Sparse Models of Lexical Variation

Jacob Eisenstein[1]

Carnegie Mellon University → Georgia Institute of Technology

October 17, 2011

[1]Collaborators: Amr Ahmed, Brendan O'Connor, Noah A. Smith, and Eric P. Xing

The department store study

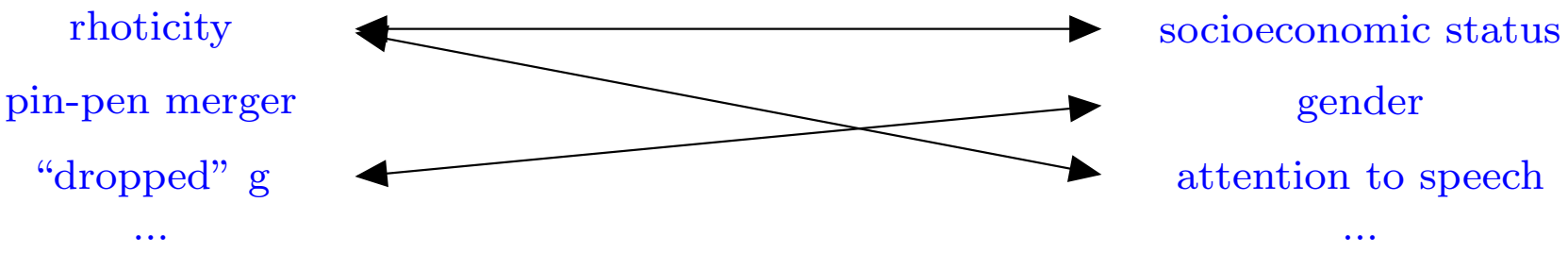
- "The Social Stratification of /r/ in New York City" (Labov, 1966)
  - Rhoticity (r-fullness) correlates with higher prices.
  - Rhoticity increases with attention to speech.
  - Dialect is not purely geographical, it depends on a range of social factors.

The department store study

- "The Social Stratification of /r/ in New York City" (Labov, 1966)
  - Rhoticity (r-fullness) correlates with higher prices.
  - Rhoticity increases with attention to speech.
  - Dialect is not purely geographical, it depends on a range of social factors.
- How did Labov know to study /r/???

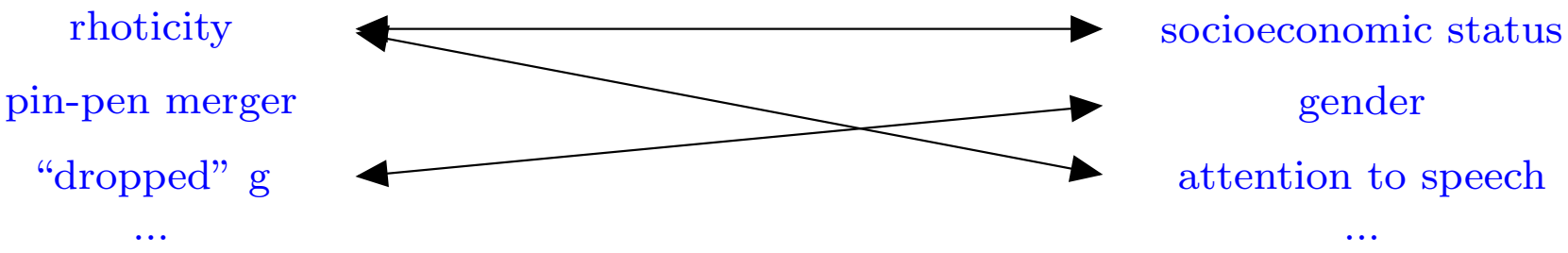
Quantitative Sociolinguistics

Quantitative sociolinguistics focuses on associations between sociolinguistic *variables* and speaker attributes.



Quantitative Sociolinguistics

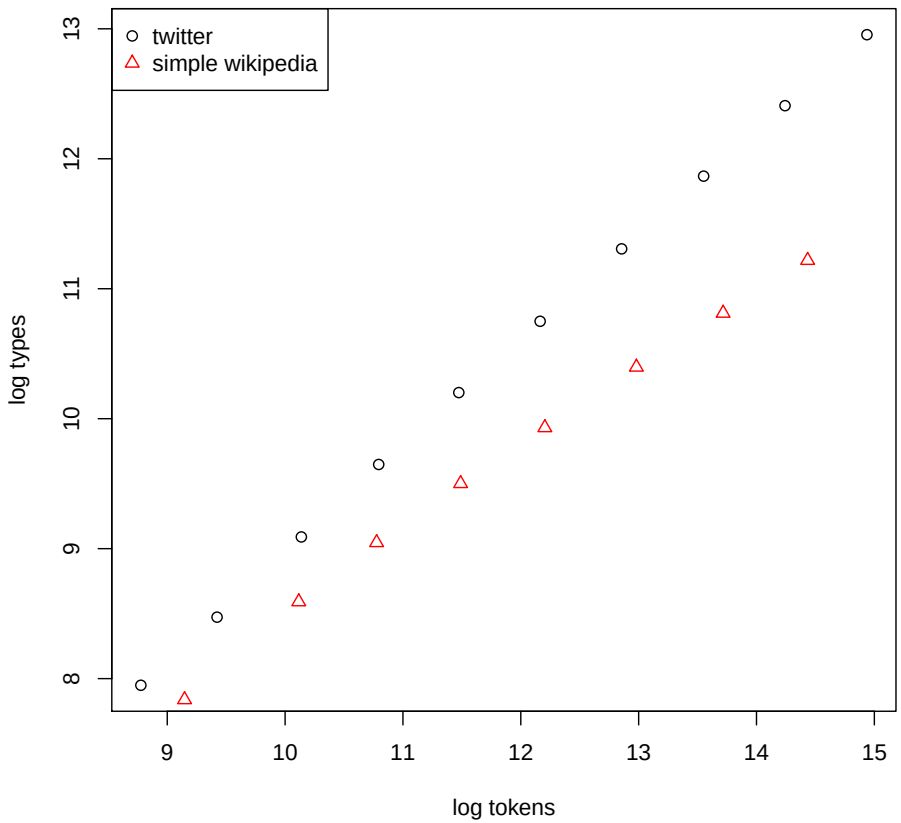
Quantitative sociolinguistics focuses on associations between sociolinguistic *variables* and speaker attributes.



- How do demographics affect language in social media?
  - There are thousands of potentially relevant variables.
  - Author attributes combine in surprising and non-linear ways.
  - This implies millions of potential associations.
    How can we find the right ones?

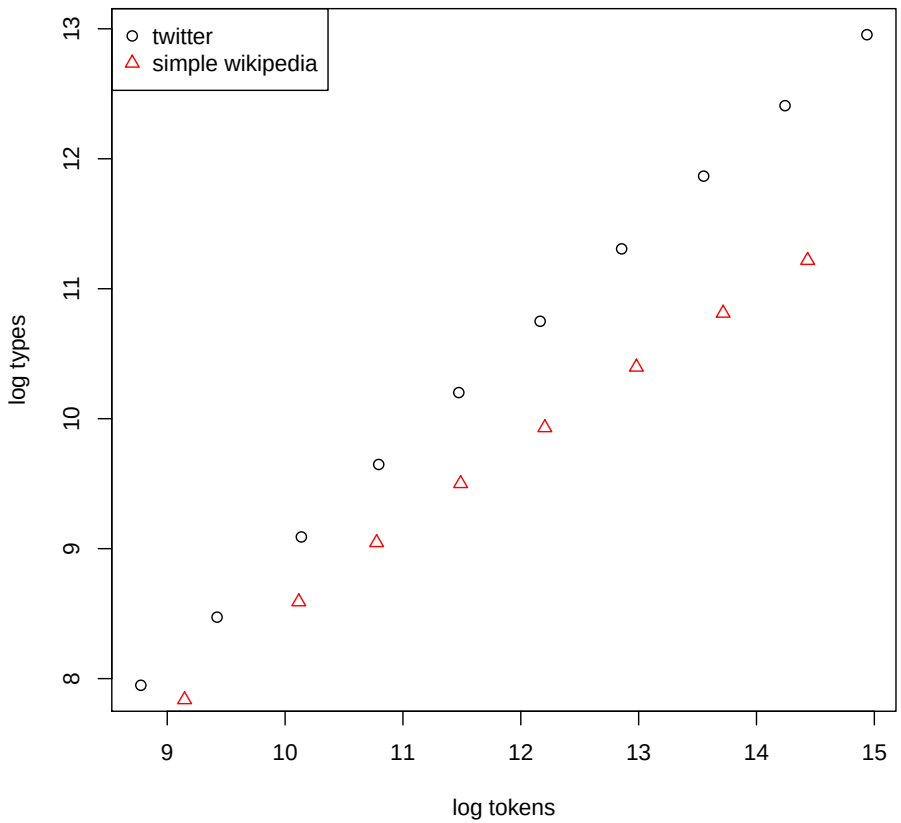
Controlling model complexity

- Language processing problems are inherently high dimensional.
- Zipf's law: no matter how much data we have, there's always a long tail of low-frequency features.



Controlling model complexity

- Language processing problems are inherently high dimensional.
- Zipf's law: no matter how much data we have, there's always a long tail of low-frequency features.



- How can control model complexity?

Sparsity

- Consider a multivariate linear model, $\mathbf{Y} = \mathbf{W}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\epsilon}$
  - $\mathbf{Y} \in \mathcal{R}^{K \times N}$ are the labels
  - $\mathbf{X} \in \mathcal{R}^{D \times N}$ are the features
  - $\mathbf{W} \in \mathcal{R}^{K \times D}$ are the weights
- In a sparse solution, $w_{i,j} = 0$ for many $\langle i, j \rangle$.
- Advantages: interpretability, robustness, low memory, fast inference

Sparsity

- Consider a multivariate linear model, $\mathbf{Y} = \mathbf{W}^{\mathsf{T}}\mathbf{X} + \epsilon$
  - $\mathbf{Y} \in \mathcal{R}^{K \times N}$ are the labels
  - $\mathbf{X} \in \mathcal{R}^{D \times N}$ are the features
  - $\mathbf{W} \in \mathcal{R}^{K \times D}$ are the weights
- In a sparse solution, $w_{i,j} = 0$ for many $\langle i, j \rangle$.
- Advantages: interpretability, robustness, low memory, fast inference

Lots of prior and ongoing work; estimation is now easy

Lange & Sinsheimer 1993, Tibshirani 1996, Tipping 2001, Figueiredo 2003, Bach et al 2011, ...

Outline

- **Application**: structured sparsity in demographic language variation
- Method: sparsity for generative models

Structured sparsity in demographic language variation[2]

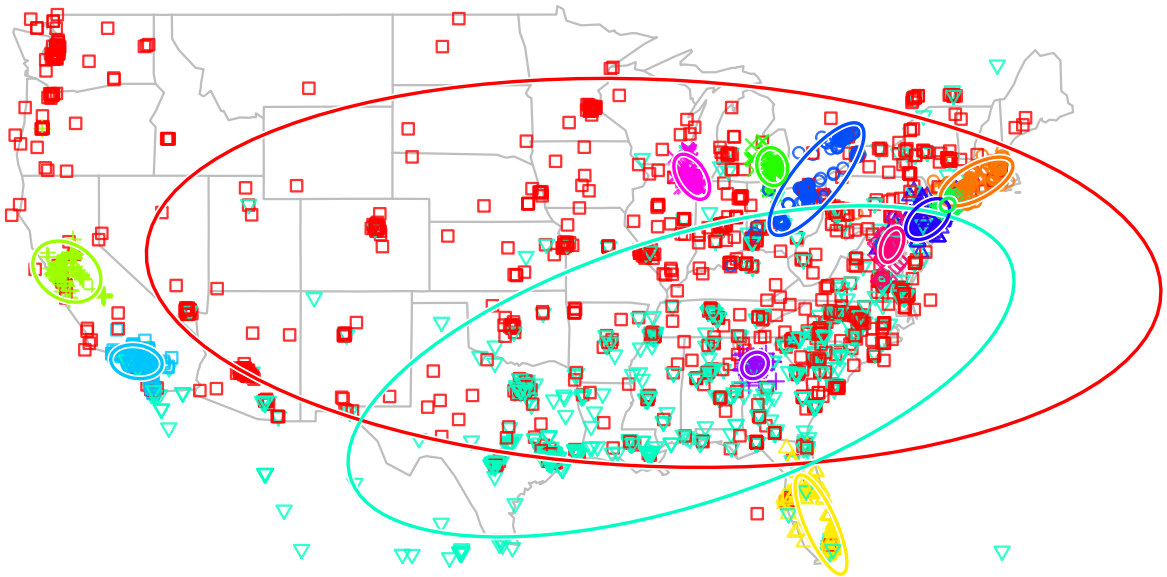
Highlights

- Language in social media displays marked demographic variation.
- *Structured* sparsity identifies a compact set words which act as demographic markers.

[2] Eisenstein, Smith and Xing. Discovering Sociolinguistic Associations with Structured Sparsity. ACL 2011.

