Topic Model

Modeling Environment

What does it mean to understand/model your environment?

Ability to *predict*

Two approaches to modeling environment of words and text

- Latent Semantic Analysis (LSA)
- Topic Model

LSA

The set up

D documents
W distinct words
F = WxD coocurrence matrix
f_{wd} = frequence of word w in document d

Transforming the coocurrence matrix

$$g_{wd} = \log\{f_{wd} + 1\}(1 - H_w) \qquad H_w = -\frac{\sum_{d=1}^{D} \frac{f_{wd}}{f_{w\cdot}} \log\{\frac{f_{wd}}{f_{w\cdot}}\}}{\log D}$$

where f_{wd}/f_w is probability that randomly chosen instance of w in corpus comes from document d

H_w = value in 0-1 where
0=word appears in only 1 doc;
1=word spread across all documents

$(1-H_w)$ = specificity:
0 = word tells you nothing about the document;
1= word tells you a lot about the document

G = WxD normalized coocurrence matrix
log transform common for word freq analysis
+1 ensures no log(0)
weighted by specificity

Representation of word i

row i of G

problem: this is high dimensional

problem: doesn’t capture similarity structure of documents

Dimensionality reduction via SVD

G = A D B

[WxD] = [WxR] [RxR] [RxD]

if R = min(W,D) reconstruction is perfect

if R < min(W,D) least squares reconstruction, i.e., capture whatever structure there is in matrix with a reduced number of parameters

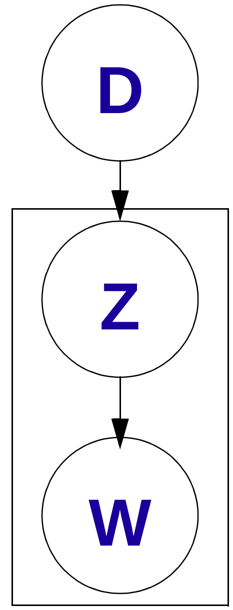
Reduced representation of word i: row i of (AD)

Can used reduced representation to determine semantic relationships

Topic Model (Hoffmann, 1999)

Probabilistic model of the way language is produced

Generative model



Select a document with probability P(D)

Select a (latent) topic with probability P(Z|D)

Generate a word with probability P(W|Z)

Produce pair $<d_i, w_i>$ on draw i

$P(D, W, Z) = P(D)\ P(Z|D)\ P(W|Z)$

$P(D, W) = \mathrm{sum}_z\ P(D)\ P(z|D)\ P(W|z)$

Learning

Minimize cross entropy (difference between distribution) of data and model

$-\ \mathrm{sum}_x\ Q(x) \log P(x)$

$= \mathrm{sum}_{w,d}\ n(d,w)\ P(d,w)$

Topic Model (Griffiths & Steyvers): Notation

$$P(w_i) = \sum_{j=1}^{T} P(w_i | z_i = j) P(z_i = j)$$

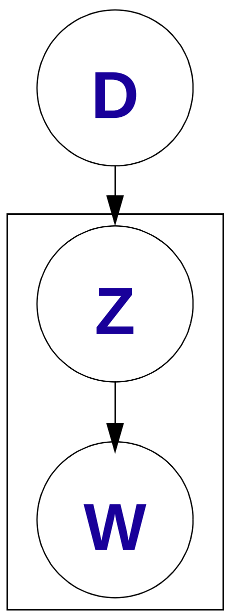
$P(w_i)$ is the same as $P(W=w_i \mid D=d_i)$

$P(z_i=j)$ is same as $P(Z=j \mid D=d_i)$ is the same as $\theta_j^{d_i}$

$P(w_i \mid z_i=j)$ is same as $P(W=w_i \mid Z=j)$ is the same as $\phi_{w_i}^{j}$

Thus, equation means

$P(W|D) = \text{sum}_j\, P(W|D, Z=j)\, P(Z=j|D) = \text{sum}_j\, P(W|Z=j)\, P(Z=j|D)$



Topic Model (Griffiths & Steyvers) Doing the Bayesian Thing

The two conditional distributions aare over *discrete alternatives.*

P(Z=j | D=d_i) or $\theta_j^{d_i}$

P(W=w_i | Z=j) or $\phi_{w_i}^{j}$

If *n* alternatives, distribution can be represented by *n–1* parameters.

But suppose you don't represent the distribution directly but rather you do the Bayesian thing of representing a whole bunch of models—a distribution of distributions...

