

Cluster Validation using the Relative Data Depth

Rebecka Jörnsten, Yehuda Vardi, Cun-Hui Zhang
Department of Statistics, Rutgers University

<http://www.stat.rutgers.edu/~rebecka>

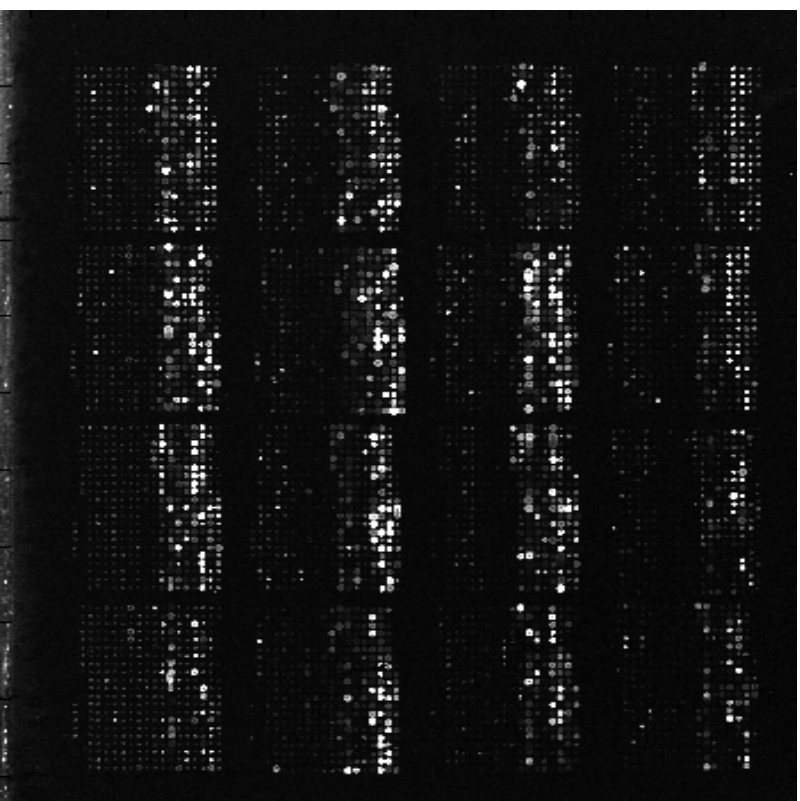
Outline

1. **Introduction**
2. **Cluster Validation**
3. **The Relative Data Depth**
4. **Results**
5. **Conclusion**

Microarray Images

Powerful tool for monitoring the expression of thousands of genes simultaneously:

- Identifying gene functions
- dynamics
- pathological context



Extracting information from the images

1. **Registration:** Where are the spots?
2. **Segmentation:** Where did hybridization occur?
3. **Background correction:** Non-specific hybridization.
4. **Data summary:** Log of background corrected mean intensity ratios.
5. **Normalization:** Remove systematic effects (spatial, between-slides).

Gene Expression Data

The extracted data from a set of arrays/slides are summarized by

a $n \times p$ data matrix.

n samples (e.g. type of cancer), p genes, $p \sim 1000 \gg n \sim 100$.

Important tasks:

1. Sample clustering -
 - validate pathological classification
 - discover new classes of cancer
2. Gene clustering -
 - indicative of genetic pathways
 - for improved sample clustering
 - dimension reduction

Cluster Validation

- Selecting the number of clusters in a data set
- Stability/robustness of generated clusters
- Identifying outliers

Most criteria are based on within and between sum of squares.

Examples: Silhouette width, Gap-statistic.

This can be problematic when the data is noisy, and many unrelated features are included.

The Silhouette width

Often used in conjunction with PAM (partitioning around mediods).

sil - Silhouette Width:

1. For each observation i , compute a_i , the average distance to all other members of the cluster i belongs to.
2. Compute b_i , the average distance to members of the nearest competing cluster.
3. $sil_i = \frac{b_i - a_i}{\max(b_i, a_i)}$. Select the number of clusters K that maximizes the average silhouette width $\sum_i sil_i / n$.

The silhouette width also functions as a visualization and outlier identification tool. Look out for negative sil_i .

