()

June 4, 2017

Unsupervised vs Supervised Methods

Unsupervised

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised

Unsupervised vs Supervised Methods

Unsupervised$\rightsquigarrow$ estimate categories and categorize documents
Supervised$\rightsquigarrow$ know categories, supervise computer with classification

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents

Supervised $\rightsquigarrow$ know categories, supervise computer with classification

There is NO sense in which there are fewer assumptions in unsupervised methods

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning

Unsupervised vs Supervised Methods

Unsupervised$\rightsquigarrow$ estimate categories and categorize documents
Supervised$\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest$\rightsquigarrow$ do supervised learning
- IF you want to

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to
  - Explore data set

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to
  - Explore data set
  - Discover new categories

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to
  - Explore data set
  - Discover new categories
  - Quickly distill documents

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to $\rightsquigarrow$ do unsupervised learning
  - Explore data set
  - Discover new categories
  - Quickly distill documents

Unsupervised vs Supervised Methods

Unsupervised$\rightsquigarrow$ estimate categories and categorize documents
Supervised$\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest$\rightsquigarrow$ do supervised learning
- IF you want to$\rightsquigarrow$ do unsupervised learning
  - Explore data set
  - Discover new categories
  - Quickly distill documents
- Debate: Unsupervised vs Supervised

Unsupervised vs Supervised Methods

Unsupervised vs Supervised Methods

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to $\rightsquigarrow$ do unsupervised learning
  - Explore data set
  - Discover new categories
  - Quickly distill documents
- Debate: Unsupervised vs Supervised
  - NOT COMPETING METHODS $\rightsquigarrow$ fruitful combination

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to $\rightsquigarrow$ do unsupervised learning
  - Explore data set
  - Discover new categories
  - Quickly distill documents
- Debate: Unsupervised vs Supervised
  - NOT COMPETING METHODS $\rightsquigarrow$ fruitful combination
  - Validate unsupervised methods $\rightsquigarrow$ supervised methods

Unsupervised vs Supervised Methods

Unsupervised $\rightsquigarrow$ estimate categories and categorize documents
Supervised $\rightsquigarrow$ know categories, supervise computer with classification
There is NO sense in which there are fewer assumptions in unsupervised methods

- IF you know categories of interest $\rightsquigarrow$ do supervised learning
- IF you want to $\rightsquigarrow$ do unsupervised learning
  - Explore data set
  - Discover new categories
  - Quickly distill documents
- Debate: Unsupervised vs Supervised
  - NOT COMPETING METHODS $\rightsquigarrow$ fruitful combination
  - Validate unsupervised methods $\rightsquigarrow$ supervised methods
  - Explore heterogeneity in coding $\rightsquigarrow$ unsupervised methods in categories

Low-Dimensional Embeddings

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Name	Party	Vote 1	Vote 2	Vote 3	
Ainsworth, Peter (E S)	Con	NA	1	NA	...
Alexander, Douglas	Lab	NA	0	0	...
Allan, Richard	LD	1	0	1	...
Allen, Graham	Lab	0	0	0	...
Amess, David	Con	1	1	NA	...
					$\ddots$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Text: n speakers, p features in the speeches (often $p > n$ for text problems)

Name	Party	'cost'	'spend'	'tax'	
Ainsworth, Peter (E S)	Con	0.00	0.01	0.30	...
Alexander, Douglas	Lab	0.32	0.20	0.86	...
Allan, Richard	LD	0.99	0.82	0.61	...
Allen, Graham	Lab	0.52	0.86	0.34	...
Amess, David	Con	0.07	0.34	0.33	...
					$\ddots$

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)

Very popular for data summary, exploration (and analysis?)

Aims:

- extract core/important information from data
- reduce the data/problem down to this information
- simplify data
- analyze data in terms of its patterns/groups

Generally: represent this information as new (and smaller number of) variables known as *principal components*

Overview

Features: these principal components will be uncorrelated (orthogonal) with (to) each other, will be linear combinations of original variables

Result: lower dimensional ‘map’ of observations in new space:

$\rightarrow$ each observation now has a value on each principal component called its (factor) score, which are projections of (original) observations onto the PCs

Interpretation of given PC: depends on correlation between component and (original) variable—known as loading

Method: (eigen-) decomposition of cov matrix or singular value decomposition of data matrix

Method

PCA performs a linear transformation on the original variables into new coordinate system, such that the first coordinate (first principal component) is the projection of the original data that contains the most information about that data

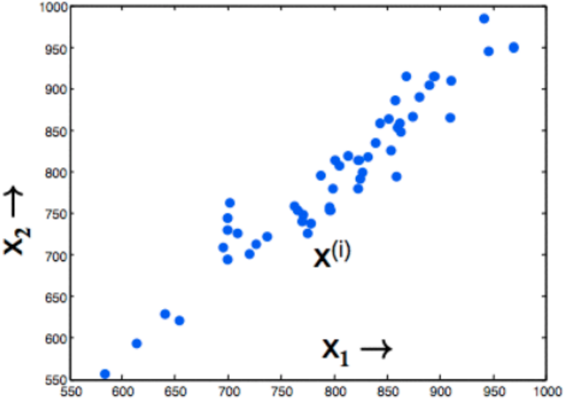
Can think of the first PC as being a line which most closely fits the data points: but, this is in terms of distance perpendicular (orthogonal) to line, not in terms of y-distance (cf OLS)

All subsequent components captures (sequentially) less variability

Assumptions: observations are independent and X is p-variate normal (may not find highest variance projection if not)

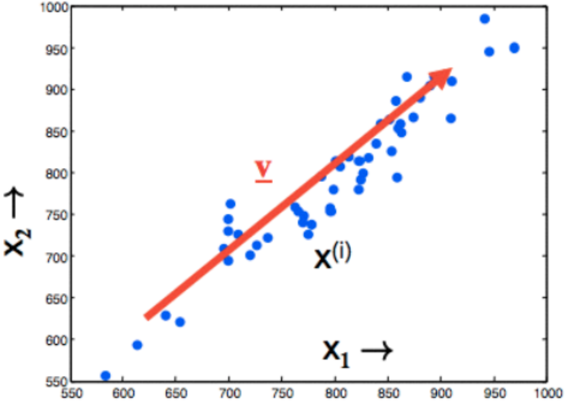
Example

Is the intrinsic dimensionality of this data: 1D; 1.5D, 2D?



Y X

Example

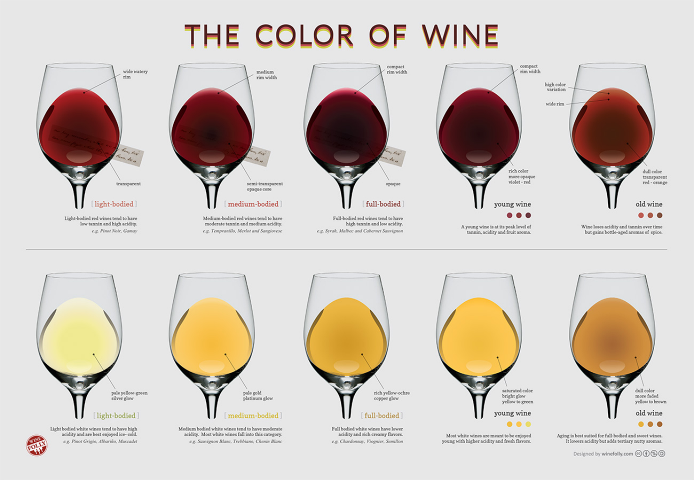


1000
950
900
850
800
750
700
650
600
550
550
600
650
700
750
800
850
900
950
1000
$x_2 \rightarrow$
$x_1 \rightarrow$
$\underline{\mathbf{v}}$
$x^{(i)}$

X Y

PCA - an analogy

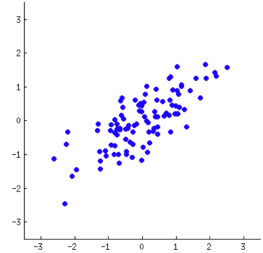
Just a method of summarizing data. Imagine N wine properties. Many are related, therefore redundant. Choose 2 to summarize all wines in your cellar.



PCA - an analogy

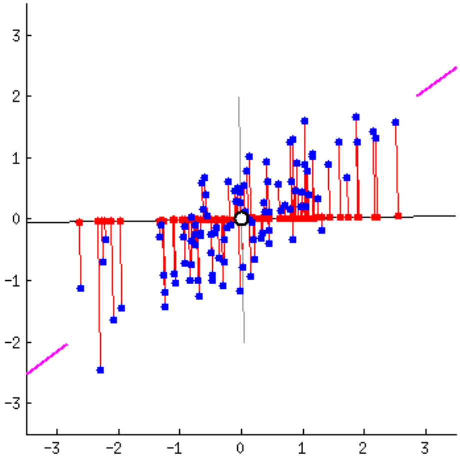
- not keeping some characteristics, discarding others: constructing linear combinations of characteristics (e.g. color = wine age + acidity level);
- PCA finds the best possible characteristics (among all possible linear combinations) to summarize wines in low dimension;
- we still want to discrimminate: we want to look for variation (i.e. properties that strongly differ across wines, that makes them look distinct)
- also looking for properties with prediction properties, that can let us reconstruct original wine characteristics;

PCA - visual intuition



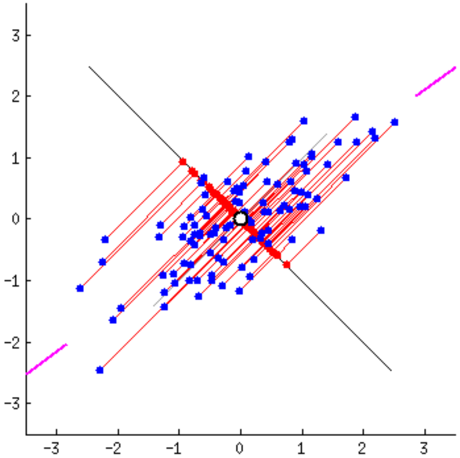
Each dot maps a particular wine onto two correlated properties (x and y). A new property can be constructed drawing a line through the center and projecting all points onto this line.
The new property will be given by linear combination $w_1 x + w_2 y$; let's visualize the projection.

PCA - visual intuition



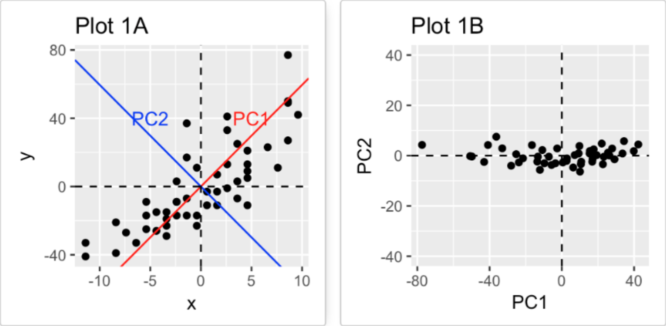
1 variation of values along this line should be maximal (pay attention to spread of red dots – can you see when it reached the maximum?)
2 if we reconstruct original characteristics, blue dots, from the new one, red dots, the reconstruction error will be given by length of the connecting red line (can you see when red line reaches minimum?)
3 Take home message: "maximum variance" and "minimum error" are reached at the same time (!!!) - when the line points to magenta ticks. This line is the new characteristic constructed by PCA - the *first principal component*;

PCA - visual intuition



- PCA will look to minimize the sum of the following square distances:
  - variance: average squared distance from the center of the distribution to each red dot;
  - total reconstruction error: average squared length of red lines;
- imagine black line as a rod and each red line as a spring: the energy of the spring is proportional to its squared length, so rod will orientated itself such as to minimize the sum of these squared distances.pc

Terminology



Y X

- PCA assumes directions with largest variance are most important; picks components that capture largest variation and that are orthogonal to each other; useful in the presence of redundancy (when variables are correlate);
- it turns out that constraining PC2 to be uncorrelated with PC2 is equivalent to constraining direction to be orthogonal;

Terminology

- Eigenvector: almost all vectors (entries in covariance matrix) change direction when multiplied by original covariance matrix S; some exceptional vectors x are in the same direction as Sx: these are eigenvectors. They fulfill property $Ax = \lambda x$, that is, they either stretch or shrink, as determined by λ eigenvalue;
- the amount of variance (spread) retained by each principal component is measured by the eigenvalues(λ); necessarily, eigenvalues for first PC are larger than for subsequent PCs, as the first PC corresponds to direction with maximal variance;

Definition

Suppose $\boldsymbol{A}$ *is an* $N \times N$ *matrix and* λ *is a scalar.*
If

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ *is an eigenvector and* λ *is the associated eigenvalue*

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose $\boldsymbol{A}$ *is an* $N \times N$ *matrix and* λ *is a scalar.*
If

$$\boldsymbol{A}\boldsymbol{x} \;=\; \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ *is an eigenvector and* λ *is the associated eigenvalue*

- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose $\boldsymbol{A}$ *is an* $N \times N$ *matrix and* λ *is a scalar.*
If

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ *is an eigenvector and* λ *is the associated eigenvalue*

- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$
- $\boldsymbol{A}$ stretches $\boldsymbol{x}$ by λ

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose $\boldsymbol{A}$ *is an* $N \times N$ *matrix and* λ *is a scalar.*
If

$$\boldsymbol{A}\boldsymbol{x} \quad = \quad \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ *is an eigenvector and* λ *is the associated eigenvalue*

- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$
- $\boldsymbol{A}$ stretches $\boldsymbol{x}$ by λ
- To find eigenvectors/values: (`eigen` in `R`)

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose $\boldsymbol{A}$ *is an* $N \times N$ *matrix and* λ *is a scalar.*
If

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ *is an eigenvector and* λ *is the associated eigenvalue*

- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$
- $\boldsymbol{A}$ stretches $\boldsymbol{x}$ by λ
- To find eigenvectors/values: (`eigen` in `R`)
  - Find λ that solves $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose $\boldsymbol{A}$ is an $N \times N$ matrix and λ is a scalar.
If

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ is an eigenvector and λ is the associated eigenvalue

- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$
- $\boldsymbol{A}$ stretches $\boldsymbol{x}$ by λ
- To find eigenvectors/values: (`eigen` in `R`)
  - Find λ that solves $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$
  - Find vectors in null space of:

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose $\boldsymbol{A}$ *is an* $N \times N$ *matrix and* λ *is a scalar.*
If

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

Then $\boldsymbol{x}$ *is an eigenvector and* λ *is the associated eigenvalue*

- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$
- $\boldsymbol{A}$ stretches $\boldsymbol{x}$ by λ
- To find eigenvectors/values: (`eigen` in `R`)
  - Find λ that solves $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$
  - Find vectors in null space of:

$$(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

PCA - visual intuition

Consider our covariance matrix:

$$\begin{pmatrix} 1.07 & 0.63 \\ 0.63 & 0.64 \end{pmatrix}$$

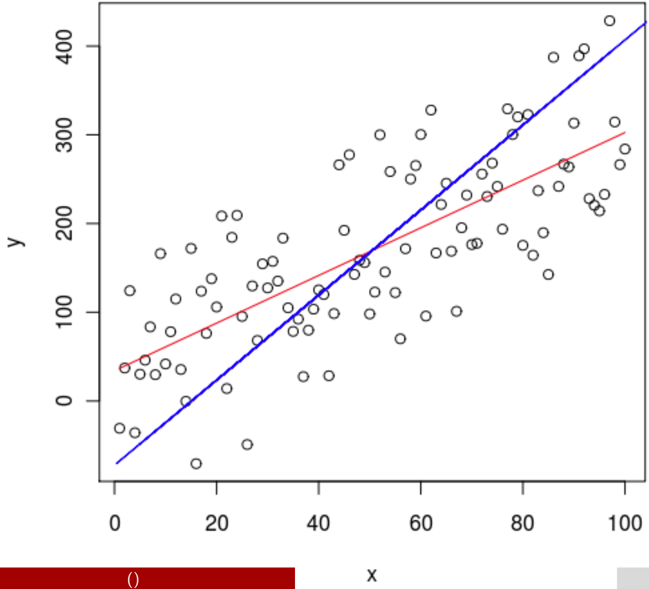
- $\sigma^2{}_x = 1.07$; $\sigma^2{}_y = 0.64$; $Cov_x y = 0.63$;
- a new orthogonal coordinate system is given by its eigenvectors, with corresponding eigenvalues located on the diagonal. In the new coordinate system, covariance matrix looks like:

$$\begin{pmatrix} 1.52 & 0 \\ 0 & 0.19 \end{pmatrix}$$

- correlation between points is now zero; also clear that variance of any projection will be given by weighted average of eigenvalues;
- direction of first component is given by first eigenvector of covariance matrix;
- visually, we can see this on the gray line that forms a rotating coordinate frame: when do blue dots become uncorrelated in this frame?

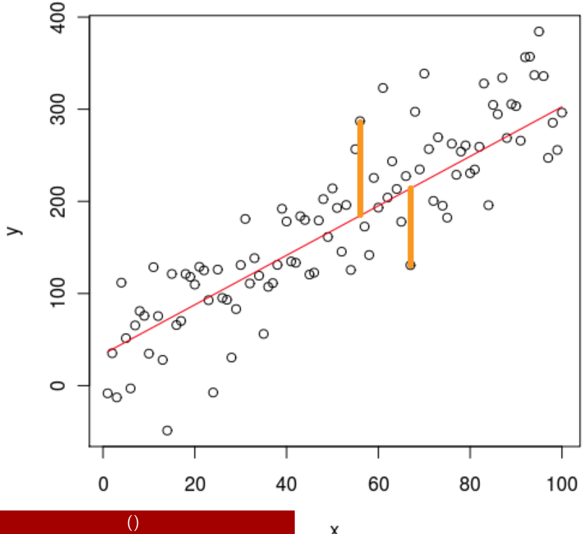
PCA v. OLS

They give different lines. Why?



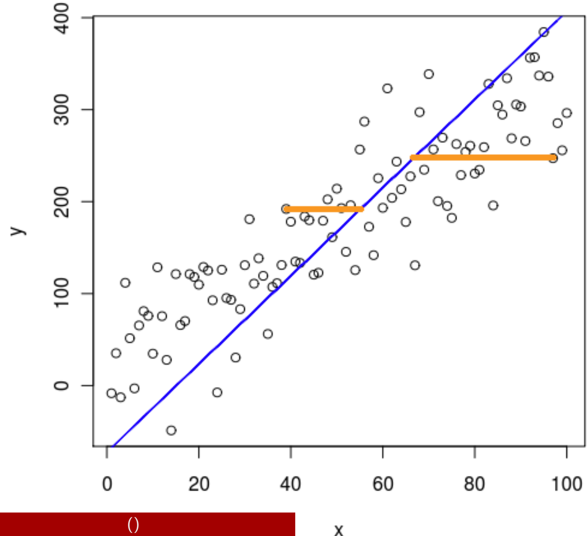
PCA v. OLS

OLS minimizes error between dependent variable and the model [line sits on original y axis of data]; PCA minimizes the error orthogonal (perpendicular) to the model line (orthodogal projection of the data).



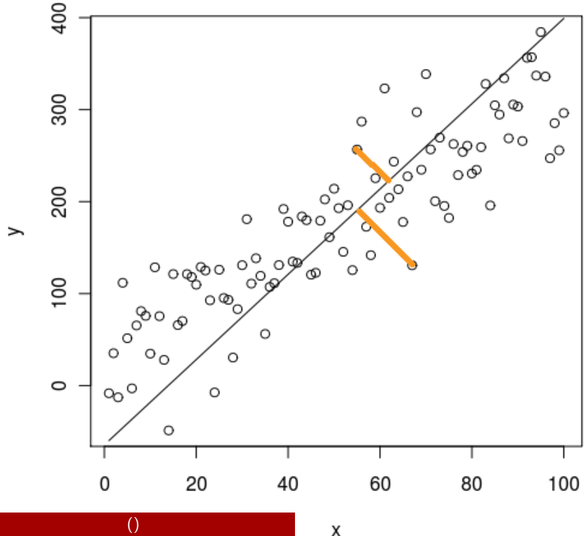
PCA v. OLS

OLS minimizes error between dependent variable and the model [line sits on original y axis of data]; PCA minimizes the error orthogonal (perpendicular) to the model line (orthodogal projection of the data).

