Topics in Empirical Economics, Part III
Bayesian Inference & Latent Dirichlet Allocation

Stephen Hansen
Universitat Pompeu Fabra

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.

Outline

Bayesian models of discrete data

LDA model specification

Graphical models / Bayesian networks

Markov chain Monte Carlo estimation of LDA

Applications in economics

Review

Recall the simple unigram model of a document in which

$$\Pr\left[\mathbf{x}_d \mid \boldsymbol{\beta}\right] = \prod_v \beta_v^{x_{d,v}}.$$

The maximum likelihood estimate for the vth categorical probability is $\widehat{\beta}_v = \frac{x_{d,v}}{N_d}$

To maximize the probability of the observed data, we get parameters that exactly match the observed frequencies.

Black Swan Paradox

What would the model predict is the probability of seeing an unobserved term?

This is sometimes called the *black swan paradox*. Europeans assumed that the fact that they had never observed a black swan implied black swans could not exist.

Since the document-term matrix is sparse, the black swan paradox will be particularly relevant for text mining.

Bottom line is we need to incorporate some additional uncertainty in our inference procedure to not drive beliefs about unobserved events to zero.

Bayesian Inference

One solution is to adopt a Bayesian inference approach, which treats β as a random variable rather than a fixed parameter.

Recall that Bayes' rule states that

$$\Pr[\beta \mid \mathbf{x}_d] = \frac{\Pr[\mathbf{x}_d \mid \beta]\Pr[\beta]}{\Pr[\mathbf{x}_d]}$$

where

- $\Pr[\beta \mid \mathbf{x}_d]$ is the posterior distribution.
- $\Pr[\mathbf{x}_d \mid \beta]$ is the likelihood function.
- $\Pr[\beta]$ is the prior distribution on the parameter vector.
- $\Pr[\mathbf{x}_d]$ is a normalizing constant sometimes called the evidence.

The prior distribution introduces initial uncertainty about the value of the parameter vector.

Dirichlet Prior

One way of ensuring Bayesian inference is tractable is to select a prior distribution from a family that ensures the posterior will be in the same family given the likelihood function. This is called a *conjugate* prior.

The Dirichlet distribution is conjugate to the categorical likelihood function, and so is a popular choice for the prior in Bayesian models of discrete data.

The Dirichlet distribution is parametrized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_V)$; is defined on the $V-1$ simplex; and has probability density function

$$\text{Dir}(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \propto \prod_v \beta_v^{\alpha_v - 1}.$$

The normalization constant is $B(\boldsymbol{\alpha}) \equiv \prod_{v=1}^{V} \Gamma(\alpha_v) / \Gamma\left(\sum_{v=1}^{V} \alpha_v\right)$.

Interpreting the Dirichlet

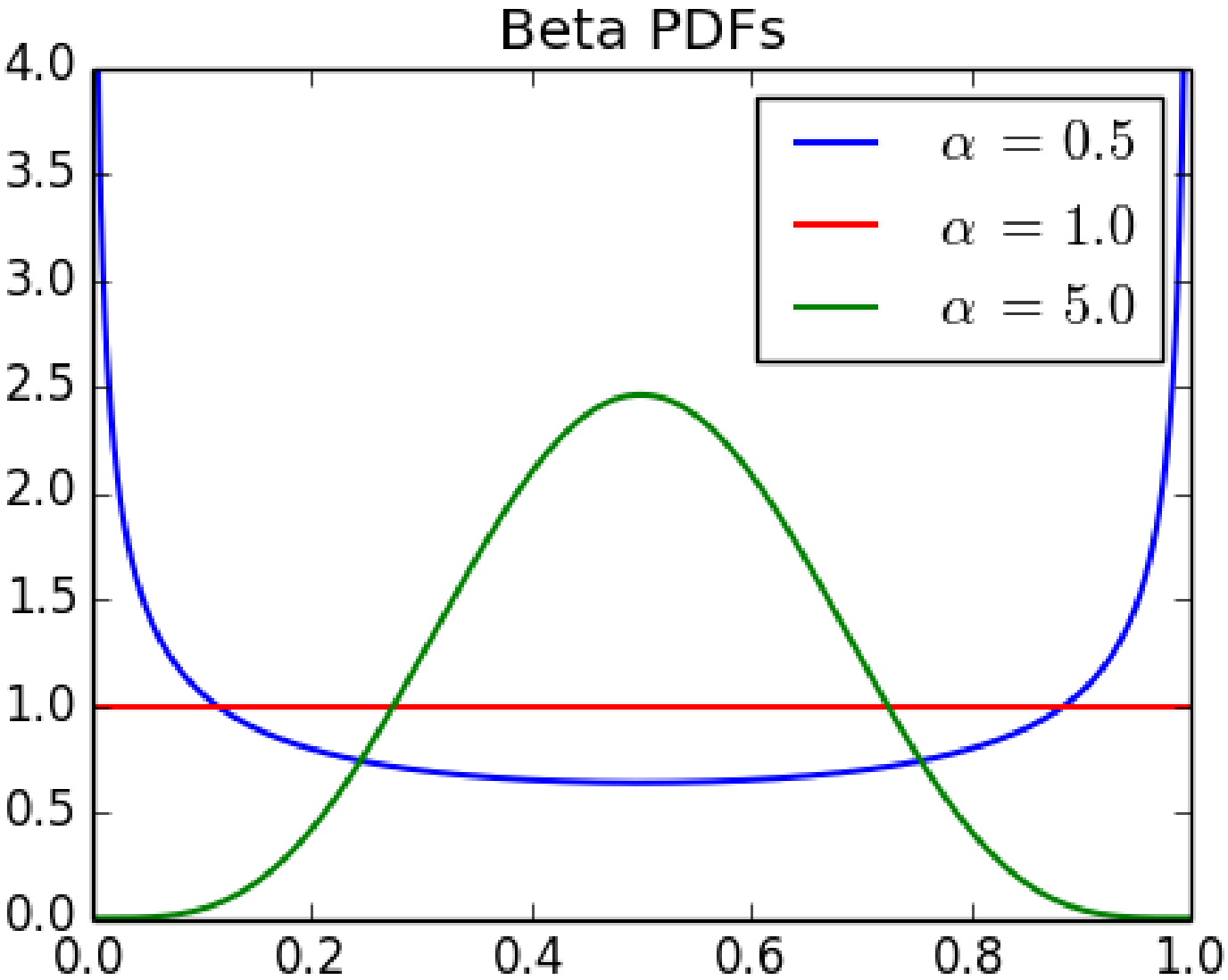
Consider a symmetric Dirichlet in which $\alpha_v = \alpha$ for all v. Agnostic about favoring one component over another.

Here the α parameter measures the concentration of distribution on the center of the simplex, where the mass on each term is more evenly spread out:

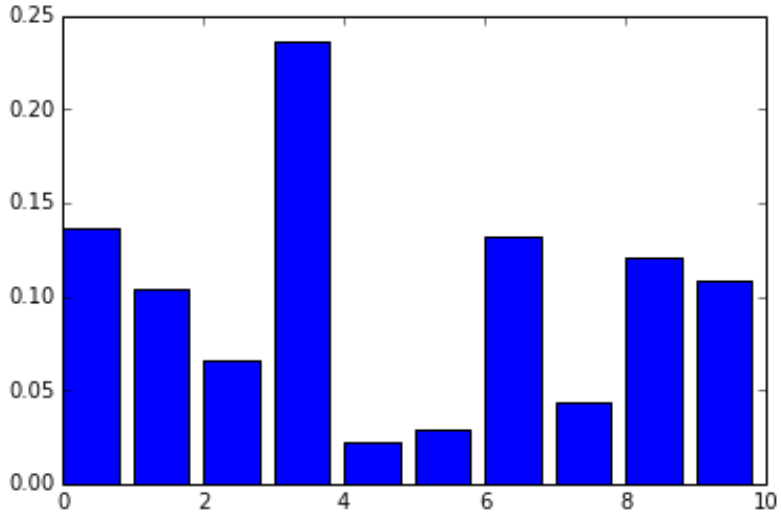
1. $\alpha = 1$ is a uniform distribution.
2. $\alpha > 1$ puts relatively more weight in center of simplex.
3. $\alpha < 1$ puts relatively more weight on corners of simplex.

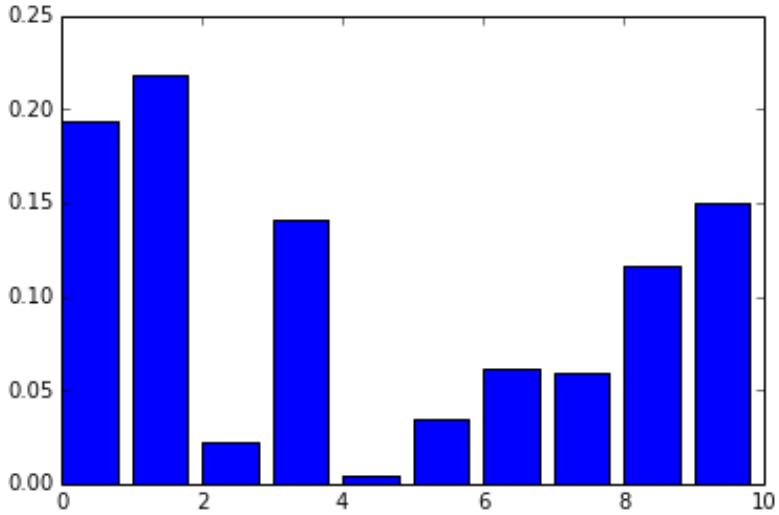
When $V = 2$, the Dirichlet is the beta distribution.