Gap - The Gap statistic:

1. Cluster the data, with number of clusters $k = 1, \dots, MAX$. Compute the within-cluster sum of squares W_k .
2. Generate B data sets under the null (no clusters), and compute $W_{k,b}^*$, $b = 1, \dots, B$, $k = 1, \dots, MAX$. Compute the Gap statistic

$$Gap(k) = \frac{1}{B} \sum_b \log(W_{k,b}^*) - \log(W_k)$$

3. Compute the standard deviation sd_k of $\log(W_{k,b}^*)$, and define as $s_k = sd_k \sqrt{1 + 1/B}$. Choose K as

$$K = \min k \text{ s.t. } Gap(k) \geq Gap(k + 1) - s_{k+1}$$

We generate null data sets by

- a) for each variable (gene) sample from a uniform distribution over the observed range, or
- b) transform the data before sampling to preserve geometry (GapPC).

k-median Clustering

More robust than k-means. Important when data is noisy, includes unrelated features.

PAM is an *approximation*. The cluster representatives are restricted to belong to the set of observations (medioids).

We present a fast exact k-median algorithm.

Given k , k-median Algorithm:

1. Modified Weiszfeld algorithm gives fast convergence to the k cluster medians, given a partition.
2. A Nearest Neighbor criterion is used to generate a partition, given the k medians.
3. We iterate between 1 and 2.
4. To avoid local minima, we apply a standard simulating annealing approach in step 2.

L1 Data Depth

Vardi and Zhang introduced a new concept of data depth; the L1 data depth generated by the multivariate median.

- For point z , and observation x_i ,

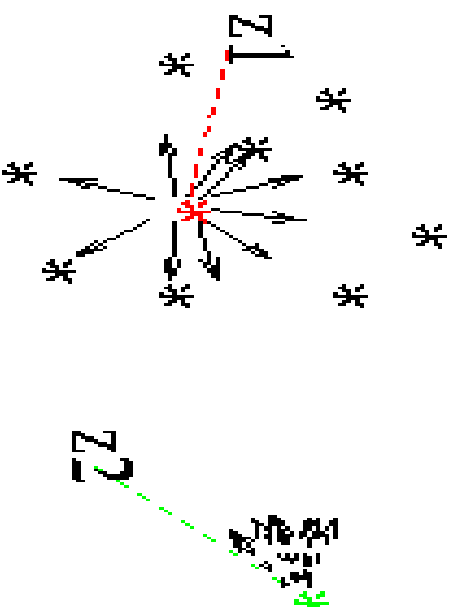
let $e_i(z) = x_i - z / \|x_i - z\|$, i.e. a unit vector pointing in the direction from z to x_i .

- Compute

$$\bar{e}(z) = \sum_{i=1}^m e_i(z) / m.$$

- Define the L1 data depth as

$D(z) = 1 - \max(0, \|\bar{e}(z)\| - f(z))$, where $f(z) = 1/m$ is $z = x_i$ some i , and 0 otherwise.



$$\|e(z_1)\| \sim 0, \|e(z_2)\| \sim 1$$

$1 - D(z)$ is the minimum additional probability mass needed at z to make *it* the multivariate median of the m observations.

The Relative Data Depth, ReD

Given k , use the k -median algorithm to generate a clustering of the n observations.

Given a k clustering, for each observation i there is a

- *within-cluster data depth* D_i^w , and
- $k - 1$ *between-cluster data depths* $D_i^b(l)$, $l = 1, \dots, k - 1$, where 1 corresponds to the order: nearest competing cluster, second nearest, etc.

