A Framework for Feature Selection in Clustering

Daniela M. Witten and Robert Tibshirani

October 10, 2011
Outline

- Problem
- Past work
- Proposed (sparse) clustering framework
- Sparse K-means clustering
- Sparse hierarchical clustering
- Applications to simulated and real data
Problem

Clustering

- given data $X_{n \times p}$ identify $K$ “clusters”
- unsupervised learning
Problem

Clustering
- given data $\mathbf{X}_{n \times p}$ identify $K$ “clusters”
- unsupervised learning

Sparse Clustering
- clusters only differ based on $q < p$ features
- benefits:
  - improved accuracy and interpretability
  - cheaper prediction
Problem - Motivating Example

- \( X_1, \ldots, X_{500} \overset{\text{indep.}}{\sim} \) bivariate normal
- clusters differ only in variable 1
Problem - set-up

Measure of dissimilarity

- additive assumption:
  \[ d(x_i, x_{i'}) = \sum_{j=1}^{P} d_{i,i',j} \]

- squared Euclidean distance:
  \[ d_{i,i',j} = (x_{ij} - x_{i'j})^2 \]
approximate $X_{n \times p} \approx A_{n \times q} B_{q \times p}$ and cluster along rows of $A$

elements:

- principal component analysis (PCA); Ghosh and Chinnaiyan (2002), Liu et al. (2003)
- nonnegative matrix factorization; Tamayo et al. (2007)

- not truly “sparse” - function of full $p$
- $A$ not guaranteed to contain signal of interest
- PCA not justified in clustering; Chang (1983)
Past Work - model-based clustering

model rows as independent multivariate observations from a mixture model

\[ LL = \sum_{i=1}^{n} \log \left[ \sum_{k=1}^{K} \pi_k f_k(\mathbf{X}_i; \mu_k, \Sigma_k) \right] \]

- \( K \): number of components
- \( f_k \): density, typically Gaussian

- use EM algorithm to fit model
- problems estimating \( \Sigma_k \) when \( p \approx n \) or \( p \gg n \)
Past Work - model-based clustering

maximize the log-likelihood subject to a sparsity inducing penalty

Pan and Shen (2007):
- assume the features are centered at 0

\[ PLL = \sum_{i=1}^{n} \log \left[ \sum_{k=1}^{K} \pi_k f_k(X_i; \mu_k, \Sigma_k) \right] - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |\mu_{kj}| \]

- sparsity induced when \( \mu_{kj} = 0 \) for all \( k \)
Past Work - model-based clustering

- Raftery and Dean (2006): recast variable selection problem as a model selection problem
- choose between models with nested subsets of variables
Past Work - model-based clustering

Friedman and Meulman (2004): clustering objects on subsets of attributes (COSA)

\[
\begin{align*}
\text{minimize} & \quad \sum_{k=1}^{K} a_k \left( \sum_{i,i' \in C_k} \sum_{j=1}^{p} (w_j d_{i,i',j} + \lambda w_j \log w_j) \right) \\
\text{subject to} & \quad \sum_{j=1}^{p} w_j = 1, \quad w_j \geq 0
\end{align*}
\]

- \( C_k \): indices of observations in cluster \( k \)
- \( a_k \): function of \( |C_k| \)
- \( \mathbf{w} \in \mathbb{R}^p \): feature weights
- \( \lambda \geq 0 \) tuning parameter

above is a simplified version, full COSA allows for \( \mathbf{w}_k \)
Proposed Framework

Clustering

\[
\text{maximize}_{\Theta \in D} \left\{ \sum_{j=1}^{p} f_j(X_j, \Theta) \right\}
\]

- \( X_j \in \mathbb{R}^n \): feature \( j \)
- \( f_j \): function of only the \( j \)th feature
Proposed Framework

Sparse Clustering

\[
\text{maximize} \quad \left\{ \sum_{j=1}^{p} w_j f_j(X_j, \Theta) \right\}
\]

subject to \( \|w\|^2 \leq 1, \|w\|_1 \leq s, w_j \geq 0 \)

