Clustering and the k-means Algorithm

David M. Blei

COS424
Princeton University

March 2, 2012

Clustering

- Goal: Automatically segment data into groups of similar points

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile
  - Search results according to topic

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile
  - Search results according to topic
  - Facebook users according to interests

Clustering

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
  - Automatically organizing data
  - Understanding hidden structure in some data
  - Representing high-dimensional data in a low-dimensional space
- Examples:
  - Customers according to purchase histories
  - Genes according to expression profile
  - Search results according to topic
  - Facebook users according to interests
  - A museum catalog according to image similarity

Clustering set-up

- Our data are

$$\mathscr{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

Clustering set-up

- Our data are
$$\mathscr{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$
- Each data point is p-dimensional, i.e.,
$$\mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle.$$

Clustering set-up

- Our data are

$$\mathscr{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

- Each data point is p-dimensional, i.e.,

$$\mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle.$$

- Define a *distance function* between data, $d(\mathbf{x}_n, \mathbf{x}_m)$.

Clustering set-up

- Our data are

$$\mathscr{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

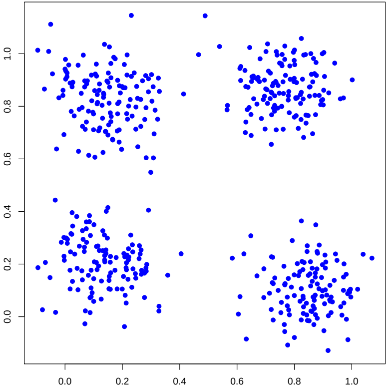
- Each data point is p-dimensional, i.e.,

$$\mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle.$$

- Define a *distance function* between data, $d(\mathbf{x}_n, \mathbf{x}_m)$.
- Goal: segment the data into k groups

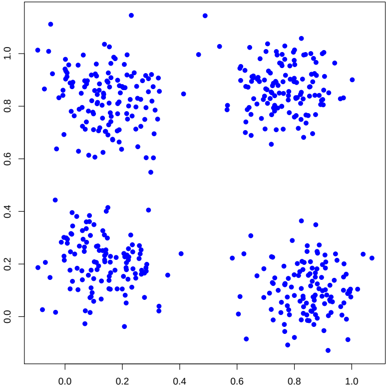
$$\{z_1, \ldots, z_N\} \quad \text{where} \quad z_i \in \{1, \ldots, K\}.$$

Example data



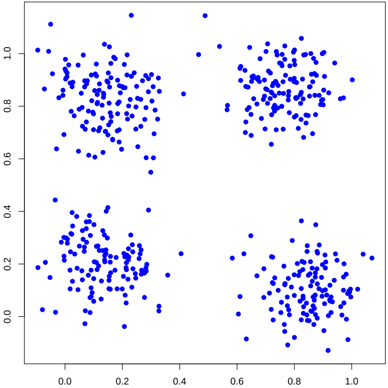
500 2-dimensional data points: $\mathbf{x}_n = \langle x_{n,1}, x_{n,2} \rangle$

Example data



- What is a good distance function here?

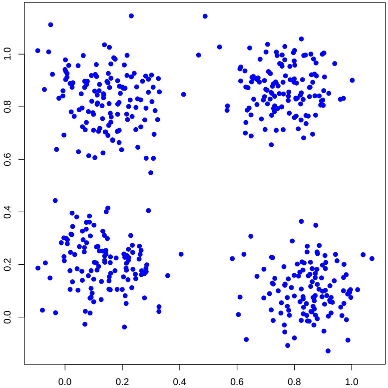
Example data



- What is a good distance function here?
- Squared Euclidean distance is reasonable

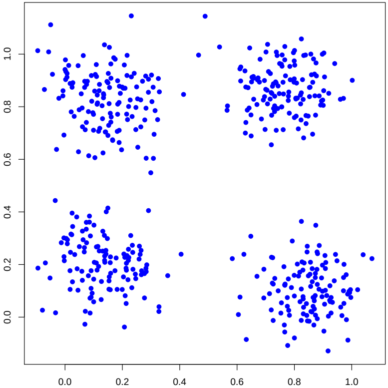
$$d(\mathbf{x}_n, \mathbf{x}_m) = \sum_{i=1}^{p} (x_{n,i} - x_{m,i})^2 = ||x_n - x_m||^2$$

Example data



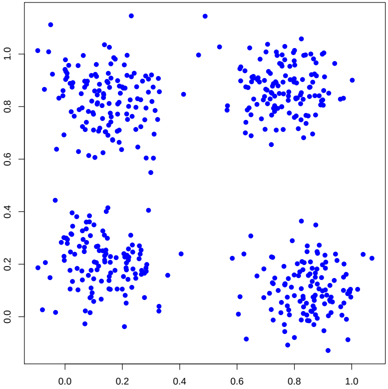
- Goal: segment this data into *k* groups.

Example data



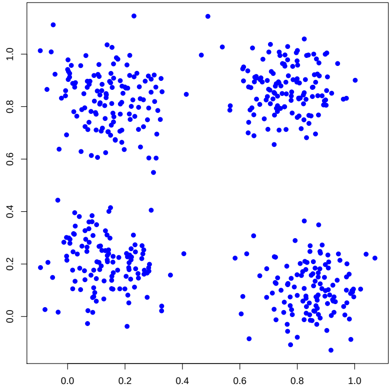
- Goal: segment this data into *k* groups.
- What should *k* be?

Example data



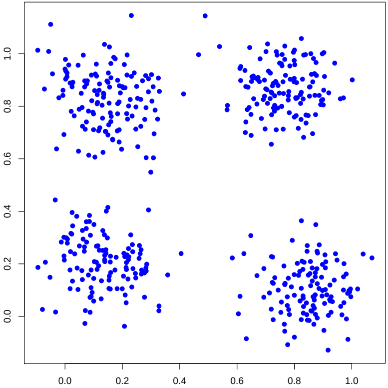
- Goal: segment this data into *k* groups.
- What should *k* be?
- Automatically choosing *k* is complicated; for now, 4.

k-means



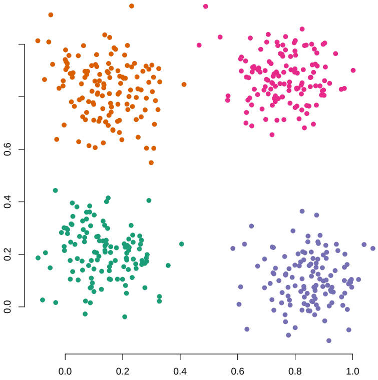
- Different clustering algorithms use data and distance in different ways

k-means



- Different clustering algorithms use data and distance in different ways
- We discuss k-means, the simplest clustering algorithm

k-means



- Different clustering algorithms use data and distance in different ways
- We discuss *k*-means, the simplest clustering algorithm

k-means

- The basic idea is to describe each cluster by its mean value.