Intuitive Example

Coin with unknown bias, ρ = probability of heads

Sequence of observations: H T T H T T T H

Maximum likelihood approach

$\rho = 3 / 8$

Bayesian approach

set of models $M = \{m_\rho\}$, where probability associated with m_ρ is ρ

e.g., $M = \{m_{0.0}, m_{0.1}, m_{0.2}, \ldots, m_{1.0}\}$

Bayesian Model Updates

Bayes rule

$$\overset{\text{posterior}}{p(m|D)} = \frac{\overset{\text{likelihood}}{p(D|m)}\,\overset{\text{prior}}{p(m)}}{p(D)}$$

Likelihood model

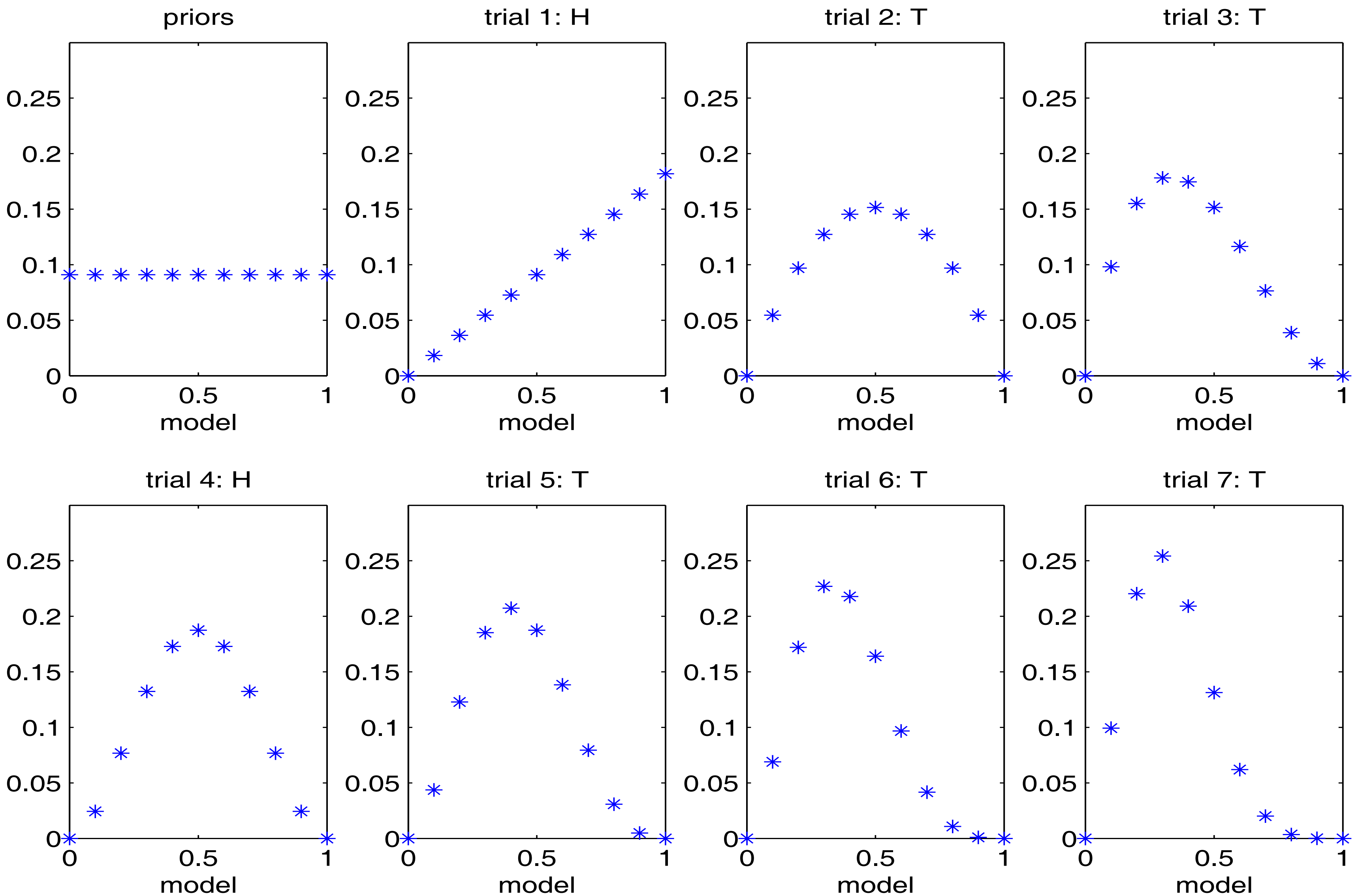
$$p(\text{head}|m_\rho) = \rho$$

$$p(\text{tail}|m_\rho) = 1-\rho$$

Priors

$$p(m_\rho) = \frac{1}{11}$$

Coin Flip Sequence: H T T H T T T



Infinite Model Spaces

This all sounds great if you have just a few models, but what if you have infinite models?

