Section 2

Masha Krupenkin[1]

April 23, 2019

[1]Stanford University

Today

Today

1. Determining the optimal K for clustering

Today

1. Determining the optimal K for clustering
2. MALLET: LDA for bulk documents

Optimal K

How can we find optimal K?

How can we find optimal K?

- **Naive:**

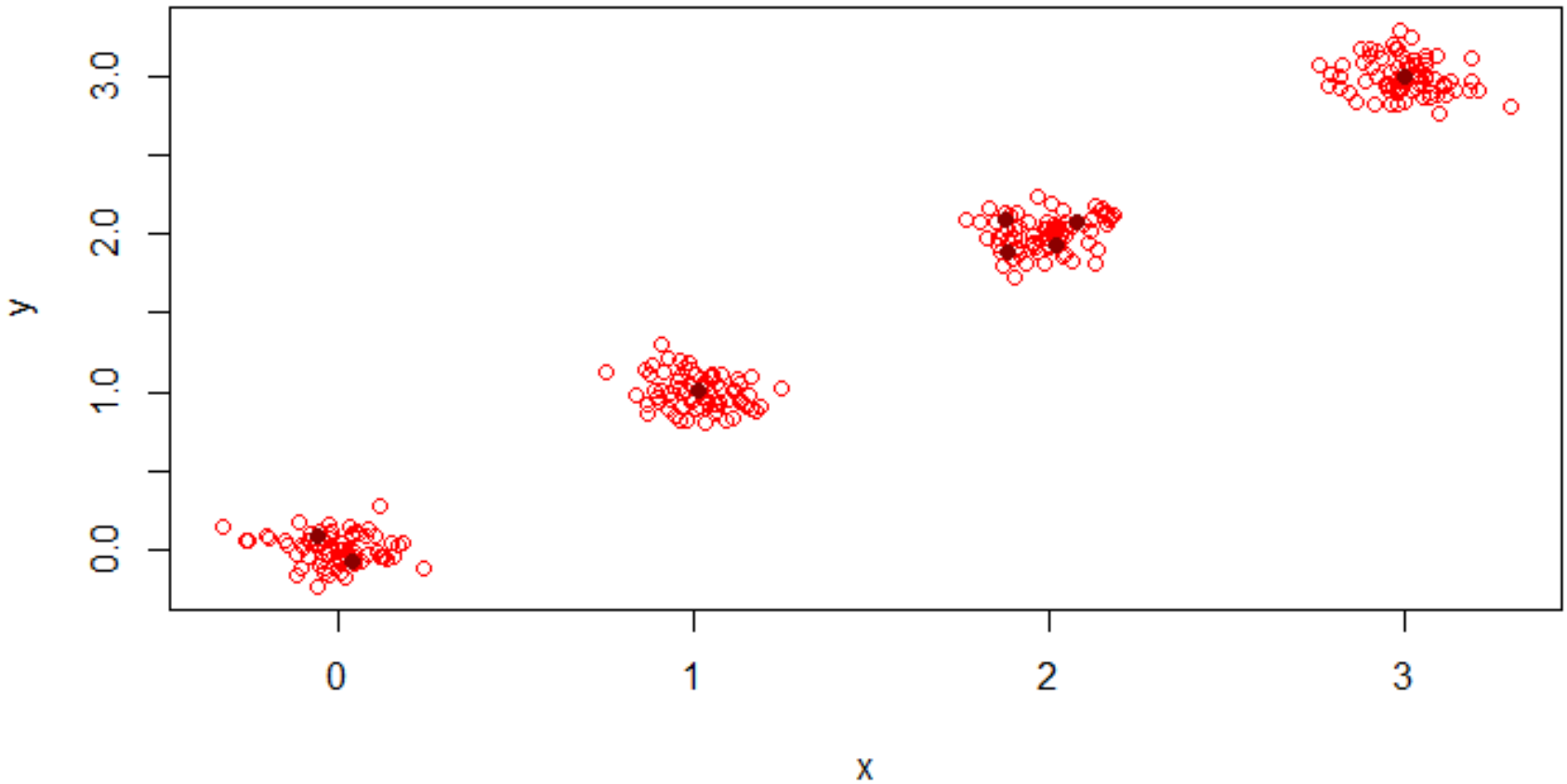
How can we find optimal K?

