Problem Statement

Let X denote an $n \times p$ data matrix, with n observations and p features. Suppose that we wish to cluster the observations and suspect that the true underlying clusters differ only with respect to some of features.

The problem here is how to find these *distinguished* features. This paper propose a method for *sparse clustering*, which could allow us to cluster the observations using only an adaptively chosen subset of the features. This method is most useful when $p \gg n$.

There are a number of advantages for this method.

1. It could result in more accurate identification of these groups using these true effective features.
2. It could yields interpretable results, as one can tell which feature is more important than others.
3. Fewer features are needed to assign a new feature to a preexisting cluster.

Before go into the formal definition, we look at a motivation example. I generated 600 independent observations from a bivariate normal distribution. A mean shift on the first feature defines two clusters. As we can see in the figure. The standard K-means will cluster the data based on the two features, thus, generates the wrong clusters. (Note in some cases, it will get the right clusters for some *lucky* initial cluster centers. The sparse version of K-means will correctly capture the right feature to do the clustering and can always have the right clusters.

The Sparse Clustering Framework

Recall that X is a $n \times p$ matrix with n observation and each observation has p features. Let $X_j \in \mathbb{R}^n$.

Many clustering methods can be expressed as an optimization of the form

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

where $f_j(X_j, \Theta$ is some function that involves only the jth feature of the data, and Θ is a parameter restricted to lie in a set D.

We define *sparse clustering* as the solution to the following problem

$$\text{maximize}_{w,\Theta \in D} \{ \sum_{j=1}^{p} w_j f_j(X_j, \Theta) \}$$
$$\text{subject to } \|w\|^2 \leq 1, \quad \|w\|_1 \leq s, \quad w_j \geq 0 \forall j$$

where w_j is a weight corresponding to feature j and s is tuning parameter, $1 \leq s \leq \sqrt{s}$,

Some observations about this formulation. 1. If $w_1 = w_2 = \ldots = w_n$, this formula reduces to the standard clustering problem. 2. The L_1, or *lasso*, penalty on w results in sparsity for small value of the tuning parameter s, i.e. some of w_j will be zero. The L_2 penalty also serves an important role, since without it, at most one element of w would be nonzero in general. 3. The value of w_j can be interpreted as the contribution of feature j to the resulting sparse clustering.

We can optimize the problem using an iterative algorithm:

- Holding w fixed, optimize with respect to Θ.
- Holding Θ fixed, optimize with respect to w.

The first step is a standard clustering procedure with a weighted version of the data. For the second step, we can rewrite the formula as

$$\text{maximize}_w \{ w^T a \}$$
$$\text{subject to } \|w\|^2 \leq 1, \quad \|w\|_1 \leq s, \quad w_j \geq 0$$

where $a_j = f_j(X_j, \Theta)$. Introducing Lagrangian multipliers λ_1 for $\|w\|^2 - 1 \leq 0$, λ_2 for $\|w\|_1 - s \leq 0$ and μ_j for $w_j \geq 0$. We obtain the KKT condition:

$$\begin{aligned}
w^T w - 1 &\leq 0 \\
\sum w_j - s &\leq 0 \\
\mu_j w_j &\leq 0 \\
w_j &\geq 0 \\
\lambda_1 (w^T w - 1) &= 0 \\
\lambda_2 (\sum w_j - s) &= 0 \\
\mu_j w_j &= 0 \\
-a + \lambda_1 w_j + \lambda_2 - \mu_j &= 0
\end{aligned}$$

First, note that from the last equation, $\mu_j = -a + \lambda_1 w_j + \lambda_2$. Thus, we have $(-a + \lambda_2 + \lambda_1 w_j) w_j = 0$.

- Case 1: $\sum w_j - s < 0$. Thus, $\lambda_2 = 0$. If $a < 0$, then $w_j = 0$. Otherwise, $w_j = \frac{a}{\lambda_1}$, where λ_1 make sure that $w^T w = 1$.
- Case 2: $\sum w_j - s = 0$. Thus, $\lambda_2 \geq 0$. If $a < 0$, then $w_j = 0$. Otherwise, $w_j = \frac{a - \lambda_2}{\lambda_1}$, where λ_1 ensures

that $w^T w = 1$ and λ_2 ensures $\sum w_j - s = 0$.

