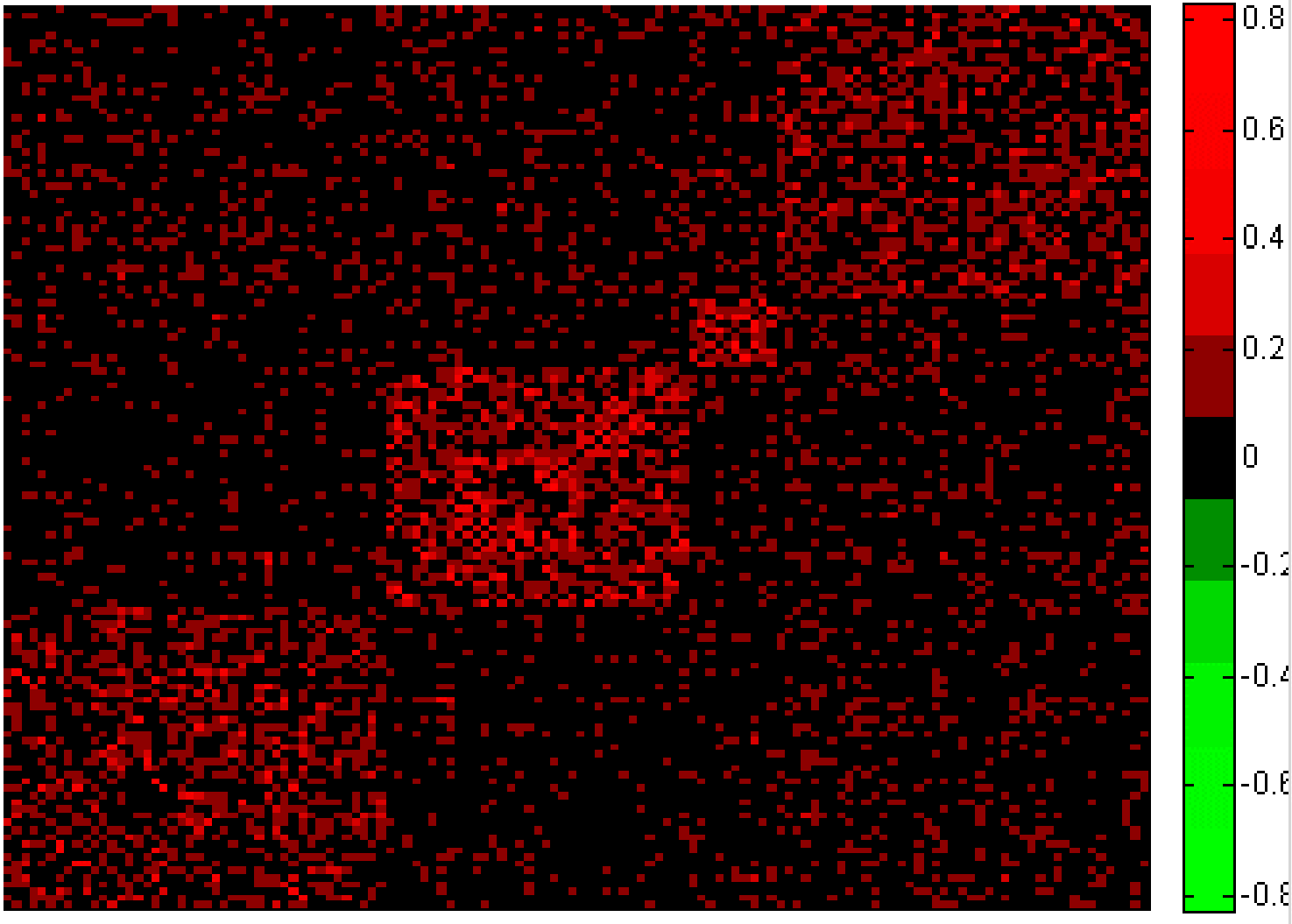
We can use the spectral clustering algorithm.

1. Form a similarity graph $\mathcal{G} = (V, E, W)$ whose nodes are the N data points and whose edges connect points x_i and x_j with a weight $w_{ij} = |c_{ij}| + |c_{ji}|$.

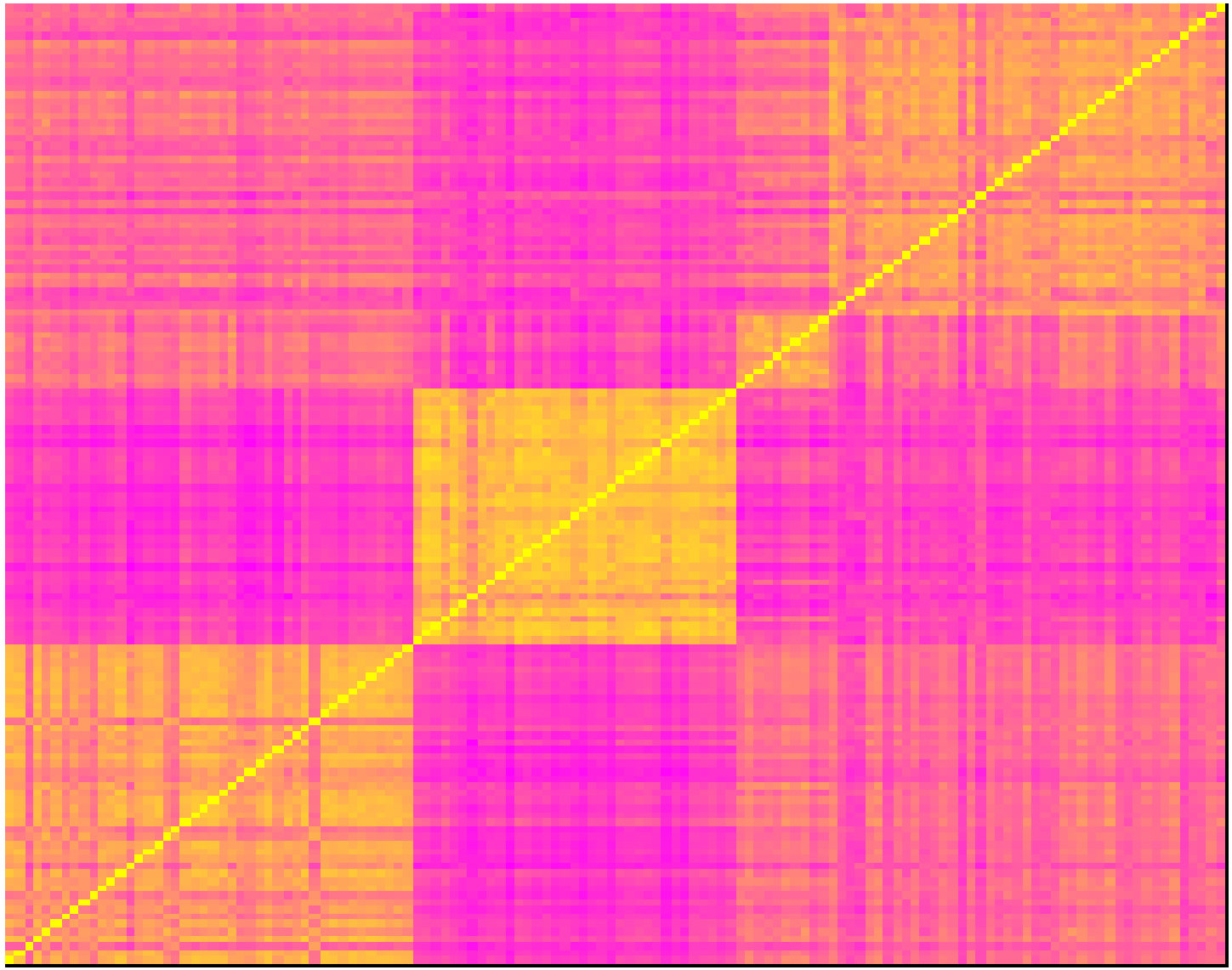
$$W = |C| + |C|^T$$

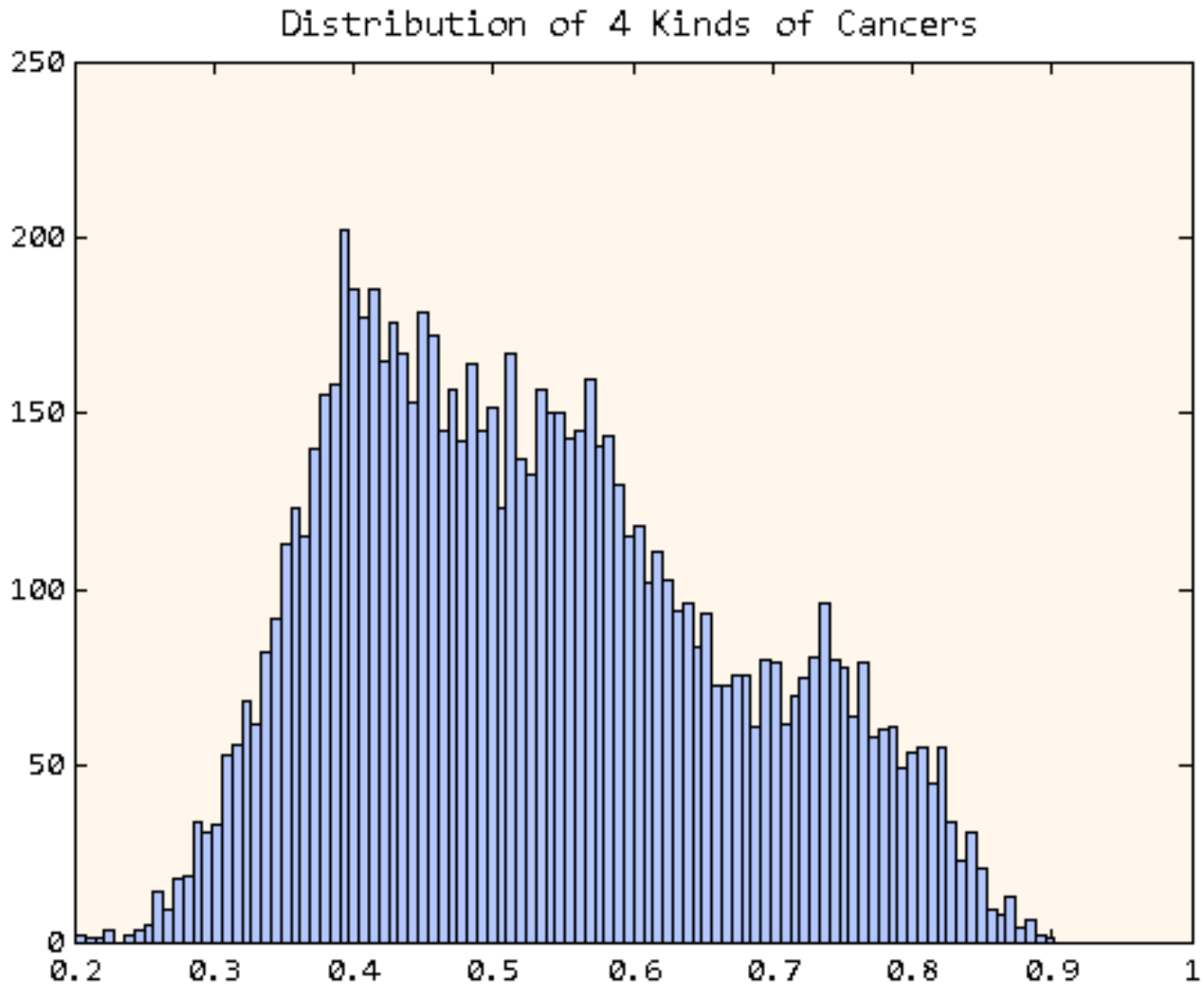
2. Apply spectral clustering algorithm to the similarity graph with weights W to obtain the segmentation of the data.