k-means

- The basic idea is to describe each cluster by its mean value.
- (Note: this works only for distances such that a mean is well-defined.)

k-means

- The basic idea is to describe each cluster by its mean value.
- (Note: this works only for distances such that a mean is well-defined.)
- The goal of *k*-means is to assign data to clusters and define these clusters with their means.

k-means algorithm

1. Initialization

k-means algorithm

1. Initialization
    - Data are $\mathbf{x}_{1:N}$

k-means algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).

k-means algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).
2. Repeat

k-means algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).
2. Repeat
   1. Assign each data point to its closest mean

$$z_n = \arg\min_{i \in \{1,\ldots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

k-means algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).
2. Repeat
   1. Assign each data point to its closest mean

      $$z_n = \arg\min_{i \in \{1,\ldots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

   2. Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,

      $$\mathbf{m}_k = \frac{1}{N_k} \sum_{\{n:\, z_n = k\}} \mathbf{x}_n$$

k-means algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).
2. Repeat
   1. Assign each data point to its closest mean

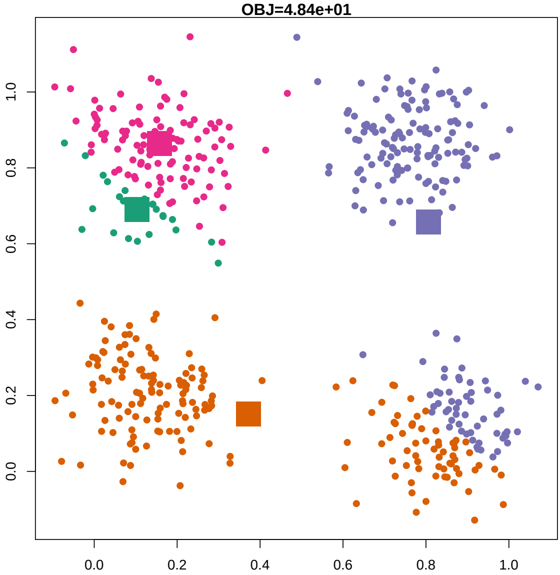
$$z_n = \arg \min_{i \in \{1,\ldots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

   2. Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,

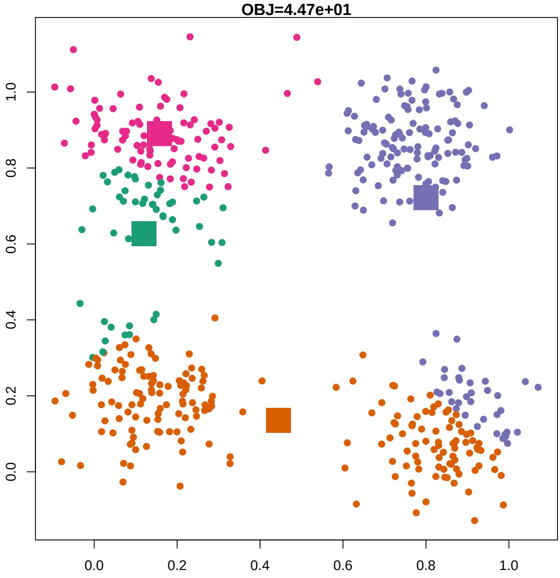
$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\{n:\, z_n = k\}} \mathbf{x}_n$$