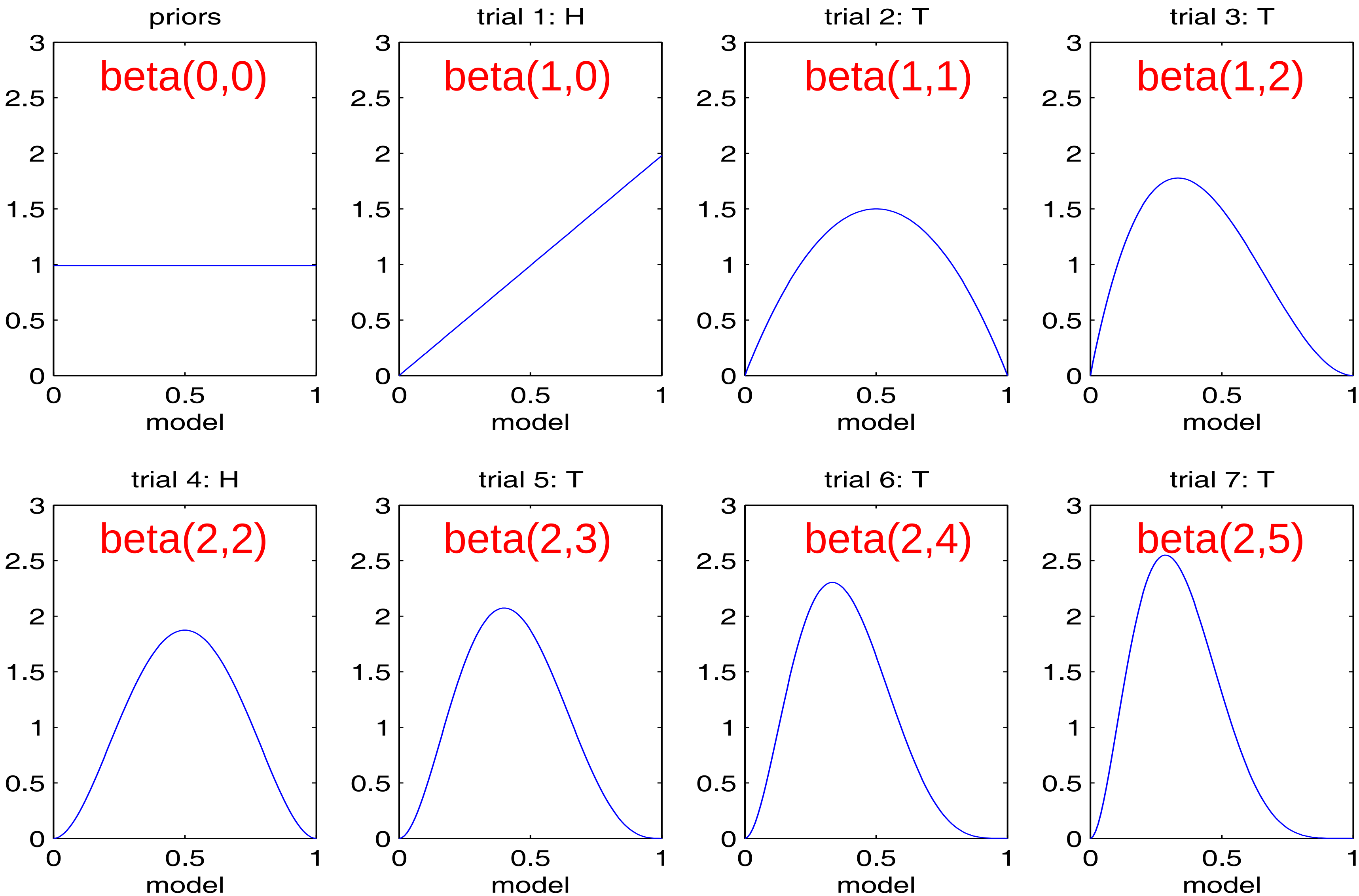
e.g., ρ continuous in [0, 1]

If you limit the form of the probability distributions, you can often do so efficiently.

