Text Mining for Economics and Finance
Latent Dirichlet Allocation

Stephen Hansen

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

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.

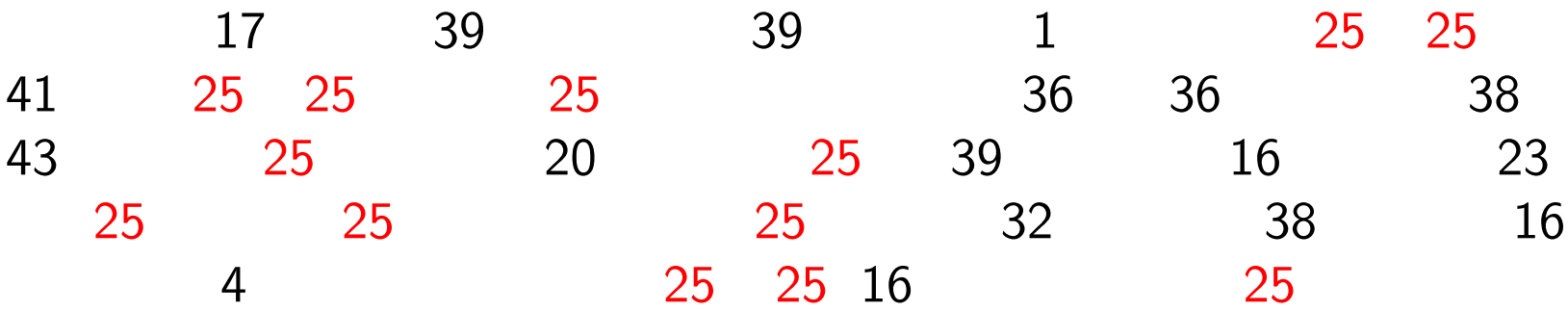
Example statement: Yellen, March 2006, #51

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

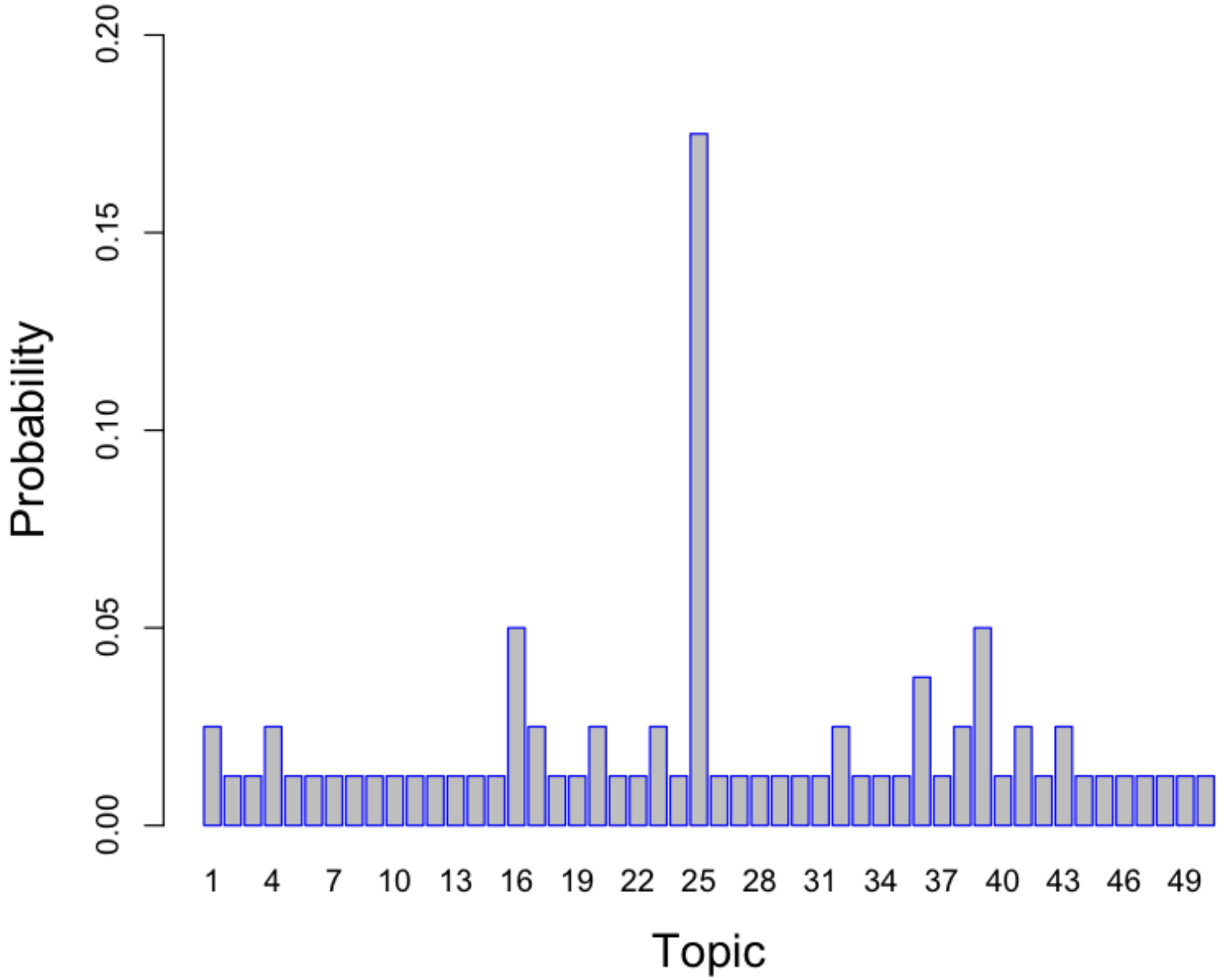
Example statement: Yellen, March 2006, #51

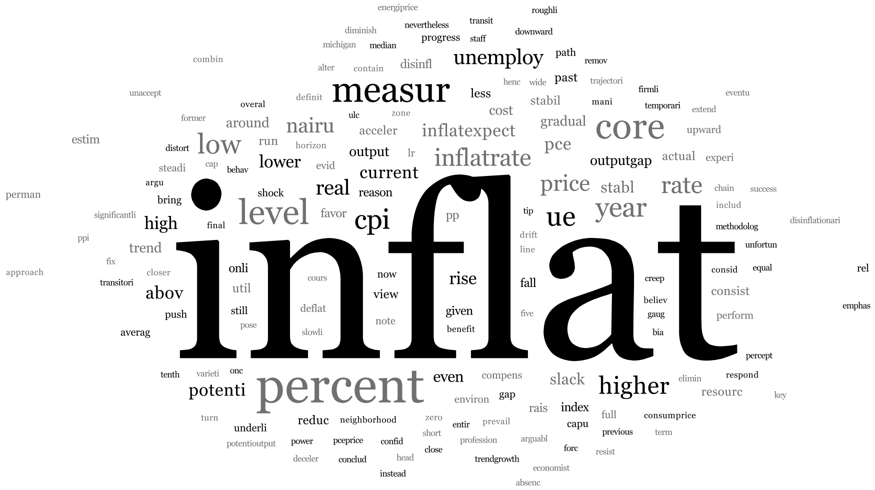
notic chang relationship between core CPI
chain core CPI suggest mayb someth go
relat substitut bia upper level index focus
nonmarket compon PCE wonder someth unusu
happen core CPI rel measur