\>

1 \leq s \leq \sqrt{p}: \text{tuning parameter}
Proposed Framework

\[
\maximize_{\Theta \in D, w} \left\{ \sum_{j=1}^{p} w_j f_j(X_j, \Theta) \right\}
\]

subject to \( \|w\|_2^2 \leq 1, \|w\|_1 \leq s, w_j \geq 0 \)

**Sparse Clustering:** observations

- \( w_1 = \cdots = w_p \) reduces to usual clustering
- \( L_1 \) penalty induces sparsity
- without \( L_2 \) penalty solution is trivial
- \( w_j \) is the contribution of feature \( j \) to the clustering
- in general, \( f_j(X_j, \Theta) > 0 \) for some or all \( j \)
Proposed Framework

Sparse Clustering: optimization (iterative)

1. holding \( \mathbf{w} \) fixed, optimize w.r.t. \( \Theta \)
   \[ \rightarrow \text{clustering on weighted data} \]

2. holding \( \Theta \) fixed, optimize w.r.t. \( \mathbf{w} \)
   \[ \rightarrow \text{rewrite as: maximize } \left\{ w^T \mathbf{a} \right\} \]
   \[ \rightarrow a_j = f_j(\mathbf{X}_j, \Theta) \]
   \[ \rightarrow \text{proposition: by KKT,} \]

\[
\mathbf{w} = \frac{(\mathbf{a}_+ - \Delta)_+}{\| (\mathbf{a}_+ - \Delta)_+ \|_2}, \quad \Delta = \begin{cases} 
0 & \text{if it results in } \| \mathbf{w} \|_1 \leq s \\
> 0 & \text{otherwise to yield } \| \mathbf{w} \|_1 = s
\end{cases}
\]
K-means Clustering

find $C_1, \ldots, C_K$ to...

- minimize the within-cluster sum of squares (WCSS)

$$\min_{C_1, \ldots, C_K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i, i' \in C_k} \sum_{j=1}^{p} d_{i, i', j}$$

- or maximize the between-cluster sum of squares (BCSS)

$$\max_{C_1, \ldots, C_K} \sum_{j=1}^{p} \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{i' = 1}^{n} d_{i, i', j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i, i' \in C_k} d_{i, i', j} \right)$$

$$\max_{\Theta \in D} \sum_{j=1}^{p} f_j(\mathbf{X}_j, \Theta)$$
Sparse K-means Clustering

\[
\text{maximize} \quad \sum_{j=1}^{p} w_j \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j} \right) \\
\text{subject to} \quad \|w\|^2 \leq 1, \quad \|w\|_1 \leq s, \quad w_j \geq 0
\]

**Sparse K-means Clustering: optimization**

- solve using earlier algorithm
  - (1) K-means using \(d_{i,i',j}^* = w_j d_{i,i',j}\) to find \(C_1, \ldots, C_K\)
  - (2) proposition to find \(w\)
- iterate until \(\frac{\sum_j |w_j^r - w_j^{r-1}|}{\sum_j |w_j^{r-1}|} < 10^{-4}\)
- tends to converge within 5 to 10 iterations
- generally will not converge to global optimum (non-convex)
Sparse K-means Clustering

**Sparse K-means Clustering: tuning** $s$

- objective function increases with $s$
- permutation approach is used:

1. Obtain permuted datasets $X_1, \ldots, X_B$ by independently permute the observations within each feature

2a. Compute $O(s)$, the objective obtained by performing sparse K-means with tuning parameter $s$

2b. Similarly compute $O_b(s)$ for $b = 1, \ldots, B$

2c. Calculate $\text{Gap}(s) = \log(O(s)) - \frac{1}{B} \sum_b \log(O_b(s))$

3. Choose $s^* = \arg\max_s \{\text{Gap}(s)\}$
   - Alternately, choose $s^*$ to equal the smallest value for which $\text{Gap}(s^*)$ is within one standard deviation of $\log(O_b(s^*))$ of the largest value of $\text{Gap}(s)$

