EC999: K-Means Clustering

Thiemo Fetzer

University of Chicago & University of Warwick

January 29, 2018

Clustering

- So far we have talked about numerical prediction/ classification and some methods to help you arrive at robust models. The focus was on understasnding $Y|X$
- The last section of the course talks about dimensionality reduction in the broadest sense.
- Dimensionality reduction can be thought of as reducing patterns in $\mathbf{X}$ and mapping them to a lower dimensonality subspace; such patterns could be group membership.
- Clustering is a form of dimensionality reduction: we want to group data together that is "similar" by some metric.

Clustering Example: Google News

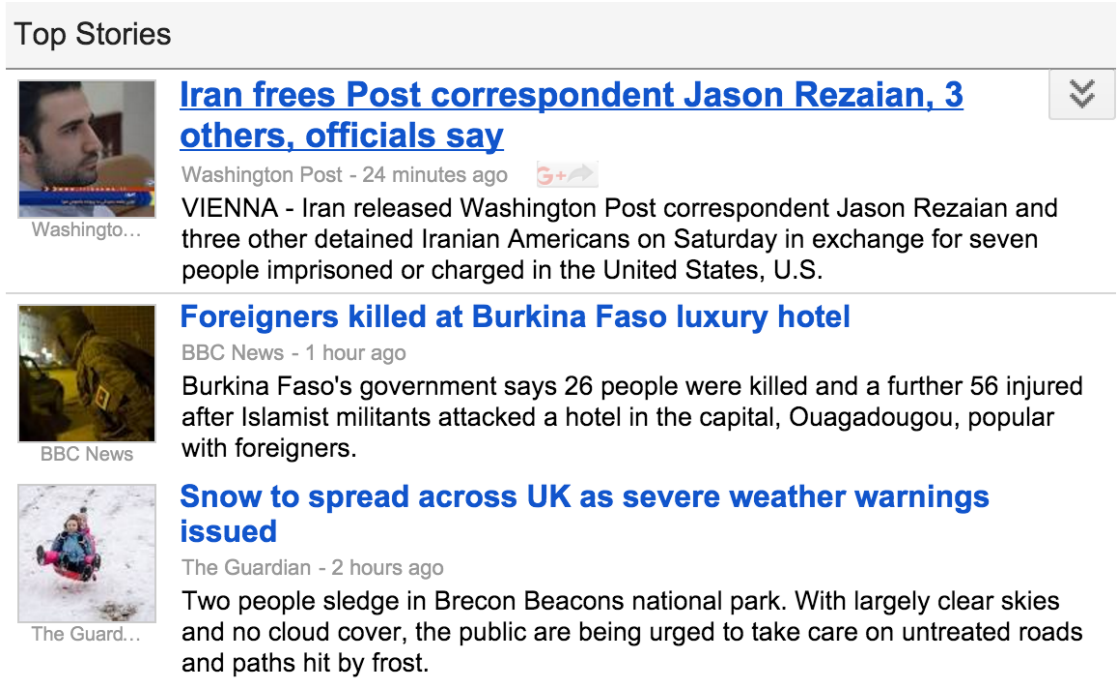


Figure: Google News as Example of Cluster

Clustering Example: Google News

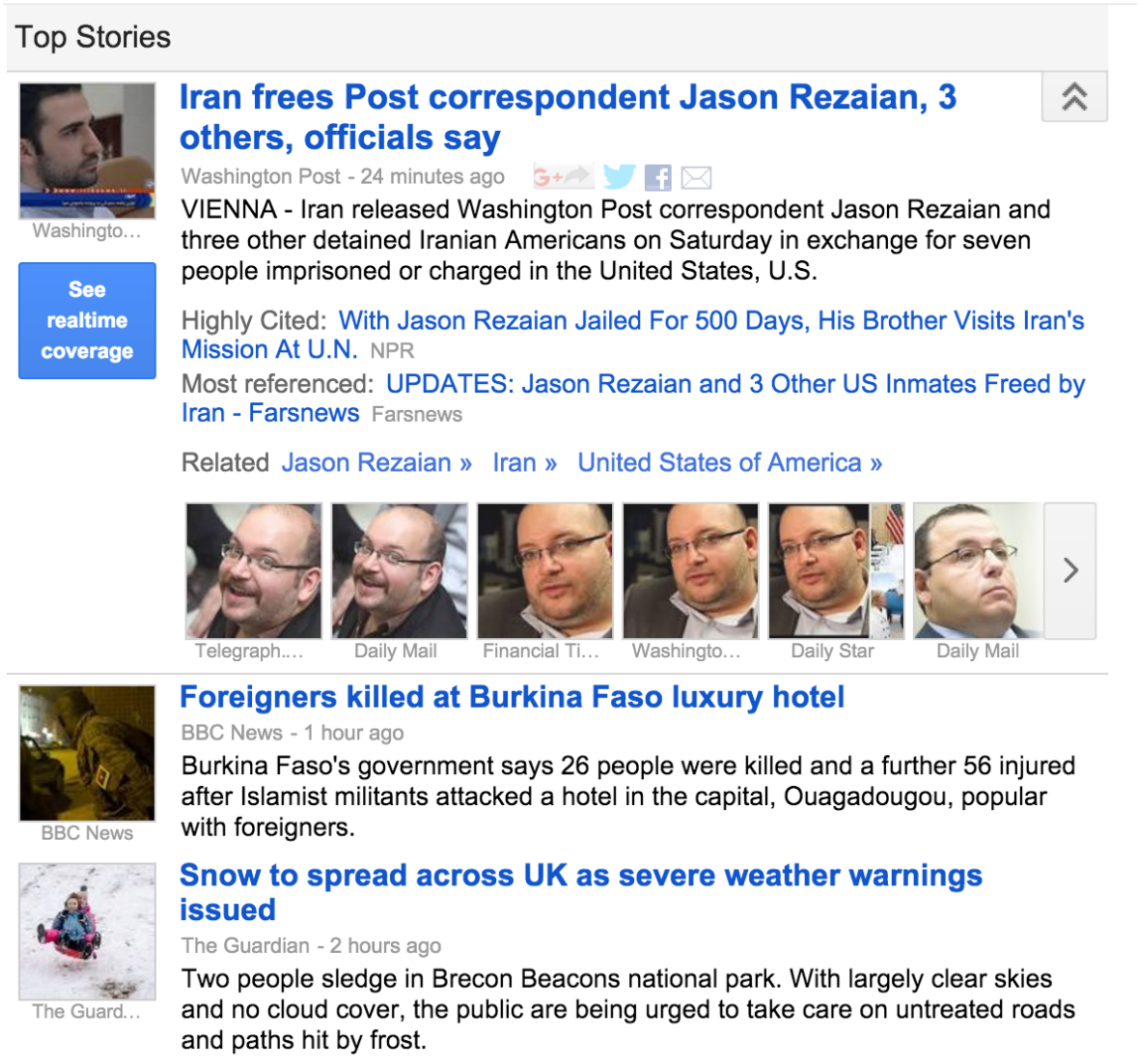
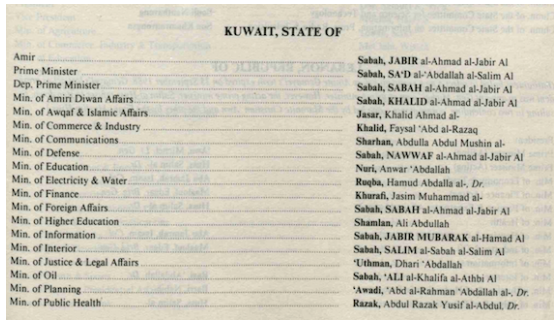


Figure: Google News as Example of Cluster

Clustering Example: Data Cleaning

KUWAIT

AMIR	AL-SABAH, SABAH AL-SALIM
PREMIER	AL-SABAH, JABIR AL-AHMAD
MIN OF AWQAF & ISLAMIC AFFAIRS	AL-RAWDHAN, ABDULLAH AL-MISHARI
MIN OF COMMERCE & INDUSTRY	AL-SABAH, ABDULLAH AL-JABIR
MIN OF EDUCATION	AL-FUHAYD, KHALID AL-MASUD
MIN OF ELECTRICITY & WATER RESOURCES	AL-SUMAYT, ABDULLAH AHMAD
MIN OF FINANCE & OIL	AL-SABAH, SABAH AL-AHMAD - ACTING
MIN OF FOREIGN AFFAIRS	AL-SABAH, SABAH AL-AHMAD
MIN OF INFORMATION & GUIDANCE	AL-SABAH, JABIR AL-ALI
MIN OF INTERIOR & DEFENSE	AL-SABAH, SAD AL-ABDULLAH
MIN OF JUSTICE	AL-JASSAR, KHALID AL-AHMAD
MIN OF POST TELEPHONE & TELEGRAPH	AL-SALIH, SALIH ABD AL-MALIK
MIN OF PUBLIC HEALTH	AL-FULAYJ, ABD AL-AZIZ IBRAHIM
MIN OF PUBLIC WORKS	AL-SALIH, KHALID AL-ISA
MIN OF SOCIAL AFFAIRS & LABOR	AL-SARAWI, ABD AL-AZIZ ABDULLAH
MIN OF ST FOR PREMIERSHIP AFFAIRS	AL-RIFAI, YUSUF HASHIM



