Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
University of Chicago

August 23rd, 2017

Discovery and Measurement

What is the research process? (Grimmer, Roberts, and Stewart 2017)

1) Discovery: a hypothesis or view of the world
2) Measurement according to some organization
3) Causal Inference: effect of some intervention

Text as data methods assist at each stage of research process

Measurement

Two approaches to measurement

1) Use an existing classification scheme to categorize documents
2) Simultaneously discover categories and measure prevalence (repurpose discovery methods)

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

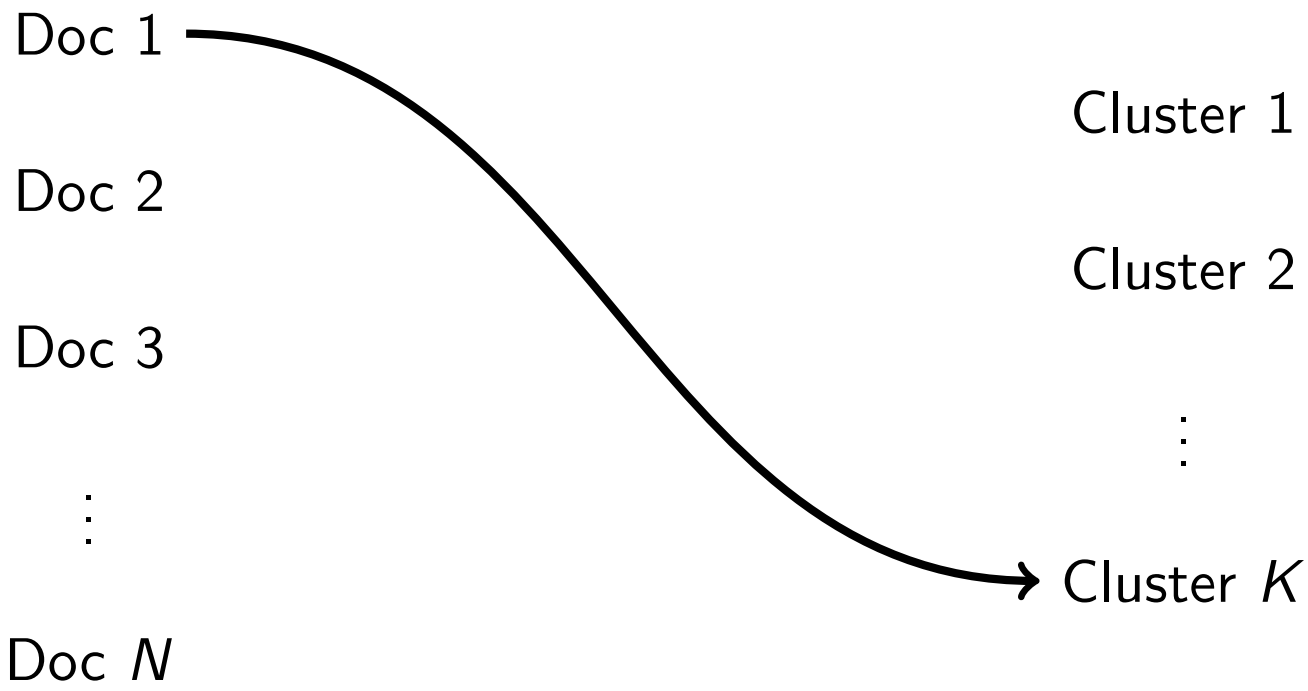
$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

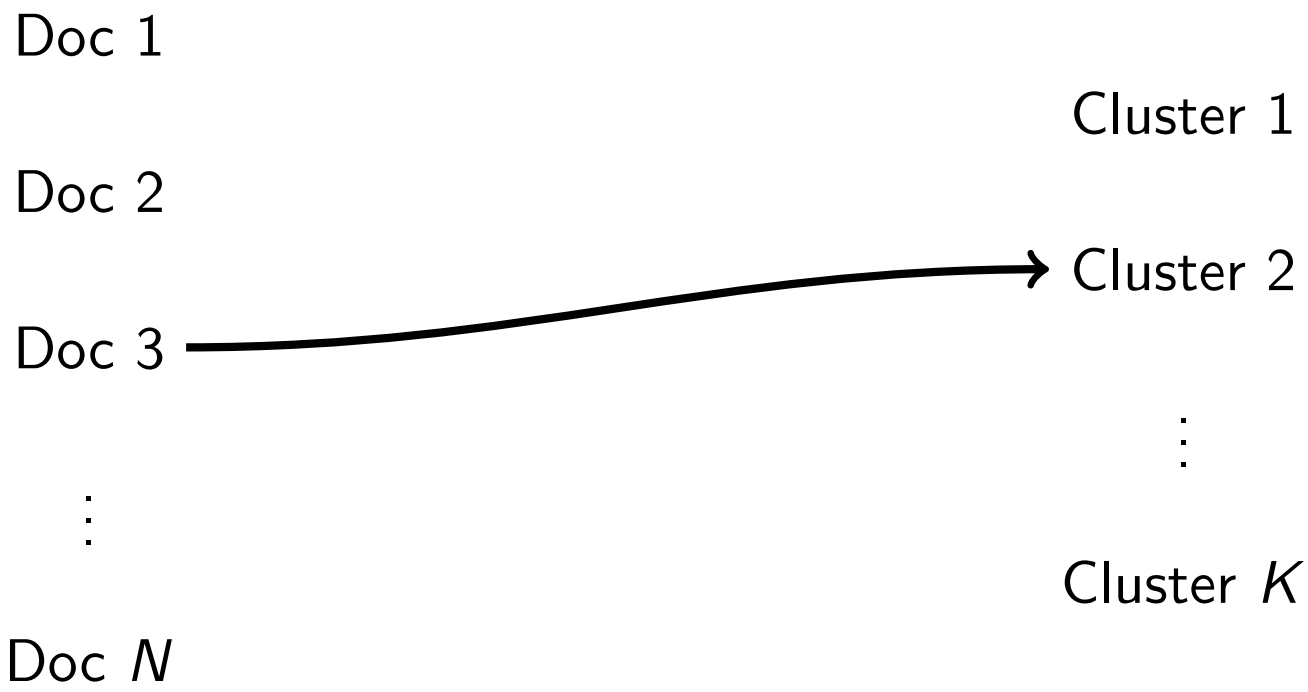
Clustering

Document $\rightsquigarrow$ One Cluster

Topic and Mixed Membership Models

Clustering

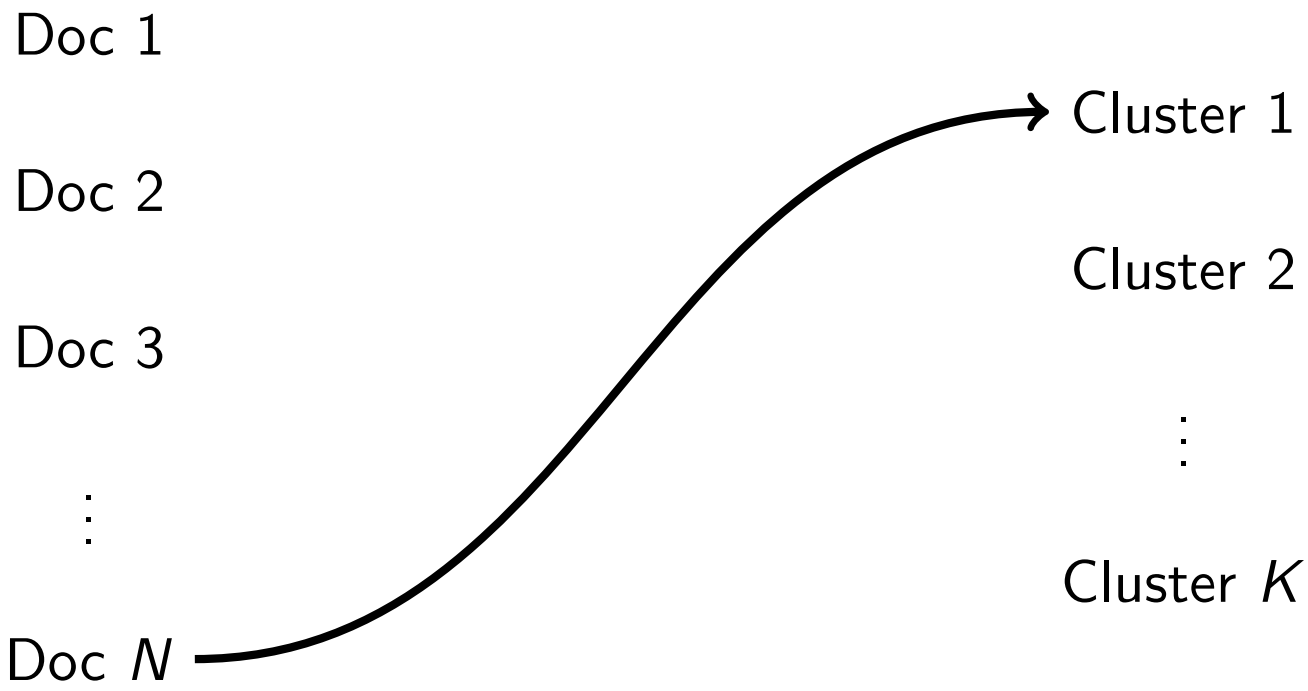
Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

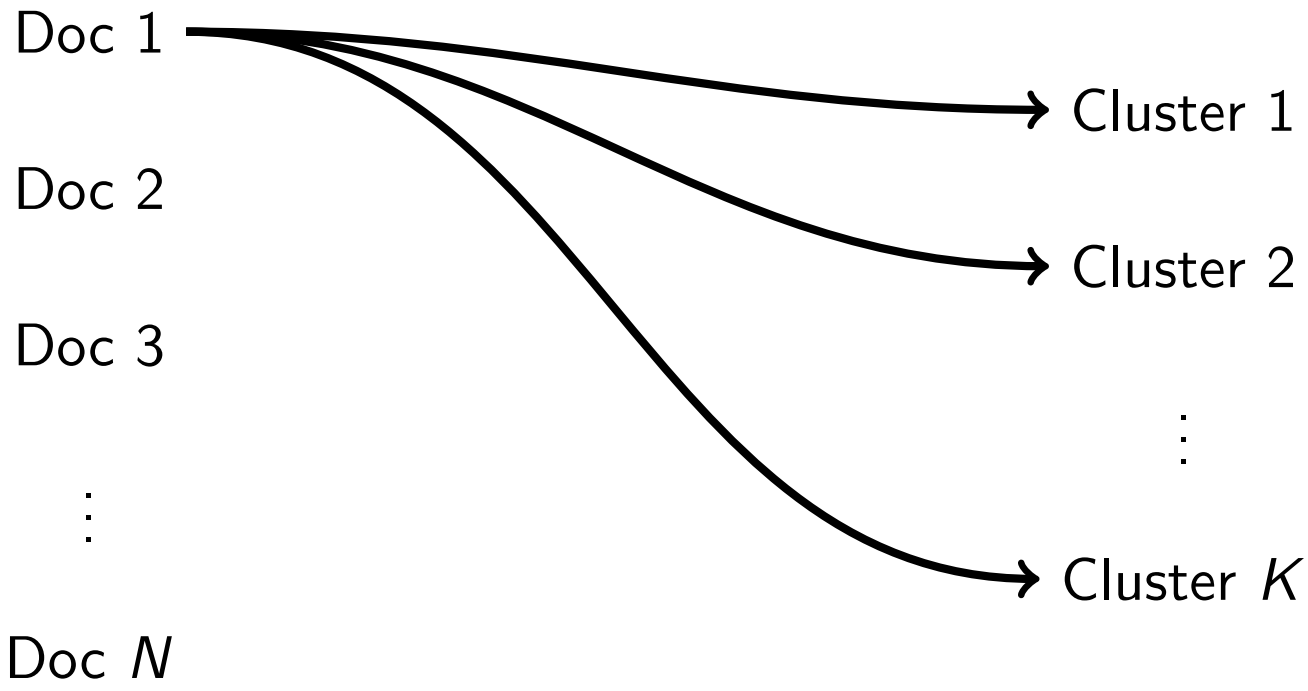
$\vdots$

Cluster K

Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

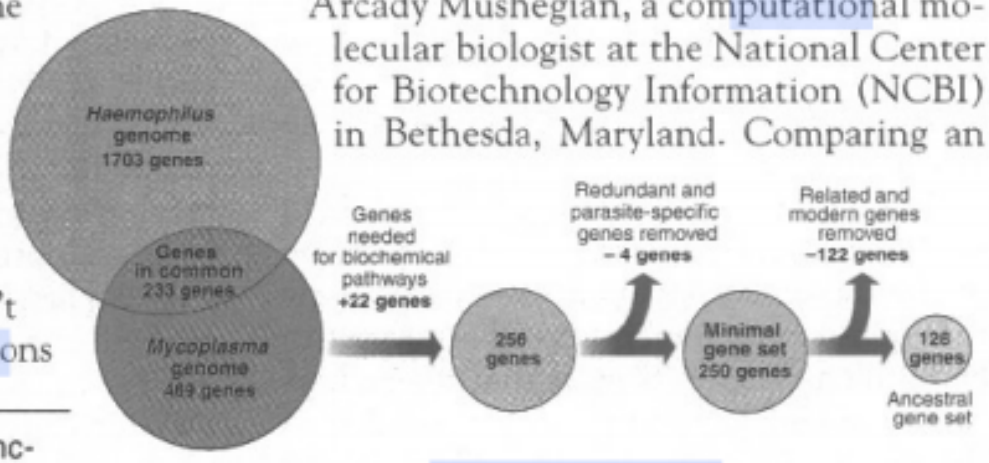


A Statistical Highlighter (With Many Colors)

Seeking Life's Bare (Genetic) Necessities

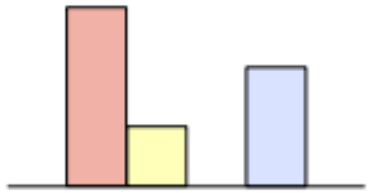
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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*.

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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*.

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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*.

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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*.

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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*.

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ 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*.

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

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

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation$\rightsquigarrow$ Find "closest" distribution

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation$\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ MCMC algorithm to approximate posterior

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\pi, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\pi | \boldsymbol{\alpha}) p(\boldsymbol{T} | \pi) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\pi_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\tau_{im} | \pi) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation$\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ MCMC algorithm to approximate posterior

Described in the slides appendix

Running a Topic Model with STM

`to the STM Code`

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	...	Word_J
Doc_1	0	1	...	0
Doc_2	2	0	...	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	...	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	...	Word_J
Doc_1	0	1	...	0
Doc_2	2	0	...	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	...	1

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

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	$\ldots$	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)

Why does this work$\rightsquigarrow$ Co-occurrence

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

Reconsider document-term matrix

	Word_1	Word_2	...	Word_J
Doc_1	0	1	...	0
Doc_2	2	0	...	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	...	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 linear algebra (provides similar results, difficult to generalize)

Why does this work$\rightsquigarrow$ Co-occurrence

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

	Word_1	Word_2	...	Word_J
Doc_1	0	1	...	0
Doc_2	2	0	...	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	...	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 linear algebra (provides similar results, difficult to generalize)
Biclustering: Models that partition documents and words simultaneously

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

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

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

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

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

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

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

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)

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

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

Validation$\rightsquigarrow$ Topic Intrusion

Wednesday$\rightsquigarrow$ discussed several validations

- 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 1: Japanese Campaign Manifestos (Catalinac 2011)

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
- Latent Dirichlet Allocation (on japanese texts)

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level

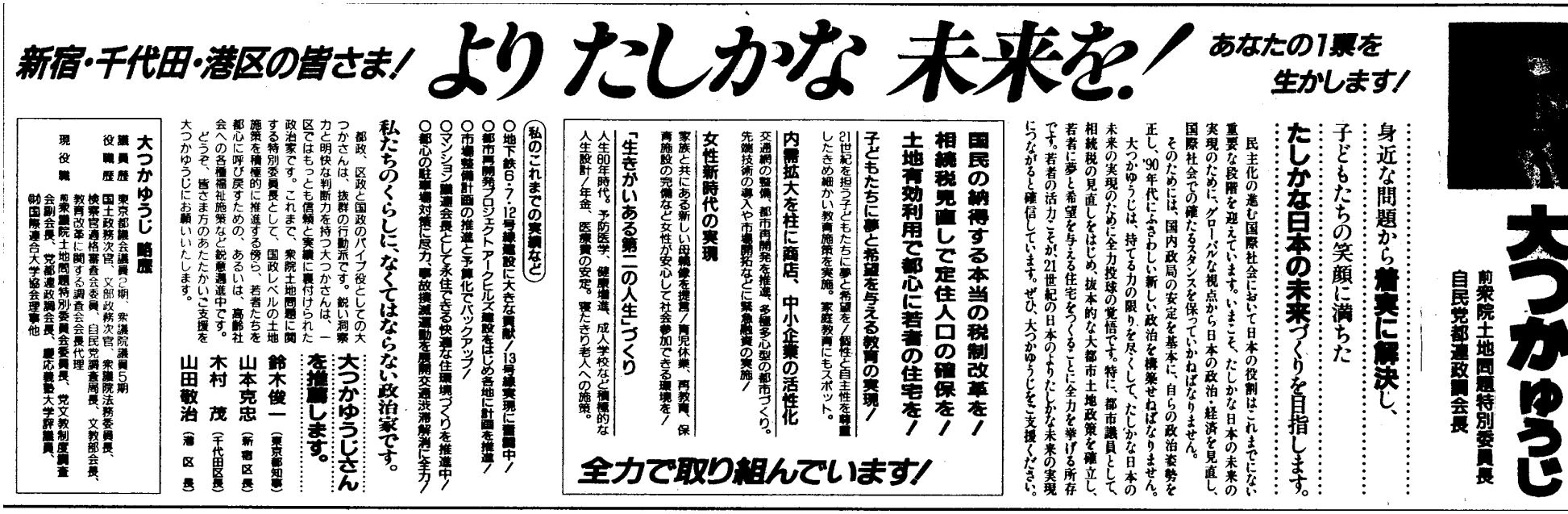
Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Typical Manifesto:



新宿・千代田・港区の皆さま！

よりたしかな未来を！

あなたの1票を生かします！

大つかゆうじ

前衆院土地問題特別委員長

自民党都連政調会長

身近な問題から着実に解決し、

子どもたちの笑顔に満ちた

たしかな日本の未来づくりを目指します。

民主化の進む国際社会において日本の役割はこれまでにない重要な段階を迎えています。いまこそ、たしかな日本の未来の実現のために、グローバルな視点から日本の政治・経済を見直し、国際社会での確たるスタンスを保っていかねばなりません。

そのためには、国内政局の安定を基本に、自らの政治姿勢を正し、'90年代にふさわしい新しい政治を構築せねばなりません。

大つかゆうじは、持てる力の限りを尽くして、たしかな日本の未来の実現のために全力投球の覚悟です。特に、都市議員として、相続税の見直しをはじめ、抜本的な大都市土地政策を確立し、若者に夢と希望を与える住宅をつくることに全力を挙げる所存です。若者の活力こそが、21世紀の日本のよりたしかな未来の実現につながると確信しています。ぜひ、大つかゆうじをご支援ください。

国民の納得する本当の税制改革を！

相続税見直しで定住人口の確保を！

土地有効利用で都心に若者の住宅を！

子どもたちに夢と希望を与える教育の実現！

21世紀を担う子どもたちに夢と希望を／個性と自主性を尊重したきめ細かい教育施策を実施。家庭教育にもスポット。

内需拡大を柱に商店、中小企業の活性化

交通網の整備、都市再開発を推進。多極多心型の都市づくり。先端技術の導入や市場開拓などに緊急融資の実施／

女性新時代の実現

家族と共にある新しい母親像を提言／育児休業、再教育、保育施設の完備など女性が安心して社会参加できる環境を／

「生きがいある第二の人生」づくり

人生80年時代。予防医学、健康増進、成人学校など積極的な人生設計／年金、医療費の安定。寝たきり老人への施策。

全力で取り組んでいます！

私のこれまでの実績など

○地下鉄6・7・12号線建設に大きな貢献／13号線実現に奮闘中！

○都市再開発プロジェクトアークヒルズ建設をはじめ各地に計画を推進！

○市場整備計画の推進と予算化でバックアップ！

○マンション建設会長として永住できる快適な住環境づくりを推進中！

○都心の駐車場対策に尽力、事故撲滅運動を展開交通渋滞解消に全力！

私たちのくらしに、なくてはならない政治家です。

都政、区政と国政のパイプ役としての大つかさんは、鋭い洞察力と明快な判断力を持つ大つかさんは、一区ではもっとも信頼と実績に裏付けられた政治家です。これまで、衆院土地問題に関する特別委員長として、国政レベルの土地施策を積極的に推進する傍ら、若者たちを都心に呼び戻すための、あるいは、高齢社会への各種福祉施策など鋭意奮進中です。

どうぞ、皆さま方のあたたかいご支援を大つかゆうじにお願いいたします。

大つかゆうじさん

を推薦します。

鈴木俊一（東京都知事）

山本克忠（新宿区長）

木村茂（千代田区長）

山田敬治（港区長）

大つかゆうじ 略歴

議員歴 東京都議会議員2期、衆議院議員5期

役職歴 国土政務次官、文部政務次官、衆議院法務委員長、検察官適格審査会委員、自民党調査局長、文教部会長、教育改革に関する調査会長代理

現役職 衆議院土地問題特別委員長、党文教制度調査会副会長、党都連政調会長、慶応義塾大学評議員、財国際連合大学協会理事他

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
    - Available only at district level

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM
- Harder for Japanese

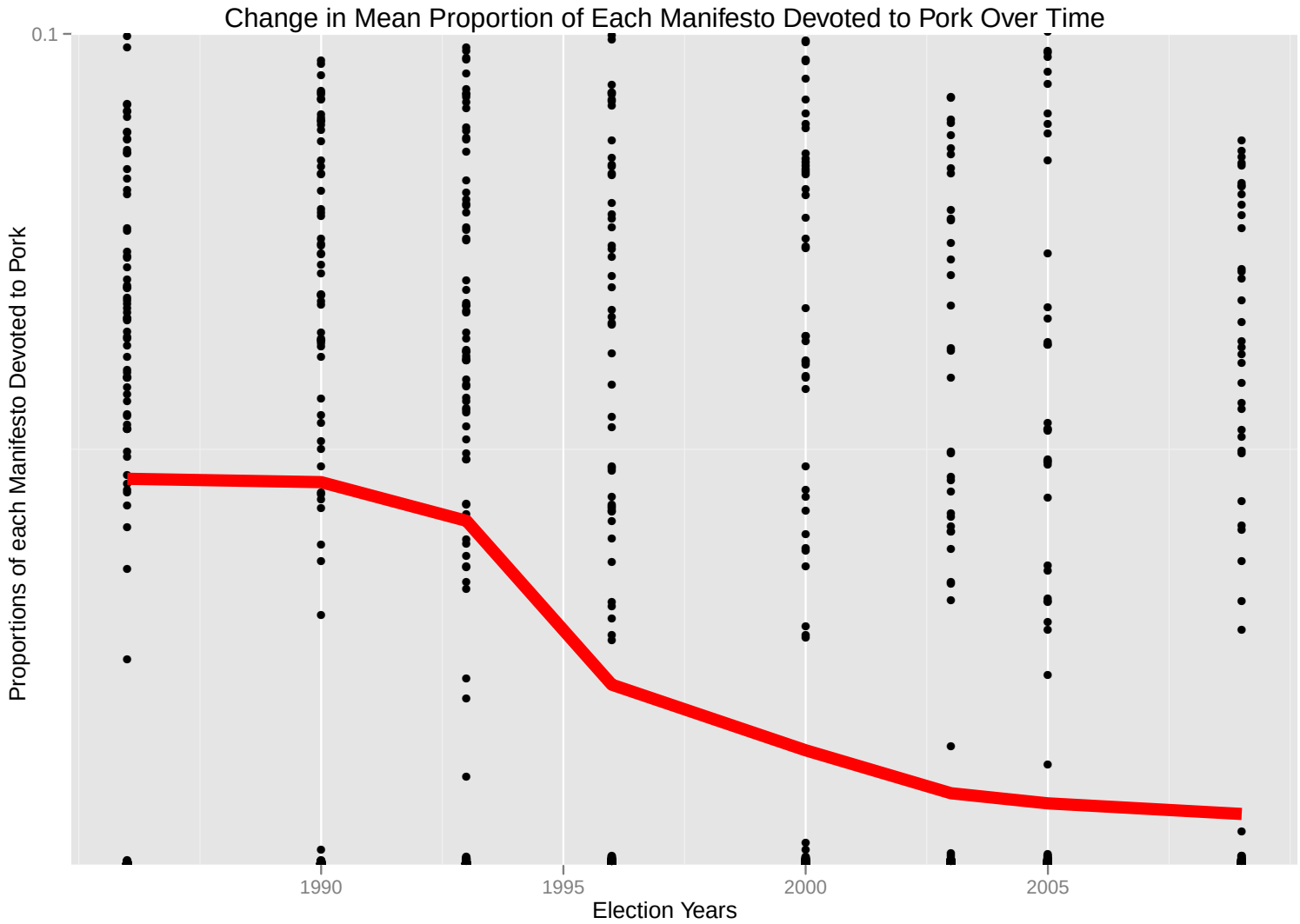
Example 1: Japanese Campaign Manifestos (Catalinac 2014)

- Applies Vanilla LDA (using R Code)
- Output: topics (with Japanese characters)

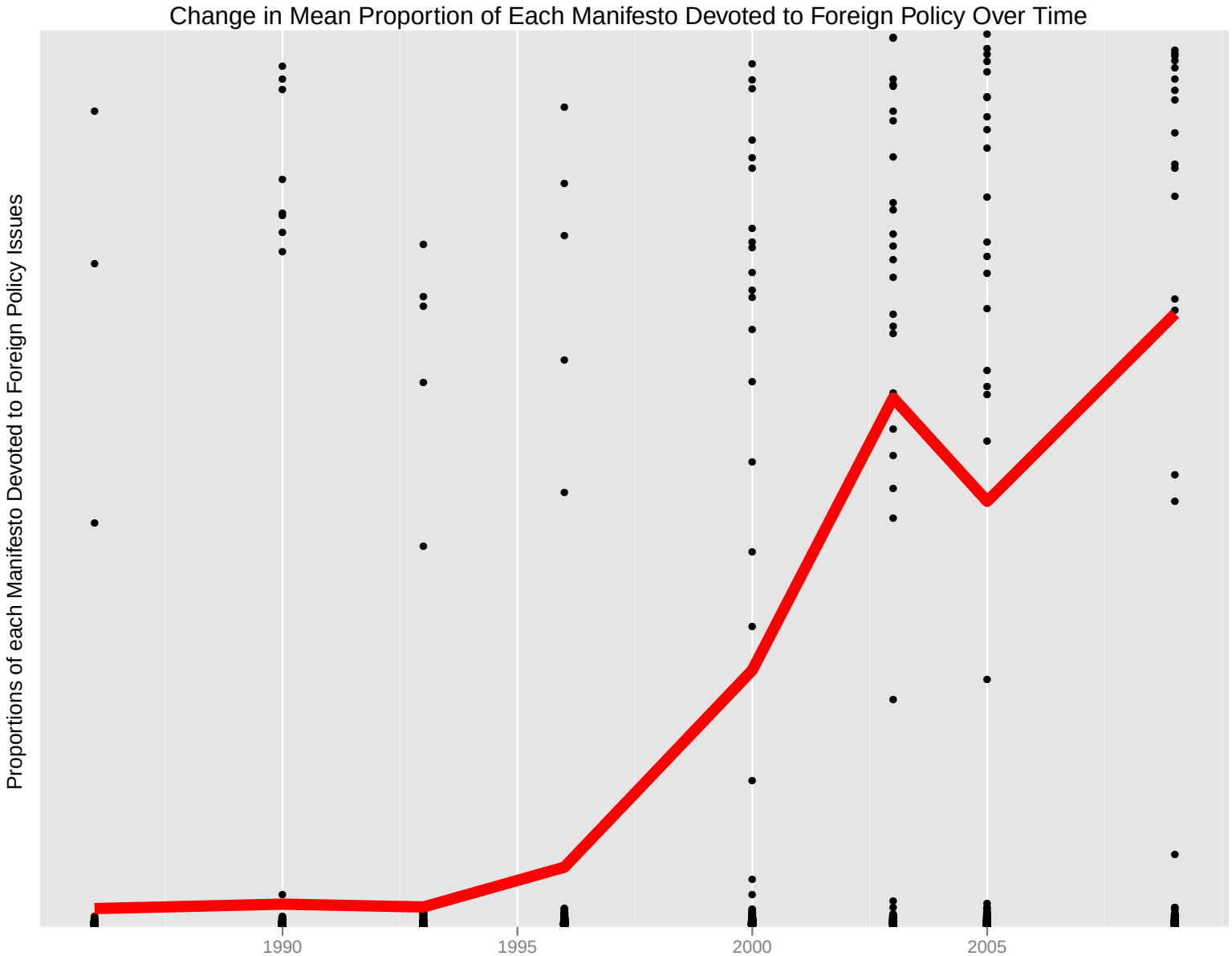
Example 1: Japanese Campaign Manifestos (Catalinac 2011)

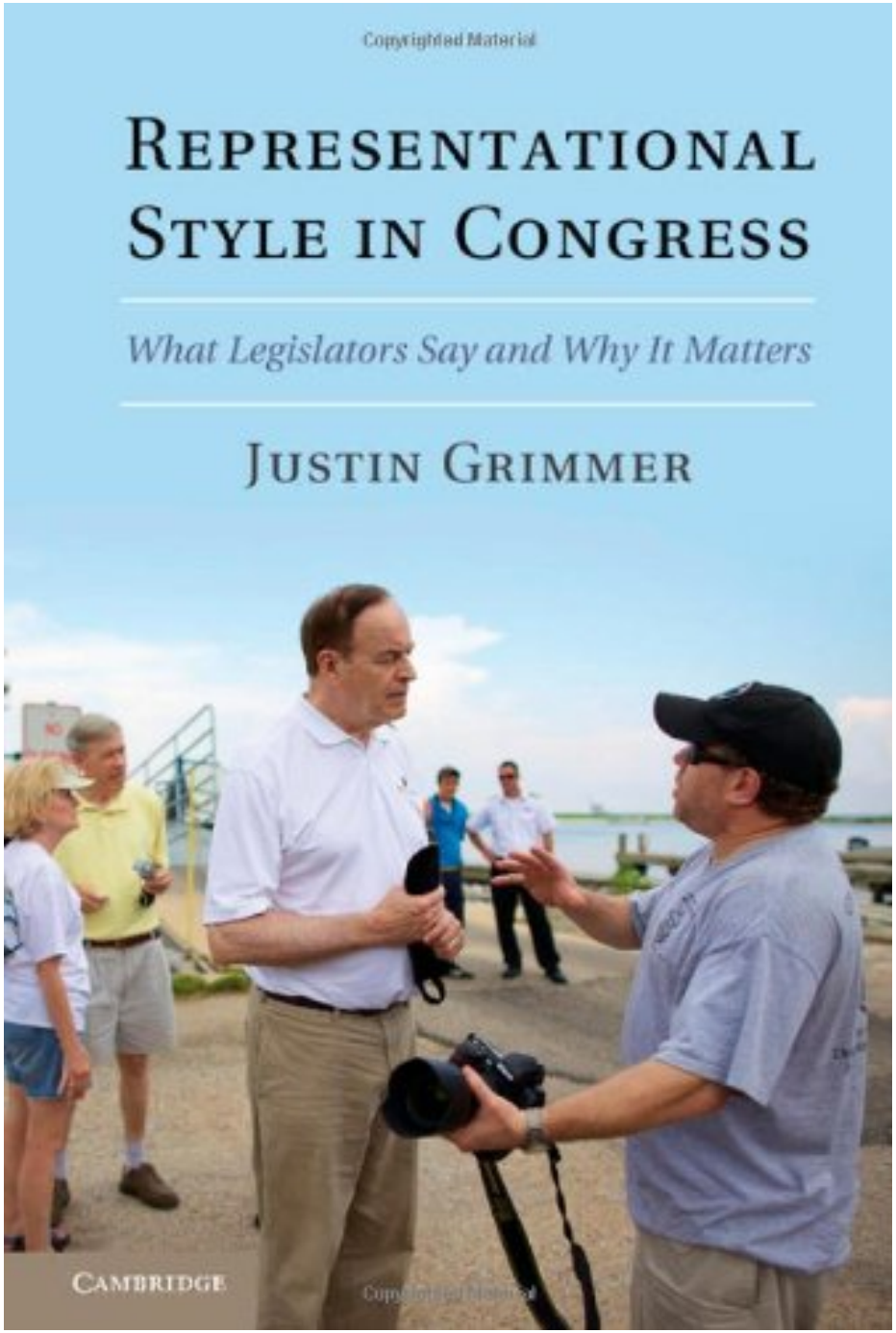
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic
改革	年金	推進	区	政治	日本
郵政	円	整備	政策	改革	国
民営	廃止	図る	地域	国民	外交
小泉	改革	つとめる	まち	企業	国家
構造	兆	社会	鹿児島	自民党	社会
政府	実現	対策	全力	日本	国民
官	無駄	振興	選挙	共産党	保障
推進	日本	充実	国政	献金	安全
民	増税	促進	作り	金権	地域
自民党	削減	安定	横浜	党	拉致
日本	一元化	確立	対策	選挙	経済
制度	政権	企業	中小	禁止	守る
民間	子供	実現	発電	憲法	問題
年金	地域	中小	推進	腐敗	北朝鮮
実現	ひと	育成	エネルギー	団体	教育
進める	サラリーマン	制度	企業	区	責任
断行	制度	政治	声	ソ連	力
地方	議員	地域	実現	守る	創る
止める	金	福祉	活性	平和	安心
保障	民主党	事業	自民党	円	目指す
財政	年間	改革	地方	反対	誇り
作る	一掃	確保	尽くす	真	憲法
賛成	郵政	強化	商店	是正	可能
社会	道路	教育	いかす	一掃	道
国民	交代	施設	全国	悪政	未来
公務員	社会保険庁	生活	政党	抜本	ひと
力	月額	支援	ひと	定数	再生
経済	手当	環境	支援	政党	将来
国	談合	発展	経済	金丸	解決
安心	支援	施策	福祉	改悪	基本
Postal privatization	Reducing Wasteful Public Spending	Pork for the District	Policies for the district	Political Reform	Nation

Example 1: Japanese Campaign Manifestos (Catalinac 2011)



Example 1: Japanese Campaign Manifestos (Catalinac 2011)




Copyrighted Material
REPRESENTATIONAL STYLE IN CONGRESS
What Legislators Say and Why It Matters
JUSTIN GRIMMER
CAMBRIDGE

Example 2: 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 2: 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)

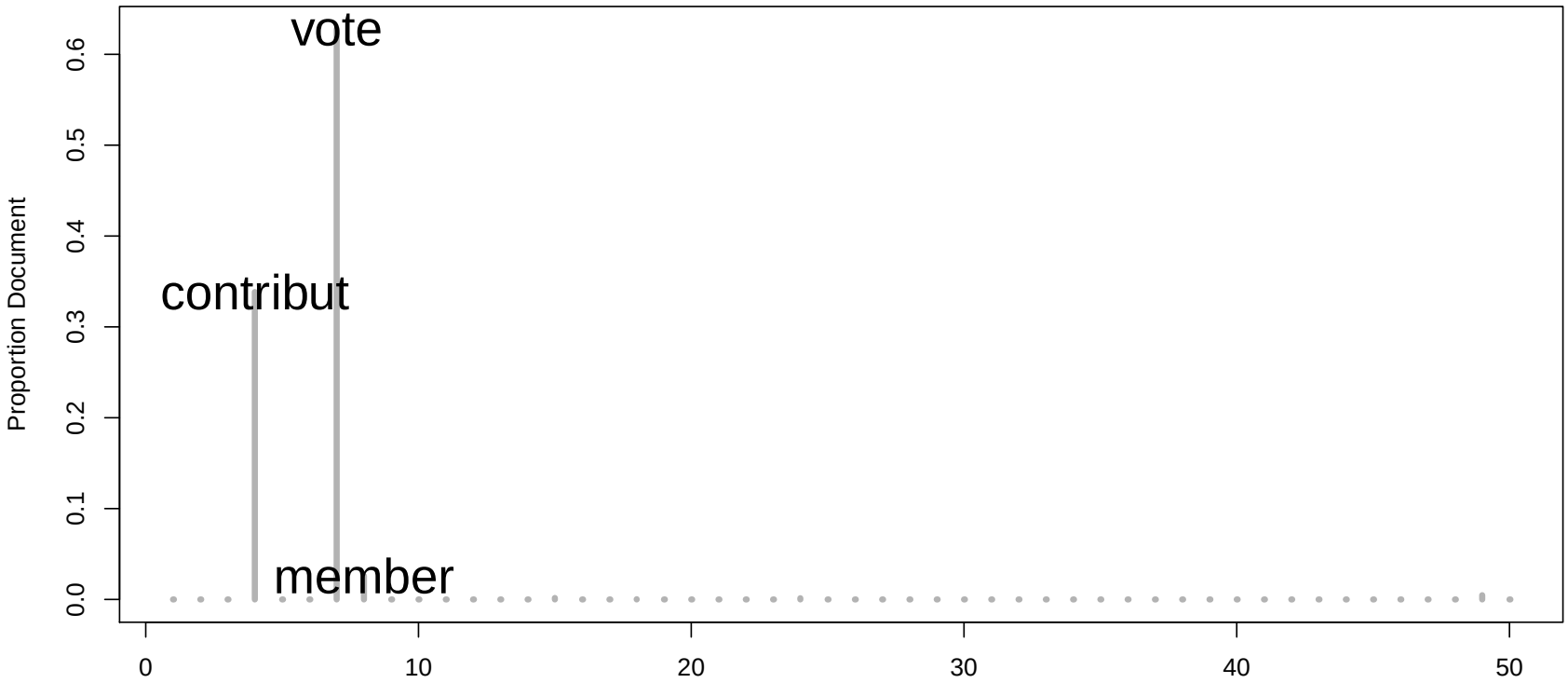
Example 2: Automated Literature Reviews

Example topics:

Label	Stems	Proportic
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

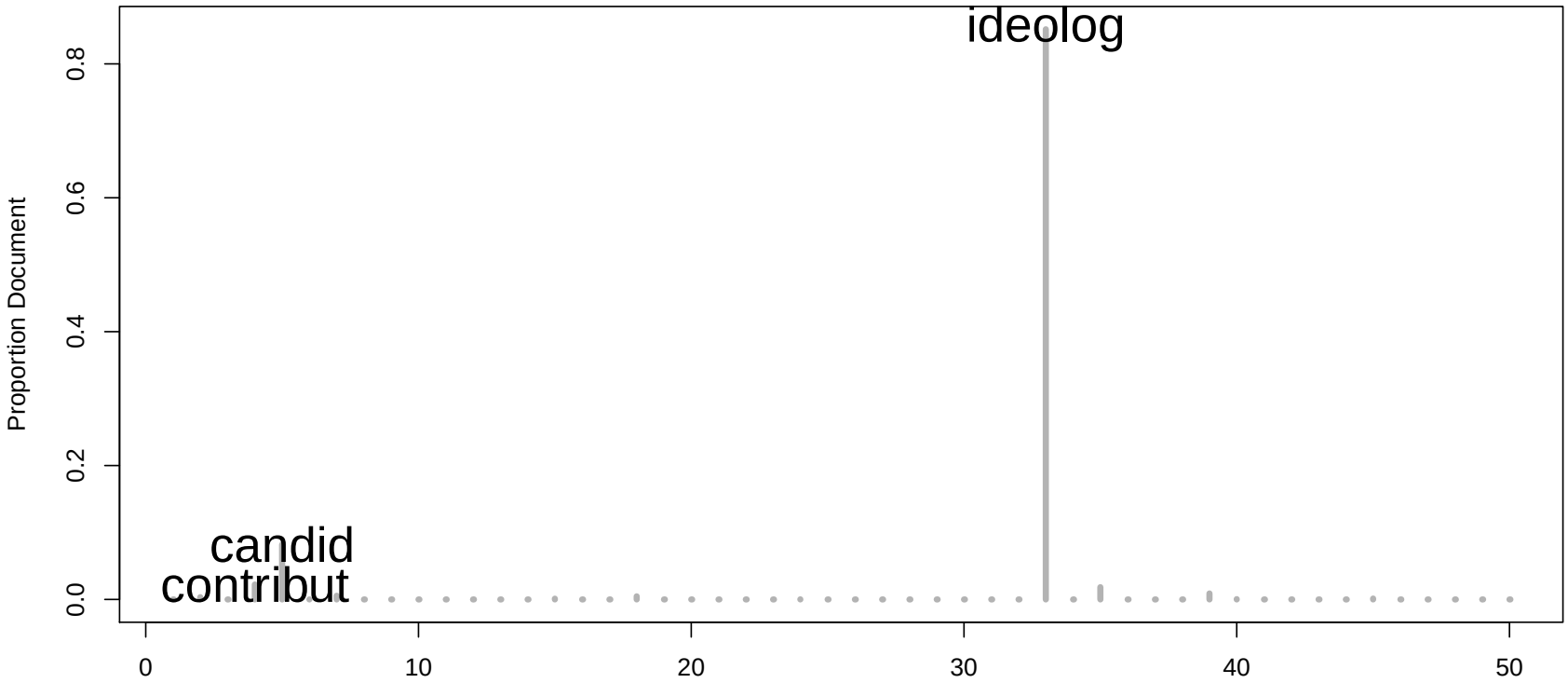
Example Document

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



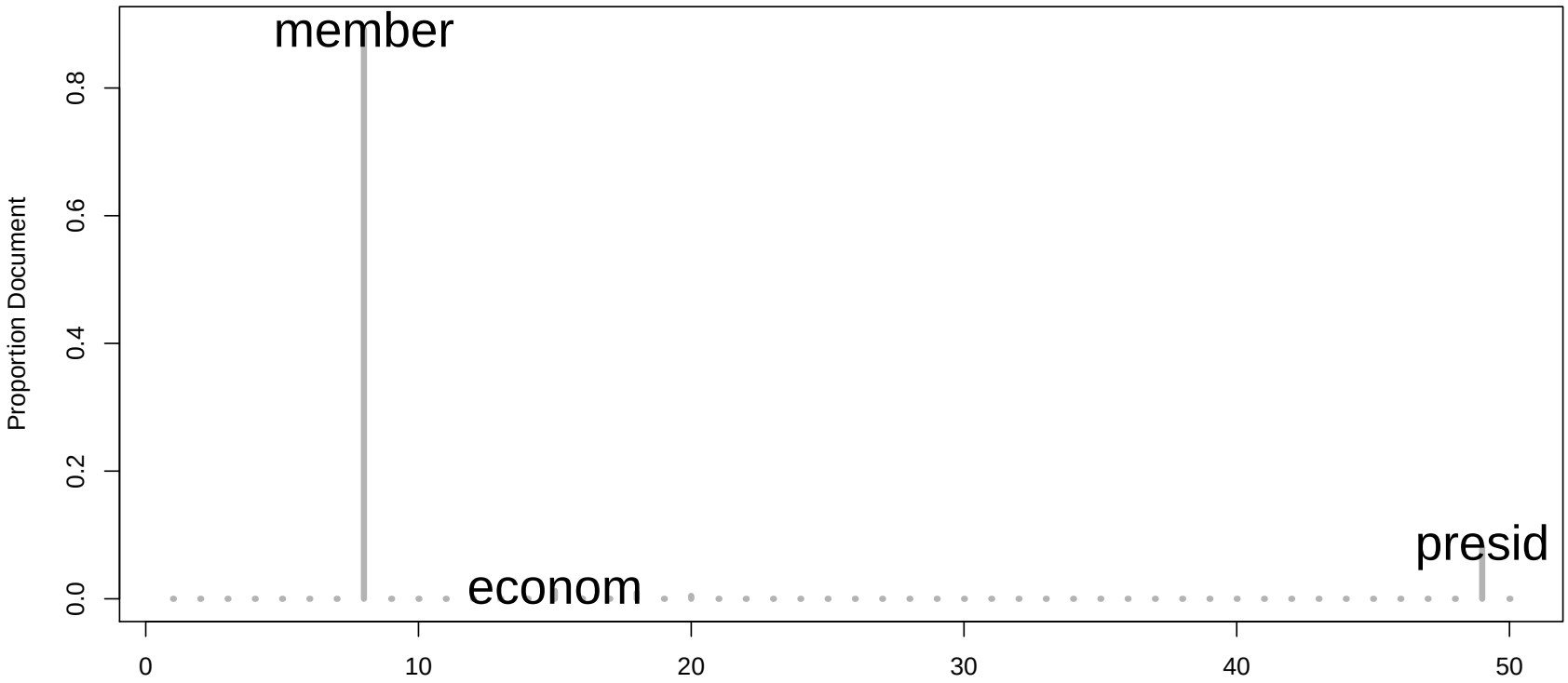
Example Document

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



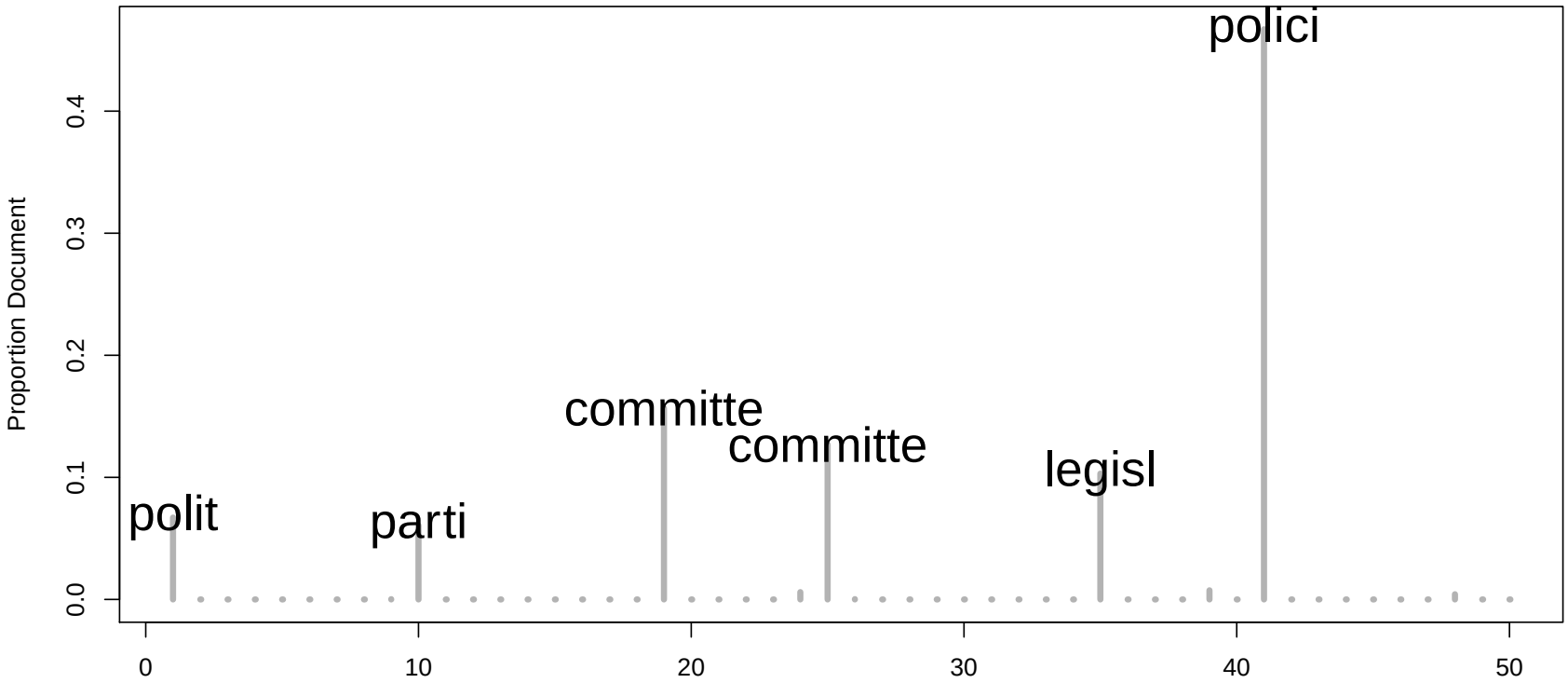
Example Document

Parker (1980) "Cycles in Congressional District Attention"

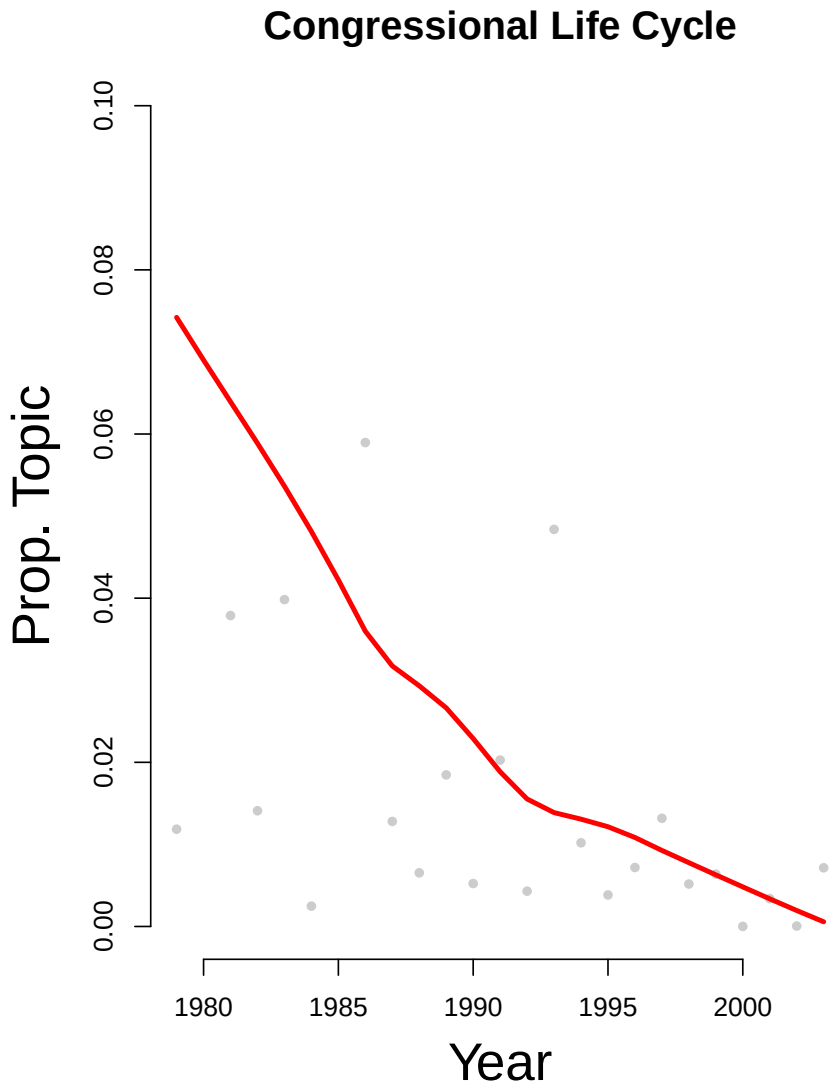


Example Document

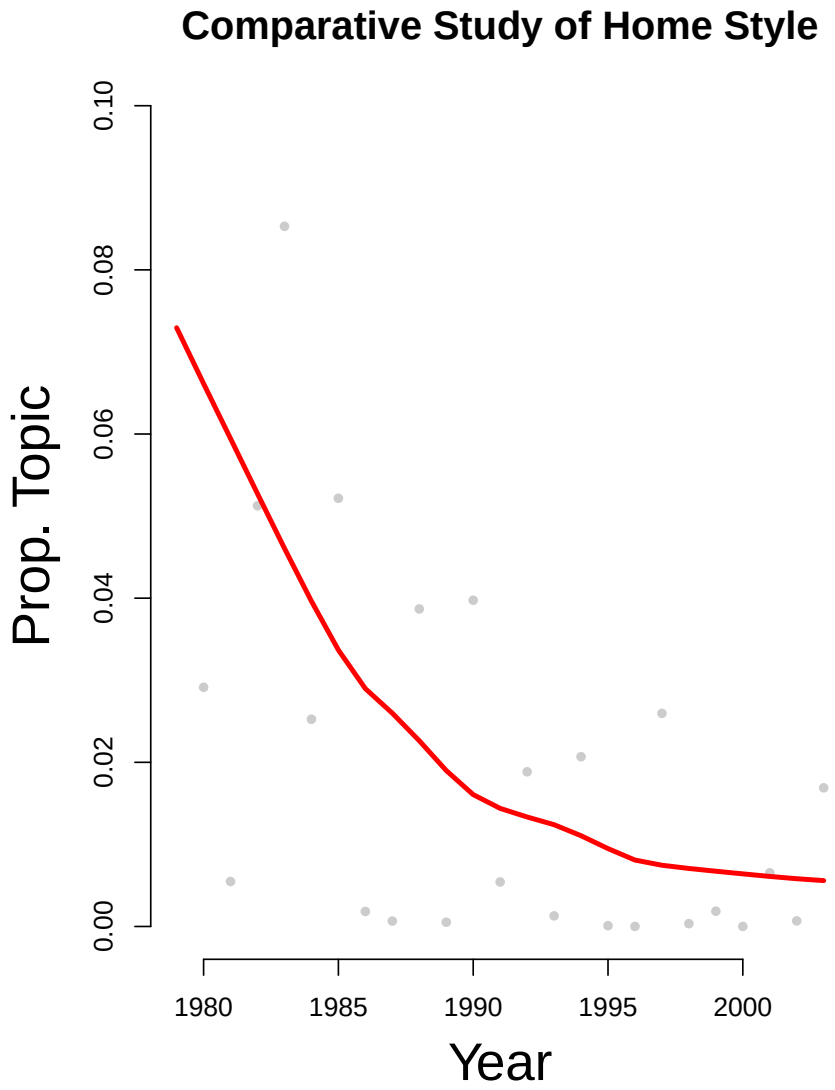
Shepsle (1985) "Policy Consequences of Government by Congressional Subcommittees"