- **Naive:**Try to minimize within cluster sum of squares

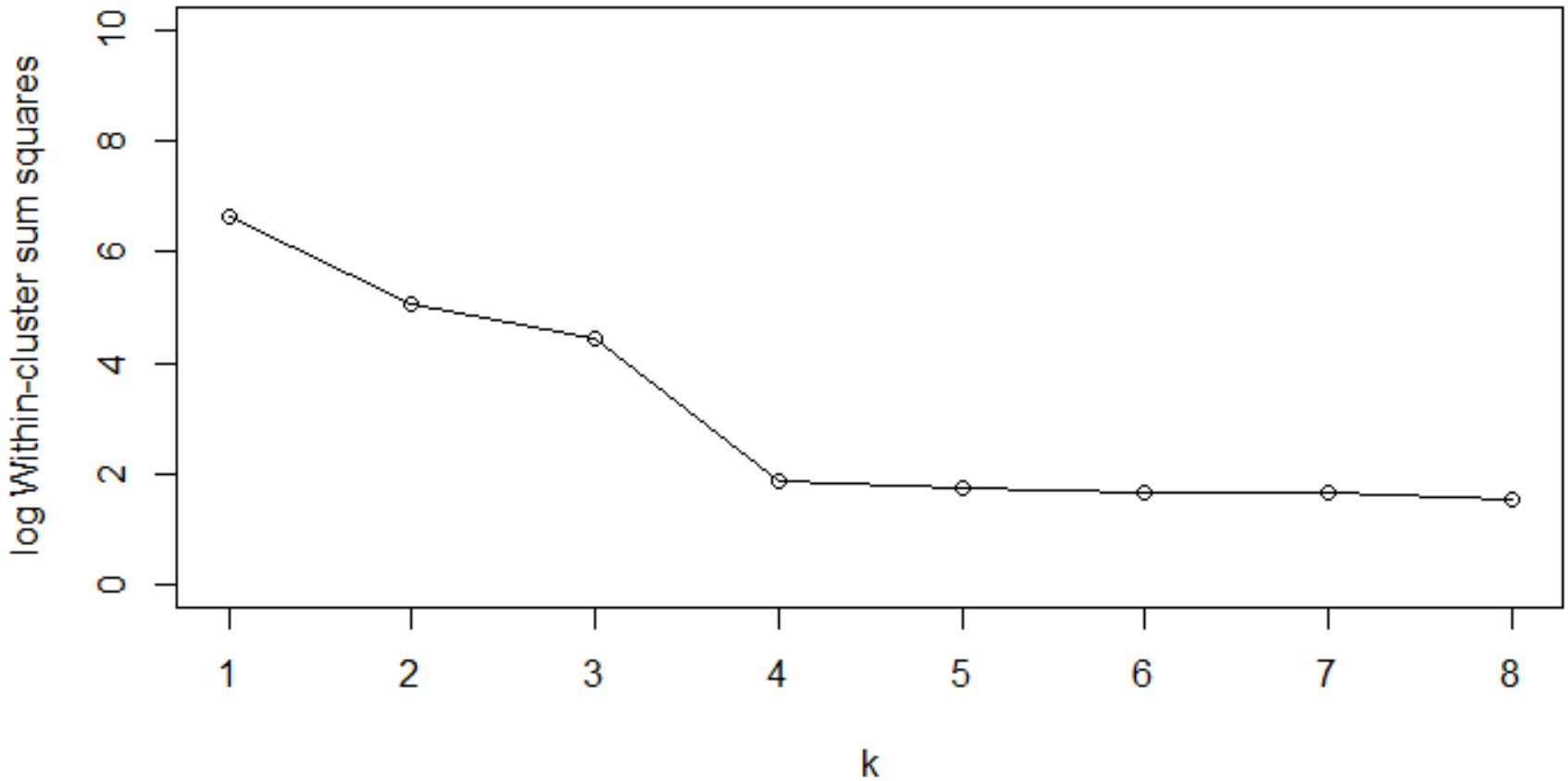
How can we find optimal K?

- **Naive:**Try to minimize within cluster sum of squares
- Doesn't work - trivial solution where each obs has its own cluster

"Eyeballing" Within-SS



"Eyeballing" Within-SS



Cluster Gap Statistic

Cluster Gap Statistic

- Statistical method to determine optimal k (Tibshirani, Walther, and Hastie 2001)

