Machine Learning Methods for Economists
Latent Dirichlet Allocation

Stephen Hansen
University of Oxford

Introduction

Recall we are interested in mixed-membership modeling, but that the pLSI model has a huge number of parameters to estimate.

One solution is to adopt a Bayesian approach; the pLSI model with a prior distribution on the document-specific mixing probabilities is called Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003).

LDA is widely used within computer science and, increasingly, social sciences.

LDA forms the basis of many, more complicated mixed-membership models.

Latent Dirichlet Allocation—Original

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Each word $w_{d,n}$ in document d is generated from a two-step process:
   - 2.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   - 2.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.

Estimate hyperparameters $\boldsymbol{\alpha}$ and term probabilities $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$.

Latent Dirichlet Allocation—Modified

1. Draw $\boldsymbol{\beta}_k$ independently for $k = 1, \ldots, K$ from Dirichlet(η).
2. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet(α).
3. Each word $w_{d,n}$ in document d is generated from a two-step process:
   - 3.1 Draw topic assignment $z_{d,n}$ from θ_d.
   - 3.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.

Fix scalar values for η and α.

Example statement: Yellen, March 2006, #51

Raw Data $\rightarrow$ Remove Stop Words $\rightarrow$ Stemming $\rightarrow$ Multi-word tokens = Bag of Words

We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

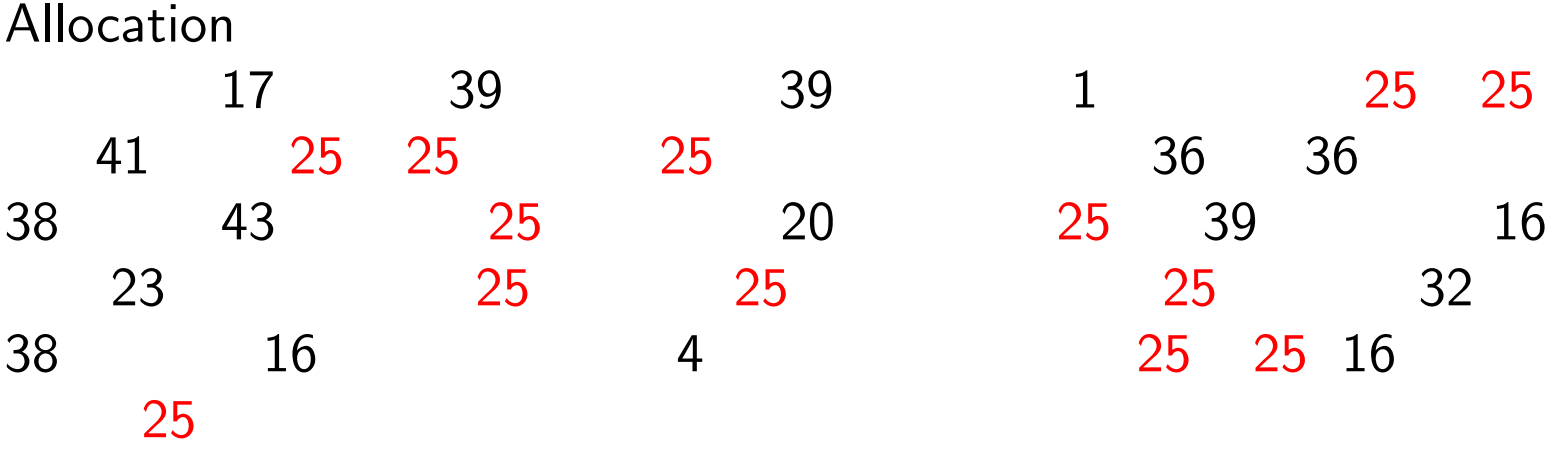
noticed change relationship between core CPI chained core CPI suggested maybe something going relating substitution bias upper level index focused nonmarket component PCE wondered something unusual happening core CPI relative measures

Example statement: Yellen, March 2006, #51

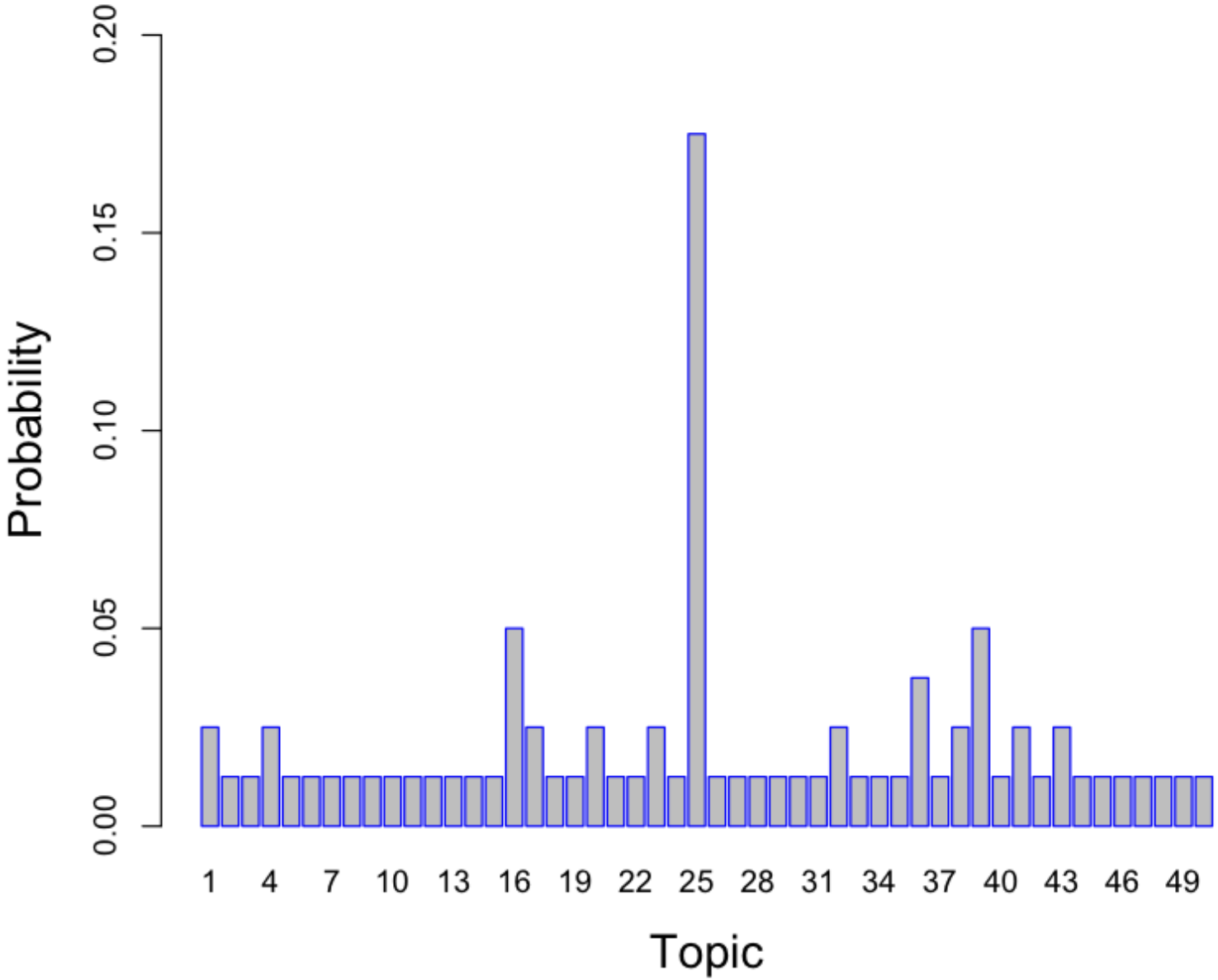
Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

notic chang relationship between core CPI chain core CPI suggest mayb someth go relat substitut bia upper level index focus nonmarket compon PCE wonder someth unusu happen core CPI rel measur

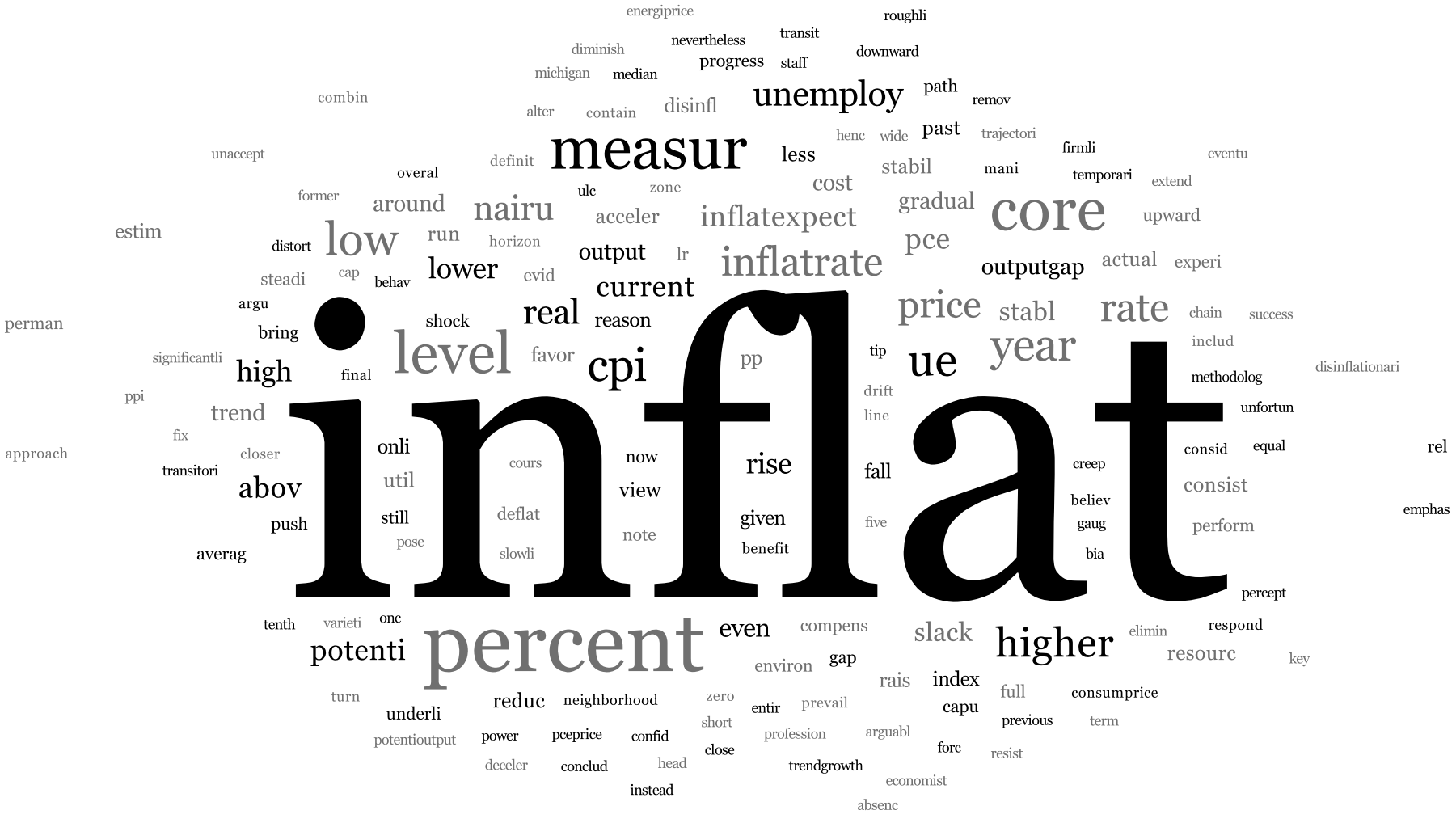
Example statement: Yellen, March 2006, #51