Example statement: Yellen, March 2006, #51



Distribution of Attention

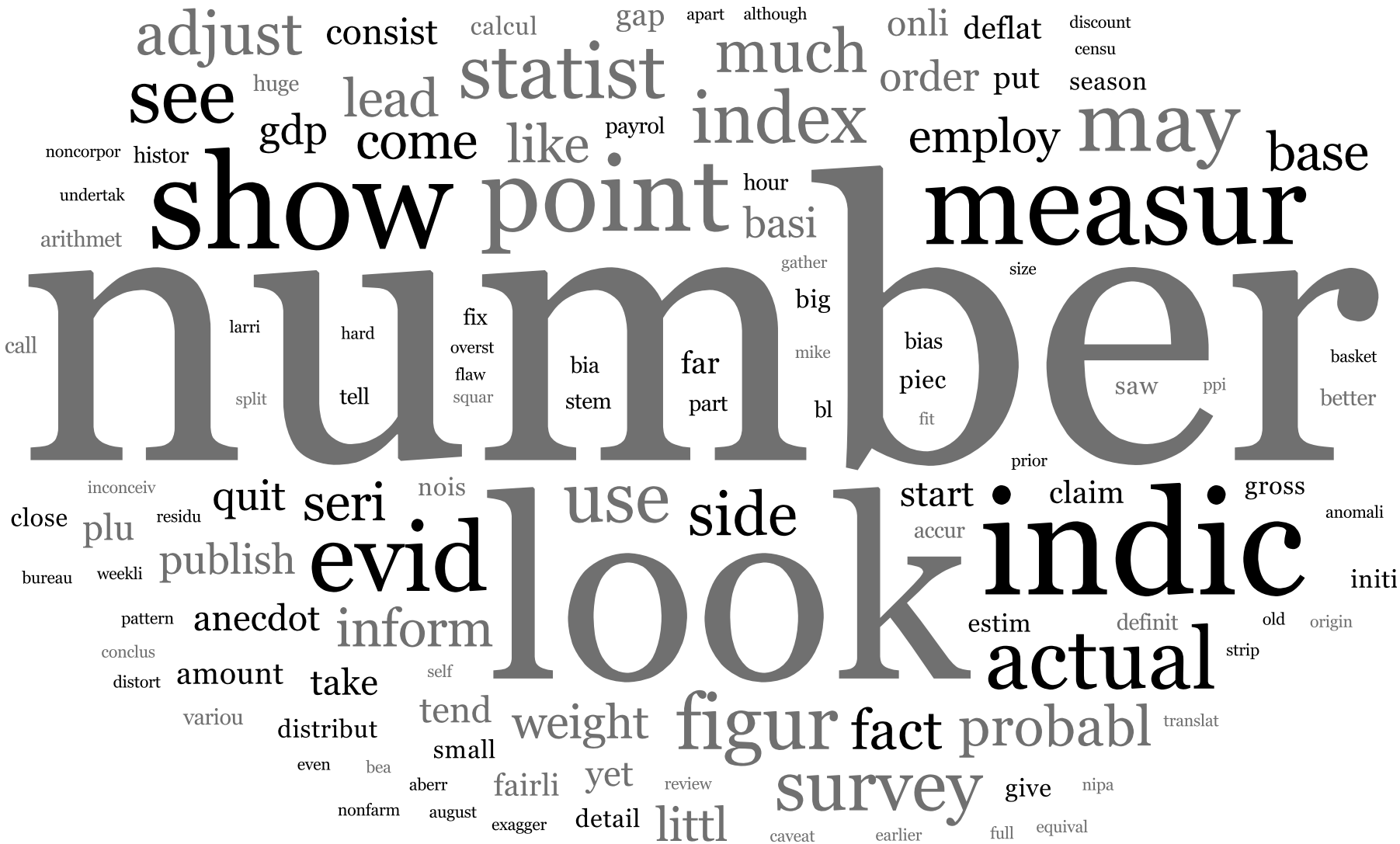



energiprice
roughli
transit
nevertheless
diminish
progress
staff
downward
michigan
median
unemploy
path
combin
alter
contain
disinfl
remov
measur
less
henc
wide
past
trajectori
unaccept
definit
firmli
eventu
overal
stabil
mani
temporari
former
ulc
zone
cost
extend
around
nairu
acceler
inflatexpect
gradual
core
upward
estim
distort
low
run
horizon
output
lr
pce
steadi
cap
behav
lower
evid
inflatrate
outputgap
actual
experi
argu
current
perman
shock
real
reason
price
stabl
rate
chain
success
bring
level
favor
cpi
pp
tip
ue
year
includ
significantli
high
final
methodolog
disinflationari
ppi
drift
trend
line
unfortun
fix
inflat
approach
closer
onli
cours
now
rise
fall
consid
equal
rel
transitori
creep
abov
util
view
consist
believ
emphas
push
still
deflat
given
five
gaug
perform
pose
note
averag
slowli
benefit
bia
percept
tenth
varieti
onc
even
compens
slack
higher
elimin
respond
potenti
percent
resourc
environ
gap
key
rais
index
turn
zero
prevail
full
consumprice
underli
reduc
neighborhood
capu
short
entir
previous
term
potentioutput
power
pceprice
confid
profession
arguabl
close
forc
resist
deceler
conclud
head
trendgrowth
instead
economist
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
base
noncorpor
histor
undertak
arithmet
show
point
hour
basi
measur
number
gather
size
big
call
larri
hard
fix
overst
flaw
squar
bia
stem
far
part
mike
bl
bias
piec
fit
saw
ppi
basket
better
split
tell
prior
inconceiv
quit
seri
nois
use
side
start
claim
gross
anomali
close
plu
residu
look
indic
accur
bureau
weekli
publish
evid
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
aberr
fairli
yet
review
survey
give
nipa
nonfarm
august
exagger
detail
littl
caveat
earlier
full
equival

Advantage of Flexibility

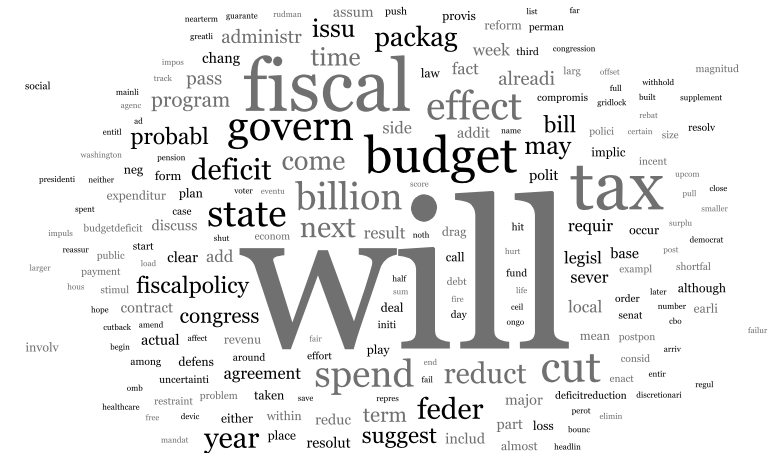
'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11.

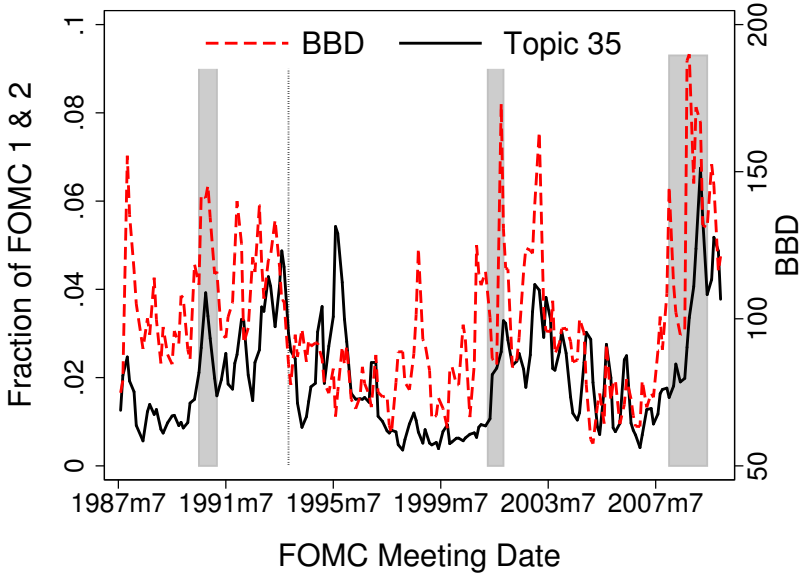
It gets assigned to 25 in this statement consistently due to the presence of other topic 25 words.

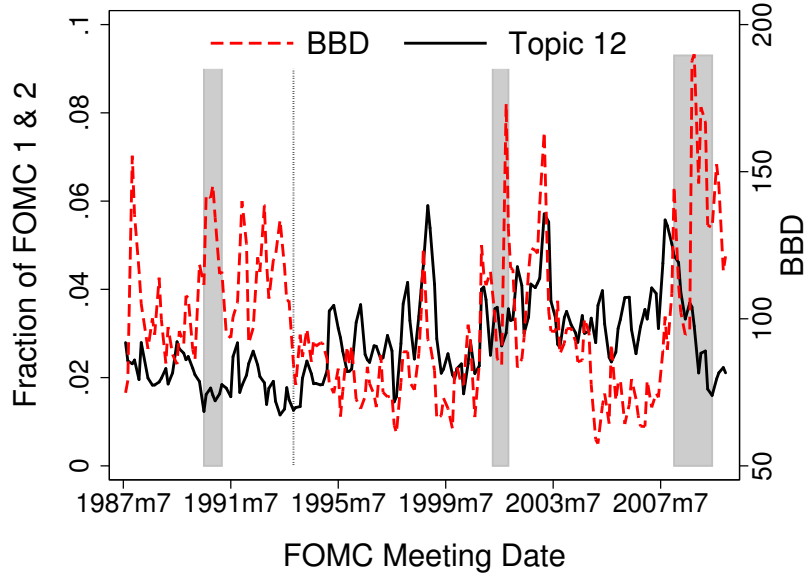
In statements containing words on evidence and numbers, it consistently gets assigned to 11.

