K-means and Hierarchical Clustering

Schwartz

September 26, 2016

Best. Music. *EVAR*

1. Elliott Smith
2. Iron and Wine
3. Damien Rice
4. Sufjan Stevens

1. Kishi Bashi
2. Beirut
3. Christine and the Queens
4. Lyle Lovett

1. Our Lady Peace
2. Oasis
3. Eve Six
4. Better than Ezra

1. Die Antwoord
2. Dan le Sac Vs Scroobius Pip
3. FuntCase & Ry Legit
4. Emalkay

1. Rage Against the Machine
2. System of a Down
3. Smashing Pumpkins
4. Jimmy Eat World

1. Beck
2. Cake
3. Beastie Boys
4. Smash Mouth

Objectives

- *Supervised* versus *Unsupervised*

Objectives

- *Supervised* versus *Unsupervised*
- K-means

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality
- K

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality
- K
  - Elbow, Silhouette, and Gap methods

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality
- K
  - Elbow, Silhouette, and Gap methods
- Hierarchical clustering

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality
- K
  - Elbow, Silhouette, and Gap methods
- Hierarchical clustering
- DBSCAN

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality
- K
  - Elbow, Silhouette, and Gap methods
- Hierarchical clustering
- DBSCAN
- Bayesian mixture models

Objectives

- *Supervised* versus *Unsupervised*
- K-means
  - and the curse of dimensionality
- K
  - Elbow, Silhouette, and Gap methods
- Hierarchical clustering
- DBSCAN
- Bayesian mixture models
- Expectation-Maximization (EM) algorithm