Application in 4 kinds of Cancers



Correlation Distributions:



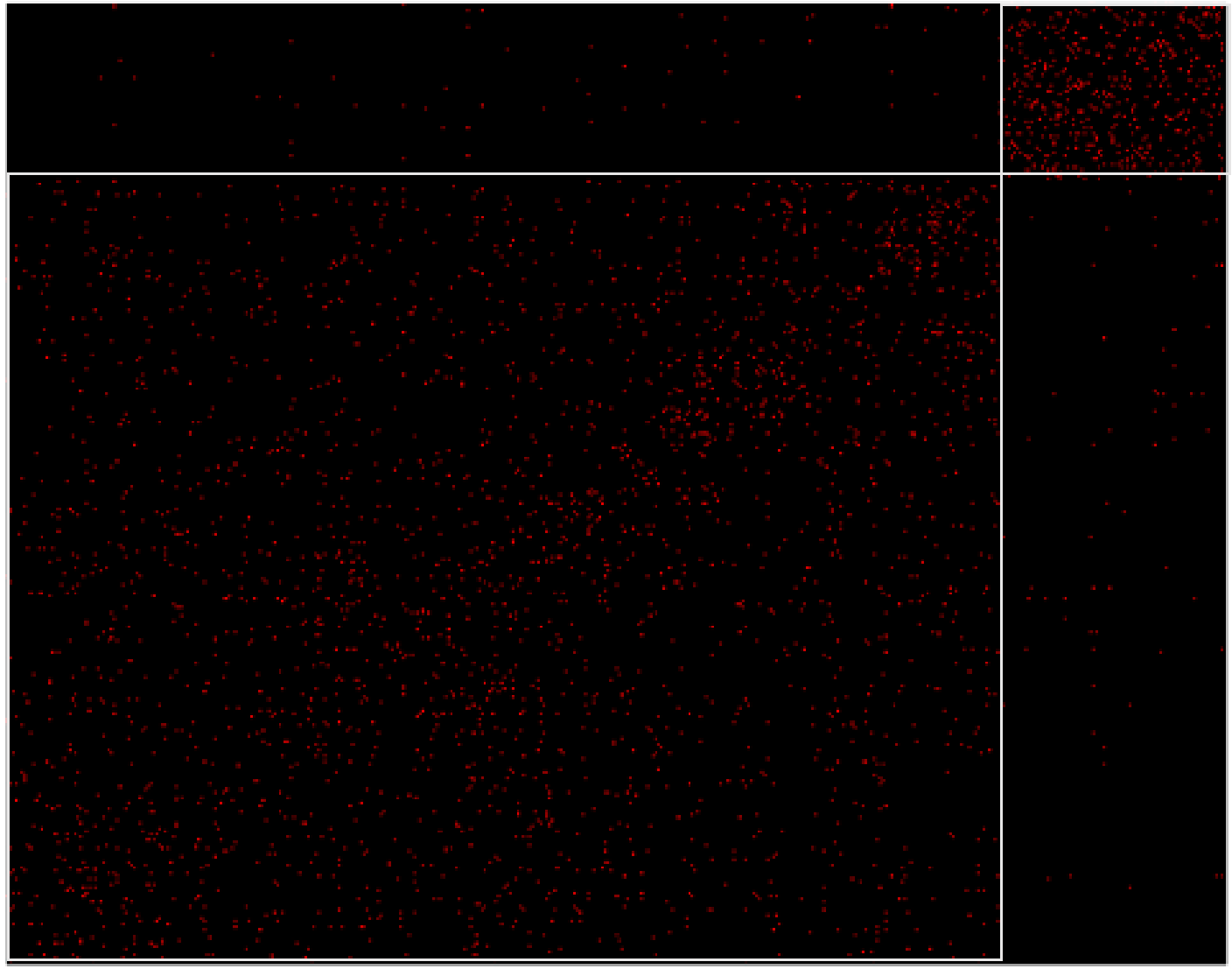

Distribution of 4 Kinds of Cancers
250
200
150
100
50
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Application in ER+/ER- samples

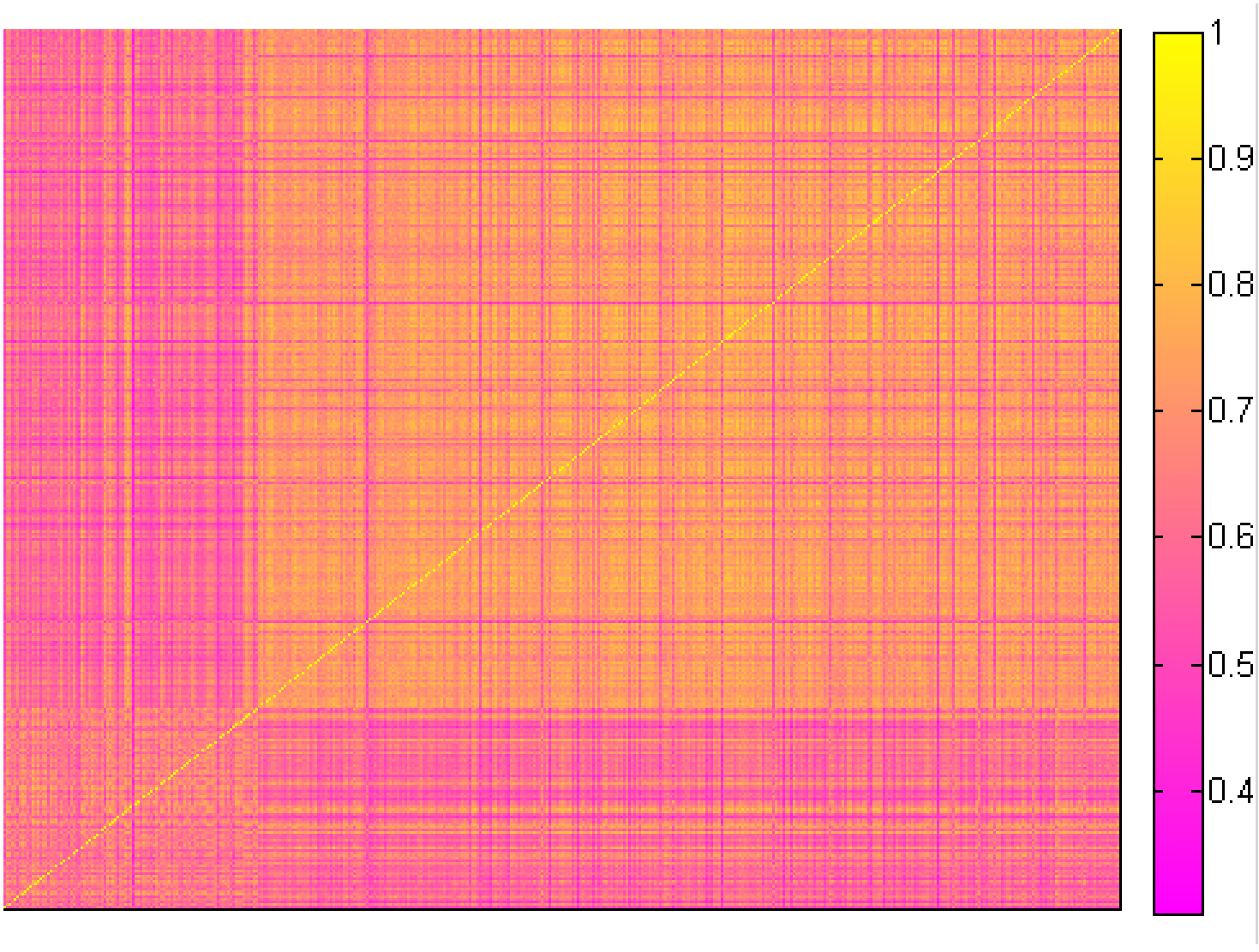
乳腺癌样本数据聚类结果，包含ER+/ER-信息

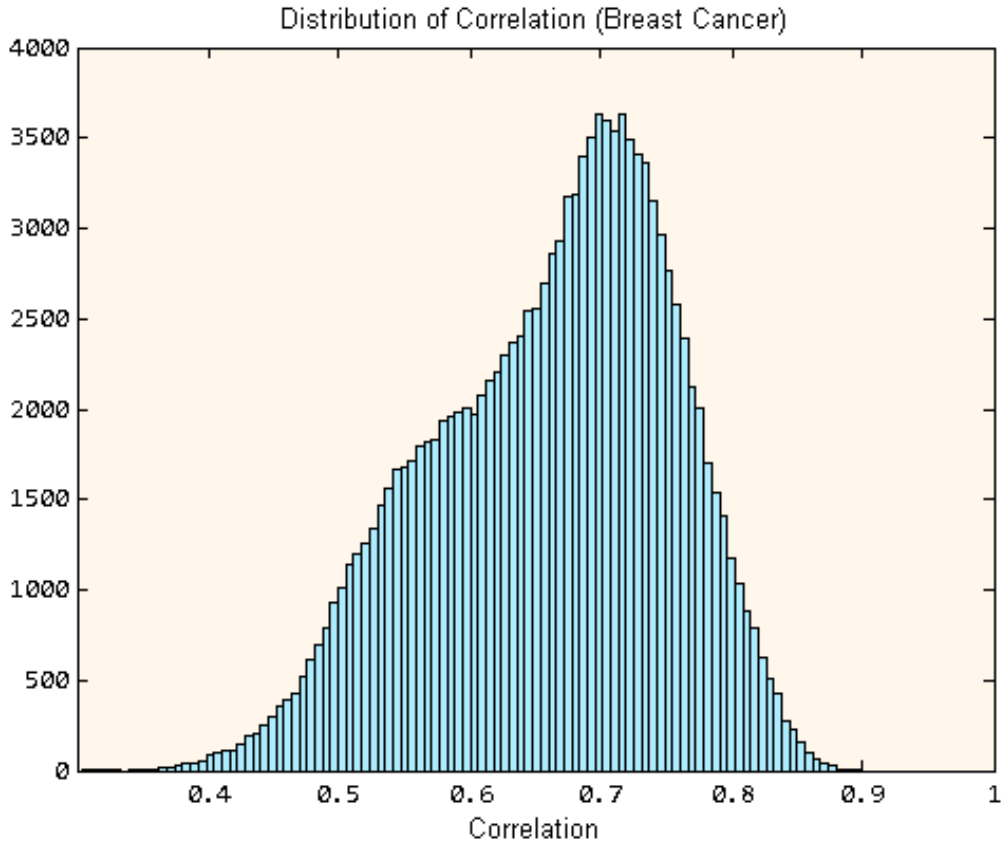
		真实值	
		Positive	Negative
预测值	Positive	370	12
	Negative	24	104

