Clustering

COLUMBIA UNIVERSITY
MAILMAN SCHOOL of PUBLIC HEALTH
Environmental Mixtures Workshop

08.23.2018

Outline

1. Introduction
2. K-means
3. Hierarchical clustering
4. Notes

What is clustering?

- Clustering → set of *unsupervised* techniques to *identify subgroups*
- Aka `clusters`

What is clustering?

- Clustering → set of *unsupervised* techniques to *identify subgroups*
- Aka `clusters`

Aim: partition observations in a dataset into homogeneous distinct groups, so that:

- Observations within group are similar to each other
- Observations in different groups are different from each other

What is clustering of environmental exposures?

- Suppose we have N study participants and exposure information on p chemicals
- We want to identify k distinct subgroups (`clusters`) of participants with distinct exposure experiences. E.g.
  - Cluster 1 includes M_1 participants with high exposures to one subset of the p chemicals (q_1) and low exposures to a different subset of p (q_2)
  - Cluster 2 includes M_2 participants with low exposures to q_1 and high exposures q_2
  - Etc...
  - Please note that the classification into *high* and *low* exposures is not necessary

An alternative formulation

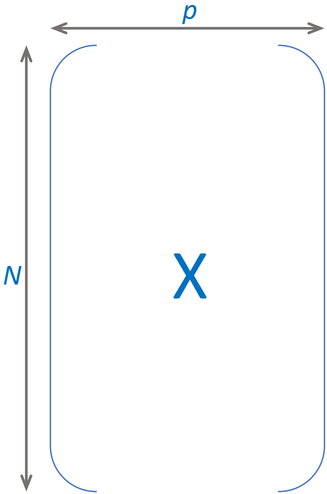
- Suppose we have N study participants and exposure information on p chemicals
- We want to identify ℓ distinct clusters of chemicals with distinct participant contributions. E.g.
  - Cluster 1 includes Q_1 chemicals to which one subset of the N participants has high exposure (m_1) and a different subset of N (m_2) has low exposure
  - Cluster 2 includes Q_2 chemicals to which m_1 participants have low exposures and m_2 participants has high exposure

An alternative formulation

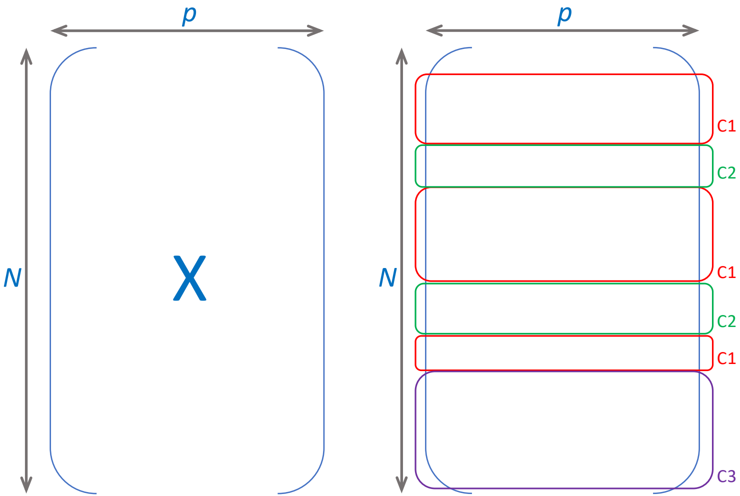
- Suppose we have N study participants and exposure information on p chemicals
- We want to identify ℓ distinct clusters of chemicals with distinct participant contributions. E.g.
  - Cluster 1 includes Q_1 chemicals to which one subset of the N participants has high exposure (m_1) and a different subset of N (m_2) has low exposure
  - Cluster 2 includes Q_2 chemicals to which m_1 participants have low exposures and m_2 participants has high exposure

- For simplicity, we will discuss clustering observations (N) on the basis of chemicals (p)
- The alternative could be performed simply by transposing the data matrix X

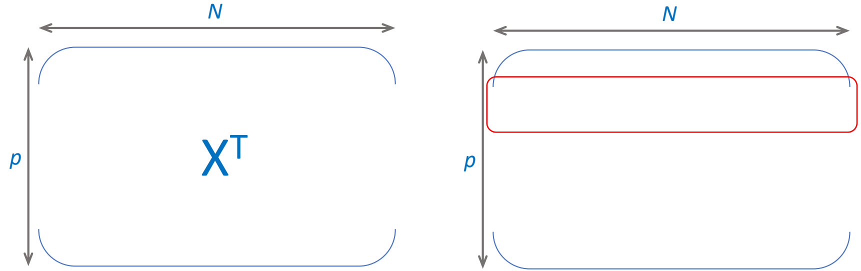
I.e.,



I.e.,



Vs.