e.g., beta distribution to represent priors and posterior in coin flip example

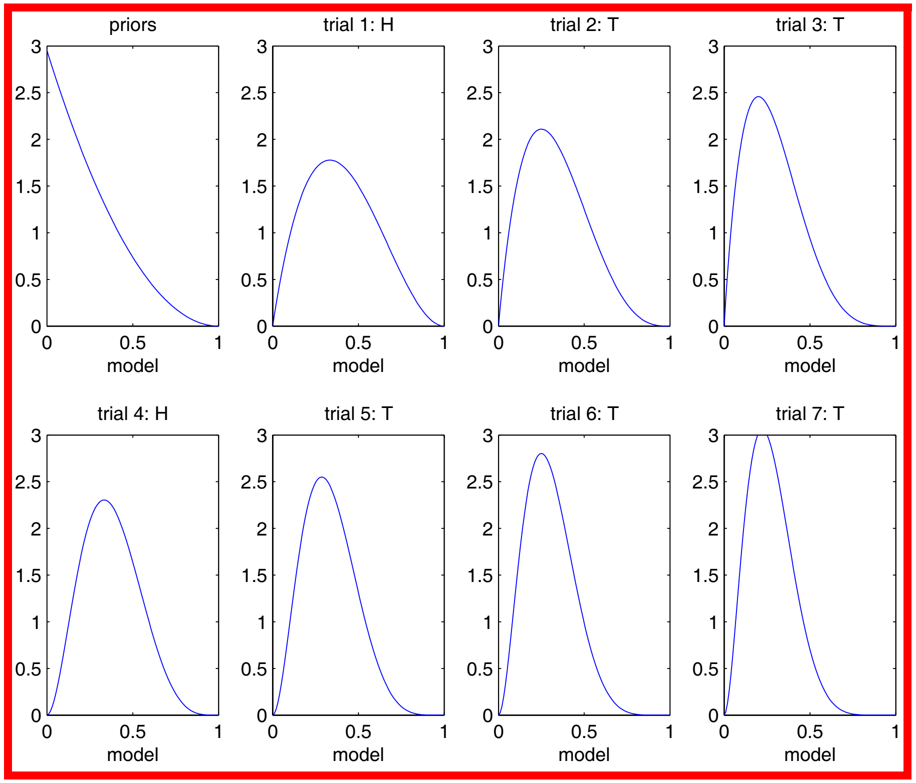
Requires only *two* parameters to update, one representing count of heads, one representing count of tails.

Coin Flip Sequence: H T T H T T T

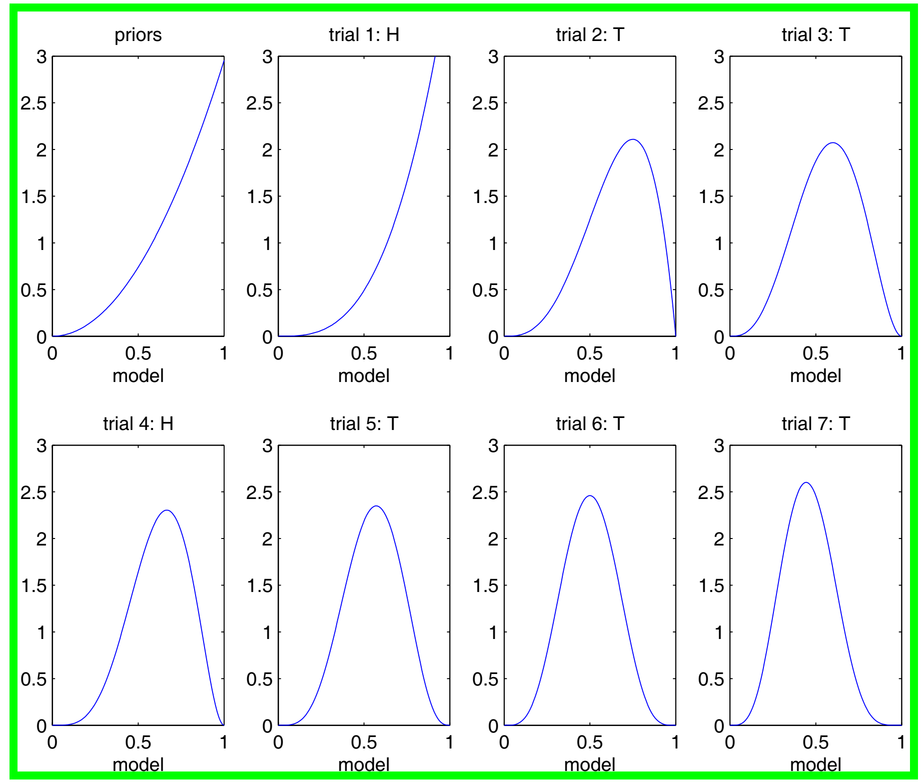


Effect of Prior Knowledge

low head-probability bias



high head-probability bias



Dirichlet Distribution

Dirichlet distribution is a generalization of beta distribution.