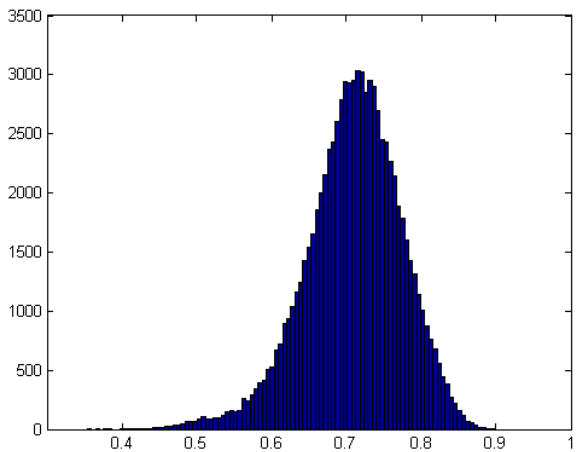
- Elapsed time is 39.406780 seconds

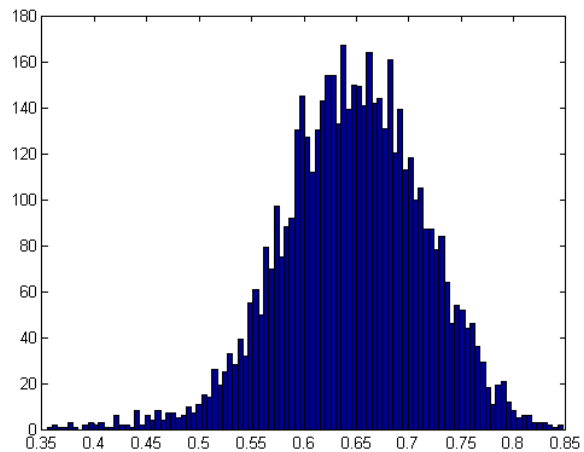


Correlation Heatmap




Distribution of Correlation (Breast Cancer)
4000
3500
3000
2500
2000
1500
1000
500
0
0.4
0.5
0.6
0.7
0.8
0.9
1
Correlation

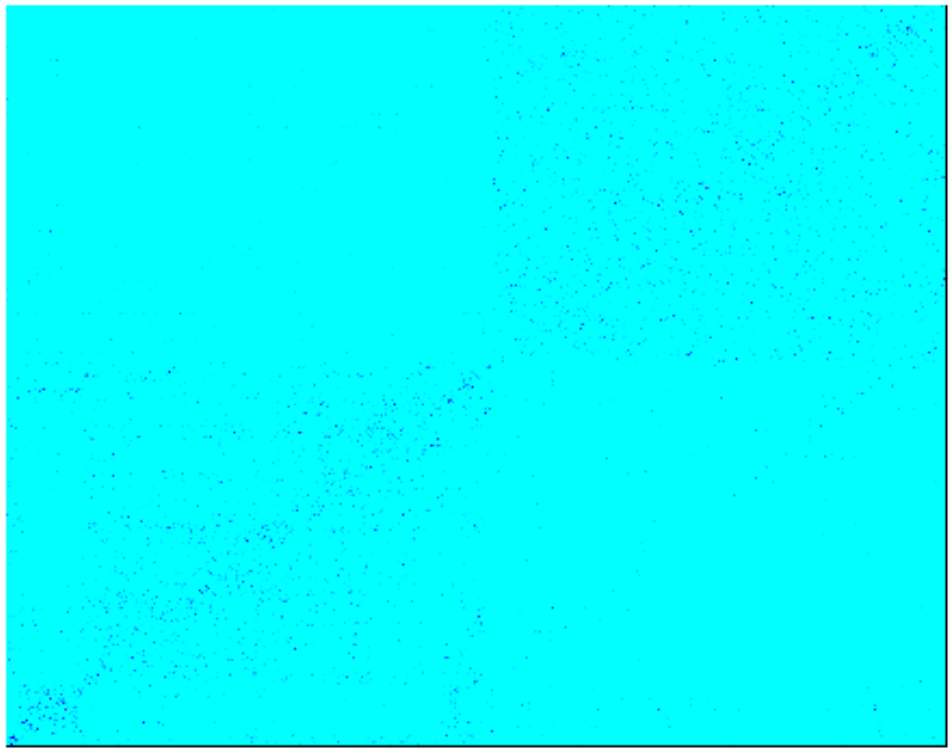

3500
3000
2500
2000
1500
1000
500
0
0.4
0.5
0.6
0.7
0.8
0.9
1


180
160
140
120
100
80
60
40
20
0
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85