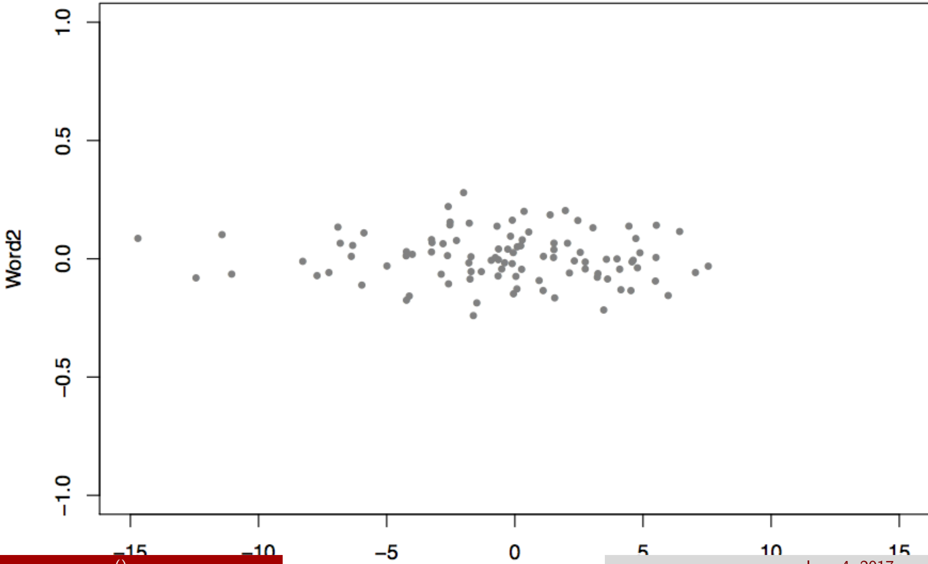


PCA v. OLS

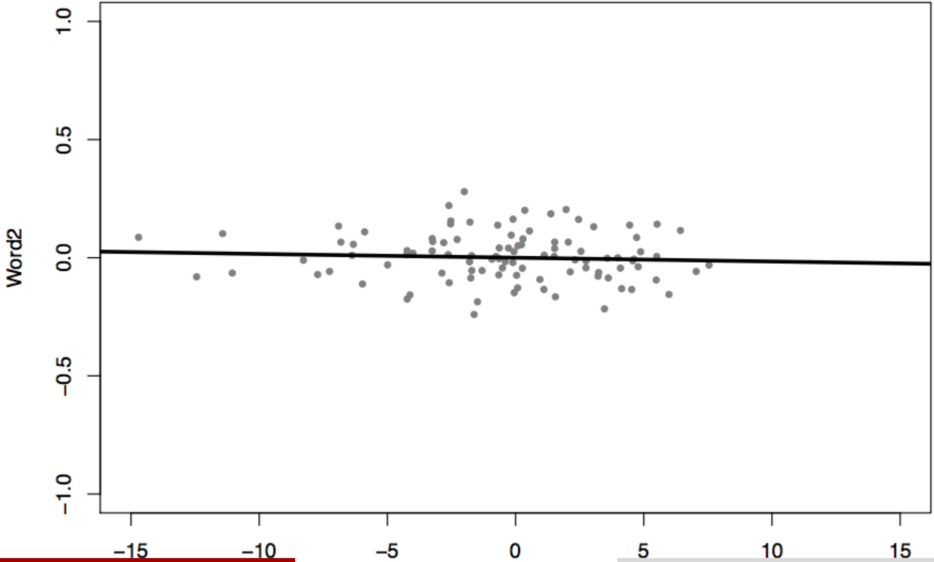
OLS minimizes error between dependent variable and the model [line sits on original y axis of data]; PCA minimizes the error orthogonal (perpendicular) to the model line (orthodogal projection of the data).



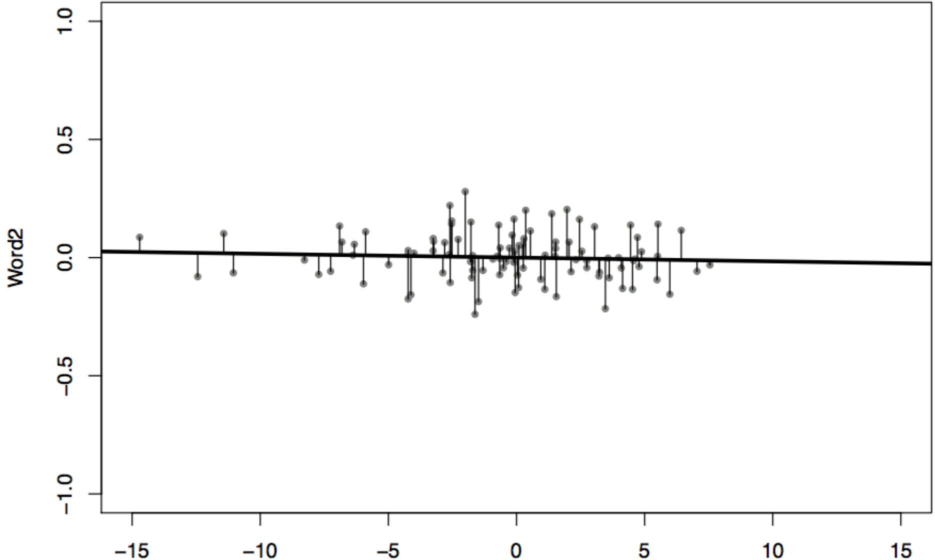
Finding a Lower Dimensional Space (Manifold Learning)



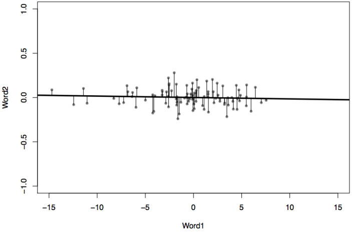
Finding a Lower Dimensional Space (Manifold Learning)



Finding a Lower Dimensional Space (Manifold Learning)

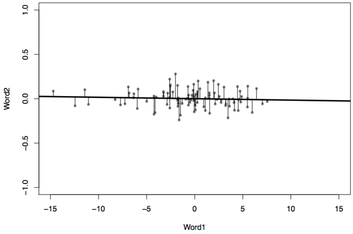


Finding a Lower Dimensional Space (Manifold Learning)



Original data:

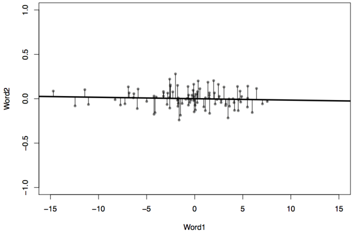
Finding a Lower Dimensional Space (Manifold Learning)



Original data:

$$\boldsymbol{x}_i \quad = \quad (x_{i1}, x_{i2})$$

Finding a Lower Dimensional Space (Manifold Learning)

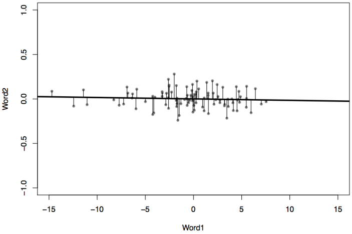


Original data:

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})$$

Which we approximate with

Finding a Lower Dimensional Space (Manifold Learning)



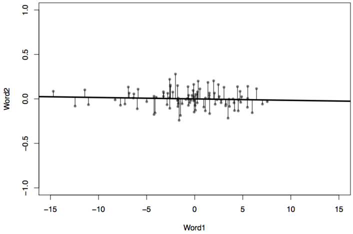
Original data:

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})$$

Which we approximate with

$$\begin{aligned} \tilde{\boldsymbol{x}}_i &= z_i \boldsymbol{w}_1 \\ &= z_i (w_{11}, w_{12}) \end{aligned}$$

Finding a Lower Dimensional Space (Manifold Learning)



Original data $\boldsymbol{x}_i \in \Re^J$

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

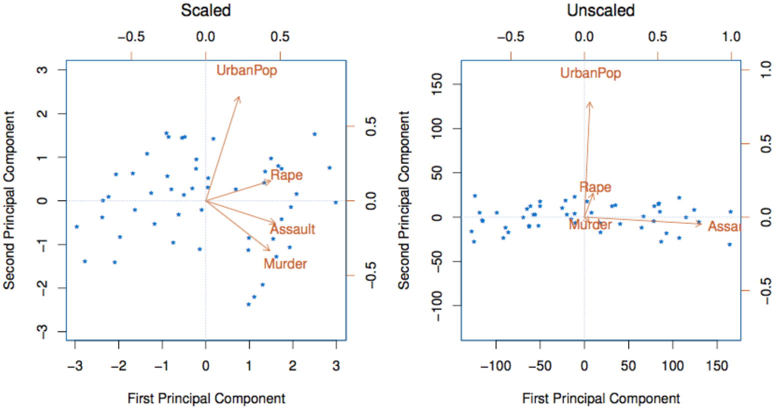
Which we approximate with $L(< J)$ weights z_{il} and vectors $\boldsymbol{w}_l \in \Re^J$

$$\tilde{\boldsymbol{x}}_i = z_{i1}\boldsymbol{w}_1 + z_{i2}\boldsymbol{w}_2 + \ldots + z_{iL}\boldsymbol{w}_L$$

Define $\boldsymbol{\theta} = (\underbrace{\boldsymbol{Z}}_{N \times L}, \underbrace{\boldsymbol{W}_L}_{L \times J})$

Application of Principal Components in R

Note: scale your variables first: $\frac{x_i - mean(x)}{sd(x)}$. It matters:



Application of Principal Components in R

Consider press releases from 2005 US Senators

Application of Principal Components in R

Consider press releases from 2005 US Senators

Define $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ as the rate senator i uses J words.

Application of Principal Components in R

Consider press releases from 2005 US Senators
Define $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ as the rate senator i uses J words.

$$x_{ij} = \frac{\text{No. Times } i \text{ uses word } j}{\text{No. words } i \text{ uses}}$$

Application of Principal Components in R

Consider press releases from 2005 US Senators

Define $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ as the rate senator i uses J words.

$$x_{ij} = \frac{\text{No. Times } i \text{ uses word } j}{\text{No. words } i \text{ uses}}$$

`dtm`: 100×2796 matrix containing word rates for senators

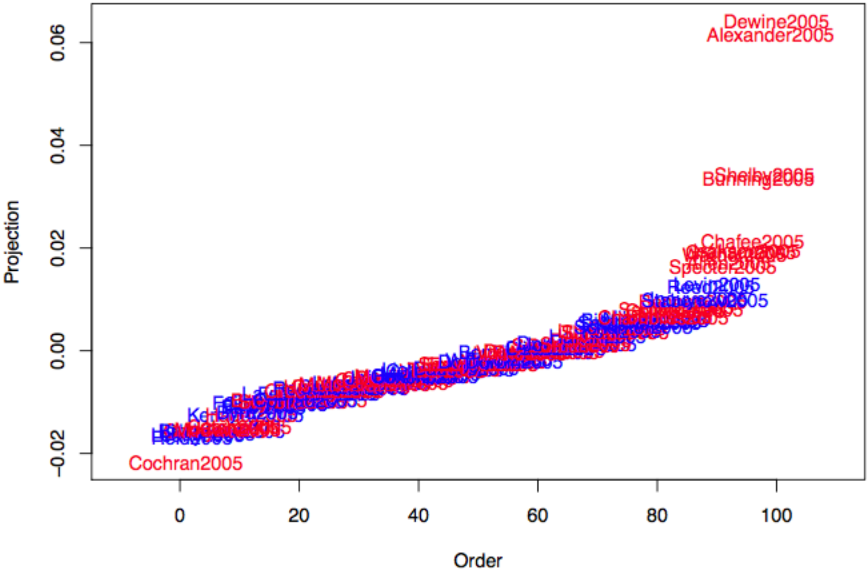
Application of Principal Components in R

Consider press releases from 2005 US Senators
Define $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ as the rate senator i uses J words.

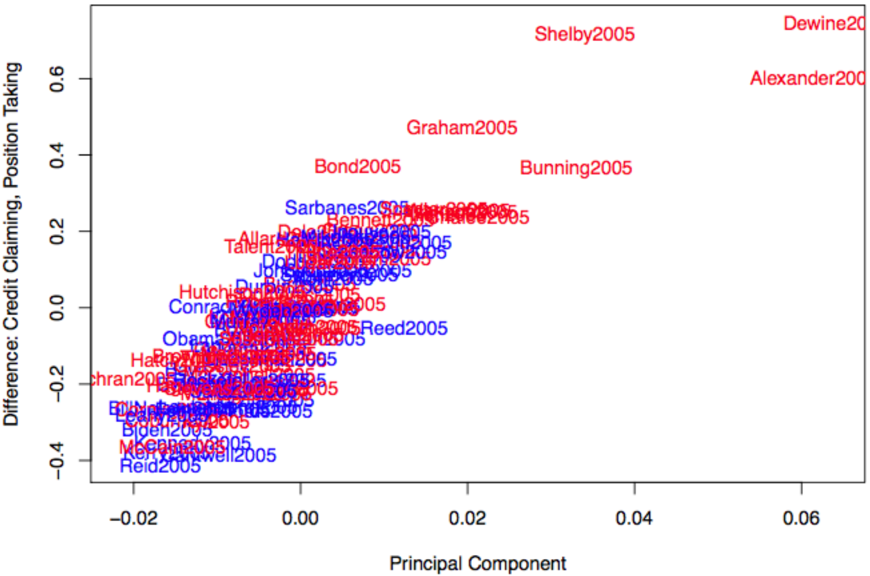
$$x_{ij} = \frac{\text{No. Times } i \text{ uses word } j}{\text{No. words } i \text{ uses}}$$

`dtm`: 100×2796 matrix containing word rates for senators
`prcomp(dtm)` applies principal components

Application of Principal Components in R



Application of Principal Components in R



How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il} \boldsymbol{w}_l||^2$$

Simplifying:

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\text{error}(L) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)$$

Four types of terms:

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)$$

Four types of terms:

1) $\boldsymbol{x}_i^{'}\boldsymbol{x}_i$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)$$

Four types of terms:

1) $\boldsymbol{x}_i^{'}\boldsymbol{x}_i$
2) $z_{ij}z_{ik}\boldsymbol{w}_j^{'}\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$ (orthogonality assumption)

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)$$

Four types of terms:

1) $\boldsymbol{x}_i^{'}\boldsymbol{x}_i$

2) $z_{ij}z_{ik}\boldsymbol{w}_j^{'}\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$ (orthogonality assumption)

3) $z_{ij}z_{ij}\boldsymbol{w}_j^{'}\boldsymbol{w}_j = z_{ij}^2$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)$$

Four types of terms:

1) $\boldsymbol{x}_i^{'}\boldsymbol{x}_i$
2) $z_{ij}z_{ik}\boldsymbol{w}_j^{'}\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$ (orthogonality assumption)
3) $z_{ij}z_{ij}\boldsymbol{w}_j^{'}\boldsymbol{w}_j = z_{ij}^2$
4) $\boldsymbol{x}_i^{'}\sum_{l=1}^{L} z_{il}\boldsymbol{w}_l = \sum_{l=1}^{L} z_{il}^2$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

We want to minimize reconstruction error $\rightsquigarrow$ how well did we do?

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l||^2$$

Simplifying:

$$\begin{aligned}\text{error}(L) &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l)^{'}(\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right)\end{aligned}$$

Four types of terms:

1) $\boldsymbol{x}_i^{'}\boldsymbol{x}_i$
2) $z_{ij}z_{ik}\boldsymbol{w}_j^{'}\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$ (orthogonality assumption)
3) $z_{ij}z_{ij}\boldsymbol{w}_j^{'}\boldsymbol{w}_j = z_{ij}^2$
4) $\boldsymbol{x}_i^{'}\sum_{l=1}^{L} z_{il}\boldsymbol{w}_l = \sum_{l=1}^{L} z_{il}^2$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$\text{error}(L) = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right)$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$\begin{aligned}
\text{error}(L) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} \boldsymbol{w}_l\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_l\right)
\end{aligned}$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$
\begin{aligned}
\text{error}(L) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} \boldsymbol{w}_l\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_l\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \frac{1}{N}\sum_{l=1}^{L}\sum_{l=1}^{N} \boldsymbol{w}_i^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_i
\end{aligned}
$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$
\begin{aligned}
\text{error}(L) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} \boldsymbol{w}_l\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_l\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \frac{1}{N}\sum_{l=1}^{L}\sum_{l=1}^{N} \boldsymbol{w}_i^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_i\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L} \boldsymbol{w}_l^{'}\boldsymbol{\Sigma}\boldsymbol{w}_l \text{where } \Sigma \text{ is an eigenvector}
\end{aligned}
$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$
\begin{aligned}
\text{error}(L) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} \boldsymbol{w}_l\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_l\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \frac{1}{N}\sum_{l=1}^{L}\sum_{l=1}^{N} \boldsymbol{w}_i^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L} \boldsymbol{w}_l^{'}\boldsymbol{\Sigma}\boldsymbol{w}_l \text{where } \Sigma \text{ is an eigenvector} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L} \lambda_l\boldsymbol{w}_l^{'}\boldsymbol{w}_l \text{variance-covariance matrix}
\end{aligned}
$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$
\begin{aligned}
\text{error}(L) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i - \sum_{l=1}^{L} \boldsymbol{w}_l\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_l\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \frac{1}{N}\sum_{l=1}^{L}\sum_{l=1}^{N} \boldsymbol{w}_i^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L} \boldsymbol{w}_l^{'}\boldsymbol{\Sigma}\boldsymbol{w}_l \text{where } \Sigma \text{ is an eigenvector} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L} \lambda_l\boldsymbol{w}_l^{'}\boldsymbol{w}_l \text{variance-covariance matrix} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L} \lambda_l \text{error depends on sum of eigenvalues}
\end{aligned}
$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

If $L = J$, i.e. $L =$ basis vectors, the same number of dimensions as our data, then we can approximate every single data point perfectly

How do we select the number of dimensions L? $\rightsquigarrow$ Model

If $L = J$, i.e. $L =$ basis vectors, the same number of dimensions as our data, then we can approximate every single data point perfectly

$$\text{error}(J) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{J}\lambda_l = 0$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

If $L = J$, i.e. $L =$ basis vectors, the same number of dimensions as our data, then we can approximate every single data point perfectly

$$\text{error}(J) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{J}\lambda_l = 0$$

So for $L < J$,

How do we select the number of dimensions L? $\rightsquigarrow$ Model

If $L = J$, i.e. $L =$ basis vectors, the same number of dimensions as our data, then we can approximate every single data point perfectly

$$\text{error}(J) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) - \sum_{l=1}^{J}\lambda_l = 0$$

So for $L < J$,

$$0 \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) - (\sum_{l=1}^{L}\lambda_l + \sum_{j=L+1}^{J}\lambda_l), \quad \text{where } \sum_{j=L+1}^{J}\lambda_l \text{ are the dimensions not included}$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

If $L = J$, i.e. $L =$ basis vectors, the same number of dimensions as our data, then we can approximate every single data point perfectly

$$\text{error}(J) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{J}\lambda_l = 0$$

So for $L < J$,

$$\begin{aligned} 0 \quad &= \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - (\sum_{l=1}^{L}\lambda_l + \sum_{j=L+1}^{J}\lambda_l), \quad \text{where } \sum_{j=L+1}^{J}\lambda_l \text{ are the dimensions not included} \\ \sum_{j=L+1}^{J}\lambda_l \quad &= \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L}\lambda_l \end{aligned}$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

If $L = J$, i.e. $L =$ basis vectors, the same number of dimensions as our data, then we can approximate every single data point perfectly

$$\text{error}(J) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{J}\lambda_l = 0$$

So for $L < J$,

$$\begin{aligned}
0 \quad &= \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - (\sum_{l=1}^{L}\lambda_l + \sum_{j=L+1}^{J}\lambda_l), \quad \text{where } \sum_{j=L+1}^{J}\lambda_l \text{ are the dimensions not included} \\
\sum_{j=L+1}^{J}\lambda_l \quad &= \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right) - \sum_{l=1}^{L}\lambda_l \\
\sum_{j=L+1}^{J}\lambda_l \quad &= \quad \text{error}(L)
\end{aligned}$$

Error becomes the sum of the remaining eigenvalues; i.e., the eigenvalues we're not using are a measure of how well we're doing

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$\sum_{j=L+1}^{J} \lambda_l \quad = \quad \text{error}(L)$$

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$\sum_{j=L+1}^{J} \lambda_l \quad = \quad \text{error}(L)$$

- Error = Sum of "remaining" eigenvalues

How do we select the number of dimensions L? $\rightsquigarrow$ Model

$$\sum_{j=L+1}^{J} \lambda_l \quad = \quad \text{error}(L)$$

- Error = Sum of "remaining" eigenvalues
- Total variance explained = (sum of included eigenvalues)/(sum of all eigenvalues)

How do we select the number of dimensions L? $\rightsquigarrow$ Model

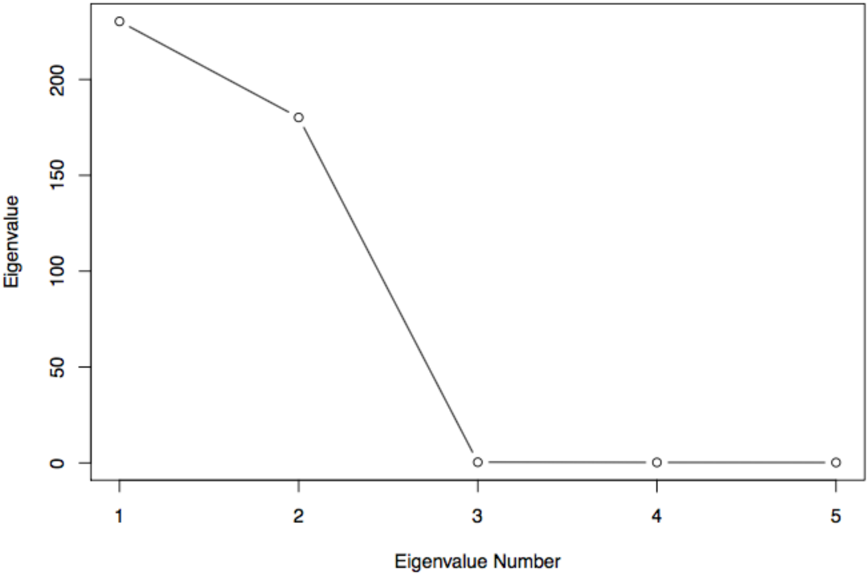
$$\sum_{j=L+1}^{J} \lambda_l \quad = \quad \text{error}(L)$$

- Error = Sum of "remaining" eigenvalues
- Total variance explained = (sum of included eigenvalues)/(sum of all eigenvalues)

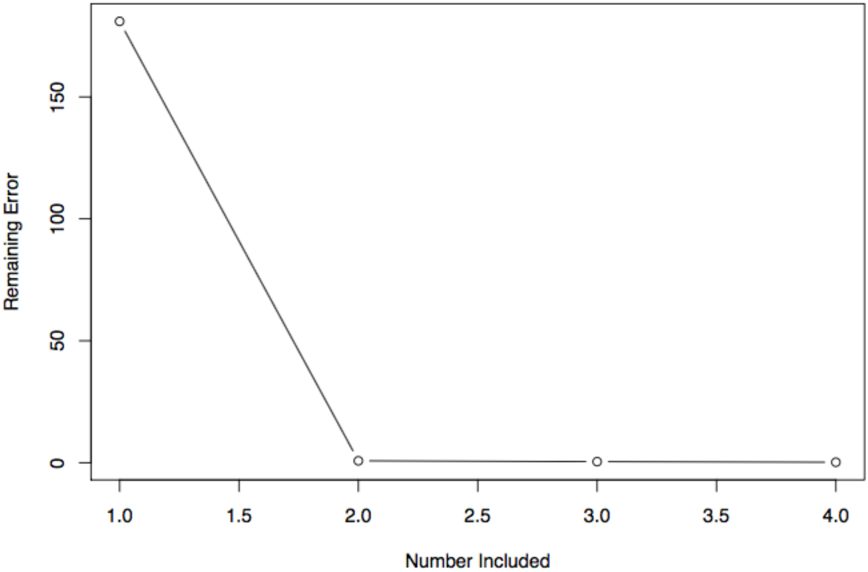
Recommendation $\rightsquigarrow$ look for Elbow

How do we select the number of dimensions L? $\rightsquigarrow$ Model

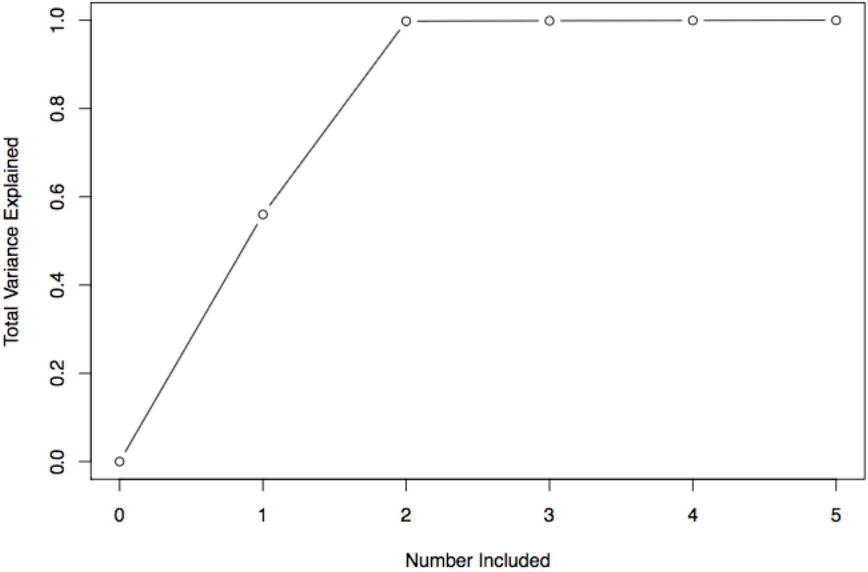
How do we select the number of dimensions L? $\rightsquigarrow$ Model



How do we select the number of dimensions L? $\rightsquigarrow$ Model



How do we select the number of dimensions L? $\rightsquigarrow$ Model



Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$?

Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$? J

Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$? J(!!!!!)

Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$? J(!!!!!)

- Attempts to assess dimensionality require a model $\rightsquigarrow$ some way to tradeoff accuracy of reconstruction with simplicity

Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$? J(!!!!!)

- Attempts to assess dimensionality require a model $\rightsquigarrow$ some way to tradeoff accuracy of reconstruction with simplicity
- Any answer (no matter how creatively obtained) supposes you have the right function to measure tradeoff

Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$? J(!!!!!)

- Attempts to assess dimensionality require a model $\rightsquigarrow$ some way to tradeoff accuracy of reconstruction with simplicity
- Any answer (no matter how creatively obtained) supposes you have the right function to measure tradeoff
- The "right" number of dimensions depends on the task you have in mind

Non-model based evaluations: What's the point?

What is the true underlying dimensionality of $\boldsymbol{X}$? J(!!!!!)

- Attempts to assess dimensionality require a model$\rightsquigarrow$ some way to tradeoff accuracy of reconstruction with simplicity
- Any answer (no matter how creatively obtained) supposes you have the right function to measure tradeoff
- The "right" number of dimensions depends on the task you have in mind

Mathematical model$\rightsquigarrow$ insufficient to make modeling decision

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Spirling and Indian Treaties

How do we preserve word order and semantic language?

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
```

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
```

```
- No Peace Between Us
```

are identical.

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?

After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.

Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

`- Peace Between Us`

`- No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

`- Peace Between Us`

`- No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?

After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.

Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ broad application

`Peace Between Us`

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.
We say $\boldsymbol{x}_i \in \mathcal{X}$

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.

We say $\boldsymbol{x}_i \in \mathcal{X}$

Spirling examines 5-character strings, $s \in \mathcal{A}$

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.
We say $\boldsymbol{x}_i \in \mathcal{X}$
Spirling examines 5-character strings, $s \in \mathcal{A}$
Define:

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.

We say $\boldsymbol{x}_i \in \mathcal{X}$

Spirling examines 5-character strings, $s \in \mathcal{A}$

Define:

$\phi_s : \mathcal{X} \rightarrow \Re$ as a function that counts the number of times string s occurs in document $\boldsymbol{x}$.

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.

We say $\boldsymbol{x}_i \in \mathcal{X}$

Spirling examines 5-character strings, $s \in \mathcal{A}$

Define:

$\phi_s : \mathcal{X} \rightarrow \Re$ as a function that counts the number of times string s occurs in document $\boldsymbol{x}$.

Define string kernel to be,

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.
We say $\boldsymbol{x}_i \in \mathcal{X}$
Spirling examines 5-character strings, $s \in \mathcal{A}$
Define:
$\phi_s : \mathcal{X} \rightarrow \Re$ as a function that counts the number of times string s occurs in document $\boldsymbol{x}$.
Define string kernel to be,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{s \in \mathcal{A}} w_s \phi_s(\boldsymbol{x}_i) \phi_s(\boldsymbol{x}_j)$$

Spirling and Indian Treaties

Consider documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where we have preserved order, punctuation, and all else.

We say $\boldsymbol{x}_i \in \mathcal{X}$

Spirling examines 5-character strings, $s \in \mathcal{A}$

Define:

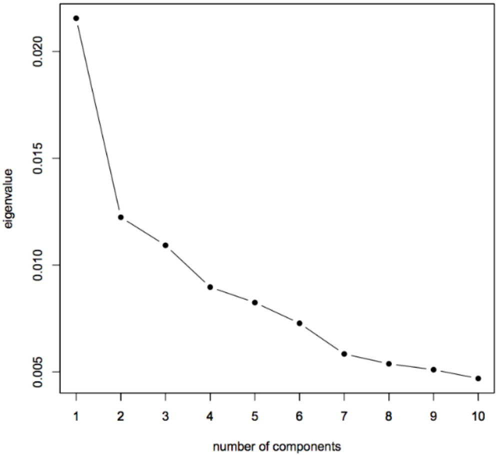
$\phi_s : \mathcal{X} \rightarrow \Re$ as a function that counts the number of times string s occurs in document $\boldsymbol{x}$.

Define string kernel to be,

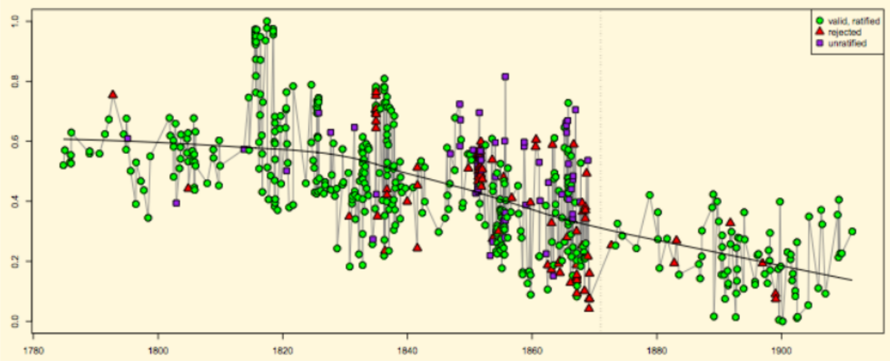
$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{s \in \mathcal{A}} w_s \phi_s(\boldsymbol{x}_i)\phi_s(\boldsymbol{x}_j)$$

$\phi(\boldsymbol{x}_i) \approx \binom{32}{5}$ element long count vector

Spirling and Indian Treaties



Spirling and Indian Treaties

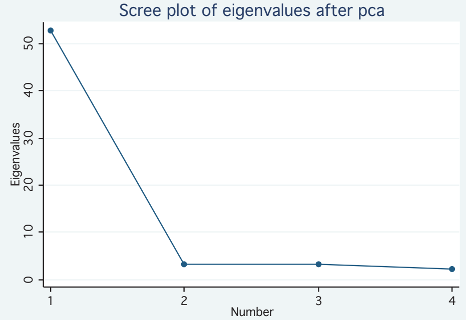


Political Speech: US Senate

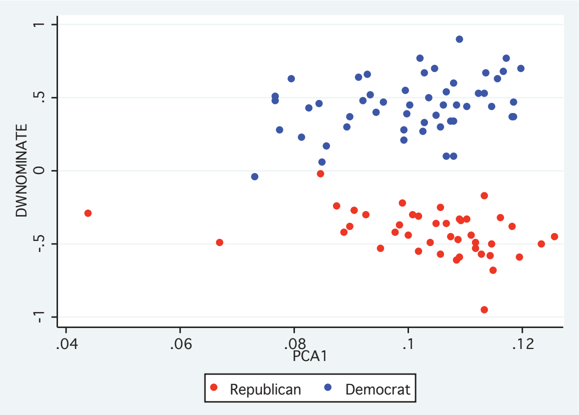
Beauchamp, 2010 (Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons)

Considers PCA of (pre-processed) 1000-top-vectors for US Senators.

Fits several components, of which 1PC model looks very good...



Partner Exercise



Strangely, in Beauchamp's work, PC1 uncorrelated with first dimension of roll calls scores.

why?

Unsupervised Clustering

Fully Automated Clustering$\rightsquigarrow$ Discovering Categories and Classifying Documents

1) Task
    a) Discovering categories and placing documents in those categories
    b) Partitioning documents into similar groups
2) Objective function
    a) What makes a pair of documents similar (dissimilar)?
    b) What makes a good clustering of texts?

$$f(\boldsymbol{X}, \boldsymbol{\theta}) \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   where:
    - $\boldsymbol{\Theta}$ = parameters that describe clusters $J \times K \rightsquigarrow$ unigram model
    - $\boldsymbol{T}$ = cluster assignments for each observation $N \times K$
3) Optimization
    - Algorithms search over $\boldsymbol{T}$ and $\boldsymbol{\Theta}$
    - Expectation-Maximization Algorithm
4) Validation
    1) Model based$\rightsquigarrow$ Exclusive/Cohesive
    2) Human based$\rightsquigarrow$ Experiments to detect properties

K-Means$\rightsquigarrow$ Objective Function

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

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

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

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

   $\boldsymbol{\theta}_k =$ exemplar for cluster k
2) $\boldsymbol{T}$ is an $N \times J$ 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 $N \times J$ 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 $N \times J$ 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

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

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

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

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

   Hard Assignment

K-Means⇝ Objective Function

Assume squared euclidean distance

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

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

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

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

- 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

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

- 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 $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sigma^2$

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 $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sigma^2$
    - 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 $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sigma^2$
    - 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 $\mathbf{\Theta}^{t-1}$ (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 $\mathbf{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$
- Conditional on $\boldsymbol{T}^t$, choose $\mathbf{\Theta}^t$

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$\rightsquigarrow$ Optimization

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

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

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

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

K-Means$\rightsquigarrow$ Optimization

Optimization algorithm:

Optimization algorithm:

- Initialize centers

K-Means $\rightsquigarrow$ Optimization

Optimization algorithm:

- Initialize centers
- Do until converged:

K-Means$\rightsquigarrow$ Optimization

Optimization algorithm:

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

K-Means$\rightsquigarrow$ Optimization

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

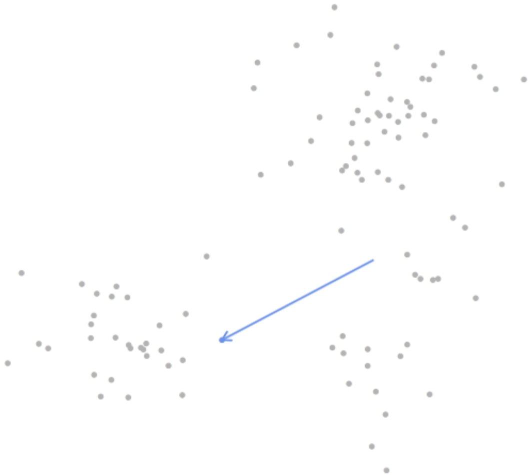
Optimization algorithm:

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

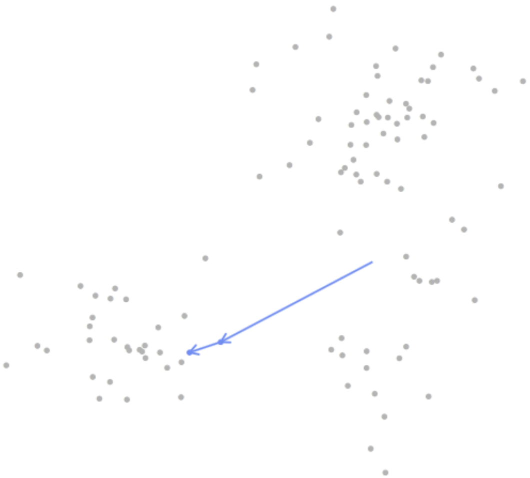
Visual Example

Visual Example

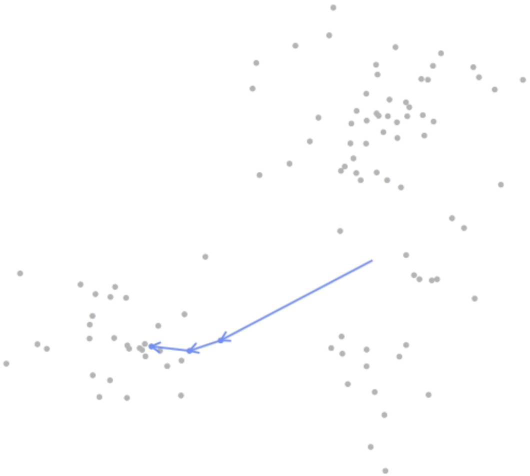
Visual Example



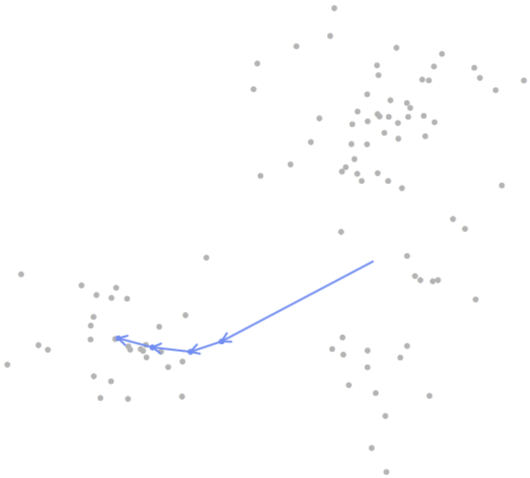
Visual Example



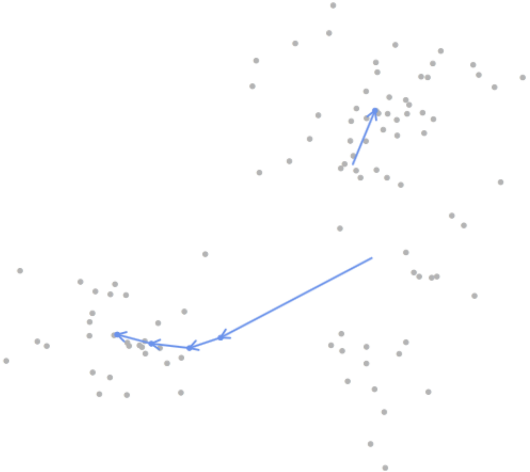
Visual Example



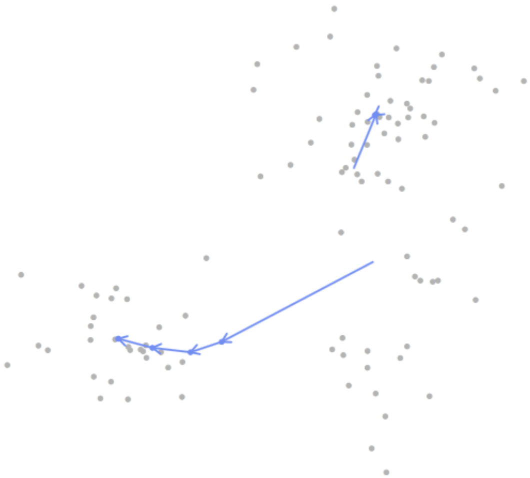
Visual Example



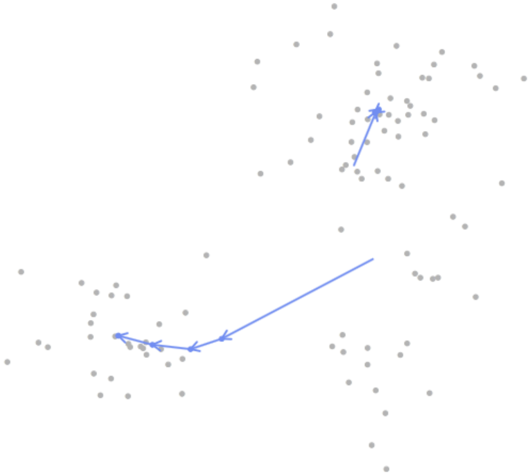
Visual Example



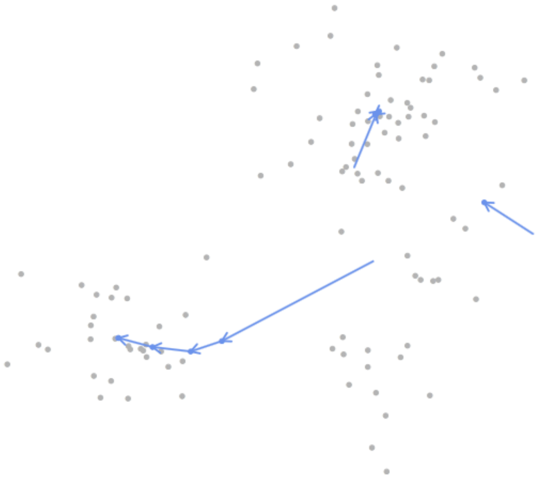
Visual Example



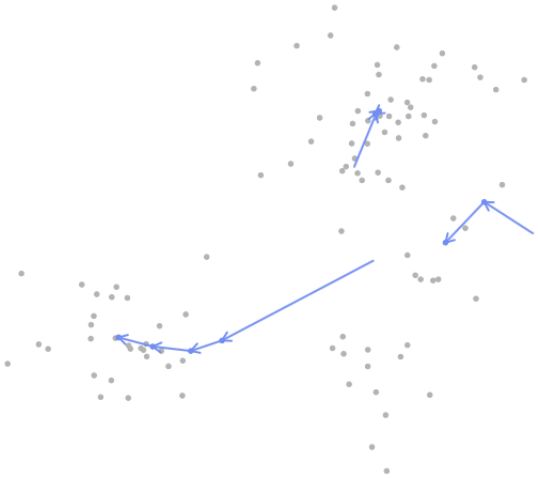
Visual Example



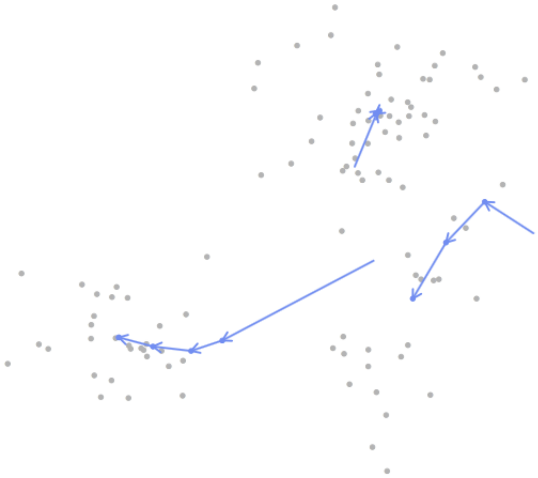
Visual Example



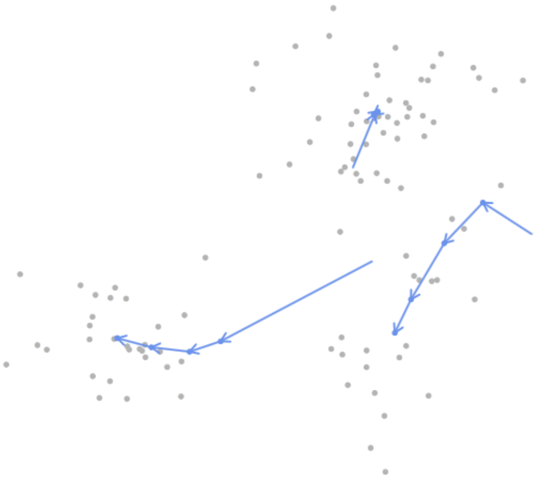
Visual Example



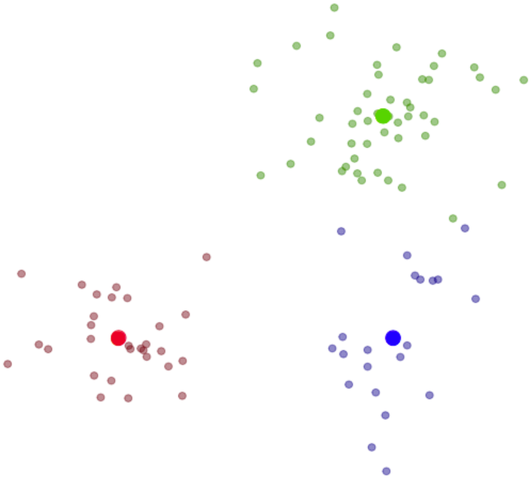
Visual Example



Visual Example



Visual Example



Data Step 1 Iteration 1, Step 2a

Instability & local optima

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 $\rightsquigarrow$ 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

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

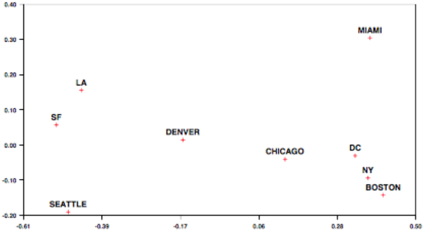
- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents
        - Assign cluster label
    - Automatic identification
    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

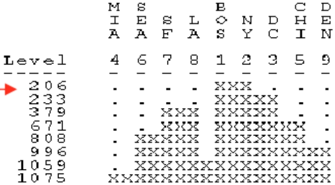
Interpreting Cluster Components

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

Hierarchical Clustering

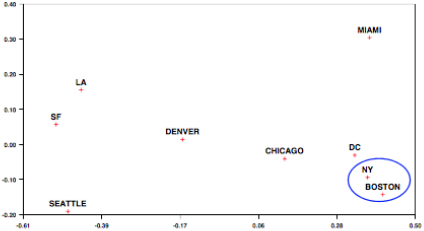




	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Closest distance is NY-BOS = 206, so merge these.