Motivation

From EMNLP 2010:[3]

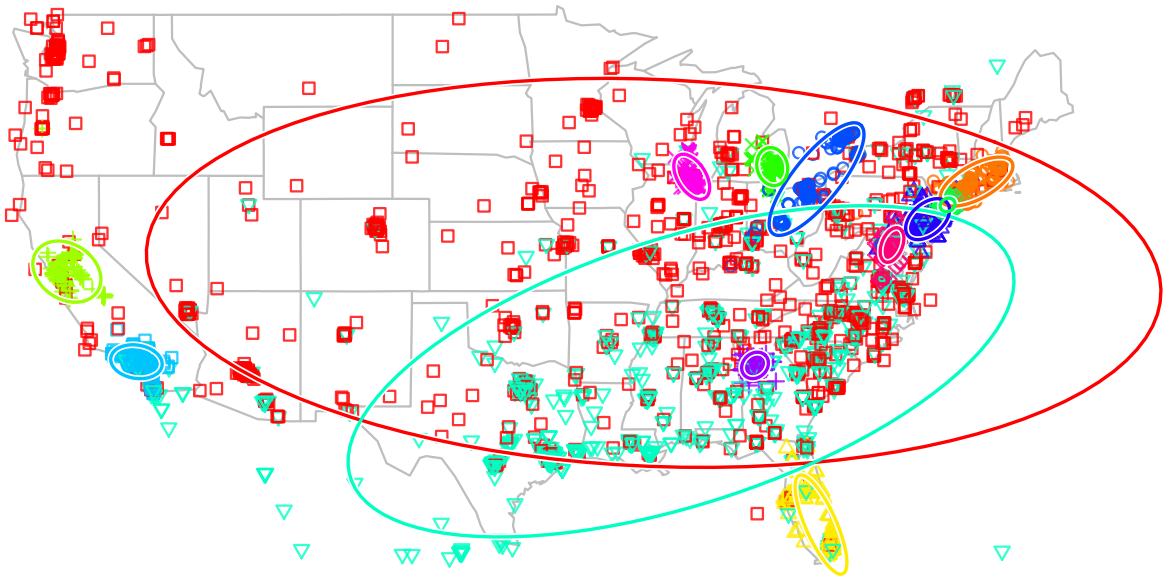


- Los Angeles dialect: coo, af, wyd, fasho, bomb

[3] Eisenstein, O'Connor, Smith, and Xing. A Latent Variable Model of Geographic Lexical Variation.

Motivation

From EMNLP 2010:[3]



- Los Angeles dialect: coo, af, wyd, fasho, bomb
- My LA in-laws don't use any of these words, ever.

[3] Eisenstein, O'Connor, Smith, and Xing. A Latent Variable Model of Geographic Lexical Variation.

Language and demographics

Dialect variation is often aligned demographically.

- Social class in New York (Labov 1966)
- Race and ethnicity in Chicago (Gordon 2005)
- Jocks vs. burnouts in suburban Detroit (Eckert 1989)
- Yuppies vs. government employees in Beijing (Zhang 2005)

Language and demographics

Dialect variation is often aligned demographically.

- Social class in New York (Labov 1966)
- Race and ethnicity in Chicago (Gordon 2005)
- Jocks vs. burnouts in suburban Detroit (Eckert 1989)
- Yuppies vs. government employees in Beijing (Zhang 2005)

Language, geography, and multiple demographic features combine through complex, non-linear interactions (e.g., Eckert 1990)...

Language and demographics

Dialect variation is often aligned demographically.

- Social class in New York (Labov 1966)
- Race and ethnicity in Chicago (Gordon 2005)
- Jocks vs. burnouts in suburban Detroit (Eckert 1989)
- Yuppies vs. government employees in Beijing (Zhang 2005)

Language, geography, and multiple demographic features combine through complex, non-linear interactions (e.g., Eckert 1990)... so it is crucial to model demographic attributes jointly.

Language and demographics

A surge of recent research on inferring demographic author properties from social media content:

- **Geographical location**: Cheng et al. 2010; Eisenstein et al. 2010; Wing & Baldridge 2011
- **Gender**: Burger et al. 2011; Mukherjee & Liu 2010; Rao et al. 2010; Schler et al. 2006
- **Age**: Nguyen et al 2011; Rao et al. 2010

Language and demographics

A surge of recent research on inferring demographic author properties from social media content:

- **Geographical location**: Cheng et al. 2010; Eisenstein et al. 2010; Wing & Baldridge 2011
- **Gender**: Burger et al. 2011; Mukherjee & Liu 2010; Rao et al. 2010; Schler et al. 2006
- **Age**: Nguyen et al 2011; Rao et al. 2010

Our contributions

- joint prediction of multiple demographic attributes from text
- joint learning and feature selection
- exploratory analysis of demographic lexical variation

Data

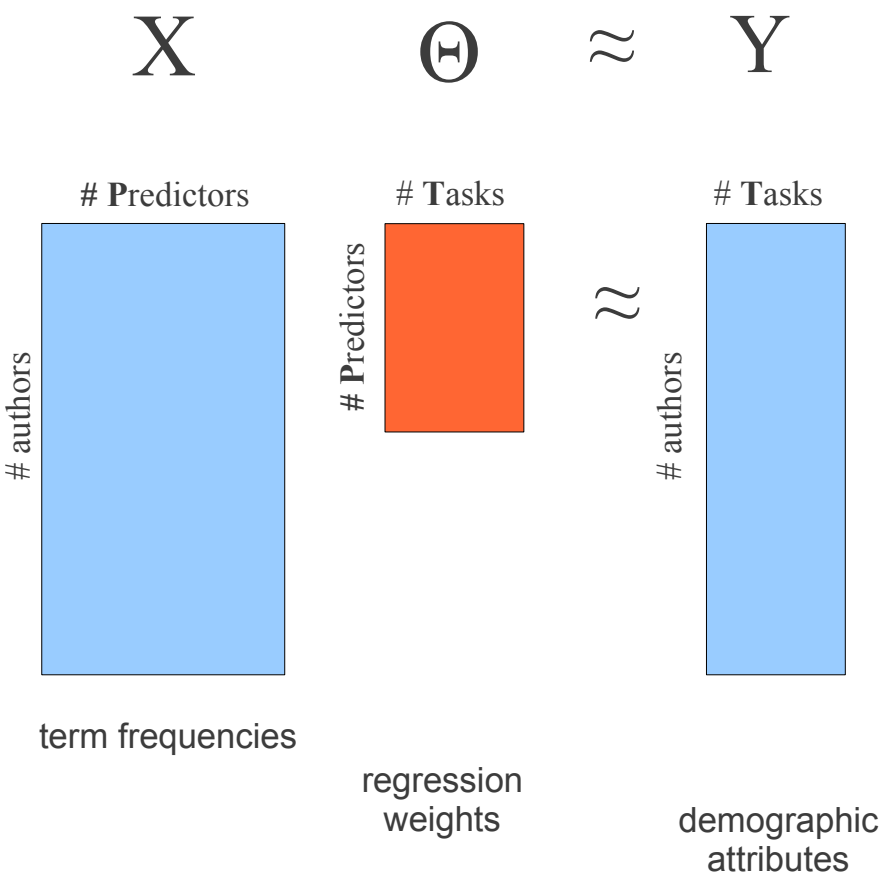
- Twitter Gardenhose feed from March 1-7, 2010
- 9250 authors, 380,000 messages, 4.7 million tokens
- Filters:
  - At least 20 messages (in Gardenhose)
  - Messages must include GPS within a USA zipcode
  - No more than 1000 followers, followees
- GPS $\rightarrow$ Zipcode $\rightarrow$ U.S. Census Demographic Statistics
  - Zipcodes commonly proxy for demographics in public health.
  - Careful! Twitter users are not an unbiased sample from a zipcode.

+

Data summary

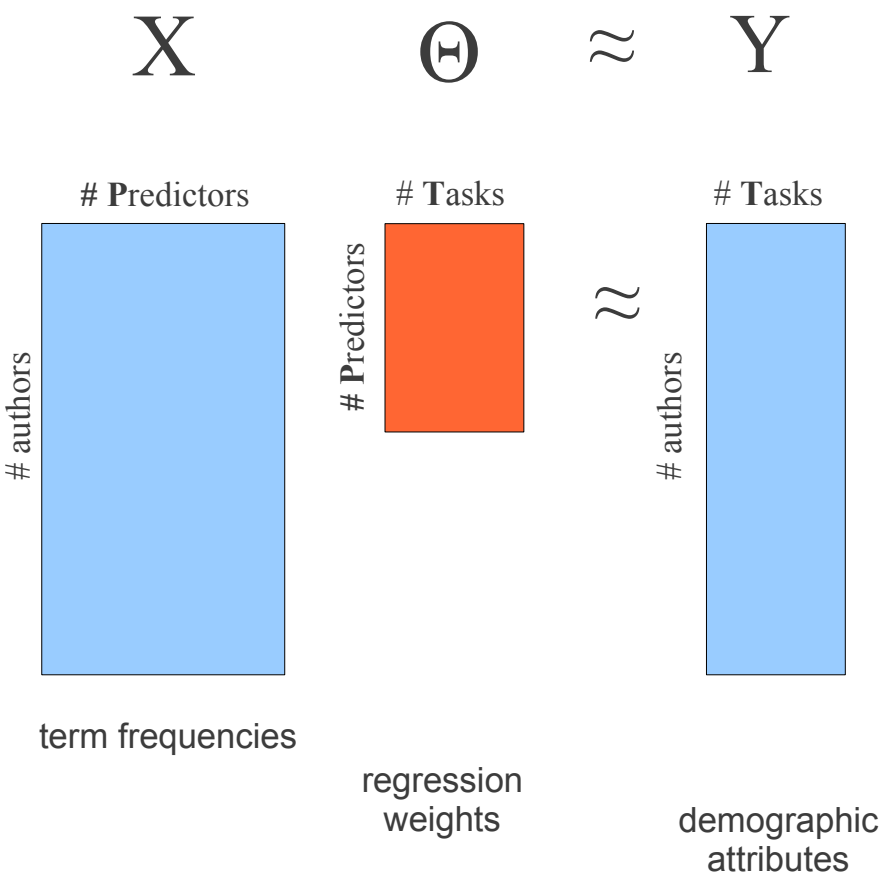
	mean	std. dev.
race & ethnicity		
% white	52.1	29.0
% African American	32.2	29.1
% Hispanic	15.7	18.3
language		
% English speakers	73.7	18.4
% Spanish speakers	14.6	15.6
% other language speakers	11.7	9.2
socioeconomic		
% urban	95.1	14.3
% with family	64.1	14.4
% renters	48.9	23.4
median income ($)	42,500	18,100

Model



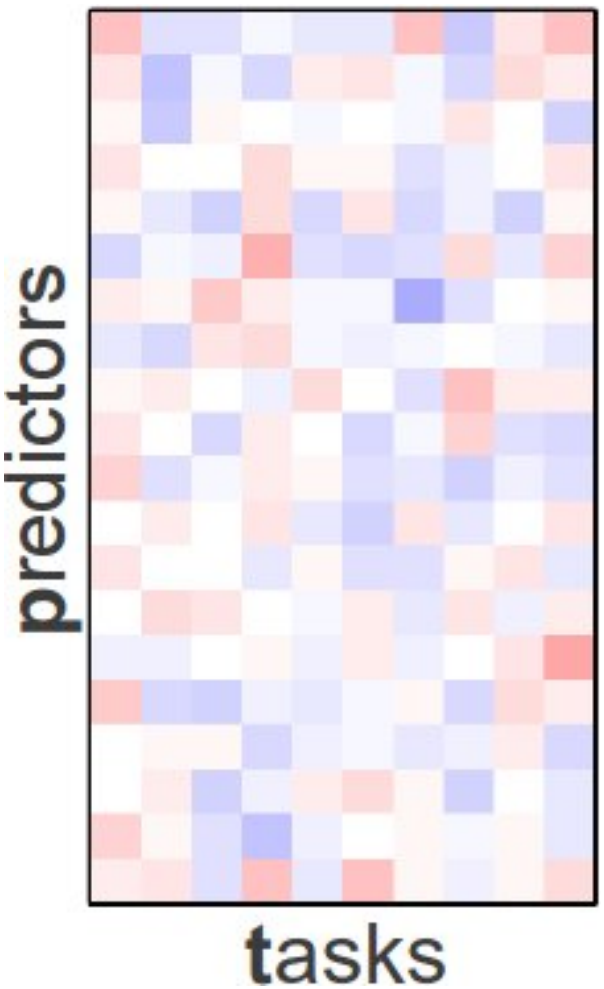
- Regress demographics against term frequencies
- $\min_{\mathbf{\Theta}} \ell(\mathbf{X\Theta} - \mathbf{Y}) + \Omega(\mathbf{\Theta})$
  - Loss: $\ell(\mathbf{X\Theta} - \mathbf{Y})$
  - Regularizer: $\Omega(\mathbf{\Theta})$

Model



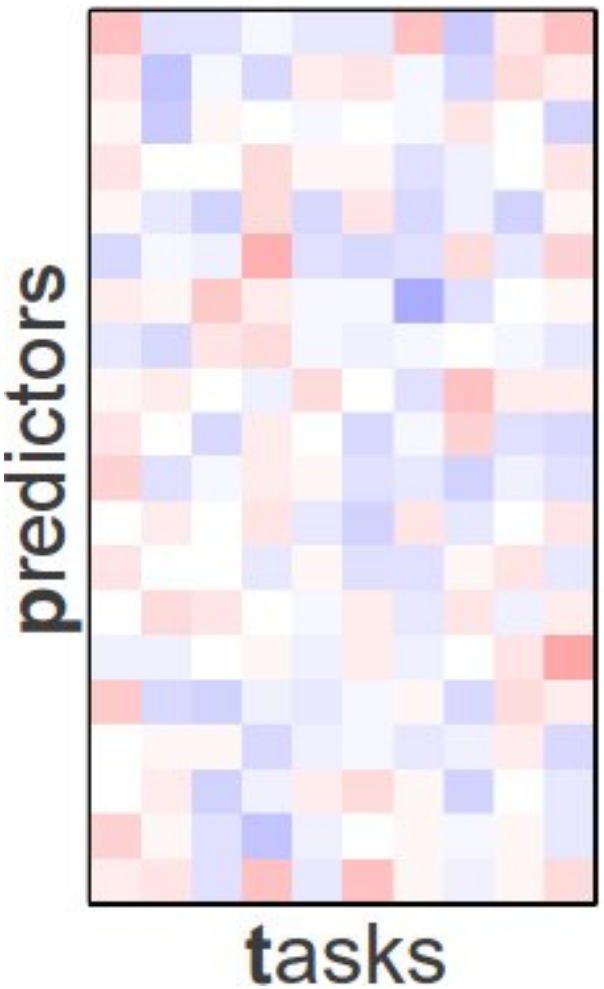
- Regress demographics against term frequencies
- $\min_{\mathbf{\Theta}} \ell(\mathbf{X\Theta} - \mathbf{Y}) + \Omega(\mathbf{\Theta})$
  - Loss: $\ell(\mathbf{X\Theta} - \mathbf{Y})$
  - Regularizer: $\Omega(\mathbf{\Theta})$
- We'll use the regularizer to make the $P \times T$ weights into an interpretable model.