3. Until assignments $\mathbf{z}_{1:N}$ do not change

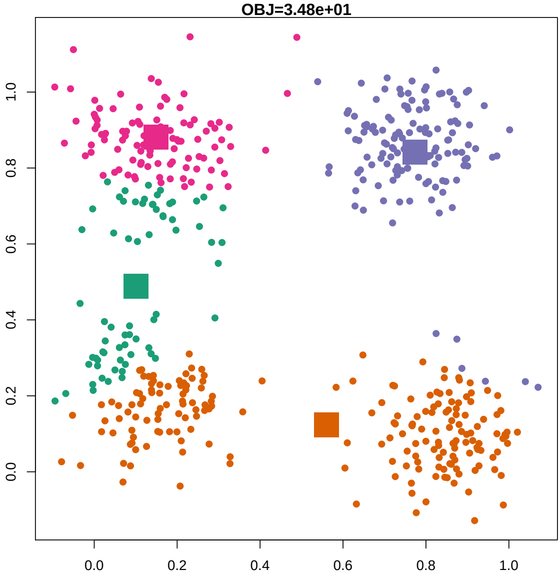
k-means example



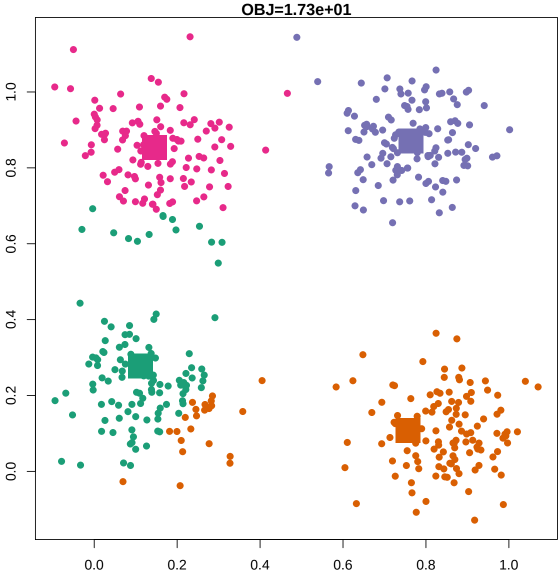
k-means example



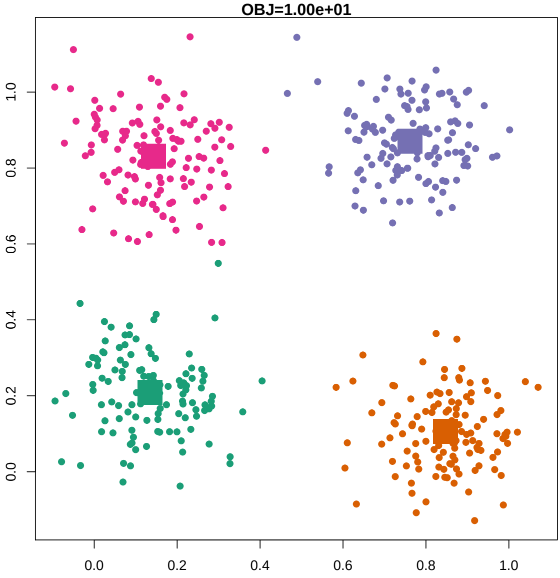
k-means example



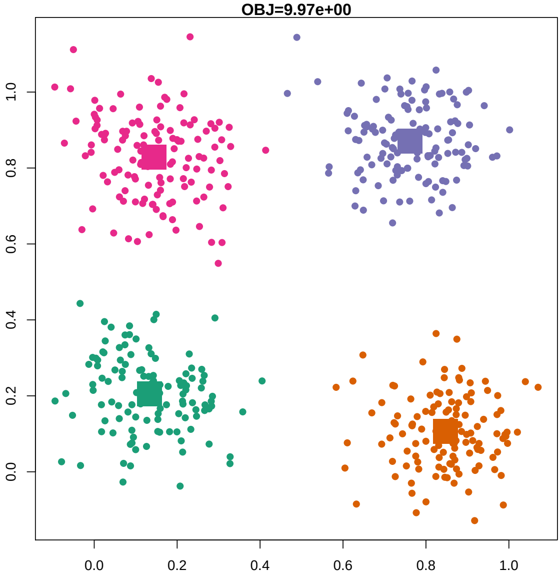
k-means example



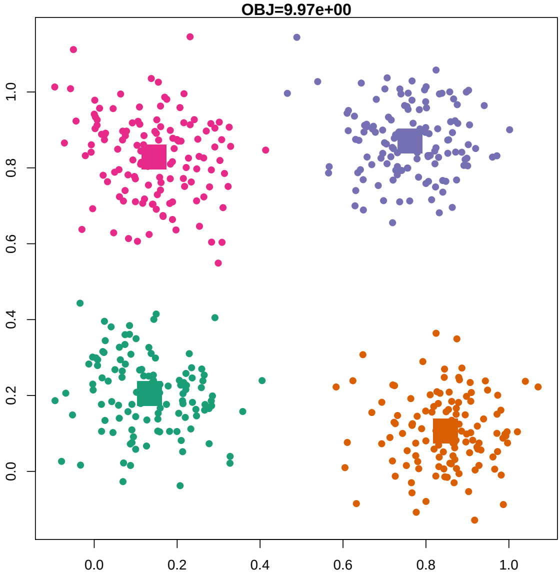
k-means example



k-means example



k-means example



Objective function

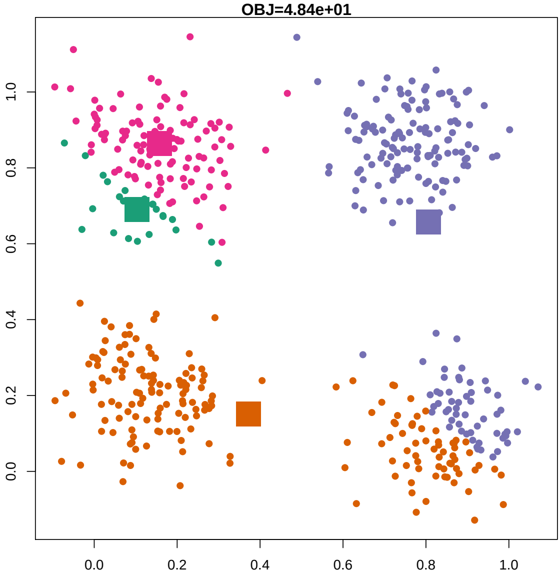
- How can we measure how well our algorithm is doing?

Objective function

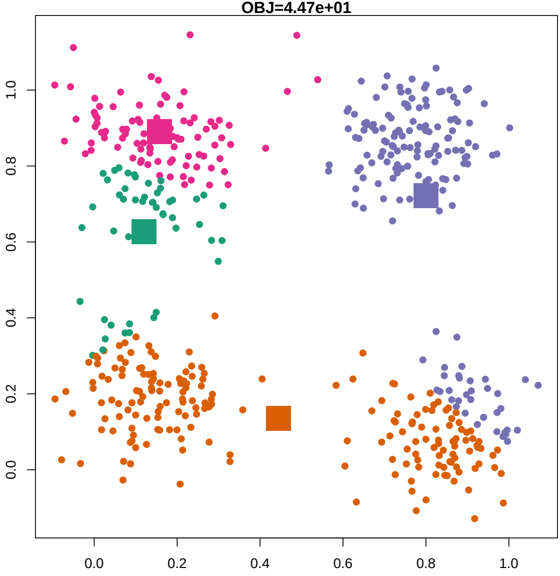
- How can we measure how well our algorithm is doing?
- The k-means objective function is the sum of the squared distances of each point to each assigned mean

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

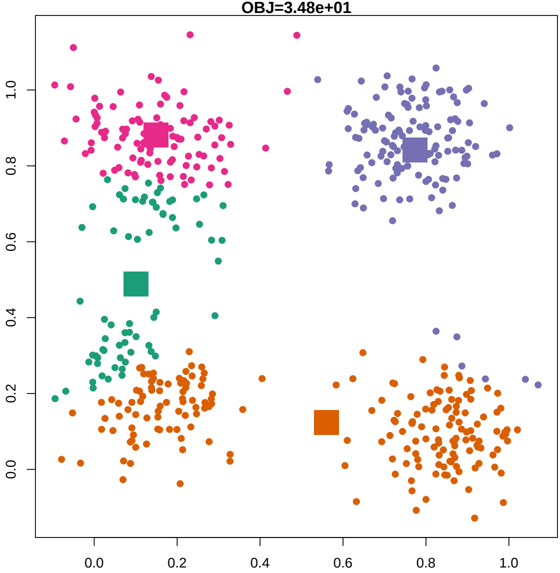
k-means example (look at the objective)



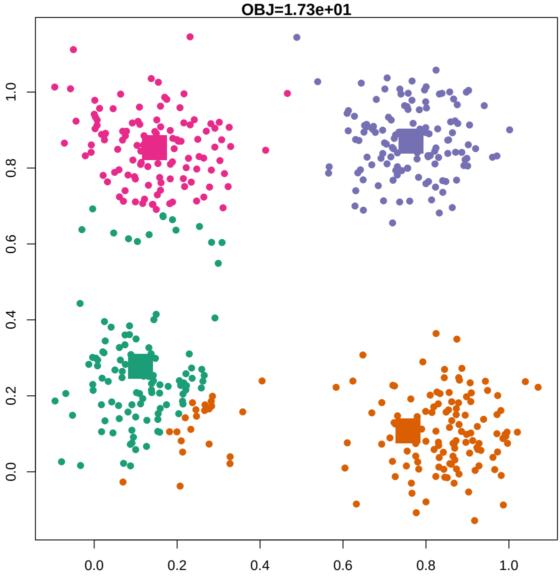
k-means example (look at the objective)



