Text Mining for the Social Sciences
Lecture 6: LDA Inference with Gibbs Sampling

Stephen Hansen

Introduction

Last time we introduced the Dirichlet-Multinomial model for Bayesian inference with discrete data.

Exact inference for simple language models is easy due to conjugacy $\rightarrow$ add empirical counts to pseudo-counts.

LDA has latent variables, which makes posterior inference less straightforward.

Need to find ways of approximating posterior distribution when we can't compute it directly.

LDA Review

Variables		Parameters	Hyperparameters
Observed	Unobserved		
w	**z**	Θ, B	α, η

Posterior Distribution

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

$$\Pr[\mathbf{z} = z' \mid \mathbf{w}, \Theta, B] = \frac{\Pr[\mathbf{w} \mid \mathbf{z} = z', \Theta, B] \Pr[\mathbf{z} = z' \mid \Theta, B]}{\sum_{z'} \Pr[\mathbf{w} \mid \mathbf{z} = z', \Theta, B] \Pr[\mathbf{z} = z' \mid \Theta, B]}.$$

Posterior Distribution

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

$$\Pr[\mathbf{z} = z' \mid \mathbf{w}, \Theta, B] = \frac{\Pr[\mathbf{w} \mid \mathbf{z} = z', \Theta, B] \Pr[\mathbf{z} = z' \mid \Theta, B]}{\sum_{z'} \Pr[\mathbf{w} \mid \mathbf{z} = z', \Theta, B] \Pr[\mathbf{z} = z' \mid \Theta, B]}.$$

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

We will consider two main paradigms for approximate posterior inference:

1. MCMC, in particular Gibbs sampling.
2. Variational inference.

The former provides a stochastic approximation to the true posterior, while the latter provides an exact solution to an approximate posterior.

Gibbs Sampling Review

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.

