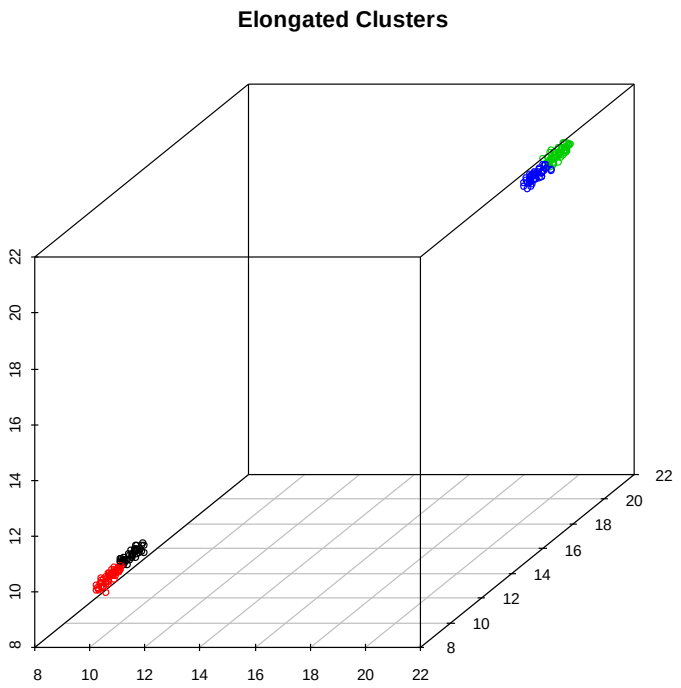
General Prediction Strength Methods for Estimating the Number of Clusters in a Dataset

Qiao Yu

April 15, 2018

Motivation

- *k*-means (medoids) clustering will happily divide any dataset into *k* clusters, regardless of whether that's appropriate or not.



Overview

- Review of previous methods
- Re-formulation and extension of Tibshirani's prediction strength method
- Contrast results for different cluster configurations
- Application to gene co-expression network

Different Methods for Deciding Number of Clusters

- Methods based on internal indices
  - Depend on between- and within- sum of squared error (BSS and WSS)
- Methods based on external indices
  - Depends on comparison between different partitionings
- Evaluate indices for different values of *k* and decide which is “best”

Internal Index Methods

Internal Indices

- Calinski & Harabasz
- Hartigan
- Krzanowski & Lai

- n = Number of samples
- p = Dimension of samples

Calinski and Harabasz (1974)

- For each number of clusters $k \geq 2$, define the index

$$I_k = \frac{\text{trace}(BSS_k)/(k-1)}{\text{trace}(WSS_k)/(n-k)}$$

- The estimated number of clusters is the k which maximizes the above.

Hartigan

- For each number of clusters $k \geq 1$, define the index

$$I_k = \left(\frac{\text{trace}(WSS_k)}{\text{trace}(WSS_{k+1})} - 1 \right)(n - k - 1)$$

- The estimated number of clusters is the smallest $k \geq 1$ such that $I_k \leq 10$.

Krzanowski and Lai (1985)

- For each number of clusters $k \geq 2$, define the indices

$$d_k = (k-1)^{2/p}\,\text{trace}(WSS_{k-1})\,\text{trace}(WSS_k), \text{ and}$$

$$I_k = |d_k||d_{k+1}|$$

- The estimated number of clusters is the k which maximizes I_k.

The silhouette width method (Kaufman and Rousseeuw, 1990)

- Silhouettes use average dissimilarity between observation *i* and other observations in the same cluster.
- Silhouette width of the observation is

$$I_{ik} = (b_i - a_i) / \max(a_i, b_i)$$

- a_i = average dissimilarity of observation *i*
- b_i = minimum dissimilarity within the cluster

The silhouette width method (cont.)

- Overall silhouette width is the average over all observations:

$$I_k = \frac{\sum_i I_{ik}}{n}$$

- The estimated number of clusters is the k for which I_k is maximized.

