The Structural Topic Model and Applied Social Science

Molly Roberts, Brandon Stewart, Dustin Tingley, Edoardo Airoldi

Harvard University, Departments of Government and Statistics

December 10, 2013

Related Work

Related Work

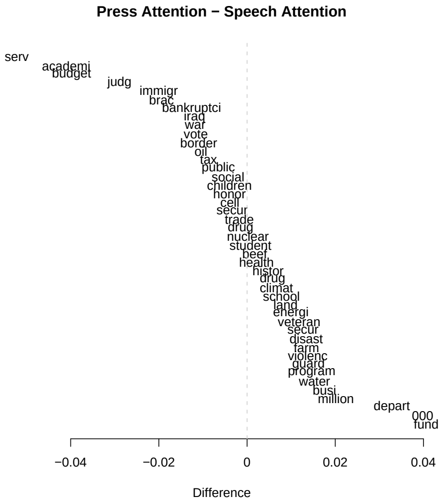
Roberts ME, Stewart BM, Airoldi EM. A Topic Model for Experimentation in the Social Sciences.

Related Work

Roberts ME, Stewart BM, Airoldi EM. A Topic Model for Experimentation in the Social Sciences. Roberts ME, Stewart BM, Tingley D, Lucas C,

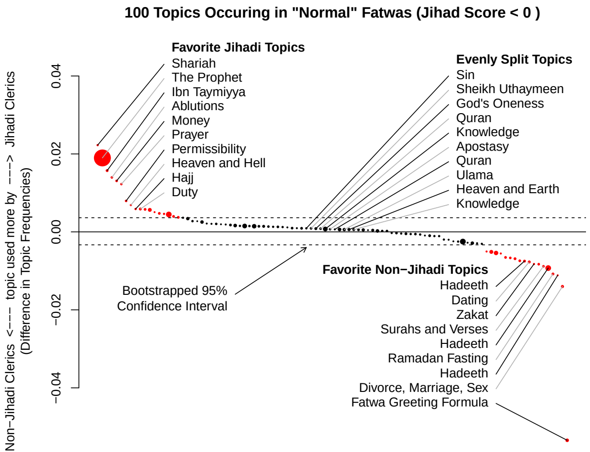
Leder-Luis J, Gadarian S, Albertson B, Rand D. Structural topic models for open-ended survey responses. Forthcoming at *American Journal of Political Science*.

How Do Senators Relate to Constituents?



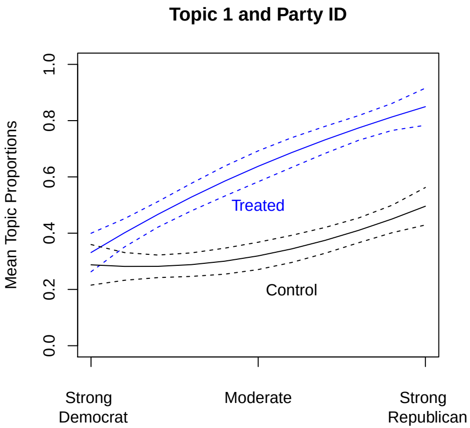
Grimmer (2010, 2013)

Why do some Muslim clerics support violent Jihad?



Nielsen (2013)

How do we analyze open-ended survey response?



Social Sciences Applications

These problems share a common structure:

Social Sciences Applications

These problems share a common structure:

- Topic models as a tool of *measurement*

Social Sciences Applications

These problems share a common structure:

- Topic models as a tool of *measurement*
  - events between countries (O'Connor et al 2013)
  - "constitutional moments" (Stewart and Young 2013)
  - media control in China (Stewart and Roberts 2014)
- Extensive "metadata" in documents

Social Sciences Applications

These problems share a common structure:

- Topic models as a tool of *measurement*
  - events between countries (O'Connor et al 2013)
  - "constitutional moments" (Stewart and Young 2013)
  - media control in China (Stewart and Roberts 2014)
- Extensive "metadata" in documents
- Topical Prevalence and Topical Content

Social Sciences Applications

These problems share a common structure:

- Topic models as a tool of *measurement*
  - events between countries (O'Connor et al 2013)
  - "constitutional moments" (Stewart and Young 2013)
  - media control in China (Stewart and Roberts 2014)
- Extensive "metadata" in documents
- Topical Prevalence and Topical Content

Primary QOI is how external variable drives topics.