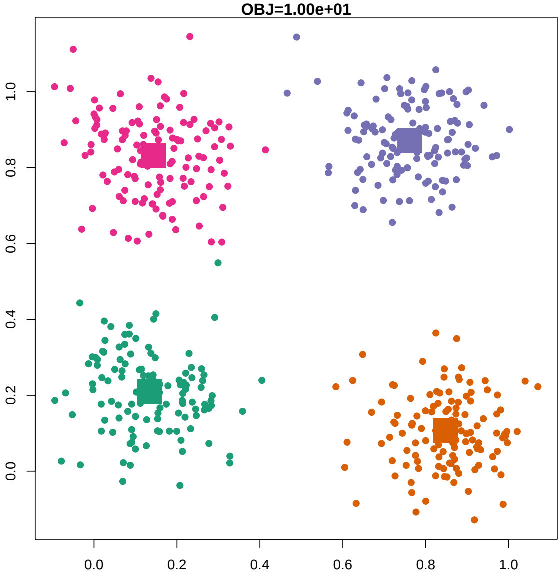
k-means example (look at the objective)



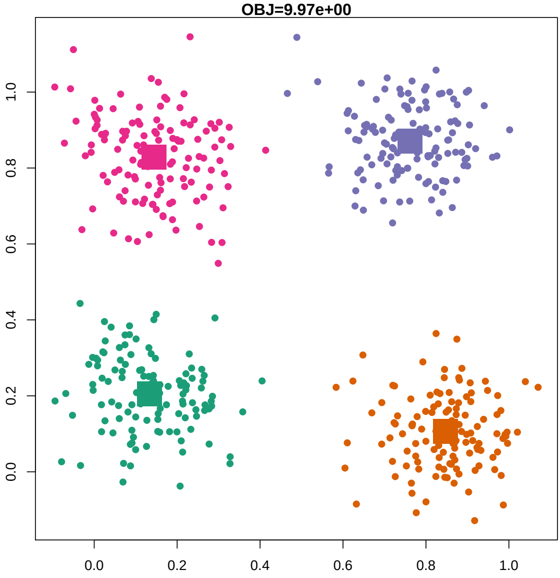
k-means example (look at the objective)



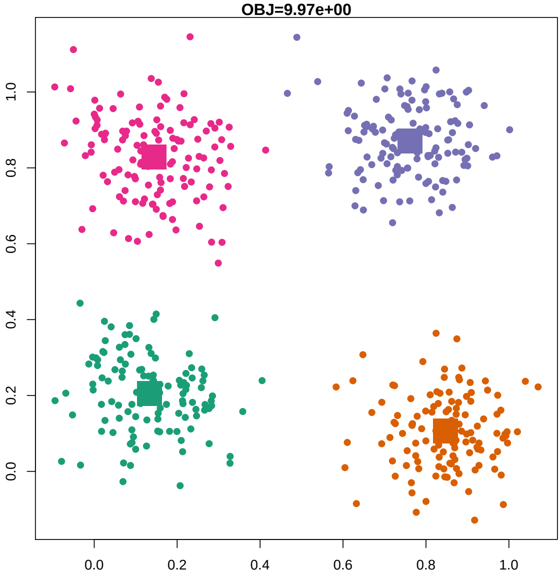
k-means example (look at the objective)



k-means example (look at the objective)



k-means example (look at the objective)



Coordinate descent

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

- Holding the means fixed, assigning each point to its closest mean minimizes F with respect to $z_{1:N}$.

Coordinate descent

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

- Holding the means fixed, assigning each point to its closest mean minimizes F with respect to $z_{1:N}$.
- Holding the assignments fixed, computing the centroids of each cluster minimizes F with respect to $\mathbf{m}_{1:k}$.

Coordinate descent

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

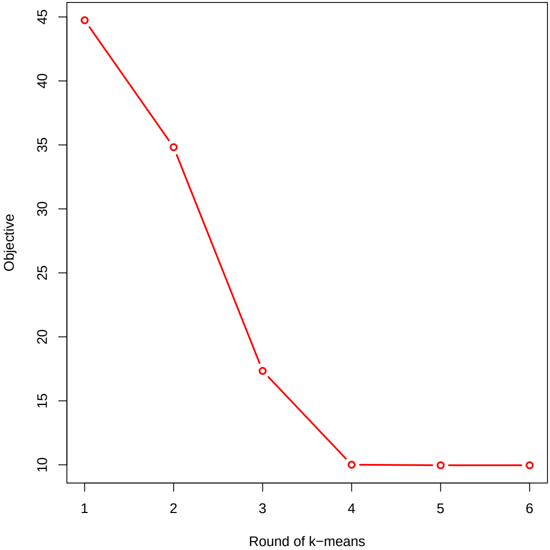
- Holding the means fixed, assigning each point to its closest mean minimizes F with respect to $z_{1:N}$.
- Holding the assignments fixed, computing the centroids of each cluster minimizes F with respect to $\mathbf{m}_{1:k}$.
- Thus, k-means is a *coordinate descent* algorithm.

Coordinate descent

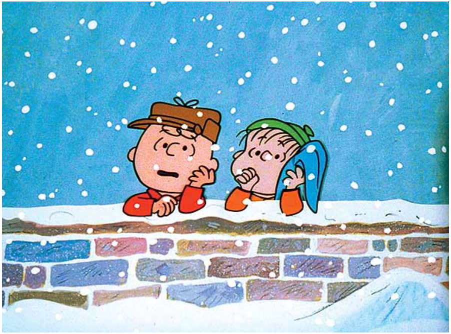
$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

- Holding the means fixed, assigning each point to its closest mean minimizes F with respect to $z_{1:N}$.
- Holding the assignments fixed, computing the centroids of each cluster minimizes F with respect to $\mathbf{m}_{1:k}$.
- Thus, k-means is a *coordinate descent* algorithm.
- It finds a *local minimum*. (Multiple restarts are often necessary.)