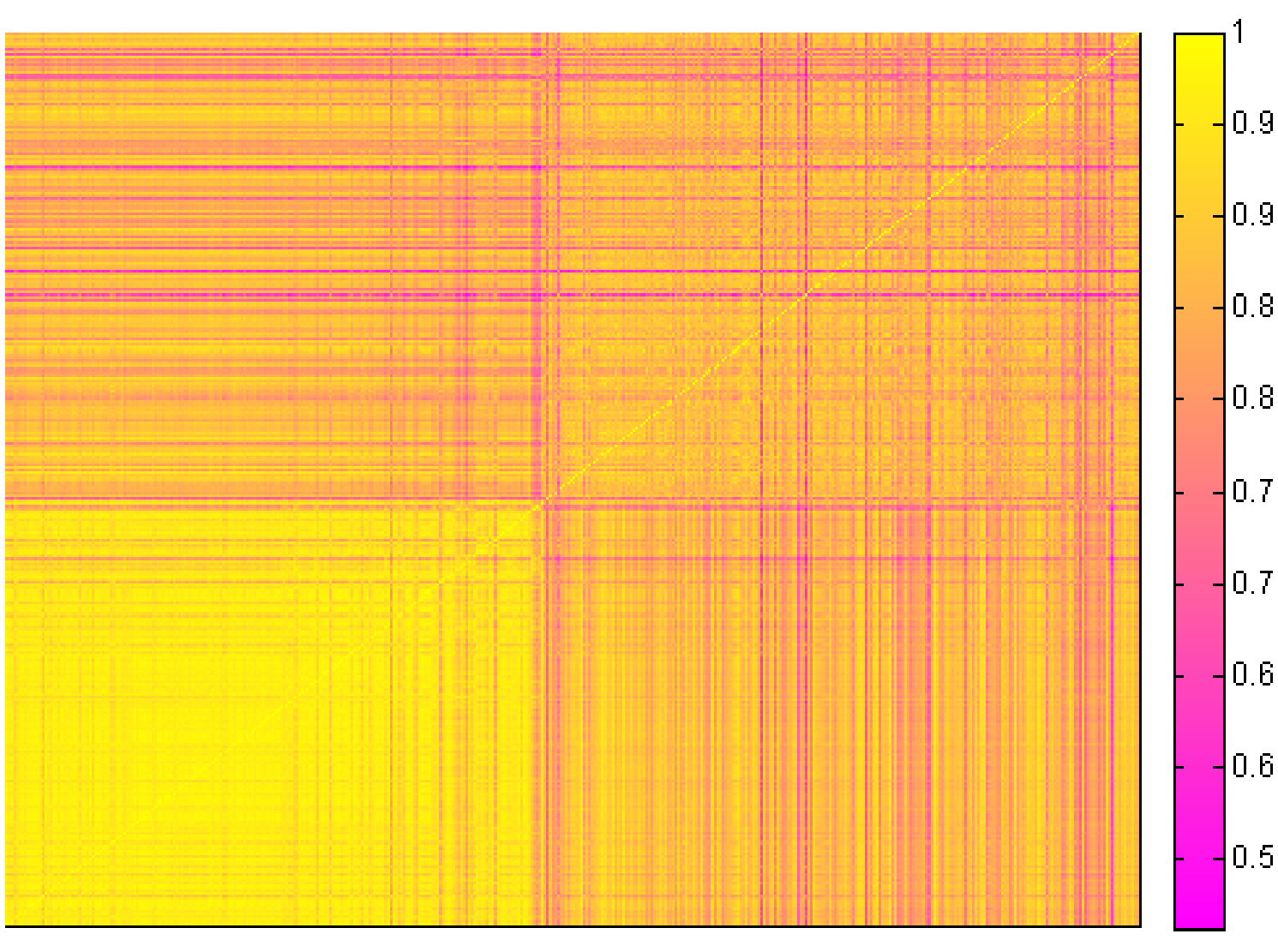
Application in HCC samples

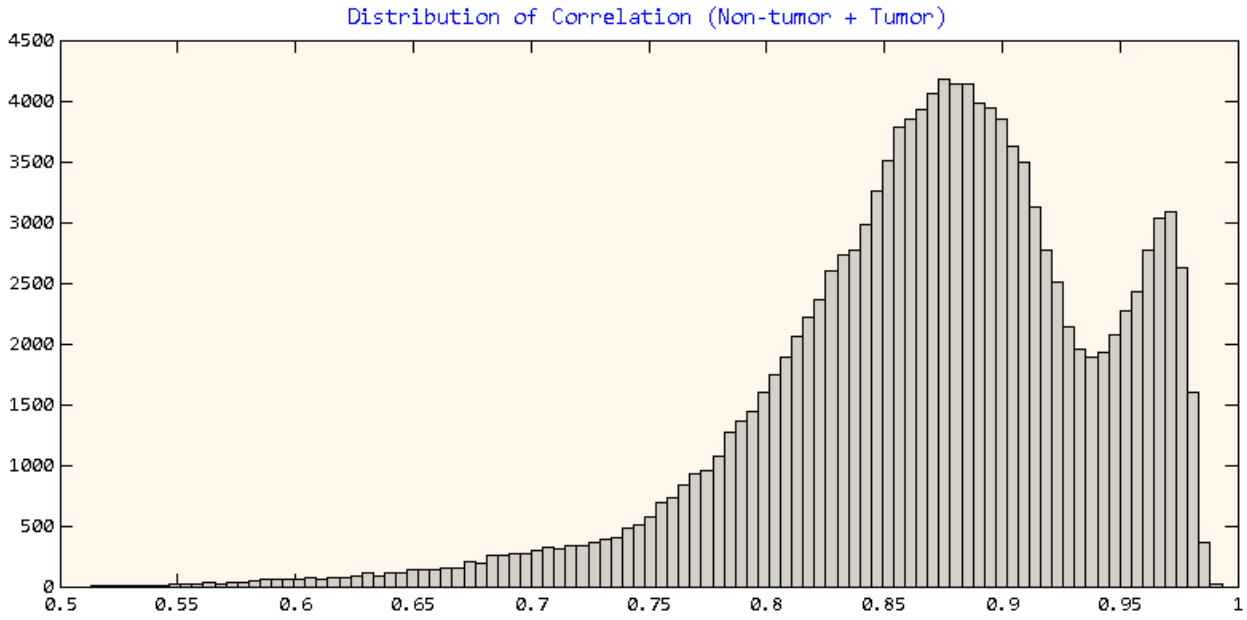
1. GSE25097 Datasets

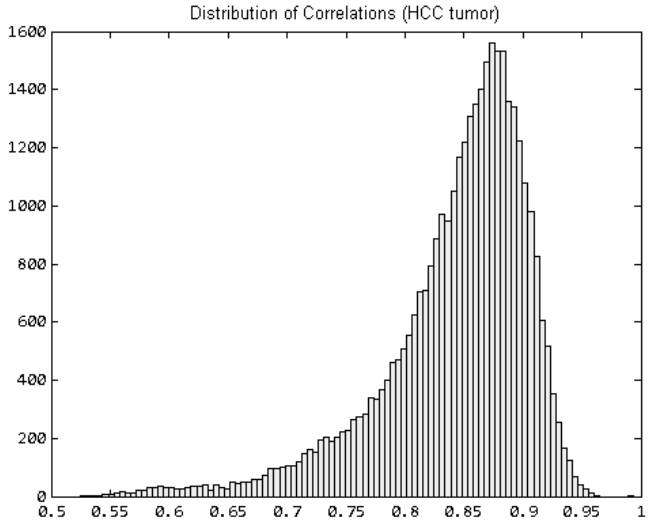
Label	Healthy	Cirrhotic	Non-tumor	Tumor
Number	6	40	243	268

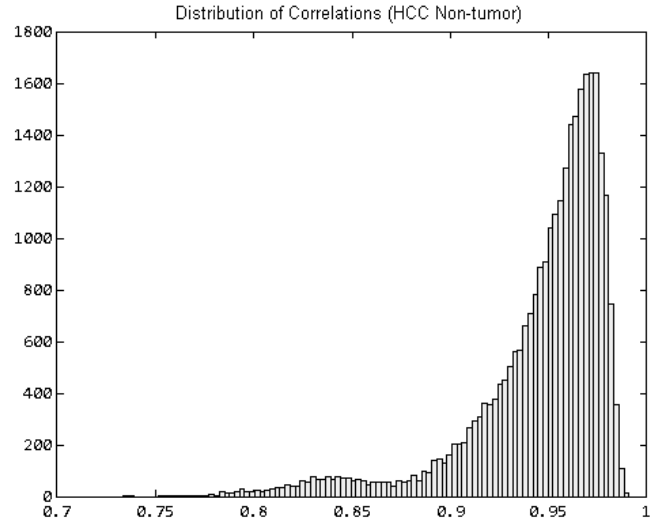


Correlation Heatmap (Non-tumor and tumor)

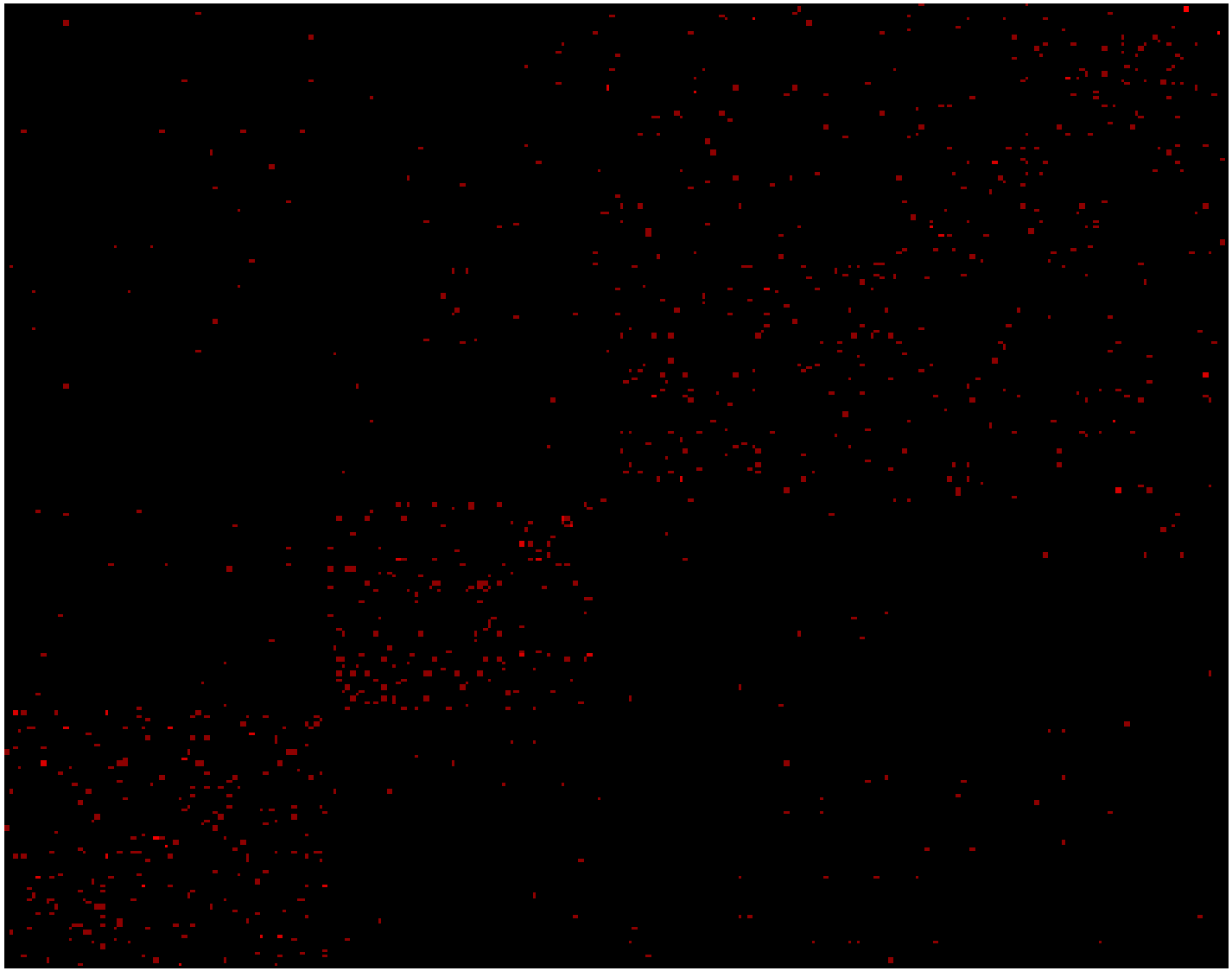


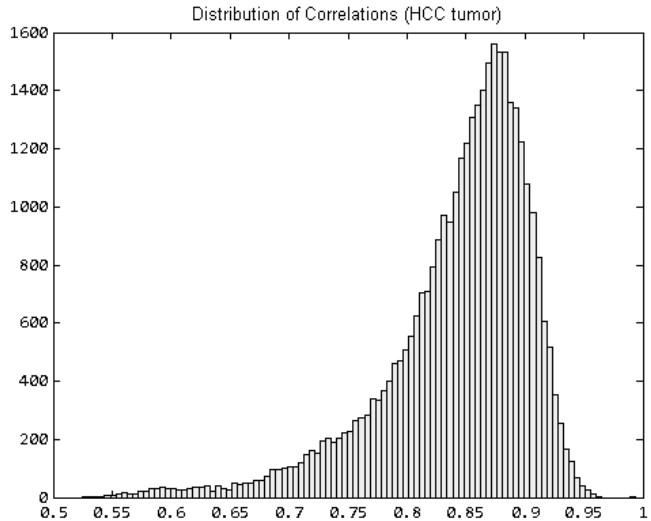

Distribution of Correlation (Non-tumor + Tumor)
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

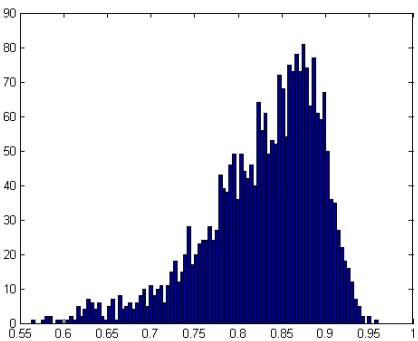

Distribution of Correlations (HCC tumor)
1600
1400
1200
1000
800
600
400
200
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

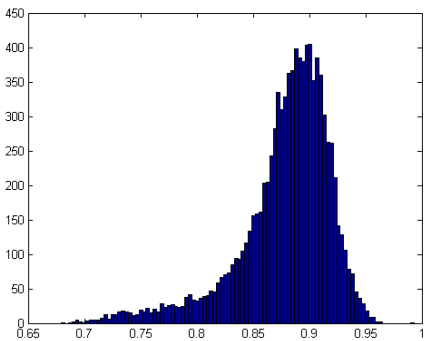

Distribution of Correlations (HCC Non-tumor)
1800
1600
1400
1200
1000
800
600
400
200
0
0.7
0.75
0.8
0.85
0.9
0.95
1

