

Machine Learning

Justin Grimmer

Professor
Department of Political Science
Stanford University

April 11th, 2019

Last class: vector space model of text
(tf-idf)

This class: statistical model of text

Goal: Discovery (some new
conceptualization)

Building Models vs Algorithms

Today we're building statistical models

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

False debate. Both are useful

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

False debate. Both are useful

We will often use statistics, because:

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

False debate. Both are useful

We will often use statistics, because:

- 1) Clarity of assumptions

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

False debate. Both are useful

We will often use statistics, because:

- 1) Clarity of assumptions
- 2) Machinery to think about optimization, uncertainty, and sensitivity \rightsquigarrow automatic objective function

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

False debate. Both are useful

We will often use statistics, because:

- 1) Clarity of assumptions
- 2) Machinery to think about optimization, uncertainty, and sensitivity \rightsquigarrow automatic objective function
- 3) Easier to extend—leverage technology to generalize the model

Building Models vs Algorithms

Today we're building statistical models

algorithms \rightsquigarrow models without a data generating process

False debate. Both are useful

We will often use statistics, because:

- 1) Clarity of assumptions
- 2) Machinery to think about optimization, uncertainty, and sensitivity \rightsquigarrow automatic objective function
- 3) Easier to extend—leverage technology to generalize the model

Probability model \rightsquigarrow form basis of statistical approaches

Unigram Model of Language

Assume we have a 3 word **vocabulary**

Unigram Model of Language

Assume we have a 3 word **vocabulary** \rightsquigarrow 3 words that we might speak.

Unigram Model of Language

Assume we have a 3 word **vocabulary** \rightsquigarrow 3 words that we might speak.

Bag of Words \rightsquigarrow each word is an independent draw over 3 words

Unigram Model of Language

Assume we have a 3 word **vocabulary** \rightsquigarrow 3 words that we might speak.

Bag of Words \rightsquigarrow each word is an independent draw over 3 words

- **Improbable model of language creation**

Unigram Model of Language

Assume we have a 3 word **vocabulary** \rightsquigarrow 3 words that we might speak.

Bag of Words \rightsquigarrow each word is an independent draw over 3 words

- **Improbable model of language creation**
- Complex dependency structure of text

Unigram Model of Language

Assume we have a 3 word **vocabulary** \rightsquigarrow 3 words that we might speak.

Bag of Words \rightsquigarrow each word is an independent draw over 3 words

- **Improbable model of language creation**
- Complex dependency structure of text
- Improbable \neq useless

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

$$p(\mathbf{X}_i = (0, 1, 0)) = \theta_2$$

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

$$p(\mathbf{X}_i = (0, 1, 0)) = \theta_2$$

$$p(\mathbf{X}_i = (0, 0, 1)) = \theta_3 = 1 - \theta_2 - \theta_1$$

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

$$p(\mathbf{X}_i = (0, 1, 0)) = \theta_2$$

$$p(\mathbf{X}_i = (0, 0, 1)) = \theta_3 = 1 - \theta_2 - \theta_1$$

The pmf for \mathbf{X}_i is,

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

$$p(\mathbf{X}_i = (0, 1, 0)) = \theta_2$$

$$p(\mathbf{X}_i = (0, 0, 1)) = \theta_3 = 1 - \theta_2 - \theta_1$$

The pmf for \mathbf{X}_i is,

$$p(\mathbf{x}_i | \boldsymbol{\theta}) = \prod_{j=1}^3 \theta_j^{x_{ij}}$$

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

$$p(\mathbf{X}_i = (0, 1, 0)) = \theta_2$$

$$p(\mathbf{X}_i = (0, 0, 1)) = \theta_3 = 1 - \theta_2 - \theta_1$$

The pmf for \mathbf{X}_i is,

$$p(\mathbf{x}_i | \boldsymbol{\theta}) = \prod_{j=1}^3 \theta_j^{x_{ij}}$$

$$\mathbf{X}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

Unigram Model of Language

Suppose we are drawing a word $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\mathbf{X}_i = (1, 0, 0)) = \theta_1$$

$$p(\mathbf{X}_i = (0, 1, 0)) = \theta_2$$

$$p(\mathbf{X}_i = (0, 0, 1)) = \theta_3 = 1 - \theta_2 - \theta_1$$

The pmf for \mathbf{X}_i is,

$$p(\mathbf{x}_i | \boldsymbol{\theta}) = \prod_{j=1}^3 \theta_j^{x_{ij}}$$

$$\mathbf{X}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

$$\mathbf{X}_i \sim \text{Categorical}(\boldsymbol{\theta})$$

Unigram Model of Language

$$p(\mathbf{x}_i | \boldsymbol{\theta}) = \prod_{j=1}^3 \theta_j^{x_{ij}}$$
$$\mathbf{x}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