Clustering Methods

- There are numerous methods for clustering
- Here we will focus on the two most popular ones

1. K-means
   - Partition N into k (pre-specified) number of clusters

Clustering Methods

- There are numerous methods for clustering
- Here we will focus on the two most popular ones

1. K-means
   - Partition N into k (pre-specified) number of clusters
2. Hierarchical clustering
   - Tree-like visual representation of all possible clusters $(1 \cdots n)$

K-means Clustering

- Iterative algorithm
- Requires pre-specification of desired number of clusters k
- Partitions each data point into k *distinct* and *non-overlapping* clusters

K-means Clustering

Aim: Minimize the within-cluster variation

$$\min_{C_1,\cdots,C_k}\left\{\sum_{k=1}^{K} W(C_k)\right\}$$

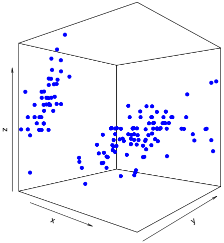
- I.e. the amount by which the observations in a cluster differ from each other

Define Within-Cluster Variation, $W(C_k)$

First, we need to define what metric of distance to use

- By far the most commonly used: Euclidean Distance
- E.g. in a 3-D space ($p = 3$), the ED between points 1 and 2 is

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$



Define Within-Cluster Variation, $W(C_k)$ (cont'd)

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

- i.e. $W(C_k)$ is the sum of all pairwise squared ED between the observations in cluster k, divided by the total number of observations in cluster k, $|C_k|$

Please note that $\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \mu_{kj})^2$, where μ_{kj} is the mean of pollutant j in cluster C_k

K-means Clustering

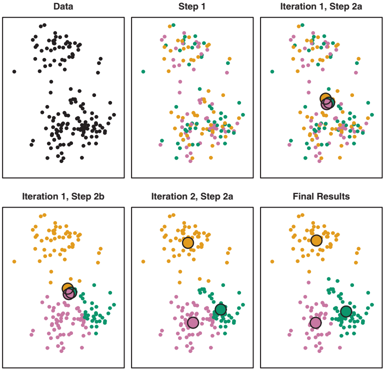
Putting all together:

Aim: Minimize the within-cluster variation

$$\min_{C_1,\cdots,C_k}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right\}$$

K-means Algorithm

1. Randomly assign a number $(1 \cdots K)$ to each of the observations → initial cluster assignments
2. Iterate until cluster assignments stop changing:
   (a) For each of the K clusters compute the cluster centroid: a vector of p means for the observations in the kth cluster
   (b) Assign each observation to each closest cluster centroid
      - *Closest* – defined using ED



Please note that the algorithm presented in ISLR (above) is not the same as the most commonly used Lloyd's algorithm, but they are equivalent

Keep in Mind

1. *Local* minimum instead of global
   - Results will depend on initial cluster assignments
   - May be unstable

   ⇒ Need to run multiple times with different initial assignments
   - Select the best run, i.e. the one that yields the minimal

$$\sum_{k=1}^{K} W(C_k)$$

Keep in Mind

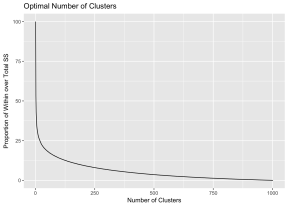
1. *Local* minimum instead of global
   - ○ Results will depend on initial cluster assignments
   - ○ May be unstable
   - ⇒ Need to run multiple times with different initial assignments
   - ○ Select the best run, i.e. the one that yields the minimal

$$\sum_{k=1}^{K} W(C_k)$$

2. *Iterative algorithm*
   - ○ Default number of iterations in R is 10
   - ○ Might need to increase if local minimum not reached within 10 iterations (i.e. if the results have not stopped changing)

How to Pick k?

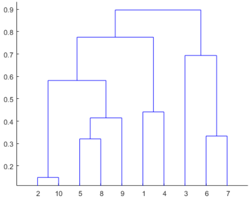
- Unfortunately, no “golden rule”
- One potential way: plot the $\sum_{k=1}^{K} W(C_k)$ against the # of clusters to identify the k above which we stop having benefit (aka find the “elbow”)



- Try several different options
- The importance of expert knowledge – aim for interpretability of the solution!

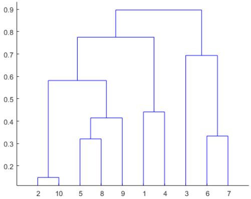
Hierarchical Clustering

- An alternative clustering approach to K-means
- Does not require pre-specification of *k* by the user
- Results in a *dendrogram*, i.e. an upside-down tree-based representation of the clustered observations
- Bottom-up or agglomerative clustering
  - The dendrogram is built starting from the leaves and combining clusters up to the trunk



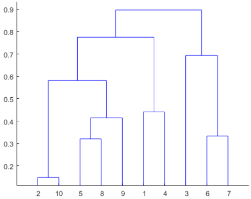
What is a Dendrogram?