Sampling algorithm can help place words in their appropriate context.

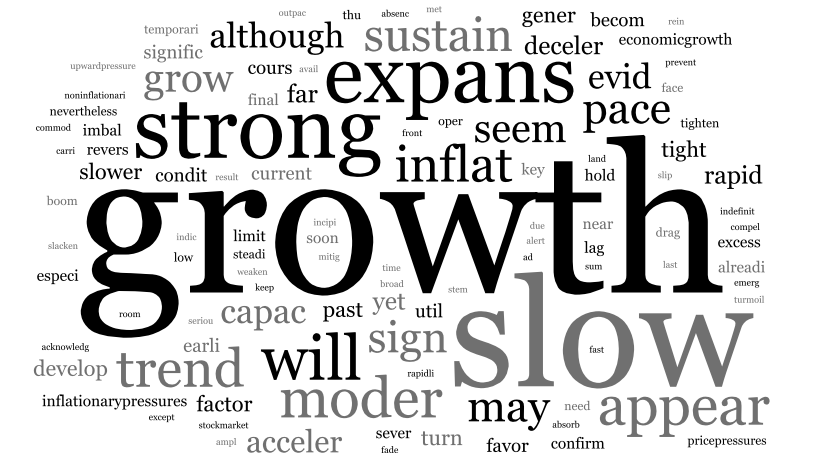
External Validation—BBD

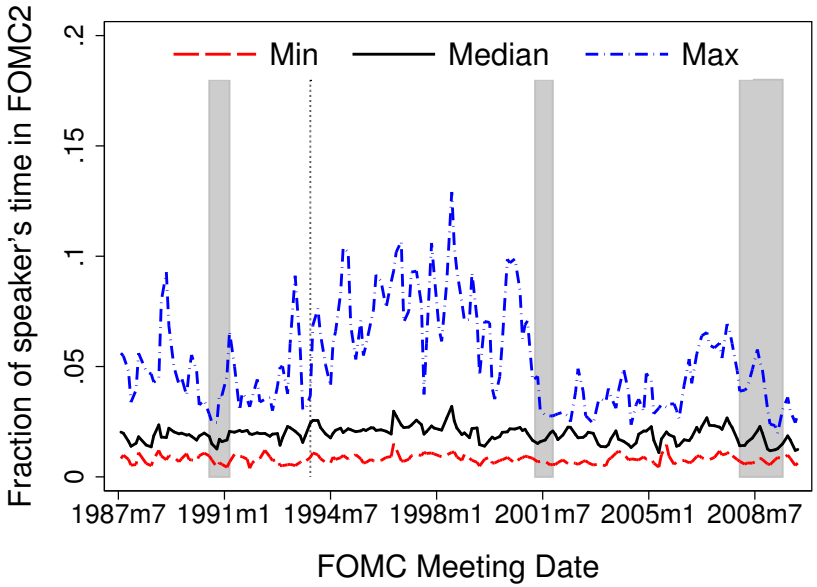


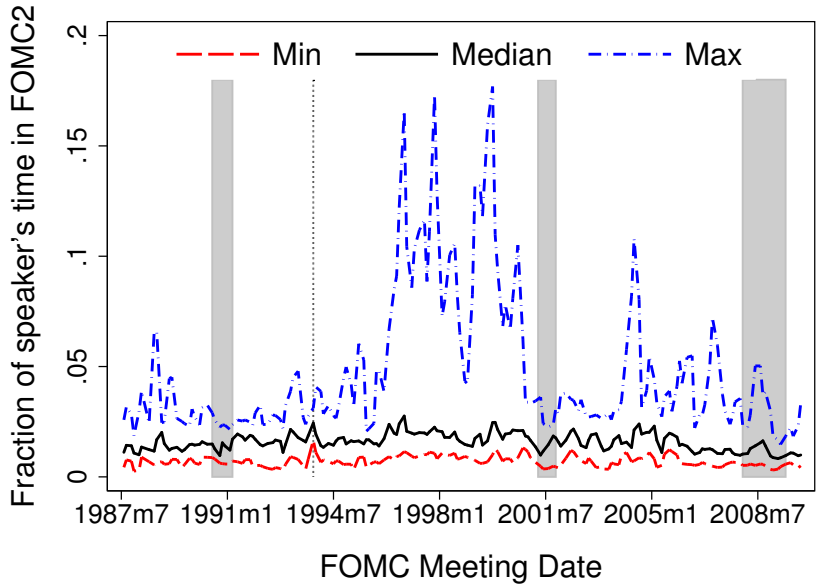




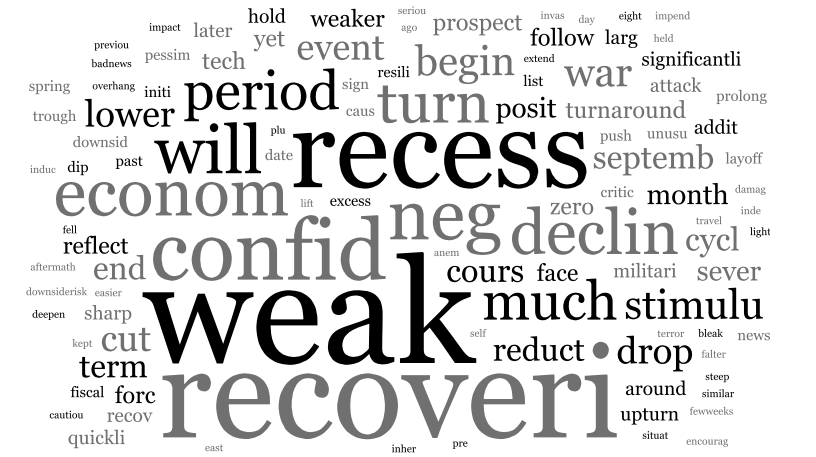
Pro-Cyclical Topics

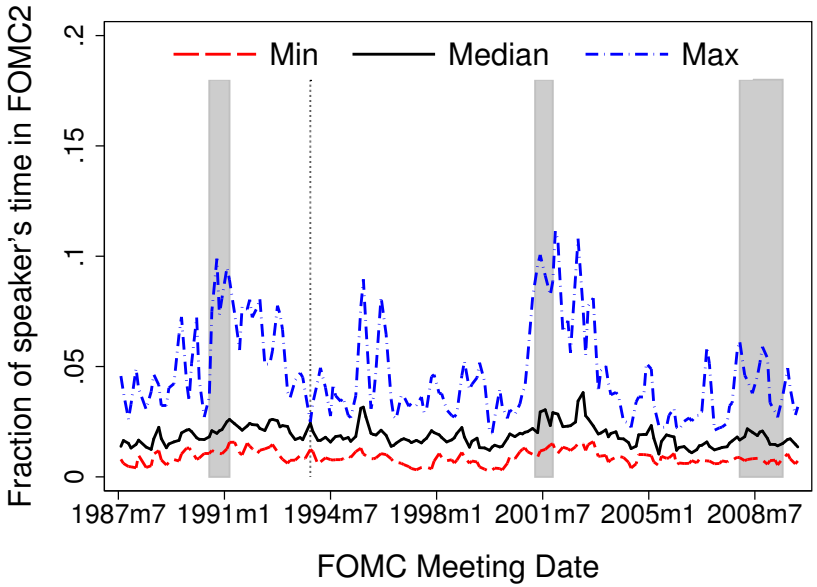


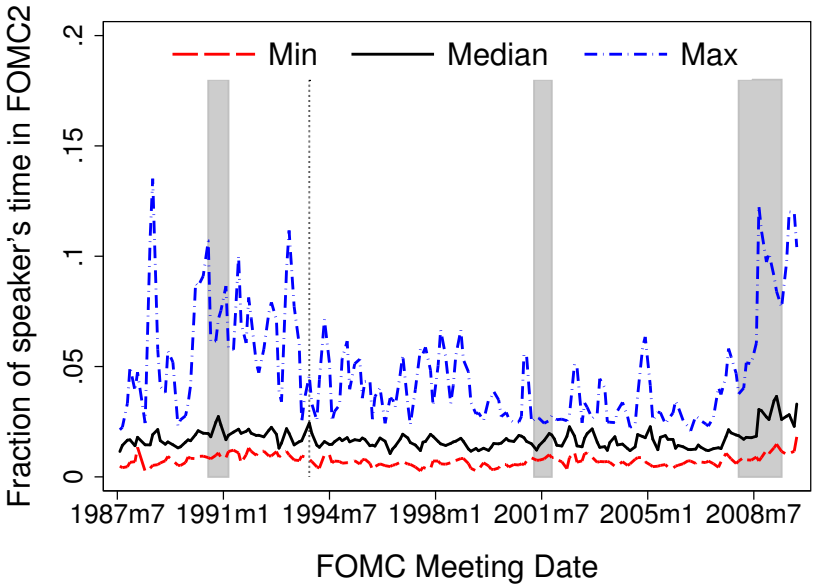




Counter-Cyclical Topics







Graphical Models

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

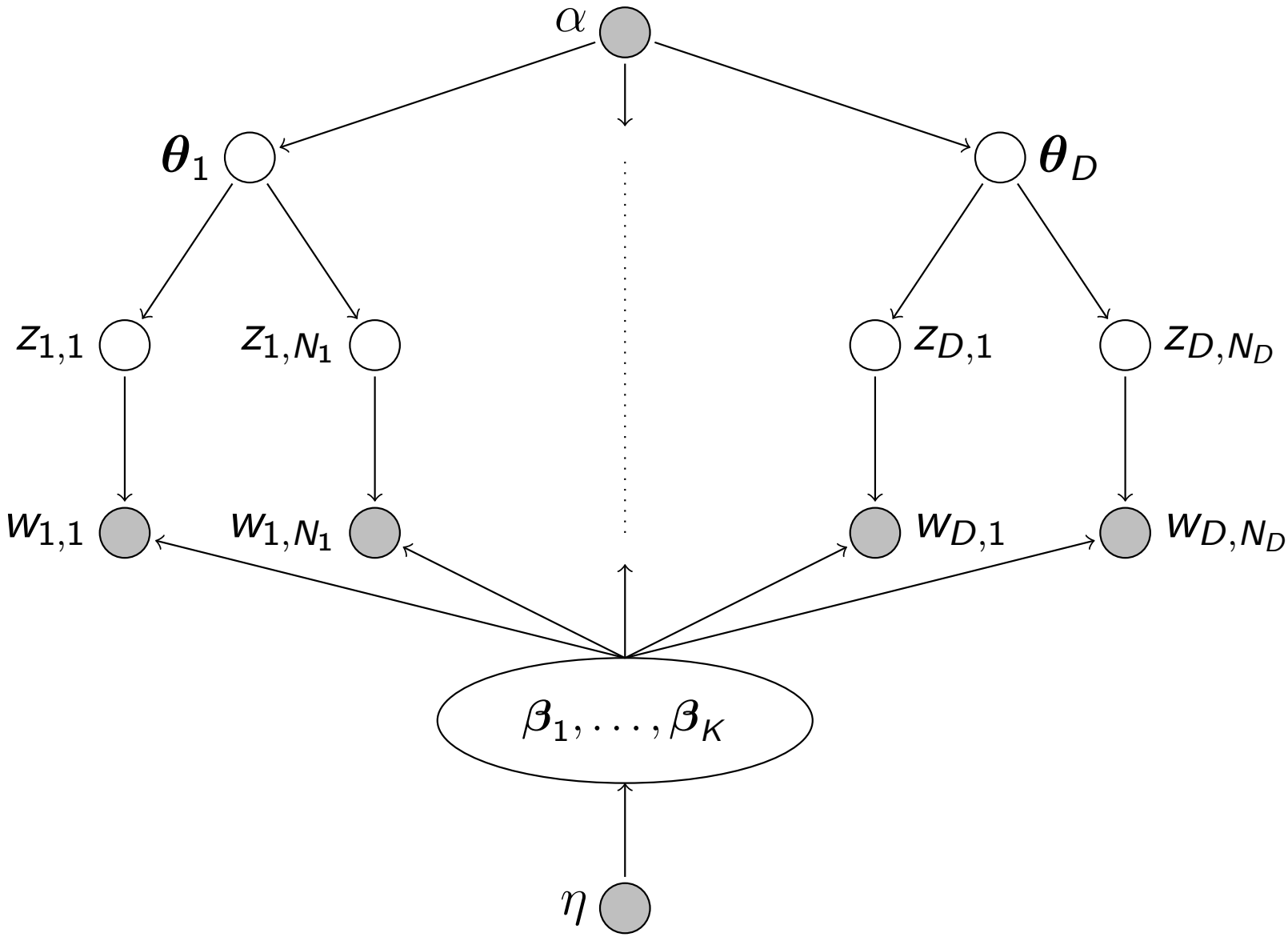
In high-dimensional models, it is useful to summarize relationships among random variables with directed graphs in which nodes represent random variables and links between nodes represent dependencies.

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.

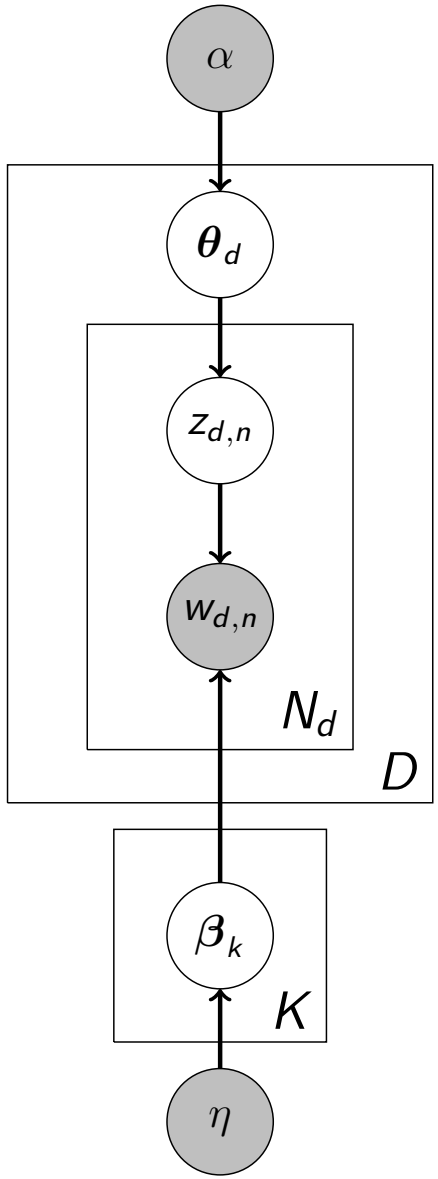
A *Bayesian network* is a probabilistic model whose joint distribution can be represented by a directed acyclic graph (DAG).

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

LDA as a Bayesian Network



LDA Plate Diagram



Graph Properties

Let $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and $\mathbf{T} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$.

Object	Parents	Children
α	$\emptyset$	$\mathbf{T}$
$\boldsymbol{\theta}_d$	α	$\mathbf{z}_d$
$z_{d,n}$	$\boldsymbol{\theta}_d$	$w_{d,n}$
$w_{d,n}$	$z_{d,n}, \mathbf{B}$	$\emptyset$
$\boldsymbol{\beta}_k$	η	$\mathbf{w}_1, \ldots, \mathbf{w}_D$
η	$\emptyset$	$\mathbf{B}$

Conditional Independence Property I

