Topic Models

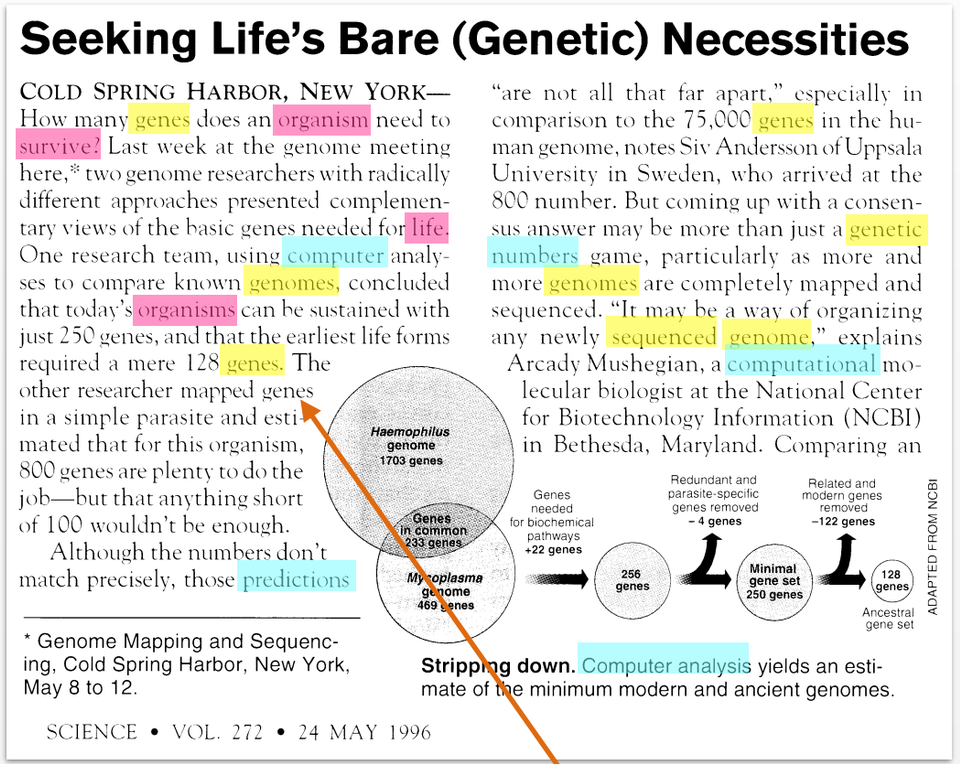
Shantanu Jain

Topic Modeling Basics

Borrowing from:
David Blei
(Columbia)

Word Mixtures

Idea: Model text as a "bag" of words (ignore order)



Seeking Life's Bare (Genetic) Necessities

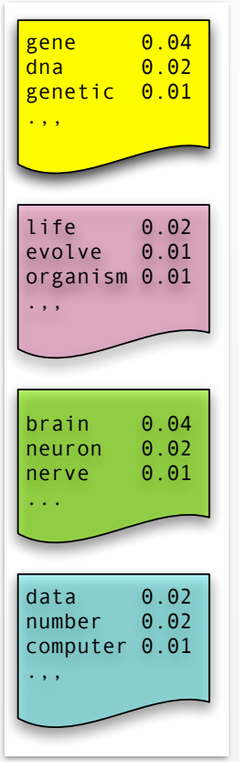
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