**Intuition:** gap statistic measures the strength of the clustering obtained on the real data relative to the clustering obtained on null data that does not contain subgroups
Sparse Hierarchical Clustering

produce a dendrogram that represents a nested set of clusters from dissimilarity matrix $U_{n \times n}$ and cut at desired height

- linkage: between cluster measure
  - complete: $d(C, C') = \max \{ d(x_i, x_i'); x_i \in C, x_i' \in C' \}$
  - average: $d(C, C') = \frac{1}{|C||C'|} \sum_{x_i, x_i'} d(x_i, x_i')$
  - single: $d(C, C') = \min \{ d(x_i, x_i'); x_i \in C, x_i' \in C' \}$
Sparse Hierarchical Clustering

place selection of dissimilarity matrix $U$ in the proposed framework

- dissimilarity matrix $U^* = \{ \sum_j d_{i,j,i'}, j \}_{i,i'}$ results in usual hierarchical clustering
- $U^*$ can be shown to be proportional to the solution of

$$\max_U \left\{ \sum_{j=1}^p \sum_{i,i'} d_{i,i',j} U_{i,i'} \right\}$$

subject to $\sum_{i,i'} U_{i,i'}^2 \leq 1$

- shape of dendrogram is invariant to scaling of features
Sparse Hierarchical Clustering

maximize $\sum_{j=1}^{p} w_j \sum_{i,i'} d_{i,i',j} U_{i,i'}$

subject to $\sum_{i,i'} U_{i,i'}^2 \leq 1$

$||w||^2 \leq 1, ||w||_1 \leq s, w_j \geq 0$

- solution is proportional to $U^{**} = \{ \sum_j w_j d_{i,i',j} \}_{i,i'}$
- clustering by reweighted dissimilarity matrix $U^{**}$ produces sparsity
Spares Hierarchical Clustering

\[
\max_{w, U} \left\{ \sum_{j=1}^{p} w_j \sum_{i, i'} d_{i, i', j} U_{i, i'} \right\}
\]

Sparse Hierarchical Clustering: optimization

- objective is bi-convex in \(U\) and \(w\)
- notation:
  - \(D \in \mathbb{R}^{n^2 \times p}: \{d_{i, i', j}\}_{i, i'}\) strung out into columns
  - \(u \in \mathbb{R}^{n^2}: U\) strung out into a vector

- rewrite:
  \[
  \max_{w, u} \{u^T Dw\} 
  \]
Spares Hierarchical Clustering

\[
\text{maximize } \{ \sum_{j=1}^{p} w_j \sum_{i,i'} d_{i,i',j} U_{i,i'} \} 
\]

Sparse Hierarchical Clustering: optimization

- solve using earlier algorithm
  1. update \( u = \frac{Dw}{||Dw||_2} \)
  2. proposition to find \( w \)
- iterate until \( \frac{\sum_j |w_j^r - w_j^{r-1}|}{\sum_j |w_j^{r-1}|} < 10^{-4} \)
- perform hierarchical clustering using \( U \)
- tends to converge within 10 iterations
- not guaranteed to converge to global optimum
Spares Hierarchical Clustering

Sparse Hierarchical Clustering: a simple underlying model
consider simple model: two classes $C_1, C_2$ only differ with respect to first $q$ features.

\[ X_{ij} \sim \begin{cases} 
\text{N}(\mu_j + c, \sigma^2), & \text{if } j \leq q, \ i \in C_1 \\
\text{N}(\mu_j, \sigma^2), & \text{otherwise}
\end{cases} \]

\[ X_{ij} - X_{i'j} \sim \begin{cases} 
\text{N}(\pm c, 2\sigma^2), & \text{if } j \leq q, \ i, i' \text{ in different classes} \\
\text{N}(0, 2\sigma^2), & \text{otherwise}
\end{cases} \]

\[ d_{i,i',j} = (X_{ij} - X_{i'j})^2, \]
\[ d_{i,i',j} \sim \begin{cases} 
2\sigma^2 \chi_1^2 \left( \frac{c^2}{2\sigma^2} \right), & \text{if } j \leq q, \ i, i' \text{ in different classes} \\
2\sigma^2 \chi_1, & \text{otherwise}
\end{cases} \]
Sparse Hierarchical Clustering: a simple underlying model