In Practice

In Practice

- ‘Vanilla” LDA with post-hoc comparison

In Practice

- 'Vanilla" LDA with post-hoc comparison
- The exchangeability paradox.

In Practice

- ‘Vanilla” LDA with post-hoc comparison
- The exchangeability paradox.
- Custom Models vs. Off the Shelf

Our Approach

General framework for including covariates

Our Approach

General framework for including covariates

- General framework for including covariates

Our Approach

General framework for including covariates

- General framework for including covariates
- Two types of covariates:

Our Approach

General framework for including covariates

- General framework for including covariates
- Two types of covariates:
  - Topical Prevalence: Logistic Normal GLM

Our Approach

General framework for including covariates

- General framework for including covariates
- Two types of covariates:
  - Topical Prevalence: Logistic Normal GLM
  - Topical Content: Multinomial Logit on Words

Our Approach

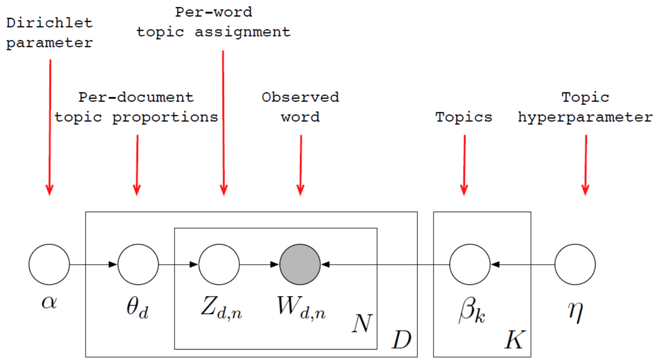
General framework for including covariates

- General framework for including covariates
- Two types of covariates:
  - Topical Prevalence: Logistic Normal GLM
  - Topical Content: Multinomial Logit on Words

Builds off: DMR (Mimno and McCallum 2008), SAGE (Eisenstein et al 2011) and the CTM (Blei and Lafferty 2007)

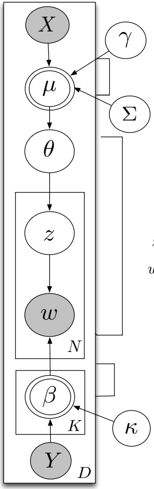
Latent Dirichlet Allocation

Figure: Plate Notation of Latent Dirichlet Allocation



Graphic from David Blei's Website: http://www.cs.princeton.edu/ blei/modeling-science.pdf

Structural Topic Model



Topic Prevalence:

$$\mu_{d,k} = X_d \gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

Language Model:

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

A Tale of Two Covariates

- Prevalence

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$
  - Documents which have similar covariates will tend to talk about the same topics.

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$
  - Documents which have similar covariates will tend to talk about the same topics.
- Content

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$
  - Documents which have similar covariates will tend to talk about the same topics.
- Content
  - Distribution over words is now document-specific

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$
  - Documents which have similar covariates will tend to talk about the same topics.
- Content
  - Distribution over words is now document-specific
  - Topics are sparse deviations from a word-specific baseline $\beta_{k,g} \propto exp(m + \kappa^{(k)} + \kappa^{(g)} + \kappa^{(k,g)})$

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$
  - Documents which have similar covariates will tend to talk about the same topics.
- Content
  - Distribution over words is now document-specific
  - Topics are sparse deviations from a word-specific baseline $\beta_{k,g} \propto exp(m + \kappa^{(k)} + \kappa^{(g)} + \kappa^{(k,g)})$
  - Documents which have similar covariates will tend to talk about topics in the same way.

A Tale of Two Covariates

- Prevalence
  - Prior on the mixture over topics is now document-specific
  - $\eta \sim \mathcal{N}(X\gamma, \Sigma)$
  - Documents which have similar covariates will tend to talk about the same topics.
- Content
  - Distribution over words is now document-specific
  - Topics are sparse deviations from a word-specific baseline $\beta_{k,g} \propto exp(m + \kappa^{(k)} + \kappa^{(g)} + \kappa^{(k,g)})$
  - Documents which have similar covariates will tend to talk about topics in the same way.
- Regularizing priors to avoid false positives

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM
- Propagating estimation uncertainty (method of composition)

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM
- Propagating estimation uncertainty (method of composition)
- Forthcoming `R` package

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM
- Propagating estimation uncertainty (method of composition)
- Forthcoming `R` package
  - Various meta-data topic models