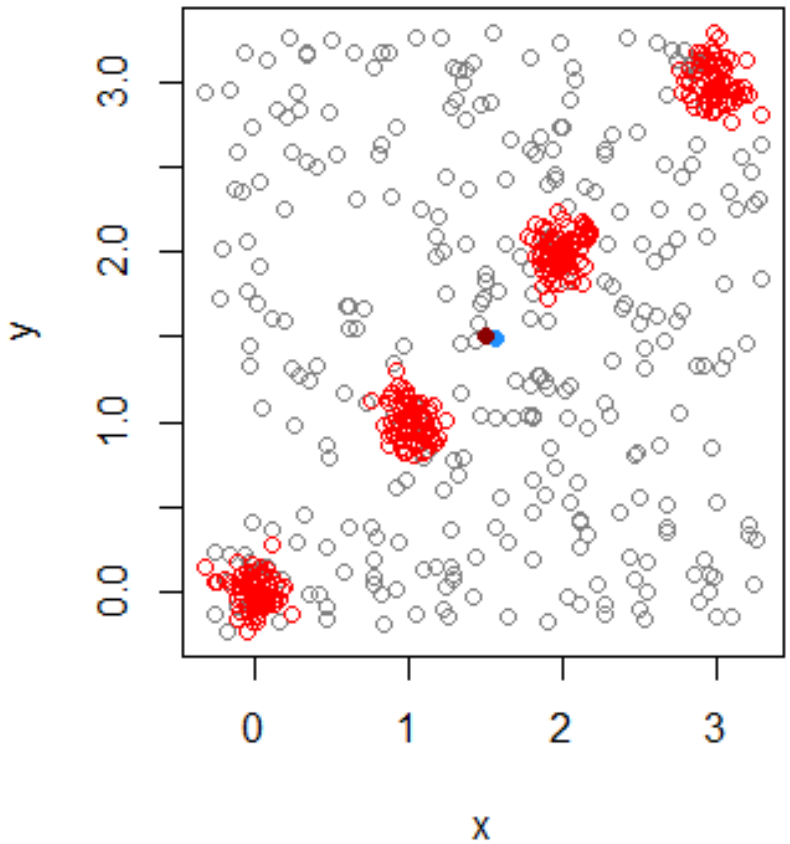
Cluster Gap Statistic

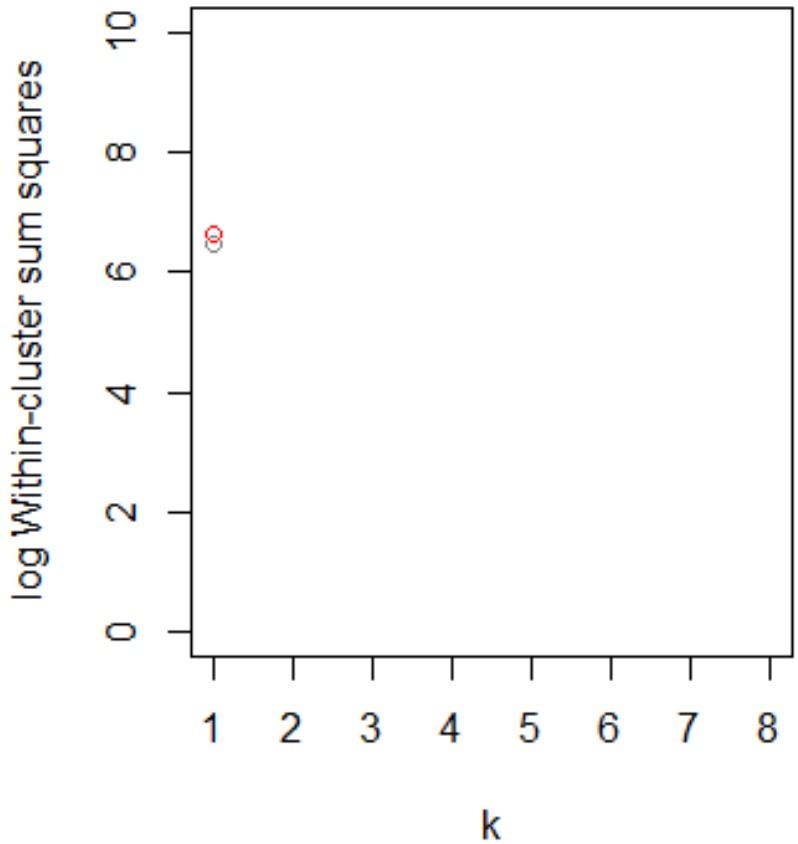
- Statistical method to determine optimal k (Tibshirani, Walther, and Hastie 2001)
- Intuition: How much does adding an unnecessary cluster reduce within-ss?

Cluster Gap Statistic

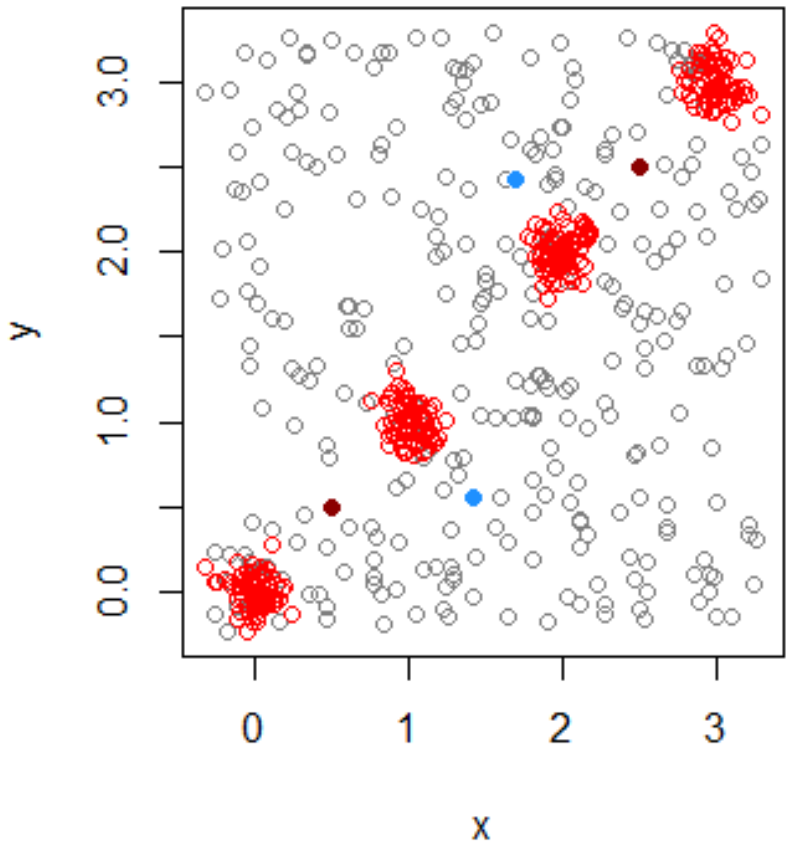
- Statistical method to determine optimal k (Tibshirani, Walther, and Hastie 2001)
- Intuition: How much does adding an unnecessary cluster reduce within-ss?
- Compare reduction in log(within-ss) by adding an additional cluster to random data vs reduction in your data