Gap (uniform) or Gap(pc) (Tibshirani et al., 2000)

- For each number of clusters *k*,

$$I_k = \frac{1}{B}\sum_b \log(\text{trace}(WSS_k^b) - \log(\text{trace}(WSS_k)$$

- *B* reference datasets generated under null distribution.

Gap statistic (cont.)

- Estimated number of clusters is smallest $k \geqslant 1$ that maximizes I_k and satisfies

$$gap_k \geq gap_{k+1} - s_{k+1}$$

- s_k = standard deviation over reference datasets.
- Uniform gap statistic samples from a uniform distribution
- “pc” (principal component) statistic samples from a uniform box aligned with the principal components of the dataset (Sarle, 1983).

External Index Methods

External Indices/Approaches

- Comparing Partitionings
- Rand Index
- Tibshirani
- Clest
- General Prediction Strength

Comparing Partitionings: The Contingency Table

- Partitionings $U=\{u_1,..., u_R\}$ and $V = \{v_1,..., v_S\}$ of n objects into R and S clusters

U/V	v_1	v_2	...	v_S
u_1	n_{11}	n_{12}	...	n_{1S}
u_2	n_{21}	n_{22}	...	...
...	...	...	...	...
u_R	n_{R1}	n_{R1}	...	n_{RS}

Comparing Partitionings: The Contingency Table

- n_{rs} = number of objects in both u_r and v_s.

U/V	v_1	v_2	...	v_S
u_1	n_{11}	n_{12}	...	n_{1S}
u_2	n_{21}	n_{22}	...	...
...	...	...	...	...
u_R	n_{R1}	n_{R1}	...	n_{RS}

Comparing Partitionings: The Contingency Table

- $n_{r.} = \sum_{s=1}^{S} n_{rs}$ = total points in cluster u_r
- $n_{.s} = \sum_{r=1}^{R} n_{rs}$ = total points in cluster v_s

U/V	v_1	v_2	...	v_S
u_1	n_{11}	n_{12}	...	n_{1S}
u_2	n_{21}	n_{22}	...	...
...	...	...	...	...
u_R	n_{R1}	n_{R1}	...	n_{RS}

Rand Index (Rand, 1971, Hubert and Arabie, 1985)

- Rand index and adjusted Rand index (m=2)

$$Rand = \frac{\binom{n}{m} - \sum_s \binom{n_{.s}}{m} - \sum_r \binom{n_{r.}}{m} + 2\sum_{r,s} \binom{n_{rs}}{m}}{\binom{n}{m}}$$

$$Adj.Rand = \frac{\sum_{r,s} \binom{n_{rs}}{m} - \sum_s \binom{n_{.s}}{m} \sum_r \binom{n_{r.}}{m} \Big/ \binom{n}{m}}{\frac{1}{2}\left(\sum_s \binom{n_{.s}}{m} + \sum_r \binom{n_{r.}}{m}\right) - \sum_s \binom{n_{.s}}{m} \sum_r \binom{n_{r.}}{m} \Big/ \binom{n}{m}}$$

Clustering as a supervised classification problem

- Input data split repeatedly into a training and a test set for a given choice of *k* (number of clusters)
- Clustering method applied to the two sets to arrive at *k* “observed” training and test set clusters.
- Use the training data to construct a classifier for predicting the training set cluster labels.
- Apply classifier to test set data -> predicted test set clusters.
- Measure of agreement calculated based on the comparison of predicted to observed test set clusters (external index).

Predicting the number of clusters

- Use cluster reproducibility measures for different *k* to estimate the true number of clusters in the data set.
- Assumes that choosing the correct number of clusters -> less random assignment of samples to clusters and to greater cluster reproducibility.

Clest
(Dudoit and Fridlyand, 2002)

- Step "A" identical to steps 1-6 of Tibshirani PS. Denote external indices computed in step A.6. by ($s_{k,1}$, $s_{k,2}$,…$s_{k,B}$). Then
  - B. Let t_k = median($s_{k,1}$, ..., $s_{k,B}$) denote observed similarity statistic for the k-cluster partition of the data.
  - C. Generate B_0 datasets under null hypothesis of k=1. Briefly, for each reference dataset, repeat the procedure described in steps A and B above, to obtain B_0 similarity statistics $t_{k,1}$, ..., $t_{k,Bo}$.
- Let t_k^0 denote average of the B_0 statistics
- Let $d_k = t_k - t_k^0$ denote the difference between the observed similarity statistic and estimated expected value under null hypothesis of k = 1.