Objective for the example data

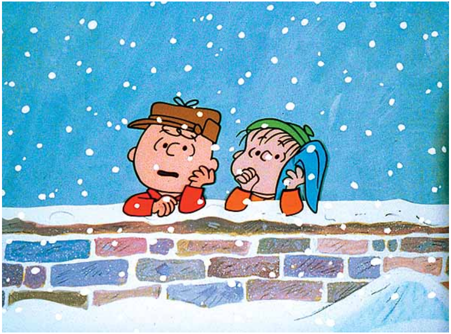


Compressing images



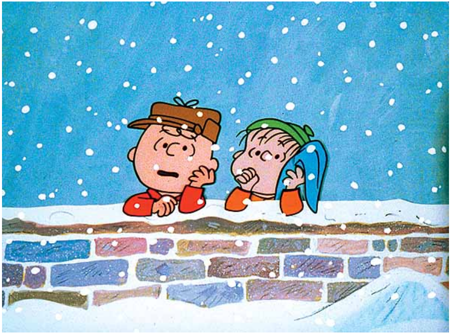
- Each pixel is associated with a red, green, and blue value

Compressing images



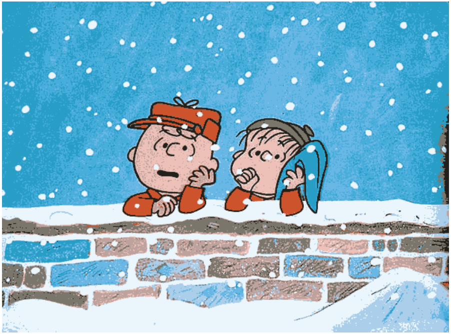
- Each pixel is associated with a red, green, and blue value
- A 1024×1024 image is a collection of 1048576 values $\langle x_1, x_2, x_3 \rangle$, which requires 3M of storage

Compressing images



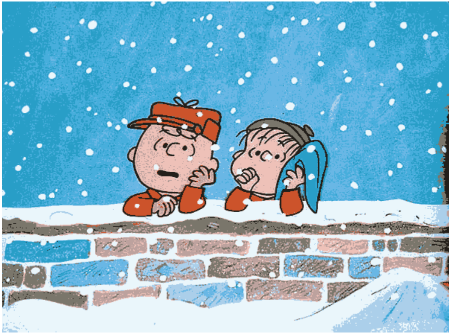
- Each pixel is associated with a red, green, and blue value
- A 1024×1024 image is a collection of 1048576 values $\langle x_1, x_2, x_3 \rangle$, which requires 3M of storage
- How can we use k-means to compress this image?

Vector quantization



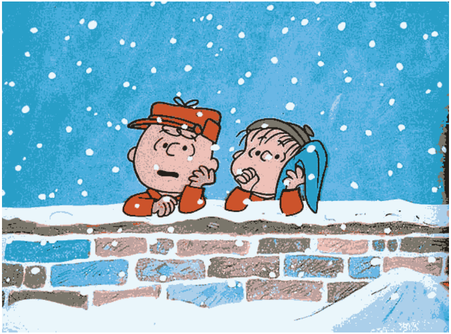
- Replace each pixel $\mathbf{x}_n$ with its assignment $\mathbf{m}_{z_n}$ (“paint by numbers”).

Vector quantization



- Replace each pixel $\mathbf{x}_n$ with its assignment $\mathbf{m}_{z_n}$ ("paint by numbers").
- The k means are called the *codebook*.

Vector quantization



- Replace each pixel $\mathbf{x}_n$ with its assignment $\mathbf{m}_{z_n}$ ("paint by numbers").
- The k means are called the *codebook*.
- With $k = 100$, we need 7 bits per pixel plus 100×3 bits ≈ 897K.

Charlie Brown and Linus VQ

2 means

Charlie Brown and Linus VQ

4 means

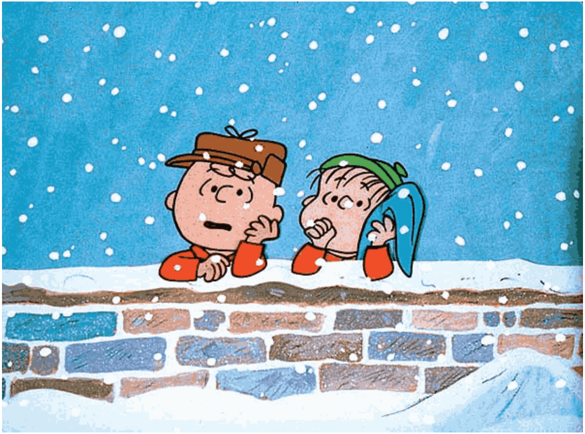
Charlie Brown and Linus VQ

8 means

Charlie Brown and Linus VQ

16 means

Charlie Brown and Linus VQ

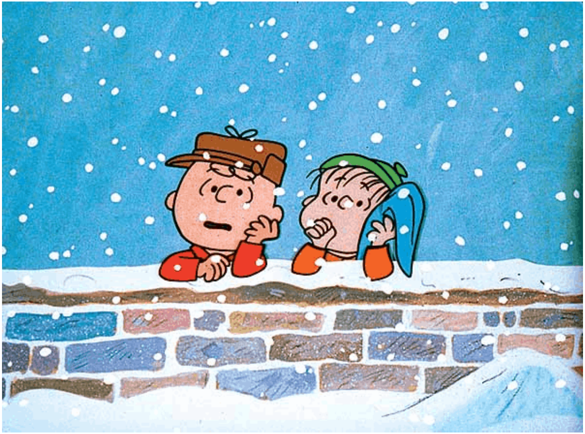


32 means

Charlie Brown and Linus VQ

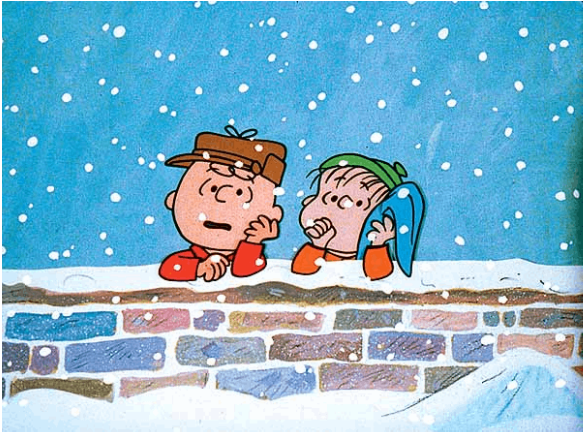
64 means

Charlie Brown and Linus VQ



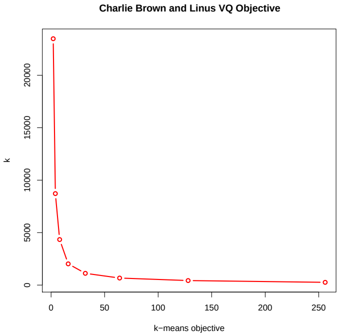
128 means

Charlie Brown and Linus VQ



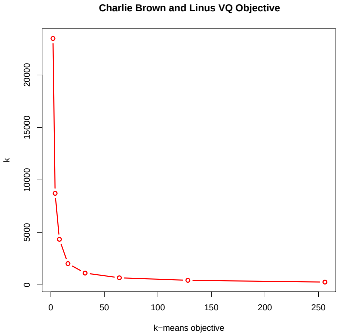
256 means

Measure of distortion



- The objective gives a measure of how distorted the compressed picture is relative to the original picture

Measure of distortion



- The objective gives a measure of how distorted the compressed picture is relative to the original picture
- For more clusters, the picture is less distorted.