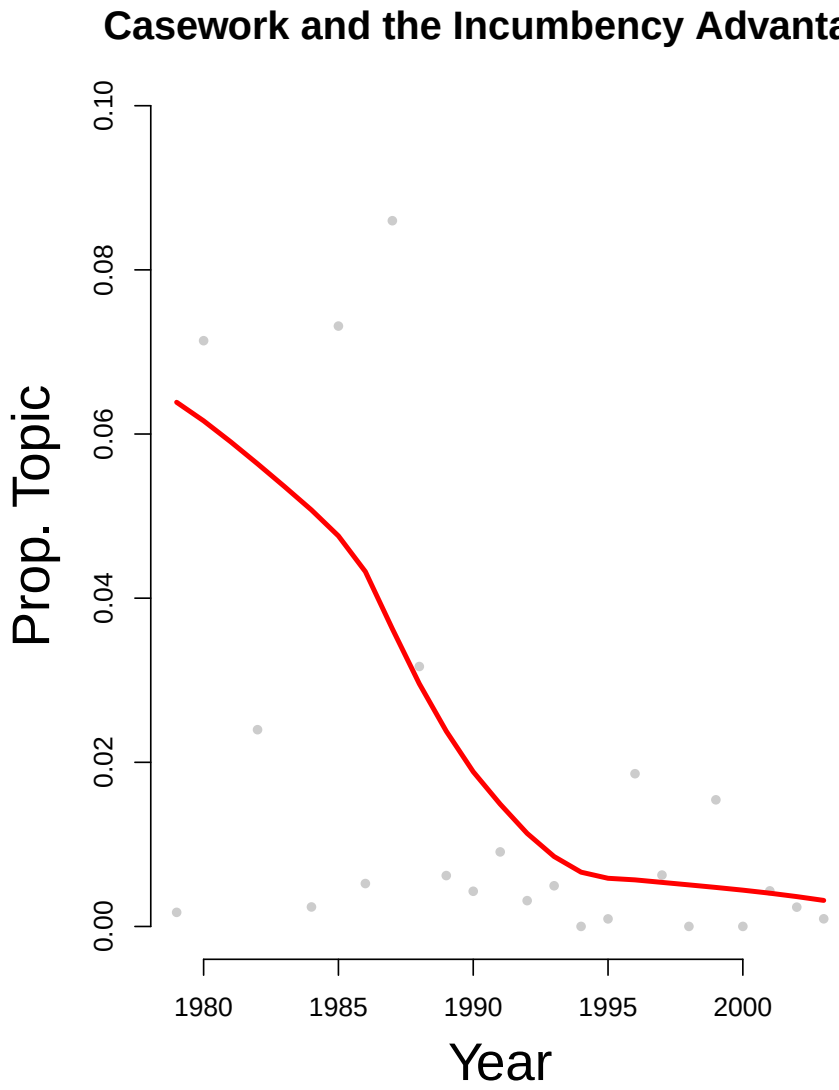
History of Home Style



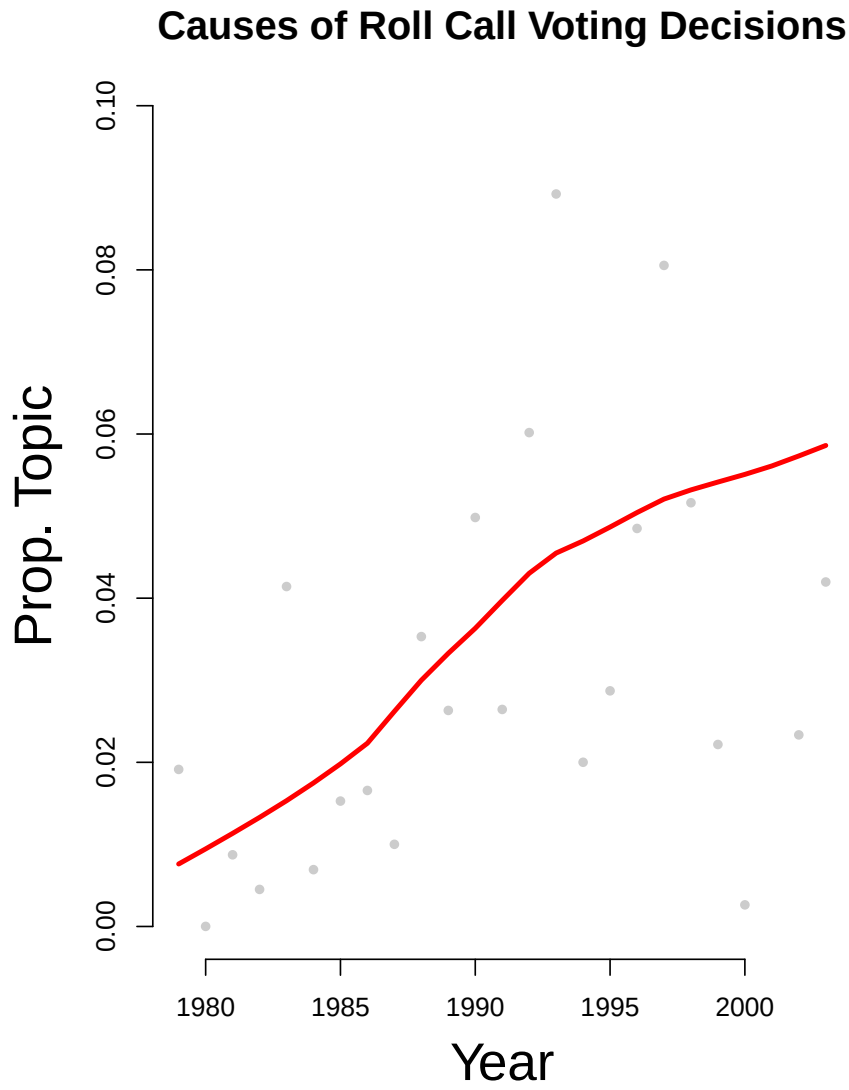
History of Home Style



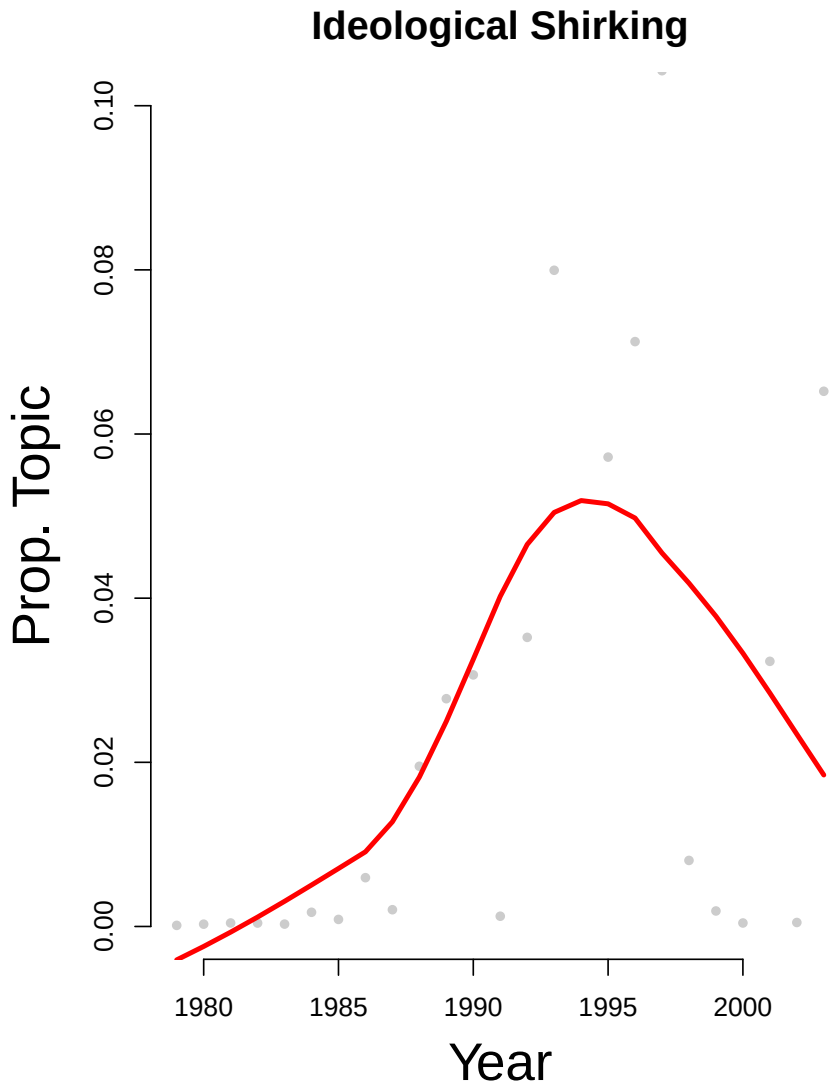
History of Home Style



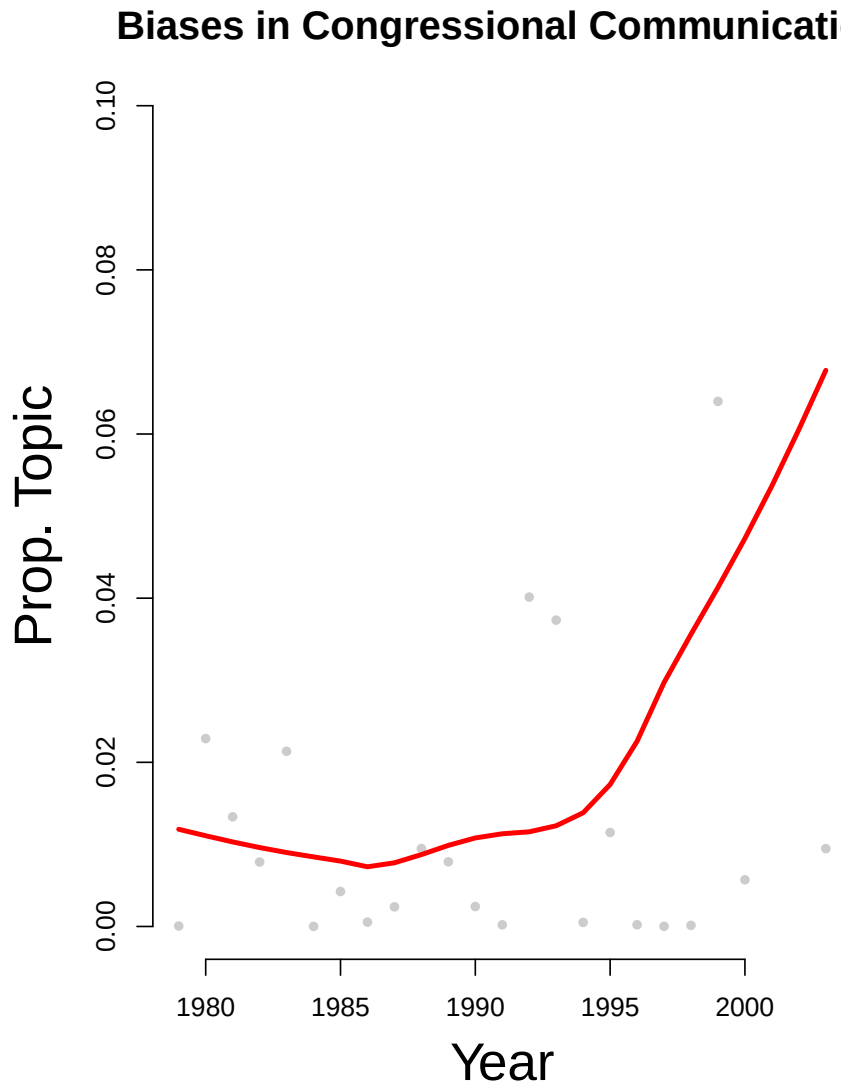
History of Home Style

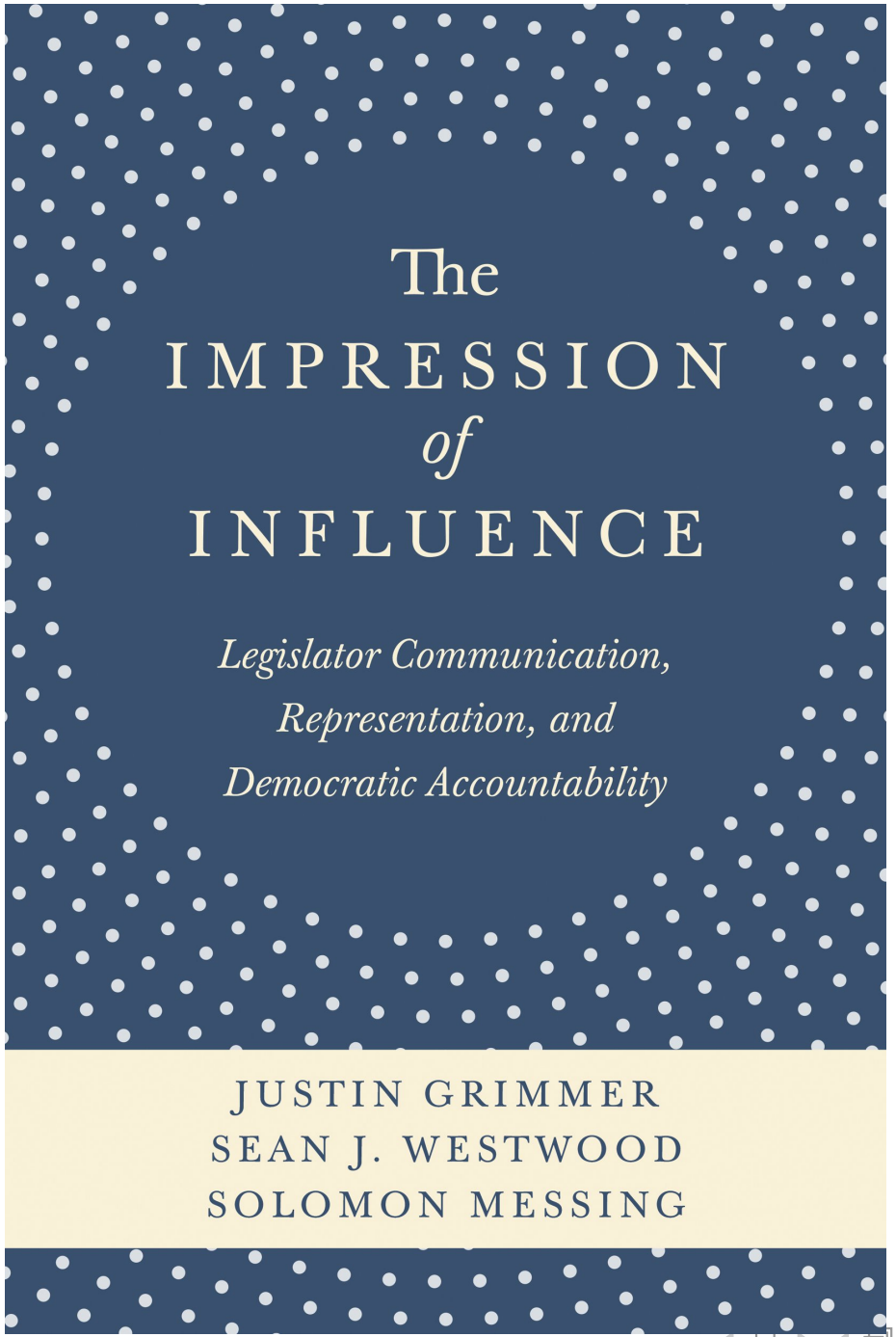


History of Home Style



History of Home Style




The
IMPRESSION
of
INFLUENCE
Legislator Communication,
Representation, and
Democratic Accountability
JUSTIN GRIMMER
SEAN J. WESTWOOD
SOLOMON MESSING

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

Correlated Topic Models

Dirichlet distribution$\rightsquigarrow$ Assumes negative covariance between topics
Logistic Normal Distribution$\rightsquigarrow$ Allows some positive covariance between topics

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\eta}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\pi}_i &= \frac{\exp(\boldsymbol{\eta}_i)}{\sum_{k=1}^{K} \exp(\eta_{ik})} \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

LDA Revisited

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

LDA Revisited

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

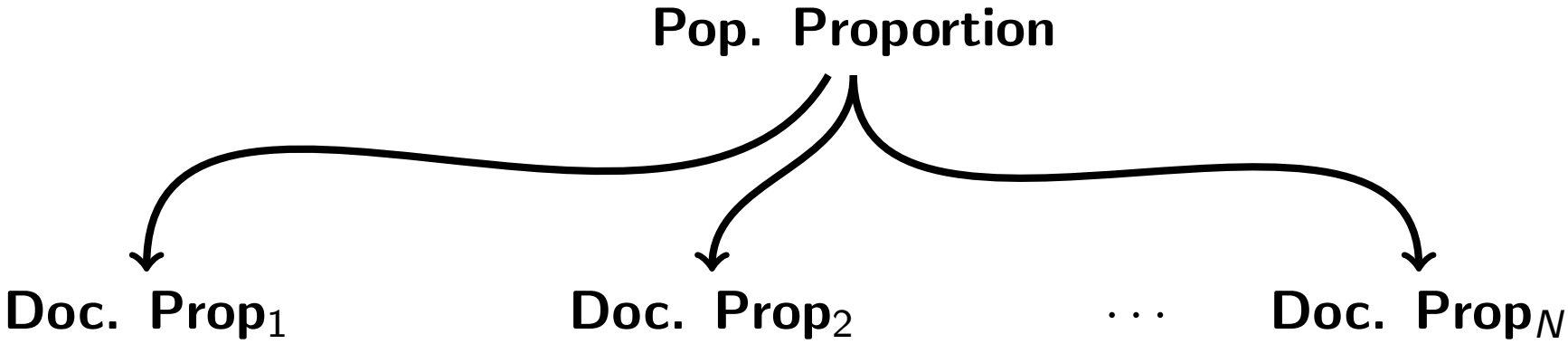
A General Hierarchical Structure

LDA:

Pop. Proportion

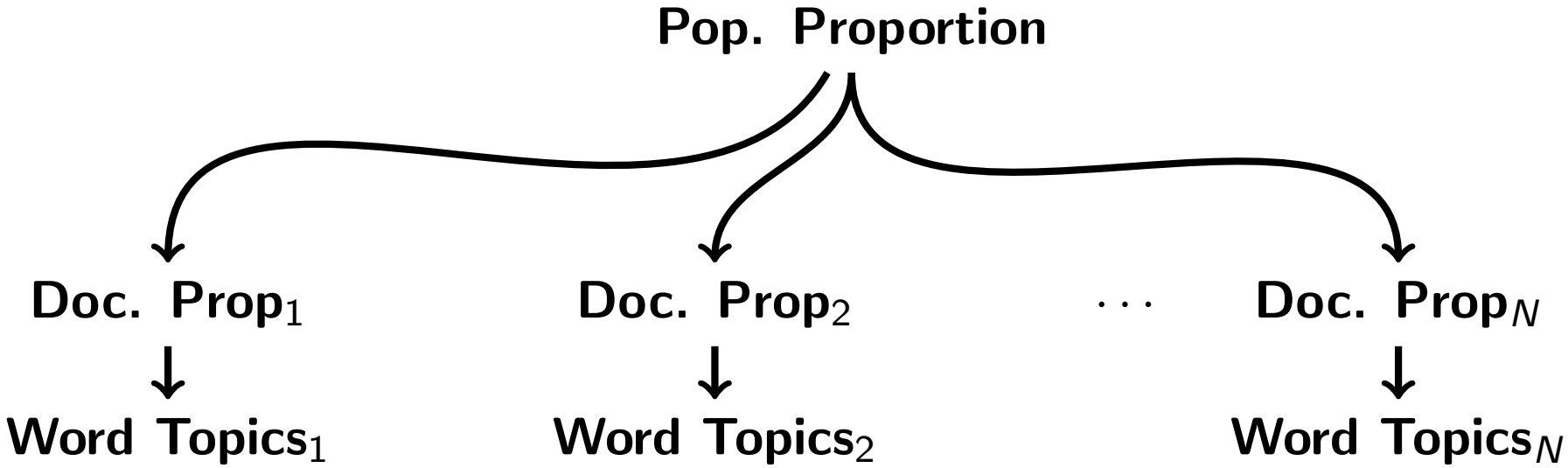
A General Hierarchical Structure

LDA:



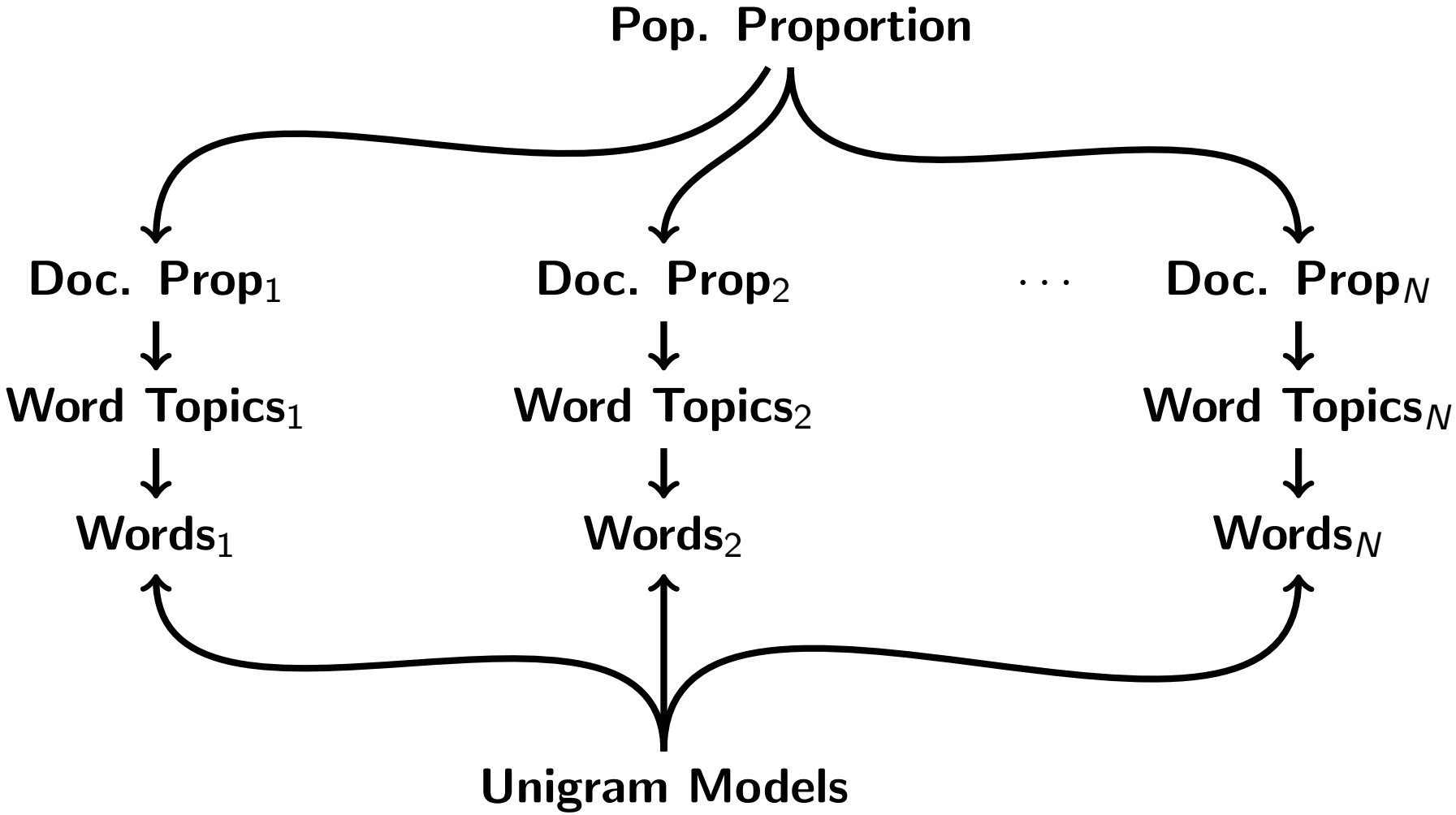
A General Hierarchical Structure

LDA:



A General Hierarchical Structure

LDA:



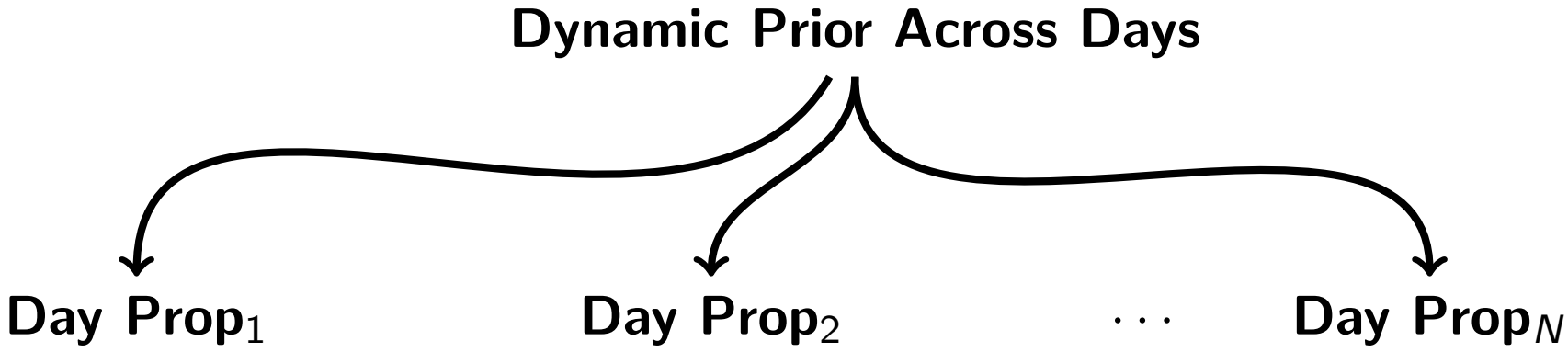
A General Hierarchical Structure

Dynamic Topic Model (Quinn et al 2010)

Dynamic Prior Across Days

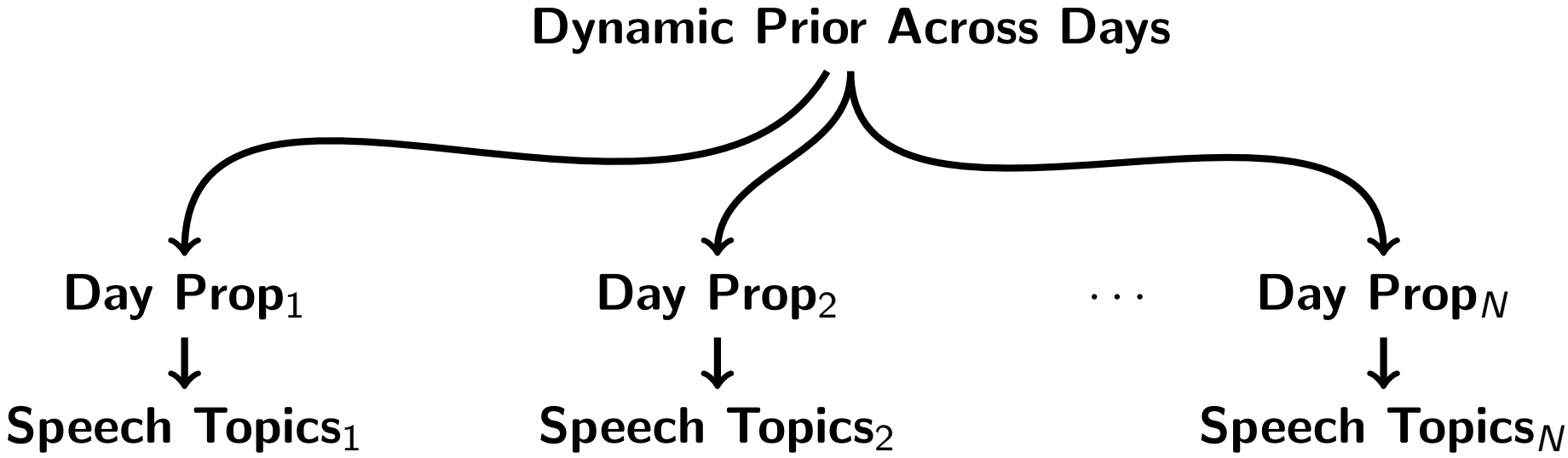
A General Hierarchical Structure

Dynamic Topic Model (Quinn et al 2010)



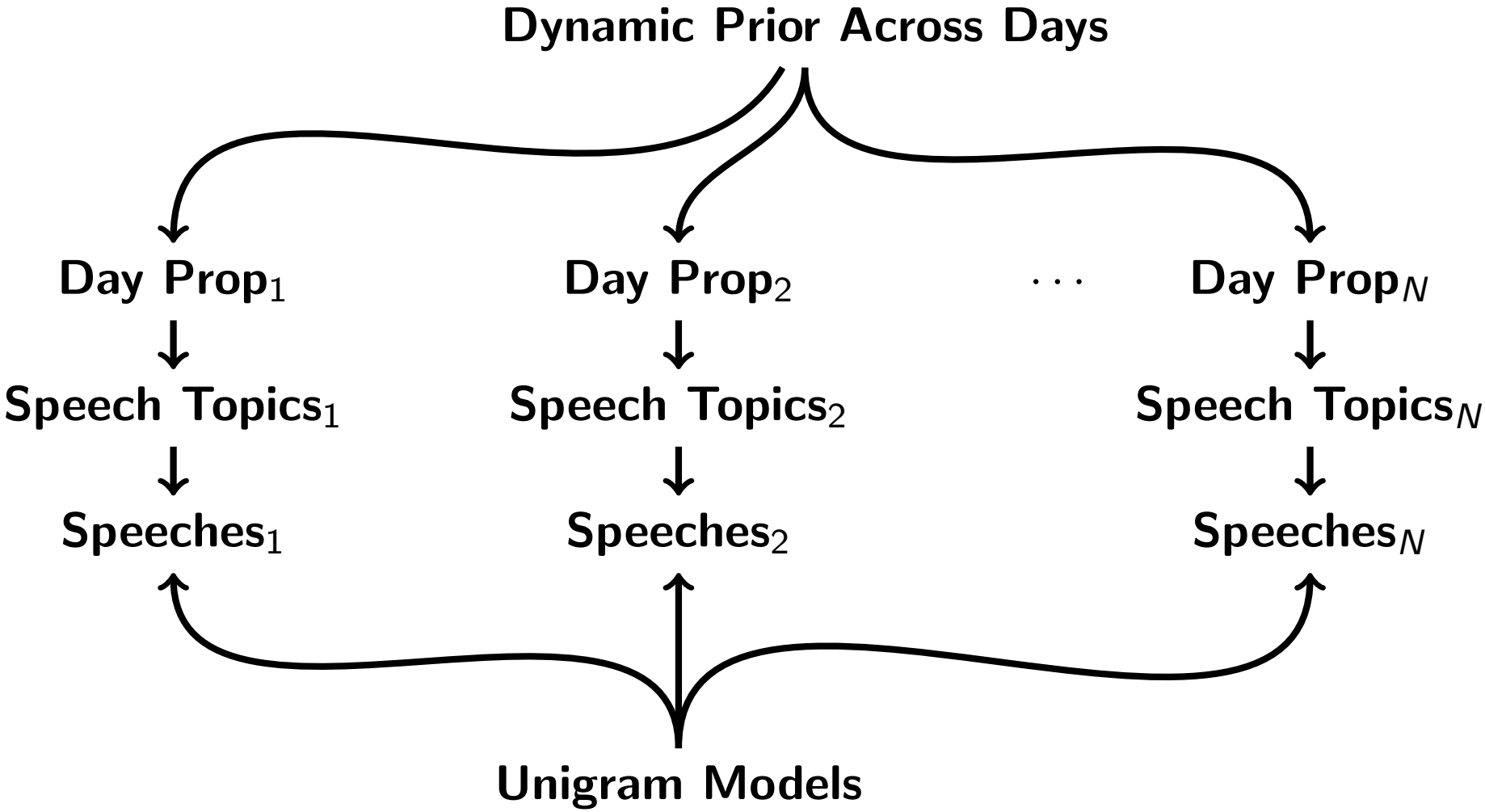
A General Hierarchical Structure

Dynamic Topic Model (Quinn et al 2010)



A General Hierarchical Structure

Dynamic Topic Model (Quinn et al 2010)



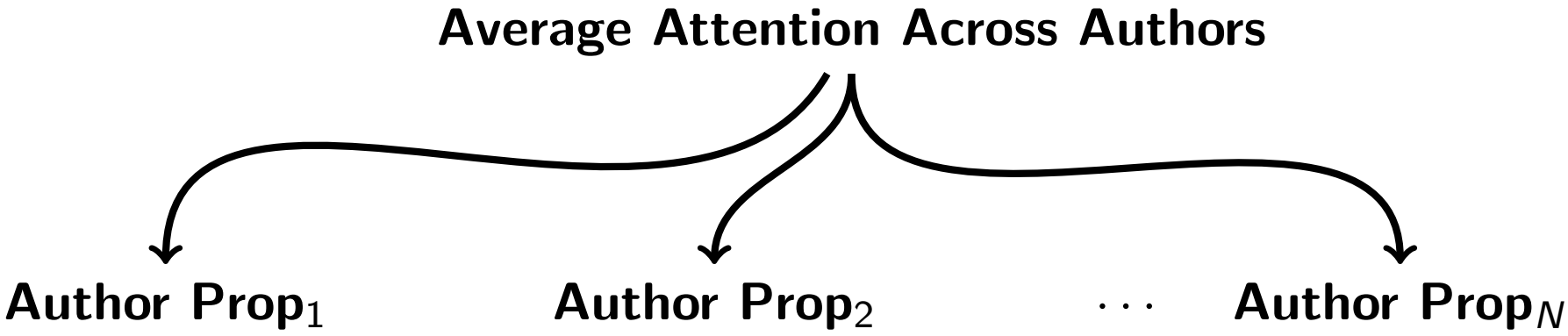
A General Hierarchical Structure

Expressed Agenda Model (Grimmer 2010)

Average Attention Across Authors

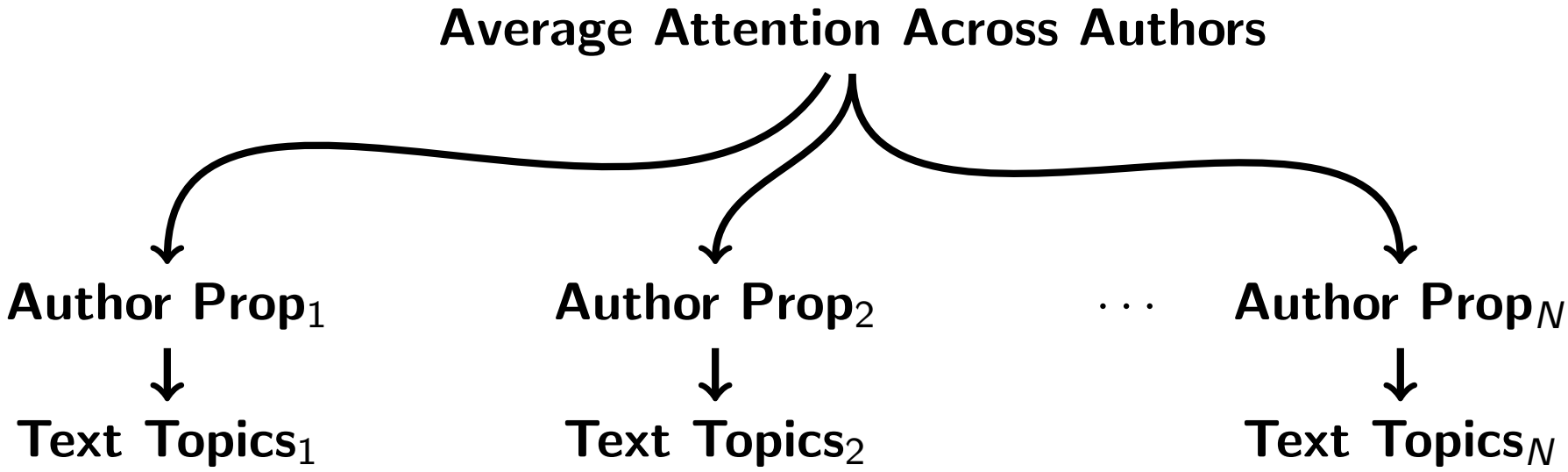
A General Hierarchical Structure

Expressed Agenda Model (Grimmer 2010)



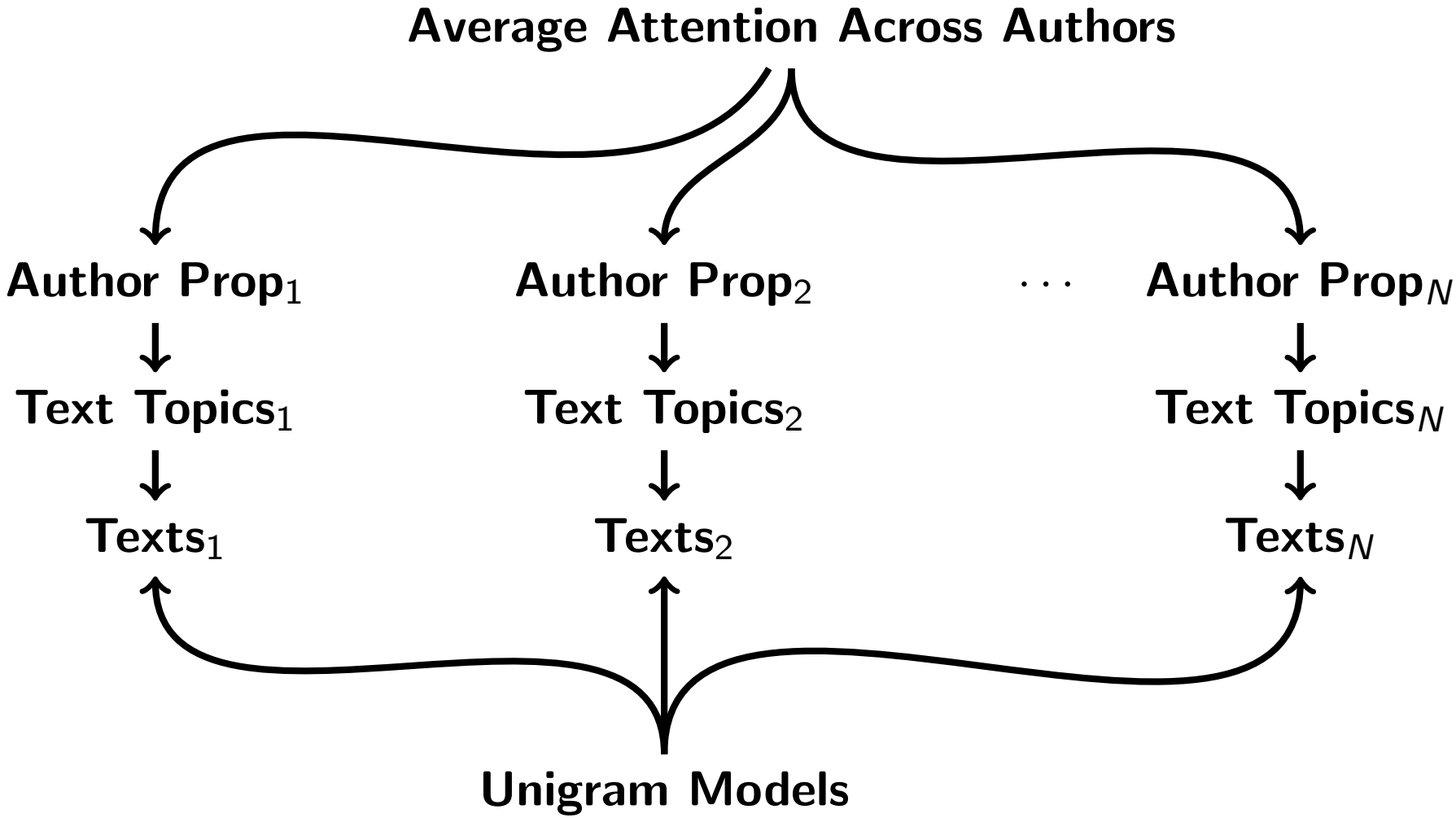
A General Hierarchical Structure

Expressed Agenda Model (Grimmer 2010)



A General Hierarchical Structure

Expressed Agenda Model (Grimmer 2010)



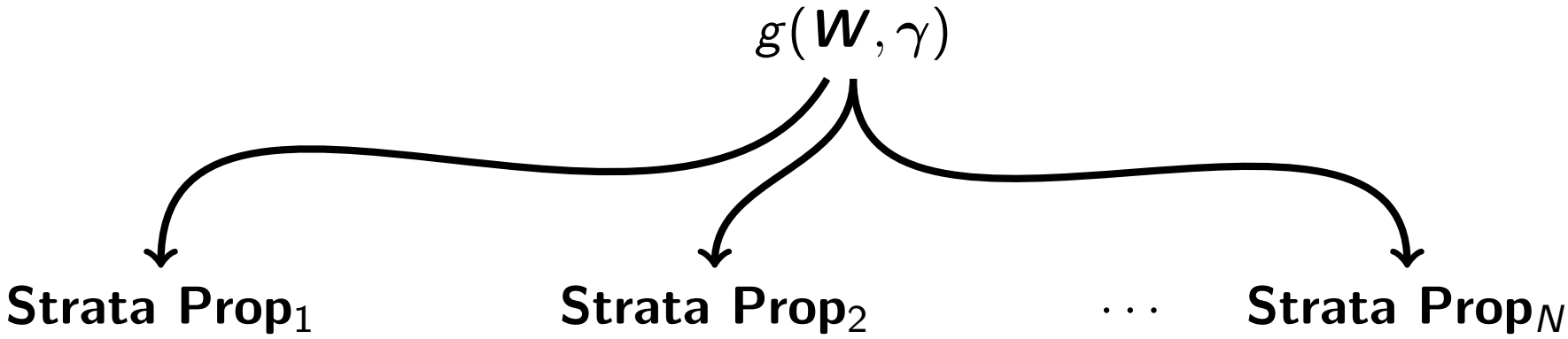
A General Hierarchical Structure

Structural Topic Model (Roberts, Stewart, Airoldi 2014)

$$g(\boldsymbol{W}, \gamma)$$

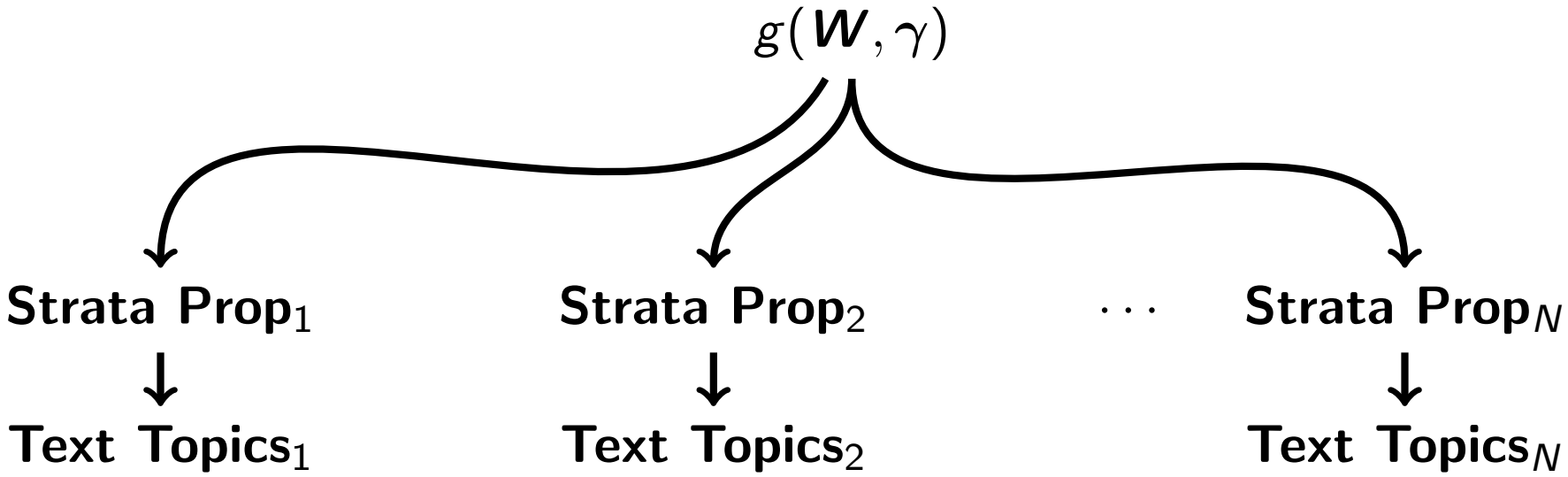
A General Hierarchical Structure

Structural Topic Model (Roberts, Stewart, Airoldi 2014)



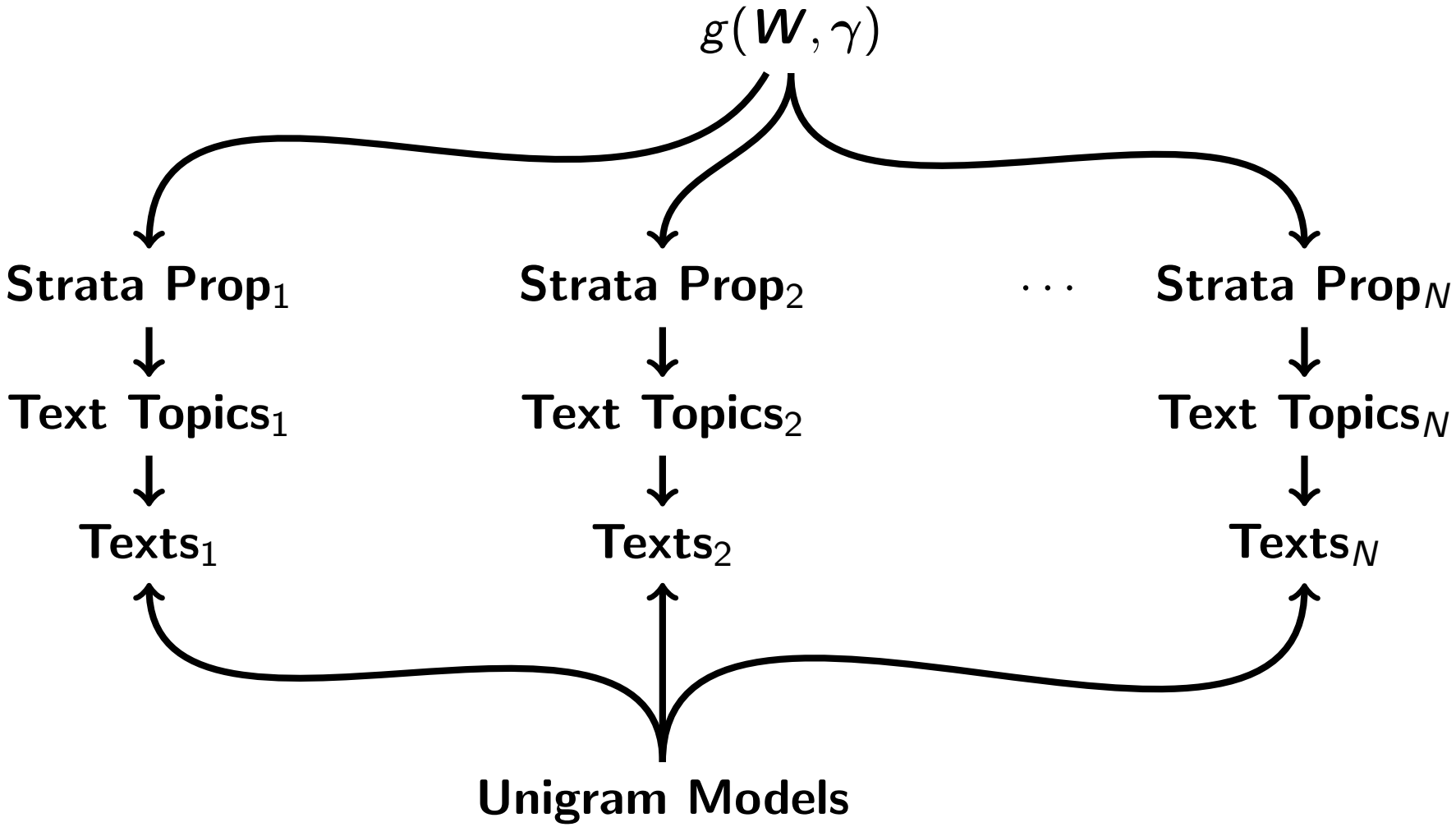
A General Hierarchical Structure

Structural Topic Model (Roberts, Stewart, Airoldi 2014)



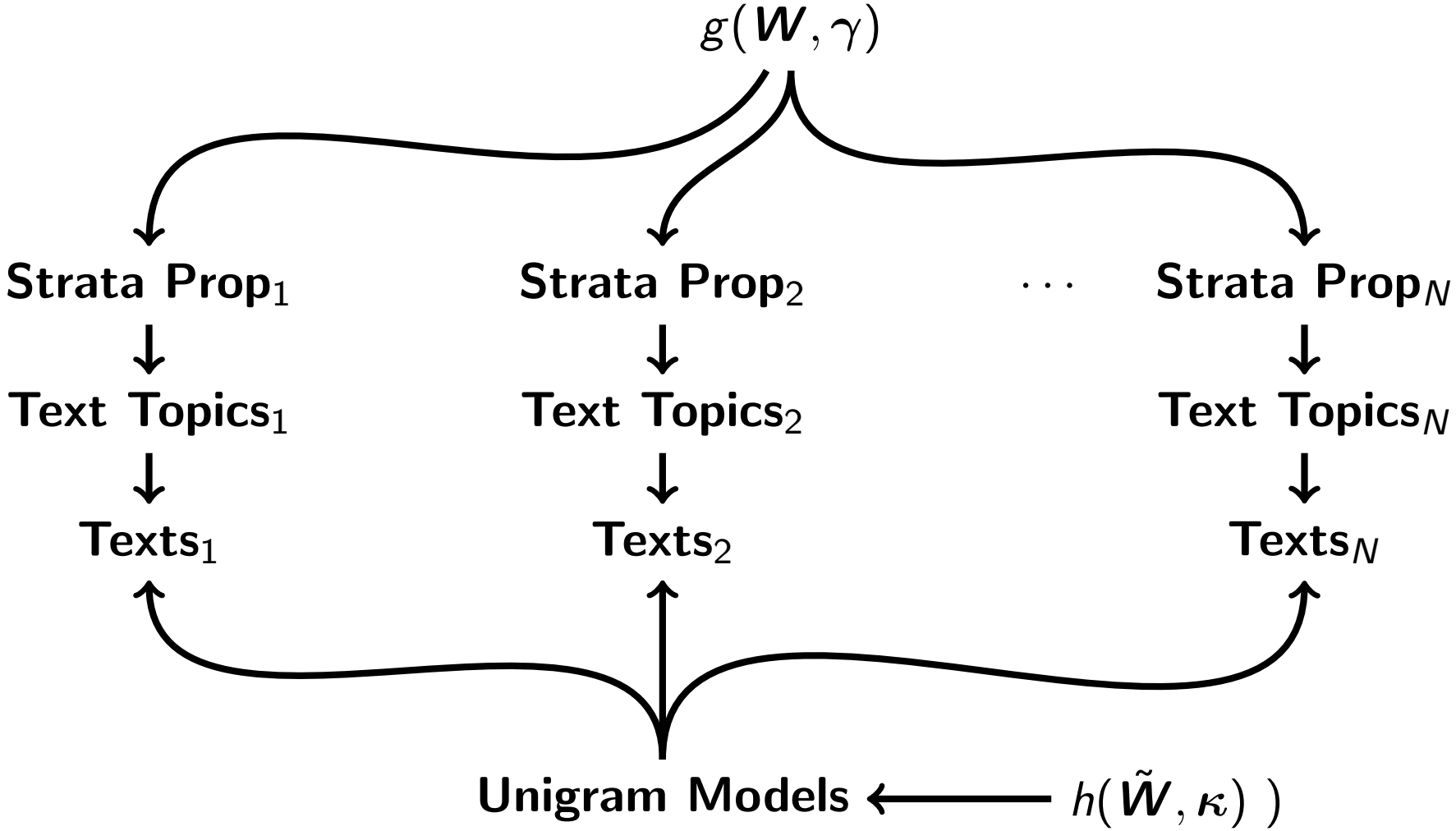
A General Hierarchical Structure

Structural Topic Model (Roberts, Stewart, Airoldi 2014)



A General Hierarchical Structure

Structural Topic Model (Roberts, Stewart, Airoldi 2014)



A General Hierarchical Structure

Conditioning on Unknown Covariates$\rightsquigarrow$ levels of mixtures at proportions (Grimmer 2013; Wallach 2008)

Mixture of Top. Attn. Models

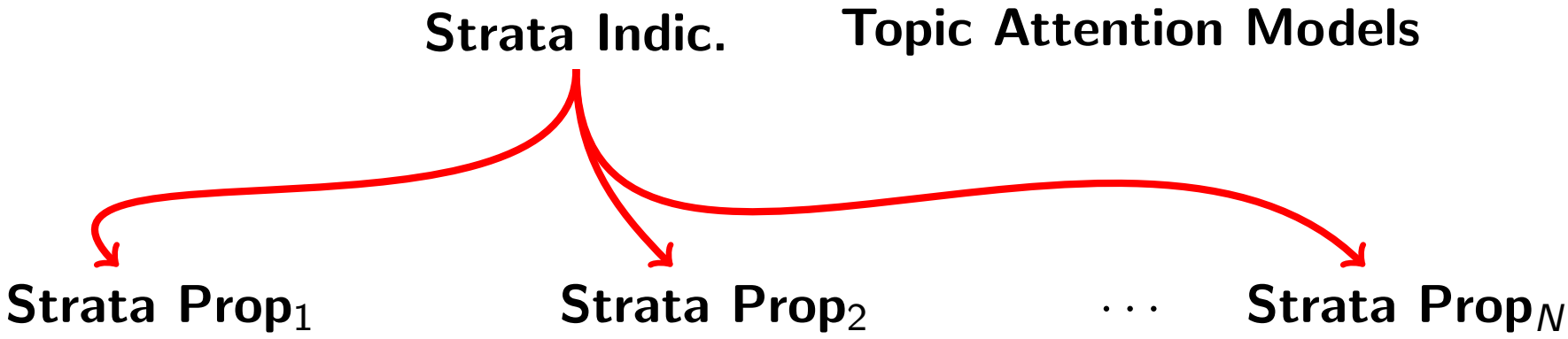
A General Hierarchical Structure

Conditioning on Unknown Covariates$\rightsquigarrow$ levels of mixtures at proportions (Grimmer 2013; Wallach 2008)

Strata Indic. **Topic Attention Models**

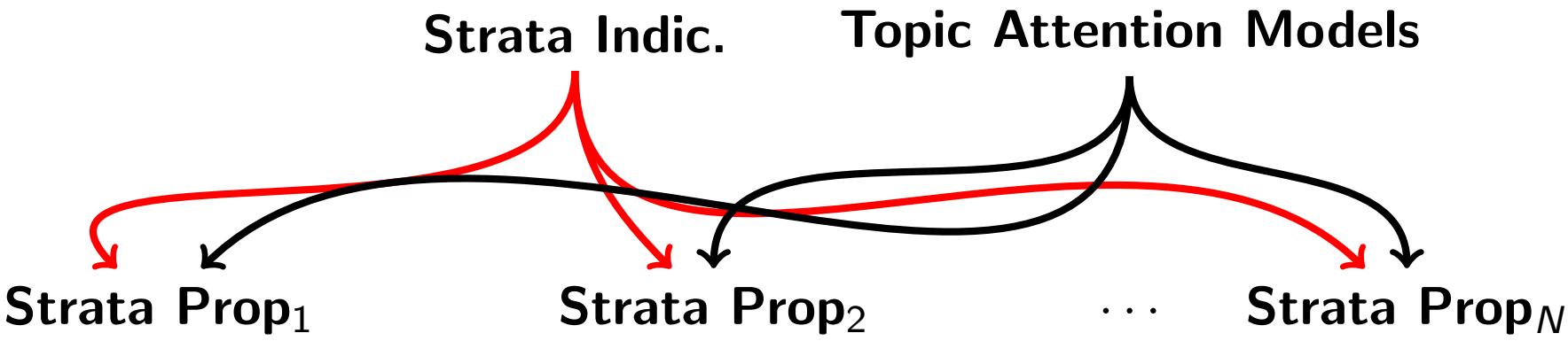
A General Hierarchical Structure

Conditioning on Unknown Covariates$\rightsquigarrow$ levels of mixtures at proportions (Grimmer 2013; Wallach 2008)