$p(x \mid z = 1)$

$p(x \mid z = 2)$

$p(x \mid z = 3)$

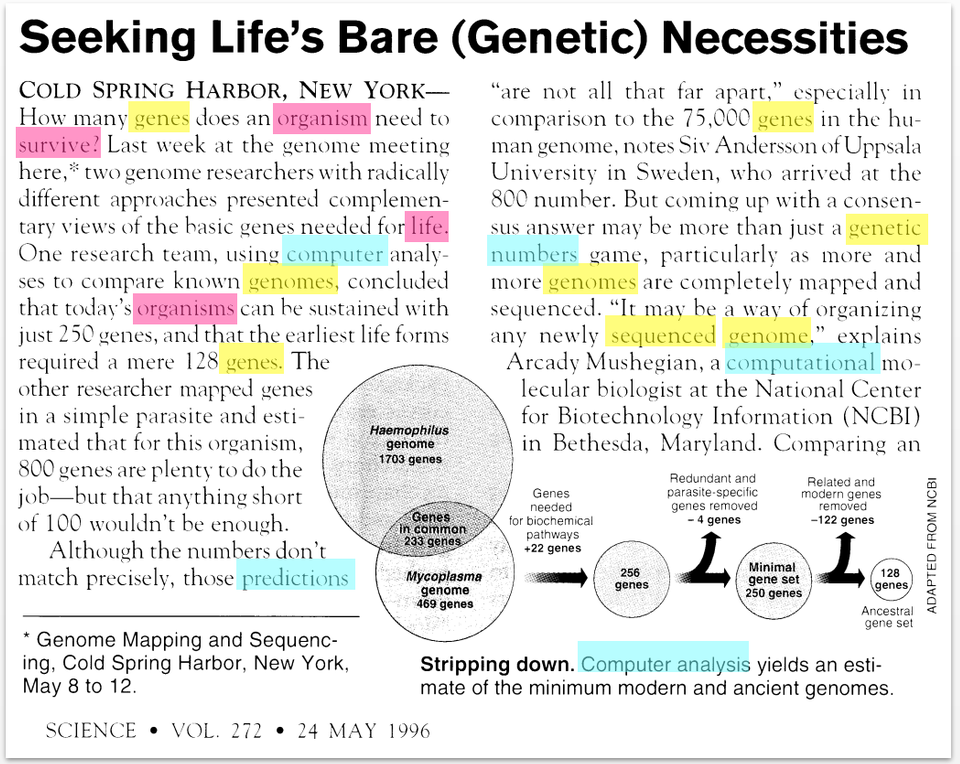
$p(x \mid z = 4)$

Word in vocabulary: $x_n \in \{1, \ldots, V\}$

Topic assignment: $z_n \in \{1, \ldots, K\}$

- Total N words in a document
- n denotes the index of the n^{th} word in the document.
- V is the number of words in the vocabulary.
- K is the number of topics.

Word Mixtures



Seeking Life's Bare (Genetic) Necessities

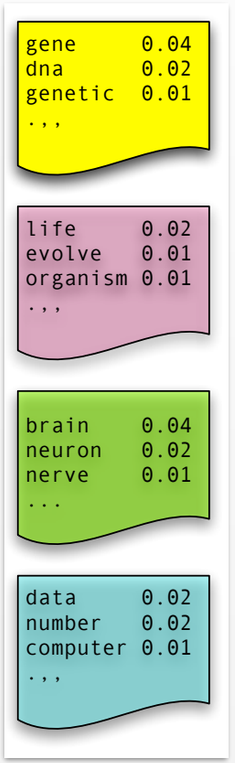
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



$p(x \mid z=1, \beta)$

$p(x \mid z=2, \beta)$

$p(x \mid z=3, \beta)$

$p(x \mid z=4, \beta)$

θ: topic proportions/probabilities, probability over the K topics

$$\theta = [\theta_1, \theta_2, \ldots \theta_K] \quad \sum_k \theta_k = 1$$

$$p(z_n = k \mid \theta) = \theta_k$$

β_k: k^{th} topic's word probabilities over the vocabulary

$$\beta_k = [\beta_{k1}, \beta_{k2}, \ldots \beta_{kV}] \quad \sum_i \beta_{ki} = 1$$