ML Cosmology

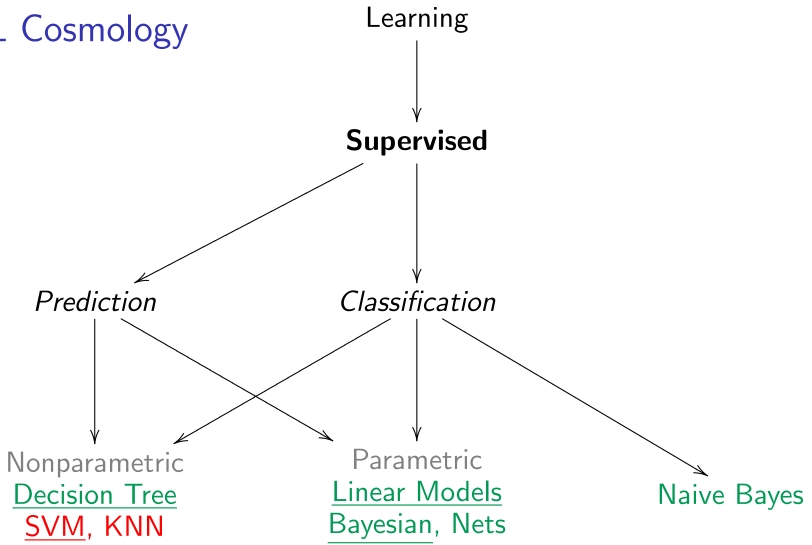
Learning

Supervised

Prediction

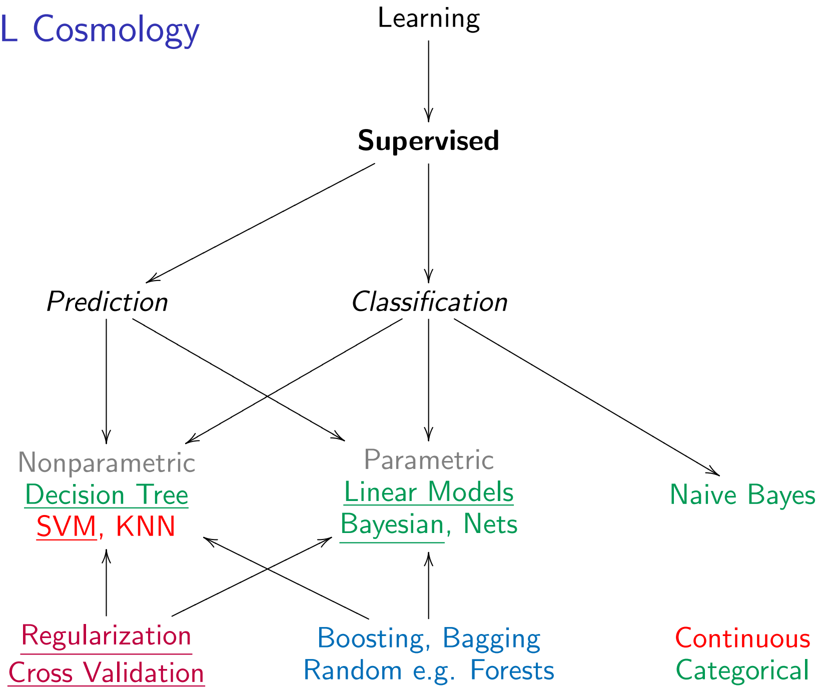
Classification

ML Cosmology

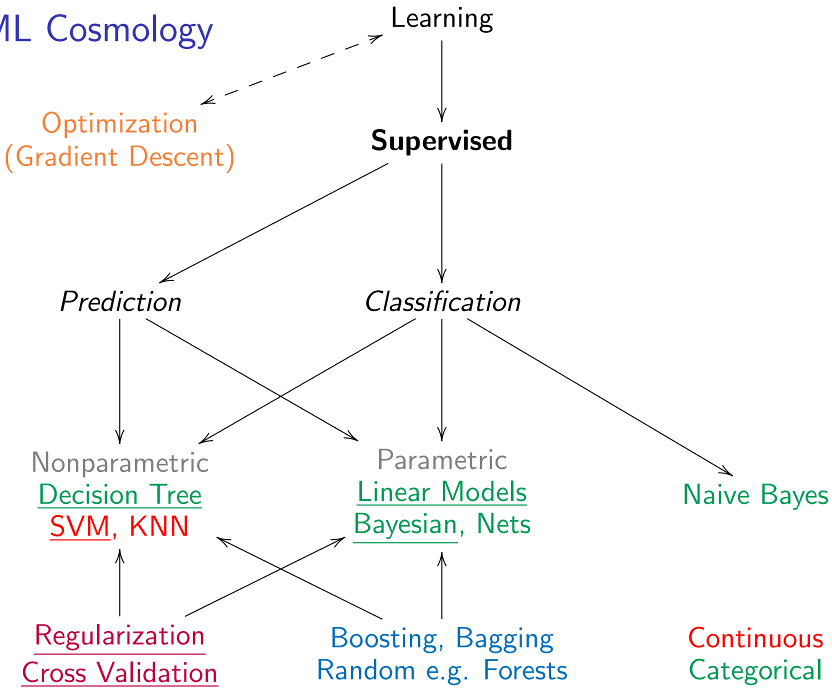


Continuous
Categorical

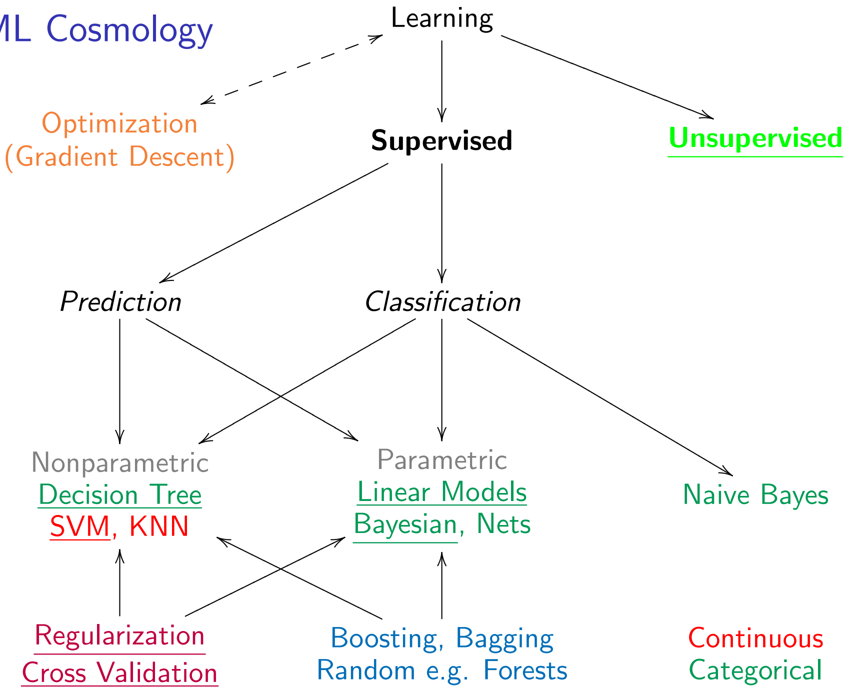
ML Cosmology



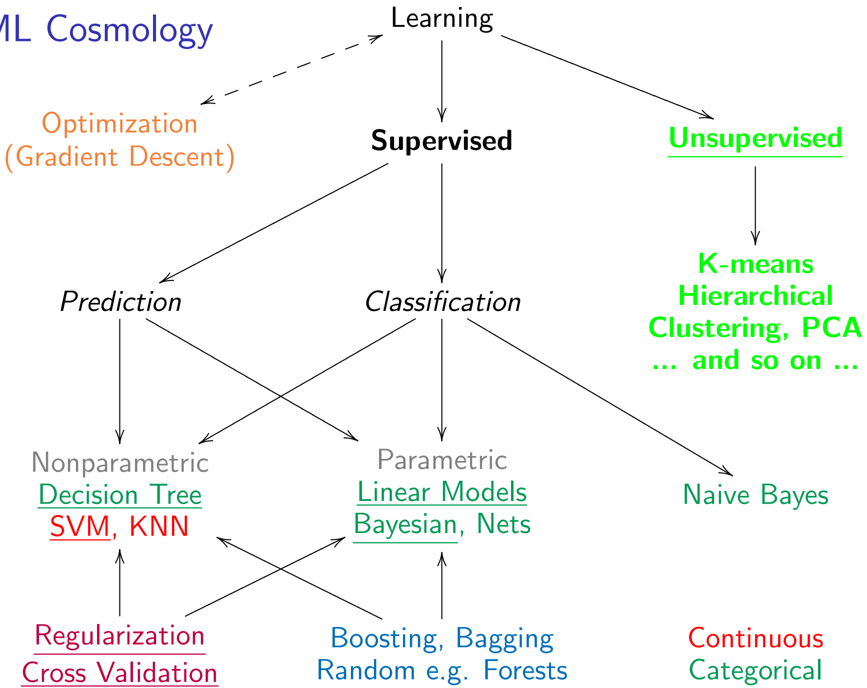
ML Cosmology



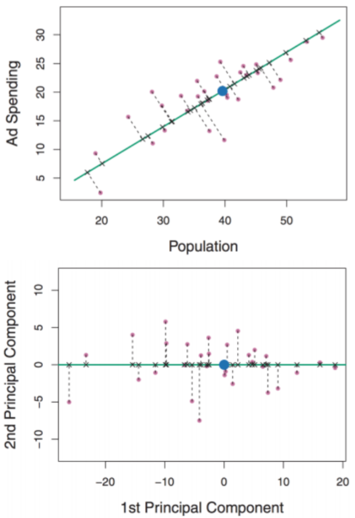
ML Cosmology

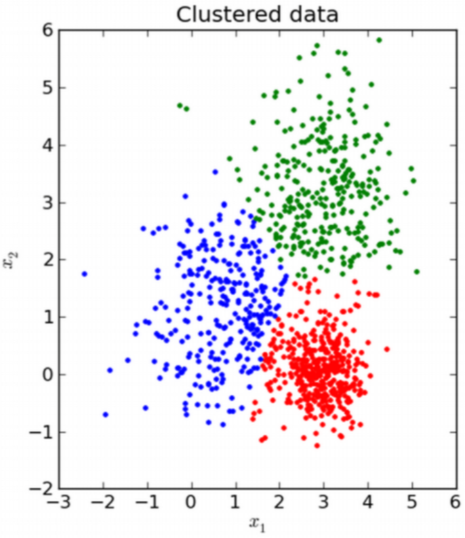


ML Cosmology



Unsupervised learning





Low dimensional representations of data capturing data variation

Homogeneous subgroups capturing data substructure

Supervised versus Unsupervised

Supervised versus Unsupervised

- Supervised learning has outcomes (labels, targets, Y) (dependent/endogenous variables)

Supervised versus Unsupervised

- Supervised learning has outcomes (labels, targets, Y)
  (dependent/endogenous variables)
- Unsupervised *doesn't*

Supervised versus Unsupervised

- Supervised learning has outcomes (labels, targets, Y) (dependent/endogenous variables)
- Unsupervised *doesn't*

- Unsupervised learning provides

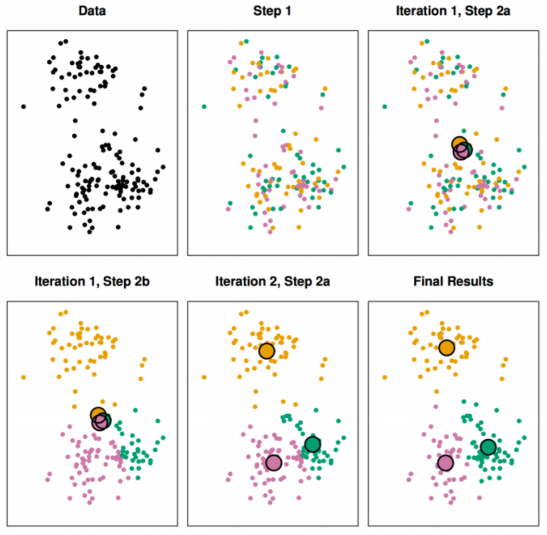
Supervised versus Unsupervised

- Supervised learning has outcomes (labels, targets, Y)
  (dependent/endogenous variables)
- Unsupervised *doesn't*

- Unsupervised learning provides
  Exploratory data analysis (EDA)
  Latent data structure discovery
  Feature engineering/Data reduction
  Quality control (QC) preprocessing

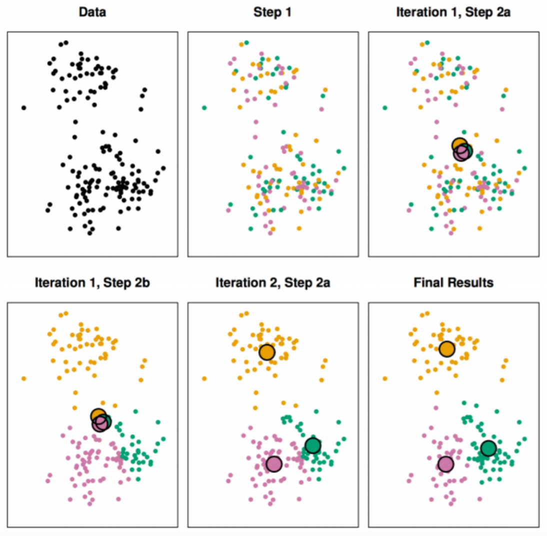
K-Means

1. Choose *number of clusters, K*
2. Randomly assign data to each cluster *k*
3. Compute the centroid for each cluster *k*
4. Assign data to the cluster with the closest centroid
5. Return to step 3, unless the centroids have stabilized



