Generative Latent Variable Models of Text

Jacob Eisenstein

Machine Learning Department, CMU

November 16, 2011

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: a probability model over text, $P(w|z)$, where z are labels or latent variables

Classification

Naive Bayes is a generative model for classification:

$$\begin{aligned}\log P(w^{(d)}|z^{(d)},\beta) &= \prod_n P(w_n^{(d)}|\beta, z_n^{(d)}) \\ &= \prod_n \beta_{z_n^{(d)}, w_n^{(d)}}\end{aligned}$$

Classification

Naive Bayes is a generative model for classification:

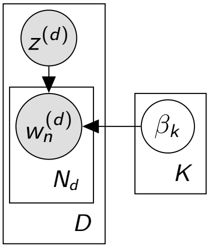
$$\log P(w^{(d)}|z^{(d)}, \beta) = \prod_n P(w_n^{(d)}|\beta, z_n^{(d)})$$
$$= \prod_n \beta_{z_n^{(d)}, w_n^{(d)}}$$

- **training**:

$$\hat{\beta} = \arg\max_{\beta} \prod_d P(w^{(d)}|z^{(d)}, \beta)$$

- **prediction**:

$$\hat{z}^{(d)} = \arg\max_{y} P(w^{(d)}|z, \beta)$$



The Dirichlet-Multinomial pair

- Each β_i is a distribution over words, typically a **multinomial** distribution.

The Dirichlet-Multinomial pair

- Each β_i is a distribution over words, typically a **multinomial** distribution.
- If we want to "be Bayesian," we can place a prior distribution on β. Then we are solving,

$$\hat{\beta} = \arg\max_{\beta} \prod_d P(w^{(d)}|z^{(d)}, \beta)P(\beta)$$

The Dirichlet-Multinomial pair

- Each β_i is a distribution over words, typically a **multinomial** distribution.
- If we want to "be Bayesian," we can place a prior distribution on β. Then we are solving,

$$\hat{\beta} = \arg\max_{\beta} \prod_{d} P(w^{(d)}|z^{(d)}, \beta)P(\beta)$$

- The conjugate prior for the multinomial is the **Dirichlet** distribution.

Conjugacy means we can do collapsed Gibbs sampling, analytically marginalizing the parameter β. This trick gets used **a lot**.

An aside

- Using priors (or not) is a key tenet of some people's world view!

An aside

- Using priors (or not) is a key tenet of some people's world view!
- But there are also practical reasons to use priors.

An aside

- Using priors (or not) is a key tenet of some people's world view!
- But there are also practical reasons to use priors.
  - They perform smoothing, improving performance when data is limited or the number of parameters is very large.

An aside

- Using priors (or not) is a key tenet of some people's world view!
- But there are also practical reasons to use priors.
  - They perform smoothing, improving performance when data is limited or the number of parameters is very large.
  - Priors also make it possible to incorporate domain knowledge.

An aside

- Using priors (or not) is a key tenet of some people's world view!
- But there are also practical reasons to use priors.
  - They perform smoothing, improving performance when data is limited or the number of parameters is very large.
  - Priors also make it possible to incorporate domain knowledge.
- Spoiler: I'll have a lot more to say about whether the Dirichlet-Multinomial pair is the best possible choice for generative models.

Naive Bayes

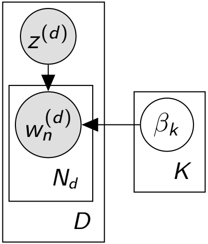
$$\log P(w^{(d)}|z^{(d)},\beta) = \prod_n P(w_n^{(d)}|\beta, z_n^{(d)})$$
$$= \prod_n \beta_{z_n^{(d)}, w_n^{(d)}}$$

- **training**:

$$\hat{\beta} = \arg\max_{\beta} \prod_d P(w^{(d)}|z^{(d)},\beta)$$

- **prediction**:

$$\hat{z}^{(d)} = \arg\max_{y} P(w^{(d)}|z,\beta)$$



Example: Political ideology classification on Twitter

Training data:

Messages containing #p2


LCranston1939 Lamont Cranston
Arrant Facism: Coordinated #OWS evictions make this clear; the military and the police exist to protect the 1% bit.ly/tpsn5m #p2
1 hour ago

rhrealitycheck RH Reality Check
Thanks to @BarbaraBCrane (of @IpasOrg) for donating & helping us stop the right-wing effort to #OccupyYourWomb! ow.ly/7uuej #p2...
2 hours ago