General Prediction Strength

- Re-formulation of Clest
- Extension to m-tuplets

Tests on Simulated Data

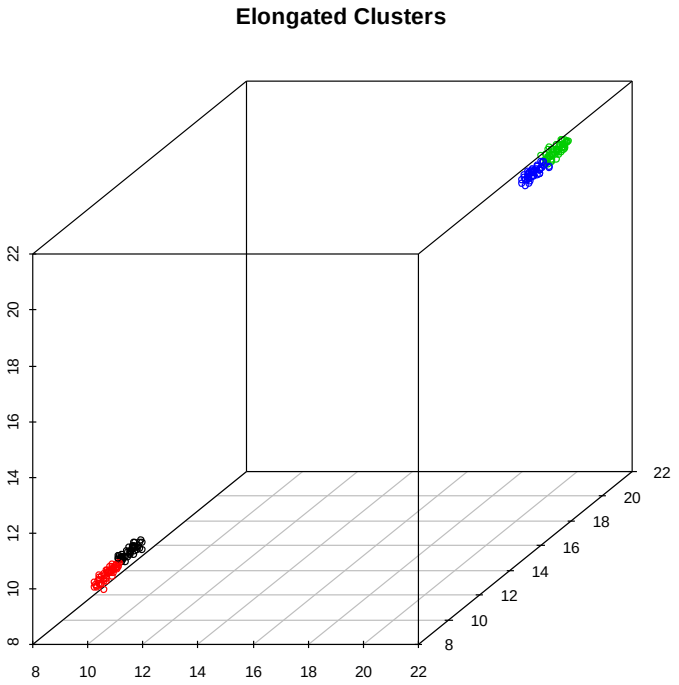
Simulations

1. A single cluster containing 200 points uniformly distributed from (-1,1) in 10-d.
2. Three normally distributed clusters in 2-d with centers at (0,0). (0,5), and (5,-3) and 25, 25, and 50 observations in each respective cluster.
3. Four normally distributed clusters in 3-d with centers randomly chosen from N(0, 5*I) and cluster size randomly chosen from {25, 50}.
4. Four normally distributed clusters in 10-d with centers randomly chosen from N(0, 1.9*I) and cluster size randomly chosen from {25, 50}.

In 3 & 4, simulations with clusters with minimum distance less than one unit were discarded.

Simulations

5. Two elongated clusters in 3-d. Generated by choosing equally spaced points between (-0.5, 0.5) and adding normal noise with sd 0.1 to each feature. Then add 10 to each feature of the points in the second cluster.
   (a) 100 points per cluster
   (b) 200 points per cluster (to illustrate effects of an increased number of observations)



	Predicted # Clusters										
Method	No Pred	1	2	3	4	5	6	7	8	9	10
sim1 (1 cluster, 10d)											
Hartigan			7	27	12	4					
Calinski		NA				3	7	8	12	12	10
Kraznowski-Lai		NA	8	8	5	7	5	3	5	7	2
Silhouette		NA	33	4	1	1	1	1	3	6	12
Gap (uniform)		50									
Gap (pc)		50									
Clest*		48	2								
sim2 (3 clusters, 2d)											
Hartigan	4				5	4	9	8	8	7	5
Calinski		NA		50							
Kraznowski-Lai		NA		29	3	4	2	1	1	5	5
Silhouette		NA	6	44							
Gap (uniform)			11	39							
Gap (pc)			12	38							
Clest*			1	49							
sim3 (4 clusters, 3d)											
Hartigan	11				2	10	10	4	4	4	5
Calinski		NA	1	6	43						
Kraznowski-Lai		NA	2	6	37	1	1			1	2
Silhouette		NA	8	13	29						
Gap (uniform)		5	7	16	20	2					
Gap (pc)		12	6	15	17						
Clest*			1	20	29						

