Machine Learning

Justin Grimmer

Professor
Department of Political Science
Stanford University

April 16th, 2019

Discovery

Search for new ways to organize text

- Complement, Not Replace, Organizations of Text
- There is No Ground Truth Conceptualization
- Once you have a conceptualization it is yours

Clustering: partition of documents

- Discover categories
- Assign documents to categories

Fully Automated Clustering

1) Notion of distance
2) Definition of "good" clustering
3) Optimization method

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

K-Means $\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal $\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k \quad = \quad (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k$ = exemplar for cluster k

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k \quad = \quad (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k =$ exemplar for cluster k
2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k =$ exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.
   If observation i is from cluster k, then

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k =$ exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.
   If observation i is from cluster k, then

$$\boldsymbol{\tau}_i = (0, 0, \ldots, 0, \underbrace{1}_{k^{th}}, 0, \ldots, 0)$$

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k$ = exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.
   If observation i is from cluster k, then

$$\boldsymbol{\tau}_i = (0, 0, \ldots, 0, \underbrace{1}_{k^{th}}, 0, \ldots, 0)$$

   Hard Assignment

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) \quad = \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \quad \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center

K-Means $\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions

K-Means $\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) \quad = \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) \quad = \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$

K-Means $\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$
  - If $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sum_{j=1}^{J} \sigma_j^2 w$

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$
  - If $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sum_{j=1}^{J} \sigma_j^2 w$
    - Each observation in same cluster

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$
  - If $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sum_{j=1}^{J} \sigma_j^2 w$
    - Each observation in same cluster
    - $\boldsymbol{\theta}_1 =$ Average across documents

K-Means$\rightsquigarrow$ Optimization

Coordinate descent

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^{t}$

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$
- Conditional on $\boldsymbol{T}^t$, choose $\boldsymbol{\Theta}^t$

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$
- Conditional on $\boldsymbol{T}^t$, choose $\boldsymbol{\Theta}^t$

Repeat until convergence$\rightsquigarrow$ as measured as change in f dropping below threshold ϵ

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$
- Conditional on $\boldsymbol{T}^t$, choose $\boldsymbol{\Theta}^t$

Repeat until convergence$\rightsquigarrow$ as measured as change in f dropping below threshold ϵ

$$\text{Change} = f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta}^t) - f(\boldsymbol{X}, \boldsymbol{T}^{t-1}, \boldsymbol{\Theta}^{t-1})$$

K-Means$\rightsquigarrow$ Optimization

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

2) Choose $\boldsymbol{T}^t$

K-Means$\leadsto$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

2) Choose $\boldsymbol{T}^t$

$$\tau_{im}^t = \begin{cases} 1 \text{ if } m = \arg\min_k \sum_{j=1}^J (x_{ij} - \theta_{kj}^t)^2 \\ 0 \text{ otherwise }, \end{cases} .$$

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

2) Choose $\boldsymbol{T}^t$

$$\tau_{im}^t = \begin{cases} 1 \text{ if } m = \arg\min_k \sum_{j=1}^{J} (x_{ij} - \theta_{kj}^t)^2 \\ 0 \text{ otherwise }, \end{cases} .$$

In words: Assign each document $\boldsymbol{x}_i$ to the closest center $\boldsymbol{\theta}_m^t$

K-Means$\rightsquigarrow$ Optimization

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

K-Means$\rightsquigarrow$ Optimization

3) Choose Θ^t $\rightsquigarrow$ Focus on the center for cluster k

$$f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k \quad = \quad \sum_{i=1}^{N} \tau_{ik}^{t} \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right)$$

K-Means⇝ Optimization

3) Choose Θ^t ⇝ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t (x_{ij} - \theta_{jk})
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk} \right) \\
0 &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk}^* \right)
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk} \right) \\
0 &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk}^* \right) \\
&= \sum_{i=1}^{N} \tau_{ij}^t x_{ij} - \theta_{jk}^* \sum_{i=1}^{N} \tau_{ij}^t
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

