Lecture 2: Latent Dirichlet Allocation

EPFL Summer School in Computational Methods for Economists

Stephen Hansen
University of Oxford / Imperial College London

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.

Review

Recall the simple unigram model of a document in which

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

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.

Bayesian Inference

In Bayesian inference, we treat $\boldsymbol{\beta}$ as a random variable rather than a fixed parameter.

Bayes' rule states that

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

where

- $\Pr[\boldsymbol{\beta} \mid \mathbf{x}_d]$ is the posterior distribution.
- $\Pr[\mathbf{x}_d \mid \boldsymbol{\beta}]$ is the likelihood function.
- $\Pr[\boldsymbol{\beta}]$ 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[1]

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

Mean is

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v}{\alpha}$$

[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

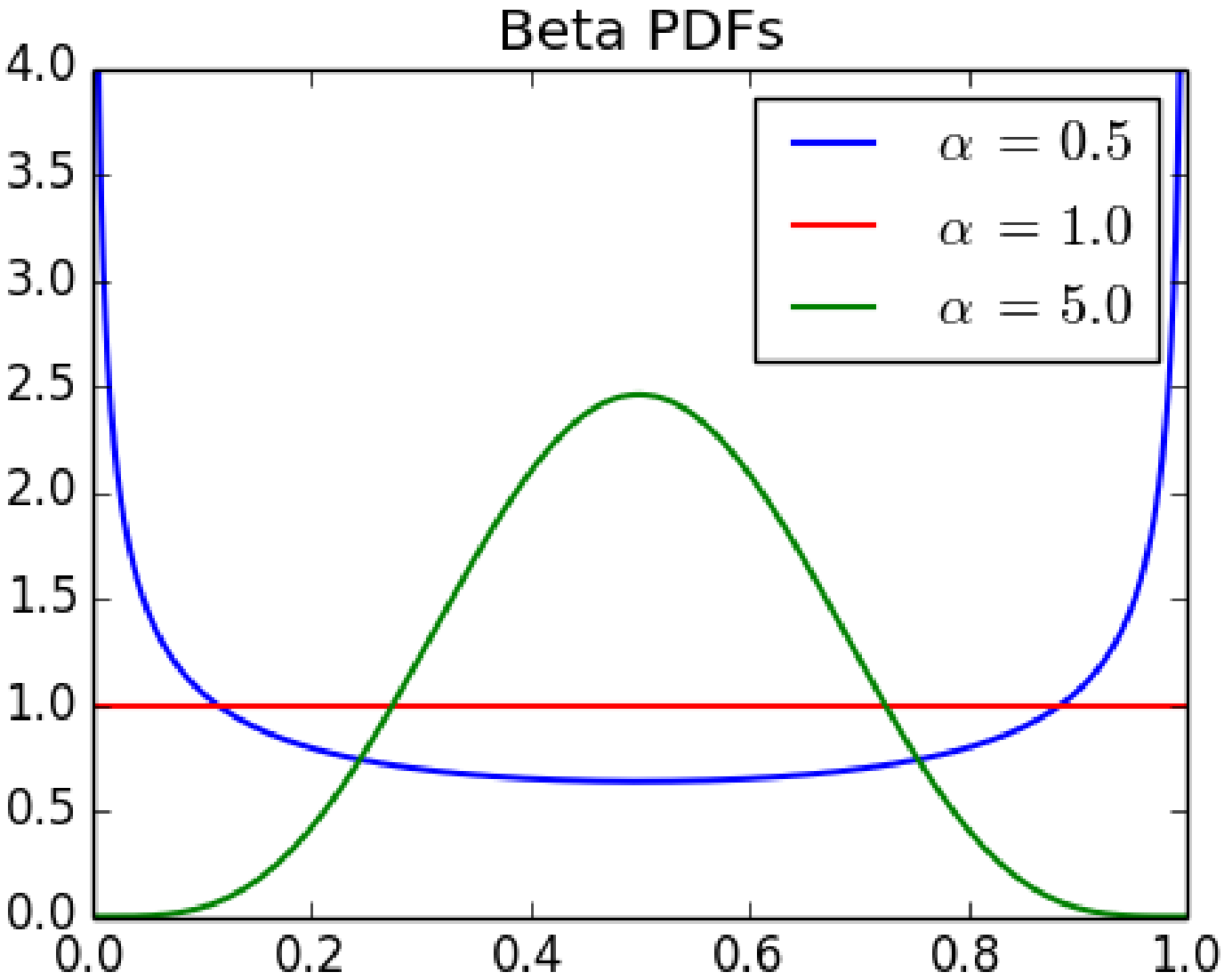
Consider a symmetric Dirichlet in which $\alpha_v = \alpha$ for all v. Agnostic about favoring one component over another.

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:

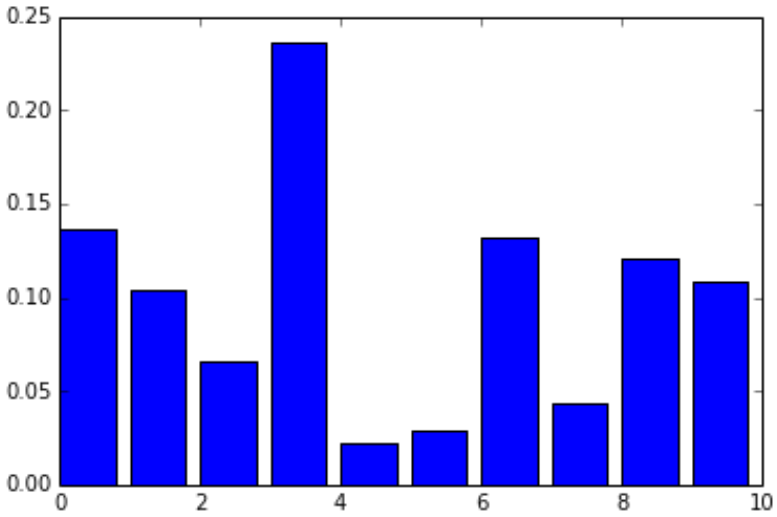
1. $\alpha = 1$ is a uniform distribution.
2. $\alpha > 1$ puts relatively more weight in center of simplex.
3. $\alpha < 1$ puts relatively more weight on corners of simplex.

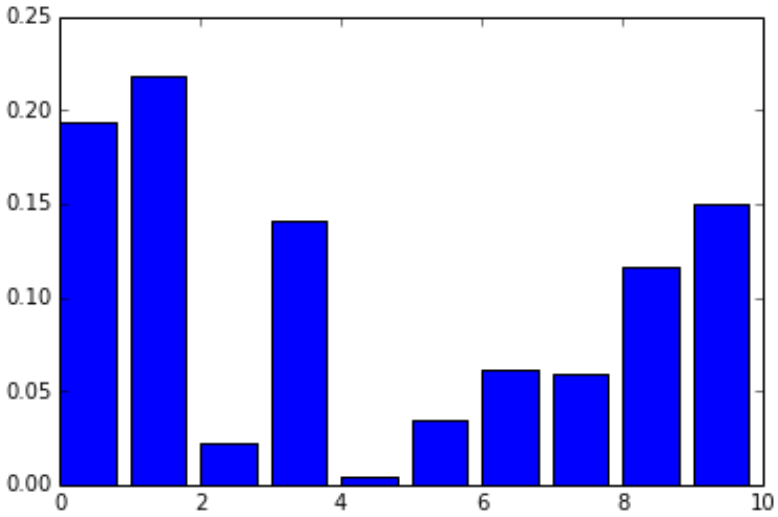
When $V = 2$, the Dirichlet is the beta distribution.