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories
  - Positive data, such as time spent on a web-page

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories
  - Positive data, such as time spent on a web-page
- k-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories
  - Positive data, such as time spent on a web-page
- k-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.
- *No need to define the mean.*

k-medoids

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
  - Discrete multivariate data, such as purchase histories
  - Positive data, such as time spent on a web-page
- k-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.
- *No need to define the mean.*
- Each of the clusters is associated with its most typical example

k-medoids algorithm

1. Initialization

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster identities $\mathbf{m}_{1:k}$

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster identities $\mathbf{m}_{1:k}$
2. Repeat

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster identities $\mathbf{m}_{1:k}$
2. Repeat
   1. Assign each data point to its closest center

$$z_n = \arg\min_{i \in \{1,\dots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster identities $\mathbf{m}_{1:k}$
2. Repeat
   1. Assign each data point to its closest center

$$z_n = \arg\min_{i \in \{1,\dots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

   2. For each cluster, find the data point in that cluster that is closest to the other points in that cluster

$$i_k = \arg\min_{\{n:\, z_n = k\}} \sum_{\{m:\, z_m = k\}} d(\mathbf{x}_n, \mathbf{x}_m)$$

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster identities $\mathbf{m}_{1:k}$
2. Repeat
   1. Assign each data point to its closest center

$$z_n = \arg \min_{i \in \{1,\ldots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

   2. For each cluster, find the data point in that cluster that is closest to the other points in that cluster

$$i_k = \arg\min_{\{n:\, z_n = k\}} \sum_{\{m:\, z_m = k\}} d(\mathbf{x}_n, \mathbf{x}_m)$$

   3. Set each cluster center equal to their closest data points

$$m_k = \mathbf{x}_{i_k}$$

k-medoids algorithm

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster identities $\mathbf{m}_{1:k}$
2. Repeat
   1. Assign each data point to its closest center

$$z_n = \arg \min_{i \in \{1,\ldots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

   2. For each cluster, find the data point in that cluster that is closest to the other points in that cluster

$$i_k = \arg\min_{\{n:\, z_n = k\}} \sum_{\{m:\, z_m = k\}} d(\mathbf{x}_n, \mathbf{x}_m)$$

   3. Set each cluster center equal to their closest data points

$$m_k = \mathbf{x}_{i_k}$$

3. Until assignments $\mathbf{z}_{1:N}$ do not change

Choosing k

- Choosing k is a nagging problem in cluster analysis

Choosing k

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k

Choosing k

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
  - A certain required compression in VQ

Choosing k

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
  - A certain required compression in VQ
  - Clustering customers for k salespeople in a business

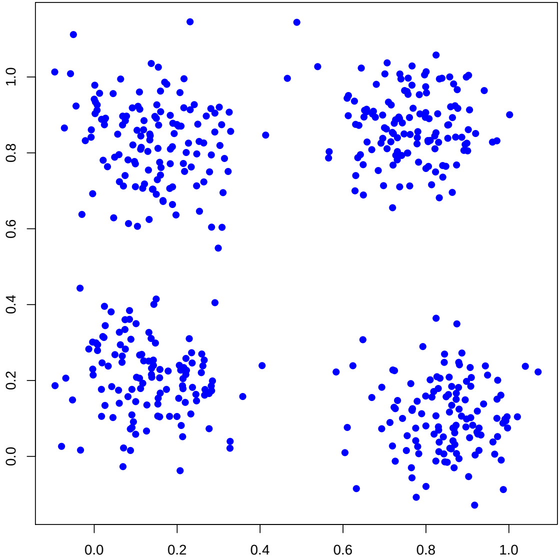
Choosing k

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
  - A certain required compression in VQ
  - Clustering customers for k salespeople in a business
- Usually, we seek the “natural” clustering, but what does this mean?

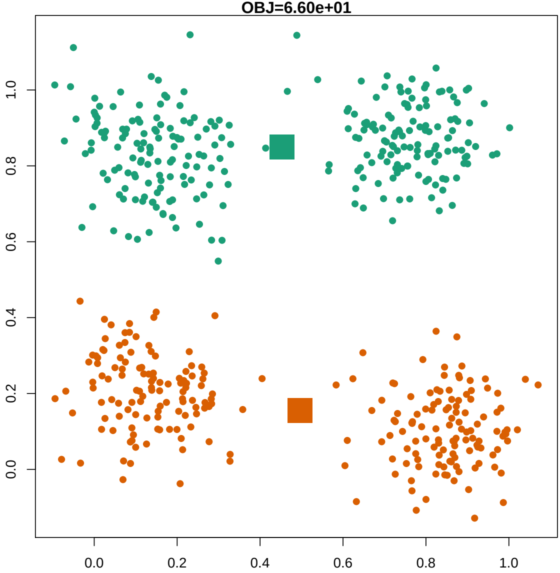
Choosing k

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
  - A certain required compression in VQ
  - Clustering customers for k salespeople in a business
- Usually, we seek the “natural” clustering, but what does this mean?
- It is not well-defined.

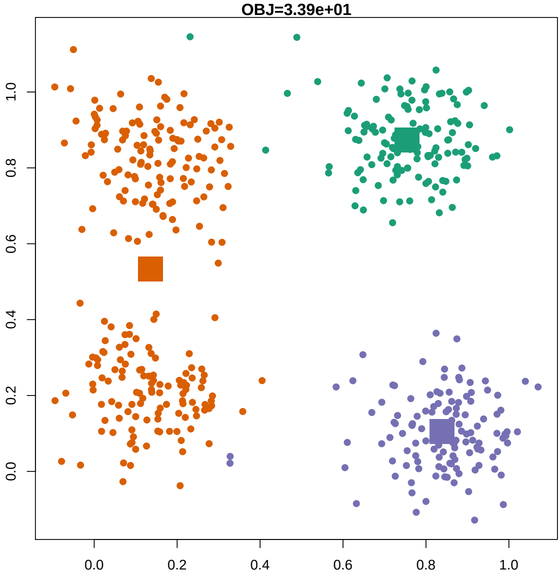
What happens as *k* increases?