- consider overall dissimilarity $\sum_j d_{i,i',j}$,
  \[
  \sum_j d_{i,i',j} \sim \begin{cases} 
  2\sigma^2 \chi_p^2 \left( \frac{qc^2}{2\sigma^2} \right), & \text{if } j \leq q, \ i, i' \text{ in different classes} \\
  2\sigma^2 \chi_p, & \text{otherwise}
  \end{cases}
  \]

- consider weighted dissimilarity $\sum_j w_j d_{i,i',j}$ under ideal situation where $w_j \propto 1_{j \leq q}$,
  \[
  \sum_j w_j d_{i,i',j} \sim \begin{cases} 
  2\sigma^2 \chi_q^2 \left( \frac{qc^2}{2\sigma^2} \right), & j \leq q, \ i, i' \text{ in different classes} \\
  2\sigma^2 \chi_q, & \text{otherwise}
  \end{cases}
  \]

- dissimilarity matrix for sparse clustering is a denoised version of the dissimilarity matrix used for standard clustering
Sparse Hierarchical Clustering: a simple underlying model

- standard clustering,
\[ E\left( \sum_j d_{i,i',j} \right) = \begin{cases} 2\sigma^2 p + c^2 q, & \text{if } j \leq q, \ i, i' \text{ in different classes} \\ 2\sigma^2 p, & \text{otherwise} \end{cases} \]

- sparse clustering,
\[ E\left( \sum_j w_j d_{i,i',j} \right) = \begin{cases} 2\sigma^2 \sum_j w_j + c^2 \sum_{j \leq q} w_j, & j \leq q, \ i, i' \text{ in different classes} \\ 2\sigma^2 \sum_j w_j, & \text{otherwise} \end{cases} \]

- note:
  1. \( w \) taken to be first SPC of \( E(D) \), not \( D \)
  2. \( w_1 = \cdots = w_q > w_{q+1} = \cdots = w_p \)
Spares Hierarchical Clustering

**Sparse Hierarchical Clustering: tuning**

- same as sparse K-means approach: permutation method
Sparse Hierarchical Clustering: complementary clustering

finding a secondary clustering after removing the signal found in primary clustering

- identify $U_1, w_1$
- solve for $U_2, w_2$:

\[
\begin{align*}
\text{maximize } & \{u_2^T Dw_2\} \\
\text{subject to } & \|u_2\|^2 \leq 1, \ u_1^T u_2 = 0, \\
& \|w_2\|^2 \leq 1, \|w_2\|_1 \leq s, \ w_j \geq 0
\end{align*}
\]
Sparse K-means - simulated data
simulation 1: comparison of sparse and standard K-means

- $K = 3$
- $q = 50$
- $X_{ij} \sim N(\mu_{ij}, 1), \mu_{ij} = \mu(1_{i \in C_1,j \leq q} - 1_{i \in C_2,j \leq q})$
- 20 observations per class
- vary values of $p$ and $\mu$

- Classification error rate (CER): $\sum_{i > i'} |1_{P(i,i')} - 1_{Q(i,i')}| / \binom{n}{2}$
Sparse K-means - simulated data

Simulation 1: comparison of sparse and standard K-means

Table 1. Standard 3-means results for Simulation 1. The reported values are the mean (and standard error) of the CER over 20 simulations. The $\mu/p$ combinations for which the CER of standard 3-means is significantly less than that of sparse 3-means (at level $\alpha = 0.05$) are shown in bold.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$p = 50$</th>
<th>$p = 200$</th>
<th>$p = 500$</th>
<th>$p = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.07 (0.01)</td>
<td>0.184 (0.015)</td>
<td>0.22 (0.009)</td>
<td>0.272 (0.006)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.023 (0.005)</td>
<td>0.077 (0.009)</td>
<td>0.16 (0.012)</td>
<td>0.232 (0.01)</td>
</tr>
<tr>
<td>0.8</td>
<td>0.013 (0.004)</td>
<td>0.038 (0.007)</td>
<td>0.08 (0.005)</td>
<td>0.198 (0.01)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.001 (0.001)</td>
<td>0.013 (0.005)</td>
<td>0.048 (0.008)</td>
<td>0.102 (0.013)</td>
</tr>
<tr>
<td>1</td>
<td>0.002 (0.002)</td>
<td>0.004 (0.002)</td>
<td>0.013 (0.004)</td>
<td>0.05 (0.006)</td>
</tr>
</tbody>
</table>