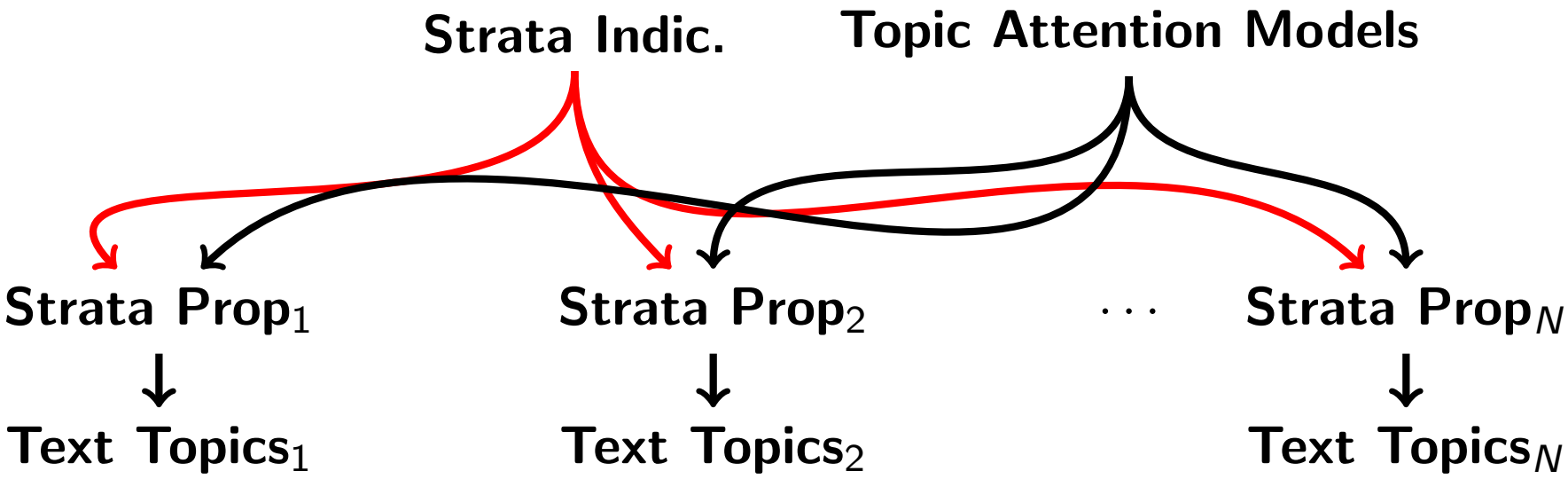
A General Hierarchical Structure

Conditioning on Unknown Covariates$\rightsquigarrow$ levels of mixtures at proportions (Grimmer 2013; Wallach 2008)



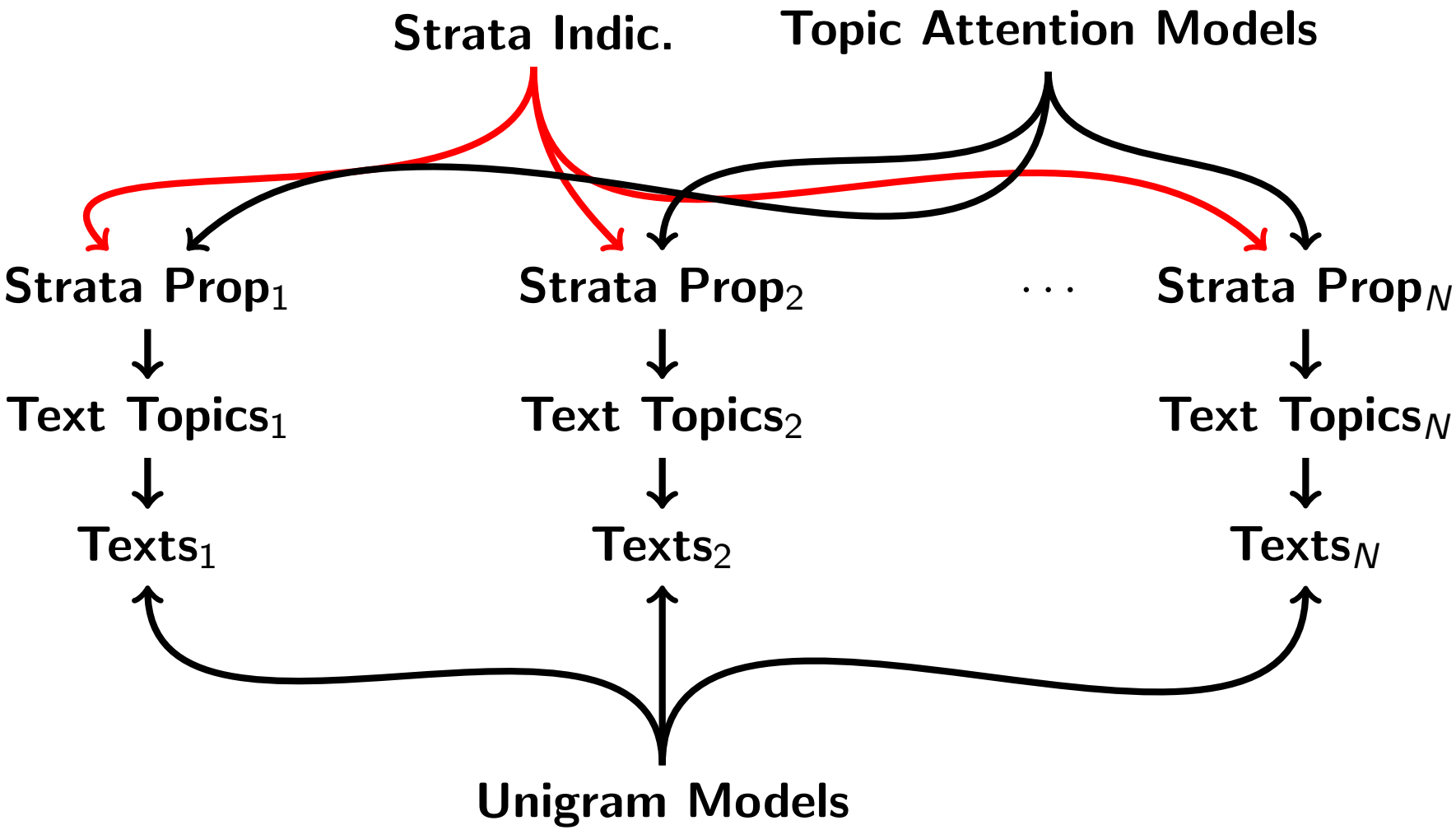
A General Hierarchical Structure

Conditioning on Unknown Covariates$\rightsquigarrow$ levels of mixtures at proportions (Grimmer 2013; Wallach 2008)



A General Hierarchical Structure

Conditioning on Unknown Covariates$\rightsquigarrow$ levels of mixtures at proportions (Grimmer 2013; Wallach 2008)



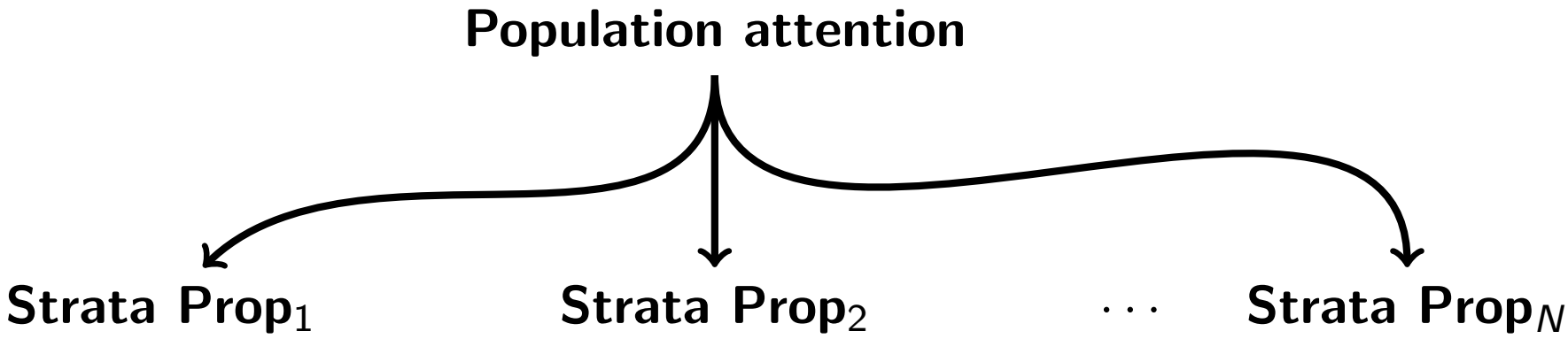
A General Hierarchical Structure

Conditioning on Unknown Covariates for Topics$\rightsquigarrow$ hierarchy of topics (Li and McCallum 2006; Blaydes, Grimmer, and McQueen 2017)

Population attention

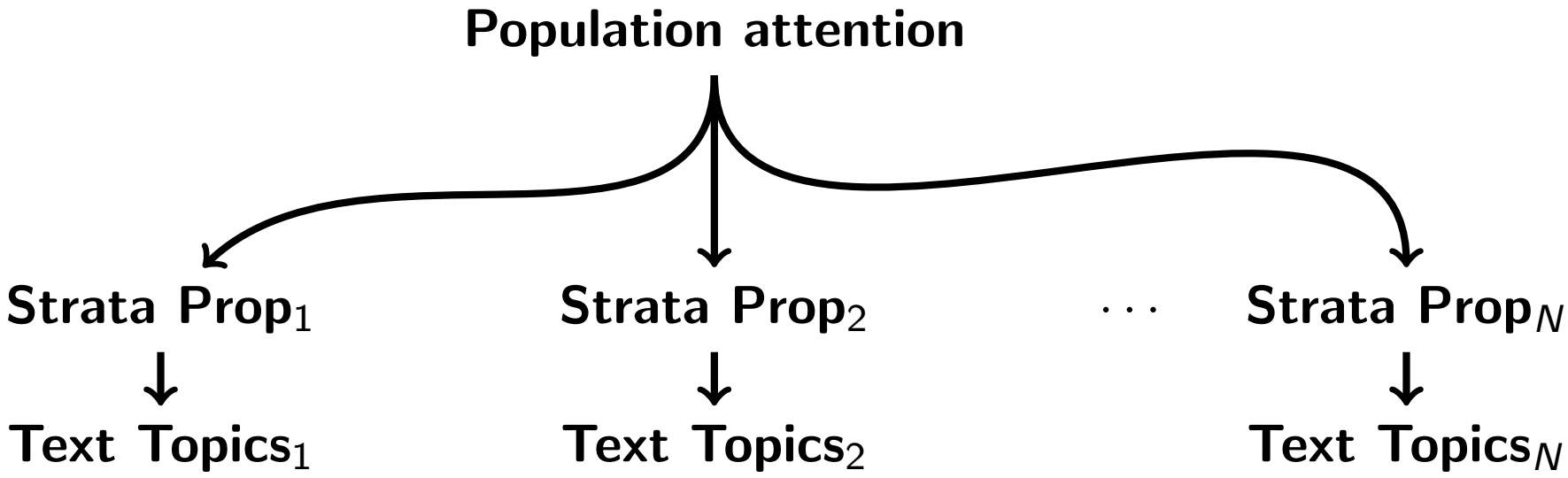
A General Hierarchical Structure

Conditioning on Unknown Covariates for Topics$\rightsquigarrow$ hierarchy of topics (Li and McCallum 2006; Blaydes, Grimmer, and McQueen 2017)



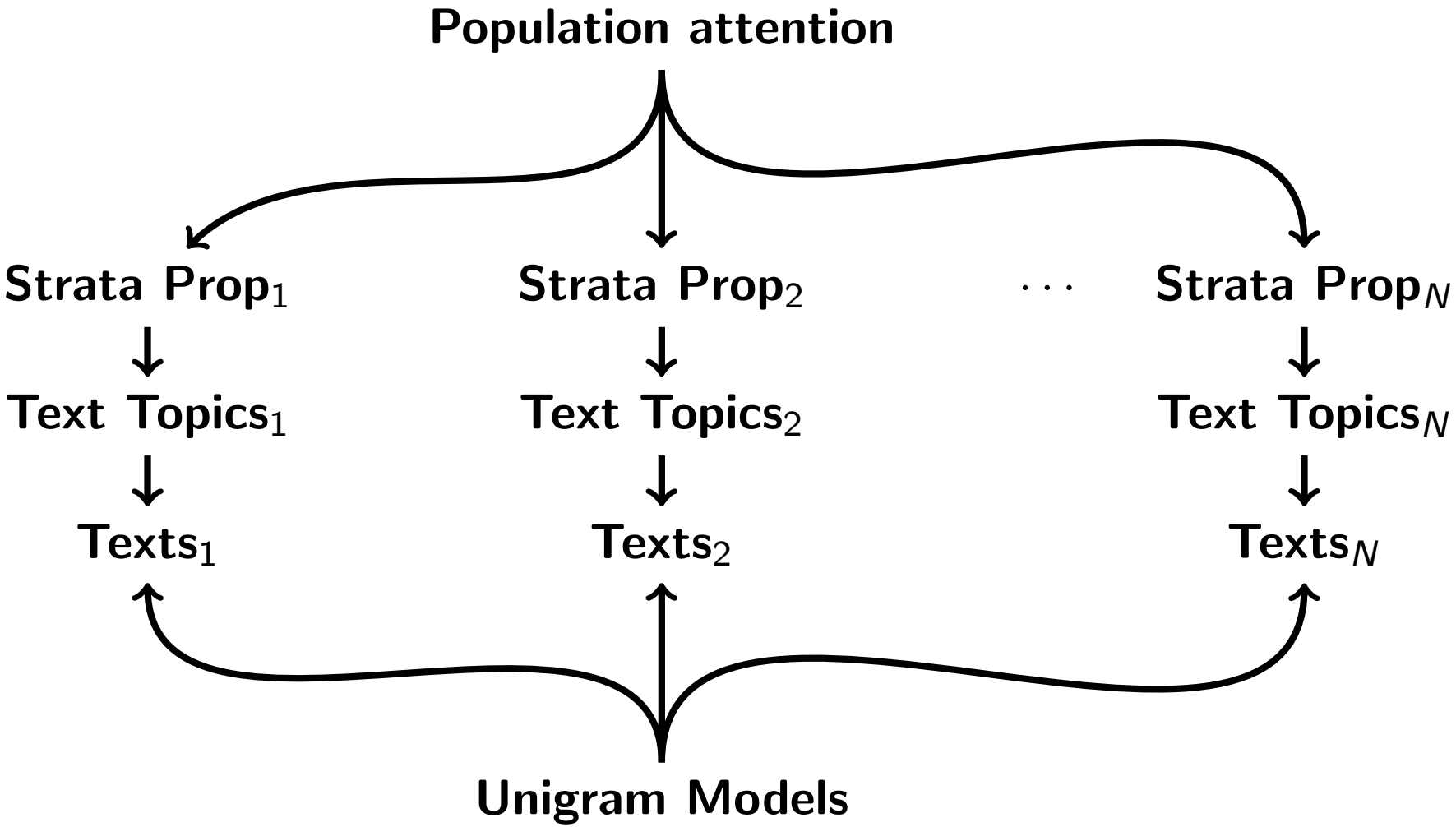
A General Hierarchical Structure

Conditioning on Unknown Covariates for Topics$\rightsquigarrow$ hierarchy of topics (Li and McCallum 2006; Blaydes, Grimmer, and McQueen 2017)



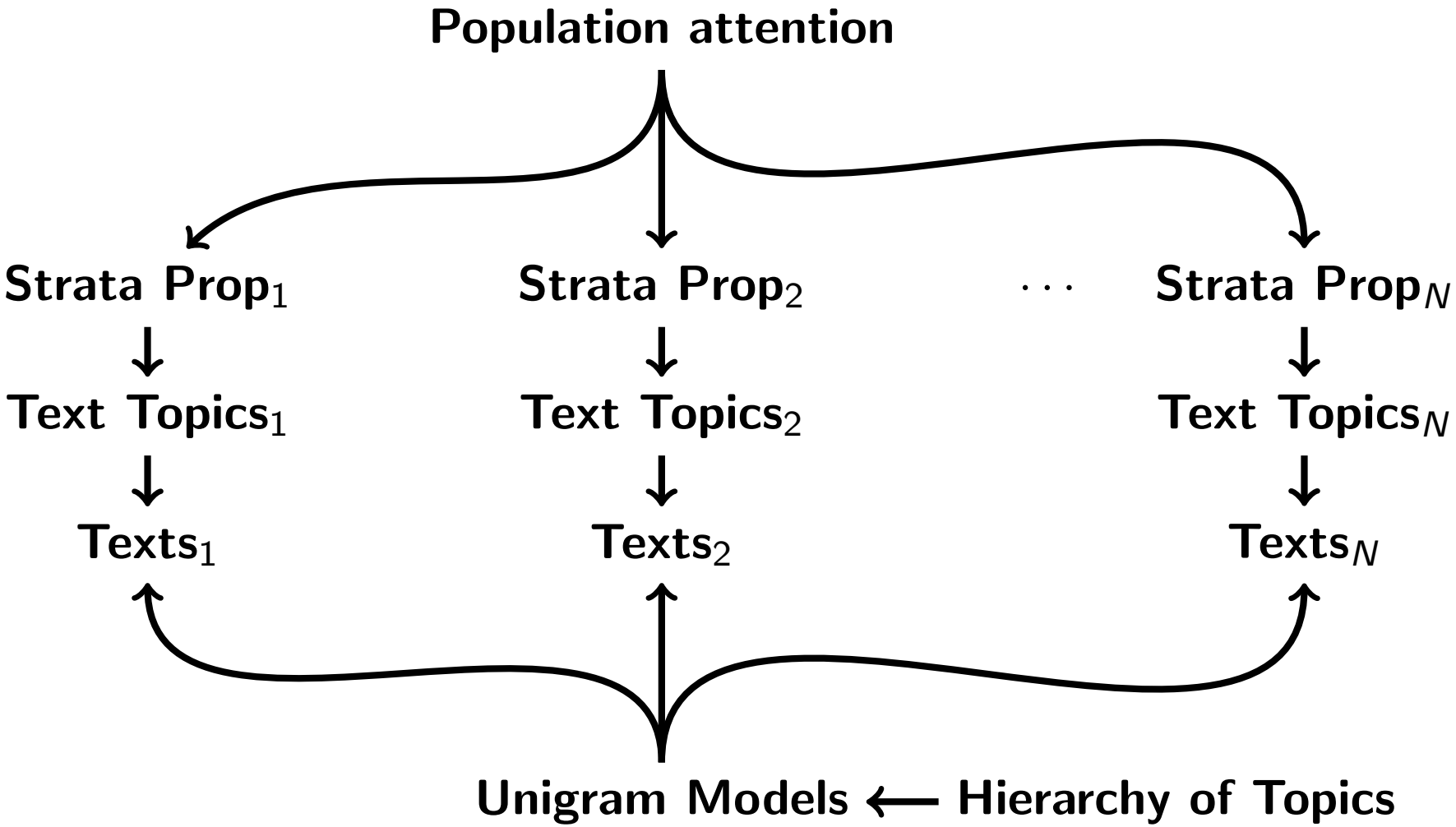
A General Hierarchical Structure

Conditioning on Unknown Covariates for Topics$\rightsquigarrow$ hierarchy of topics (Li and McCallum 2006; Blaydes, Grimmer, and McQueen 2017)



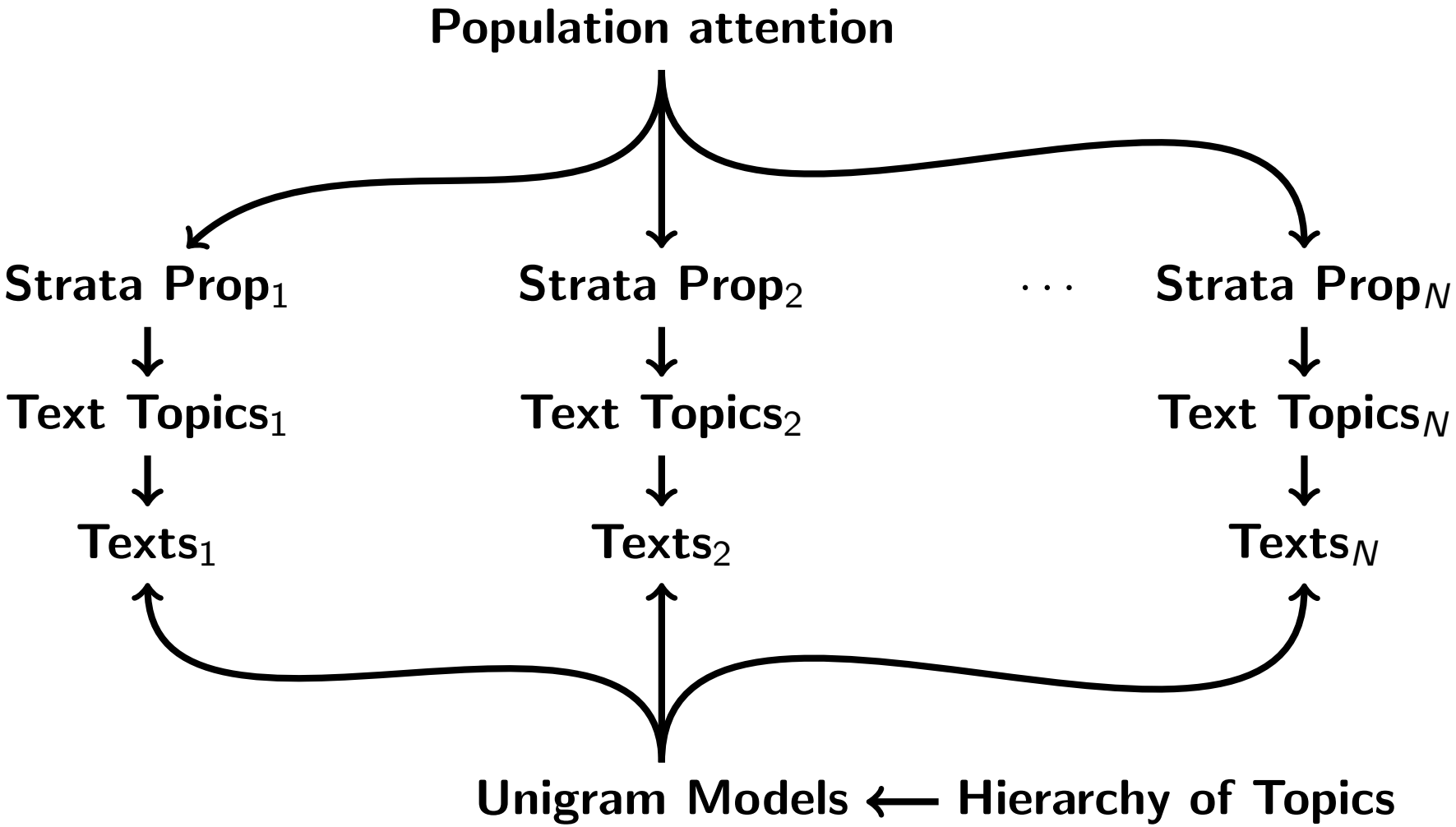
A General Hierarchical Structure

Conditioning on Unknown Covariates for Topics$\rightsquigarrow$ hierarchy of topics (Li and McCallum 2006; Blaydes, Grimmer, and McQueen 2017)



A General Hierarchical Structure

Conditioning on Unknown Covariates for Topics$\rightsquigarrow$ hierarchy of topics (Li and McCallum 2006; Blaydes, Grimmer, and McQueen 2017)



Why Encode Structure in Extensions of LDA?

Why Encode Structure in Extensions of LDA?

- Substantive reasons

Why Encode Structure in Extensions of LDA?

- Substantive reasons
    - Additional structure corresponds to substantively interesting content

Why Encode Structure in Extensions of LDA?

- Substantive reasons
  - Additional structure corresponds to substantively interesting content
  - Avoids potential ad-hoc secondary analysis

Why Encode Structure in Extensions of LDA?

- Substantive reasons
  - Additional structure corresponds to substantively interesting content
  - Avoids potential ad-hoc secondary analysis
  - Clear data generating process

Why Encode Structure in Extensions of LDA?

- Substantive reasons
  - Additional structure corresponds to substantively interesting content
  - Avoids potential ad-hoc secondary analysis
  - Clear data generating process
- Statistical reasons

Why Encode Structure in Extensions of LDA?

- Substantive reasons
  - Additional structure corresponds to substantively interesting content
  - Avoids potential ad-hoc secondary analysis
  - Clear data generating process
- Statistical reasons
  - Smoothing$\rightsquigarrow$ borrow information across groups intelligently

Why Encode Structure in Extensions of LDA?

- Substantive reasons
  - Additional structure corresponds to substantively interesting content
  - Avoids potential ad-hoc secondary analysis
  - Clear data generating process
- Statistical reasons
  - Smoothing$\rightsquigarrow$ borrow information across groups intelligently
  - Uncertainty$\rightsquigarrow$ potential for better uncertainty estimates

Why Encode Structure in Extensions of LDA?

- Substantive reasons
  - Additional structure corresponds to substantively interesting content
  - Avoids potential ad-hoc secondary analysis
  - Clear data generating process
- Statistical reasons
  - Smoothing$\rightsquigarrow$ borrow information across groups intelligently
  - Uncertainty$\rightsquigarrow$ potential for better uncertainty estimates
  - Improved topics$\rightsquigarrow$ small word conditions, structure could help

Plan for the Class

1) Discuss model with unknown covariates for strata proportions$\rightsquigarrow$ presentational style
2) Discuss model with hierarchy of topics$\rightsquigarrow$ mirrors genre

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:
Senators (representatives) regularly engage the public $\rightarrow$ presentational style
But we know little about this engagement

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:
Senators (representatives) regularly engage the public $\rightarrow$ presentational style
But we know little about this engagement
Why? Hard to Measure

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:
Senators (representatives) regularly engage the public $\rightarrow$ presentational style
But we know little about this engagement
Why? Hard to Measure
Describe model that facilitates estimation of presentational styles in Senate press releases

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:
Senators (representatives) regularly engage the public $\rightarrow$ presentational style
But we know little about this engagement
Why? Hard to Measure
Describe model that facilitates estimation of presentational styles in Senate press releases

- Characterize representation provided to constituents

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:
Senators (representatives) regularly engage the public $\rightarrow$ presentational style
But we know little about this engagement
Why? Hard to Measure
Describe model that facilitates estimation of presentational styles in Senate press releases

- Characterize representation provided to constituents
- Divide attention over a set of topics

Unknown Covariates for Issue Attention: Measuring Attention in Senate Press Releases

Substantive problem:
Senators (representatives) regularly engage the public $\rightarrow$ presentational style
But we know little about this engagement
Why? Hard to Measure
Describe model that facilitates estimation of presentational styles in Senate press releases

- Characterize representation provided to constituents
- Divide attention over a set of topics
- Given attention to topics, write press releases

Presentational Styles$\rightsquigarrow$ Objective Function

- $\pi_{itk} \equiv$ Attention senator i allocates to issue k in year t
- $\pi_{itk} \equiv$ Probability press release is about issue k
- $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it44})$

Presentational Styles$\rightsquigarrow$ Objective Function

- $\pi_{itk} \equiv$ Attention senator i allocates to issue k in year t
- $\pi_{itk} \equiv$ Probability press release is about issue k
- $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it44})$

Press release-level parameters (press release j from senator i in year t)

Presentational Styles$\rightsquigarrow$ Objective Function

- $\pi_{itk} \equiv$ Attention senator i allocates to issue k in year t
- $\pi_{itk} \equiv$ Probability press release is about issue k
- $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it44})$

Press release-level parameters (press release j from senator i in year t)

- Assume: Each press release j assigned to one topic.
- Let $\boldsymbol{\tau}_{ijt}$ indicate press release j's topic.

Presentational Styles$\rightsquigarrow$ Objective Function

- $\pi_{itk} \equiv$ Attention senator i allocates to issue k in year t
- $\pi_{itk} \equiv$ Probability press release is about issue k
- $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it44})$

Press release-level parameters (press release j from senator i in year t)

- Assume: Each press release j assigned to one topic.
- Let $\boldsymbol{\tau}_{ijt}$ indicate press release j's topic.

$$\boldsymbol{\tau}_{ijt} \sim \text{Multinomial}(1, \boldsymbol{\pi}_{it})$$

Presentational Styles$\rightsquigarrow$ Objective Function

- $\pi_{itk} \equiv$ Attention senator i allocates to issue k in year t
- $\pi_{itk} \equiv$ Probability press release is about issue k
- $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it44})$

Press release-level parameters (press release j from senator i in year t)

- Assume: Each press release j assigned to one topic.
- Let $\boldsymbol{\tau}_{ijt}$ indicate press release j's topic.

$$\boldsymbol{\tau}_{ijt} \sim \text{Multinomial}(1, \boldsymbol{\pi}_{it})$$

- Conditional on topic, draw document's content.

Presentational Styles$\rightsquigarrow$ Objective Function

- $\pi_{itk} \equiv$ Attention senator i allocates to issue k in year t
- $\pi_{itk} \equiv$ Probability press release is about issue k
- $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it44})$

Press release-level parameters (press release j from senator i in year t)

- Assume: Each press release j assigned to one topic.
- Let $\boldsymbol{\tau}_{ijt}$ indicate press release j's topic.

$$\boldsymbol{\tau}_{ijt} \sim \text{Multinomial}(1, \boldsymbol{\pi}_{it})$$

- Conditional on topic, draw document's content.
- If $\tau_{ijtk} = 1$ then

$$\boldsymbol{x}_{ijt} \sim \text{Multinomial}(n_{ijt}, \boldsymbol{\theta}_k).$$

Priors

Each π_{it} is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions

Priors

Each π_{it} is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions .

$$\boldsymbol{\sigma}_{it} \sim \text{Multinomial}(1, \boldsymbol{\beta}).$$

Priors

Each $\boldsymbol{\pi}_{it}$ is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions .

$$\begin{aligned} \boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}). \\ \boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \end{aligned}$$

Priors

Each $\boldsymbol{\pi}_{it}$ is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions .

$$\begin{aligned} \boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}). \\ \boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\ \alpha_{ks} &\sim \text{Gamma}(0.25, 1) \end{aligned}$$

Priors

Each $\boldsymbol{\pi}_{it}$ is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions .

$$\begin{aligned} \boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}). \\ \boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\ \alpha_{ks} &\sim \text{Gamma}(0.25, 1) \end{aligned}$$

Other priors:

Priors

Each $\boldsymbol{\pi}_{it}$ is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions .

$$\begin{aligned} \boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}). \\ \boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\ \alpha_{ks} &\sim \text{Gamma}(0.25, 1) \end{aligned}$$

Other priors:

$$\boldsymbol{\theta}_k \sim \text{Multinomial}(\boldsymbol{\lambda})$$

Priors

Each $\boldsymbol{\pi}_{it}$ is a draw from one-of-S styles$\rightsquigarrow$ mixture of Dirichlet distributions .

$$\begin{aligned} \boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}). \\ \boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\ \alpha_{ks} &\sim \text{Gamma}(0.25, 1) \end{aligned}$$

Other priors:

$$\begin{aligned} \boldsymbol{\theta}_k &\sim \text{Multinomial}(\boldsymbol{\lambda}) \\ \boldsymbol{\beta} &\sim \text{Multinomial}(\mathbf{1}) \end{aligned}$$

Presentational Styles$\rightsquigarrow$ Objective Function

Presentational Styles$\rightsquigarrow$ Objective Function

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \\
\alpha_{ks} &\sim \text{Gamma}(0.25, 1)
\end{aligned}
$$

Presentational Styles$\rightsquigarrow$ Objective Function

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \\
\alpha_{ks} &\sim \text{Gamma}(0.25, 1) \\
\boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta})
\end{aligned}
$$

Presentational Styles$\rightsquigarrow$ Objective Function

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \\
\alpha_{ks} &\sim \text{Gamma}(0.25, 1) \\
\boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}) \\
\boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s)
\end{aligned}
$$

Presentational Styles$\rightsquigarrow$ Objective Function

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \\
\alpha_{ks} &\sim \text{Gamma}(0.25, 1) \\
\boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}) \\
\boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\
\boldsymbol{\tau}_{ijt} | \boldsymbol{\pi}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\pi}_{it})
\end{aligned}
$$

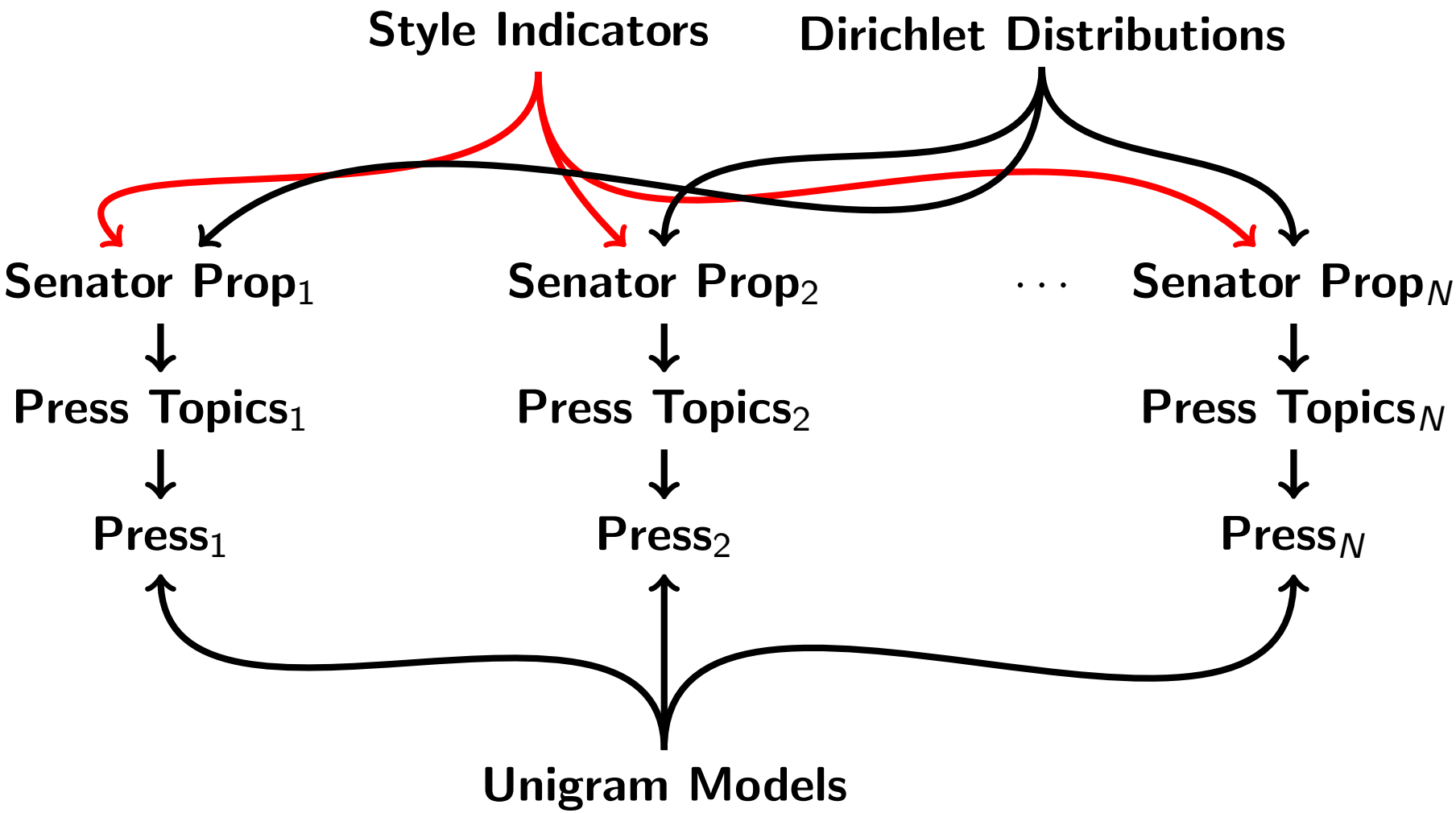
Presentational Styles$\rightsquigarrow$ Objective Function

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \\
\alpha_{ks} &\sim \text{Gamma}(0.25, 1) \\
\boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}) \\
\boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\
\boldsymbol{\tau}_{ijt} | \boldsymbol{\pi}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\pi}_{it}) \\
\boldsymbol{x}_{ijt} | \tau_{ijtk} = 1, \boldsymbol{\theta}_k &\sim \text{Multinomial}(n_{ijt}, \boldsymbol{\theta}_k)
\end{aligned}
$$

Presentational Styles$\rightsquigarrow$ Objective Function

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \\
\alpha_{ks} &\sim \text{Gamma}(0.25, 1) \\
\boldsymbol{\sigma}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\beta}) \\
\boldsymbol{\pi}_{it} | \sigma_{its} = 1, \boldsymbol{\alpha}_s &\sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \\
\boldsymbol{\tau}_{ijt} | \boldsymbol{\pi}_{it} &\sim \text{Multinomial}(1, \boldsymbol{\pi}_{it}) \\
\boldsymbol{x}_{ijt} | \tau_{ijtk} = 1, \boldsymbol{\theta}_k &\sim \text{Multinomial}(n_{ijt}, \boldsymbol{\theta}_k)
\end{aligned}
$$

Mixture of Styles, Mixture of Topics



Posterior:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\tau} | \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{s=1}^{S} \frac{\exp(-\frac{\alpha_{ks}}{1/4})}{1/4} \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times$$

$$\prod_{i=1}^{N} \prod_{t=2005}^{2007} \prod_{s=1}^{S} \left[\beta_s \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ks})}{\prod_{k=1}^{K} \Gamma(\alpha_{ks})} \prod_{k=1}^{K} \pi_{itk}^{\alpha_{ks} - 1} \prod_{j=1}^{D_{it}} \prod_{k=1}^{K} \left[\pi_{itk} \prod_{w=1}^{W} \theta_{kw}^{x_{ijtw}} \right]^{\tau_{ijtk}} \right]^{\sigma_{its}}$$

Posterior:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\tau} | \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{s=1}^{S} \frac{\exp(-\frac{\alpha_{ks}}{1/4})}{1/4} \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times$$

$$\prod_{i=1}^{N} \prod_{t=2005}^{2007} \prod_{s=1}^{S} \left[\beta_s \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ks})}{\prod_{k=1}^{K} \Gamma(\alpha_{ks})} \prod_{k=1}^{K} \pi_{itk}^{\alpha_{ks} - 1} \prod_{j=1}^{D_{it}} \prod_{k=1}^{K} \left[\pi_{itk} \prod_{w=1}^{W} \theta_{kw}^{x_{ijtw}} \right]^{\tau_{ijtk}} \right]^{\sigma_{its}}$$

1) Estimate with Variational Approximation

Posterior:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\tau} | \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{s=1}^{S} \frac{\exp(-\frac{\alpha_{ks}}{1/4})}{1/4} \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times$$
$$\prod_{i=1}^{N} \prod_{t=2005}^{2007} \prod_{s=1}^{S} \left[\beta_s \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ks})}{\prod_{k=1}^{K} \Gamma(\alpha_{ks})} \prod_{k=1}^{K} \pi_{itk}^{\alpha_{ks} - 1} \prod_{j=1}^{D_{it}} \prod_{k=1}^{K} \left[\pi_{itk} \prod_{w=1}^{W} \theta_{kw}^{x_{ijtw}} \right]^{\tau_{ijtk}} \right]^{\sigma_{its}}$$

1) Estimate with Variational Approximation
2) Determining number of clusters at top? (Grimmer, Shorey, Wallach, and Zlotnick, In Progress)

Posterior:

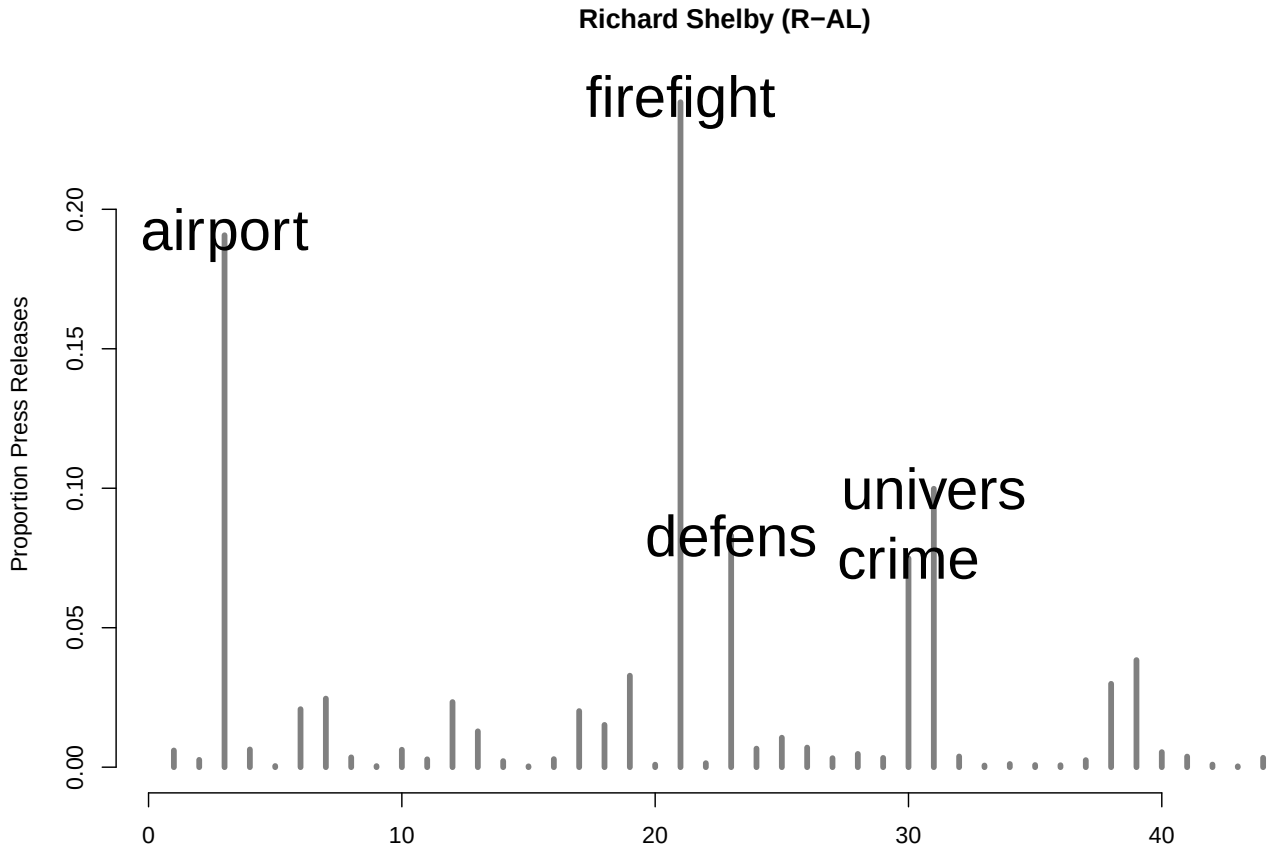
$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\tau} | \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{s=1}^{S} \frac{\exp(-\frac{\alpha_{ks}}{1/4})}{1/4} \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times$$
$$\prod_{i=1}^{N} \prod_{t=2005}^{2007} \prod_{s=1}^{S} \left[\beta_s \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ks})}{\prod_{k=1}^{K} \Gamma(\alpha_{ks})} \prod_{k=1}^{K} \pi_{itk}^{\alpha_{ks} - 1} \prod_{j=1}^{D_{it}} \prod_{k=1}^{K} \left[\pi_{itk} \prod_{w=1}^{W} \theta_{kw}^{x_{ijtw}} \right]^{\tau_{ijtk}} \right]^{\sigma_{its}}$$

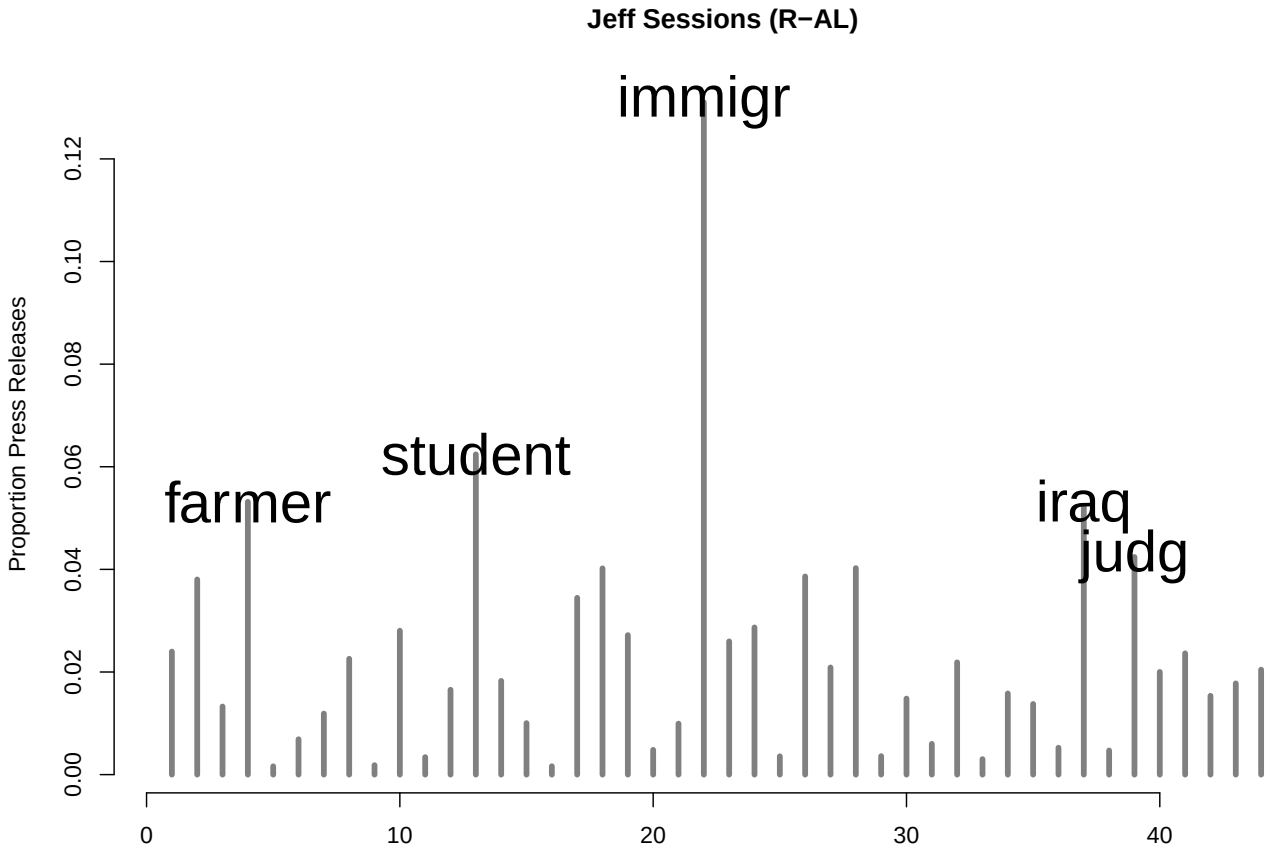
1) Estimate with Variational Approximation
2) Determining number of clusters at top? (Grimmer, Shorey, Wallach, and Zlotnick, In Progress)
    - Non-parametric model$\rightsquigarrow$ statistical selection

Posterior:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\tau} | \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{s=1}^{S} \frac{\exp(-\frac{\alpha_{ks}}{1/4})}{1/4} \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times$$
$$\prod_{i=1}^{N} \prod_{t=2005}^{2007} \prod_{s=1}^{S} \left[\beta_s \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ks})}{\prod_{k=1}^{K} \Gamma(\alpha_{ks})} \prod_{k=1}^{K} \pi_{itk}^{\alpha_{ks} - 1} \prod_{j=1}^{D_{it}} \prod_{k=1}^{K} \left[\pi_{itk} \prod_{w=1}^{W} \theta_{kw}^{x_{ijtw}} \right]^{\tau_{ijtk}} \right]^{\sigma_{its}}$$

1) Estimate with Variational Approximation
2) Determining number of clusters at top? (Grimmer, Shorey, Wallach, and Zlotnick, In Progress)
   - Non-parametric model$\rightsquigarrow$ statistical selection
   - Experiments/Coding Exercises to assess


Richard Shelby (R−AL)
firefight
airport
univers
defens
crime
Proportion Press Releases
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40


Jeff Sessions (R−AL)
immigr
student
farmer
iraq
judg
Proportion Press Releases
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30
40

Notions of validity: From Quinn, Monroe, et al (2010)

- Semantic Validity: All categories are coherent and meaningful
- Convergent Construct Validity: Measures concur with existing measures in critical details.
- Discriminant Construct Validity: Measures differ from existing measures in productive ways.
- Predictive Measure: Measures from the model corresponds to external events in expected ways.
- Hypothesis Validity: Measures generated from the model can be used to test substantive hypotheses.

To establish utility of new measures, demonstrate variety of validations
None of these validations are performed using a canned statistic
All: require substantive knowledge on areas (and what we expect!) [

Home Style Measures, Semantic Validity

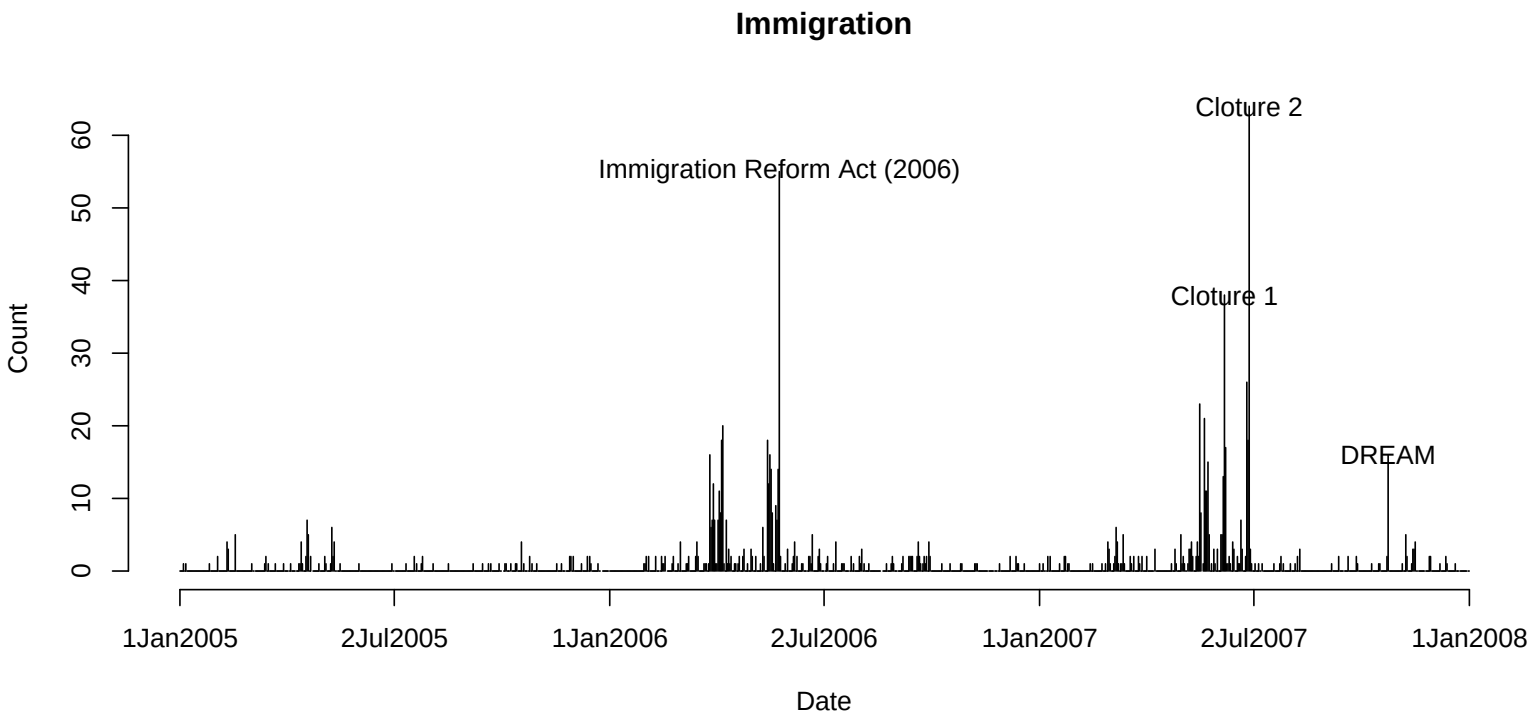
Must: Demonstrate to reader that topics are coherent and semantically meaningful

Description	Stems	%
Honorary	honor,prayer,rememb,fund,tribut	5.0
Transp. Grants	airport,transport,announc,urban,hud	4.8
Iraq	iraq,iraqi,troop,war,sectarian	4.7
DHS Policy	homeland,port,terrorist,dh,fema	4.1
Judicial Nom.	judg,court,suprem,nomin,nomine	3.8
Fire Dept. Grant	firefight,homeland,afgp,award,equip	3.7

How: examples in text are also useful.