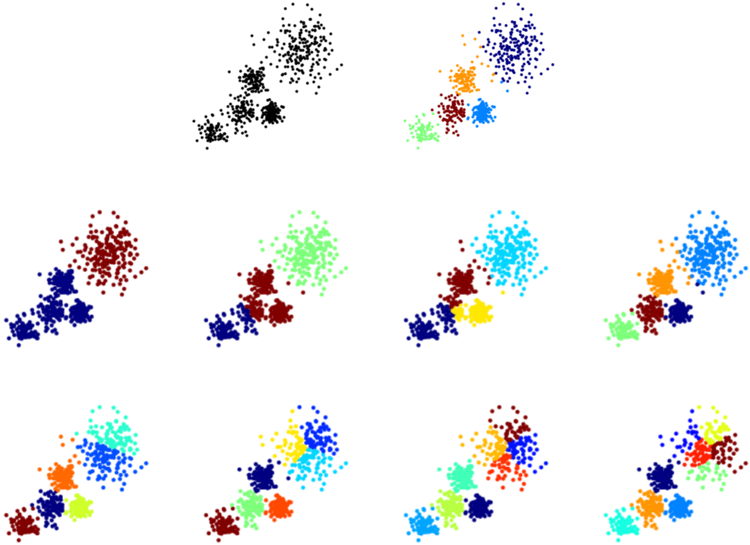
K-Means

1. Choose *number of clusters, K*
2. Randomly assign data to each cluster k
3. Compute the centroid for each cluster k
4. Assign data to the cluster with the closest centroid
5. Return to step 3, unless the centroids have stabilized



Any other ideas?

K?



Elbow and Silhouette methods

For some clustering

$C_K \mapsto \{1, 2, \cdots K\}$

clustering fit can be measured as

$$W(C_K) = \frac{1}{K} \sum_{C_K(i), C_K(j) = k} ||x_i - x_j||^2$$

Elbow and Silhouette methods

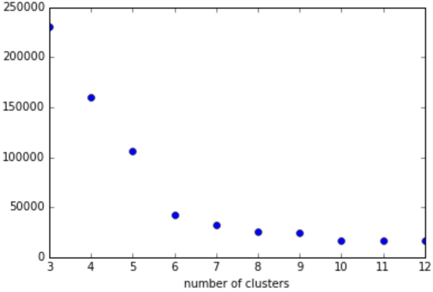
For some clustering

$C_K \mapsto \{1, 2, \cdots K\}$

clustering fit can be measured as

$$W(C_K) = \frac{1}{K} \sum_{C_K(i), C_K(j)=k} ||x_i - x_j||^2$$

Select based on diminishing returns



Elbow and Silhouette methods

For some clustering

$C_K \mapsto \{1, 2, \cdots K\}$

clustering fit can be measured as

$\bar{\Delta}_i^k$: x_i's dissimilarity in cluster k

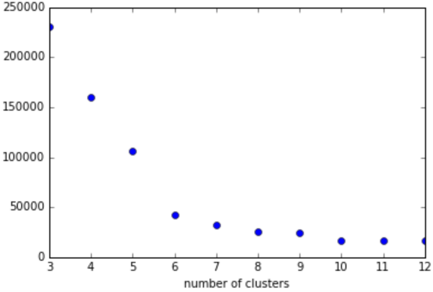
$\bar{\Delta}_i^{k'}$: x_i's dissimilarity to cluster k'

(x_i in, & closest to clusters k, k')

$$W(C_K) = \frac{1}{K} \sum_{C_K(i), C_K(j)=k} ||x_i - x_j||^2$$

$$Silhouette(i) = \frac{\bar{\Delta}_i^{k'} - \bar{\Delta}_i^k}{\max(\bar{\Delta}_i^{k'}, \bar{\Delta}_i^k)}$$

Select based on diminishing returns



Elbow and Silhouette methods

For some clustering

$C_K \mapsto \{1, 2, \cdots K\}$

clustering fit can be measured as

$$W(C_K) = \frac{1}{K} \sum_{C_K(i), C_K(j)=k} ||x_i - x_j||^2$$

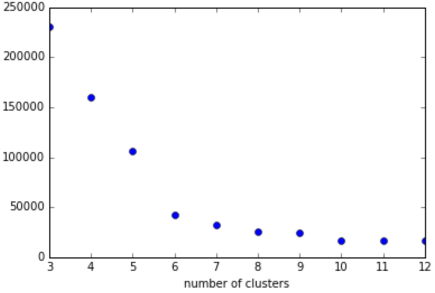
$\bar{\Delta}_i^k$: x_i's dissimilarity in cluster k

$\bar{\Delta}_i^{k'}$: x_i's dissimilarity to cluster k'

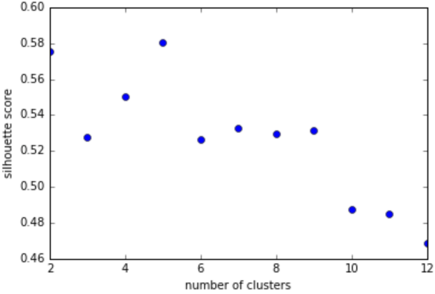
(x_i in, & closest to clusters k, k')

$$Silhouette(i) = \frac{\bar{\Delta}_i^{k'} - \bar{\Delta}_i^k}{\max(\bar{\Delta}_i^{k'}, \bar{\Delta}_i^k)}$$

Select based on diminishing returns



Compare average silhouette scores



Elbow and Silhouette methods

For some clustering

$C_K \mapsto \{1, 2, \cdots K\}$

clustering fit can be measured as

$\bar{\Delta}_i^k$: x_i's dissimilarity in cluster k

$\bar{\Delta}_i^{k'}$: x_i's dissimilarity to cluster k'

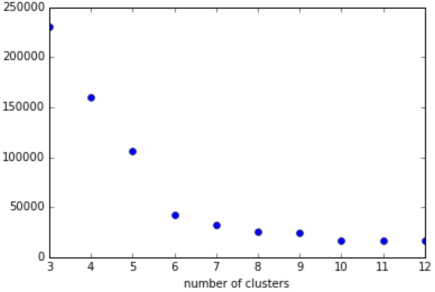
(x_i in, & closest to clusters k, k')

$$W(C_K) = \frac{1}{K} \sum_{C_K(i), C_K(j) = k} ||x_i - x_j||^2$$

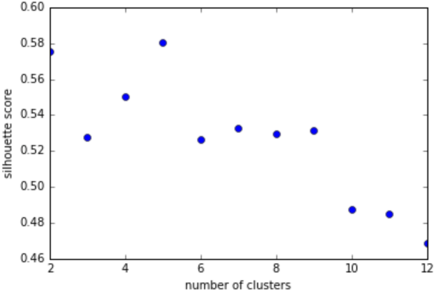
$$Silhouette(i) = \frac{\bar{\Delta}_i^{k'} - \bar{\Delta}_i^k}{\max(\bar{\Delta}_i^{k'}, \bar{\Delta}_i^k)}$$

Does the scaling of the x matter?

Select based on diminishing returns



Compare average silhouette scores



Elbow and Silhouette methods

For some clustering

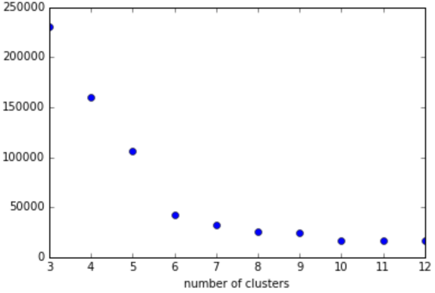
$C_K \mapsto \{1, 2, \cdots K\}$

clustering fit can be measured as

$$W(C_K) = \frac{1}{K} \sum_{C_K(i), C_K(j)=k} ||x_i - x_j||^2$$

Does the scaling of the x matter?