Table 2. Sparse 3-means results for Simulation 1. The reported values are the mean (and standard error) of the CER over 20 simulations. The $\mu/p$ combinations for which the CER of sparse 3-means is significantly less than that of standard 3-means (at level $\alpha = 0.05$) are shown in bold.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$p = 50$</th>
<th>$p = 200$</th>
<th>$p = 500$</th>
<th>$p = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.146 (0.014)</td>
<td>$\mathbf{0.157 (0.016)}$</td>
<td>$\mathbf{0.183 (0.015)}$</td>
<td>0.241 (0.017)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.081 (0.011)</td>
<td>$\mathbf{0.049 (0.008)}$</td>
<td>$\mathbf{0.078 (0.013)}$</td>
<td>$\mathbf{0.098 (0.013)}$</td>
</tr>
<tr>
<td>0.8</td>
<td>0.043 (0.008)</td>
<td>$\mathbf{0.031 (0.007)}$</td>
<td>$\mathbf{0.031 (0.005)}$</td>
<td>$\mathbf{0.037 (0.006)}$</td>
</tr>
<tr>
<td>0.9</td>
<td>0.015 (0.006)</td>
<td>$\mathbf{0.005 (0.003)}$</td>
<td>$\mathbf{0.014 (0.004)}$</td>
<td>$\mathbf{0.014 (0.004)}$</td>
</tr>
<tr>
<td>1</td>
<td>0.009 (0.004)</td>
<td>0.004 (0.002)</td>
<td>$\mathbf{0.001 (0.001)}$</td>
<td>$\mathbf{0.002 (0.002)}$</td>
</tr>
</tbody>
</table>
Sparse K-means - simulated data

Simulation 1: comparison of sparse and standard K-means

Table 3. Sparse 3-means results for Simulation 1. The mean number of nonzero feature weights resulting from the method for tuning parameter selection of Section 3.2 is shown; standard errors are given in parentheses. Note that 50 features differ between the three classes.

<table>
<thead>
<tr>
<th></th>
<th>$p = 50$</th>
<th>$p = 200$</th>
<th>$p = 500$</th>
<th>$p = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = 0.6$</td>
<td>41.35 (0.895)</td>
<td>167.4 (7.147)</td>
<td>243.1 (31.726)</td>
<td>119.45 (41.259)</td>
</tr>
<tr>
<td>$\mu = 0.7$</td>
<td>40.85 (0.642)</td>
<td>195.65 (2.514)</td>
<td>208.85 (19.995)</td>
<td>130.15 (17.007)</td>
</tr>
<tr>
<td>$\mu = 0.8$</td>
<td>38.2 (0.651)</td>
<td>198.85 (0.654)</td>
<td>156.35 (13.491)</td>
<td>106.7 (10.988)</td>
</tr>
<tr>
<td>$\mu = 0.9$</td>
<td>38.7 (0.719)</td>
<td>200 (0)</td>
<td>204.75 (19.96)</td>
<td>83.7 (9.271)</td>
</tr>
<tr>
<td>$\mu = 1$</td>
<td>36.95 (0.478)</td>
<td>200 (0)</td>
<td>222.85 (20.247)</td>
<td>91.65 (14.573)</td>
</tr>
</tbody>
</table>

- fewer features with nonzero weights would result from selecting the tuning parameter at the smallest value that is within one standard deviation of the maximal gap statistic.
Sparse K-means - simulated data
simulation 2: comparison of sparse and other approaches