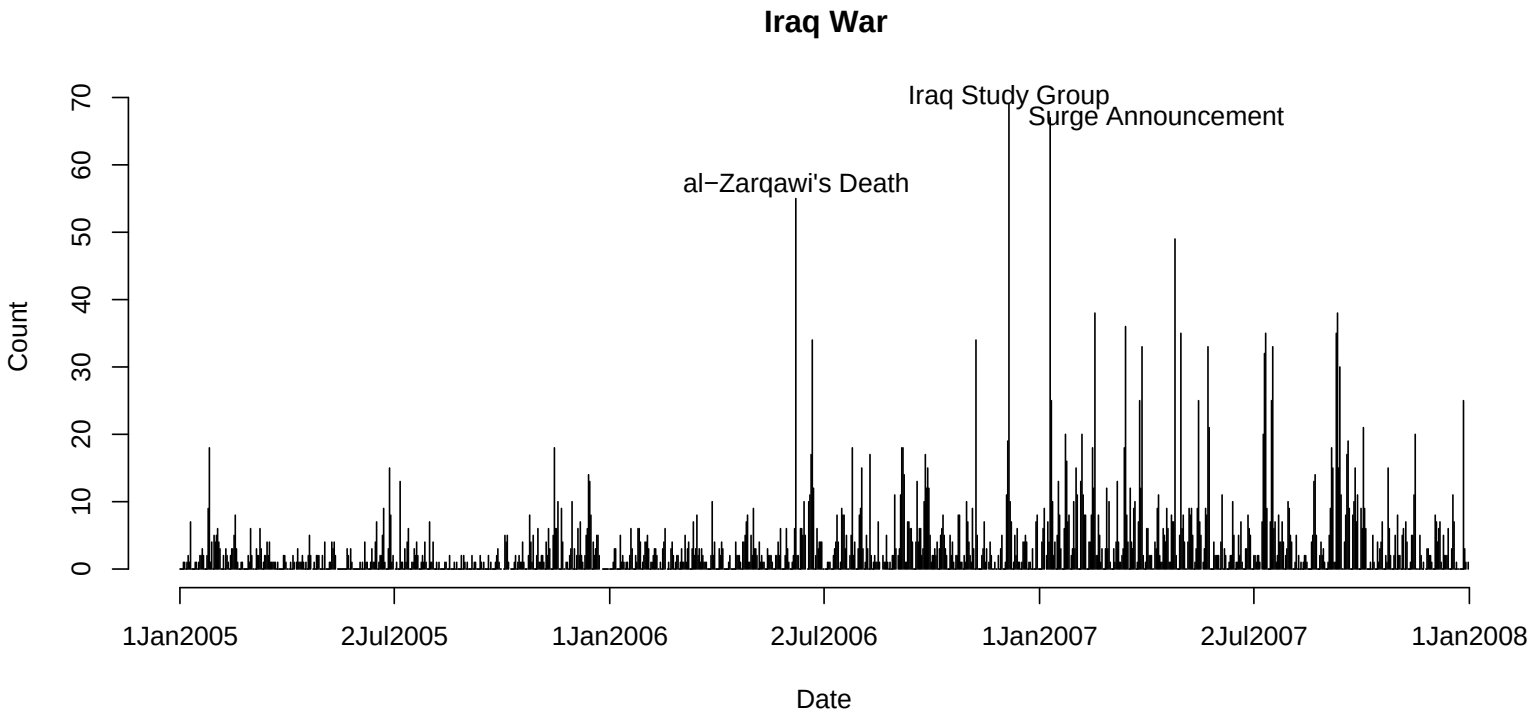
Home Style Measures, Convergent Validity

Over time variation



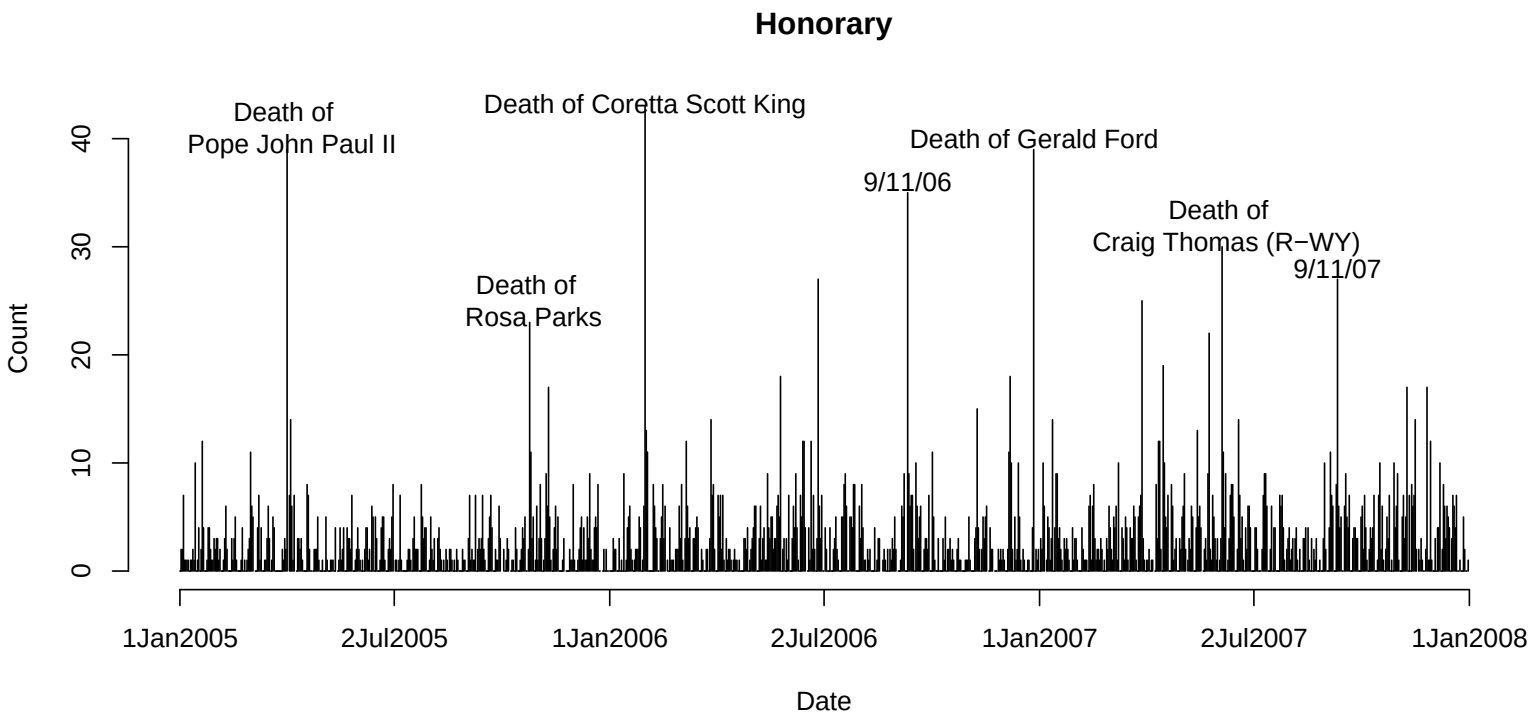
Home Style Measures, Convergent Validity

Over time variation



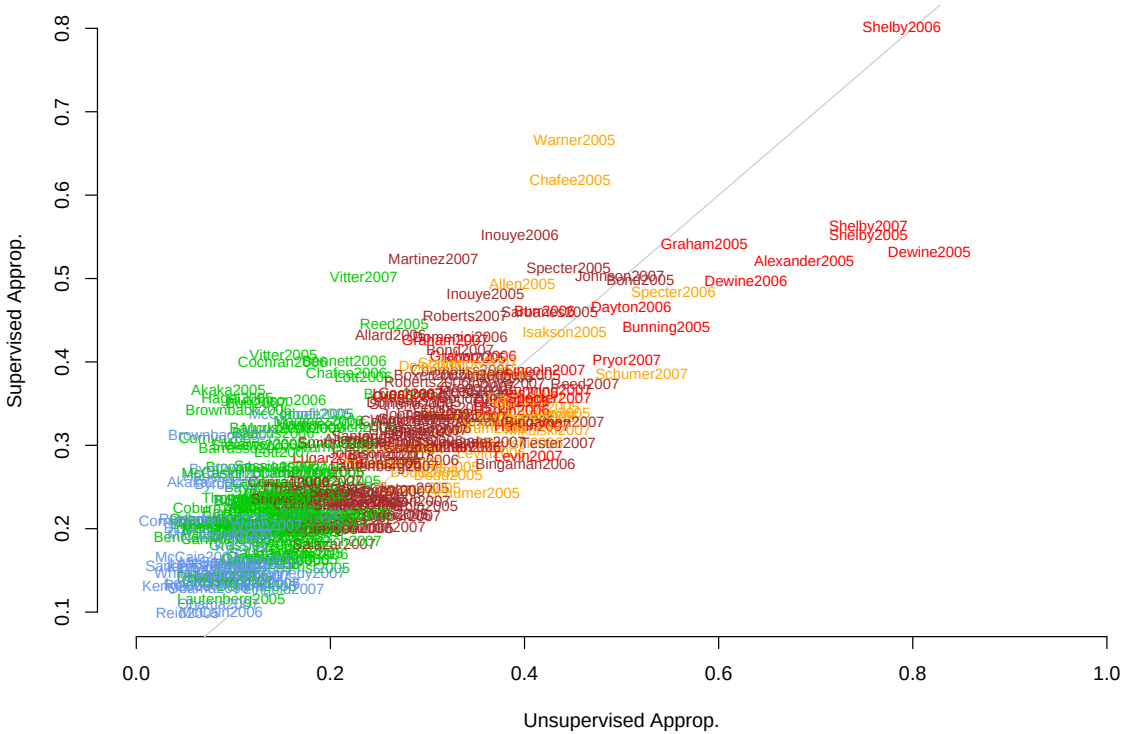
Home Style Measures, Convergent Validity

Over time variation

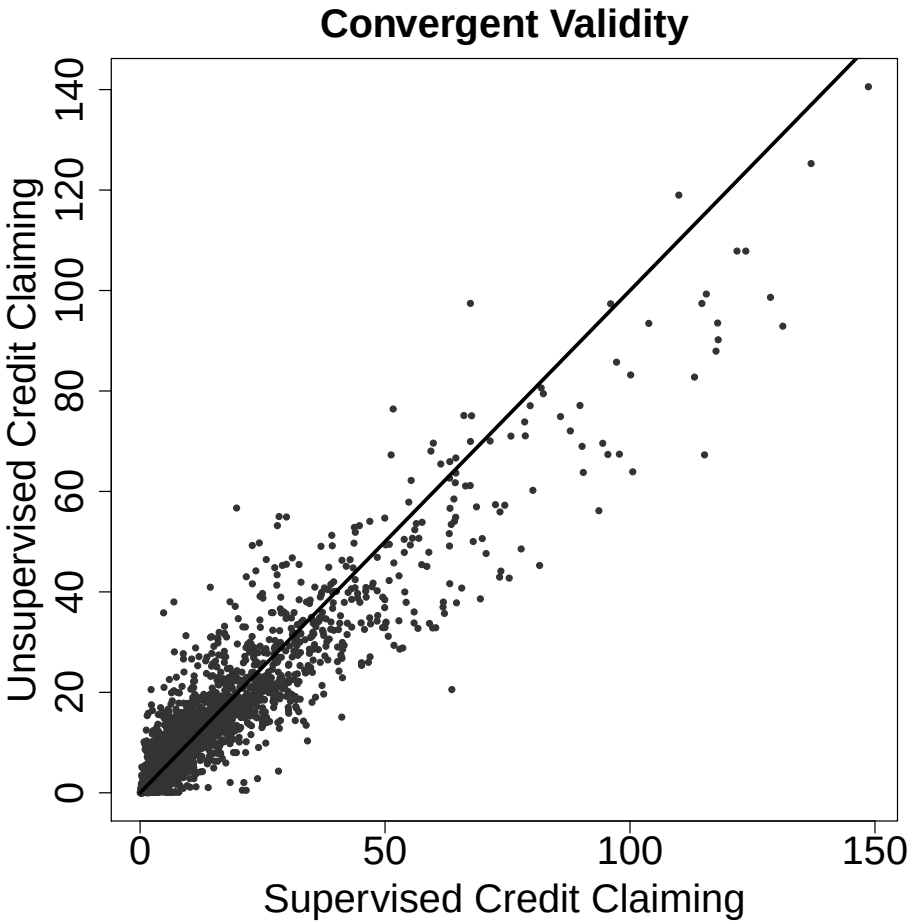


Home Style Measures, Convergent Validity

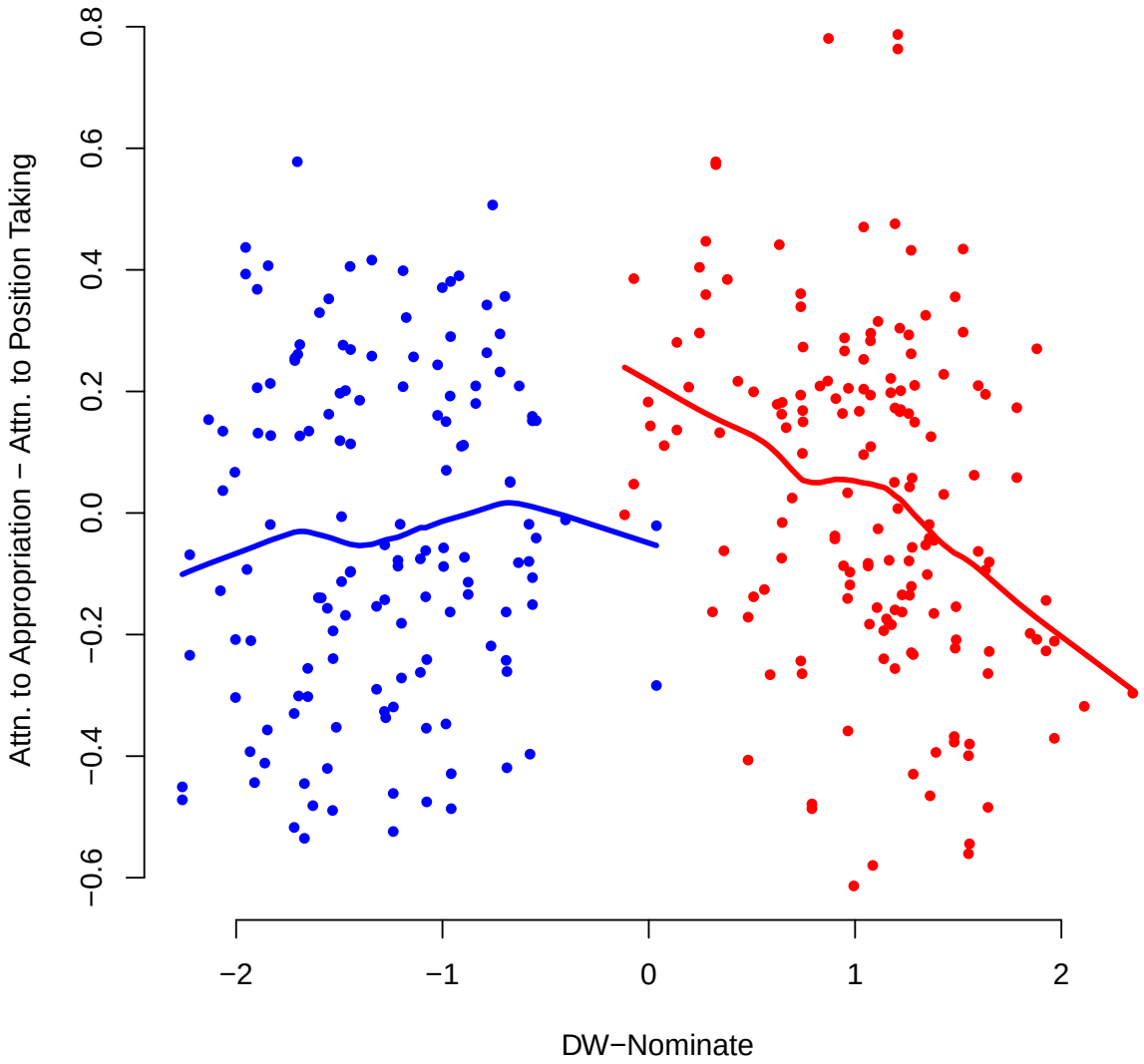
Supervised/Unsupervised Convergence



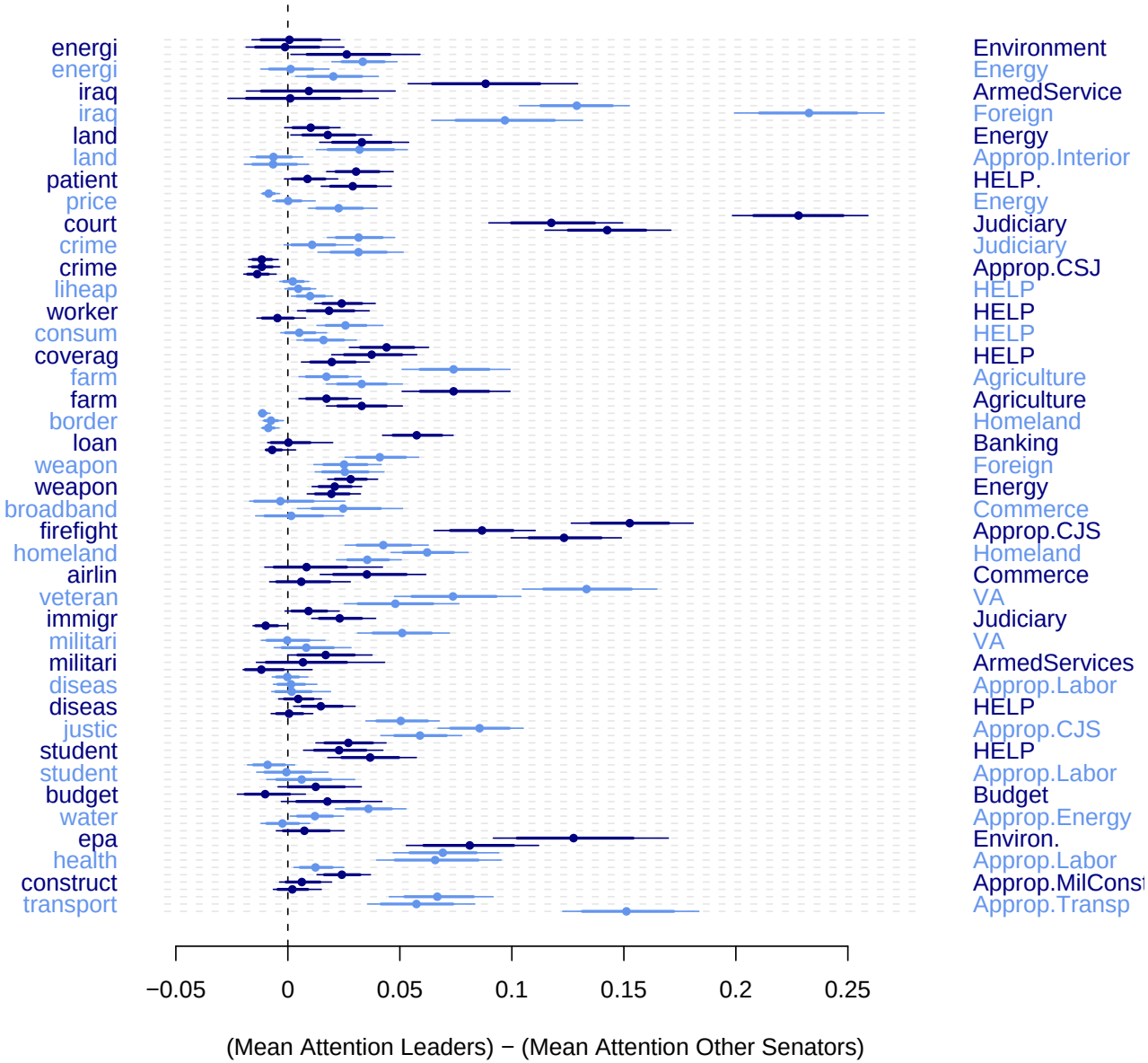
Home Style Measures, Convergent Validity



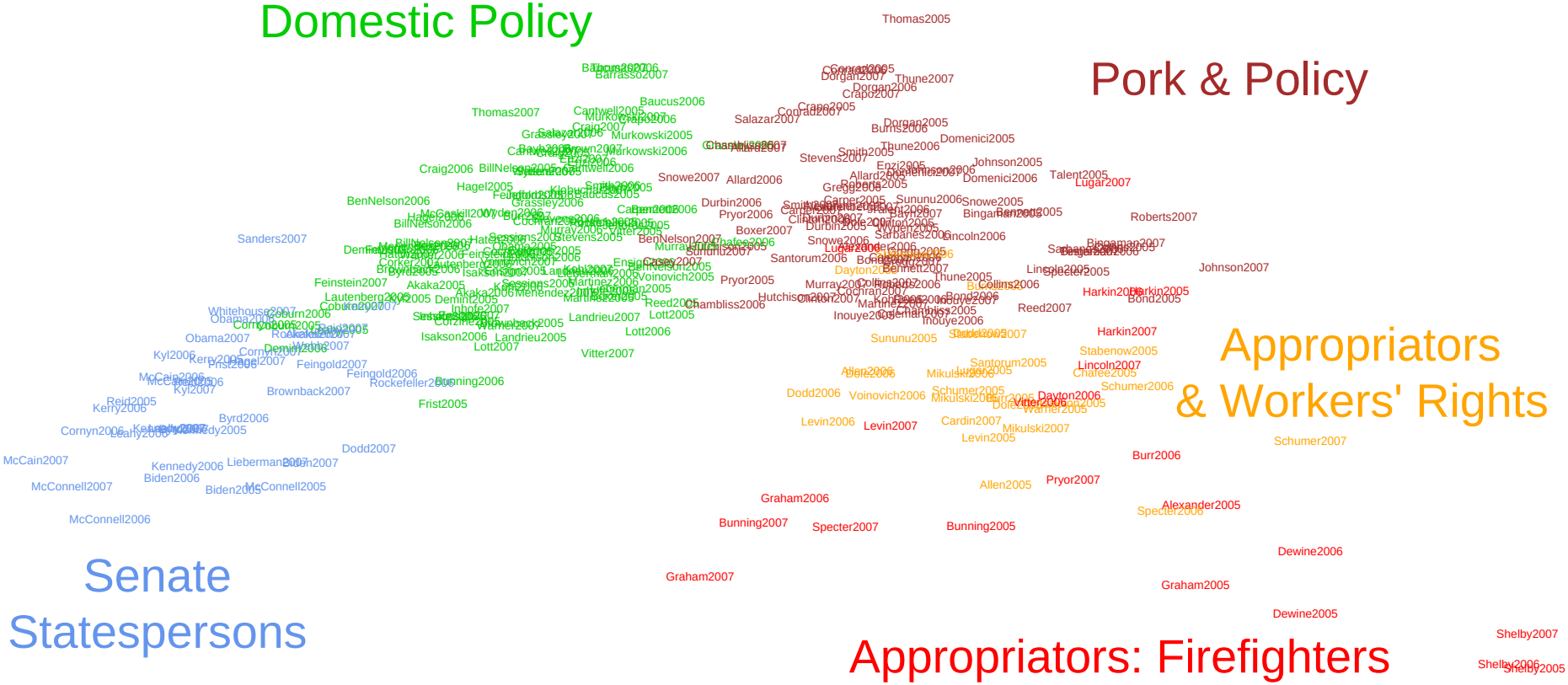
Discriminant Construct Validity



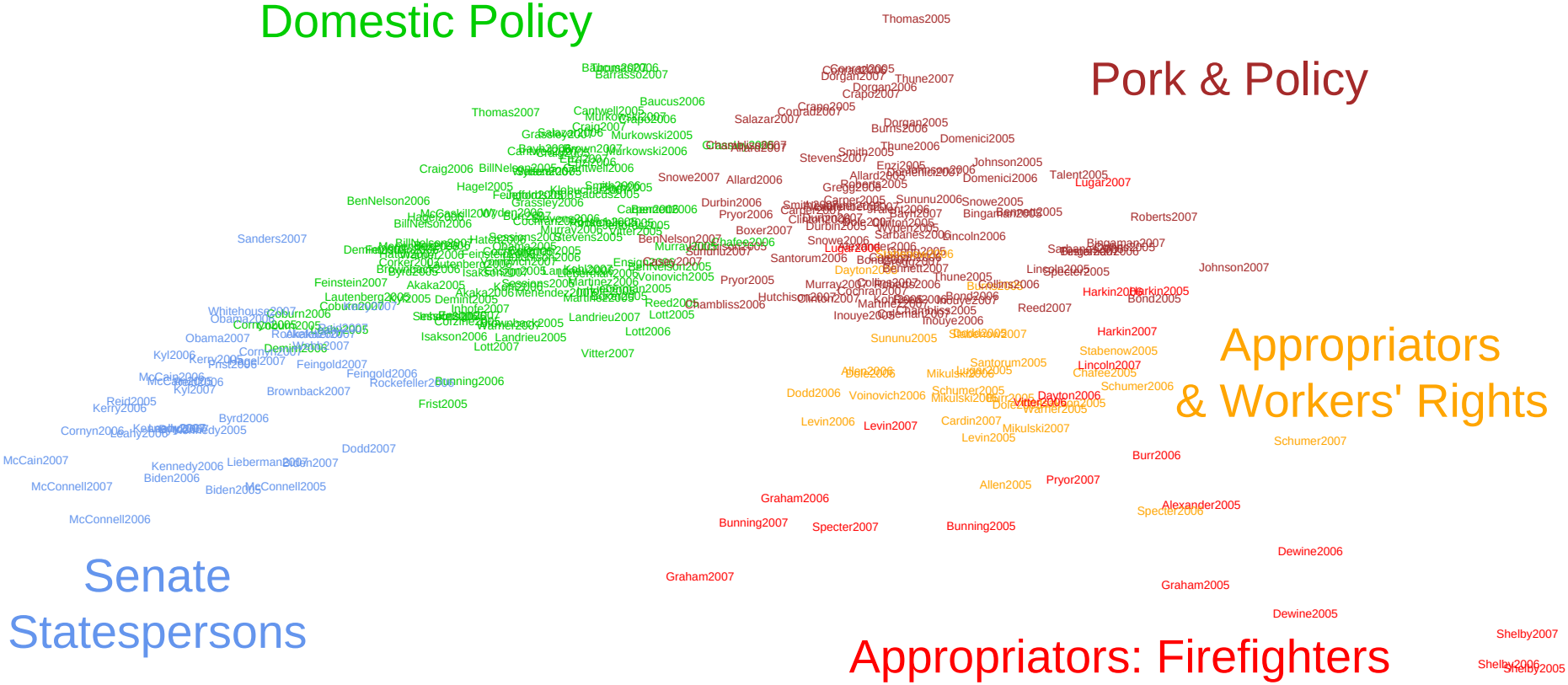
Predictive Validity



Hypothesis Validity



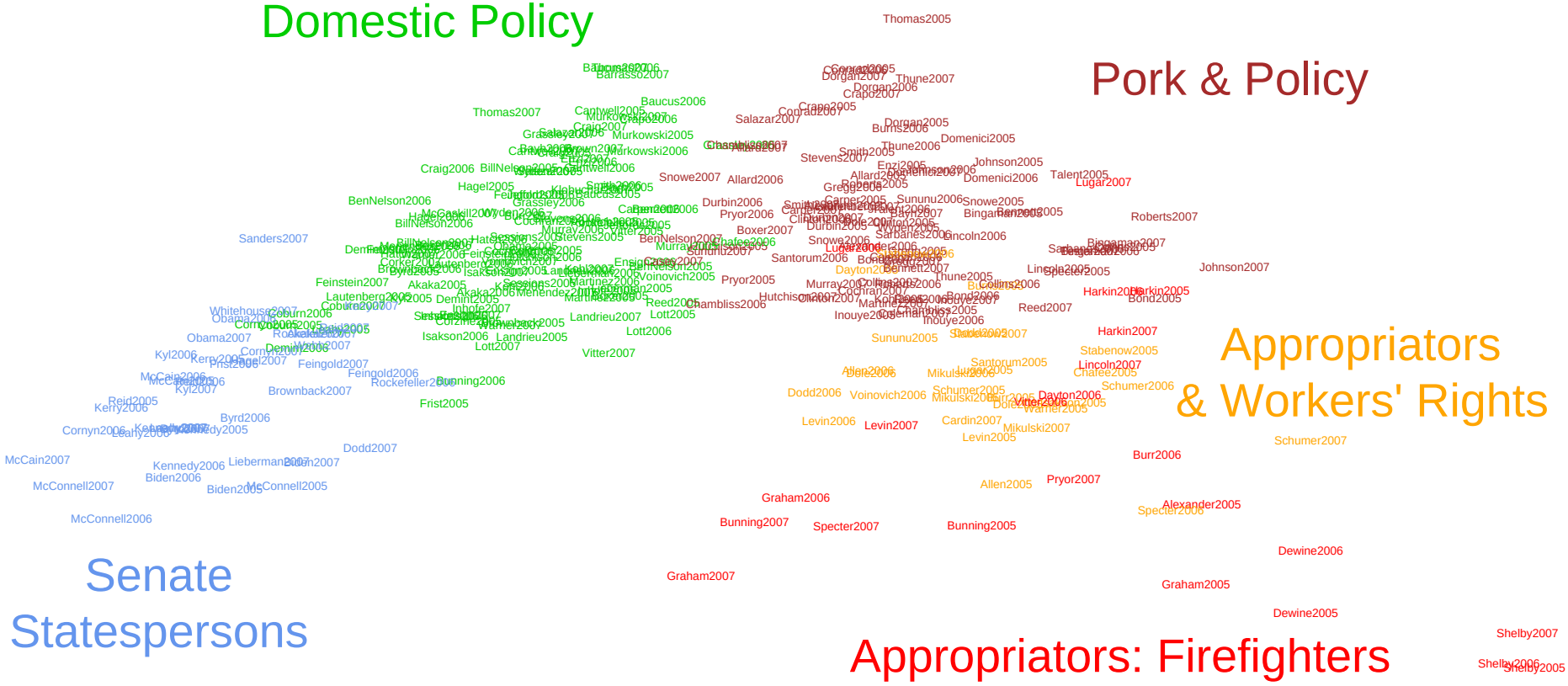
Hypothesis Validity



Senate Statesperson

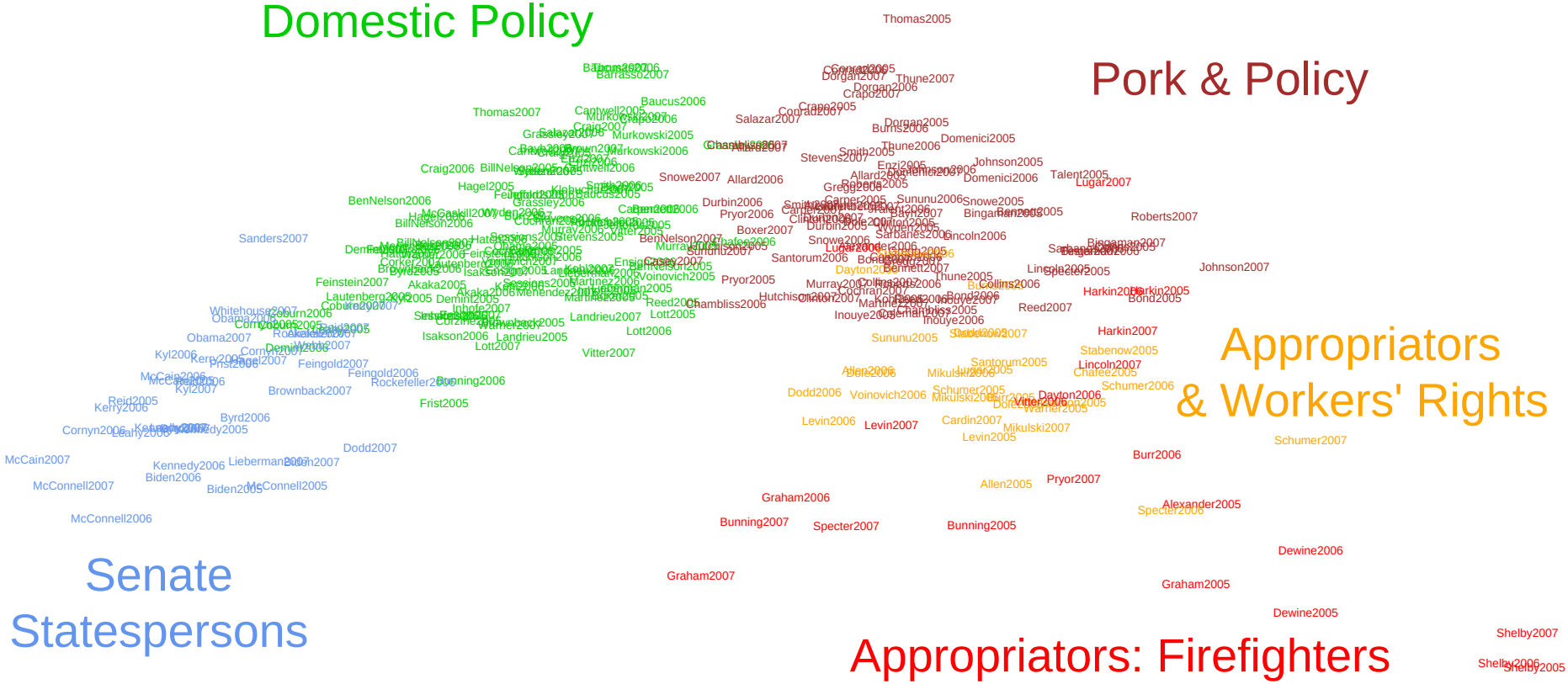
- Iraq War
- Intelligence
- Intl. Relations

Hypothesis Validity



Senate Statesperson	Domestic Policy
- Iraq War	- Environment
- Intelligence	- Gas prices
- Intl. Relations	- DHS
	- Consumer

Hypothesis Validity



Senate Statesperson

- Iraq War
- Intelligence
- Intl. Relations

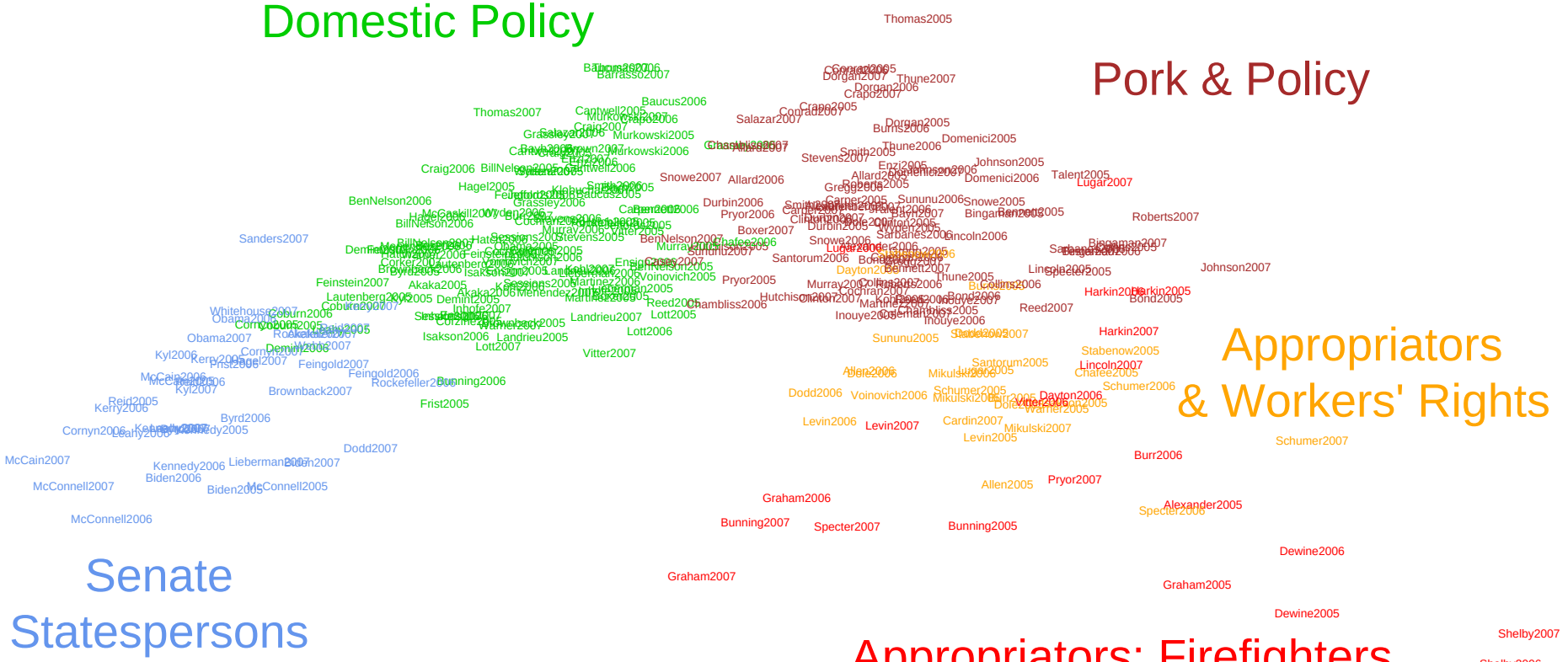
Domestic Policy

- Environment
- Gas prices
- DHS
- Consumer

Pork & Policy

- WRDA grants
- Farming
- Health Care
- Education

Hypothesis Validity



Senate Statesperson

- Iraq War
- Intelligence
- Intl. Relations

Domestic Policy

- Environment
- Gas prices
- DHS
- Consumer

Pork & Policy

- WRDA grants
- Farming
- Health Care
- Education

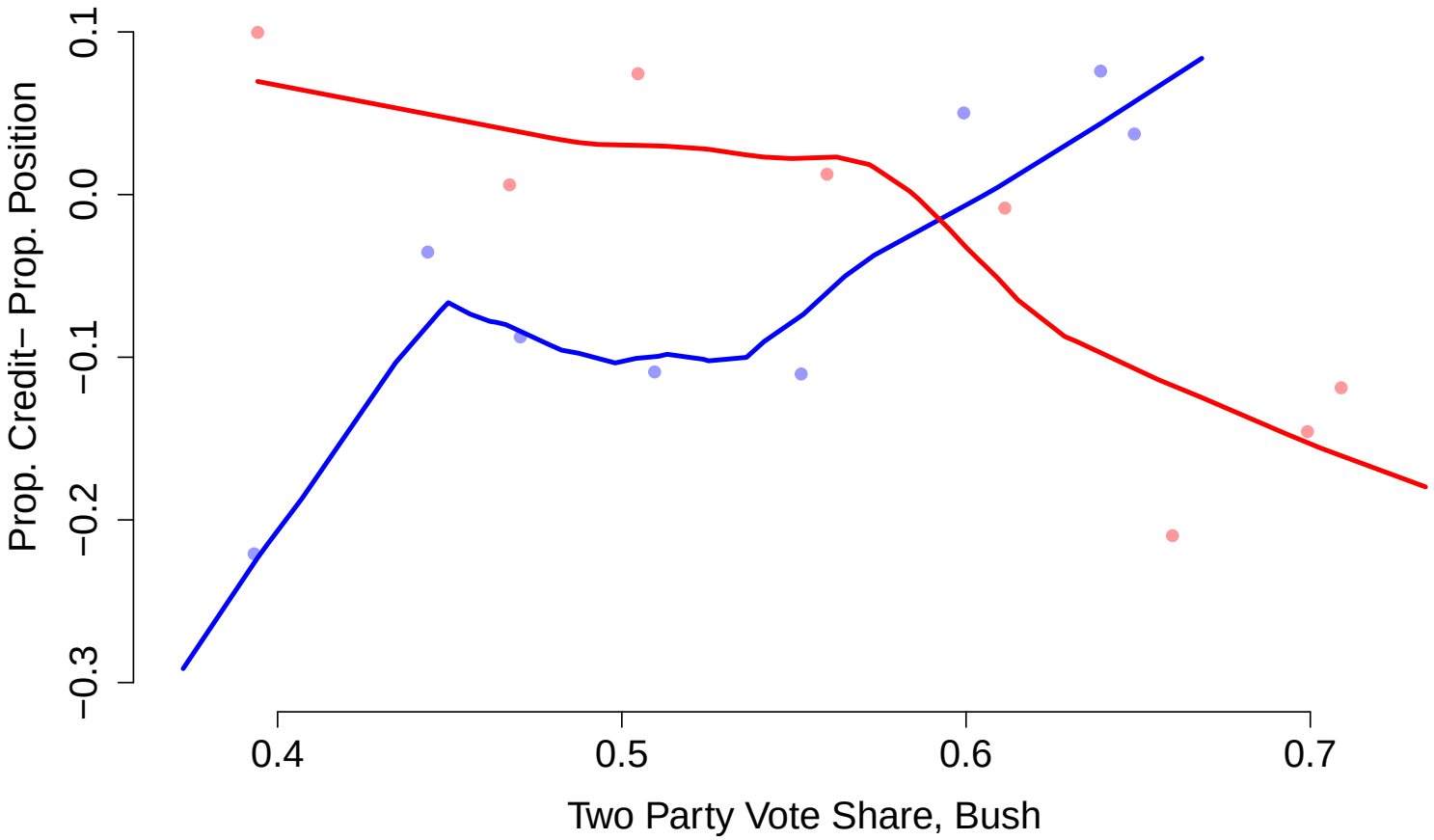
Appropriators

- Fire Grants
- Airport Grants
- University Money

Hypothesis Validity

Why do senators adopt different styles?

District Fit



What are the right number of topics?

What are the right number of topics?

- Number of topics$\rightsquigarrow$ depends on task at hand

What are the right number of topics?

- Number of topics$\rightsquigarrow$ depends on task at hand
- Coarse$\rightsquigarrow$ broad comparisons, lose distinctions

What are the right number of topics?

- Number of topics$\rightsquigarrow$ depends on task at hand
- Coarse$\rightsquigarrow$ broad comparisons, lose distinctions
- Granular$\rightsquigarrow$ specific insights, lose broader picture

What are the right number of topics?

- Number of topics$\rightsquigarrow$ depends on task at hand
- Coarse$\rightsquigarrow$ broad comparisons, lose distinctions
- Granular$\rightsquigarrow$ specific insights, lose broader picture
- Hierarchy of topics$\rightsquigarrow$ Pachinko Allocation, Hierarchies of von-Mises Fisher Distributions

What are the right number of topics?

- Number of topics$\rightsquigarrow$ depends on task at hand
- Coarse$\rightsquigarrow$ broad comparisons, lose distinctions
- Granular$\rightsquigarrow$ specific insights, lose broader picture
- Hierarchy of topics$\rightsquigarrow$ Pachinko Allocation, Hierarchies of von-Mises Fisher Distributions

Blaydes, Grimmer, and McQueen 2017$\rightsquigarrow$ estimate nested topics to explore the Mirros for Princes

The Mirrors Genre (BGM 2017)

26 Christian mirrors

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)
- Kalila wa Dimna (748 CE)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)
- Kalila wa Dimna (748 CE)
- The Sultan's Register of Laws (1632-1633 CE)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)
- Kalila wa Dimna (748 CE)
- The Sultan's Register of Laws (1632-1633 CE)

Work with translations

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)
- Kalila wa Dimna (748 CE)
- The Sultan's Register of Laws (1632-1633 CE)

Work with translations$\rightsquigarrow$ little evidence of selection

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)
- Kalila wa Dimna (748 CE)
- The Sultan's Register of Laws (1632-1633 CE)

Work with translations$\rightsquigarrow$ little evidence of selection

- Collect data on collection of 98 (51 Christian, 47 Islamic, some not translated)

The Mirrors Genre (BGM 2017)

26 Christian mirrors

- The Prince (1513 CE)
- Advice to Justinian (527 CE)
- The Adventures of Telemachus (1699 CE)

21 Islamic texts

- Advice on the Art of Governance (1612 CE)
- Kalila wa Dimna (748 CE)
- The Sultan's Register of Laws (1632-1633 CE)

Work with translations$\rightsquigarrow$ little evidence of selection

- Collect data on collection of 98 (51 Christian, 47 Islamic, some not translated)
- No difference on Year/Region

Preprocessing Texts

47 books

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
    - almighty, god

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
  - almighty, god
  - monarch, prince, king, ruler

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
  - almighty, god
  - monarch, prince, king, ruler
  - Lord $\neq$ lord

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
  - almighty, god
  - monarch, prince, king, ruler
  - Lord $\neq$ lord

Result: short segment j in book i is a count vector

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
  - almighty, god
  - monarch, prince, king, ruler
  - Lord $\neq$ lord

Result: short segment j in book i is a count vector

$$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ij2124})$$

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
  - almighty, god
  - monarch, prince, king, ruler
  - Lord $\neq$ lord

Result: short segment j in book i is a count vector

$$\mathbf{x}_{ij} \quad = \quad (x_{ij1}, x_{ij2}, \ldots, x_{ij2124})$$

We work with a normalized version of the documents,

Preprocessing Texts

47 books $\rightsquigarrow$ Each divided into paragraphs

Create feature space

- Bag of words, stem, discard punctuation, stop words
- Translate words left in Arabic (allah) and discard proper nouns
- Identified synonyms
  - almighty, god
  - monarch, prince, king, ruler
  - Lord $\neq$ lord

Result: short segment j in book i is a count vector

$$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ij2124})$$

We work with a normalized version of the documents,

$$\mathbf{x}^*_{ij} = \frac{\mathbf{x}_{ij}}{\sqrt{\mathbf{x}'_{ij}\mathbf{x}_{ij}}}$$

Measuring Themes in the Mirrors

Model built around two hierarchies:

Measuring Themes in the Mirrors

Model built around two hierarchies:

1) Books $\rightsquigarrow$ paragraphs (Blei, Ng, Jordan 2003; Wallach, 2008; Quinn et al 2010; Grimmer 2010; Roberts et al 2014)

Measuring Themes in the Mirrors

Model built around two hierarchies:

1) Books $\rightsquigarrow$ paragraphs (Blei, Ng, Jordan 2003; Wallach, 2008; Quinn et al 2010; Grimmer 2010; Roberts et al 2014)
2) Coarse topics $\rightsquigarrow$ granular topics (Li and McCallum 2006; Gopal and Yang 2014)

Measuring Themes in the Mirrors

Estimate four quantities of interest

Measuring Themes in the Mirrors

Estimate four quantities of interest

1) Granular topics (60)

Measuring Themes in the Mirrors

Estimate four quantities of interest

1) Granular topics (60)
2) Coarse (broad) topics (3)

Measuring Themes in the Mirrors

Estimate four quantities of interest

1) Granular topics (60)
2) Coarse (broad) topics (3)
    - Each granular topic classified into one coarse topic

Measuring Themes in the Mirrors

Estimate four quantities of interest

1) Granular topics (60)
2) Coarse (broad) topics (3)
    - Each granular topic classified into one coarse topic
3) Each book i's $\mathbf{themes}_i$

$$\mathbf{themes}_i \quad = \quad (\text{theme}_{i1}, \text{theme}_{i2}, \ldots, \text{theme}_{i60})$$

Measuring Themes in the Mirrors

Estimate four quantities of interest

1) Granular topics (60)
2) Coarse (broad) topics (3)
    - Each granular topic classified into one coarse topic
3) Each book i's $\textbf{themes}_i$
4) Each short segment's granular (and coarse) topic

A Hierarchy of Topics

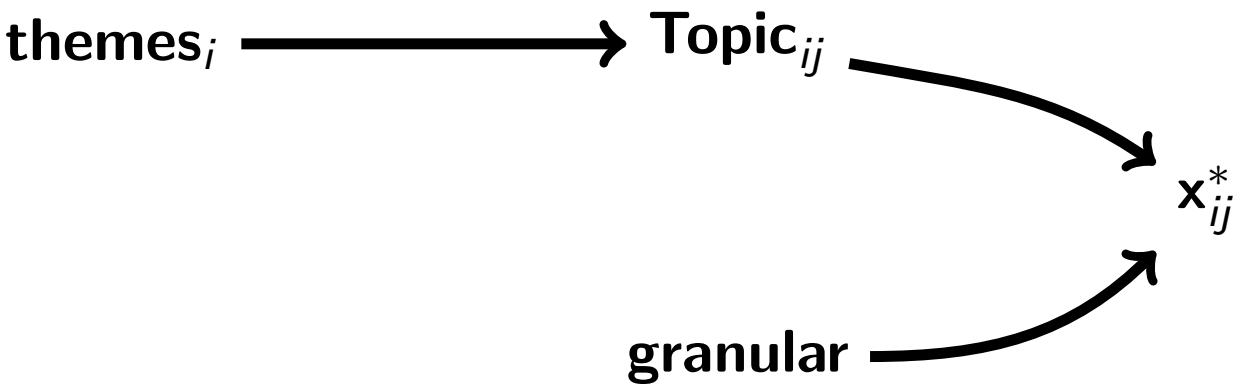
themes$_j$

A Hierarchy of Topics

$$\textbf{themes}_i \longrightarrow \textbf{Topic}_{ij}$$

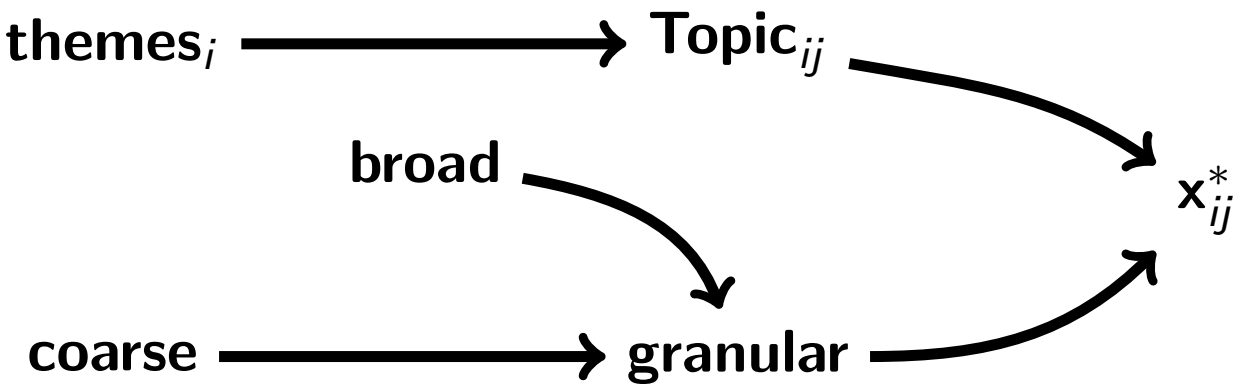
$$\textbf{Topic}_{ij} \sim \text{Multinomial}(1, \textbf{themes}_i)$$

A Hierarchy of Topics



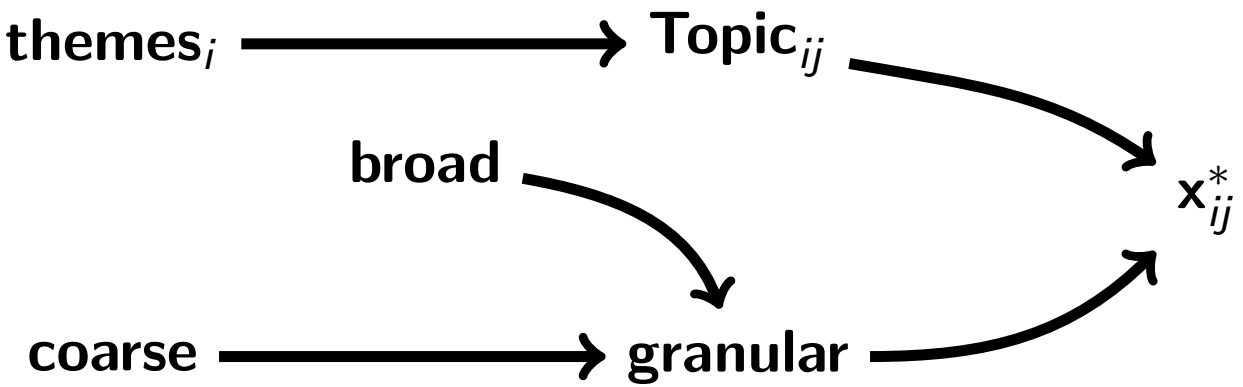
$$\begin{aligned} \textbf{Topic}_{ij} &\sim \text{Multinomial}(1, \textbf{themes}_i) \\ \mathbf{x}^*_{ij} | \text{Topic}_{ijk} = 1 &\sim \text{vMF}(\kappa, \textbf{granular}_k) \end{aligned}$$

A Hierarchy of Topics



$$
\begin{aligned}
\textbf{Topic}_{ij} &\sim \text{Multinomial}(1, \textbf{themes}_i) \\
\mathbf{x}^*_{ij} | \text{Topic}_{ijk} = 1 &\sim \text{vMF}(\kappa, \textbf{granular}_k) \\
\textbf{broad}_k &\sim \text{Multinomial}(1, \textbf{Broad Theme Prior}) \\
\textbf{granular}_k | \text{broad}_{km} = 1 &\sim \text{vMF}(\kappa, \textbf{coarse}_m)
\end{aligned}
$$

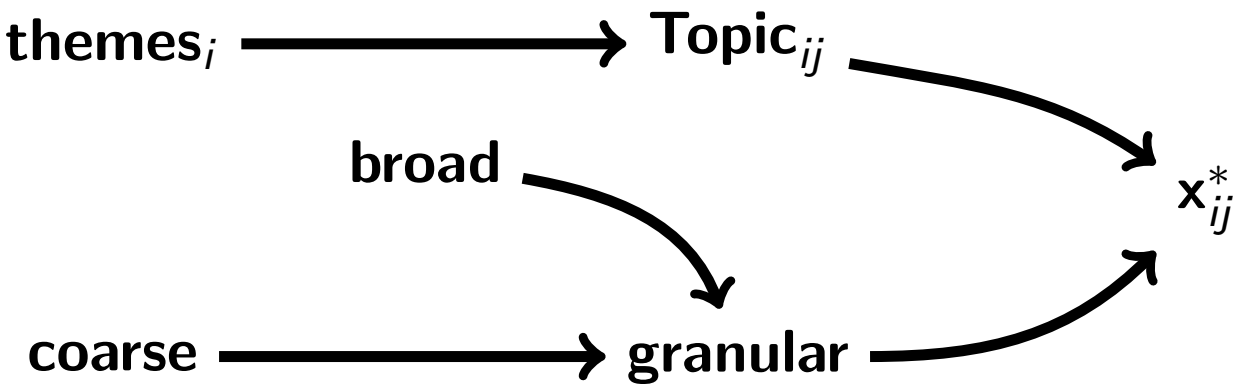
A Hierarchy of Topics



$$
\begin{aligned}
\textbf{Topic}_{ij} &\sim \text{Multinomial}(1, \textbf{themes}_i) \\
\mathbf{x}^*_{ij} | \text{Topic}_{ijk} = 1 &\sim \text{vMF}(\kappa, \textbf{granular}_k) \\
\textbf{broad}_k &\sim \text{Multinomial}(1, \textbf{Broad Theme Prior}) \\
\textbf{granular}_k | \text{broad}_{km} = 1 &\sim \text{vMF}(\kappa, \textbf{coarse}_m)
\end{aligned}
$$

Estimate model with Variational Approximation

A Hierarchy of Topics



$$
\begin{aligned}
\mathbf{Topic}_{ij} &\sim \text{Multinomial}(1, \mathbf{themes}_i) \\
\mathbf{x}^*_{ij} | \text{Topic}_{ijk} = 1 &\sim \text{vMF}(\kappa, \mathbf{granular}_k) \\
\mathbf{broad}_k &\sim \text{Multinomial}(1, \mathbf{Broad\ Theme\ Prior}) \\
\mathbf{granular}_k | \text{broad}_{km} = 1 &\sim \text{vMF}(\kappa, \mathbf{coarse}_m)
\end{aligned}
$$

Estimate model with Variational Approximation
Model selection: automatic model fit, qualitative evaluation

