Sparse Additive Generative Models of Text

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing

Carnegie Mellon University

June 29, 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: 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

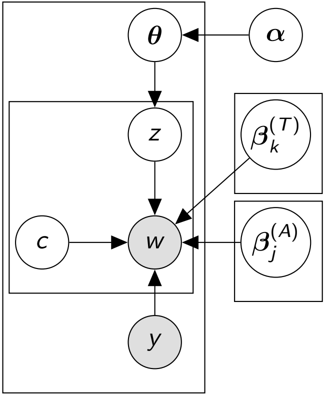
- An LDA **model** with K topics and V words requires $K \times V$ parameters.
- An LDA **paper** shows 10 words per topic.
- What about the other $V - 10$ words per topic??

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

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**:
  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

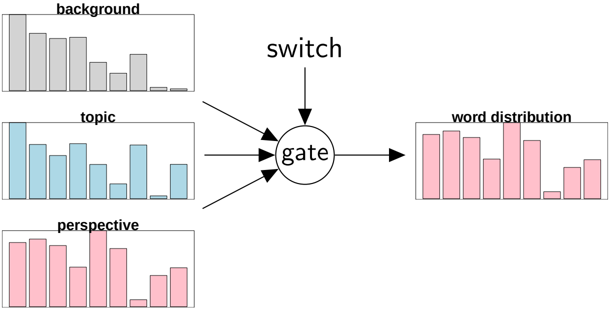
- **Multinomial generative models**: each class or latent theme is represented by a distribution over tokens, $P(w|y) = \boldsymbol{\beta}_y$
- **Sparse Additive GEnerative models**:
  each class or latent theme is represented by its deviation from a background distribution.

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

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

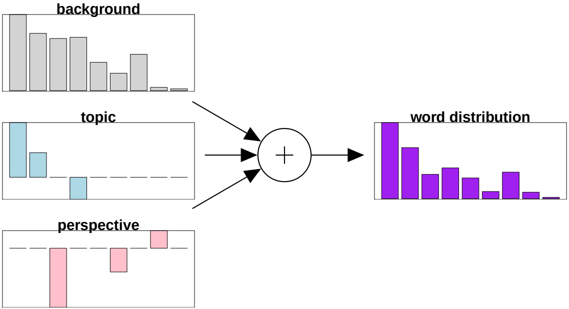
Sparse Additive Generative Models

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



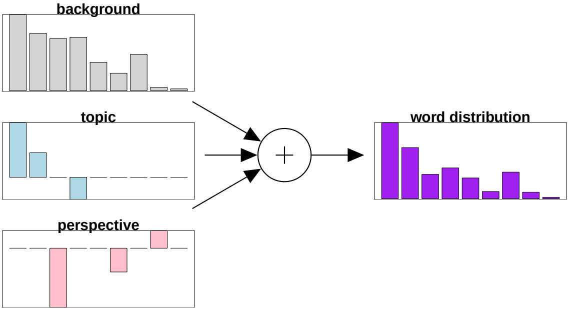
Sparse Additive Generative Models

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 induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$

Sparsity through integration

- The Laplace distribution induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$
  - 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 induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$
  - 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 induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$
  - 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 induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$
    - We can apply 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, updating:
    - The variational distribution $Q(\boldsymbol{\tau})$

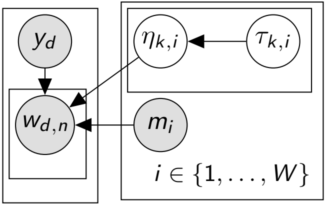
Sparsity through integration

- The Laplace distribution induces sparsity: $\eta \sim \mathcal{L}(0, \sigma)$
    - 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 $\beta_i = \frac{\exp(\eta_{y_d,i} + m_i)}{\sum_j \exp(\eta_{y_d,j} + m_j)}$
- 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,