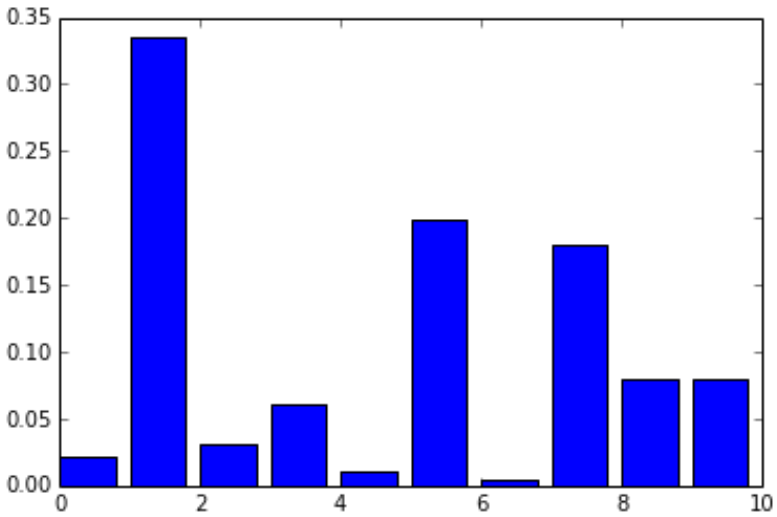
Beta with Different Parameters

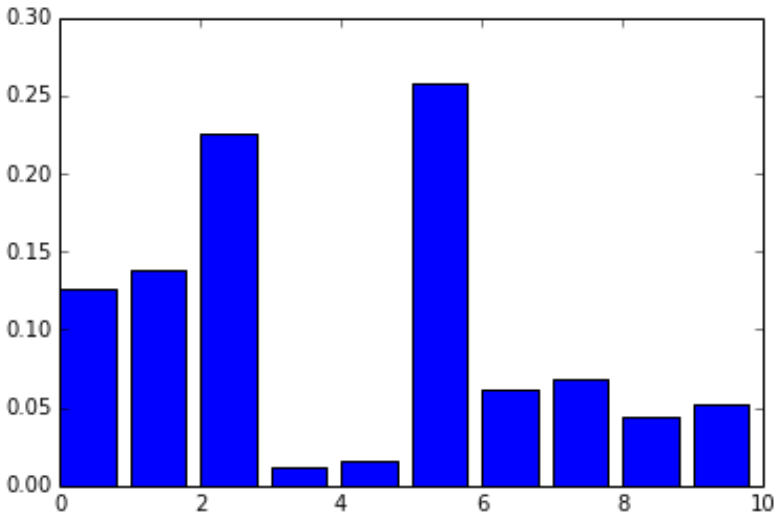


Draws from Dirichlet with $\alpha = 1$

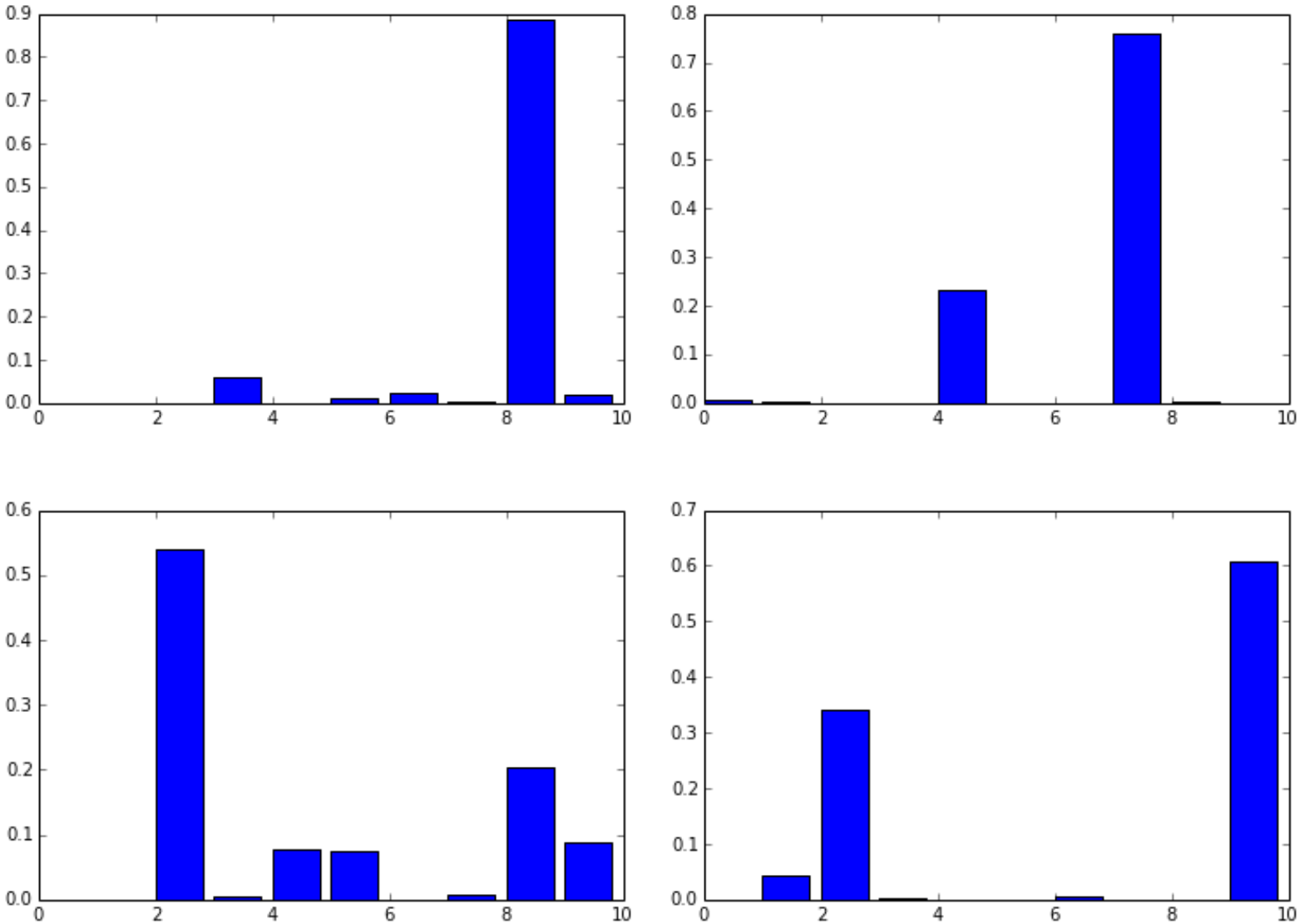




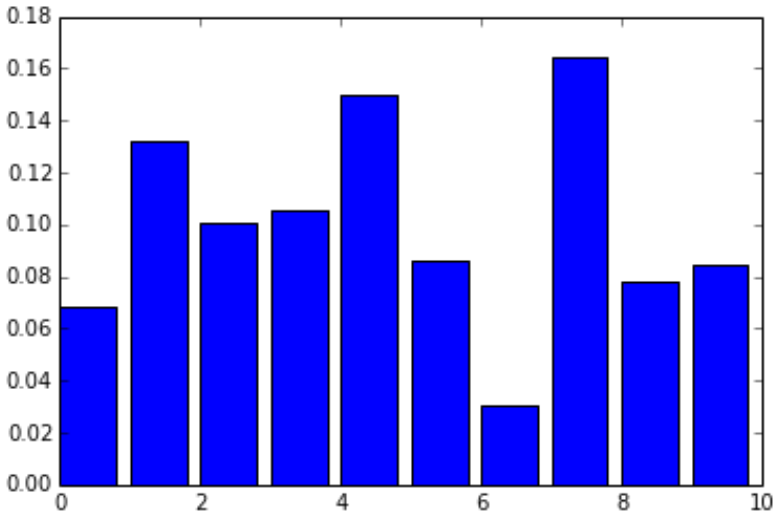


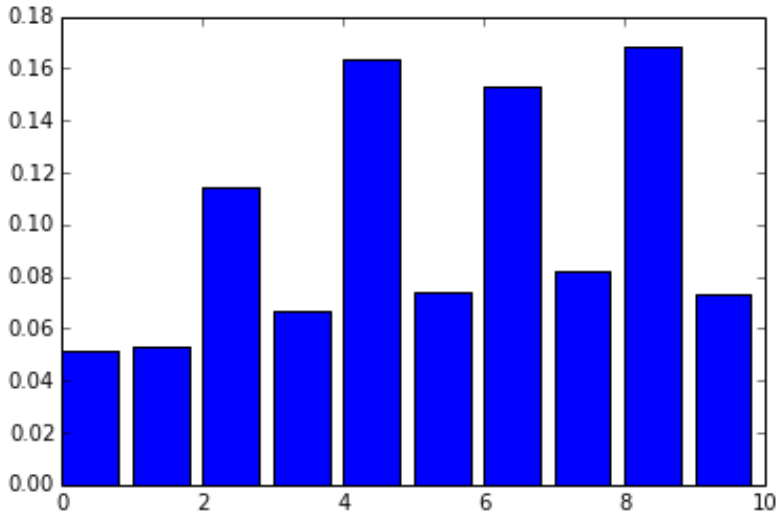


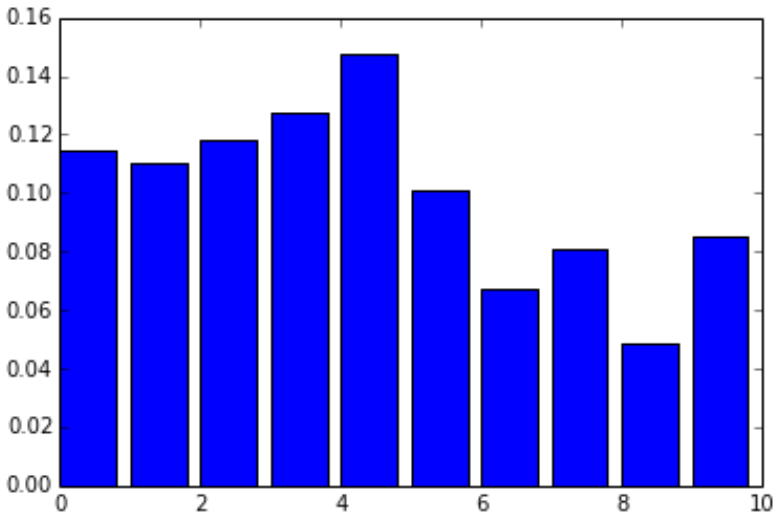
Draws from Dirichlet with $\alpha = 0.1$

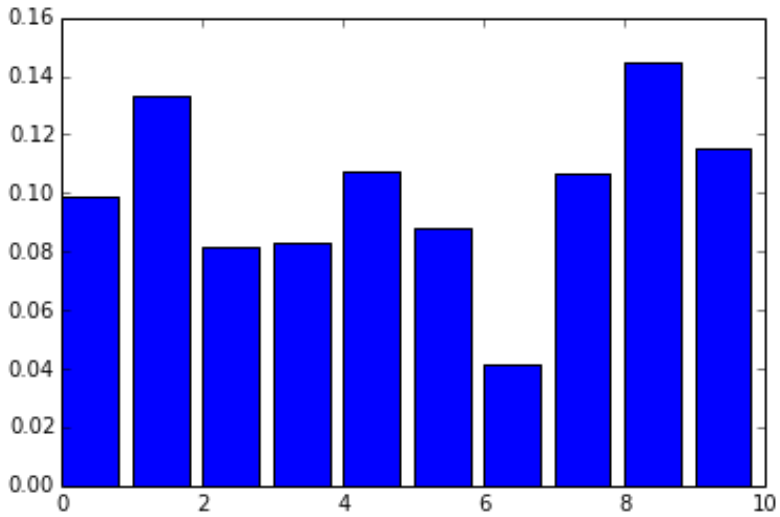


Draws from Dirichlet with $\alpha = 10$









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 $\mathbf{x}_d$.