Distribution of Attention



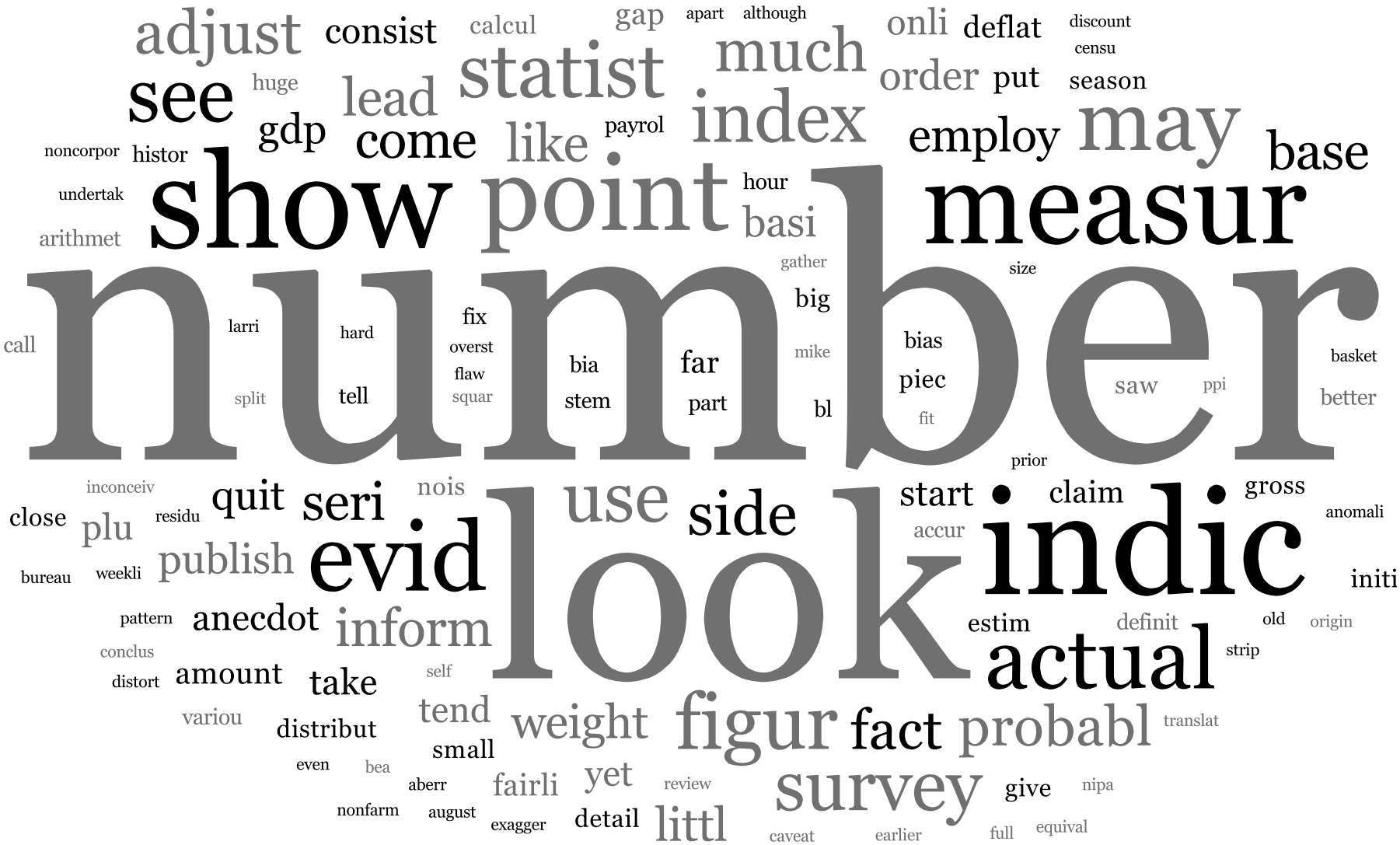
Topic 25


energiprice
roughli
nevertheless
transit
diminish
progress
staff
downward
michigan
median
combin
unemploy
path
alter
contain
disinfl
remov
measur
past
henc
wide
trajectori
unaccept
definit
less
firmli
eventu
overal
stabil
mani
temporari
former
ulc
zone
cost
extend
around
nairu
acceler
inflatexpect
gradual
core
upward
estim
low
run
horizon
pce
distort
output
lr
inflatrate
outputgap
actual
experi
steadi
cap
behav
lower
evid
current
argu
price
stabl
rate
chain
success
perman
bring
shock
real
reason
level
favor
cpi
pp
tip
ue
year
includ
significantli
high
final
methodolog
disinflationari
ppi
drift
line
trend
unfortun
fix
onli
consid
equal
rel
approach
closer
cours
now
rise
fall
creep
transitori
util
view
consist
abov
believ
deflat
given
five
gaug
emphas
push
still
note
perform
pose
benefit
averag
slowli
bia
percept
tenth
varieti
onc
even
compens
slack
higher
elimin
respond
potenti
percent
resourc
key
environ
gap
rais
index
full
consumprice
turn
reduc
neighborhood
zero
entir
prevail
capu
underli
short
profession
arguabl
previous
term
potentioutput
power
pceprice
confid
close
forc
resist
deceler
conclud
head
trendgrowth
economist
instead
absenc

Advantage of Flexibility

'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11

Topic 11


adjust
consist
calcul
gap
apart
although
onli
deflat
discount
censu
see
huge
lead
statist
much
order
put
season
gdp
come
like
payrol
index
employ
may
noncorpor
histor
base
undertak
show
point
hour
basi
measur
arithmet
number
gather
size
big
call
larri
hard
fix
overst
flaw
bia
far
mike
bias
basket
split
tell
squar
stem
part
bl
piec
saw
ppi
better
fit
prior
inconceiv
quit
seri
nois
use
side
start
claim
gross
close
plu
residu
look
accur
anomali
bureau
weekli
publish
evid
indic
initi
pattern
anecdot
inform
estim
definit
old
origin
conclus
actual
strip
distort
amount
take
self
variou
distribut
tend
weight
figur
fact
probabl
translat
small
even
bea
survey
aberr
fairli
yet
review
give
nipa
nonfarm
august
exagger
detail
littl
caveat
earlier
full
equival

Advantage of Flexibility

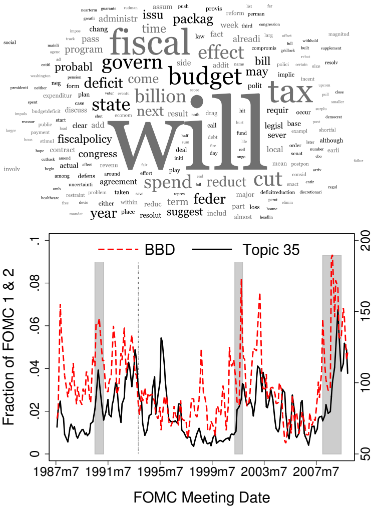
'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11.

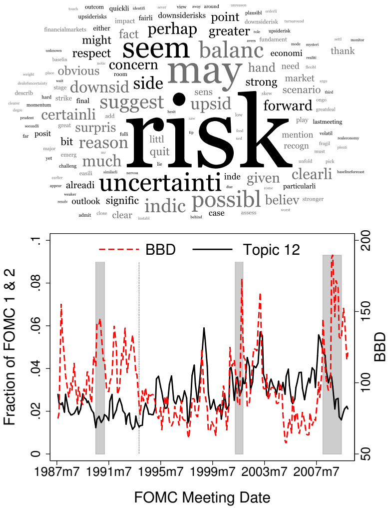
It gets assigned to 25 in this statement consistently due to the presence of other topic 25 words.

In statements containing words on evidence and numbers, it consistently gets assigned to 11.

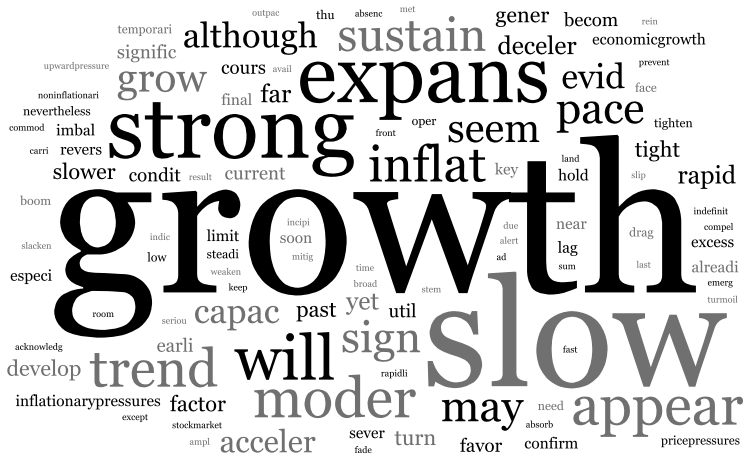
Sampling algorithm can help place words in their appropriate context.