Hierarchical Clustering

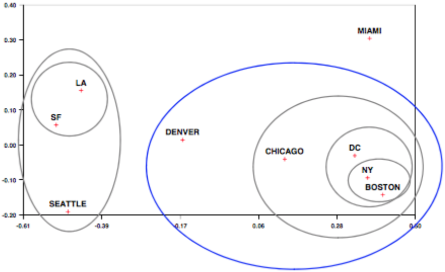



```
           M  S           B           C  D
           I  E  S  L     O  N  D     H  E
           A  A  F  A     S  Y  C     I  N
Level      4  6  7  8     1  2  3     5  9
-----      -  -  -  -     -  -  -     -  -
  206      .  .  .  .     XXX         .  .
  233      .  .  .  .     XXXXX       .  .
  379      .  .  XXX      XXXXX       .  .
  671      .  .  XXX      XXXXXXX        .
  808      .  XXXXX       XXXXXXX        .
  996      .  XXXXX       XXXXXXXXXX
 1059      .  XXXXXXXXXXXXXXXXXXX
 1075      XXXXXXXXXXXXXXXXXXXXXXX
```

	BOS NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/ NY	0	233	1308	802	2815	2934	2786	1771
DC	233	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

Closest pair is DC to BOSNY combo @ 233. So merge these.

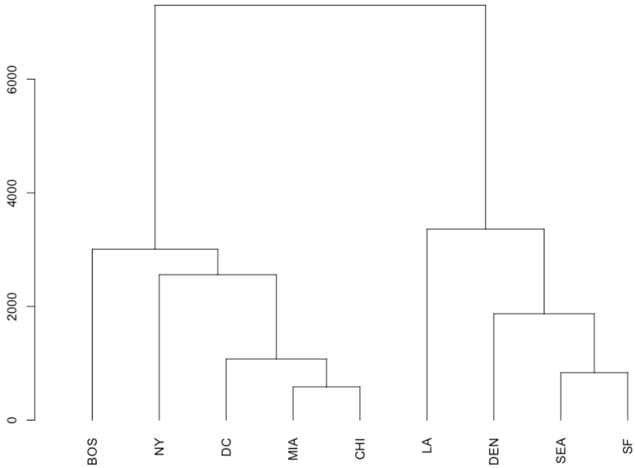
Hierarchical Clustering




```
       M S       B       C D
       I E S L   O N D   H E
       A A F A   S Y C   I N
Level  4 6 7 8   1 2 3   5 9
-----  - - - -   - - -   - -
 206   . . . .   XXX .   . .
 233   . . . .   XXXXX   . .
 379   . . XXX   XXXXX   . .
 671   . . XXX   XXXXXXXXX .
 808   . XXXXX   XXXXXXXXX .
 996   . XXXXX   XXXXXXXXXXX
1059   . XXXXXXXXXXXXXXXXXXX
1075   XXXXXXXXXXXXXXXXXXXXX
```

	BOS/NY/DC/CHI/DEN	MIA	SF/LA/SEA
BOS/NY/DC/CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0

Hierarchical Clustering



Hierarchical Clustering

Linkage	*Description*
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

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
- 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
- Humans should be the final judge
  - Compare insights across clusterings

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

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
  - 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
  - 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
  - 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
  - Validation: model based fit statistics
- How do we know we have the “right” model?

YOU DON’T! $\rightsquigarrow$ And never will

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization
    - Select the best model.

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization
    - Select the best model.
        - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization
    - Select the best model.
        - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
        - Stats + Substance to select model + K
4) Validation

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization
    - Select the best model.
        - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
        - Stats + Substance to select model + K
4) Validation
    - Is our statistic capturing what we want from the clustering?

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization
    - Select the best model.
        - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
        - Stats + Substance to select model + K
4) Validation
    - Is our statistic capturing what we want from the clustering?
    - Are there features we're missing

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K
4) Validation
   - Is our statistic capturing what we want from the clustering?
   - Are there features we're missing
   - Very Open Research Question

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i'\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

A Motivating Clustering Model$\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* \quad = \quad \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\boldsymbol{\tau}_i \quad \sim \quad \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}}$$

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

A Motivating Clustering Model$\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

Provides:

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment
- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \rightsquigarrow$ Proportion of documents in each component

A Motivating Clustering Model$\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* \quad = \quad \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned} \boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\ \boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}} \end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment
- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \rightsquigarrow$ Proportion of documents in each component
- $\boldsymbol{\mu}_k \rightsquigarrow$ Exemplar document for cluster k

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned} \boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\ \boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}} \end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment
- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \rightsquigarrow$ Proportion of documents in each component
- $\boldsymbol{\mu}_k \rightsquigarrow$ Exemplar document for cluster k

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?

Problem

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\leadsto$ predict new documents?
Problem $\leadsto$ in sample evaluation leads to overfit.

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?

Problem$\rightsquigarrow$ in sample evaluation leads to overfit.

Solution$\rightsquigarrow$ evaluate performance on held out data

For held out document $\boldsymbol{x}^*_{\text{out}}$

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?
Problem$\rightsquigarrow$ in sample evaluation leads to overfit.
Solution$\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^*_{\text{out}}$

$$\log p(\boldsymbol{x}^*_{\text{out}}|\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{X}) \quad = \quad \log \sum_{k=1}^{K} p(\boldsymbol{x}^*_{\text{out}}, \tau_{ik}|\boldsymbol{\mu}_k, \boldsymbol{\pi}, \boldsymbol{X})$$

Measuring Cluster Performance: Out of Sample Prediction

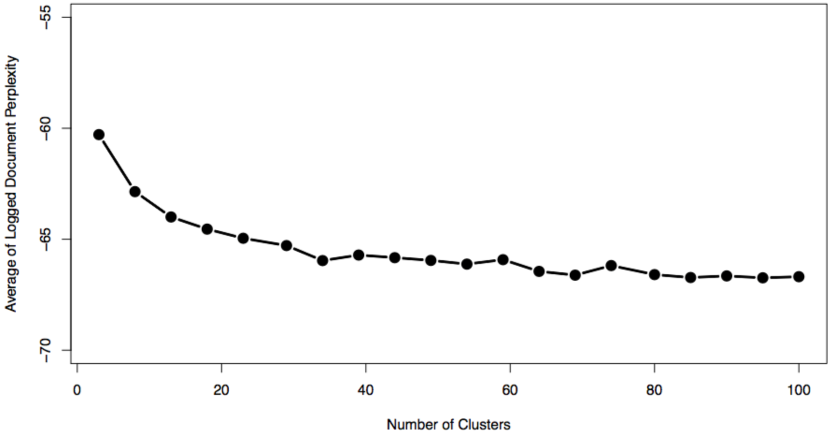
How well does our model perform?$\rightsquigarrow$ predict new documents?
Problem$\rightsquigarrow$ in sample evaluation leads to overfit.
Solution$\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^*_{\text{out}}$

$$\begin{aligned}
\log p(\boldsymbol{x}^*_{\text{out}}|\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{X}) &= \log \sum_{k=1}^{K} p(\boldsymbol{x}^*_{\text{out}}, \tau_{ik}|\boldsymbol{\mu}_k, \boldsymbol{\pi}, \boldsymbol{X}) \\
&= \log \sum_{k=1}^{K} \left[\pi_k \exp(\kappa \boldsymbol{\mu}_k^{'} \boldsymbol{x}^*_{\text{out}})\right]
\end{aligned}$$

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^*_{\text{out}}$

$$
\begin{aligned}
\log p(\boldsymbol{x}^*_{\text{out}} | \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{X}) &= \log \sum_{k=1}^{K} p(\boldsymbol{x}^*_{\text{out}}, \tau_{ik} | \boldsymbol{\mu}_k, \boldsymbol{\pi}, \boldsymbol{X}) \\
&= \log \sum_{k=1}^{K} \left[\pi_k \exp(\kappa \boldsymbol{\mu}_k^{'} \boldsymbol{x}^*_{\text{out}}) \right] \\
\text{Perplexity}_{\text{word}} &= \exp\left(- \log p(\boldsymbol{x}^*_{\text{out}} | \boldsymbol{\mu}, \boldsymbol{\pi})\right)
\end{aligned}
$$


Average of Logged Document Perplexity
−55
−60
−65
−70
0
20
40
60
80
100
Number of Clusters

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it?

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction$\rightsquigarrow$ One Task
- Do we care about it?$\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction$\leadsto$ One Task
- Do we care about it?$\leadsto$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy$\leadsto$ measure quality in topics and clusters

(Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy $\rightsquigarrow$ measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al AJPS 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy $\rightsquigarrow$ measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al AJPS 2014)
- Experiments: measure topic and cluster quality

Measuring Cohesiveness and Exclusivity

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic

Measuring Cohesiveness and Exclusivity

- Consider the output of clustering model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic
- For example $\boldsymbol{v}_3 =$ (earmark , egregious , pork , fiscal , today)

Measuring Cohesiveness and Exclusivity

To measure cohesiveness we examine the extent to which two words that indicate a document belongs to a cluster actually co-occur in the documents that belong to that cluster. $D(m_1, m_2)$ will count the number of times the words m_1 and m_2 co-occur in documents, where $D(m_1)$ counts the number of documents in which the word m_1 appears.
Define the function D as a function that counts the number of times its argument occurs:

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$D(\text{earmark}, \text{egregious}) \quad = \quad \text{No. times earmark and egregious co-occur}$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log\left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})}\right)$$

Define overall cohesiveness as:

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$
\begin{aligned}
D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\
D(\text{egregious}) &= \text{Number of times Egregious occurs}
\end{aligned}
$$

Define cohesiveness for topic k as

$$
\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)
$$

Define overall cohesiveness as:

$$
\text{Cohesive} = (\sum_{k=1}^{K} \text{Cohesive}_k)/K
$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$
\begin{aligned}
D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\
D(\text{egregious}) &= \text{Number of times Egregious occurs}
\end{aligned}
$$

Define cohesiveness for topic k as

$$
\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)
$$

Define overall cohesiveness as:

$$
\begin{aligned}
\text{Cohesive} &= (\sum_{k=1}^{K} \text{Cohesive}_k)/K \\
&= \left(\sum_{k=1}^{K} \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right) \right) /K
\end{aligned}
$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive

Suppose that each cluster has a center vector $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k}, \ldots, mu_{Jk}$ where μjk describes the weight attached to the j^{th} word in cluster k. For each cluster, we want to select the M largest weights. For each word $m \in M$ we can define exclusivity as the ratio between the weight of word m in topic k and the um of weight of word m across all topics:

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

$$\text{Exclusivity}_k = \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

$$\begin{aligned}\text{Exclusivity}_k &= \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \\ \text{Exclusivity} &= \left(\sum_{k=1}^{K} \text{Exclusivity}_k \right) / K\end{aligned}$$

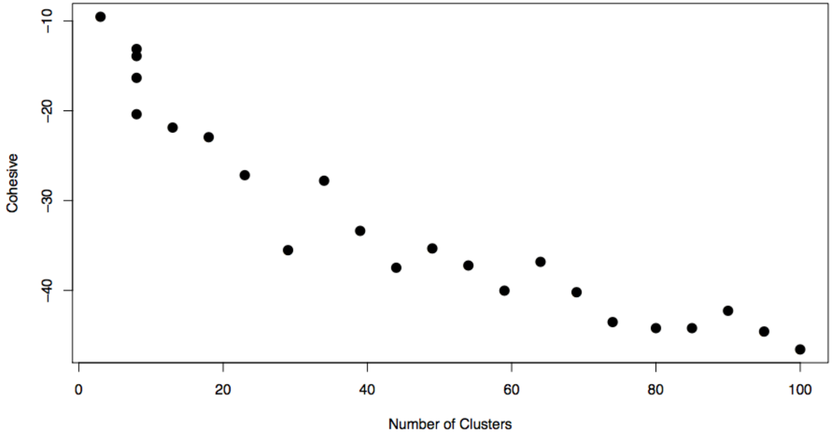
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

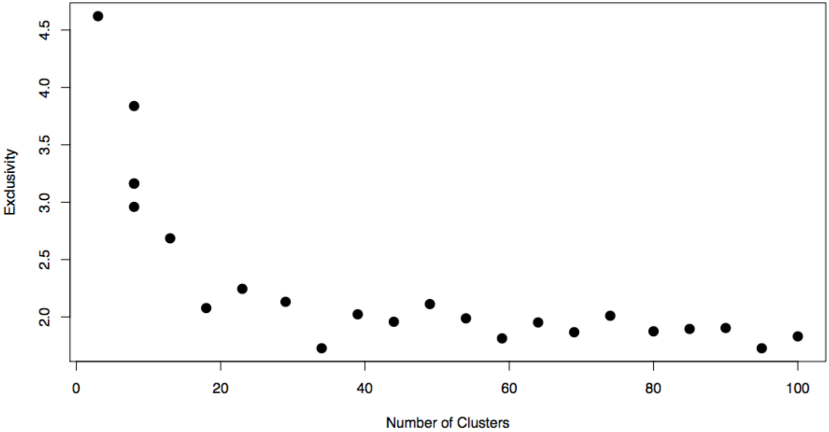
$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

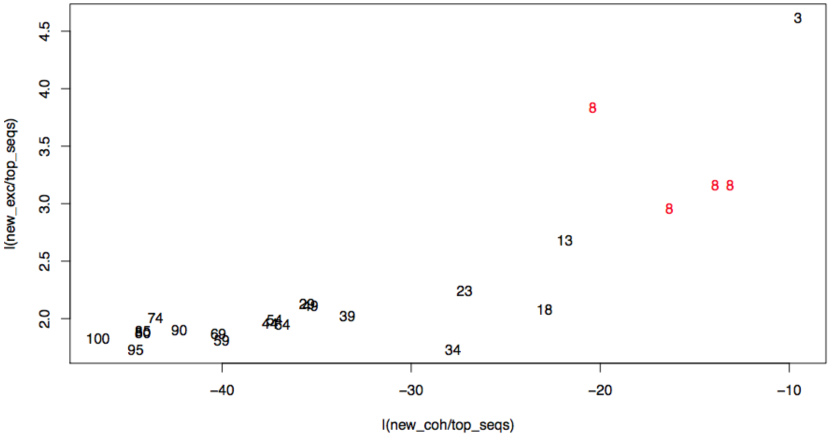
Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

$$\begin{aligned}
\text{Exclusivity}_k &= \sum_{j: v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \\
\text{Exclusivity} &= \left(\sum_{k=1}^{K} \text{Exclusivity}_k \right) / K \\
&= \left(\sum_{k=1}^{K} \sum_{j: v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \right) / K
\end{aligned}$$



Cohesive
−10
−20
−30
−40
0
20
40
60
80
100
Number of Clusters


Exclusivity
4.5
4.0
3.5
3.0
2.5
2.0
0
20
40
60
80
100
Number of Clusters


I(new_exc/top_seqs)
4.5
4.0
3.5
3.0
2.5
2.0
3
8
8
8
8
13
23
18
39
34
74
90
100
95
−40
−30
−20
−10
I(new_coh/top_seqs)

Experimental Approaches

Mathematical approaches

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation

Experimental Approaches

Mathematical approaches$\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model$\rightsquigarrow$ including text representation
Humans$\rightsquigarrow$ read texts

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation
Humans $\rightsquigarrow$ read texts
Humans $\rightsquigarrow$ use cluster output

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers

assumes we're in the model $\rightsquigarrow$ including text representation

Humans $\rightsquigarrow$ read texts

Humans $\rightsquigarrow$ use cluster output

Do humans think the model is performing well?

Experimental Approaches

Mathematical approaches$\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model$\rightsquigarrow$ including text representation
Humans$\rightsquigarrow$ read texts
Humans$\rightsquigarrow$ use cluster output
Do humans think the model is performing well?

1) Topic Quality

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation
Humans $\rightsquigarrow$ read texts
Humans $\rightsquigarrow$ use cluster output
Do humans think the model is performing well?

1) Topic Quality
2) Cluster Quality

Experimental Approaches

1) Take M top words for a topic
2) Randomly select a top word from another topic
   2a) Sample the topic number from l from $K - 1$ (uniform probability)
   2b) Sample word j from the M top words in topic l
   2c) Permute the words and randomly insert the intruder:
      - List:

$$\text{test} = (v_{k,3}, v_{k,1}, v_{l,j}, v_{k,2}, v_{k,4}, v_{k,5})$$

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`bowl, flooding, olympic, olympics, nfl, coach`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`bowl, flooding, olympic, olympics, nfl, coach`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`stocks, investors, fed, guns, trading, earning`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`stocks, investors, fed, guns, trading, earning`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

Higher rate of intruder identification $\rightsquigarrow$ more exclusive/cohesive topics

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

Higher rate of intruder identification $\rightsquigarrow$ more exclusive/cohesive topics

Deploy on Mechanical Turk

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)
- Select clustering with highest cluster quality

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)
- Select clustering with highest cluster quality
- Can be used to compare any clusterings, regardless of source

How do we Choose K?

Generate many candidate models

1) Assess Cohesiveness/Exclusivity, select models on frontier
2) Use experiments
3) Read
4) Final decision $\rightsquigarrow$ combination

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):

k-means , Mixture of multinomials

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):

k-means , Mixture of multinomials , k-medoids , affinity propagation

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation , agglomerative Hierarchical

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation , agglomerative Hierarchical fuzzy k-means, trimmed k-means, k-Harmonic means, fuzzy k-medoids, fuzzy k modes, maximum entropy clustering, model based hierarchical (agglomerative), proximus, ROCK, divisive hierarchical, DISMEA, Fuzzy, QTClust, self-organizing map, self-organizing tree, unnormalized spectral, MS spectral, NJW Spectral, SM Spectral, Dirichlet Process Multinomial, Dirichlet Process Normal, Dirichlet Process von-mises Fisher, Mixture of von mises-Fisher (EM), Mixture of von Mises Fisher (VA), Mixture of normals, co-clustering mutual information, co-clustering SVD, LLAhclust, CLUES, bclust, c-shell, qtClustering, LDA, Express Agenda Model, Hierarchical Dirichlet process prior, multinomial, uniform process mulitinomial, Chinese Restaurant Distance Dirichlet process multinomial, Pitmann-Yor Process multinomial, LSA, ...

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method —

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
    - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
    - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,. . .
    - Well-defined statistical, data analytic, or machine learning foundations
    - How to add substantive knowledge: With few exceptions, unclear

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible

Deep problem in cluster analysis literature: full automation requires more information

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
    - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
    - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering (local ensemble aggregates different clustering methods to create a single clustering).