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

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

Applying this formula to LDA yields

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

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

Posterior Distribution

The inference problem in LDA is to compute the posterior distribution over $\mathbf{z}$, $\mathbf{T}$, and $\mathbf{B}$ 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

$$\Pr[\mathbf{z} = z' \mid \mathbf{w}, \mathbf{T}, \mathbf{B}] = \frac{\Pr[\mathbf{w} \mid \mathbf{z} = z', \mathbf{T}, \mathbf{B}] \Pr[\mathbf{z} = z' \mid \mathbf{T}, \mathbf{B}]}{\sum_{z'} \Pr[\mathbf{w} \mid \mathbf{z} = z', \mathbf{T}, \mathbf{B}] \Pr[\mathbf{z} = z' \mid \mathbf{T}, \mathbf{B}]}.$$

Posterior Distribution

The inference problem in LDA is to compute the posterior distribution over $\mathbf{z}$, $\mathbf{T}$, and $\mathbf{B}$ 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

$$\Pr[\mathbf{z} = z' \mid \mathbf{w}, \mathbf{T}, \mathbf{B}] = \frac{\Pr[\mathbf{w} \mid \mathbf{z} = z', \mathbf{T}, \mathbf{B}] \Pr[\mathbf{z} = z' \mid \mathbf{T}, \mathbf{B}]}{\sum_{z'} \Pr[\mathbf{w} \mid \mathbf{z} = z', \mathbf{T}, \mathbf{B}] \Pr[\mathbf{z} = z' \mid \mathbf{T}, \mathbf{B}]}.$$

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 $\approx$ 7.88e169 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

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

Suppose we can compute the conditional distribution $f_i \equiv f(x_i \mid \mathbf{x}_{-i})$.