Unigram Model of Language

$$\begin{aligned} p(\mathbf{x}_i | \boldsymbol{\theta}) &= \prod_{j=1}^3 \theta_j^{x_{ij}} \\ \mathbf{x}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\ E[x_{ij}] &= \theta_j \end{aligned}$$

Unigram Model of Language

$$p(\mathbf{x}_i | \boldsymbol{\theta}) = \prod_{j=1}^3 \theta_j^{x_{ij}}$$

$$\mathbf{X}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

$$E[x_{ij}] = \theta_j$$

$$\text{Var}(X_{ij}) = \theta_j(1 - \theta_j)$$

Unigram Model of Language

$$p(\mathbf{x}_i | \boldsymbol{\theta}) = \prod_{j=1}^3 \theta_j^{x_{ij}}$$

$$\mathbf{X}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

$$E[x_{ij}] = \theta_j$$

$$\text{Var}(X_{ij}) = \theta_j(1 - \theta_j)$$

$$\text{Cov}(x_{ij}, x_{ik}) = -\theta_j \theta_k$$

Unigram Model of Language

Suppose we make N independent draws:

$$\mathbf{x}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

Then:

$$\begin{aligned}\mathbf{x} &= \sum_{i=1}^N \mathbf{x}_i \\ &= \left(\sum_{i=1}^N X_{i1}, \sum_{i=1}^N X_{i2}, \sum_{i=1}^N X_{i3} \right)\end{aligned}$$

$$\mathbf{x} \sim \text{Multinomial}(N, \boldsymbol{\theta})$$

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

Unigram Model of Language

Obtaining maximum-likelihood estimates:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \sum_{j=1}^3 x_j \log \theta_j + c$$

Include constraint that $\sum_{j=1}^3 \theta_j = 1$

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \sum_{j=1}^3 x_j \log \theta_j + \lambda \left(\sum_{j=1}^3 \theta_j - 1 \right) + c$$

Unigram Model of Language

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \sum_{j=1}^3 x_j \log \theta_j + \lambda \left(\sum_{j=1}^3 \theta_j - 1 \right) + c$$

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1} = \frac{x_1}{\theta_1} + \lambda$$

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2} = \frac{x_2}{\theta_2} + \lambda$$

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_3} = \frac{x_3}{\theta_3} + \lambda$$

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x})}{\partial \lambda} = \sum_{j=1}^3 \theta_j - 1$$

Unigram Model of Language

$$0 = \frac{x_1}{\theta_1^*} + \lambda^*$$

$$0 = \frac{x_2}{\theta_2^*} + \lambda^*$$

$$0 = \frac{x_3}{\theta_3^*} + \lambda^*$$

$$0 = \sum_{j=1}^3 \theta_j^* - 1$$

Unigram Model of Language

$$\begin{aligned}\theta_1^* &= \frac{x_1}{x_1 + x_2 + x_3} \\ \theta_2^* &= \frac{x_2}{x_1 + x_2 + x_3} \\ \theta_3^* &= \frac{x_3}{x_1 + x_2 + x_3}\end{aligned}$$

Unigram Model of Language

$$\begin{aligned}\theta_1^* &= \frac{x_1}{x_1 + x_2 + x_3} \\ \theta_2^* &= \frac{x_2}{x_1 + x_2 + x_3} \\ \theta_3^* &= \frac{x_3}{x_1 + x_2 + x_3}\end{aligned}$$

Maximum likelihood estimates \rightsquigarrow Rates words are used

Unigram Model of Language

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

Unigram Model of Language

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates \rightsquigarrow our summary of the document/speaker

Unigram Model of Language

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates \rightsquigarrow our summary of the document/speaker

$$\text{- } \sum_{j=1}^3 \theta_j = 1$$

Unigram Model of Language

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates \rightsquigarrow our summary of the document/speaker