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering (local ensemble aggregates different clustering methods to create a single clustering).
   - New Clustering: weighted average of clusterings from methods

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
    - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering (local ensemble aggregates different clustering methods to create a single clustering).
    - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering (local ensemble aggregates different clustering methods to create a single clustering).
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)
7) $\rightsquigarrow$ Millions of clusterings easily comprehended

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering (local ensemble aggregates different clustering methods to create a single clustering).
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)
7) $\rightsquigarrow$ Millions of clusterings easily comprehended
8) (Or, our new strategy: represent entire Bell space directly; no need to examine document contents)

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)

- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking

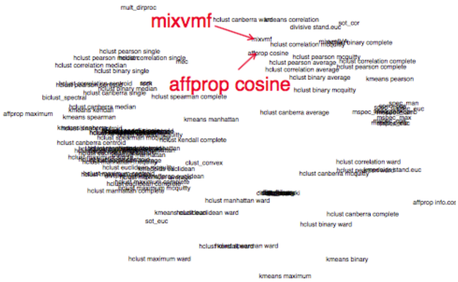
Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)
- Apply our method (relying on many clustering algorithms)

Example CAC discovery



Each point is a clustering

Affinity Propagation-Cosine (Dueck and Frey 2007)

Close to:

Mixture of von Mises-Fisher distributions (Banerjee et. al. 2005)

⇒ Similar clustering of documents

Example CAC discovery

Found a region with clusterings that all reveal the same important insight

Example CAC discovery

Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering Random Walk (Metrics 1-6)

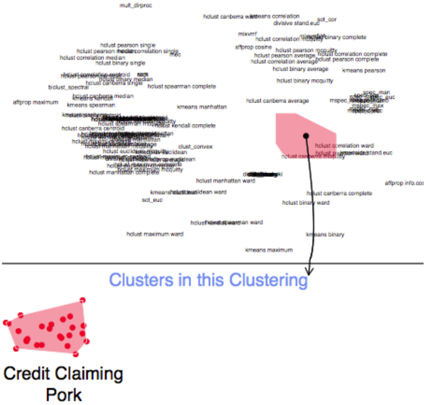
0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

0.05 Kmediods-Cosine

0.04 Spectral clustering Symmetric (Metrics 1-6)

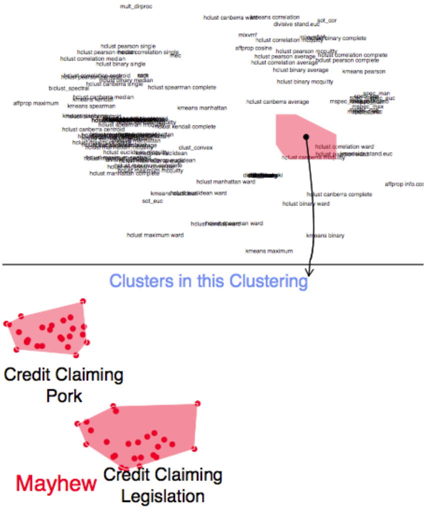
Example CAC discovery



Credit Claiming, Pork:
"Sens. Frank R. Lautenberg (D-NJ) and Robert Menendez (D-NJ) announced that the U.S. Department of Commerce has awarded a $100,000 grant to the South Jersey Economic Development District"

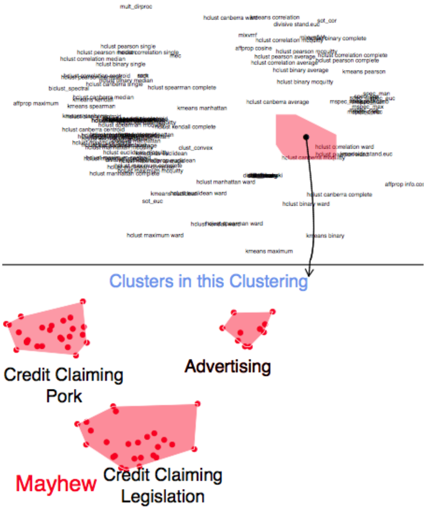
Mayhew

Example CAC discovery



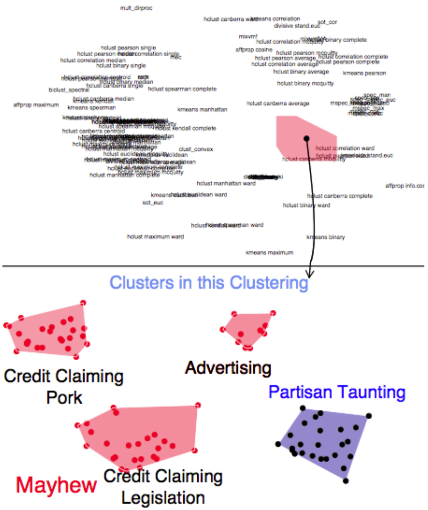
Credit Claiming, Legislation:
"As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period"

Example CAC discovery



Advertising:
"Senate Adopts Lautenberg/Menendez Resolution Honoring Spelling Bee Champion from New Jersey"

Example CAC discovery



Partisan Taunting:
"Republicans Selling Out Nation on Chemical Plant Security"

Topic Models

Goal

> *Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.*
>
> Blei, 2012

Note that in social science we often use the outputs from topic models as a measurement strategy:

"who pays more attention to education policy, conservatives or liberals?"

Recall: Clustering

Document 1

Document 2

Document 3

$\vdots$

Document N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Recall: Clustering

Document 1

Document 2 $\longrightarrow$ Cluster 1

Document 3

Cluster 2

$\vdots$

$\vdots$

Cluster K

Document N

Recall: Clustering

Document 1

Document 2

Document 3

$\vdots$

Document N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Recall: Clustering

Document 1

Document 2

Document 3

$\vdots$

Document N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic Modeling

Document 1

Document 2

Document 3

$\vdots$

Document N

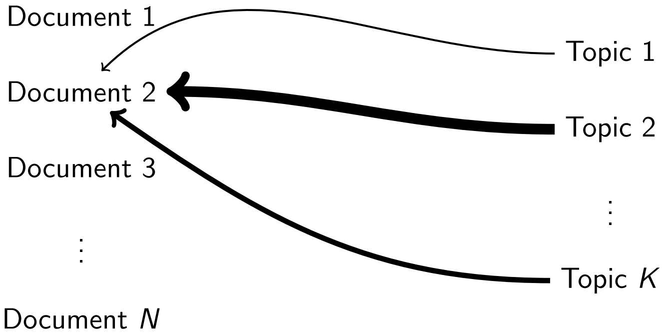
Topic 1

Topic 2

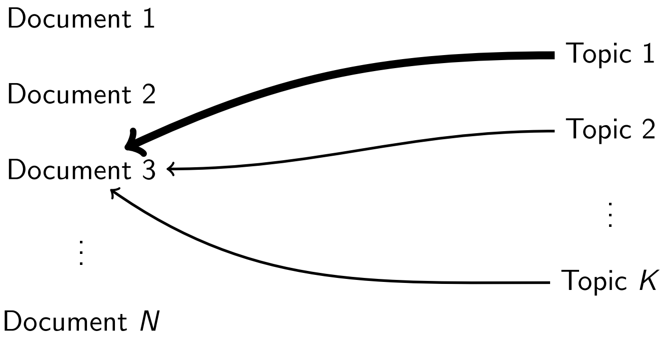
$\vdots$

Topic K

Topic Modeling



Topic Modeling



"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{X}, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

   Where:
   - $\boldsymbol{\pi} = N \times K$ matrix with row $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\Theta} = K \times J$ matrix, with row $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\pi}$.
3) Optimization
   - Variational Approximation$\rightsquigarrow$ EM Algorithm where every step is an "E"
   - Collapsed Gibbs Sampling$\rightsquigarrow$ MCMC algorithm
   - Many other variants
4) Validation$\rightsquigarrow$ many of the same methods from clustering

Binomial and Multinomial

Binomial distribution: the number of successes in a sequence of independent *yes/no* experiments (Bernoulli trials).

$$P(X = x \mid n, p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

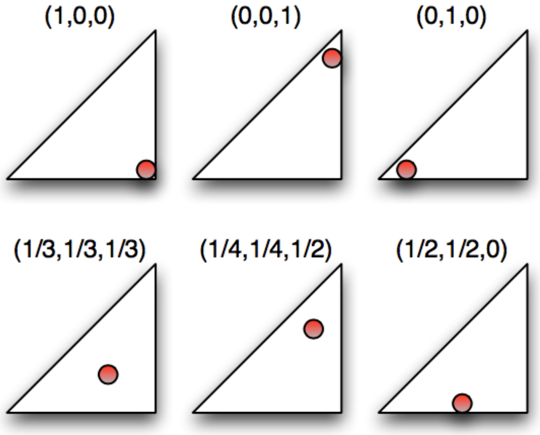
Multinomial: suppose that each experiment results in one of k *possible outcomes* with probabilities $p_1, \ldots, p_k$; Multinomial models the distribution of the histogram vector which indicates how many time each outcome was observed over N trials of experiments.

$$P(x_1, \ldots, x_k \mid n, p_1, \ldots, p_k) = \frac{N!}{\prod_{i=1}^k x_i!} p_i^{x_i}, \; \sum_i x_i = N, x_i \geq 0$$

the Multinomial

- distribution over discrete outcomes;
- represented by non-negative vector that sums to one;
- now imagine a distribution over multinomial distributions: that's a Dirichlet distribution. What does the distribution look like?
- breaking sticks analogy: draw prob parameters from Beta on breaking sticks, conditional on the previous one

the Dirichlet



- Simplex triangle plot: there is a density distribution superimposed on the triangle (probability SIMPLEX).
- If $\alpha = (1, 1, 1)$ then we have the Uniform distribution.

Beta distribution

$$p(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- $p \in [0, 1]$: considering p as the parameter of a Binomial distribution, we can think of Beta is a "distribution over distributions" (binomials).
- Beta function simply defines binomial coefficient for continuous variables. (likewise, Gamma function defines factorial in continuous domain.)

$$B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \simeq \begin{pmatrix} \alpha - 1 \\ \alpha + \beta - 2 \end{pmatrix}$$

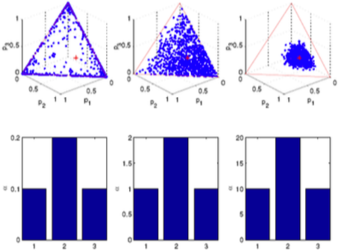
- Beta is the conjugate prior of Binomial.

Dirichlet I - Multivariate generalization of β distribution

$$p(P = \{p_i\} \mid \alpha_i) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \prod_i p_i^{\alpha_i - 1}$$

- $\sum_i p_i = 1, p_i \geq 0$
- Two parameters: the scale (or concentration) $\sigma = \sum_i \alpha_i$, and the base measure $(\alpha'_1, \ldots, \alpha'_k), \alpha'_i = \alpha_i / \sigma$.
- A generalization of Beta:
  - Beta is a distribution over binomials (in an interval $p \in [0, 1]$);
  - Dirichlet is a distribution over Multinomials (in the so-called simplex $\sum_i p_i = 1$; $p_i \geq 0$).
- Dirichlet is the conjugate prior of multinomial.

Dirichlet I - Multivariate generalization of β distribution



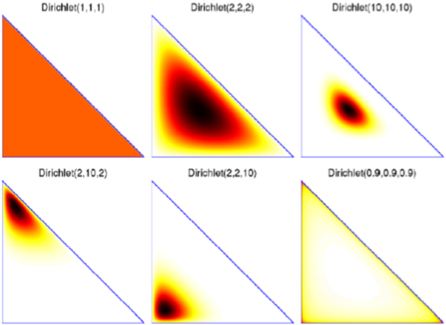
- The base measure determines the mean distribution;
- Altering the scale affects the variance.

$$E(p_i) = \frac{\alpha_i}{\sigma} = \alpha_i' \tag{1}$$

$$Var(p_i) = \frac{\alpha_i(\sigma - \alpha)}{\sigma^2(\sigma + 1)} = \frac{\alpha_i'(1 - \alpha_i')}{(\sigma + 1)} \tag{2}$$

$$Cov(p_i, p_j) = \frac{-\alpha_i \alpha_j}{\sigma^2(\sigma + 1)} \tag{3}$$

Dirichlet I - Multivariate generalization of β distribution



- A Dirichlet with small concentration σ favors extreme distributions, but this prior belief is very weak and is easily overwritten by data.
- As $\sigma \to \infty$, the covariance $\to 0$ and the samples $\to$ base measure.

Dirichlet I - Multivariate generalization of β distribution

Suppose that we are interested in a simple generative model (monogram) for English words. If asked "what is the next word in a newly-discovered work of Shakespeare?", our model must surely assign non-zero probability for words that Shakespeare never used before. Our model should also satisfy a consistency rule called exchangeability: the probability of finding a particular word at a given location in the stream of text should be the same everywhere in thee stream.

α concentration parameter

- simplest and most common Dirichlet prior is the symmetric Dirichlet distribution, where all parameters are equal (no prior information favoring one component/word over any other;
- intuitively the concentration parameter can be thought of as determining how "concentrated" the probability mass of a sample of Dirichlet distributions is likely to be;
- values above 1 prefer variates that are dense, evenly distributed distributions, i.e. all the values within a single sample are similar to each other.
- values below 1 prefer sparse distributions, i.e. most of the values within a single sample will be close to 0, and the vast majority of the mass will be concentrated in a few of the values.

α concentration parameter

- consider a topic model, which is used to learn the topics that are discussed in a set of documents, where each "topic" is described using a categorical distribution over a vocabulary of words.
- A typical vocabulary might have 100,000 words, leading to a 100,000-dimensional categorical distribution.
- the prior distribution for the parameters of the categorical distribution would likely be a symmetric Dirichlet distribution
- However, a coherent topic might only have a few hundred words with any significant probability mass
- a reasonable setting for the concentration parameter might be 0.01 or 0.001. (standard packages set $\alpha = \frac{1}{50}$)

Unigram Model of Language

Suppose we have several speakers (authors/clusters/topics/categories/ ...)
Speaker i produces document $\boldsymbol{x}_i$,

$$\boldsymbol{X}_i \quad \sim \quad \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i \rightsquigarrow$ Speaker specific word rates
Build hierarchical model:

$$\boldsymbol{\theta}_i \quad \sim \quad \text{Distribution on Simplex}$$

Hierarchical Models as a Modeling Paradigm

Why Build a Hierarchical Model?

1) Borrow strength across documents$\rightsquigarrow$ Improved and granular inferences
2) Shrink estimates$\rightsquigarrow$ regularization
3) Incorporate further covariate information
   i) Author
   ii) Time
   iii) ...
3) Learn additional structure
   i) Hierarchies of word rates
   ii) Clusters of similar word rates
   iii) Low dimensional approximations of word rates
4) Encodes complicated dependencies between documents/speakers

Dirichlet-Multinomial Unigram Language Model

For N observations we observe a 3-element long count vector

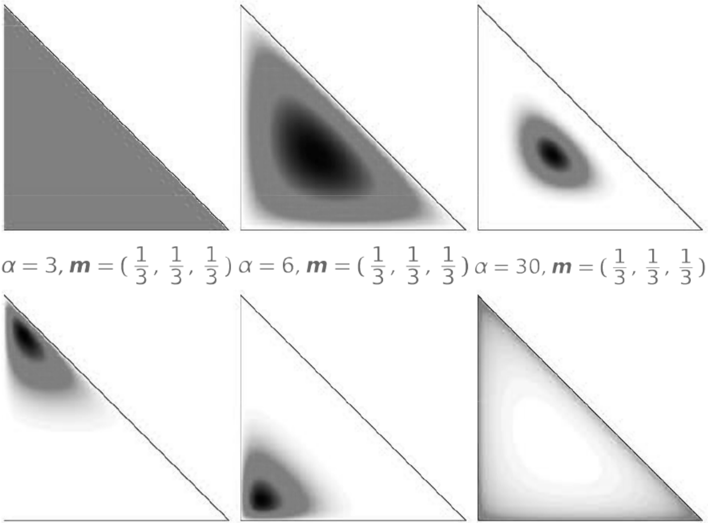
$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$$

Where $N_i = \sum_{j=1}^{3} x_{ij}$.
Suppose

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

- Dirichlet distribution$\rightsquigarrow$ assumption about population of word rates
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ describes population use of words and variation
- Just one distribution simplex

α parameterisation

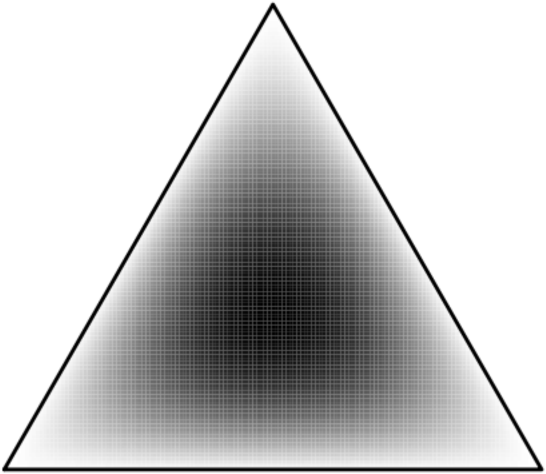


$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 30, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

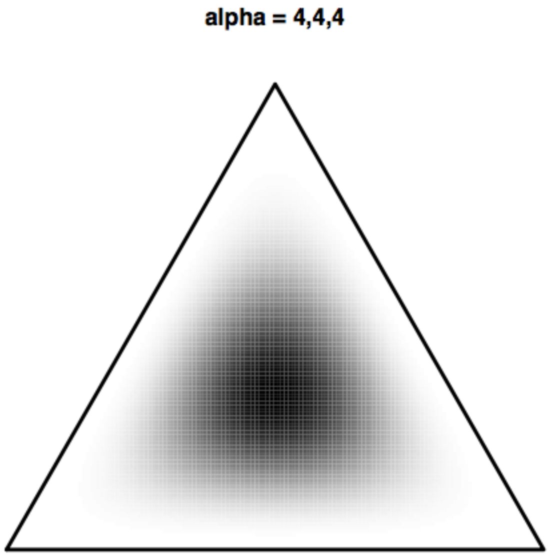
$\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7})$ $\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7})$ $\alpha = 2.7, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