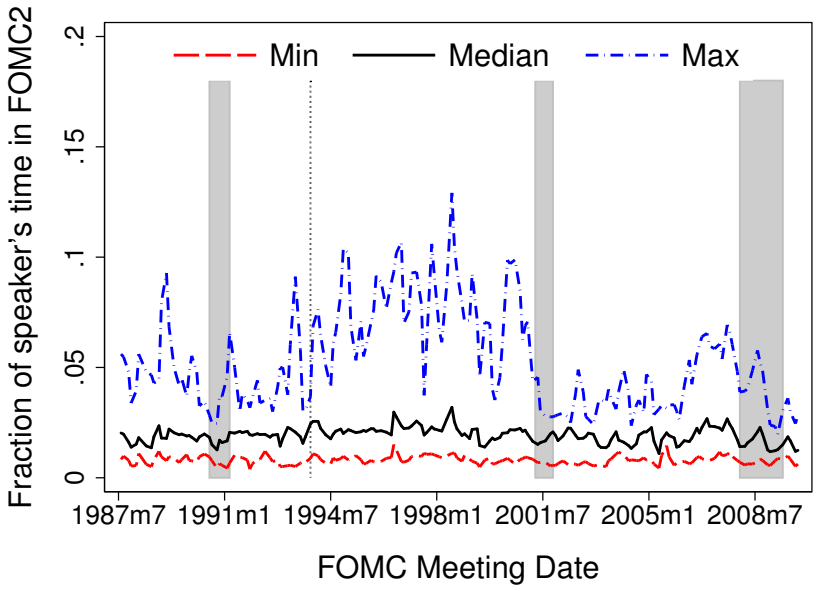
External Validation—BBD

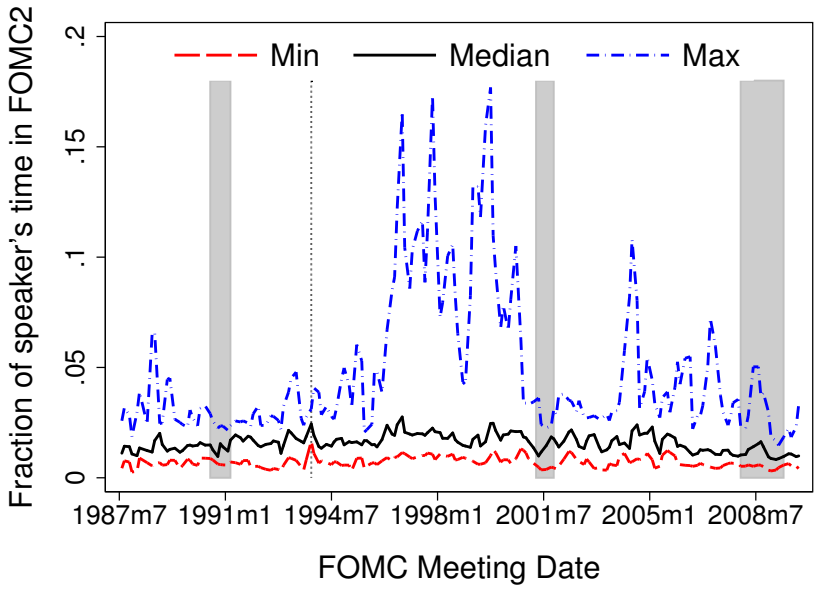




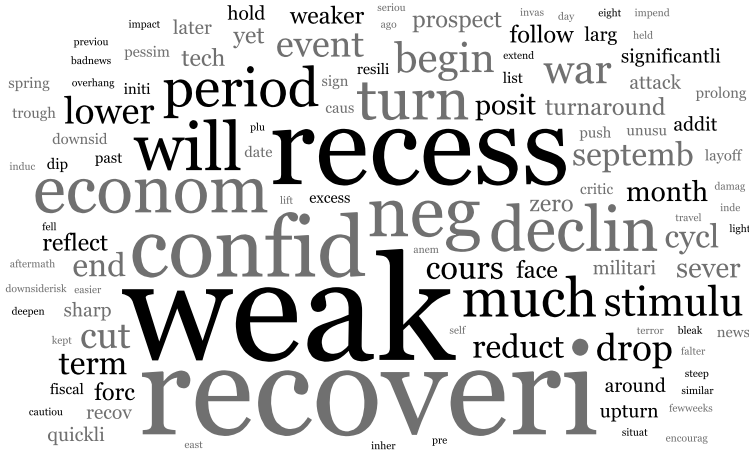
Pro-Cyclical Topics

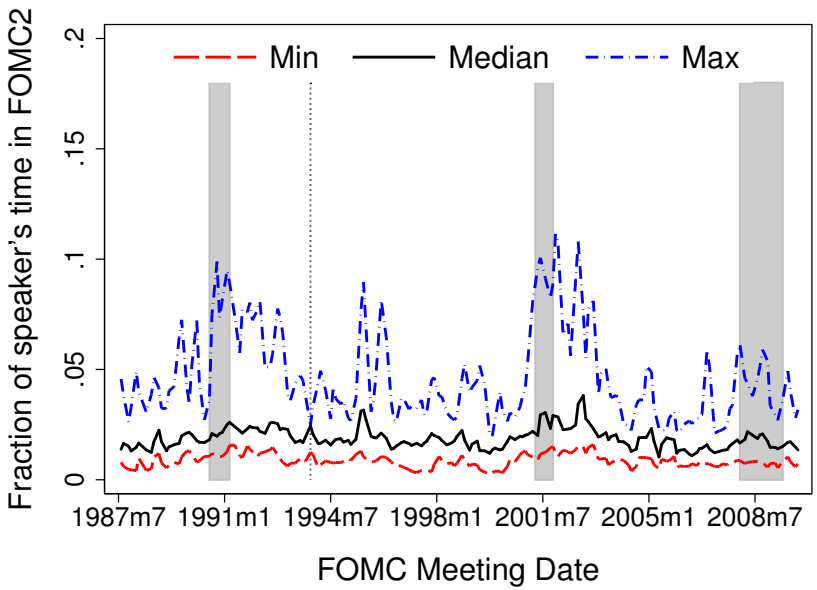


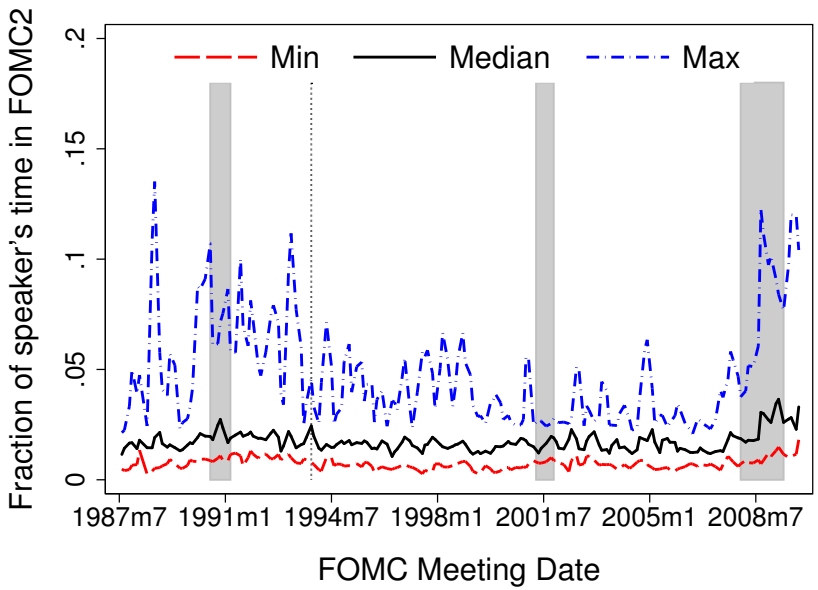




Counter-Cyclical Topics



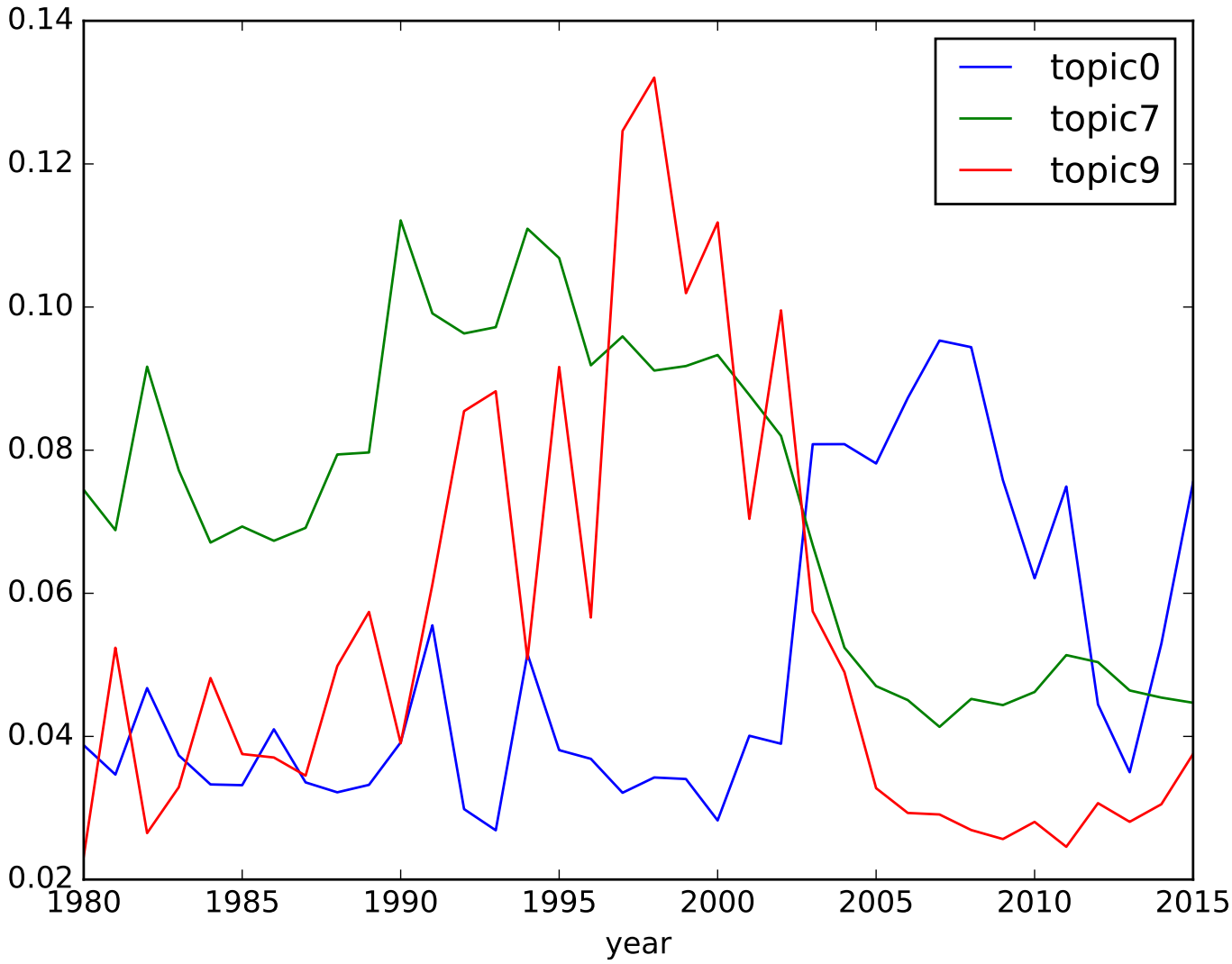




Topics on NYT Data (Iraq, Iran, Syria from mid-1980s)

Topic	Top Terms
0	american.forc.militari.troop.command.iraqi.gener.armi.iraq.offic
2	shiit.mr.govern.sunni.polit.parti.leader.iraqi.elect.minist
3	iranian.attack.air.iraqi.gulf.report.today.missil.forc.fire
4	iran.iranian.islam.ayatollah.presid.leader.teheran.govern.polit.revolut
6	iran.nuclear.iranian.program.sanction.negoti.enrich.agenc.uranium.deal
7	iraq.iraqi.hussein.baghdad.war.saddam.kuwait.nation.today.countri
8	govern.compani.bank.state.money.work.million.billion.project.contract
9	weapon.intellig.report.use.inspector.chemic.nation.site.program.offici
10	syria.israel.syrian.arab.isra.mr.lebanon.assad.saudi.presid
11	oil.percent.year.price.countri.export.million.econom.day.trade
13	kill.american.attack.baghdad.bomb.iraqi.polic.offici.al.insurg
14	unit.nation.council.secur.mr.resolut.diplomat.meet.foreign.franc
16	mr.report.prison.releas.charg.case.court.arrest.accus.investig
18	govern.syria.group.kurdish.syrian.turkey.forc.opposit.border.rebel

Distribution of Topics in Iraq Articles



Applications of LDA

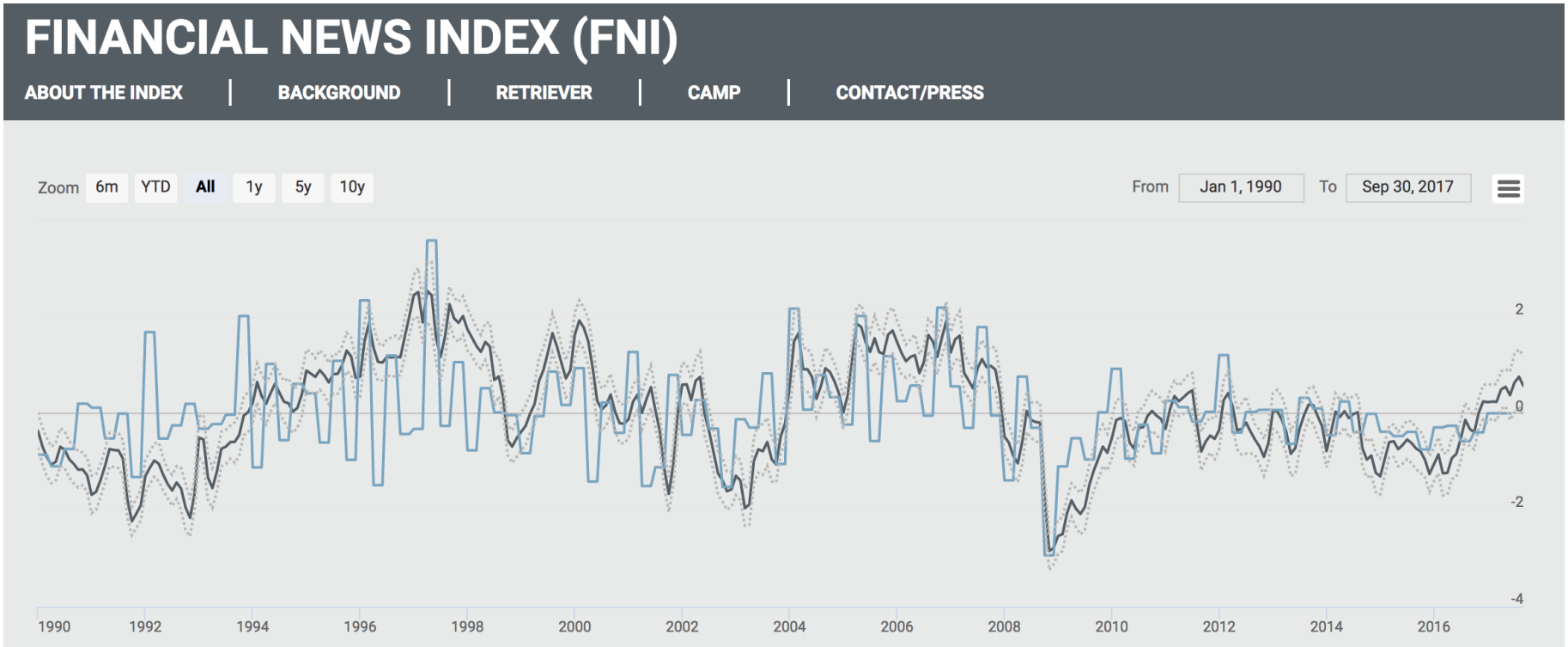
1. Forecasting: Mueller and Rauh (2018);[1] Larsen and Thorsrud (2018).[2]
2. Transparency: Hansen et. al. (2017).
3. Information Processing: Nimark and Pitschner (2017).[3]
4. Basis for structural estimation?

[1] APSR, "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text".

[2] Journal of Econometrics, "The Value of News for Economic Developments".

[3] WP, "News Media and Delegated Information Choice".

Financial News Index



Graphical Models

Consider a probabilistic model with joint distribution $p(\mathbf{x})$ over the random variables $\mathbf{x} = (x_1, \ldots, x_N)$.

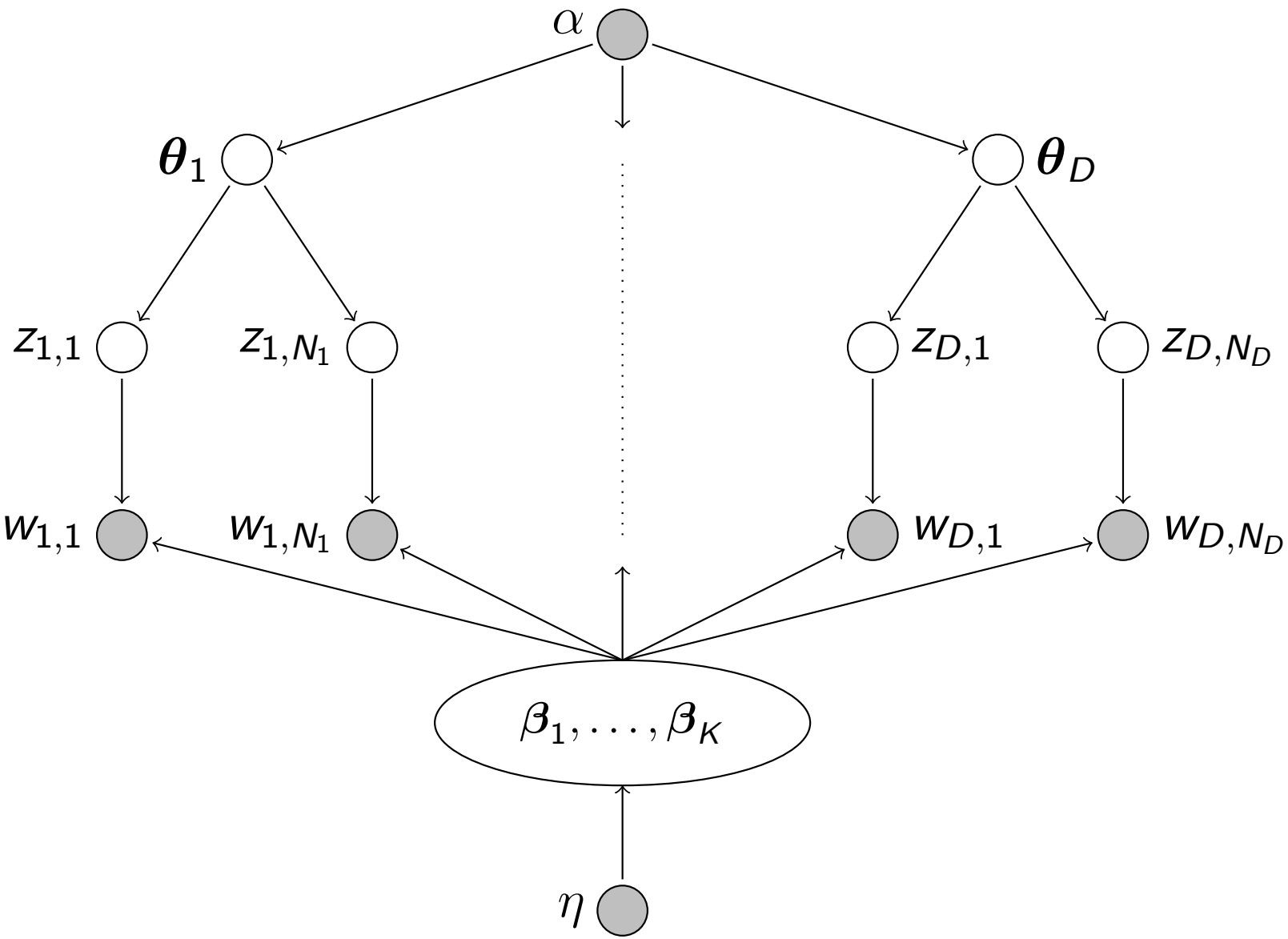
In high-dimensional models, it is useful to summarize relationships among random variables with directed graphs in which nodes represent random variables and links between nodes represent dependencies.