KUWAIT, STATE OF

Amir	**Sabah, JABIR** al-Ahmad al-Jabir Al
Prime Minister	**Sabah, SA'D** al-'Abdallah al-Salim Al
Dep. Prime Minister	**Sabah, SABAH** al-Ahmad al-Jabir Al
Min. of Amiri Diwan Affairs	**Sabah, KHALID** al-Ahmad al-Jabir Al
Min. of Awqaf & Islamic Affairs	**Jasar,** Khalid Ahmad al-
Min. of Commerce & Industry	**Khalid,** Faysal 'Abd al-Razaq
Min. of Communications	**Sharhan,** Abdulla Abdul Mushin al-
Min. of Defense	**Sabah, NAWWAF** al-Ahmad al-Jabir Al
Min. of Education	**Nuri,** Anwar 'Abdallah
Min. of Electricity & Water	**Ruqba,** Hamud Abdalla al-, *Dr.*
Min. of Finance	**Khurafi,** Jasim Muhammad al-
Min. of Foreign Affairs	**Sabah, SABAH** al-Ahmad al-Jabir Al
Min. of Higher Education	**Shamlan,** Ali Abdullah
Min. of Information	**Sabah, JABIR MUBARAK** al-Hamad Al
Min. of Interior	**Sabah, SALIM** al-Sabah al-Salim Al
Min. of Justice & Legal Affairs	**'Uthman,** Dhari 'Abdallah
Min. of Oil	**Sabah, 'ALI** al-Khalifa al-Athbi Al
Min. of Planning	**'Awadi,** 'Abd al-Rahman 'Abdallah al-, *Dr.*
Min. of Public Health	**Razak,** Abdul Razak Yusif al-Abdul, *Dr.*

Figure: Raw Data

MIN OF FOREIGN AFFAIRS AL SABAH JABIR AL AHMAD AL JABIR
Min of Foreign Affairs Sabah Jabir al Ahmad al Jabir al
Min of Foreign Sabah Sabah al Ahmad al Jabir al
Min of Foreign Affairs Sabah Sabah Ahmad Jabir al
Min of Foreign Affairs Sabah Sabah Ahmad Jabir Al
Min of Foreign Affairs Sabah sabah al Ahmad al Jabir Al
Min of Foreign Affairs Sabah Sabah al Ahmad al Jabir
Min of Foreign Affairs Sabah SABAH al Ahmad al Jabir Al
Min of Foreign Affairs Sabah SABAH al Ahmad al Jabir al

Clustering Example: Data Cleaning

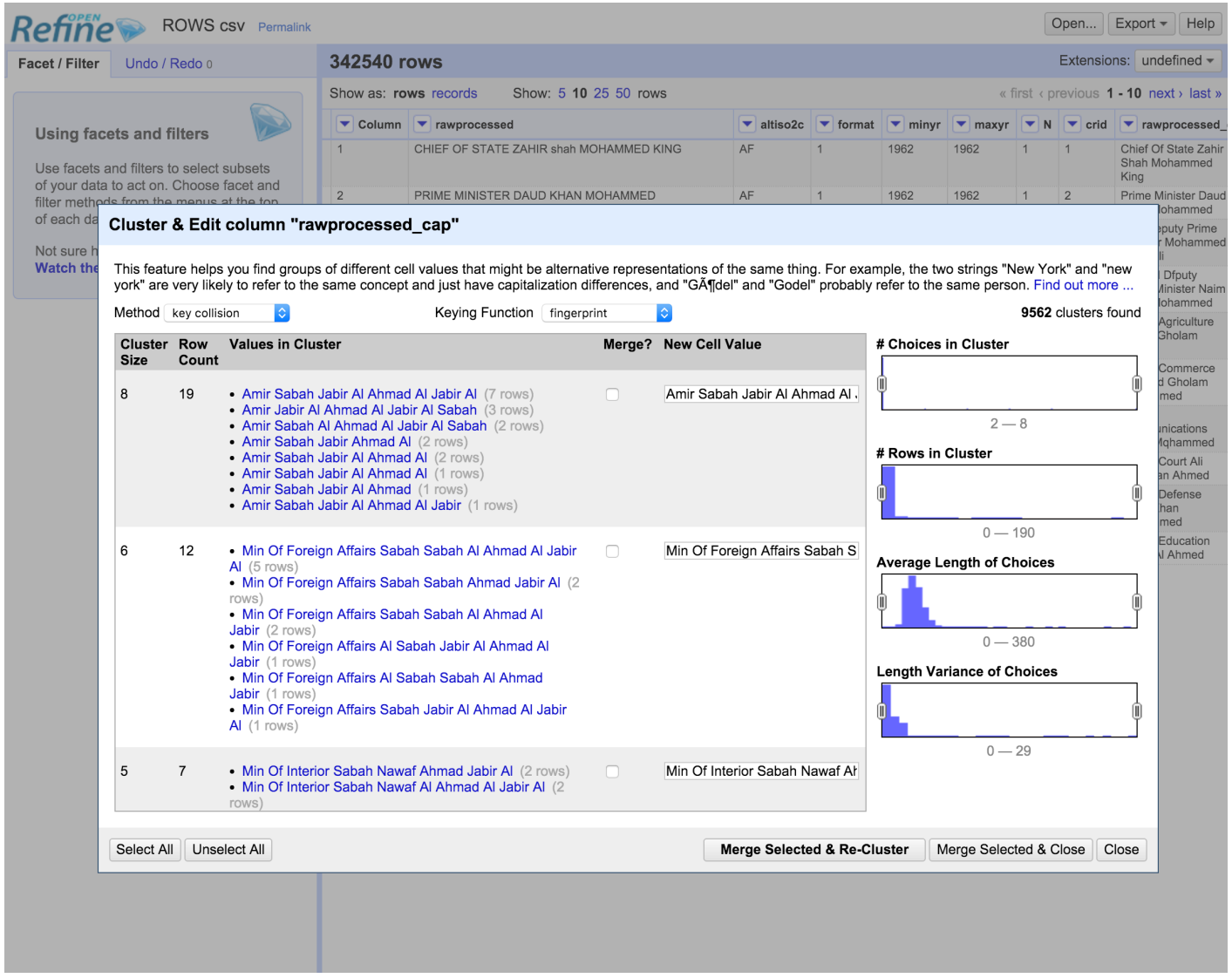
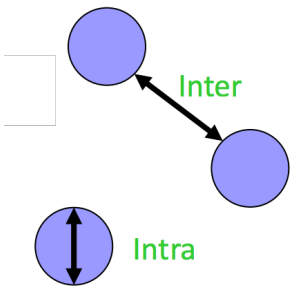


Figure: Open Refine for Working with Messy Data: Clustering Text

What are good clusters?

- A good clustering method will provide...
  - High Intra cluster similarity: cohesive within clusters
  - Low Inter class similarity: distinctive between clusters.



- The quality of a clustering method depends on
  - the similarity measure used by the method
  - its implementation, and
  - Its innate ability to discover hidden patterns.

Plan

K-Means Clustering

K-medoids Clustering

K-Means: One dimensional case

- Suppose $\mathbf{X}$ is one dimensional.

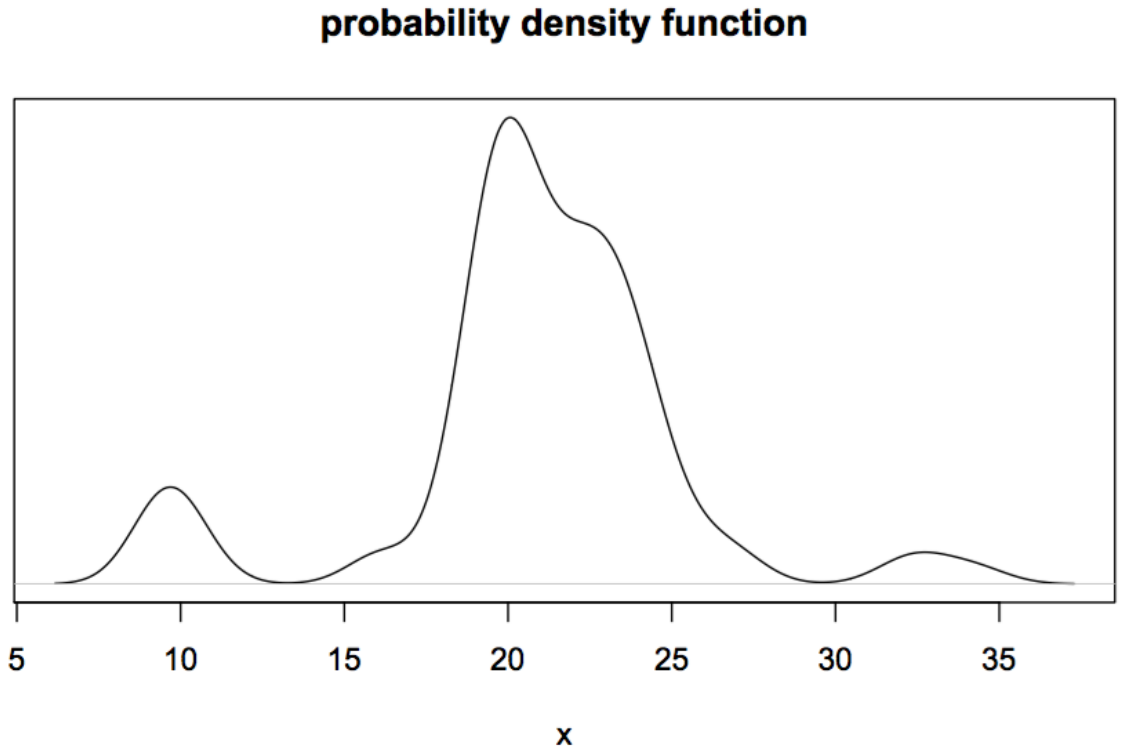


Figure: Kernel Density of Hypothetical Mixture Distribution of five Normals, taken from Matt Taddy.

K-Means: One dimensional case

- Suppose $\mathbf{X}$ is one dimensional.