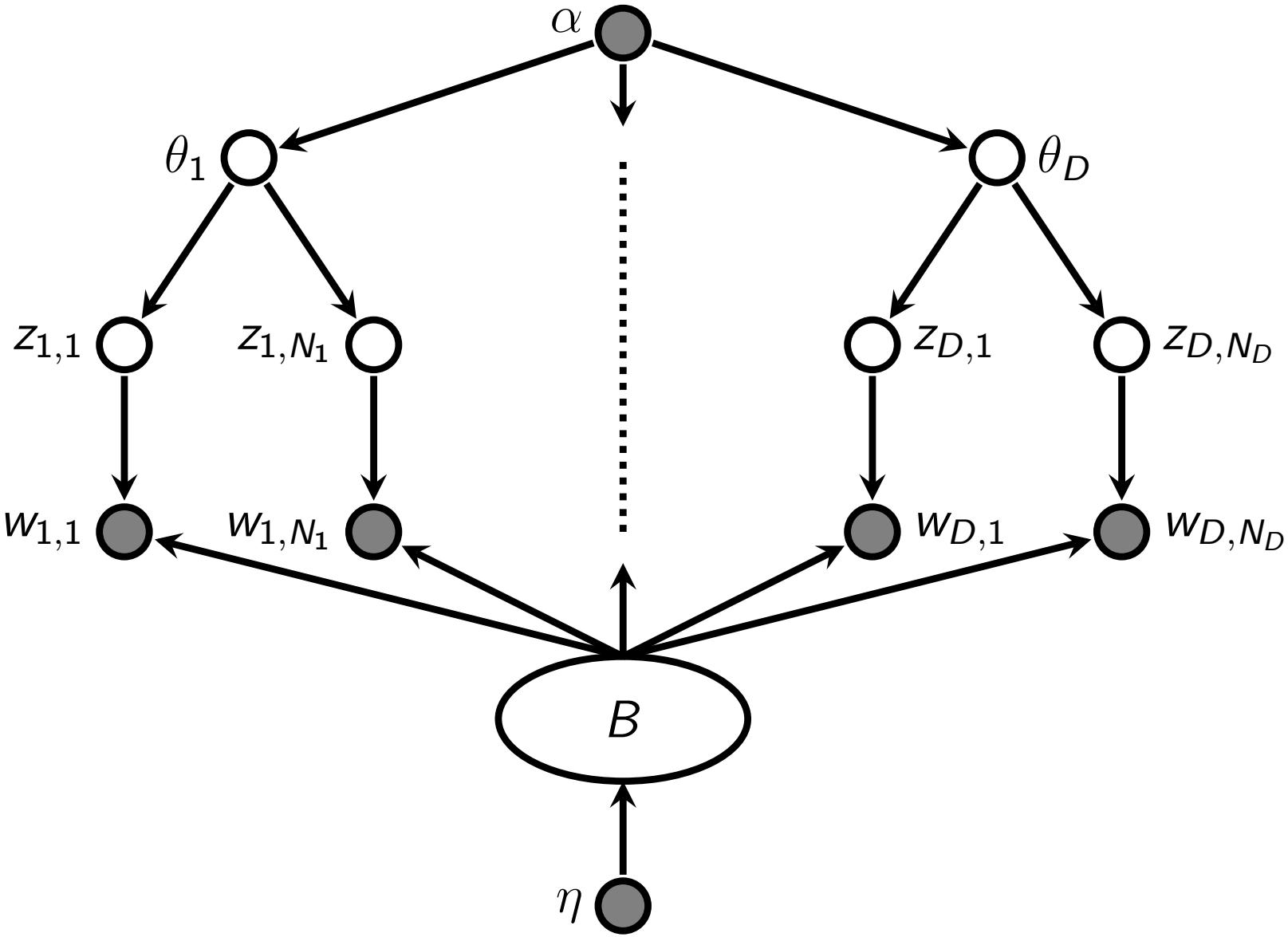
Markov Blanket

The *Markov blanket* of a node x in a Bayesian network is the set of nodes consisting of x's parents, children, and children's parents.

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

This can greatly simplify the expression for $f(x_i \mid \mathbf{x}_{-i})$.

LDA as a Directed Graph



Sampling Equations for Allocations

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

- The parent θ_d.
- The child $w_{d,n}$.
- The child's parents $\beta_1, \ldots, \beta_K$.

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

Sampling Equations for θ_d

The Markov blanket of θ_d is:

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

$\Pr[\theta_d \mid \alpha, \mathbf{z}_d]$ should look familiar from last lecture.

Let $n_d(k)$ is the number of allocations to k in document d.

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

Sampling Equations for β_k

The Markov blanket of β_k is:

- The parent η.
- The children $\mathbf{w}$.
- The children's parents $\mathbf{z}$.

$\Pr[\beta_k \mid \eta, \mathbf{w}, \mathbf{z}]$ should also be familiar.

Let $n_k(v)$ be the number of times topic k allocation variables generate term v.

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

Collapsed Sampling

Collapsed sampling refers to analytically integrating out some variables 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 Θ and B parameters and just samples topic allocations.

Recall that we can factor the LDA joint likelihood as

$$\Pr[B \mid \eta] \prod_d \Pr[\mathbf{w}_d \mid \mathbf{z}_d, B] \left(\prod_d \Pr[\mathbf{z}_d \mid \theta_d] \Pr[\theta_d \mid \alpha] \right)$$

Integrating out Θ

We can express $\Pr[\mathbf{z}_d \mid \theta_d] = \prod_k \theta_{d,k}^{n_d(k)}$.

From results in the previous lecture on evidence in Dirichlet-Multinomial model

$$\int_{\theta_d} \prod_k \theta_{d,k}^{n_d(k)} \operatorname{Dir}(\theta_d \mid \alpha)\, d\theta_d = \frac{\Gamma(\alpha K)}{\Gamma(\alpha K + N_d)} \frac{\prod_k \Gamma(\alpha + n_d(k))}{\prod_k \Gamma(\alpha)}.$$

So we can replace the term in parentheses in the factored likelihood by

$$\left[\frac{\Gamma(\alpha K)}{\Gamma^K(\alpha)}\right]^D \prod_d \frac{\prod_k \Gamma(\alpha + n_d(k))}{\Gamma(\alpha K + N_d)}.$$

Integrating out B

We can express $\Pr[\mathbf{w}_d \mid \mathbf{z}_d, B] = \prod_v \prod_k \beta_{k,v}^{n_{d,k}(v)}$ where $n_{d,k}(v)$ is the number of times topic k allocation variables in document d generate term v.

Then $\prod_d \Pr[\mathbf{w}_d \mid \mathbf{z}_d, B] = \prod_v \prod_k \beta_{k,v}^{n_k(v)}$.