3) Choose Θ^t $\rightsquigarrow$ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk} \right) \\
0 &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk}^* \right) \\
&= \sum_{i=1}^{N} \tau_{ij}^t x_{ij} - \theta_{jk}^* \sum_{i=1}^{N} \tau_{ij}^t \\
\frac{\sum_{i=1}^{N} \tau_{ik}^t x_{ij}}{\sum_{i=1}^{N} \tau_{ik}^t} &= \theta_{jk}^*
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}}$$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers

K-Means $\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers
- Do until converged:

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers
- Do until converged:
    - For each document, find closest center$\rightsquigarrow$ $\boldsymbol{\tau}_i^t$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers
- Do until converged:
  - For each document, find closest center$\rightsquigarrow$ $\boldsymbol{\tau}_i^t$
  - For each center, take average of assigned documents$\rightsquigarrow$ $\boldsymbol{\theta}_k^t$

K-Means$\rightsquigarrow$ Optimization

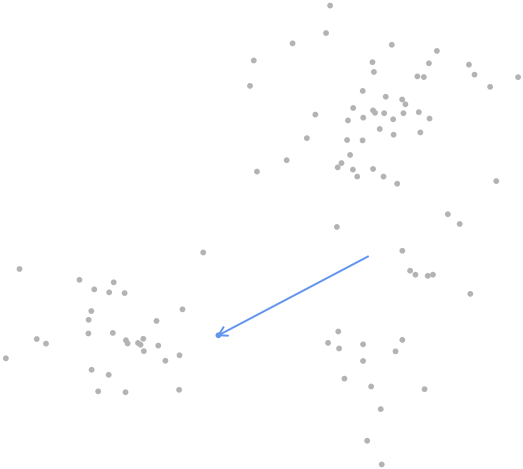
$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

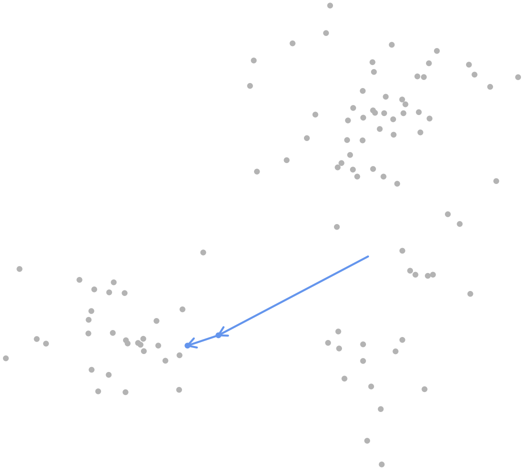
- Initialize centers
- Do until converged:
  - For each document, find closest center$\rightsquigarrow$ $\boldsymbol{\tau}_i^t$
  - For each center, take average of assigned documents$\rightsquigarrow$ $\boldsymbol{\theta}_k^t$
  - Update change $f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta}^t) - f(\boldsymbol{X}, \boldsymbol{T}^{t-1}, \boldsymbol{\Theta}^{t-1})$