	Predicted # Clusters										
Method	No Pred	1	2	3	4	5	6	7	8	9	10
sim4 (4 clusters, 10d)											
Hartigan					50						
Calinski		NA	2	9	39						
Kraznowski-Lai		NA			42	1	2	1		2	2
Silhouette		NA	5	11	34						
Gap (uniform)		1	2	20	27						
Gap (pc)		5	6	7	32						
Clest*				1	49						
sim5a (2 clusters, 3d, 100pts/clus)											
Hartigan	35									1	14
Calinski		NA			15		29	4	2		
Kraznowski-Lai		NA	49						1		
Silhouette		NA	50								
Gap (uniform)				31	9	4	6				
Gap (pc)			50								
Clest*			44		6						

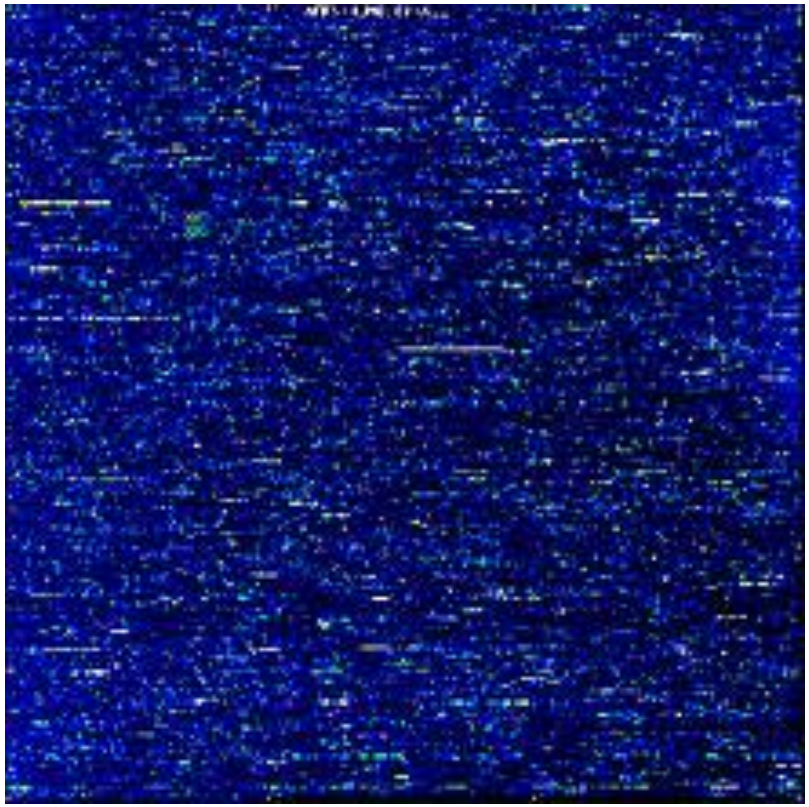
Results of Simulations

- Clest performs consistently well
- Not all values of *m* perform equally well in all the simulations (*m*=3 and *m*=5 do best overall)
- Performance especially noticeable on elongated cluster simulation.
- Of internal index methods, **Hartigan** seems least robust
- **Calinski** and **Kraznowski-Lai indices** and the silhouette width method cannot predict a single cluster.

Application to Gene Co-Expression Networks

DNA Microarrays

- Expression level of thousands of genes at once
- Lots of processing and normalization