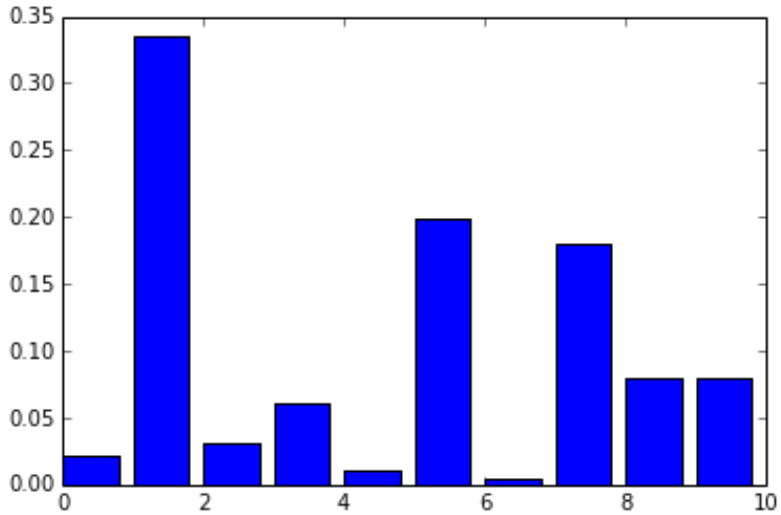
Beta with Different Parameters

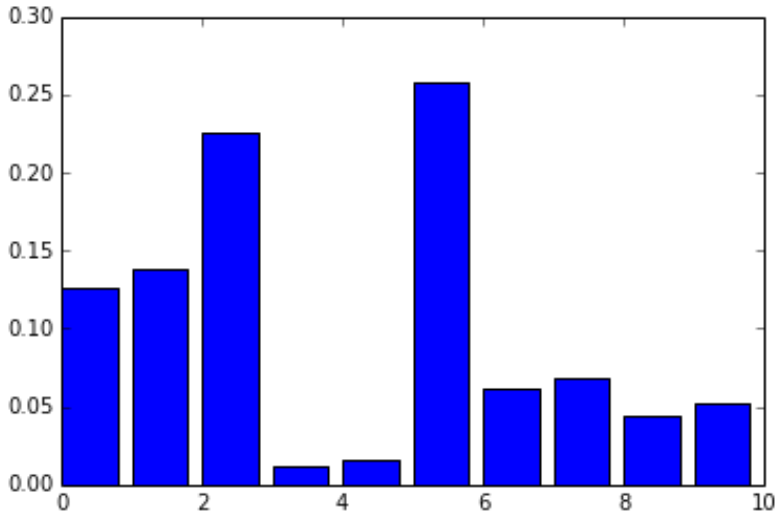


Draws from Dirichlet with $\alpha = 1$

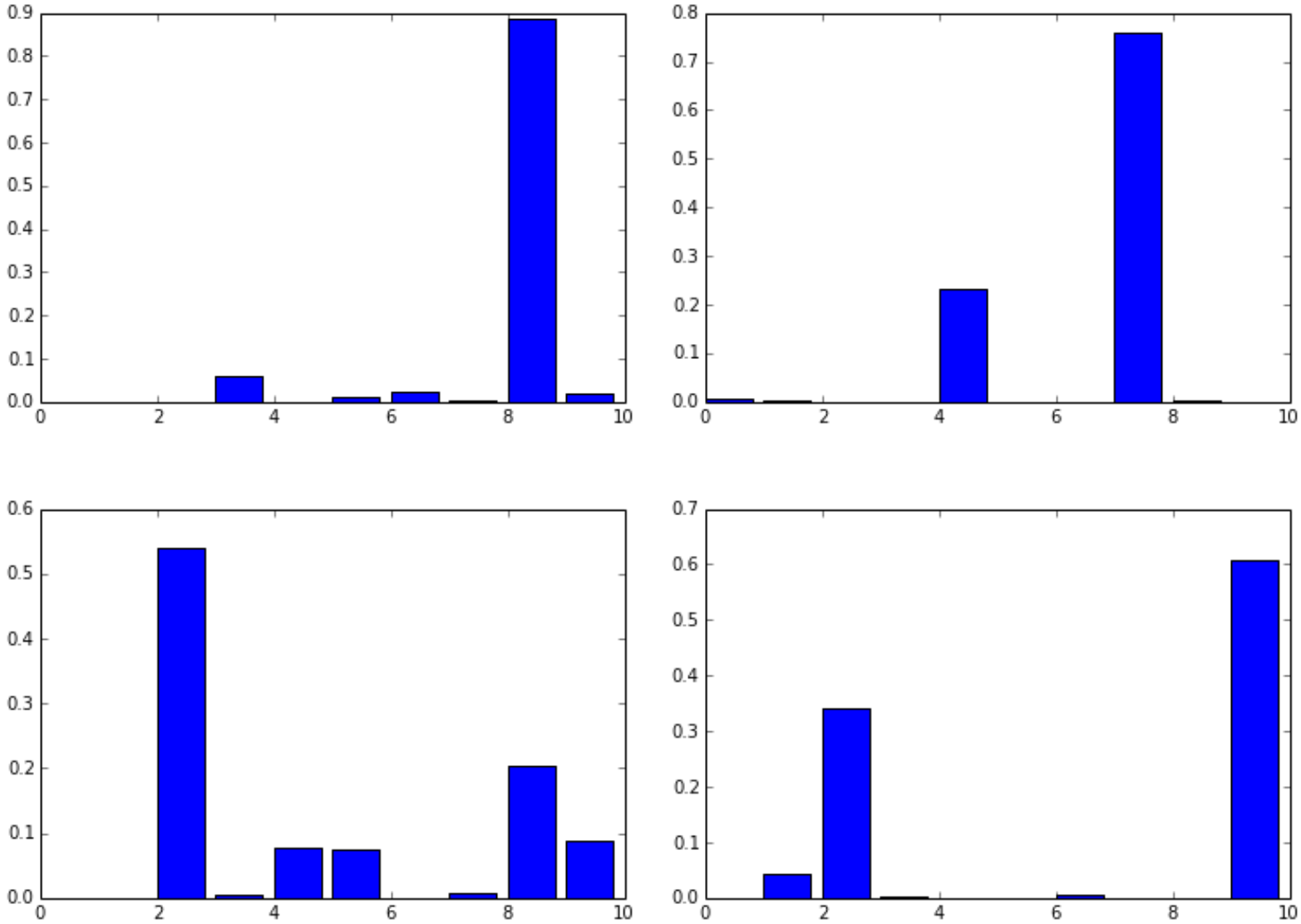




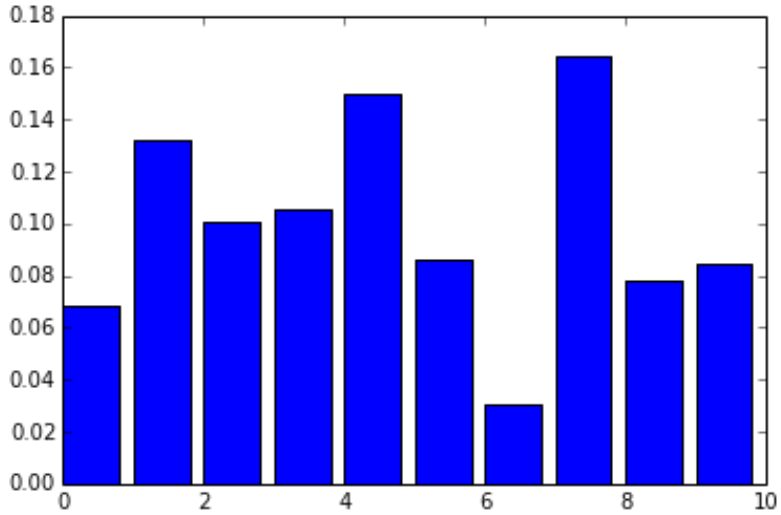


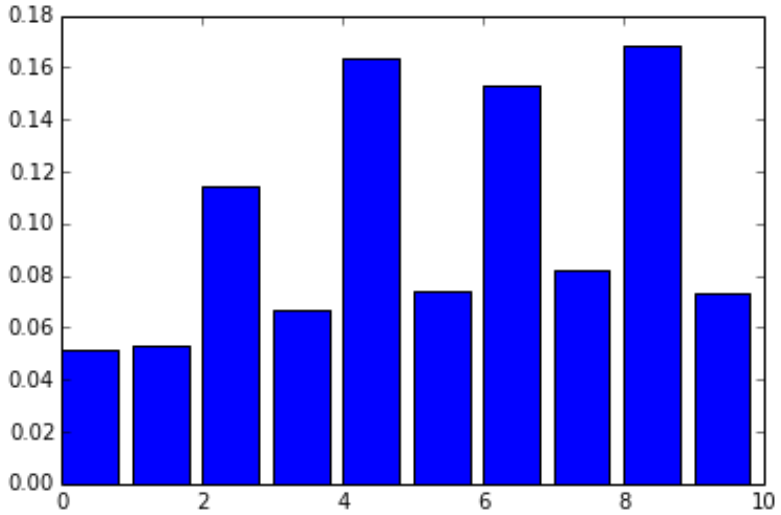


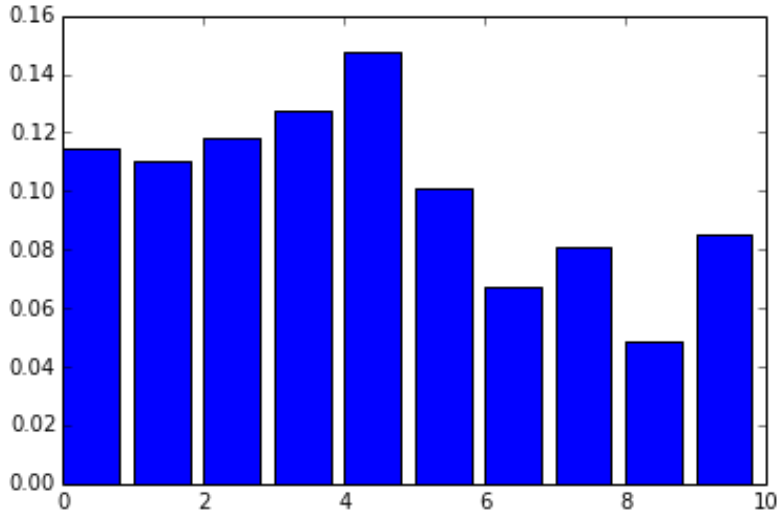
Draws from Dirichlet with $\alpha = 0.1$

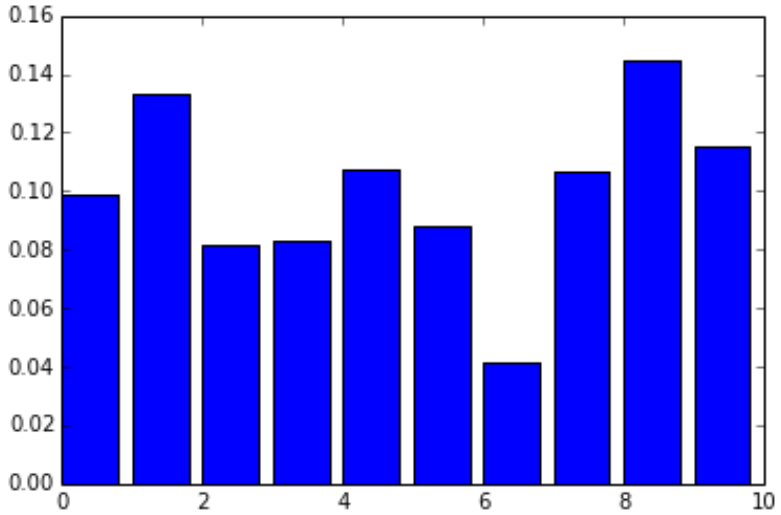


Draws from Dirichlet with $\alpha = 10$









Posterior Distribution

$$\Pr[\boldsymbol{\beta} \mid \mathbf{x}_d] \propto \Pr[\mathbf{x}_d \mid \boldsymbol{\beta}] \Pr[\boldsymbol{\beta}] \propto \prod_{v=1}^{V} \beta_v^{x_{d,v}} \prod_{v=1}^{V} \beta_v^{\alpha_v - 1} = \prod_{v=1}^{V} \beta_v^{x_{d,v} + \alpha_v - 1}.$$

Posterior is a Dirichlet with parameters $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_V)$ where $\widehat{\alpha}_v \equiv \alpha_v + x_{d,v}$.

Add term counts to the prior distribution's parameters to form posterior distribution.

The parameters in the prior distribution are sometimes called *pseudo-counts*, and can be viewed as observations made before $\mathbf{x}_d$.

Moments of the Posterior

One can show that a Dirichlet with parameter vector $\boldsymbol{\alpha}$ satisfies

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v}{\alpha} \text{ and } V[\,\beta_v\,] = \frac{\alpha_v(\alpha - \alpha_v)}{\alpha^2(\alpha+1)}, \text{ where } \alpha \equiv \sum_v \alpha_v.$$

If we apply these formulas to the posterior distribution we obtain

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v + x_{d,v}}{\alpha + N_d} \text{ and } V[\,\beta_v\,] = \frac{(\alpha_v + x_{d,v})(\alpha + N_d - \alpha_v - x_{d,v})}{(\alpha + N_d)^2(\alpha + N_d + 1)}.$$

One can also show that the mean corresponds to the predictive distribution for an additional word.