Interpreting Unsupervised Models

Two approaches to labeling output

Interpreting Unsupervised Models

Two approaches to labeling output

1) Computational: identify discriminating words

Interpreting Unsupervised Models

Two approaches to labeling output

1) Computational: identify discriminating words
2) Manual: Segments classified to coarse, granular topics. Read, discuss, and label

Interpreting Unsupervised Models

Two approaches to labeling output

1) Computational: identify discriminating words
2) Manual: Segments classified to coarse, granular topics. Read, discuss, and label

Unsupervised models structure and guide our reading

Art of Rulership

Practices and ideals of political rule

Art of Rulership

Practices and ideals of political rule

king

Art of Rulership

Practices and ideals of political rule

king,princ

Art of Rulership

Practices and ideals of political rule

king,princ,citi

Art of Rulership

Practices and ideals of political rule

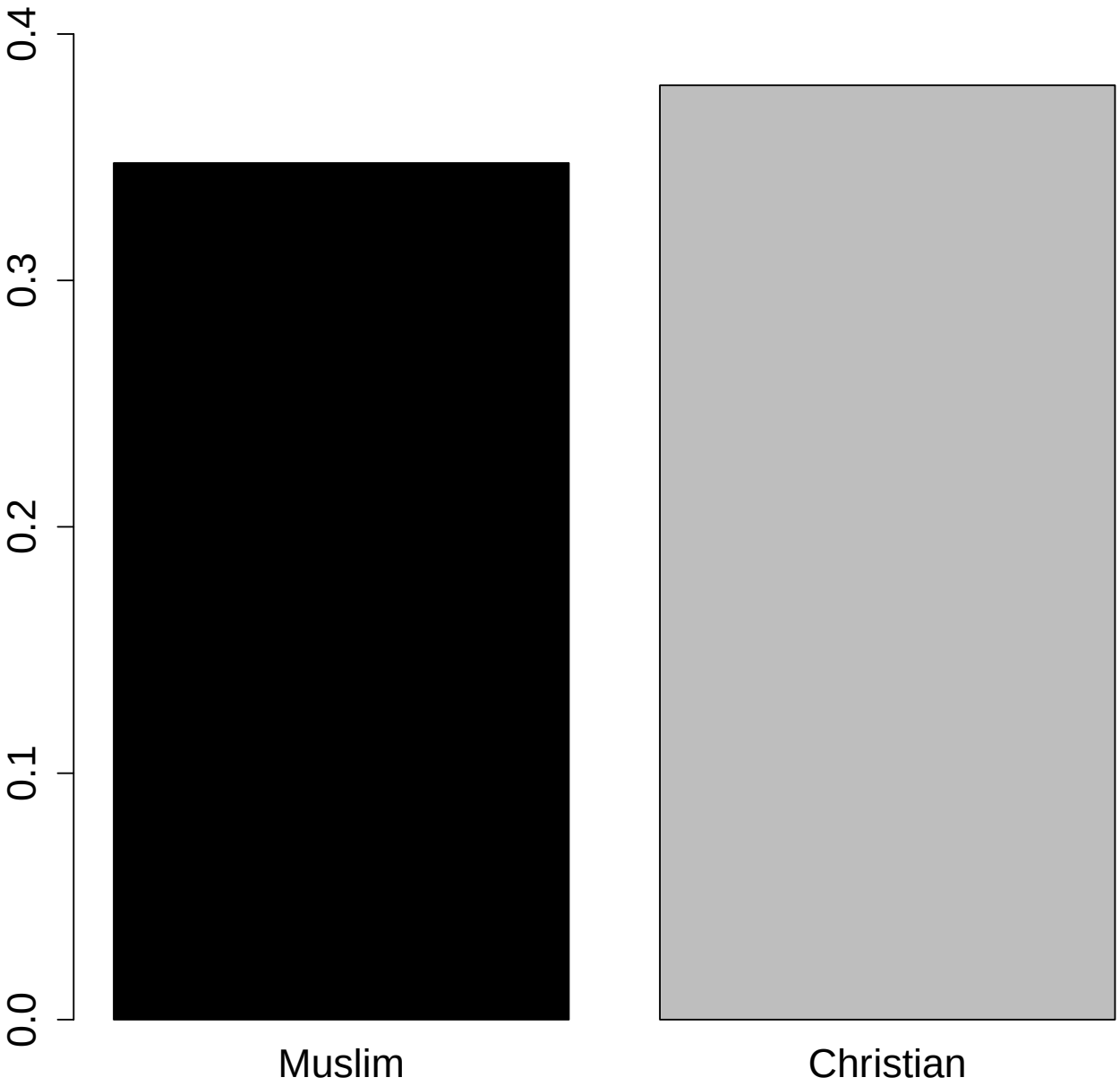
king,princ,citi,great,place,work,emperor,enemi,armi,letter

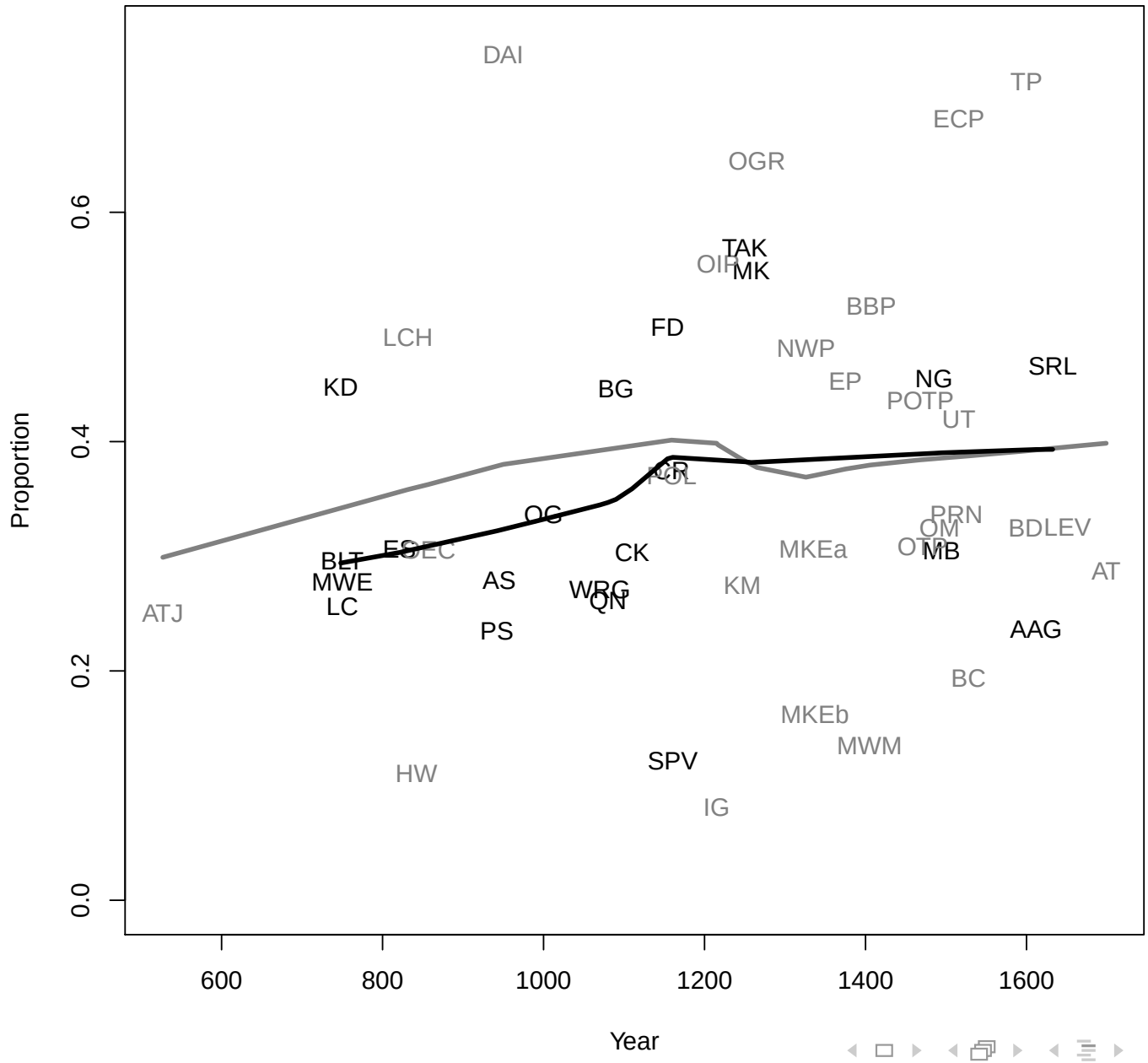
Art of Rulership

Practices and ideals of political rule

king,princ,citi,great,place,work,emperor,enemi,armi,letter

36.5% of paragraphs


0.0
0.1
0.2
0.3
0.4
Muslim
Christian


Coarse Topic 1
DAI
TP
ECP
OGR
TAK
OIP
MK
BBP
FD
LCH
NWP
SRL
KD
BG
EP
NG
POTP
UT
PRL
PRN
OG
OM
BDLEV
MKEa
OTP
MB
CK
BLT
MWE
LC
AS
WRG
QN
KM
AT
ATJ
PS
AAG
BC
MKEb
MWM
SPV
HW
IG
Proportion
0.0
0.2
0.4
0.6
600
800
1000
1200
1400
1600
Year

Religion and Virtue

Connection between religion, virtue, justice and political rule

Religion and Virtue

Connection between religion, virtue, justice and political rule

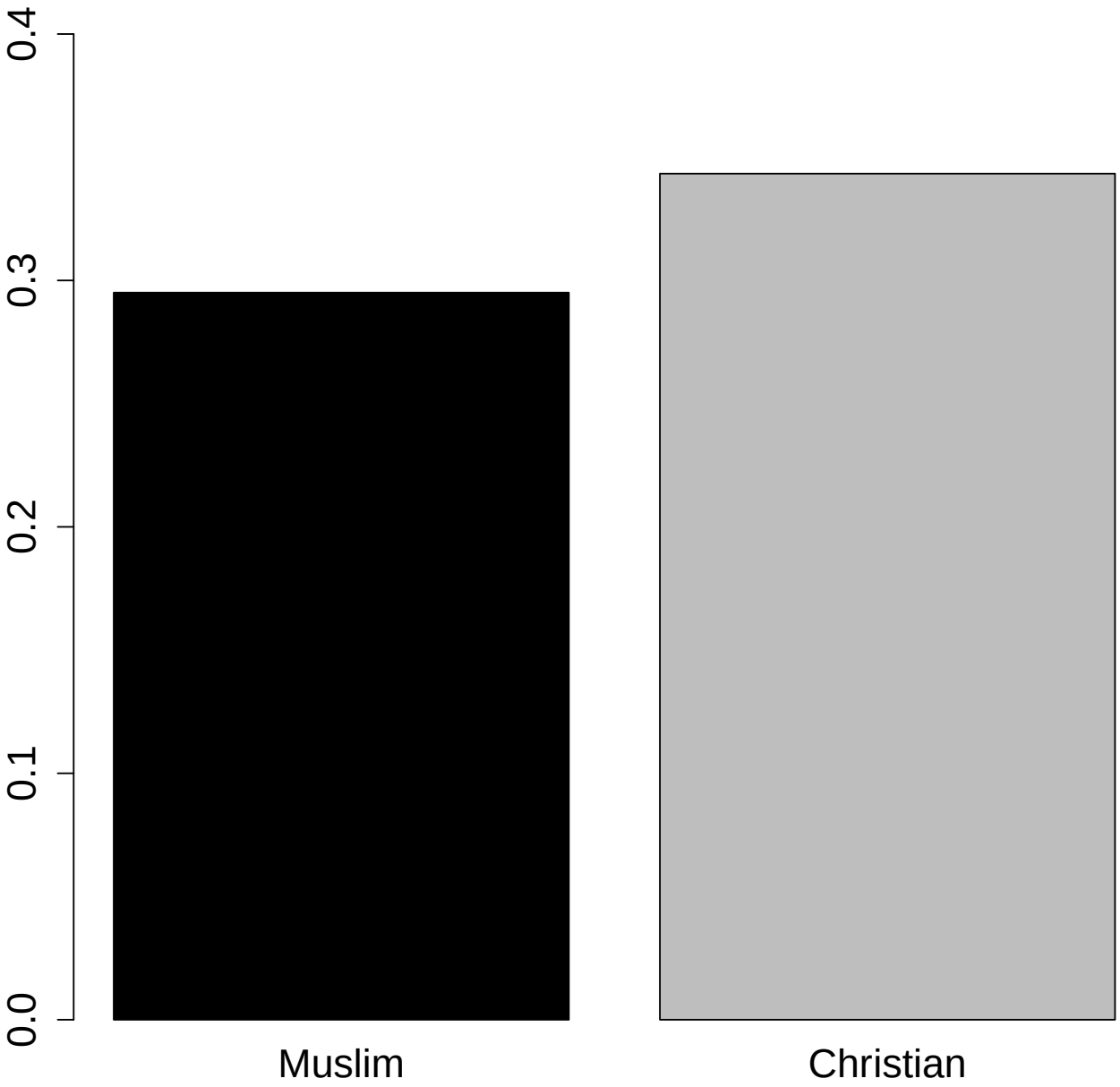
almighti,good,virtu,power,ruler,justic,prayer,rule,prophet,mena

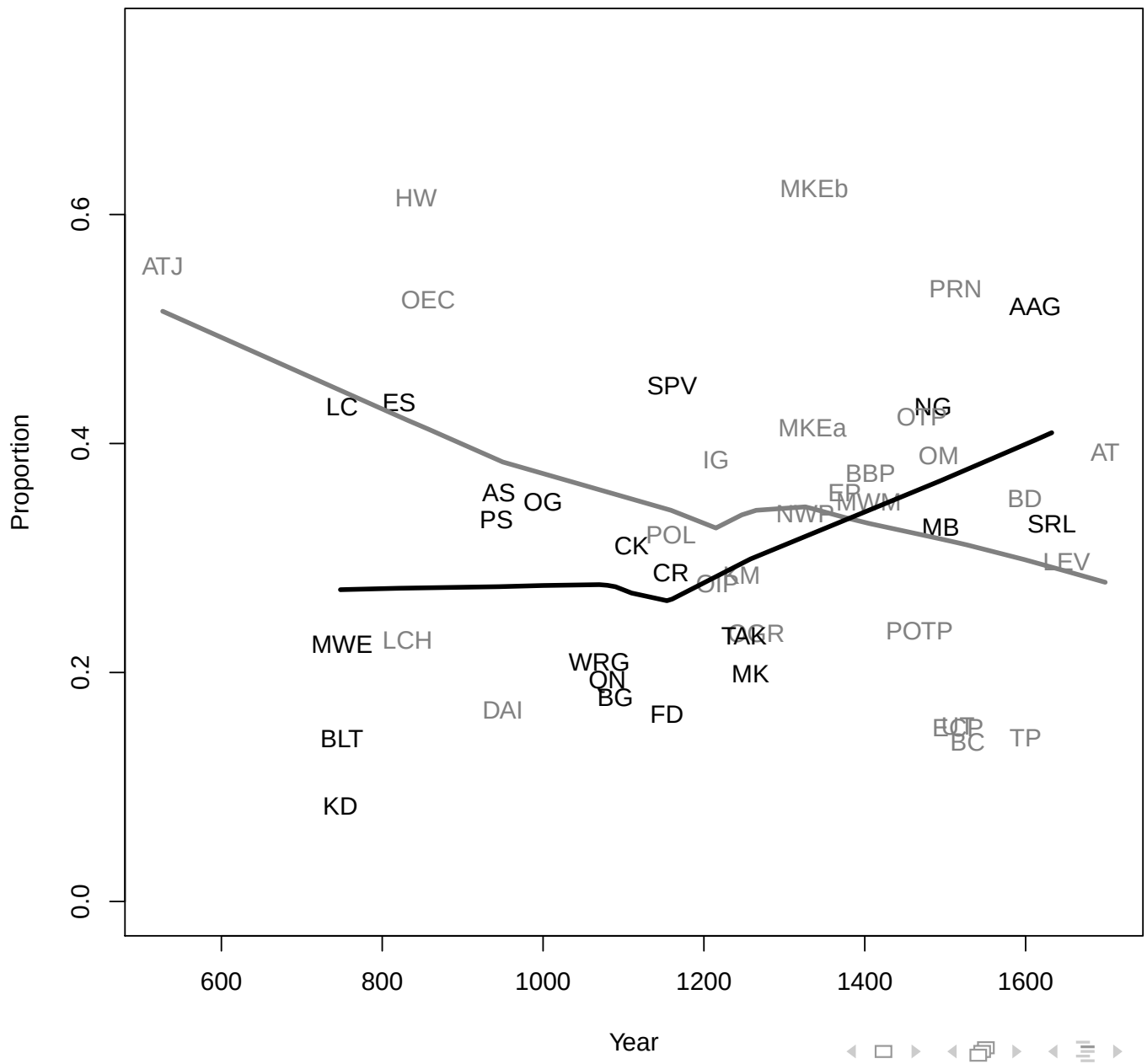
Religion and Virtue

Connection between religion, virtue, justice and political rule

almighti,good,virtu,power,ruler,justic,prayer,rule,prophet,mena

32.2% of pargraphs


0.0
0.1
0.2
0.3
0.4
Muslim
Christian


Coarse Topic 2
Proportion
0.0
0.2
0.4
0.6
600
800
1000
1200
1400
1600
Year
HW
MKEb
ATJ
OEC
PRN
AAG
SPV
LC
ES
NG
OTP
MKEa
IG
OM
AT
BBP
AS
OG
PS
BD
MB
SRL
POL
CK
CR
LEV
MWE
LCH
TAK
POTP
WRG
ON
MK
BG
DAI
FD
BLT
TP
BC
KD

Inner Life of the Ruler

Personal relationships, care for and practices of the self, and ultimate fate of the soul

Inner Life of the Ruler

Personal relationships, care for and practices of the self, and ultimate fate of the soul

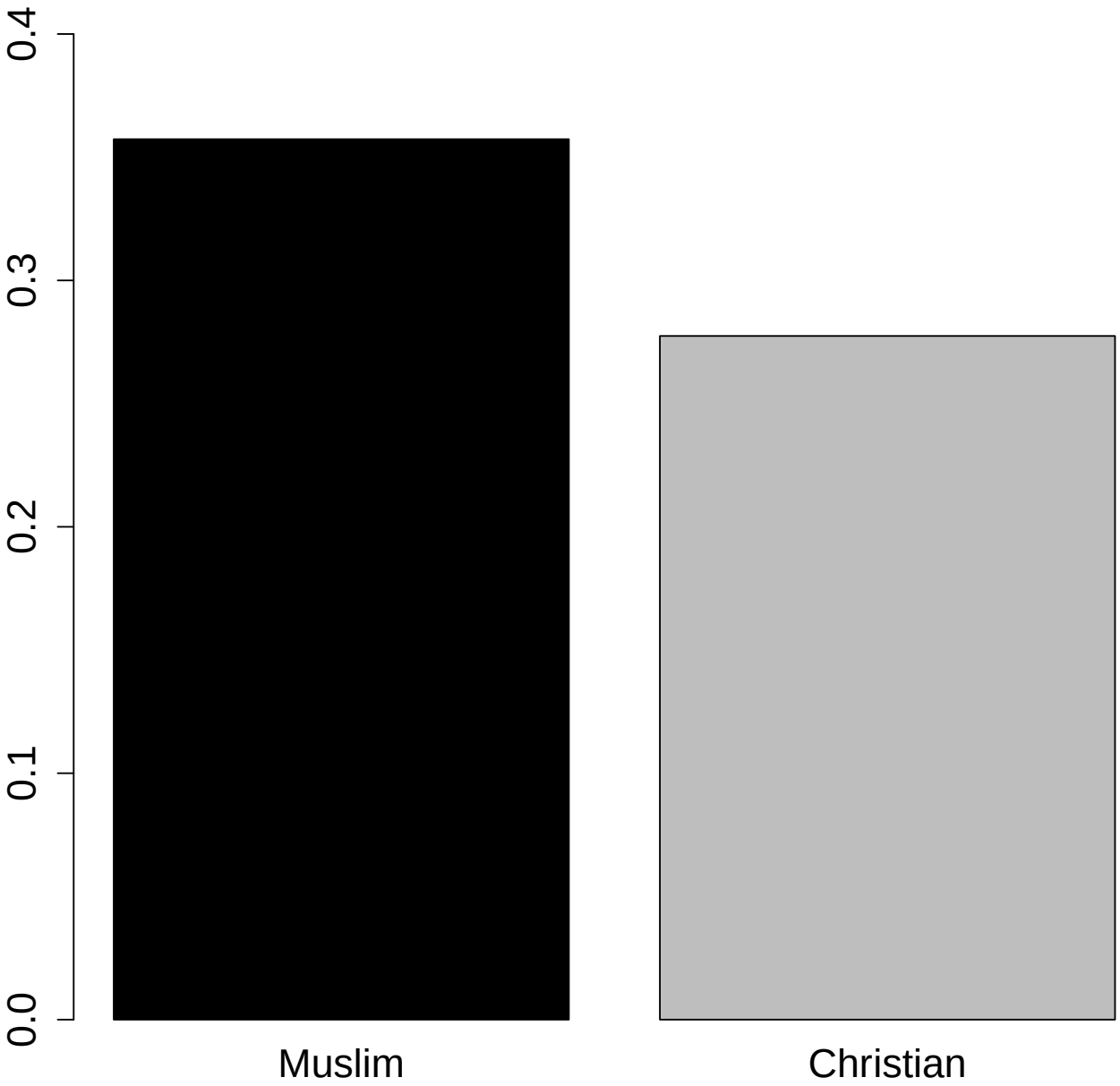
man,land,woman,know,bodi,eye,ladi,love,faculti,old

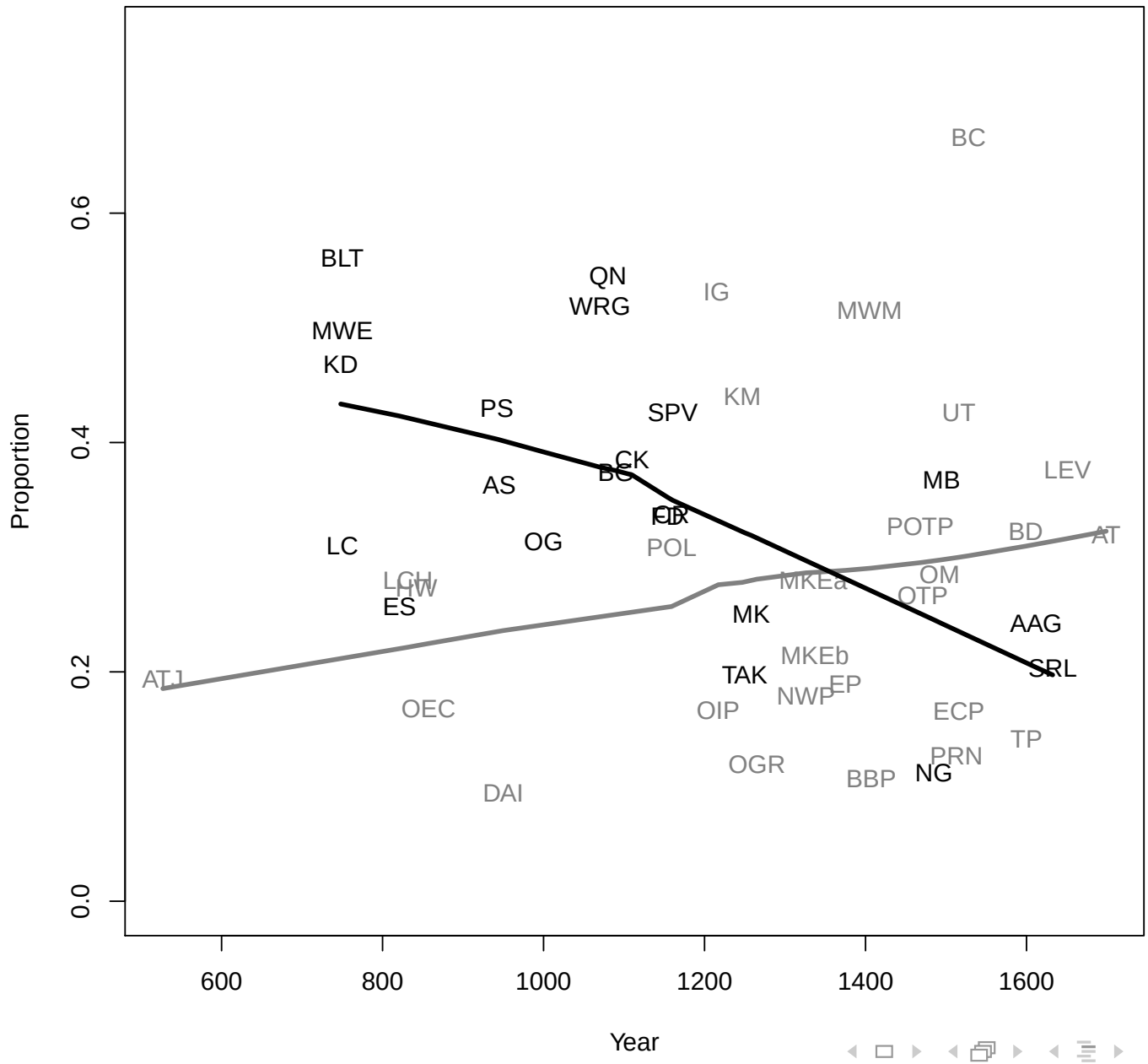
Inner Life of the Ruler

Personal relationships, care for and practices of the self, and ultimate fate of the soul

man,land,woman,know,bodi,eye,ladi,love,faculti,old

31.2% of paragraphs

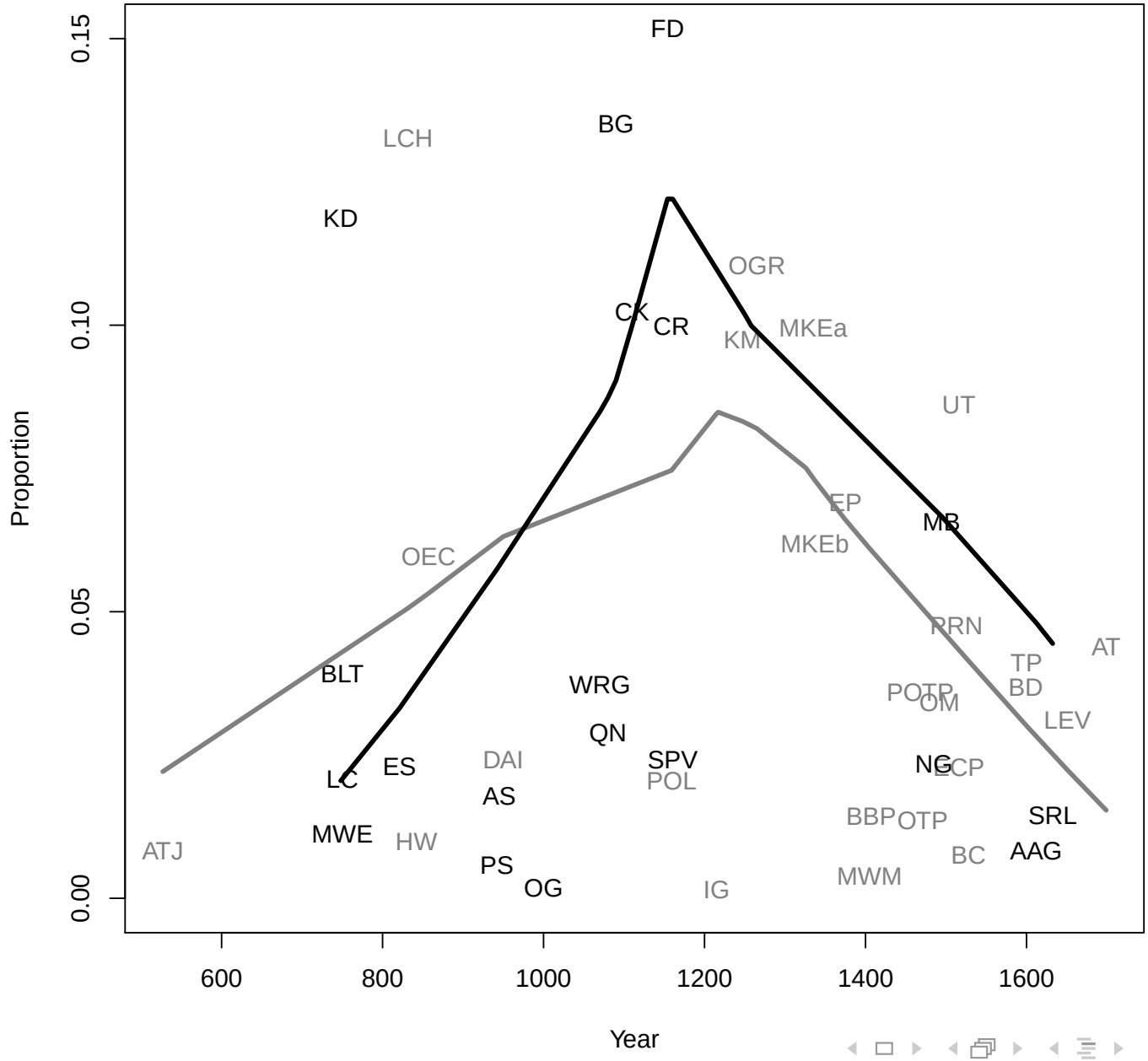

0.0
0.1
0.2
0.3
0.4
Muslim
Christian


Coarse Topic 3
Proportion
0.0
0.2
0.4
0.6
600
800
1000
1200
1400
1600
Year
BC
BLT
QN
WRG
IG
MWM
MWE
KD
KM
PS
SPV
UT
CK
BC
MB
LEV
AS
POTP
BD
AT
LC
OG
POL
OM
MKEa
OTP
ES
MK
AAG
MKEb
SRL
TAK
EP
NWP
OEC
OIP
ECP
TP
PRN
OGR
BBP
NG
DAI

Granular: Best Practices for Ruling

king,princ,citi,great,place,work,emperor,enemi,armi,letter
king,kingdom,royal,minist,reign,father,court,majesti,presenc,war

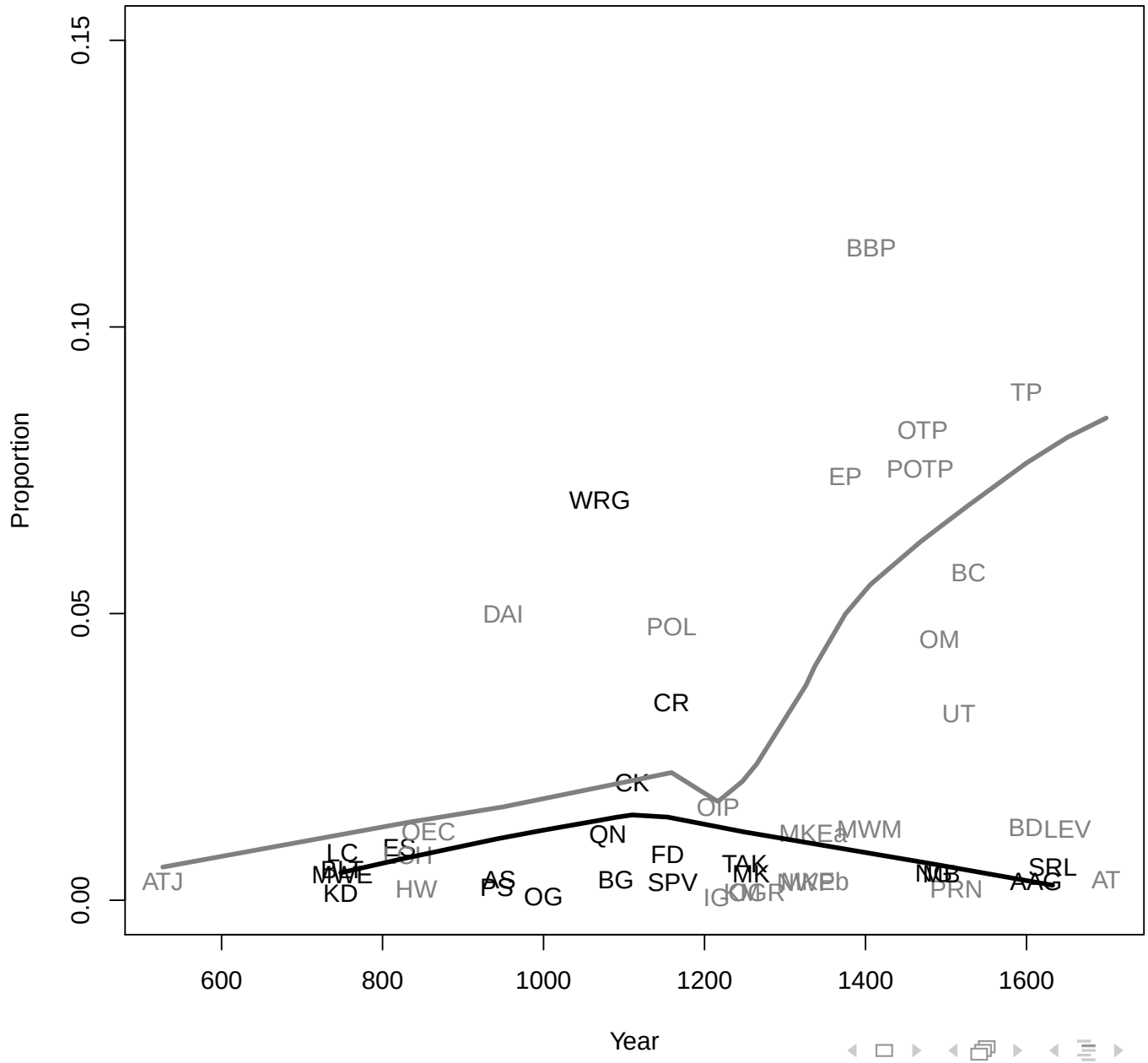
6.2% of paragraphs


Coarse Topic 1 Granular Topic 1
Proportion
0.00
0.05
0.10
0.15
600
800
1000
1200
1400
1600
Year
FD
BG
LCH
KD
OGR
CK
CR
KM
MKEa
UT
EP
MB
MKEb
OEC
RRN
AT
TP
BD
BLT
WRG
POL
OM
LEV
QN
DAI
SPV
NG
LCP
LC
ES
AS
POL
BBP
OTP
SRL
MWE
HW
ATJ
PS
BC
AAG
OG
IG
MWM

Granular: Characteristics that distinguish Just Ruler from Tyrant

king,princ,citi,great,place,work,emperor,enemi,armi,letter
king,kingdom,royal,minist,reign,father,court,majesti,presenc,war
princ,good,peopl,christian,tyranni,war,mind,ought,state,public

3.1% of paragraphs

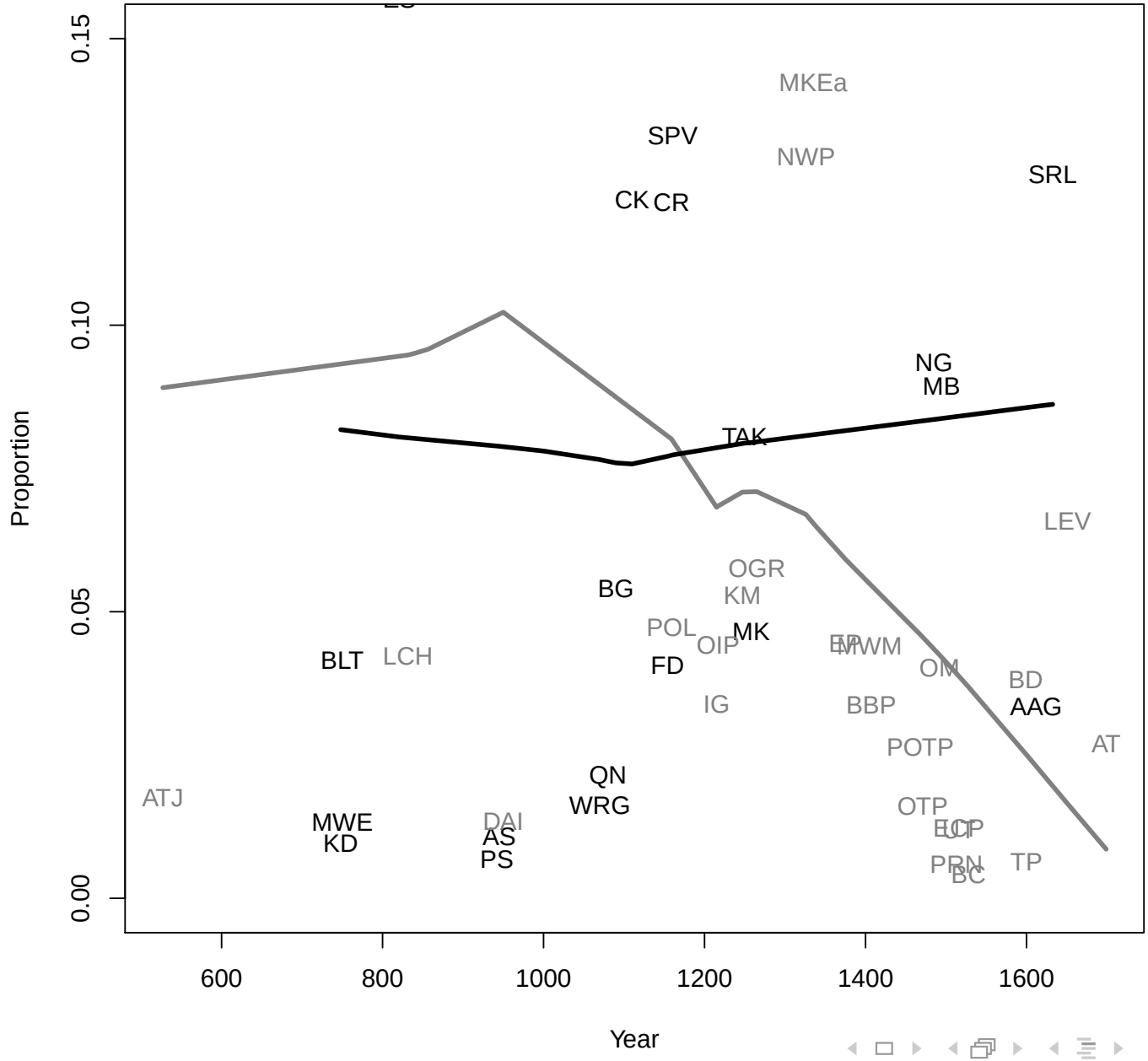

Coarse Topic 1 Granular Topic 2
Proportion
0.00
0.05
0.10
0.15
600
800
1000
1200
1400
1600
Year
ATJ
KD
HW
PS
OG
BG
SPV
IG
PRN
AAG
AT
WRG
DAI
POL
CR
CK
QN
OIP
FD
BBP
TP
OTP
POTP
EP
BC
OM
UT
MWM
BDLEV
SRL

Granular: Religious Virtues and Political Ideals

almighti,good,virtu,power,ruler,justic,prayer,rule,prophet,mena
almighti,bless,grant,peac,messeng,prophet,merci,holi,command,grace

6.9% of paragraphs

Coarse Topic 2 Granular Topic 1

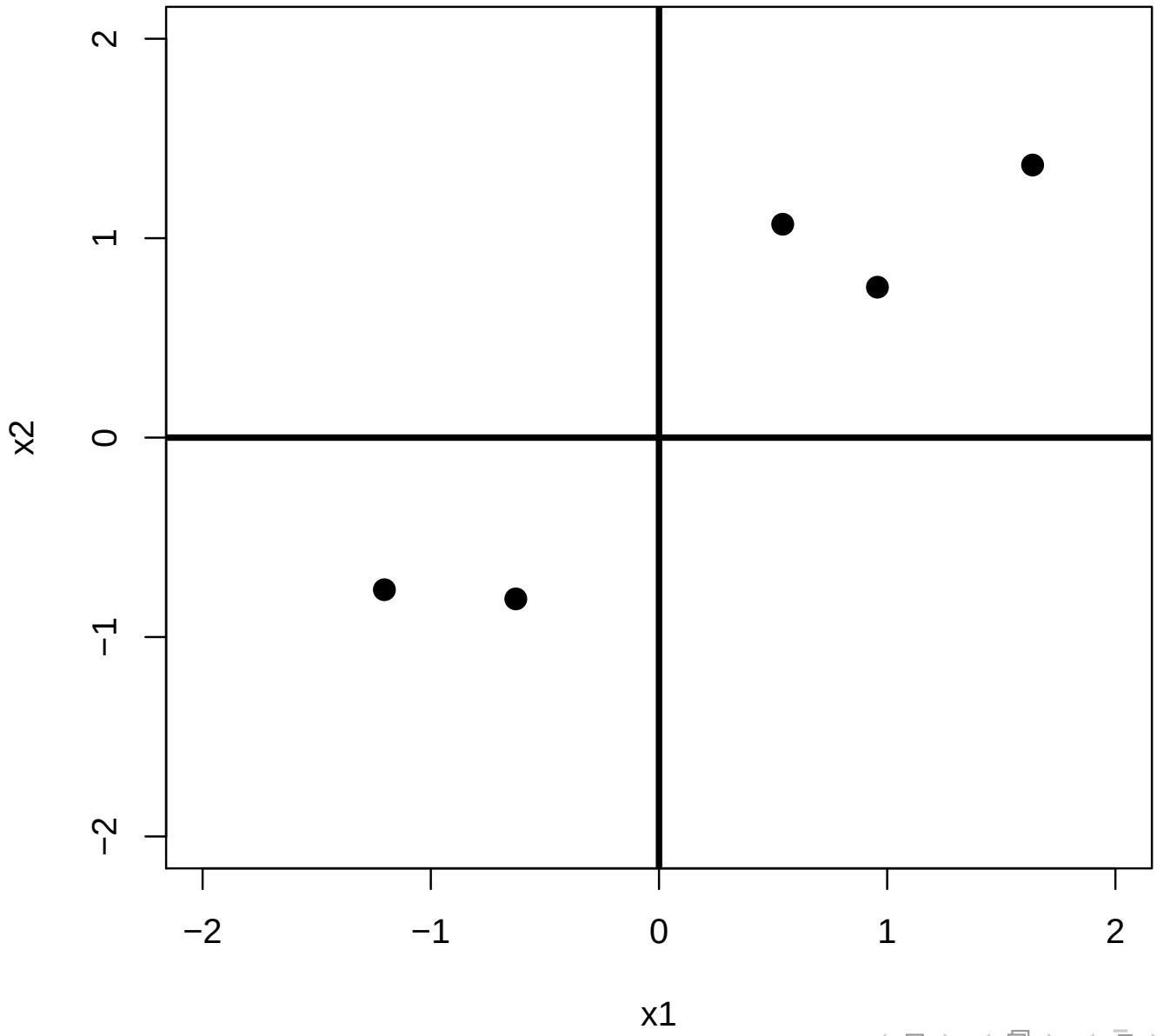


Principal Component Analysis

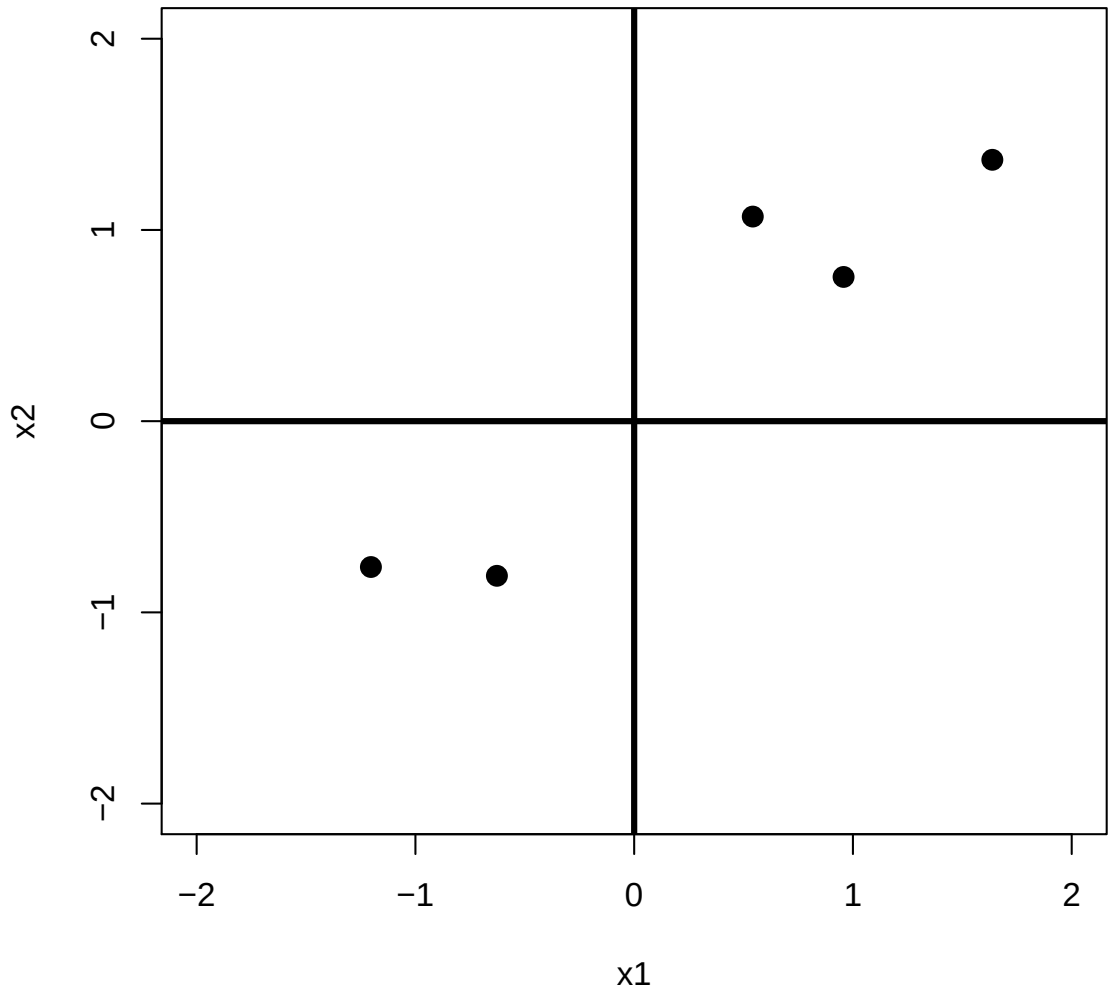
A Simple Two-Dimensional Example

Suppose we have the following observations:

$$\begin{aligned} x_1 &= (0.54, 1.07) \\ x_2 &= (-1.20, -0.76) \\ x_3 &= (-0.63, -0.81) \\ x_4 &= (0.96, 0.75) \\ x_5 &= (1.64, 1.37) \end{aligned}$$

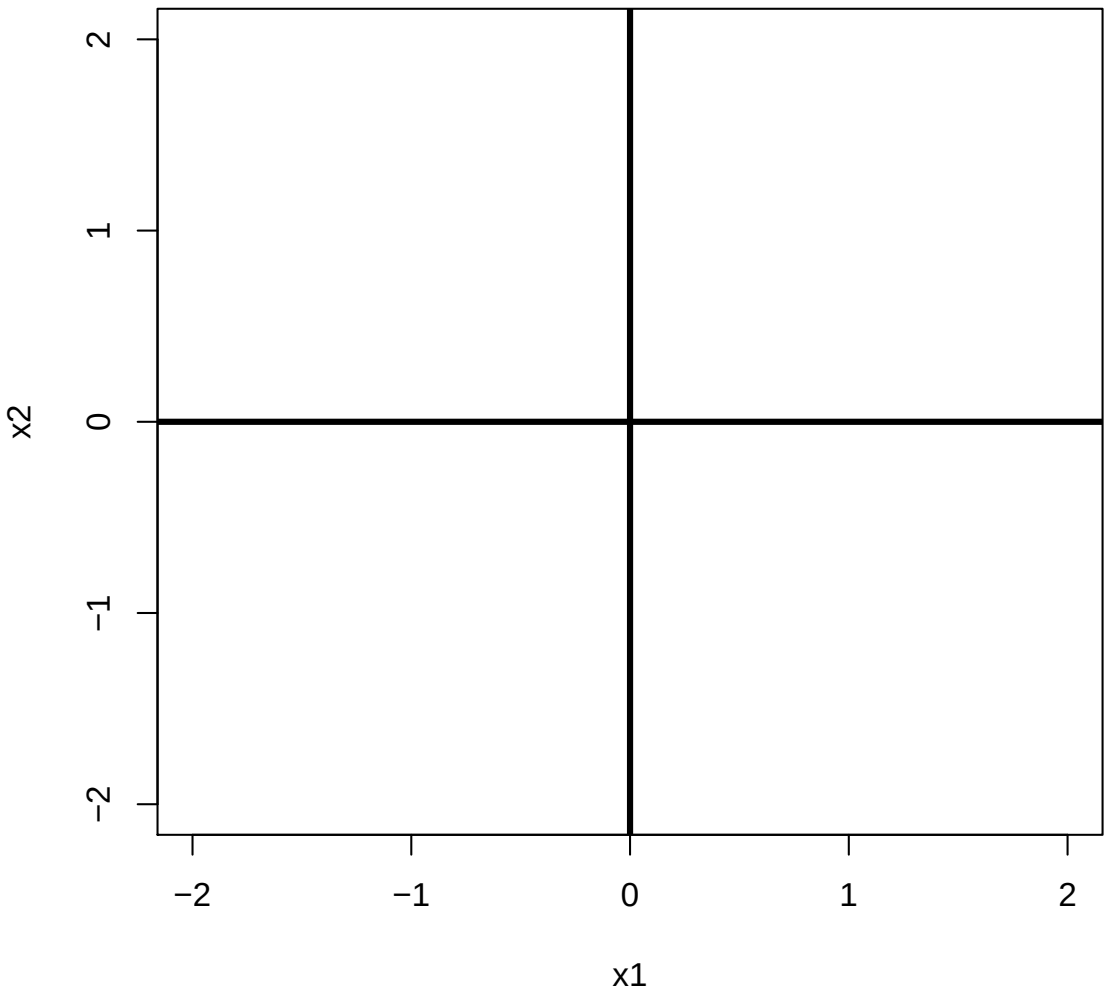

x2
2
1
0
−1
−2
−2
−1
0
1
2
x1

Goal: find line that summarizes bivariate information



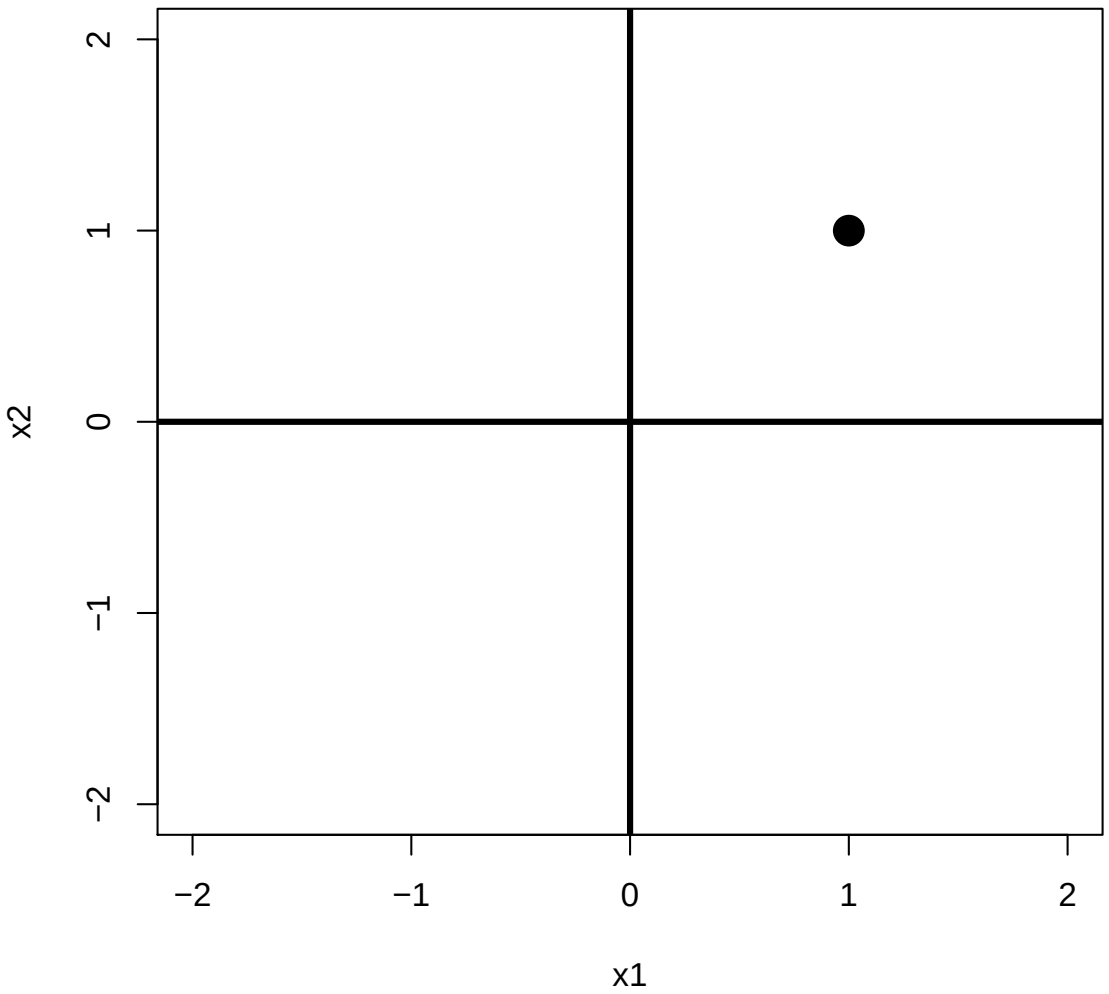
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1)$