A node's *parents* are the set of nodes that link to it; a node's *children* are the the set of nodes that it links to.

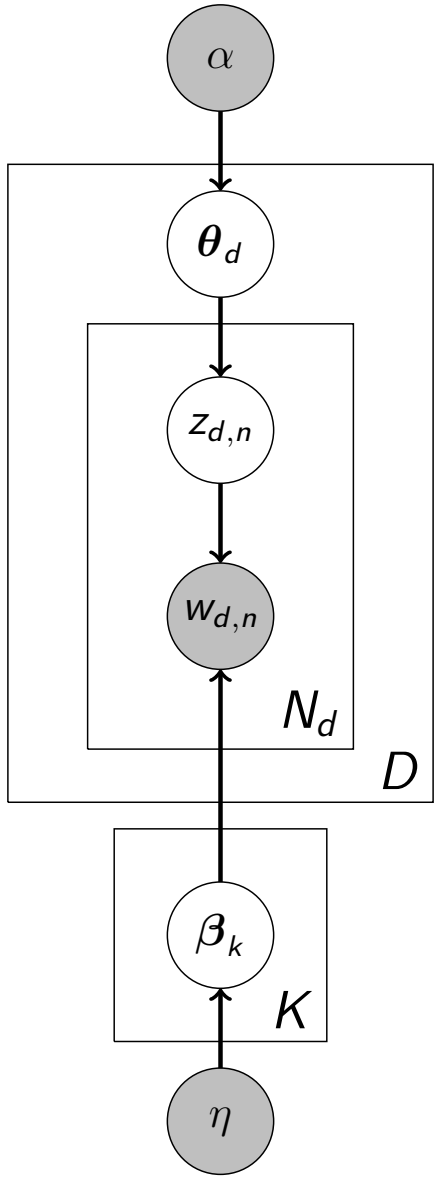
A *Bayesian network* is a probabilistic model whose joint distribution can be represented by a directed acyclic graph (DAG).

The nodes in a DAG can be ordered so that parents precede children.

LDA as a Bayesian Network



LDA Plate Diagram



Conditional Independence Property I

In a Bayesian network, nodes are independent of their ancestors conditional on their parents.

This means we can write $p(\mathbf{x}) = \prod_{i=1}^{N} p\left(x_i \mid \text{parents}(x_i) \right)$, which can greatly simplify joint distributions.

Applying this formula to LDA yields

$$\left(\prod_d \prod_n p\left(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta} \right) \right) \left(\prod_d \prod_n p\left(z_{d,n} \mid \boldsymbol{\theta}_d \right) \right) \times \left(\prod_d p\left(\boldsymbol{\theta}_d \mid \alpha \right) \right) \left(\prod_k p\left(\boldsymbol{\beta}_k \mid \eta \right) \right)$$

Conditional Independence Property II

The *Markov blanket* $\mathrm{MB}(x_i)$ of a node x_i in a Bayesian network is the set of nodes consisting of x_i's parents, children, and children's parents.

Conditional on its Markov blanket, the node x_i is independent of all nodes outside its Markov blanket.

So $p\left(x_i \mid \mathbf{x}_{-i}\right)$ has the same distribution as $p\left(x_i \mid \mathrm{MB}(x_i)\right)$.

Posterior Distribution

The inference problem in LDA is to compute the posterior distribution over $\mathbf{z}$, $\boldsymbol{\theta}$, and $\boldsymbol{\beta}$ given the data $\mathbf{w}$ and Dirichlet hyperparameters.

Let's consider the simpler problem of inferring the latent variables taking the parameters as given. Posterior distribution is

$$p(\mathbf{z} = z' \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{p(\mathbf{w} \mid \mathbf{z} = z', \boldsymbol{\theta}, \boldsymbol{\beta})\, p(\mathbf{z} = z' \mid \boldsymbol{\theta}, \boldsymbol{\beta})}{\sum_{z'} p(\mathbf{w} \mid \mathbf{z} = z', \boldsymbol{\theta}, \boldsymbol{\beta})\, p(\mathbf{z} = z' \mid \boldsymbol{\theta}, \boldsymbol{\beta})}.$$

Posterior Distribution

The inference problem in LDA is to compute the posterior distribution over $\mathbf{z}$, $\boldsymbol{\theta}$, and $\boldsymbol{\beta}$ given the data $\mathbf{w}$ and Dirichlet hyperparameters.

Let's consider the simpler problem of inferring the latent variables taking the parameters as given. Posterior distribution is

$$p(\mathbf{z} = z' \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{p(\mathbf{w} \mid \mathbf{z} = z', \boldsymbol{\theta}, \boldsymbol{\beta})\, p(\mathbf{z} = z' \mid \boldsymbol{\theta}, \boldsymbol{\beta})}{\sum_{z'} p(\mathbf{w} \mid \mathbf{z} = z', \boldsymbol{\theta}, \boldsymbol{\beta})\, p(\mathbf{z} = z' \mid \boldsymbol{\theta}, \boldsymbol{\beta})}.$$

We can compute the numerator easily, and each element of denominator.

But $z' \in \{1, \ldots, K\}^N \Rightarrow$ there are K^N terms in the sum $\Rightarrow$ intractable problem.

For example, a 100 word corpus with 50 topics has $\approx 7.88 \times 10^{169}$ terms.

Approximate Inference

Instead of obtaining a closed-form solution for the posterior distribution, we must approximate it.

Markov chain Monte Carlo methods provide a stochastic approximation to the true posterior.

The general idea is to define a Markov chain whose stationary distribution is equivalent to the posterior distribution, which we then draw samples from.

There are several MCMC methods, but we will consider Gibbs sampling.

Gibbs Sampling Review

We want to draw samples from some joint distribution over $\mathbf{x} = (x_1, \ldots, x_N)$ given by $p(\mathbf{x})$ (e.g. a posterior distribution).

Suppose we can compute the conditional distribution $p\left(x_i \mid \mathbf{x}_{-i}\right)$.

Then we can use the following algorithm:

1. Randomly allocate an initial value for $\mathbf{x}$, say $\mathbf{x}^0$
2. Let S be the number of iterations to run chain. For each $s \in \{1, \ldots, S\}$, draw x_i^s according to

$$x_i^s \sim p\left(x_i \mid x_1^s, \ldots, x_{i-1}^s, x_{i+1}^{s-1}, \ldots, x_N^{s-1}\right)$$

3. Discard initial iterations (burn in), and collect samples from every mth (thinning interval) iteration thereafter.
4. Use collected samples to approximate joint distribution, or related distributions and moments.

Sampling Equations for $\boldsymbol{\theta}_d$

The Markov blanket of $\boldsymbol{\theta}_d$ is:

- The parent α.
- The children $\mathbf{z}_d$.

So we need to draw samples from $p(\boldsymbol{\theta}_d \mid \alpha, \mathbf{z}_d)$. This is the posterior distribution for $\boldsymbol{\theta}_d$ given a fixed value for the vector of allocation variables $\mathbf{z}_d$.

Let $n_{d,k} \equiv \sum_n \mathbb{1}(z_{d,n} = k)$ be the number of words in document d that have topic allocation k.

Then $p(\boldsymbol{\theta}_d \mid \alpha, \mathbf{z}_d) = \text{Dir}(\alpha + n_{d,1}, \ldots, \alpha + n_{d,K})$.

More Detailed Derivation

By Bayes' Rule we have $p(\boldsymbol{\theta}_d \mid \alpha, \mathbf{z}_d) \propto p(\mathbf{z}_d \mid \boldsymbol{\theta}_d)\, p(\boldsymbol{\theta}_d \mid \alpha)$.

$p(\mathbf{z}_d \mid \boldsymbol{\theta}_d)$ is essentially the same likelihood function we saw in previous slides. It is

$$p(\mathbf{z}_d \mid \boldsymbol{\theta}_d) = \prod_n \sum_k \mathbb{1}(z_{d,n} = k)\theta_{d,k} = \prod_k \theta_{d,k}^{n_{d,k}}.$$

Putting this together, we arrive at

$$p(\boldsymbol{\theta}_d \mid \alpha, \mathbf{z}_d) \propto \prod_k \theta_{d,k}^{n_{d,k}} \prod_k \theta_{d,k}^{\alpha-1} = \prod_k \theta_{d,k}^{n_{d,k}+\alpha-1}$$

which is exactly the Dirichlet posterior we saw in the previous slide.

Sampling Equations for $\boldsymbol{\beta}_k$

The Markov blanket of $\boldsymbol{\beta}_k$ is:

- The parent η.
- The children $\mathbf{w}$.
- The children's parents $\mathbf{z}$ and $\boldsymbol{\beta}_{-k}$.

Let $m_{k,v} \equiv \sum_n \sum_d \mathbb{1}(z_{d,n} = k)\mathbb{1}(w_{d,n} = v)$ be the number of times topic k allocation variables generate term v.

Only the allocation variables assigned to k—and their associated words—are informative about $\boldsymbol{\beta}_k$.

$$p\left(\boldsymbol{\beta}_k \mid \eta, \mathbf{w}, \mathbf{z}, \boldsymbol{\beta}_{-k}\right) = \text{Dir}\left(\eta + m_{k,1}, \ldots, \eta + m_{k,V}\right).$$

More Detailed Derivation

By Bayes' Rule we have $p\left(\boldsymbol{\beta}_k \mid \mathbf{z}, \mathbf{w}, \eta, \boldsymbol{\beta}_{-k}\right) \propto p\left(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}_k \mid \eta\right)$.

The likelihood function $p\left(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\beta}\right)$ takes the form

$$p\left(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\beta}\right) = \prod_d \prod_n \sum_v \sum_{k'} \mathbb{1}(w_{d,n} = v)\mathbb{1}(z_{d,n} = k')\beta_{k',v} =$$

$$\prod_v \prod_{k'} \beta_{k',v}^{m_{k',v}} = \prod_v \beta_{k,v}^{m_{k,v}} \prod_v \prod_{k' \neq k} \beta_{k',v}^{m_{k',v}} \propto \prod_v \beta_{k,v}^{m_{k,v}}.$$

Putting this together, we arrive at

$$p\left(\boldsymbol{\beta}_k \mid \mathbf{z}, \mathbf{w}, \eta, \boldsymbol{\beta}_{-k}\right) \propto \prod_v \beta_{k,v}^{m_{k,v}} \prod_v \beta_{k,v}^{\eta-1} = \prod_v \beta_{k,v}^{m_{k,v}+\eta-1}$$

which is exactly the Dirichlet posterior we saw in the previous slide.

Sampling Equations for Allocations

The Markov blanket of $z_{d,n}$ is:

- The parent $\boldsymbol{\theta}_d$.
- The child $w_{d,n}$.
- The child's parents $\boldsymbol{\beta}$.

$$\Pr[z_{d,n} = k \mid w_{d,n} = v, \boldsymbol{\beta}, \boldsymbol{\theta}_d] =$$

$$\frac{\Pr[w_{d,n} = v \mid z_{d,n} = k, \boldsymbol{\beta}, \boldsymbol{\theta}_d] \Pr[z_{d,n} = k \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d]}{\sum_k \Pr[w_{d,n} = v \mid z_{d,n} = k, \boldsymbol{\beta}, \boldsymbol{\theta}_d] \Pr[z_{d,n} = k \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d]} = \frac{\theta_d^k \beta_k^v}{\sum_k \theta_d^k \beta_k^v}$$

Summary

To complete one iteration of Gibbs sampling, we need to:

1. Sample from a multinomial distribution N times for the topic allocation variables.
2. Sample from a Dirichlet D times for the document-specific mixing probabilities.
3. Sample from a Dirichlet K times for the topic-specific term probabilities.

Sampling from these distributions is standard, and implemented in many programming languages.

Collapsed Sampling

Collapsed sampling refers to analytically integrating out some variables in the joint likelihood and sampling the remainder.

This tends to be more efficient because we reduce the dimensionality of the space we sample from.

Griffiths and Steyvers (2004)[4] proposed a collapsed sampler for LDA that integrates out the $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ terms and samples only $\mathbf{z}$.

For details see Heinrich (2009)[5] and technical appendix of Hansen, McMahon, and Prat (2015).

[4] PNAS, "Finding Scientific Topics".

[5] Technical Report, "Parameter estimation for text analysis".

Collapsed Sampling Equation for LDA

The sampling equation for the nth allocation variable in document d is:

$$\Pr\left[z_{d,n} = k \;\middle|\; \mathbf{z}_{-(d,n)}, \mathbf{w}, \alpha, \eta \right] \propto \frac{m^{-}_{k,w_{d,n}} + \eta}{\sum_v m^{-}_{k,v} + \eta V} \left(n^{-}_{d,k} + \alpha \right)$$

where the $-$ superscript denotes counts excluding (d, n) term.