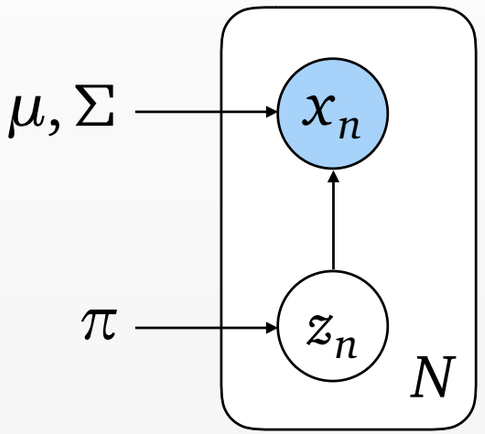
$$p(x_n = i \mid z_n = k, \beta) = \beta_{ki}$$

$$z_n \sim \text{Discrete}(\theta)$$ *Pick a topic*

$$x_n \mid z_n = k \sim \text{Discrete}(\beta_k)$$ *Pick a word given topic*

Gaussian Mixtures vs Word Mixtures

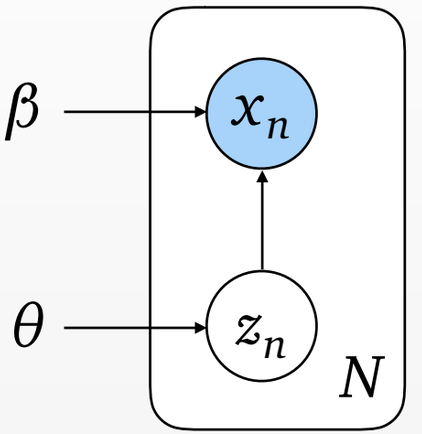
Gaussian Mixture Model



$$z_n \sim \text{Discrete}(\pi_1, \ldots, \pi_K)$$

$$x_n \mid z_n = k \sim \text{Normal}(\mu_k, \Sigma_k)$$

Discrete Mixture Model



$$z_n \sim \text{Discrete}(\theta_1, \ldots, \theta_K)$$

$$x_n \mid z_n = k \sim \text{Discrete}(\beta_{k,1}, \ldots, \beta_{k,V})$$

Difference: Replace Gaussian with Discrete

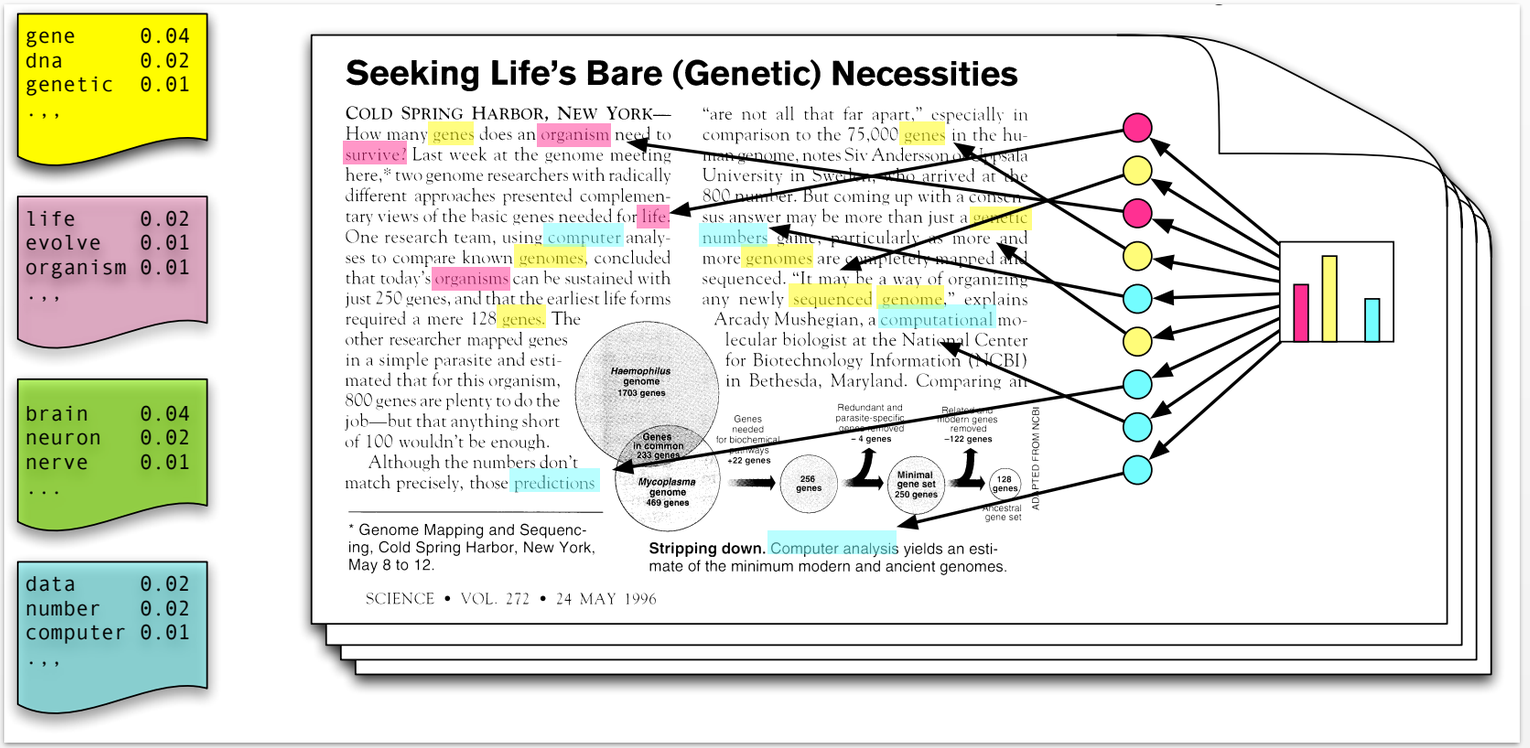
Using the term Discrete distribution to mean Categorical distribution