- $K = 3$
- $X_{ij} \sim N(\mu_{ij}, 1), \mu_{ij} = \mu(1_{i \in C_1} \cdot j \leq q - 1_{i \in C_2} \cdot j \leq q)$
- two scenarios:
  - small simulation: $p = 25, q = 5$, 10 observations per class
  - larger simulation: $p = 500, q = 50$, 20 observations per class
- other methods:
  1. COSA + PAM/heirarchical clustering, Friedman and Meulman (2004)
  2. model-based clustering, Raftery and Dean (2006)
  3. penalized log-likelihood, Pan and Sen (2007)
  4. PCA + 3-means clustering
Sparse K-means - simulated data
simulation 2: comparison of sparse and other approaches

Table 4. Results for Simulation 2. The quantities reported are the mean and standard error (given in parentheses) of the CER, and of the number of nonzero coefficients, over 25 simulated datasets

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Method</th>
<th>CER</th>
<th>Num. nonzero coef.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small simulation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 25$, $q = 5$, 10 obs. per class</td>
<td>Sparse K-means</td>
<td>0.112 (0.019)</td>
<td>8.2 (0.733)</td>
</tr>
<tr>
<td></td>
<td>$K$-means</td>
<td>0.263 (0.011)</td>
<td>25 (0)</td>
</tr>
<tr>
<td></td>
<td>Pan and Shen</td>
<td>0.126 (0.017)</td>
<td>6.72 (0.334)</td>
</tr>
<tr>
<td></td>
<td>COSA w/Hier. Clust.</td>
<td>0.381 (0.016)</td>
<td>25 (0)</td>
</tr>
<tr>
<td></td>
<td>COSA w/K-medoids</td>
<td>0.369 (0.012)</td>
<td>25 (0)</td>
</tr>
<tr>
<td></td>
<td>Raftery and Dean</td>
<td>0.514 (0.031)</td>
<td>22 (0.86)</td>
</tr>
<tr>
<td></td>
<td>PCA w/$K$-means</td>
<td>0.16 (0.012)</td>
<td>25 (0)</td>
</tr>
<tr>
<td>Large simulation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 500$, $q = 50$, 20 obs. per class</td>
<td>Sparse K-means</td>
<td>0.106 (0.019)</td>
<td>141.92 (9.561)</td>
</tr>
<tr>
<td></td>
<td>$K$-means</td>
<td>0.214 (0.011)</td>
<td>500 (0)</td>
</tr>
<tr>
<td></td>
<td>Pan and Shen</td>
<td>0.134 (0.013)</td>
<td>76 (3.821)</td>
</tr>
<tr>
<td></td>
<td>COSA w/Hier. Clust.</td>
<td>0.458 (0.011)</td>
<td>500 (0)</td>
</tr>
<tr>
<td></td>
<td>COSA w/K-medoids</td>
<td>0.427 (0.004)</td>
<td>500 (0)</td>
</tr>
<tr>
<td></td>
<td>PCA w/$K$-means</td>
<td>0.058 (0.006)</td>
<td>500 (0)</td>
</tr>
</tbody>
</table>
Sparse Hierarchical - breast cancer data, Perou et al. (2000)

- 65 surgical specimens of human breast tumors
- 1753 genes, 496 “intrinsic genes”
- hierarchical clustering using “Eigen” linkage
- 4 classes identified, 3 samples unassigned

Figure 4. Using the intrinsic gene set, hierarchical clustering was performed on all 65 observations (left panel) and on only the 62 observations that were assigned to one of the four classes (right panel). Note that the classes identified using all 65 observations are largely lost in the dendrogram obtained using just 62 observations. The four classes are basal-like (dashed gray), Erb-B2 (solid gray), normal breast-like (dashed black), and ER+ (solid black). In the left-hand panel, observations that do not belong to any class are shown in dotted black. The online version of this figure is in color.