BuddyRoemer Gov. Buddy Roemer
Bloomberg's actions in the midnight hours against Occupy Wall Street protestors were unjust, uncalled for, and unconstitutional. #p2 #ows
3 hours ago Favorite Retweet Reply

peterrothberg Peter Rothberg
Occupy Everywhere on November 17: A Guide to Action. thenation.com/blog/164612/oc… #ows #p2
3 hours ago

mmfa Media Matters
Don't worry, folks: #FoxNews has enough conspiracy theorists to go around for anyone willing to believe them! bit.ly/utOfhM #p2
3 hours ago

Messages containing #tcot


AG_Conservative AG
I miss having a president who loves America and Americans. RT if you agree. #tcot
13 Nov
Retweeted 100+ times

SaintRPh Matt Dawson
Febreeze should film a commercial at Zucotti Park. #ows #tcot
11 minutes ago

PatDollard Patrick Dollard
Breaking: JUDGE RULES #OCCUPYWALLSTREET CAMP CANNOT RETURN TO ZUCCOTTI PARK bit.ly/uXynP7 #tcot
16 minutes ago Favorite Retweet Reply

PatDollard Patrick Dollard
Team Obama Pressured Solyndra To Hide Layoffs Until After Elections bit.ly/vweqaK #tcot
16 minutes ago

FoxieNews Debbie
Dear #Occupy Protesters, You can't have your Sleepovers at Zuccott Park anymore. Go Home and Please Occupy a Shower and a Job STAT! #tcot
18 minutes ago

Example: Political ideology classification on Twitter

Training data:

Messages containing #p2

Messages containing #tcot

- $\beta_{\#\mathrm{p2}}$ emphasizes *protest*, *unconstitutional*, *fascism*
- $\beta_{\#\mathrm{tcot}}$ emphasizes *nobama*, *solyndra*, *socialism*

Naive Bayes for Ideology Prediction

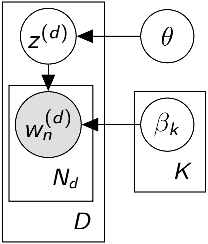
Lin et al (2006) applied Naive Bayes to the "bitter lemons" corpus of text about the Palestinian-Israeli conflict:

Model	Data Set	Accuracy	Reduction
Baseline		0.5	
SVM	Editors	0.9724	
NB-M	Editors	0.9895	61%
NB-B	Editors	0.9909	67%
SVM	Guests	0.8621	
NB-M	Guests	0.8789	12%
NB-B	Guests	0.8859	17%

Palestinian	palestinian, israel, state, politics, peace, international, people, settle, occupation, sharon, right, govern, two, secure, end, conflict, process, side, negotiate
Israeli	israel, palestinian, state, settle, sharon, peace, arafat, arab, politics, two, process, secure, conflict, lead, america, agree, right, gaza, govern

Unsupervised Naive Bayes

When the label z is not observed, it can be imputed.
This is a method for probabilistic clustering:

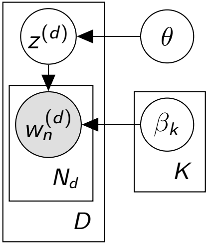


$$P(w|\theta,\beta) = \sum_z P(z|\theta) \prod_n P(w_n|\beta_z)$$

where θ is a prior on z.

Unsupervised Naive Bayes

When the label z is not observed, it can be imputed.
This is a method for probabilistic clustering:



$$P(w|\theta,\beta) = \sum_z P(z|\theta) \prod_n P(w_n|\beta_z)$$

where θ is a prior on z.

Typically we optimize using expectation-maximization:

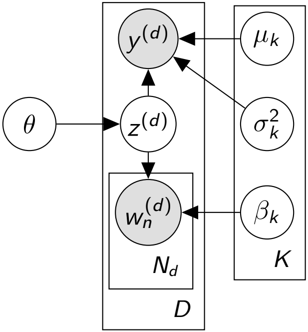
- In the **e-step** we compute the distribution $Q(z)$
- In the **m-step** we update the parameter β

Latent Variable Models

- Imagine we have additional data $y^{(d)}$: for each author on Twitter,
  - $y^{(d)}$ is their geographical location,
  - $w^{(d)}$ is the set of all words in all their tweets,
  - $z^{(d)}$ is a latent variable which must explain both $y^{(d)}$ and $w^{(d)}$.
- We want to learn to predict y from w. (Eisenstein, O'Connor, Smith, and Xing. EMNLP 2010)