- $\sum_{j=1}^3 \theta_j = 1$
- $\theta_j \geq 0$

Unigram Model of Language

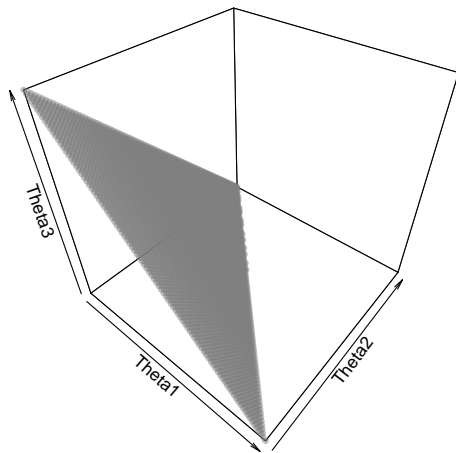
$$p(\mathbf{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^3 \theta_j^{x_j}$$

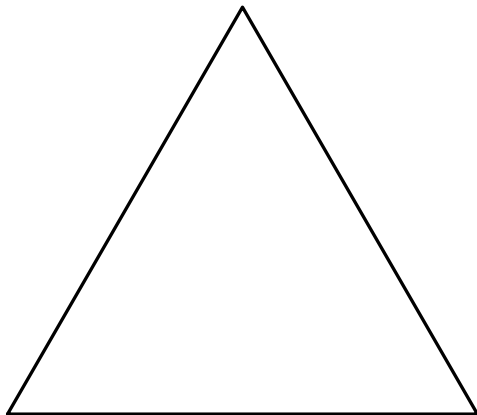
$\boldsymbol{\theta}$: encodes information about word rates \rightsquigarrow our summary of the document/speaker

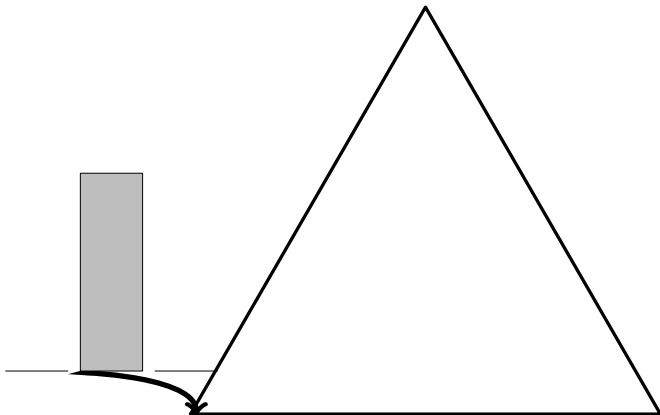
- $\sum_{j=1}^3 \theta_j = 1$

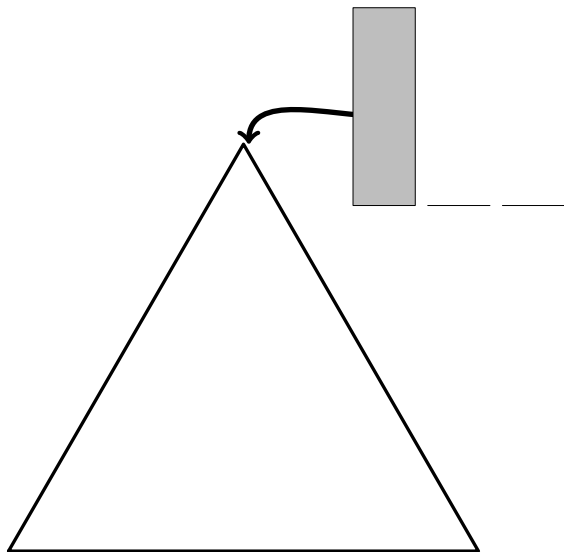
- $\theta_j \geq 0$

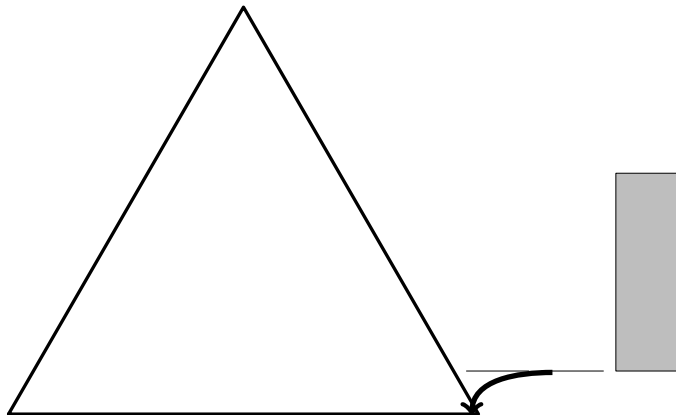
$\boldsymbol{\theta} \in \Delta^2$ (2-dimensional simplex)