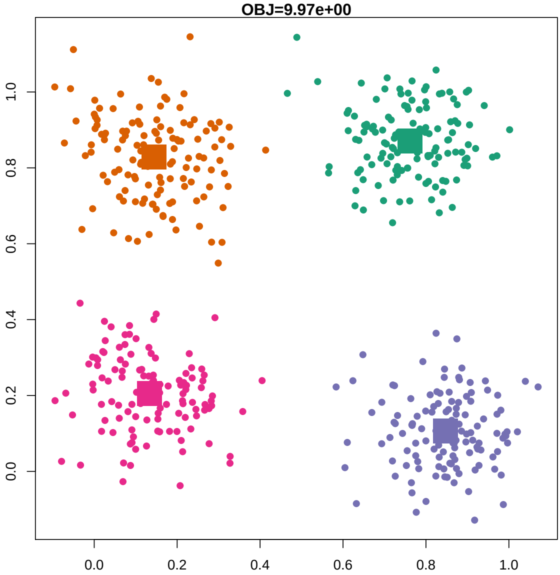
What happens as k increases?



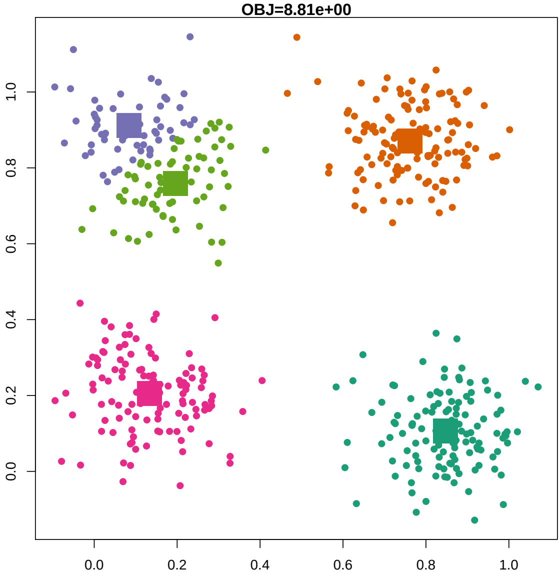
What happens as k increases?



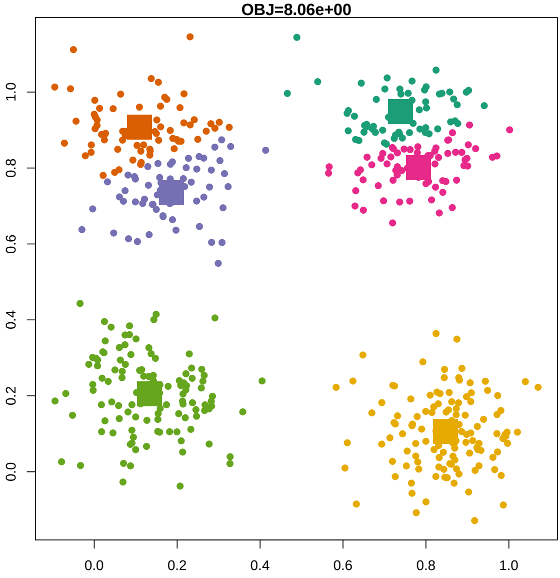
What happens as *k* increases?



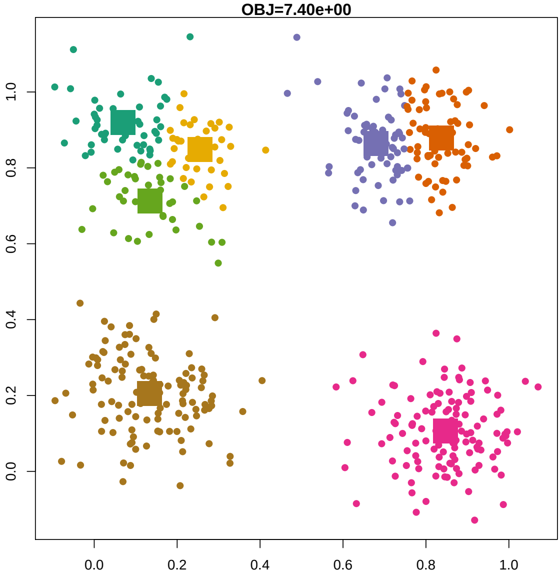
What happens as k increases?



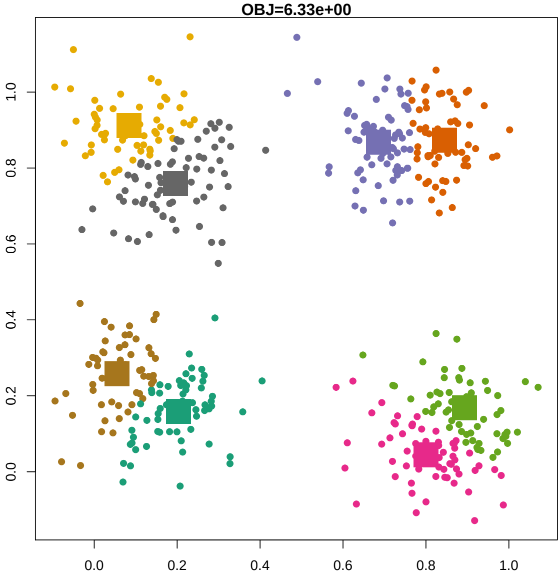
What happens as k increases?



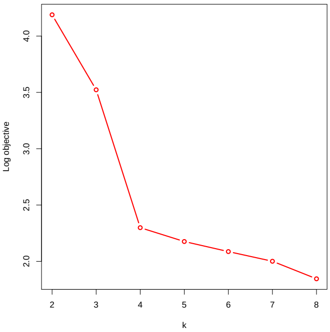
What happens as k increases?



What happens as *k* increases?

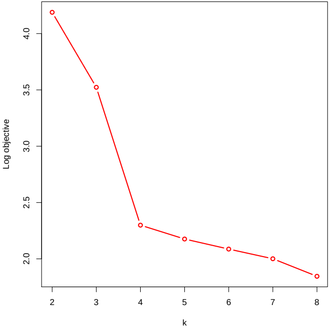


Heuristic: A kink in the objective



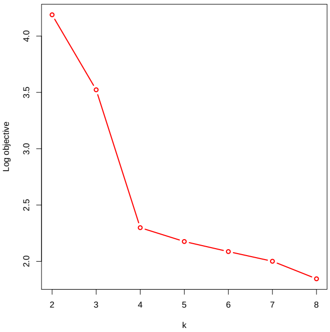
- Notice the “kink” in the objective between 3 and 5.

Heuristic: A kink in the objective



- Notice the “kink” in the objective between 3 and 5.
- This suggests that 4 is the right number of clusters.

Heuristic: A kink in the objective



- Notice the “kink” in the objective between 3 and 5.
- This suggests that 4 is the right number of clusters.
- Tibshirani (2001) presents a method for finding this kink.