Collapsed Sampling Equation for LDA

The sampling equation for the nth allocation variable in document d is:

$$\Pr\left[z_{d,n} = k \mid \mathbf{z}_{-(d,n)}, \mathbf{w}, \alpha, \eta \right] \propto \frac{m^{-}_{k,w_{d,n}} + \eta}{\sum_v m^{-}_{k,v} + \eta V} \left(n^{-}_{d,k} + \alpha \right)$$

where the $-$ superscript denotes counts excluding (d, n) term.

Probability term n in document d is assigned to topic k is increasing in:

1. The number of other terms in document d that are currently assigned to k.
2. The number of other occurrences of the term $w_{d,n}$ in the entire corpus that are currently assigned to k.

Both mean that terms that regularly co-occur in documents will be grouped together to form topics.

Property 1 means that terms within a document will tend to be grouped together into few topics rather than spread across many separate topics.

Predictive Distributions

Collapsed sampling gives the distribution of the allocation variables, but we also care about variables we integrated out.

Their predictive distributions are easy to form given topic assignments:

$$\widehat{\beta}_{k,v} = \frac{m_{k,v} + \eta}{\sum_{v=1}^{V} (m_{k,v} + \eta)} \quad \text{and} \quad \widehat{\theta}_{d,k} = \frac{n_{d,k} + \alpha}{\sum_{k=1}^{K} (n_{d,k} + \alpha)}.$$

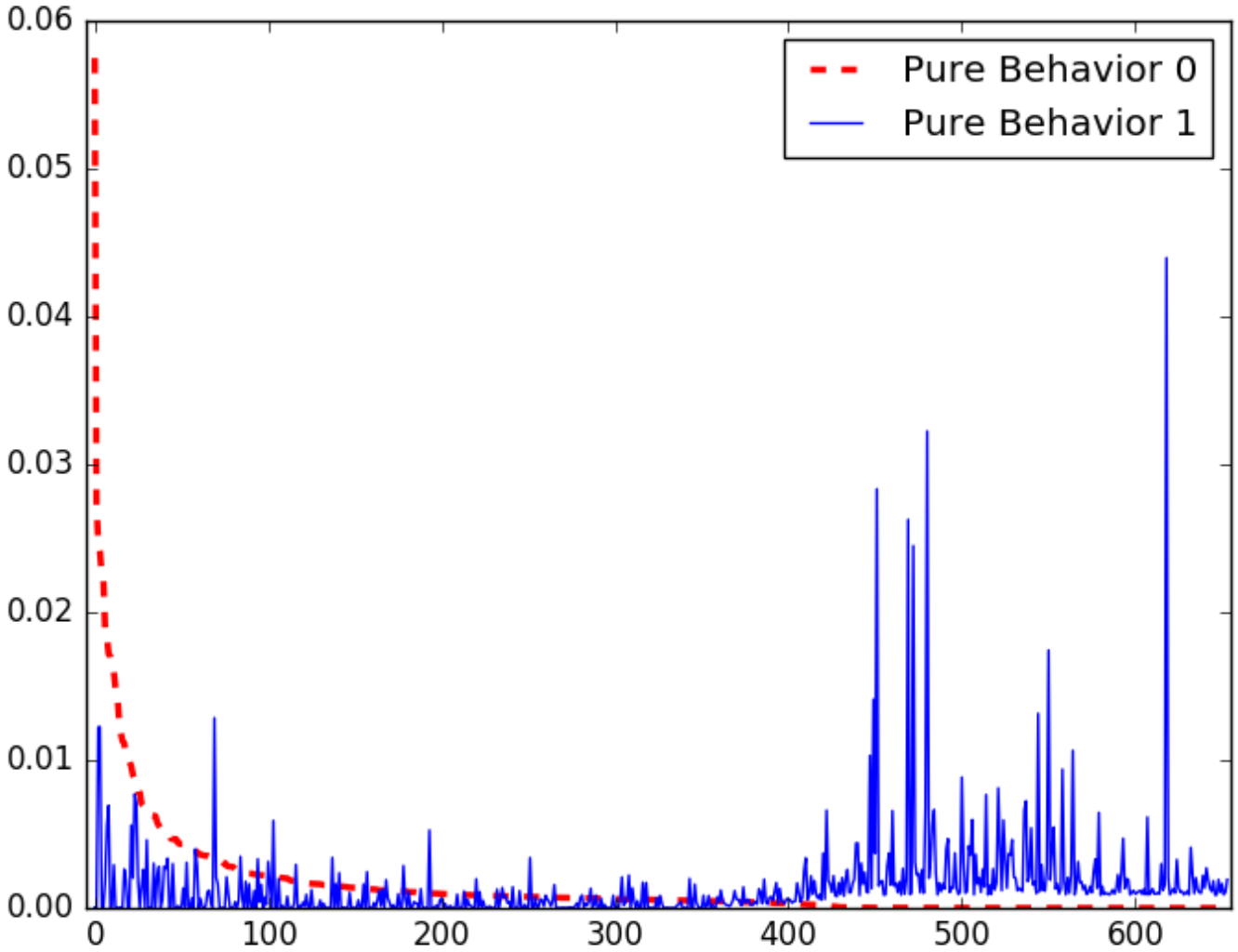
LDA on Survey Data

Recall from the first lecture that text data is one instance of count data.

Although typically applied to natural language, LDA is in principle applicable to *any* count data.

We recently applied it to CEO survey data to estimate management "behaviors" with $K = 2$.

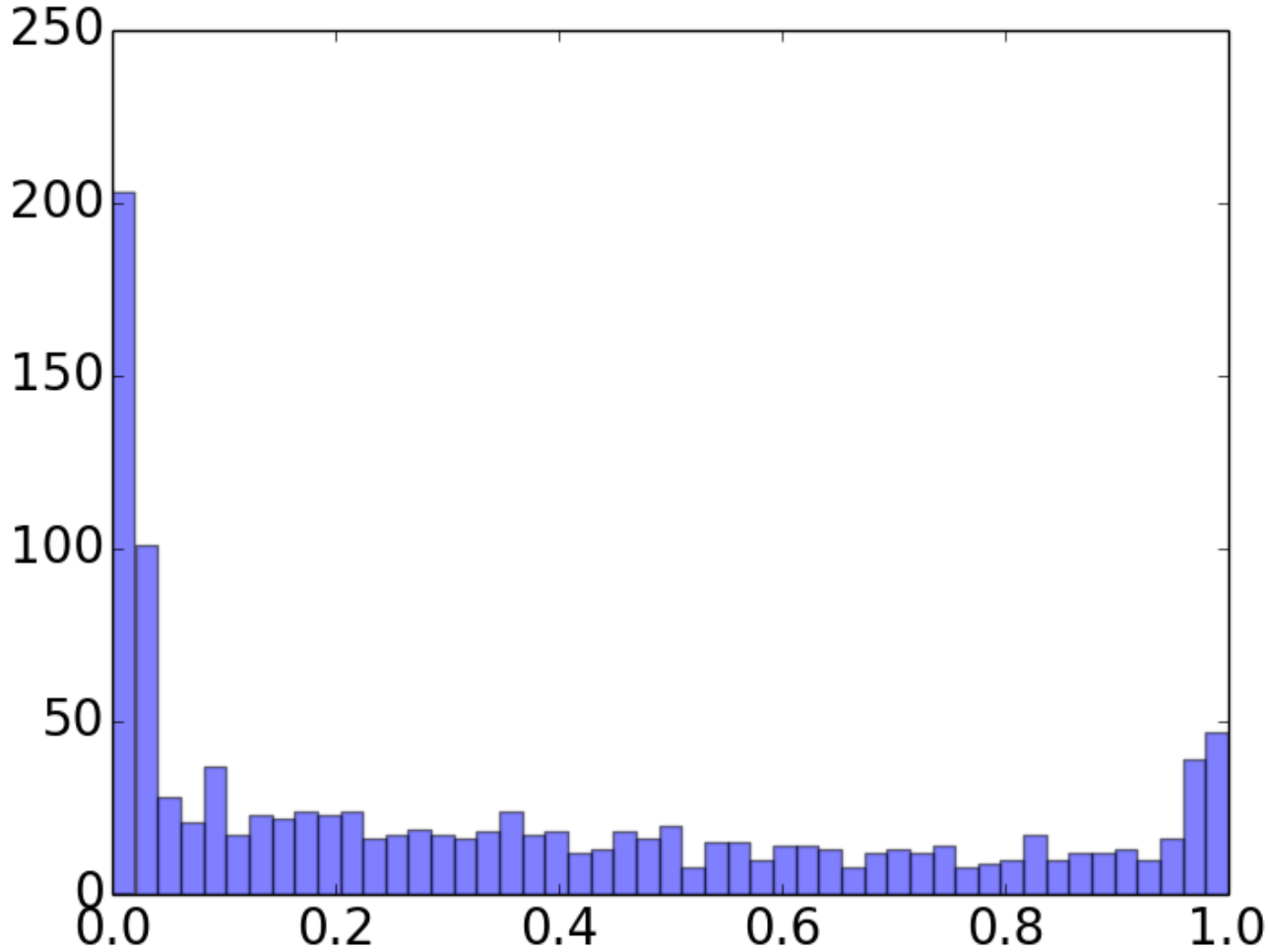
Pure Behaviors are Sharply Distinct



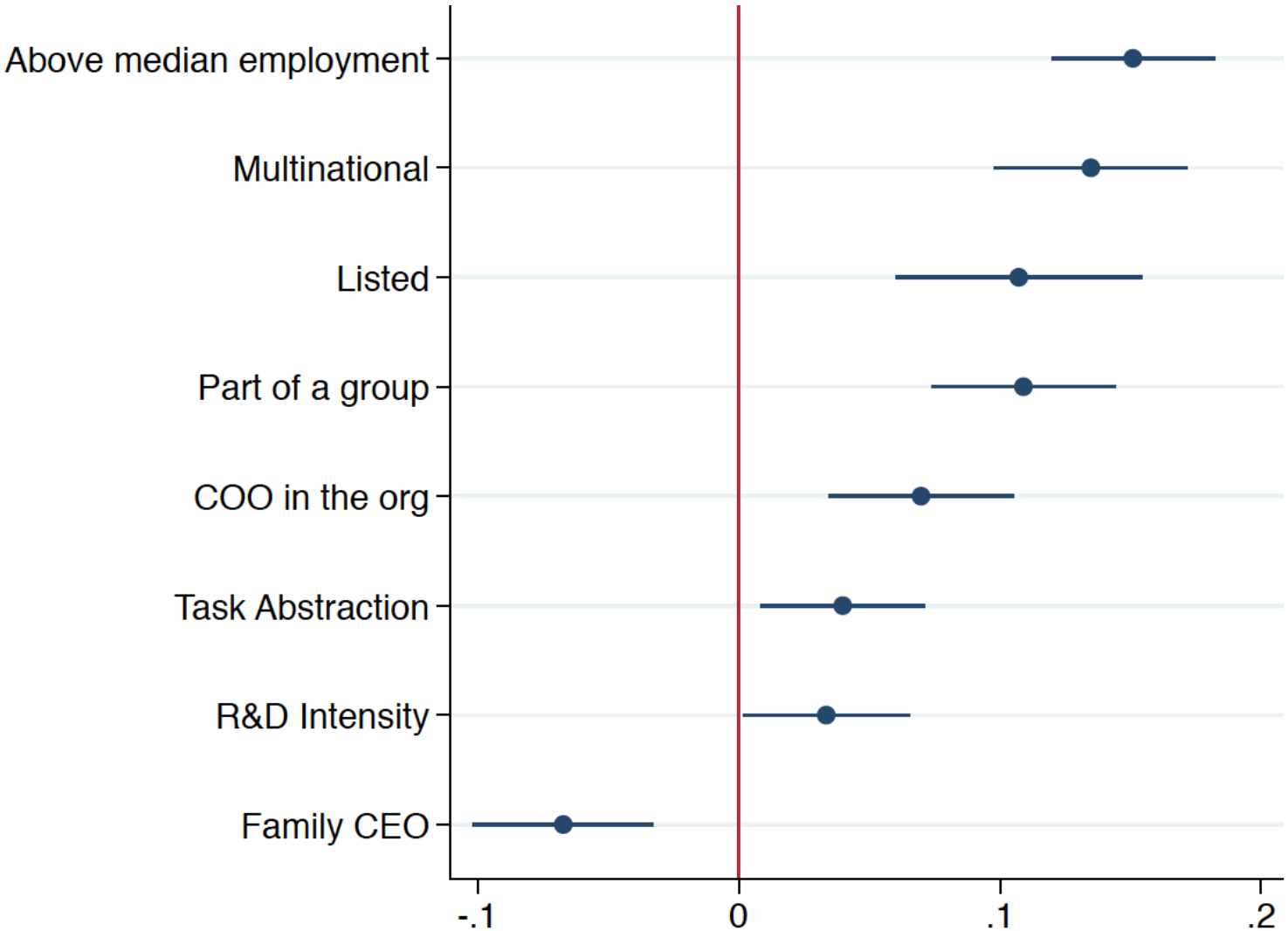
Differences across Pure Behaviors

Feature	X times less likely in Behavior 1	Feature	X times more likely in Behavior 1
Plant Visits	0.11	Communications	1.9
Just Outsiders	0.5	Outsiders + Insiders	1.9
Production	0.5	C-suite	34
Suppliers	0.3	Multifunction	1.5

Estimated Behavior Indices



Correlates of Behavior Index



Managers vs. Leaders

Kotter (1999) emphasizes a behavioral distinction between "management" and "leadership".