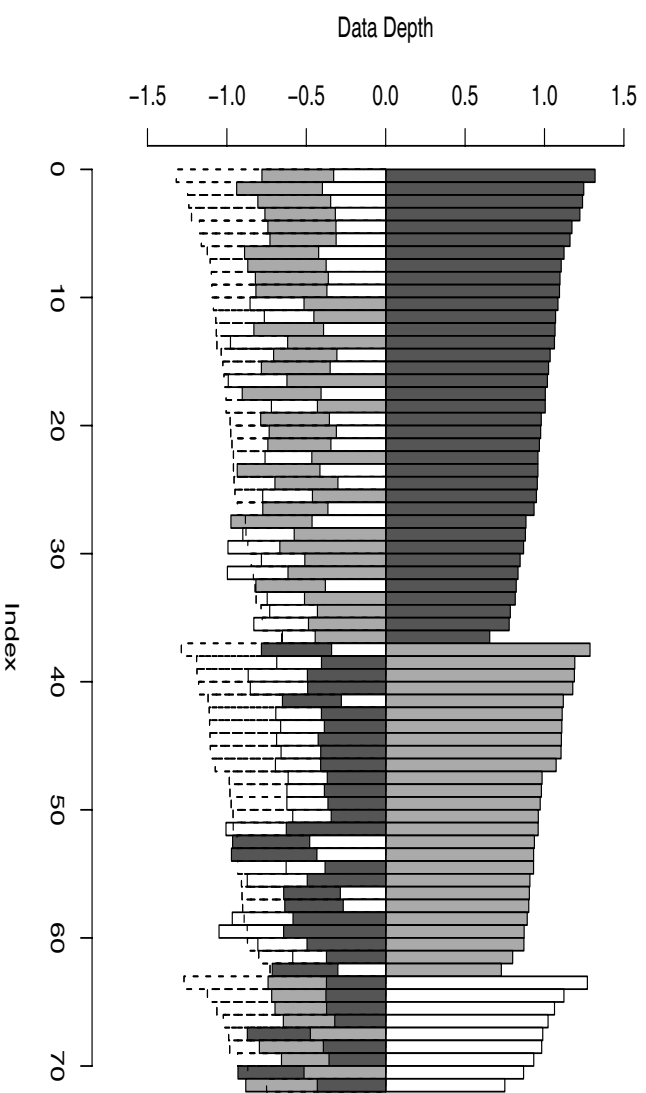
- The *Relative Data Depth* ReD_i is given by $D_i^w - \tilde{D}_i^b$

Basic: Use $\tilde{D}_i^b = D_i^b(1)$ - uses only the nearest competing cluster.

Alternative: Use $\tilde{D}_i^b = \sum_{l <= q} D_i^b(l)$ for some i , $\tilde{D}_i^b = 0$ for other i .

Idea is to put more weight on the most informative observations.

Visualization



Cluster Validation: Look for color patters in the lower panel.

Abrupt drops in the within-cluster depths.

Observations that are “deep” wrt many clusters.