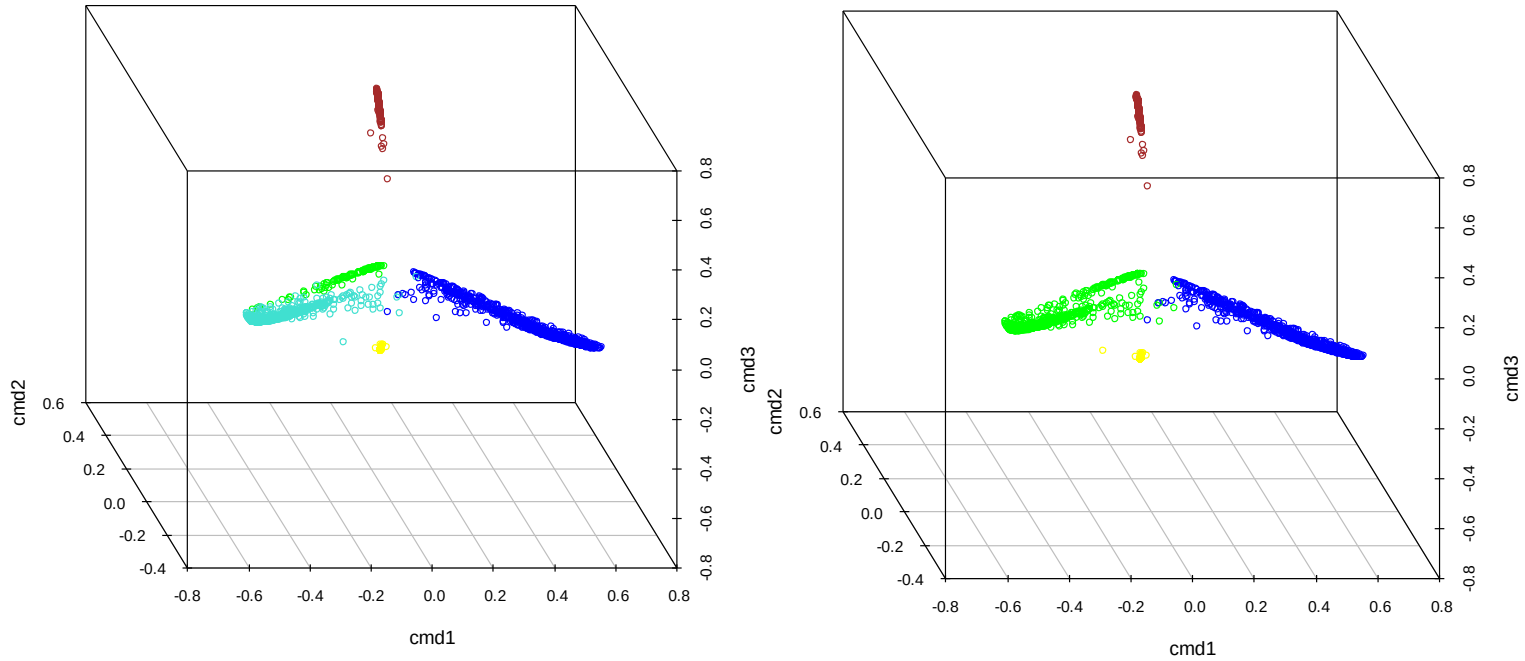


Use of Microarrays

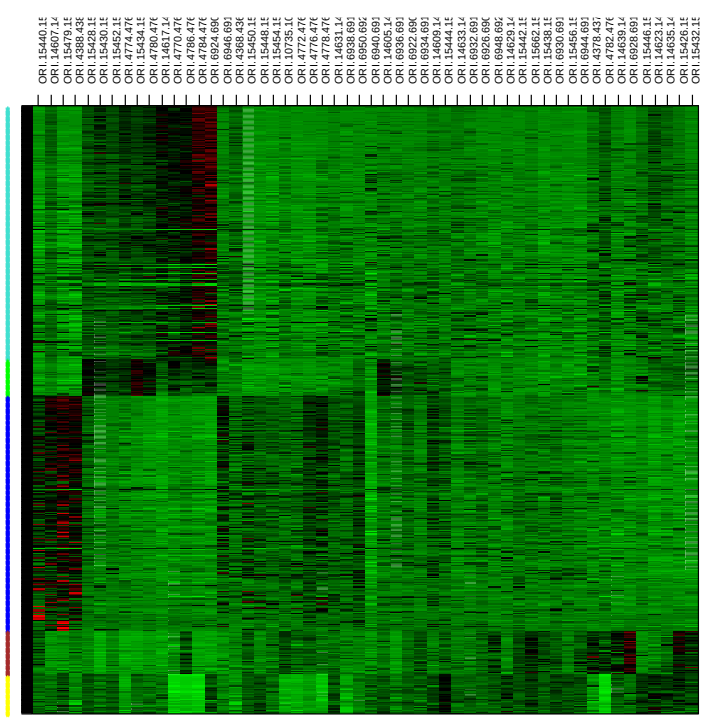
- Within an experiment use “normal” and “diseased” cell types (e.g.).
- Generally examined for differences in expression levels between cell types.
- Look for genes that characteristically vary with disease.