Management involves monitoring and implementing tasks, i.e. "setting up systems to ensure that plans are implemented precisely and efficiently."

Leadership aims primarily at the creation of organizational alignment, and involves significant investments in interpersonal communication.

> *The knowledge worker makes much greater time demands than the manual worker on his superiors as well as on his associates...One has to sit down with a knowledge worker and think through with him what should be done and why, before when knowing whether he is doing a satisfactory job or not (Drucker 1967).*

Out-of-Sample Documents

We are sometimes interested in obtaining the document-topic distribution for out-of-sample documents.

We can perform Gibbs sampling treating estimated topics as fixed

$$\Pr\left[z_{d,n} = k \;\middle|\; \mathbf{z}_{-(d,n)}, \mathbf{w}_d, \alpha, \eta \right] \propto \widehat{\beta}_{k,v_{d,n}} \left[n^{-}_{d,k} + \alpha \right]$$

for each out-of-sample document d.

Only 10-20 iterations necessary since topics already estimated.

Model Selection

There are three parameters to set to run the Gibbs sampling algorithm: number of topics K and hyperparameters α, η.

Priors don't receive too much attention in literature. Griffiths and Steyvers recommend $\eta = 200/V$ and $\alpha = 50/K$. Smaller values will tend to generate more concentrated distributions. (See also Wallach et. al. 2009 [6]).

Methods to choose K:

1. Predict text well $\rightarrow$ out-of-sample goodness-of-fit.
2. Information criteria.
3. Cohesion (focus on interpretability).

[6] NIPS, "Rethinking LDA: Why Priors Matter".

Cross Validation

Fit LDA on training data, obtain estimates of $\widehat{\beta}_1, \ldots, \widehat{\beta}_K$.

For test data, obtain θ_d distributions via sampling as above, or else use uniform distribution.

Compute log-likelihood of held-out data as

$$\ell\left(\mathbf{w} \mid \widehat{\Theta}\right) = \sum_{d=1}^{D} \sum_{v=1}^{V} x_{d,v} \log \left(\sum_{k=1}^{K} \widehat{\theta}_{d,k} \widehat{\beta}_{k,v} \right)$$

Higher values indicate better goodness-of-fit.

Information Criteria

Information criteria trade off goodness-of-fit with model complexity.

There are various forms: AIC, BIC, DIC, etc.

Erosheva et. al. (2007)[7] compare several in the context of an LDA-like model for survey data, and find that AICM is optimal.

Let $\mu_\ell = \frac{1}{S}\sum_s \ell\left(\mathbf{w} \mid \widehat{\Theta}^s\right)$ be the average value of the log-likelihood across S draws of a Markov chain and

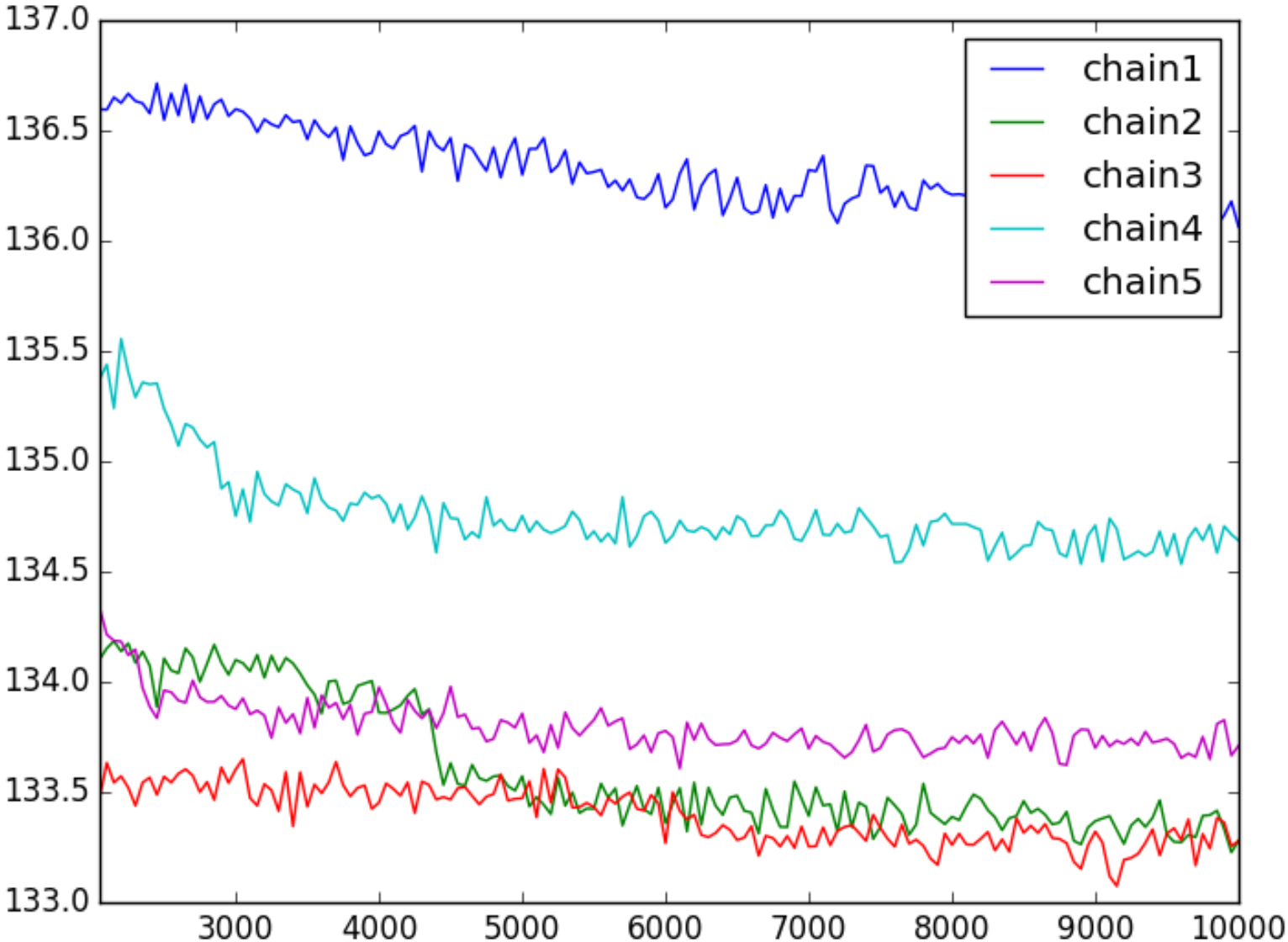
Let $\sigma^2_\ell = \frac{1}{S}\sum_s \left(\ell\left(\mathbf{w} \mid \widehat{\Theta}^s\right) - \mu_\ell\right)^2$ be the variance.

The AICM is $2(\mu_\ell - \sigma^2_\ell)$.

[7] Annals of Applied Statistics, "Describing Disability through Individual-Level Mixture Models for Multivariate Binary Data"

Perplexities with $K = 2$ on CEO Survey Data

Perplexities with $K = 10$ on CEO Survey Data



Formalizing Interpretablility

Chang et. al. (2009)[8] propose an objective way of determining whether topics are interpretable.

Two tests:

1. *Word intrusion*. Form set of words consisting of top five words from topic k + word with low probability in topic k. Ask subjects to identify inserted word.
2. *Topic intrusion*. Show subjects a snippet of a document + top three topics associated to it + randomly drawn other topic. Ask to identify inserted topic.

Estimate LDA and other topic models on NYT and Wikipedia articles for $K = 50, 100, 150$.

[8] NIPS, "Reading Tea Leaves: How Humans Interpret Topic Models".

Results

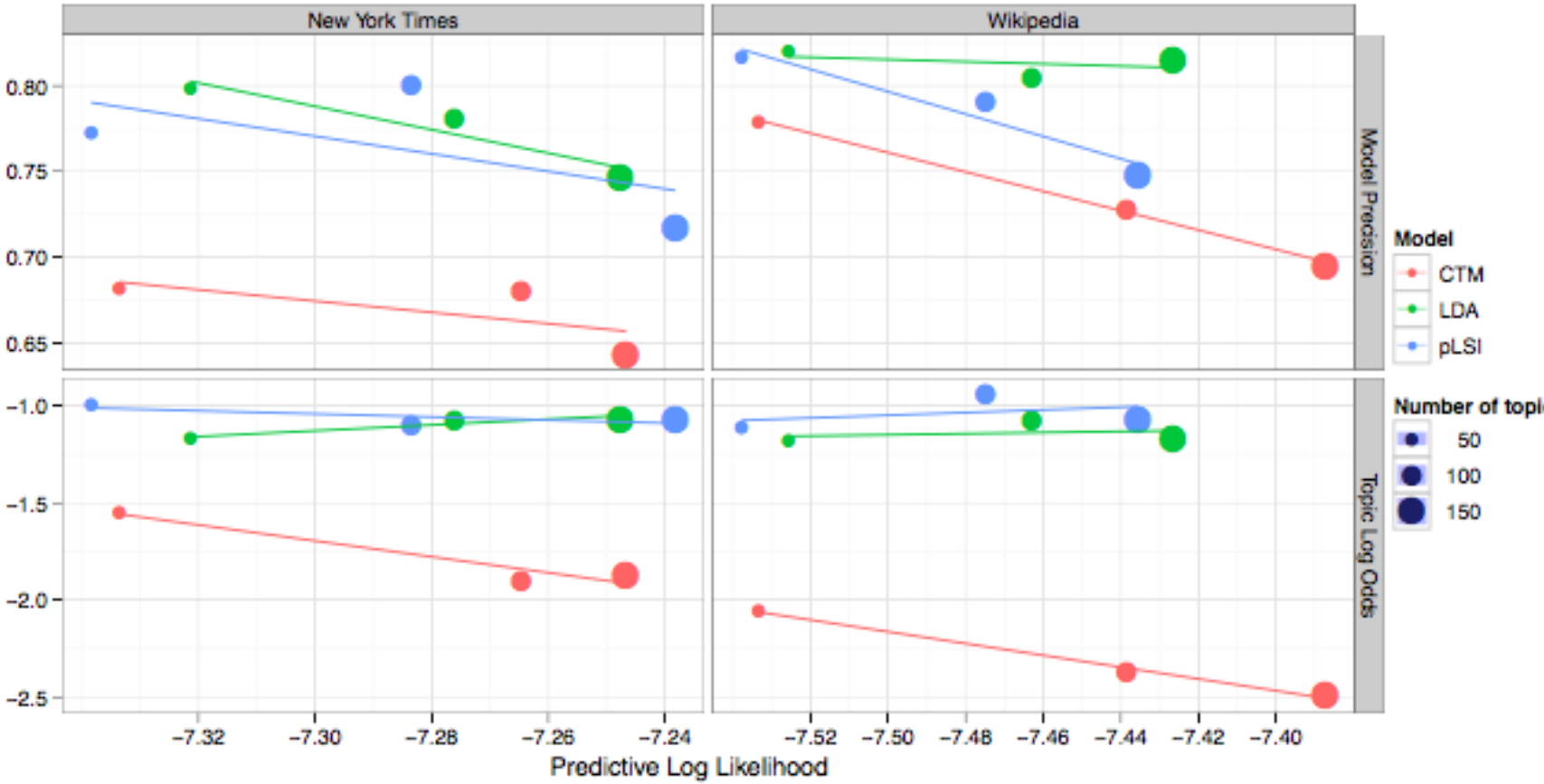


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase the agreement between human subjects and the model for either task (as shown by the downward-sloping regression lines).

Takeaway

Topics seem objectively interpretable in many contexts.

Tradeoff between goodness-of-fit and interpretablility, which is generally more important in social science.

Active area of research assessing LDA models in terms of topic coherence.

Newman et. al. (2010)[9] propose a method based on mutual pointwise information between top words in topics as computed via co-occurrence in Wikipedia.

[9] ACL, "Automatic Evaluation of Topic Coherence".

Dictionary Methods + LDA

The terms in dictionaries come labeled, so can be seen as a type of supervised approach to information retrieval.

One can combine dictionary methods with the output of LDA to weight words counts by topic.