Three regularizers



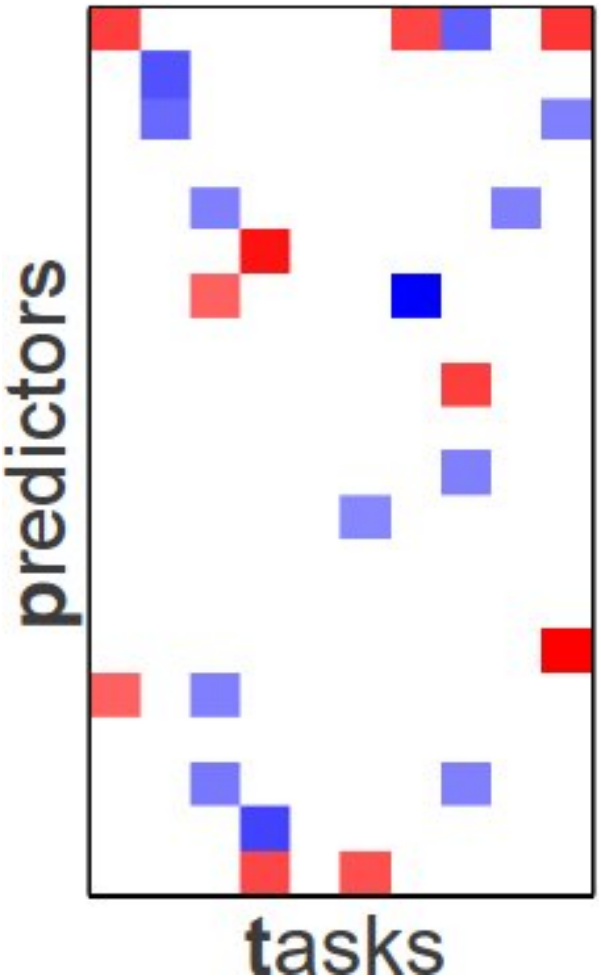
- L2 regularizer: $\Omega(\mathbf{\Theta}) = \sum_t \sum_p \theta_{pt}^2$
- encourages small regression coefficients
- equivalent to running T separate ridge regressions

Three regularizers



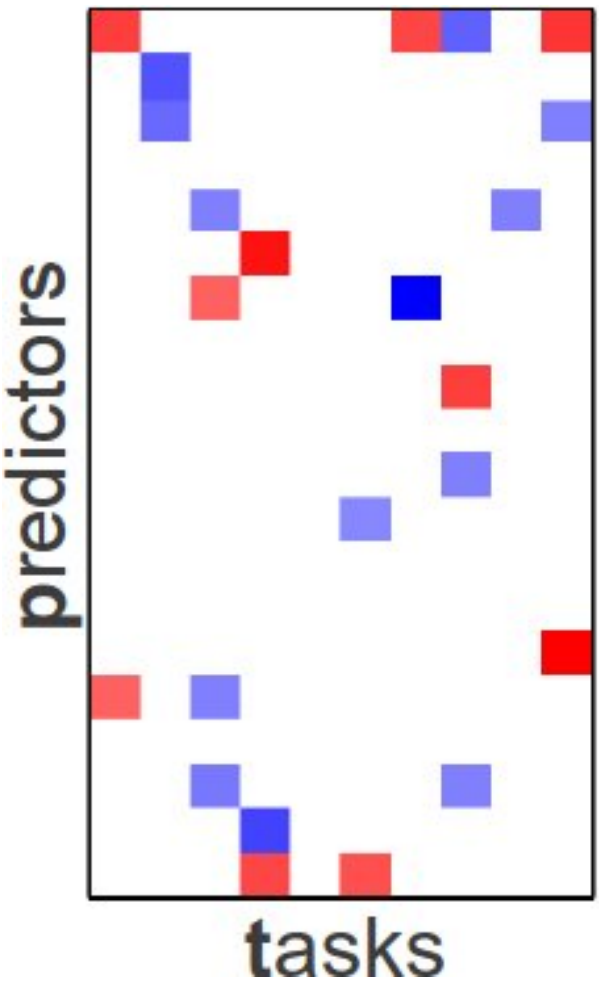
- L2 regularizer:
  $\Omega(\mathbf{\Theta}) = \sum_t \sum_p \theta_{pt}^2$
- encourages small regression coefficients
- equivalent to running T separate ridge regressions
- Interpretable? No.
  Every word-demographic association has a non-zero weight.

Three regularizers



- $L1$ regularizer:[a] $\Omega(\mathbf{\Theta}) = \sum_t \sum_p |\theta_{pt}|$
- encourages regression coefficients to go to zero
- equivalent to running T separate lasso regressions

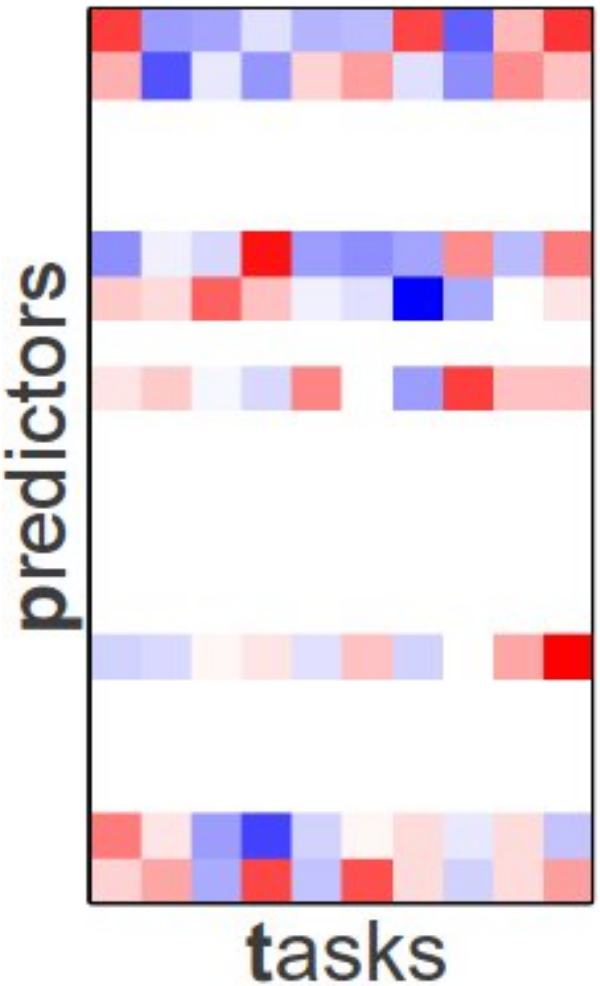
Three regularizers



- $L1$ regularizer:[a]
  $\Omega(\mathbf{\Theta}) = \sum_t \sum_p |\theta_{pt}|$
- encourages regression coefficients to go to zero
- equivalent to running T separate lasso regressions
- Interpretable? Sort of.
  L1 isolates a few robust associations, but it forces us to think about each demographic attribute separately.

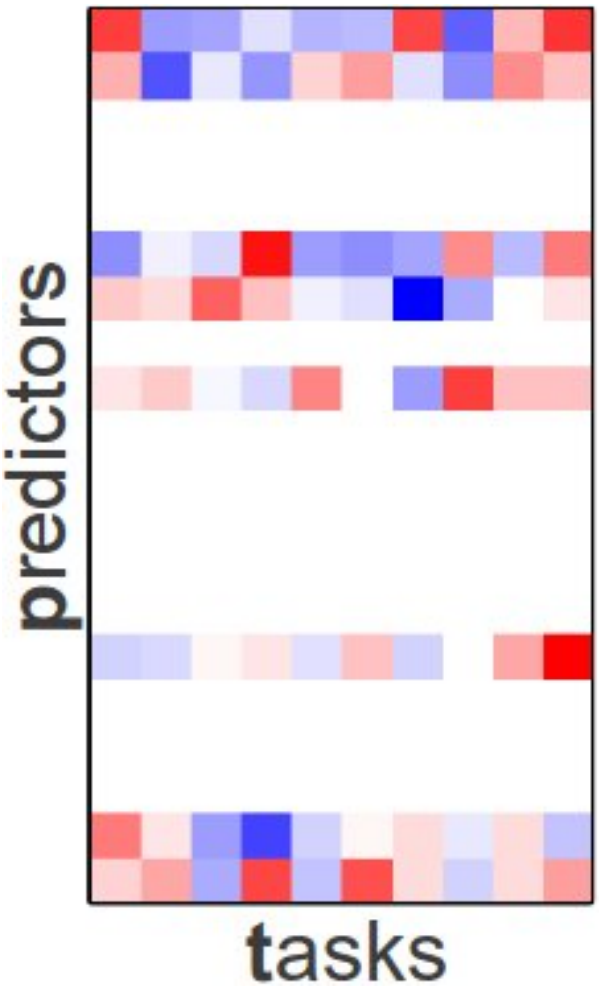
[a] Tibshirani, 1996

Three regularizers



- $L1/L\infty$ regularizer:[a] $\Omega(\mathbf{\Theta}) = \sum_t \max_p \theta_{pt}$
- pushes entire rows to zero
- **multi-output lasso**; not equivalent to any decomposition of original problem

Three regularizers

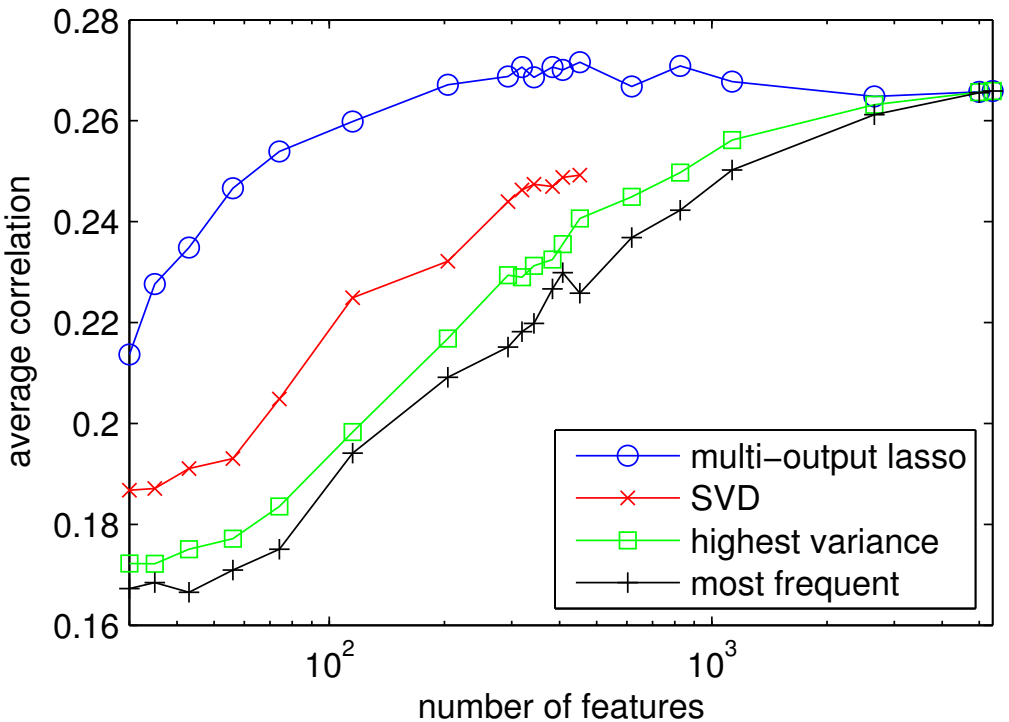


- $L1/L\infty$ regularizer:[a]
  $\Omega(\mathbf{\Theta}) = \sum_t \max_p \theta_{pt}$
- pushes entire rows to zero
- **multi-output lasso**;
  not equivalent to any decomposition of original problem
- Interpretable? Yes!
  Isolates a few words which have strong associations across a range of demographic attributes.

[a] Turlach, Venables, and Wright, 2005

Regularization path

By increasing the strength of the regularizer, we disgard more and more terms from the vocabulary.



Predictive accuracy remains high even when the vocabulary is reduced by more than an order of magnitude.