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

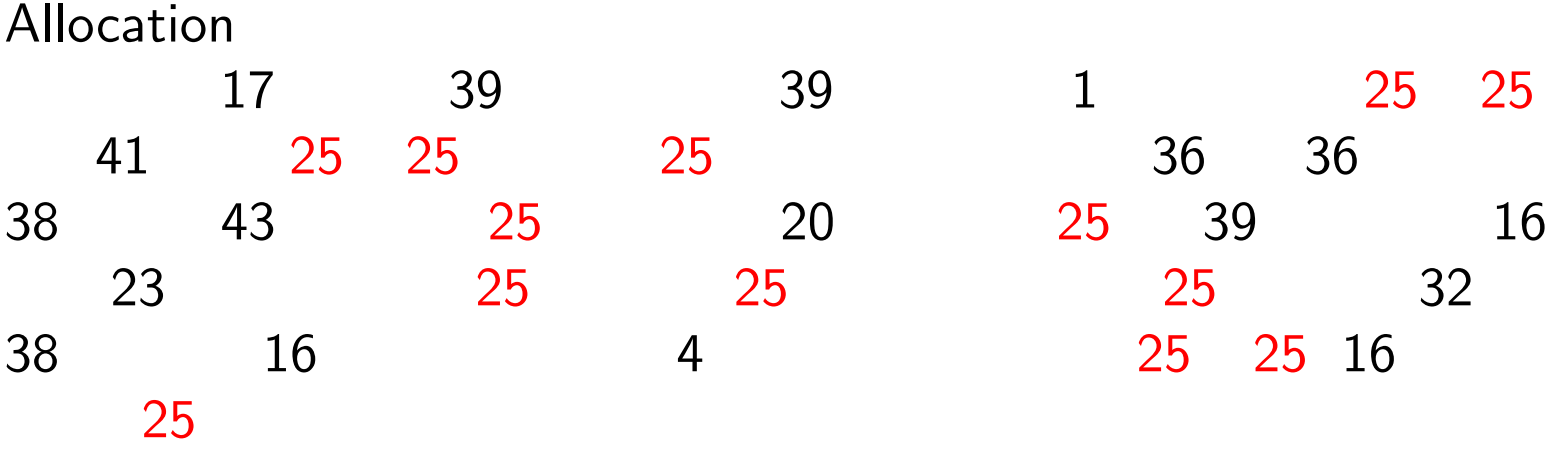
noticed change relationship between core CPI chained core CPI suggested maybe something going relating substitution bias upper level index focused nonmarket component PCE wondered something unusual happening core CPI relative measures

Example statement: Yellen, March 2006, #51

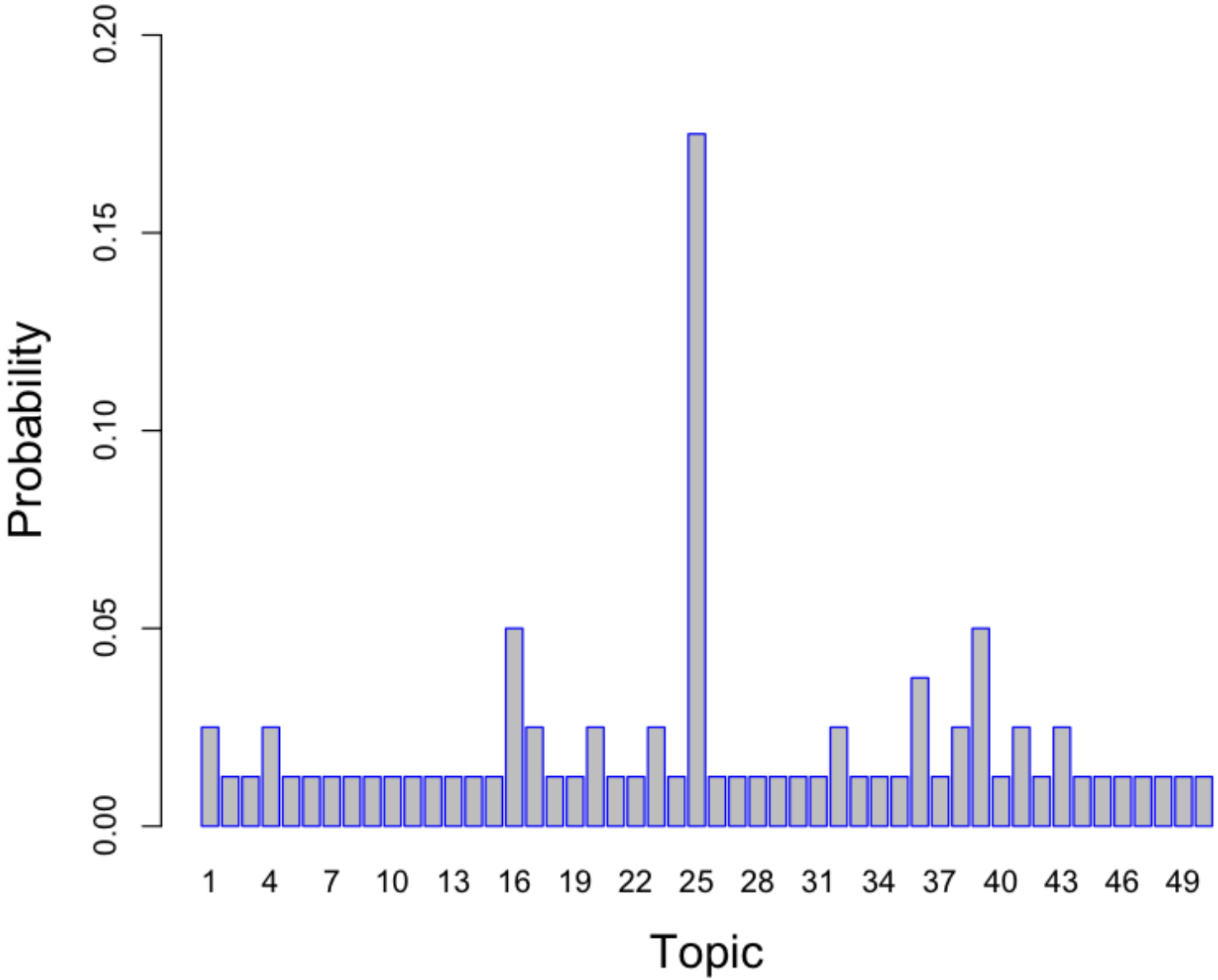
Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