Evidence is

$$\int_{\beta_k} \prod_v \beta_{k,v}^{n_k(v)} \operatorname{Dir}(\beta_k \mid \eta)\, d\beta_k = \frac{\Gamma(\eta V)}{\Gamma(\eta V + \sum_v n_k(v))} \frac{\prod_v \Gamma(\eta + n_k(v))}{\prod_v \Gamma(\alpha)} \Rightarrow$$

$$\Pr[B \mid \eta] \prod_d \Pr[\mathbf{w}_d \mid \mathbf{z}_d, B] = \left[\frac{\Gamma(\eta V)}{\Gamma^V(\eta)}\right]^K \prod_k \frac{\prod_v \Gamma(\eta + n_k(v))}{\Gamma(\eta V + \sum_v n_k(v))}.$$

Collapsed Sampling Equation

We can now express the joint likelihood of **w** and **z** alone.

Heinrich (2009) and technical appendix of Hansen, McMahon, and Prat (2015) derive sampling equation for $z_{d,n}$ from the joint likelihood:

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

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

Interpretation

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.

2. 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 joint distribution of allocation variables, but we usually care more about parameters we integrated out.

Their predictive distributions are easy to form given topic assignments (recall results in last lecture):

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

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. Standard cross-validation exercise to choose K.
2. Interpretability. General versus specific.

In FOMC transcript analysis, choose $K = 50$.

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.

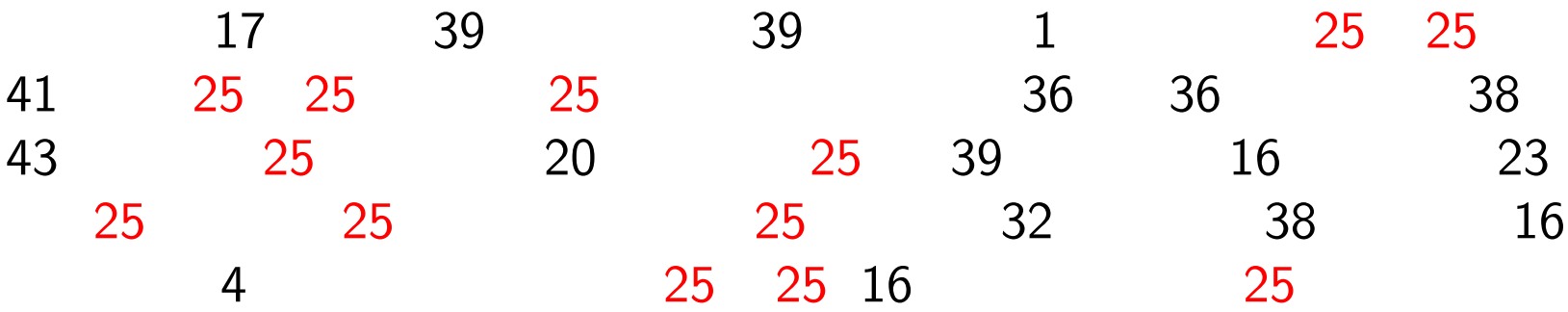
Example statement: Yellen, March 2006, #51