Select based on diminishing returns



$\bar{\Delta}_i^k$: x_i's dissimilarity in cluster k

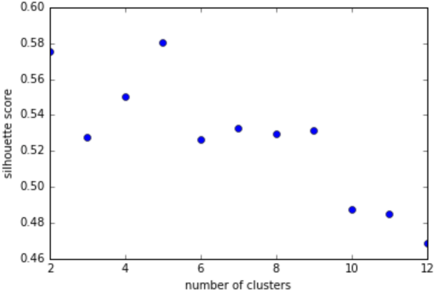
$\bar{\Delta}_i^{k'}$: x_i's dissimilarity to cluster k'

(x_i in, & closest to clusters k, k')

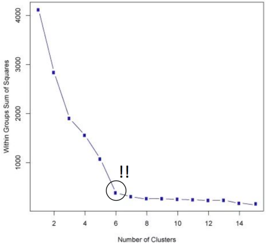
$$Silhouette(i) = \frac{\bar{\Delta}_i^{k'} - \bar{\Delta}_i^k}{\max(\bar{\Delta}_i^{k'}, \bar{\Delta}_i^k)}$$

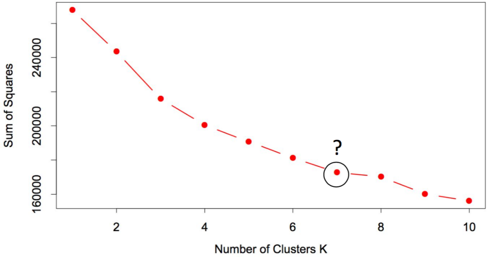
Interpret silhouette scores?

Compare average silhouette scores

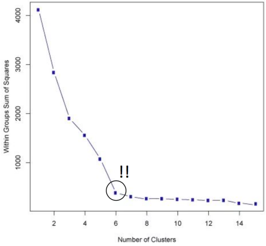


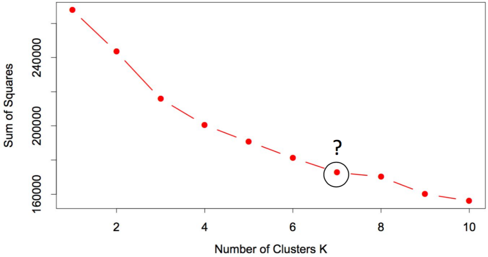
Elbow and Silhouette methods

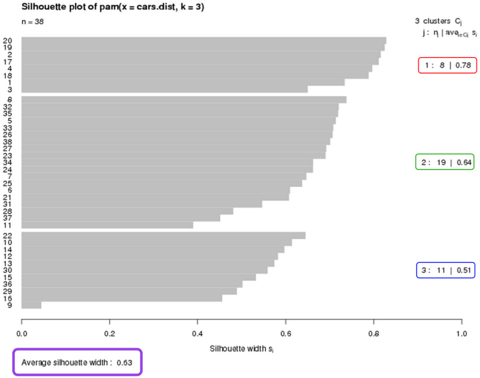




Elbow and Silhouette methods







Guidelines for Overall Avg Silhouette

Range	Interpretation
0.71 – 1.0	Strong structure found
0.51 – 0.7	Reasonable structure
0.26 – 0.5	Structure weak/artificial
< 0.25	No substantial structure

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n+1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n+1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

- The classes have the same tests...
  How could we test which class is doing better?

```
array([59, 92, 93, 83, 92, 61, 84, 92, 70, 93, 76, 70, 84, 61, 75, 91, 73,
       67, 65, 64, 75, 80, 66, 56, 62, 53, 82, 69, 85, 94, 80, 86, 97, 99,
       77, 84, 94, 70, 80, 87, 77, 85, 95, 80, 88, 90, 85, 84, 92, 75, 83,
       89, 76, 95, 91, 97, 80, 70, 71, 94])
```

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n+1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

- The classes have the same tests...
  How could we test which class is doing better?

```
array([59, 92, 93, 83, 92, 61, 84, 92, 70, 93, 76, 70, 84, 61, 75, 91, 73,
       67, 65, 64, 75, 80, 66, 56, 62, 53, 82, 69, 85, 94, 80, 86, 97, 99,
       77, 84, 94, 70, 80, 87, 77, 85, 95, 80, 88, 90, 85, 84, 92, 75, 83,
       89, 76, 95, 91, 97, 80, 70, 71, 94])
```

- We could *permute* the index

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n + 1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

- The classes have the same tests...
  How could we test which class is doing better?

```
array([59, 92, 93, 83, 92, 61, 84, 92, 70, 93, 76, 70, 84, 61, 75, 91, 73,
       67, 65, 64, 75, 80, 66, 56, 62, 53, 82, 69, 85, 94, 80, 86, 97, 99,
       77, 84, 94, 70, 80, 87, 77, 85, 95, 80, 88, 90, 85, 84, 92, 75, 83,
       89, 76, 95, 91, 97, 80, 70, 71, 94])
```

- We could *permute* the index
  (The classes doing equally well $\iff$ index doesn't matter)

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n + 1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

- The classes have the same tests...
  How could we test which class is doing better?

```
array([59, 92, 93, 83, 92, 61, 84, 92, 70, 93, 76, 70, 84, 61, 75, 91, 73,
       67, 65, 64, 75, 80, 66, 56, 62, 53, 82, 69, 85, 94, 80, 86, 97, 99,
       77, 84, 94, 70, 80, 87, 77, 85, 95, 80, 88, 90, 85, 84, 92, 75, 83,
       89, 76, 95, 91, 97, 80, 70, 71, 94])
```

- We could *permute* the index
  (The classes doing equally well $\iff$ index doesn't matter)

1. Repeatedly permute the index and recalculating the test statistic each time

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n+1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

- The classes have the same tests...
  How could we test which class is doing better?

```
array([59, 92, 93, 83, 92, 61, 84, 92, 70, 93, 76, 70, 84, 61, 75, 91, 73,
       67, 65, 64, 75, 80, 66, 56, 62, 53, 82, 69, 85, 94, 80, 86, 97, 99,
       77, 84, 94, 70, 80, 87, 77, 85, 95, 80, 88, 90, 85, 84, 92, 75, 83,
       89, 76, 95, 91, 97, 80, 70, 71, 94])
```

- We could *permute* the index
  (The classes doing equally well $\iff$ index doesn't matter)

1. Repeatedly permute the index and recalculating the test statistic each time
2. These samples approximate the test statistic distribution under the null

The *permutation test*

- Students enrolling for a popular class get ids $i = 1, \cdots n$. As the class quickly fills, another section of the class is opened and students for that class are given ids $i = n + 1, \cdots 2n$.

```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60])
```

- The classes have the same tests...
  How could we test which class is doing better?