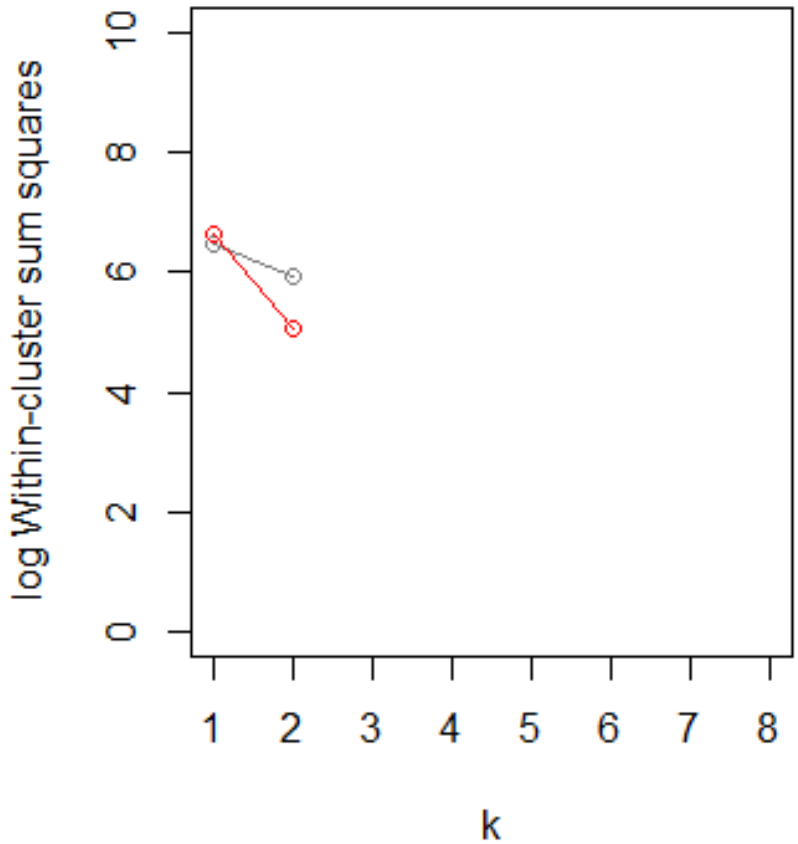
Intuition


y
0.0
1.0
2.0
3.0
0
1
2
3
x

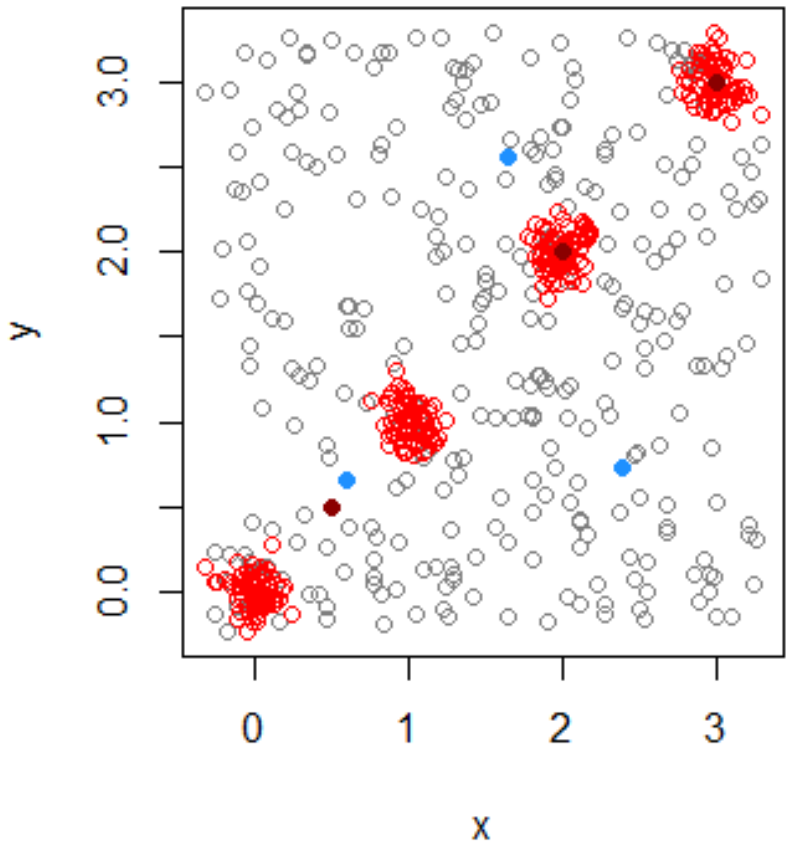

log Within-cluster sum squares
0
2
4
6
8
10
1
2
3
4
5
6
7
8
k

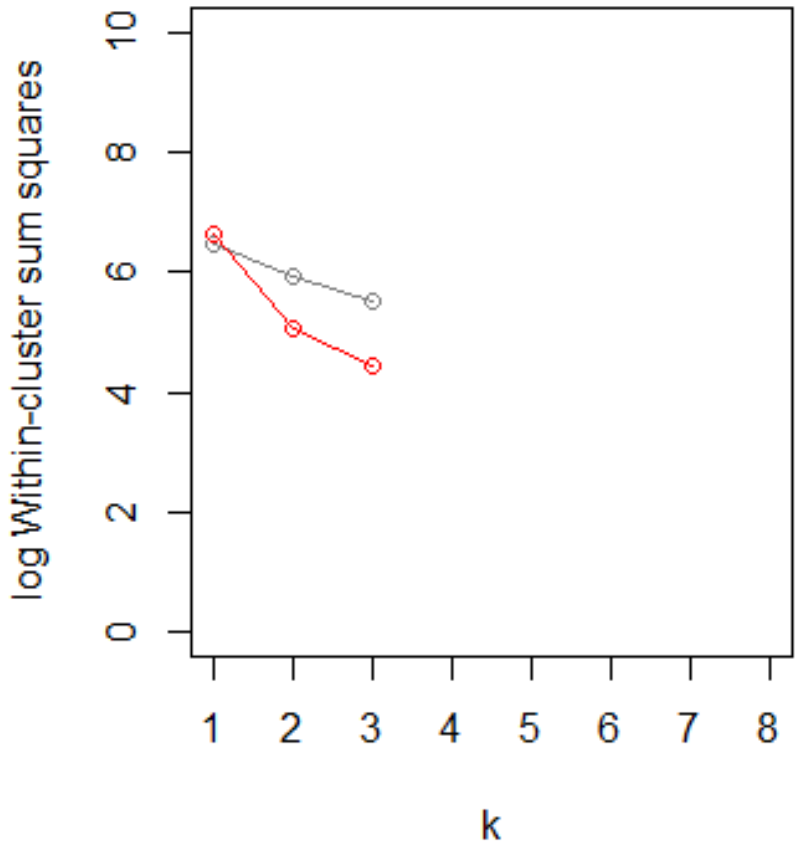
Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3