α parameterisation

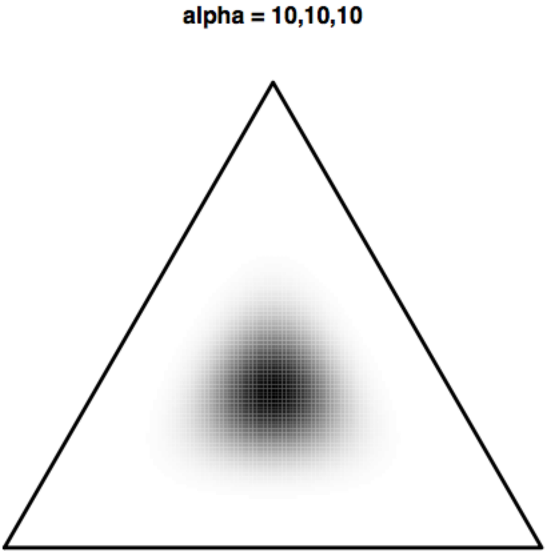
alpha = 2,2,2



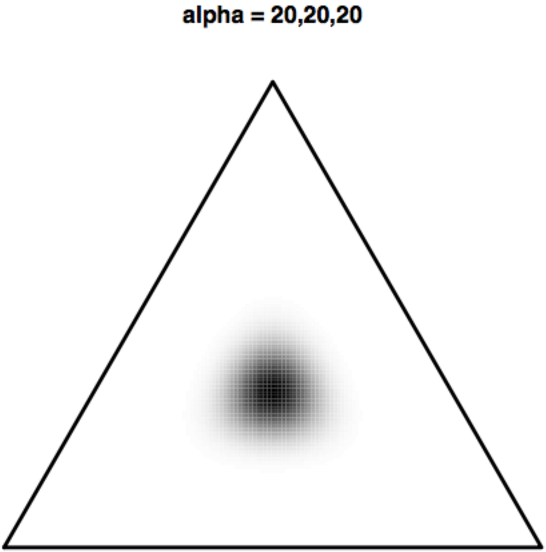
α parameterisation



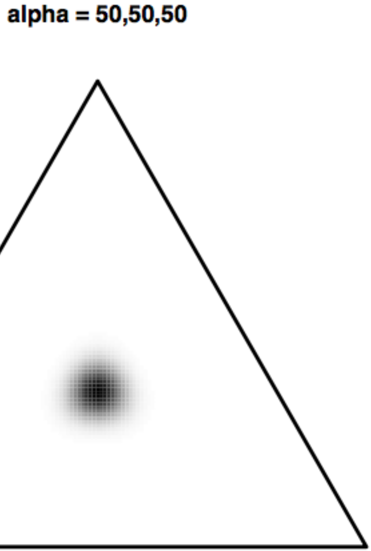
α parameterisation



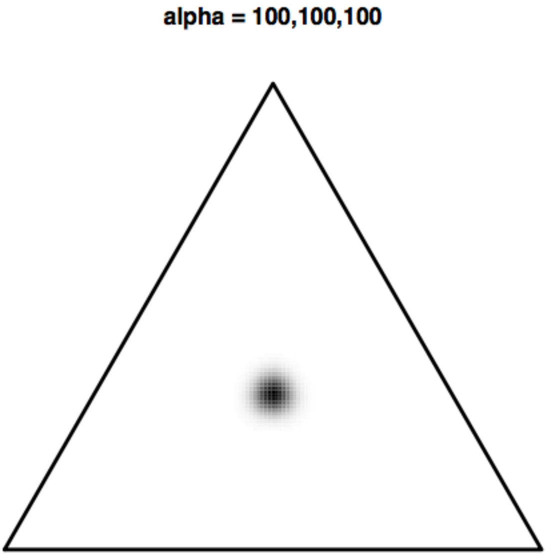
α parameterisation



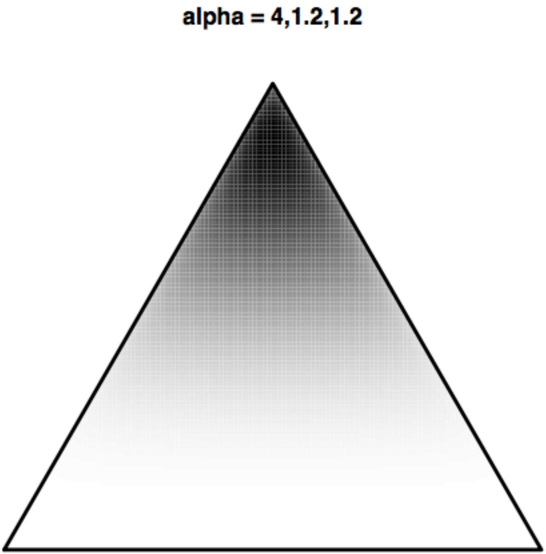
α parameterisation



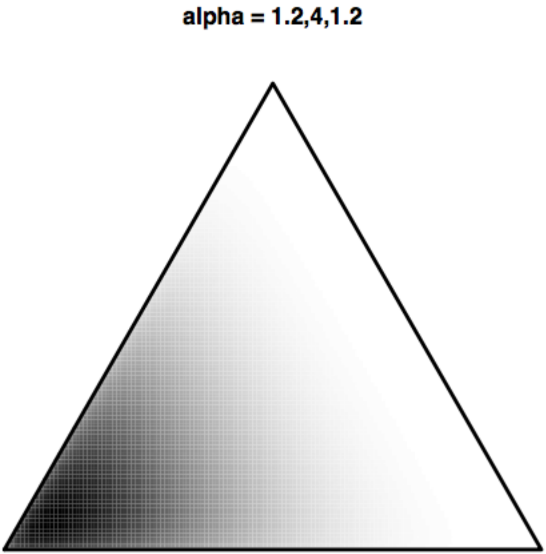
α parameterisation



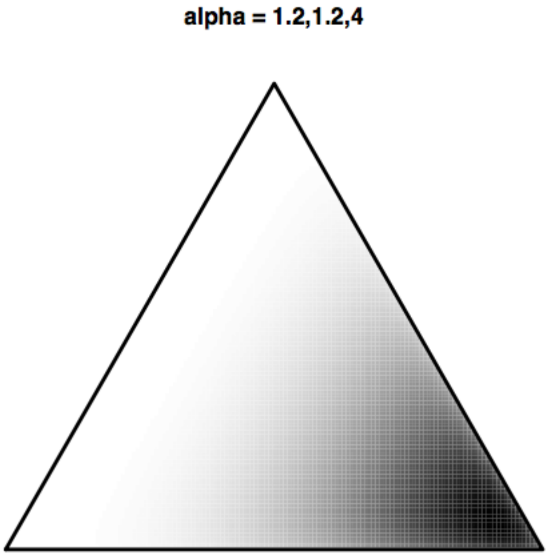
α parameterisation



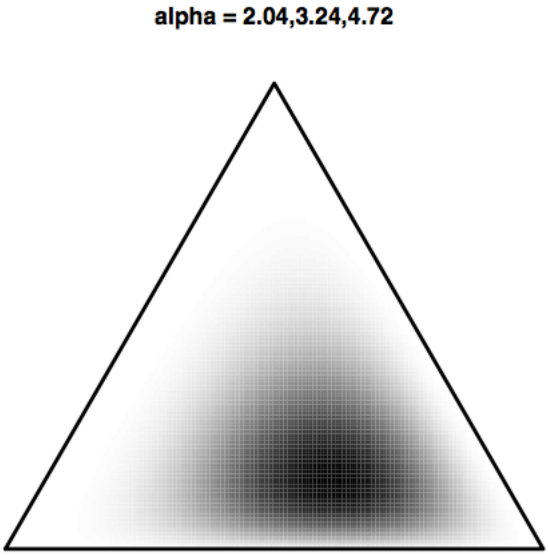
α parameterisation



α parameterisation



α parameterisation



- Important Facts

$$
\begin{aligned}
E[\boldsymbol{\theta}_i] &= \left(\frac{\alpha_1}{\sum_{j=1}^{3} \alpha_j}, \frac{\alpha_2}{\sum_{j=1}^{3} \alpha_j}, \frac{\alpha_3}{\sum_{j=1}^{3} \alpha_j} \right) \\
\text{var}(\theta_{ij}) &= \frac{\alpha_i \left(\sum_{j=1}^{3} \alpha_j - \alpha_i \right)}{\left(\sum_{j=1}^{3} \alpha_j \right)^2 \left(\sum_{j=1}^{3} \alpha_j + 1 \right)} \\
\text{cov}(\theta_{ik}, \theta_{ij}) &= \frac{-\alpha_k \alpha_j}{\left(\sum_{j=1}^{3} \alpha_j \right)^2 \left(\sum_{j=1}^{3} \alpha_j + 1 \right)} \\
\text{Mode}(\theta_j) &= \frac{\alpha_j - 1}{\sum_{k=1}^{3} \alpha_k - 3}
\end{aligned}
$$

Unigram Model of Language

Assume we have a 3 word vocabulary

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.
Bag of Words $\rightsquigarrow$ each word is an independent draw over 3 words

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.
Bag of Words $\rightsquigarrow$ each word is an independent draw over 3 words

- Improbable model of language creation

Unigram Model of Language

Assume we have a 3 word vocabulary$\rightsquigarrow$ 3 words that we might speak.
Bag of Words$\rightsquigarrow$ each word is an independent draw over 3 words

- Improbable model of language creation
- Complex dependency structure of text

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.

Bag of Words $\rightsquigarrow$ each word is an independent draw over 3 words

- Improbable model of language creation
- Complex dependency structure of text
- Improbable $\neq$ useless

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\boldsymbol{X}_i = (1,0,0)) \quad = \quad \theta_1$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$\begin{aligned} p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\ p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \end{aligned}$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\
p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1
\end{aligned}
$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\
p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1
\end{aligned}
$$

The pmf for $\boldsymbol{X}_i$ is,

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\
p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1
\end{aligned}
$$

The pmf for $\boldsymbol{X}_i$ is,

$$
p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \prod_{j=1}^{3} \theta_j^{x_{ij}}
$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\
p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1
\end{aligned}
$$

The pmf for $\boldsymbol{X}_i$ is,

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta})
\end{aligned}
$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\
p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1
\end{aligned}
$$

The pmf for $\boldsymbol{X}_i$ is,

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\
\boldsymbol{X}_i &\sim \text{Categorical}(\boldsymbol{\theta})
\end{aligned}
$$

Unigram Model of Language

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta})
\end{aligned}
$$

Unigram Model of Language

$$
\begin{aligned}
p(\boldsymbol{x}_i | \boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\
E[x_{ij}] &= \theta_j
\end{aligned}
$$

Unigram Model of Language

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\
E[x_{ij}] &= \theta_j \\
\text{Var}(X_{ij}) &= \theta_j(1 - \theta_j)
\end{aligned}
$$

Unigram Model of Language

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\
E[x_{ij}] &= \theta_j \\
\text{Var}(X_{ij}) &= \theta_j(1-\theta_j) \\
\text{Cov}(X_{ij}, x_{ik}) &= -\theta_j\theta_k
\end{aligned}
$$

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \quad \propto \quad \prod_{j=1}^{3} \theta_j^{x_j}$$

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates$\rightsquigarrow$ our summary of the document/speaker

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates $\rightsquigarrow$ our summary of the document/speaker

- $\sum_{j=1}^{3} \theta_j = 1$

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \;\; \propto \;\; \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates$\rightsquigarrow$ our summary of the document/speaker

- $\sum_{j=1}^{3} \theta_j = 1$
- $\theta_j \geq 0$

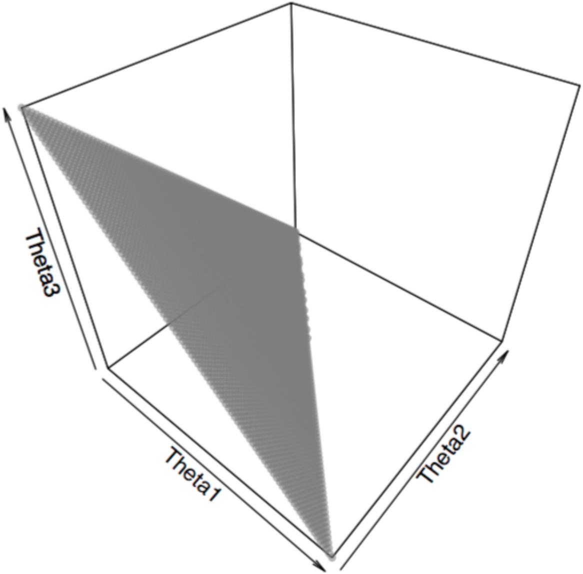
Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \;\propto\; \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates$\rightsquigarrow$ our summary of the document/speaker

- $\sum_{j=1}^{3} \theta_j = 1$
- $\theta_j \geq 0$

$\boldsymbol{\theta} \in \Delta^2$ (2-dimensional simplex)


Theta3
Theta1
Theta2

Dirichlet-Multinomial Unigram Model of Language

$$\begin{aligned}\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)\end{aligned}$$

let's say we want to make inferences about the word rates; multiply dirichlet distribution component with multinomial distribution component. Dirichlet kernel (signature component of the probability distribution that gives a realization/the value of the random variable) of gives us the new parameters (α and x_ij.

Dirichlet-Multinomial Unigram Model of Language

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

let's say we want to make inferences about the word rates; multiply dirichlet distribution component with multinomial distribution component. Dirichlet kernel (signature component of the probability distribution that gives a realization/the value of the random variable) of gives us the new parameters (α and $x_i j$.

$$p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \boldsymbol{x}_i) \quad \propto \quad p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\ p(\boldsymbol{x}_i|\boldsymbol{\theta}_i)$$

Dirichlet-Multinomial Unigram Model of Language

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

let's say we want to make inferences about the word rates; multiply dirichlet distribution component with multinomial distribution component. Dirichlet kernel (signature component of the probability distribution that gives a realization/the value of the random variable) of gives us the new parameters (α and $x_i j$.

$$
\begin{aligned}
p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \boldsymbol{x}_i) &\propto p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\ p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) \\
&\propto \frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} \prod_{j=1}^{3} \theta_j^{\alpha_j - 1} \prod_{j=1}^{3} \theta_{ij}^{x_{ij}}
\end{aligned}
$$

Dirichlet-Multinomial Unigram Model of Language

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

let's say we want to make inferences about the word rates; multiply dirichlet distribution component with multinomial distribution component. Dirichlet kernel (signature component of the probability distribution that gives a realization/the value of the random variable) of gives us the new parameters (α and x_ij.

$$\begin{aligned} p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \boldsymbol{x}_i) &\propto p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\ p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) \\ &\propto \frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\prod_{j=1}^{3}\theta_j^{\alpha_j-1}\prod_{j=1}^{3}\theta_{ij}^{x_{ij}} \\ &\propto \frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\underbrace{\prod_{j=1}^{3}\theta_j^{\alpha_j+x_{ij}-1}}_{\text{Dirichlet Kernel}} \end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

the posterior distribution of theta is Dirichlet has parameters alpha (things we assume before hand) and x (data we observe);

$$\begin{aligned} \boldsymbol{\theta}_i | \boldsymbol{\alpha}, \boldsymbol{x}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{x}) \\ \text{E}[\theta_{ij} | \boldsymbol{\alpha}, \boldsymbol{x}_i] &= \frac{\alpha_j + x_{ij}}{\sum_{j=1}^{3}(x_{ij} + \alpha_j)} \end{aligned}$$

- $\alpha_j \rightsquigarrow$ "pseudo" data that smooth the estimates toward $\frac{\alpha_j}{\alpha_1+\alpha_2+\alpha_3}$
- as $N_i \to \infty$ data $(\boldsymbol{x}_i)$ overwhelm $\boldsymbol{\alpha}$

Alternative Priors on the Simplex

Dirichlet distribution

- Imposes specific form on variance
- Imposes negative correlation between all components.
- We might expect some word rates to positively covary.

Alternative $\rightsquigarrow$ Logistic-Normal distribution

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

*Notice: this is a different representation than a document-term matrix. x_{im} is a number that says which of the J words are used. The difference is for clarity and we'll this representation is closely related to document-term matrix

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$\boldsymbol{\pi}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

- $\boldsymbol{\pi}_i$ in LDA, is an N (documents) x K (topics) matrix representing the proportion of a document i in each topic.
- in short, the extent to which document i attention to topics differs from all documents in the population, as governed by a Dirichlet distribution;

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i)
\end{aligned}
$$

$\boldsymbol{\tau}_{im}$: conditional on document-specific attention to documents, for each word we will draw the word's topic from a multinomial distribution with the rate at which a topic occurs given by $\boldsymbol{\pi}_k$, the document's attention to the topics

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

x_{im}: conditional on each word's topic in the unigram model for that specific topic, we will draw the mth word in our data.

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$\boldsymbol{\theta}_k \sim \text{Dirichlet}(\mathbf{1})$$

$$\begin{aligned} \boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\ x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k) \end{aligned}$$

$\boldsymbol{\theta}_k$: K x V word probability matrix for each topic, aka our unigram model for each topic: a PMF giving prob of obtaining word from that document; if some components of $\boldsymbol{\theta}_k$ are big, it means they occur more frequently and that they are indicative of respective topic; think about this in terms of triangle simplex: we draw words rates from a particular area of the triangle (that with the highest density)

Back to Vanilla LDA $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

$\boldsymbol{\alpha}$ prior: comes from a gamma distribution, $(\alpha_1, \alpha_1, \alpha_1)$ describes population use of words and variation;

LDA Summary

$$\begin{aligned}
\textbf{Unigram Model}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\textbf{Doc. Prop}_i &\sim \text{Dirichlet}(\textbf{Pop. Proportion}) \\
\textbf{Word Topic}_{im} &\sim \text{Multinomial}(1, \textbf{Doc. Prop}_i) \\
\text{Word}_{im} &\sim \text{Multinomial}(1, \textbf{Unigram Model}_k)
\end{aligned}$$

Aside: Dirichlet distribution

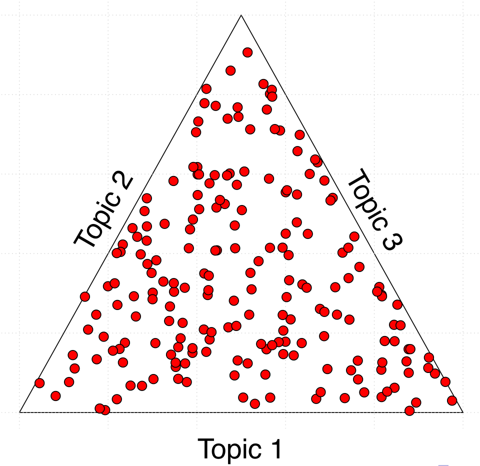
The Dirichlet distribution is a conjugate prior for the multinomial ('categorical' if you only have one trial) distribution. Makes certain calculations easier.

It is parameterized by a vector of positive real numbers $\boldsymbol{\alpha}$. In principle, one can have $\alpha_1, \ldots, \alpha_k$ be different concentration parameters, but LDA uses special symmetric Dirichlet where all the values of $\boldsymbol{\alpha}$ are the same.

Larger values of α (assuming we are in symmetric case) mean we think (*a priori*) that documents are generally an even mix of the topics. If α is small (less than 1) we think a given document is generally from one or a few topics.

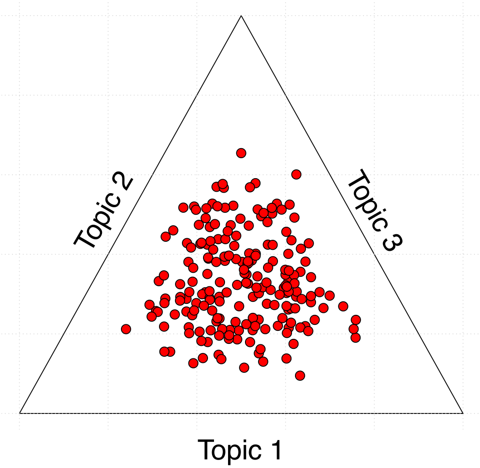
Example of Dirichlet

200 documents, 3 topics, $\alpha = 1$
(uniform)



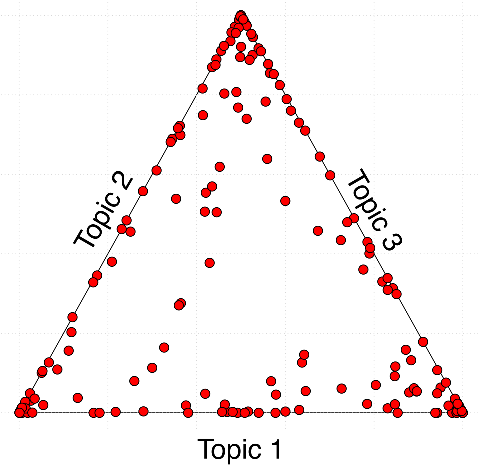
Example of Dirichlet

200 documents, 3 topics, $\alpha = 5$



Example of Dirichlet

200 documents, 3 topics, $\alpha = 0.2$



And actually. . .

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words. A low concentration parameter means each topic is a mixture of a few of the words.

In practice, one can estimate the concentration parameters, or simple set them at suggested values.

We want topic models to be similar as we increase number of topics. Can use asymmetric priors for per-document topic distributions (the θs). Asymmetric priors on per-topic word distributions don't do much. Wallach et al "Rethinking LDA: Why Priors Matter"

A General Hierarchical Structure

LDA:

Pop. Proportion

A General Hierarchical Structure

LDA:

Pop. Proportion

Doc. Prop$_1$ **Doc. Prop**$_2$ $\cdots$ **Doc. Prop**$_N$

A General Hierarchical Structure

LDA:

Pop. Proportion

Doc. Prop$_1$ ↓ **Word Topics**$_1$

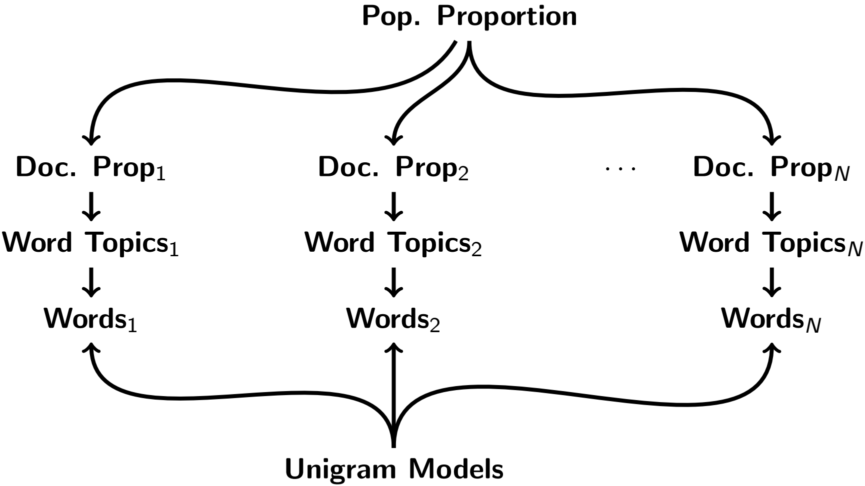
Doc. Prop$_2$ ↓ **Word Topics**$_2$

$\cdots$

Doc. Prop$_N$ ↓ **Word Topics**$_N$

A General Hierarchical Structure

LDA:



LDA as a generative model

- Each topic is a multinomial distribution over words; each topic's multinomial distribution over words will be drawn from a Dirichlet distribution;
- Each document is a multinomial distribution over topics; each document's multinomial distribution over topics will be drawn from a Dirichlet distribution; For every document, we have a Dirichlet distribution over all the topics it could use and then it selects what topics it will talk about in the document

Generative model for LDA

TOPIC 1

computer, technology, system, service, site, phone, internet, machine

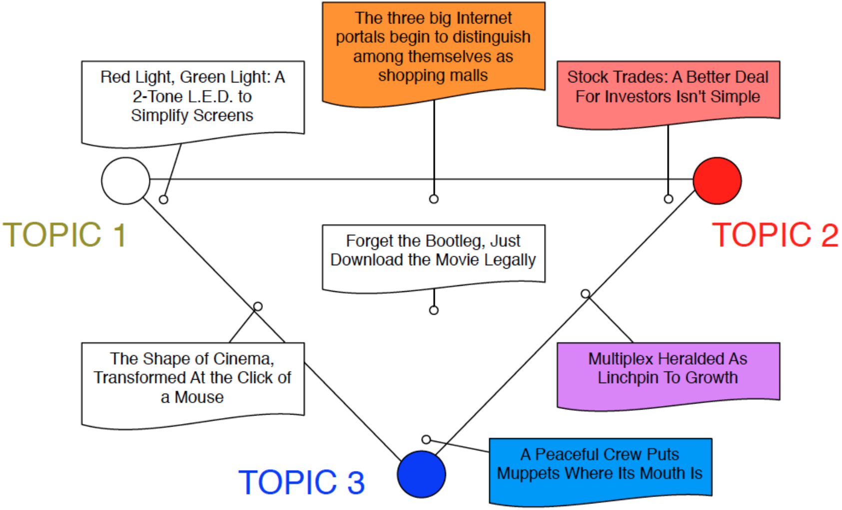
TOPIC 2

sell, sale, store, product, business, advertising, market, consumer

TOPIC 3

play, film, movie, theater, production, star, director, stage

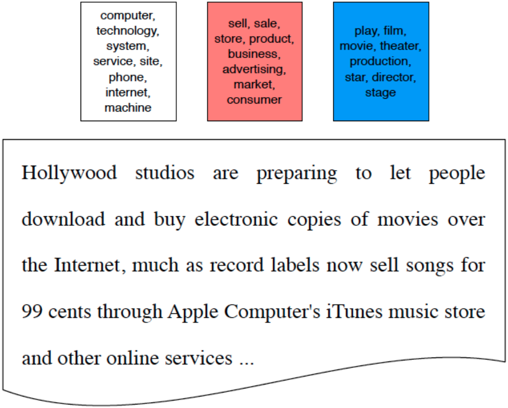
Generative model for LDA



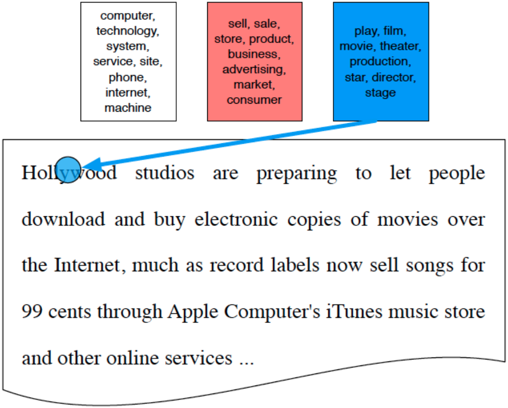
Generative model for LDA

- Once we're in a document, we need to select the words we will use;
- Each word will select a topic it will use which comes from the multinomial distribution governing the language model;
- If the first word chooses the entertainment topic, we go into that topic, which is itself a multinomial distribution, and we select which word to use.

