LDA

Remember

Generative process $\neq$ inference algorithm

- Examples from the course so far?

Today

LDA – we've learned generative process so far

- In words, what are the parameters we want to estimate?

Notation

Known quantities:

N documents

V unique number of words

M_i words in document i

w_{ij} indicates jth word in document i

Unknown:

$z_{ij} \in 1, ..., K$ indicates topic of word j in document i

θ_i is length K vector indicating topic proportions in document i

ϕ_k is length V vector of word probabilities, aka topic k

Another look at LDA generative process

- For each topic $k \in [1, K]$ draw $\phi_k \sim \text{Dirichlet}(\beta)$
- For each document $i \in [1, N]$:
  - Draw a distribution over topics $\theta_i \sim \text{Dirichlet}(\alpha)$
  - For each word index $j \in [1, M_i]$:
    - Draw a topic assignment $z_{ij} \sim \text{Multinomial}(1, \theta_i)$
    - Draw a word $w_{ij} \sim \text{Multinomial}(1, \phi_{z_{ij}})$

Multinomial

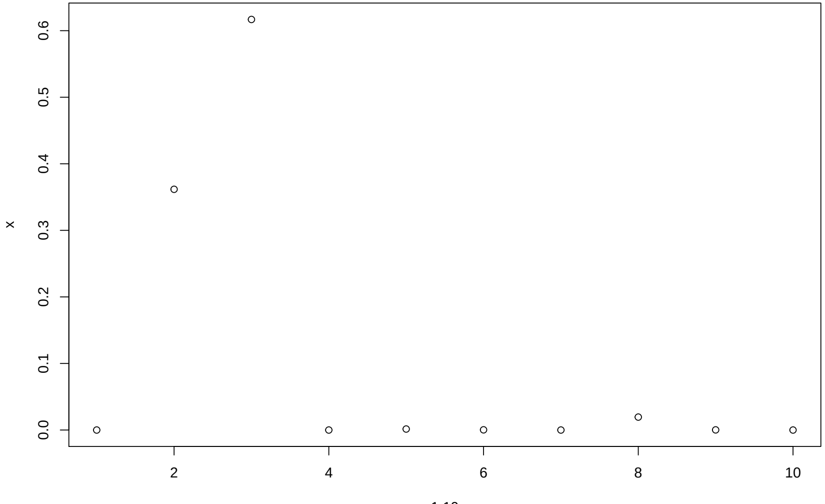
```r
rmultinom(n = 1, size = 1, prob = rep(1/3, 3))
```

```
##      [,1]
## [1,]    0
## [2,]    1
## [3,]    0
```

Dirichelt

Think of it as a distribution over probability distributions

```
library(MCMCpack)
x <- rdirichlet(n = 1, alpha = rep(.1, 10))
plot(x = 1:10, y = x)
```



Use in LDA

We use dirichlet distribution twice in LDA DGP—what for?

Use in LDA

We use dirichlet distribution twice in LDA DGP—what for?

1. To draw ϕ_k – distribution over words in vocab (aka topic k)
2. To draw θ_i – distribution over k topics in document i

Helpful note

Hyperparameters are really *vectors*, but since what's mainstream is to use symmetric Dirichlet distributions, notation is abused and shown as a scalar

- $\phi_k \sim \text{Dirichlet}(\beta)$
  - $\boldsymbol{\beta}$ actually length V vector
- $\theta_i \sim \text{Dirichlet}(\alpha)$
  - $\boldsymbol{\alpha}$ actually length K vector

Example

Assume $V = 500$, $K = 3$, $\alpha = .1$, and $\beta = .01$

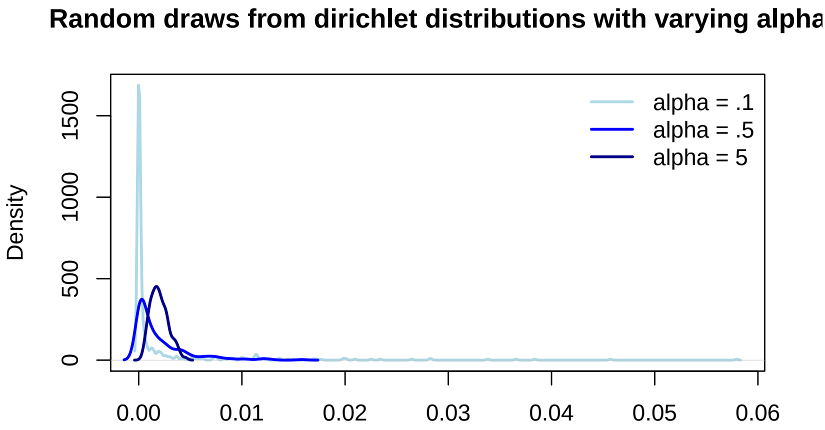
```
## how do we draw a topic? (phi_k)
#rdirichlet(n = 1, alpha = ???)

## how do we draw a distribution over topis? (theta_i)
#rdirichlet(n = 1, alpha = ???)
```

Why small hyperparameters?

Look at this for a second and describe what we're seeing. Recall $V = 500$

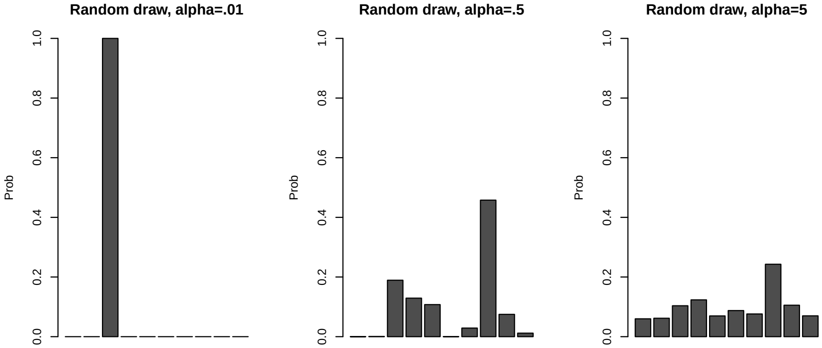
```
set.seed(109123)
phi1 <- rdirichlet(n = 1, alpha = rep(.1, 500))
phi2 <- rdirichlet(n = 1, alpha = rep(.5, 500))
phi3 <- rdirichlet(n = 1, alpha = rep(5, 500))
```



Why small hyperparameters?

Look at this for a second and describe what we're seeing. Let's say $K = 10$.

```
set.seed(109123)
theta1 <- rdirichlet(n = 1, alpha = rep(.01, 10))
theta2 <- rdirichlet(n = 1, alpha = rep(.5, 10))
theta3 <- rdirichlet(n = 1, alpha = rep(5, 10))
```



Last thing

The meaning of hyperparameters varies with the the dimensions of the data and K.

- In other words, $\alpha = .1$ means something different depending on the data. You'll sometimes see advise/defaults as 1/K.
- Think Bayesian
  - Dirichlet(α) – our prior beliefs about how the topics in our documents are distributed (... dominated by one topic? ... a mixture over most topics?)
  - Dirichlet(β) – our prior beliefs about how our topics are defined (... a few distinctive words? ... a mixture of most of the words?)
- But of course, everything we know about priors applies, like priors an be overwhelmed with enough information from the data