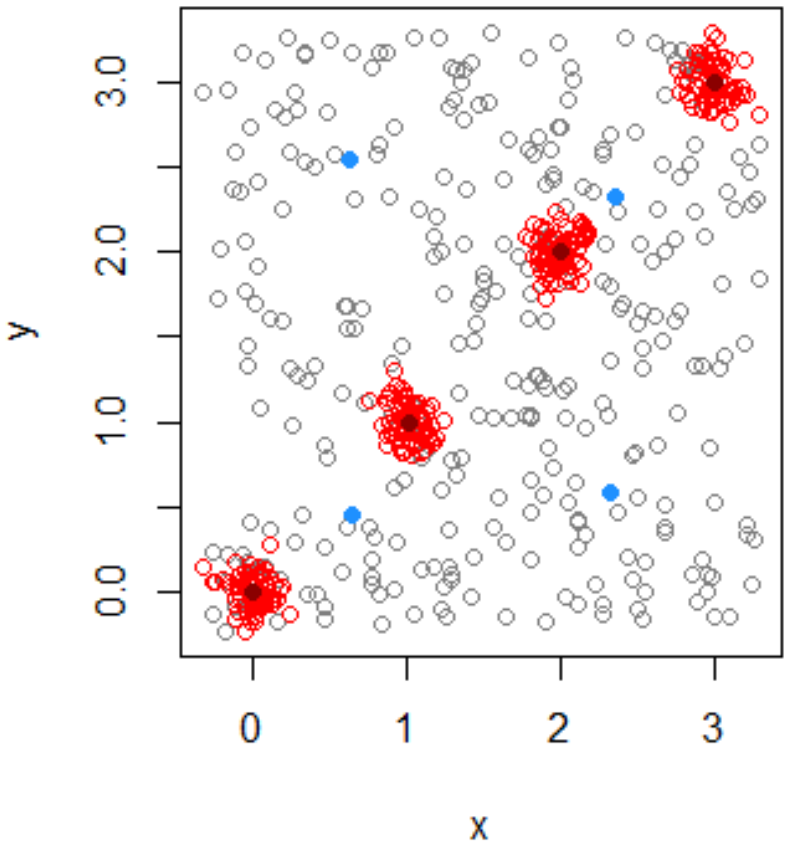

log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

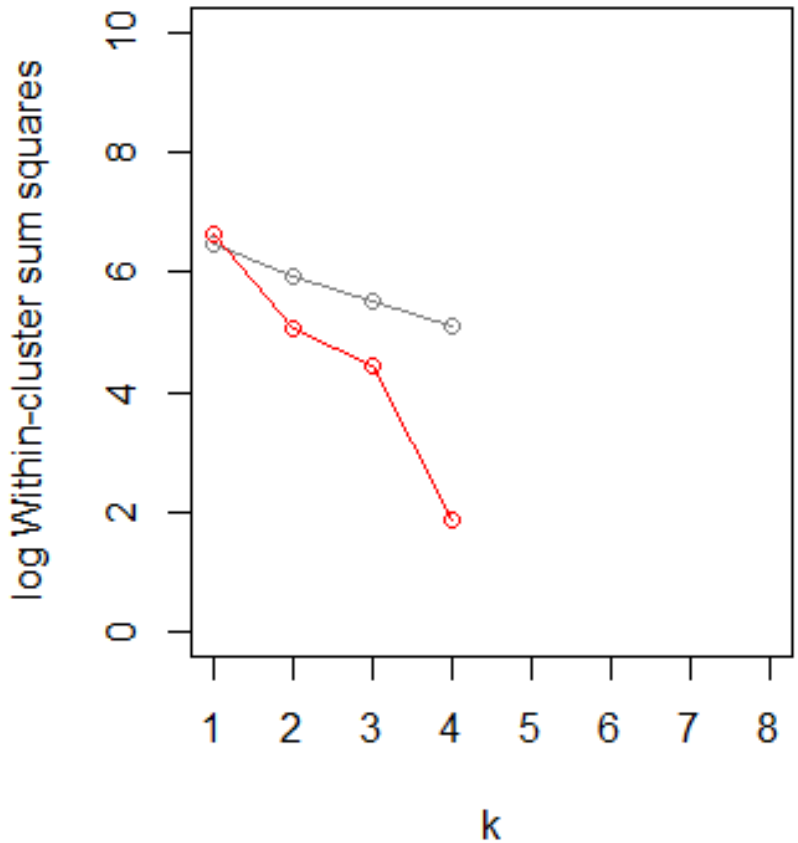
Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3