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

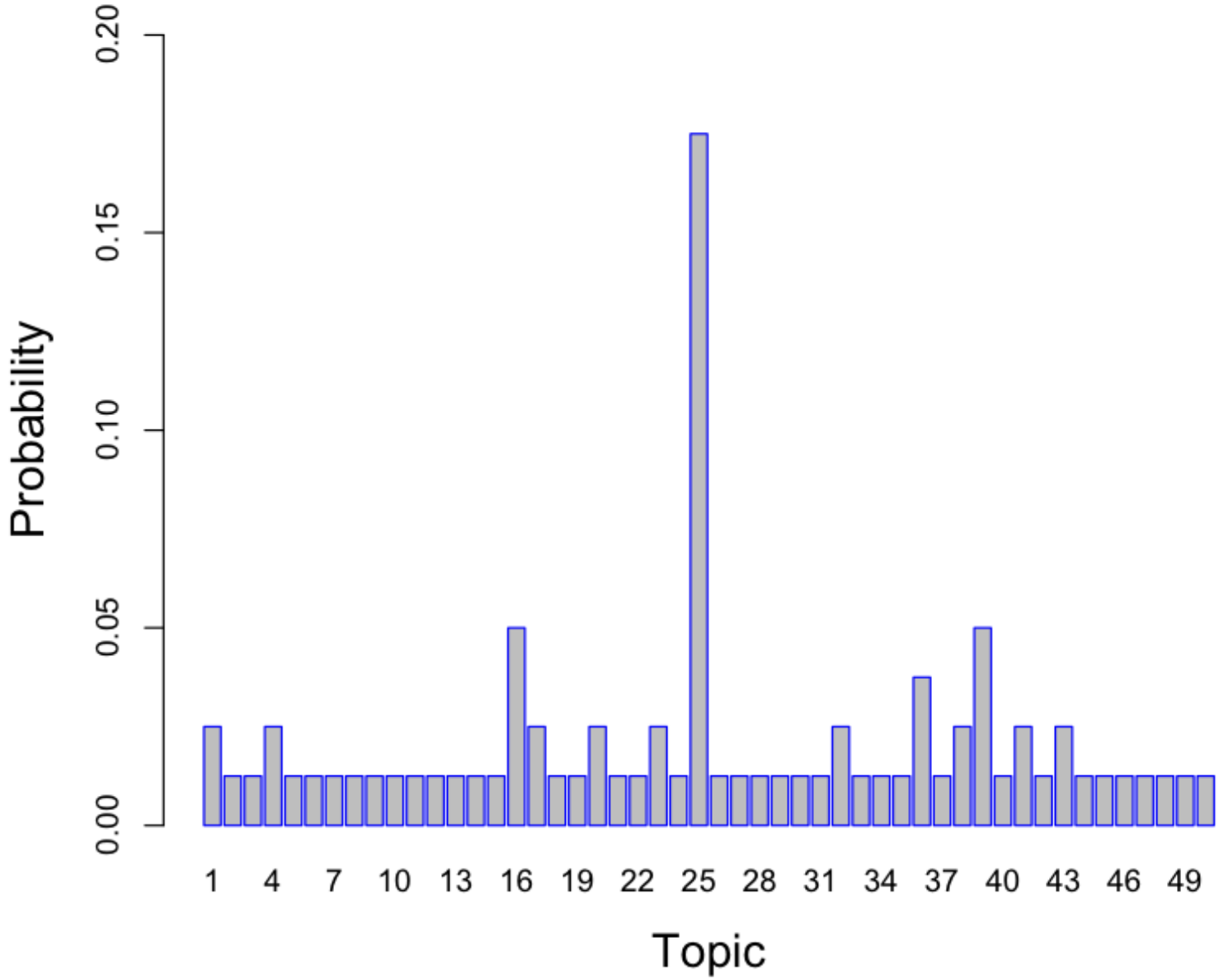
Example statement: Yellen, March 2006, #51

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

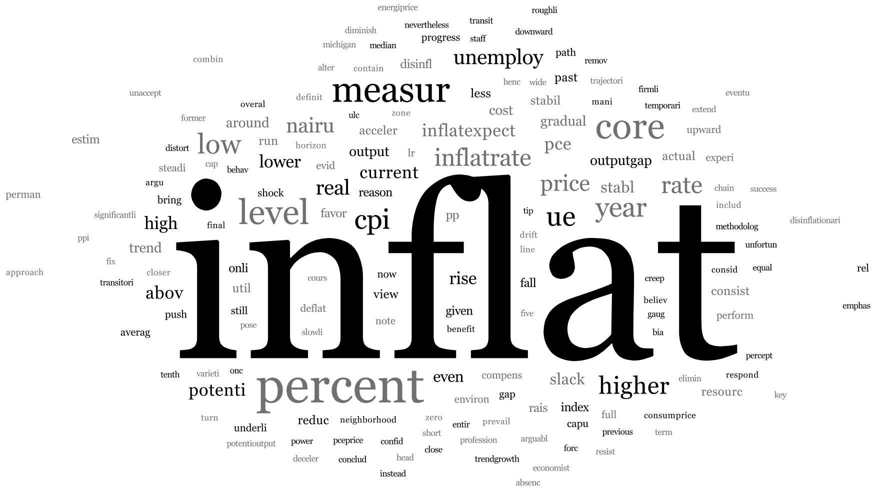
Example statement: Yellen, March 2006, #51