Data Overwhelming the Prior

Recall the MLE estimates for $\widehat{\beta}_v$ satisfies $N_d \widehat{\beta}_v = x_{d,v}$. We then have

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v + N_d \widehat{\beta}_v}{\alpha + N_d} \text{ and } V[\,\beta_v\,] = \frac{(\alpha_v + N_d \widehat{\beta}_v)(\alpha + N_d - \alpha_v - N_d \widehat{\beta}_v)}{(\alpha + N_d)^2(\alpha + N_d + 1)}.$$

If we take the limit as $N_d \to \infty$, we obtain a degenerate posterior distribution concentrated fully on the MLE parameter estimates.

Intuition: the more data we see, the less our priors should influence our beliefs.

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.

Graphical Models

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

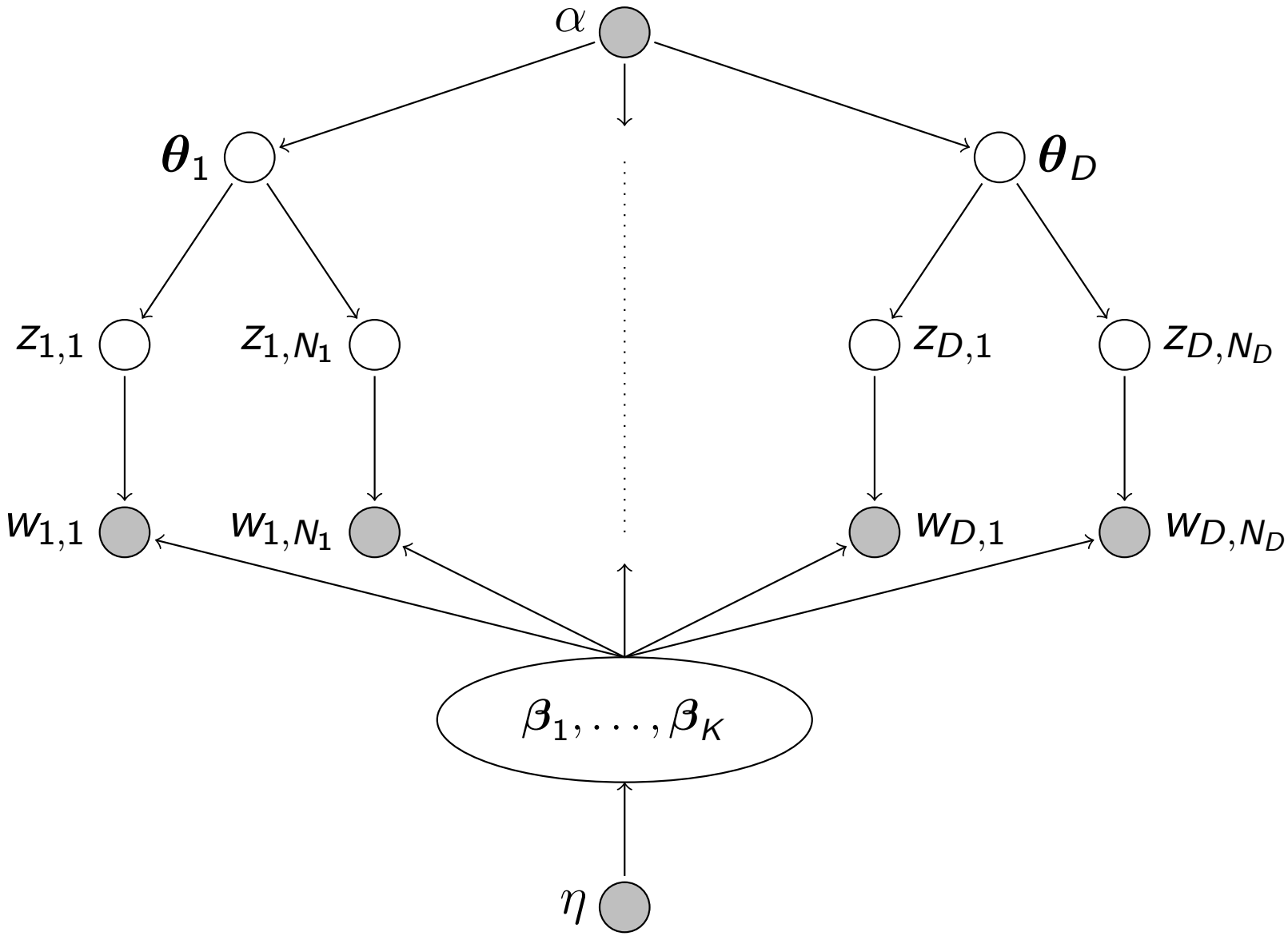
In high-dimensional models, it is useful to summarize relationships among random variables with directed graphs in which nodes represent random variables and links between nodes represent dependencies.

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.

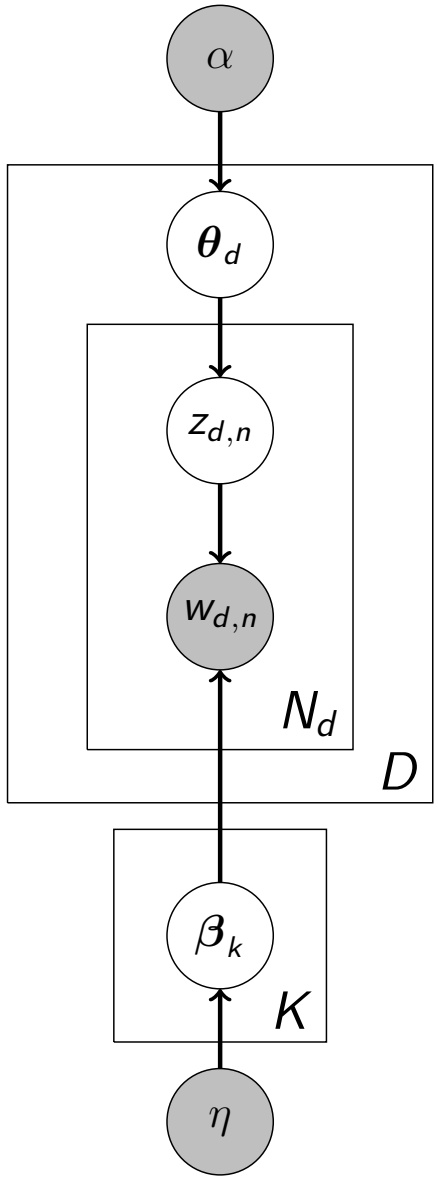
A *Bayesian network* is a probabilistic model whose joint distribution can be represented by a directed acyclic graph (DAG).

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 β_k

The Markov blanket of β_k is:

- The parent η.
- The children $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$.
- The children's parents $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$.

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

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

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

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

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

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

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:

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

Example: Seed

	term1		term2		term3		term4		
Doc	T0	T1	T0	T1	T0	T1	T0	T1	$\widehat{\theta}_i$
A	0	0	1	3	0	2	82	80	0.506
B	9	4	1	0	5	4	12	19	0.5
C	35	43	0	0	38	30	0	0	0.5
	0.24	0.254	0.011	0.017	0.235	0.195	0.513	0.535	
	$\widehat{\beta}_0^1$	$\widehat{\beta}_1^1$	$\widehat{\beta}_0^2$	$\widehat{\beta}_1^2$	$\widehat{\beta}_0^3$	$\widehat{\beta}_1^3$	$\widehat{\beta}_0^4$	$\widehat{\beta}_1^4$	

Example: Iteration 2

	term1		term2		term3		term4		
Doc	T0	T1	T0	T1	T0	T1	T0	T1	$\widehat{\theta}_i$
A	0	0	4	0	2	0	35	127	0.753
B	10	3	1	0	5	4	4	27	0.625
C	73	5	0	0	63	5	0	0	0.074
	0.421	0.047	0.026	0.001	0.355	0.053	0.198	0.899	
	$\widehat{\beta}_0^1$	$\widehat{\beta}_1^1$	$\widehat{\beta}_0^2$	$\widehat{\beta}_1^2$	$\widehat{\beta}_0^3$	$\widehat{\beta}_1^3$	$\widehat{\beta}_0^4$	$\widehat{\beta}_1^4$	

Example: Iteration 5

	term1		term2		term3		term4		
Doc	T0	T1	T0	T1	T0	T1	T0	T1	$\widehat{\theta}_i$
A	0	0	0	4	2	0	0	162	0.982
B	13	0	0	1	9	0	0	31	0.589
C	78	0	0	0	68	0	0	0	0.007
	0.535	0.001	0.001	0.026	0.464	0.001	0.001	0.973	
	$\widehat{\beta}_0^1$	$\widehat{\beta}_1^1$	$\widehat{\beta}_0^2$	$\widehat{\beta}_1^2$	$\widehat{\beta}_0^3$	$\widehat{\beta}_1^3$	$\widehat{\beta}_0^4$	$\widehat{\beta}_1^4$	

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.

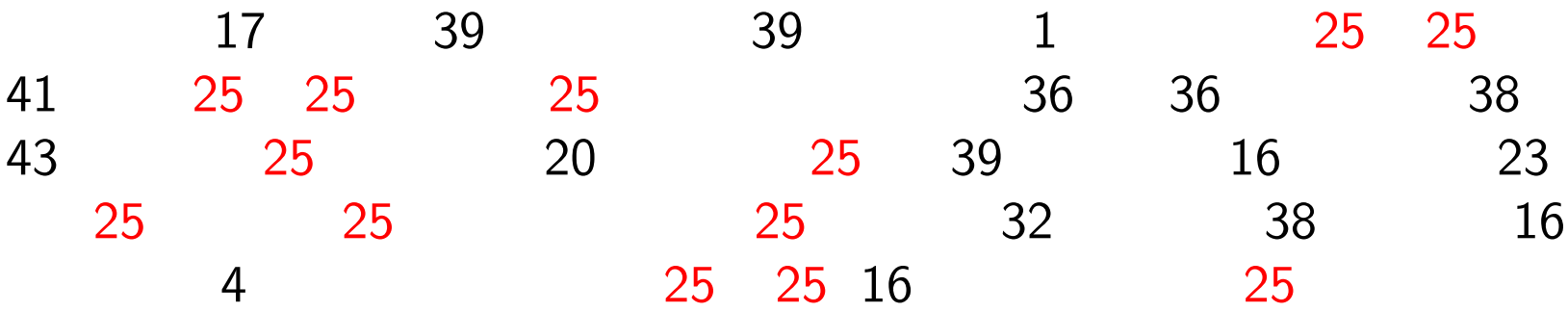
Example statement: Yellen, March 2006, #51