In summary, we have $w = \frac{S(a\_+,\Delta)}{\|S(a\_+,\Delta)\|^2}$, where $S(x, c) = \text{sign}(x)(|x| - c)\_+$. $\Delta = 0$ if that results in $\|w\|_1 \leq s$, otherwise, $\Delta > 0$ is chosen to yield $\|w\|_1 = s$.

Sparse K-Means Clustering

The Sparse K-Means Method

K-means clustering minimize the *within-cluster sum of squares*(WCSS), i.e. it seeks to partition the n observation into K sets, such that WCSS

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

is minimal, where n_k is the number of observations in cluster k and C_k contains the indices of the observations in cluster k. d_i, i', j denote the dissimilarity measure between observations i and i' along feature j. In this paper, they take $d\_i, i', j = (X\_ij - X\_i'j)^2$.

Defin the *between-cluster sum of squares*(BCSS) as

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

Note that BCSS can be rewrite as

$$\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{j=1}^{p} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j}$$

For a given dataset, the first part is constant and the second part equal to WCSS. So, minimize WCSS is equivalent to maximize BCSS.

The sparse K-means clustering criterion is as follows:

$$\text{maximize}_{C_1,\ldots,C_K,w} \sum_{j=1}^{p} w_j (\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j})$$
$$\text{subject to } \|w\|^2 \leq 1, \quad \|w\|_1 \leq s, \quad w_j \geq 0$$

This criterion assigns a weight to each feature based on the increases in BCSS that the feature can contribute.

Similar as the general sparse clustering algorithm, an iterative algorithm for this is presented.

Algorithm for sparse K-Means clustering

1. Initialize w as $w_i = \frac{1}{\sqrt{p}}$, where $i = 1, \ldots, p$.
2. Iterative until convergence $\frac{\sum_{j=1}^{p} |w_j^r - w_j^{r-1}|}{\sum_j = 1^p |w_j^{r-1}|} < 10^{-4}$.

    1. Holding w fixed, optimize with respect to $C_1, \ldots, C_K$. That is

    $$\text{maximize}_{C_1,\ldots,C_K} \{\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} w_j d_{i,i',j}\}$$

    by applying the standard K-means algorithm to the $n \times n$ dissimilarity matrix with (i, i') element $\sum_j w_j d_{i,i',j}$.
    2. Holding $C_1, \ldots, C_K$ fixed, optimize with respect to w by applying

    $$w = \frac{S(a_+, \Delta)}{\|S(a_+, \Delta)\|^2}$$

    where

    $$a_j = \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j}$$

    and $\Delta = 0$ if $\|w\|_1 < s$, otherwise $\Delta > 0$ is chosen so that $\|w\|_1 = s$.

3. The clusters are given by $C_1, \ldots, C_K$ and the feature weights corresponding to this clustering are given by $w_1, \ldots, w_p$.

Note that we cannot find the global minimum using the above algorithm. As the problem is non-convex and step 2.1 is not guaranteed to find a global minimum.

Selection of Tuning Parameter for Sparse K-Means

Given K, this algorithm has one tuning parameter s. Note that we cannot simply select s to maximize the objective function, since as s is increased, the objective will increase as well. Instead, they apply a permutation approach that is related to the gap statistics of Tibshirani, Walther and Hastie for selecting the number of clusters K in standard K-means clustering.

Algorithm to select tuning parameter s for sparse K-means

1. Obtain permuted datasets $X_1, \ldots, X_B$ by independently permuting the observations within each feature.
2. For each candidate tuning parameter value s:

    1. Compute

$$O(s) = \sum_j w_j \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i',j}\right)$$

    the objective obtained by performing sparse K-means with tuning parameter s on the data X.
    2. For $b = 1, \ldots, B$, compute $O_b(s)$, the objective obtained by performing sparse K-means with tuning parameter value s on the data X_b.
    3. Calculate $gap(s) = \log\left(O(s)\right) - \frac{1}{B}\sum_{b=1}^{B}\log\left(O_b(s)\right)$.