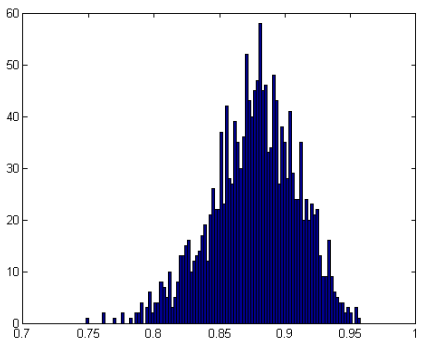
2. Tumor samples only



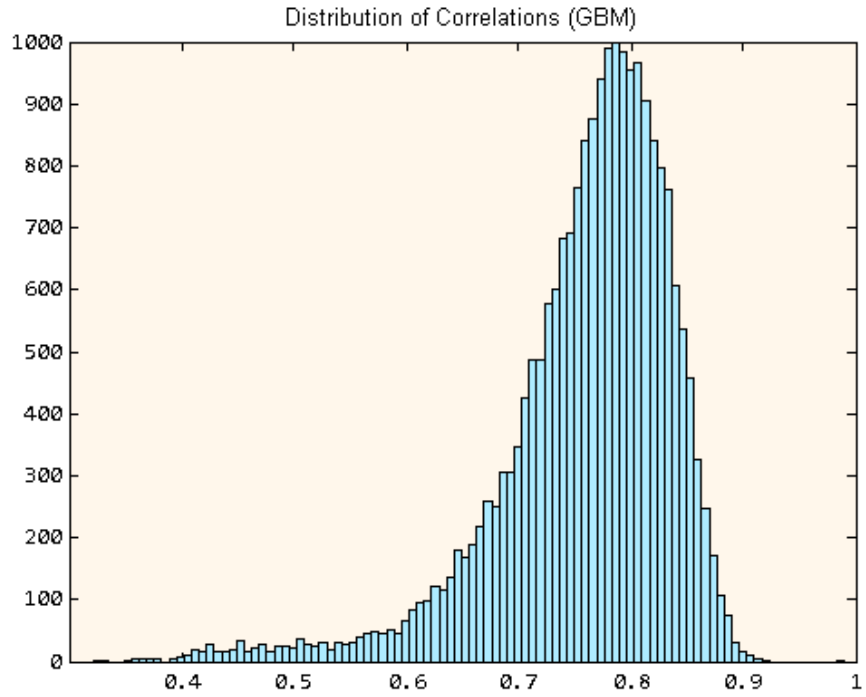

Distribution of Correlations (HCC tumor)
1600
1400
1200
1000
800
600
400
200
0
0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1

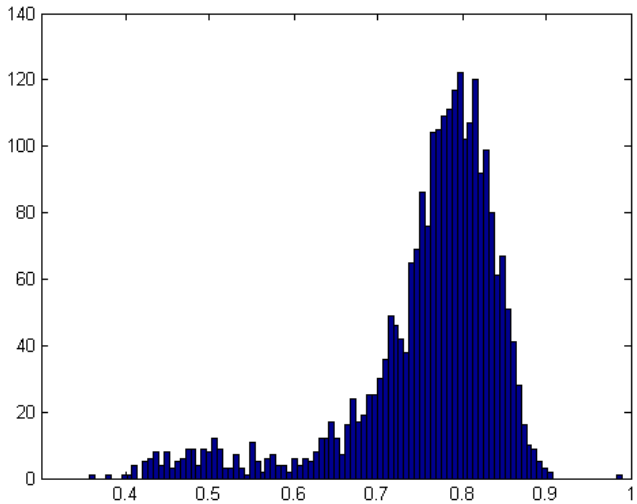

90
80
70
60
50
40
30
20
10
0
0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1

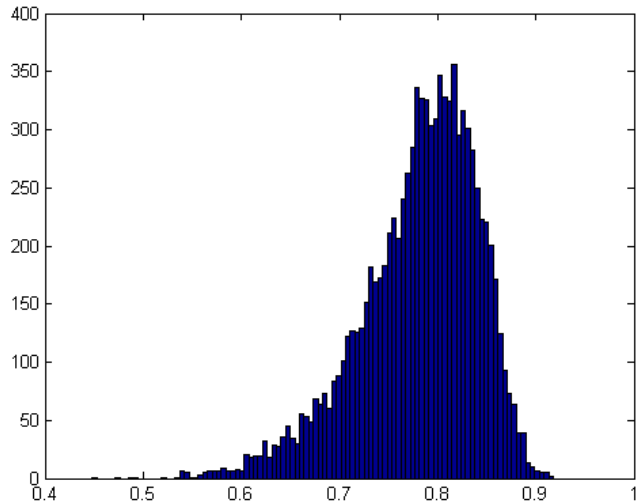

450
400
350
300
250
200
150
100
50
0
0.65 0.7 0.75 0.8 0.85 0.9 0.95 1

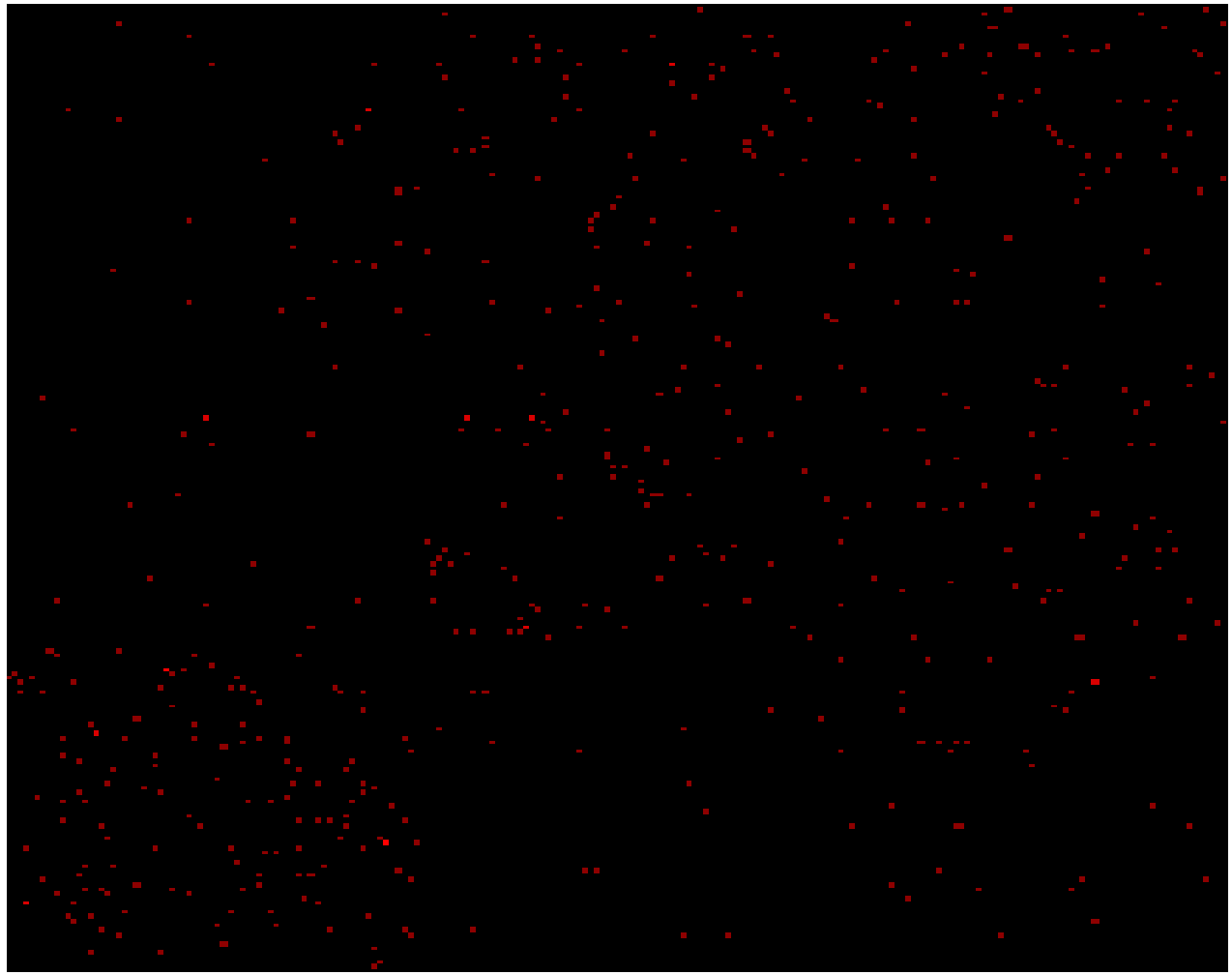

60
50
40
30
20
10
0
0.7 0.75 0.8 0.85 0.9 0.95 1

Application in GBM TCGA Dataset




140
120
100
80
60
40
20
0
0.4
0.5
0.6
0.7
0.8
0.9
1


400
350
300
250
200
150
100
50
0
0.4
0.5
0.6
0.7
0.8
0.9
1



分类结果 / 生存期标签	1	2
1	33	55
2	38	84
	71	139

Partial Least Squares

PCA in a Statistical View:

Suppose a random variable $x \in R^D$ with zero-mean $\mathbb{E}(x) = 0$, and try to find $d < D$ principle components $y \in R^d$:

- y_i's are 'uncorrelated' linear combinations of x:

$$y_i = u_i^T x \in R, u_i \in R^D$$

- The variance of y_i is maximized subject to

$$u_i^T u_i = 1, i = 1, \cdots, d \quad Var(y_1) \geq Var(y_2) \geq \cdots \geq Var(y_d) > 0$$

Principle Components Regression

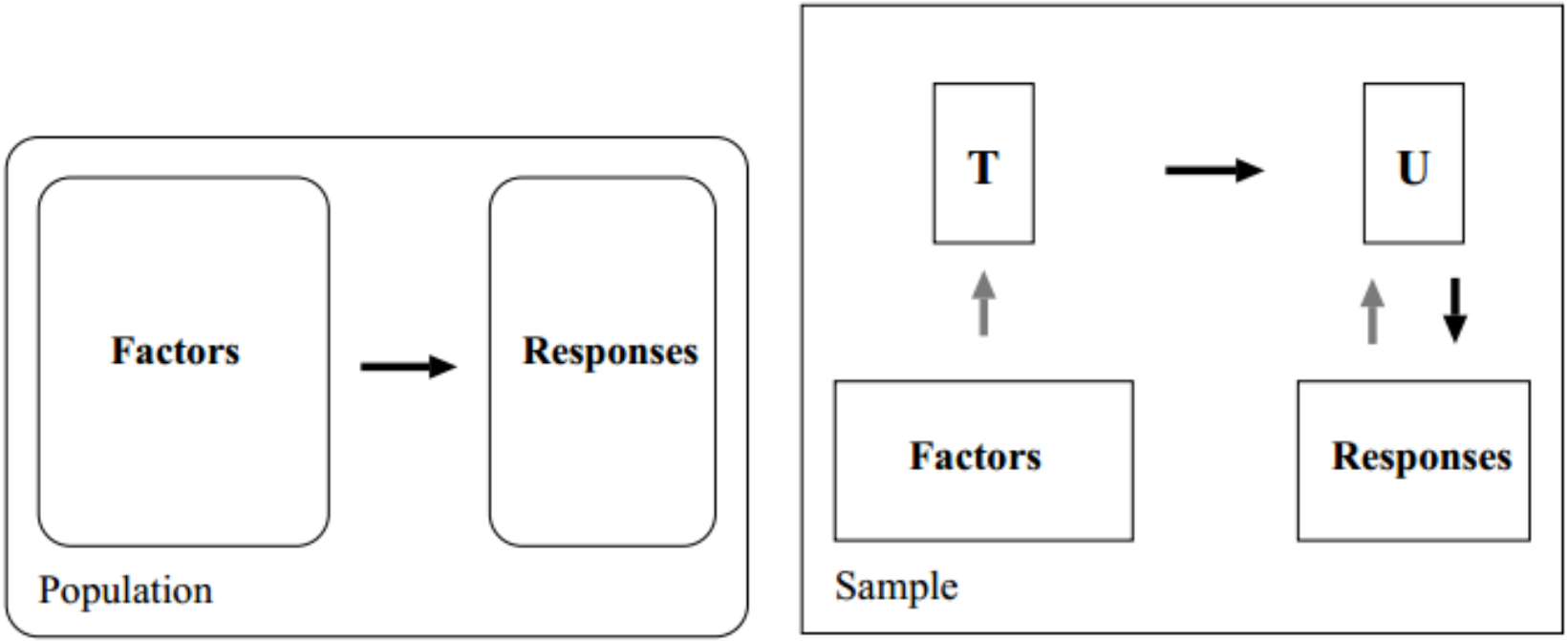
- Form the derived input columns $z_m = Xu_m, m = 1, \cdots, M (M \leq p)$
- Regress the response y on $z_1, \cdots, z_M$