Predictive accuracy

vocabulary	**# features**	average	white	Afr. Am.	Hisp.
full	5418	0.260	0.337	0.318	0.296
multi-output lasso	394.6	0.260	0.326	0.308	0.304
SVD		0.237	0.321	0.299	0.269
highest variance		0.220	0.309	0.287	0.245
most frequent		0.204	0.294	0.264	0.222

- **metric**: Pearson correlation between predicted and true values of demographic attributes
- **baselines**:
  - SVD of author-term matrix
  - highest variance words
  - most frequent words

Predictive accuracy

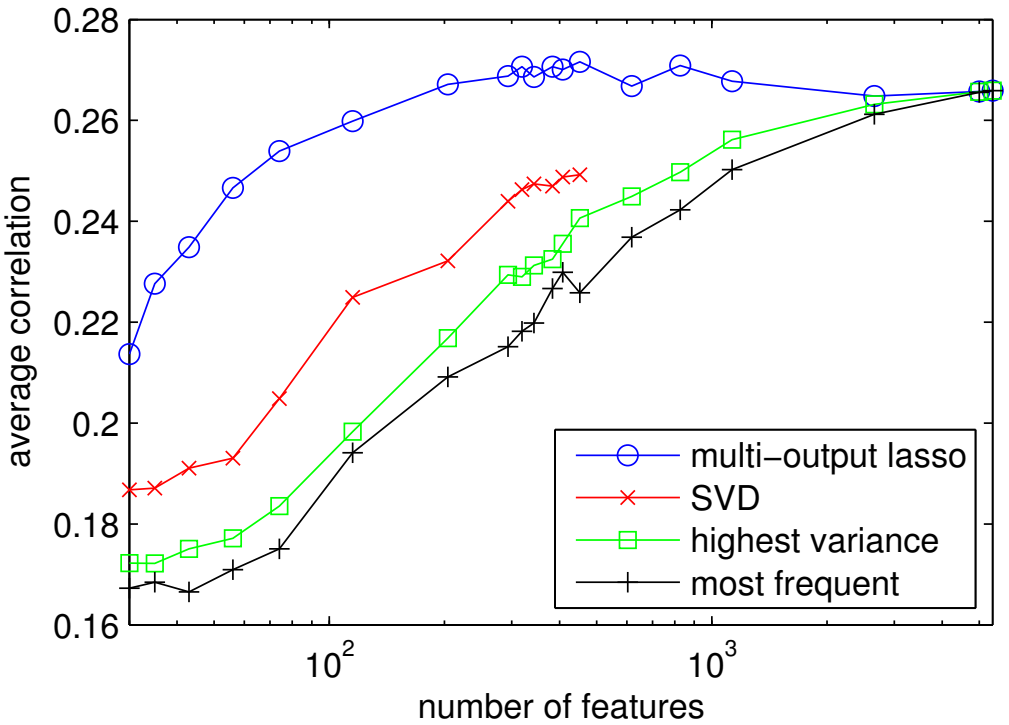
vocabulary	**# features**	average	Eng. lang.	Span. lang.	other lang.
full	5418	0.260	0.384	0.296	0.256
multi-output lasso	394.6	0.260	0.383	0.303	0.249
SVD		0.237	0.352	0.272	0.226
highest variance		0.220	0.315	0.248	0.199
most frequent		0.204	0.293	0.229	0.178

Predictive accuracy

vocabulary	**# features**	average	urban	family	renter	med. inc.
full	5418	0.260	0.155	0.113	0.295	0.152
multi-output lasso	394.6	0.260	0.153	0.113	0.302	0.156
SVD		0.237	0.138	0.081	0.278	0.136
highest variance		0.220	0.132	0.085	0.250	0.135
most frequent		0.204	0.129	0.073	0.228	0.126

- All confidence intervals are tighter than ± 0.02.
- 93% reduction in model size, no loss in performance.
- Demographic attributes are more difficult to predict than race, ethnicity, and language.

Qualitative analysis



- In predictive experiments, regularization chosen on dev set
- Now we tune it to identify a more compact set of 69 terms

Place names and foreign language

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
atlanta			-	+	-	-				
famu		+	-	+	-	-				-
harlem				-					+	
con			+	-	+				+	
la		-	+	-	+					
si		-	+	-	+					

The symbols + and - indicate significant positive or negative association, as measured by a Wald Test, with Bonferroni correction.

Standard English

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
as			-	+	-					
awesome	+	-					-		-	+
break			-	+	-	-				
campus			-	+	-	-				
dead	-	+		-	+		+		+	
hell			-	+	-	-				
shit	-								+	
train				-	+				+	
will			-	+	-					
would				+					-	

Abbreviations

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
bbm	-	+		-		+	+		+	
lls		+	-	+	-	-				
lmaoo	-	+	+	-	+	+	+		+	
lmaooo	-	+	+	-	+	+	+		+	
lmaoooo	-	+	+	-	+	+			+	
lmfaoo	-		+	-	+	+			+	
lmfaooo	-		+	-	+	+			+	
lml	-	+	+	-	+	+	+		+	-
odee	-		+	-	+		+		+	
omw	-	+	+	-	+	+	+		+	
smfh	-	+	+	-	+	+	+		+	
smh	-	+					+		+	
w	-		+	-	+	+	+		+	

:-p	:)	:(	:-)	:-|	
+				'	white
'	'	'	'		Afr. Am.
+			+	+	Hisp.
'			'	'	Eng. lang.
+			+	+	Span. lang.
				+	other lang.
				+	urban
					family
					renter
					med. inc.

Other

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
dats	-	+		-					+	-
deadass	-	+	+	-	+	+	+		+	
haha	+	-							-	
hahah	+	-								
hahaha	+	-							-	+
ima	-		+	-	+				+	
madd	-			-		+		+		
nah	-		+	-	+	+			+	
ova	-	+		-					+	
sis	-	+							+	
skool	-	+		-		+	+		+	-
wassup	-	+	+	-	+	+	+		+	-
wat	-	+	+	-	+	+	+		+	-
ya	-	+							+	
yall	-	+								
yep			-	+	-	-	-		-	
yoo	-	+	+	-	+	+	+		+	
yooo	-	+		-	+				+	

Overview

- Marked difference in how paralinguistic commentary is performed
  - **Emoticons** correlate with % Whites, English speakers, Hispanics
  - **Abbreviations** correlate % African Americans, Hispanics, renters
- **Phoneticization** correlates strongly with % African American
  - May reflect phonological features of AAE (“dats”, “ima”, “wassup”)

Summary of part 1

- **Data**: By mashing up geotagged messages with census data, we can study the demographics of lexical variation in social media on a large scale
- **Method**: Linear regression with structured sparsity
  - yields robust predictions with 93% compression of the vocabulary
  - enables exploratory qualitative analysis of demographic lexical variation

Outline

- Structured sparsity in demographic language variation
- **Sparsity for generative models**

Generative models of text

Generative models are a powerful tool for understanding document collections.

- Classfication/clustering (Naive Bayes)
- Discover latent themes (LDA)
- Distinguish latent and observed factors (e.g. Topic-aspect models)

Generative models of text

Generative models are a powerful tool for understanding document collections.

- Classfication/clustering (Naive Bayes)
- Discover latent themes (LDA)
- Distinguish latent and observed factors (e.g. Topic-aspect models)

Unifying idea: each class or latent theme is represented by a distribution over tokens, $P(w|y)$

Redundancy

- A naïve Bayes classifier must estimate the parameter $Pr(w = \text{“the”} | y)$ for every class y.

Redundancy

- A naïve Bayes classifier must estimate the parameter $Pr(w = \text{“the”} | y)$ for every class y.
- The probability $Pr(w = \text{“the”})$ is a fact about English, not about any of the classes (usually).

Redundancy

- A naïve Bayes classifier must estimate the parameter $Pr(w = \text{“the”} | y)$ for every class y.
- The probability $Pr(w = \text{“the”})$ is a fact about English, not about any of the classes (usually).
- Heuristic solutions like stopword pruning are hard to generalize to new domains.

Redundancy

- A naïve Bayes classifier must estimate the parameter $Pr(w = \text{“the”} | y)$ for every class y.
- The probability $Pr(w = \text{“the”})$ is a fact about English, not about any of the classes (usually).
- Heuristic solutions like stopword pruning are hard to generalize to new domains.
- It would be better to focus computation on parameters that distinguish the classes.

Overparametrization

- An LDA **model** with K topics and V words requires $K \times V$ parameters.
- An LDA **paper** shows 10 words per topic.

Overparametrization

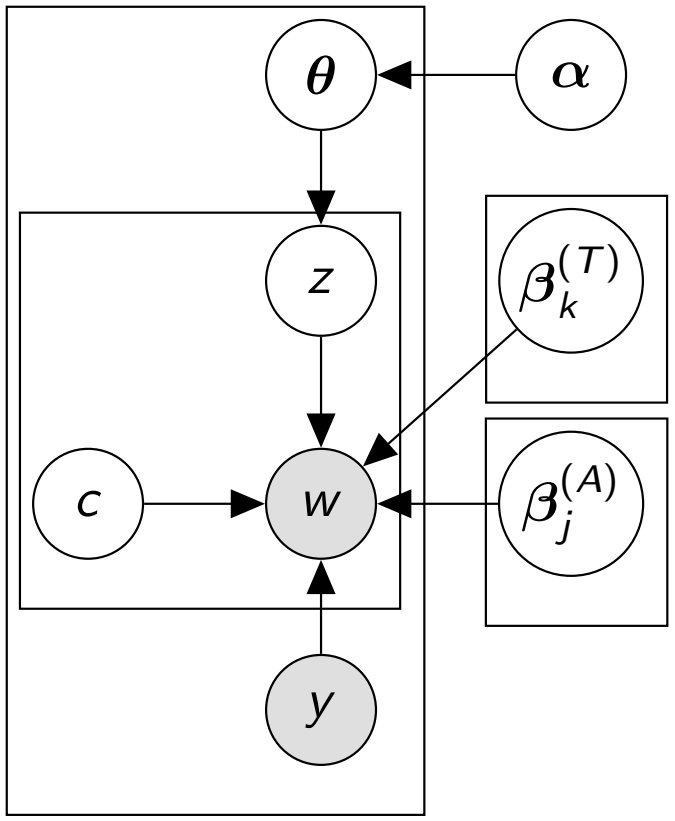
- An LDA **model** with K topics and V words requires $K \times V$ parameters.
- An LDA **paper** shows 10 words per topic.
- What about the other $V - 10$ words per topic??

Overparametrization

- An LDA **model** with K topics and V words requires $K \times V$ parameters.
- An LDA **paper** shows 10 words per topic.
- What about the other $V - 10$ words per topic??
  - These parameters affect the assignment of documents...
  - But they may be unnoticed by the user.
  - And there may not be enough data to estimate them accurately.

Inference complexity

- Latent topics may be combined with additional facets, such as sentiment and author perspective.
- "Switching" variables decide if a word is drawn from a topic or from another facet.
- Twice as many latent variables per document!



Sparse Additive Generative Models

- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \beta_y$

Sparse Additive Generative Models

- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \beta_y$
- **Sparse Additive GEnerative models**:
  each class or latent theme is represented by its deviation from a background distribution.

$$P(w|y, \mathbf{m}) \propto \exp\left(\mathbf{m} + \eta_y\right)$$

Sparse Additive Generative Models

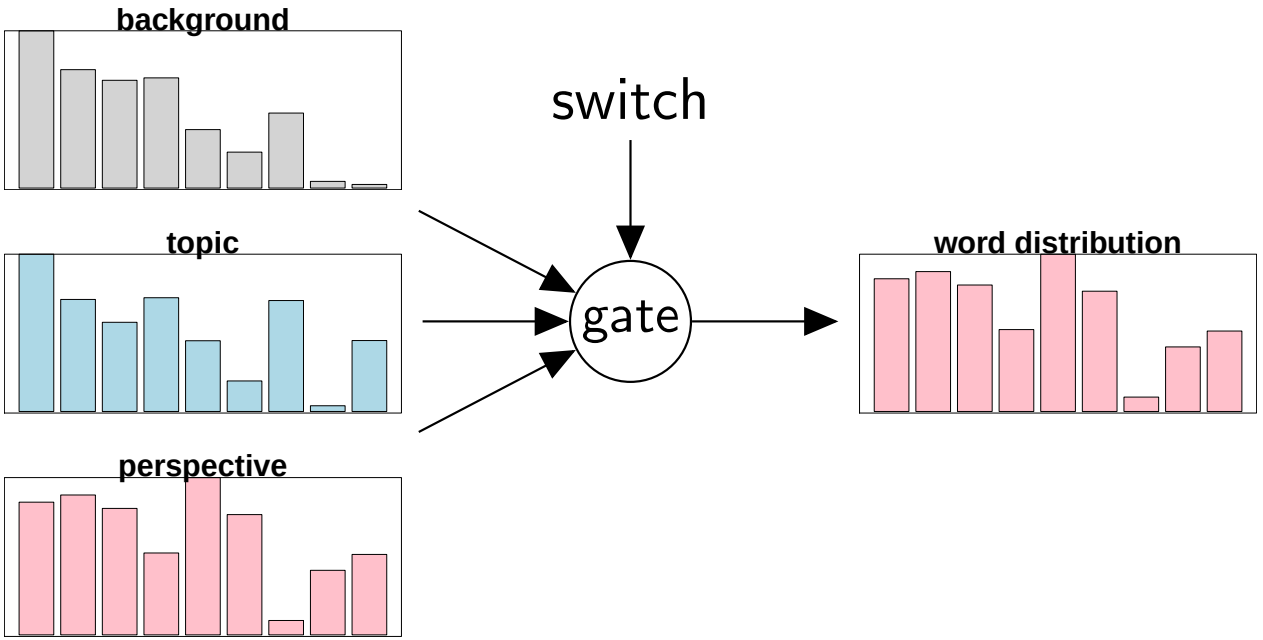
- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \beta_y$
- **Sparse Additive GEnerative models**:
  each class or latent theme is represented by its deviation from a background distribution.

