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

Moira Regelson, Ph.D.
September 15, 2005
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

\[
I_{ik} = \frac{(b_i - a_i)}{\max(a_i, b_i)}
\]

- \( a_i \) = 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:

\[
I_k = \frac{\sum I_{ik}}{n}
\]

• 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 $k$,

\[ I_k = \frac{1}{B} \sum_b \log(\text{trace}(WSS_b^k)) - \log(\text{trace}(WSS_k)) \]

• $B$ reference datasets generated under null distribution.
Gap statistic (cont.)

- Estimated number of clusters is smallest $k \geq 1$ 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 $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_S\}$ of $n$ objects into $R$ and $S$ clusters

<table>
<thead>
<tr>
<th>$U/V$</th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>...</th>
<th>$v_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>...</td>
<td>$n_{1S}$</td>
</tr>
<tr>
<td>$u_2$</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_R$</td>
<td>$n_{R1}$</td>
<td>$n_{R1}$</td>
<td>...</td>
<td>$n_{RS}$</td>
</tr>
</tbody>
</table>
Comparing Partitionings: The Contingency Table

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

<table>
<thead>
<tr>
<th>$U/V$</th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>...</th>
<th>$v_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>...</td>
<td>$n_{1S}$</td>
</tr>
<tr>
<td>$u_2$</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_R$</td>
<td>$n_{R1}$</td>
<td>$n_{R1}$</td>
<td>...</td>
<td>$n_{RS}$</td>
</tr>
</tbody>
</table>
Comparing Partitionings: The Contingency Table

- \( n_{r.*} = \sum_{s=1}^{S} n_{rs} = \text{total points in cluster } u_r \)

- \( n_{.*s} = \sum_{r=1}^{R} n_{rs} = \text{total points in cluster } v_s \)

<table>
<thead>
<tr>
<th>( U/V )</th>
<th>( v_1 )</th>
<th>( v_2 )</th>
<th>...</th>
<th>( v_S )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
<td>...</td>
<td>( n_{1S} )</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( u_R )</td>
<td>( n_{R1} )</td>
<td>( n_{R1} )</td>
<td>...</td>
<td>( n_{RS} )</td>
</tr>
</tbody>
</table>
Rand Index (Rand, 1971, Hubert and Arabie, 1985)

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

\[
\text{Rand} = \frac{\binom{n}{m} - \sum_{s} \binom{n_s}{m} - \sum_{r} \binom{n_r}{m} + 2 \sum_{r,s} \binom{n_{rs}}{m}}{\binom{n}{m}}
\]

\[
\text{Adj.Rand} = \frac{\sum_{r,s} \binom{n_{rs}}{m} - \sum_{s} \binom{n_s}{m} \sum_{r} \binom{n_r}{m} / \binom{n}{m}}{\frac{1}{2} \left( \sum_{s} \binom{n_s}{m} + \sum_{r} \binom{n_r}{m} \right) - \sum_{s} \binom{n_s}{m} \sum_{r} \binom{n_r}{m} / \binom{n}{m}}
\]
Clustering as a supervised classification problem

1. Input data split repeatedly into a training and a test set for a given choice of \( k \) (number of clusters).
2. Clustering method applied to the two sets to arrive at \( k \) “observed” training and test set clusters.
3. Use the training data to construct a classifier for predicting the training set cluster labels.
4. Apply classifier to test set data -> predicted test set clusters.
5. 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 $\rightarrow$ less random assignment of samples to clusters and to greater cluster reproducibility.
Tibshirani Prediction Strength
(Tibshirani et al., 2001)

• Specify $k_{max}$ and max number of iterations, $B$.
• For $k$ in $\{2 \ldots k_{max}\}$, repeat $B$ times:
  1. Split data set into a training set and a test set
  2. Apply clustering procedure to partition training set into $k$ clusters, record cluster labels as outcomes.
  3. Construct classifier using training set and cluster labels.
  4. Apply resulting classifier to test set -> “predicted” labels.
  5. Apply the clustering procedure to the test set to arrive at the “observed” labels.
  6. Compute a measure of agreement (external index) $ps1(k,b)$ comparing the sets of labels obtained in steps 4 and 5.
7. Switch the role of the test and training sets to arrive at another estimate of the index, $ps2(k,b)$.