Visual Example

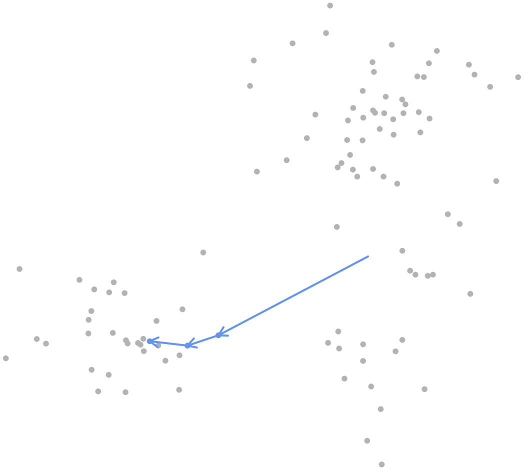
Visual Example



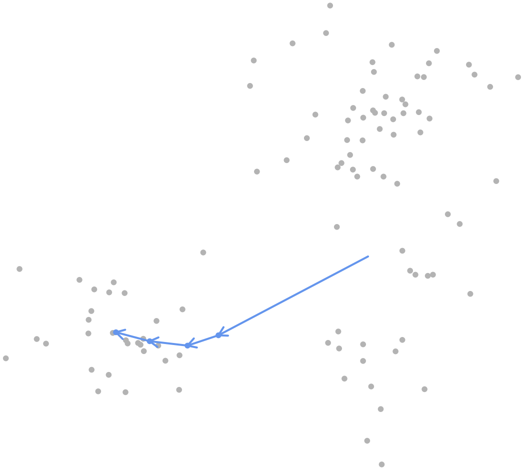
Visual Example



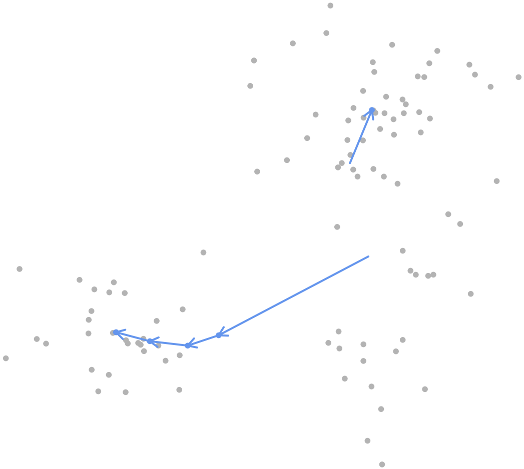
Visual Example



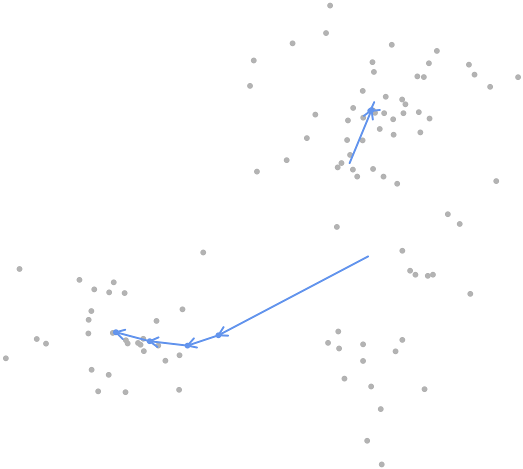
Visual Example



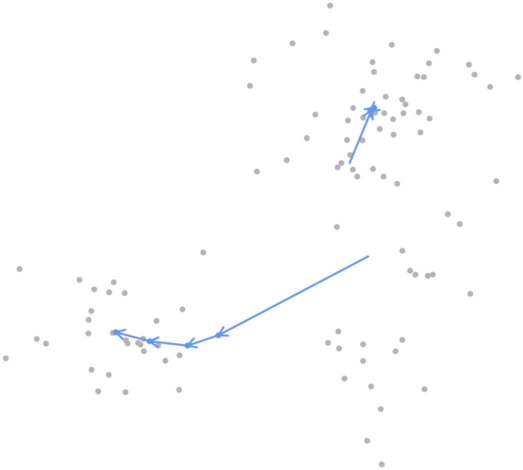
Visual Example



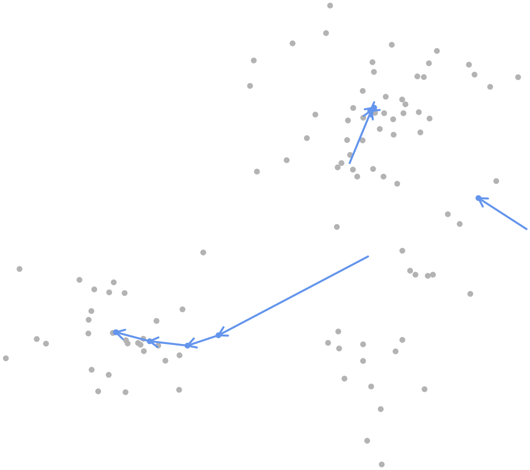
Visual Example



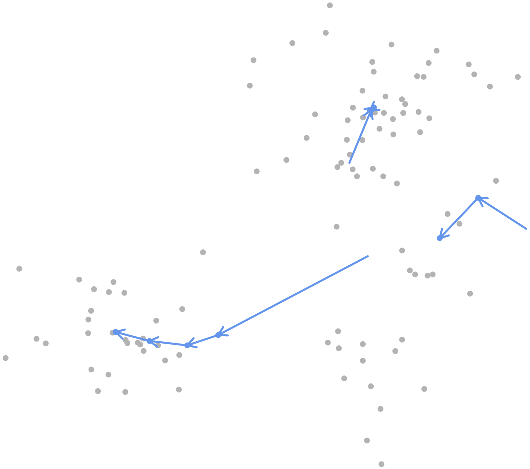
Visual Example



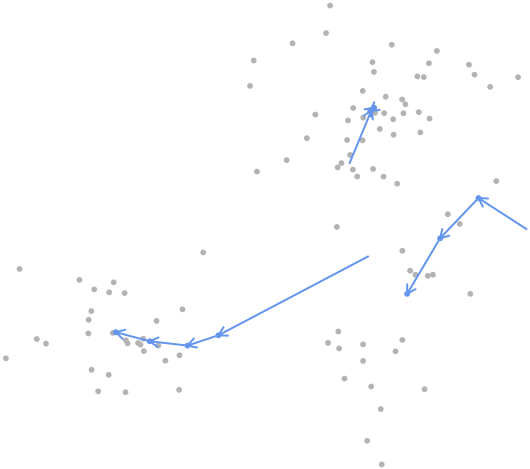
Visual Example



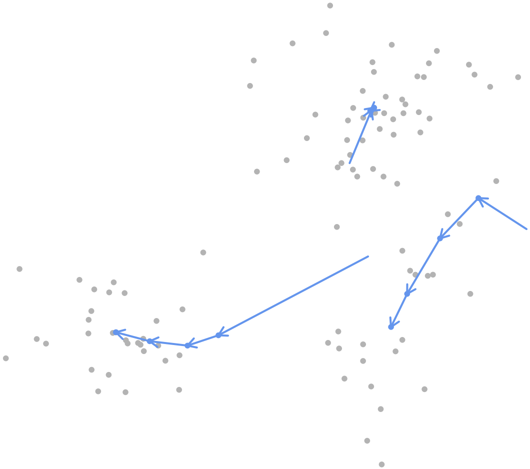
Visual Example



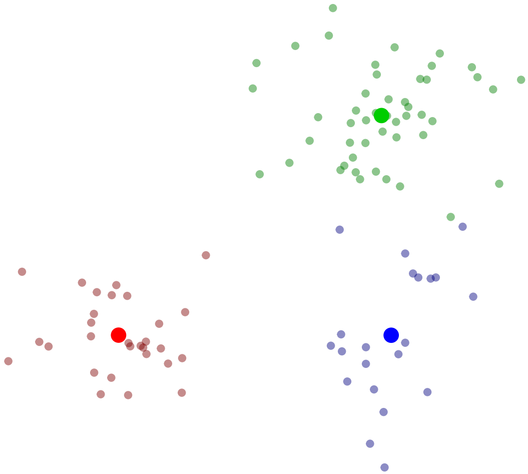
Visual Example



Visual Example



Visual Example



An Example: Jeff Flake

`To the R Code!`

Interpreting Cluster Components

Unsupervised methods

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)

Interpreting Cluster Components

Unsupervised methods$\leadsto$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents
        - Assign cluster label
    - Automatic identification
        - Know label classes
        - Use methods to identify separating words

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents
        - Assign cluster label
    - Automatic identification
        - Know label classes
        - Use methods to identify separating words
        - Use these to help infer differences across clusters
- Transparency

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean
  - Provide documents + organizations

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents
        - Assign cluster label
    - Automatic identification
        - Know label classes
        - Use methods to identify separating words
        - Use these to help infer differences across clusters
- Transparency
    - Debate what clusters are
    - Debate what they mean
    - Provide documents + organizations

`back to the R code!`

How Do We Choose K?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
    - Sum squared errors decreases as K increases
    - Trivial answer: each document in own cluster (useless)