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

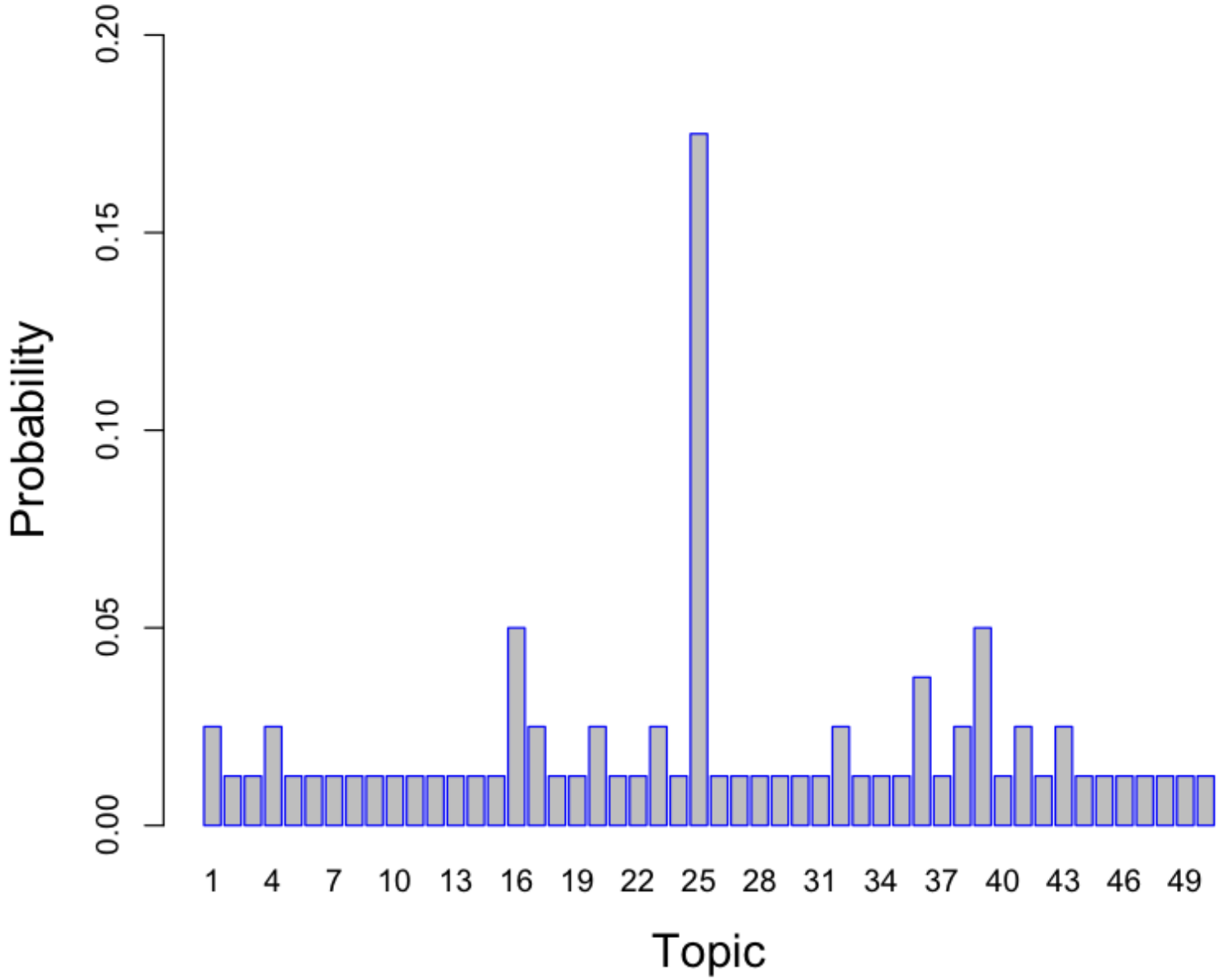
Example statement: Yellen, March 2006, #51

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

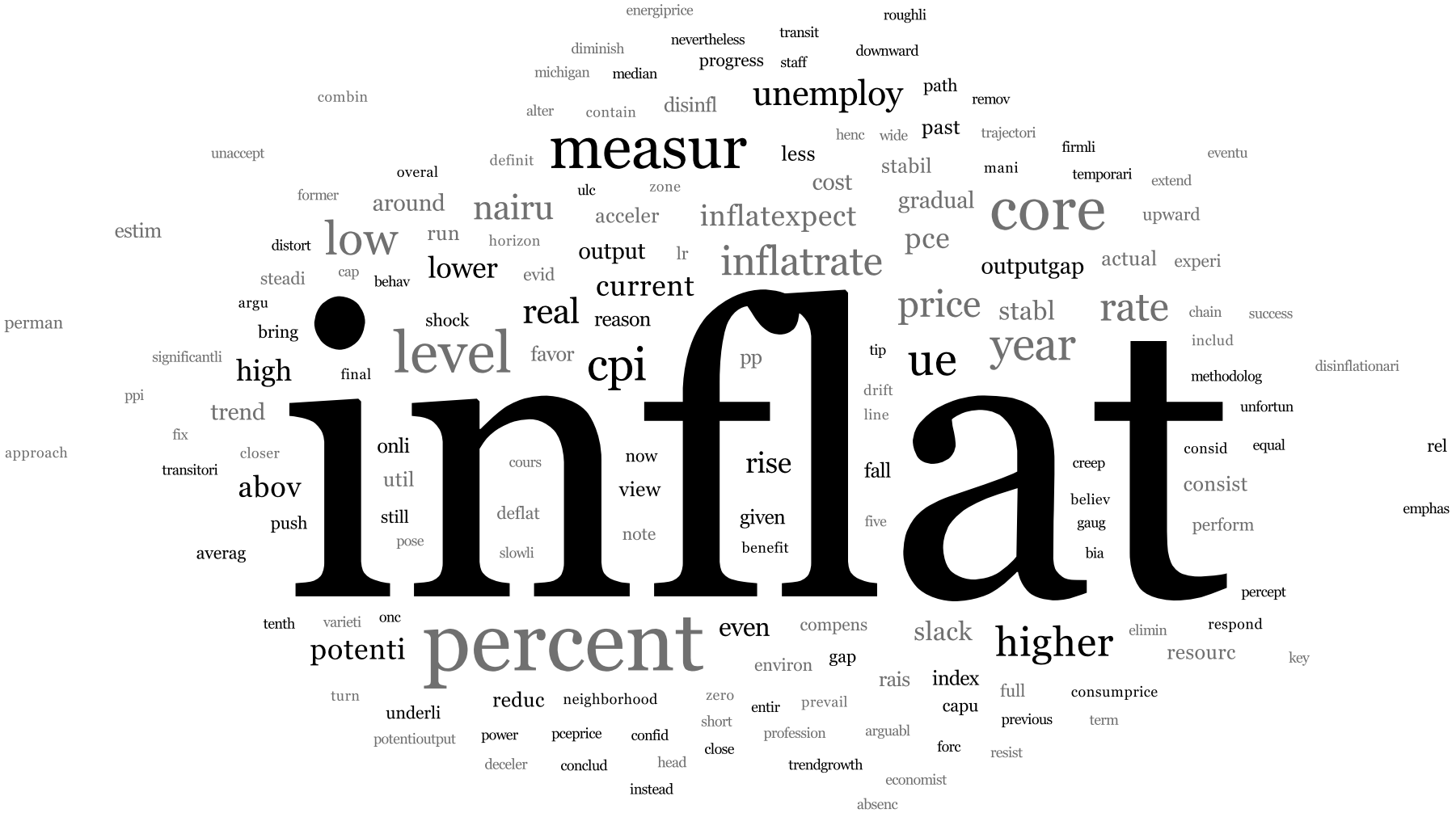
Example statement: Yellen, March 2006, #51



Distribution of Attention



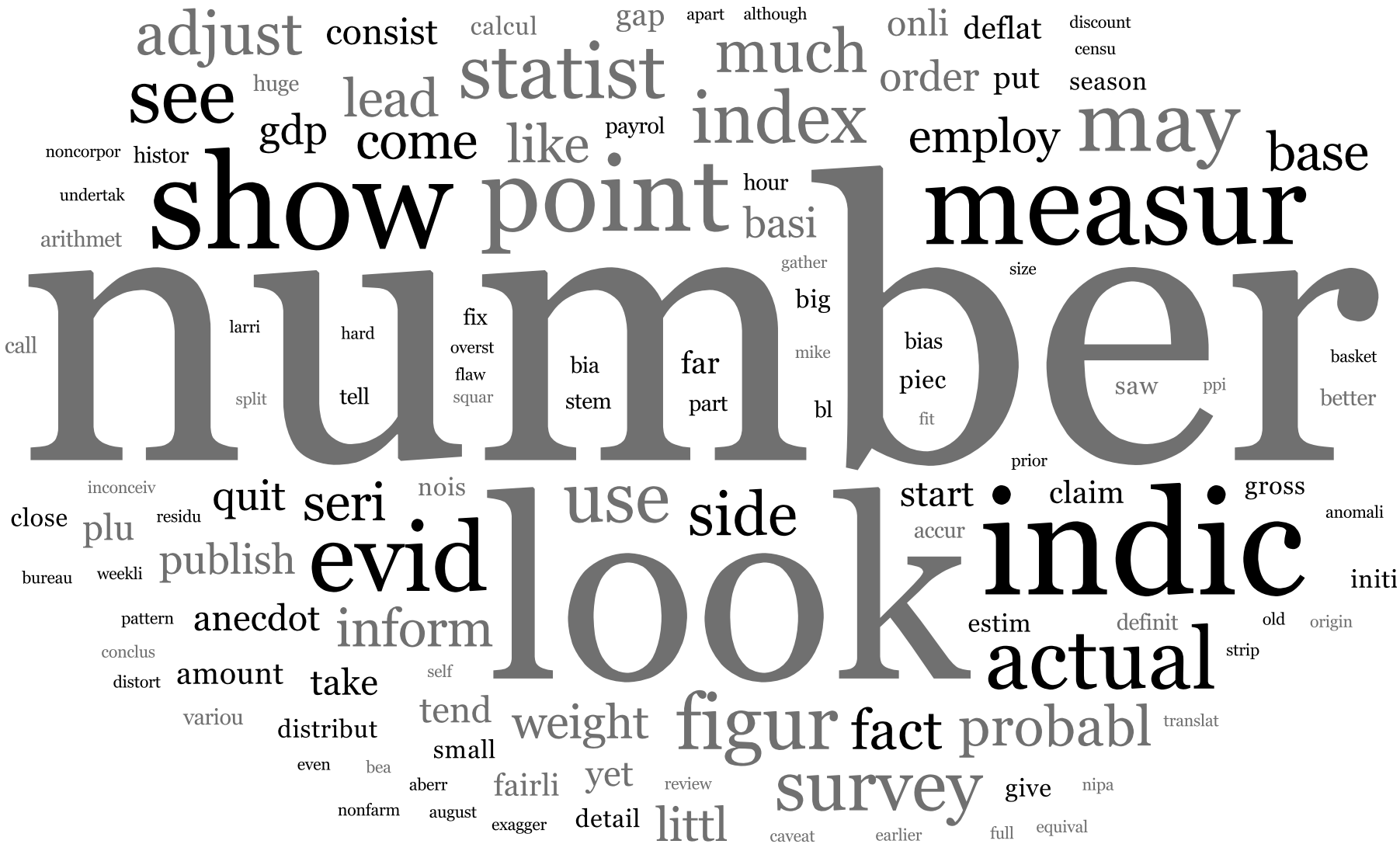
Topic 25


energiprice
roughli
nevertheless
transit
diminish
downward
progress
staff
michigan
median
combin
unemploy
path
alter
contain
disinfl
remov
measur
henc
wide
past
trajectori
unaccept
definit
less
firmli
eventu
overal
stabil
mani
temporari
extend
former
around
nairu
ulc
zone
cost
gradual
core
estim
low
run
acceler
inflatexpect
upward
distort
horizon
output
lr
pce
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
push
still
deflat
given
five
gaug
emphas
perform
pose
note
benefit
averag
slowli
bia
percent
percept
tenth
varieti
onc
even
compens
slack
higher
elimin
respond
potenti
environ
gap
resourc
key
rais
index
turn
reduc
neighborhood
zero
full
consumprice
underli
entir
prevail
capu
short
previous
term
potentioutput
power
pceprice
confid
profession
arguabl
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
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
strip
actual
amount
take
self
distort
variou
distribut
tend
weight
figur
fact
probabl
translat
small
even
bea
survey
give
nipa
aberr
fairli
yet
review
nonfarm
august
exagger
detail
littl
caveat
earlier
full
equival