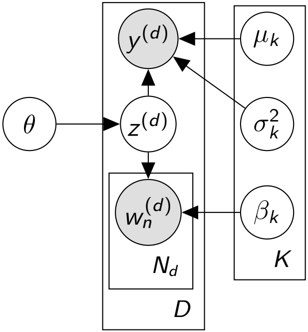
Latent Variable Models

- Imagine we have additional data $y^{(d)}$: for each author on Twitter,
  - $y^{(d)}$ is their geographical location,
  - $w^{(d)}$ is the set of all words in all their tweets,
  - $z^{(d)}$ is a latent variable which must explain both $y^{(d)}$ and $w^{(d)}$.
- We want to learn to predict y from w. (Eisenstein, O'Connor, Smith, and Xing. EMNLP 2010)



Latent Variable Models

- Imagine we have additional data $y^{(d)}$: for each author on Twitter,
  - $y^{(d)}$ is their geographical location,
  - $w^{(d)}$ is the set of all words in all their tweets,
  - $z^{(d)}$ is a latent variable which must explain both $y^{(d)}$ and $w^{(d)}$.
- We want to learn to predict y from w. (Eisenstein, O'Connor, Smith, and Xing. EMNLP 2010)

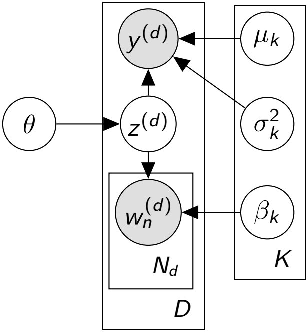


In training, we maximize:

$$P(y, w|\theta, \beta, \mu, \sigma^2) = \sum_z P(z|\theta) P(y|\mu_z, \sigma_z^2) \prod_n P(w_n|\beta_z)$$

Latent Variable Models

- **training**: Expectation-maximization, alternating between updates to $Q(z)$ and the parameters $\{\beta, \theta, \mu, \sigma^2\}$



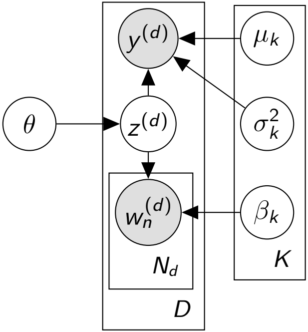
Latent Variable Models

- **training**: Expectation-maximization, alternating between updates to $Q(z)$ and the parameters $\{\beta, \theta, \mu, \sigma^2\}$
- **prediction**:

$$\hat{y} = \arg\max_y P(y|w)$$

$$P(y|w) = \sum_z P(y|\mu_z, \sigma_z^2) P(z|w,\theta)$$

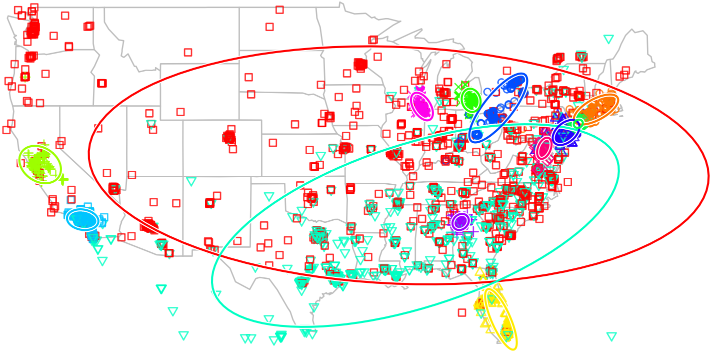
$$P(z|w,\theta) = P(w|\beta_z) P(z|\theta) / P(w)$$



Quantitative Results

error in km:	mean	median
mean location	1148	1018
text regression	948	712
mixture model	947	644

Qualitative Results



Each author in our dataset is a point;
cluster membership is indicated by color and shape.[1]

[1]Figure by Brendan O'Connor

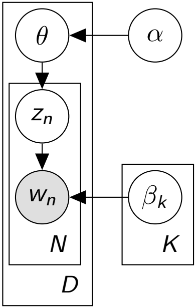
Qualitative results

For each cluster, we rank words by log-odds: $\log \beta_i - \log \frac{1}{K} \sum_j \beta_j$:

- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Pac. Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

Discovering latent themes

Topic models like latent Dirichlet allocation discover latent **themes** or **topics** in document collections:



- Each β_k is a topic, a distribution over words.
- Each θ_d represents the topic proportions for document d.
- Each z_n is the latent topic which generates the word w_n.

$$P(w|\theta, \beta) = \prod_n P(z_n|\theta) P(w_n|\beta_{z_n})$$

Topics in Twitter

"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp

Key point is that individual authors are **admixtures** of these topics, e.g., my Twitter feed is 60% chit-chat, 30% basketball, 10% emoticons.