Results

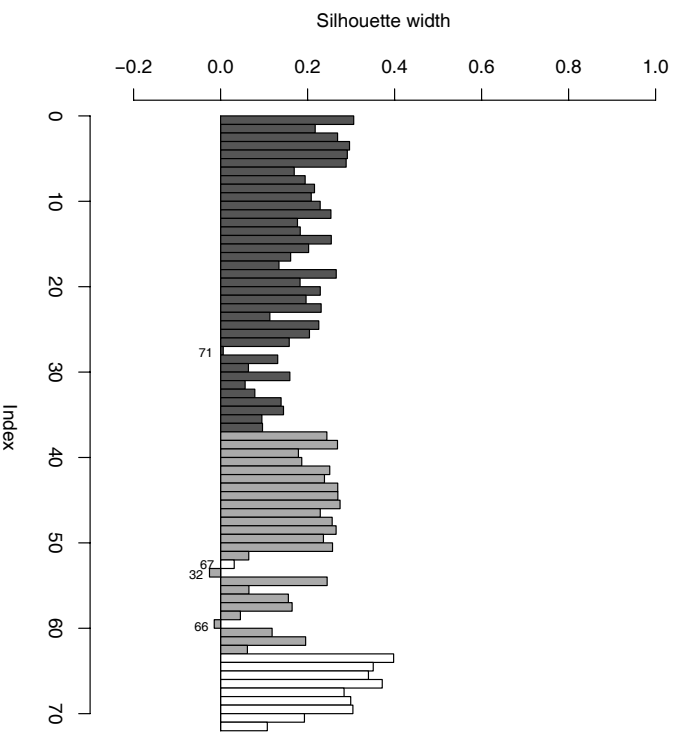
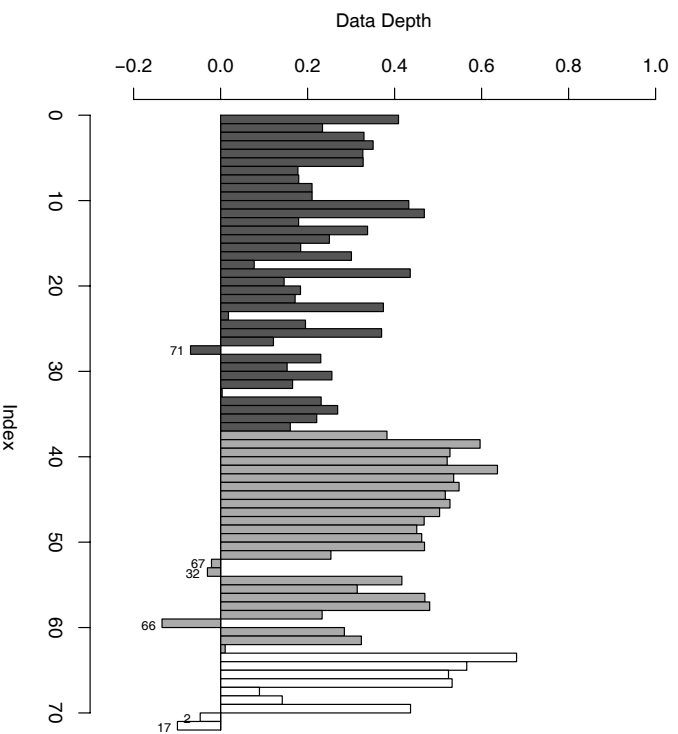
We apply the ReD cluster validation and selection criterion to

- a) The Acute Leukemia data set (Golub et al).
- b) The NCI60 data set (Ross et al).
- c) Noisy Leukemia data set (adding noise to data).
- d) Leukemia, including unrelated features.
- e) Several simulated data sets.

We use ReD with PAM/ k -median, and compare with sil.

We also use GapReD, where ReD(k) replaces W_k in the Gap algorithm.

Leukemia Results



ReD

sil

Both ReD and sil identify observations that are difficult to classify in a supervised fashion.

Leukemia Results

All methods select $K = 3$ clusters, using the 50, 100 or 200 genes with highest variance across samples.

(NCI60 data set; GapRed, sil and ReD select $K = 3$ for 50, 100, 200 genes, whereas Gap selects $K = 3$ for 50, 200 and $K = 6$ for 100.)

Simulation study:

a) Including unrelated features. 100 high-variance genes and 100 unrelated (random normal, mean 0, same variance as the high-variance genes).

Method	Number of clusters				
100 unrelated	1	2	3	4	5
sil	0	38	10	2	0
ReD	0	0	35	11	4
GapPC	0	48	2	0	0
GapReD	0	19	29	2	0

Leukemia Results

b) Noisy data. SNR 4, 2 and 1.25.

Method	Number of clusters				
SNR=4	1	2	3	4	5
sil	0	28	22	0	0
ReD	0	1	49	0	0
GapPC	0	22	28	0	0
GapReD	0	1	49	0	0
SNR=2	1	2	3	4	5
sil	0	29	21	0	0
ReD	0	0	46	4	0
GapPC	0	39	11	0	0
GapReD	0	8	42	0	0
SNR=1.25	1	2	3	4	5
sil	0	40	10	0	0
ReD	0	7	33	10	0
GapPC	0	43	7	0	0
GapReD	0	18	32	2	0

Leukemia Results

We also compare the k-median and PAM at these SNR levels.

With no added noise - PAM generates 1 error, k-median 2 errors.

If we standardize the genes - PAM generates 3 errors but the k-median is unaffected.

We simulate 50 noisy data sets and compare the clusterings generated by PAM and the k-median, to the known labels.

SNR	Errors	
	PAM	k-median
SNR=4	3.6	2.7
SNR=2	5.9	3.4
SNR=1.25	9.2	4.6

Simulation Results

Model 1: 3 clusters in 2 dimensions. 25, 25, and 50 observations in each cluster. Cluster means (0,0), (0,5) and (5,3), identity covariance matrix, Gaussian distribution.