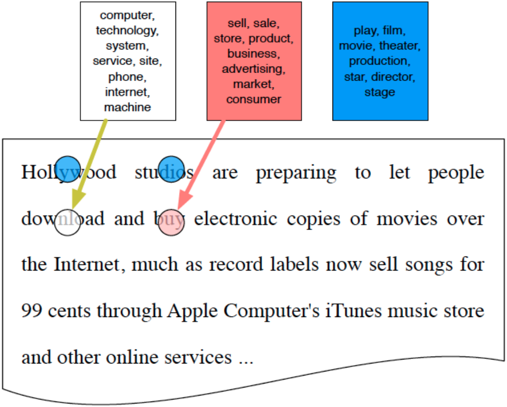
Generative model for LDA



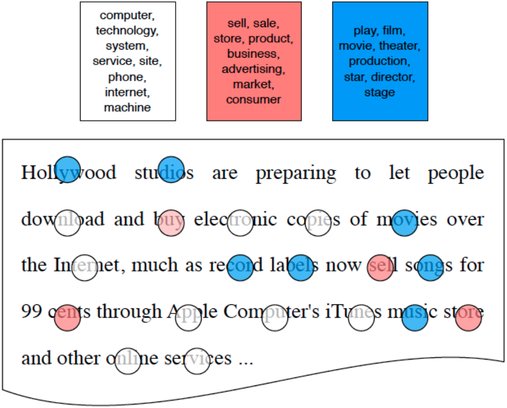
Generative model for LDA



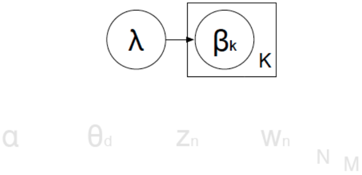
Generative model for LDA



Generative model for LDA

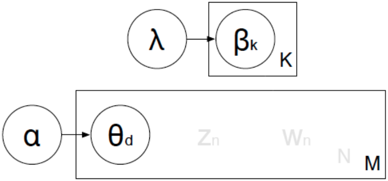


Generative model for LDA



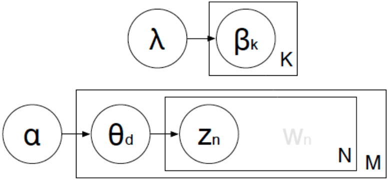
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Generative model for LDA



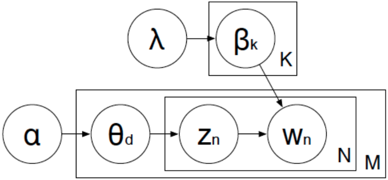
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Generative model for LDA



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

Generative model for LDA



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

Drawbacks

- topic instability, K and Multi-Modality: the way the LDA algorithm follows the gradient function, so it's on the surface and is trying to maximise it based on where it was before. This leads to only finding the local maximum; This means that the topic we find in one run may not exist in another!
- Also, since there are several local maxima, we do not even know what's the best one;
- Roberts, Stewart and Tingley (2016) offer a framework for choosing between local maxima: semantic coherence & exclusivity.

Running a Topic Model with Mallet

`to the Mallet website!!`

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\mathbf{Doc}_i^{'}\mathbf{Doc}_l$

Inner product of Terms (columns): $\mathbf{Word}_j^{'}\mathbf{Word}_k$

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\dots$	Word_J
Doc_1	0	1	$\dots$	0
Doc_2	2	0	$\dots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\dots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Allows: measure of correlation of term usage across documents

(heuristically: partition words, based on usage in documents)

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$
Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)
Latent Semantic Analysis: Reduce information in matrix using singular value decomposition (provides similar results, difficult to generalize - not probabilistic)

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\mathbf{Doc}_i^{'}\mathbf{Doc}_l$

Inner product of Terms (columns): $\mathbf{Word}_j^{'}\mathbf{Word}_k$
Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)
Latent Semantic Analysis: Reduce information in matrix using singular value decomposition (provides similar results, difficult to generalize - not probabilistic)
Biclustering: Models that partition documents and words simultaneously

Why does this work$\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha|\boldsymbol{X}) \quad \propto \quad p(\alpha)p(\boldsymbol{\pi}|\alpha)p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})$$

Why does this work$\leadsto$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{T}|\boldsymbol{\pi}) \underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \leadsto$ Greater weight on terms that occur together

Why does this work$\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) \underbrace{p(\boldsymbol{T} | \boldsymbol{\pi})}_{2} \underbrace{p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together
2) $\boldsymbol{\pi} \rightsquigarrow$ Greater weight on indicators that appear more regularly

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) \underbrace{p(\boldsymbol{\pi}|\boldsymbol{\alpha})}_{3} \underbrace{p(\boldsymbol{T}|\boldsymbol{\pi})}_{2} \underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together
2) $\boldsymbol{\pi} \rightsquigarrow$ Greater weight on indicators that appear more regularly
3) $\boldsymbol{\alpha} \rightsquigarrow$ Emphasis on $\boldsymbol{\pi}$ with greater weight

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\alpha) p(\boldsymbol{\pi} | \alpha) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \theta, \boldsymbol{T})$$

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})\underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

- implies that making $\boldsymbol{\theta}$ a sparse matrix will increase the probability of certain words – remember that the $\boldsymbol{\theta}$ values for a given topic must sum to one, so the more terms we assign a non-zero $\boldsymbol{\theta}$ value the thinner we have to spread our probability for the topic;
- implies that having sparsely distributed topics can result in a high probability for a document, where the ideal way to form the sparse components is to make them non-overlapping clusters of co-occurring words in different documents
- wants to form sparse, segregated word cluster

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \underbrace{p(\boldsymbol{T}|\boldsymbol{\pi})}_{2} \underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

- implies that making $\boldsymbol{\pi}$ have concentrated components will increase the probability
- encourages a sparse $\boldsymbol{\pi}$ matrix so that the probability of choosing a given $\boldsymbol{T}$ value will be large, e.g. $\boldsymbol{\pi} = (0.25,0.25,0.25,0.25)$ would yield smaller probabilities than $\boldsymbol{\pi} = (0.5,0.5,0,0)$
- penalizes documents for having too many possible topics

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) \underbrace{p(\boldsymbol{\pi}|\boldsymbol{\alpha})}_{3} \underbrace{p(\boldsymbol{T}|\boldsymbol{\pi})}_{2} \underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

- implies that using a small $\boldsymbol{\alpha}$ will increase the probability
- also penalizes using a large number of possible topics for a given document – small $\boldsymbol{\alpha}$ values yield sparse $\boldsymbol{\pi}$s.

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

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

- But if we only have a few topics to choose from and each topic has a small number of non-zero word probabilities, then we surely better form meaningful clusters that could represent a diverse number of documents. How should we do this? Form clusters of co-occurring terms, which is largely what LDA accomplishes.

Validation $\rightsquigarrow$ Topic Intrusion

- Labeling paragraphs
  - Identify separating words automatically
  - Label topics manually (read!)
- Statistical methods
  1) Entropy
  2) Exclusivity
  3) Cohesiveness
- Experiment Based Methods
  - Word intrusion $\rightsquigarrow$ topic validity
  - Topic intrusion $\rightsquigarrow$ model fit

Validation $\rightsquigarrow$ Topic Intrusion

1) Ask research assistant to read paragraph
2) Construct experiment
   - For the document, select top three topics
   - Select a fourth topic
   - Show participant, ask her/him to identify intruder

Higher identification $\rightsquigarrow$ topics are a better model of text

Example: Automated Literature Reviews

Recall: literature reviews are hard to conduct

LDA: developed (in part) to help structure `JSTOR` database

Use `JSTOR`'s research service to obtain data to analyze

Question: How do scholars use classic text: Home Style

Analysis: all articles that cite Home Style in JSTOR's data

Example: Automated Literature Reviews

Output: topic estimates

- Obtain $\log \theta_k$ from model
- One method to summarize a topic:
  - $\exp(\log \theta_k)$ (select 10-20 biggest words)
  - $\exp(\log \theta_k) - \text{Average}_{j \neq k} \exp(\log \theta_j)$ (select 10-20 biggest words)

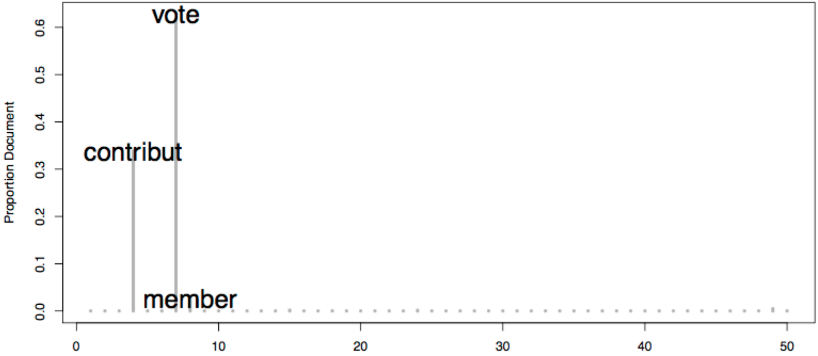
Example: Automated Literature Reviews

Example topics:

Label	Stems	Proportio
Life Style	member,district,attent,congress,time,cohort,retir	0.03
Comp.Home	constitu,mp,member,parti,role,local,british	0.02
Casework	casework,district,constitu,variabl,staff,congression,fiorina	0.03
Votes	vote,variabl,model,estim,measur,legisl,constitu	0.04
Id. Shirk	ideolog,vote,shirk,constitu,parti,senat,voter	0.03
C. letters	mail,govern, activ,respond,commun,offic	0.02

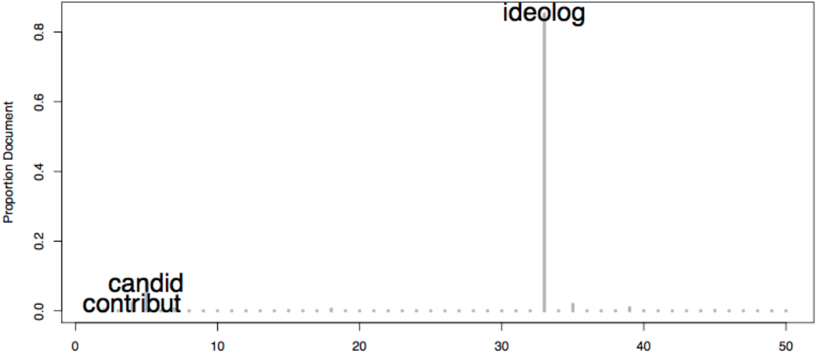
Example Document

Wawro (2001) “A Panel Probit Analysis of Campaign Contributions and Roll Call Votes”



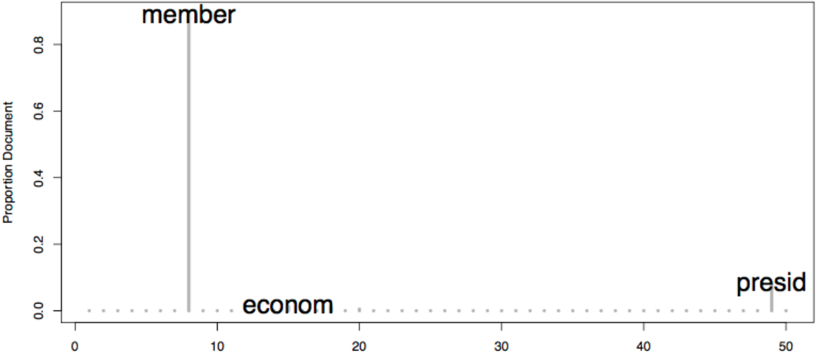
Example Document

Bender (1996) "Legislator Voting and Shirking A Critical Review of the Literature"



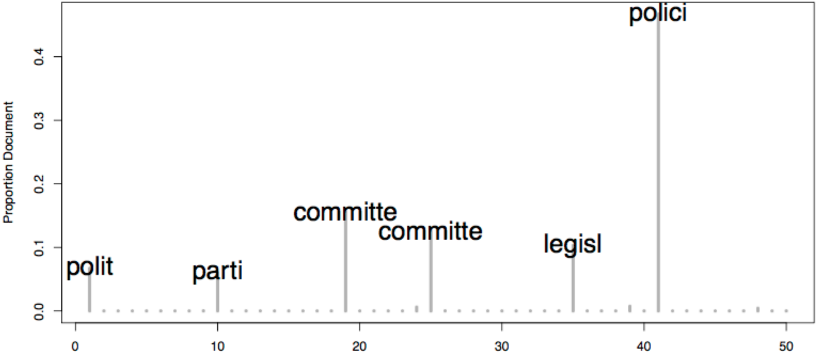
Example Document

Parker (1980) “Cycles in Congressional District Attention”



Example Document

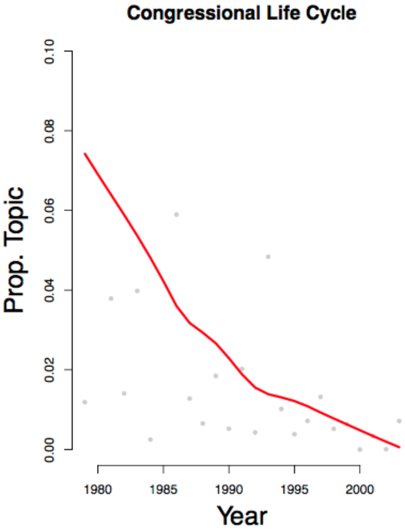
Shepsle (1985) "Policy Consequences of Government by Congressional Subcommittees"



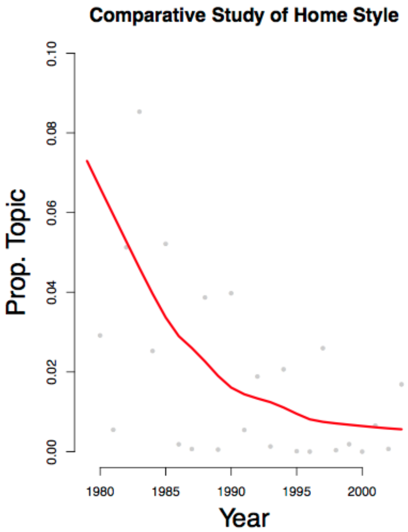
History of Home Style (Fenno 1978)

Fenno (1978) tries to identify the "home styles" that each members of Congress uses to help them secure their first goal (re-election)

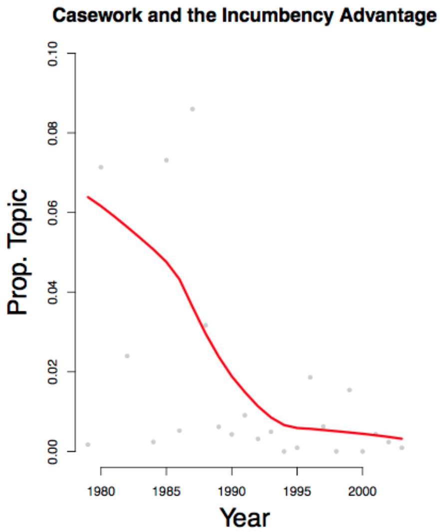
History of Home Style (Fenno 1978)



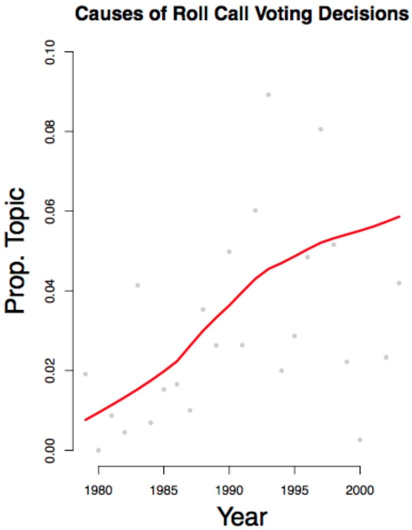
History of Home Style (Fenno 1978)



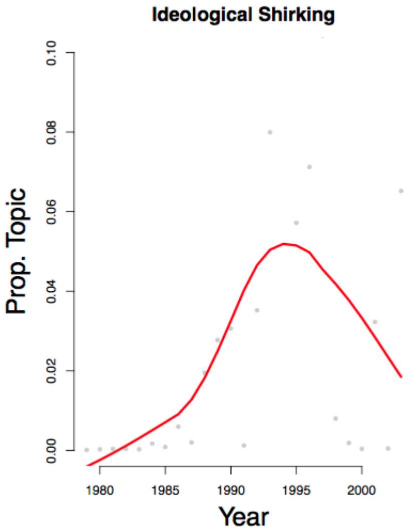
History of Home Style (Fenno 1978)



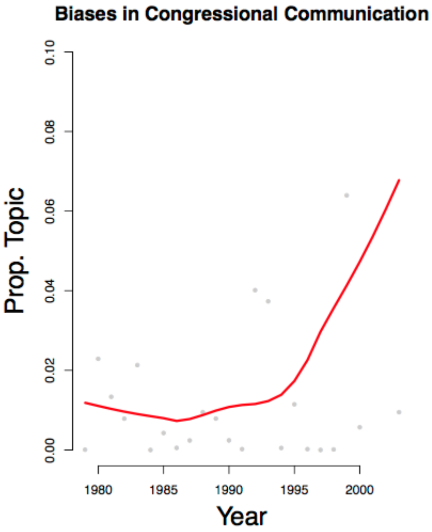
History of Home Style (Fenno 1978)



History of Home Style (Fenno 1978)



History of Home Style (Fenno 1978)



What legislators claim (Grimmer, Westwood, Messing 2014)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08

"Dave Camp announced today that he was able to secure $2.5 million for widening M-72 from US-31 easterly 7.2 miles to Old M-72. The bill will now head to the Senate for consideration...We have two more hurdles to clear to make sure the money is in the bill when it hits the President's desk: a vote in the Senate and a conference committee" (Camp, 2005)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08

"Congressman Doc Hastings has boosted federal funding for work on the Columbia Basin water supply for next year. Hastings has added $400,000 for work on the Odessa Subaquifer, which when combined with the funding in the President's budget request, totals $1 million for Fiscal Year 2009"…"Hastings' funding for the Odessa Subaquifer and Potholes Reservoir was included in the Fiscal Year 2009 Energy and Water Appropriations bill which was approved today by the full House Appropriations Committee. (Hastings, 2008)"

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08

"Maurice Hinchey (D-NY) today announced that the West Endicott Fire Company has been awarded a $17,051 federal grant to purchase approximately 10 sets of protective clothing, as well as radio equipment and air packs for its volunteer firefighters" (Hinchey, 2008)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08

"Congressman Pete Visclosky today announced that the Crown Point Fire Department will receive a $16,550 Department of Homeland Security (DHS) grant to purchase a modular portable video system" (Visclosky, 2008)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08
Stimulus	recovery,funding,jobs,information, act,	0.06

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08
Stimulus	recovery,funding,jobs,information, act,	0.06
Transportation	transportation,project,airport,transit,million	0.06

A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic. Sum down the columns is one.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
conservative	0.00188	0.00088	0.00185	0.00221	0.00168
party	0.00145	0.00067	0.00066	0.00577	0.00093
general	0.00073	0.00033	0.00018	0.00192	0.00040
election	0.00079	0.00053	0.00022	0.00235	0.00076
manifesto	0.00059	0.00078	0.00032	0.00099	0.00048
⋮	⋮	⋮	⋮	⋮	⋮

Continued. . .

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to analyst to label the topics!

Meaningless 'junk' topics not unusual: debate as to whether one has to interpret every topic.

Continued

The topic distribution for each document...

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
doc 1	0.00009	0.00009	0.00009	0.00009	0.99965
doc 2	0.00011	0.00011	0.00011	0.00011	0.99954
doc 3	0.00010	0.00010	0.00010	0.00010	0.99959
doc 4	0.00006	0.00006	0.00006	0.00006	0.99978
doc 5	0.00002	0.00002	0.00002	0.00002	0.99991
doc 6	0.00019	0.00019	0.00019	0.00019	0.99924
⋮	⋮	⋮	⋮	⋮	⋮

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the number of topics, K, is not picked automatically. Analysts select various Ks and check that their results are 'robust'. But see over.

As with all unsupervised learning, interpretation is non-trivial, and requires a lot of validation. Rant: 'just-so' stories abound. Lazy analysts conclude whatever they want.

Practical Notes II: Picking k

Crudely: in social science, researchers fit 'enough' topics until they see what they think they should. E.g. a certain topic—like `finance` suddenly peels off—so stop there.

$\rightarrow$ Check findings are robust in the neighborhood: if best model has $k = 35$, check $k = 30 - 40$ yields similar inferences.

NB: social scientists typically fit far fewer topics than CS, even to same data.

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

where $\mathbf{w}$ are the words in the test set. Higher $\mathcal{L}$ implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

Do this for all k.

In practice. . .