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

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

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 θ_d

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

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

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

We computed this posterior above. Let $n_{d,k}$ be the number of words in document d that have topic allocation k.

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

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

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

- The parent η.
- The children $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$.
- The children's parents:
  1. $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$.
  2. $\mathbf{B}_{-k}$.

Consider $\Pr[\boldsymbol{\beta}_k \mid \eta, \mathbf{w}, \mathbf{z}]$. Only the allocation variables assigned to k—and their associated words—are informative about $\boldsymbol{\beta}_k$.

Let $m_{k,v}$ be the number of times topic k allocation variables generate term v.

Then $\Pr[\boldsymbol{\beta}_k \mid \eta, \mathbf{w}, \mathbf{z}] = \text{Dir}(\eta + m_{k,1}, \ldots, \eta + m_{k,V})$.

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 $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$.

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

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

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) proposed a collapsed sampler for LDA that integrates out the $\mathbf{T}$ and $\mathbf{B}$ terms and samples only $\mathbf{z}$.

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

Collapsed Sampling Equation for LDA

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

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

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:

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

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 $v_{d,n}$ 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

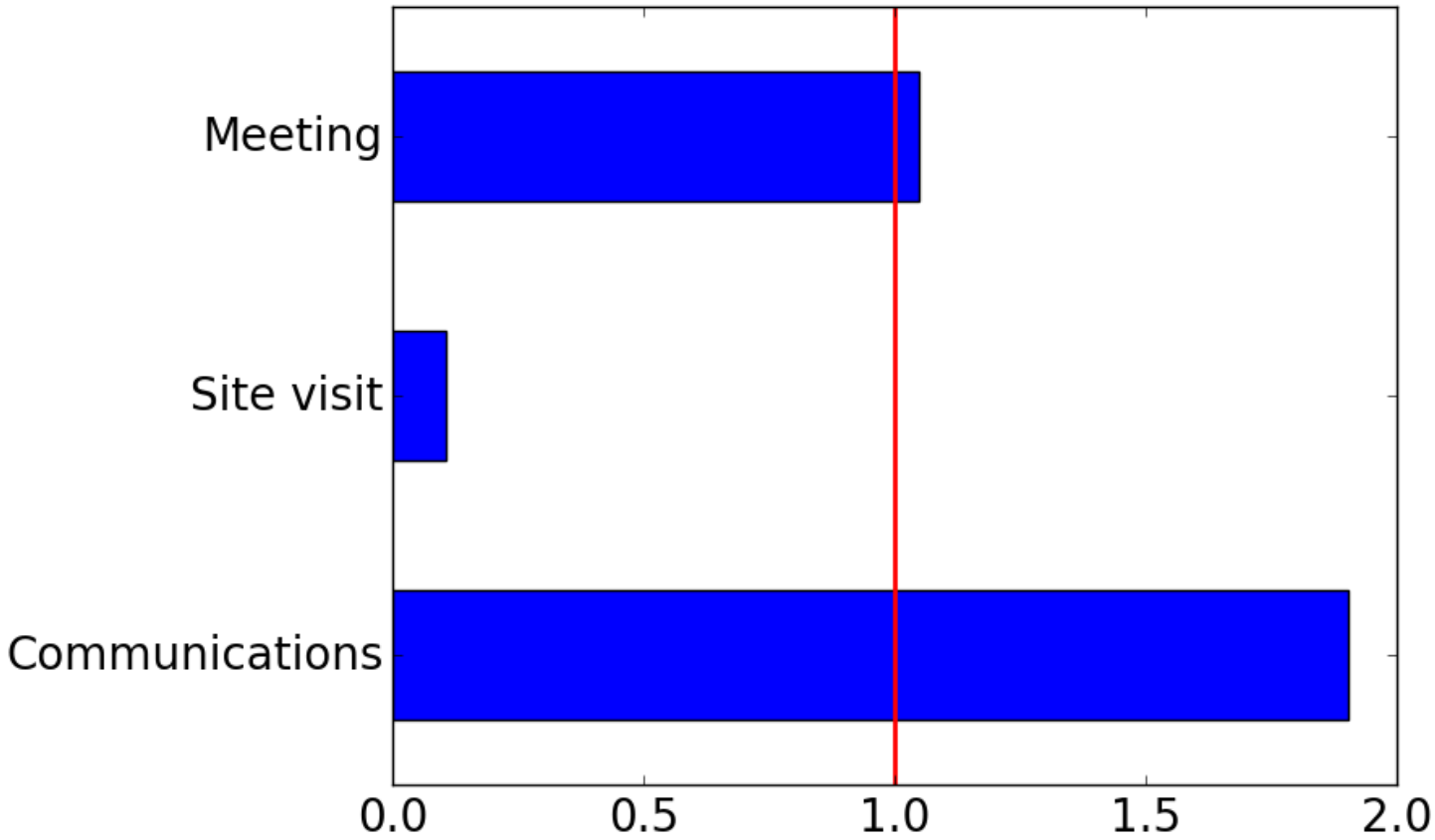
Recall from the first lecture that text data is one instance of count data.

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

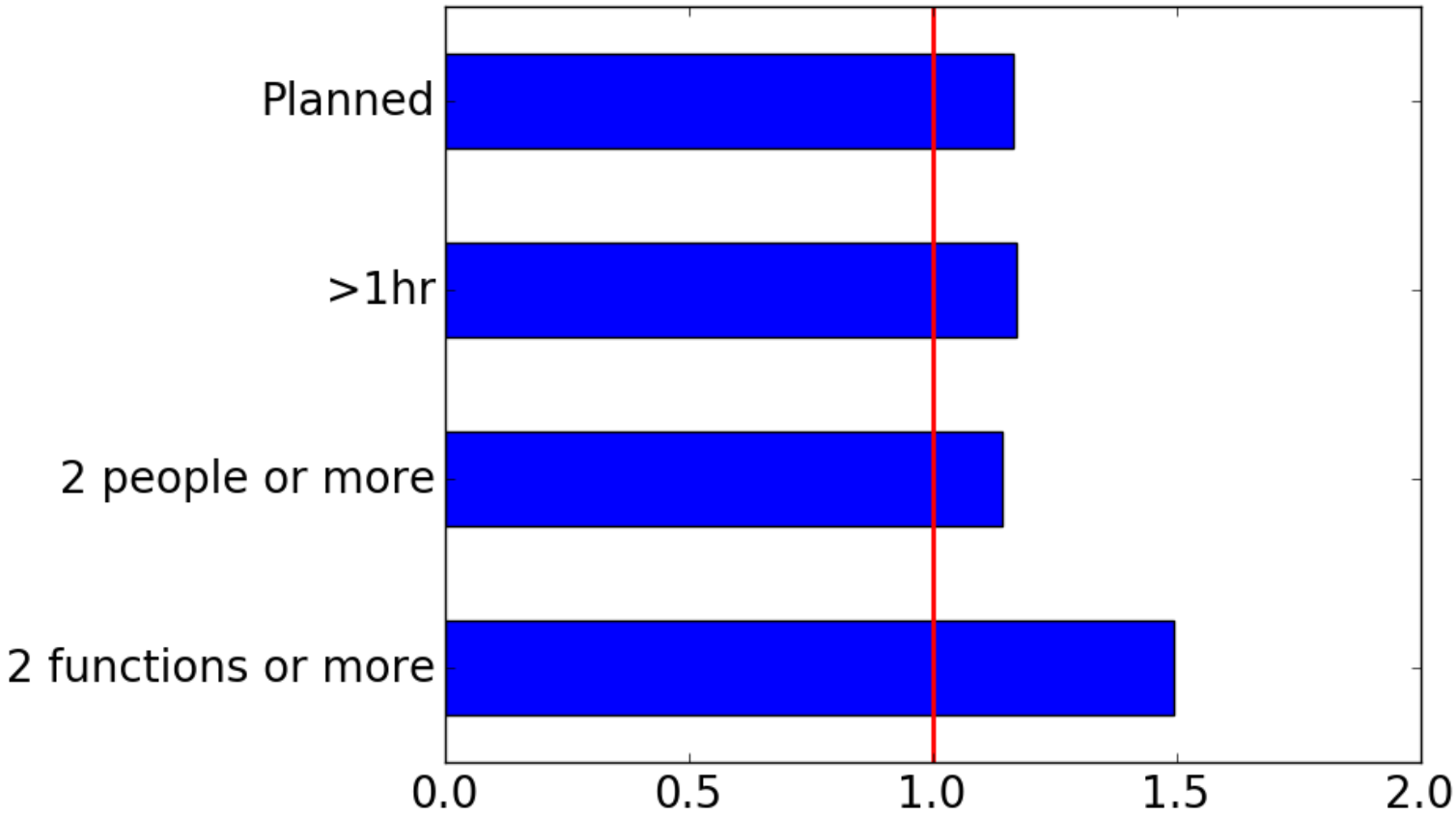
We recently applied it to CEO survey data to estimate management "behaviors" with $K = 2$.