Partial Least Squares

Consider both high variance and high correlation with the response:

The m-th PLS direction ψ_m:

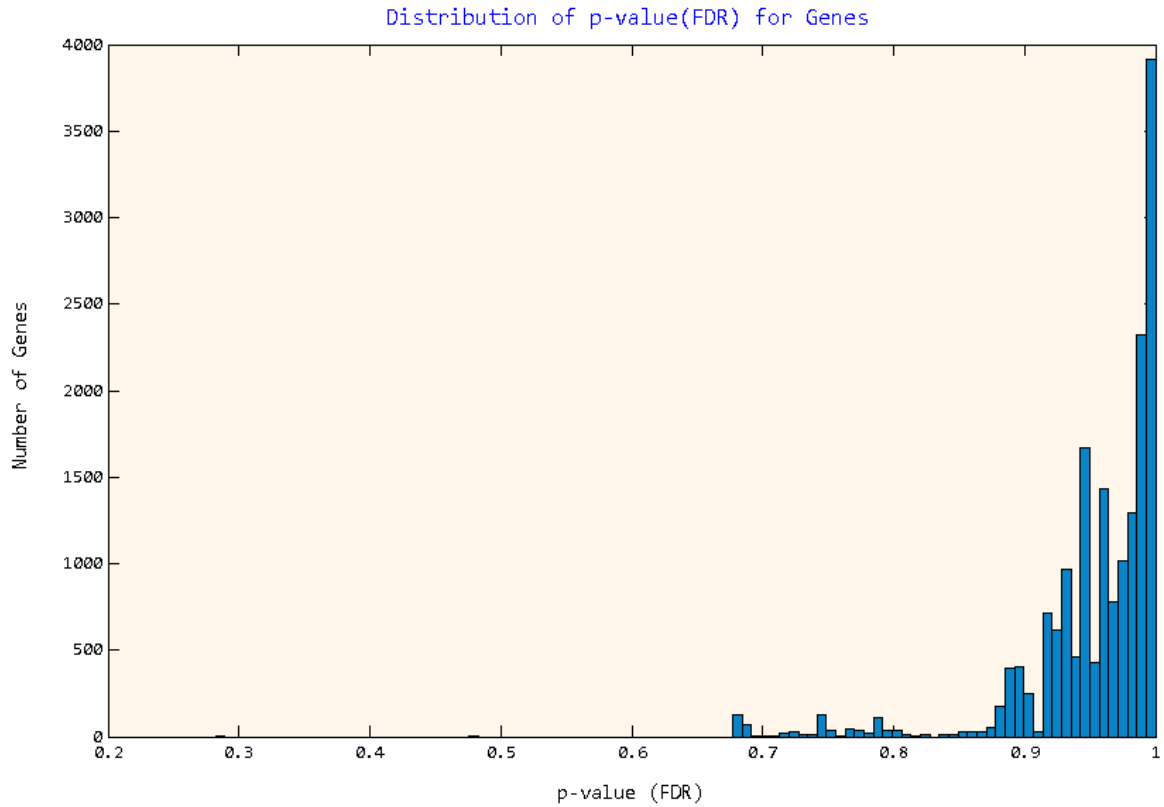
$$\begin{aligned} \max_{\alpha} \quad & \mathsf{Corr}^2(y, X\alpha)\mathsf{Var}(X\alpha) \text{ or } \mathsf{Cov}(y, X\alpha) \\ \text{s.t.} \quad & ||\alpha|| = 1, \alpha^T \Sigma_N \hat{\psi}_l = 0, l = 1, \cdots, m-1 \end{aligned}$$

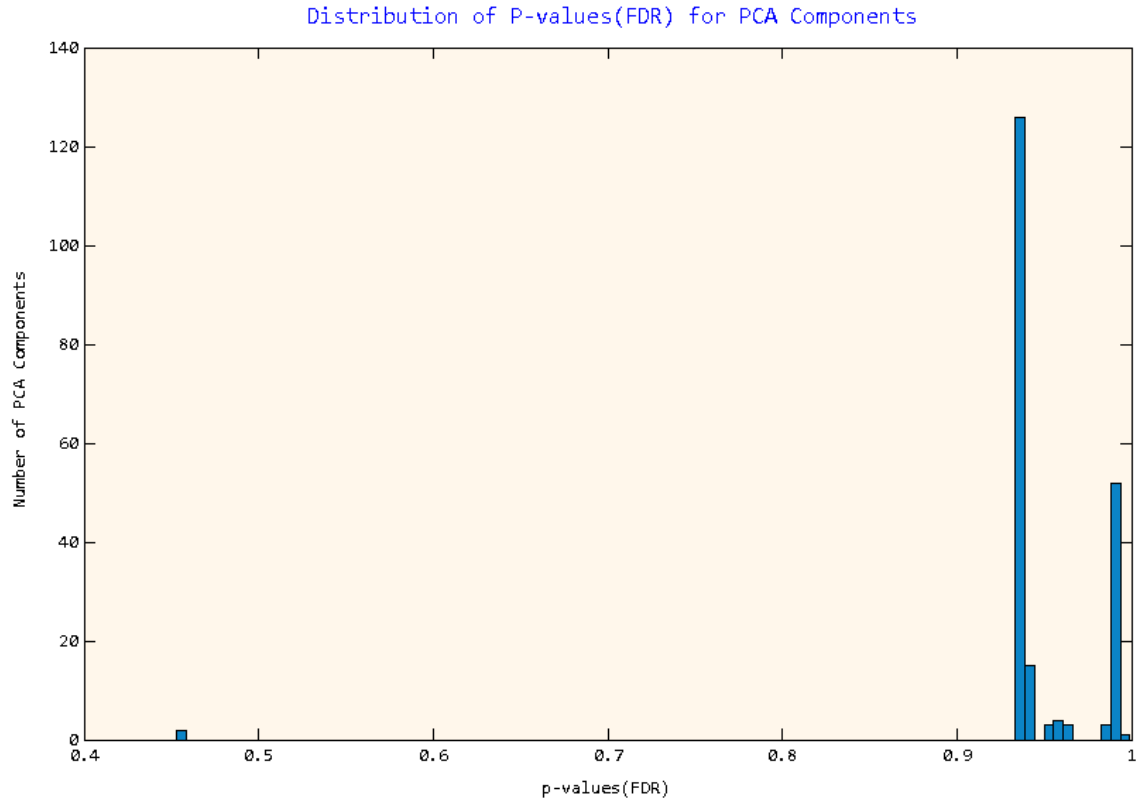


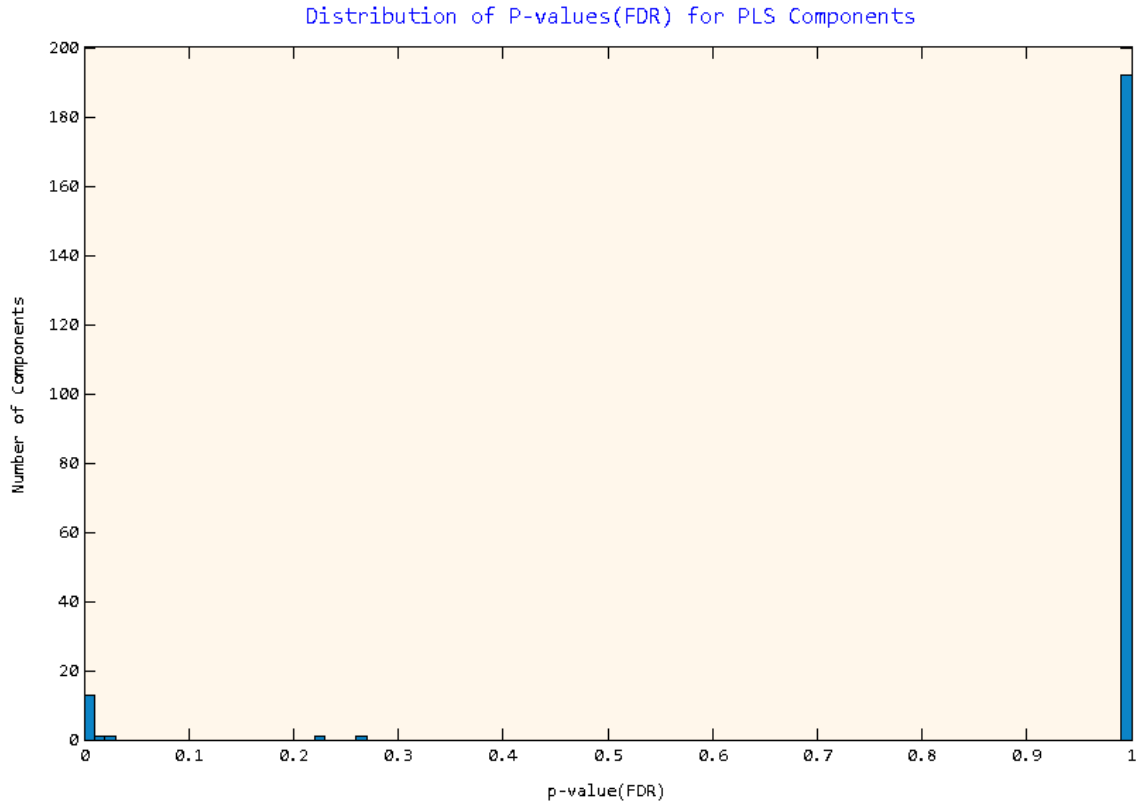
$$
\begin{aligned}
X &= TP^T + E, \\
Y &= UQ^T + F, \\
U &= TB
\end{aligned}
$$