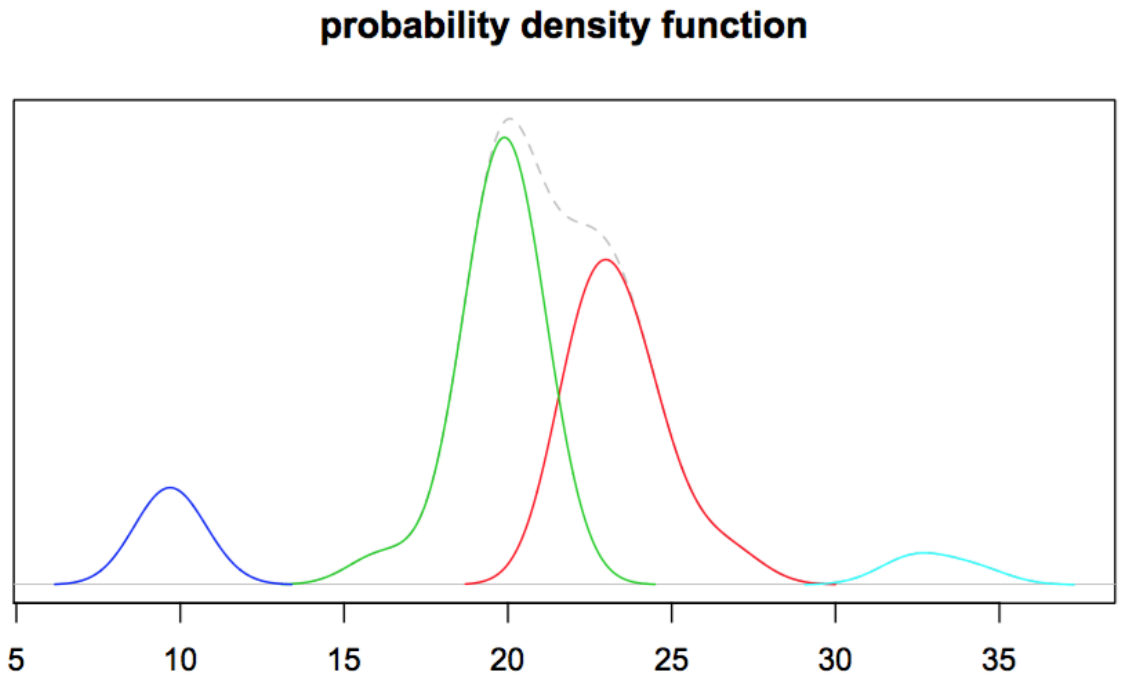


Figure: Kernel Density of Hypothetical Mixture Distribution of five Normals, taken from Matt Taddy.

K-Means: One dimensional case

- Suppose $\mathbf{X}$ is one dimensional.

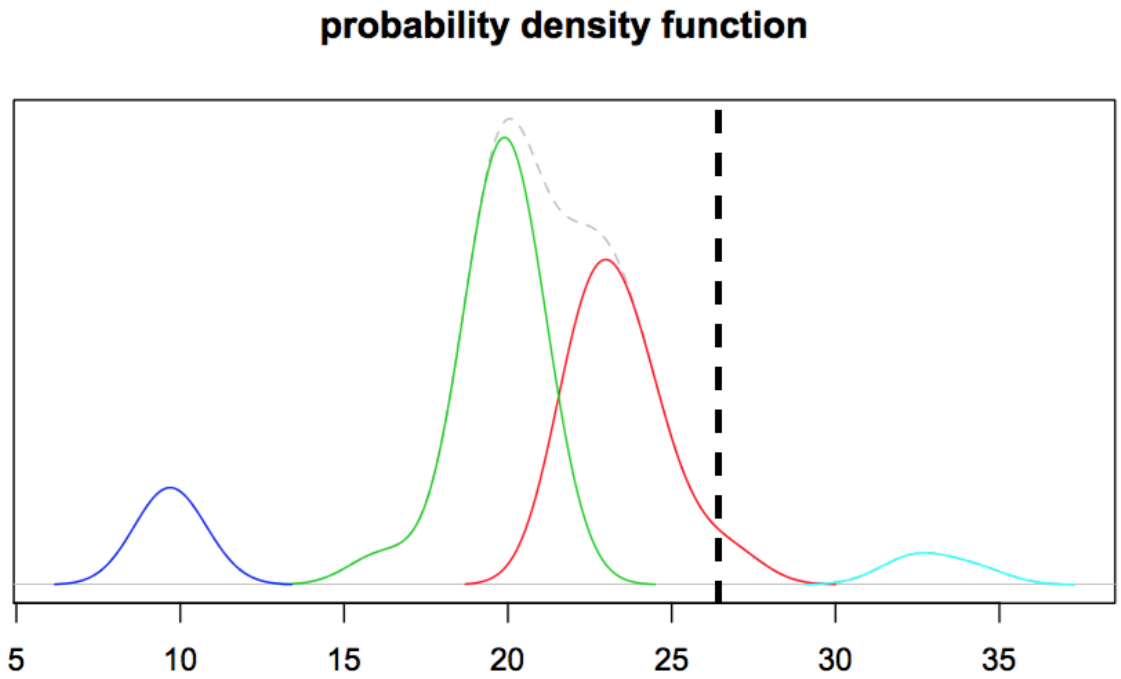


Figure: Kernel Density of Hypothetical Mixture Distribution of five Normals, taken from Matt Taddy.

K-Means Clustering

- Let $C_1, ...C_K$ denote sets containing the indices of the observations in each cluster, i.e. each $\mathbf{x_j}$ is assigned to a cluster C_i.
- the partitioning of the data is non-overlapping and exhaustive:

$$C_1 \cup C_2 \cup ... \cup C_K = \{1, ..., n\}$$
$$\text{for } k \neq k' : C_k \cap C_{k'} = \emptyset$$

- K-Means clustering attempts to minimize the *within-cluster* variation, that is, choose an assignment rule $C_1, ...C_K$ such that:

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

- where we can measure distance between points within a cluster using squared Euclidian distance ($\mathcal{L}_2$ norm).

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

K-Means Within Cluster Distance

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

- Compute all pairwise distances between all points $i, i' \in C_k$ for all k.
- There is a trick, you can show that:

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

[Proof (idea) is simple/ draw]: Suppose two points on uniform line belonging to a single cluster, i.e. x_1, x_2, then the above relationship is:

$$\underbrace{\frac{1}{2}[(x_1 - x_2)^2 + (x_2 - x_1)^2]}_{\text{all pairwise combinations}} = 2[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2]$$

K-Means Clustering

- Overall objective function can be rewritten as

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

- Using $\mathcal{L}_2$ norm notation

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \bar{x}_k\|_2^2$$

- So what are we trying to optimize? By choosing assignments of points $\mathbf{x_1}, \ldots \mathbf{x_n}$ to K clusters, we want to minimize the across cluster sum of squared deviations.
- This is a very difficult problem to solve... for a given number of K topics, there are K^n ways of partitioning the data into K clusters.
- However, there are algorithms that help find *locally optimal* solutions.

K-Means Clustering Algorithm

Algorithm (*K-Means Clustering Algorithm*)

1. *Initialize by randomly assigning each observation to a cluster K, i.e. create at random $C_1, ..., C_K$.*
2. *Continue the following steps, and only stop when the assignments of observations to clusters $C_1, ..., C_K$ does not change anymore:*

   a) For each cluster $k \in K$, compute the centroid $\bar{\mathbf{x}}_\mathbf{k}$.

   b) Assign each observation to the cluster whose centroid is closest (as defined by Euclidian distance).

K-Means: Two dimensional algorithm illustration

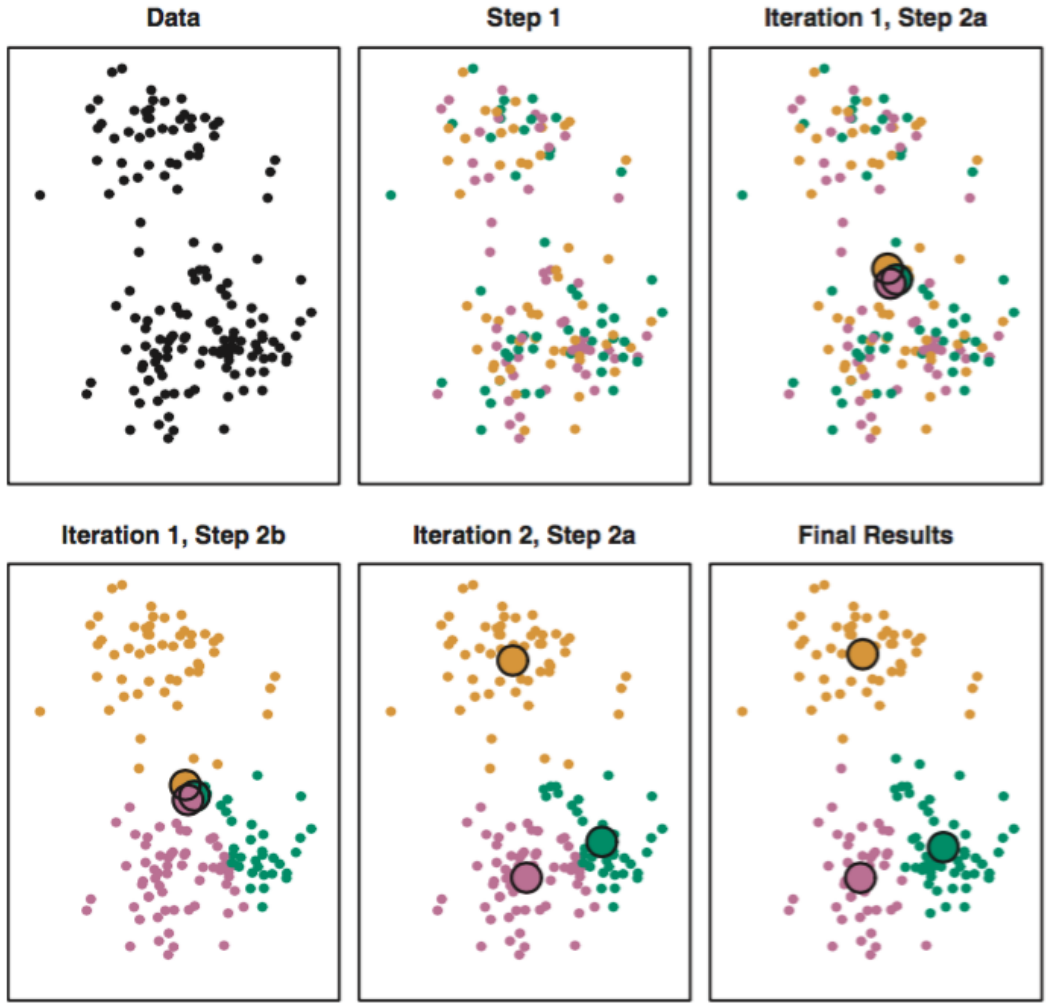


Figure: Two dimensional example illustrating sequence of iterations.