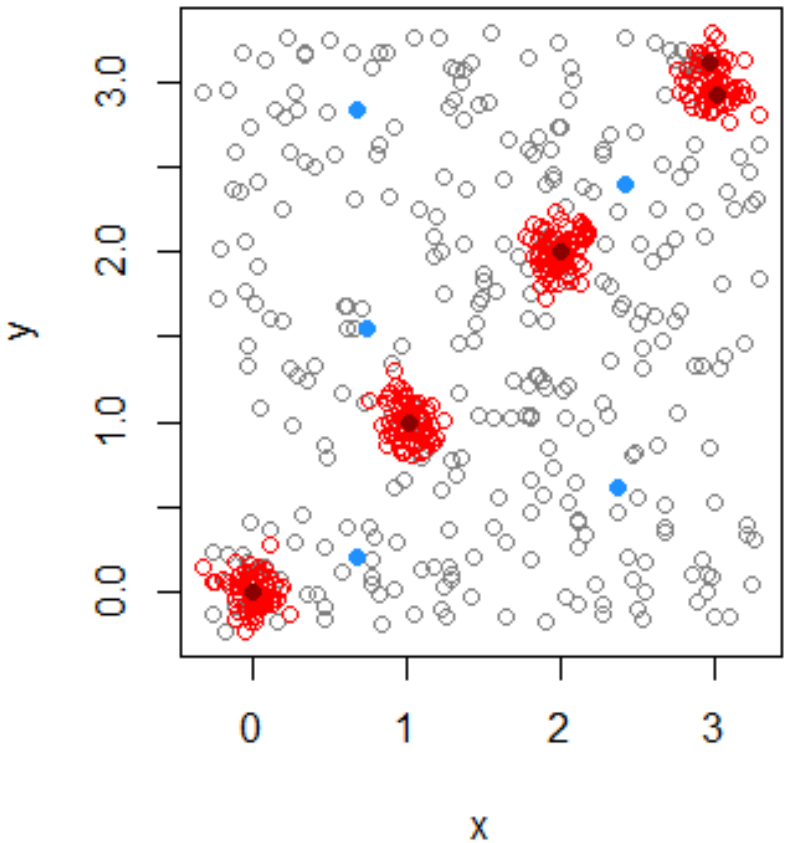

log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

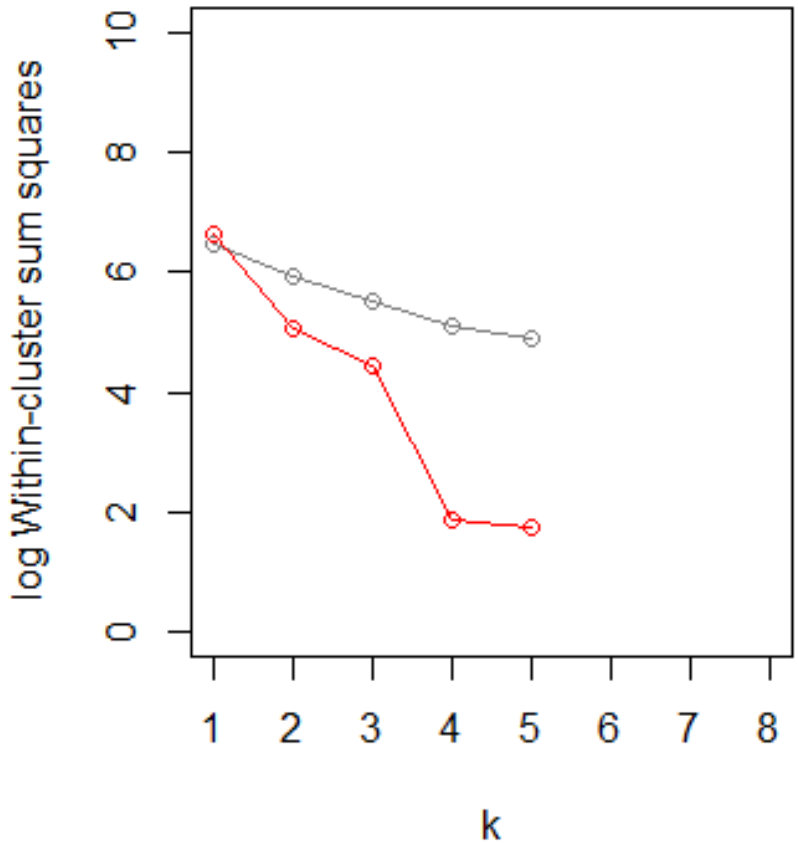
Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3