Distribution of Attention



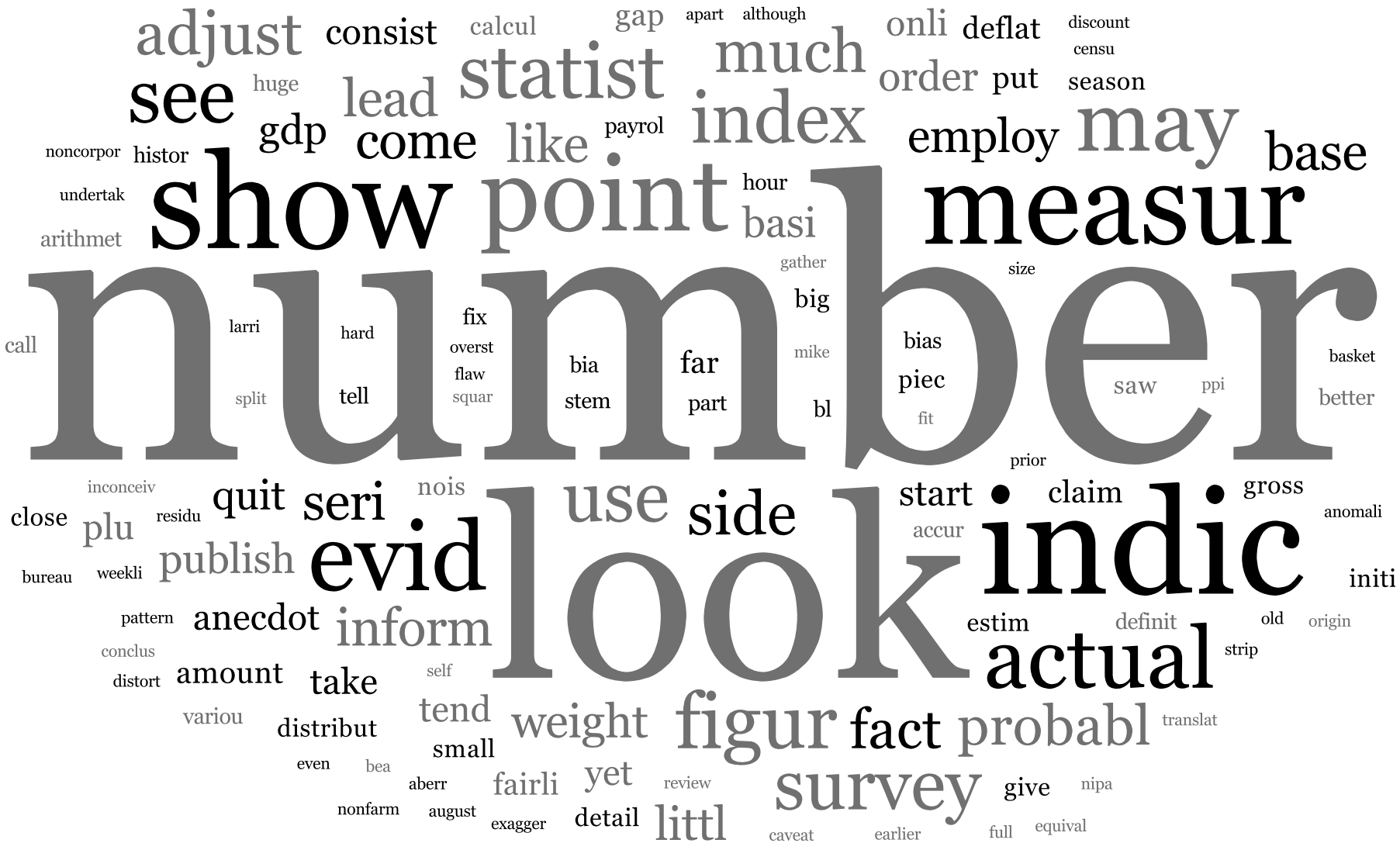
Topic 25


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

Chain Convergence and Selection

Determining when a chain has converged can be tricky.