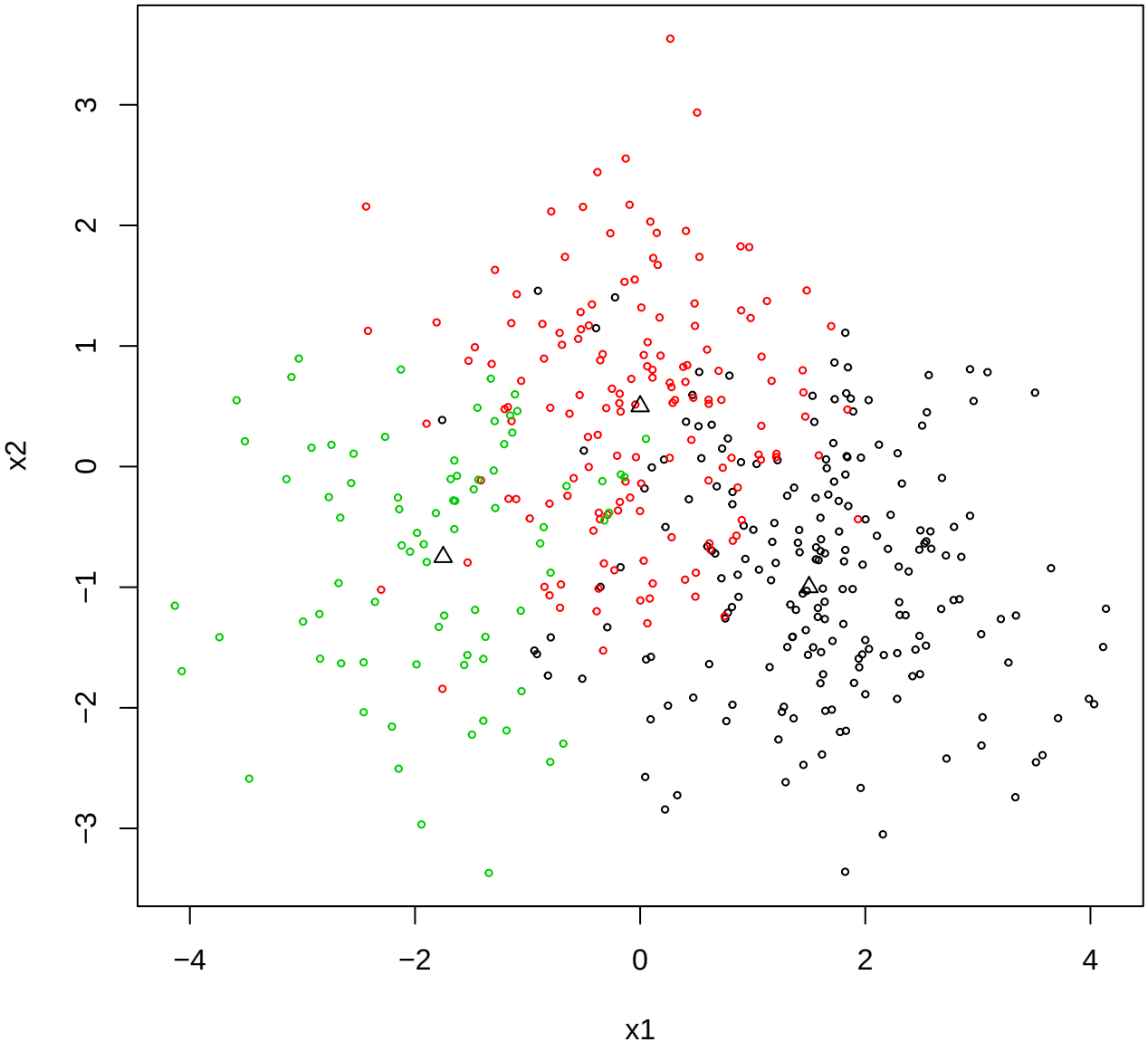
What guarantees convergence?

- At each iteration step, the objective function will improve or stay the same, why?

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

- In Step 2(a) the cluster means for each feature are the constants that minimize the sum-of-squared deviations, and in Step 2(b), reallocating the observations can only improve.
- This means that as the algorithm is run, the clustering obtained will continually improve until the result no longer changes; the objective function will never increase.
- However, the convergence is to a *local minimum*, not a global minimum: what should we do?

K-Means: Choosing K



K-Means: Choosing K

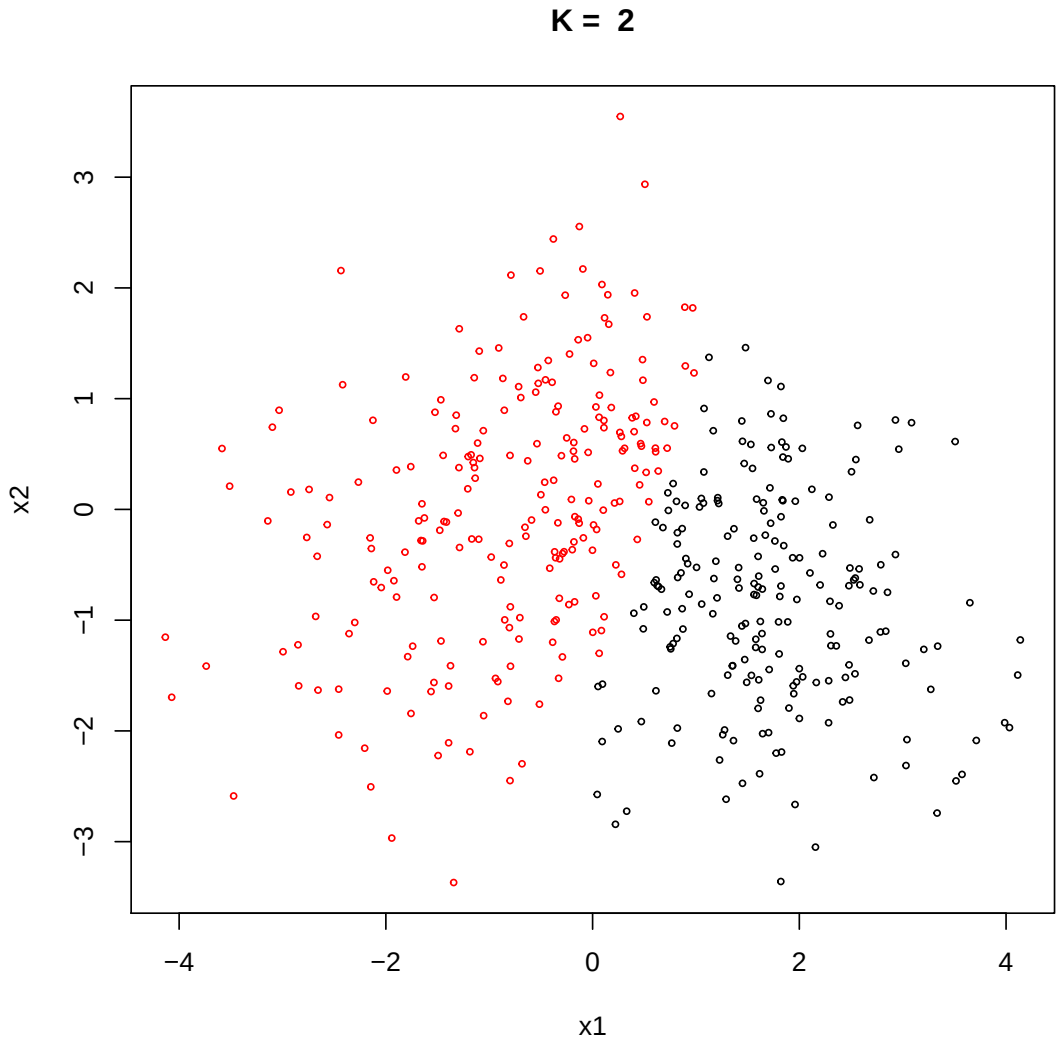


Figure: Example K-Means on simulated data with different values for K.