Can visualize results in terms of likelihood ratios of marginals over specific data features.

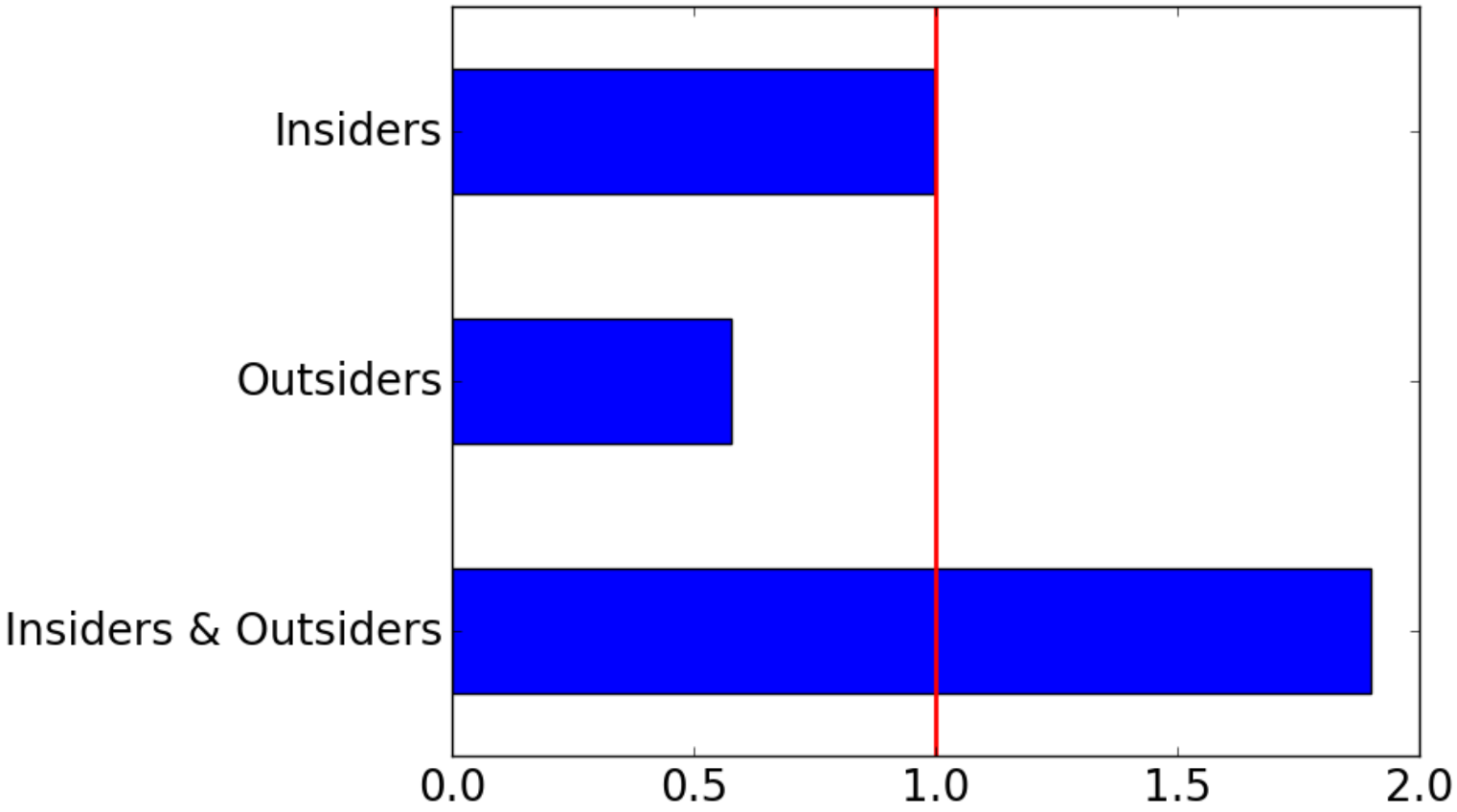
Activity Type



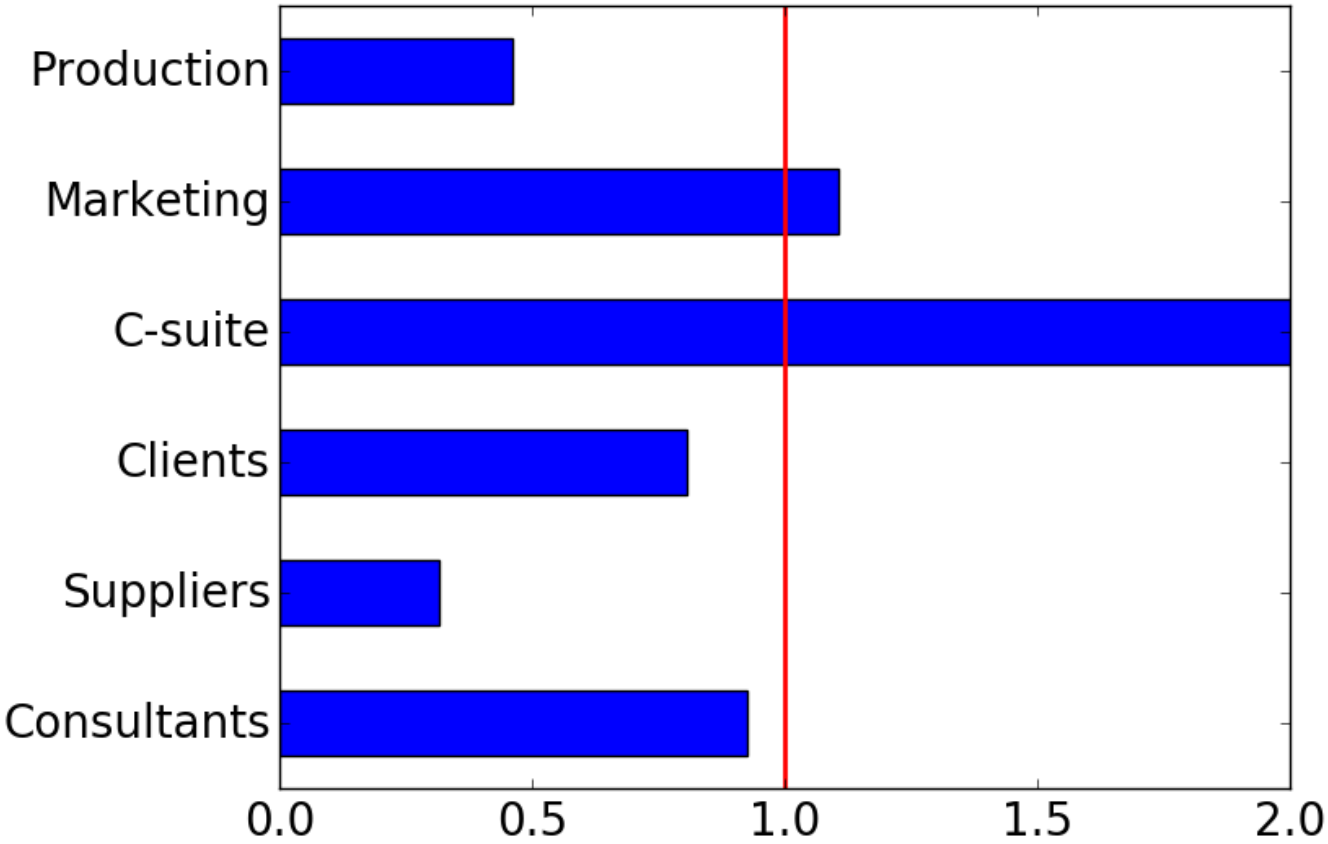
Planning; Size; Number of Functions



Insiders vs Outsiders



Functions


Production
Marketing
C-suite
Clients
Suppliers
Consultants
0.0
0.5
1.0
1.5
2.0

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).