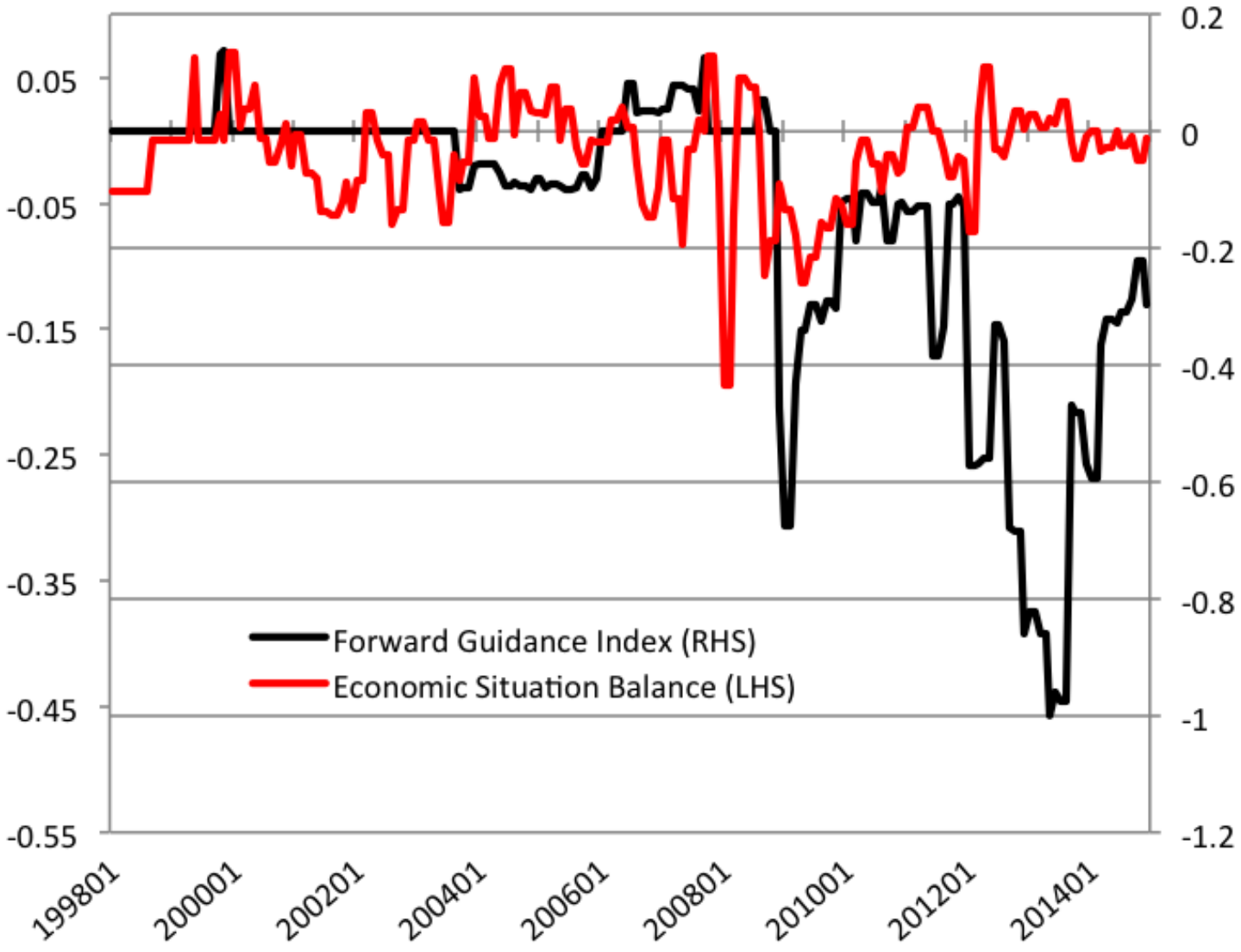
Recent application to minutes of the Federal Reserve to extract index of economic situation and forward guidance.

First step is to run 15-topic model and identify two separate kinds of topic.

Monetary Measures of Tone

Contraction	**Expansion**
decreas*	increas*
decelerat*	accelerat*
slow*	fast*
weak*	strong*
low*	high*
loss*	gain*
contract*	expand*

Indices


0.05
-0.05
-0.15
-0.25
-0.35
-0.45
-0.55
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
199801
200001
200201
200401
200601
200801
201001
201201
201401
Forward Guidance Index (RHS)
Economic Situation Balance (LHS)

Extensions

One of the main advantages LDA is that is provides a basis for more complex probabilistic models for text.

Some of the best known of these drop the Dirichlet prior on term or topic distributions and replace it with the logistic normal, which makes introducing structure easier.

Examples:

1. Dynamic topic model (Blei and Lafferty 2006[10]).
2. Structural topic model (Roberts et. al. 2016 [11]).

[10] ICML, "Dynamic Topic Models"
[11] JASA, "A Model of Text for Experiments in the Social Sciences"

Dynamic Topic Model

In LDA, the count of terms within documents are independent and identically distributed given the Dirichlet prior on θ_d.

This rules out all dependencies across texts, which is an unnatural assumption in many social science models (more on this later).

The dynamic topic model introduces time-series dependencies into the data generating process: each time period has a separate topic model, and time periods are linked via smoothly evolving parameters.

These dependencies can be present in the distributions over terms or in the prior distribution from which document-topic distributions are drawn.

Data Generating Process

Suppose all documents carry a time stamp t (daily, monthly, quarterly, yearly).

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Generate parameter vector $b_{t,k} \in \mathbb{R}^V$ where $b_{t,k} \sim \mathcal{N}(b_{t-1,k}, \sigma^2 I_V)$.
3. Form topics according to $\beta_{t,k,v} = \frac{\exp(b_{t,k,v})}{\sum_v \exp(b_{t,k,v})}$.
4. Each word $w_{d,n}$ in document d is generated from a two-step process:
   - 4.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   - 4.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.

This model captures changes in the words used when discussing topics; topical content of documents remains iid.

Example with Journal Abstracts

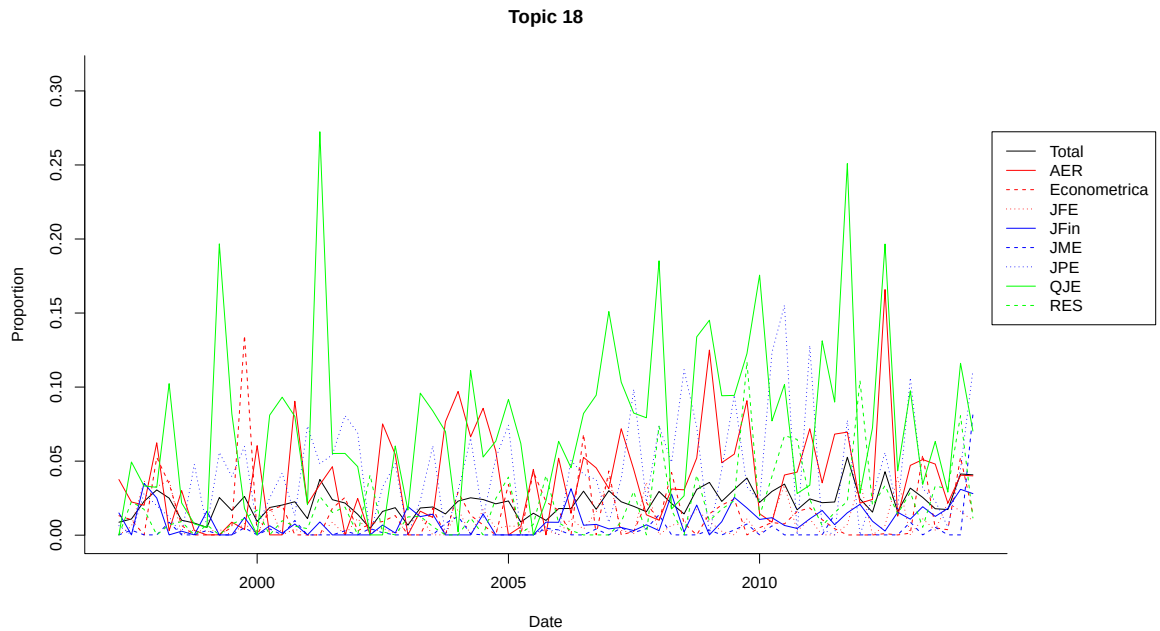
The following results are for a 20-topic version of DTM I on the abstracts of eight top economics journals from 1997:II to 2014:II (thanks to Julian Ashwin for collecting data and estimating):

- The Quarterly Journal of Economics
- Journal of Political Economy
- American Economic Review
- Econometrica
- Journal of Financial Economics
- Journal of Finance
- Review of Economic Studies
- Journal of Monetary Economics

Example with Journal Abstracts

Education economics topic

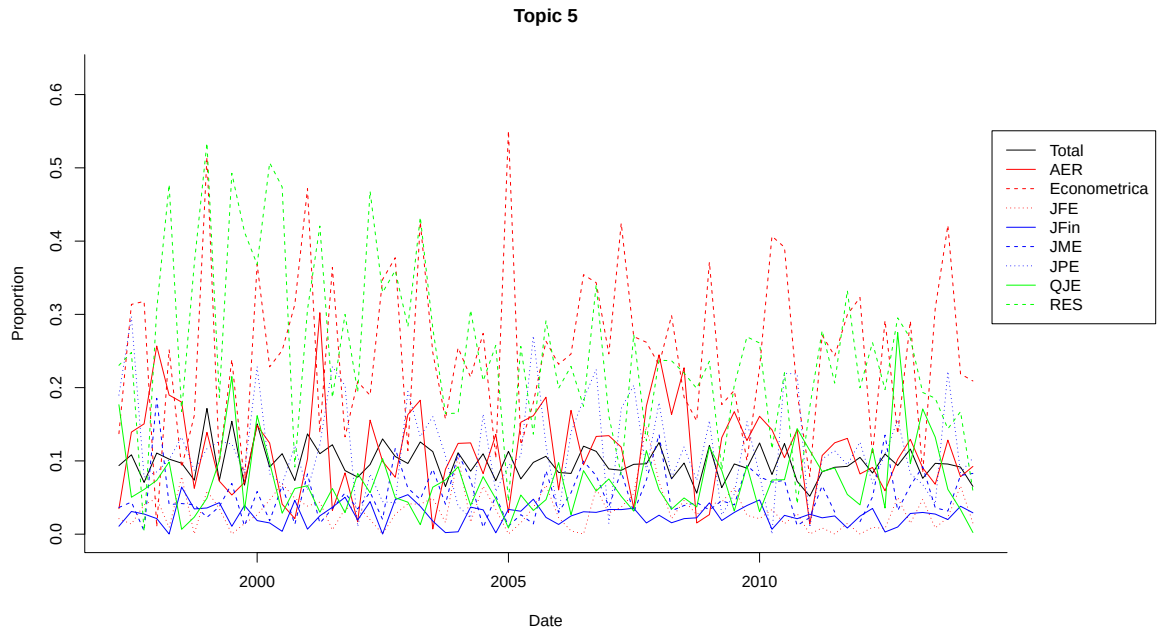
school	school	school	school	school
educ	educ	educ	student	group
network	network	famili	network	network
program	famili	group	group	student
student	student	student	famili	peer
famili	program	network	educ	social
score	colleg	program	social	educ
year	class	black	peer	year
black	group	score	score	score
class	year	white	year	famili



Example with Journal Abstracts

Game Theory topic

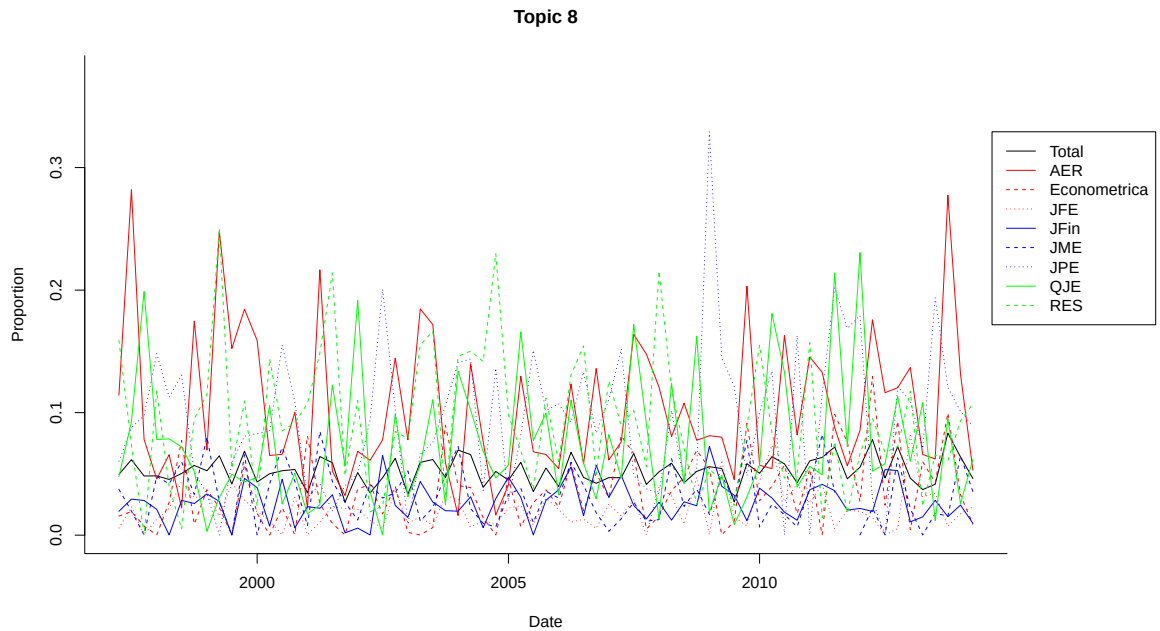
equilibrium	inform	agent	inform	inform
agent	equilibrium	equilibrium	equilibrium	agent
inform	agent	inform	agent	equilibrium
contract	game	game	game	game
game	contract	optim	contract	privat
auction	optim	contract	optim	optim
optim	auction	auction	effici	contract
may	equilibria	equilibria	auction	equilibria
equilibria	effici	effici	set	player
effici	player	privat	player	effici