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + manual search

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + manual search$\rightsquigarrow$discuss statistical methods/experimental methods on Thursday
- Humans should be the final judge

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + manual search$\rightsquigarrow$discuss statistical methods/experimental methods on Thursday
- Humans should be the final judge
  - Compare insights across clusterings

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\leadsto$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \quad \sim \quad \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\boldsymbol{\tau}_i|\boldsymbol{\pi} \quad \sim \quad \text{Multinomial}(1, \boldsymbol{\pi})$$

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{\tau}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i|\tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned}\boldsymbol{\tau}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i|\tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k)\end{aligned}$$

In words:

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{\tau}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | \tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

In words:

- Draw a cluster label

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \quad \sim \quad \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned}
\boldsymbol{\tau}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\
\boldsymbol{x}_i | \tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k)
\end{aligned}$$

In words:

- Draw a cluster label
- Given distribution, draw realization

Mixture of Unigram Models (Mixture of Multinomials)

A mixture of unigram-language models

$$
\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\tau}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\
\boldsymbol{x}_i|\tau_{ik} = 1, \boldsymbol{\theta}_k &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_k)
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X})$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) \quad \propto \quad \overbrace{p(\boldsymbol{\pi})p(\theta)}^{1} \quad \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \theta)}_{\text{Complete data likelihood}}$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) \propto \overbrace{p(\boldsymbol{\pi})p(\boldsymbol{\theta})}^{1} \quad \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \boldsymbol{\theta})}_{\text{Complete data likelihood}}$$

$$\propto \underbrace{\prod_{i=1}^{N} p(\boldsymbol{\tau}_i, \boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}_{\text{Complete data likelihood}}$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) & \propto \overbrace{p(\boldsymbol{\pi}) p(\boldsymbol{\theta})}^{1} \quad \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \boldsymbol{\theta})}_{\text{Complete data likelihood}} \\
& \propto \underbrace{\prod_{i=1}^{N} p(\boldsymbol{\tau}_i, \boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}_{\text{Complete data likelihood}} \\
& \propto \prod_{i=1}^{N} p(\boldsymbol{\tau}_i | \boldsymbol{\pi}) p(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\tau}_i)
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) &\propto \overbrace{p(\boldsymbol{\pi}) p(\boldsymbol{\theta})}^{1} \quad \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \boldsymbol{\theta})}_{\text{Complete data likelihood}} \\
&\propto \underbrace{\prod_{i=1}^{N} p(\boldsymbol{\tau}_i, \boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}_{\text{Complete data likelihood}} \\
&\propto \prod_{i=1}^{N} p(\boldsymbol{\tau}_i | \boldsymbol{\pi}) p(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\tau}_i) \\
&\propto \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right]^{\tau_{ik}}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates$\rightsquigarrow$ EM Algorithm

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i | \boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data} | \boldsymbol{\theta}_k, \boldsymbol{\pi}] \quad = \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t | \boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t | \boldsymbol{\Theta}, \pi_k)$$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] \quad = \quad \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates$\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

   Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates$\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k)p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

$$\begin{aligned}\text{Change} &= \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}^{t+1}] \\ &\quad - \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t}, \boldsymbol{\pi}^{t}]\end{aligned}$$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

$$\begin{aligned}\text{Change} &= \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}^{t+1}] \\ &\quad - \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t}, \boldsymbol{\pi}^{t}]\end{aligned}$$

Our update steps will be strikingly similar to the K-Means algorithm

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

$$\begin{aligned}\text{Change} &= \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}^{t+1}] \\ &\quad - \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t}, \boldsymbol{\pi}^{t}]\end{aligned}$$

Our update steps will be strikingly similar to the K-Means algorithm

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X})$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) = \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}}$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^{K}\left(\pi_m^t \prod_{j=1}^{J}(\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^{K}\left(\pi_m^t \prod_{j=1}^{J}(\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Define:

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^K \left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^J (\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^K \left(\pi_m^t \prod_{j=1}^J (\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Define:

$$
r_{ik}^t \equiv \frac{\pi_k^t \prod_{j=1}^J (\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^K \left(\pi_m^t \prod_{j=1}^J (\theta_{jm}^t)^{x_{ij}}\right)}
$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^K \left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^J (\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^K \left(\pi_m^t \prod_{j=1}^J (\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Define: Avoid underflow

$$
r_{ik}^t = \left[1 + \sum_{k' \neq k} \frac{\pi_{k'} \prod_{j=1}^J (\theta_{jk'}^t)^{x_{ij}}}{\pi_k \prod_{j=1}^J (\theta_{jk}^t)^{x_{ij}}}\right]^{-1}
$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log\left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}}\right)$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E[log Complete data} | \boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N} \sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log\left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

$$\pi_k^{t+1} = \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N}$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

$$
\begin{aligned}
\pi_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N} \\
\theta_{jk}^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t} x_{ij}}{\sum_{m=1}^{J}\sum_{i=1}^{N} r_{ik}^{t} x_{im}}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N} \sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

$$
\begin{aligned}
\pi_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N} \\
\theta_{jk}^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t} x_{ij}}{\sum_{m=1}^{J} \sum_{i=1}^{N} r_{ik}^{t} x_{im}} \propto \sum_{i=1}^{N} r_{ik}^{t} \boldsymbol{x}_i
\end{aligned}
$$

Example: Jeff Flake Again!

`To the R Code!`

Fully Automated Clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
    - Spectral clustering
    - Affinity Propagation

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics

Fully Automated Clustering

- Notion of similarity and "good" partition$\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality$\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T!

Fully Automated Clustering

- Notion of similarity and "good" partition$\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality$\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T!$\rightsquigarrow$ **And never will**

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T! $\rightsquigarrow$ And never will $\rightsquigarrow$ but still useful(!!!!)

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T! $\rightsquigarrow$ **And never will** $\rightsquigarrow$ **but still useful(!!!!)**

Appendix: Why EM Works

Goal:

$$\text{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{X}|\theta) \quad = \quad \sum_{\boldsymbol{T}} p(\boldsymbol{X}, \boldsymbol{T}|\theta)$$

Define:

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{T}|\theta)}{q(\boldsymbol{T})} \right] \\
K(q||p) &= -\sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log \left[\frac{p(\boldsymbol{T}|\boldsymbol{X}, \theta)}{q(\boldsymbol{T})} \right]
\end{aligned}$$

Then:

$$\log p(\boldsymbol{X}|\theta) \quad = \quad \mathcal{L}(q, \boldsymbol{\theta}) + K(q||p)$$

Appendix: Why EM Works

$$
\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) + K(q||p) &= \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{T}|\boldsymbol{\theta})}{q(\boldsymbol{T})}\right] - \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log \left[\frac{p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{T})}\right] \\
&= \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log(p(\boldsymbol{X}|\boldsymbol{\theta})) + \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log(p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{\theta})) \\
&\quad - \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log q(\boldsymbol{T}) - \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{\theta}) + \sum_{\boldsymbol{T}} q(\boldsymbol{T}) \log q(\boldsymbol{T}) \\
&= \log p(\boldsymbol{X}|\boldsymbol{\theta})
\end{aligned}
$$

Collect terms that cancel and recognize $\sum_{\boldsymbol{T}} q(\boldsymbol{T}) = 1$ and we see equivalence

Appendix: Why EM Works

$K(q||p) \geq 0$ with $K(q||p) = 0$ only if $q = p$. So, $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower-bound on the log-likelihood.
E-step

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) - K(q||p) \quad = \quad \mathcal{L}(q, \boldsymbol{\theta})$$

$\mathcal{L}(q, \boldsymbol{\theta}) \rightsquigarrow$ biggest when $K(q||p) = 0$, so set

$$q(\boldsymbol{T}) \quad = \quad p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{\theta})$$

M-step:
Given the new value of q, maximize parameters (expectation of the log complete data likelihood)
Change in log-likelihood will be greater$\rightsquigarrow$ because new maximum induces non-zero KL-divergence. Changes in log-likelihood are greater than changes in lower bound.