Figure 5. Four hierarchical clustering methods were used to cluster the 62 observations that were assigned to one of four classes in Perou et al. (2000). Sparse clustering results in the best separation between the four classes. The classes are indicated as in Figure 4. The online version of this figure is in color.
Sparse Hierarchical - breast cancer data, Perou et al. (2000)

Figure 4. Using the intrinsic gene set, hierarchical clustering was performed on all 65 observations (left panel) and on only the 62 observations that were assigned to one of the four classes (right panel). Note that the classes identified using all 65 observations are largely lost in the dendrogram obtained using just 62 observations. The four classes are basal-like (dashed gray), Erb-B2 (solid gray), normal breast-like (dashed black), and ER\(^+\) (solid black). In the left-hand panel, observations that do not belong to any class are shown in dotted black. The online version of this figure is in color.

The 65 samples were hierarchically clustered using what we will refer to as “Eisen” linkage; this is a centroid-based linkage that is implemented in Michael Eisen’s Cluster program (Eisen et al. 1998). Two sets of genes were used for the clustering: the full set of 1753 genes, and an intrinsic gene set consisting of 496 genes. The intrinsic genes were defined as having the greatest level of variation in expression between different tumors relative to variation in expression between paired samples taken from the same tumor before and after chemotherapy. The dendrogram obtained using the intrinsic gene set was used to identify four classes—basal-like, Erb-B2, normal breast-like, and ER\(^+\)—to which 62 of the 65 samples belong. It was determined that the remaining three observations did not belong to any of the four classes.

Figure 5. Four hierarchical clustering methods were used to cluster the 62 observations that were assigned to one of four classes in Perou et al. (2000). Sparse clustering results in the best separation between the four classes. The classes are indicated as in Figure 4. The online version of this figure is in color.
Figure 6. The gap statistic was used to determine the optimal value of the tuning parameter for sparse hierarchical clustering. *Left:* The largest value of the gap statistic corresponds to 93 genes with nonzero weights. *Right:* The dendrogram corresponding to 93 nonzero weights. The classes are indicated as in Figure 4. The online version of this figure is in color.

---

**6. SNP DATA EXAMPLE**

We wondered whether one could use sparse clustering in order to identify distinct populations in single nucleotide polymorphism (SNP) data, and also to identify the SNPs that differ between the populations. A SNP is a nucleotide position in a DNA sequence at which genetic variability exists in the population. We used the publicly available Haplotype Map ("HapMap") data of the International HapMap Consortium (International HapMap Consortium 2005, 2007). We used the Phase III SNP data for chromosome 22. We restricted the analysis to three of the populations: African ancestry in southwest U.S.A., Utah residents with European ancestry, and Han Chinese from Beijing. We used the SNPs for which measurements
Sparse Hierarchical - breast cancer data, Perou et al. (2000)

Figure 6. The gap statistic was used to determine the optimal value of the tuning parameter for sparse hierarchical clustering. Left: The largest value of the gap statistic corresponds to 93 genes with nonzero weights. Right: The dendrogram corresponding to 93 nonzero weights. The classes are indicated as in Figure 4. The linear version of this figure is in color.

3. Standard hierarchical clustering using the 496 genes with highest marginal variance.
4. COSA hierarchical clustering using all 1753 genes. The resulting dendrograms are shown in Figure 5. Sparse clustering of all 1753 genes with the tuning parameter chosen to yield 496 nonzero genes does best at capturing the four classes; in fact, a comparison with Figure 4 reveals that it does quite abit better than clustering based on the intrinsic genes only!

Figure 6 displays the result of performing the automated tuning parameter selection method. This resulted in 93 genes having nonzero weights. Figure 7 shows that the gene weights obtained using sparse clustering are highly correlated with the marginal variances of the genes. However, the results obtained from sparse clustering are different from the results obtained by simply clustering on the high-variance genes (Figure 5). The reason for this lies in the form of the criterion (15). Though the nonzero $w_j$'s tend to correspond to genes with high marginal variances, sparse clustering does not simply cluster the genes with highest marginal variances. Rather, it weights each gene-wise dissimilarity matrix by a different amount.