Combining topics and labels

Recall the Twitter political ideology problem:

Messages containing #p2

Messages containing #tcot

Adding topics

- Authors don't just express ideological viewpoints, they discuss topics: health care, taxes, regulation, ...

Adding topics

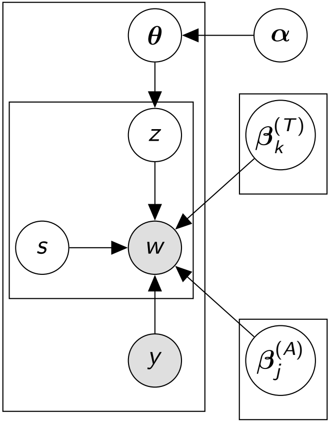
- Authors don't just express ideological viewpoints, they discuss topics: health care, taxes, regulation, ...
- In **prediction**, these topical differences make learning harder. Left-wing and right-wing perspectives on a single topic may share more words than a single perspective on multiple topics.

Adding topics

- Authors don't just express ideological viewpoints, they discuss topics: health care, taxes, regulation, ...
- In **prediction**, these topical differences make learning harder. Left-wing and right-wing perspectives on a single topic may share more words than a single perspective on multiple topics.
- In **analysis**, we often want to understand topic-specific differences: e.g., how do the left-wing and right-wing perspectives differ with respect to foreign policy

Switching models

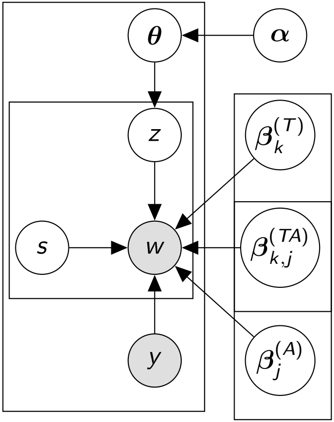
We can combine topics and labels by adding a "switch" for each word, which determines if the word is generated from a topic or the label:



- Each s_n determines whether w_n is generated by the topic z_n or the label y.
- Each $\beta_k^{(T)}$ is a word distribution associated a latent topic.
- Each $\beta_j^{(A)}$ is a word distribution associated with a label.

Switching models

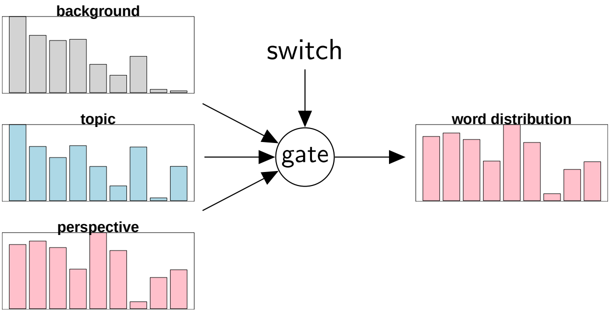
We can combine topics and labels by adding a “switch” for each word, which determines if the word is generated from a topic or the label:



- Each s_n determines whether w_n is generated by the topic z_n or the label y.
- Each $\beta_k^{(T)}$ is a word distribution associated a latent topic.
- Each $\beta_j^{(A)}$ is a word distribution associated with a label.
- Each $\beta_{k,j}^{(TA)}$ is a word distribution associated with a topic-label **interaction**.

Switching models: a schematic

A topic-perspective-background model:



Example output: topics and cultures

Topic 1		
fashion style look dress wear new collection accessories black		
UK	**India**	**Singapore**
shoes	fashion	price
fashion	women	posted
clothing	indian	earrings
high	designer	length
designer	sarees	item
style	leather	sgd
love	girls	silver
london	china	clothes
shirts	jewellery	shop
bag	jewelry	code

Topic 2		
food add chicken recipe cooking taste rice recipes sugar soup		
UK	**India**	**Singapore**
food	recipe	coffee
wine	recipes	cup
restaurant	powder	oil
coffee	indian	comments
cheese	salt	fried
soup	tsp	add
eat	rice	restaurant
chef	masala	rice
english	oil	tea
drink	coriander	seafood