3. Choose s^* corresponding to the largest value of $gap(s)$. Alternately, choose s^* to equal the smallest value for which $gap(s^*)$ is within a standard deviation of $\log(O_b(s^*))$ of the largest value of $gap(s)$.

Note that while there may be strong correlations between the features in the original data X, the features in the permuted datasets $X_1, \ldots, X_B$ are uncorrelated with each other. The gap statistics measures the strength of the clustering obtained on the real data relative to the clustering obtained on null data that dose not contain subgroups.

Sparse Hierarchical Clustering

The Sparse Hierarchical Clustering Method

Hierarchical clustering produces a dendrogram that represents a nested set of clusters: depending on where the dendrogram is cut, between 1 and n clusters can result.

Hierarchical clustering takes as input a $n \times n$ dissimilarity matrix U. If U is the overall dissimilarity matrix $\{\sum_j d\_i, i', j\}\_i, i'$, then the standard hierarchical clustering results. Since scaling the dissimilarity matrix by a factor does not affect the shape of the resulting dendrogram, we ignore proportionality constants in the following discussion. Consider the criterion:

$$\begin{aligned}&\text{maximize}_U \left\{\sum_j \sum_{i,i'} d_{i,i',j} U_{i,i'}\right\}\\ &\text{subject to } \sum_{i,i'} U_{i,i'}^2 \leq 1\end{aligned}$$

Let U^* optimize the above equation. It is easy to see that $U\_i, i' \propto \sum_j d\_i, i'j$. and so performing hierarchical clustering on U^* results in the standard hierarchical clustering.

In order to obtain the sparsity in the features, we consider the following problem:

$$\begin{aligned} &\text{maximize}_{w,U} \{ \sum_j w_j \sum_{i,i'} d_{i,i',j} U_{i,i'} \} \\ &\text{subject to } \sum_{i,i'} U_{i,i'}^2 \leq 1, \ \ \|w\|^2 \leq 1, \ \ \|w\|_1 \leq s, \ \ w_j \geq 0 \end{aligned}$$

Let U^{**} optimize the above equation. Thus, $U^{**}\_i, i' \propto \sum_j w_j d\_i, i'$. Since w is sparse for small values of parameter s, U^{**} involves only a subset of the features, thus results in sparse hierarchical clustering.

Note, this problem is *bi-convex* in U and w: with w fixed, it is convex in U, and with U fixed, it is convex in w.

Let $D \in \mathbb{R}^{n^2 \times p}$ be the matrix in which column j consists of the elements $\{d\_i, i', j\}\_i, i'$, strung out into a vector. Then, $\sum_j w_j \sum \_i, i' d\_i, i', jU\_i, i' = u^T Dw$ where $u \in \mathbb{R}^{n^2}$ is obtained by stringing out U into a vector. Thus, we can rewrite the problem as

$$\begin{aligned} &\text{maximize}_{w,u} \{u^T Dw\} \\ &\quad \text{subject to } \|u\|^2 \leq 1, \ \ \|w\|^2 \leq 1, \ \ \|w\|_1 \leq s, \ \ w_j \geq 0 \end{aligned}$$

Algorithm for sparse hierarchical clustering

1. Initialize w as $w_1 = \ldots = w_p = \frac{1}{\sqrt{p}}$.
2. Iterate until convergence:
    1. Update $u = \frac{Dw}{\|Dw\|_2}$.
    2. Update $w = \frac{S(a\_+, \Delta)}{\|S(a\_+, \Delta)\|_2}$ where $a = D^T u$ and $\Delta = 0$ if this results in $\|w\|_1 \leq s$; otherwise, $\Delta > 0$ is chosen such that $\|w\|_1 = s$.
3. Rewrite u as a $n \times n$ matrix U.
4. Perform hierarchical clustering on the $n \times n$ dissimilarity matrix U.