Advantage of Flexibility

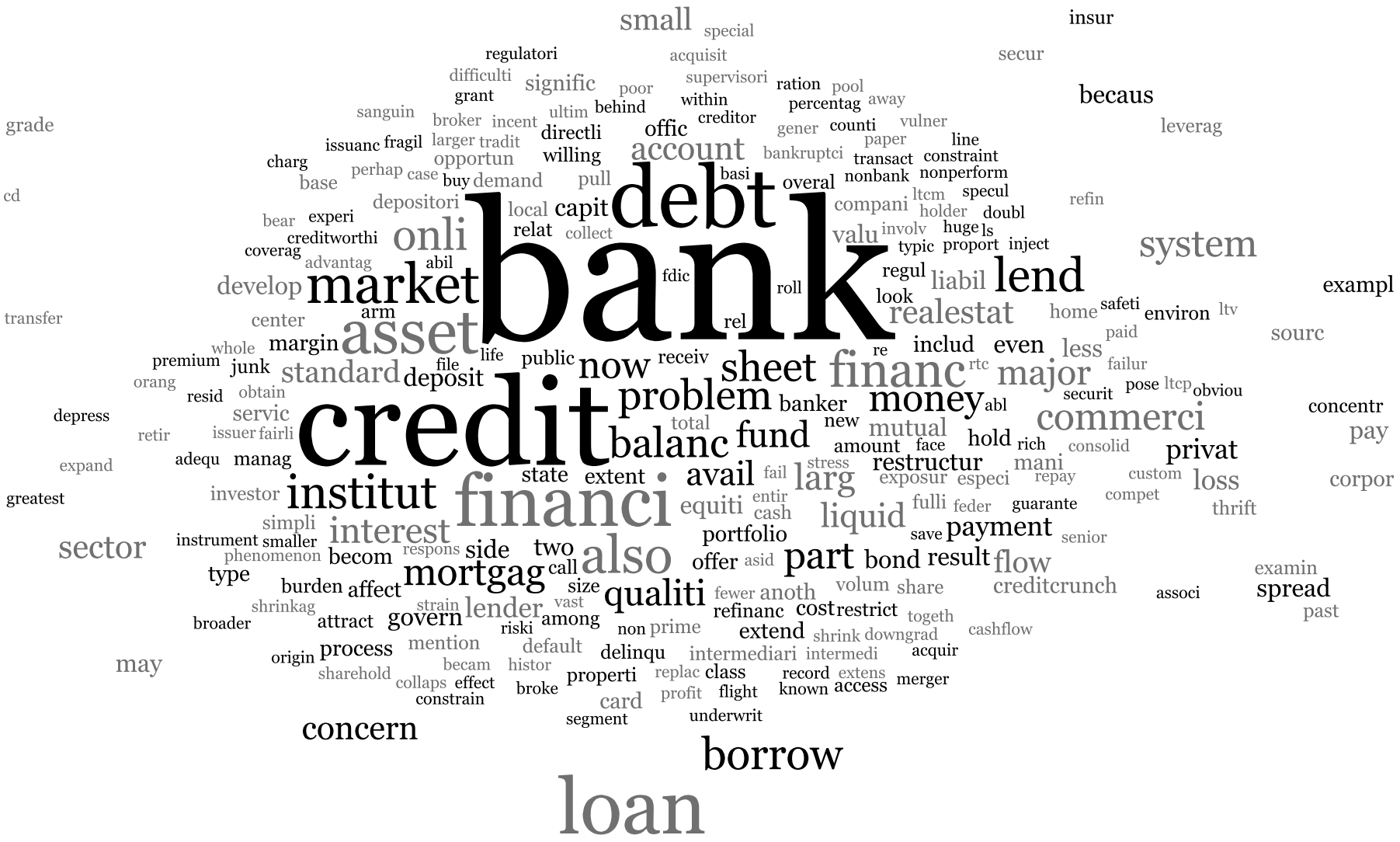
'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11.

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.

Topic 38


small
special
insur
regulatori
acquisit
secur
difficulti
signific
supervisori
ration
pool
poor
grant
within
percentag
away
becaus
sanguin
ultim
behind
creditor
grade
broker
incent
directli
offic
gener
counti
vulner
leverag
issuanc
fragil
larger
tradit
account
bankruptci
paper
line
charg
opportun
willing
transact
constraint
perhap
case
basi
nonbank
nonperform
cd
base
buy
demand
pull
overal
depositori
ltcm
specul
refin
compani
holder
doubl
bear
experi
local
capit
debt
relat
collect
involv
huge
ls
valu
typic
proport
inject
creditworthi
coverag
advantag
abil
onli
bank
system
develop
market
fdic
regul
liabil
lend
exampl
look
roll
realestat
transfer
center
arm
rel
home
safeti
environ
ltv
asset
paid
sourc
whole
margin
life
public
now
receiv
re
includ
even
less
premium
junk
file
standard
deposit
sheet
financ
rtc
major
failur
orang
resid
obtain
problem
banker
money
abl
securit
pose
ltcp
obviou
depress
servic
total
new
mutual
commerci
concentr
retir
issuer
fairli
balanc
fund
hold
rich
consolid
pay
expand
adequ
manag
amount
face
restructur
mani
privat
state
extent
avail
fail
larg
exposur
especi
repay
custom
loss
corpor
greatest
investor
institut
financi
equiti
entir
cash
liquid
full
feder
guarante
compet
thrift
simpli
interest
portfolio
save
payment
senior
sector
instrument
smaller
phenomenon
becom
respons
side
two
also
offer
asid
part
bond
result
flow
examin
type
mortgag
call
size
qualiti
fewer
anoth
volum
share
creditcrunch
associ
spread
burden
affect
refinanc
cost
restrict
shrinkag
strain
vast
lender
govern
attract
among
broader
riski
non
prime
extend
shrink
downgrad
cashflow
togeth
past
may
origin
process
mention
default
delinqu
intermediari
intermedi
acquir
sharehold
becam
histor
collaps
effect
properti
replac
class
record
extens
merger
constrain
broke
profit
flight
known
access
card
underwrit
segment
concern
borrow
loan

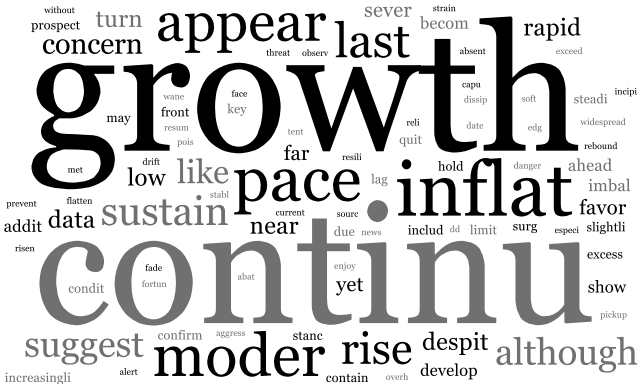
Topic 23

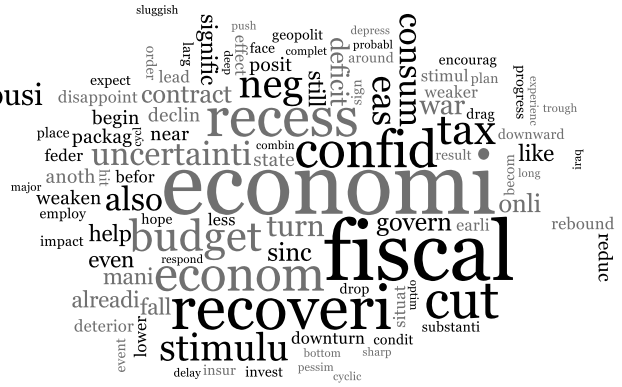

inventori
invest
stockmkt
profit
high
busi
spend
incom
capit
consumpt
sector
level
wealth
stock
effect
final
demand
earn
still
declin
equip
accumul
rel
corpor
equiti
mortgag
real
strong
hold
fixd
fundament
purchas
singl
improv
show
good
dispos
ratio
share
return
gain
sale
fall
see
neg
swing
factor
quit
firm
adjust
tax
basic
tech
expenditur
exist
signific
cashflow
lag
though
analyst
rapidli
person
start
import
low
cyclic
substanti
saverate
histor
valu
margin
big
famili
play
overhang
current
relat
rise

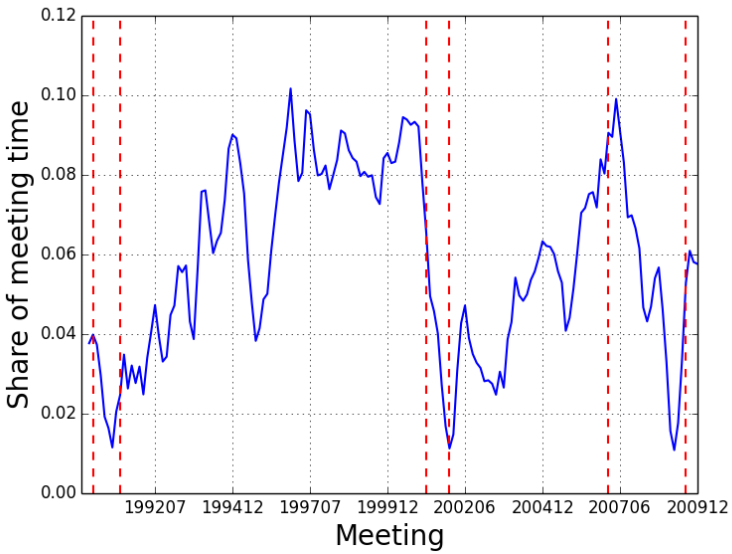
Topic 29

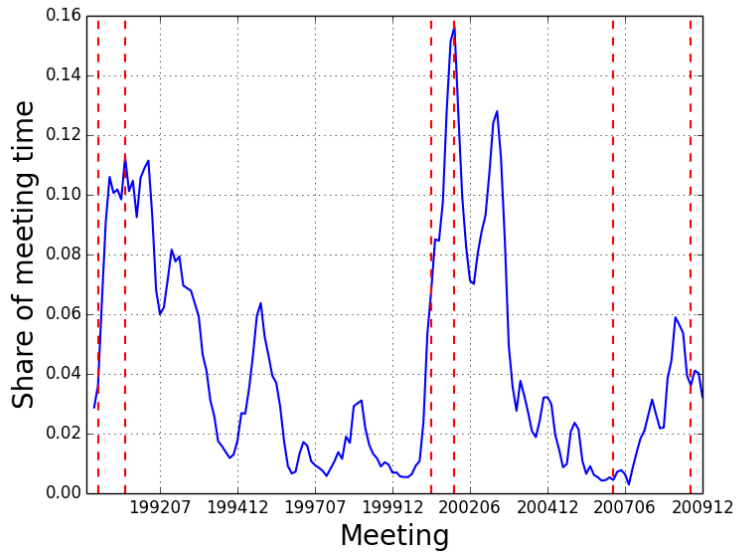

deceler
bargain
chain
consumprice
soon
annual
surg
exert
develop
boom
element
innov
perform
ultim
remark
much
occur
higher
improv
analysi
decad
combin
found
per
enough
clear
servic
fall
result
firm
seen
unit
declin
popul
workforc
capu
insur
benefit
side
also
earn
medic
resourc
competit
enjoy
output
begin
difficult
individu
prefl
stori
simpli
run
markup
compens
old
worker
payrol
slow
gain
averag
capit
bonus
collect
pattern
ulc
employ
employe
rate
shift
ad
offset
live
acceler
wage
enhanc
profit
extraordinari
return
pay
suppli
may
part
turn
larg
factor
mfp
ue
rapid
qualiti
contribut
import
rise
product
trend
productgrowth
estim
effici
power
shortag
recent
job
labor
effect
past
deriv
asid
line
still
manag
demandgrowth
name
new
risen
upsid
escal
introduct
total
cyclic
attribut
led
given
salari
rais
late
allow
normal
less
share
slower
pickup
hike
translat
creation
advanc
virtual
pass
compon
associ
potentioutput
onli
measur
structur
price
tight
margin
labmkt
age
believ
pressur
unemploy
workweek
input
premium
avail
behind
ongo
faster
hire
focu
sum
rapidli
attract
demand
cost
minimum
technolog
train
materi
hypothesi
possibl
perman
know
even
evid
welfar
longer
rather
overal
growth
work
becom
catch
forc
eventu
particip
significantli
outputgrowth
despit
eci
account
inde
optimist
insecur
hourli
produc
prevent
harder
health
reconcil
unusu
explan
upward
deepen
previous
short
inflationari
statist
skill
lack
relat
employgrowth
plenti
multifactor
nonfarm
back
squeez
help
experienc
lower
effort
qualifi
match
transitori
dispar
latter
stem
busicycl