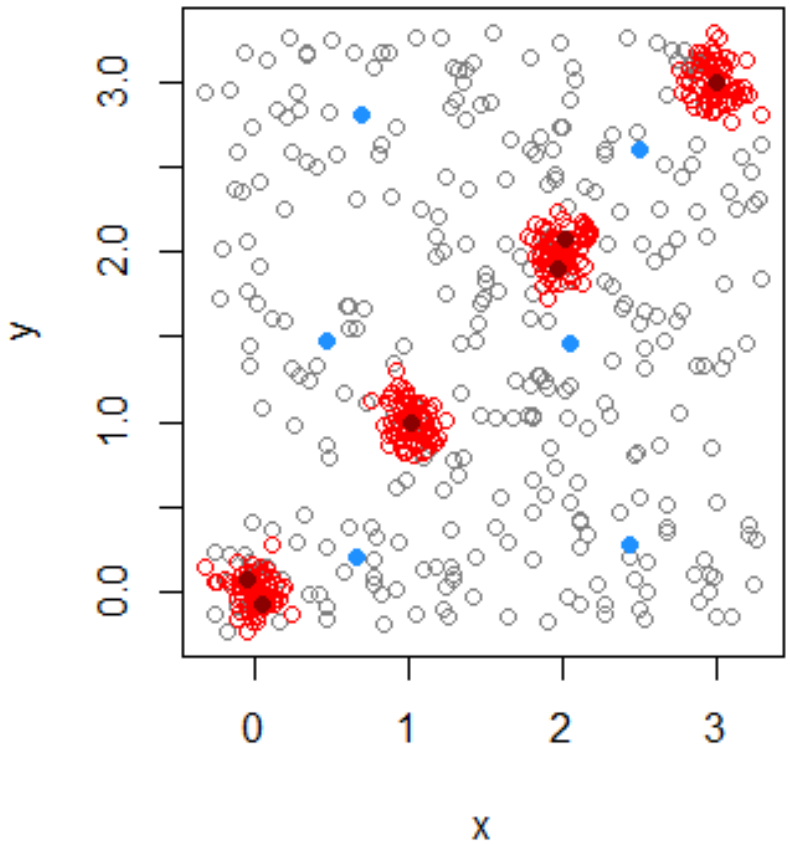

log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

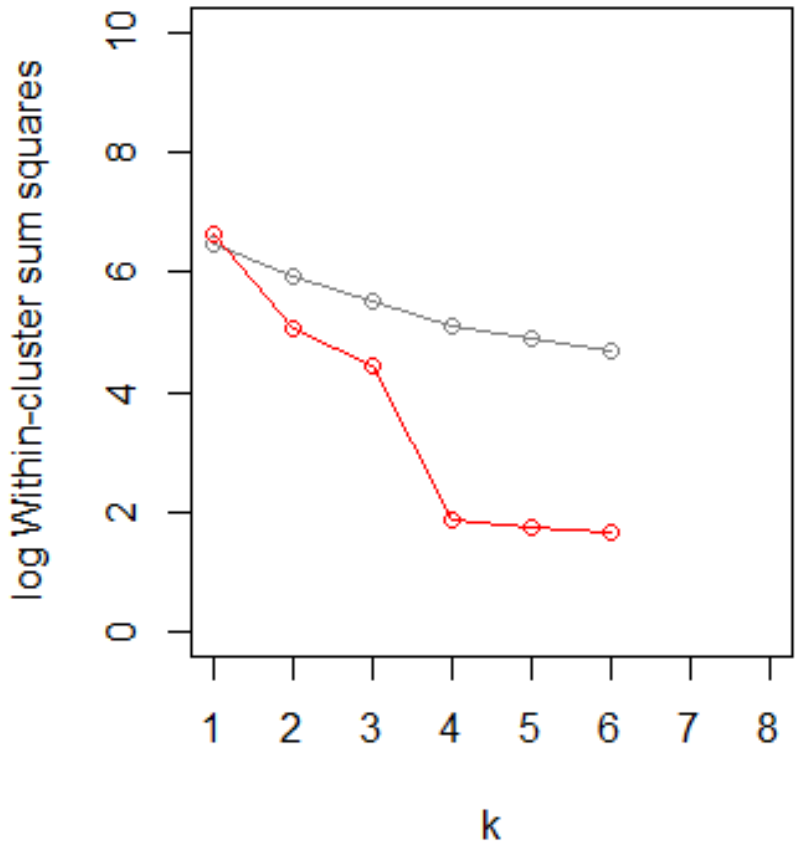
Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3

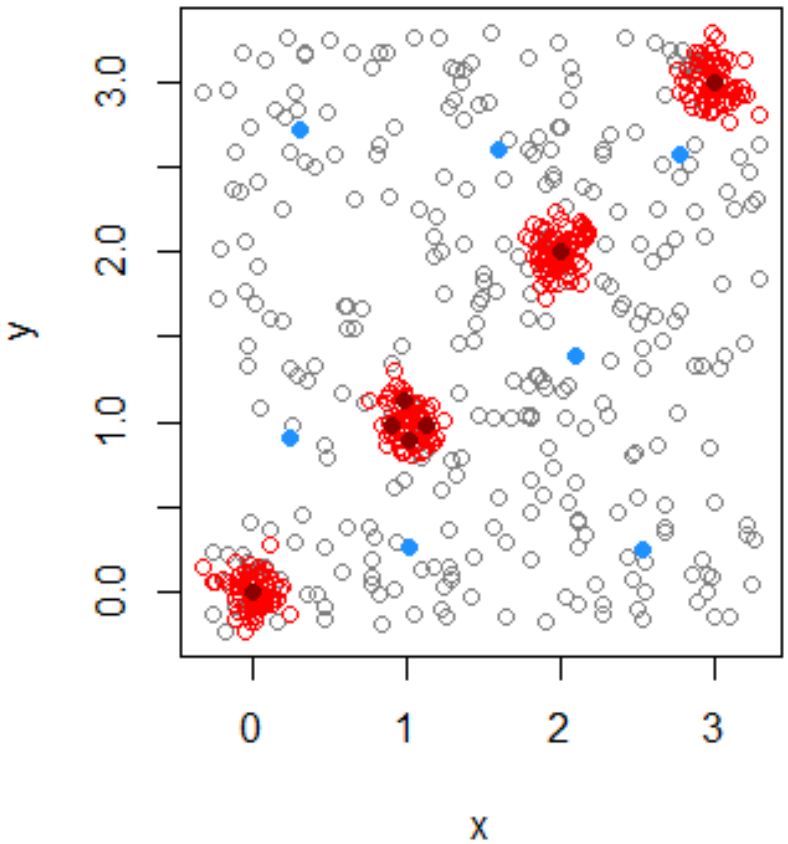

log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

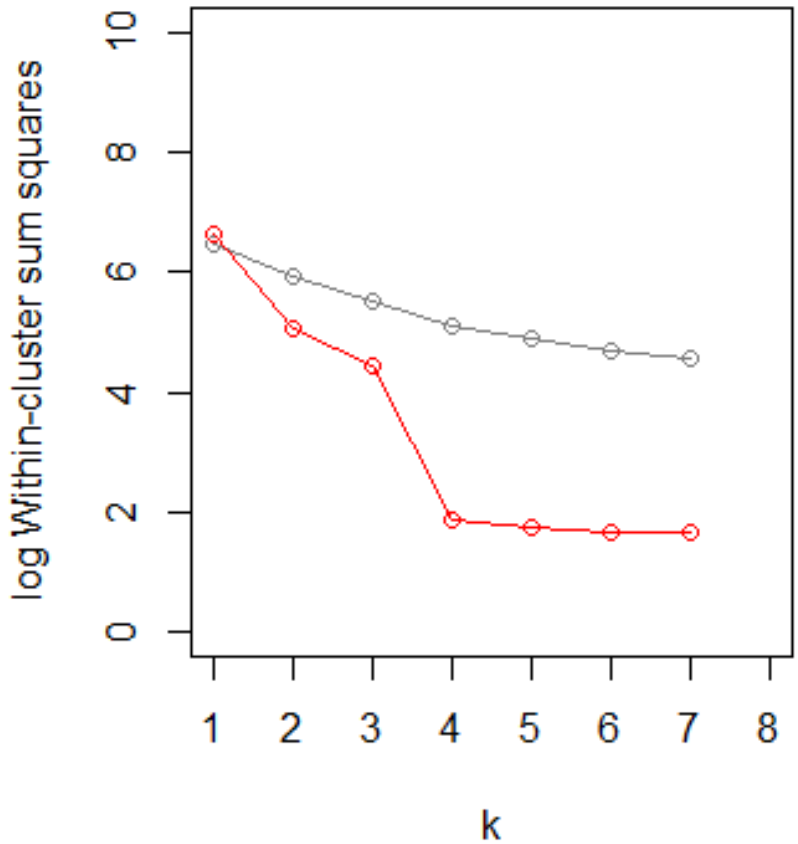
Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3