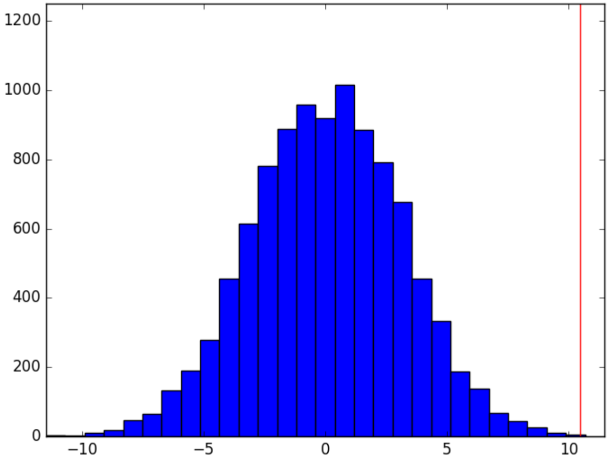
```
array([59, 92, 93, 83, 92, 61, 84, 92, 70, 93, 76, 70, 84, 61, 75, 91, 73,
       67, 65, 64, 75, 80, 66, 56, 62, 53, 82, 69, 85, 94, 80, 86, 97, 99,
       77, 84, 94, 70, 80, 87, 77, 85, 95, 80, 88, 90, 85, 84, 92, 75, 83,
       89, 76, 95, 91, 97, 80, 70, 71, 94])
```

- We could *permute* the index
  (The classes doing equally well $\iff$ index doesn't matter)

1. Repeatedly permute the index and recalculating the test statistic each time
2. These samples approximate the test statistic distribution under the null
3. Compare the test statistic to this null distribution
   to suggest how "strange" the actual observed test statistic is if the null is true

The *permutation test (This is Scott S's **favorite** test!!)*

1. Permute the ids (i.e., believe the null is true: ids don't matter)
2. Recalculate the test statistic each time (under null)
3. See how strange your observed statistic is compared to nulls



The *permutation test* and the *null distribution*

- What was needed in the permutation test was really the distribution of the test statistic under the null hypothesis

The *permutation test* and the *null distribution*

- What was needed in the permutation test was really the distribution of the test statistic under the null hypothesis
- We used permutation to get it

The *permutation test* and the *null distribution*

- What was needed in the permutation test was really the distribution of the test statistic under the null hypothesis
- We used permutation to get it
  but that was just a means to an end

The *permutation test* and the *null distribution*

- What was needed in the permutation test was really the distribution of the test statistic under the null hypothesis

- We used permutation to get it
  but that was just a means to an end

- We can test against *ANY* null distribution we wish to propose

The Gap statistic

412 *R. Tibshirani, G. Walther and T. Hastie*

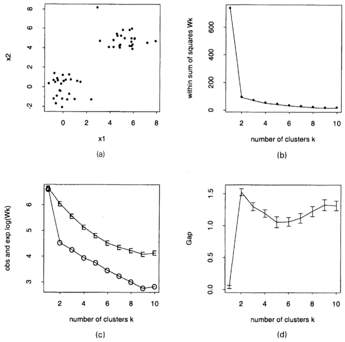


Fig. 1. Results for the two-cluster example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

416 *R. Tibshirani, G. Walther and T. Hastie*

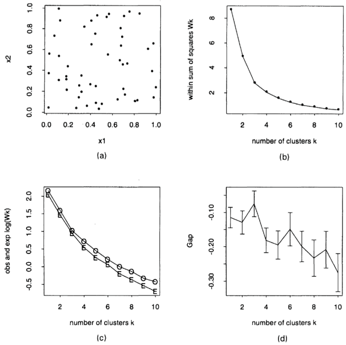


Fig. 2. Results for the uniform data example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

The Gap statistic

For M null distribution samples, calculate

$$\mathrm{Gap}(K) = \bar{l} - \log W(C_K) \quad \text{and} \quad s_K = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\log W_j(C_K) - \bar{l}\right)^2}$$

$$\text{where } W_j(C_K) = \frac{1}{K}\sum_{C_K(i),C_K(j)=k} ||x_i^{(j)} - x_j^{(j)}||^2 \quad \text{and} \quad \bar{l} = \frac{1}{M}\sum_{j=1}^{M}\log W_j(C_K)$$

412 R. Tibshirani, G. Walther and T. Hastie

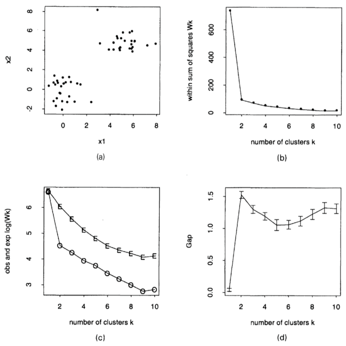


Fig. 1. Results for the two-cluster example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

416 R. Tibshirani, G. Walther and T. Hastie

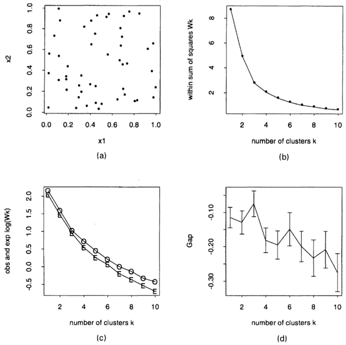


Fig. 2. Results for the uniform data example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

The Gap statistic

For M null distribution samples, calculate

$$\text{Gap}(K) = \bar{l} - \log W(C_K) \quad \text{and} \quad s_K = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\log W_j(C_K) - \bar{l}\right)^2}$$

$$\text{where } W_j(C_K) = \frac{1}{K}\sum_{C_K(i),C_K(j)=k} ||x_i^{(j)} - x_j^{(j)}||^2 \quad \text{and} \quad \bar{l} = \frac{1}{M}\sum_{j=1}^{M}\log W_j(C_K)$$

Then choose the smallest K such that $\text{Gap}(K) \geq \text{Gap}(K+1) - s_{K+1}$

412 *R. Tibshirani, G. Walther and T. Hastie*

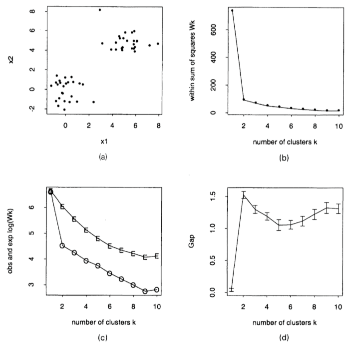


Fig. 1. Results for the two-cluster example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

416 *R. Tibshirani, G. Walther and T. Hastie*

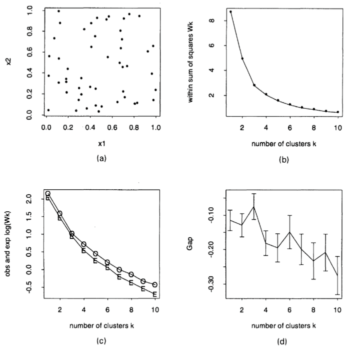


Fig. 2. Results for the uniform data example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

Curse of dimensionality

Just as nearest neighbors breaks down in high dimensional space...

Distance based clustering breaks down in high dimensional space...

Curse of dimensionality

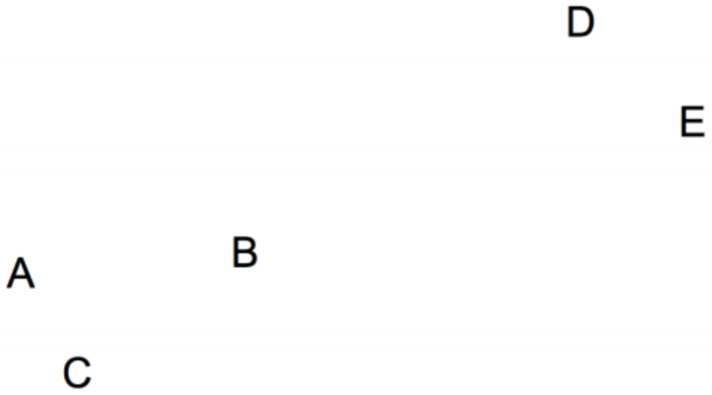
Just as nearest neighbors breaks down in high dimensional space...

Distance based clustering breaks down in high dimensional space...

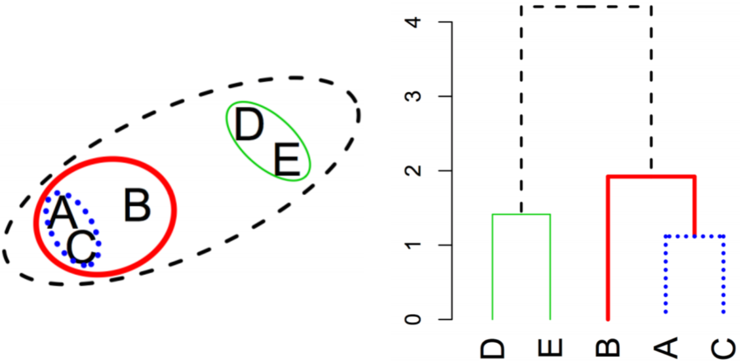
(Go see Ryan's great slides motivating the curse of dimensionality!)