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

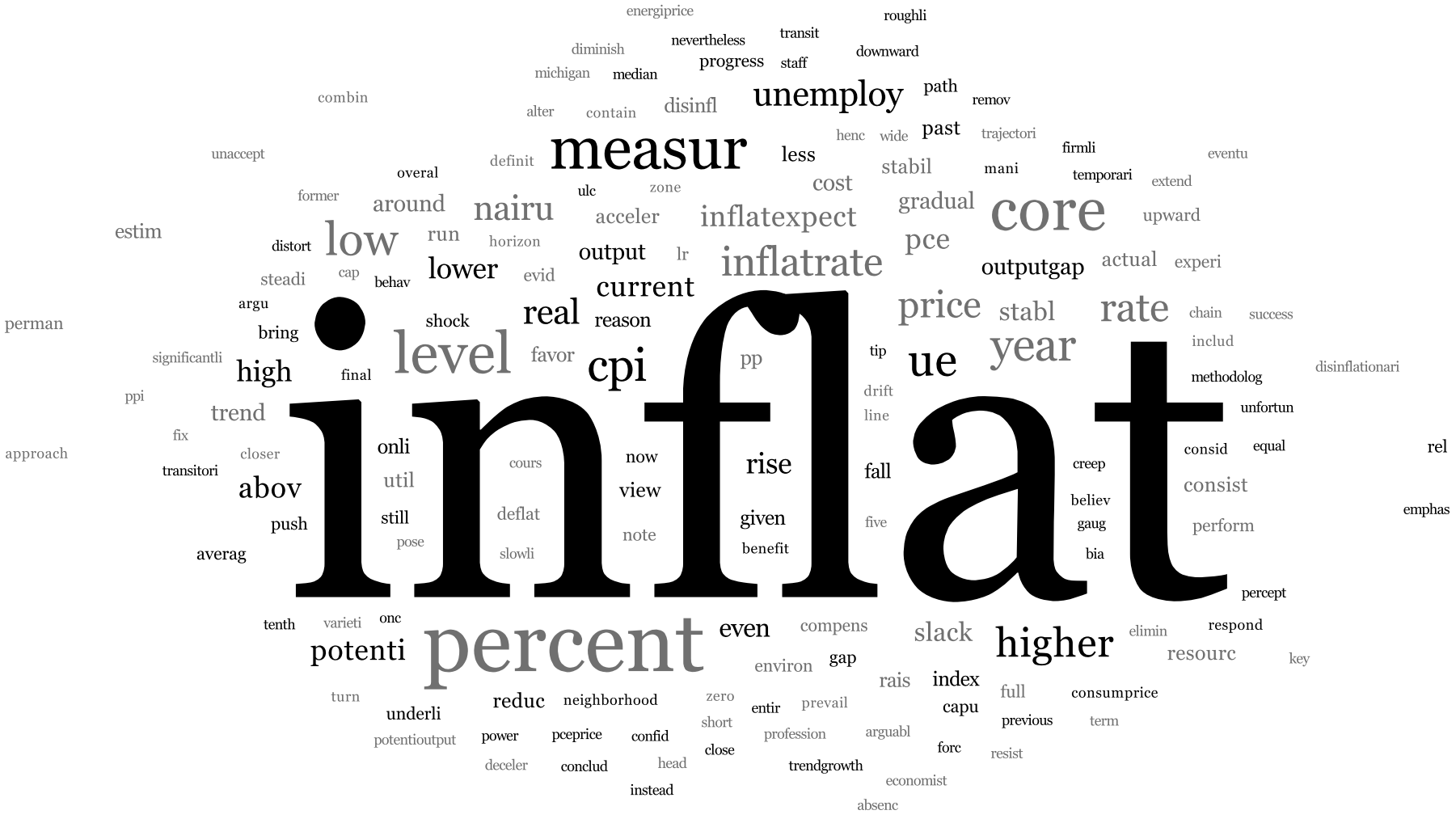
Example statement: Yellen, March 2006, #51