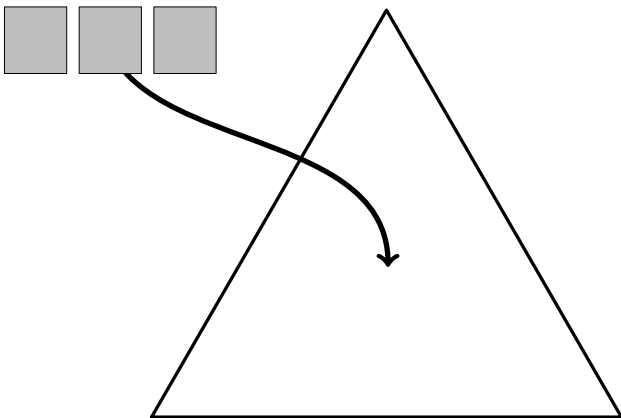












Unigram Model of Language

Suppose we have several speakers
(authors/clusters/topics/categories/ ...)
Speaker i produces document \mathbf{X}_i ,

$$\mathbf{X}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i \rightsquigarrow$ Speaker specific word rates
Build hierarchical model:

$$\boldsymbol{\theta}_i \sim \text{Distribution on Simplex}$$

Hierarchical Models as a Modeling Paradigm

Why Build a Hierarchical Model?

- 1) Borrow strength across documents \rightsquigarrow Improved and granular inferences
- 2) Shrink estimates \rightsquigarrow regularization
- 3) Incorporate further covariate information
 - i) Author
 - ii) Time
 - iii) ...
- 3) Learn additional structure
 - i) Hierarchies of word rates
 - ii) Clusters of similar word rates
 - iii) Low dimensional approximations of word rates
- 4) Encodes complicated dependencies between documents/speakers

Dirichlet-Multinomial Unigram Language Model

For N observations we observe a 3-element long count vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$$

Where $N_i = \sum_{j=1}^3 x_{ij}$.

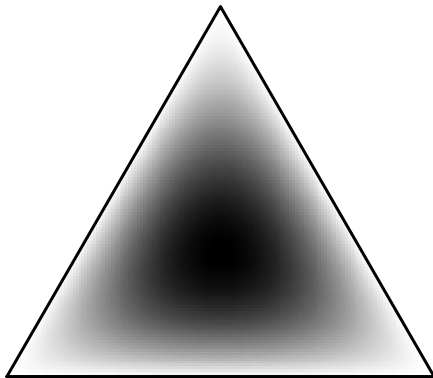
Suppose

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

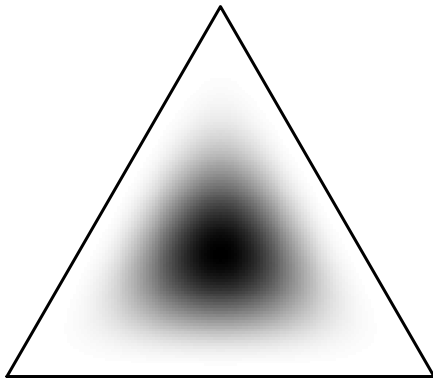
$$\mathbf{x}_i | \boldsymbol{\theta}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

- Dirichlet distribution \rightsquigarrow assumption about **population** of word rates
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ describes population use of words and variation
- **Just one distribution simplex**