We also performed complementary sparse clustering on the full set of 1753 genes, using the method of Section 4.4. Tuning parameters for the initial and complementary sparse clusterings were selected to yield 496 genes with nonzero weights. The complementary sparse clustering dendrogram is shown in Figure 8, along with $w_1$ and $w_2$ (the feature weights for the initial and complementary clusterings). The dendrogram obtained using complementary sparse clustering suggests a previously unknown pattern in the data. Recall that the dendrogram for the initial sparse clustering can be found in Figure 5.

In response to a reviewer's inquiry about the robustness of sparse hierarchical clustering, we repeatedly resampled the 62 observations that belong to one of the four classes and performed sparse hierarchical clustering on the resampled datasets. Asmall jitter was added to other resampled in order to avoid samples having correlation one with each other. We found that for the most part, the resulting clusters accurately reflected the true class labels.

6. SNP DATA EXAMPLE

We wondered whether one could use sparse clustering in order to identify distinct populations in single nucleotide polymorphism (SNP) data, and also to identify the SNPs that differ between the populations. A SNP is a nucleotide position in a DNA sequence at which genetic variability exists in the population. We used the publicly available Haplotype Map ("HapMap") data of the International HapMap Consortium (International HapMap Consortium 2005, 2007). We used the Phase III SNP data for chromosome 22. We restricted the analysis to three of the populations: African ancestry in southwest U.S.A., Utah residents with European ancestry, and Han Chinese from Beijing. We used the SNPs for which measurements...
Sparse Hierarchical - SNP data

- Phase III SNP data from chromosome 22
- 3 populations:
  - African ancestry in southwest U.S.A.
  - Utah residents with European ancestry
  - Han Chinese from Beijing
- $n = 315, \ p = 17,026$
Sparse Hierarchical - SNP data

Figure 8. Complementary sparse clustering was performed. Tuning parameters for the initial and complementary clusterings were selected to yield 496 genes with nonzero weights.

Left: A plot of \(w_1\) against \(w_2\).

Right: The dendrogram for complementary sparse clustering. The classes are indicated as in Figure 4. The online version of this figure is in color.

The resulting data have dimension \(315 \times 17,026\). We coded \(AA\) as 2, \(Aa\) as 1, and \(aa\) as 0. Missing values were imputed using 5-nearest neighbors (Troyanskaya et al. 2001). Sparse and standard 3-means clustering were performed on the data. The CERs obtained using standard 3-means and sparse 3-means are shown in Figure 9.

CER was computed by comparing the clustering class labels to the true population identity for each sample. When the tuning parameter in sparse clustering was chosen to yield between 198 and 2809 SNPs with nonzero weights, sparse clustering resulted in slightly lower CER than standard 3-means clustering.

The main improvement of sparse clustering over standard clustering is in interpretability, since the nonzero elements of \(w\) determine the SNPs involved in the sparse clustering. We can use the weights obtained from sparse clustering to identify SNPs on chromosome 22 that distinguish between the populations (Figure 9). SNPs in a few genomic regions appear to be responsible for the clustering obtained.

In this example, the tuning parameter selection method of Section 3.2 does not perform well. Rather than selecting a tuning parameter that yields between 198 and 2809 SNPs with nonzero weights (resulting in the lowest CER), the highest gap statistic is obtained when all SNPs are used. The one standard deviation rule described in Section 3.2 results in a tuning parameter that yields 7160 genes with nonzero weights. The fact that the gap statistic seemingly overestimates the number of features with nonzero weights may reflect the need for a more accurate method for tuning parameter selection, or it may suggest

Figure 9. Left: The gap statistics obtained as a function of the number of SNPs with nonzero weights. Center: The CERs obtained using sparse and standard 3-means clustering, for a range of values of the tuning parameter. Right: Sparse clustering was performed using the tuning parameter that yields 198 nonzero SNPs (this is the smallest number of SNPs that resulted in minimal CER in the center panel). Chromosome 22 was split into 500 segments of equal length. The average weights of the SNPs in each segment are shown, as a function of the nucleotide position of the segments.