8. Use $ps1(k,b)$ and $ps2(k,b)$ to compute the mean value, $ps(k,b) = (ps1(k,b) + ps2(k,b))/2$, and standard error, $se(k,b) = |ps1(k,b) - ps2(k,b)|/2$. Use these to define $pse(k,b) = ps(k,b) + se(k,b) = \max(ps1(k,b), ps2(k,b))$. 
Tibshirani PS (cont.)

• $pse(k) = \text{median of } pse(k,b) \text{ over all random splits}$

• Values of $pse(k)$ used to estimate the number of clusters in the dataset using a threshold rule
**Clest**
*(Dudoit and Fridlyand, 2002)*

- Step “A” identical to steps 1-6 of Tibshirani PS. Denote external indices computed in step A.6. by \((s_{k,1}, s_{k,2}, \ldots s_{k,B})\). Then
  
  B. Let \(t_k = \text{median}(s_{k,1}, \ldots, s_{k,B})\) 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 \(t_{k,1}, \ldots, t_{k,B_0}\).

- Let \(t^0_k\) denote average of the \(B_0\) statistics

- Let \(d_k = t_k - t^0_k\) denote the difference between the observed similarity statistic and estimated expected value under null hypothesis of \(k = 1\).
General Prediction Strength

• Re-formulation of Tibshirani PS
• Extension to m-tuplets
Tibshirani re-formulation

• Originally, measure of agreement formulated as

\[ ps(k) = \min_{1 \leq s \leq k} \frac{1}{n_s(n_s - 1)} \sum_{x_i \neq x_i' \in v_s} I(D[U(X_{te})]_{ii'} = 1) \]

\[ U(X_{te}) = \left\{ \arg\min_r (\text{dist}(\text{medoid}(u_r), x_j \in X_{te})) \right\} \]

\[ D[U]_{ij} = \begin{cases} 1 & x_i, x_j \in u_r \text{ for some cluster } u_r \in U \\ 0 & \text{otherwise} \end{cases} \]

• Note: partitioning around medoid (PAM) clustering used
General Prediction Strength

• Re-formulation of Tibshirani PS & extension to m-tuplets:

\[
PS(k,m) = \min_{1 \leq s \leq k} p_m(s), \text{ where } p_m(s) = \frac{\sum_{r=1}^{k} \binom{n_{rs}}{m}}{\binom{n_s}{m}}
\]

• Add a standard error in cross-validation: \(PSE(k,m)\) used with threshold.

• Intuitive interpretation: fraction of \(m\)-tuplets in test set that co-cluster in the training set
Asymmetry in PS

• Difference between Rand index and PSE(k,m):

  • Rand = 0.95, adjusted Rand = 0.86.

  • \( PSE(k,m=2) = 0.17 \), but when role of \( U \) and \( V \) reversed,
    \( PSE(k,m=2) = 0.83 \).