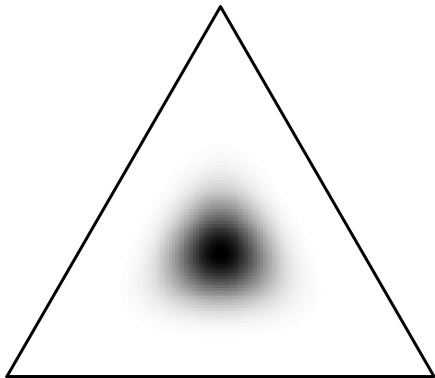
$\alpha = 2,2,2$



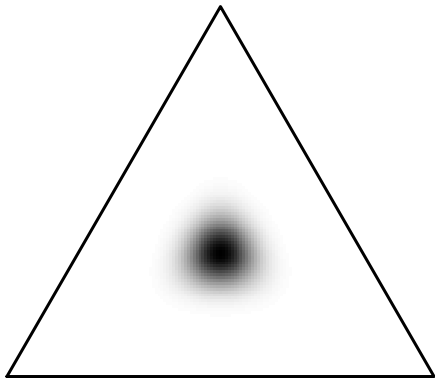
$\alpha = 4,4,4$



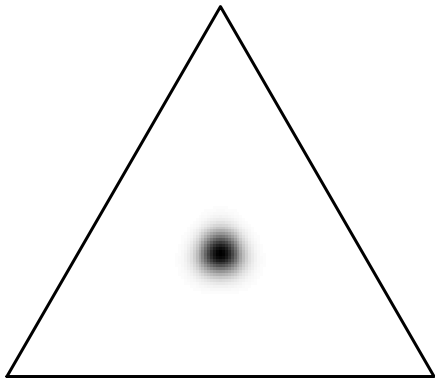
$\alpha = 10, 10, 10$



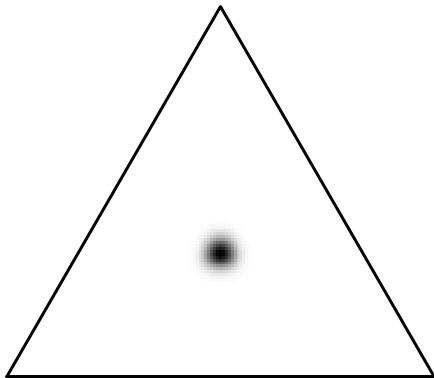
$\alpha = 20, 20, 20$



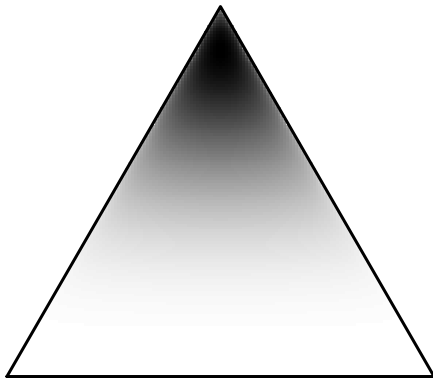
$\alpha = 50, 50, 50$



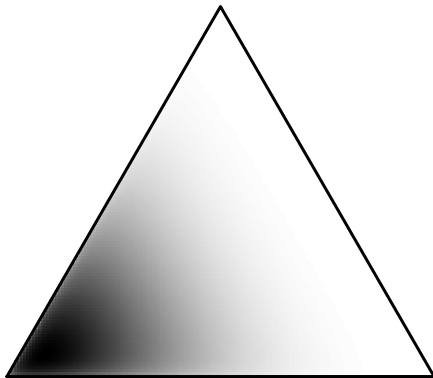
$\alpha = 100,100,100$



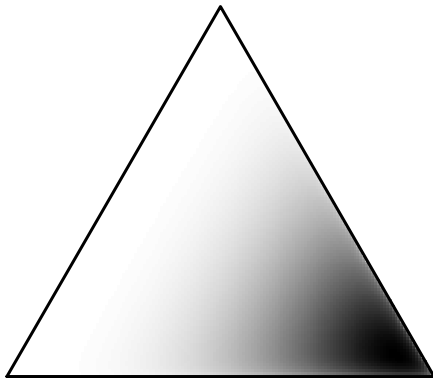
$\alpha = 4, 1.2, 1.2$



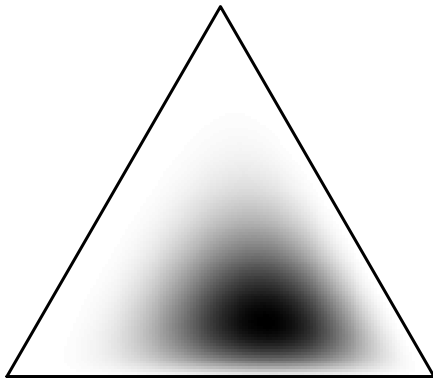
$\alpha = 1.2, 4, 1.2$



$\alpha = 1.2, 1.2, 4$



$\alpha = 2.04, 3.24, 4.72$



Dirichlet Distribution

Suppose

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

Then,

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \prod_{j=1}^3 \theta_{ij}^{\alpha_j-1}$$