(They are in the K nearest neighbors (KNN) lecture!!)

Challange

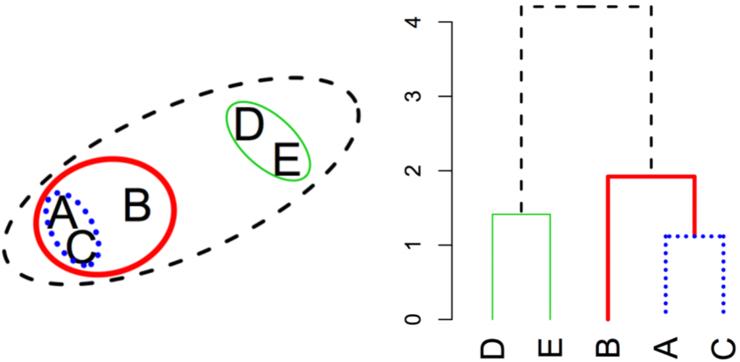


Hierarchical clustering



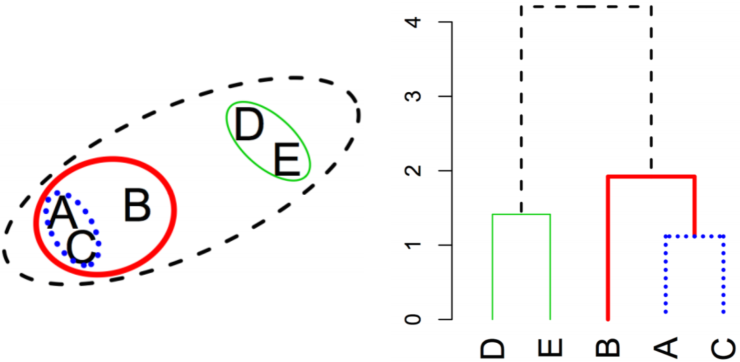
1. Assign each point to a cluster

Hierarchical clustering



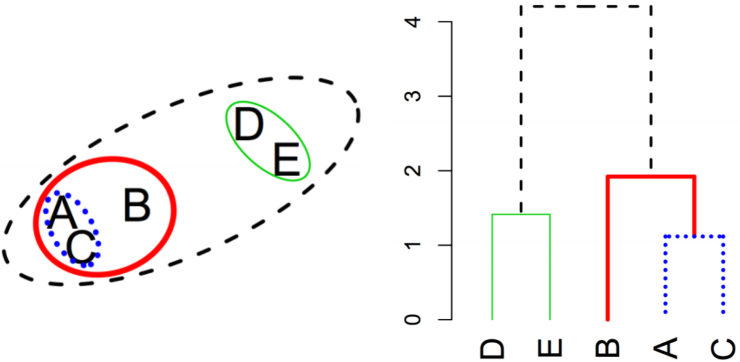
1. Assign each point to a cluster
2. Computer pairwise cluster distances

Hierarchical clustering



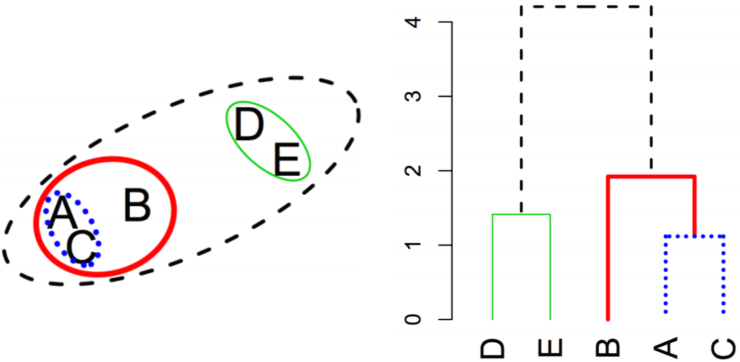
1. Assign each point to a cluster
2. Computer pairwise cluster distances
3. Merge *closest two* clusters

Hierarchical clustering



1. Assign each point to a cluster
2. Computer pairwise cluster distances
3. Merge *closest two* clusters
4. Return to 2 until all clusters merged

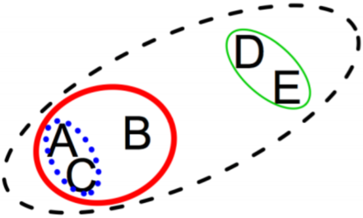
Hierarchical clustering

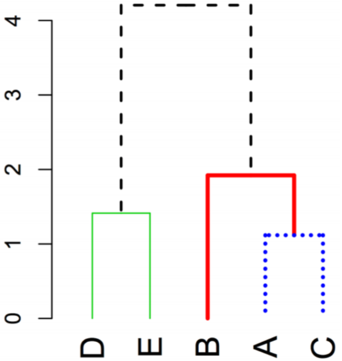


1. Assign each point to a cluster
2. Computer pairwise cluster distances
3. Merge *closest two* clusters
4. Return to 2 until all clusters merged

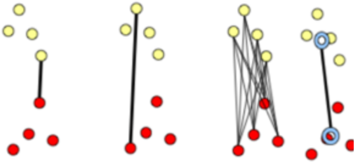
- Single: minimum pairwise point dissimilarity
- Complete: maximum pairwise point dissimilarity
- Average: average pairwise point dissimilarity
- Centroid: centroid dissimilarity

Hierarchical clustering

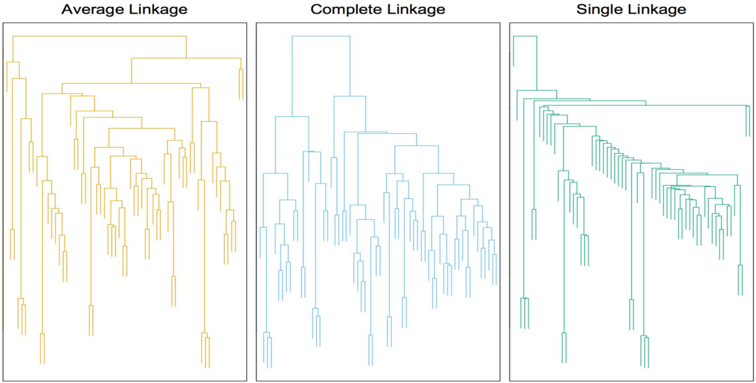




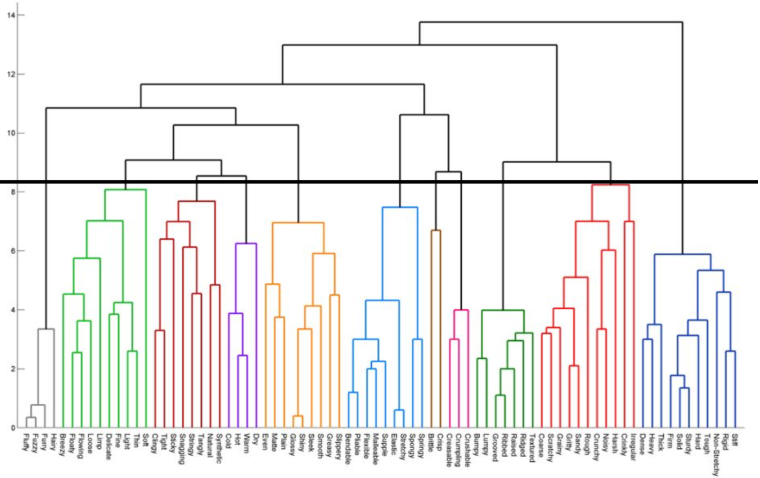
1. Assign each point to a cluster
2. Computer pairwise cluster distances
3. Merge *closest two* clusters
4. Return to 2 until all clusters merged