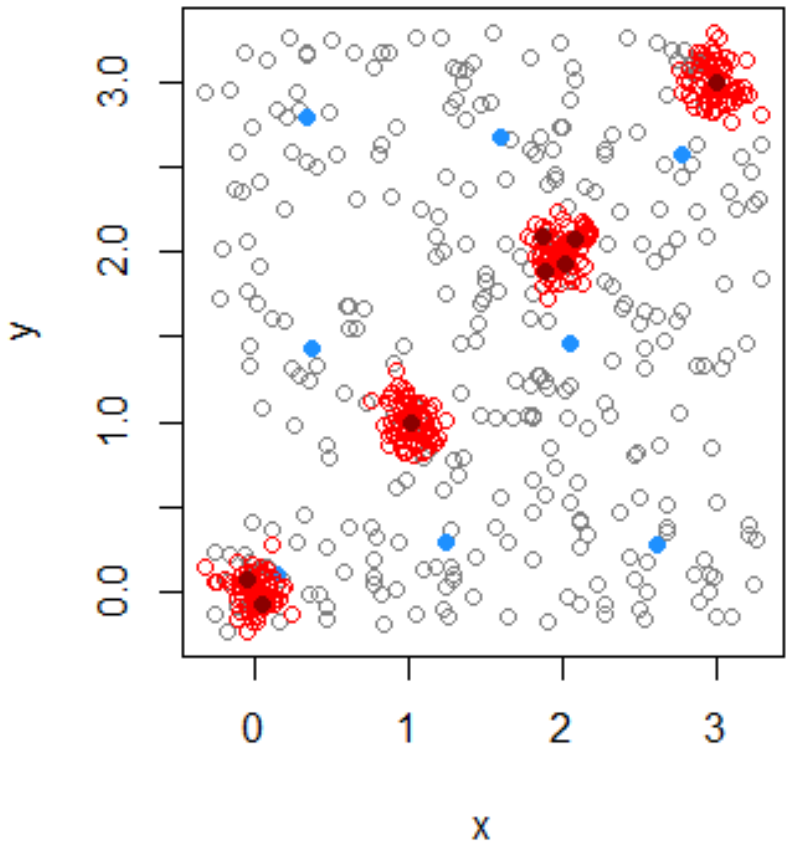

log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

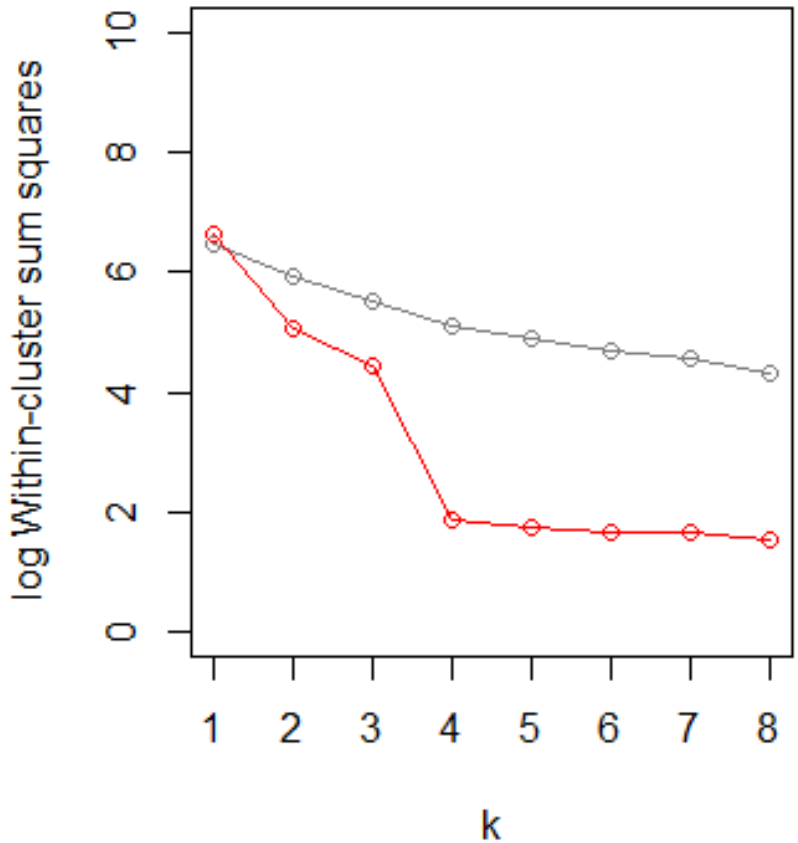
Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3


log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

Intuition


y
x
0.0
1.0
2.0
3.0
0
1
2
3


log Within-cluster sum squares
k
0
2
4
6
8
10
1
2
3
4
5
6
7
8

MALLET

Why MALLET?

Why MALLET?

- MAchine Learning for LanguagE Toolkit

Why MALLET?

- MAchine Learning for LanguagE Toolkit
- One-stop shop for machine learning needs

Why MALLET?

- MAchine Learning for LanguagE Toolkit
- One-stop shop for machine learning needs
- Chews through hundred of thousands of files easily

Download

`http://mallet.cs.umass.edu/`

Tutorial

https://electricarchaeology.ca/2011/08/30/getting-started-with-mallet-and-topic-modeling/

Demo

Demo

- What topics did people search for about Hillary Clinton

Demo

- What topics did people search for about Hillary Clinton
- (Data from Peterson, Goel and Iyengar 2018)

Devil's in the details...

Devil's in the details...

- What happens if I keep stopwords?

Devil's in the details...

- What happens if I keep stopwords?
- What happens if I use bigrams?

Devil's in the details...

- What happens if I keep stopwords?
- What happens if I use bigrams?
- What happens if I vary number of topics?