- Each observation is represented by a leaf, i.e. start with N leaves at the bottom
- As we move up the tree, the leaves fuse into branches
- At the top, all branches fuse into the trunk



What is a Dendrogram?

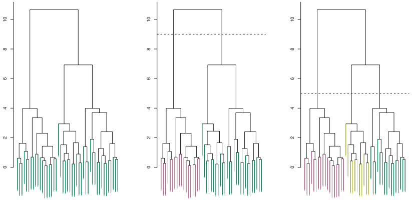
- ○ Each observation is represented by a leaf, i.e. start with N leaves at the bottom
- ○ As we move up the tree, the leaves fuse into branches
- ○ At the top, all branches fuse into the trunk



- ○ Early fusions (lower): more similar the (groups of) observations to each other
- ○ Later fusions (higher): observations can be quite different

Number of Clusters

- A single tree can be used to obtain different numbers of clusters
- *Hierarchical*: clusters obtained by cutting the tree at a given height are *nested* within the clusters obtained by cutting the tree at any greater height



Hierarchical Clustering Algorithm

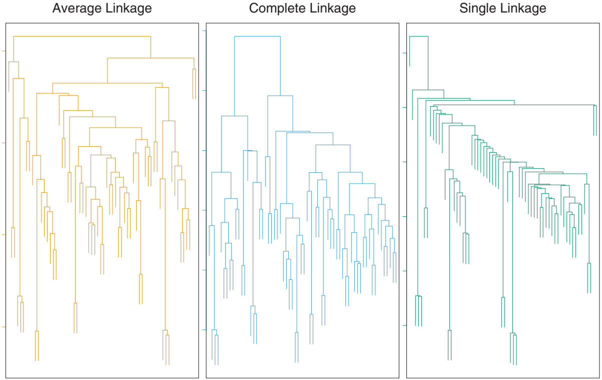
1. Begin with N observations and estimate the ED distance of all the $\binom{N}{2} = N(N-1)/2$ pairwise dissimilarities
   - Each observation is its own cluster
2. For $i = n, n-1, \cdots, 2$:
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of the clusters that are the *least* dissimilar (i.e. most similar). Fuse these two clusters.
      - The dissimilarity between these two clusters indicates the height in the tree at which the fusion is placed

   (b) Compute the new pairwise dissimilarities among the $i-1$ remaining clusters
      - Linkage: Dissimilarity between two groups of observations

Linkage

1. Average
   - Mean inter-cluster dissimilarity
   - Computes all pairwise dissimilarities between observations in Cluster A and in Cluster B and records the *average* of the two
   - Robust against noise
2. Complete
   - Max inter-cluster dissimilarity
   - Computes all pairwise dissimilarities between observations in Cluster A and in Cluster B and records the *largest*
   - Compact clusters
3. Single
   - Min inter-cluster dissimilarity
   - Computes all pairwise dissimilarities between observations in Cluster A and in Cluster B and records the *smallest*
   - Results in extended, trailing clusters – single observations are fused one at a time

Linkage (cont'd)



- Average and complete linkage tend to yield more balanced clusters

Keep in Mind

1. What type of linkage should be used?
   - Results can vary a lot based on this

2. Where should we cut the dendrogram to obtain the clusters?

Final thoughts

- Should we scale the observations? E.g. standardize
    - What if some pollutants we have are in ng/m^3 and others in mg/m^3? or even in μg/g?
    - This can impact the solution
- Clustering should be used as an exploratory tool
- Try several different options
- Aim for interpretable solutions

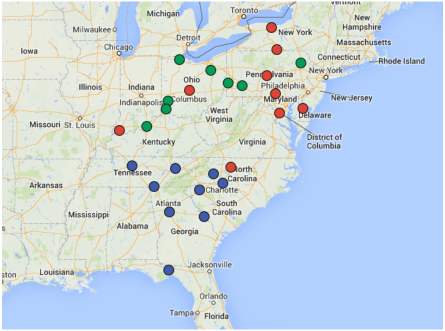
Clusters in Health Models

- Clustering is unsupervised
- The resulting clusters could be included in a health model as a categorical variable
- Or could be used to assess effect modification in the association between one of the mixture members and the outcome of interest
- Or could be used to identify population subgroups e.g. based on neighborhood-level factors – and subsequently assess as modifiers in the exposure–outcome relationship

An Example: Cluster-Specific $PM_{2.5}$ Effects

One way to assess long-term exposure to pollution mixtures:

- Is the effect of $PM_{2.5}$ the same across locations with compositional differences in $PM_{2.5}$?
  - Group cities together
  - Given common pollution profiles