K-Means: Choosing K

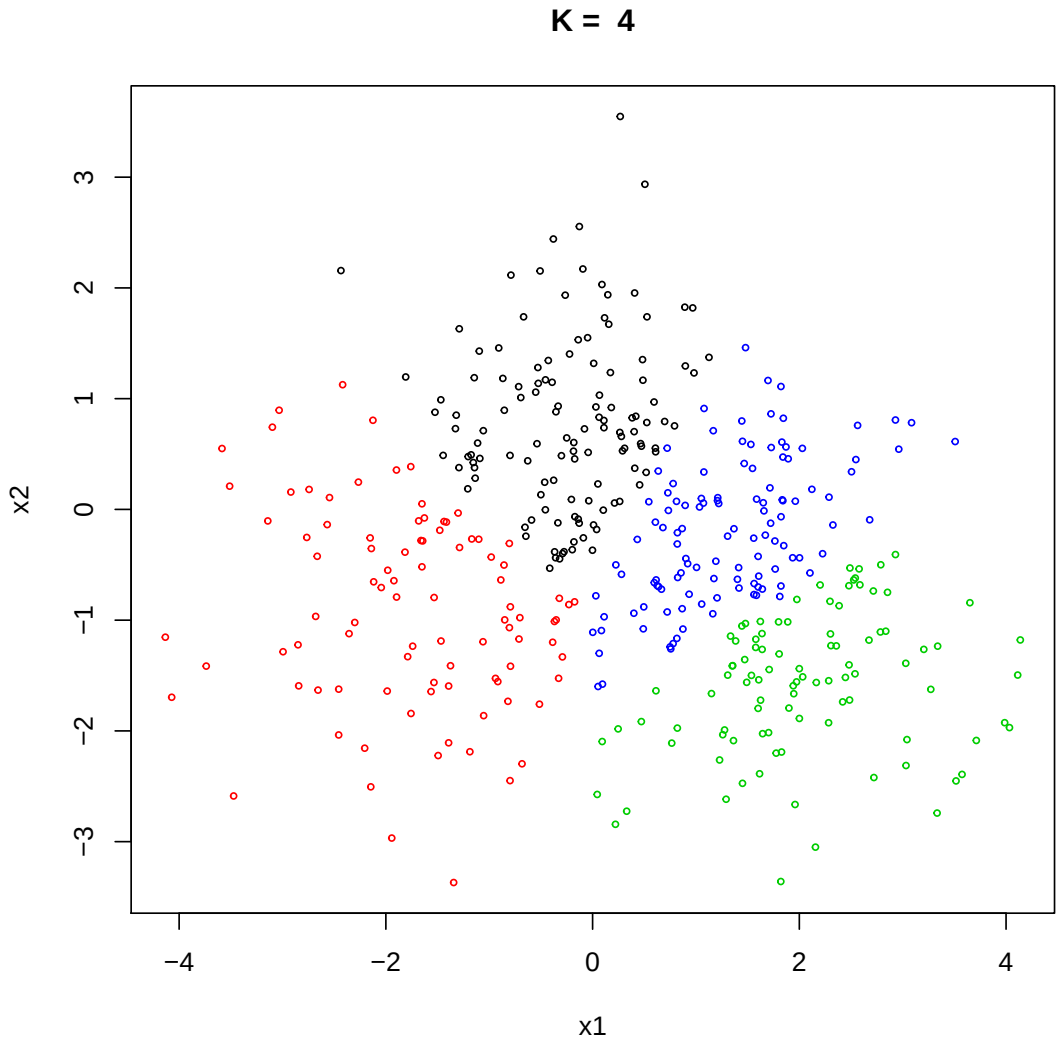


Figure: Example K-Means on simulated data with different values for K.

K-Means: Choosing K

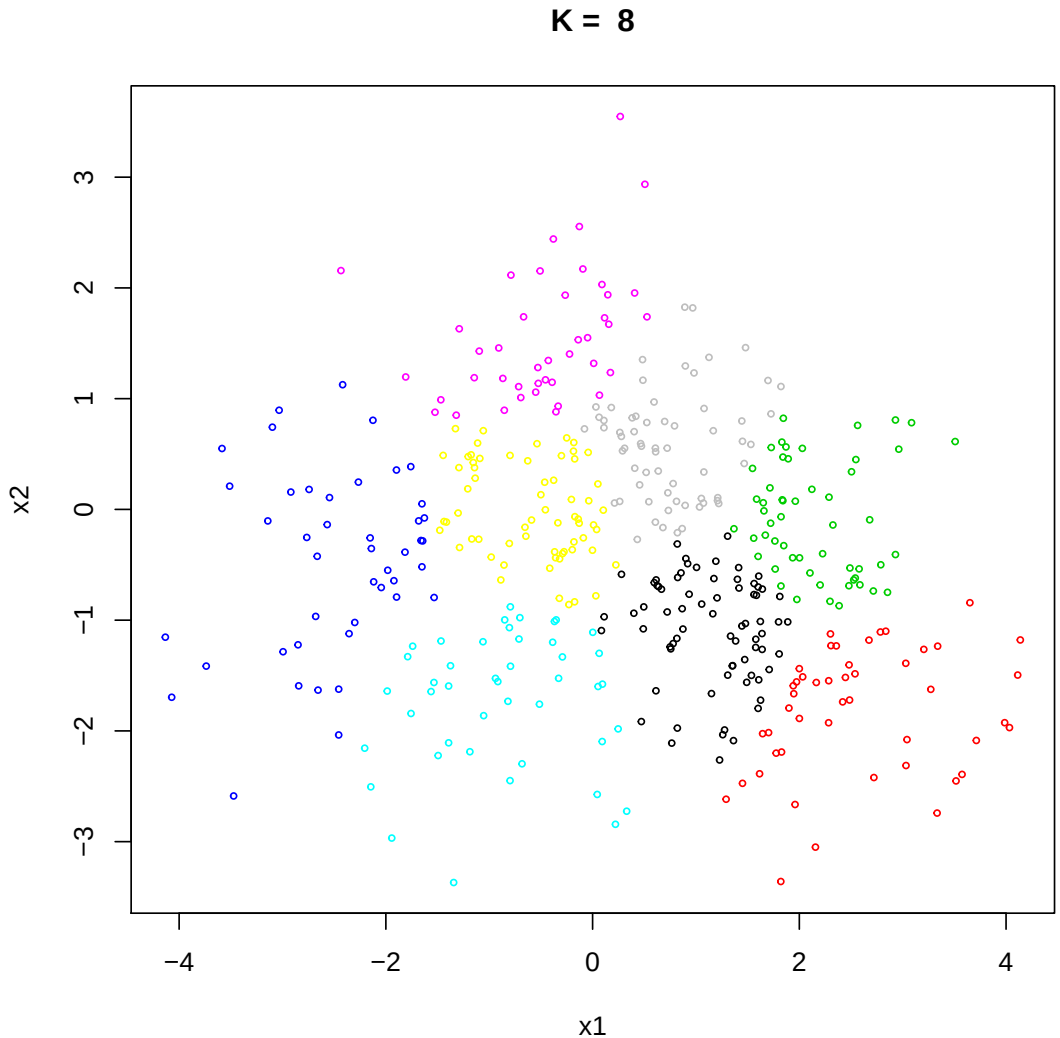


Figure: Example K-Means on simulated data with different values for K.

How to optimally choose the number of clusters K?

- Suppose you set $K = N$: What is our objective function value?

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

How to optimally choose the number of clusters K?

- Suppose you set $K = N$: What is our objective function value?

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

- It would be zero!

How to optimally choose the number of clusters K?

- Suppose you set $K = N$: What is our objective function value?

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

- It would be zero!
- Is there a way that we can validate the choice of K? No validation set, because we dont know the true number of clusters or have a good benchmark.
- Visual selection using the previous scree plot
- Gap statistic is theoretically motivated in Hastie et al. (2001). "Estimating the number of clusters in a data set via the gap statistic."

How to optimally choose the number of clusters K?

- Suppose you set $K = N$: What is our objective function value?

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

- It would be zero!
- Is there a way that we can validate the choice of K? No validation set, because we dont know the true number of clusters or have a good benchmark.
- Visual selection using the previous scree plot
- Gap statistic is theoretically motivated in Hastie et al. (2001). "Estimating the number of clusters in a data set via the gap statistic."
- Some rules of thumb suggest $K = \sqrt{n/2}$, where n is the number of observations.