- If $\alpha = (1, 1, 1)$ **Uniform distribution**

Dirichlet Distribution

- Important Facts

$$\begin{aligned}E[\theta_i] &= \left(\frac{\alpha_1}{\sum_{j=1}^3 \alpha_j}, \frac{\alpha_2}{\sum_{j=1}^3 \alpha_j}, \frac{\alpha_3}{\sum_{j=1}^3 \alpha_j} \right) \\ \text{var}(\theta_{ij}) &= \frac{\alpha_i \left(\sum_{j=1}^3 \alpha_j - \alpha_i \right)}{\left(\sum_{j=1}^3 \alpha_j \right)^2 \left(\sum_{j=1}^3 \alpha_j + 1 \right)} \\ \text{cov}(\theta_{ik}, \theta_{ij}) &= \frac{-\alpha_k \alpha_j}{\left(\sum_{j=1}^3 \alpha_j \right)^2 \left(\sum_{j=1}^3 \alpha_j + 1 \right)} \\ \text{Mode}(\theta_j) &= \frac{\alpha_j - 1}{\sum_{k=1}^3 \alpha_k - 3}\end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

$$\mathbf{x}_i | \theta_i \sim \text{Multinomial}(N_i, \theta_i)$$

Dirichlet-Multinomial Unigram Model of Language

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

$$\mathbf{x}_i | \theta_i \sim \text{Multinomial}(N_i, \theta_i)$$

$$p(\theta_i | \alpha, \mathbf{x}_i) \propto p(\theta_i | \alpha) p(\mathbf{x}_i | \theta_i)$$

Dirichlet-Multinomial Unigram Model of Language

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\mathbf{x}_i | \boldsymbol{\theta}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

$$\begin{aligned} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}, \mathbf{x}_i) &\propto p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) p(\mathbf{x}_i | \boldsymbol{\theta}_i) \\ &\propto \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \prod_{j=1}^3 \theta_j^{\alpha_j - 1} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

$$\mathbf{x}_i | \theta_i \sim \text{Multinomial}(N_i, \theta_i)$$

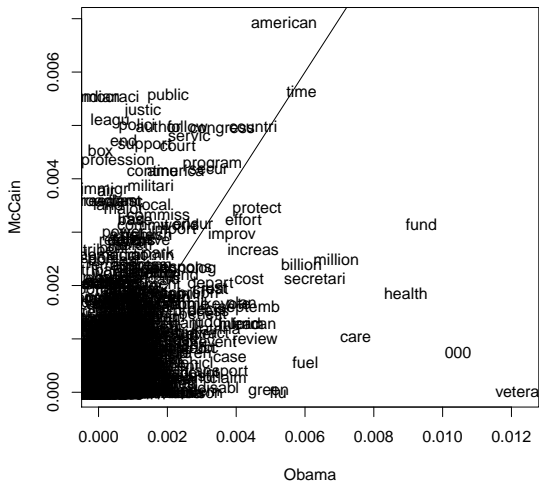
$$\begin{aligned} p(\theta_i | \alpha, \mathbf{x}_i) &\propto p(\theta_i | \alpha) p(\mathbf{x}_i | \theta_i) \\ &\propto \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \prod_{j=1}^3 \theta_j^{\alpha_j - 1} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \\ &\propto \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \underbrace{\prod_{j=1}^3 \theta_j^{\alpha_j + x_{ij} - 1}}_{\text{Dirichlet Kernel}} \end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

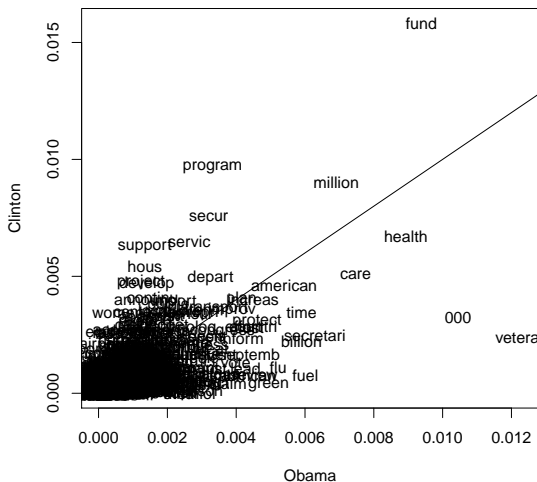
$$\begin{aligned}\boldsymbol{\theta}_i | \boldsymbol{\alpha}, \mathbf{x}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{x}) \\ \mathbb{E}[\theta_{ij} | \boldsymbol{\alpha}, \mathbf{x}_i] &= \frac{\alpha_j + x_{ij}}{\sum_{j=1}^3 (x_{ij} + \alpha_j)}\end{aligned}$$