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM
- Propagating estimation uncertainty (method of composition)
- Forthcoming `R` package
  - Various meta-data topic models
  - Post-estimation tools (labeling, evaluation statistics, plotting)

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM
- Propagating estimation uncertainty (method of composition)
- Forthcoming `R` package
    - Various meta-data topic models
    - Post-estimation tools (labeling, evaluation statistics, plotting)
    - Automated model selection

Inference and Implementation

- Semi-collapsed, non-conjugate, mean-field variational EM
- Propagating estimation uncertainty (method of composition)
- Forthcoming `R` package
  - Various meta-data topic models
  - Post-estimation tools (labeling, evaluation statistics, plotting)
  - Automated model selection
  - Covariate uncertainty calculation

Applications

In This Paper:

- Open-Ended Survey Response (1 of 3)
- Media Coverage of China (short example from longer paper)

Open-Ended Response

Researchers opt for **closed ended** responses.

Open-Ended Response

Researchers opt for **closed ended** responses.This requires,

- Choosing an arbitrary scale
- Choosing **researcher defined** *categories*. Sometimes putting an "other" open ended option.

Open-Ended Response

Researchers opt for **closed ended** responses.This requires,

- Choosing an arbitrary scale
- Choosing **researcher defined** *categories*. Sometimes putting an "other" open ended option.

A debate exists on whether this is a good idea.

Open-Ended Response

Researchers opt for **closed ended** responses.This requires,

- Choosing an arbitrary scale
- Choosing **researcher defined** *categories*. Sometimes putting an "other" open ended option.

A debate exists on whether this is a good idea.

There are workflow advantages to closed ended responses.

Open-Ended Response

Researchers opt for **closed ended** responses.This requires,

- Choosing an arbitrary scale
- Choosing **researcher defined** *categories*. Sometimes putting an "other" open ended option.

A debate exists on whether this is a good idea.

There are workflow advantages to closed ended responses.

- In 10 minutes I can move from a mTurk survey, get 100 closed ended responses to questions, put the data in R and type lm()

Open-Ended Response

Researchers opt for **closed ended** responses.This requires,

- Choosing an arbitrary scale
- Choosing **researcher defined** *categories*. Sometimes putting an "other" open ended option.

A debate exists on whether this is a good idea.

There are workflow advantages to closed ended responses.

- In 10 minutes I can move from a mTurk survey, get 100 closed ended responses to questions, put the data in R and type lm()

We want open-ended analysis to be (almost) that easy.

Survey Experiment

Rand et al., *Nature*, "Spontaneous giving and calculated greed."

Survey Experiment

Rand et al., *Nature*, “Spontaneous giving and calculated greed.”

- Gut responses are cooperative

Survey Experiment

Rand et al., *Nature*, "Spontaneous giving and calculated greed."

- Gut responses are cooperative
- calculated responses lead to defection in prisoner's dilemma

Survey Experiment

Rand et al., *Nature*, "Spontaneous giving and calculated greed."

- Gut responses are cooperative
- calculated responses lead to defection in prisoner's dilemma
- Subjects were told to

Survey Experiment

Rand et al., *Nature*, “Spontaneous giving and calculated greed.”

- Gut responses are cooperative
- calculated responses lead to defection in prisoner’s dilemma
- Subjects were told to
    1. Write about when they have acted out of intuition, or feeling

Survey Experiment

Rand et al., *Nature*, "Spontaneous giving and calculated greed."

- Gut responses are cooperative
- calculated responses lead to defection in prisoner's dilemma
- Subjects were told to
  1. Write about when they have acted out of intuition, or feeling
  2. Write about a time when they reflected and thought a lot about something.

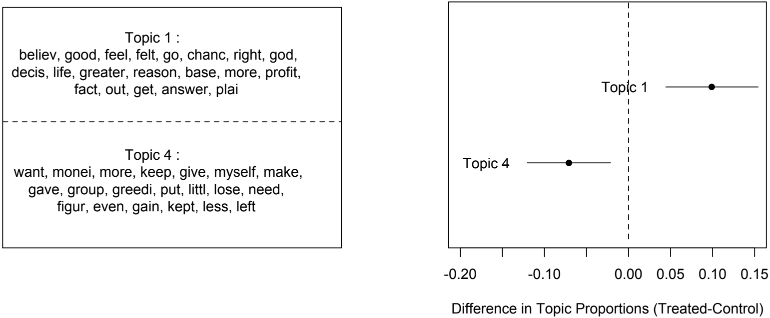
Survey Experiment

Rand et al., *Nature*, "Spontaneous giving and calculated greed."