Archeology

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)

Archeology

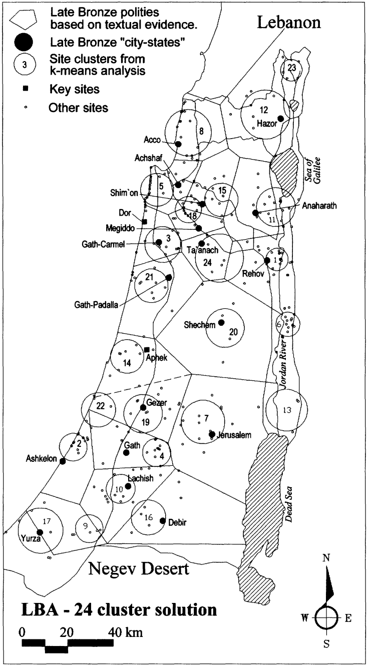
- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel

Archeology

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
- Make inferences about political history based on the clusters

Archeology

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
- Make inferences about political history based on the clusters
- Choose k very carefully, with a complicated computational technique.



Late Bronze polities
based on textual evidence.
Late Bronze "city-states"
Site clusters from
k-means analysis
Key sites
Other sites
Lebanon
Hazor
Acco
Achshaf
Sea of Galilee
Shim`on
Anaharath
Dor
Megiddo
Gath-Carmel
Ta`anach
Rehov
Gath-Padalla
Shechem
Aphek
Jordan River
Gezer
Jerusalem
Gath
Ashkelon
Lachish
Dead Sea
Debir
Yurza
Negev Desert
N
W
E
S
LBA - 24 cluster solution
0
20
40 km

- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)

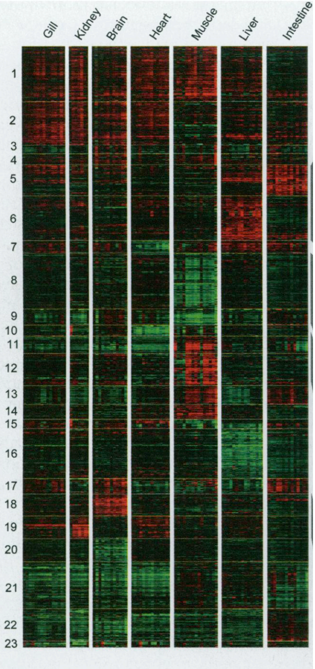
- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)
- Exposed carp to different levels of cold

Computational Biology

- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)
- Exposed carp to different levels of cold
- Clustered genes based on their response in different tissues

Computational Biology

- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)
- Exposed carp to different levels of cold
- Clustered genes based on their response in different tissues
- (No mention of how $k = 23$ was chosen.)


Gill
Kidney
Brain
Heart
Muscle
Liver
Intestine
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Education

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)

Education

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students

Education

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students
- Used the clusters to identify groups to buttress an analysis of what affects motivation.

Education

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students
- Used the clusters to identify groups to buttress an analysis of what affects motivation.
- I.e., the levels of encouragement are corrected for

Education

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students
- Used the clusters to identify groups to buttress an analysis of what affects motivation.
- I.e., the levels of encouragement are corrected for
- Chose the number of clusters to get nice results

TABLE 3. Five-Cluster Solution: Z scores on Each Clustering Variable

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Teacher caring	−.5	−.5 to .5	−.5 to .5	−.5	1.0
Peers' academic support	1.0	−.5	1.0	−.5	−.5 to .5
Parents' academic support	.5	−1.0	−.5 to .5	−.5 to .5	1.0

TABLE 4. Means and Standard Deviations for Each Cluster on Grade 8 Motivational Variables

	Academic Self-Efficacy		Intrinsic Valuing of Education		Teacher-Rated Effort	
Cluster	*M*	SD	*M*	SD	*M*	SD
1. All positive	3.59	.48[a]	2.99	.55[a]	3.74	.26[a]
2. Peer negative, parents very negative	2.44	.66[b]	2.16	.51[b]	3.05	.61[b]
3. Peer positive	3.01	.73[c]	2.43	.66[b]	3.26	.66[b]
4. Negative teacher and peer	2.47	.63[b]	2.24	.51[b]	3.17	.59[b]
5. Positive teacher and parents	3.19	.65[c]	2.89	.62[a]	3.54	.47[a]

Sociology

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)

Sociology

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity

Sociology

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
- K chosen for interpretability and "stability," which means that they could interpret multiple k-means runs on different data in the same way.

Sociology

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
- K chosen for interpretability and "stability," which means that they could interpret multiple k-means runs on different data in the same way.
- Draw the conclusion that patterns exist. What's wrong with this?

Sociology

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
- K chosen for interpretability and "stability," which means that they could interpret multiple k-means runs on different data in the same way.
- Draw the conclusion that patterns exist. What's wrong with this?
- k-means will find patterns everywhere!

TABLE 2. Percentage distribution of participants, by cluster, and behavioral patterns defining each cluster

Cluster type and behavioral patterns	%
Light substance dabblers—infrequent or no current use of substances† None have had sex	24.4
Abstainers—none have ever used substances† or had sex	22.7
Sex dabblers—all have had sex Median no. of partners=1 60% used a condom at last sex Infrequent use of substances†	14.5
Drinkers—all consumed alcohol in past 12 mos. 49% report binge drinking Infrequent or no illicit drug use None have had sex	7.4
Smokers—all smoke cigarettes daily Infrequent use of alcohol/illicit drugs 62% have had sex	7.3
Alcohol-and-sex dabblers—all drink occasionally; all have had sex Infrequent tobacco/illicit drug use	5.4
Binge drinkers—all binge frequently Infrequent cigarette, marijuana and other drug use 60% binge ≥1 time/wk. 45% have had sex	4.4
Heavy dabblers—all smoke, drink and binge drink with moderate frequency 45% use marijuana; few use other illicit drugs 91% have had sex	3.6
Combination sex and drug use—all have had sex; all used alcohol/illicit drug at last sex	3.4
Marijuana users—all use marijuana frequently; few have used other illicit drugs 94% use alcohol 79% smoke cigarettes 74% have had sex	1.7
Multiple partners—all report ≥14 sexual partners 75% report low or moderate use of substances†	1.3
Sex for drugs or money—all have had sex for drugs or money 50% report low or moderate use of substances† Median no. of partners=3	1.2
High marijuana use and sex—all use marijuana frequently; all have had sex All used alcohol/other drug at last sex 82% have had >1 partner (median=6)	1.1
Marijuana and other drug users—95% report heavy marijuana use; all use other illicit drugs 68% have had sex 28% used alcohol/other drug at last sex	0.6
Injection-drug users—all have injected drugs 82% have had sex Median no. of partners=4	0.6
Males who have sex with males—all are males who have had sex with another male 78% have had multiple partners (median=5) 40% used marijuana in past 30 days 50% use alcohol ≥1 time/mo. 17% have had sex for drugs or money	0.3

Summary