Evolution of Objective function as function of K

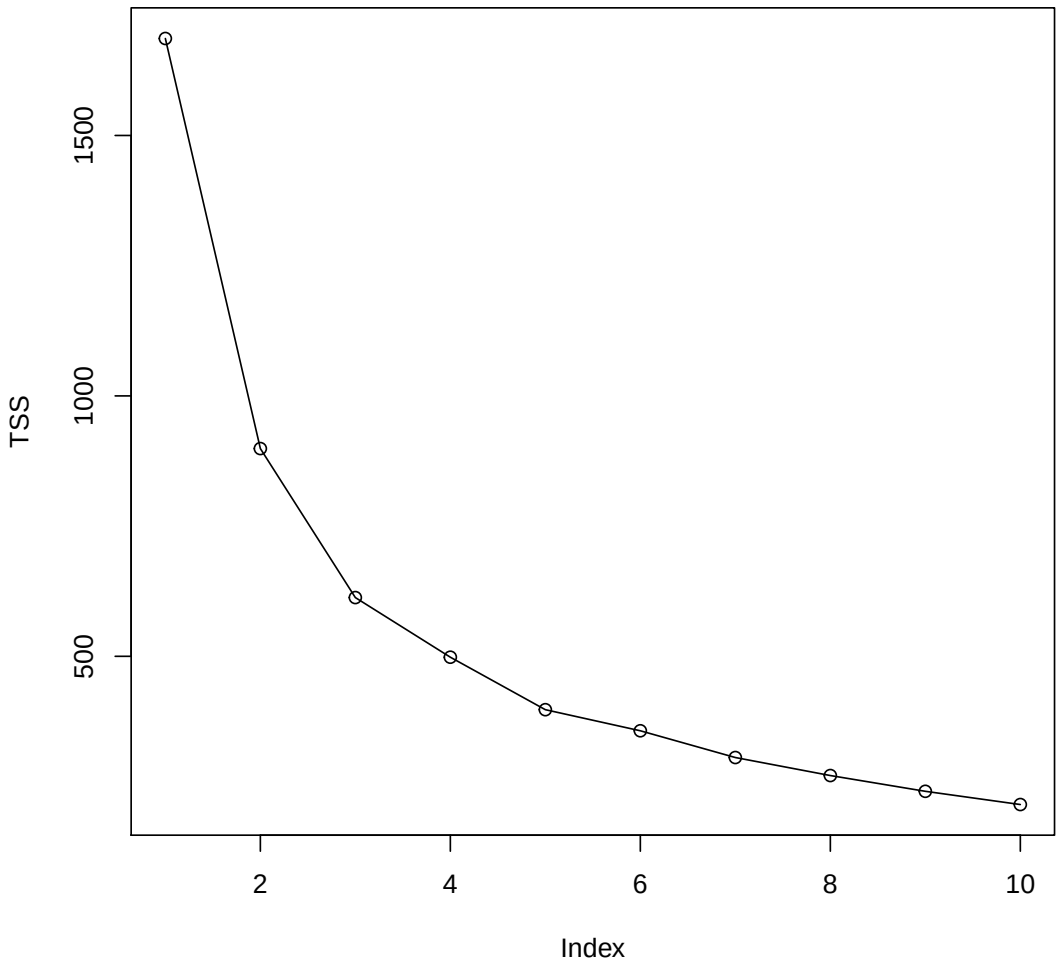


Figure: Scree Plot and Elbow points.

Evolution of Objective function as function of K

Figure: Scree Plot and Elbow points.

Evolution of Objective function as function of K

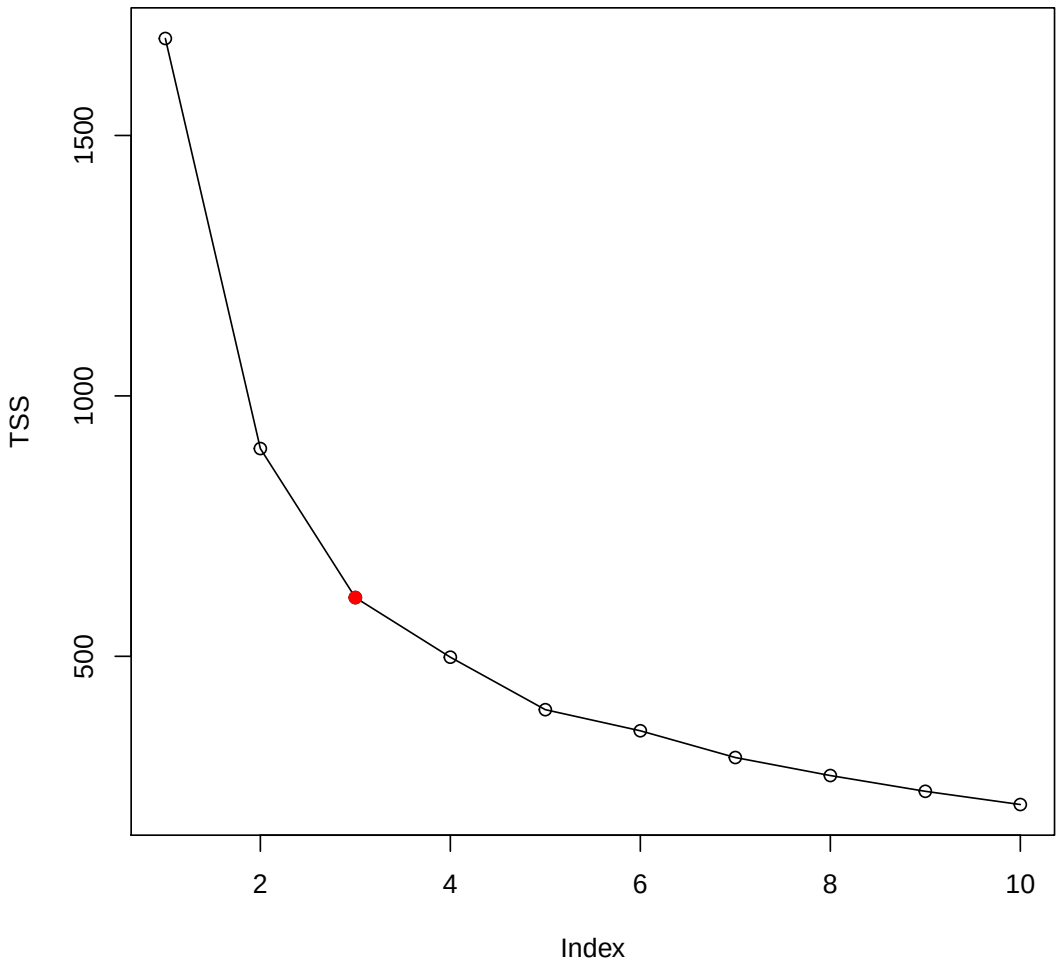


Figure: Scree Plot and Elbow points.

Formal(ish) intuition for Elbow points

- Scree plot: Mountain peak stops, rubble starts.
- At elbow point, here $K = 3$, the rate at which objective function changes, changes a lot.
- Think of the second derivative of objective function" elbow point is the point where the second derivative reaches a local maximum.

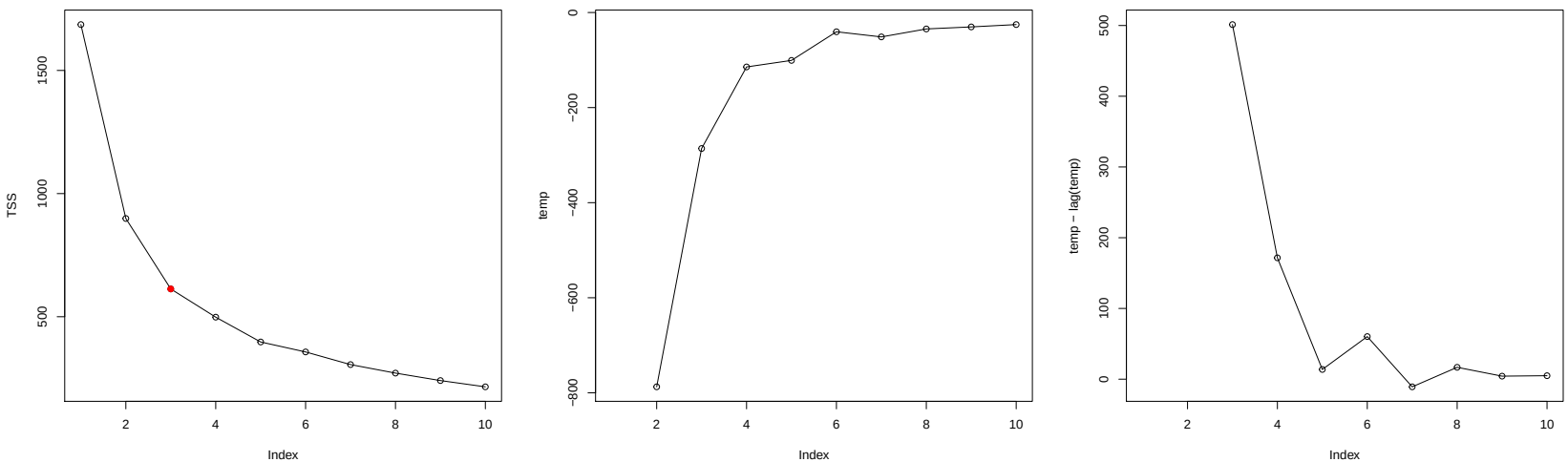


Figure: Scree Plot and Elbow points.

K-Means Clustering: An application from research

- In Fetzer and Marden (2016), we study the impact of conservation units in Brazil on land related conflict.
- The idea in a nutshell: conservation units assign property rights; once property rights are assigned, people need not fight over property titles to land.
- Categorize land into three classes $\{F, S, C\}$, Forested, Shrubland, Cropland.

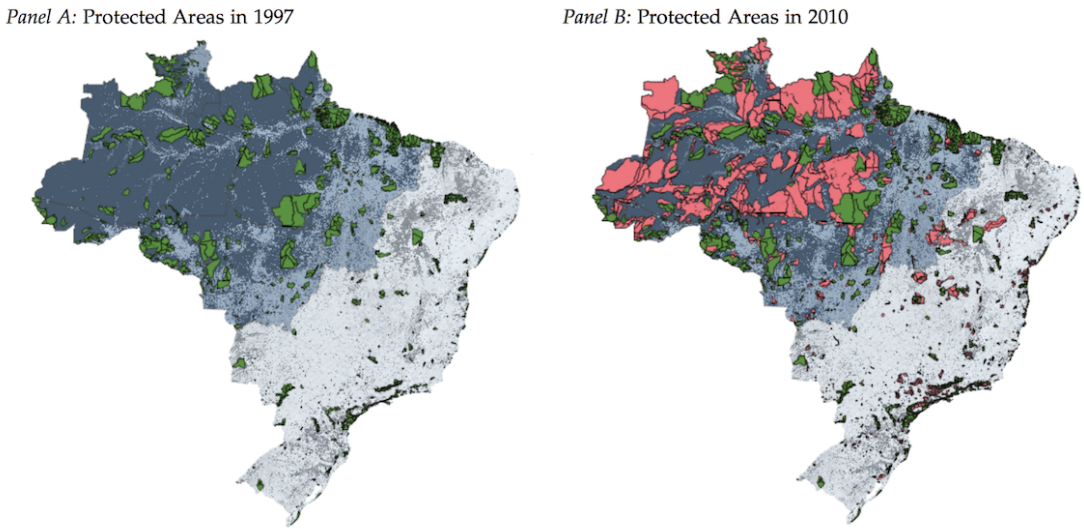


Figure:

K-Means Clustering: An application from research

- In Fetzer and Marden (2016), we study the impact of conservation units in Brazil on land related conflict.
- The idea in a nutshell: conservation units assign property rights; once property rights are assigned, people need not fight over property titles to land.
- Categorize land into three classes $\{F, S, C\}$, Forested, Shrubland, Cropland.