K is less clear. Two potential goals:

1. Predict text well. Statistical criteria to select K.
2. Interpretability. General versus specific.

Formalizing Interpretablility

Chang et. al. (2009) propose an objective way of determining whether topics are indeed 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 $K = 50, 100, 150$.

Results

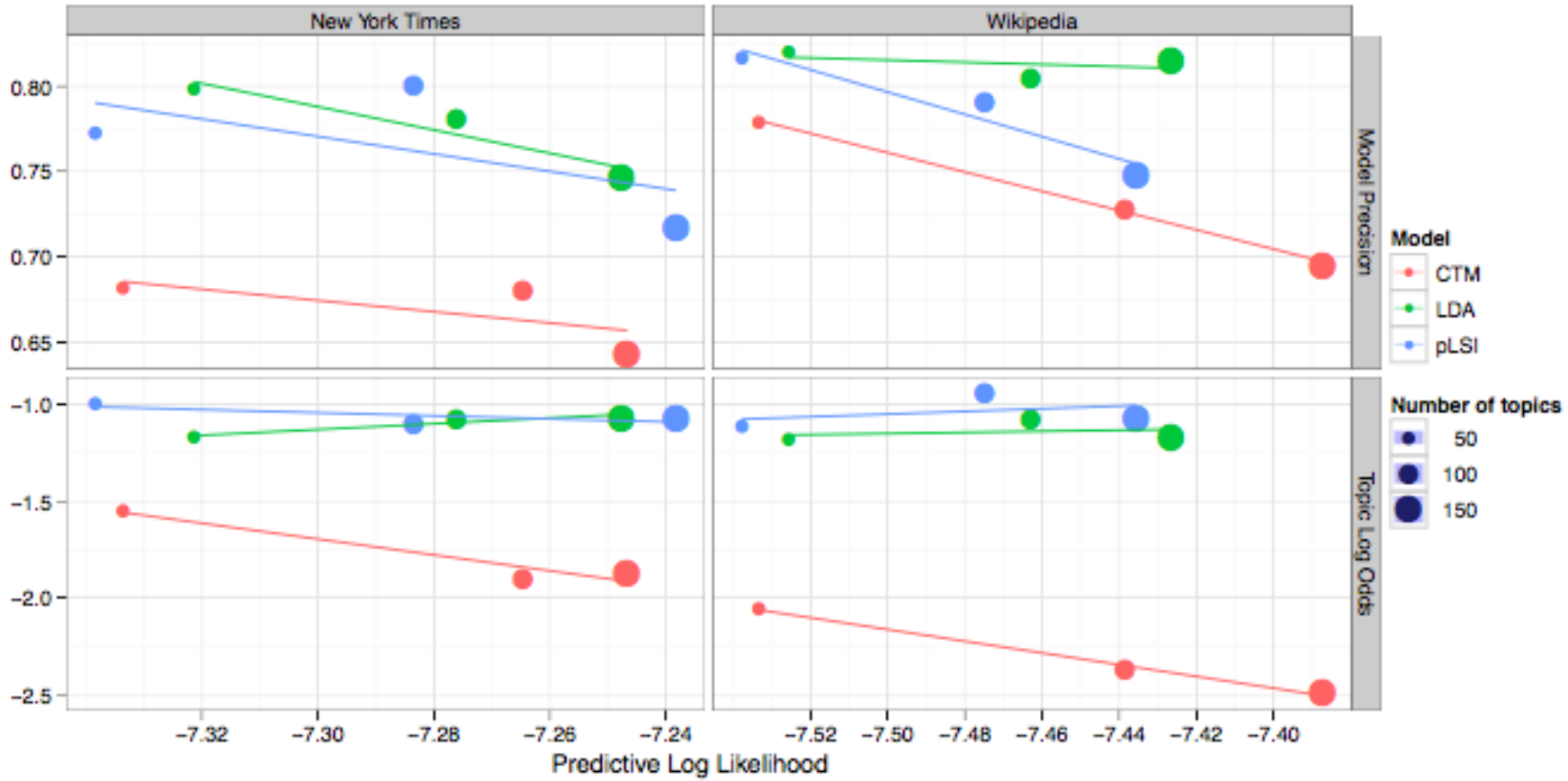


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase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

Implications

Topics seem objectively interpretable in many contexts.

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

Potential development of statistical models in future to explicitly maximize interpretablility.

Chain Convergence and Selection

Determining when a chain has converged can be tricky.

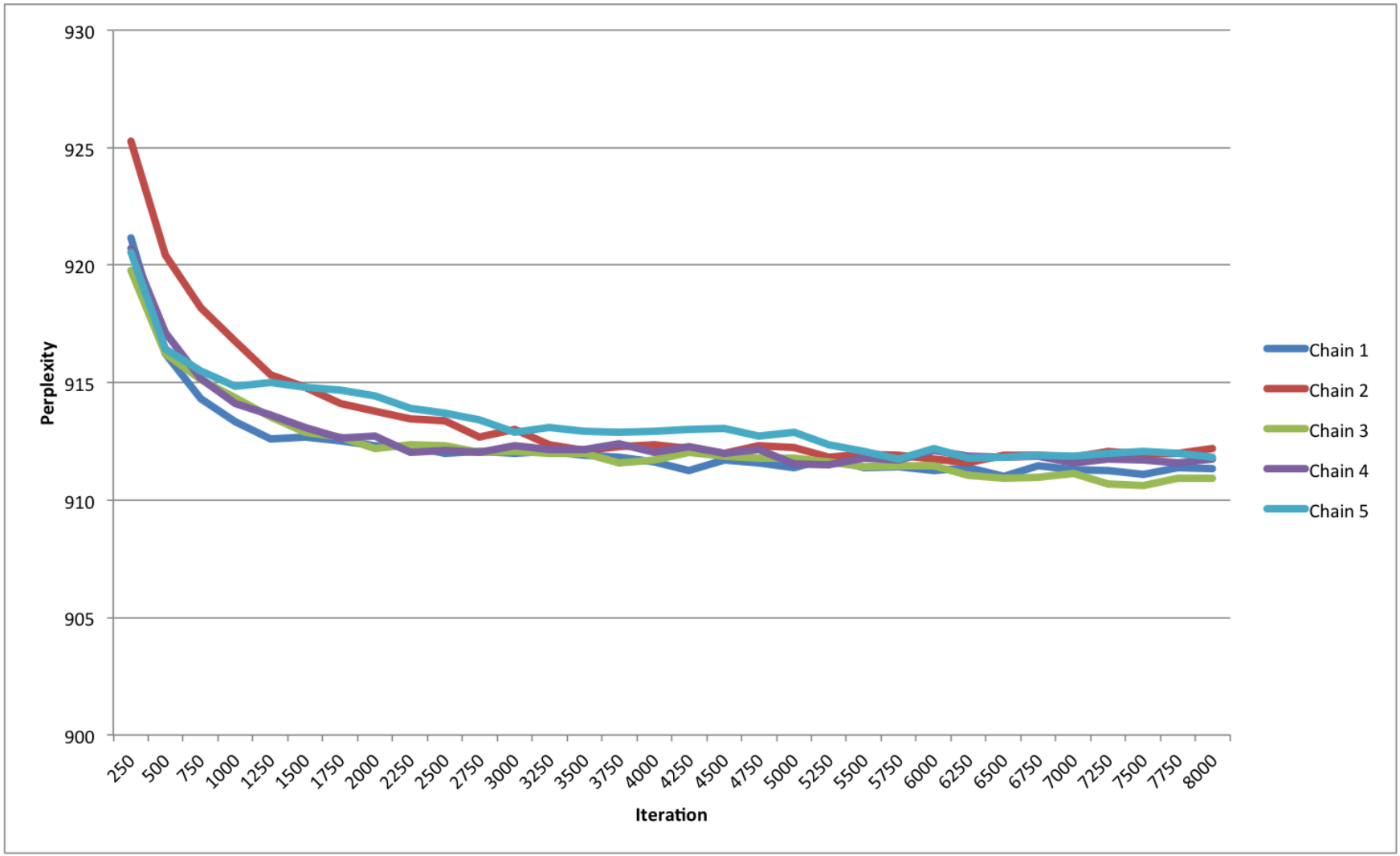
One approach is to measure how well different states of the chain predict the data, and determine convergence in terms of its stability.

Standard practice is run chains from different starting values, after which you can select the best-fit chain for analysis.