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

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

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

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

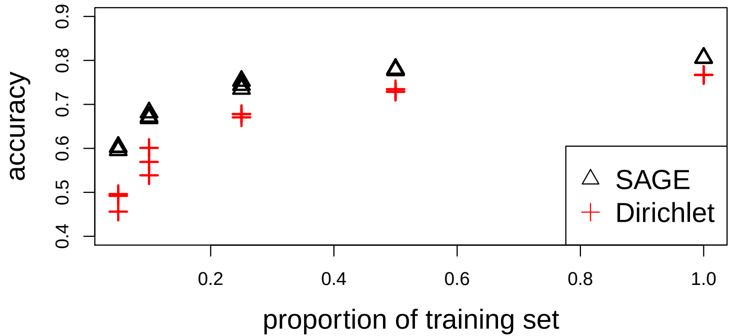
- We maximize the variational bound

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

- 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

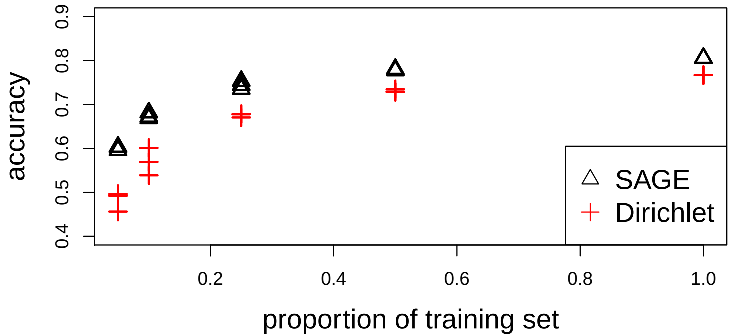
Document classification evaluation

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



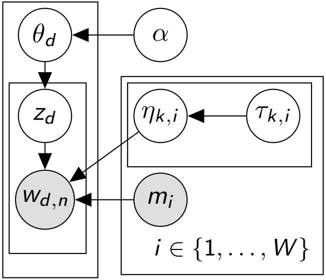
Document classification evaluation

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

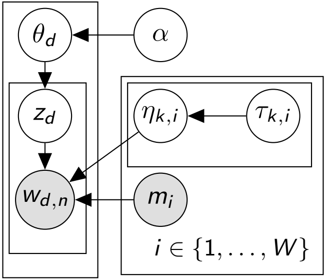


- Adaptive sparsity:
  - 10% non-zeros for full training set (11K docs)
  - 2% non-zeros for minimal training set (550 docs)

SAGE in latent variable models



SAGE in latent variable models



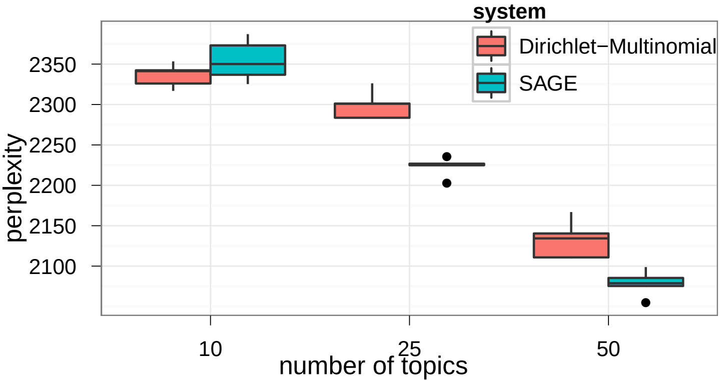
The gradient for $\boldsymbol{\eta}$ now includes **expected** counts:

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

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

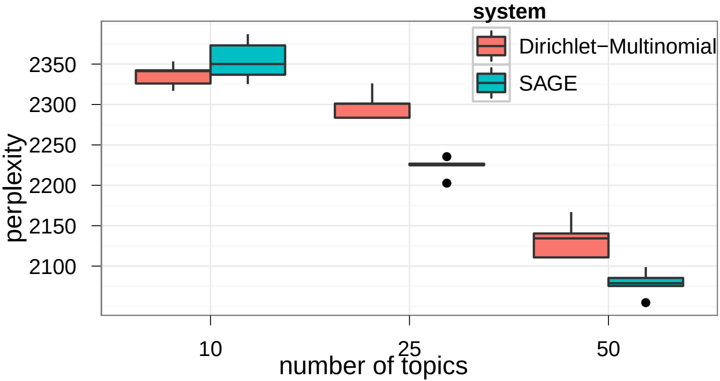
Sparse topic model results

- NIPS dataset: 1986 training docs, 10K vocabulary



Sparse topic model results

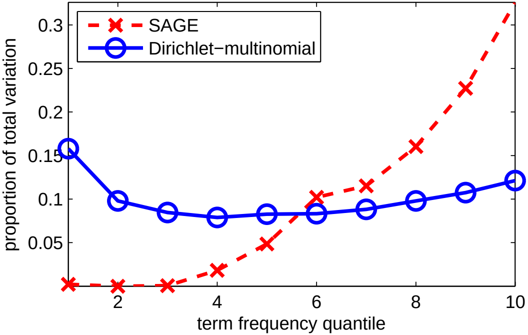
- NIPS dataset: 1986 training docs, 10K vocabulary