Distribution of Attention



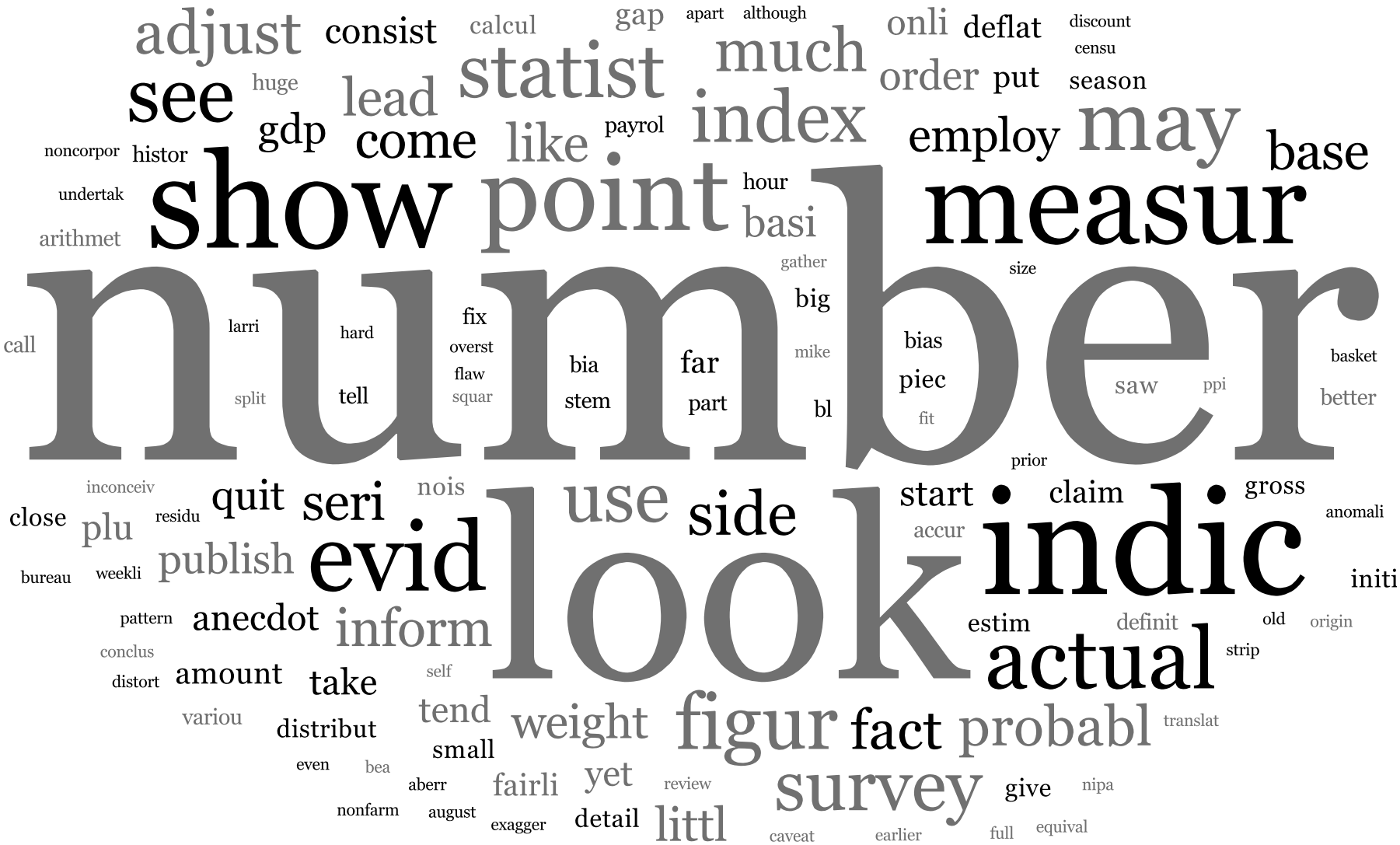
Topic 25


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

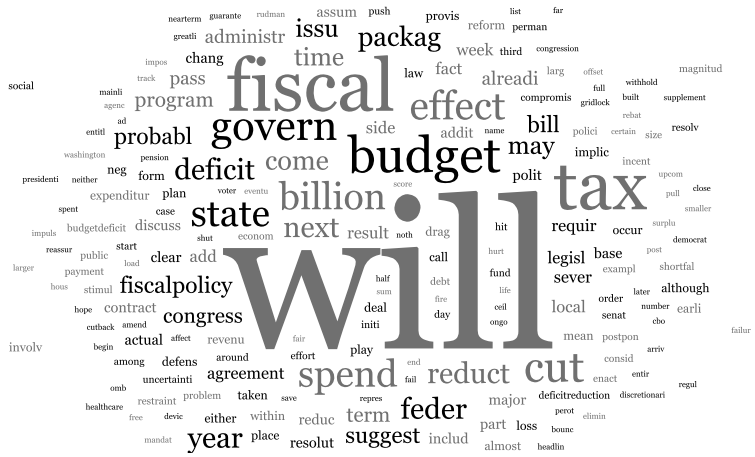
'measur' has probability 0.026 in topic 25, and probability 0.021 in topic 11.

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

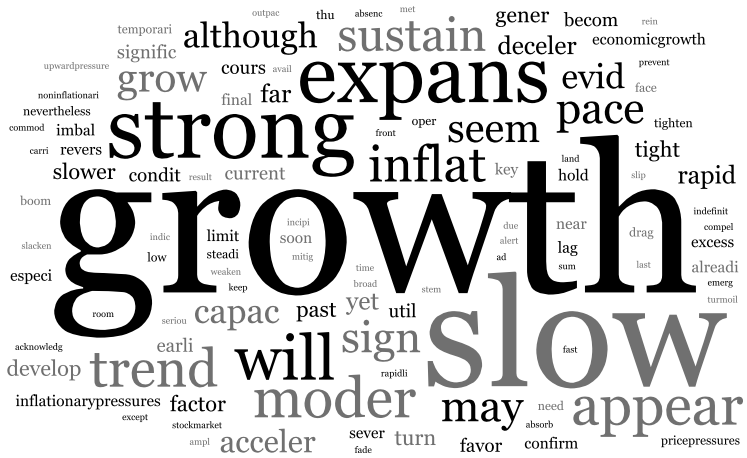
In statements containing words on evidence and numbers, it consistently gets assigned to 11.

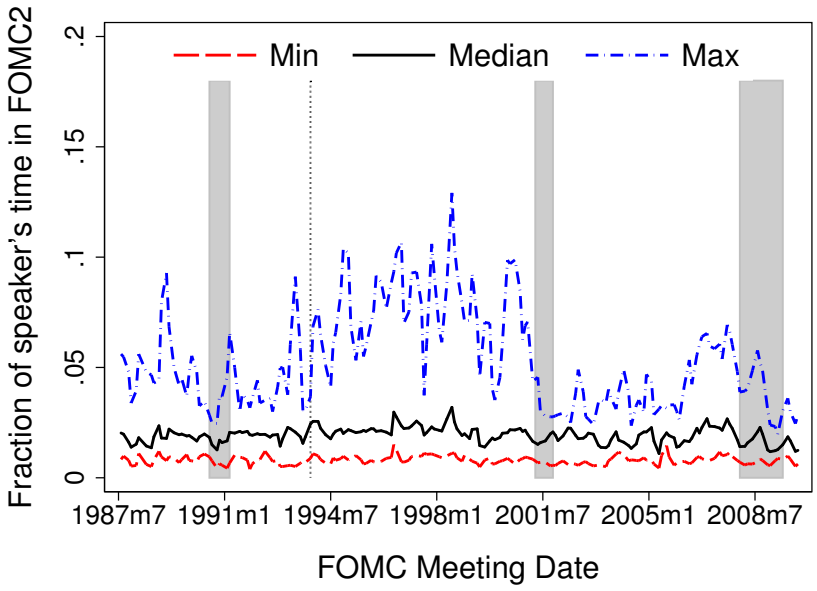
Sampling algorithm can help place words in their appropriate context.