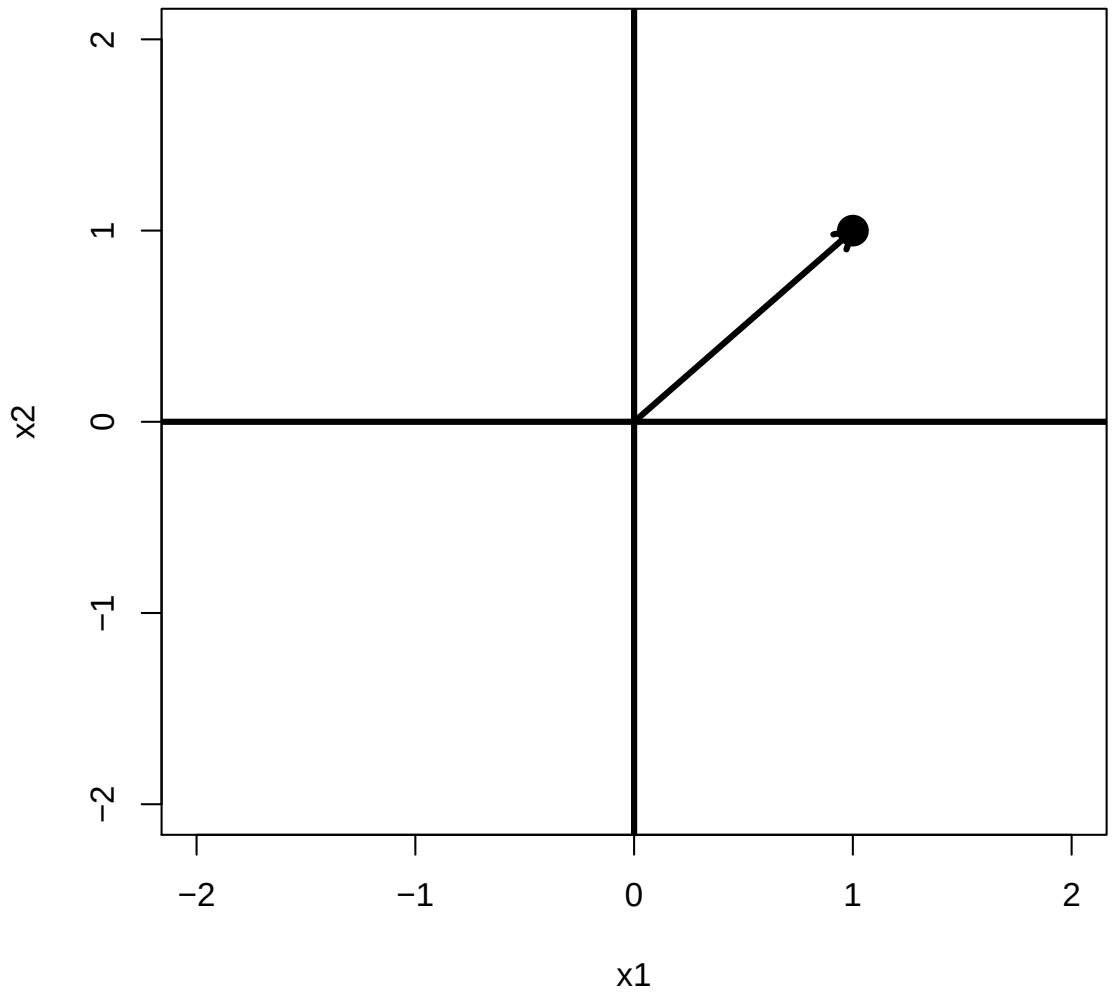
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1)$



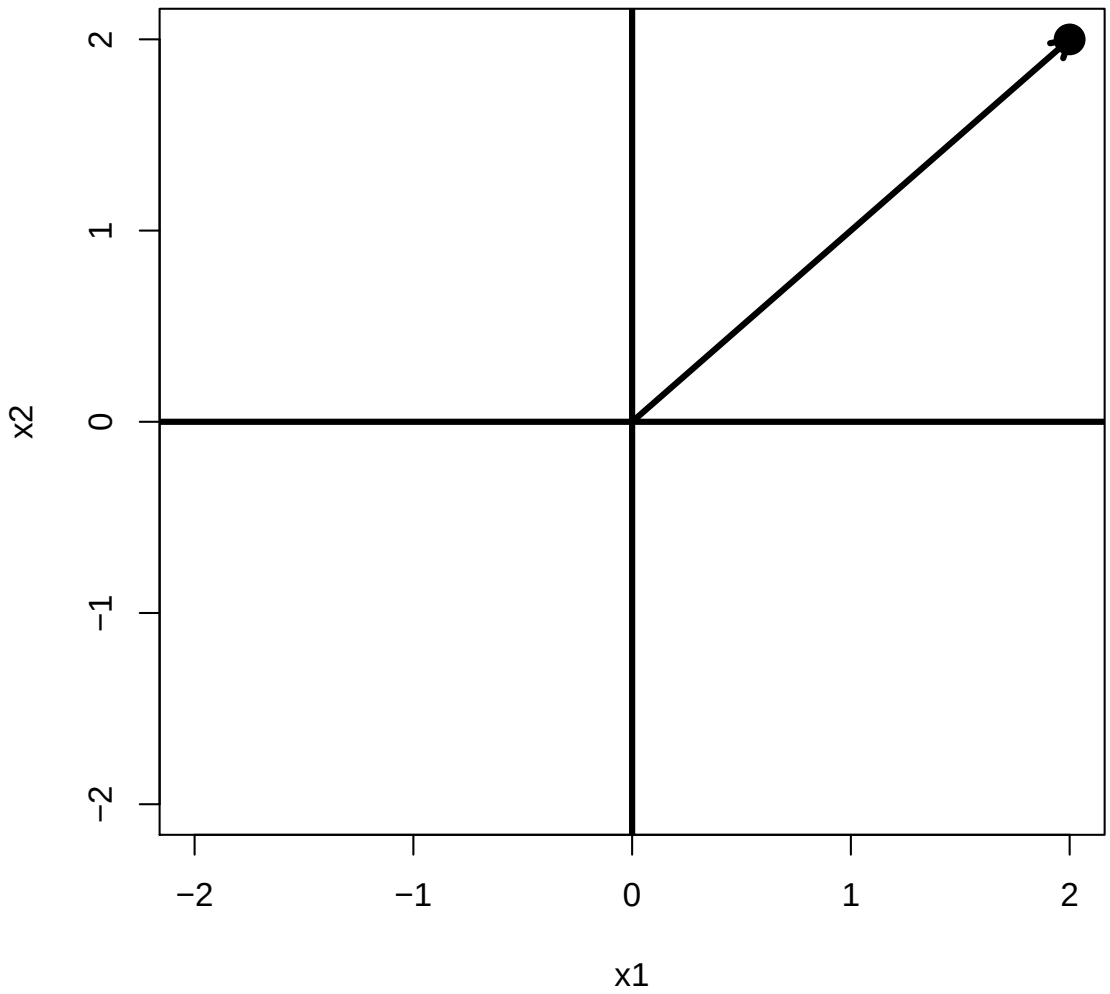
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1)$



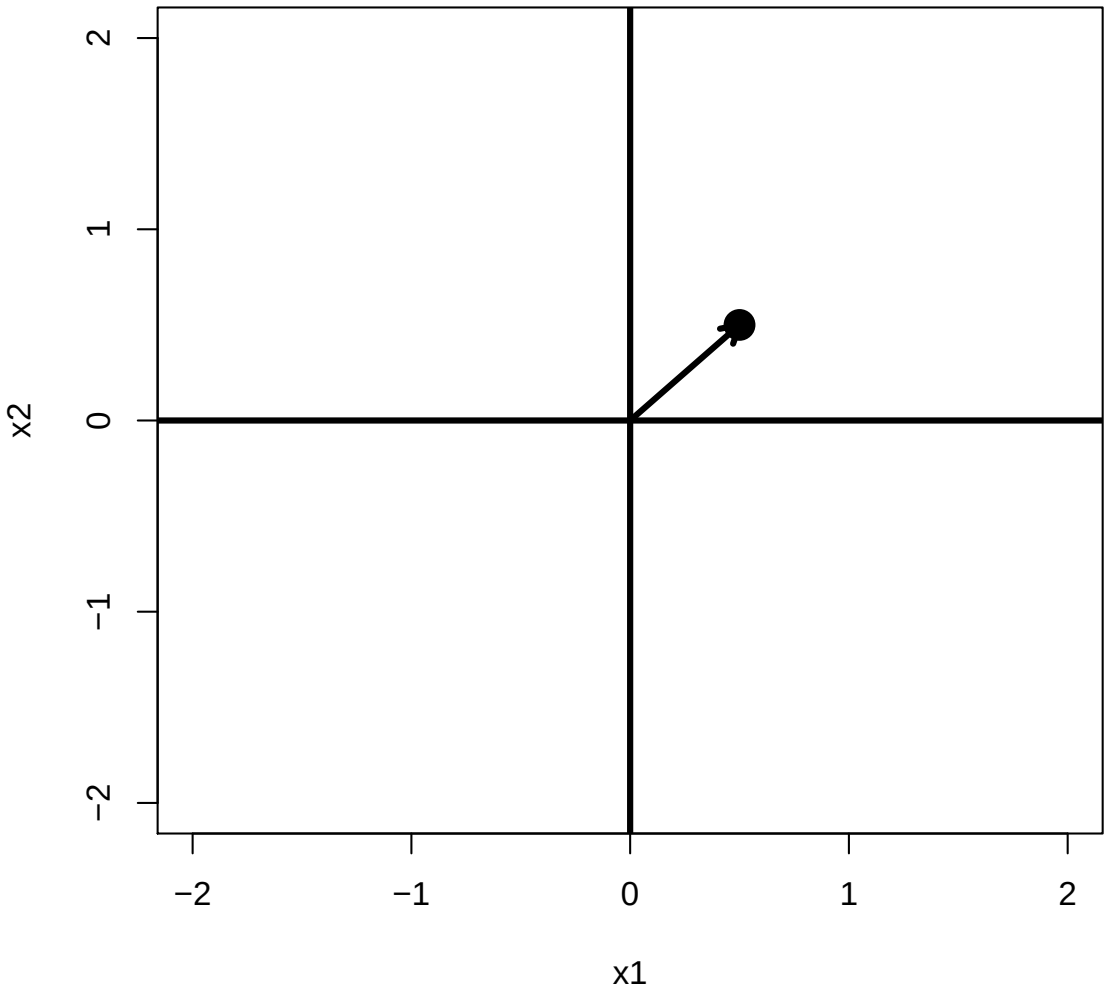
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1)$ $2\boldsymbol{w}_1 = (2, 2)$



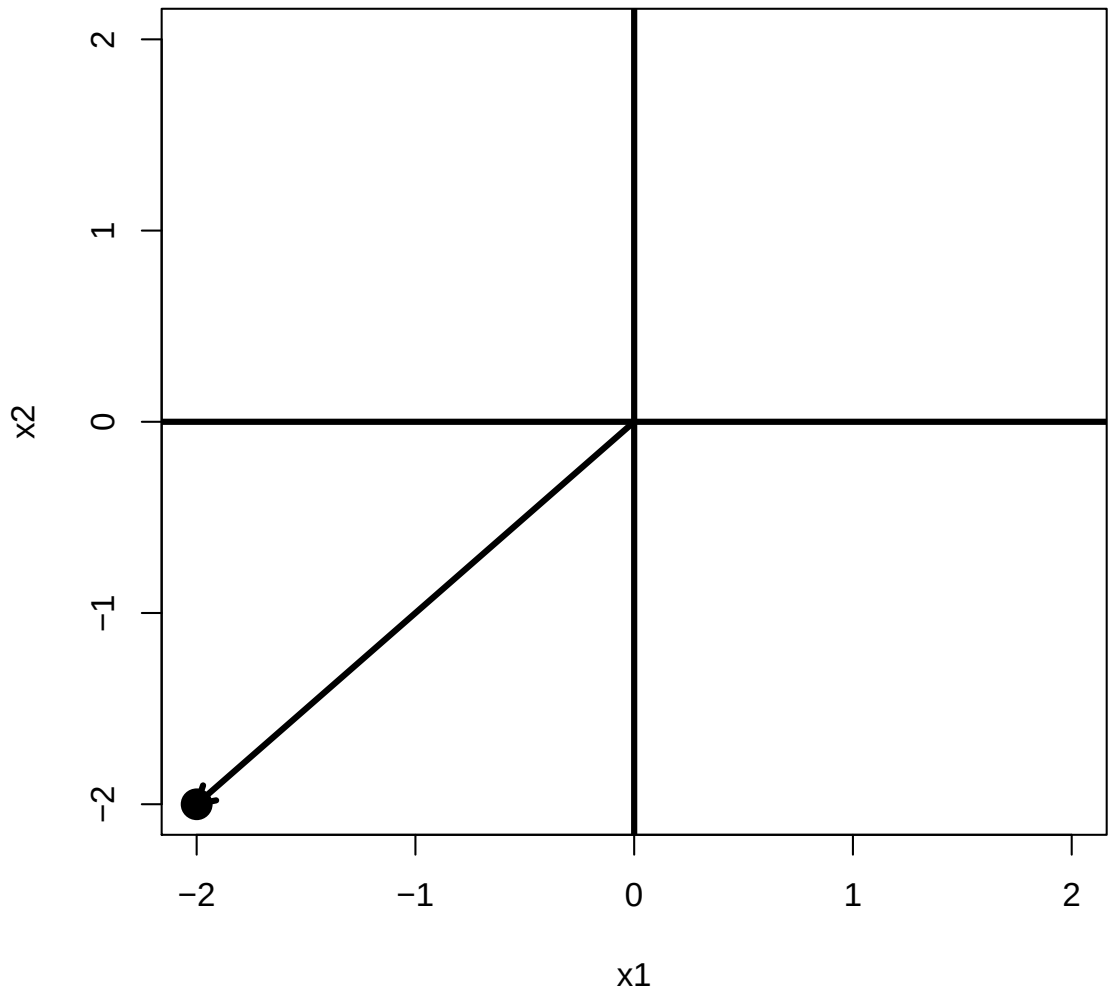
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1)$ $\frac{1}{2}\boldsymbol{w}_1 = (1/2, 1/2)$



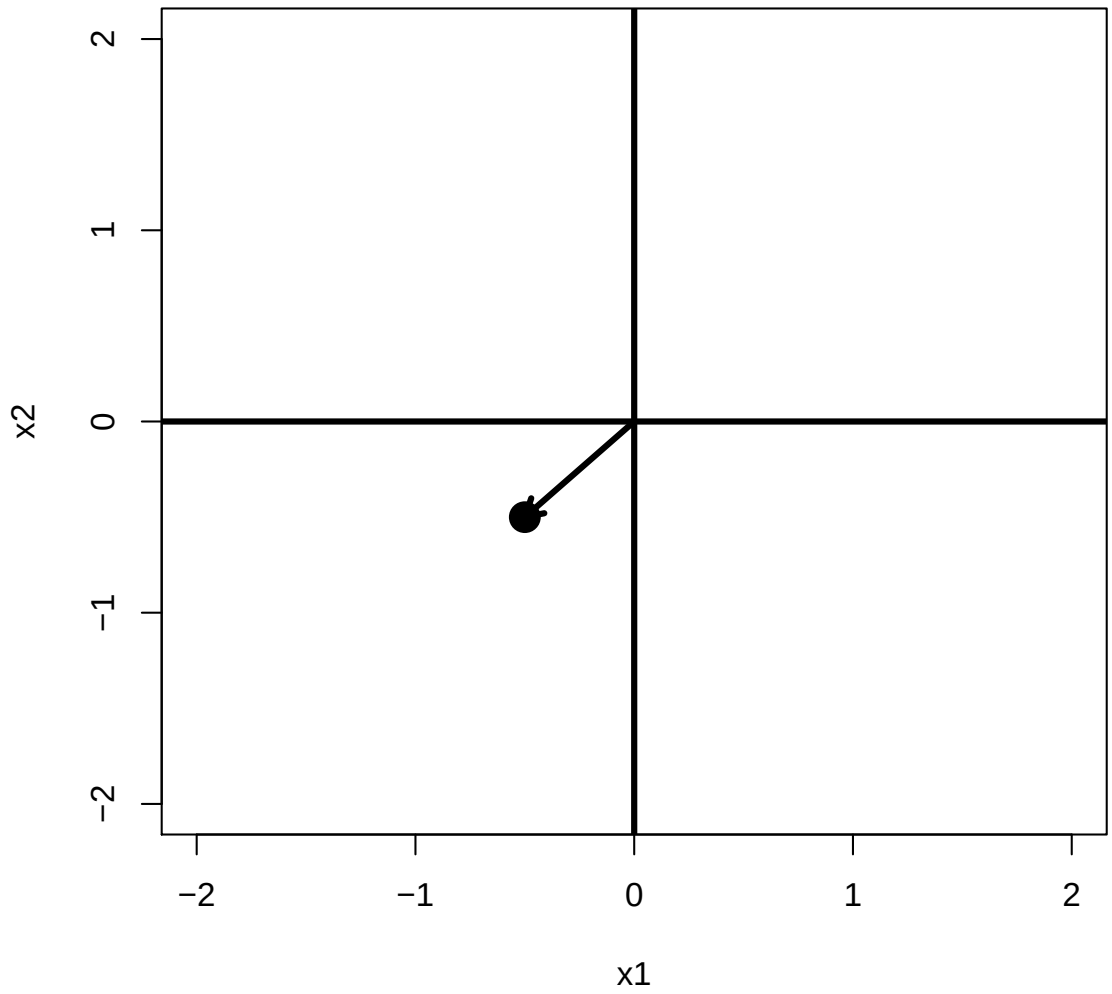
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1) \quad -2\boldsymbol{w}_1 = (-2, -2)$



Vectors to Draw a Line

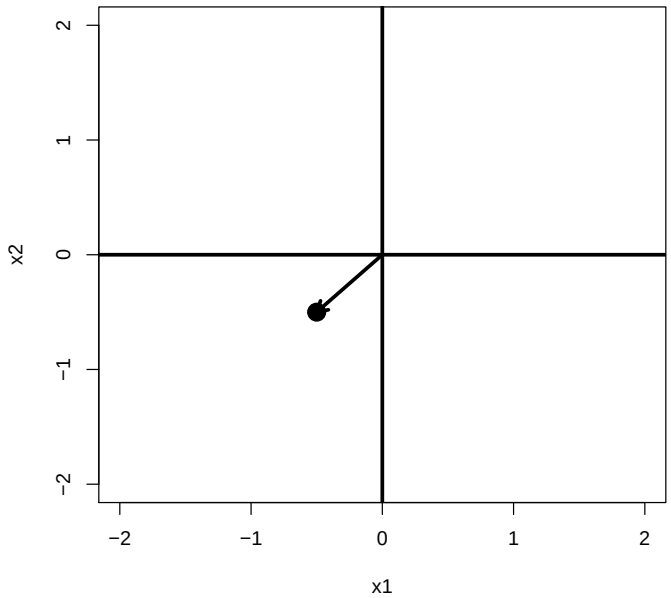
Suppose $\boldsymbol{w}_1 = (1, 1)$ $\quad -\frac{1}{2}\boldsymbol{w}_1 = (-1/2, -1/2)$



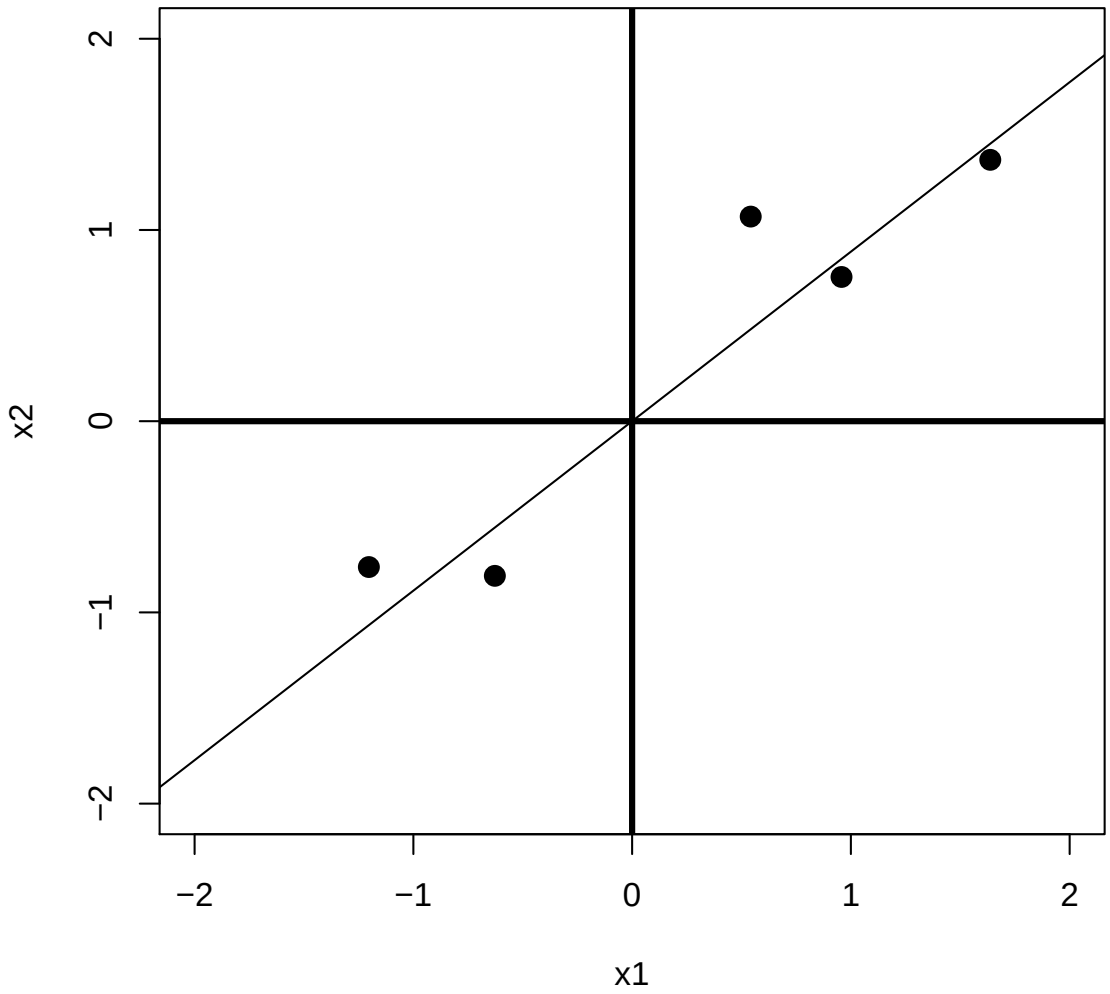
Vectors to Draw a Line

Suppose $\boldsymbol{w}_1 = (1, 1)$

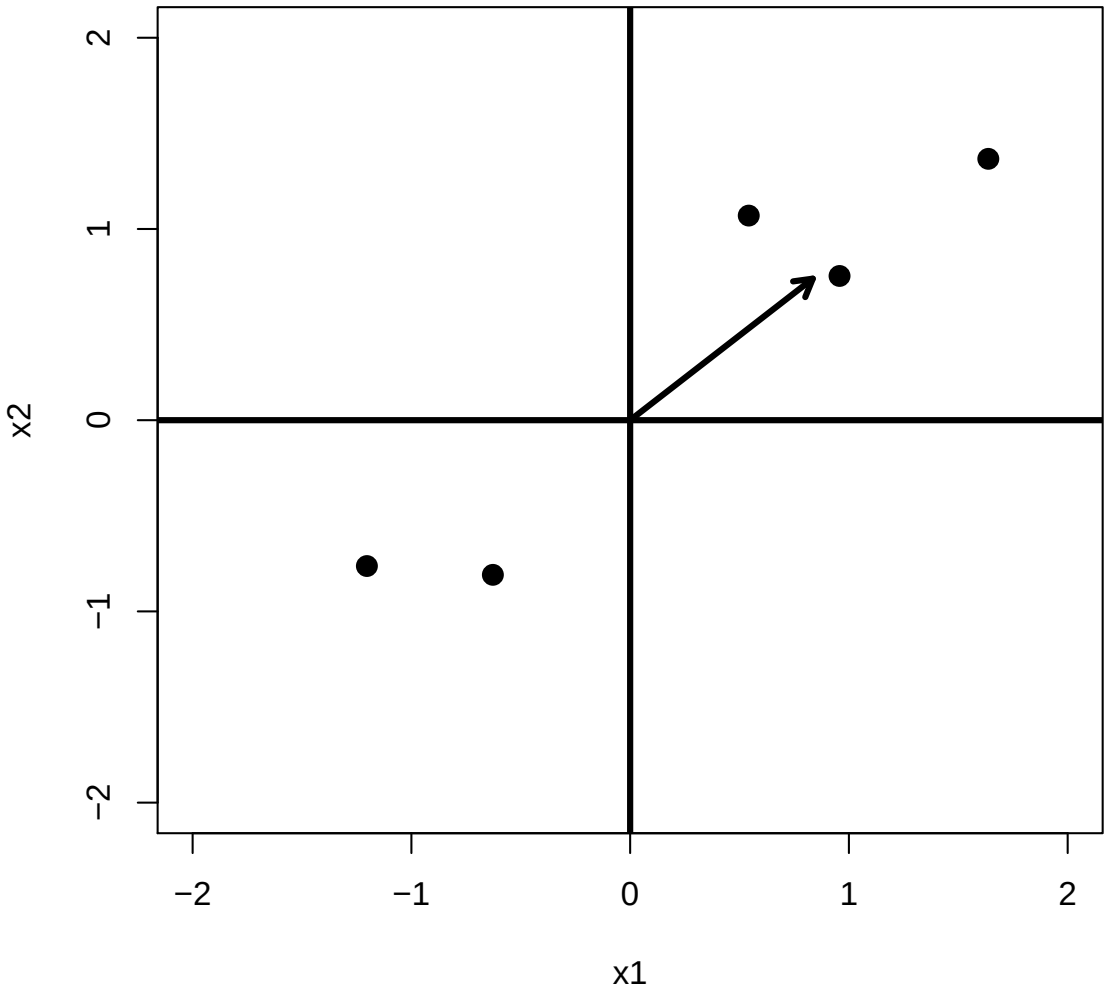
z_i = amount we shrink/flip $\boldsymbol{w}_1$ to approximate point i.



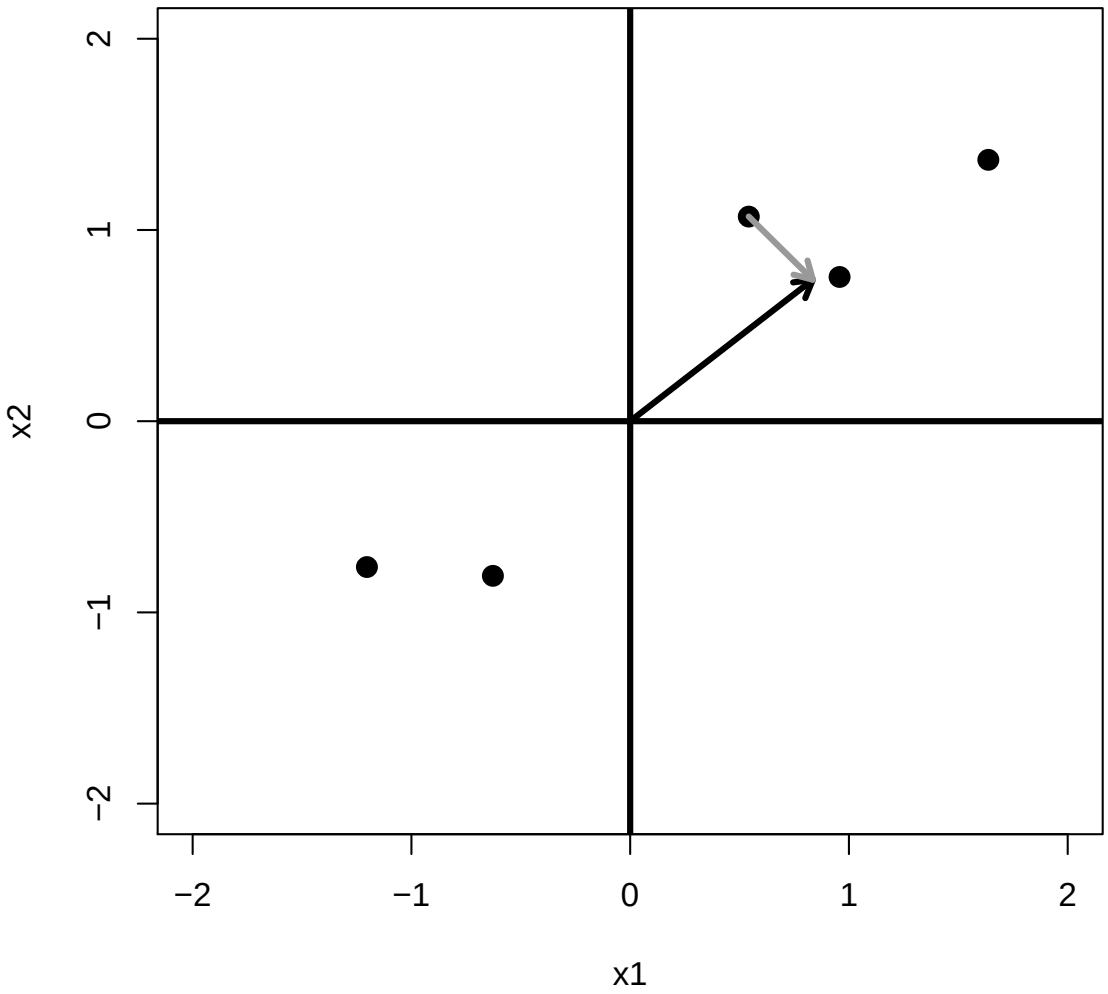
$\boldsymbol{w}_1 = (0.75, 0.66)$



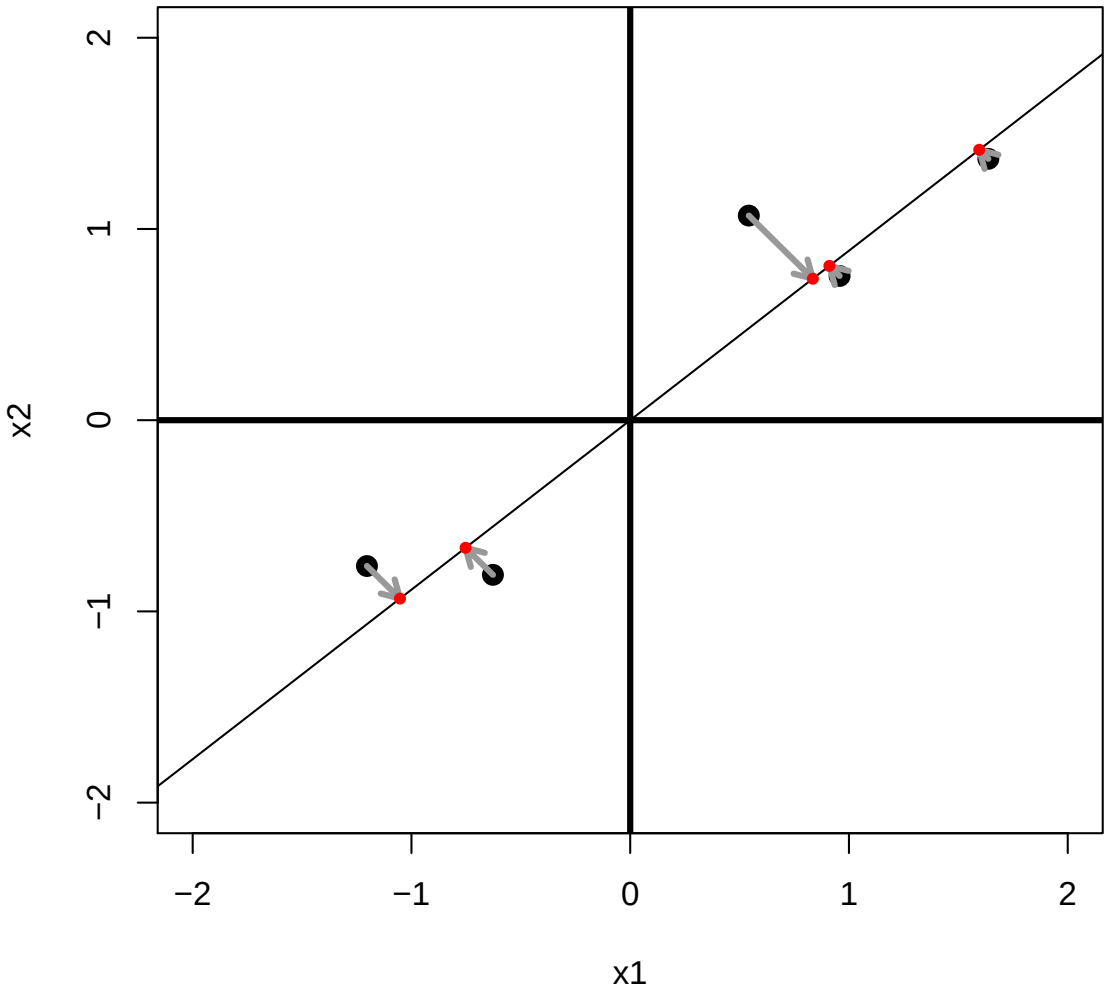
$\boldsymbol{w}_1 = (0.75, 0.66)$ $z_1 = 1.12$



$\boldsymbol{w}_1 = (0.75, 0.66)$ $z_1 = 1.12$



$\boldsymbol{w}_1 = (0.75, 0.66)$ $z_1 = 1.12$



Algebraic Representation

$$\boldsymbol{x}_i = z_i \boldsymbol{w}_1 + \boldsymbol{e}_i$$

Algebraic Representation

$$\begin{aligned} \boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\ (x_{i1}, x_{i2}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}) \end{aligned}$$

Algebraic Representation

$$\begin{aligned} \boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\ (x_{i1}, x_{i2}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}) \end{aligned}$$

Find $\boldsymbol{w}_1 = (w_{11}, w_{12})$ and z_i to minimize the error

Algebraic Representation

$$\begin{aligned} \boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\ (x_{i1}, x_{i2}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}) \end{aligned}$$

Find $\boldsymbol{w}_1 = (w_{11}, w_{12})$ and z_i to minimize the error

$$\text{error} = \frac{1}{N} \sum_{i=1}^{N} \big((x_{i1}, x_{i2}) - z_i(w_{11}, w_{12})\big)' \big((x_{i1}, x_{i2}) - z_i(w_{11}, w_{12})\big)$$

Algebraic Representation

$$\begin{aligned} \boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\ (x_{i1}, x_{i2}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}) \end{aligned}$$

Find $\boldsymbol{w}_1 = (w_{11}, w_{12})$ and z_i to minimize the error

$$\begin{aligned} \text{error} &= \frac{1}{N} \sum_{i=1}^{N} \big((x_{i1}, x_{i2}) - z_i(w_{11}, w_{12})\big)' \big((x_{i1}, x_{i2}) - z_i(w_{11}, w_{12})\big) \\ &= \frac{1}{N} \sum_{i=1}^{N} (x_{i1} - z_i w_{11})^2 + (x_{i2} - z_i w_{12})^2 \end{aligned}$$

Three Dimensional Approximation

$$
\begin{aligned}
\boldsymbol{x}_1 &= (0.09, -1.02, -0.10) \\
\boldsymbol{x}_2 &= (0.09, 1.41, 0.67) \\
\boldsymbol{x}_3 &= (-0.81, -1.46, -0.54) \\
\boldsymbol{x}_4 &= (1.43, 0.26, 0.61) \\
\boldsymbol{x}_5 &= (1.23, 0.87, 1.33
\end{aligned}
$$

Find $\boldsymbol{w}_1 = (w_{11}, w_{12}, w_{13})$ and z_i to provide best one dimensional approximation.

Three-Dimensional Visualization

Three-Dimensional Visualization

$\boldsymbol{w}_1 = (0.48, 0.75, 0.46)$

$$\boldsymbol{x}_i = z_i \boldsymbol{w}_1 + \boldsymbol{e}_i$$

$$
\begin{aligned}
\boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\
(x_{i1}, x_{i2}, x_{i3}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}, z_i w_{13} + e_{i3})
\end{aligned}
$$

$$\begin{aligned} \boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\ (x_{i1}, x_{i2}, x_{i3}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}, z_i w_{13} + e_{i3}) \end{aligned}$$

Find $\boldsymbol{w}_1 = (w_{11}, w_{12}, w_{13})$ and z_i to minimize the error

$$
\begin{aligned}
\boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\
(x_{i1}, x_{i2}, x_{i3}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}, z_i w_{13} + e_{i3})
\end{aligned}
$$

Find $\boldsymbol{w}_1 = (w_{11}, w_{12}, w_{13})$ and z_i to minimize the error

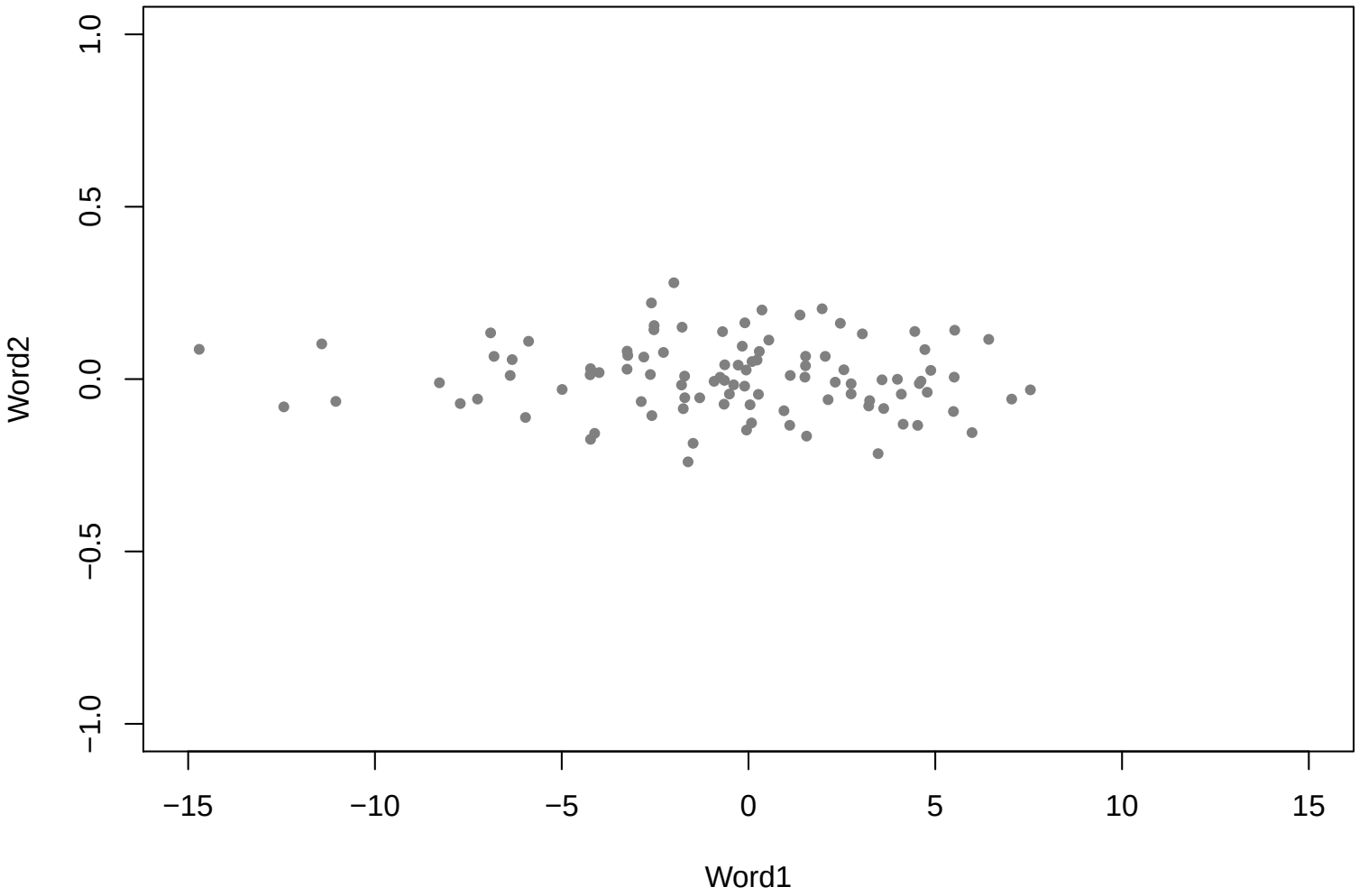
$$
\begin{aligned}
\text{error} &= \frac{1}{N} \sum_{i=1}^{N} \big((x_{i1}, x_{i2}, x_{13}) - z_i (w_{11}, w_{12}, w_{13})\big)' \\
&\quad \big((x_{i1}, x_{i2}, x_{i3}) - z_i (w_{11}, w_{12}, w_{13})\big)
\end{aligned}
$$

$$\begin{aligned} \boldsymbol{x}_i &= z_i \boldsymbol{w}_1 + \boldsymbol{e}_i \\ (x_{i1}, x_{i2}, x_{i3}) &= (z_i w_{11} + e_{i1}, z_i w_{12} + e_{i2}, z_i w_{13} + e_{i3}) \end{aligned}$$

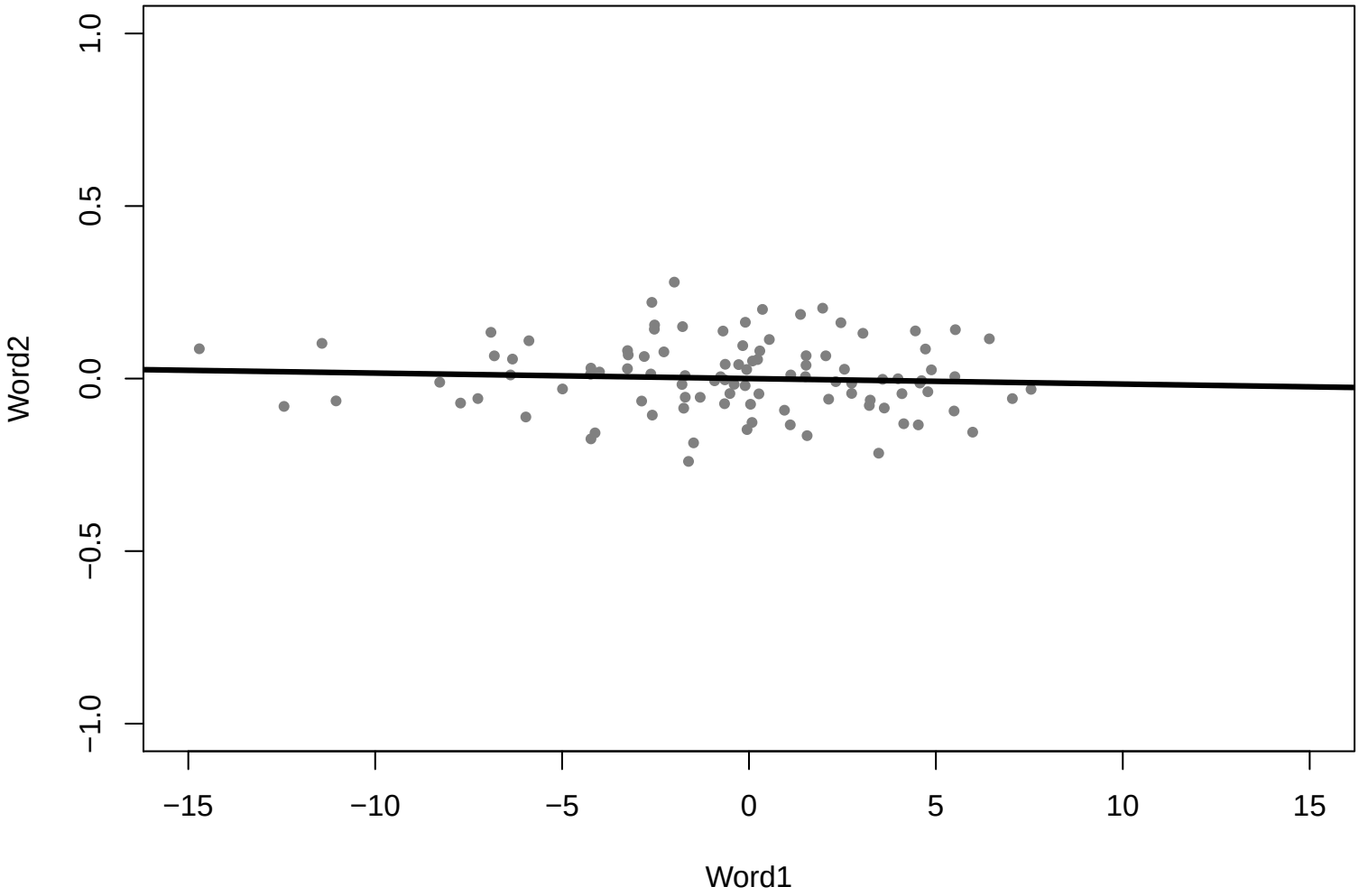
Find $\boldsymbol{w}_1 = (w_{11}, w_{12}, w_{13})$ and z_i to minimize the error

$$\begin{aligned} \text{error} &= \frac{1}{N} \sum_{i=1}^{N} \big((x_{i1}, x_{i2}, x_{13}) - z_i (w_{11}, w_{12}, w_{13})\big)' \\ &\quad \big((x_{i1}, x_{i2}, x_{i3}) - z_i (w_{11}, w_{12}, w_{13})\big) \\ &= \frac{1}{N} \sum_{i=1}^{N} (x_{i1} - z_i w_{11})^2 + (x_{i2} - z_i w_{12})^2 + (x_{i3} - z_i w_{13})^2 \end{aligned}$$

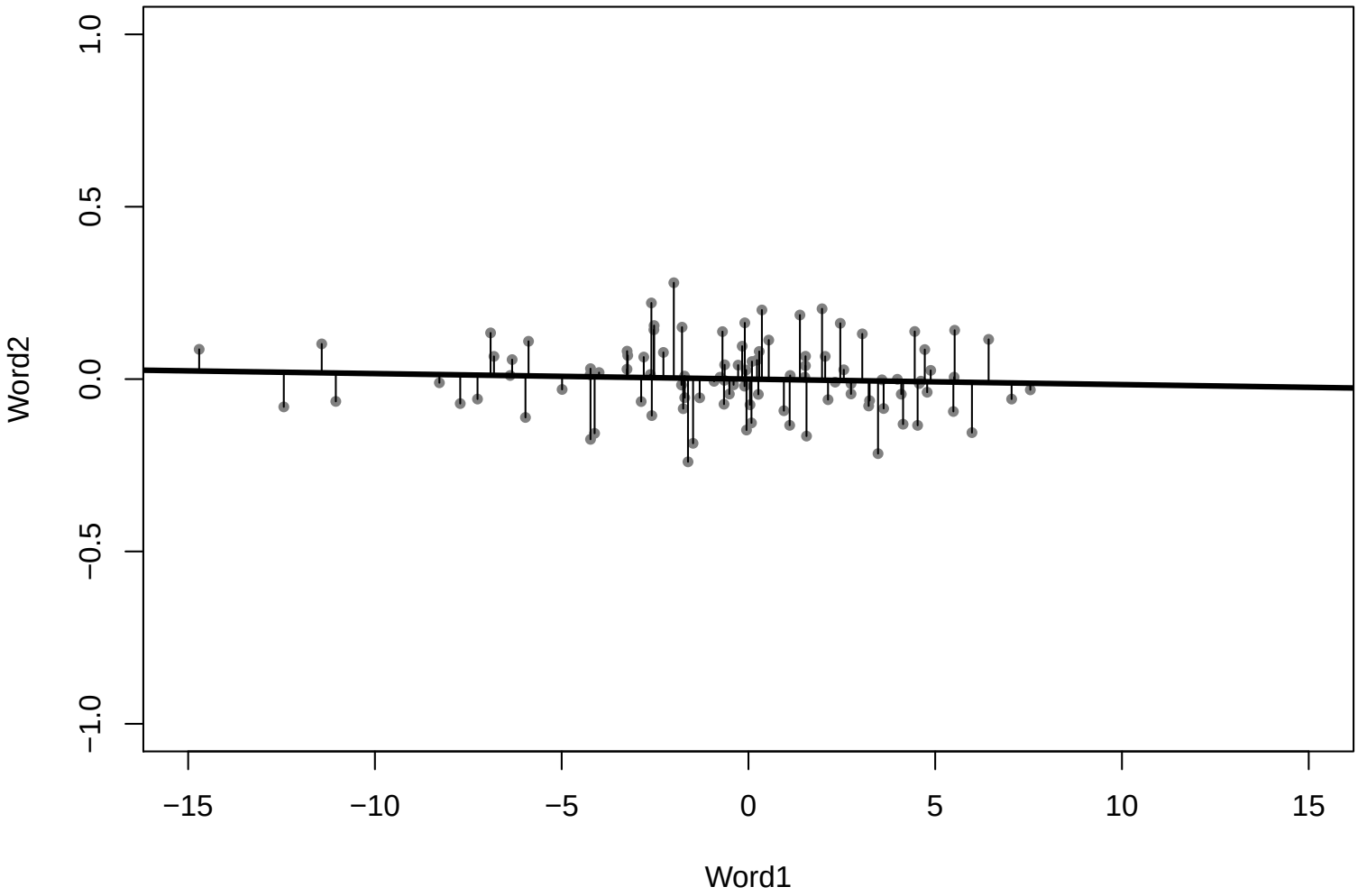
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



PCA Output

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

PCA Output

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

PCA Output

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

1) K Principal Components $\boldsymbol{w}_k$

PCA Output

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

1) K Principal Components $\boldsymbol{w}_k$

$$\boldsymbol{w}_k = (w_{1k}, w_{2k}, \ldots, w_{Jk})$$

PCA Output

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

1) K Principal Components $\boldsymbol{w}_k$

$$\boldsymbol{w}_k = (w_{1k}, w_{2k}, \ldots, w_{Jk})$$

2) K component vector describing loadings on principal components for each document

PCA Output

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

1) K Principal Components $\boldsymbol{w}_k$

$$\boldsymbol{w}_k = (w_{1k}, w_{2k}, \ldots, w_{Jk})$$

2) K component vector describing loadings on principal components for each document

$$\boldsymbol{z}_i = (z_{1i}, z_{2i}, \ldots, z_{Ki})$$

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

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

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

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*

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

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

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

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

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

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

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

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

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

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

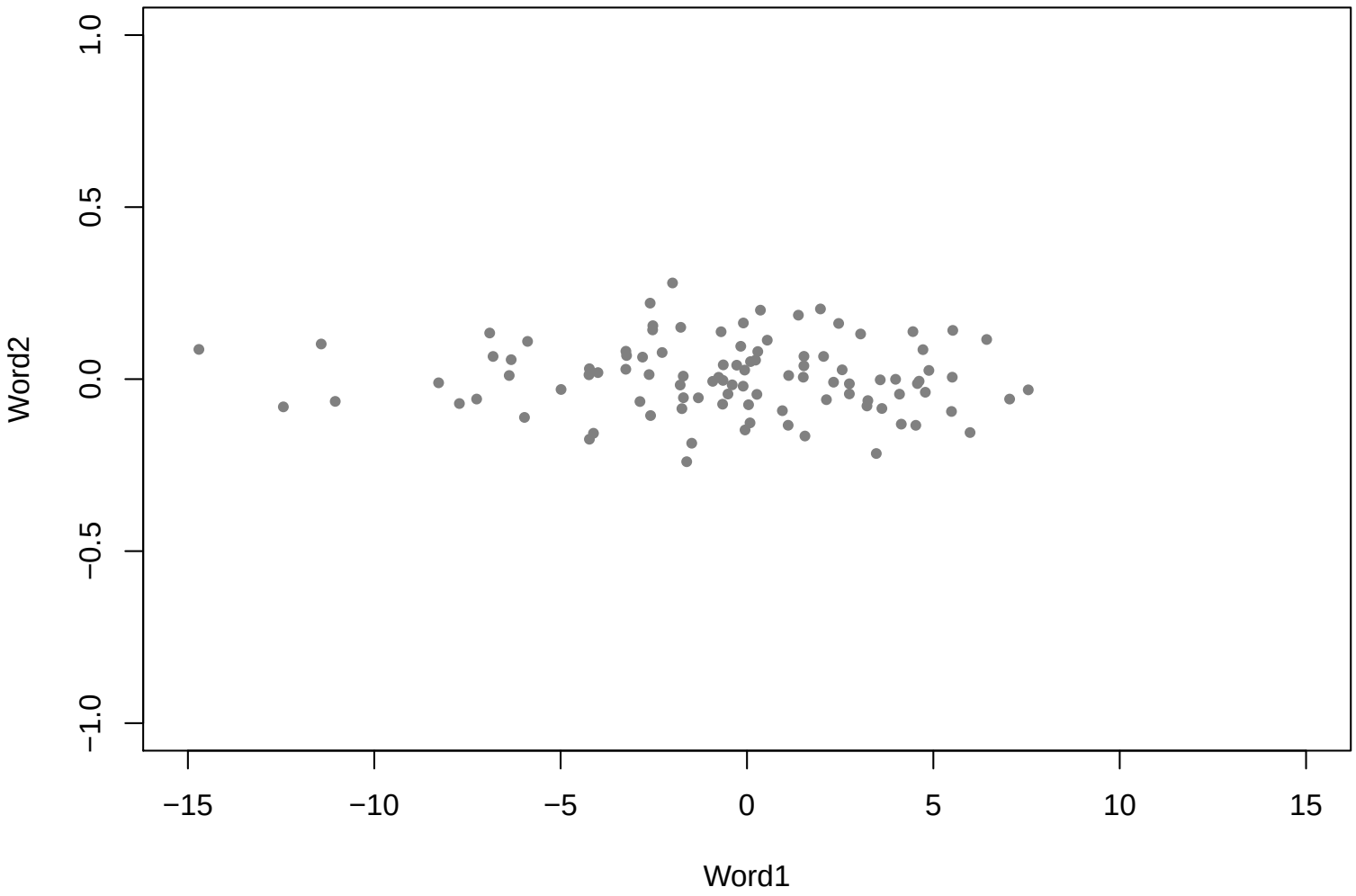
$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