$$P(w|y, \mathbf{m}) \propto \exp\left(\mathbf{m} + \boldsymbol{\eta}_y\right)$$

  - $\mathbf{m}$ captures the background word log-probabilities
  - $\boldsymbol{\eta}$ contains sparse deviations for each topic or class
  - additional facets can be added in log-space

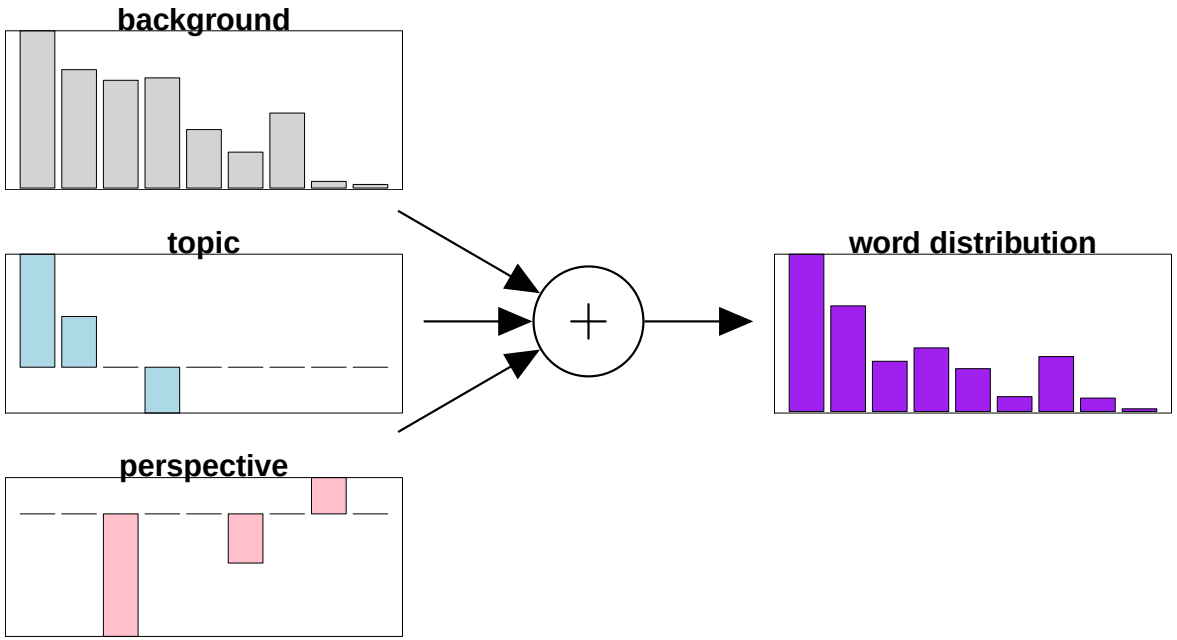
Sparse Additive Generative Models

A topic-perspective-background model using Dirichlet-multinomials:



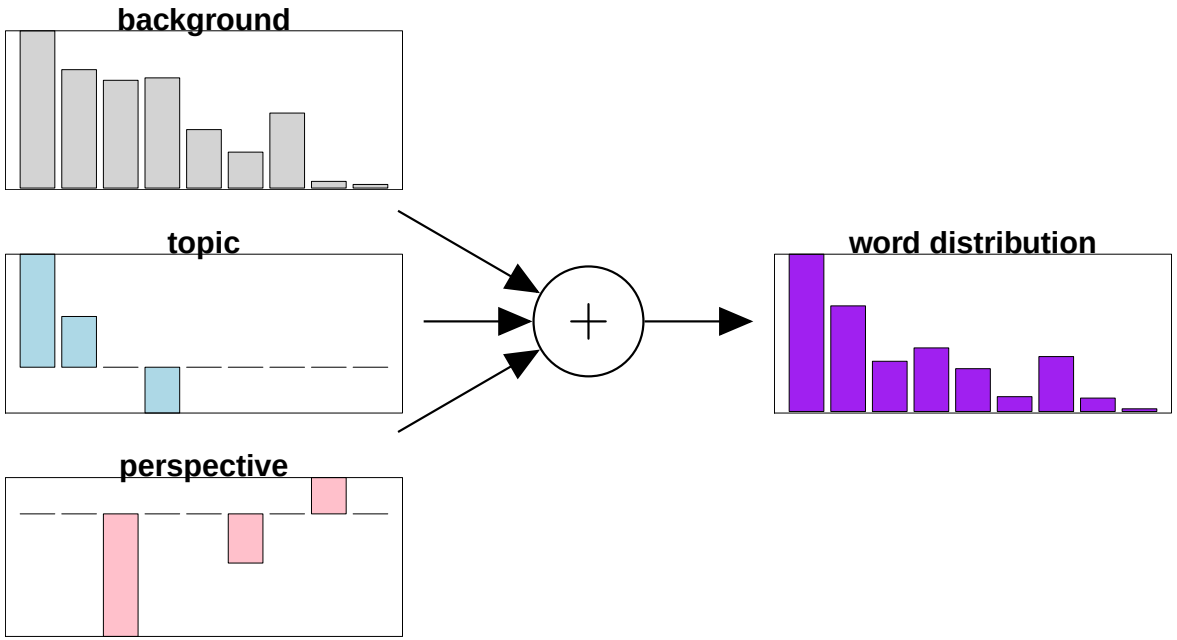
Sparse Additive Generative Models

A topic-perspective-background model using SAGE:



Sparse Additive Generative Models

A topic-perspective-background model using SAGE:



Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many i

Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many i
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i|\boldsymbol{\eta} + \mathbf{m}) \neq Pr(w = i|\mathbf{m})$, even if $\eta_i = 0$.

Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many i
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i|\boldsymbol{\eta} + \mathbf{m}) \neq Pr(w = i|\mathbf{m})$, even if $\eta_i = 0$.
- But for most pairs of words, $\frac{Pr(w=i|\boldsymbol{\eta}+\mathbf{m})}{Pr(w=j|\boldsymbol{\eta}+\mathbf{m})} = \frac{Pr(w=i|\mathbf{m})}{Pr(w=j|\mathbf{m})}$

Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many i
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i|\boldsymbol{\eta} + \mathbf{m}) \neq Pr(w = i|\mathbf{m})$, even if $\eta_i = 0$.
- But for most pairs of words, $\frac{Pr(w=i|\boldsymbol{\eta}+\mathbf{m})}{Pr(w=j|\boldsymbol{\eta}+\mathbf{m})} = \frac{Pr(w=i|\mathbf{m})}{Pr(w=j|\mathbf{m})}$

Sparsity deviation of log probabilities

- Sparsity: $\eta_i = 0$ for many i
- Due to normalization, the generative probabilities will not be identical, $Pr(w = i|\boldsymbol{\eta} + \mathbf{m}) \neq Pr(w = i|\mathbf{m})$, even if $\eta_i = 0$.
- But for most pairs of words, $\frac{Pr(w=i|\boldsymbol{\eta}+\mathbf{m})}{Pr(w=j|\boldsymbol{\eta}+\mathbf{m})} = \frac{Pr(w=i|\mathbf{m})}{Pr(w=j|\mathbf{m})}$

Different notion of sparsity from sparseTM (Wang & Blei, 2009), which sets $Pr(w = i|y) = 0$ for many i.

Sparsity through integration

- The Laplace distribution is equivalent to an $L1$ regularizer: $\eta \sim \mathcal{L}(0, \sigma)$

Sparsity through integration

- The Laplace distribution is equivalent to an $L1$ regularizer: $\eta \sim \mathcal{L}(0, \sigma)$
  - We can apply the integral: $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\text{Exp}(\tau; \sigma)d\tau$ (Lange & Simsheimer, 1993)

Sparsity through integration

- The Laplace distribution is equivalent to an $L1$ regularizer:
  $\eta \sim \mathcal{L}(0, \sigma)$
  - We can apply the integral:
    $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\text{Exp}(\tau; \sigma)d\tau$ (Lange & Simsheimer, 1993)
  - Other integrals also induce sparsity, e.g.
    $\int \mathcal{N}(\eta; 0, \tau)\frac{1}{\tau}d\tau$ (Figueiredo, 2001; Guan & Dy, 2009)

Sparsity through integration

- The Laplace distribution is equivalent to an $L1$ regularizer:
  $\eta \sim \mathcal{L}(0, \sigma)$
  - We can apply the integral:
    $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\text{Exp}(\tau; \sigma)d\tau$ (Lange & Simsheimer, 1993)
  - Other integrals also induce sparsity, e.g.
    $\int \mathcal{N}(\eta; 0, \tau)\frac{1}{\tau}d\tau$ (Figueiredo, 2001; Guan & Dy, 2009)
- We solve this integral through coordinate ascent, updating:

Sparsity through integration

- The Laplace distribution is equivalent to an $L1$ regularizer: $\eta \sim \mathcal{L}(0, \sigma)$
  - We can apply the integral: $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\text{Exp}(\tau; \sigma)d\tau$ (Lange & Simsheimer, 1993)
  - Other integrals also induce sparsity, e.g. $\int \mathcal{N}(\eta; 0, \tau)\frac{1}{\tau}d\tau$ (Figueiredo, 2001; Guan & Dy, 2009)
- We solve this integral through coordinate ascent, updating:
  - The variational distribution $Q(\boldsymbol{\tau})$

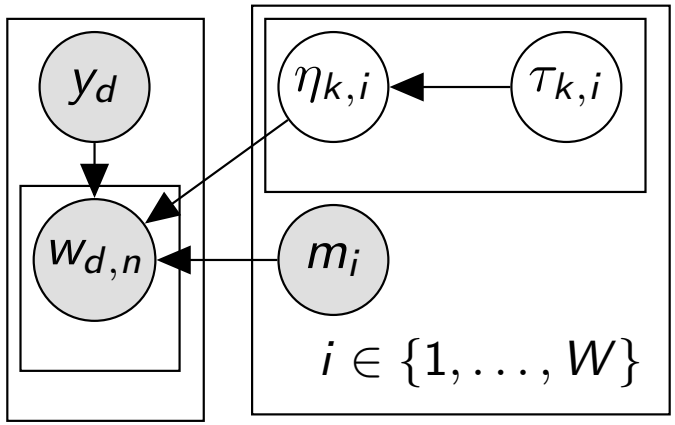
Sparsity through integration

- The Laplace distribution is equivalent to an $L1$ regularizer: $\eta \sim \mathcal{L}(0, \sigma)$
  - We can apply the integral: $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\text{Exp}(\tau; \sigma)d\tau$ (Lange & Simsheimer, 1993)
  - Other integrals also induce sparsity, e.g. $\int \mathcal{N}(\eta; 0, \tau)\frac{1}{\tau}d\tau$ (Figueiredo, 2001; Guan & Dy, 2009)
- We solve this integral through coordinate ascent, updating:
  - The variational distribution $Q(\boldsymbol{\tau})$
  - A **point estimate** of $\boldsymbol{\eta}$

Applications

- Document classification
- Topic models
- Multifaceted topic models

SAGE in document classification



- Each document d has a label y_d
- Each token $w_{d,n}$ is drawn from a multinomial distribution $\boldsymbol{\beta}$, where $\beta_i = \frac{\exp\left(\eta_{y_d,i} + m_i\right)}{\sum_j \exp\left(\eta_{y_d,j} + m_j\right)}$
- Each parameter $\eta_{k,i}$ is drawn from a distribution equal to $\mathcal{N}(0, \tau_{k,i})$, with $P(\tau_{k,i}) \sim 1/\tau_{k,i}$

Inference

- We maximize the variational bound

$$\ell = \sum_d \sum_n^{N_d} \log P(w_n^{(d)} | \mathbf{m}, \boldsymbol{\eta}_{y_d}) + \sum_k \langle \log P(\boldsymbol{\eta}_k | \mathbf{0}, \boldsymbol{\tau}_k) \rangle$$
$$+ \sum_k \langle \log P(\boldsymbol{\tau}_k | \gamma) \rangle - \sum_k \langle \log Q(\boldsymbol{\tau}_k) \rangle ,$$

Inference

- We maximize the variational bound

$$\ell = \sum_d \sum_n^{N_d} \log P(w_n^{(d)} | \mathbf{m}, \boldsymbol{\eta}_{y_d}) + \sum_k \langle \log P(\boldsymbol{\eta}_k | \mathbf{0}, \boldsymbol{\tau}_k) \rangle + \sum_k \langle \log P(\boldsymbol{\tau}_k | \gamma) \rangle - \sum_k \langle \log Q(\boldsymbol{\tau}_k) \rangle ,$$

- The gradient wrt $\boldsymbol{\eta}$ is,

$$\frac{\partial \ell}{\partial \boldsymbol{\eta}_k} = \mathbf{c}_k - C_k \boldsymbol{\beta}_k - \text{diag}\left(\left\langle \boldsymbol{\tau}_k^{-1} \right\rangle\right) \boldsymbol{\eta}_k,$$

where

  - $\mathbf{c}_k$ are the observed counts for class k
  - $C_k = \sum_i c_{ki}$
  - $\boldsymbol{\beta}_k \propto \exp(\boldsymbol{\eta}_k + \mathbf{m})$

Inference

- We maximize the variational bound

$$\ell = \sum_d \sum_n^{N_d} \log P(w_n^{(d)} | \mathbf{m}, \boldsymbol{\eta}_{y_d}) + \sum_k \langle \log P(\boldsymbol{\eta}_k | \mathbf{0}, \boldsymbol{\tau}_k) \rangle$$
$$+ \sum_k \langle \log P(\boldsymbol{\tau}_k | \gamma) \rangle - \sum_k \langle \log Q(\boldsymbol{\tau}_k) \rangle ,$$