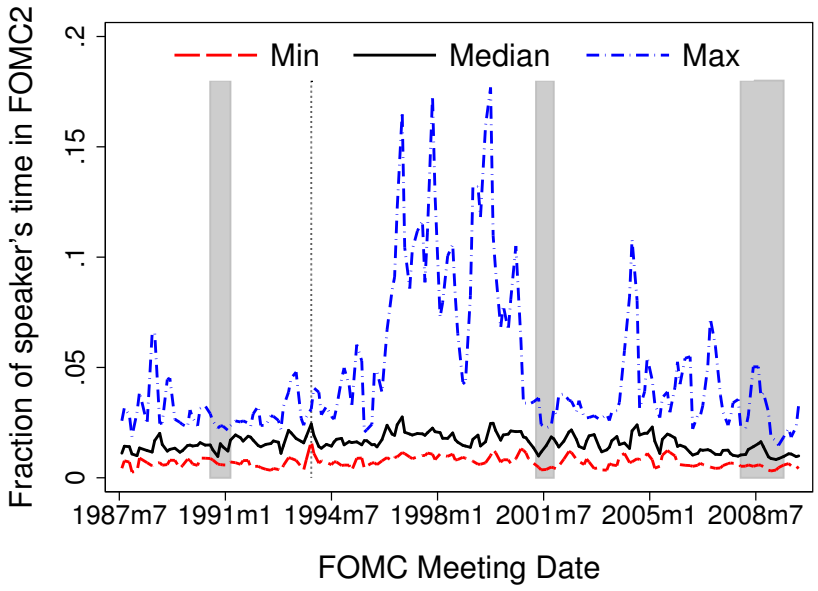
External Validation—BBD



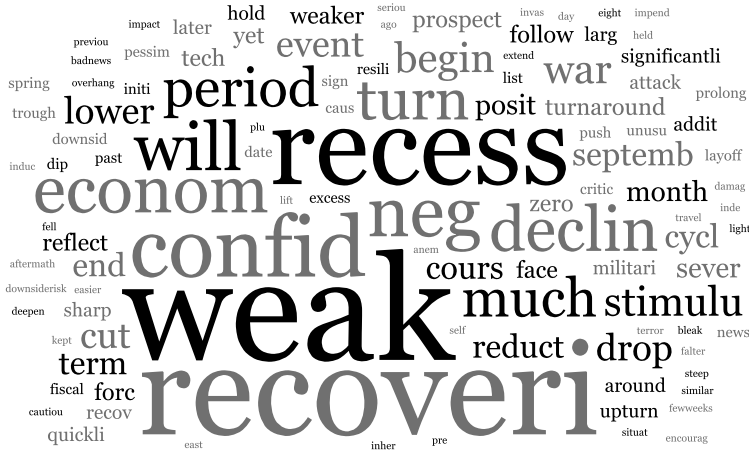
Pro-Cyclical Topics

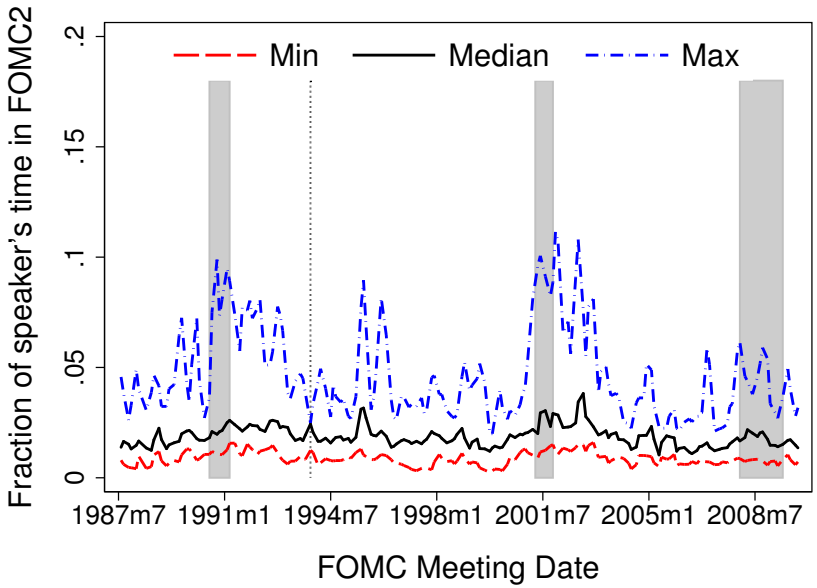


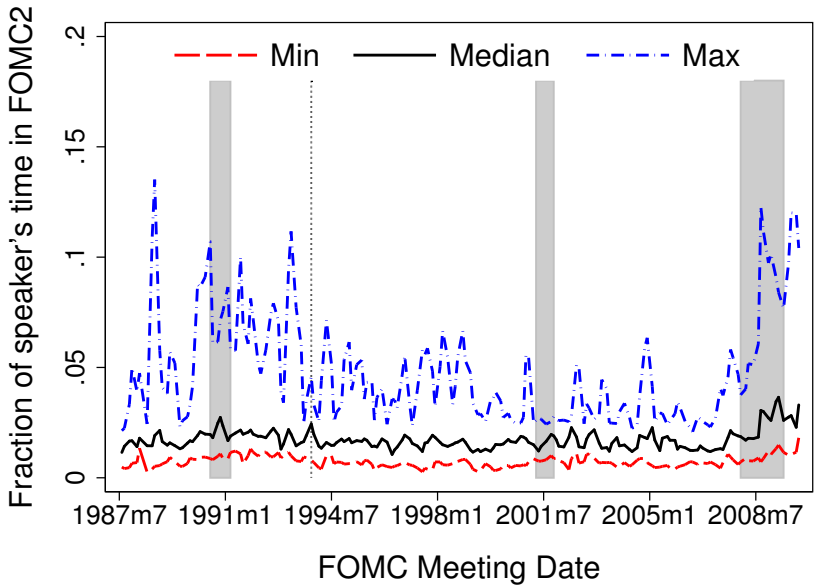




Counter-Cyclical Topics



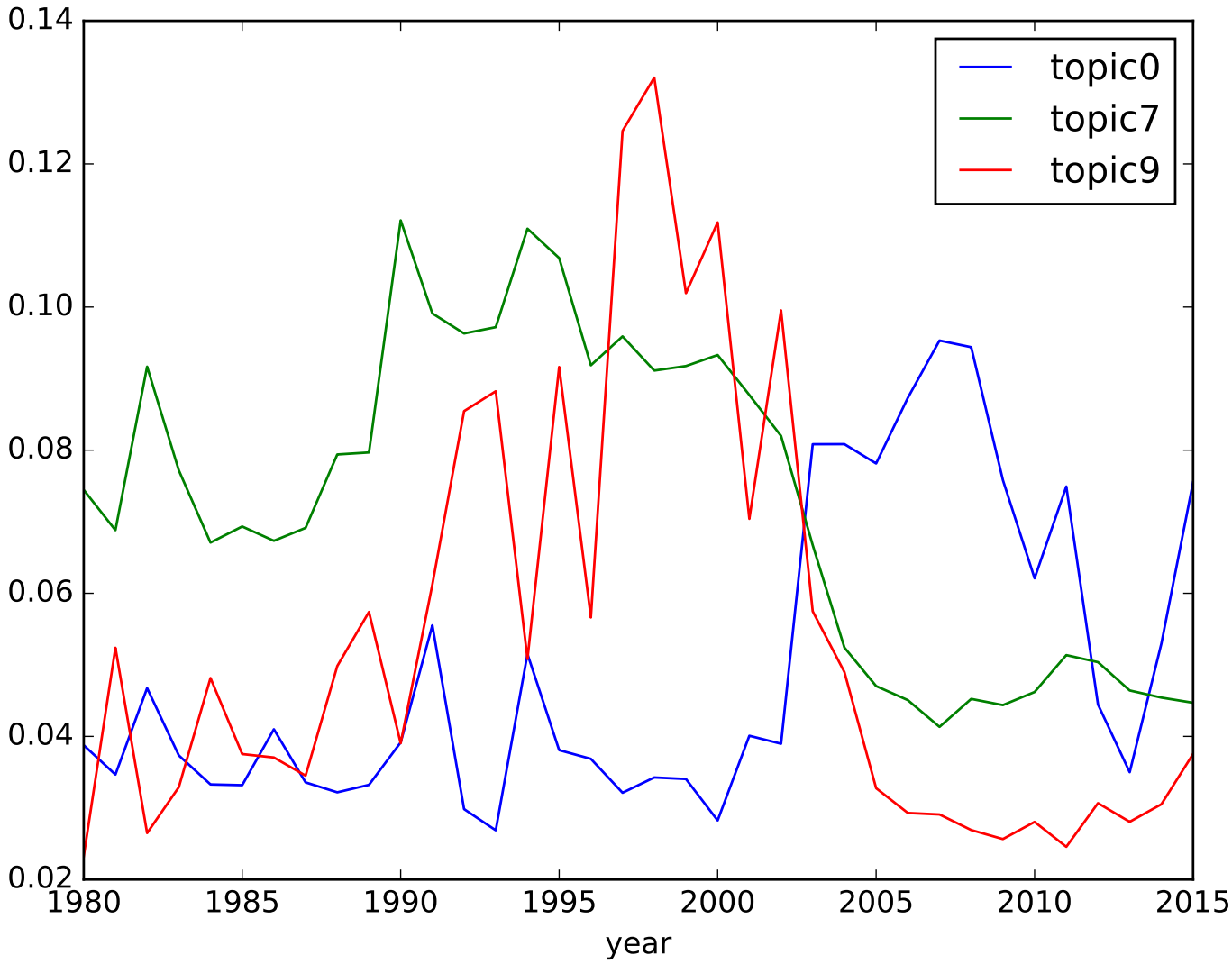




Topics on NYT Data (Iraq, Iran, Syria from mid-1980s)

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



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. Two main methods:

1. Markov Chain Monte Carlo (MCMC) provides a stochastic approximation to the true posterior.

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

2. Variational inference provides a deterministic solution for an approximate prior.

   General idea: define a family of simplified posteriors that approximate the true posterior, find optimal posterior in the simplified class.

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.

Gibbs Sampling for LDA

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.

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 $\boldsymbol{\beta}$.

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

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

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)[2] 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)[3] and technical appendix of Hansen, McMahon, and Prat (2015).

[2] PNAS, "Finding Scientific Topics".

[3] 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.

Variational Inference

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.

True and Approximate Distributions

Suppose we have observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$ (treat any parameters as fixed for now).

Let $p(\mathbf{x}, \mathbf{z})$ be their joint distribution.

Assume that $p(\mathbf{z} \mid \mathbf{x})$ is intractable to compute, for example because the latent space is too high-dimensional.