Method	Number of clusters				
Model 1	1	2	3*	4	5
sil	0	5	45	0	0
ReD	0	0	49	1	0
GapPC	0	0	49	1	0
GapReD	0	0	50	0	0

Simulation Results

Model 2: 4 clusters in 10 dimensions, 7 noise variables. Cluster sizes randomly chosen 25 or 50. Cluster means for first 3 variables drawn from $N(0_3, 5I_3)$, for the last 7 from $N(0_7, I_7)$.

Method	Number of clusters				
Model 2	1	2	3	4*	5
sil	0	20	18	12	0
ReD	0	11	16	23	0
GapPC	0	1	12	37	0
GapReD	0	2	18	30	0

Simulation Results

Model 3: 4 clusters in 10 dimensions. Cluster sizes drawn from (25,50).

Cluster means drawn from $N(0_{10}, 1.9I_{10})$.

Method	Number of clusters				
Model 3	1	2	3	4*	5
sil	0	4	10	36	0
ReD	0	1	16	33	0
GapPC	0	0	3	47	0
GapReD	0	1	6	43	0

Simulation Results

Model 4: 3 clusters in 3 dimensions. Cluster sizes (25,25,50). Cluster means (5,0,0), (0,-5,5), and (0,5,5). Covariance matrix $4I_3$.

Method	Number of clusters				
Model 4	1	2	3*	4	5
sil	0	33	17	0	0
ReD	0	3	45	2	0
GapPC	0	7	43	0	0
GapReD	0	3	42	5	0

Simulation Results

Model 5: Same as model 4, but including 3 noise variables with means $(0,0,0)$. Covariance matrix $1.5^2 I_3$ for all variables.

Method	Number of clusters				
Model 5	1	2	3*	4	5
sil	0	34	16	0	0
ReD	0	4	43	3	0
GapPC	0	3	46	1	0
GapReD	0	0	44	6	0

Simulation Results

Model 6: Linearly dependent cluster means. Cluster sizes (25,25,50). Cluster means (0,0,0), (0,-5,5), (0,5,-5). Variance along z-axis .5, variance in x and y 1.5.

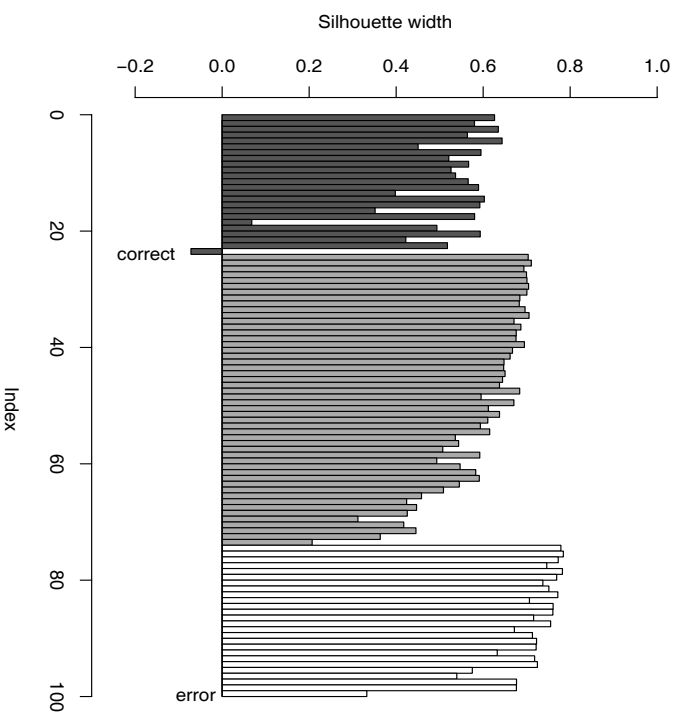
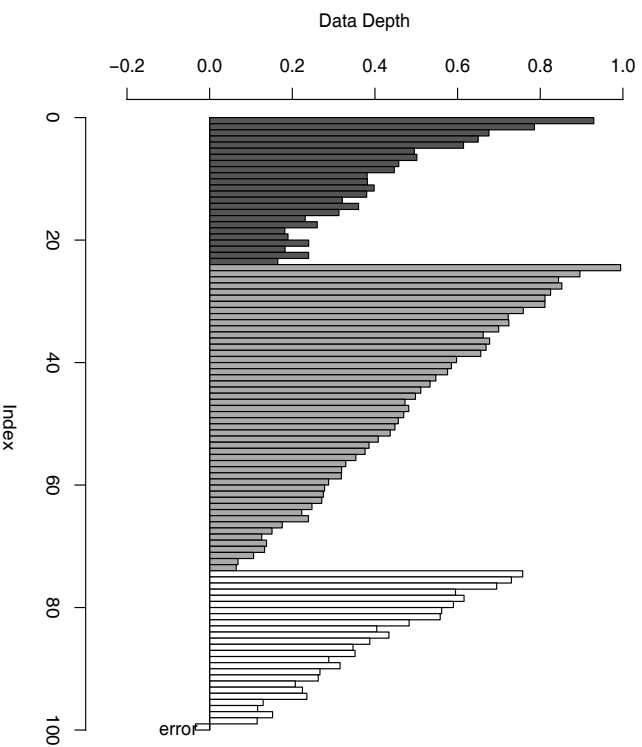
Method	Number of clusters				
Model 6	1	2	3*	4	5
sil	0	0	50	0	0
ReD	0	26	24	0	0
GapPC	0	0	49	1	0
GapReD	0	0	47	3	0