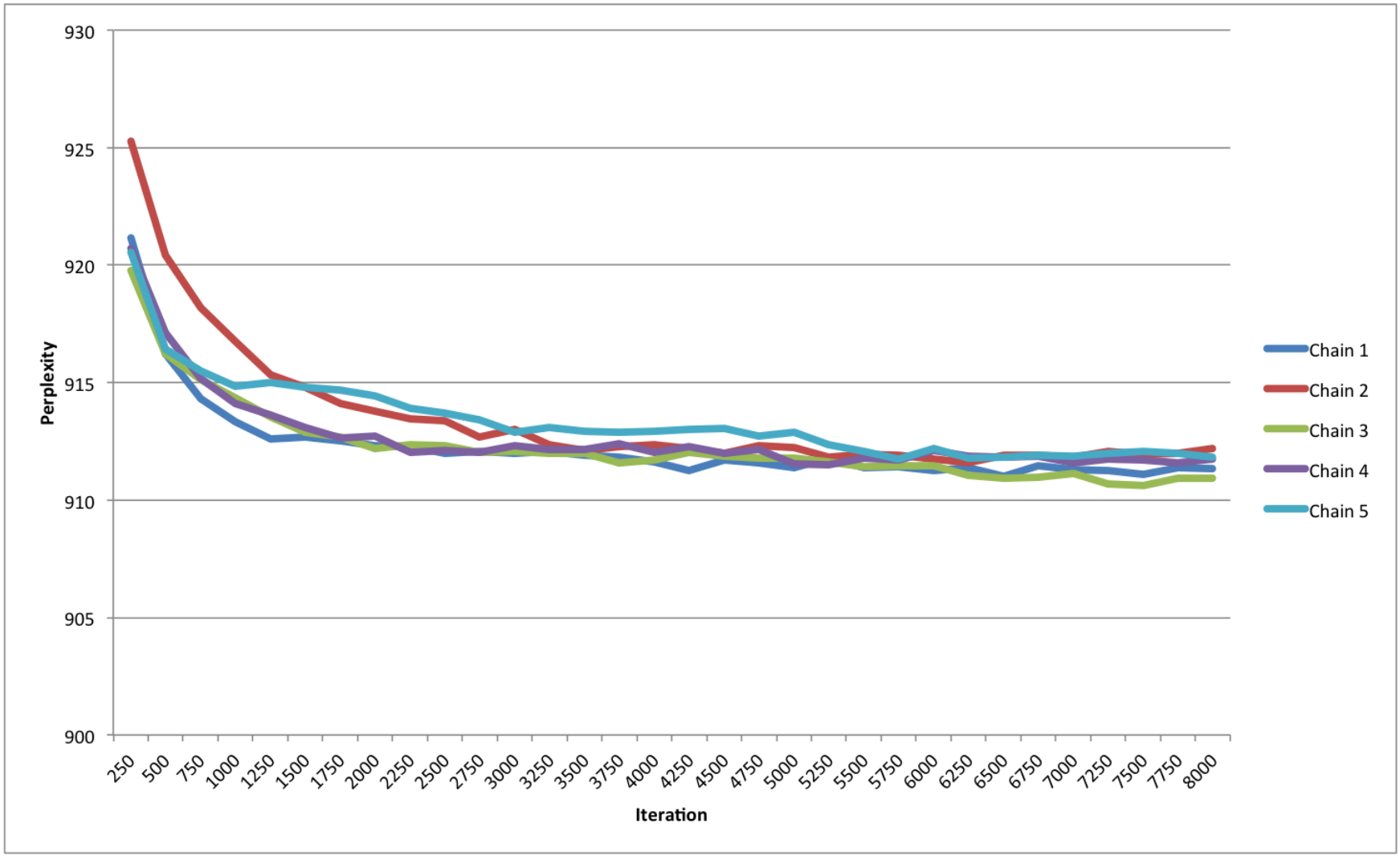
One approach is to measure how well different states of the chain predict the data, and determine convergence in terms of its stability.