From ccLDA (Paul and Girju, 2009)

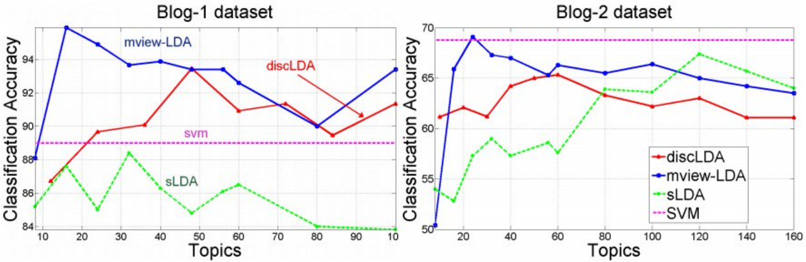
Example output: topics and perspectives

palestinian israeli israel military civilians attacks	
Aspect A	**Aspect B**
war	violence
public	palestinians
government	occupation
media	resistance
society	intifada
terrorist	violent
soldiers	non
incitement	force

state israel solution palestine palestinian states borders	
Israeli	**Palestinian**
jewish	palestinians
arab	return
israeli	right
jews	refugees
population	problem
jordan	refugee
west	rights
south	resolution

From TAM (Paul and Girju, 2010);
added unsupervised and semi-supervised learning to ccLDA.

Results: ideology prediction



From Multiview-LDA (Ahmed and Xing, 2010)

Results: geography prediction

error in km:	mean	median
mean location	1148	1018
text regression	948	712
mixture model	947	644
mixture model + topics	900	494

Overview

Capabilities of generative models:

- Classification and clustering (Naive Bayes)
- Discovering latent topics (LDA)
- Combining topics and labels (ccLDA, TAM, Multiview-LDA)

Overview

Capabilities of generative models:

- Classification and clustering (Naive Bayes)
- Discovering latent topics (LDA)
- Combining topics and labels (ccLDA, TAM, Multiview-LDA)

We have focused on text, but there are many, many applications of these models to vision and computational biology.

Taking stock

Generative models models have many advantages:

- Interpretability
- Can combine multiple modalities
- Relatively simple semi-supervised extensions
- Easy to incorporate domain-specific insights in model design

Taking stock

Generative models models have many advantages:

- Interpretability
- Can combine multiple modalities
- Relatively simple semi-supervised extensions
- Easy to incorporate domain-specific insights in model design

But they also have problems! (Eisenstein et al., ICML 2011)

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

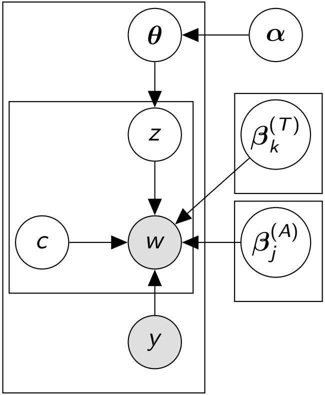
- An LDA **model** with K topics and V words requires $K \times V$ parameters.
- An LDA **paper** shows 10 words per topic.
- What about the other $V - 10$ words per topic??

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, $P(w|y) = \boldsymbol{\beta}_y$

Sparse Additive Generative Models

- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \boldsymbol{\beta}_y$
- **Sparse Additive Generative models (SAGE)**:
  each class or latent theme is represented by its deviation from a background distribution.

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

Sparse Additive Generative Models

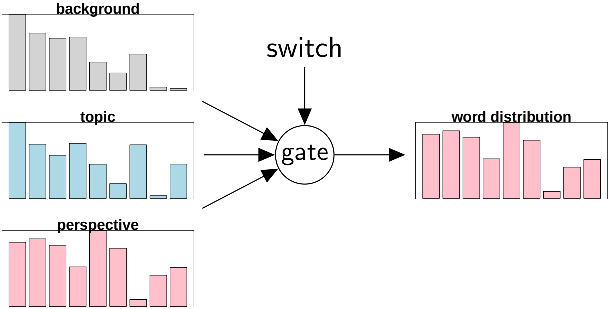
- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \boldsymbol{\beta}_y$
- **Sparse Additive Generative models (SAGE)**:
  each class or latent theme is represented by its deviation from a background distribution.