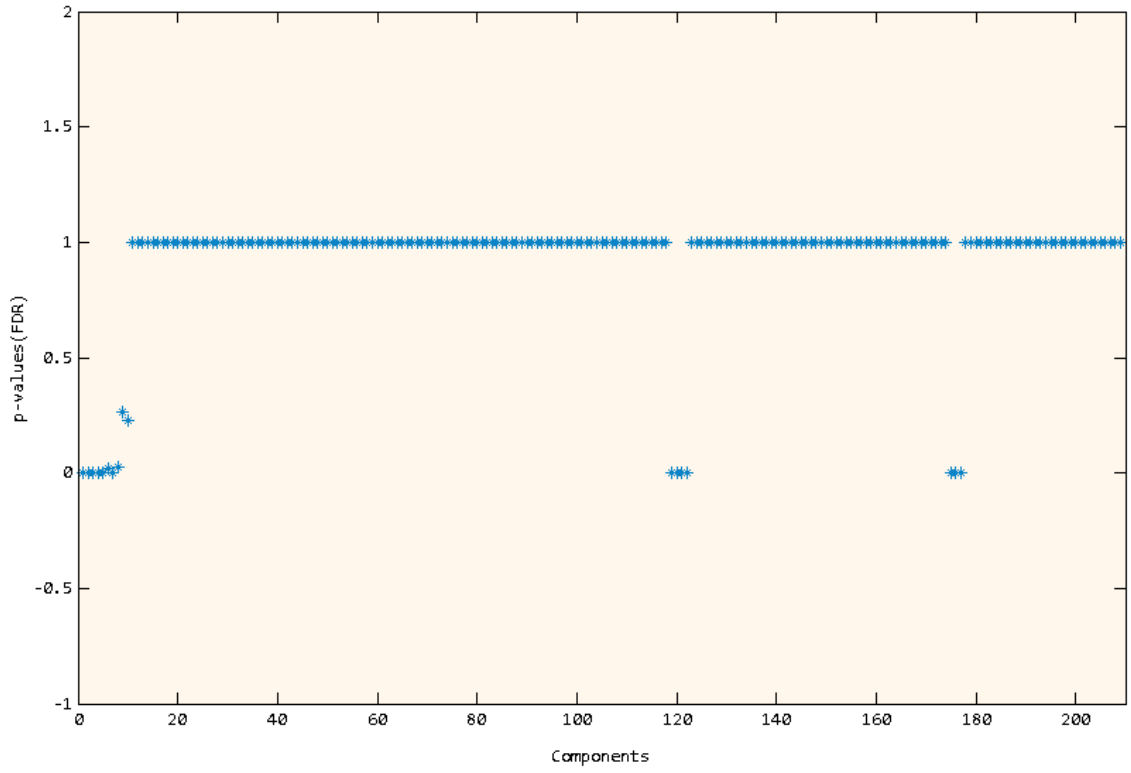
GBM mRNA Data

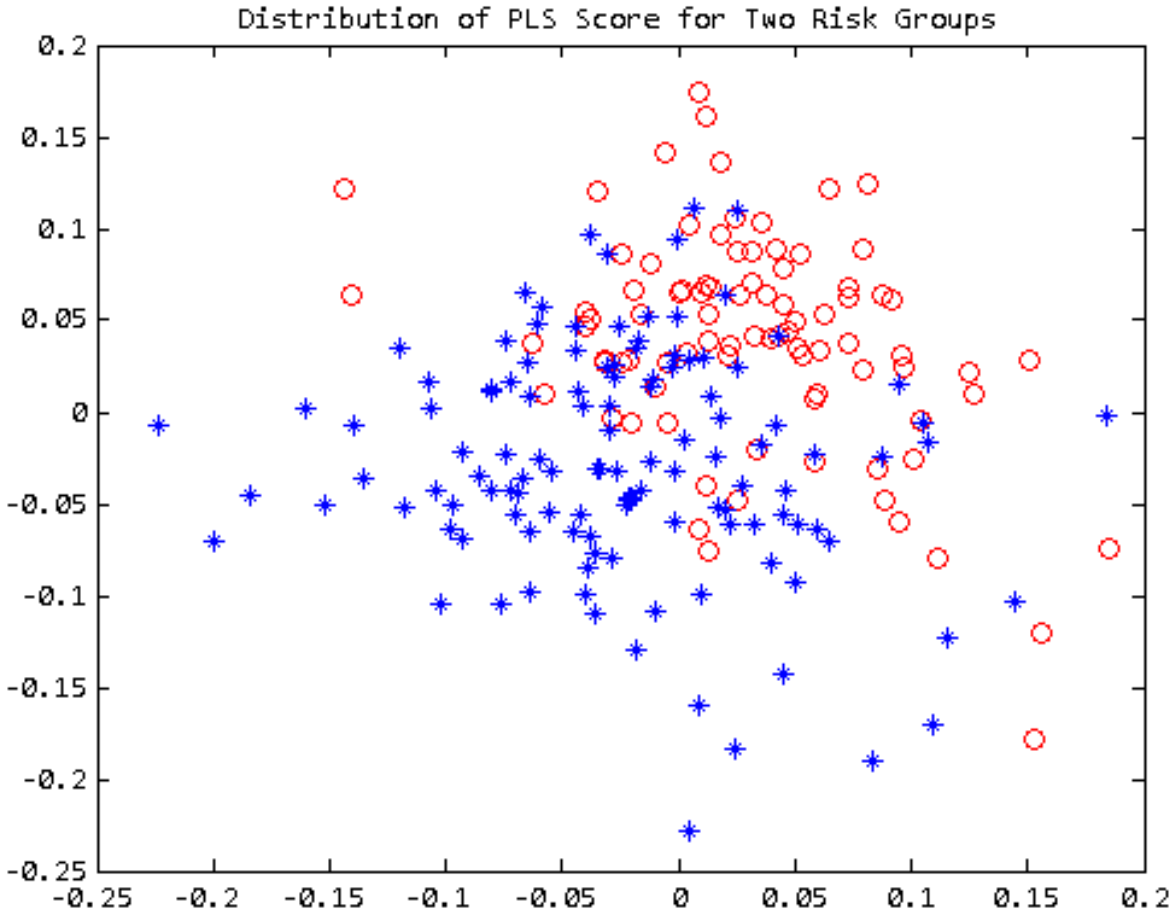
1. Wilcoxon rank sum testing (Survivial $\sim$ Gene,PCA,PLS components)

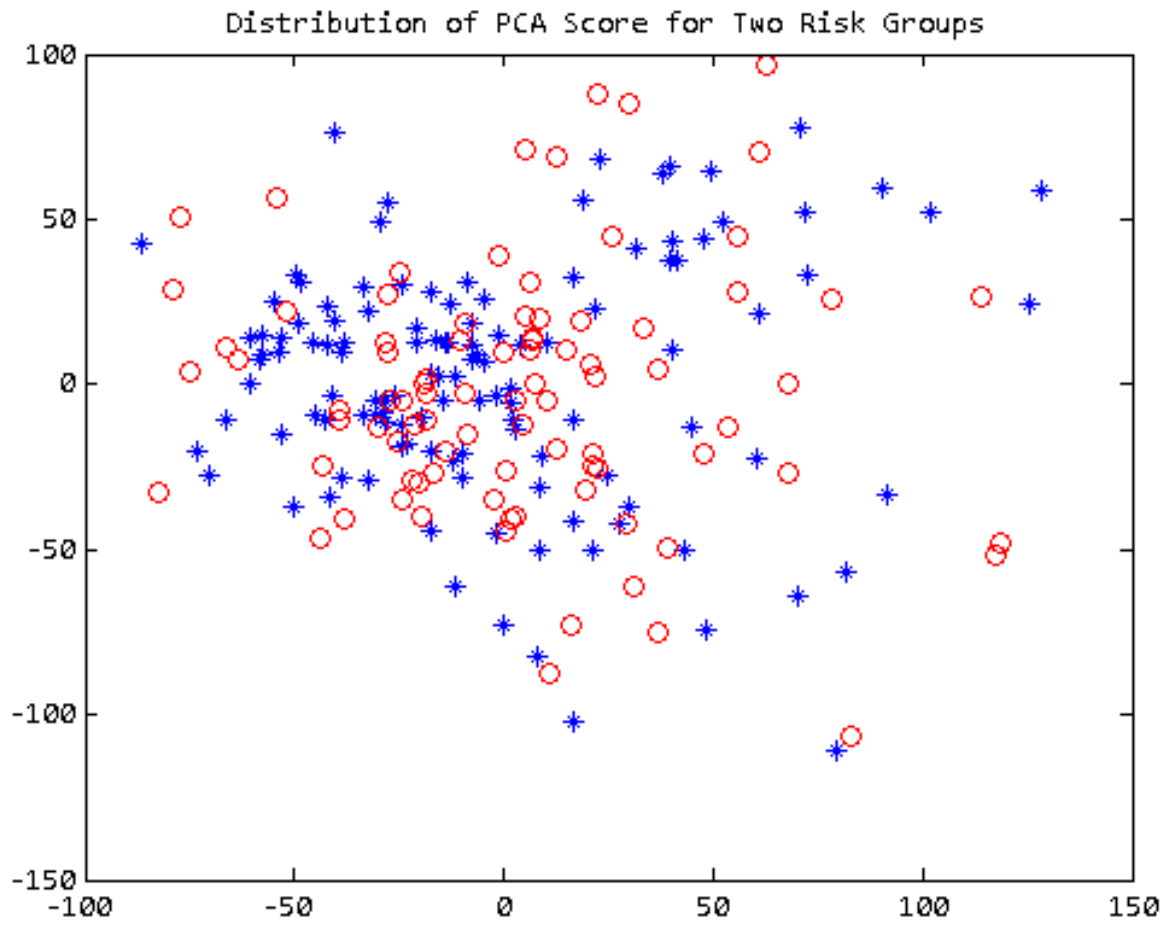



Distribution of P-values(FDR) for PCA Components
Number of PCA Components
p-values(FDR)
0
20
40
60
80
100
120
140
0.4
0.5
0.6
0.7
0.8
0.9
1