Inference

- We maximize the variational bound

$$\ell = \sum_d \sum_n^{N_d} \log P(w_n^{(d)} | \mathbf{m}, \boldsymbol{\eta}_{y_d}) + \sum_k \langle \log P(\boldsymbol{\eta}_k | \mathbf{0}, \boldsymbol{\tau}_k) \rangle$$
$$+ \sum_k \langle \log P(\boldsymbol{\tau}_k | \gamma) \rangle - \sum_k \langle \log Q(\boldsymbol{\tau}_k) \rangle ,$$

- We choose $Q(\tau_{k,i}) = \text{Gamma}(\tau_{k,i}; a_{k,i}, b_{k,i})$

Inference

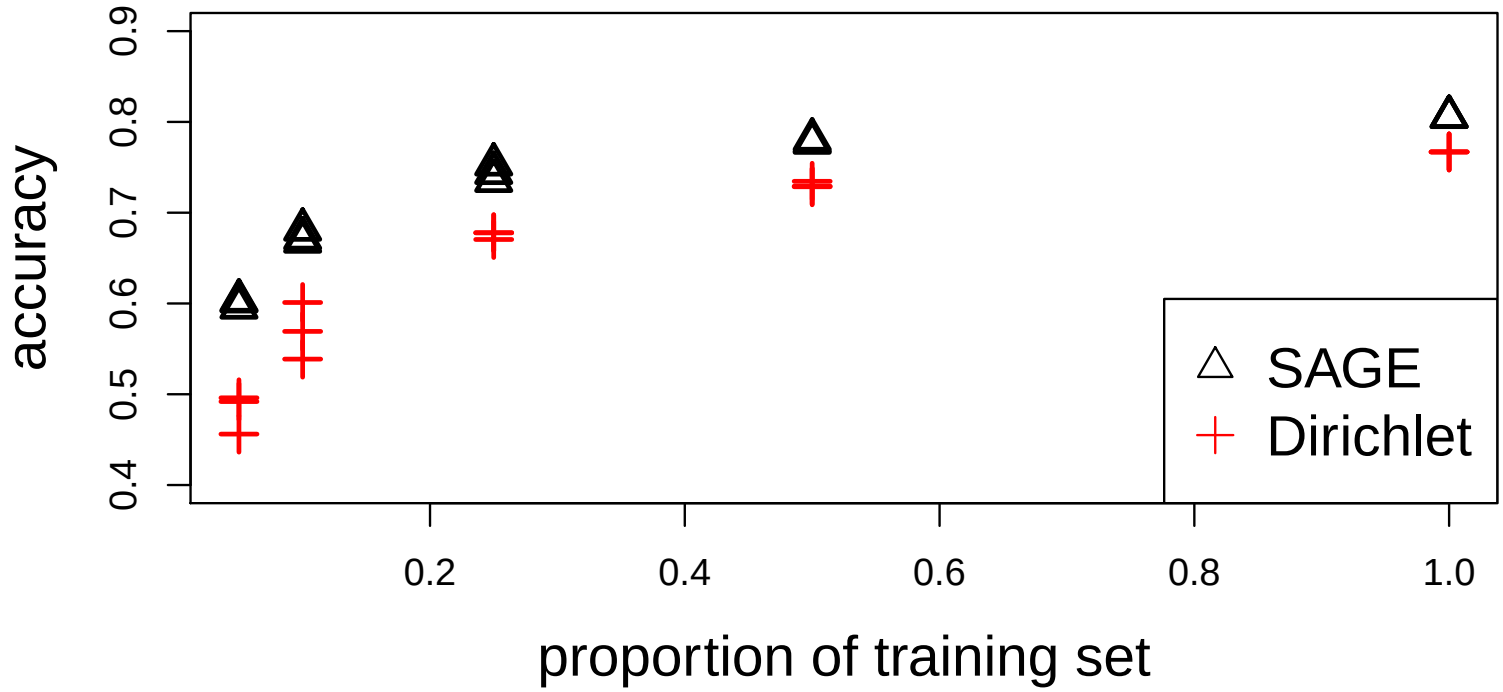
- We maximize the variational bound

$$\ell = \sum_d \sum_n^{N_d} \log P(w_n^{(d)} | \mathbf{m}, \boldsymbol{\eta}_{y_d}) + \sum_k \langle \log P(\boldsymbol{\eta}_k | \mathbf{0}, \boldsymbol{\tau}_k) \rangle$$
$$+ \sum_k \langle \log P(\boldsymbol{\tau}_k | \gamma) \rangle - \sum_k \langle \log Q(\boldsymbol{\tau}_k) \rangle ,$$

- We choose $Q(\tau_{k,i}) = \text{Gamma}(\tau_{k,i}; a_{k,i}, b_{k,i})$
- Iterate between a Newton update to a and a closed-form update to b

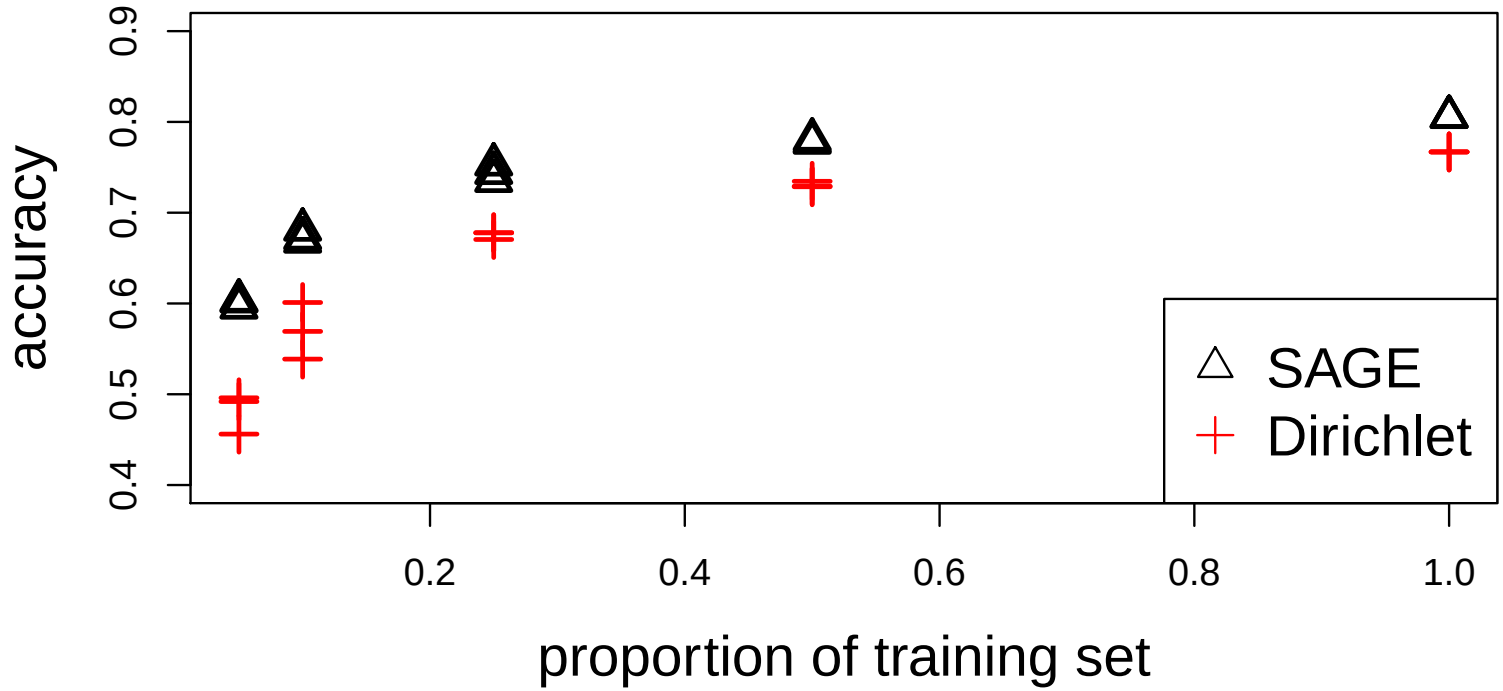
Document classification evaluation

- 20 newsgroups data: 11K training docs, 50K vocab



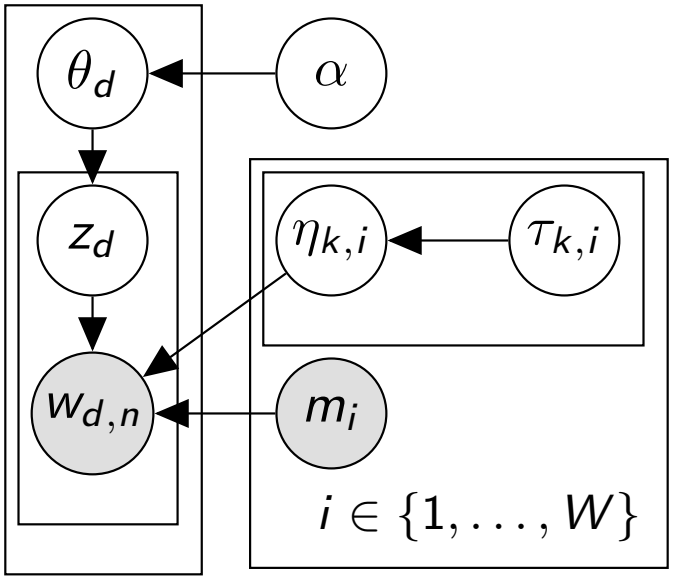
Document classification evaluation

- 20 newsgroups data: 11K training docs, 50K vocab

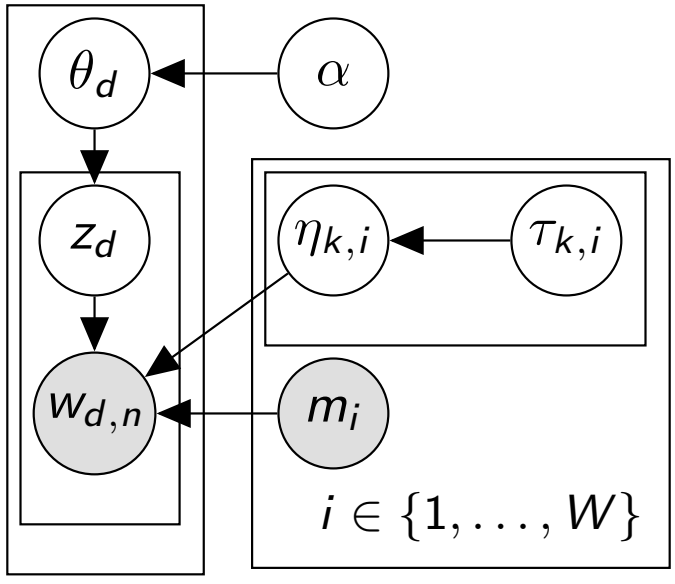


- Adaptive sparsity:
  - 10% non-zeros for full training set (11K docs)
  - 2% non-zeros for minimal training set (550 docs)

SAGE in latent variable models



SAGE in latent variable models



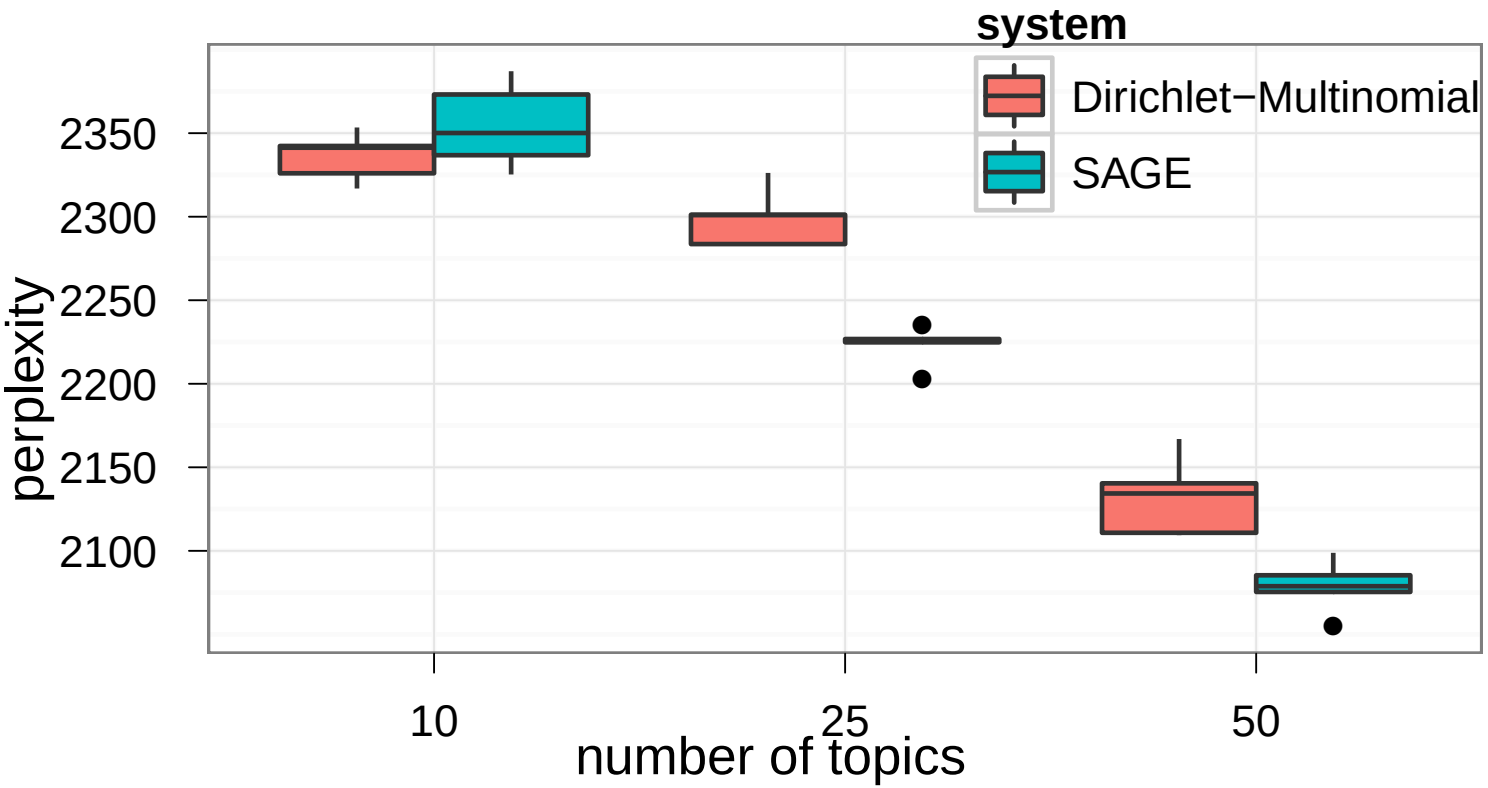
The gradient for $\boldsymbol{\eta}$ now includes **expected** counts:

$$\frac{\partial \ell}{\partial \boldsymbol{\eta}_k} = \langle \mathbf{c}_k \rangle - \langle C_k \rangle \boldsymbol{\beta}_k - \text{diag}\left(\left\langle \boldsymbol{\tau}_k^{-1} \right\rangle\right) \boldsymbol{\eta}_k,$$

where $\langle c_{ki} \rangle = \sum_n Q_{z_n}(k)\delta(w_n = i)$.