Let $q(\mathbf{z})$ be an approximate distribution over the latent variables. It will depend on variational parameters we suppress for now.

Kullback-Leibler Divergence

To measure the closeness of $p(\mathbf{z} \mid \mathbf{x})$ and $q(\mathbf{z})$, we can use the Kullback-Leibler divergence:

$$\mathbb{KL}(p \parallel q) = \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x}) \log \left[\frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right] \qquad \text{(forwards KL)}$$

$$\mathbb{KL}(q \parallel p) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(reverse KL)}$$

Forwards KL:

1. "Zero-avoiding"
2. Used in expectation propagation

Reverse KL:

1. "Zero-forcing" $\rightarrow$ better when multi-modal posterior
2. Used in variational inference

Forward vs Reverse KL (Bishop 2006)

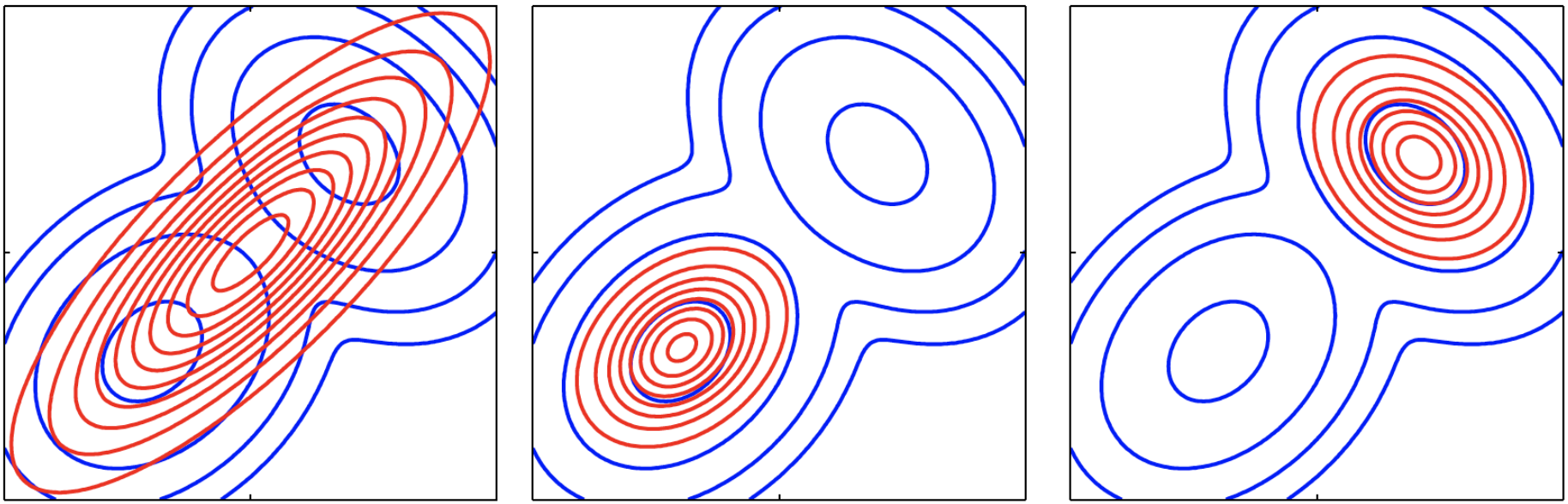


Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

KL Divergence and Evidence Lower Bound

Recall from the EM algorithm discussion that

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x}) / p(\mathbf{x})} \right] =$$

$$\underbrace{\log \left[p(\mathbf{x}) \right]}_{\text{log evidence}} - \underbrace{\left\{ \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{z}, \mathbf{x}) \right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[q(\mathbf{z}) \right] \right\}}_{\text{evidence lower bound (ELB)}} \geq 0.$$

KL Divergence and Evidence Lower Bound

Recall from the EM algorithm discussion that

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x}) / p(\mathbf{x})} \right] =$$

$$\underbrace{\log \left[p(\mathbf{x}) \right]}_{\text{log evidence}} - \underbrace{\left\{ \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{z}, \mathbf{x}) \right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[q(\mathbf{z}) \right] \right\}}_{\text{evidence lower bound (ELB)}} \geq 0.$$

$\log \left[p(\mathbf{x}) \right]$ is hard to compute, but does not depend on $q(\mathbf{z})$.

Minimize KL divergence = maximize ELB, which we can usually compute.

ELB is expected complete data log-likelihood plus entropy of approximating distribution.

The formal optimization problem is

$$\max_{q(\mathbf{z})} \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{z}, \mathbf{x}) \right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[q(\mathbf{z}) \right] \text{ such that } \sum_{\mathbf{z}} q(\mathbf{z}) = 1.$$

Comparison to EM Algorithm

In the EM algorithm, we take the expectation of the complete data log-likelihood with respect to the posterior distribution over $\mathbf{z}$ given fixed parameter values.

We use the true $p(\mathbf{z} \mid \mathbf{x})$ rather than the approximation $q(\mathbf{z})$, so the KL divergence is zero.

The ELB computed using true $p(\mathbf{z} \mid \mathbf{x})$ equals $\log [p(\mathbf{x})]$.

By contrast, with variational inference the ELB is not tight, but we want to make it as tight as possible.

Mean Field Approximation

The space of potential approximating distributions is large, so in practice some restrictions are made.

In mean-field approximation, we assume that q factorizes as

$$q(\mathbf{z}) = \prod_i q_i(z_i).$$

For simplicity, we assume each latent variable is independent, but can also have independent blocks of latent variables.

Independence assumptions implicit in mean field approximation generally not present in true posterior.

One can show that the optimal approximating distribution for z_i satisfies

$$q_i^*(z_i) \propto \exp\left(\mathbb{E}_{\mathbf{z}_{-i}}\left\{\log\left[p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})\right]\right\}\right).$$

Inference

The optimal update equation is a function of $q_j(z_j)$ for $j \neq i$.

Coordinate ascent algorithm: update each q_i term holding constant the current values of q_{-i}.

Use optimized q_i to approximate posterior distribution.

Interpretation

Form true conditional posterior $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$, then take expectation with respect to approximate distribution over the conditioning variables.

Close relationship to Gibbs sampling:

- In Gibbs sampling, we repeatedly sample values from $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$ to simulate true joint distribution.
- In variational inference, we instead average over $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$ rather than take samples.
- Benefit is that analytical averaging "stands in" for collecting many samples.
- But when z_i is strongly correlated with neighboring nodes, averaging distorts the estimated marginal $q_i(z_i)$.

Variational Bayes

Now suppose we wish to approximate posterior over both latent variables $\mathbf{z}$ and parameters $\boldsymbol{\theta}$ given data $\mathbf{x}$.

Mean field assumption is to approximate $p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{x})$ with $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{z}}(\mathbf{z})$, or $q(\boldsymbol{\theta}) \prod_i q_i(z_i)$ given conditional independence of latent variables.

Can implement VBEM algorithm by alternating between updating $q_i(z_i)$ given $q(\boldsymbol{\theta})$ (VB E-step), and updating $q(\boldsymbol{\theta})$ given $q_i(z_i)$ (VB M-step).

Distinction between latent variables and parameters becomes rather artificial, both are treated as unknown quantities and iteratively updated.

Variational Bayes and LDA

We can estimate LDA via Variational Bayes using the mean field approximation

$$p(\Theta, B, \mathbf{z} \mid \mathbf{w}) \approx \prod_{k=1}^{K} q(\beta_k \mid \lambda_k) \prod_{d=1}^{D} \left[q(\theta_d \mid \gamma_d) \prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n}) \right]$$

where:

- β_k is Dirichlet with parameters λ_k
- θ_d is Dirichlet with parameters γ_d
- $z_{d,n}$ is multinomial with parameters $\phi_{d,n}$

λ_k, γ_d, and $\phi_{d,n}$ are variational parameters we iteratively update according to the mean-field formula above.

Placing variational distributions in the same family as their priors is without loss of generality within exponential family (see Wainwright and Jordan 2008).