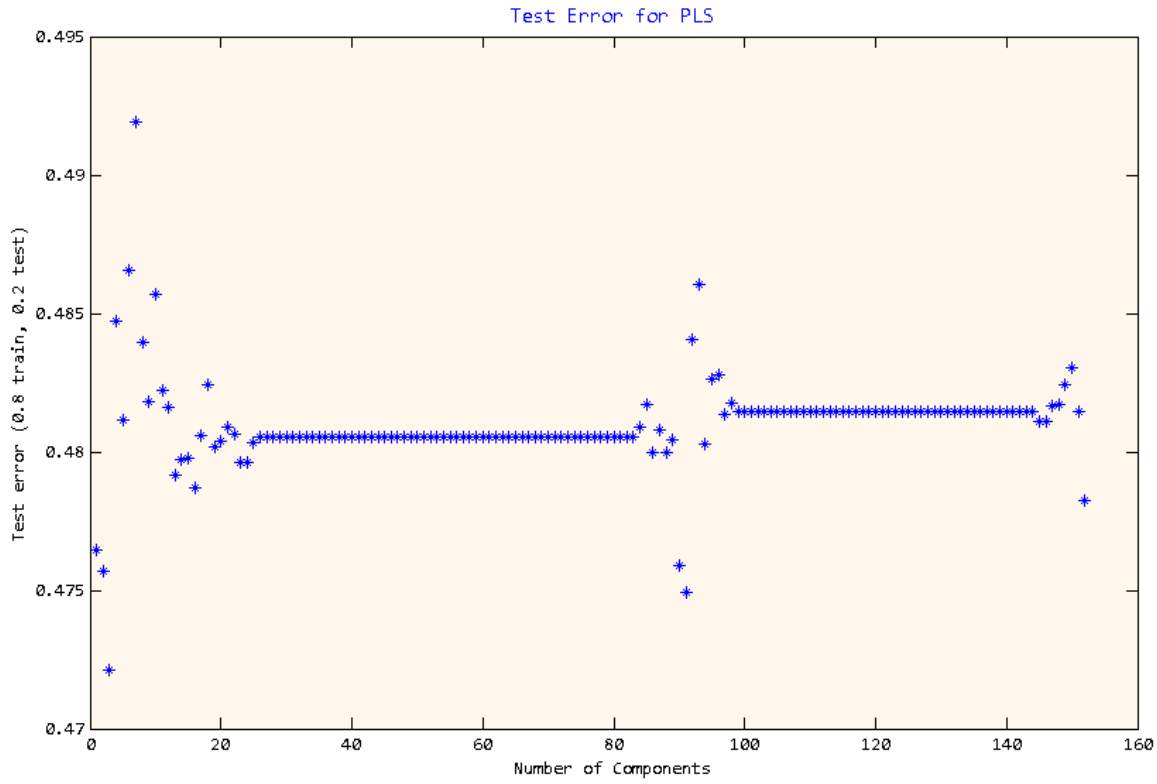

Distribution of P-values(FDR) for PLS Components
Number of Components
p-value(FDR)
0
20
40
60
80
100
120
140
160
180
200
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


2
1.5
1
0.5
0
-0.5
-1
p-values(FDR)
0
20
40
60
80
100
120
140
160
180
200
Components


Distribution of PLS Score for Two Risk Groups
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2


Distribution of PCA Score for Two Risk Groups
100
50
0
-50
-100
-150
-100
-50
0
50
100
150

2. Classification error (overfitting)

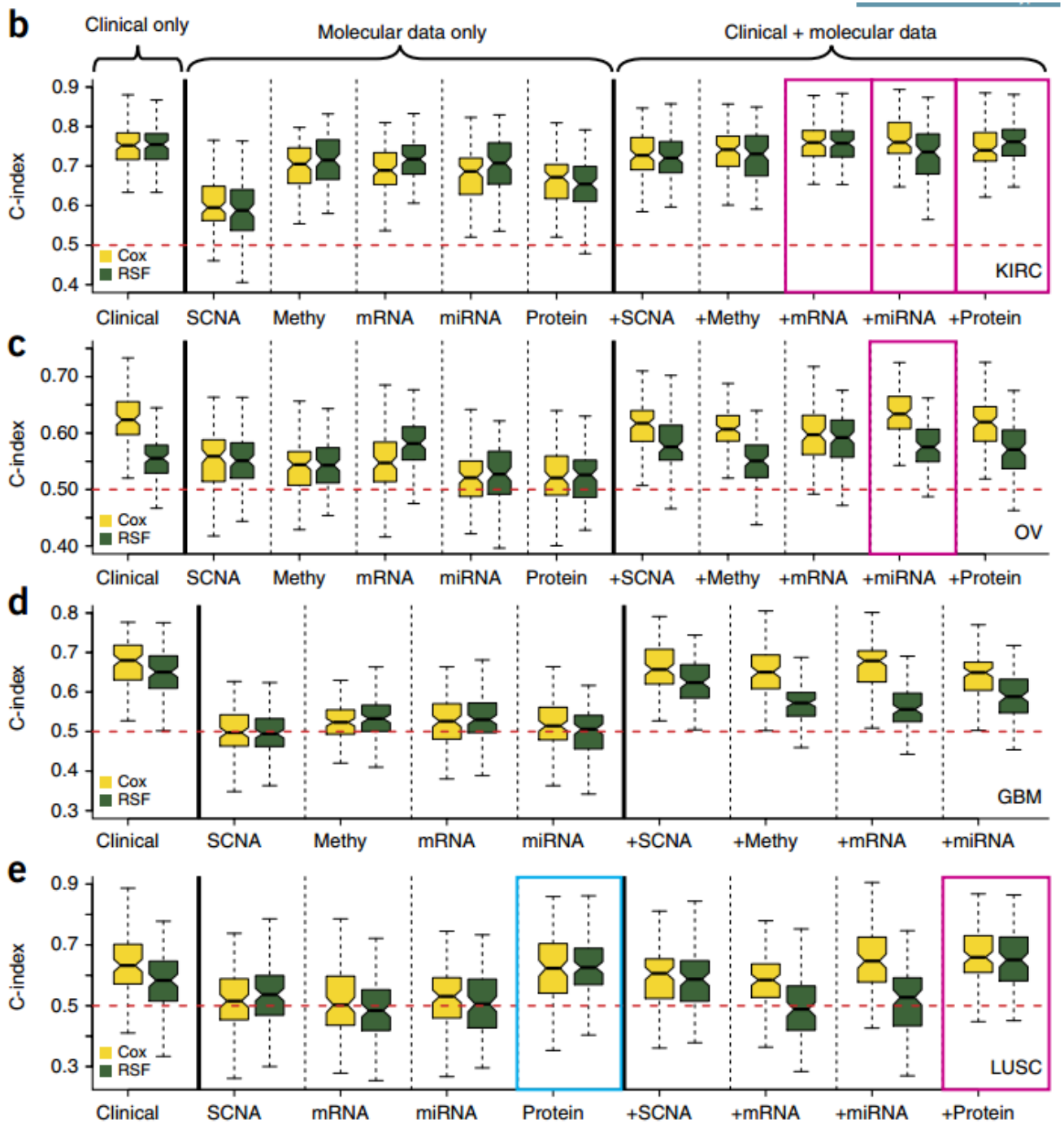


(Mean error of 100 times)

Results of TCGA Survival Data Analysis

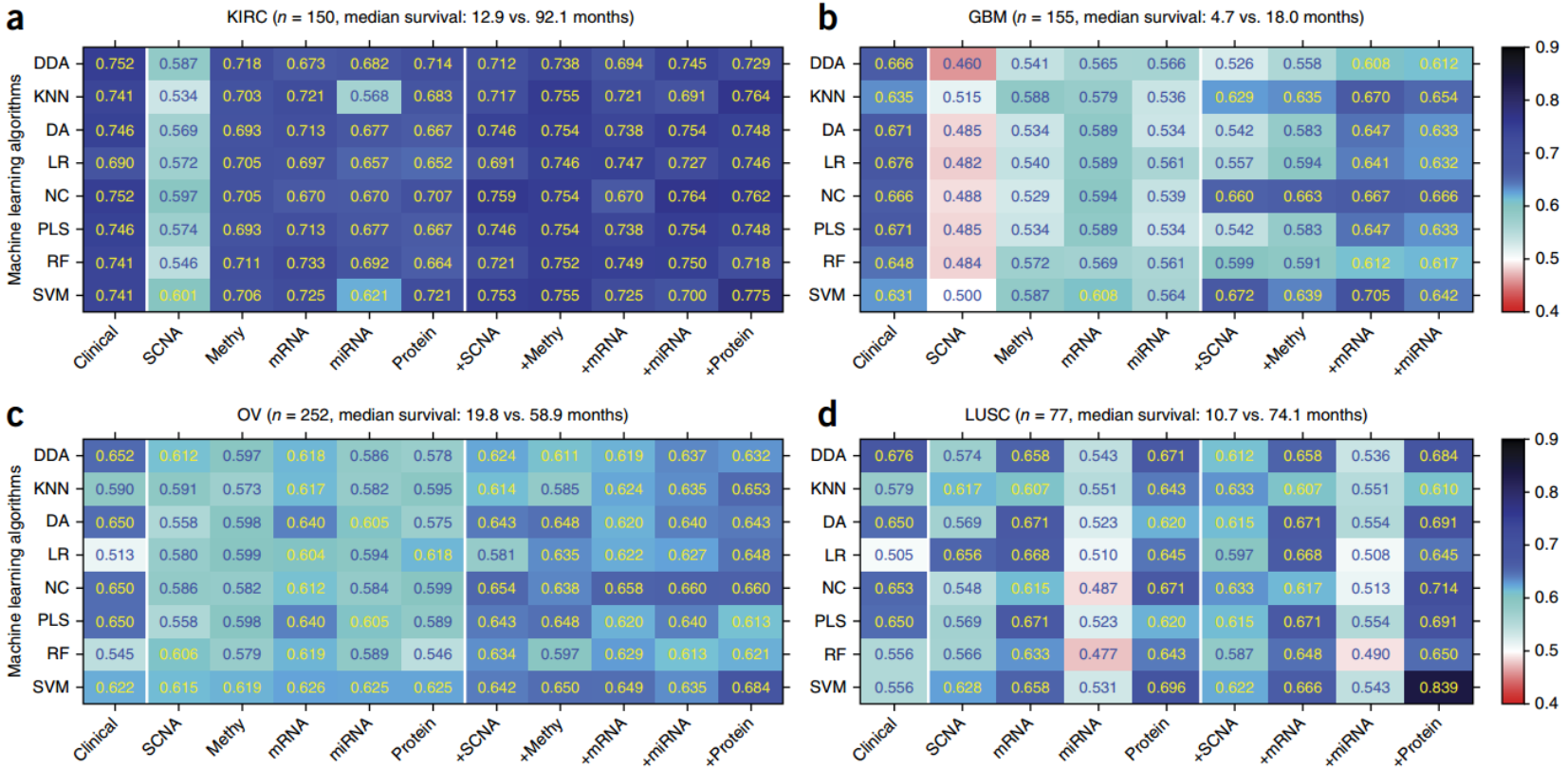
how and to what extent TCGA molecular data can affect oncology practice.

Assessing the clinical utility of cancer genomic and proteomic data across tumor types


b
Clinical only
Molecular data only
Clinical + molecular data
C-index
Cox
RSF
KIRC
Clinical
SCNA
Methy
mRNA
miRNA
Protein
+SCNA
+Methy
+mRNA
+miRNA
+Protein
c
OV
d
GBM
e
LUSC

- Clinical-variables-only model: show substaintial predictive power
- The relative predictive power of individual molecular data sets strongly depended on the cancer type
- Clinical variables + molecular data: significantly improve predictive power but the increase is limited.

Prediction of Dichotomized Survival data



Predictive power of molecular data strongly depends on the cancer type.