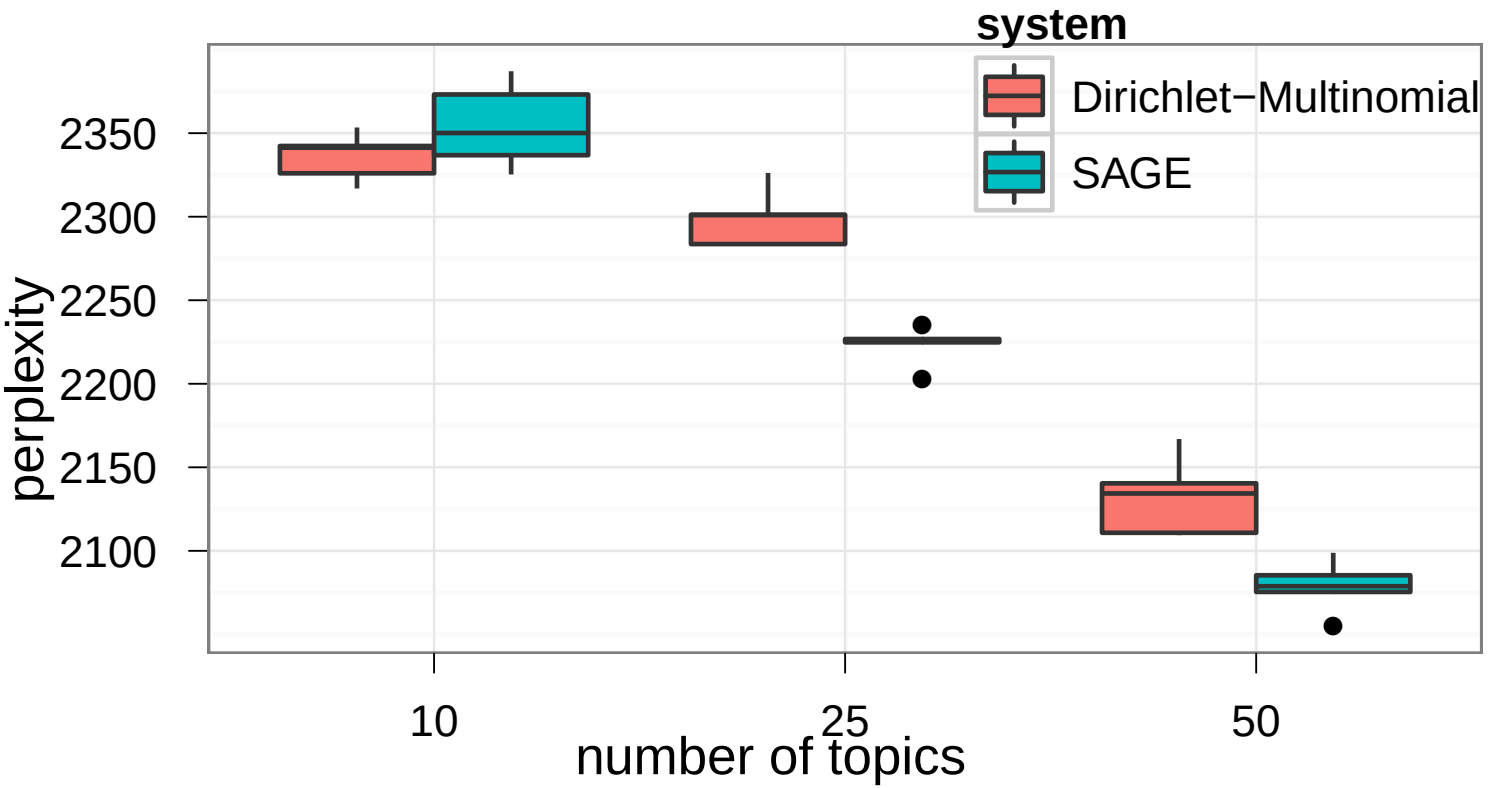
Sparse topic model results

- NIPS dataset: 1986 training docs, 10K vocabulary



Sparse topic model results

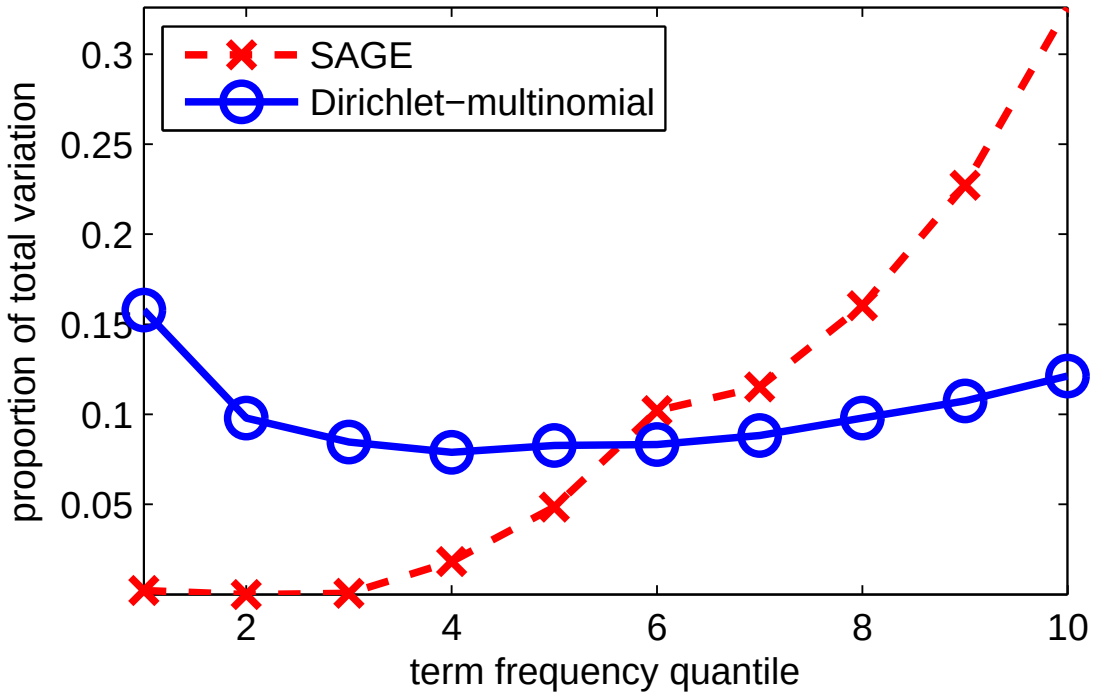
- NIPS dataset: 1986 training docs, 10K vocabulary



- Adaptive sparsity:
  - 5% non-zeros for 10 topics
  - 1% non-zeros for 50 topics

Sparse topic model analysis

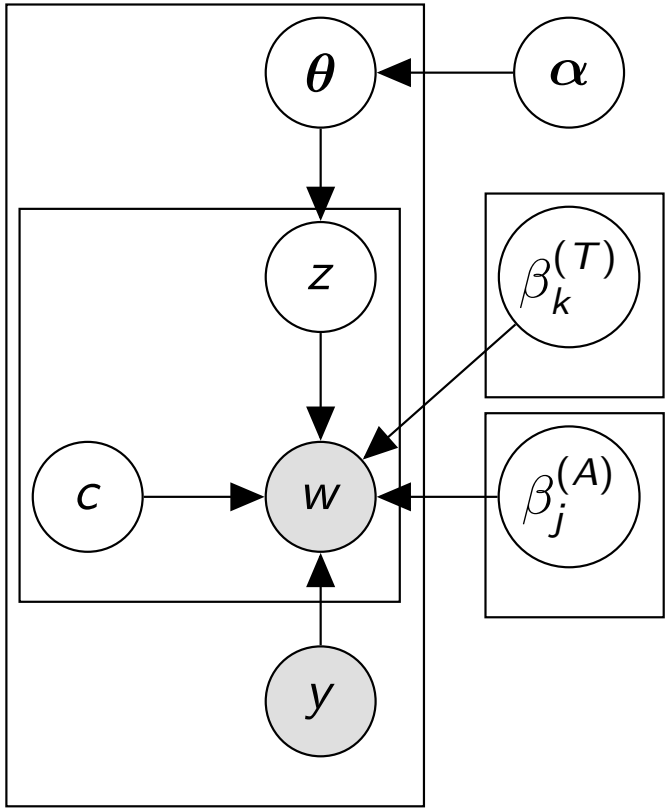
Total variation $= \sum_i |\beta_{k,i} - \overline{\beta}_i|$



Standard topic models assign the greatest amount of variation for the probabilities of the words with the least evidence!

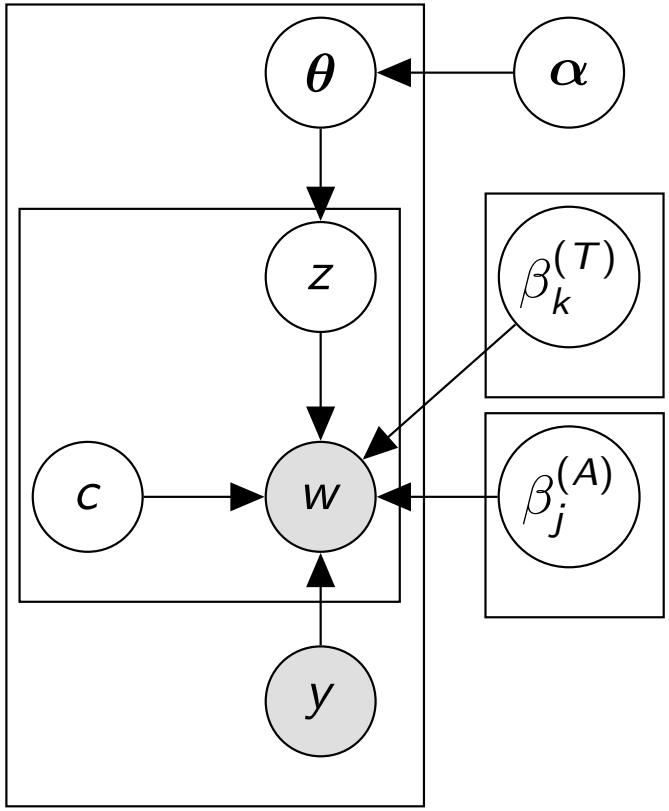
Multifaceted generative models

- Combines latent topics $\beta^{(T)}$ with other facets $\beta^{(A)}$, e.g. ideology, dialect, sentiment



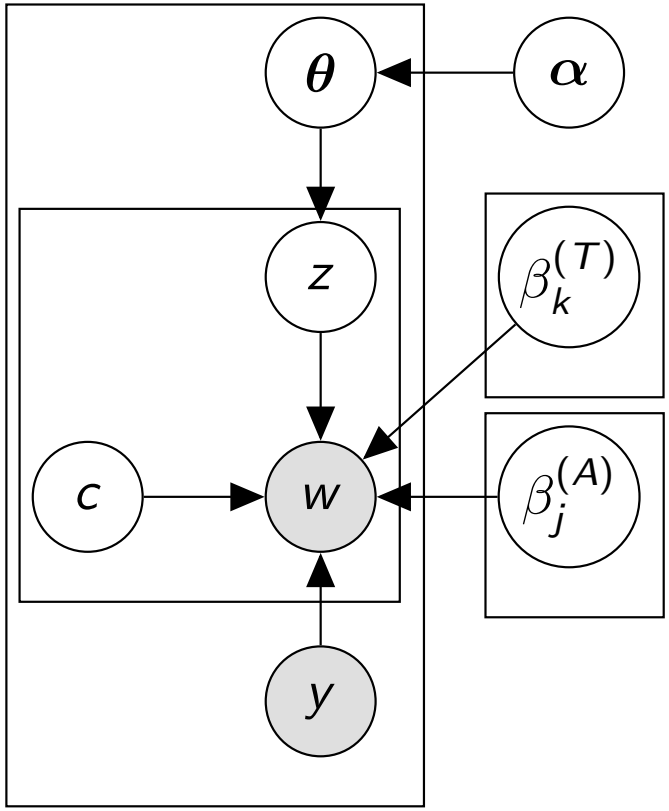
Multifaceted generative models

- Combines latent topics $\beta^{(T)}$ with other facets $\beta^{(A)}$, e.g. ideology, dialect, sentiment
- Typically, a **switching variable** determines which generative facet produces each token (Paul & Girju, 2010; Ahmed & Xing, 2010).