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

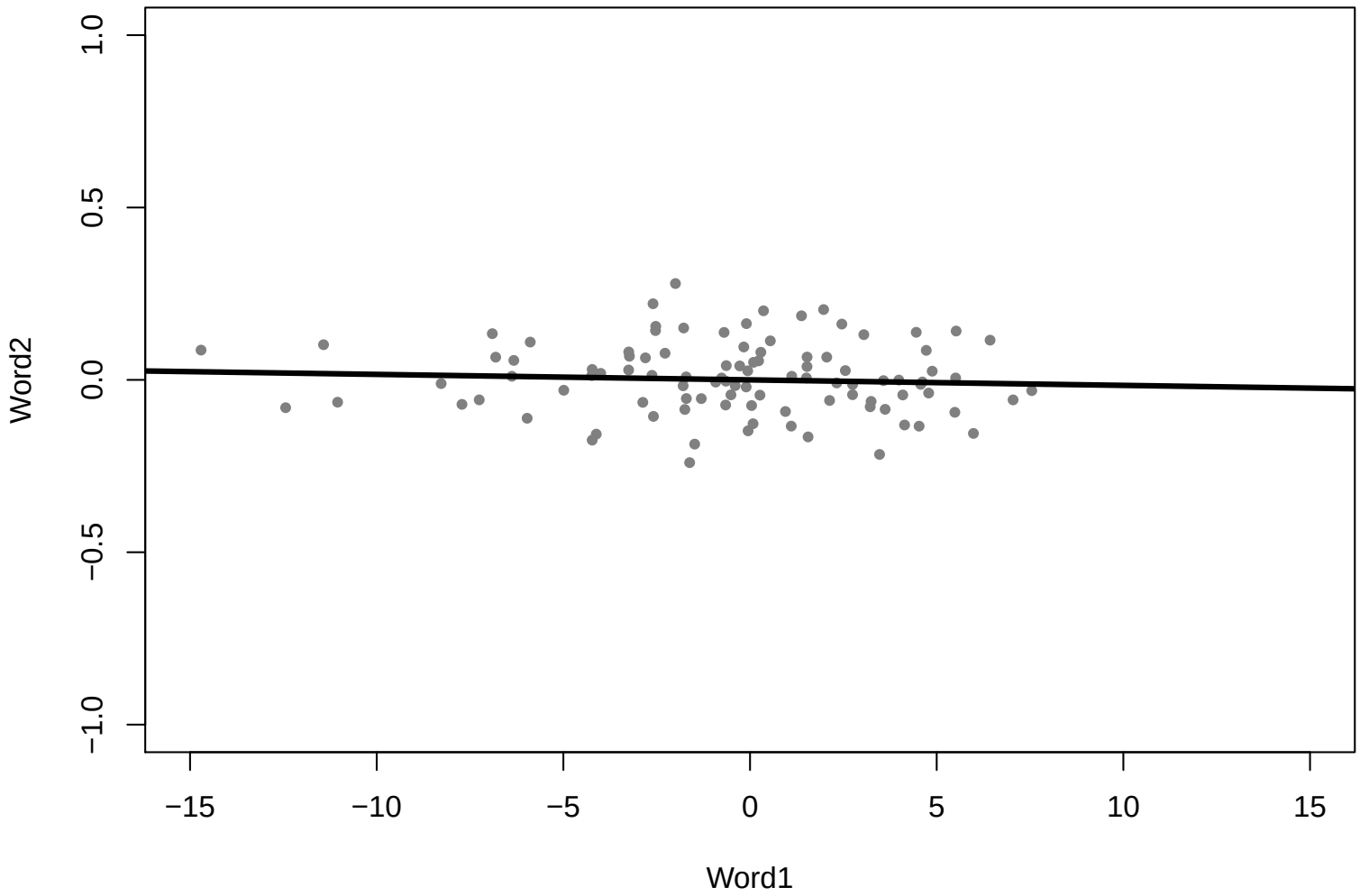
- $\boldsymbol{A}$ stretches the eigenvector $\boldsymbol{x}$
- $\boldsymbol{A}$ stretches $\boldsymbol{x}$ by λ
- To find eigenvectors/values: (`eigen` in `R`)
  - Find λ that solves $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$
  - Find vectors in null space of:

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

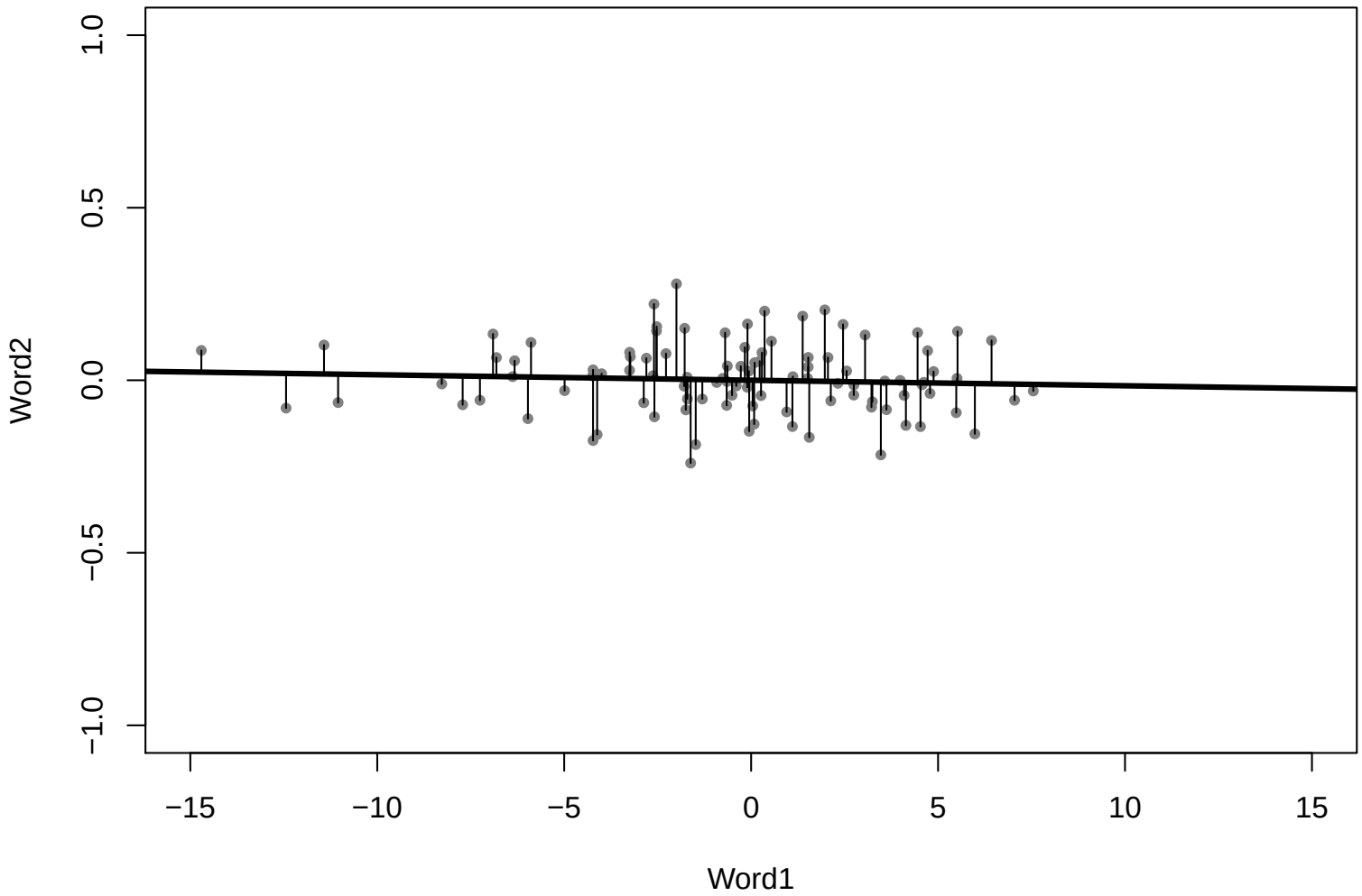
Finding a Lower Dimensional Space (Manifold Learning)



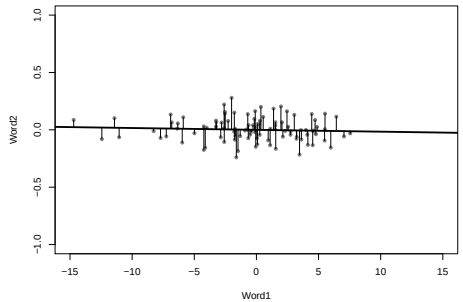
Finding a Lower Dimensional Space (Manifold Learning)



Finding a Lower Dimensional Space (Manifold Learning)

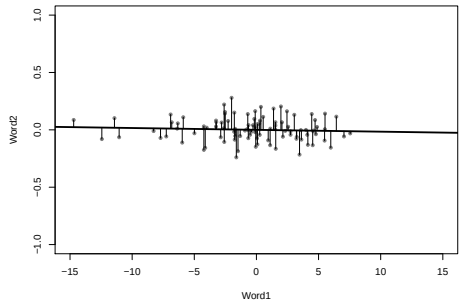


Finding a Lower Dimensional Space (Manifold Learning)



Original data:

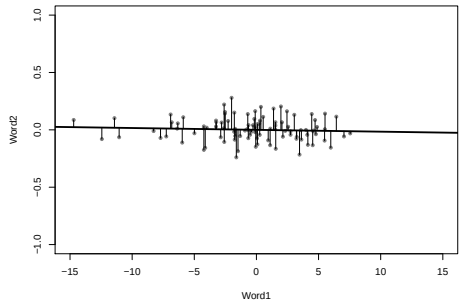
Finding a Lower Dimensional Space (Manifold Learning)



Original data:

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

Finding a Lower Dimensional Space (Manifold Learning)

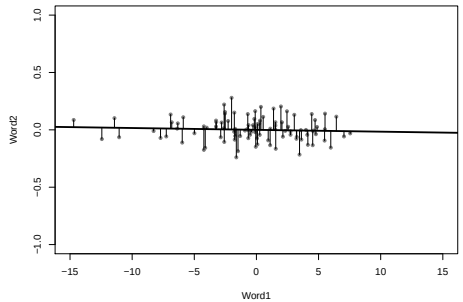


Original data:

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

Which we approximate with

Finding a Lower Dimensional Space (Manifold Learning)



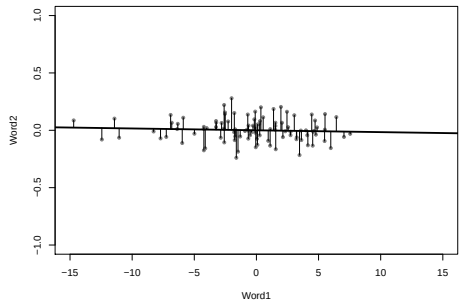
Original data:

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

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

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

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

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

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

Principal Component Analysis$\rightsquigarrow$ Objective function

Consider 1-dimensional case ($L = 1$), centered data, and $||\boldsymbol{w}_1|| = 1$.

Principal Component Analysis$\rightsquigarrow$ Objective function

Consider 1-dimensional case ($L = 1$), centered data, and $||\boldsymbol{w}_1|| = 1$.

$$f(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{x}_i - z_{i1} \boldsymbol{w}_1||^2$$

Principal Component Analysis$\rightsquigarrow$ Objective function

Consider 1-dimensional case ($L = 1$), centered data, and $||\boldsymbol{w}_1|| = 1$.

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{X}) &= \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1) \end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Objective function

Consider 1-dimensional case ($L = 1$), centered data, and $||\boldsymbol{w}_1|| = 1$.

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{X}) &= \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1) \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{x}_i'\boldsymbol{x}_i - 2z_{i1}\boldsymbol{w}_1'\boldsymbol{x}_i + z_{i1}^2\right) \end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Objective function

Consider 1-dimensional case ($L = 1$), centered data, and $||\boldsymbol{w}_1|| = 1$.

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{X}) &= \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}\boldsymbol{w}_1) \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{x}_i'\boldsymbol{x}_i - 2z_{i1}\boldsymbol{w}_1'\boldsymbol{x}_i + z_{i1}^2\right) \end{aligned}$$

$\boldsymbol{w}_1'\boldsymbol{w}_1 = 1$

Principal Component Analysis$\rightsquigarrow$ Optimization

Optimization:

Principal Component Analysis$\rightsquigarrow$ Optimization

Optimization:

$$\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{X})}{\partial z_{i1}} = -\frac{2\boldsymbol{w}_1'\boldsymbol{x}_i + 2z_{i1}}{N}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Optimization:

$$\begin{aligned} \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{X})}{\partial z_{i1}} &= -\frac{2\boldsymbol{w}_1' \boldsymbol{x}_i + 2z_{i1}}{N} \\ 0 &= -\frac{2\boldsymbol{w}_1' \boldsymbol{x}_i + 2z_{i1}^*}{N} \end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Optimization:

$$\begin{aligned}
\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{X})}{\partial z_{i1}} &= -\frac{2\boldsymbol{w}_1^{'}\boldsymbol{x}_i + 2z_{i1}}{N} \\
0 &= -\frac{2\boldsymbol{w}_1^{'}\boldsymbol{x}_i + 2z_{i1}^*}{N} \\
z_{i1}^* &= \boldsymbol{w}_1^{'}\boldsymbol{x}_i
\end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Substituting in z_{i1}^*

Principal Component Analysis$\rightsquigarrow$ Optimization

Substituting in z_{i1}^*

$$= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1)$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Substituting in z_{i1}^*

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\underbrace{\boldsymbol{x}_i'\boldsymbol{x}_i}_{\text{Constant}} - 2z_{i1}^* \underbrace{\boldsymbol{w}_1'\boldsymbol{x}_i}_{z_{i1}^*} + (z_{i1}^*)^2 \underbrace{\boldsymbol{w}_1'\boldsymbol{w}_1}_{1})
\end{aligned}
$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Substituting in z_{i1}^*

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\underbrace{\boldsymbol{x}_i'\boldsymbol{x}_i}_{\text{Constant}} - 2z_{i1}^* \underbrace{\boldsymbol{w}_1'\boldsymbol{x}_i}_{z_{i1}^*} + (z_{i1}^*)^2 \underbrace{\boldsymbol{w}_1'\boldsymbol{w}_1}_{1}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(z_{i1}^*)^2 + c
\end{aligned}
$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Substituting in z_{i1}^*

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\underbrace{\boldsymbol{x}_i'\boldsymbol{x}_i}_{\text{Constant}} - 2z_{i1}^* \underbrace{\boldsymbol{w}_1'\boldsymbol{x}_i}_{z_{i1}^*} + (z_{i1}^*)^2 \underbrace{\boldsymbol{w}_1'\boldsymbol{w}_1}_{1}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(z_{i1}^*)^2 + c \\
&= -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_1'\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{w}_1
\end{aligned}
$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Substituting in z_{i1}^*

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1)'(\boldsymbol{x}_i - z_{i1}^*\boldsymbol{w}_1) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\underbrace{\boldsymbol{x}_i'\boldsymbol{x}_i}_{\text{Constant}} - 2z_{i1}^* \underbrace{\boldsymbol{w}_1'\boldsymbol{x}_i}_{z_{i1}^*} + (z_{i1}^*)^2 \underbrace{\boldsymbol{w}_1'\boldsymbol{w}_1}_{1}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(z_{i1}^*)^2 + c \\
&= -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_1'\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{w}_1 \\
&= -\boldsymbol{w}_1'\boldsymbol{\Sigma}\boldsymbol{w}_1
\end{aligned}
$$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

$$\boldsymbol{z}_1 \quad = \quad (\boldsymbol{w}_1\boldsymbol{x}_1, \boldsymbol{w}_1\boldsymbol{x}_2, \ldots, \boldsymbol{w}_1\boldsymbol{x}_N)$$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

$$\begin{aligned} \boldsymbol{z}_1 &= (\boldsymbol{w}_1\boldsymbol{x}_1, \boldsymbol{w}_1\boldsymbol{x}_2, \ldots, \boldsymbol{w}_1\boldsymbol{x}_N) \\ \text{var}(\boldsymbol{z}_1) &= E[\boldsymbol{z}_1^2] - E[\boldsymbol{z}_1]^2 \end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

$$\begin{aligned} \boldsymbol{z}_1 &= (\boldsymbol{w}_1\boldsymbol{x}_1, \boldsymbol{w}_1\boldsymbol{x}_2, \ldots, \boldsymbol{w}_1\boldsymbol{x}_N) \\ \text{var}(\boldsymbol{z}_1) &= E[\boldsymbol{z}_1^2] - E[\boldsymbol{z}_1]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} z_{i1}^2 - 0 \end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

$$\begin{aligned}
\boldsymbol{z}_1 &= (\boldsymbol{w}_1\boldsymbol{x}_1, \boldsymbol{w}_1\boldsymbol{x}_2, \ldots, \boldsymbol{w}_1\boldsymbol{x}_N) \\
\text{var}(\boldsymbol{z}_1) &= E[\boldsymbol{z}_1^2] - E[\boldsymbol{z}_1]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} z_{i1}^2 - 0 \\
&= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_1^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_1 = \boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1
\end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

$$\begin{aligned}
\boldsymbol{z}_1 &= (\boldsymbol{w}_1\boldsymbol{x}_1, \boldsymbol{w}_1\boldsymbol{x}_2, \ldots, \boldsymbol{w}_1\boldsymbol{x}_N) \\
\text{var}(\boldsymbol{z}_1) &= E[\boldsymbol{z}_1^2] - E[\boldsymbol{z}_1]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} z_{i1}^2 - 0 \\
&= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_1^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_1 = \boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1
\end{aligned}$$

Minimize reconstruction error

Principal Component Analysis$\rightsquigarrow$ Optimization

$$= \quad -\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1$$

where $\boldsymbol{\Sigma}$ is the :

- Empirical covariance matrix$\rightsquigarrow$ $\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X}$
- Variance of the projected data. Define

$$\begin{aligned}
\boldsymbol{z}_1 &= (\boldsymbol{w}_1\boldsymbol{x}_1, \boldsymbol{w}_1\boldsymbol{x}_2, \ldots, \boldsymbol{w}_1\boldsymbol{x}_N) \\
\text{var}(\boldsymbol{z}_1) &= E[\boldsymbol{z}_1^2] - E[\boldsymbol{z}_1]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} z_{i1}^2 - 0 \\
&= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_1^{'}\boldsymbol{x}_i\boldsymbol{x}_i^{'}\boldsymbol{w}_1 = \boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1
\end{aligned}$$

Minimize reconstruction error $\rightsquigarrow$ maximize variance of projected data

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) \quad = \quad \boldsymbol{w}_1^{'} \boldsymbol{\Sigma} \boldsymbol{w}_1 - \lambda_1 (\boldsymbol{w}_1^{'} \boldsymbol{w}_1 - 1)$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$\begin{aligned} g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'} \boldsymbol{\Sigma} \boldsymbol{w}_1 - \lambda_1 (\boldsymbol{w}_1^{'} \boldsymbol{w}_1 - 1) \\ \frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma} \boldsymbol{w}_1 - 2\lambda_1 \boldsymbol{w}_1 \end{aligned}$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$
\begin{aligned}
g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'} \boldsymbol{\Sigma} \boldsymbol{w}_1 - \lambda_1 (\boldsymbol{w}_1^{'} \boldsymbol{w}_1 - 1) \\
\frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma} \boldsymbol{w}_1 - 2\lambda_1 \boldsymbol{w}_1 \\
\boldsymbol{\Sigma} \boldsymbol{w}_1^* &= \lambda_1 \boldsymbol{w}_1^*
\end{aligned}
$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$
\begin{aligned}
g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'} \boldsymbol{\Sigma} \boldsymbol{w}_1 - \lambda_1 (\boldsymbol{w}_1^{'} \boldsymbol{w}_1 - 1) \\
\frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma} \boldsymbol{w}_1 - 2\lambda_1 \boldsymbol{w}_1 \\
\boldsymbol{\Sigma} \boldsymbol{w}_1^* &= \lambda_1 \boldsymbol{w}_1^*
\end{aligned}
$$

$\boldsymbol{w}_1^* =$ Eigenvector of $\boldsymbol{\Sigma}$

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$\begin{aligned} g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'} \boldsymbol{\Sigma} \boldsymbol{w}_1 - \lambda_1 (\boldsymbol{w}_1^{'} \boldsymbol{w}_1 - 1) \\ \frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma} \boldsymbol{w}_1 - 2\lambda_1 \boldsymbol{w}_1 \\ \boldsymbol{\Sigma} \boldsymbol{w}_1^* &= \lambda_1 \boldsymbol{w}_1^* \end{aligned}$$

$\boldsymbol{w}_1^*$ = Eigenvector of $\boldsymbol{\Sigma}$ (!!!!!!)

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$\begin{aligned} g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'} \boldsymbol{\Sigma} \boldsymbol{w}_1 - \lambda_1 (\boldsymbol{w}_1^{'} \boldsymbol{w}_1 - 1) \\ \frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma} \boldsymbol{w}_1 - 2\lambda_1 \boldsymbol{w}_1 \\ \boldsymbol{\Sigma} \boldsymbol{w}_1^* &= \lambda_1 \boldsymbol{w}_1^* \end{aligned}$$

$\boldsymbol{w}_1^*$ = Eigenvector of $\boldsymbol{\Sigma}$ (!!!!!!)

We want $\boldsymbol{w}_1$ to maximize variance and

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$\begin{aligned} g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1 - \lambda_1(\boldsymbol{w}_1^{'}\boldsymbol{w}_1 - 1) \\ \frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma}\boldsymbol{w}_1 - 2\lambda_1\boldsymbol{w}_1 \\ \boldsymbol{\Sigma}\boldsymbol{w}_1^* &= \lambda_1\boldsymbol{w}_1^* \end{aligned}$$

$\boldsymbol{w}_1^*$ = Eigenvector of $\boldsymbol{\Sigma}$ (!!!!!!)
We want $\boldsymbol{w}_1$ to maximize variance and

$$\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1 = \lambda_1$$

Principal Component Analysis$\rightsquigarrow$ Optimization

Maximize variance, subject to constraints

$$\begin{aligned} g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X}) &= \boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1 - \lambda_1(\boldsymbol{w}_1^{'}\boldsymbol{w}_1 - 1) \\ \frac{\partial g(\boldsymbol{z}^*, \boldsymbol{w}_1, \boldsymbol{X})}{\partial \boldsymbol{w}_1} &= 2\boldsymbol{\Sigma}\boldsymbol{w}_1 - 2\lambda_1\boldsymbol{w}_1 \\ \boldsymbol{\Sigma}\boldsymbol{w}_1^* &= \lambda_1\boldsymbol{w}_1^* \end{aligned}$$

$\boldsymbol{w}_1^*$ = Eigenvector of $\boldsymbol{\Sigma}$ (!!!!!!)
We want $\boldsymbol{w}_1$ to maximize variance and

$$\boldsymbol{w}_1^{'}\boldsymbol{\Sigma}\boldsymbol{w}_1 = \lambda_1$$

So $\boldsymbol{w}_1$ is eigenvector associated with the largest eigenvalue λ_1

An Introduction to Eigenvectors, Values, and Diagonalization

Theorem

Suppose $\boldsymbol{A}$ *is an invertible* $N \times N$ *matrix with* N *linearly independent eigenvectors. Then we can write* $\boldsymbol{A}$ *as,*

$$\boldsymbol{A} = \boldsymbol{W}' \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{pmatrix} \boldsymbol{W}$$

where $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_N)$ *is an* $N \times N$ *matrix with the* N *eigenvectors as column vectors.*

An Introduction to Eigenvectors, Values, and Diagonalization

Definition

Suppose A is a covariance matrix. Then, we can write A as

$$\boldsymbol{A} = \boldsymbol{W}' \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{pmatrix} \boldsymbol{W}$$

Where $\lambda_1 > \lambda_2 > \ldots > \lambda_N \geq 0$.
We will call $\boldsymbol{w}_1$ *the first eigenvector,* $\boldsymbol{w}_2$ *the second eigenvector, ...,* $\boldsymbol{w}_j$ *the* j^{th} *eigenvector.*

Back to Principal Components

Theorem

Suppose we want to approximate N observations $\boldsymbol{x}_i \in \Re^J$ with $L < J$ orthogonal-unit length vectors $\boldsymbol{w}_l \in \Re^J$ with associated scores z_{il} to minimize reconstruction error:

Back to Principal Components

Theorem

Suppose we want to approximate N observations $\boldsymbol{x}_i \in \Re^J$ with $L < J$ orthogonal-unit length vectors $\boldsymbol{w}_l \in \Re^J$ with associated scores z_{il} to minimize reconstruction error:

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

Back to Principal Components

Theorem

Suppose we want to approximate N observations $\boldsymbol{x}_i \in \Re^J$ with $L < J$ orthogonal-unit length vectors $\boldsymbol{w}_l \in \Re^J$ with associated scores z_{il} to minimize reconstruction error:

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

The optimal solution sets each $\boldsymbol{w}_l$ to be the l^{th} eigenvector of the empirical covariance matrix.

Back to Principal Components

Theorem

Suppose we want to approximate N observations $\boldsymbol{x}_i \in \Re^J$ *with* $L < J$ *orthogonal-unit length vectors* $\boldsymbol{w}_l \in \Re^J$ *with associated scores* z_{il} *to minimize reconstruction error:*

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

The optimal solution sets each $\boldsymbol{w}_l$ *to be the* l^{th} *eigenvector of the empirical covariance matrix.Further* $z^*_{il} = \boldsymbol{w}'_l \boldsymbol{x}_i$ *so that the L dimensional representation is:*

Back to Principal Components

Theorem

Suppose we want to approximate N observations $\boldsymbol{x}_i \in \Re^J$ *with* $L < J$ *orthogonal-unit length vectors* $\boldsymbol{w}_l \in \Re^J$ *with associated scores* z_{il} *to minimize reconstruction error:*

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

The optimal solution sets each $\boldsymbol{w}_l$ *to be the* l^{th} *eigenvector of the empirical covariance matrix.Further* $z_{il}^* = \boldsymbol{w}_l^{'} \boldsymbol{x}_i$ *so that the L dimensional representation is:*

$$\boldsymbol{x}_i^L = (\boldsymbol{w}_1^{'} \boldsymbol{x}_i, \boldsymbol{w}_2^{'} \boldsymbol{x}_i, \ldots, \boldsymbol{w}_L^{'} \boldsymbol{x}_i)$$

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.

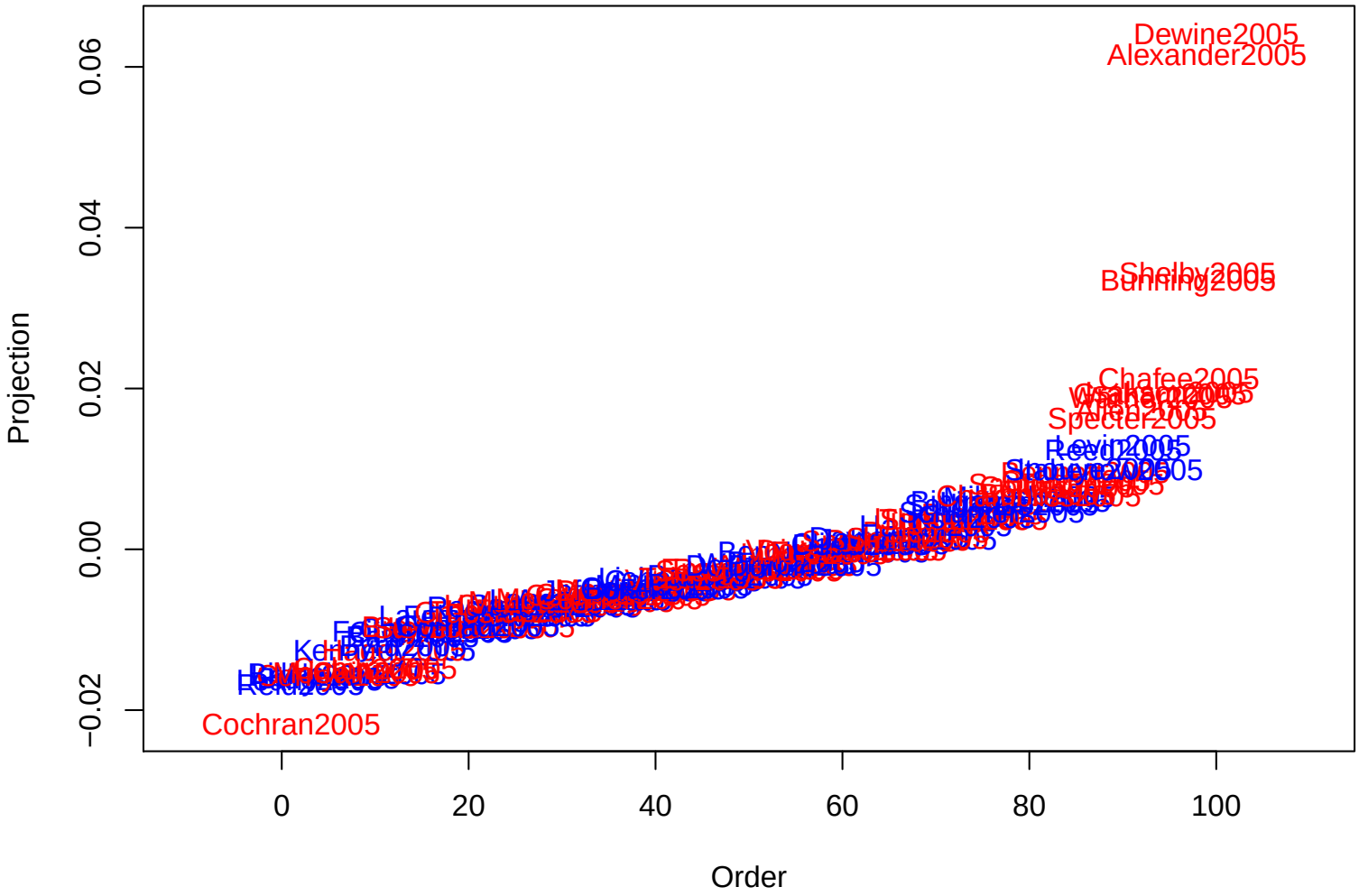
$$x_{ij} = \frac{\text{No. Times } i \text{ uses word } j}{\text{No. words } i \text{ uses}}$$

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

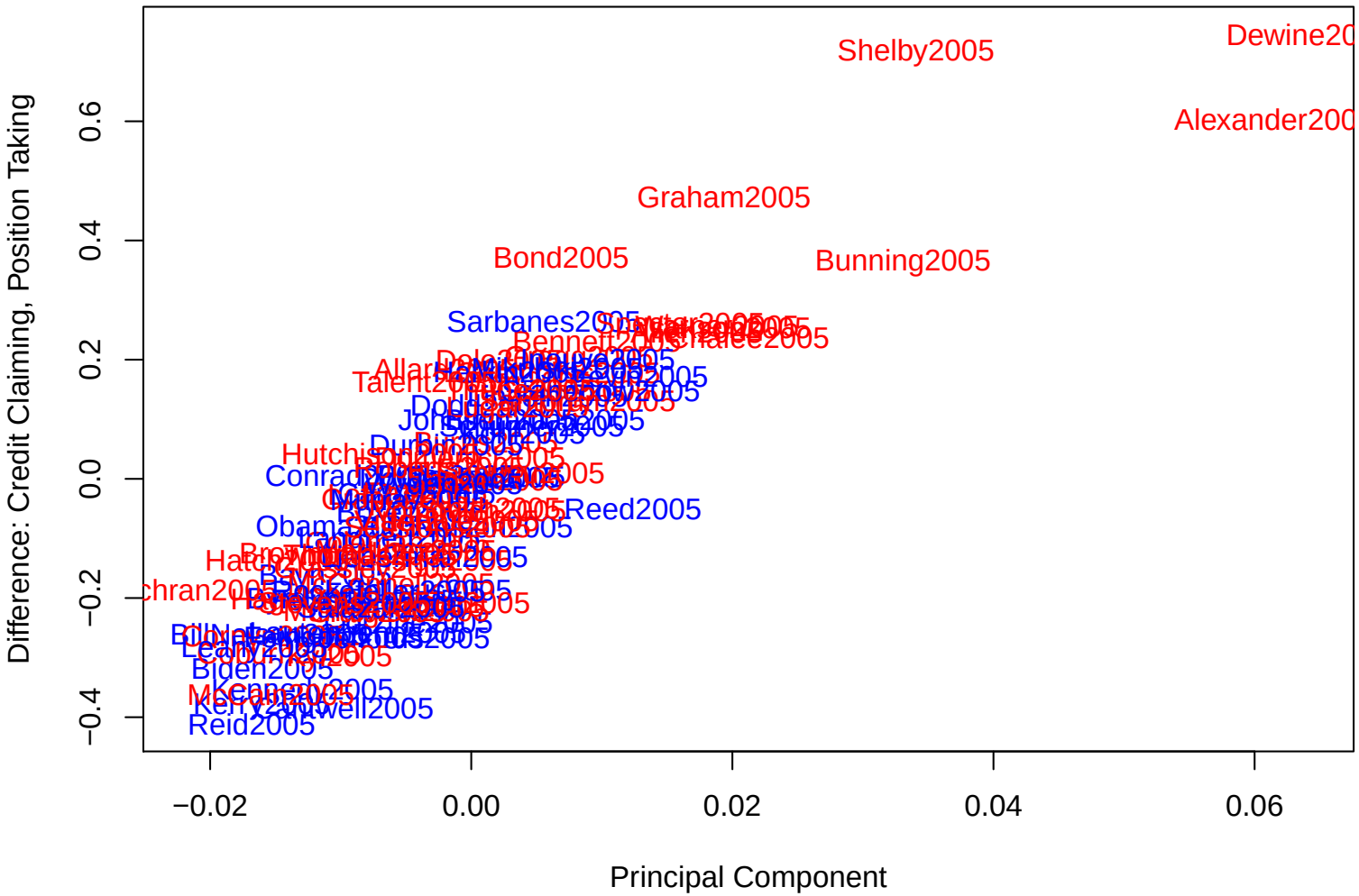
```
load("SenateTDM.RData")
dtm<- t(tdm)
for(z in 1:100){
dtm[z,]<- dtm[z,]/sum(dtm[z,])
}

store<- prcomp(dtm, scale = F)
scores<- store$x[,1]
```

Application of Principal Components in R



Application of Principal Components in R



Probabilistic Principal Components (Tipping and Bishop 1999)

$$
\begin{aligned}
\boldsymbol{x}|\boldsymbol{w} &\sim \text{Multivariate Normal}(\boldsymbol{Z}\boldsymbol{W} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) \\
\boldsymbol{w} &\sim \text{Multivariate Normal}(\boldsymbol{0}, \boldsymbol{I}) \\
\boldsymbol{x} &\sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \boldsymbol{W}\boldsymbol{W}^{'} + \sigma^2\boldsymbol{I}
\end{aligned}
$$

1) Log-likelihood $\rightsquigarrow$ straightforward
2) Optimization via EM-Algorithm
3) Corresponds to traditional PCA is $\lim_{\sigma^2} \to 0$
4) Closely related to Factor analysis.

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?

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

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

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

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

Simplifying:

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

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

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

Simplifying:

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

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) $z_{ij}z_{ik}\boldsymbol{w}_j'\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$

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) $z_{ij}z_{ik}\boldsymbol{w}_j'\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$
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) $z_{ij}z_{ik}\boldsymbol{w}_j'\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$
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) $z_{ij}z_{ik}\boldsymbol{w}_j'\boldsymbol{w}_k = z_{ij}z_{ik}0 = 0$
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

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

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

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

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

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

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

If $L = J$

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

If $L = J$

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

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

If $L = J$

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

So for $L < J$,

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

If $L = J$

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

So for $L < J$,

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

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

If $L = J$

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

So for $L < J$,

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

$$\sum_{j=L+1}^{J}\lambda_l = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) - \sum_{l=1}^{L}\lambda_l$$

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

If $L = J$

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

So for $L < J$,

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) - (\sum_{l=1}^{L}\lambda_l + \sum_{j=L+1}^{J}\lambda_l) \\ \sum_{j=L+1}^{J}\lambda_l &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) - \sum_{l=1}^{L}\lambda_l \\ \sum_{j=L+1}^{J}\lambda_l &= \text{error}(L) \end{aligned}$$

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)

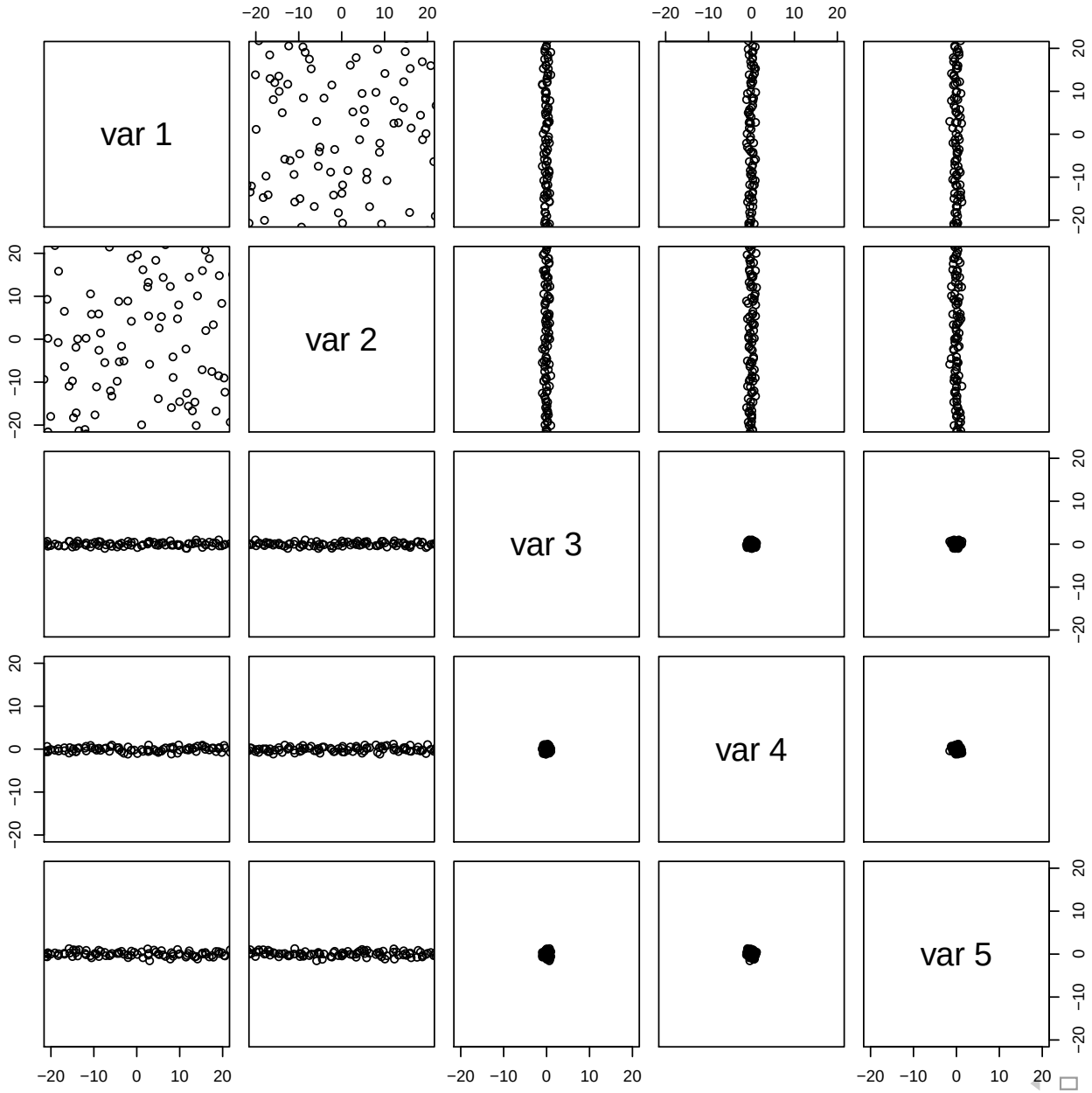
How do we select the number of dimensions L?$\rightsquigarrow$ Model

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

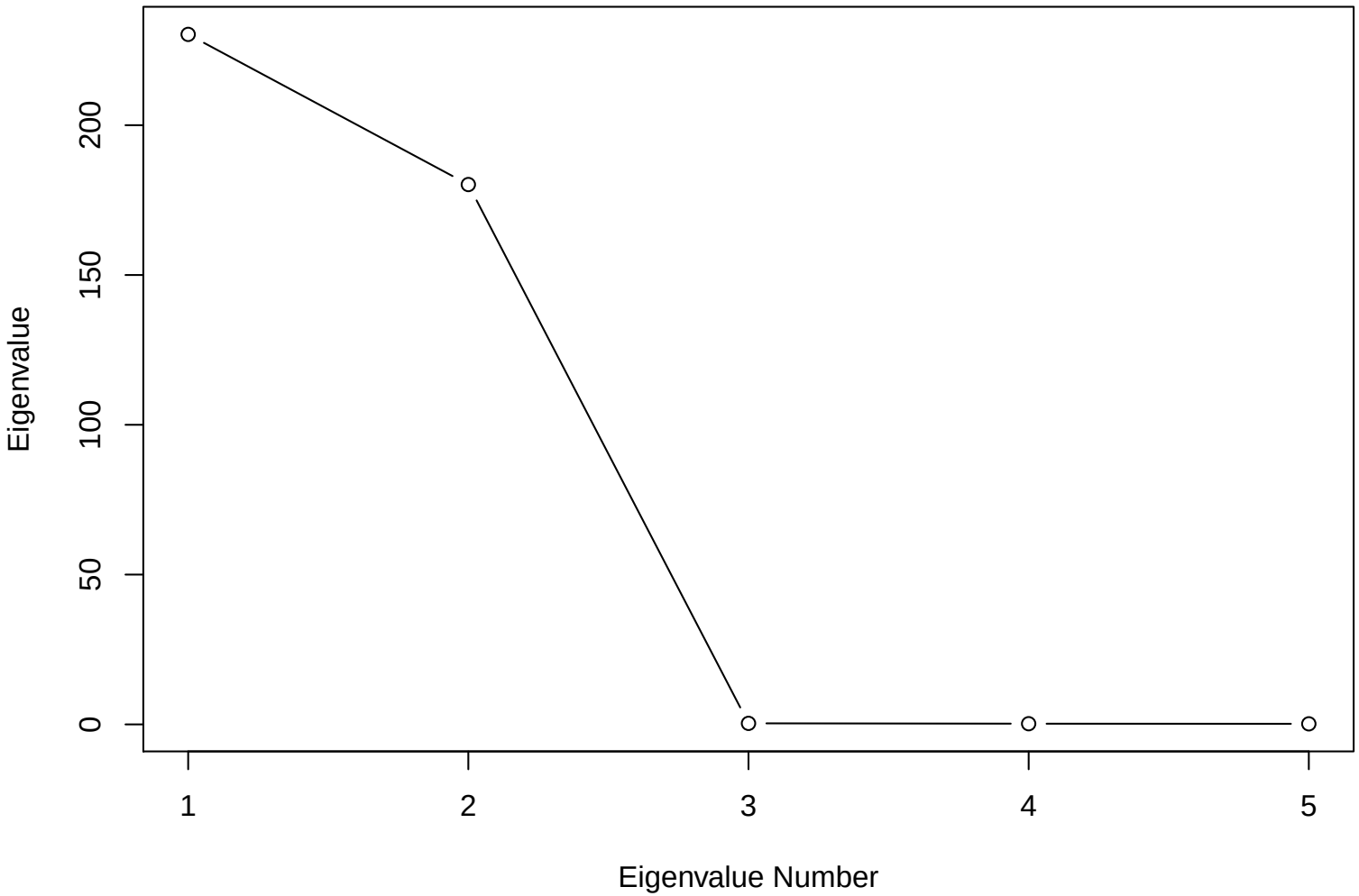
- Error = Sum of "remaining" eigenvalues
- Total variance explained = (sum of included eigenvalues)/(sum of all eigenvalues)

Recommendation$\rightsquigarrow$ look for Elbow

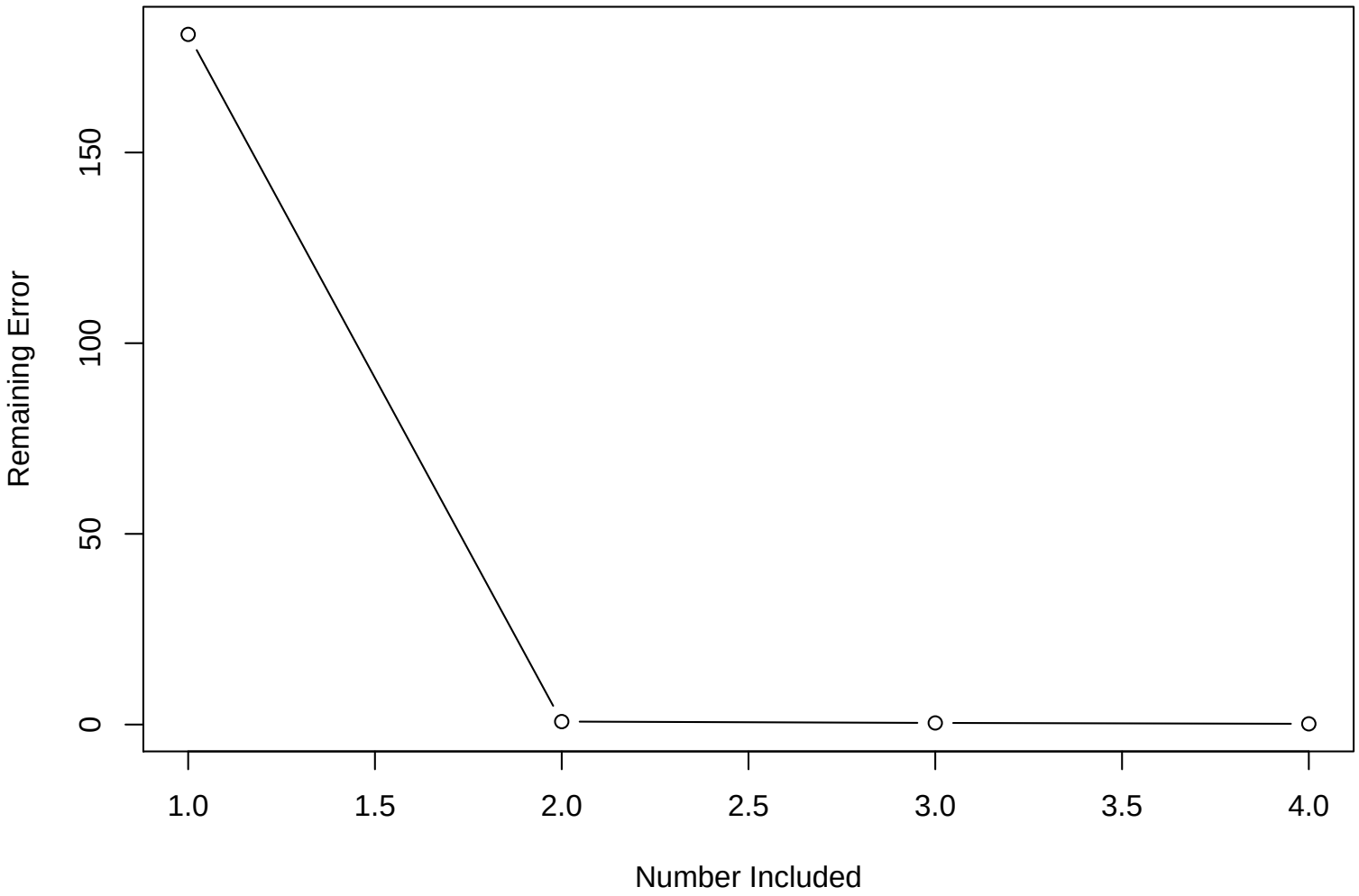
How do we select the number of dimensions L?$\rightsquigarrow$ Model



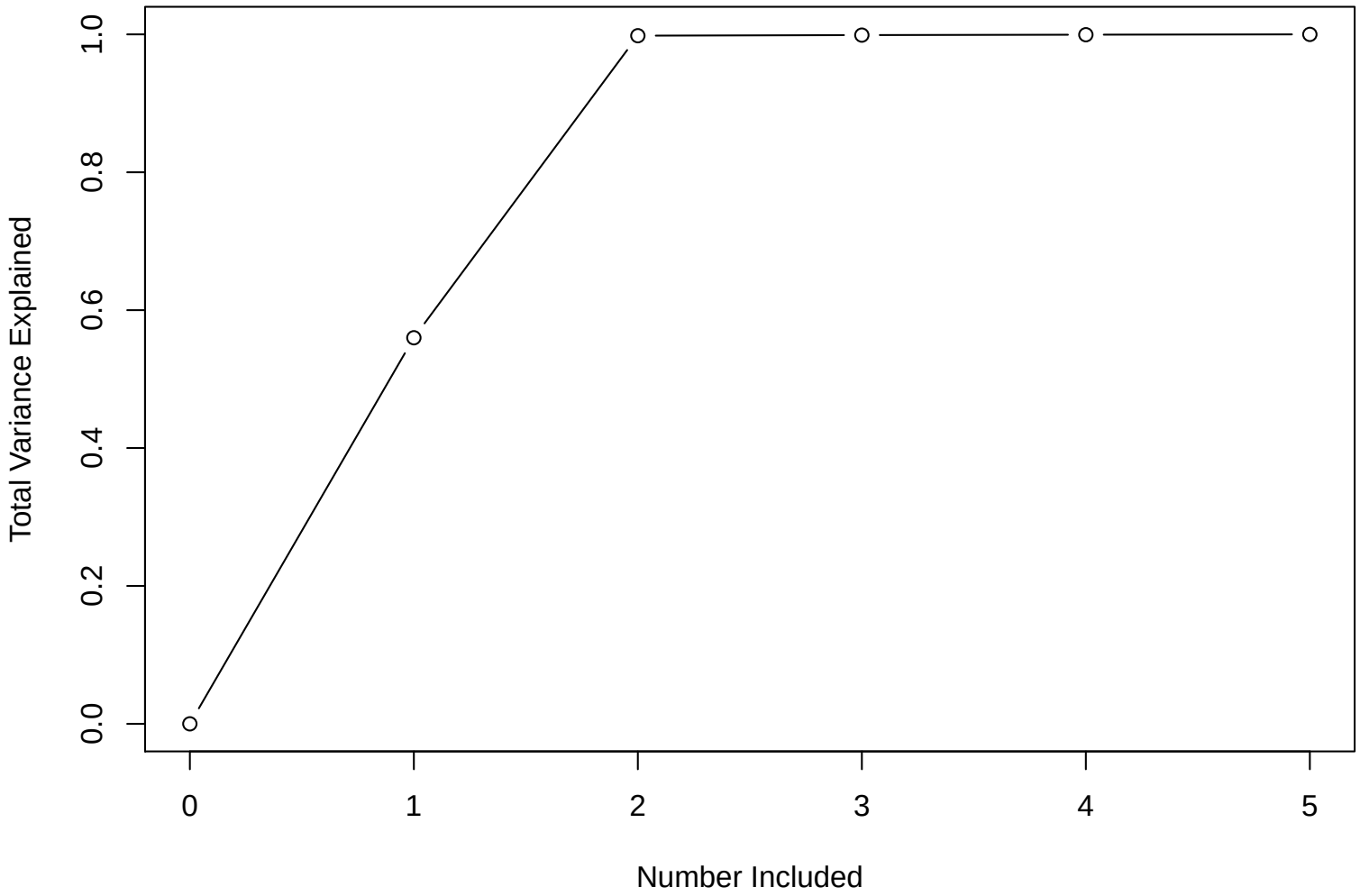
How do we select the number of dimensions L? $\rightsquigarrow$ Model



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

Appendix

Kernel Principal Component Analysis

Define a Kernel $(N \times N)$ matrix as:

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_1, x_N) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & k(\boldsymbol{x}_N, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

where $k(\cdot, \cdot)$ is a function that behaves like a similarity function.

Kernel Principal Component Analysis

Define a Kernel $(N \times N)$ matrix as:

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_1, x_N) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & k(\boldsymbol{x}_N, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

where $k(\cdot, \cdot)$ is a function that behaves like a similarity function. Where we suppose this matrix emerges from applying $\phi : \Re^J \to \Re^M$ to the data and then taking the inner product:

Kernel Principal Component Analysis

Define a Kernel $(N \times N)$ matrix as:

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_1, x_N) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & k(\boldsymbol{x}_N, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

where $k(\cdot, \cdot)$ is a function that behaves like a similarity function. Where we suppose this matrix emerges from applying $\phi : \Re^J \to \Re^M$ to the data and then taking the inner product:

$$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}' \text{ (The inner product matrix)}$$

Kernel Principal Component Analysis

Define a Kernel $(N \times N)$ matrix as:

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_1, x_N) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & k(\boldsymbol{x}_N, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

where $k(\cdot, \cdot)$ is a function that behaves like a similarity function. Where we suppose this matrix emerges from applying $\phi : \Re^J \to \Re^M$ to the data and then taking the inner product:

$$\begin{aligned} \boldsymbol{K} &= \boldsymbol{\Phi}\boldsymbol{\Phi}' \text{ (The inner product matrix)} \\ &= \begin{pmatrix} \phi(\boldsymbol{x}_1)'\phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_1)'\phi(\boldsymbol{x}_2) & \dots & \phi(\boldsymbol{x}_1)'\phi(\boldsymbol{x}_N) \\ \phi(\boldsymbol{x}_2)'\phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_2)'\phi(\boldsymbol{x}_2) & \dots & \phi(\boldsymbol{x}_2)'\phi(\boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\boldsymbol{x}_N)'\phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_N)'\phi(\boldsymbol{x}_2) & \dots & \phi(\boldsymbol{x}_N)'\phi(\boldsymbol{x}_N) \end{pmatrix} \end{aligned}$$

Kernel Principal Component Analysis

Define a Kernel $(N \times N)$ matrix as:

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_1, x_N) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & k(\boldsymbol{x}_N, \boldsymbol{x}_2) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

where $k(\cdot, \cdot)$ is a function that behaves like a similarity function. Where we suppose this matrix emerges from applying $\phi : \Re^J \rightarrow \Re^M$ to the data and then taking the inner product:

$$\begin{aligned} \boldsymbol{K} &= \boldsymbol{\Phi}\boldsymbol{\Phi}' \text{ (The inner product matrix)} \\ &= \begin{pmatrix} \phi(\boldsymbol{x}_1)'\phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_1)'\phi(\boldsymbol{x}_2) & \dots & \phi(\boldsymbol{x}_1)'\phi(\boldsymbol{x}_N) \\ \phi(\boldsymbol{x}_2)'\phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_2)'\phi(\boldsymbol{x}_2) & \dots & \phi(\boldsymbol{x}_2)'\phi(\boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\boldsymbol{x}_N)'\phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_N)'\phi(\boldsymbol{x}_2) & \dots & \phi(\boldsymbol{x}_N)'\phi(\boldsymbol{x}_N) \end{pmatrix} \end{aligned}$$

Compute PCA of $\boldsymbol{\Phi}$ from $\boldsymbol{\Phi}\boldsymbol{\Phi}'$

Kernel PCA

PCA of $\boldsymbol{X}$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1.

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1)\end{aligned}$$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

But $\boldsymbol{v}_1$ needs unit length, and

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

But $\boldsymbol{v}_1$ needs unit length, and

$$||\boldsymbol{v}_1||^2 = \boldsymbol{v}_1'\boldsymbol{v}_1$$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

But $\boldsymbol{v}_1$ needs unit length, and

$$\begin{aligned}||\boldsymbol{v}_1||^2 = \boldsymbol{v}_1'\boldsymbol{v}_1 & \\ &= \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{u}_1\end{aligned}$$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

But $\boldsymbol{v}_1$ needs unit length, and

$$\begin{aligned}||\boldsymbol{v}_1||^2 = \boldsymbol{v}_1'\boldsymbol{v}_1 & \\ &= \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{u}_1 \\ &= \lambda_1\boldsymbol{u}_1'\boldsymbol{u}_1 = \lambda_1\end{aligned}$$

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

But $\boldsymbol{v}_1$ needs unit length, and

$$\begin{aligned}||\boldsymbol{v}_1||^2 = \boldsymbol{v}_1'\boldsymbol{v}_1 & \\ &= \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{u}_1 \\ &= \lambda_1\boldsymbol{u}_1'\boldsymbol{u}_1 = \lambda_1\end{aligned}$$

So first eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ is

Kernel PCA

PCA of $\boldsymbol{X}$ Eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ ($\frac{1}{N}$ doesn't affect eigenvectors)

Suppose $\boldsymbol{u}_1$ is an eigenvector for $\boldsymbol{X}\boldsymbol{X}'$, with value λ_1. Then

$$\begin{aligned}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{u}_1 &= \lambda_1\boldsymbol{u}_1 \\ (\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{u}_1) &= \lambda_1(\boldsymbol{X}'\boldsymbol{u}_1) \\ &= \lambda_1\boldsymbol{v}_1\end{aligned}$$

But $\boldsymbol{v}_1$ needs unit length, and

$$\begin{aligned}||\boldsymbol{v}_1||^2 = \boldsymbol{v}_1'\boldsymbol{v}_1 & \\ &= \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{u}_1 \\ &= \lambda_1\boldsymbol{u}_1'\boldsymbol{u}_1 = \lambda_1\end{aligned}$$

So first eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ is

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}}\boldsymbol{X}'\boldsymbol{u}_1$$

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$.