Topic Modeling

Topics (shared)

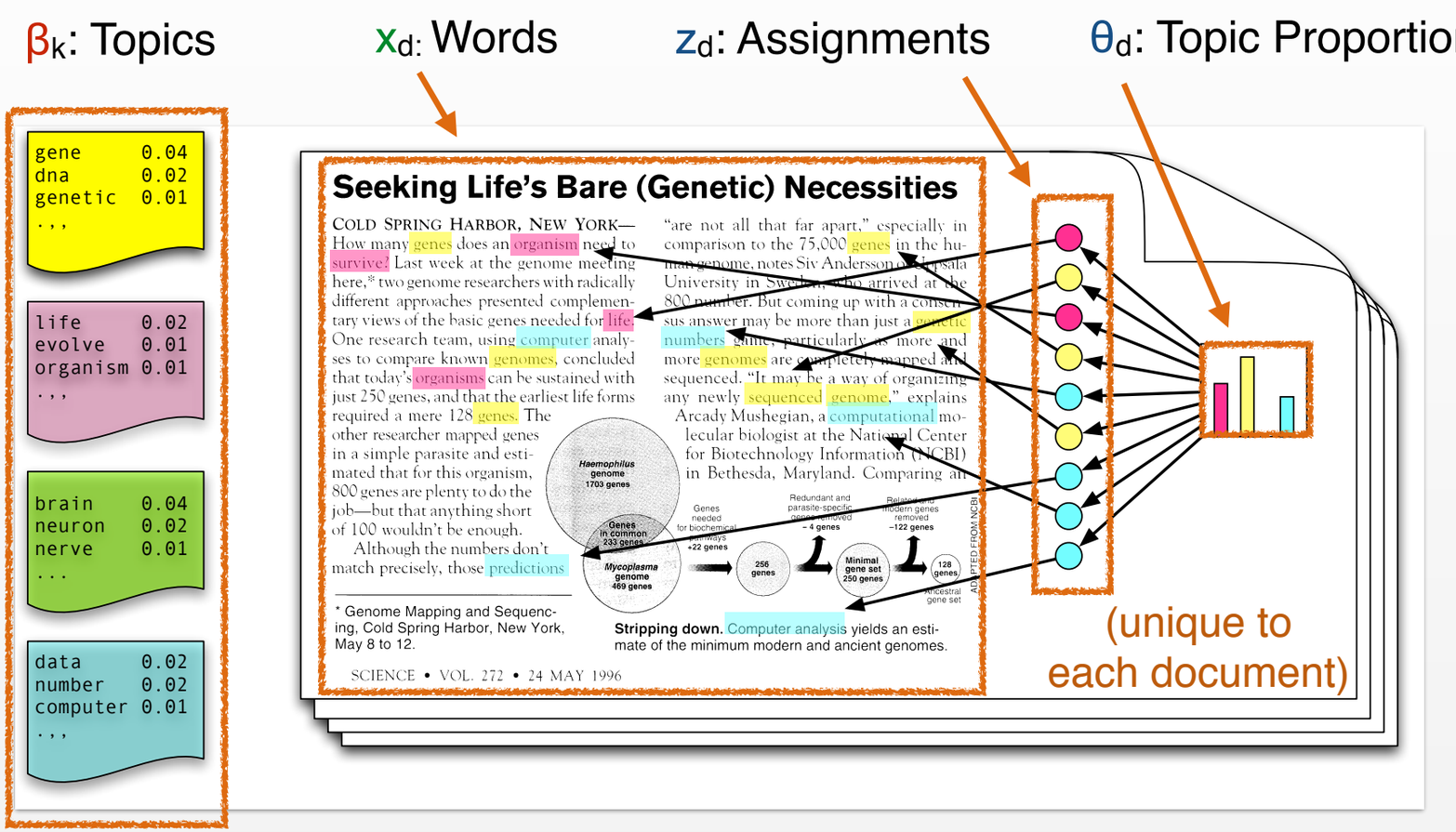
Words in Document (mixture over topics)