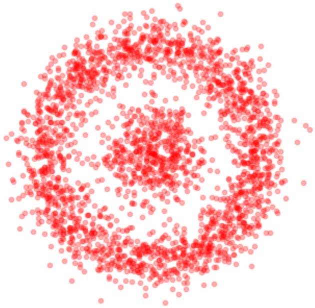
Hierarchical clustering *distance characteristics*



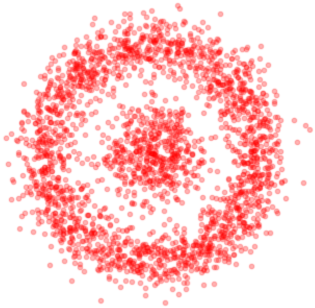
Hierarchical clustering *cluster count*



DBSCAN

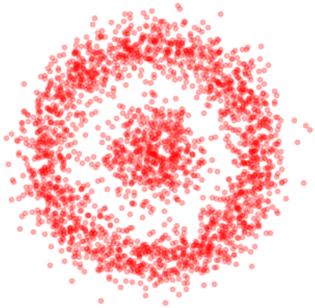


DBSCAN



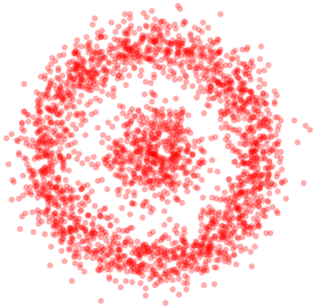
- Specify connection distance ϵ and minimum number of points for core m

DBSCAN



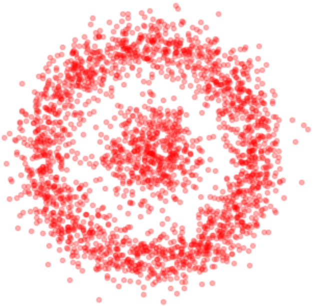
- Specify connection distance ϵ and minimum number of points for core m
- A cluster is all connected *core points*

DBSCAN

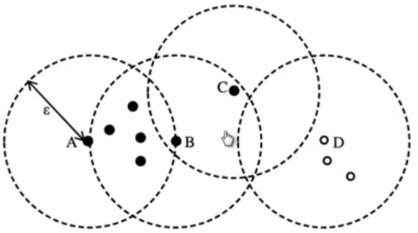


- Specify connection distance ϵ and minimum number of points for core m
- A cluster is all connected *core points*
- All other points are *noise*

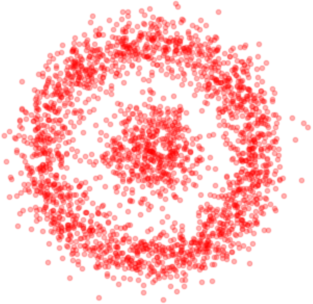
DBSCAN



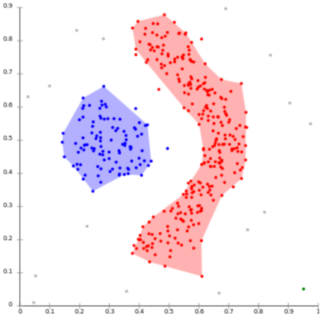
- Specify connection distance ϵ and minimum number of points for core m
- A cluster is all connected *core points*
- All other points are *noise*

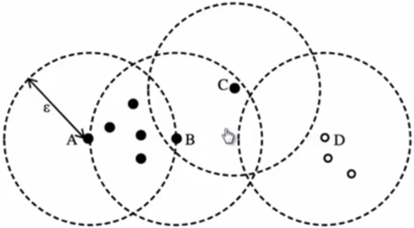


DBSCAN

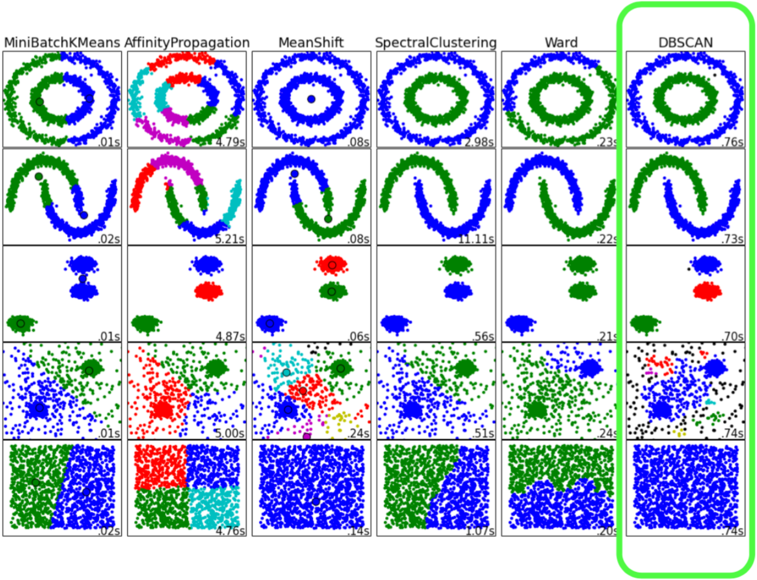


- Specify connection distance ϵ and minimum number of points for core m
- A cluster is all connected *core points*
- All other points are *noise*





DBSCAN


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
DBSCAN
.01s
4.79s
.08s
2.98s
.23s
.76s
.02s
5.21s
.08s
11.11s
.22s
.73s
.01s
4.87s
.06s
.56s
.21s
.70s
.01s
5.00s
.24s
.51s
.24s
.74s
.02s
4.76s
.14s
1.07s
.20s
.74s

Bayesian Mixture Models *(xtra: Scott S's grad school jam)*

$$f(X_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)$$

Bayesian Mixture Models *(xtra: Scott S's grad school jam)*

$$f(X_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)$$

$$\begin{aligned}
f(X_i|V_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}) &= N(\mu_k, \sigma_k^2) \\
\Pr(\boldsymbol{V}_i) &= \mathit{Multinomial}(\boldsymbol{\pi}, n = 1) \\
f(\boldsymbol{\pi}) &= \mathit{Dirichlet}(\boldsymbol{\omega}) \\
f(\mu_k) &= N(\mu_0, \sigma_0^2) \\
f(\sigma_k^{-2}) &= \mathit{Gamma}(\alpha, \beta)
\end{aligned}$$

$$\begin{tikzcd}
\omega \arrow[d] & & \\
\pi \arrow[d] & \mu_0, \sigma_0^2 \arrow[d] & \alpha, \beta \arrow[d] \\
V_i = k \arrow[dr] & \mu_k \arrow[d] & \sigma_k \arrow[dl] \\
& X_i &
\end{tikzcd}$$

Bayesian Mixture Models *(xtra: Scott S's grad school jam)*

$$f(X_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)$$

$$\begin{aligned}
f(X_i|V_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}) &= N(\mu_k, \sigma_k^2) \\
\Pr(\boldsymbol{V}_i) &= \mathit{Multinomial}(\boldsymbol{\pi}, n = 1) \\
f(\boldsymbol{\pi}) &= \mathit{Dirichlet}(\boldsymbol{\omega}) \\
f(\mu_k) &= N(\mu_0, \sigma_0^2) \\
f(\sigma_k^{-2}) &= \mathit{Gamma}(\alpha, \beta)
\end{aligned}$$

$$\begin{tikzcd}
\omega \arrow[d] & & \\
\pi \arrow[d] & \mu_0, \sigma_0^2 \arrow[d] & \alpha, \beta \arrow[d] \\
V_i = k \arrow[dr] & \mu_k \arrow[d] & \sigma_k \arrow[dl] \\
& X_i &
\end{tikzcd}$$

$$f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}|\mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega}) =$$

$$\prod_{i=1}^{n}\left[\left(\sum_{k=1}^{K} 1_{[V_{ik}=1]} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{X_i-\mu_k}{\sigma_k}\right)^2}\right)\left(\prod_{k=1}^{K} \pi_k^{V_{ik}}\right)\right]$$

$$\times \left(\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\left(\frac{\mu_k-\mu_0}{\sigma_0}\right)^2} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha-1} e^{-\beta\frac{1}{\sigma^2}}\right)\left(\frac{1}{\mathbf{B}(\omega)} \prod_{k=1}^{K} \pi_k^{\omega_k-1}\right)$$