Update for γ_d

Recall from Gibbs sampling slides that

$$\theta_d \mid \mathbf{z}_d \sim \text{Dir}\left([\alpha + \sum_n \mathbb{1}(z_{d,n} = k)]_{k=1}^K \right)$$

so

$$\log\left[p(\theta_d \mid \cdot)\right] = \sum_k \left[\alpha - 1 + \sum_n \mathbb{1}(z_{d,n} = k) \right] \log\left(\theta_{d,k}\right) + \text{constant}.$$

Update for γ_d

Recall from Gibbs sampling slides that

$$\theta_d \mid \mathbf{z}_d \sim \text{Dir}\left([\alpha + \sum_n \mathbb{1}(z_{d,n} = k)]_{k=1}^K \right)$$

so

$$\log\left[p(\theta_d \mid \cdot)\right] = \sum_k \left[\alpha - 1 + \sum_n \mathbb{1}(z_{d,n} = k) \right] \log\left(\theta_{d,k}\right) + \text{constant}.$$

Taking expectations (ignoring constant) gives

$$\mathbb{E}_{\mathbf{z}_d}[\log\left[p(\theta_d \mid \cdot)\right]] = \sum_k \left[\alpha - 1 + \sum_n \phi_{d,n,k} \right] \log\left(\theta_{d,k}\right)$$

so optimal update is

$$\gamma^*_{d,k} = \alpha + \sum_n \phi_{d,n,k}.$$

Update for λ_k

Recall from Gibbs sampling slides that

$$\beta_k \mid \mathbf{z}, \mathbf{w} \sim \text{Dir}\left([\eta + \sum_d \sum_n \mathbb{1}(z_{d,n} = k)\mathbb{1}(w_{d,n} = v)]_{v=1}^V\right)$$

Again taking expectations of log probability, optimal update is

$$\lambda_{k,v}^* = \eta + \sum_d \sum_n \phi_{d,n,k} \mathbb{1}(w_{d,n} = v).$$

Updates for both θ_d and β_k very similar to those for Gibbs sampling, but replacing actual with expected counts.

Update for $\phi_{d,n}$

From the mean field formula and previous results on Gibbs sampling,

$$\phi_{d,n,k} \propto \exp\left(\mathbb{E}\left[\log(\beta_{k,v_{d,n}}\theta_{d,k})\right]\right).$$

Result on Dirichlet: $\mathbb{E}[\log(\theta_i)] = \Psi(\alpha_i) - \Psi(\sum_i \alpha_i)$, so

$$\phi^*_{d,n,k} \propto \exp\left(\Psi\left(\lambda_{k,v_{d,n}}\right) - \Psi\left(\sum_v \lambda_{k,v}\right) + \Psi\left(\gamma_{d,k}\right)\right).$$

(Ψ function is derivative of $\log(\Gamma)$, implemented in many scientific computing packages).

Overall Algorithm

Seed $\phi^1_{d,n,k} = 1/k$. Then at iteration s:

1. For each topic k (or randomly seed if $s = 1$)

$$\lambda^{s+1}_{k,v} = \eta + \sum_d \sum_n \phi^s_{d,n,k} \mathbb{1}(w_{d,n} = v)$$

2. For each document d

   2.1 $\gamma^{s+1}_{d,k} = \alpha + \sum_n \phi^s_{d,n,k}$

   2.2 For each word n in document d

$$\phi^{s+1}_{d,n,k} \propto \exp\left(\Psi\left(\lambda^{s+1}_{k,v_{d,n}}\right) - \Psi\left(\sum_v \lambda^{s+1}_{k,v}\right) + \Psi\left(\gamma^{s+1}_{d,k}\right)\right)$$

3. Check convergence of ELB, if not then proceed to iteration $s + 1$

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 [4]).

Methods to choose K:

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

[4] 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)[5] 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 $2(\mu_\ell - \sigma^2_\ell)$.

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

Formalizing Interpretablility

Chang et. al. (2009)[6] 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 $K = 50, 100, 150$.

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

Results

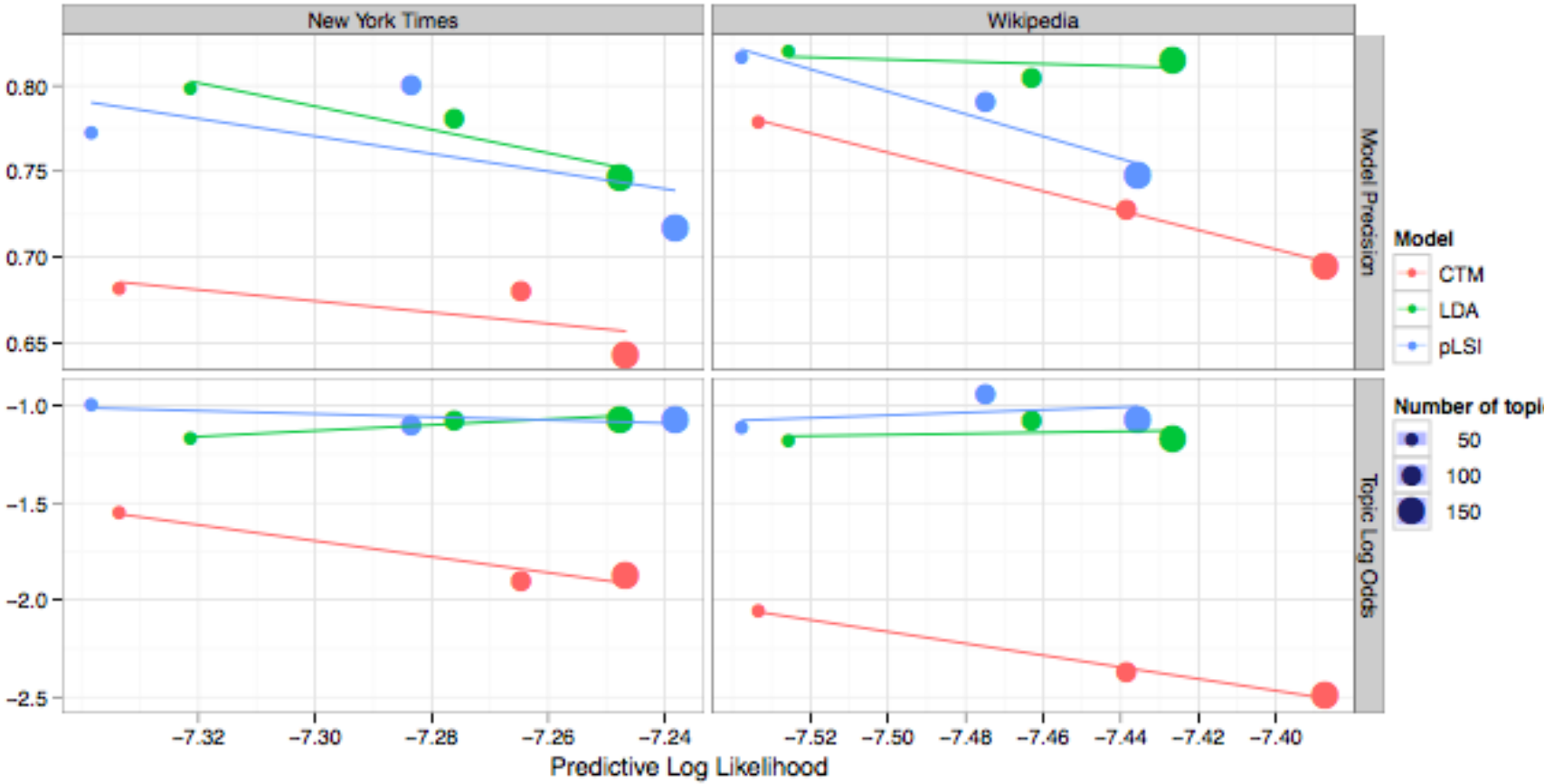


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase 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)[7] propose a method based on mutual pointwise information between top words in topics as computed via co-occurrence in Wikipedia.

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

Conclusion

LDA is a Bayesian model for modeling documents as mixtures of topics.

Provides nice example of approximate Bayesian inference in tractable framework.

Many extensions we have not had time to cover: correlated topic model, dynamic topic model, etc.

Inference algorithms become more complicated, but basic elements stay the same.