An Example: Cluster-Specific $PM_{2.5}$ Effects

One way to assess long-term exposure to pollution mixtures:

- Is the effect of $PM_{2.5}$ the same across locations with compositional differences in $PM_{2.5}$?
  - Group cities together
  - Given common pollution profiles

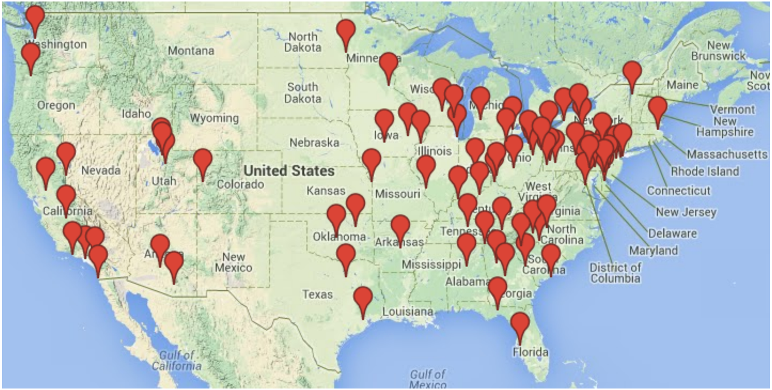
Study Goal:

Is the long-term $PM_{2.5}$-mortality association modified by $PM_{2.5}$ composition?

Cluster-Specific PM$_{2.5}$ Effects

Data Collection & Methods

- ○ 81 cities across the US, 2000 – 2010
- ○ Medicare enrollees ($\geq$ 65 years)
- ○ All-cause mortality



Cluster-Specific $PM_{2.5}$ Effects

Data Collection & Methods

$PM_{2.5}$ and speciated $PM_{2.5}$ data from EPA's AQS

- 24 $PM_{2.5}$ species: (NO_3^-, Na^+, K^+, SO_4^{2-}, NH_4^+, EC, OC, S, Cu, Fe, Zn, Ni, V, Ti, Mg, K, Si, Na, Cl, Ca, Br, Sr, Pb, Mn)

Group together cities with similar pollution profiles:

- K-means clustering
  - Assign cities to clusters based on $PM_{2.5}$ species concentrations

Cluster-Specific $PM_{2.5}$ Effects

Data Collection & Methods

$PM_{2.5}$ and speciated $PM_{2.5}$ data from EPA's AQS

- 24 $PM_{2.5}$ species: (NO_3^-, Na^+, K^+, SO_4^{2-}, NH_4^+, EC, OC, S, Cu, Fe, Zn, Ni, V, Ti, Mg, K, Si, Na, Cl, Ca, Br, Sr, Pb, Mn)

Group together cities with similar pollution profiles:

- K-means clustering
  - Assign cities to clusters based on $PM_{2.5}$ species concentrations

Health Models:

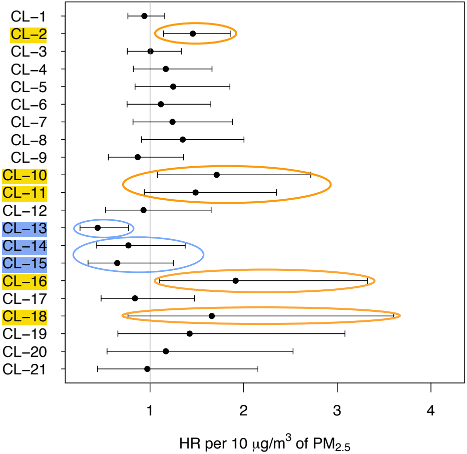
- City-specific Cox models
- In the second stage random effects meta-analysis
  - Indicators for cluster membership

  → Cluster-specific effects

Cluster-Specific $PM_{2.5}$ Effects

Results

- Followed >19M, observed >6M deaths
- HR = 1.11 (1.01, 1.23) per 10 $\mu g/m^3$ annual $PM_{2.5}$



- Cl-13: Southwest – oceanic
- Cl-14 & 15: Rocky Mounts – crustal

- Cl-2: Southeast – regional
- Cl-10: Harbors (South & Southeast) – residual oil combustion and regional
- Cl-11: Industrial Midwest – metals & regional
- Cl-16: Harbors (Northwest) – residual oil combustion
- Birmingham, AL (CL-18): traffic & metals

Thank you!

Questions?

mk3961@cumc.columbia.edu

Lloyd's Algorithm

Given an initial set of k means $m_1^{(1)}, \cdots, m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

1. Assignment Step:
   Assign each observation to the cluster whose mean has the least squared Euclidean distance, i.e. the "nearest" mean
2. Update Step:
   Calculate the new means to be the centroids of the observations in the new clusters