- $\alpha_j \rightsquigarrow$ “pseudo” data that smooth the estimates toward $\frac{\alpha_j}{\alpha_1 + \alpha_2 + \alpha_3}$
- as $N_i \rightarrow \infty$ data (\mathbf{x}_i) **overwhelm** $\boldsymbol{\alpha}$

Senate Press Releases



Senate Press Releases



Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\mathbf{x}_i \rightsquigarrow$
marginalize over $\boldsymbol{\theta}$

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\mathbf{x}_i | \boldsymbol{\theta}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\mathbf{x}_i \rightsquigarrow$
marginalize over $\boldsymbol{\theta}$

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\mathbf{x}_i | \boldsymbol{\theta}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

This implies:

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\mathbf{x}_i \rightsquigarrow$ marginalize over θ

$$\begin{aligned}\theta_i &\sim \text{Dirichlet}(\alpha) \\ \mathbf{x}_i | \theta_i &\sim \text{Multinomial}(N_i, \theta_i)\end{aligned}$$

This implies:

$$p(\mathbf{x}_i | \alpha) = \int_{\Delta^2} p(\mathbf{x}_i, \theta_i | \alpha) d\theta$$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\mathbf{x}_i \rightsquigarrow$ marginalize over θ

$$\begin{aligned}\theta_i &\sim \text{Dirichlet}(\alpha) \\ \mathbf{x}_i | \theta_i &\sim \text{Multinomial}(N_i, \theta_i)\end{aligned}$$

This implies:

$$\begin{aligned}p(\mathbf{x}_i | \alpha) &= \int_{\Delta^2} p(\mathbf{x}_i, \theta_i | \alpha) d\theta \\ &= \int_{\Delta^2} p(\mathbf{x}_i | \theta_i) p(\theta_i | \alpha) d\theta\end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\mathbf{x}_i \rightsquigarrow$
marginalize over $\boldsymbol{\theta}$

$$\begin{aligned}\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \mathbf{x}_i | \boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)\end{aligned}$$

This implies:

$$\begin{aligned}p(\mathbf{x}_i | \boldsymbol{\alpha}) &= \int_{\Delta^2} p(\mathbf{x}_i, \boldsymbol{\theta}_i | \boldsymbol{\alpha}) d\boldsymbol{\theta} \\ &= \int_{\Delta^2} p(\mathbf{x}_i | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) d\boldsymbol{\theta} \\ &= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \int_{\Delta^2} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j - 1} d\boldsymbol{\theta}\end{aligned}$$

$$p(\mathbf{x}_i|\boldsymbol{\alpha}) = \binom{N_i}{n_1!n_2!n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \int_{\Delta^2} \underbrace{\prod_{j=1}^3 \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j-1}}_{\text{Dirichlet Kernel}} d\boldsymbol{\theta}$$

$$\begin{aligned}
p(\mathbf{x}_i|\boldsymbol{\alpha}) &= \binom{N_i}{n_1!n_2!n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \underbrace{\int_{\Delta^2} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j-1} d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1!n_2!n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \frac{\prod_{j=1}^3 \Gamma(x_{ij} + \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}_i | \boldsymbol{\alpha}) &= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \underbrace{\int_{\Delta^2} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j - 1} d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \frac{\prod_{j=1}^3 \Gamma(x_{ij} + \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \\
&= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \prod_{j=1}^3 \frac{\Gamma(x_{ij} + \alpha_j)}{\Gamma(\alpha_j)}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}_i | \boldsymbol{\alpha}) &= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \underbrace{\int_{\Delta^2} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j - 1} d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \frac{\prod_{j=1}^3 \Gamma(x_{ij} + \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \\
&= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \prod_{j=1}^3 \frac{\Gamma(x_{ij} + \alpha_j)}{\Gamma(\alpha_j)}
\end{aligned}$$

Has some intuitive properties

$$\begin{aligned}
p(\mathbf{x}_i|\boldsymbol{\alpha}) &= \binom{N_i}{n_1!n_2!n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \underbrace{\int_{\Delta^2} \prod_{j=1}^3 \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j-1} d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1!n_2!n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)} \frac{\prod_{j=1}^3 \Gamma(x_{ij} + \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \\
&= \binom{N_i}{n_1!n_2!n_3!} \frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \prod_{j=1}^3 \frac{\Gamma(x_{ij} + \alpha_j)}{\Gamma(\alpha_j)}
\end{aligned}$$