Classical Multi-Dimensional Scaling

- Used to visualize abstract TOM dissimilarity
- “Principal component analysis”



Inspection of Heatmap



- Red for highly expressed genes
- Green for low expression
- Consistent expression across genes (rows) in clusters

=> Either 4 or 5 clusters justified

Conclusion

- There are several indices for evaluating clusterings
  - External compare different partitionings, internal do not
- Indices can be used to predict number of clusters
- Prediction Strength index method works across different cluster configurations
- Fairly simple and intuitive
- Effective on elongated clusters
- Results of varying *m* reflect hierarchical structure in data

Acknowledgements

- Steve Horvath
- Meghna Kamath
- Fred Fox and Tumor Cell Biology Training Grant (USHHS Institutional National Research Service Award #T32 CA09056)
- Stan Nelson and the UCLA Microarray Core Facility
- NIH Program Project grant #1U19AI063603-01.

References

- http://www.genetics.ucla.edu/labs/horvath/GeneralPredictionStrength.
- CALINSKI, R. & HARABASZ, J. (1974). A dendrite method for cluster analysis. *Commun Statistics*, 1-27.
- DUDOIT, S. & FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* **3**, RESEARCH0036.
- EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. & BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-8.
- FREIJE, W. A., CASTRO-VARGAS, F. E., FANG, Z., HORVATH, S., HARTIGAN, J. A. (1985). Statistical theory in clustering. *J. Classification*, 63-76.
- HUBERT, L. & ARABIE, P. (1985). Comparing Partitions. *Journal of Classification* **2**, 193-218.
- KAUFMAN, L. & ROUSSEEUW, P. J. (1990). *Finding groups in data: an introduction to cluster analysis.* New York: Wiley.
- KRZANOWSKI, W. & LAI, Y. (1985). A criterion for determining the number of groups in a dataset using sum of squares clustering. *Biometrics*, 23-34.
- MISCHEL, P., ZHANG, B., CARLSON, M., FANG, Z., FREIJE, W., CASTRO, E., SCHECK, A., LIAU, L., KORNBLUM, H., GESCHWIND, D., CLOUGHESY, T., HORVATH, S. & NELSON, S. (2005). Hub Genes Predict Survival for Brain Cancer Patients.
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 846-850.
- RAVASZ, E., SOMERA, A. L., MONGRU, D. A., OLTVAI, Z. N. & BARABASI, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-5.
- SARLE, W. (1983). Cubic Clustering Criterion. SAS Institute, Inc.
- TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. & GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96**, 2907-12.
- TIBSHIRANI, R., WALTHER, G., BOTSTEIN, D. & BROWN, P. (2001). Cluster validation by prediction strength. Stanford.
- TIBSHIRANI, R., WALTHER, G. & HASTIE, T. (2000). Estimating the number of clusters in a dataset via the gap statistic. Department of Biostatistics, Stanford.
- YEUNG, K. Y., HAYNOR, D. R. & RUZZO, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics* **17**, 309-18.
- ZHANG, B. & HORVATH, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*.

Extra Slides

WSS and BSS

$$TSS_{ij} = \sum_{s=1}^{n} (x_{is} - \bar{x}_i)(x_{js} - \bar{x}_j)$$

$$= \sum_{g} WSS_{ij}^{g} + BSS_{ij}^{g}$$

$$BSS_{ij}^{g} = n_g (\bar{x}_i^{g} - \bar{x}_i)(\bar{x}_j^{g} - \bar{x}_j)$$