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}^{'}}_{\text{Unknown}} \boldsymbol{u}_1$$

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}'}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}'}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

$$\phi(\boldsymbol{x}_i)' \boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \phi(\boldsymbol{x}_i)' \boldsymbol{\Phi}' \boldsymbol{u}_1$$

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}^{'}}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

$$\begin{aligned} \phi(\boldsymbol{x}_i)^{'}\boldsymbol{w}_1 &= \frac{1}{\sqrt{\lambda_1}}\phi(\boldsymbol{x}_i)^{'}\boldsymbol{\Phi}^{'}\boldsymbol{u}_1 \\ \phi(\boldsymbol{x}_i)^{'}\boldsymbol{\Phi}^{'} &= \left[\phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_N)\right] \end{aligned}$$

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}^{'}}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

$$\begin{aligned}
\phi(\boldsymbol{x}_i)^{'}\boldsymbol{w}_1 &= \frac{1}{\sqrt{\lambda_1}}\phi(\boldsymbol{x}_i)^{'}\boldsymbol{\Phi}^{'}\boldsymbol{u}_1 \\
\phi(\boldsymbol{x}_i)^{'}\boldsymbol{\Phi}^{'} &= \left[\phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_N)\right] \\
&= [k(\boldsymbol{x}_i, \boldsymbol{x}_1), k(\boldsymbol{x}_i, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}_i, \boldsymbol{x}_N)]
\end{aligned}$$

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}^{'}}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

$$\begin{aligned}
\phi(\boldsymbol{x}_i)^{'}\boldsymbol{w}_1 &= \frac{1}{\sqrt{\lambda_1}}\phi(\boldsymbol{x}_i)^{'}\boldsymbol{\Phi}^{'}\boldsymbol{u}_1 \\
\phi(\boldsymbol{x}_i)^{'}\boldsymbol{\Phi}^{'} &= \left[\phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_i)^{'}\phi(\boldsymbol{x}_N)\right] \\
&= [k(\boldsymbol{x}_i, \boldsymbol{x}_1), k(\boldsymbol{x}_i, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}_i, \boldsymbol{x}_N)] = \boldsymbol{k}(\boldsymbol{x}_i, *)
\end{aligned}$$

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}'}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

$$\begin{aligned}
\phi(\boldsymbol{x}_i)'\boldsymbol{w}_1 &= \frac{1}{\sqrt{\lambda_1}}\phi(\boldsymbol{x}_i)'\boldsymbol{\Phi}'\boldsymbol{u}_1 \\
\phi(\boldsymbol{x}_i)'\boldsymbol{\Phi}' &= \left[\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_N)\right] \\
&= [k(\boldsymbol{x}_i, \boldsymbol{x}_1), k(\boldsymbol{x}_i, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}_i, \boldsymbol{x}_N)] = \boldsymbol{k}(\boldsymbol{x}_i, *)
\end{aligned}$$

Then, we can obtain projection for observation i using Kernel with

Kernel PCA

$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$ (assume $\boldsymbol{\Phi}$ is mean-centered, for now)

We can obtain $\boldsymbol{u}_1$ and λ_1 from $\boldsymbol{K}$. We know that

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}} \underbrace{\boldsymbol{\Phi}'}_{\text{Unknown}} \boldsymbol{u}_1$$

But suppose we want to project a point $\phi(\boldsymbol{x}_i)$, then

$$\begin{aligned}
\phi(\boldsymbol{x}_i)'\boldsymbol{w}_1 &= \frac{1}{\sqrt{\lambda_1}}\phi(\boldsymbol{x}_i)'\boldsymbol{\Phi}'\boldsymbol{u}_1 \\
\phi(\boldsymbol{x}_i)'\boldsymbol{\Phi}' &= \left[\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_N)\right] \\
&= [k(\boldsymbol{x}_i, \boldsymbol{x}_1), k(\boldsymbol{x}_i, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}_i, \boldsymbol{x}_N)] = \boldsymbol{k}(\boldsymbol{x}_i, *)
\end{aligned}$$

Then, we can obtain projection for observation i using Kernel with

$$\phi(\boldsymbol{x}_i)'\boldsymbol{w}_1 = \frac{1}{\sqrt{\lambda_1}}\boldsymbol{k}(\boldsymbol{x}_i, *)\boldsymbol{u}_1$$

Kernel PCA

Center $\boldsymbol{K}$?
Use centering matrix $\boldsymbol{H}$

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{I}_N - \frac{(\mathbf{1}_N \mathbf{1}_N')}{N} \\ \boldsymbol{K}_{\text{center}} &= \boldsymbol{H}\boldsymbol{K}\boldsymbol{H} \end{aligned}$$

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`

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 Betweeen 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,

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

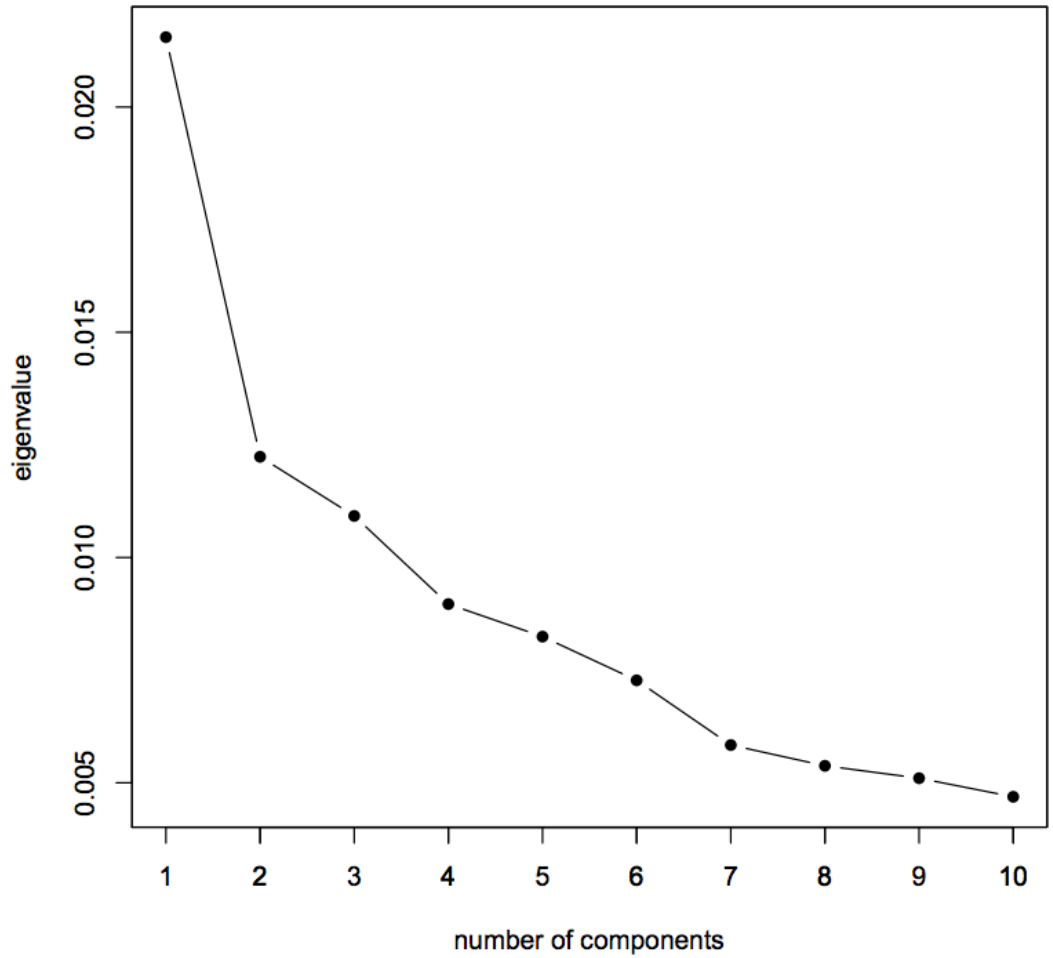
Spirling and Indian Treaties

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,

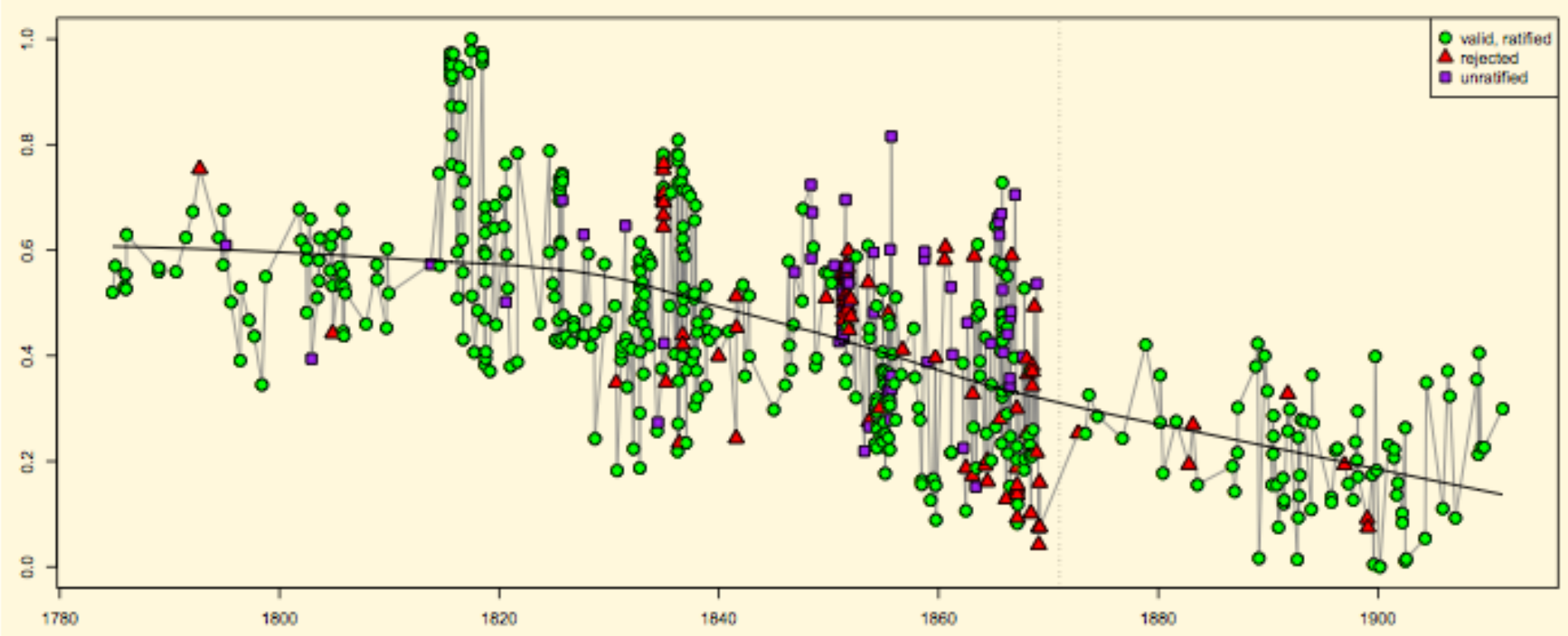
$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{s \in \mathcal{A}} w_s \phi_s(\boldsymbol{x}_i) \phi_s(\boldsymbol{x}_j)$$

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

Spirling and Indian Treaties



Spirling and Indian Treaties



The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$

The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$
- Actor has ideal point $\boldsymbol{\theta}_i \in \Re^M$

The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$
- Actor has ideal point $\boldsymbol{\theta}_i \in \Re^M$
- We describe actor i's utility from proposal $\boldsymbol{p} \in \Re^M$ with utility function

The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$
- Actor has ideal point $\boldsymbol{\theta}_i \in \Re^M$
- We describe actor i's utility from proposal $\boldsymbol{p} \in \Re^M$ with utility function

$$u_i(\boldsymbol{\theta}_i, \boldsymbol{p}) \quad = \quad -d(\boldsymbol{\theta}_i, \boldsymbol{p})$$

The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$
- Actor has ideal point $\boldsymbol{\theta}_i \in \Re^M$
- We describe actor i's utility from proposal $\boldsymbol{p} \in \Re^M$ with utility function

$$\begin{aligned} u_i(\boldsymbol{\theta}_i, \boldsymbol{p}) &= -d(\boldsymbol{\theta}_i, \boldsymbol{p}) \\ &= -\sum_{l=1}^{L}(\underbrace{\theta_{il}}_{\text{ideal policy}} - p_l)^2 \end{aligned}$$

The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$
- Actor has ideal point $\boldsymbol{\theta}_i \in \Re^M$
- We describe actor i's utility from proposal $\boldsymbol{p} \in \Re^M$ with utility function

$$\begin{aligned} u_i(\boldsymbol{\theta}_i, \boldsymbol{p}) &= -d(\boldsymbol{\theta}_i, \boldsymbol{p}) \\ &= -\sum_{l=1}^{L}(\underbrace{\theta_{il}}_{\text{ideal policy}} - p_l)^2 \end{aligned}$$

Estimation goal: $\widehat{\boldsymbol{\theta}}_i$

The Spatial Model

- Suppose we have actor i $(i = 1, 2, 3, \ldots, N)$
- Actor has ideal point $\boldsymbol{\theta}_i \in \Re^M$
- We describe actor i's utility from proposal $\boldsymbol{p} \in \Re^M$ with utility function

$$\begin{aligned} u_i(\boldsymbol{\theta}_i, \boldsymbol{p}) &= -d(\boldsymbol{\theta}_i, \boldsymbol{p}) \\ &= -\sum_{l=1}^{L}(\underbrace{\theta_{il}}_{\text{ideal policy}} - p_l)^2 \end{aligned}$$

Estimation goal: $\widehat{\boldsymbol{\theta}}_i$
Scaling$\rightsquigarrow$ placing actors in low-dimensional space (like principal components!)

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present
  - `NOMINATE` methods$\rightsquigarrow$ place legislators on one dimension, estimate of ideology

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present
  - `NOMINATE` methods$\rightsquigarrow$ place legislators on one dimension, estimate of ideology
- Wildly successful:

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present
  - `NOMINATE` methods$\rightsquigarrow$ place legislators on one dimension, estimate of ideology
- Wildly successful:
  - Estimates are accurate: face validity Congressional scholars agree upon

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present
  - `NOMINATE` methods$\rightsquigarrow$ place legislators on one dimension, estimate of ideology
- Wildly successful:
  - Estimates are accurate: face validity Congressional scholars agree upon
  - Insightful$\rightsquigarrow$ unidimensional Congress

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present
  - `NOMINATE` methods$\rightsquigarrow$ place legislators on one dimension, estimate of ideology
- Wildly successful:
  - Estimates are accurate: face validity Congressional scholars agree upon
  - Insightful$\rightsquigarrow$ unidimensional Congress
  - Extensible: insight of IRT allows model to be embedded in many forms

Estimating Ideal Points: Roll Call Data and the US Congress

US Congress and Roll Call

- Poole and Rosenthal `voteview`
  - Roll Call Data$\rightsquigarrow$ 1789-Present
  - `NOMINATE` methods$\rightsquigarrow$ place legislators on one dimension, estimate of ideology
- Wildly successful:
  - Estimates are accurate: face validity Congressional scholars agree upon
  - Insightful$\rightsquigarrow$ unidimensional Congress
  - Extensible: insight of IRT allows model to be embedded in many forms
  - Widely used: hard to write a paper on American political institutions with ideal points

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes
   - Parliamentary systems$\rightsquigarrow$ no discretion, no variation.

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes
   - Parliamentary systems$\rightsquigarrow$ no discretion, no variation.
   - Spirling and Quinn (2011)$\rightsquigarrow$ mixture model like models for blocs in UK Parliament

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes
   - Parliamentary systems$\rightsquigarrow$ no discretion, no variation.
   - Spirling and Quinn (2011)$\rightsquigarrow$ mixture model like models for blocs in UK Parliament
2) Not everyone votes!

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes
   - Parliamentary systems$\rightsquigarrow$ no discretion, no variation.
   - Spirling and Quinn (2011)$\rightsquigarrow$ mixture model like models for blocs in UK Parliament
2) Not everyone votes!
   - Voters$\rightsquigarrow$ survey responses (but problems with that)

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes
   - Parliamentary systems$\rightsquigarrow$ no discretion, no variation.
   - Spirling and Quinn (2011)$\rightsquigarrow$ mixture model like models for blocs in UK Parliament
2) Not everyone votes!
   - Voters$\rightsquigarrow$ survey responses (but problems with that)
   - Challengers$\rightsquigarrow$ NPAT surveys (but they don't fill those out anymore)

Estimating Ideal Points in General

Two Limitations with the NOMINATE project:

1) US Congress is distinct$\rightsquigarrow$ roll call votes fail to measure ideology in other settings
   - Weak party pressure$\rightsquigarrow$ individual discretion on votes
   - Parliamentary systems$\rightsquigarrow$ no discretion, no variation.
   - Spirling and Quinn (2011)$\rightsquigarrow$ mixture model like models for blocs in UK Parliament
2) Not everyone votes!
   - Voters$\rightsquigarrow$ survey responses (but problems with that)
   - Challengers$\rightsquigarrow$ NPAT surveys (but they don't fill those out anymore)
   - Bonica (2013, 2014)$\rightsquigarrow$ estimate ideology from donations (but not everyone donates)

Estimating Ideal Points in General

But Everyone talks!

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress
    - Hard not to find ideology

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress
  - Hard not to find ideology
  - Behavior that is primarily ideological

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress
  - Hard not to find ideology
  - Behavior that is primarily ideological
- Reality: scaling is much more difficult than roll call voting examples

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress
  - Hard not to find ideology
  - Behavior that is primarily ideological
- Reality: scaling is much more difficult than roll call voting examples
  - Hard to find ideology

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress
  - Hard not to find ideology
  - Behavior that is primarily ideological
- Reality: scaling is much more difficult than roll call voting examples
  - Hard to find ideology
  - Much of political speech reveals little about position on ideological spectrum$\rightsquigarrow$ advertising, regional

Estimating Ideal Points in General

But Everyone talks!

- If we could scale based on conversation, we can measure ideology anywhere
- Much of the literature$\rightsquigarrow$ relies upon intuition from US Congress
  - Hard not to find ideology
  - Behavior that is primarily ideological
- Reality: scaling is much more difficult than roll call voting examples
  - Hard to find ideology
  - Much of political speech reveals little about position on ideological spectrum$\rightsquigarrow$ advertising, regional

Healthy skepticism!

Our plan

Our plan

1) Wordscores$\rightsquigarrow$ "supervised" scaling

Our plan

1) Wordscores$\rightsquigarrow$ "supervised" scaling
2) Wordfish$\rightsquigarrow$ single dimension

Wordscores$\rightsquigarrow$ Big in Europe

Wordscores, Like the Hoff$\rightsquigarrow$ Big in Europe

Wordscores: Objective Function

For each legislator i, suppose we observe D_i documents.

Wordscores: Objective Function

For each legislator i, suppose we observe D_i documents.
Define:

Wordscores: Objective Function

For each legislator i, suppose we observe D_i documents.
Define:

$$\boldsymbol{x}_i = \sum_{l=1}^{D_i} \boldsymbol{x}_{il}$$

Wordscores: Objective Function

For each legislator i, suppose we observe D_i documents.
Define:

$$\begin{aligned} \boldsymbol{x}_i &= \sum_{l=1}^{D_i} \boldsymbol{x}_{il} \\ &= \sum_{l=1}^{D_i} (x_{il1}, x_{il2}, \ldots, x_{ilJ}) \end{aligned}$$

Wordscores: Objective Function

For each legislator i, suppose we observe D_i documents.
Define:

$$\begin{aligned} \boldsymbol{x}_i &= \sum_{l=1}^{D_i} \boldsymbol{x}_{il} \\ &= \sum_{l=1}^{D_i} (x_{il1}, x_{il2}, \ldots, x_{ilJ}) \end{aligned}$$

$\boldsymbol{x}_i \rightsquigarrow$ aggregation across documents.

Wordscores: Objective Functions

Choose two legislators as exemplars

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

For each word j we can define:

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

For each word j we can define:

$$P_{jL} = \text{Probability of word from Liberal}$$

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

For each word j we can define:

$$\begin{aligned} P_{jL} &= \text{Probability of word from Liberal} \\ P_{jC} &= \text{Probability of word from Conservative} \end{aligned}$$

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

For each word j we can define:

$$\begin{aligned} P_{jL} &= \text{Probability of word from Liberal} \\ P_{jC} &= \text{Probability of word from Conservative} \end{aligned}$$

Define the score for word j

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

For each word j we can define:

$$\begin{aligned} P_{jL} &= \text{Probability of word from Liberal} \\ P_{jC} &= \text{Probability of word from Conservative} \end{aligned}$$

Define the score for word j

$$S_j = Y_C P_{jC} + Y_L P_{jL}$$

Wordscores: Objective Functions

Choose two legislators as exemplars

- Legislator $L \in \{1, 2, \ldots, N\}$ is liberal. $Y_L = -1$
- For example, might select Elizabeth Warren
- Legislator $C \in \{1, 2, \ldots, N\}$ is Conservative. $Y_C = 1$
- For example, might select Ted Cruz

For each word j we can define:

$$\begin{aligned} P_{jL} &= \text{Probability of word from Liberal} \\ P_{jC} &= \text{Probability of word from Conservative} \end{aligned}$$

Define the score for word j

$$\begin{aligned} S_j &= Y_C P_{jC} + Y_L P_{jL} \\ &= P_{jC} - P_{jL} \end{aligned}$$

Wordscores: Objective Functions

Scale other legislators:

Wordscores: Objective Functions

Scale other legislators:

$$N_i = \sum_{j=1}^{J} \boldsymbol{x}_i$$

Wordscores: Objective Functions

Scale other legislators:

$$N_i = \sum_{j=1}^{J} \boldsymbol{x}_i$$

$\hat{\theta}_i$ is

Wordscores: Objective Functions

Scale other legislators:

$$N_i = \sum_{j=1}^{J} \boldsymbol{x}_i$$

$\hat{\theta}_i$ is

$$\hat{\theta}_i = \sum_{j=1}^{J} \left(\frac{x_{ij}}{N_i} \right) S_j$$

Wordscores: Objective Functions

Scale other legislators:

$$N_i = \sum_{j=1}^{J} \boldsymbol{x}_i$$

$\hat{\theta}_i$ is

$$\begin{aligned} \hat{\theta}_i &= \sum_{j=1}^{J} \left(\frac{x_{ij}}{N_i} \right) S_j \\ &= \frac{\boldsymbol{x}_i^{'}}{N_i} \boldsymbol{S} \end{aligned}$$

Wordscores: Optimization

$$
\begin{aligned}
N_L &= \sum_{m=1}^{J} x_{mL} \\
N_C &= \sum_{m=1}^{J} x_{mC}
\end{aligned}
$$

Wordscores: Optimization

$$N_L = \sum_{m=1}^{J} x_{mL}$$

$$N_C = \sum_{m=1}^{J} x_{mC}$$

Estimate P_{jL}, P_{jC}, and S_j

Wordscores: Optimization

$$N_L = \sum_{m=1}^{J} x_{mL}$$

$$N_C = \sum_{m=1}^{J} x_{mC}$$

Estimate P_{jL}, P_{jC}, and S_j

$$P_{jL} = \frac{\frac{x_{jL}}{N_L}}{\frac{x_{jL}}{N_L} + \frac{x_{jC}}{N_C}}$$

Wordscores: Optimization

$$N_L = \sum_{m=1}^{J} x_{mL}$$

$$N_C = \sum_{m=1}^{J} x_{mC}$$

Estimate P_{jL}, P_{jC}, and S_j

$$P_{jL} = \frac{\frac{x_{jL}}{N_L}}{\frac{x_{jL}}{N_L} + \frac{x_{jC}}{N_C}}$$

$$P_{jC} = 1 - P_{jL} = \frac{\frac{x_{jC}}{N_C}}{\frac{x_{jL}}{N_L} + \frac{x_{jC}}{N_C}}$$

Wordscores: Optimization

$$
\begin{aligned}
N_L &= \sum_{m=1}^{J} x_{mL} \\
N_C &= \sum_{m=1}^{J} x_{mC}
\end{aligned}
$$

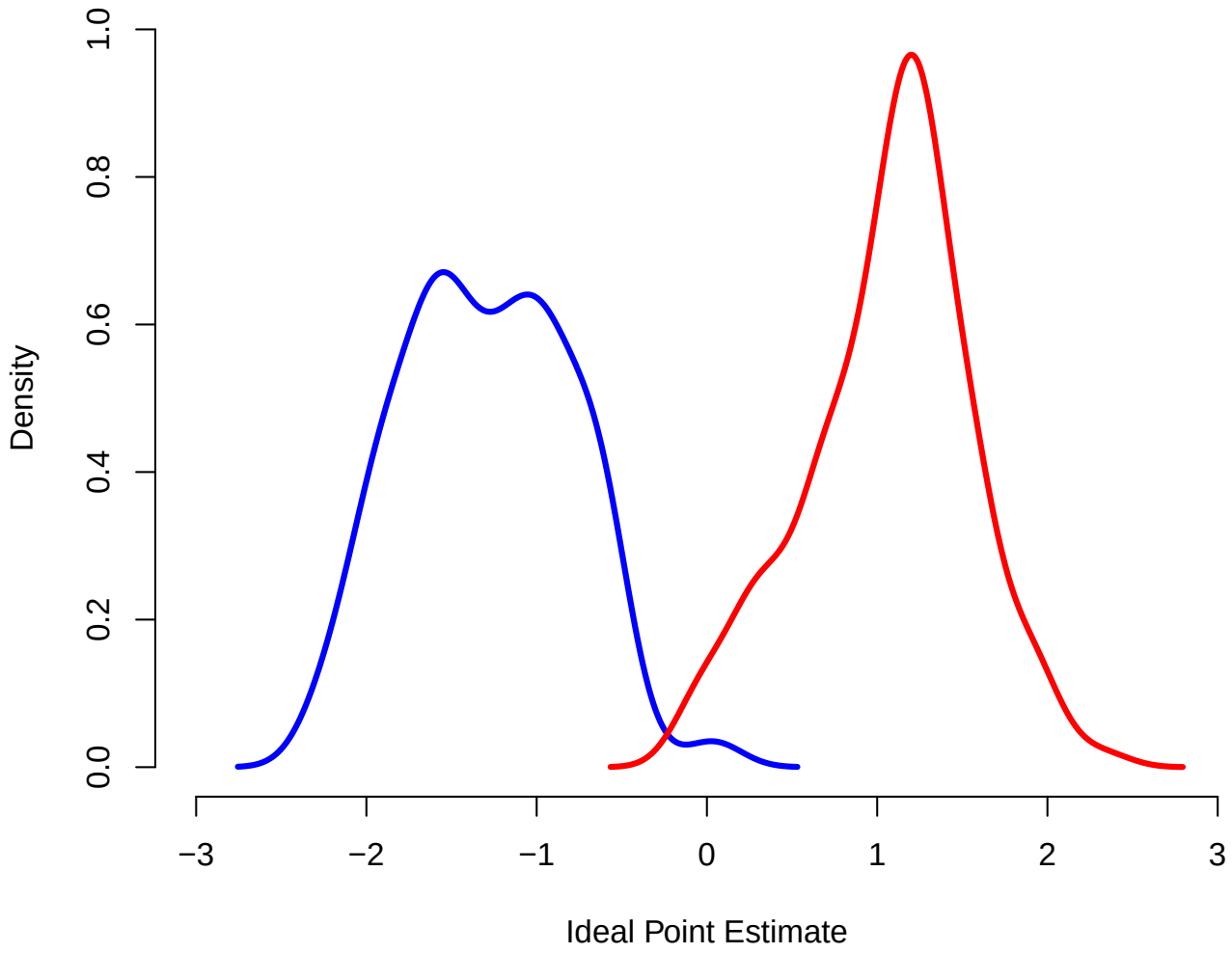
Estimate P_{jL}, P_{jC}, and S_j

$$
\begin{aligned}
P_{jL} &= \frac{\frac{x_{jL}}{N_L}}{\frac{x_{jL}}{N_L} + \frac{x_{jC}}{N_C}} \\
P_{jC} &= 1 - P_{jL} = \frac{\frac{x_{jC}}{N_C}}{\frac{x_{jL}}{N_L} + \frac{x_{jC}}{N_C}} \\
S_j &= P_{jC} - P_{jL}
\end{aligned}
$$

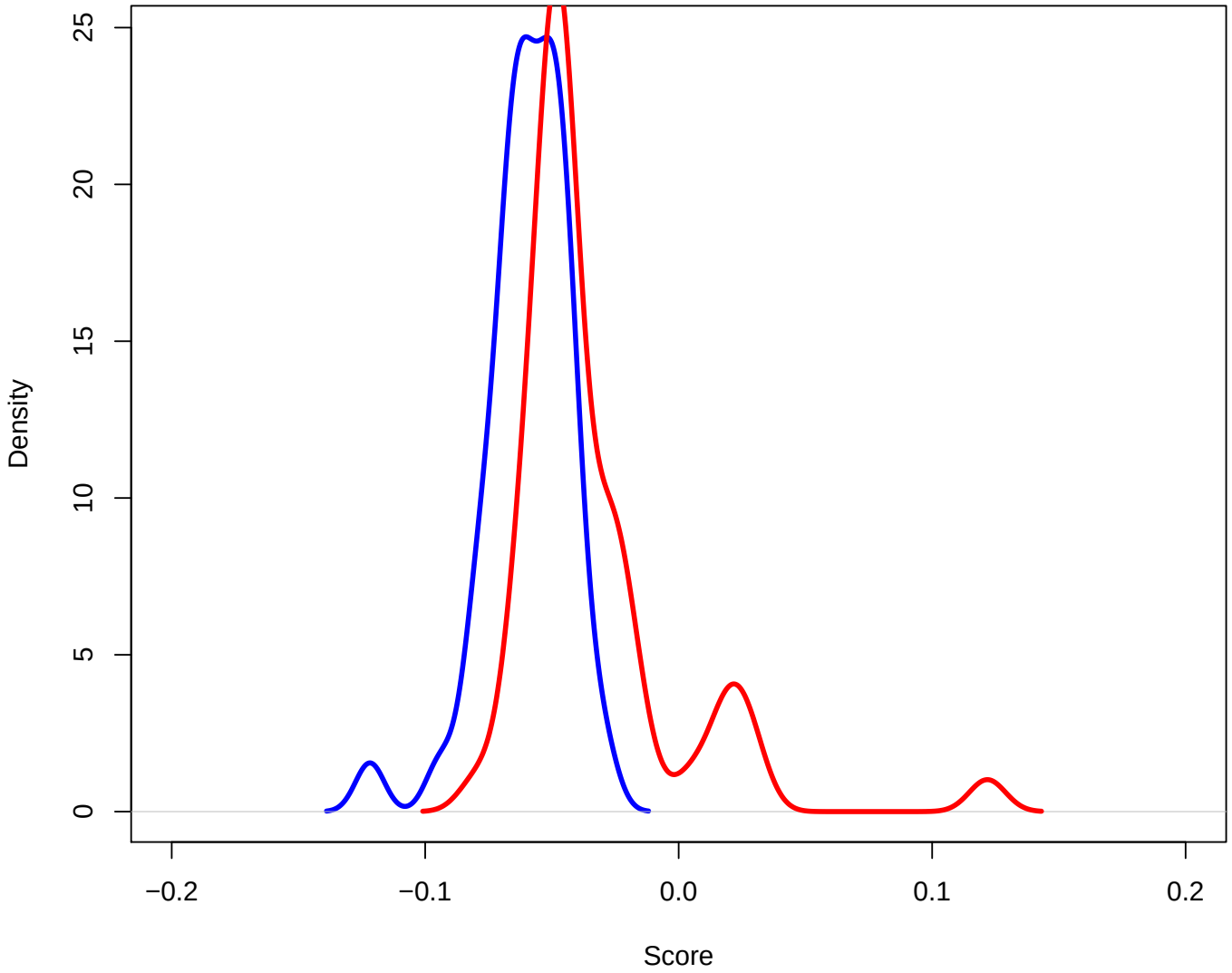
Applied to the Senate Press Releases

$L =$ Ted Kennedy
$C =$ Tom Coburn
Apply to other senators.

Applying to Senate Press Releases⇝ Gold Standard Scaling from NOMINATE



Applying to Senate Press Releases⇝ WordScores



Supervised Scaling

Wordscores is one example of supervised embedding:

Supervised Scaling

Wordscores is one example of supervised embedding:

1) Linear Discriminant Analysis (LDA)$\rightsquigarrow$ the Federalist Papers

Supervised Scaling

Wordscores is one example of supervised embedding:

1) Linear Discriminant Analysis (LDA)$\rightsquigarrow$ the Federalist Papers
2) Multinomial Inverse Regression

Supervised Scaling

Wordscores is one example of supervised embedding:

1) Linear Discriminant Analysis (LDA)$\rightsquigarrow$ the Federalist Papers
2) Multinomial Inverse Regression
3) Any method that separates words

Supervised Scaling

Wordscores is one example of supervised embedding:

1) Linear Discriminant Analysis (LDA)$\rightsquigarrow$ the Federalist Papers
2) Multinomial Inverse Regression
3) Any method that separates words

The findings will depend heavily on supervision

Supervised Scaling

Wordscores is one example of supervised embedding:

1) Linear Discriminant Analysis (LDA)$\rightsquigarrow$ the Federalist Papers
2) Multinomial Inverse Regression
3) Any method that separates words

The findings will depend heavily on supervision

- Sensitive to who is chosen

Supervised Scaling

Wordscores is one example of supervised embedding:

1) Linear Discriminant Analysis (LDA)$\rightsquigarrow$ the Federalist Papers
2) Multinomial Inverse Regression
3) Any method that separates words

The findings will depend heavily on supervision

- Sensitive to who is chosen
- False prediction problem$\rightsquigarrow$ speech accomplishes many goals, only some of which are ideological

What Can Go Wrong with Wordscores?

Lowe (2008): Discusses potentially problematic wordscores properties

What Can Go Wrong with Wordscores?

Lowe (2008): Discusses potentially problematic wordscores properties

1) Each word is weighted equally (fixable with different scoring procedure)

What Can Go Wrong with Wordscores?

Lowe (2008): Discusses potentially problematic wordscores properties

1) Each word is weighted equally (fixable with different scoring procedure)
2) Unique words are conflated with centrist (fixable with MCQ fightin' words style algorithm)

What Can Go Wrong with Wordscores?

Lowe (2008): Discusses potentially problematic wordscores properties

1) Each word is weighted equally (fixable with different scoring procedure)
2) Unique words are conflated with centrist (fixable with MCQ fightin' words style algorithm)
3) General problem: hard to interpret$\rightsquigarrow$ lack of a model

What Can Go Wrong with Wordscores?

Lowe (2008): Discusses potentially problematic wordscores properties

1) Each word is weighted equally (fixable with different scoring procedure)
2) Unique words are conflated with centrist (fixable with MCQ fightin' words style algorithm)
3) General problem: hard to interpret$\rightsquigarrow$ lack of a model

To be fair: fast, nonparametric, and novel [trailblazing] method for scoring documents (starts conversation)

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

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

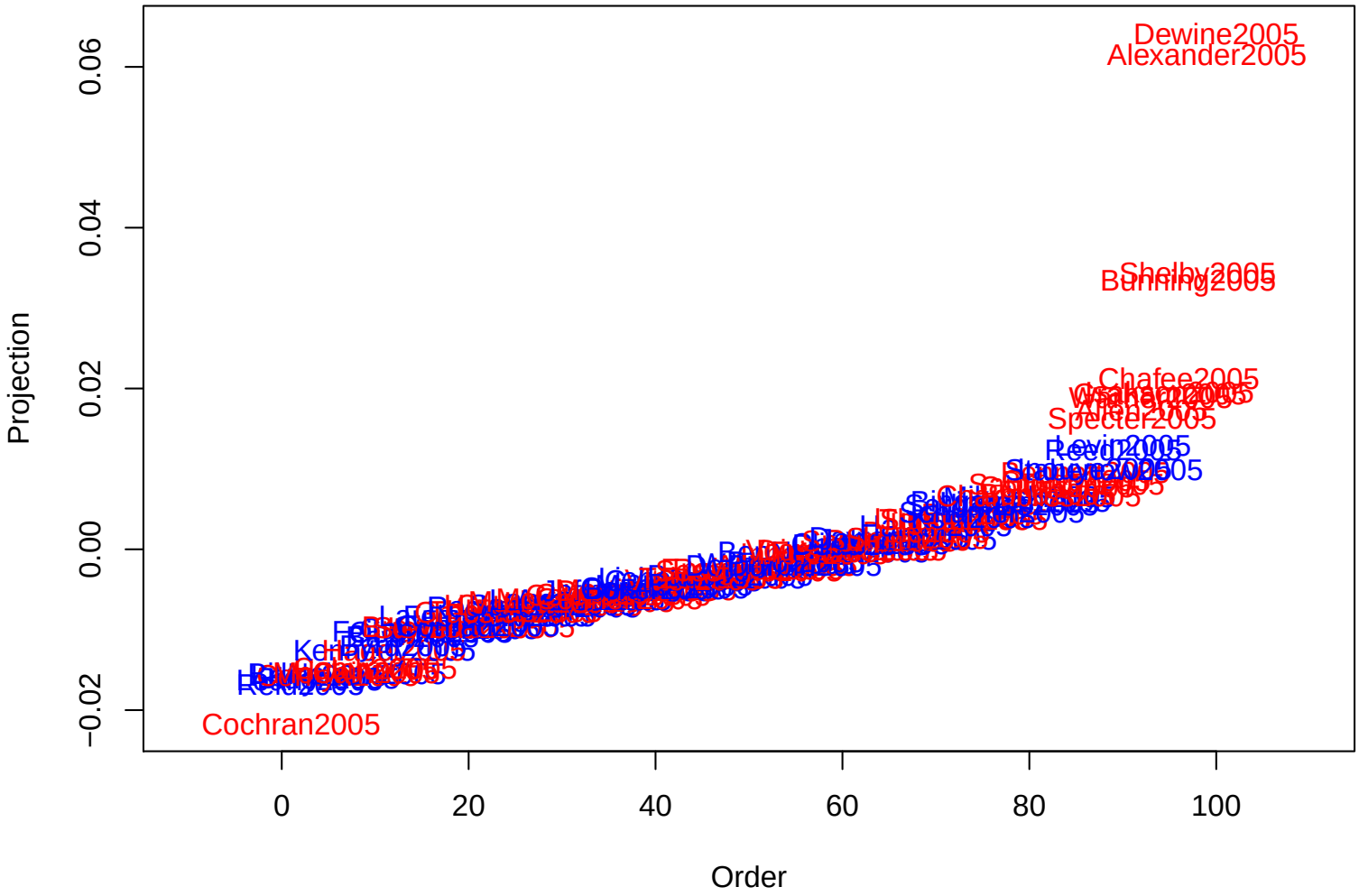
Application of Principal Components in R

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.

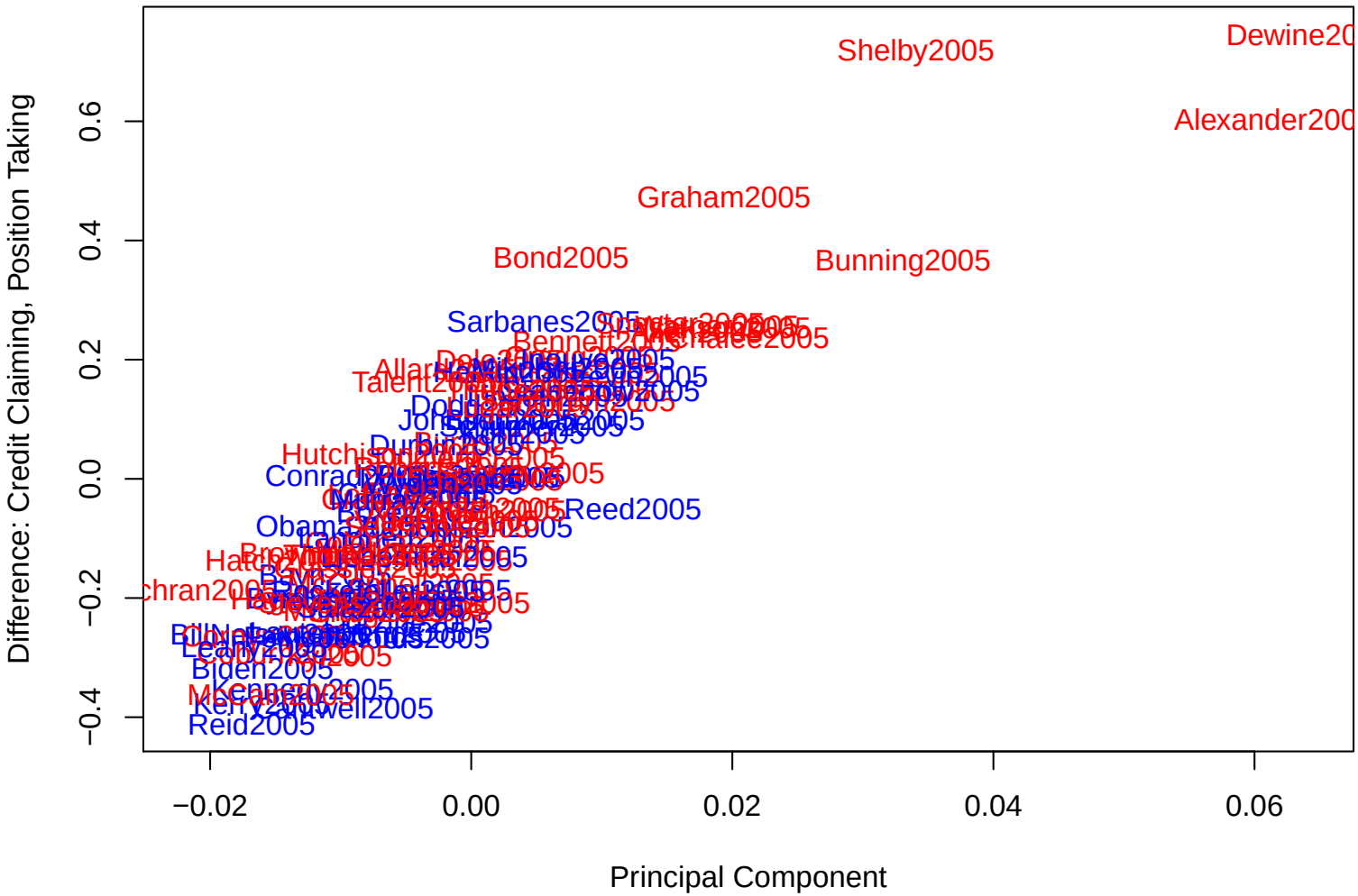
$$x_{ij} = \frac{\text{No. Times } i \text{ uses word } j}{\text{No. words } i \text{ uses}}$$

`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



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
  c) Tuesday$\rightsquigarrow$ IRT and topic models to scale with votes + text

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
  c) Tuesday$\rightsquigarrow$ IRT and topic models to scale with votes + text
- 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
  c) Tuesday$\rightsquigarrow$ IRT and topic models to scale with votes + text
- 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
  c) Tuesday$\rightsquigarrow$ IRT and topic models to scale with votes + text
- 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
  c) Tuesday$\rightsquigarrow$ IRT and topic models to scale with votes + text
- 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

Monroe and Maeda (2005) and Slopkin and Proksch (2008) develop similar algorithms

WordFish Objective Function

Suppose we have legislator i.

WordFish Objective Function

Suppose we have legislator i.

$$x_{ij} \sim \text{Poisson}(\lambda_{ij})$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

$$\lambda_{ij} = \text{Rate individual } i \text{ uses word } j$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

$$\begin{aligned} \lambda_{ij} &= \text{Rate individual } i \text{ uses word } j \\ \alpha_i &= \text{Individual } i \text{ loquaciousness} \end{aligned}$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

$$\begin{aligned} \lambda_{ij} &= \text{Rate individual } i \text{ uses word } j \\ \alpha_i &= \text{Individual } i \text{ loquaciousness} \\ \psi_j &= \text{Word } j\text{'s frequency} \end{aligned}$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

$$\begin{aligned} \lambda_{ij} &= \text{Rate individual } i \text{ uses word } j \\ \alpha_i &= \text{Individual } i \text{ loquaciousness} \\ \psi_j &= \text{Word } j\text{'s frequency} \\ \beta_j &= \text{Word } j\text{'s discrimination} \end{aligned}$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

$$\begin{aligned} \lambda_{ij} &= \text{Rate individual } i \text{ uses word } j \\ \alpha_i &= \text{Individual } i \text{ loquaciousness} \\ \psi_j &= \text{Word } j\text{'s frequency} \\ \beta_j &= \text{Word } j\text{'s discrimination} \\ \theta_i &= \text{Legislator } i\text{'s latent positions} \end{aligned}$$

WordFish Objective Function

Suppose we have legislator i.

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \theta_i) \end{aligned}$$

Where,

$$\begin{aligned} \lambda_{ij} &= \text{Rate individual } i \text{ uses word } j \\ \alpha_i &= \text{Individual } i \text{ loquaciousness} \\ \psi_j &= \text{Word } j\text{'s frequency} \\ \beta_j &= \text{Word } j\text{'s discrimination} \\ \theta_i &= \text{Legislator } i\text{'s latent positions} \end{aligned}$$

"Regression" of x_{ij} on ideal points θ_i, where we have to learn θ_i

WordFish Objective Function/Optimization

Implies the following posterior distribution:

WordFish Objective Function/Optimization

Implies the following posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\psi})p(\boldsymbol{\theta}) \times \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right]\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{x_{ij}}}{x_{ij}!}$$

WordFish Objective Function/Optimization

Implies the following posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\psi})p(\boldsymbol{\theta}) \times \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right]\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{x_{ij}}}{x_{ij}!}$$

Estimate parameters:

WordFish Objective Function/Optimization

Implies the following posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\psi})p(\boldsymbol{\theta}) \times \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right]\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{x_{ij}}}{x_{ij}!}$$

Estimate parameters:

- EM-algorithm

WordFish Objective Function/Optimization

Implies the following posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\psi})p(\boldsymbol{\theta}) \times \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right]\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{x_{ij}}}{x_{ij}!}$$

Estimate parameters:

- EM-algorithm
- MCMC algorithm

WordFish Objective Function/Optimization

Implies the following posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\psi})p(\boldsymbol{\theta}) \times \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right]\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{x_{ij}}}{x_{ij}!}$$

Estimate parameters:

- EM-algorithm
- MCMC algorithm
- Variational Approximation

WordFish Objective Function/Optimization

Implies the following posterior distribution:

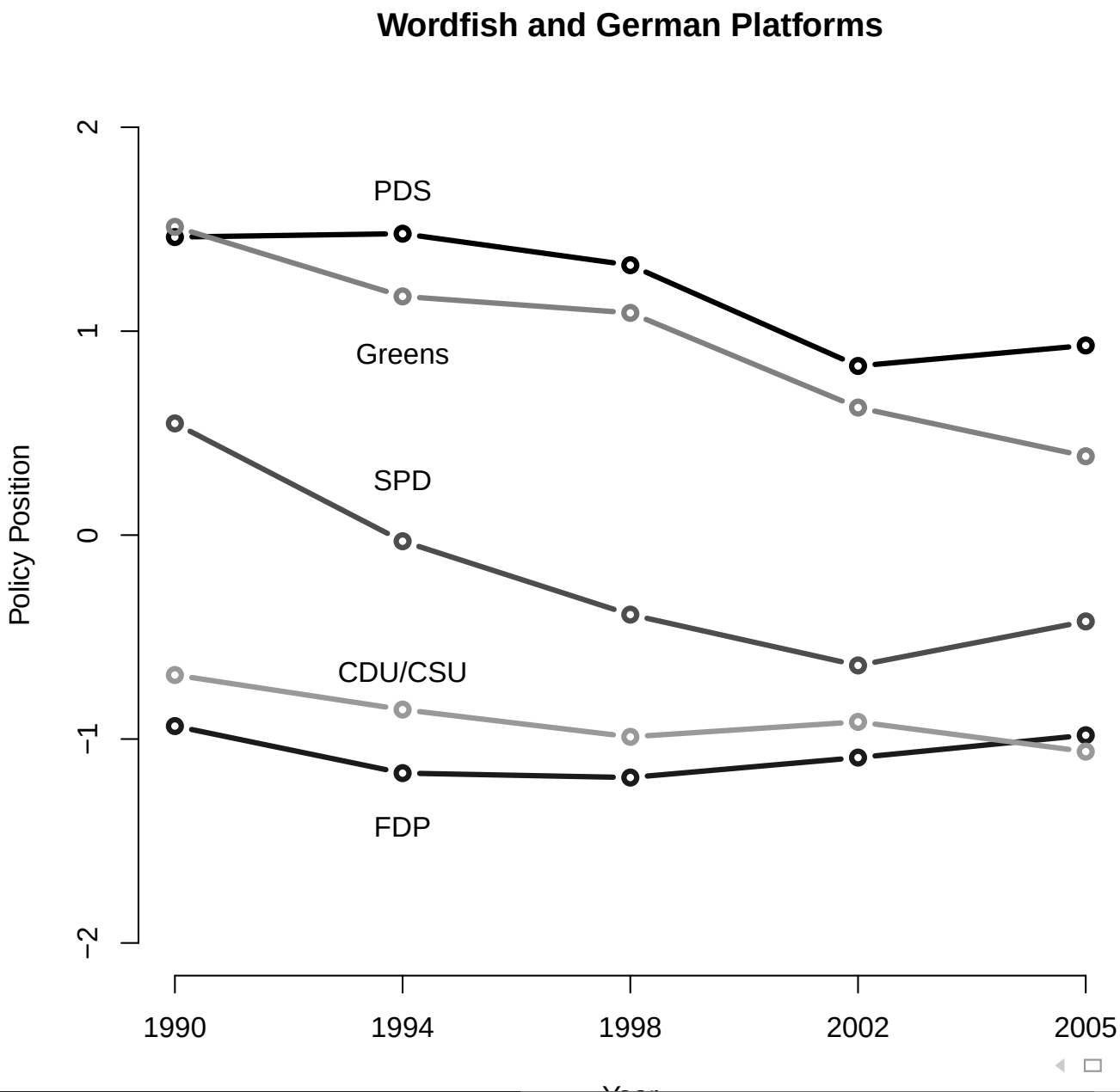
$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\psi})p(\boldsymbol{\theta}) \times \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\left[-\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)\right]\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)^{x_{ij}}}{x_{ij}!}$$

Estimate parameters:

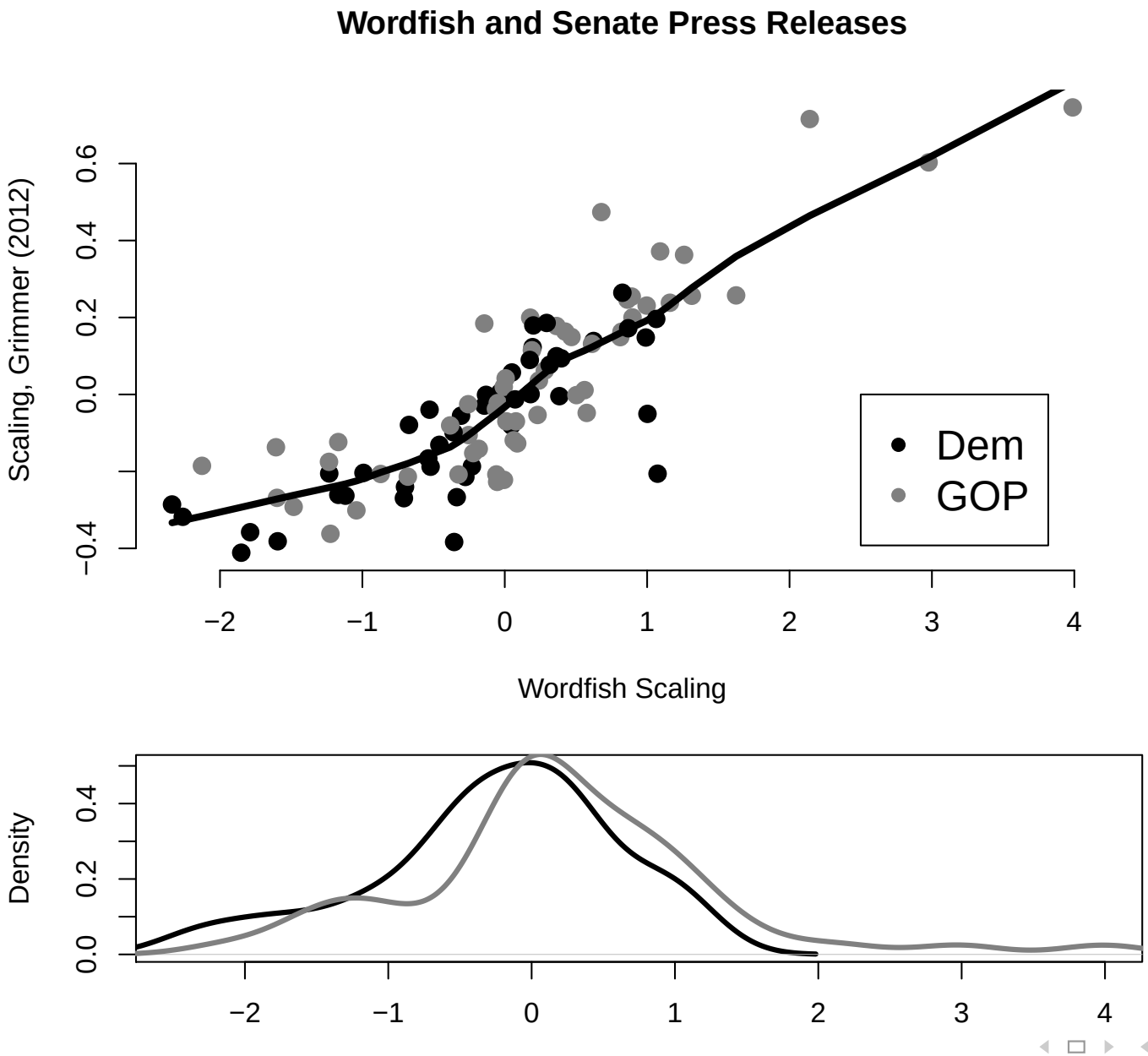
- EM-algorithm
- MCMC algorithm
- Variational Approximation

`Wordfish` package in `R`

Applications: German Party Manifestos



Applications: German Party Manifestos



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