\[
\begin{array}{|c|ccccc|}
\hline
U/V & v_1 & v_2 & v_3 & v_4 & v_5 \\
\hline
u_1 & 5   & 0   & 0   & 0   & 0   \\
 u_2 & 5   & 50  & 0   & 0   & 0   \\
 u_3 & 5   & 0   & 50  & 0   & 0   \\
 u_4 & 5   & 0   & 0   & 50  & 0   \\
 u_5 & 5   & 0   & 0   & 0   & 50  \\
\hline
\end{array}
\]
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)
<table>
<thead>
<tr>
<th>Method</th>
<th>Predicted # Clusters</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Predicted # Clusters</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Pred</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>sim1 (1 cluster, 10d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartigan</td>
<td>7 27 12 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>m=2</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calinski</td>
<td>NA</td>
<td>3 7 8 12 12</td>
<td></td>
<td></td>
<td></td>
<td>m=3</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kraznowski-Lai</td>
<td>NA 8 8 5 7 5 3 5 7 2</td>
<td>m=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silhouette</td>
<td>NA 33 4 1 1 1 3 6 12</td>
<td>m=10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (uniform)</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (pc)</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clest*</td>
<td>48 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sim2 (3 clusters, 2d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartigan</td>
<td>4 5 4 9 8 8 7</td>
<td>m=2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>m=3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calinski</td>
<td>NA 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kraznowski-Lai</td>
<td>NA 29 3 4 2 1 1 5 5</td>
<td>m=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silhouette</td>
<td>NA 6 44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (uniform)</td>
<td>11 39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (pc)</td>
<td>12 38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clest*</td>
<td>1 49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sim3 (4 clusters, 3d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartigan</td>
<td>11 2 10 10 4 4 4</td>
<td>m=2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>m=3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calinski</td>
<td>NA 1 6 43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kraznowski-Lai</td>
<td>NA 2 6 37 1 1 1 2 2</td>
<td>m=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silhouette</td>
<td>NA 8 13 29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (uniform)</td>
<td>5 7 16 20 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (pc)</td>
<td>12 6 15 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clest*</td>
<td>1 20 29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Predicted # Clusters</td>
<td>Prediction Strength (cutoff=0.8)</td>
<td>Predicted # Clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------</td>
<td>----------------------------------</td>
<td>----------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No Pred 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>sim4 (4 clusters, 10d)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartigan</td>
<td>NA 2 9 39</td>
<td>m=2 3 34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calinski</td>
<td>NA 42 1 2 1 2 2</td>
<td>m=3 5 3 38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kraznowski-Lai</td>
<td>NA 5 11 34</td>
<td>m=5 6 4 5 35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silhouette</td>
<td>NA 19</td>
<td>m=10 19 4 2 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (uniform)</td>
<td>1 2 20 27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (pc)</td>
<td>5 6 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleft*</td>
<td>1 49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>sim5a (2 clusters, 3d, 100pts/clus)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartigan</td>
<td>35</td>
<td>m=2 46 3 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calinski</td>
<td>NA 15 29 4 2</td>
<td>m=3 48 2 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kraznowski-Lai</td>
<td>NA 1 1</td>
<td>m=5 49 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silhouette</td>
<td>NA 31 9 4 6</td>
<td>m=10 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (uniform)</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (pc)</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleft*</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>sim5b (2 clusters, 3d, 200pts/clus)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartigan</td>
<td>50</td>
<td>m=2 27 23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calinski</td>
<td>NA 20 29 1</td>
<td>m=3 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kraznowski-Lai</td>
<td>NA 50</td>
<td>m=5 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silhouette</td>
<td>NA 50</td>
<td>m=10 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (uniform)</td>
<td>34 5 6 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap (pc)</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results of Simulations

- PS 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.
- Clest performs comparably to PS.
- 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.
Gene co-expression network

- Use DNA microarray data to annotate genes by clustering them on the basis of their expression profiles across several microarrays.

- Studying co-expression patterns can provide insight into the underlying cellular processes (Eisen et al., 1998, Tamayo et al., 1999).
Building a network

• Use Pearson correlation coefficient as a co-expression measure.
• Threshold correlation coefficient to arrive at gene co-expression network.
Building a network (cont.)

• Node corresponds to expression profile of a given gene.
• Nodes connected \((a_{ij}=1)\) if they have significant pairwise expression profile association across perturbations (cell or tissue samples).
Topological Overlap Matrix

• TOM given by (Ravasz et al, 2002)

\[
\omega_{ij} = \sum_{u} a_{iu} a_{uj} + a_{ij} \over \min(k_i, k_j) + 1 - a_{ij}
\]

• \(k_i\) is the connectivity of node \(i\):

\[
k_i = \sum_{j=1}^{n} a_{ij}
\]
• $\omega_{ij} = 1$ if the node with fewer links satisfies two conditions:
  – all of its neighbors are also neighbors of the other node and
  – it is linked to the other node. In contrast,
• $\omega_{ij} = 0$ if nodes i and j are unlinked and the two nodes have no common neighbors.
• similarity measure, associated dissimilarity measure $d_{ij} = 1 - \omega_{ij}$.
Gene Co-Expression Network for Brain Tumor Data

• Brain tumor (glioblastoma) microarray data (previously described in Freije et al (2004), Mischel et al (2005), and Zhang and Horvath (2005)).