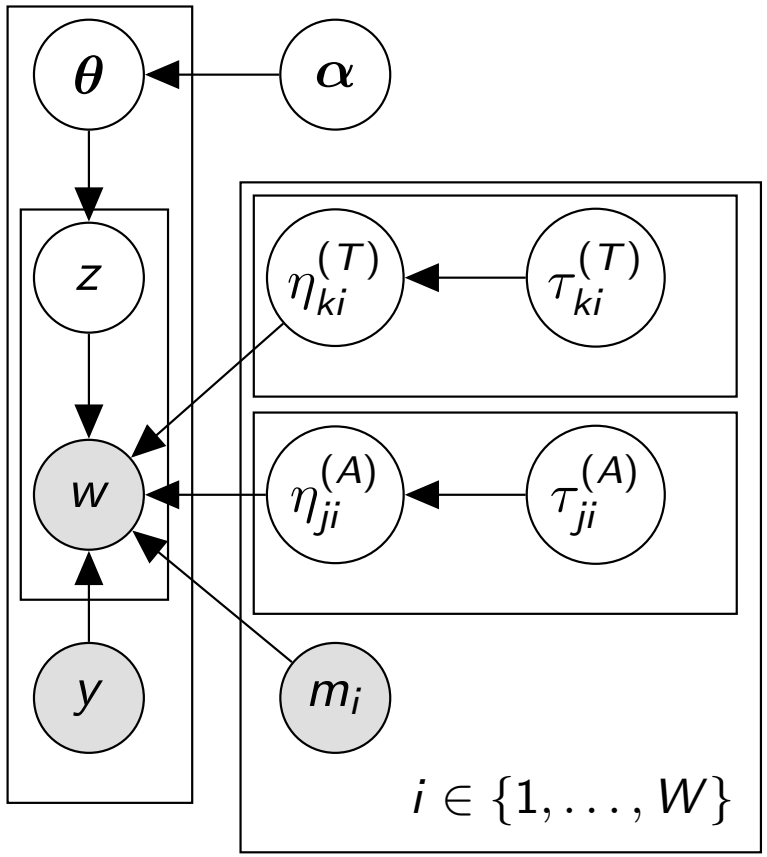
Multifaceted generative models

- Combines latent topics $\beta^{(T)}$ with other facets $\beta^{(A)}$, e.g. ideology, dialect, sentiment
- Typically, a **switching variable** determines which generative facet produces each token (Paul & Girju, 2010; Ahmed & Xing, 2010).
- There is one switching variable per token, complicating inference.



Multifaceted generative models in SAGE

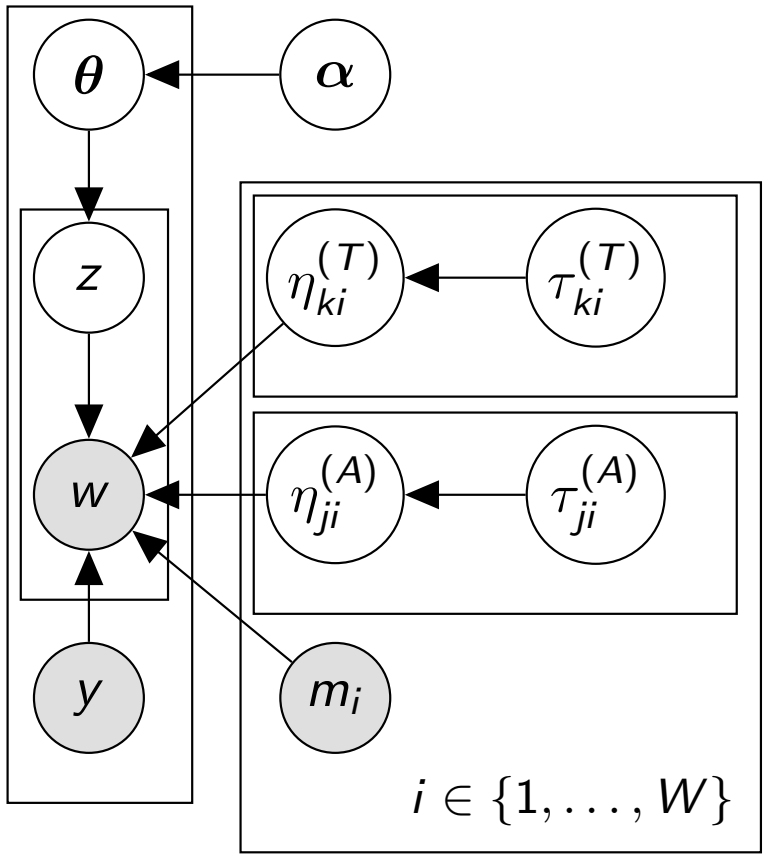
- In SAGE, switching variables are not needed



Multifaceted generative models in SAGE

- In SAGE, switching variables are not needed
- Instead, we just sum all the facets in log-space:

$$P(w|z,y) \propto \exp\left(\boldsymbol{\eta}_z^{(T)} + \boldsymbol{\eta}_y^{(A)} + \mathbf{m}\right)$$



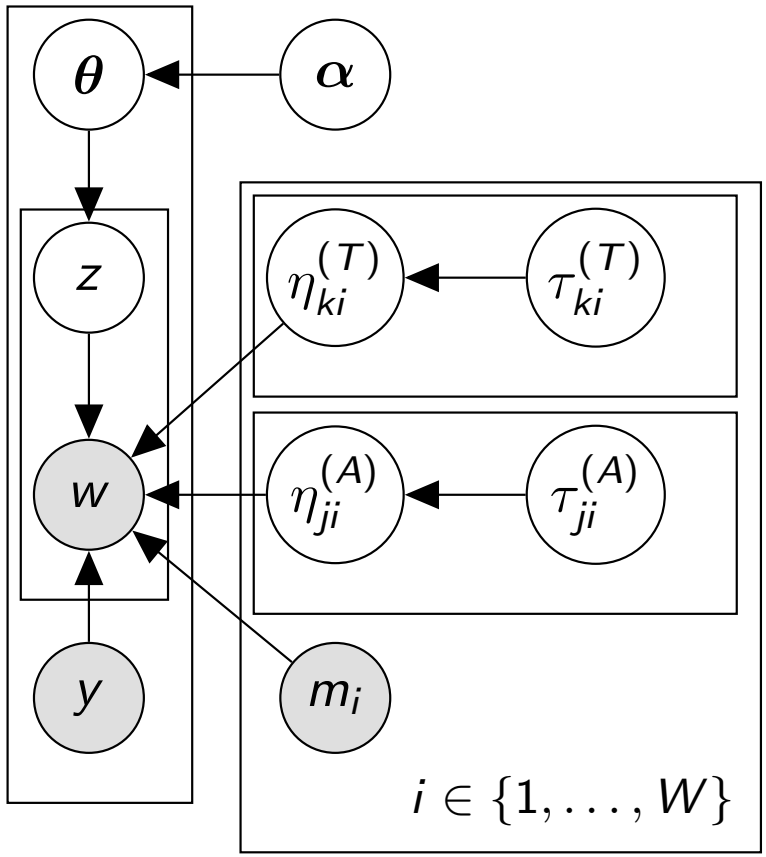
Multifaceted generative models in SAGE

- In SAGE, switching variables are not needed
- Instead, we just sum all the facets in log-space:

$$P(w|z,y) \propto \exp\left(\boldsymbol{\eta}_z^{(T)} + \boldsymbol{\eta}_y^{(A)} + \mathbf{m}\right)$$

- The gradient for $\eta^{(T)}$ is now

$$\frac{\partial \ell}{\partial \boldsymbol{\eta}_k^{(T)}} = \left\langle \mathbf{c}_k^{(T)} \right\rangle - \sum_j \langle C_{jk} \rangle \boldsymbol{\beta}_{jk} - \operatorname{diag}\left(\left\langle \boldsymbol{\tau}_k^{-1} \right\rangle\right) \boldsymbol{\eta}_k,$$

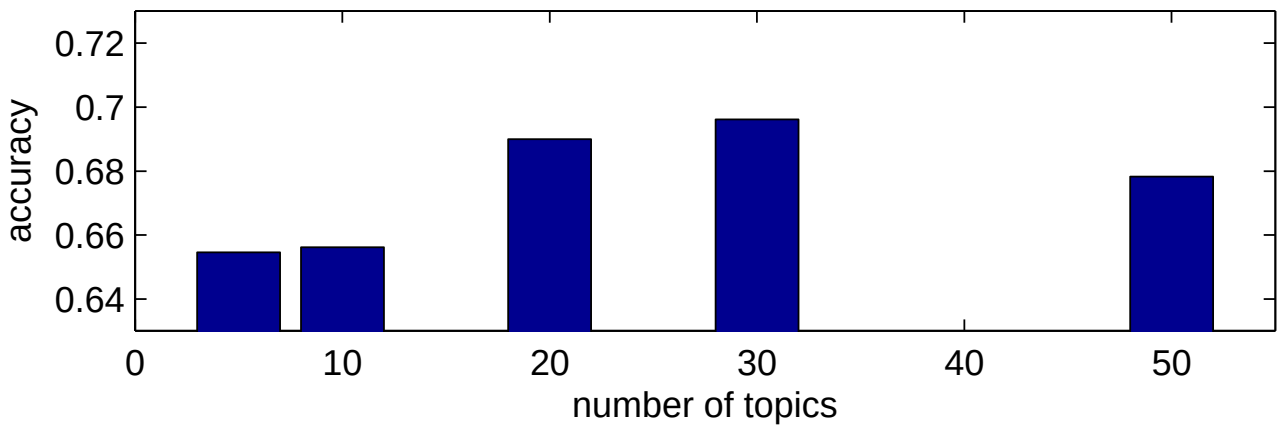


Evaluation: Ideology prediction

- Task: predict blog ideology
- Model: latent topics, observed ideology labels
- Data: six blogs total (two held out), 21K documents, 5.1M tokens

Evaluation: Ideology prediction

- Task: predict blog ideology
- Model: latent topics, observed ideology labels
- Data: six blogs total (two held out), 21K documents, 5.1M tokens



Results match previous best of 69% for Multiview LDA and support vector machine (Ahmed & Xing, 2010).

Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent "region" generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent "region" generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

error in kilometers:	median	mean
Eisenstein et al, 2010 (5K word vocabulary)	494	900
Wing & Baldridge, 2011 (22K word vocabulary)	479	967

Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent "region" generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

error in kilometers:	median	mean
Eisenstein et al, 2010 (5K word vocabulary)	494	900
Wing & Baldridge, 2011 (22K word vocabulary)	479	967
SAGE (5K)	501	845

Evaluation: Geographical Topic Model

- Task: location prediction from Twitter text
- Model: latent "region" generates text and locations
- 9800 weeklong twitter transcripts; 380K messages; 4.9M tokens

error in kilometers:	median	mean
Eisenstein et al, 2010 (5K word vocabulary)	494	900
Wing & Baldridge, 2011 (22K word vocabulary)	479	967
SAGE (5K)	501	845
SAGE (22K)	**461**	**791**

Summary of part 2

- The Dirichlet-multinomial pair is computationally convenient, but does not adequately control model complexity.

Summary of part 2

- The Dirichlet-multinomial pair is computationally convenient, but does not adequately control model complexity.
- The **S**parse **A**dditive **GE**nerative model (SAGE):
  - gracefully handles extraneous parameters,
  - adaptively controls sparsity without a regularization constant,
  - facilitates inference in multifaceted models.

Conclusion

- Language is inherently high-dimensional,
  controlling model complexity is the name of the game
  - robust performance on held-out data
  - interpretable models that yield linguistic insights
- The $L1/L_\infty$ norm performs vocabulary compression jointly with multi-task linear regression on demographics.
- SAGE brings sparsity to generative models,
  estimating sparse deviations from background word probabilities.

Conclusion

- Language is inherently high-dimensional,
  controlling model complexity is the name of the game
  - robust performance on held-out data
  - interpretable models that yield linguistic insights
- The $L1/L_\infty$ norm performs vocabulary compression jointly with multi-task linear regression on demographics.
- SAGE brings sparsity to generative models,
  estimating sparse deviations from background word probabilities.

Thanks!

Example Topics

20 Newsgroups, Vocab=20000, K=25

LDA (perplexity = 1131)

- health insurance smokeless tobacco smoked infections care meat
- wolverine punisher hulk mutants spiderman dy timucin bagged marvel
- gaza gazans glocks glock israeli revolver safeties kratz israel
- homosexuality gay homosexual homosexuals promiscuous optilink male
- god turkish armenian armenians gun atheists armenia genocide firearms

Example Topics

20 Newsgroups, Vocab=20000, K=25

LDA (perplexity = 1131)

- health insurance smokeless tobacco smoked infections care meat
- wolverine punisher hulk mutants spiderman dy timucin bagged marvel
- gaza gazans glocks glock israeli revolver safeties kratz israel
- homosexuality gay homosexual homosexuals promiscuous optilink male
- god turkish armenian armenians gun atheists armenia genocide firearms

SAGE (Perplexity = 1090)

- ftp pub anonymous faq directory uk cypherpunks dcr loren
- disease msg patients candida dyer yeast vitamin infection syndrome
- car cars bike bikes miles tires odometer mavenry altcit
- jews israeli arab arabs israel objective morality baerga amehdi hossien
- god jesus christians bible faith atheism christ atheists christianity