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

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

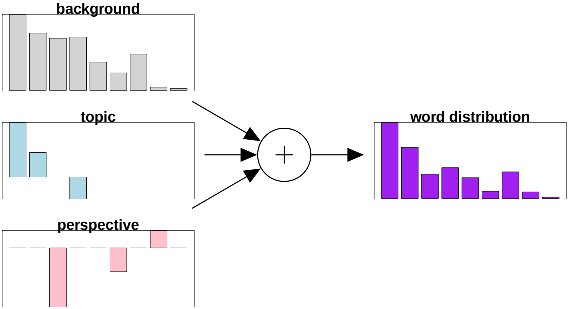
Sparse Additive Generative Models

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



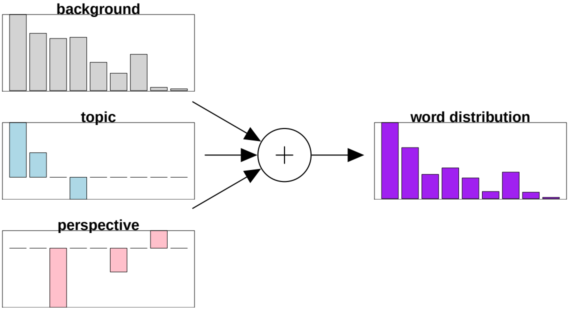
Sparse Additive Generative Models

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

Sparsity through integration

- The $L1$ regularizer is equivalent to a Laplace prior distribution: $\eta \sim \mathcal{L}(0, \sigma)$
  - The Laplace distribution is equal to 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 $L1$ regularizer is equivalent to a Laplace prior distribution: $\eta \sim \mathcal{L}(0, \sigma)$
  - The Laplace distribution is equal to the integral: $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\mathrm{Exp}(\tau; \sigma) d\tau$ (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 $L1$ regularizer is equivalent to a Laplace prior distribution: $\eta \sim \mathcal{L}(0, \sigma)$
  - The Laplace distribution is equal to the integral: $\mathcal{L}(\eta; 0, \sigma) = \int \mathcal{N}(\eta; 0, \tau)\mathrm{Exp}(\tau; \sigma) d\tau$ (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 (EM), updating:

Sparsity through integration

- The $L1$ regularizer is equivalent to a Laplace prior distribution: $\eta \sim \mathcal{L}(0, \sigma)$
  - The Laplace distribution is equal to 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 (EM), updating:
  - The distribution $Q(\boldsymbol{\tau})$

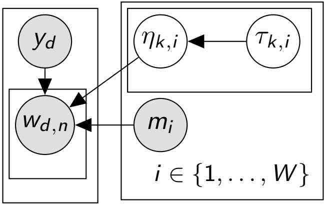
Sparsity through integration

- The $L1$ regularizer is equivalent to a Laplace prior distribution: $\eta \sim \mathcal{L}(0, \sigma)$
  - The Laplace distribution is equal to 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 (EM), updating:
  - The 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 $\beta_i = \frac{\exp(\eta_{y_d,i} + m_i)}{\sum_j \exp(\eta_{y_d,j} + m_j)}$
- 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

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

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,

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

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

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

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 $Q(\tau_{k,i}) = \mathrm{Gamma}(\tau_{k,i}; a_{k,i}, b_{k,i})$