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

Sparse topic model analysis

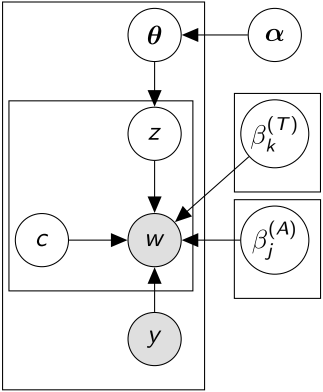
Total variation $= \sum_i |\beta_{k,i} - \overline{\beta}_i|$



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

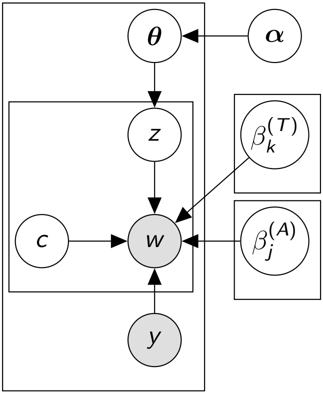
Multifaceted generative models

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



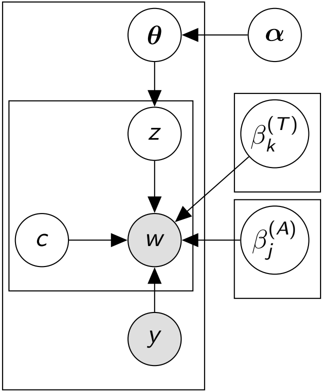
Multifaceted generative models

- 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).



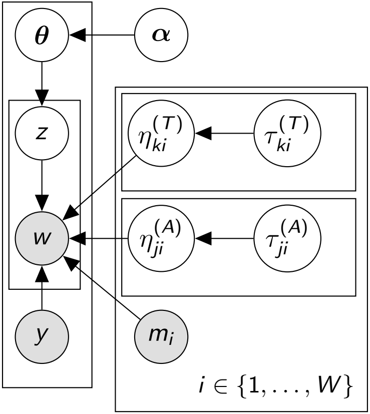
Multifaceted generative models

- 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

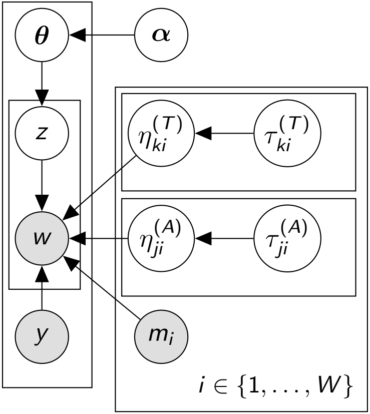
- In SAGE, switching variables are not needed



Multifaceted generative models in SAGE

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

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



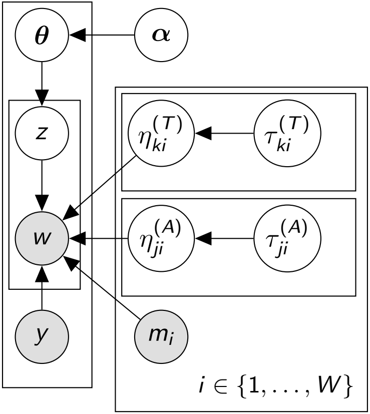
Multifaceted generative models in SAGE

- 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

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

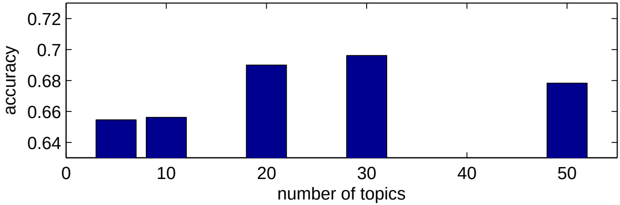


Evaluation: Ideology prediction

- 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 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

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

Summary

- 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.

Summary

- 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.

- 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