External Validation









LDA on Survey Data

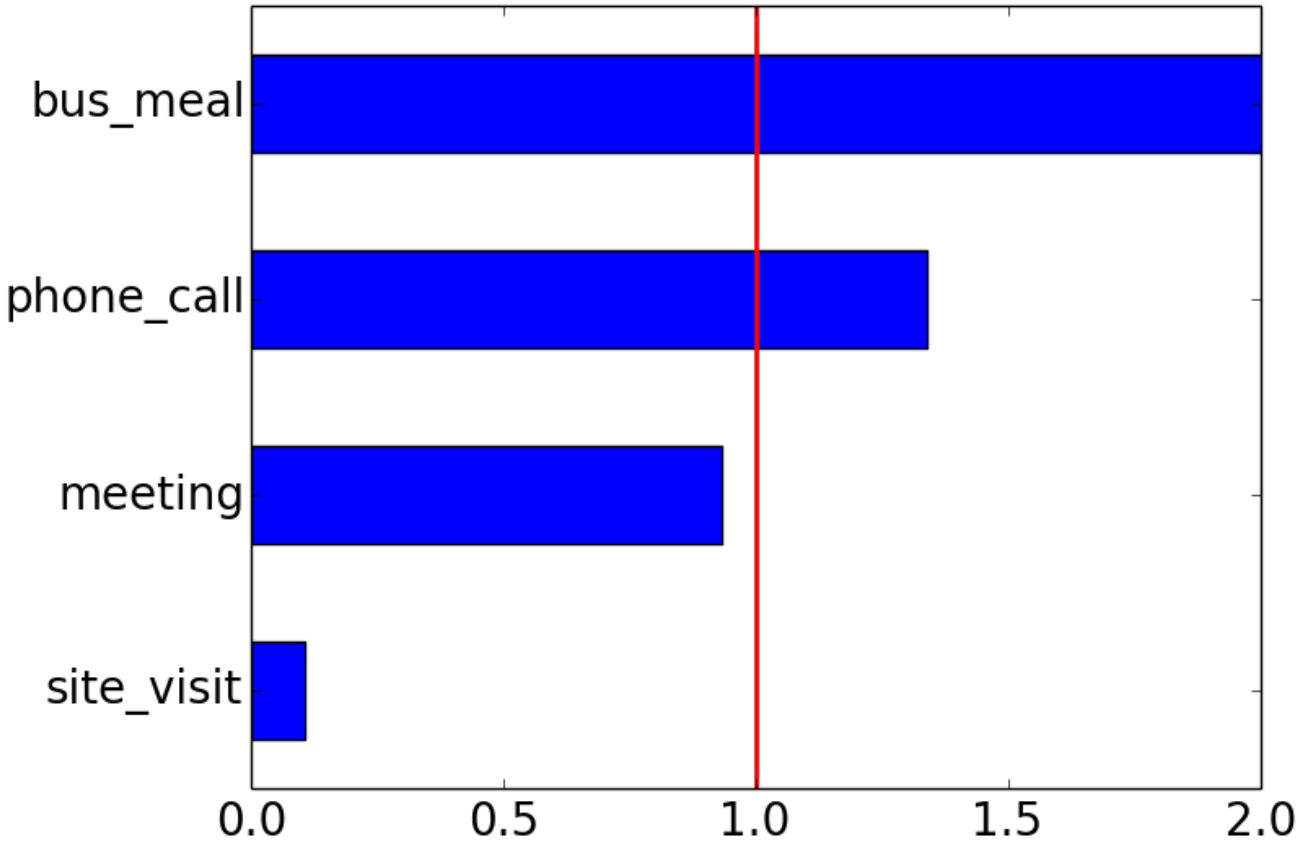
Recall from the first lecture that text data is one instance of count data.

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

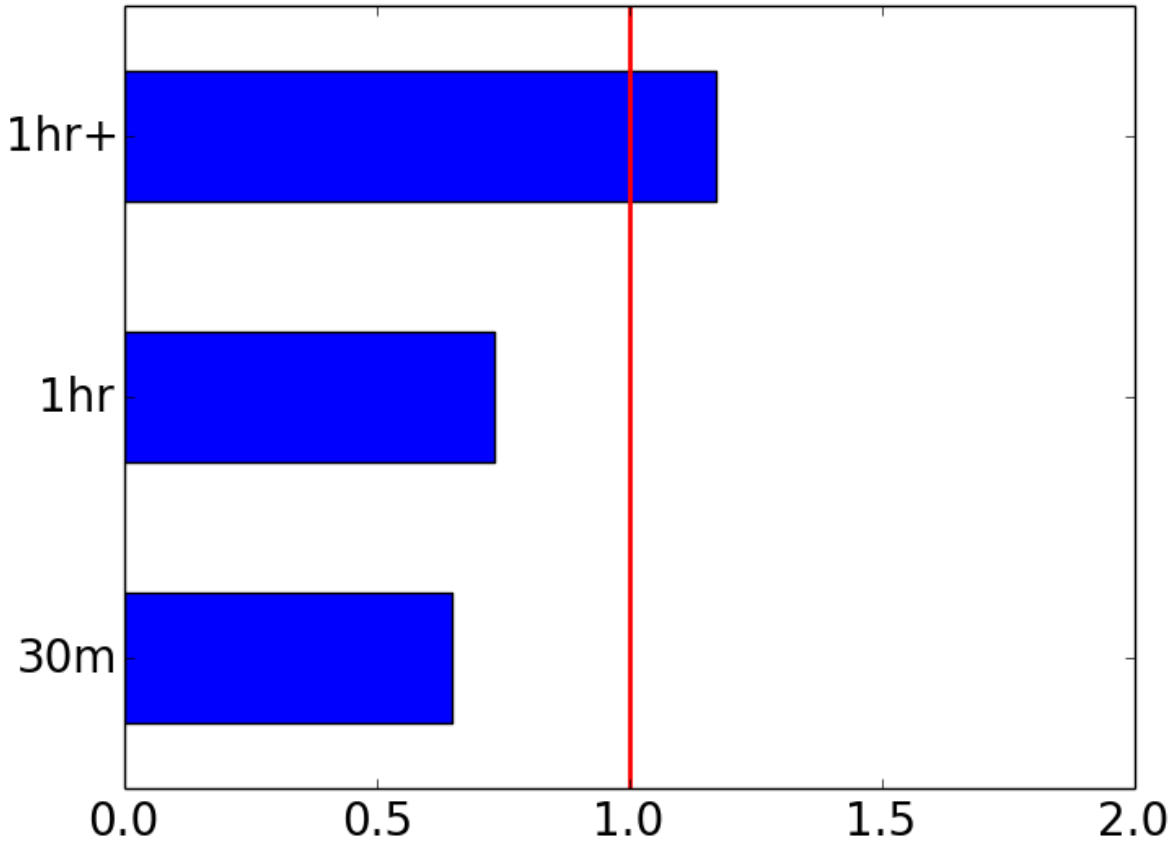
We recently applied it to CEO survey data to estimate management "behaviors" with $K = 2$.