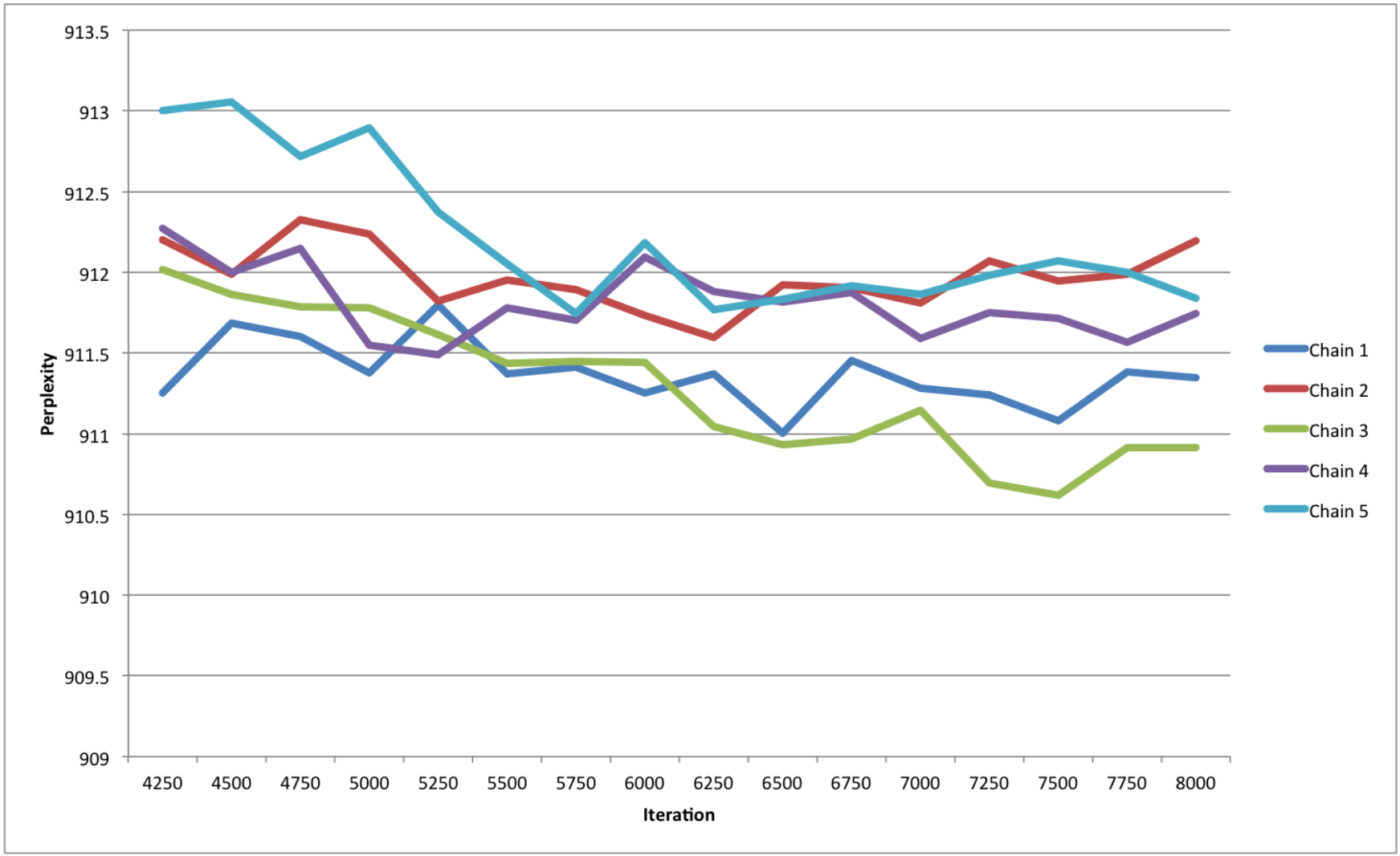
For LDA, a typical goodness-of-fit measure is *perplexity*, given by

$$\exp\left[-\frac{\sum_{d=1}^{D}\sum_{v=1}^{V} x_{d,v}\log\left(\sum_{k=1}^{K}\widehat{\theta}_{d,k}\widehat{\beta}_{k,v}\right)}{\sum_{d=1}^{D} N_d}\right].$$

Perplexity 1



Perplexity 2



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.

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.

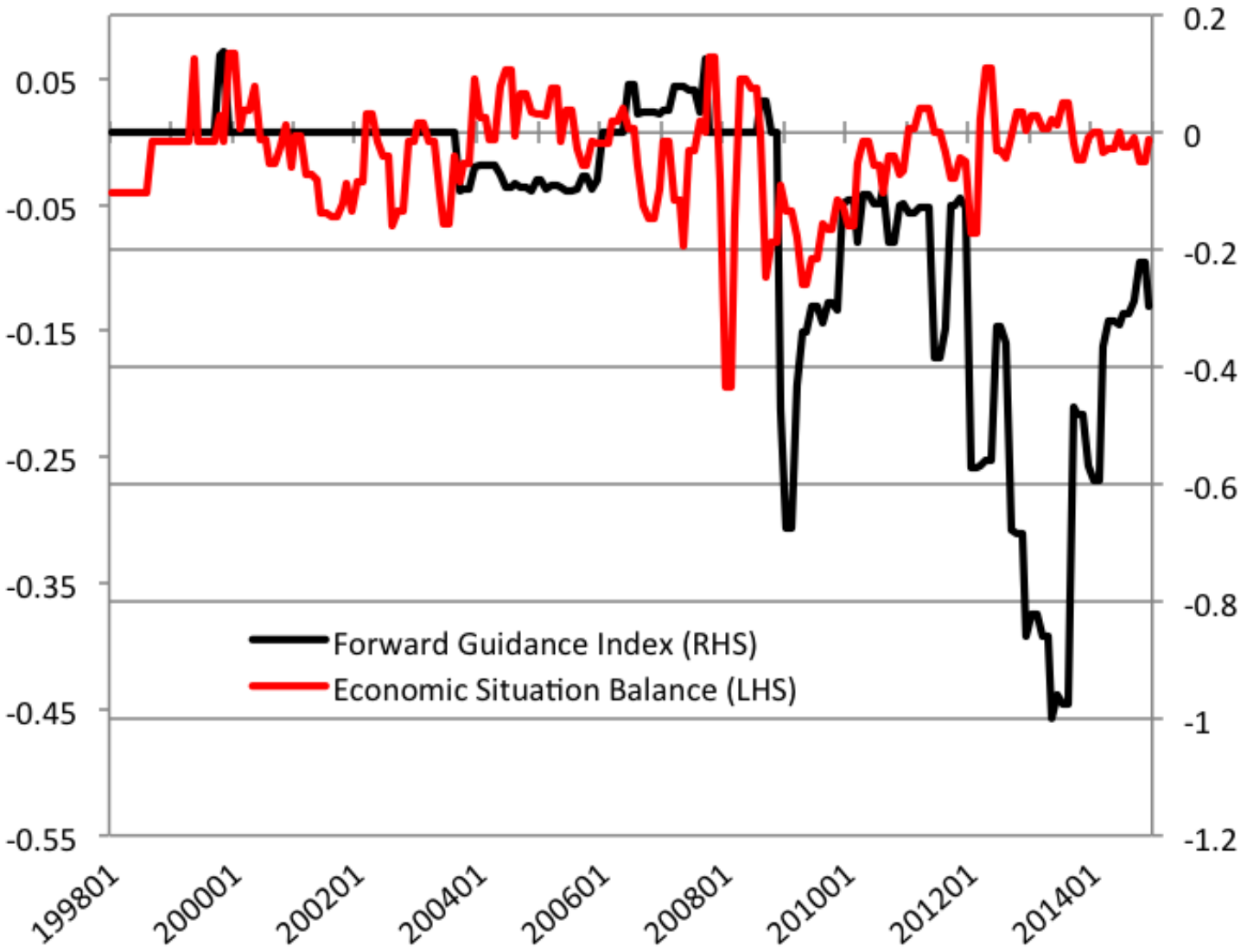
First step is to run 15-topic model and identify two separate kinds of topic.

Example 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
Forward Guidance Index (RHS)
Economic Situation Balance (LHS)
199801
200001
200201
200401
200601
200801
201001
201201
201401

Conclusion

Key ideas from this lecture:

1. Latent Dirichlet Allocation. Influential in computer science, many potential applications in economics.
2. Graphical models to simplify and visualize complex joint likelihoods.
3. Gibbs sampling to stochastically approximate posterior.
4. Interpretable output in several domains.

Topic for future: incorporate economic structure into LDA.