Inference

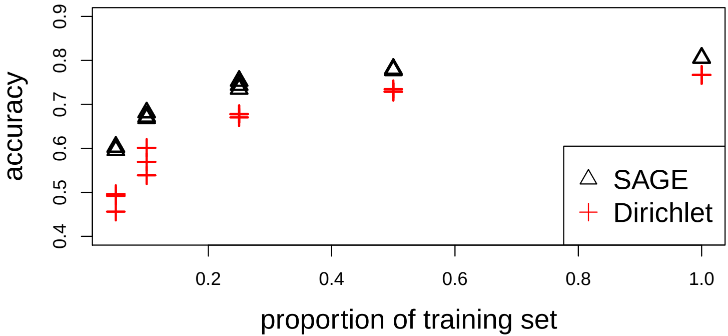
- We maximize the variational bound

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

- 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

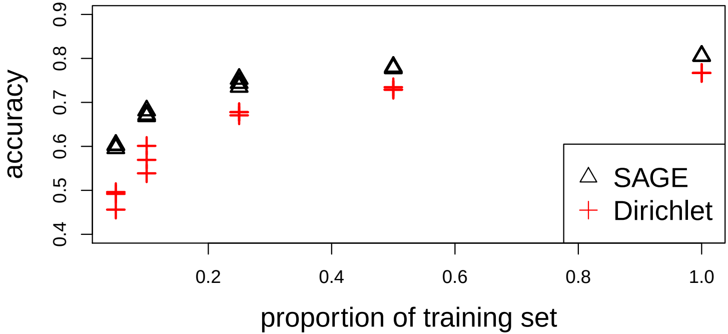
Document classification evaluation

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



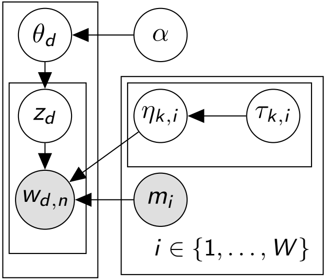
Document classification evaluation

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

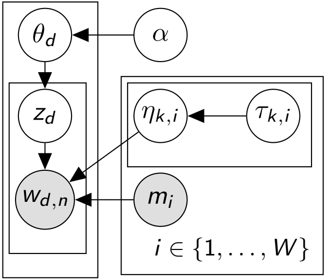


- Adaptive sparsity:
  - 10% non-zeros for full training set (11K docs)
  - 2% non-zeros for minimal training set (550 docs)

SAGE in latent variable models



SAGE in latent variable models



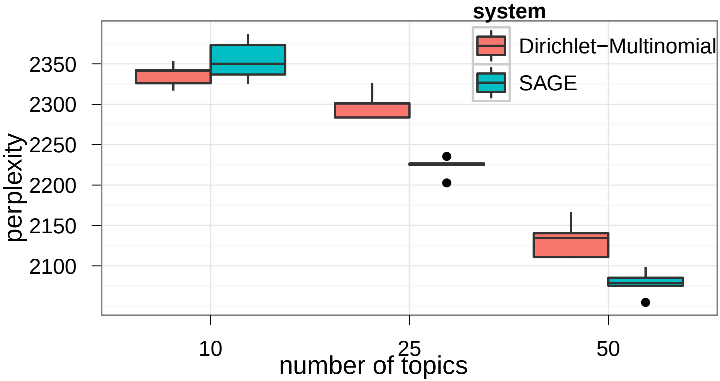
The gradient for $\boldsymbol{\eta}$ now includes **expected** counts:

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

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

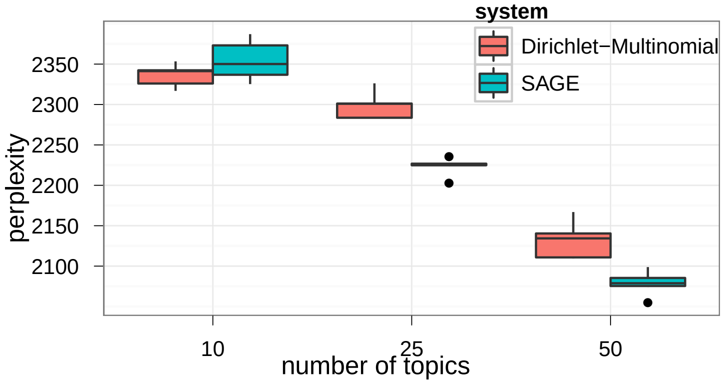
Sparse topic model results

- NIPS dataset: 1986 training docs, 10K vocabulary



Sparse topic model results

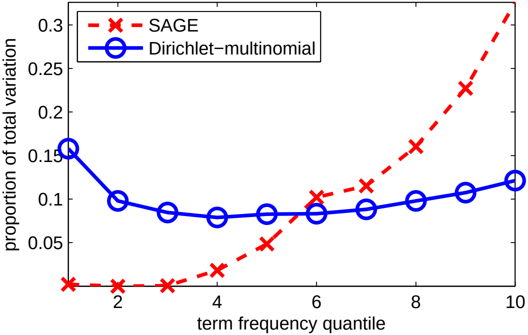
- NIPS dataset: 1986 training docs, 10K vocabulary



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

Sparse topic model analysis

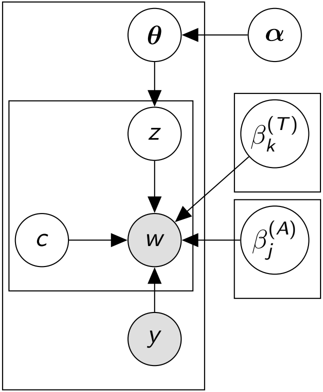
Total variation $= \sum_i |\beta_{k,i} - \overline{\beta}_i|$