Figure:

K-Means Clustering: An application from research

- In Fetzer and Marden (2016), we study the impact of conservation units in Brazil on land related conflict.
- The idea in a nutshell: conservation units assign property rights; once property rights are assigned, people need not fight over property titles to land.
- Categorize land into three classes $\{F, S, C\}$, *F*orested, *S*hrubland, *C*ropland.

Figure:

K-Means Clustering: An application from research

- We have around 12 years of data and the time series tells us something about how each land pixel is being used; we want to look how conservation units affect *within pixel* land use.
- Always forested FFFFFFF, Illegal logging FFFSFFF, Land clearing then rotating crop use FFFSCSCSC, etc.
- Conservation units, without enforcement, should only affect stationary banditry.
- We want to cluster the data into three groups: roving bandits, stationary bandits and unused. So dimensionality reduction from $4^5 = 1,024$ possible combinations to three clusters.
- We create some numeric features representing five letter length series: dummy variable if pixel ever C, length of repeating state, length of repeating state pair (e.g. SCSC has SC repeating twice), number of times pixel is forested.

K-Means Clustering: An application from research

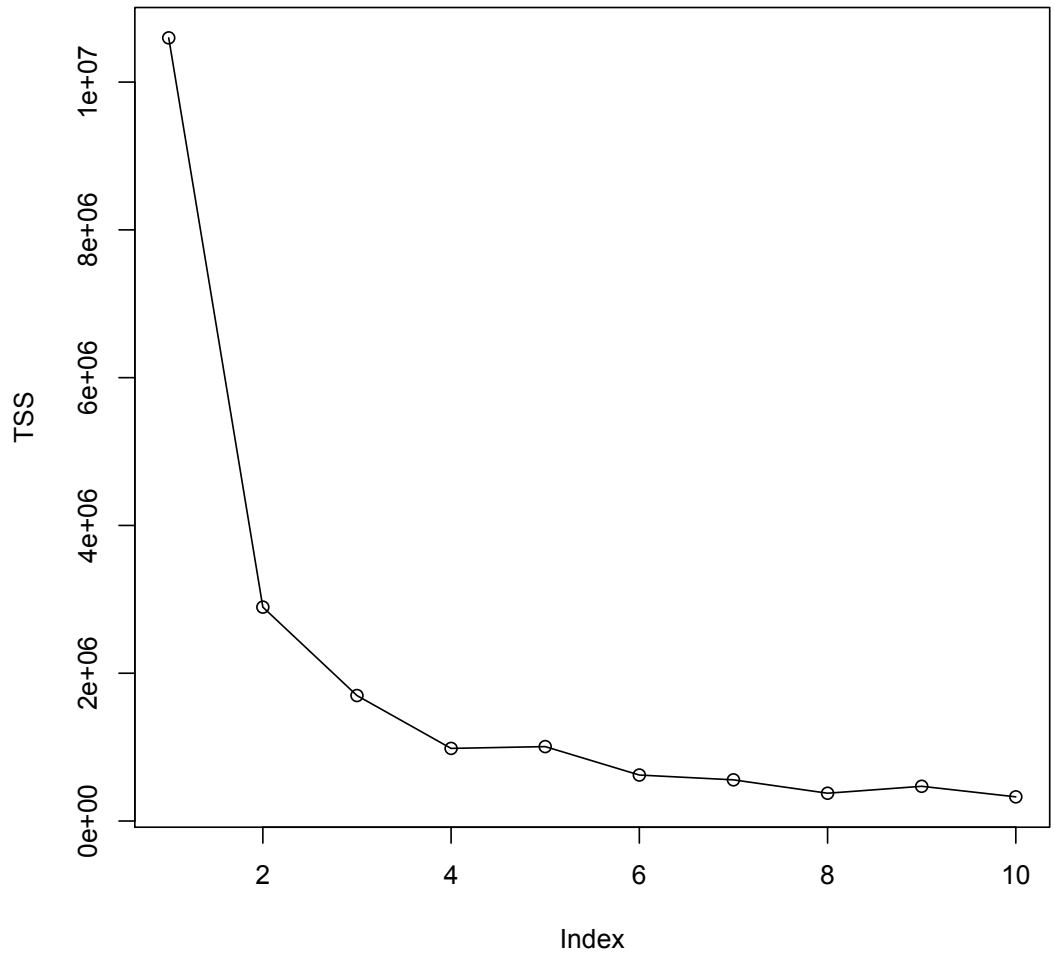


Figure: Separation Achieved to Cluster Sequences of Land Use Patterns

K-Means Clustering: An application from research

Table A1: K-Means Clusters and Separation on Features

		Clustering Variables									
Cluster	Interpretation	to F	from F	Ever C	len(F)	len(S)	len(repeat S, C)	len(repeat F)	repeat pairs	Freq	
1	Stationary	0.056	0.204	1.000	0.157	3.361	4.668	0.000	0.229	SC	1.4%
2	Stationary	0.155	0.416	1.000	0.495	0.676	4.467	0.216	0.147	C	1.9%
3	Stationary	0.000	0.083	0.000	0.000	5.000	5.000	0.000	0.000	S	20.5%
4	Roving/Stationary	0.493	1.012	0.064	1.545	3.387	3.400	0.653	0.215	FS	0.7%
5	Roving	0.466	0.984	0.222	3.430	1.236	1.454	3.151	0.057	FS	1.1%
6	Forested	0.001	0.002	0.002	4.998	0.000	0.000	4.998	0.000	F	74.4%

Figure: Separation Achieved to Cluster Sequences of Land Use Patterns

K-Means Clustering: An application from research

Table A2: K-Means Clusters and common class sequences

Cluster	Interpretation	Sequence	Share within cluster	Overall share	N
1	Stationary	SSSSC	18.4%	0.8%	11081
1	Stationary	CSSSS	15.5%	0.6%	9344
1	Stationary	SSSCC	12.8%	0.5%	7699
1	Stationary	SSSCS	12.0%	0.5%	7219
1	Stationary	CCSSS	11.4%	0.5%	6864
1	Stationary	SCSSS	6.1%	0.3%	3677
1	Stationary	SSCSS	3.7%	0.2%	2226
1	Stationary	SCSCS	3.7%	0.2%	2206
1	Stationary	CSSSC	2.7%	0.1%	1649
1	Stationary	SSCCS	2.6%	0.1%	1585
2	Stationary	SSSSS	100.0%	20.5%	297609
3	Stationary	CCCCC	39.5%	1.9%	27451
3	Stationary	CCCSS	8.0%	0.4%	5537
3	Stationary	SSCCC	7.5%	0.4%	5216
3	Stationary	SCCCC	6.9%	0.3%	4785
3	Stationary	CCCCS	6.3%	0.3%	4413
3	Stationary	CCCSC	4.5%	0.2%	3161
3	Stationary	SCCCS	2.9%	0.1%	1991
3	Stationary	CCCFF	2.4%	0.1%	1691
3	Stationary	CSCSC	2.3%	0.1%	1604
3	Stationary	CSCCC	2.3%	0.1%	1585
4	Roving/Stationary	FSSSS	20.8%	0.4%	5474
4	Roving/Stationary	FFSSS	16.8%	0.3%	4413
4	Roving/Stationary	SSSSF	15.8%	0.3%	4163
4	Roving/Stationary	SSSFF	11.3%	0.2%	2957
4	Roving/Stationary	SSSFS	5.2%	0.1%	1374
4	Roving/Stationary	FSSSF	4.0%	0.1%	1053
4	Roving/Stationary	FSFSF	3.2%	0.1%	830
4	Roving/Stationary	SFSSS	3.0%	0.1%	775
4	Roving/Stationary	SSFSS	2.9%	0.1%	760
4	Roving/Stationary	SSFSF	2.3%	0.0%	598
5	Roving	FFFSS	18.3%	0.6%	8153
5	Roving	FFFFS	16.5%	0.5%	7351
5	Roving	SFFFF	11.2%	0.3%	4997
5	Roving	SSFFF	7.5%	0.2%	3360
5	Roving	FFFSF	7.0%	0.2%	3111
5	Roving	CCFFF	4.8%	0.1%	2143
5	Roving	FFFCC	4.1%	0.1%	1814
5	Roving	FSFFF	4.1%	0.1%	1804
5	Roving	FFSFF	2.2%	0.1%	977
5	Roving	FFFSC	1.9%	0.1%	848
6	Forested	FFFFF	99.5%	74.4%	1081436
6	Forested	CFFFF	0.2%	0.2%	2605
6	Forested	FFFFC	0.2%	0.1%	2090
6	Forested	FFCFF	0.0%	0.0%	271

Figure: Separation Achieved to Cluster Sequences of Land Use Patterns

K-Means Clustering: An application from research

Table 4: Protection and Inferred Land Use pattern changes

Panel A: Full Panel

	Full Panel		
	(1) Forested	(2) Roving	(3) Stationary
Protected	0.008**	0.008***	-0.016***
	(0.004)	(0.002)	(0.004)
Mean of DV	.686	.0281	.286
N	1584634	1584634	1584634
Pixel FE	Yes	Yes	Yes
State x Year FE	Yes	Yes	Yes
Matched Pair-by-Year FE			