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

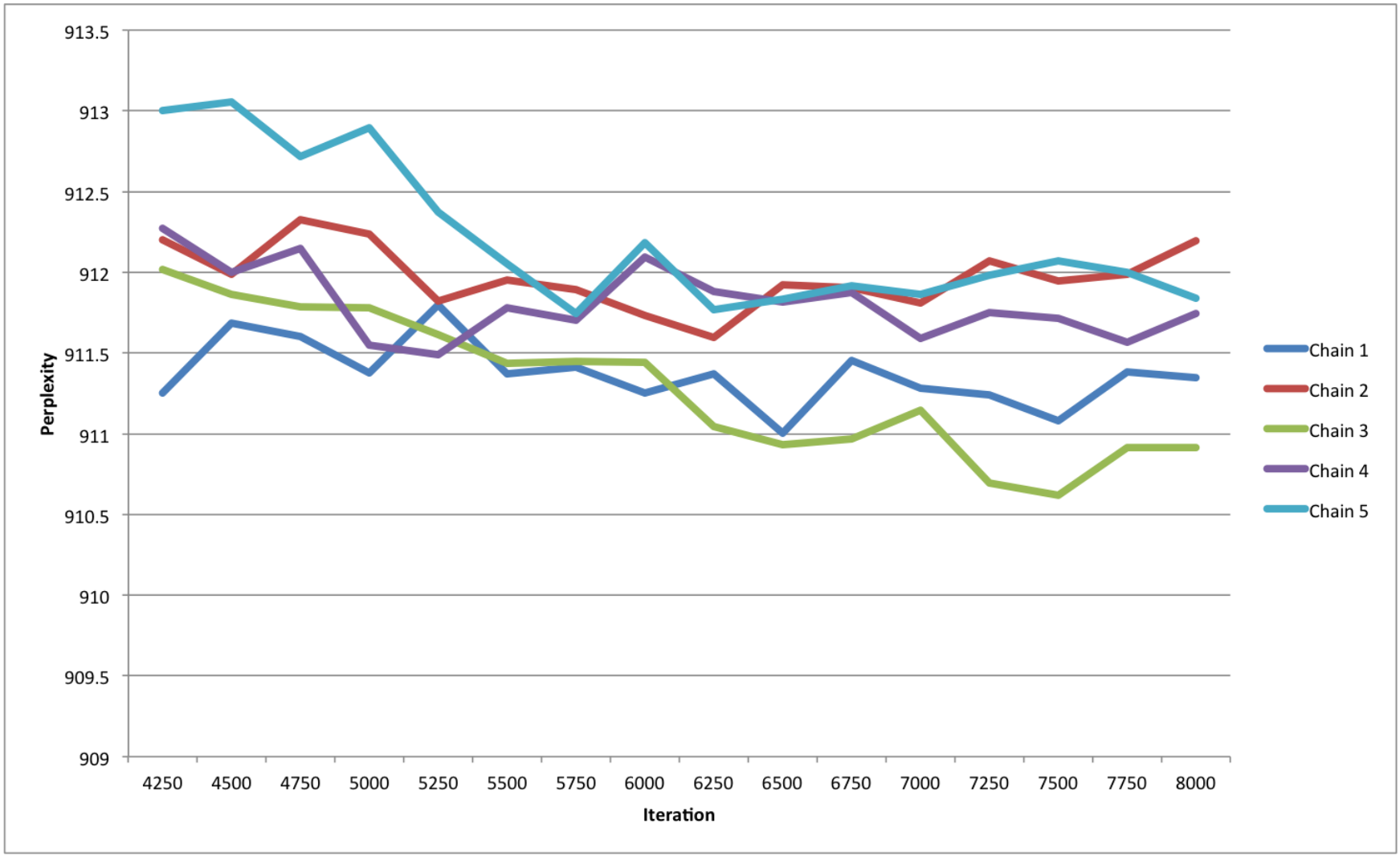
For LDA, a typical goodness-of-fit measure is *perplexity*, given by

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

Perplexity 1



Perplexity 2



Out-of-Sample Documents

As with the vector space model and LSA, we may be interested in out-of-sample documents.

Under the assumption that a document lies in the same topic space as the within-sample documents, can perform Gibbs sampling treating estimated topics as fixed

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

for each out-of-sample document d'.

Only 10-20 iterations necessary since topics already estimated.

Conclusion

The posterior distribution for LDA cannot be computed directly.

But we can nevertheless exploit the conjugacy of the Dirichlet-Multinomial model from the previous lecture to build a simple sampling algorithm.

The downside is that we abandon exact solutions for stochastic approximations.

More sophisticated algorithms exploit sparseness in the count matrices to speed up sampling substantially.