Perplexity is popular option

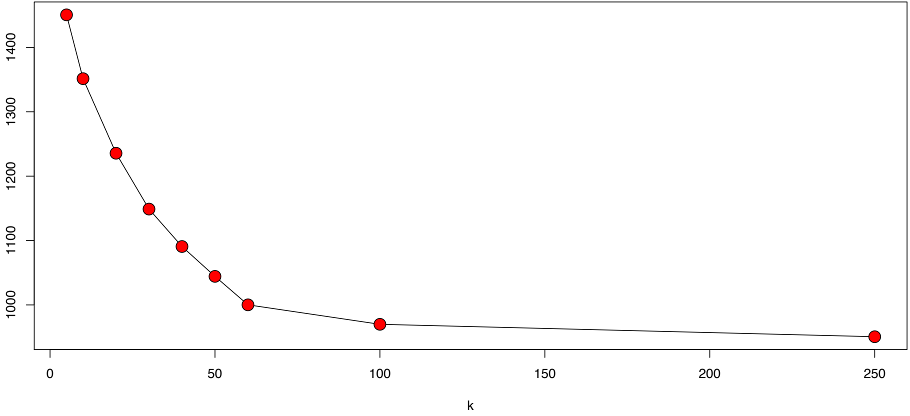
$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In general, $\mathcal{L}(\mathbf{w})$ is intractable, but there are ways to approximate it.

But: the topic models that hold-out calculations suggest are optimal and not much liked by humans! "Reading Tea Leaves: How Humans Interpret Topic Models" by Chang et al.

Perplexity Likes a Lot of Topics (manifestos)



Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

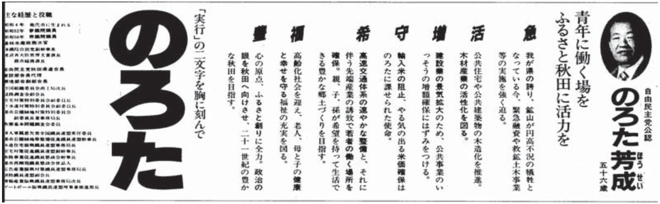
1. Rise of China? Need to focus on security.

vs.

2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of Westminster-style polity.

To decide, we need data source that covers all lower house legislators where they set out their policy priorities over time. See if/when they shift priorities.

Manifestos



自由民主党公認
のろた芳成（ほうせい）
五十六歳

青年に働く場を
ふるさと秋田に活力を

急　我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活　公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増　建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守　輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希　高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福　高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

豊　心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

「実行」の二文字を胸に刻んで

のろた

主な経歴と役職

7,497. 1986–2009. Standardized form.

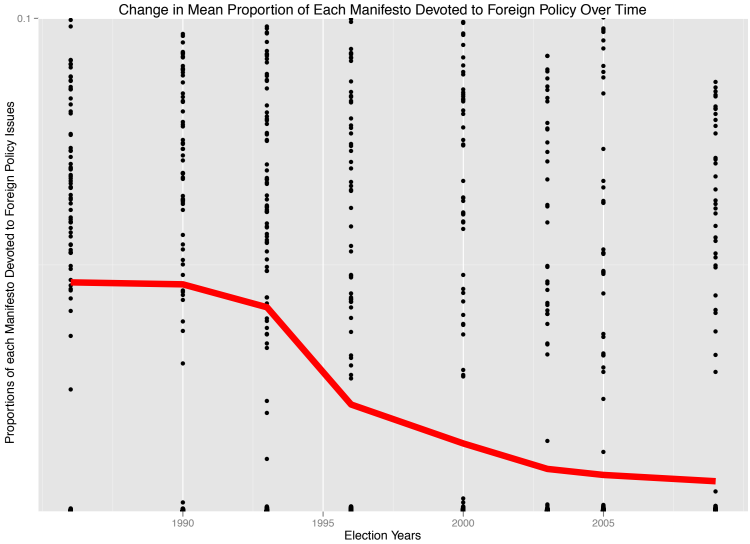
"... instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters"

Manifestos were hand transcribed from microfilm. Japanese install of Windows/R used to fit LDA.

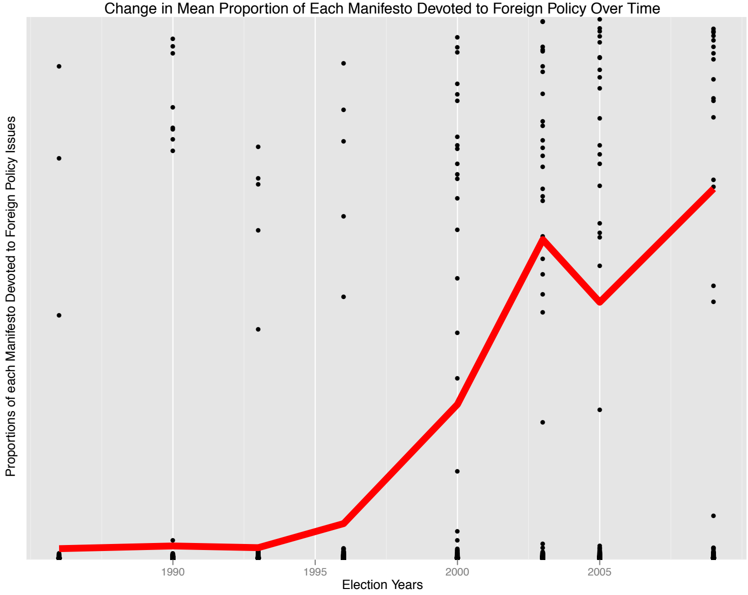
Topic Distribution over Words

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	改革	年金	推進	区	政治	日本
2	郵政	円	整備	政策	改革	国
3	民営	廃止	図る	地域	国民	外交
4	小泉	改革	つとめる	まち	企業	国家
5	構造	兆	社会	鹿児島	自民党	社会
6	政府	実現	対策	全力	日本	国民
7	官	無駄	振興	選挙	共産党	保障
8	推進	日本	充実	国政	献金	安全
9	民	増税	促進	作り	金権	地域
10	自民党	削減	安定	横浜	党	拉致
11	日本	一元化	確立	対策	選挙	経済
12	制度	政権	企業	中小	禁止	守る
13	民間	子供	実現	発電	憲法	問題
14	年金	地域	中小	推進	腐敗	北朝鮮
15	実現	ひと	育成	エネルギー	団体	教育
16	進める	サラリーマン	制度	企業	区	責任
17	断行	制度	政治	声	ソ連	力
18	地方	議員	地域	実現	守る	創る
19	止める	金	福祉	活性	平和	安心
20	保障	民主党	事業	自民党	円	目指す
21	財政	年間	改革	地方	反対	誇り
22	作る	一掃	確保	尽くす	真	憲法
23	賛成	郵政	強化	商店	是正	可能
24	社会	道路	教育	いかす	一掃	道
25	国民	交代	施設	全国	悪政	未来
26	公務員	社会保険庁	生活	政党	抜本	ひと
27	力	月額	支援	ひと	定数	再生
28	経済	手当	環境	支援	政党	将来
29	国	談合	発展	経済	金丸	解決
30	安心	支援	施策	福祉	改革	基本

Change in proportion of ‘Pork’ Topic



Change in proportion of ‘Foreign Policy’ Topic



Correlated Topic Models

it makes sense that knowing the prevalence of one topic in a document tells us something about distribution over the other topics

Dirichlet distribution $\rightsquigarrow$ Assumes negative covariance between topics

Logistic Normal Distribution (not conjugate to multinomial topic mixing)

$\rightsquigarrow$ Allows some positive covariance between topics

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\eta}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\pi}_i &= \frac{\exp(\boldsymbol{\eta}_i)}{\sum_{k=1}^{K} \exp(\eta_{ik})} \\
\boldsymbol{\tau}_{im}|\boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Structural Topic Models

Allows content and prevalence of topics to vary with covariates.

- Content (distribution of words over topics): content can vary with binary variable (Liberal v Conservative); with normal LDA, we would need for example 2 topics (Liberal-Guns and Conservative-Guns), but here we can see it is the same topic but approached differently depending on whether document is Liberal or Conservative;
- Prevalence (distribution of topics over documents): can vary with both categorical and continuous variables (e.g. time).
- Ameliorates the problems of multimodality through spectral initialisation (if they can find some anchor words for each topic and assign that word only to one topic, all of the other terms in matrix of words over topics are a combination of anchor terms); result is deterministic (not dependent on starting value).

Structural Topic Model

In general, we have lots of metadata: e.g. author covariates, like gender or party membership.

But this it non-trivial to include in LDA.

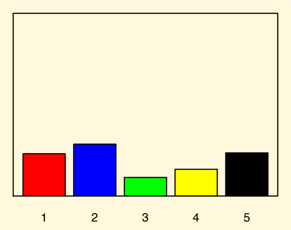
$\rightarrow$ STM = LDA + contextual information

This allows more accurate estimation and more interpretable results.

Also allows us to 'test' hypothesis in more sensible way (though be careful!)

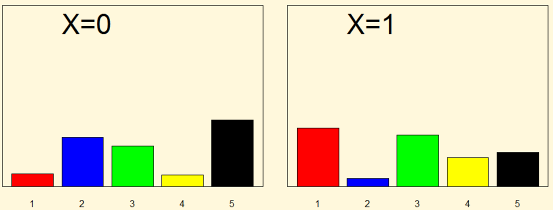
Compare: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.



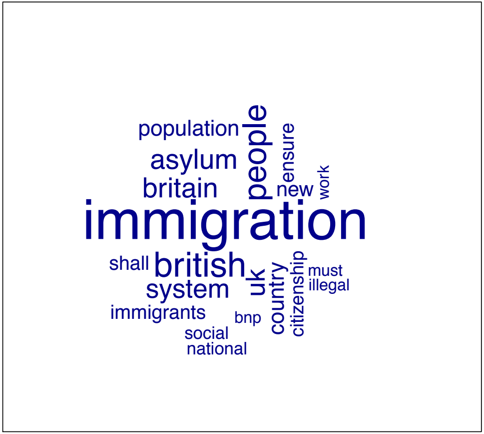
STM, that topic distribution is a function of the document metadata.

e.g. perhaps male author ($X = 0$) documents have different topics relative to female ($X = 1$) author docs.



Compare: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.



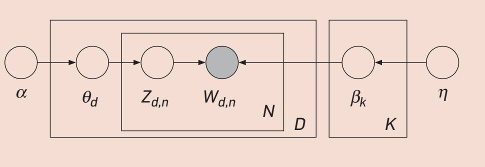
STM: that word distribution is a function of the document metadata.

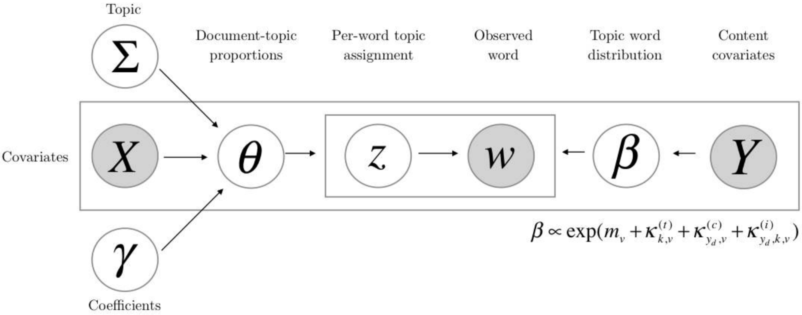
e.g. perhaps right parties ($X = 0$) talk about a given topic differently to left ($X = 1$) parties.

X=0

X=1

Compare: Plate Diagram





$$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$$

- Content (distribution of words over topics): content can vary with binary variable (Liberal v Conservative); with normal LDA, we would need for example 2 topics (Liberal-Guns and Conservative-Guns), but here we can see it is the same topic but approached differently depending on whether document is Liberal or Conservative
- Prevalence (distribution of topics over documents): can vary with both categorical and continuous variables (e.g. time).
- Ameliorates the problems of multimodality through spectral initialisation (if they can find some anchor words for each topic and assign that word only to one topic, all of the other terms in matrix of words over topics are a combination of anchor terms); result is deterministic (not dependent on starting value).

Scaling: Wordfish

Unsupervised Embedding

Basic idea:

Unsupervised Embedding

Basic idea:

- Actors have underlying latent position

Unsupervised Embedding

Basic idea:

- Actors have underlying latent position
- Actors articulate that latent position in their speech

Unsupervised Embedding

Basic idea:

- Actors have underlying latent position
- Actors articulate that latent position in their speech
- This is associated with word usage, so high discriminating words correspond to ideological speech

Unsupervised Embedding

Basic idea:

- Actors have underlying latent position
- Actors articulate that latent position in their speech
- This is associated with word usage, so high discriminating words correspond to ideological speech
- Some words discriminate better than others $\rightsquigarrow$ encode that in our model

Unsupervised Embedding

Basic idea:

- Actors have underlying latent position
- Actors articulate that latent position in their speech
- This is associated with word usage, so high discriminating words correspond to ideological speech
- Some words discriminate better than others $\rightsquigarrow$ encode that in our model

Simplest model: Principal Components

Probabilistic Unsupervised Embeddings

Principal components is powerful

Probabilistic Unsupervised Embeddings

Principal components is powerful$\rightsquigarrow$ statistical model for unsupervised scaling

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing

Probabilistic Unsupervised Embeddings

Principal components is powerful$\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:
  a) Estimate ideal points with few observations

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:
  a) Estimate ideal points with few observations
  b) Makes clear how to extend models

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:
  a) Estimate ideal points with few observations
  b) Makes clear how to extend models
- Clinton, Jackman, and Rivers (2004) $\rightsquigarrow$ intuition about IRT

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:
  a) Estimate ideal points with few observations
  b) Makes clear how to extend models
- Clinton, Jackman, and Rivers (2004) $\rightsquigarrow$ intuition about IRT
- Rivers (2002) $\rightsquigarrow$ Identification conditions

Probabilistic Unsupervised Embeddings

Principal components is powerful $\rightsquigarrow$ statistical model for unsupervised scaling

Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:
  a) Estimate ideal points with few observations
  b) Makes clear how to extend models
- Clinton, Jackman, and Rivers (2004) $\rightsquigarrow$ intuition about IRT
- Rivers (2002) $\rightsquigarrow$ Identification conditions
- Bonica (2014a, 2014b) $\rightsquigarrow$ uses IRT (like the one we're about to use) to scale donors

Probabilistic Unsupervised Embeddings

Principal components is powerful$\rightsquigarrow$ statistical model for unsupervised scaling
Item Response Theory (IRT)

- Origins: educational testing
- Jackman (2002), Clinton, Jackman, and Rivers (2004) apply to roll call voting
- Power of IRT:
  a) Estimate ideal points with few observations
  b) Makes clear how to extend models
- Clinton, Jackman, and Rivers (2004)$\rightsquigarrow$ intuition about IRT
- Rivers (2002)$\rightsquigarrow$ Identification conditions
- Bonica (2014a, 2014b)$\rightsquigarrow$ uses IRT (like the one we're about to use) to scale donors

Scaling

Time Series Problems

We suspect that the German Greens and Social Democrats have moved steadily rightwards, post-reunification.

$\rightarrow$ This is a time series problem, but extant techniques struggle. . .

i.e. hand-coding is expensive,

and hard to find reference texts for Wordscores over time

$\rightarrow$ need to assume lexicon is pretty stable, and that you can identify texts that contain all relevant terms.

Slapin & Proksch (2008)

Would be helpful to have an unsupervised approach, which is not dependent on reference texts

Suggest WORDFISH scaling technique ("A Scaling Model for Estimating Time-Series Party Positions from Text")

1 Begin with naive Bayes assumption: idea that each word's occurrence is independent of all other words in the text.

→ surely false, but convenient starting point.

2 Need a (parametric) model for frequencies of words.

→ Choose *Poisson*: extremely simple because it has only one parameter—λ (which is mean and variance!).

Poisson set up

Recall the density function for Poisson:

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

And in a 'typical' GLM context, we would make

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$

with log-likelihood (dropping constant part),

$$l(\lambda; y) = \sum_{i=1}^{n} y_i \log \lambda - n\lambda.$$

$\rightarrow$ the λ which maximizes this is the MLE.

Here. . .

The count of word j from party i, in year t,

$$y_{ijt} \sim \mathcal{P}(\lambda_{ijt})$$

and

$$\log(\lambda_{ijt}) = \alpha_{it} + \psi_j + \beta_j \times \omega_{it}$$

or

$$\boxed{\lambda_{ijt} = \exp(\alpha_{it} + \psi_j + \beta_j \times \omega_{it})}$$

Notes

One dimensional: which is assumed to be left-right.

→ can limit analysis to given issue area to obtain dimensional scaling in that space.

Parties 'move' to the extent that the words they use look more or less like the words that other parties use.

No over time smoothing/constraints: party manifesto position in t is not modeled as function of party manifesto position in $t-1$

So. . .

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

α_{it} fixed effect(s) for party i in time t: some parties have longer manifestos in certain years (which boosts all counts)

ψ_j word fixed effect: some parties just use certain words more (e.g. their own name)

β_j word specific weight: importance of this word in discriminating between party positions.

ω_{it} estimate of party's position in a given year (so, this applies to specific manifesto)

Problem

NB

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

Nothing on RHS is known: everything needs to be estimated.

$\rightarrow$ unlike GLM arrangement, where Xs are known.

but similar to ideal point estimation wherein the legislators' ideal points are not known: $\Phi(\beta_j' \mathbf{x}_i - \alpha_j)$.

Solution I

NB

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

Suppose we knew the word parameters , ψ_j and β_j.

$\rightarrow$ then we could use a Poisson GLM to estimate α_{it} (a constant/fixed effect) and ω_{it} which is the position.

Or Suppose we knew the party parameters, ω_{it} and α_{it}. Then we could use a Poisson GLM to estimate ψ_j (a constant/fixed effect) and β_j which is a word specific 'effect'.

Solution II: Intuition

first start with good guesses (starting values) of both sets of parameters,

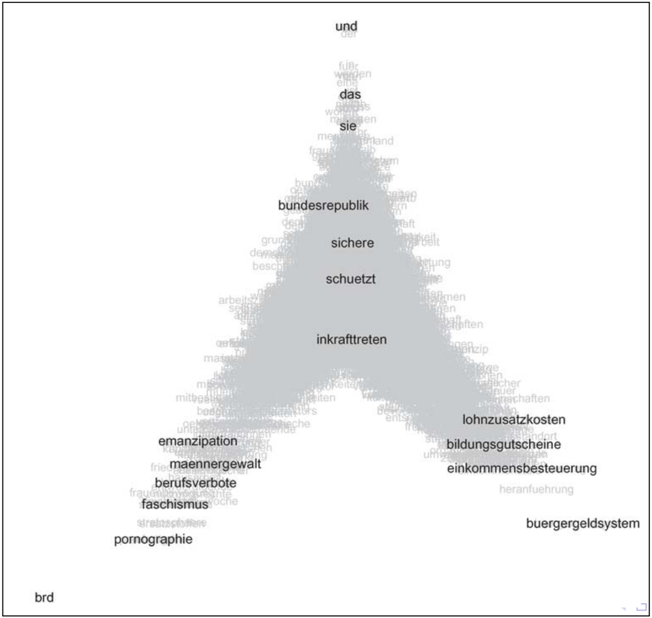
then run a Poisson regression holding word parameters fixed, and estimating the party parameters,

then run a Poisson regression holding party parameters fixed, and estimating the word parameter,

and iterate across these steps until confident we have correct answers (EM algorithm).

btw can use parametric bootstrap for uncertainty estimates.

Results



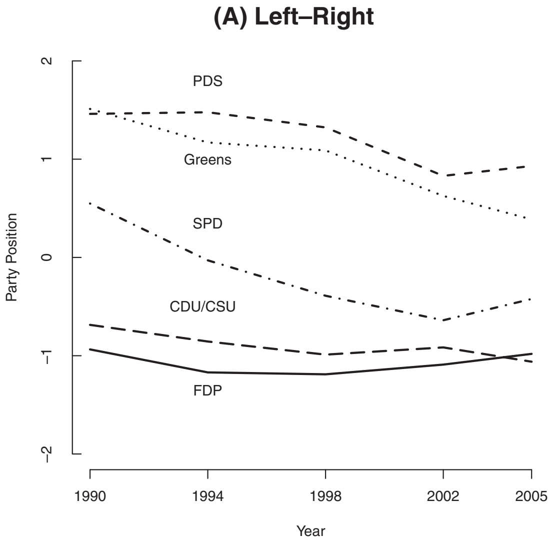
y is word fixed effects: words with high fixed effects have zero weight (v common).

x is word weights: those with high (absolute) weights discriminate well.

Results II

	Top 10 Words Placing Parties on the. . .	
Dimension	**Left**	**Right**
Left-Right	Federal Republic of Germany (BRD)	general welfare payments (Bürgergeldsystem)
	immediate (sofortiger)	introduction (Heranführung)
	pornography (Pornographie)	income taxation (Einkommensbesteuerung)
	sexuality (Sexualität)	non-wage labor costs (Lohnzusatzkosten)
	substitute materials (Ersatzstoffen)	business location (Wirtschaftsstandort)
	stratosphere (Stratosphäre)	university of applied sciences (Fachhochschule)
	women's movement (Frauenbewegung)	education vouchers (Bildungsgutscheine)
	fascism (Faschismus)	mobility (Beweglichkeit)
	Two thirds world (Zweidrittelwelt)	peace tasks (Friedensaufgaben)
	established (etablierten)	protection (Protektion)

Results III, the ω_{it}s



The Problem with Text-Based Scaling

What does validation mean?

1) Replicate NOMINATE, DIME, or other gold standards?
2) Agreement with experts
3) Prediction of other behavior

Must answer this to make progress on pure text scaling