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

(which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mathbf{X},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k|\mathbf{V},\boldsymbol{\sigma}^2,\mathbf{X},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2|\mathbf{V},\boldsymbol{\sigma}^2,\mathbf{X},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik}=1|\mathbf{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi}|\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mathbf{X},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k|\mathbf{V},\boldsymbol{\sigma}^2,\mathbf{X},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2|\mathbf{V},\boldsymbol{\sigma}^2,\mathbf{X},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik}=1|\mathbf{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi},\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi}|\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$

$$\propto f(\mathbf{X},\mathbf{V},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}|\mu_0,\sigma_0^2,\alpha,\beta,\boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *posterior sampling*

A Gibbs sampler for the posterior

$$f(\mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Is made by cycling through the *full conditional distributions*

$$f(\mu_k | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$f(\sigma_k^2 | \mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(V_{ik} = 1 | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

$$\Pr(\boldsymbol{\pi} | \mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$

$$\propto f(\mathbf{X}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} | \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega})$$ (which is proportional to the joint distribution)

Markov Chain Monte Carlo (MCMC) *full conditionals*

$$\Pr(V_{ik} = 1|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega}) \propto \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{X_i - \mu_k}{\sigma_k}\right)^2}$$

$$f(\pi|\mathbf{V}, \boldsymbol{mu}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega}) = Dirichlet(\{\omega_k + n_k : k = 1, \cdots K\})$$

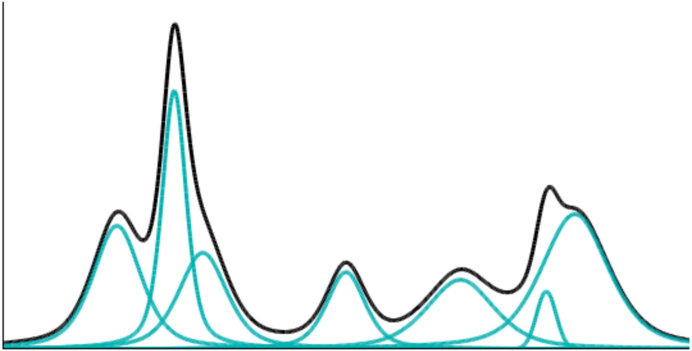
$$n_k = \sum_{V_{ik}=1} 1$$

$$f(\sigma_k^2|\mathbf{V}, \boldsymbol{\mu}, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega}) = Gamma\left(\frac{n_k}{2} + \alpha, \frac{1}{2}\sum_{V_{ik}=1}(X_i - \mu_k)^2 + \beta\right)$$

$$f(\mu_k|\mathbf{V}, \boldsymbol{\sigma}^2, \mathbf{X}, \mu_0, \sigma_0^2, \alpha, \beta, \boldsymbol{\omega}) =$$

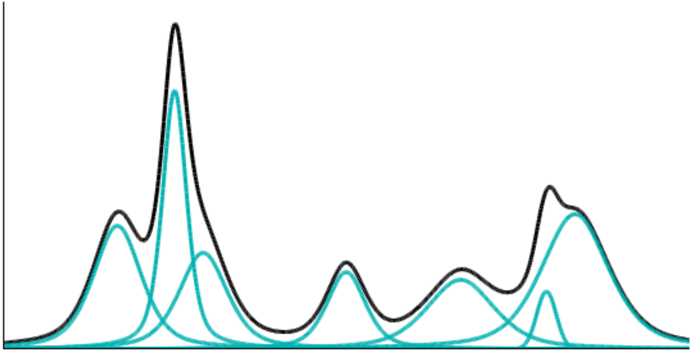
$$\mathrm{N}\left(\hat{\sigma}_k^2\left(\frac{\sum_{V_{ik}=1} X_i}{\sigma_k^2} + \frac{\mu_0}{\sigma_0^2}\right), \hat{\sigma}_k^2 = \left(\frac{n_k}{\sigma_k^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right)$$

Mixture Models!



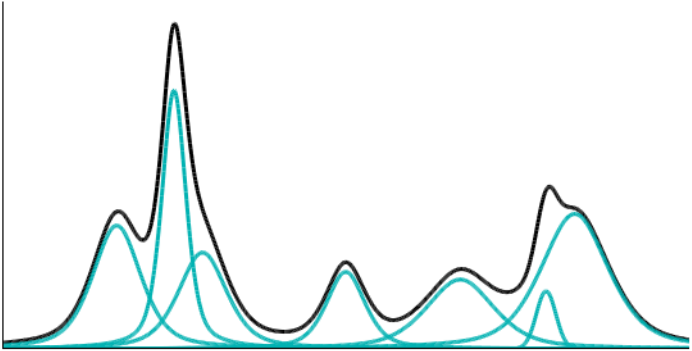
- Can be used to model "subpopulations"

Mixture Models!



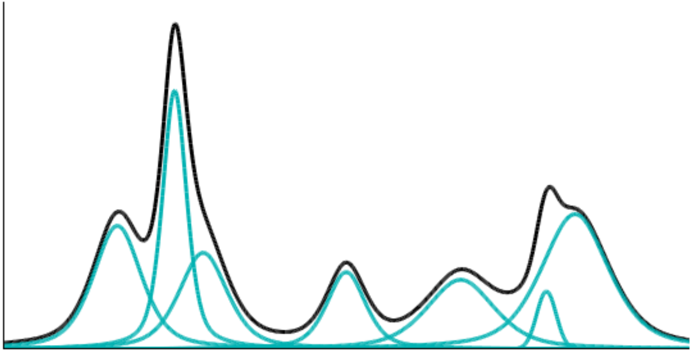
- Can be used to model "subpopulations"
- Or simply complex distributional shapes

Mixture Models!



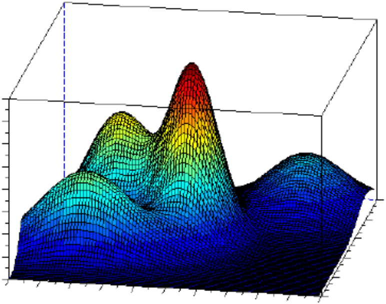
- Can be used to model "subpopulations"
- Or simply complex distributional shapes
- It's *almost nonparametric* like kernel density estimation

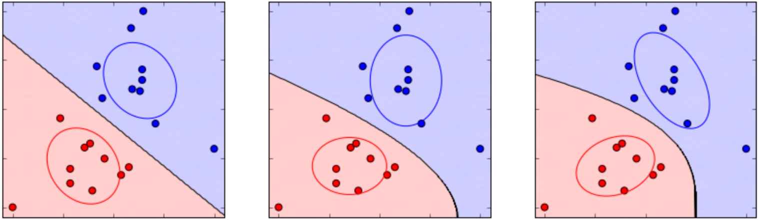
Mixture Models!



- Can be used to model “subpopulations”
- Or simply complex distributional shapes
- It’s *almost nonparametric* like kernel density estimation
- Also can be fit with Expectation-Maximization (EM) alg.

Mixture Models!





Byee!!

Scott's Last Lecture Party!!

<Party Dance, Party Dance>