Example with Journal Abstracts

Industrial Organisation topic

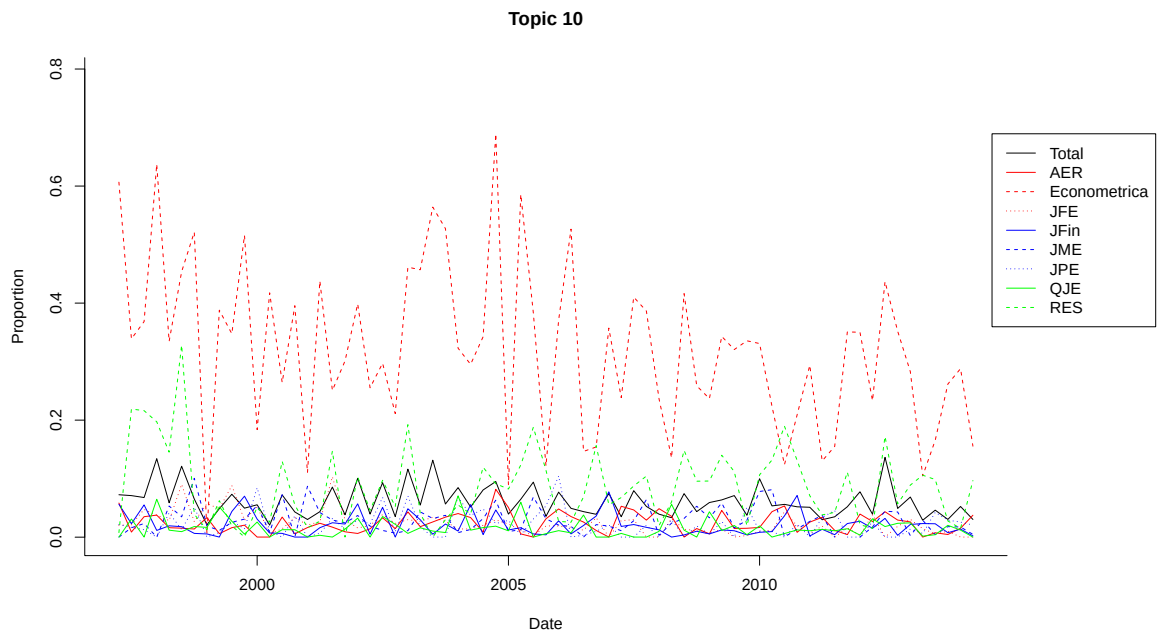
cost	product	product	cost	product
product	cost	cost	product	cost
competit	competit	competit	competit	competit
incent	good	consum	consum	good
effici	industri	good	industri	demand
contract	demand	industri	demand	consum
demand	profit	demand	entri	entri
may	qualiti	qualiti	welfar	qualiti
effort	effort	entri	qualiti	export
welfar	consum	welfar	good	industri



Example with Journal Abstracts

Econometric Theory topic

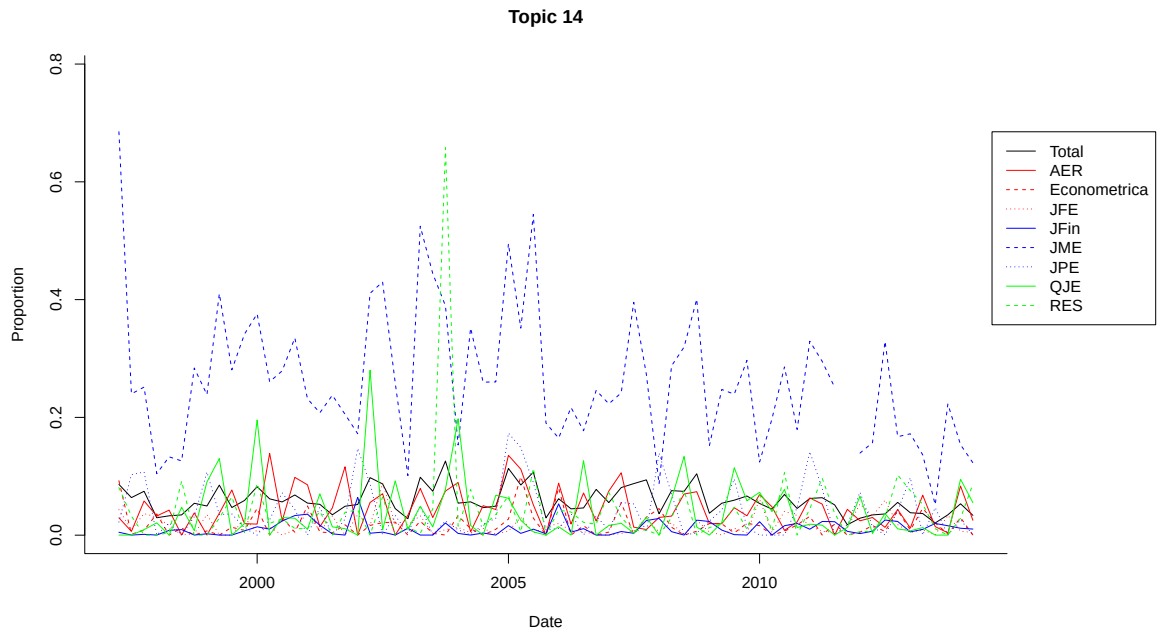
estim	estim	estim	estim	estim
test	distribut	method	distribut	method
distribut	function	distribut	function	distribut
paramet	test	function	paramet	condit
condit	paramet	condit	condit	test
asymptot	method	test	method	paramet
method	asymptot	paramet	identif	function
sampl	condit	asymptot	test	identif
function	sampl	variabl	asymptot	asymptot
regress	consist	consist	variabl	propos



Example with Journal Abstracts

Monetary policy topic

polici	polici	polici	polici	polici
monetari	rate	monetari	monetari	rate
rate	monetari	inflat	inflat	monetari
inflat	inflat	rate	rate	exchang
exchang	exchang	money	shock	inflat
money	interest	exchang	interest	nomin
currenc	money	shock	real	money
real	real	rule	nomin	shock
chang	scienc	interest	optim	interest
nomin	currenc	real	exchang	real



Structural Topic Model

A typical application of topic modeling in the social sciences first estimates LDA, then uses estimates of θ_d as the dependent variable in an regression on covariates to test whether different types of documents have different content.

This is contradictory because documents are assumed to be generated by a statistical process that we subsequently reject.

The structural topic model (STM) of Roberts et. al. (2016) explicitly introduces covariates into a topic model, and allows one to estimate the impact of document-level covariates on topic content and prevalence as part of the topic model itself.

There is a user-friendly R package stm for implementing it and additional information on `www.structuraltopicmodel.com`.

Topic Prevalence vs. Content

The process for generating individual words is the same as for plain LDA conditional on the $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_d$ terms.

However both objects can depend on potentially different sets of document-level covariates:

1. Topic Prevalence. Each document has P attributes $\mathbf{s}_d$ that affect the likelihood of discussing topic k.
2. Topic Content. Each document has an A-level categorical attribute t_d that affects the likelihood of discussing term v overall, and of discussing it within topic k.

The generation of the $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_d$ terms is via multinomial logistic regression, similar to the dynamic topic model.

Topic Prevalence Model

For each topic, draw $\gamma_k \sim \mathcal{N}_P(\mathbf{0}, \sigma_k^2 \mathbf{I}_P)$.

$\gamma_{k,p}$ is a coefficient that determines the effect of covariate p on the use of topic k. Its prior shrinks it towards 0 but does not induce sparsity.

$\boldsymbol{\mu}_d \in \mathbb{R}^{K-1}$ is the mean of the logistic normal from which $\boldsymbol{\theta}_d$ is drawn, where $\mu_{d,k} = \sum_p \gamma_{k,p} s_{d,p}$.

First draw $\boldsymbol{\eta}_d \sim \mathcal{N}_{K-1}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma})$, then form $\theta_{d,k} = \frac{\exp(\eta_{d,k})}{\sum_k \exp(\eta_{d,k})}$ where $\eta_{d,K}$ is normalized to 0.

When the covariates are constants, this becomes the correlated topic model (Blei and Lafferty 2007). The $\boldsymbol{\Sigma}$ matrix captures that some topics may be more or less likely to be discussed jointly.

Topic Content Model

$$\beta_{d,k,v} = \frac{\exp(m_v + \kappa'_{k,v} + \kappa''_{t_d,v} + \kappa'''_{t_d,k,v})}{\sum_v \exp(m_v + \kappa'_{k,v} + \kappa''_{t_d,v} + \kappa'''_{t_d,k,v})}$$

1. m_v is the baseline log-transformed rate of term v. Effect of covariates will be to generate deviations from these baseline frequencies.
2. κ' terms capture propensity of term v to appear in topic k across all documents.
3. κ'' terms capture propensity of covariates to generate term v.
4. κ''' terms capture propensity of covariates to generate term v in topic k.

Laplace priors placed on all κ terms to promote sparsity in the estimates.

STM vs LDA

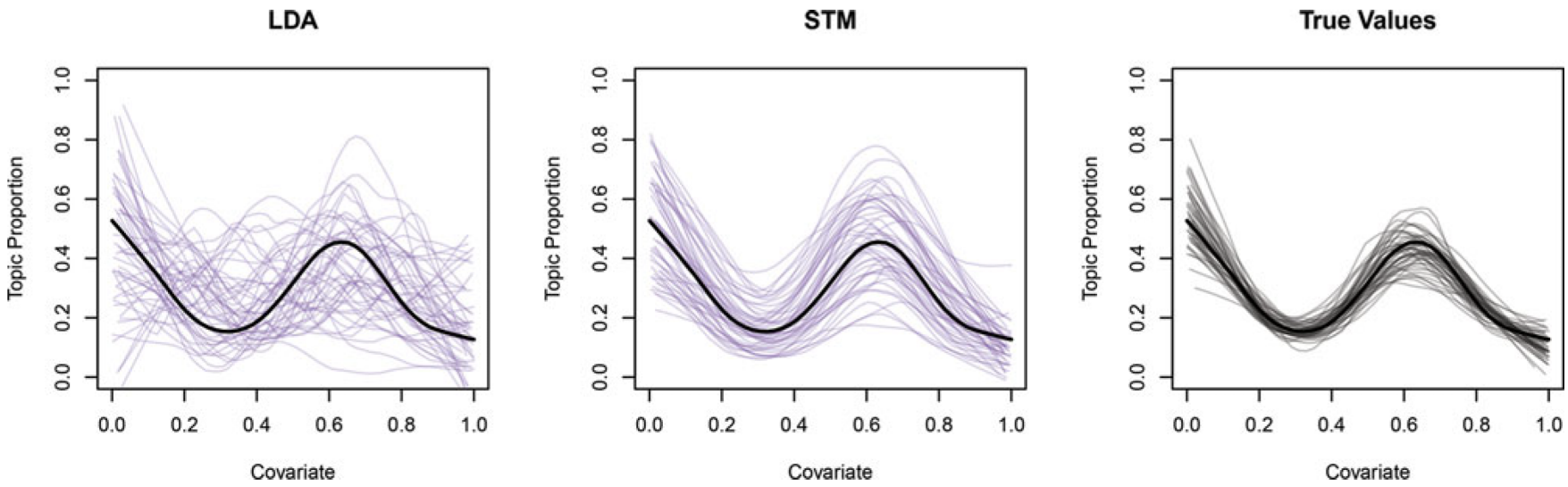


Figure 2. Plot of fitted covariate-topic relationships from 50 simulated datasets using LDA and the proposed structural topic model of text. The third panel shows t estimated relationship using the true values of the topic and thus only reflects sampling variability in the data-generating process.

Example

Gadarian and Albertson (2014)[12] conduct an experiment in which they cue a treatment group to worry about immigration and a control group to think about immigration, and record open-ended responses.

Estimate three-topic structural topic model on the corpus of survey responses.

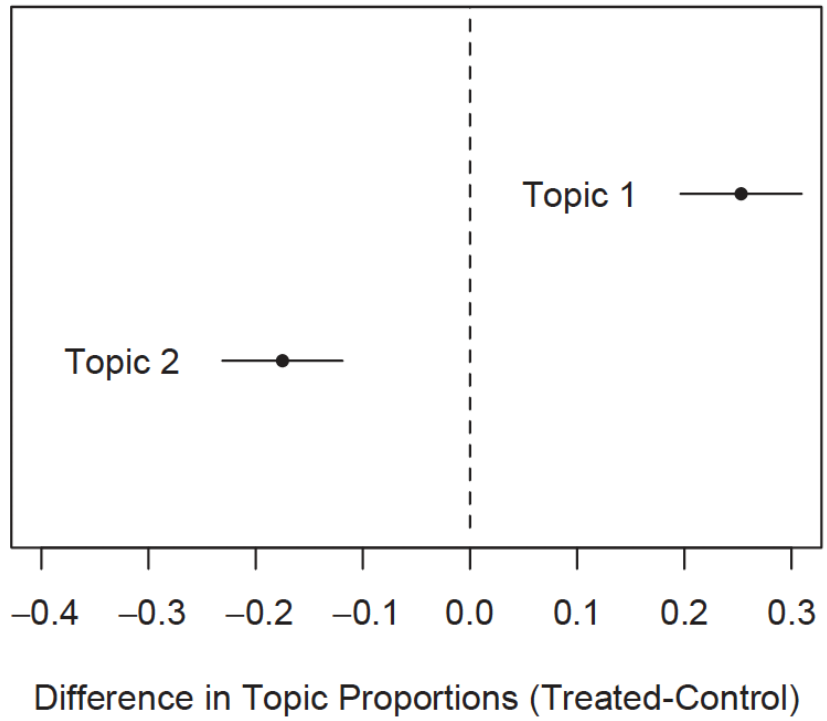
Covariates: (i) political party affiliation; (ii) treatment dummy; (iii) interaction effect.

[12] Political Psychology, "Anxiety, Immigration, and the Search for Information"

Topic Output

Topic 1:
illeg, job, immigr, tax, pai, american,
care, welfar, crime, system, secur,
social, cost, health, servic, school,
languag, take, us, free

Topic 2:
immigr, illeg, legal, border, need,
worri, mexico, think, countri, law,
mexican, make, america, worker, those,
american, fine, concern, long, fenc



Effect of Covariate