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

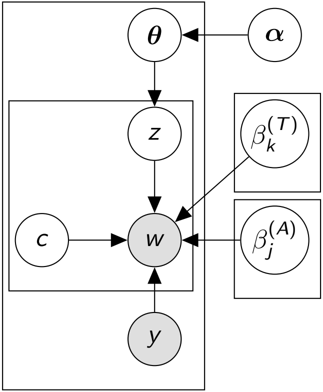
Multifaceted generative models

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



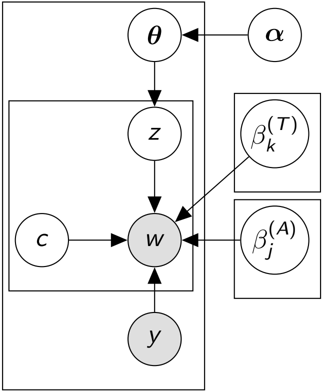
Multifaceted generative models

- 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).



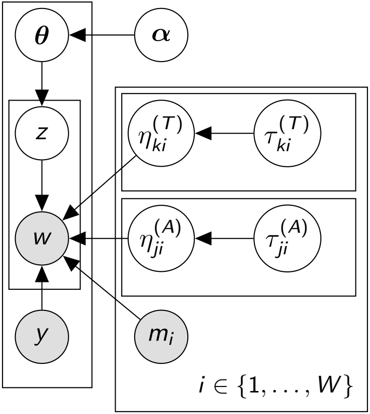
Multifaceted generative models

- 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

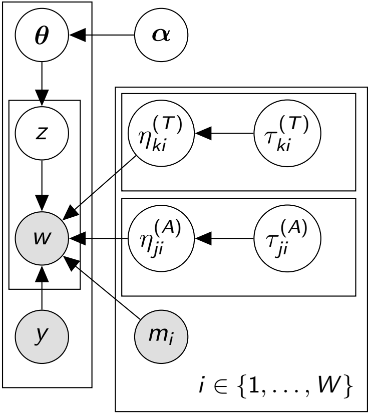
- In SAGE, switching variables are not needed



Multifaceted generative models in SAGE

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

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



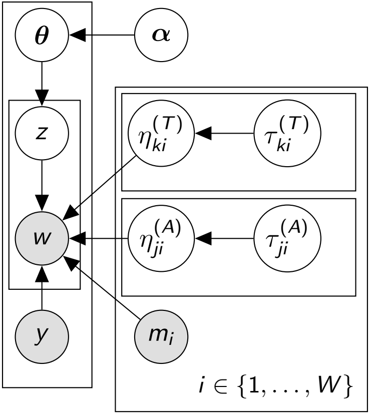
Multifaceted generative models in SAGE

- 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

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

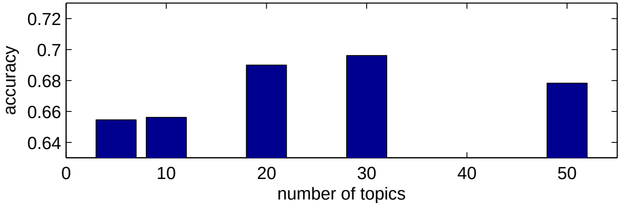


Evaluation: Ideology prediction

- 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 km:	mean	median
mean location	1148	1018
text regression	948	712
mixture model	947	644
mixture model + topics	900	494

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 km:	mean	median
mean location	1148	1018
text regression	948	712
mixture model	947	644
mixture model + topics	900	494
SAGE (5K vocab)	845	501

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 km:	mean	median
mean location	1148	1018
text regression	948	712
mixture model	947	644
mixture model + topics	900	494
SAGE (5K vocab)	845	501
SAGE (22K vocab)	**791**	**461**

Summary of SAGE

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

Summary of SAGE

- 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

- Generative models provide powerful tools for understanding natural language data.
- Capabilities include prediction, clustering, and discovering latent topics, as well as more exotic models that combine latent and observed aspects.
- As always, controlling model complexity is critical.
  - SAGE improves on the Dirichlet-Multinomial pair by modeling sparse deviations in log-odds.

Conclusion

- Generative models provide powerful tools for understanding natural language data.
- Capabilities include prediction, clustering, and discovering latent topics, as well as more exotic models that combine latent and observed aspects.
- As always, controlling model complexity is critical.
    - SAGE improves on the Dirichlet-Multinomial pair by modeling sparse deviations in log-odds.

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