• Network constructed using correlation threshold of 0.7 and the 1800 most highly connected genes.
Gene Co-Expression Network for Brain Tumor Data (cont.)

- Used PS method (with PAM) on TOM with $m=(2,5,10)$
- $m=2$, $m=5$ -> 5 clusters
  - Same as Mischel et al.
- $m=10$ -> 4 clusters
  - Reasonable interpretation?
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.
Extra Slides
WSS and BSS

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

\[ = \sum_{g} WSS_{ij}^g + BSS_{ij}^g \]

\[ BSS_{ij}^g = n_g (\bar{x}_i - x_i) (\bar{x}_j - x_j) \]
Tibshirani re-formulation

• Originally, measure of agreement formulated as

\[ ps(k) = \min_{1 \leq s \leq k} \frac{1}{n_s(n_s - 1)} \sum_{x_i \neq x_i' \in v_s} I(D[U(X_{te})]_{ii'} = 1) \]

\[ U(X_{te}) = \left\{ \arg\min_r (\text{dist}(\text{medoid}(u_r), x_j \in X_{te})) \right\} \]

\[ D[U]_{ij} = \begin{cases} 1 & x_i, x_j \in u_r \text{ for some cluster } u_r \in U \\ 0 & \text{otherwise} \end{cases} \]

• Note: partitioning around medoid (PAM) clustering used
Tibshirani re-formulation detail

• Since the matrix $D[U]$ is already the indicator matrix, this can be re-written as:

$$ps(k) = \min_{1 \leq s \leq k} \frac{1}{n_s (n_s - 1)} \sum_{s} \sum_{x_i \in v_s; x_i \neq x_j} D[U(X_{te})]_{ii'}$$

• Now we can partition the observations $x_i$ in cluster $v_s$ by the cluster $u_r$ to which they belong:

$$ps(k) = \min_{1 \leq s \leq k} \frac{1}{n_s (n_s - 1)} \sum_{s} \left\{ \sum_{r=1}^{k} \sum_{x_i \in v_s; x_i \neq x_j; x_i \in u_r} D[U(X_{te})]_{ii'} \right\}$$

• We can also divide the observations $x_i$ in cluster $v_s$ by whether or not they belong to cluster $u_r$:

$$ps(k) = \min_{1 \leq s \leq k} \frac{1}{n_s (n_s - 1)} \sum_{s} \left\{ \sum_{r=1}^{k} \sum_{x_i \in v_s; x_i \neq x_j; x_i \in u_r} D[U(X_{te})]_{ii'} + \sum_{x_i \in v_s; x_i \in u_r} D[U(X_{te})]_{ii'} \right\}$$
Tibshirani re-formulation detail (cont.)

• By the definition of $D$, if $x_i \in u_r$ and $x_i' \not\in u_r$, then $D[U(X_{te})]_{ii'}=0$ and if $x_i \in u_r$ and $x_i' \in u_r$, $D[U(X_{te})]_{ii'}=1$. Therefore,

$$ps(k) = \min_{1 \leq s \leq k} \frac{1}{n_s(n_s-1)} \sum_{r=1}^{k} \sum_{x_i \in u_r, x_i \not\in u_r} (1) = \min_{1 \leq s \leq k} \frac{1}{n_s(n_s-1)} \sum_{r=1}^{k} \sum_{i \neq i'} (1)$$

$$= \min_{1 \leq s \leq k} \frac{1}{n_s(n_s-1)} \sum_{r=1}^{k} n_{rs}(n_{rs}-1) = \min_{1 \leq s \leq k} \frac{\sum_{r=1}^{k} \binom{n_{rs}}{2}}{\binom{n_s}{2}}$$

• Thus, $ps(k)$ corresponds to $PS(k,m=2)$:

$$PS(k,m) = \min_{1 \leq s \leq k} p_m(s), \text{ where } p_{m=2}(s) = \frac{\sum_{r=1}^{k} \binom{n_{rs}}{2}}{\binom{n_s}{2}}$$