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

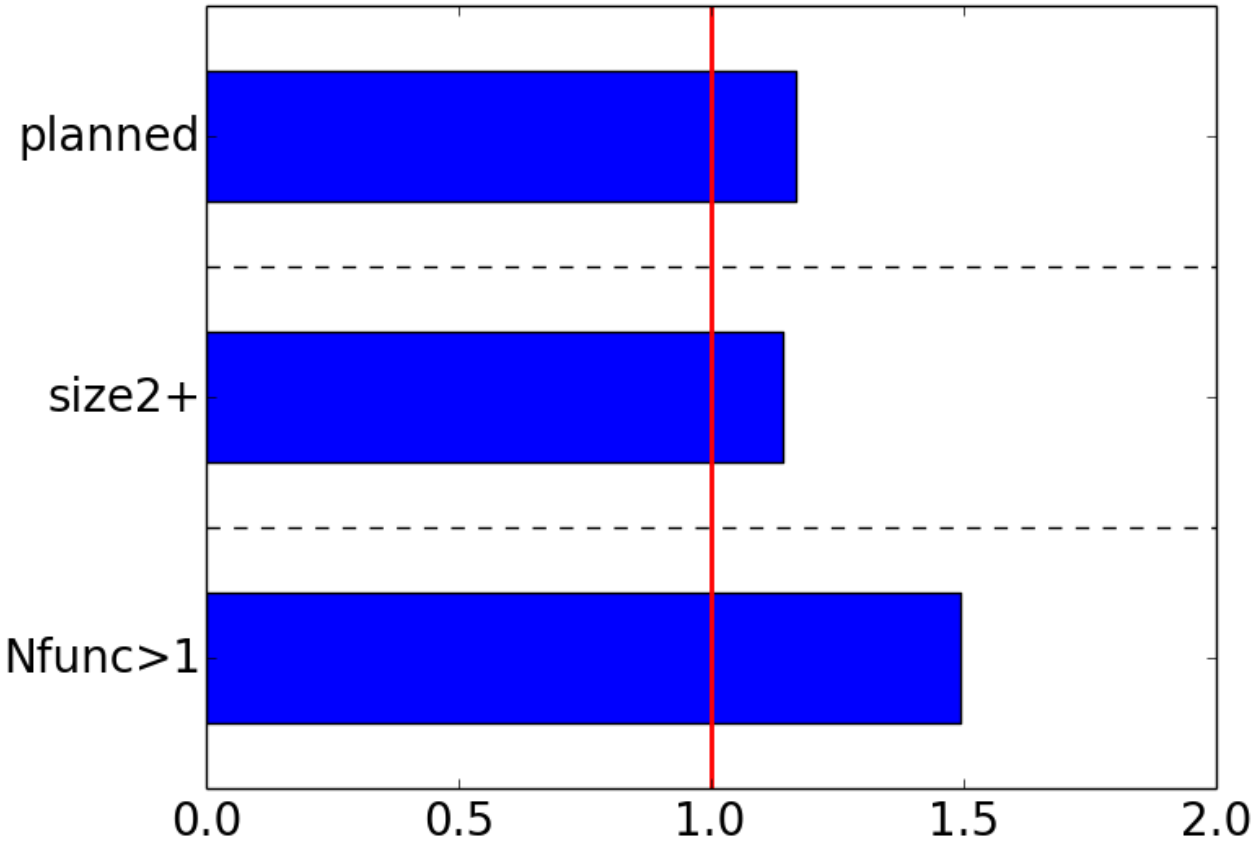
Activity Type



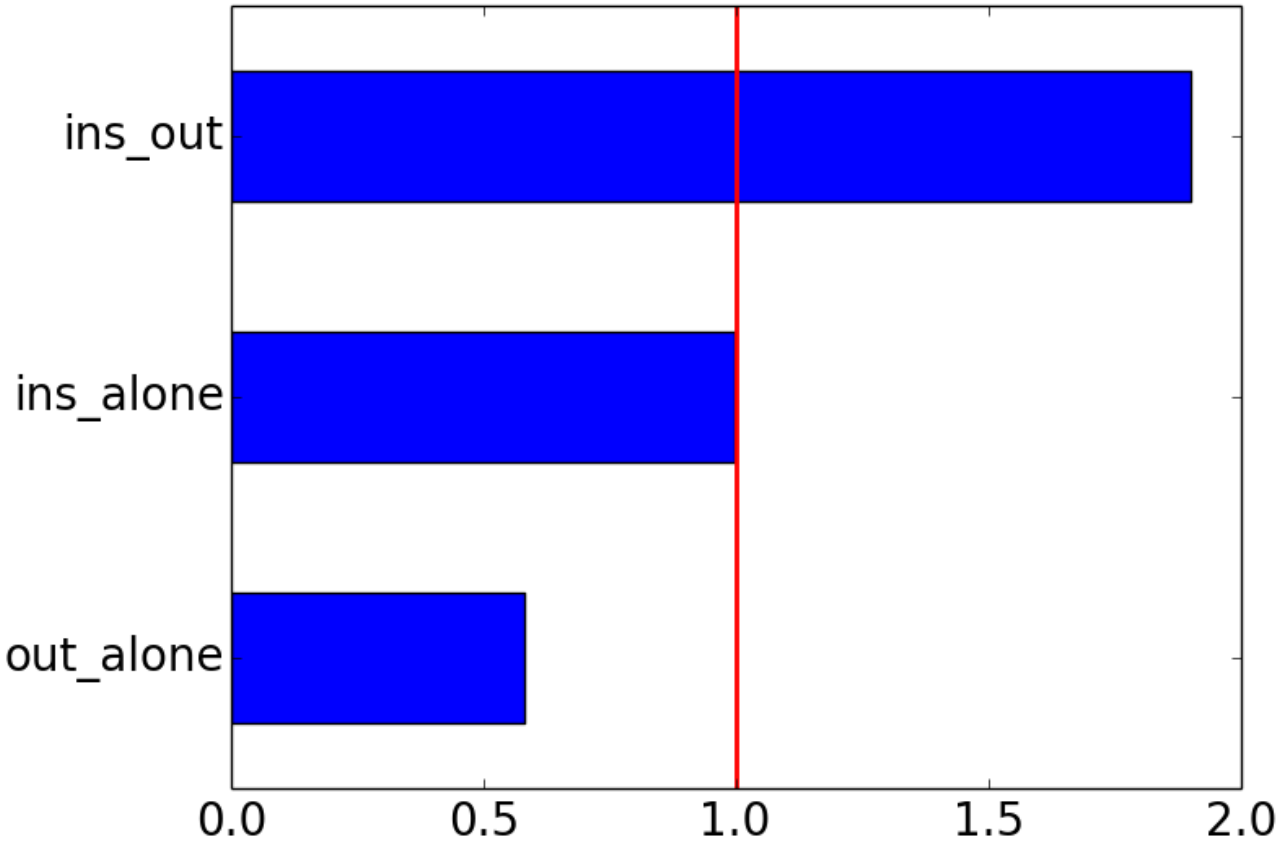
Activity Duration



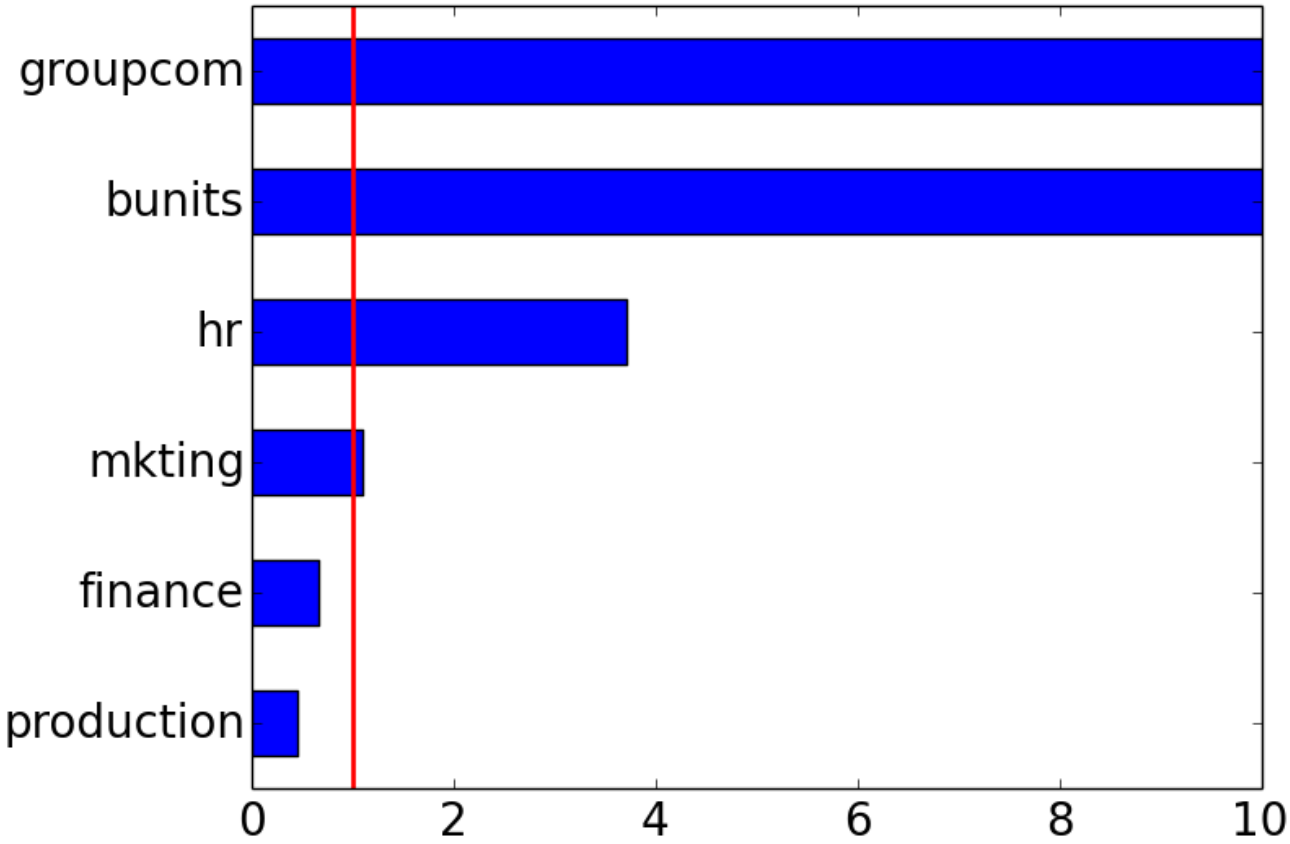
Planning; Size; Number of Functions



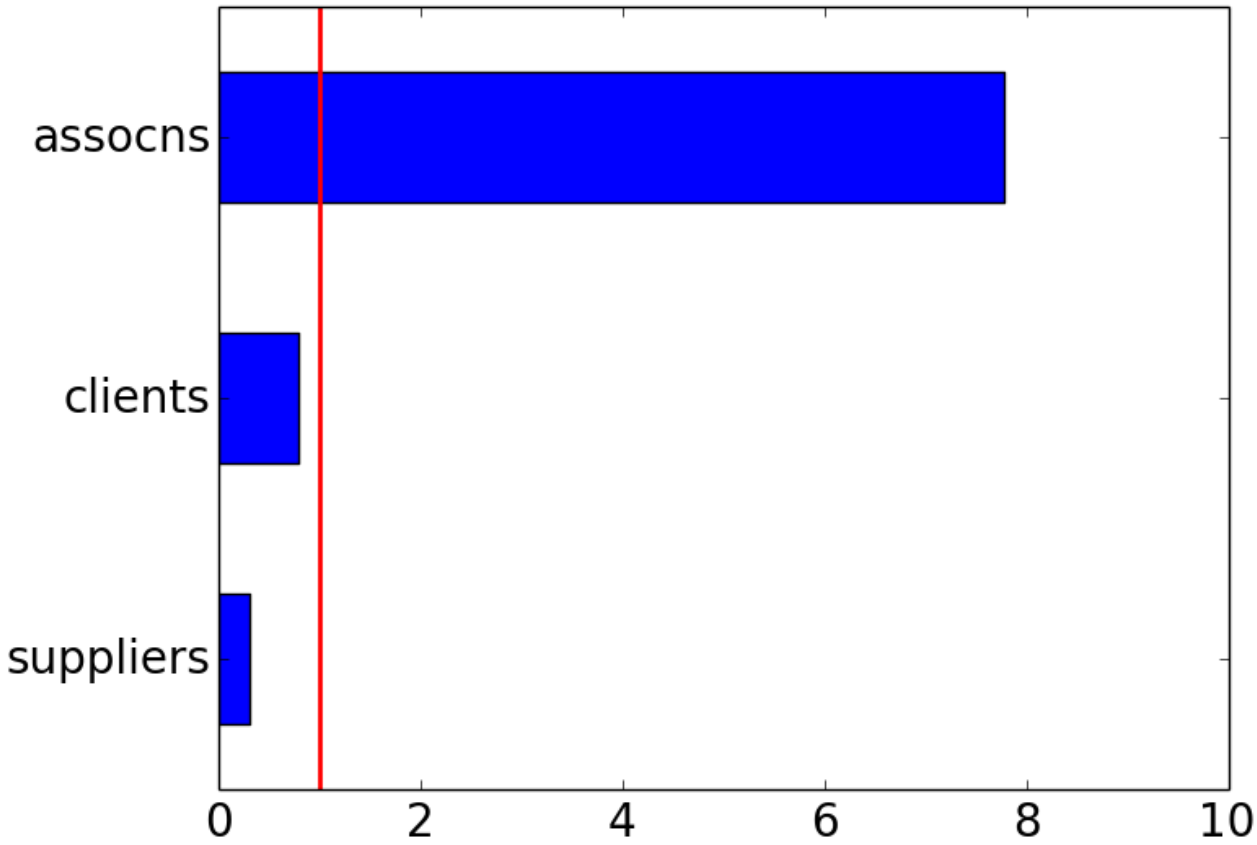
Insiders vs Outsiders



Inside Functions



Outside Functions



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

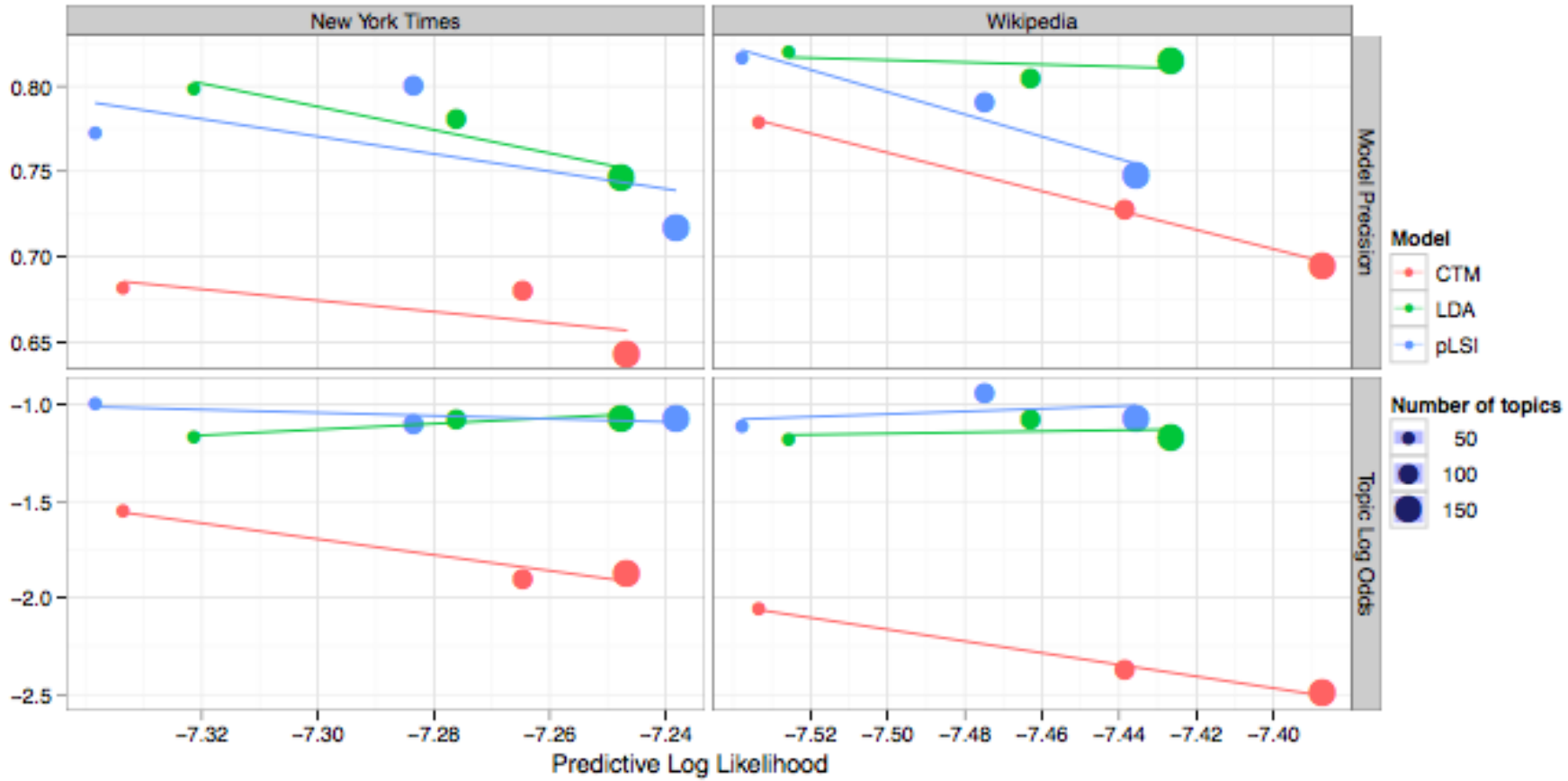


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase 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.

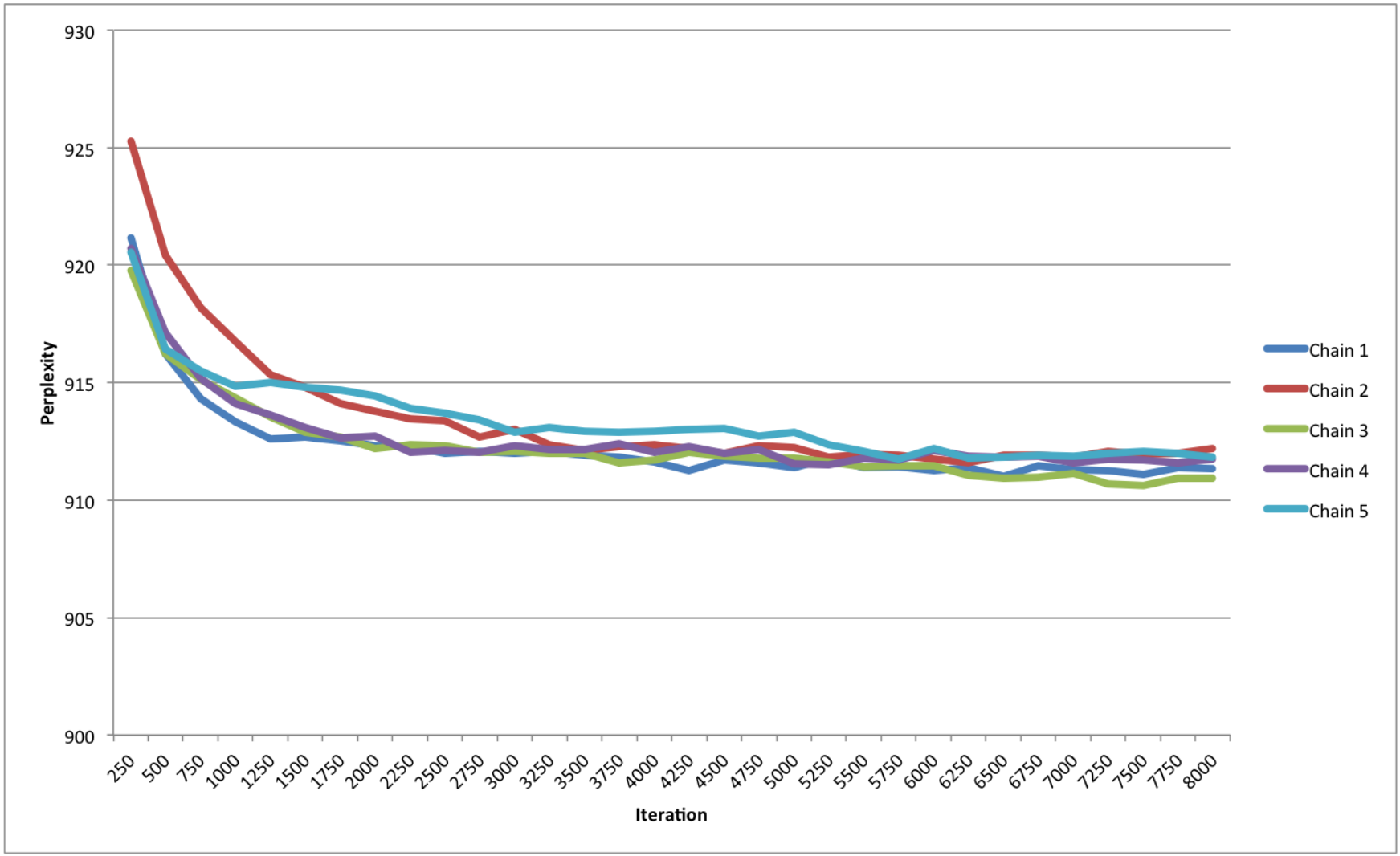
One approach is to measure how well different states of the chain predict the data, and determine convergence in terms of its stability.

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

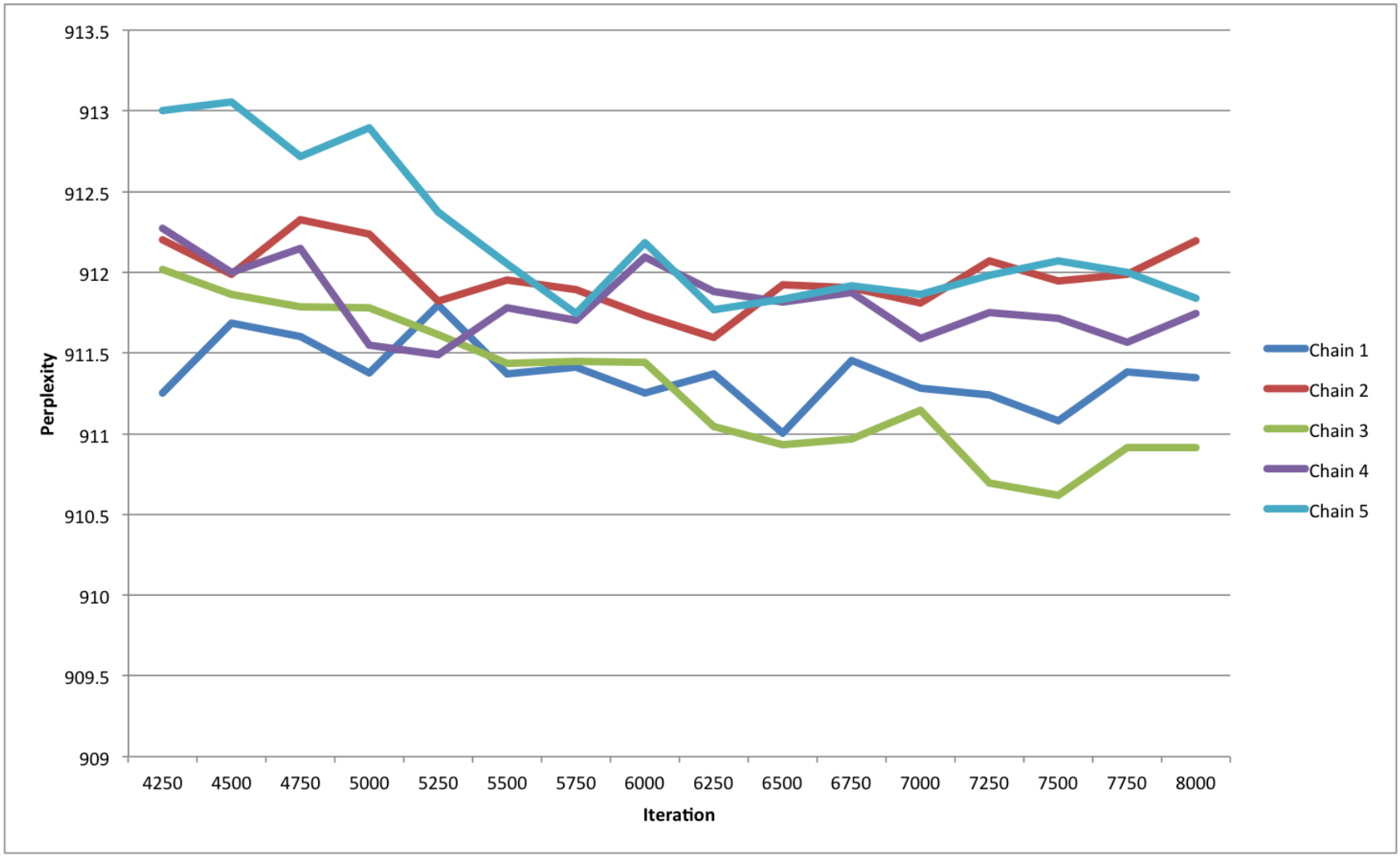
For LDA, a typical goodness-of-fit measure is *perplexity*, given by

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

Perplexity 1



Perplexity 2



Variational Inference

Blei, Ng, and Jordan (2003) use variational inference rather than MCMC to approximate the posterior.

Approximate the true posterior distribution with a simpler functional form that depends on a set of variational parameters.

Then optimize the approximate posterior with respect to the variational parameters so that it lies "close to" the true posterior.

The inference problem becomes an optimization problem.

But note that the family of distributions used to approximate the posterior typically does not include the true posterior.

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.

Gibbs Sampling / Variational Inference

Advantages of sampling:

1. Typically easier to derive sampling algorithms
2. More accurate, especially for approximating features of posterior distribution beyond the mode

Advantages of variational inference:

1. Faster, especially when optimized
2. Deterministic
3. Convergence easy to assess

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.