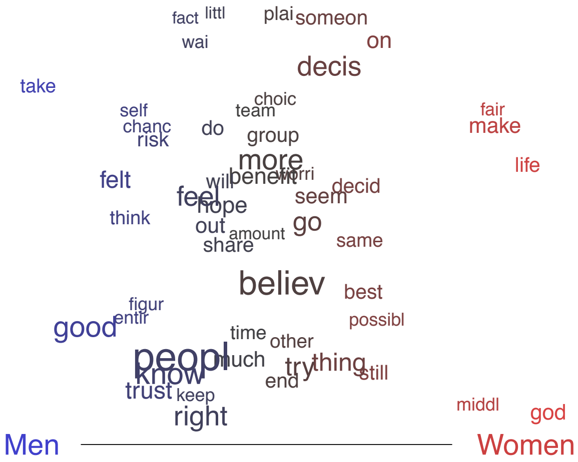
- Gut responses are cooperative
- calculated responses lead to defection in prisoner's dilemma
- Subjects were told to
  1. Write about when they have acted out of intuition, or feeling
  2. Write about a time when they reflected and thought a lot about something.

Afterward, subjects asked to describe their reasoning.

Intuition Priming Effects



Different Intuitive Strategy: Women vs. Men



Conclusion

- Applied Social Science

Conclusion

- Applied Social Science
  - Explanation vs. prediction/exploration

Conclusion

- Applied Social Science
  - Explanation vs. prediction/exploration
  - Background covariates on documents

Conclusion

- Applied Social Science
  - Explanation vs. prediction/exploration
  - Background covariates on documents
  - Need off-the-shelf tools

Conclusion

- Applied Social Science
  - Explanation vs. prediction/exploration
  - Background covariates on documents
  - Need off-the-shelf tools
- Our Contribution

Conclusion

- Applied Social Science
  - Explanation vs. prediction/exploration
  - Background covariates on documents
  - Need off-the-shelf tools
- Our Contribution
  - A new topic model for incorporating covariate info

Conclusion

- Applied Social Science
  - Explanation vs. prediction/exploration
  - Background covariates on documents
  - Need off-the-shelf tools
- Our Contribution
  - A new topic model for incorporating covariate info
  - New software tools (releasing in the next few weeks)

Conclusion

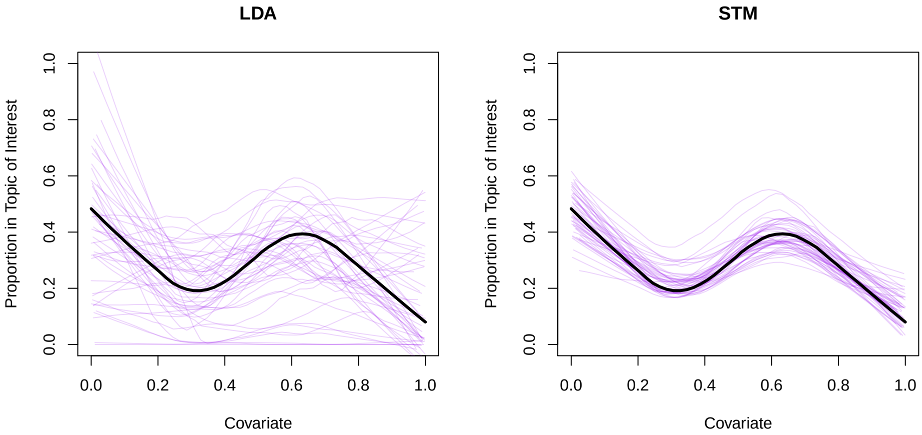
- Applied Social Science
  - Explanation vs. prediction/exploration
  - Background covariates on documents
  - Need off-the-shelf tools
- Our Contribution
  - A new topic model for incorporating covariate info
  - New software tools (releasing in the next few weeks)
  - Methods for model selection, labeling topics and others

Thanks!

Papers at:

scholar.harvard.edu/~bstewart

LDA and STM


LDA
STM
Proportion in Topic of Interest
0.0
0.2
0.4
0.6
0.8
1.0
Covariate