Note the problem is the *sparse principal components*(SPC) criterion. The first SPC of the $n^2 \times p$ matrix D is denoted as w. Then $u \propto Dw$ can be rewritten as a $n \times n$ matrix U which is a weighted linear combination of the features wise dissimilarity matrices. When s is small, then some of the w_j will be zero, and so U depends on a subset of the features. We then perform hierarchical clustering on U in order to obtain a dendrogram that is based only on an adaptively chosen subset of features.

A simple underlying model for Sparse Hierarchical Clustering

Suppose that the n observations fall into two clusters, C_1 and C_2, which differ only with respect to the first q features. The elements X_{ij} are independent and normally distributed with a mean shift between the two

classes in the first q features:

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

Note that for $i \neq i'$,

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

Let $d\_i, i', j = (X\_ij - X\_i'j)^2$, that is, the dissimilarity measure is squared Euclidean distance. Then, for $i \neq i'$,

$$d_{i,i',j} \sim \begin{cases} 2\sigma^2 \chi_1^2(\frac{c^2}{2\sigma^2}) & \text{if } j \leq q, i \text{ and } i' \text{ in different classes} \\ 2\sigma^2 \chi_1^2 & \text{otherwise} \end{cases}$$

where $\chi_1^2(\lambda)$ denotes the noncentral χ_1^2 distribution with noncentrality parameter λ. Thus, the overall dissimilarity matrix used by standard hierarchical clustering has off-diagonal elements

$$\sum_j d_{i,i',j} \sim \begin{cases} 2\sigma^2 \chi_p^2(\frac{qc^2}{2\sigma^2}) & \text{if } i, i' \text{ in different classes} \\ 2\sigma^2 \chi_p^2 & \text{otherwise} \end{cases}$$

and so for $i \neq i'$,

$$E(d_{i,i',j}) = \begin{cases} 2\sigma^2 + c^2 & \text{if } j \leq q, i \text{ and } i' \text{ in different classes} \\ 2\sigma^2 & \text{otherwise} \end{cases}$$

and

$$E(\sum_j d_{i,i',j}) = \begin{cases} 2p\sigma^2 + qc^2 & \text{if } i \text{ and } i' \text{ in different classes} \\ 2p\sigma^2 & \text{otherwise} \end{cases}$$

Now consider the sparse hierarchical clustering. Suppose $w_j \propto 1_{j \leq q}$. this is corresponding to the ideal situation where the importatn features have equal nonzero weights and the unimportant features have zero weights. Thus, we have

$$\sum_j w_j d_{i,i',j}) \propto \begin{cases} 2\sigma^2 \chi_q^2(\frac{qc^2}{2\sigma^2}) & \text{if } i, i' \text{ in different classes} \\ 2\sigma^2 \chi_q^2 & \text{otherwise} \end{cases}$$

So in the ideal setting, the dissimilarity matrix used for sparse hierarchical clustering is a denoised version of the dissimilarity matrix used for standard hierarchical clustering.

In real case, w is the first SPC of D. To simplify the discussion, suppose w is the first SPC of $E(D)$, rather than D. Thus,

$$w_1 = \ldots = w_q > w_{q+1} = \ldots = w_p$$

This is because initially $w_1 = \ldots = w_p$ and in each iteration of the algorithm, the above equation always holds. Finally, the expectations of the off-diagonal elements of the dissimilarity matrix used for sparse hierarchical clustering is

$$E(\sum_j w_j d_{i,i',j}) = \sum_j w_j E(d_{i,i',j}) = \begin{cases} 2\sigma^2 \sum_j w_j + c^2 \sum_{j \le q} w_j & \text{if } i \text{ and } i' \text{ in different classes} \\ 2\sigma^2 \sum_j w_j & \text{otherwise} \end{cases}$$

We can see that the expected dissimilarity between observations in different classes relative to observations in the same class is greater for sparse hierarchical clustering than for standard hierarchical clustering.

Selection of Tuning parameter for Sparse Hierarchical Clustering

Let $O(s) = \sum_j w_j \sum \_i, i' d\_i, i', jU\_i, i'$. Then the remaining approach is similar as the sparse K-means clustering.