Has some intuitive properties

$$E[X_{ij}] = N \frac{\alpha_j}{\sum_{k=1}^3 \alpha_k}$$

Dirichlet-Multinomial Unigram Model

We can also generate a predictive distribution \rightsquigarrow probability next word is j

$$\begin{aligned}P(\tilde{x} = 1 | \mathbf{x}_i, \boldsymbol{\alpha}) &= \int_{\Delta^2} p(\tilde{x} = 1 | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \boldsymbol{\alpha}, \mathbf{x}_i) d\boldsymbol{\theta} \\&= \int_{\Delta^2} \theta_j \text{Dir}(\boldsymbol{\theta} | \mathbf{x}_i + \boldsymbol{\alpha}) d\boldsymbol{\theta} \\&= \frac{x_{ij} + \alpha_j}{\sum_{j=1}^3 (x_{ij} + \alpha_j)}\end{aligned}$$

Dirichlet-Multinomial Unigram Model

Where does α come from?

Extend the model \rightsquigarrow **infer** it

$$\alpha_j \sim \text{Gamma}(0.25, 1)$$

$$\theta_i | \alpha \sim \text{Dirichlet}(\theta_i)$$

$$\mathbf{x}_i | \theta_i \sim \text{Multinomial}(N_i, \theta_i)$$

Dirichlet-Multinomial Unigram Model

Which yields

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\alpha} | \mathbf{X}) &\propto p(\boldsymbol{\alpha}) \prod_{i=1}^N p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) p(\mathbf{x}_i | \boldsymbol{\alpha}) \\ p(\boldsymbol{\alpha} | \mathbf{X}) &= \int_{\Delta^2} p(\boldsymbol{\theta}, \boldsymbol{\alpha} | \mathbf{X}) d\boldsymbol{\theta} \\ &\propto p(\boldsymbol{\alpha}) \prod_{i=1}^N \int_{\Delta^2} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) p(\mathbf{x}_i | \boldsymbol{\theta}_i, \boldsymbol{\alpha}) d\boldsymbol{\theta} \\ &\propto \prod_{j=1}^3 4 \exp(-4\alpha_j) \times \prod_{i=1}^N \left[\frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\Gamma(\sum_{j=1}^3 (x_{ij} + \alpha_j))} \times \prod_{j=1}^3 \frac{\Gamma(x_{ij} + \alpha_j)}{\Gamma(\alpha_j)} \right] \end{aligned}$$

Unigram Model of Language

Suppose $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iJ})$

- x_{ij} = Number of times word j occurs in document i .
- $N_i = \sum_{j=1}^J x_{ij}$ total number of words in document i

Assume a generation process

$$\mathbf{x}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_J)$$

$$p(\mathbf{x}_i | \boldsymbol{\theta}) \propto \prod_{j=1}^J \theta_j^{x_{ij}}$$

Alternative Priors on the Simplex

Dirichlet distribution

- Imposes specific form on variance
- Imposes negative correlation between all components.
- We might expect some word rates to positively covary.

Alternative \rightsquigarrow **Logistic-Normal** distribution

Logistic-Normal Distribution

Suppose $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$.

Define:

$$\mathbf{y}_i = \left(\log \left(\frac{x_{i1}}{x_{i3}} \right), \log \left(\frac{x_{i2}}{x_{i3}} \right) \right)$$

$$\mathbf{y}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \text{cov}(y_{i1}, y_{i2}) \\ \text{cov}(y_{i1}, y_{i2}) & \sigma_2^2 \end{pmatrix}$$

Logistic-Normal Distribution

To get back original data apply:

$$x_{i1} = \left(\frac{\exp(y_{i1})}{\exp(y_{i1}) + \exp(y_{i2}) + 1} \right)$$

$$x_{i2} = \left(\frac{\exp(y_{i2})}{\exp(y_{i1}) + \exp(y_{i2}) + 1} \right)$$

$$x_{i3} = \left(\frac{1}{\exp(y_{i1}) + \exp(y_{i2}) + 1} \right)$$

$$\mathbf{x}_i = g(\mathbf{y}_i)$$

Logistic-Normal Distribution

An alternative hierarchical model:

$$\boldsymbol{\eta}_i \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\theta}_i = g(\boldsymbol{\eta}_i)$$

$$\mathbf{x}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

Widely used:

- Correlated models
- Natural way to encode **regressions** in prior

Next week: clustering!