Beta distribution is a conjugate prior for a binomial RV; Dirichlet is a conjugate prior for a multinomial RV

Rather than representing uncertainty in the probabilities over *two* alternatives, a Dirichlet represents uncertainty in the probabilities over *n* alternatives.

You can think of the uncertainty space over *n* probabilities constrained such that $P(x) = 0$ if $(\text{sum}_i\ x_i) \mathrel{!}= 1$ or if $x_i < 0$...

...or the representational space over *n–1* probabilities constrained such that $P(x)=0$ if $(\text{sum}_i\ x_i) > 1$ or if $x_i < 0$.

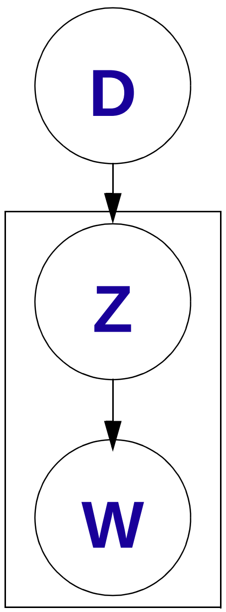
Dirichlet for multinomial RV with *n* alternatives has *n* parameters (beta has 2).

Each parameter is a count of the number of occurrences.

Back to the Topic Model (Griffiths & Steyvers)

To learn P(Z|D) and P(W|Z), we need to estimate latent var. Z

Computing P(Z|D,W)



P(D, W, Z) = P(D) P(Z|D) P(W|Z)

P(D, W) = sum_z P(D) P(z|D) P(W|z)

P(Z|D,W) = P(D, W, Z) / P(D, W)

= P(Z|D) P(W|Z) / [sum_z P(z|D) P(W|z)]

Doing the Bayesian thing

Treat the θ and ϕ as random variables with a Dirichlet distribution.

i.e., numerator P(Z|D)P(W|D) = $\text{integral}_{\theta,\phi}$ P(Z|D,θ) P(W|D,ϕ) P(θ) P(ϕ) and similarly for denominator

So you don't need to represent θ and ϕ explicitly, but instead just the *parameters* of the Dirichlet

These parameters are *counts* of occurrence.

Equation 3

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Ignore α and β for the moment

- First term: proportion of topic j draws in which w_i picked
- Second term: proportion of words in document d_i assigned to topic j
- This formula integrates out the Dirichlet uncertainty over the multinomial probabilities!

What are α and β?

- Uniform ("symmetric") prior over multinomial alternatives
- Larger value of α and β means to trust the prior more
- "...how heavily the distributions are smoothed"

Procedure

MCMC: procedure for obtaining samples from a complicated distribution, e.g., from P(Z|D,W)

1. Randomly assign each <d_i, w_i> pair a z_i value.
2. For each i resample according to Equation 3 (one *iteration*)
3. Repeat for a 1000 iteration "burn in"
4. Use current z's as "sample": assignment of each <d_i, w_i> pair to topic z_i
5. Run for another 100 iterations
6. Repeat steps 4 and 5 for a total of 10 times
7. Repeat steps 1-6 for a total of 10 times.

-> 100 samples of the z's

Use the 100 x 5628867 assignments to determine the "n"s in equation 4

$$P(w|z=j,\mathbf{z},\mathbf{w}) = \int P(w|z=j,\phi^{(j)})P(\phi^{(j)}|\mathbf{z},\mathbf{w})\ d\phi^{(j)} = \frac{n_j^{(w)}+\beta}{n_j^{(\cdot)}+W\beta}$$

Results

- **Table 1**
- **Predicting word association norms**

  “the” -> ?

  “dog” -> ?

  Figure 1: median rank of k’th associate

- **Combining syntax and semantics in a more complex generative model**

  HMM to generate tokens from different syntactic categories

  One category produces words from topic model

  Table 2

- **Details of work**

  Found optimal dimensionality for LSA; used same dim. for Topic Model