Panel B: Matched Panel Results

	Matched Panel		
	(1) Forested	(2) Roving	(3) Stationary
Protected	0.002	0.006**	-0.008*
	(0.006)	(0.003)	(0.004)
	.879	.0166	.105
	311370	311370	311370
	Yes	Yes	Yes
	Yes	Yes	Yes

Figure: Separation Achieved to Cluster Sequences of Land Use Patterns

K-Means Clustering Issues

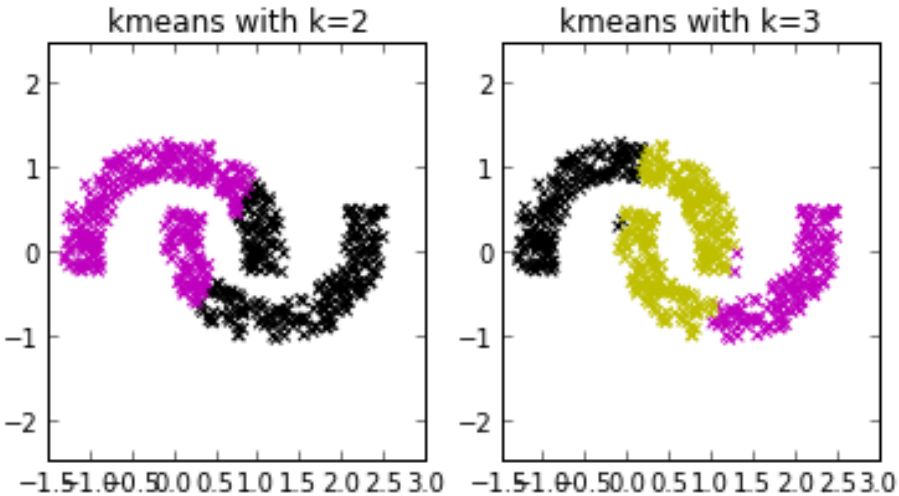
- K-Means works with quantitative data only, but some data may be categorical.

K-Means Clustering Issues

- K-Means works with quantitative data only, but some data may be categorical.
- Similarity is hard to define and typically, assessment requires human judgement of the sensibility of detected clusters. Euclidian distance may not be a good way to measure distance, and may not be defined for non -numeric features!

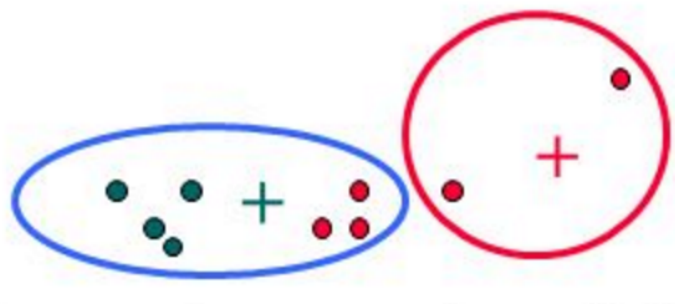
K-Means Clustering Issues

- K-Means works with quantitative data only, but some data may be categorical.
- Similarity is hard to define and typically, assessment requires human judgement of the sensibility of detected clusters. Euclidian distance may not be a good way to measure distance, and may not be defined for non -numeric features!
- Its not suitable to detect clusters in non convex shapes.



K-Means Clustering Issues

- K-Means works with quantitative data only, but some data may be categorical.
- Similarity is hard to define and typically, assessment requires human judgement of the sensibility of detected clusters. Euclidian distance may not be a good way to measure distance, and may not be defined for non -numeric features!
- Its not suitable to detect clusters in non convex shapes.
- K-Means as presented here is a *hard clustering* approach; it forces every observation into a cluster; this is problematic if there are outliers, as means change a lot.



Some hands-on R code on clustering

```
library(cluster)
library(quanteda)


## quanteda version 0.9.9.65
## Using 3 of 4 cores for parallel computing
##
## Attaching package:  'quanteda'
## The following object is masked from 'package:utils':
##
##     View

load(file="../../Data/trumpstweets.rdata")
# remove retweet entities
tw.user.df$text = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tw.user.df$text)
# remove at people
tw.user.df$text = gsub("@\\w+", "", tw.user.df$text)
# remove punctuation
tw.user.df$text = gsub("[[:punct:]]", "", tw.user.df$text)
# remove numbers
tw.user.df$text = gsub("[[:digit:]]", "", tw.user.df$text)
# remove html links
tw.user.df$text = gsub("http\\w+", "", tw.user.df$text)
# remove unnecessary spaces
tw.user.df$text = gsub("[ \t]{2,}", "", tw.user.df$text)
tw.user.df$text = gsub("^\\s+|\\s+$", "", tw.user.df$text)

trump.dfm<-dfm(tw.user.df$text, remove=stopwords(), stem=TRUE)
```

K-Means on Trump tweets

```
## k-means clustering
set.seed(18022017)
trump.dfm.trim <- dfm_trim(trump.dfm, min_count = 3, min_docfreq = 2)

# some guesses for "optimal k is sqrt of number of documents
k <- round(sqrt(ndoc(trump.dfm.trim)/2))

#a bit too many
k
```

```
## [1] 34
```

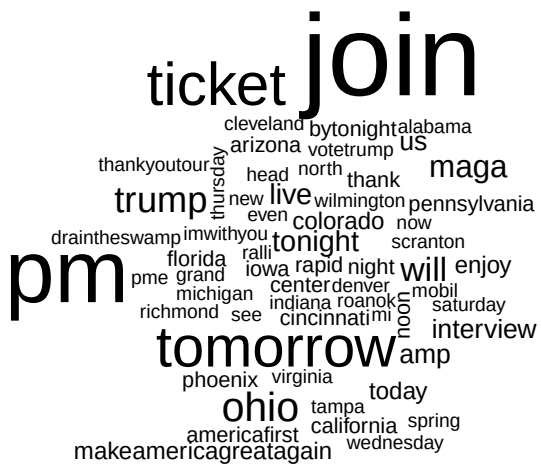
```
#computing words as shares of words in total document is like normalising length of documents, so can use Eucl

#tf function converts word count dfm to share
clusterk5 <- kmeans(tf(trump.dfm.trim, "prop"), 5)
splits<-split(docnames(trump.dfm.trim), clusterk5$cluster)
unlist(lapply(splits, length))
```

```
##    1    2    3    4    5
##   56  347  100 1690  135
```

K-Means on Trump tweets

```
## k-means clustering
textplot_wordcloud(dfm_trim(trump.dfm.trim[splits[[3]]], min_count = 1))
```



Plan

K-Means Clustering

K-medoids Clustering

Improving on K-Means Clustering

- Sensitivity of K-Means to outliers can distort the results dramatically: why? the inclusion of an outlier in any cluster, for small number of N can change the mean a lot.
- What can you do? Rather than computing a centroid, we can just pick a "representative observation".
- In the K-Means clustering step, means are computed in step 2a).
- Alternatively to computing the means, you can simply select a *most representative* point out of the points that are allocated to a specific cluster $\rightarrow$ *medoids*.
- Select a "representative point", rather than compute a mean.

K-Medoids Clustering Algorithm

Algorithm (*K-Medoids Clustering Algorithm*)

1. *Initialize by randomly selecting K points from N as the medoids $\mathbf{m} = \{\mathbf{m}_1, ..., \mathbf{m}_K\}$ and assign each observation to the closest medoid, where closest is defined by some dissimilarity metric $D(\mathbf{x}_i, \mathbf{x}_{i'})$.*
2. **Medoid selection step**: *For a given cluster assignment $C_1, ..., C_K$, find for each cluster k the observation $\mathbf{m}_k$, that minimizes the total distance to the other points belonging to this cluster, i.e. find for each k:*

$$i_k^* = argmin_{\{i \in C_k\}} \sum_{i' \in C_k} D(\mathbf{x}_i, \mathbf{x}_{i'})$$

3. **Cluster assignment step**: *Given a set of medoids $\{\mathbf{m_1}, .., \mathbf{m_K}\}$*
4. *Continue step 2,3 until the assigments do not change anymore.*

How do we measure distance?

- This is not straightforward. Typically, we want a distance function $D(\mathbf{x}, \mathbf{y})$ to satisfy three properties.
  1. Nonnegativity: $D(\mathbf{x}, \mathbf{y}) \geq 0$
  2. Symmetry: $D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x})$
  3. Triangle Inequality: $D(\mathbf{x}, \mathbf{z}) \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z})$
- There are many functions that satisfy this property. In the clustering section, instead of working in feature space X, we look at distance matrices.

Dissimilarity Matrix

Suppose you have some matrix $\mathbf{X}$ with dimension $n \times p$ and you compute all pairwise distances between any points, using *any distance metric you like*.

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Dissimilarity Matrix

Suppose you have some matrix $\mathbf{X}$ with dimension $n \times p$ and you compute all pairwise distances between any points, using *any distance metric you like*.

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Dissimilarity Matrix

Suppose you have some matrix $\mathbf{X}$ with dimension $n \times p$ and you compute all pairwise distances between any points, using *any distance metric you like*.

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Can we perform k-medoids clustering on distance matrices?

- The answer is: Yes!
- All we do is assign points as medoids and then, reassign points to their closest medoids.
- We do not need to know the information about the actual features X that underly the distance matrix, but rather, we *only* need to know their pairwise distances, i.e. the data dissimilarity matrix D which has dimensions $n \times n$.
- This greatly improves the set of potential applications for k-Means clusteirng, as you can define *any distance*.