Simulation Results

Model 7: 3 clusters. Cluster 1 and 3 are clearly separated from Cluster 2. Cluster 1 is loose, Cluster 3 is tight. There are 50 observations in each.

Method	Number of clusters				
Model 7	1	2	3*	4	5
sil	0	38	12	0	0
ReD	0	2	47	1	0
GapPC	0	2	6	37	5
GapReD	0	4	37	9	0

Identifying outliers



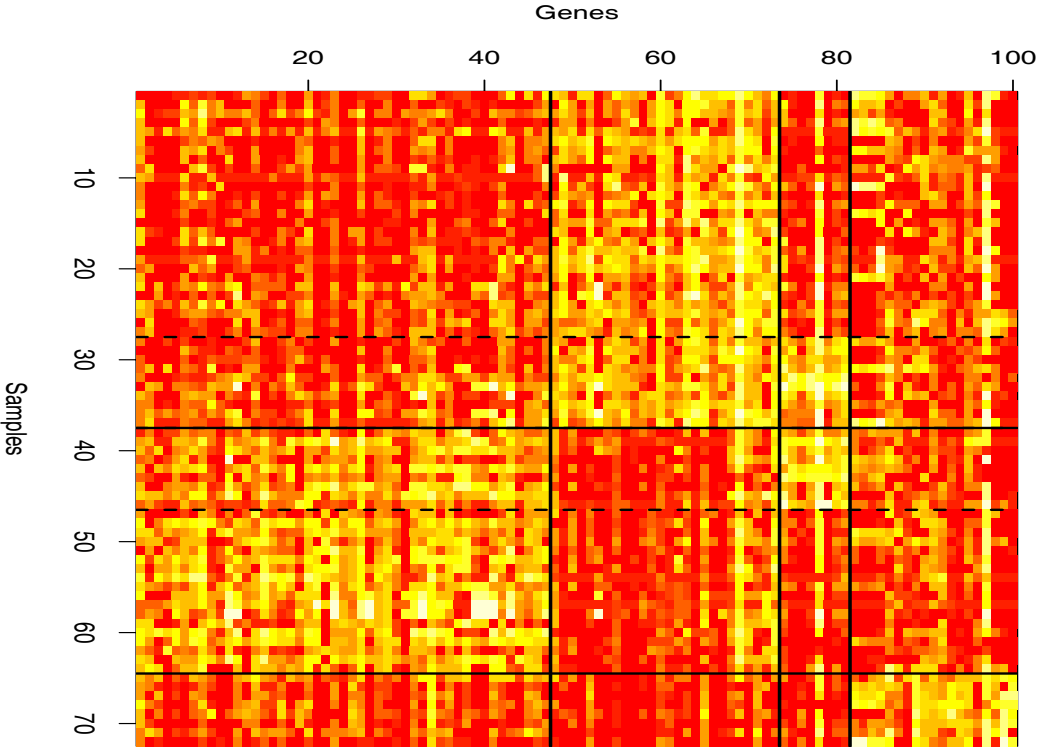
ReD

sil

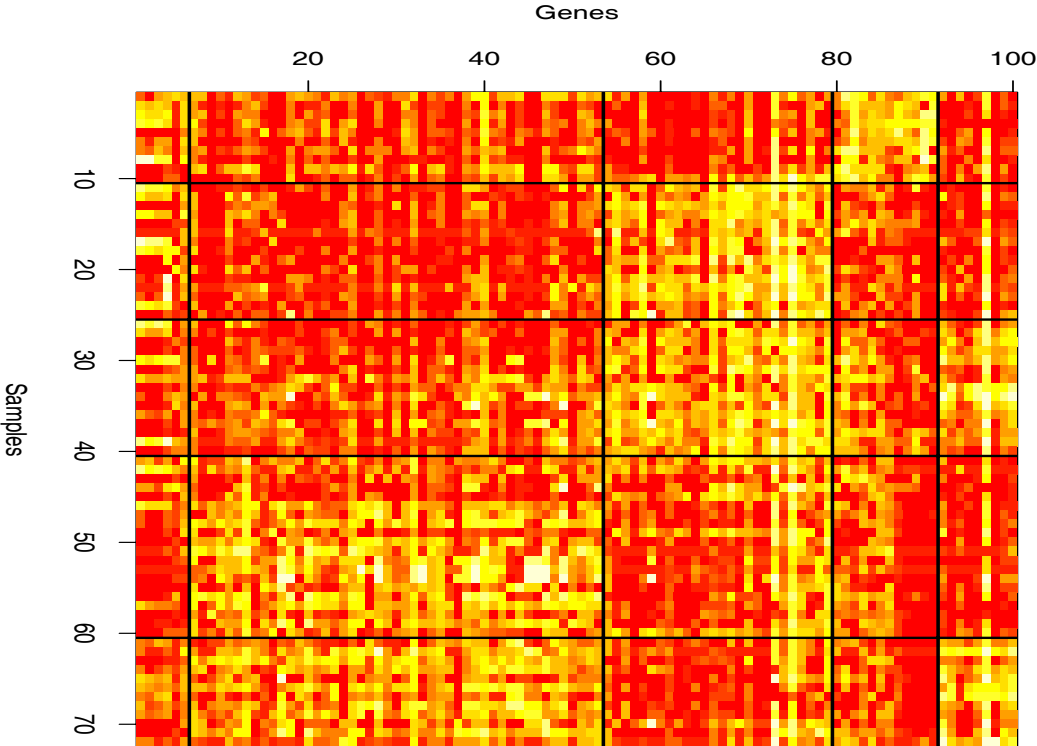
ReD correctly identifies the error, whereas there is a negative sil for a correctly clustered observation. This observation falls in the loose cluster, and is compared with the nearest tight cluster.

Gene Clustering followed by Sample Clustering

k-median



PAM



Gene Clustering followed by Sample Clustering

- PAM did not identify a subdivision of cancer samples, but the k-median did. The AML and ALL-B sample classes were split in two.
- If we don't cluster the genes, and generate 5 sample clusters - we do not get this subdivision. Gene clustering allows us to discover structure in the data that is otherwise occluded.
- The ALL-B, ALL-T and AML classes correspond to a local minimum of the k-median algorithm ($k=3$) applied to the gene cluster data. This local minimum has a much higher average data depth.

Conclusion

- ReD is a useful tool for outlier identification, visualization.
- ReD is a robust cluster validation and selection criterion.
- ReD outperforms sil and Gap when the within-cluster variances differ.
- For some noisy, high-dimensional data sets, using the PAM approximation of the k-median can be risky.
- Future work: Can use a ReD penalty in k-median clustering.
Will this generate tighter clusters, more meaningful clusters?