Topic Proportions (document-specific)



Idea: Model ***corpus*** of documents with ***shared*** topics

Topic Modeling



$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

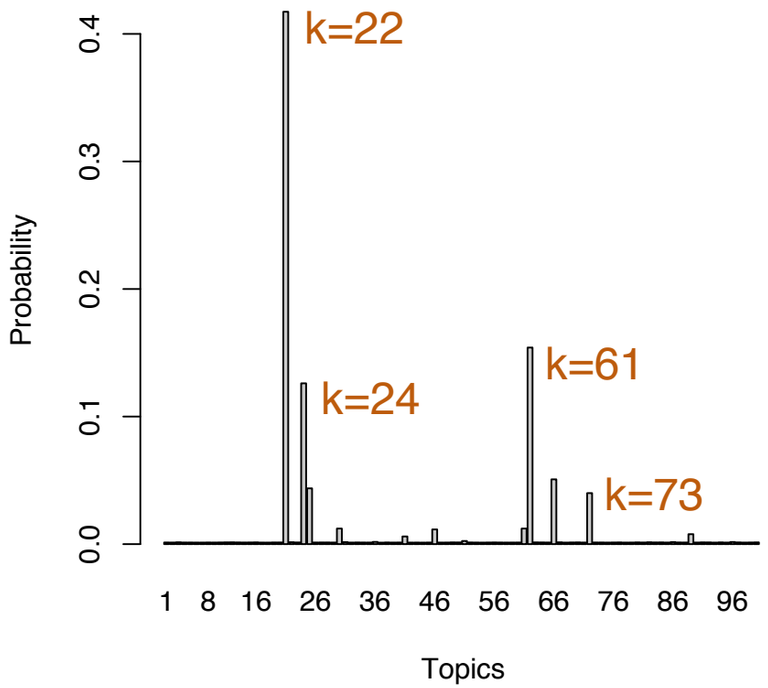
$$x_{d,n} \mid z_{d,n} = k \sim \text{Discrete}(\beta_k)$$

Distribution over Topic Assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The

$z_{d,n} \sim \text{Discrete}(\theta_d)$



Next Slide: Frequent words in these topics

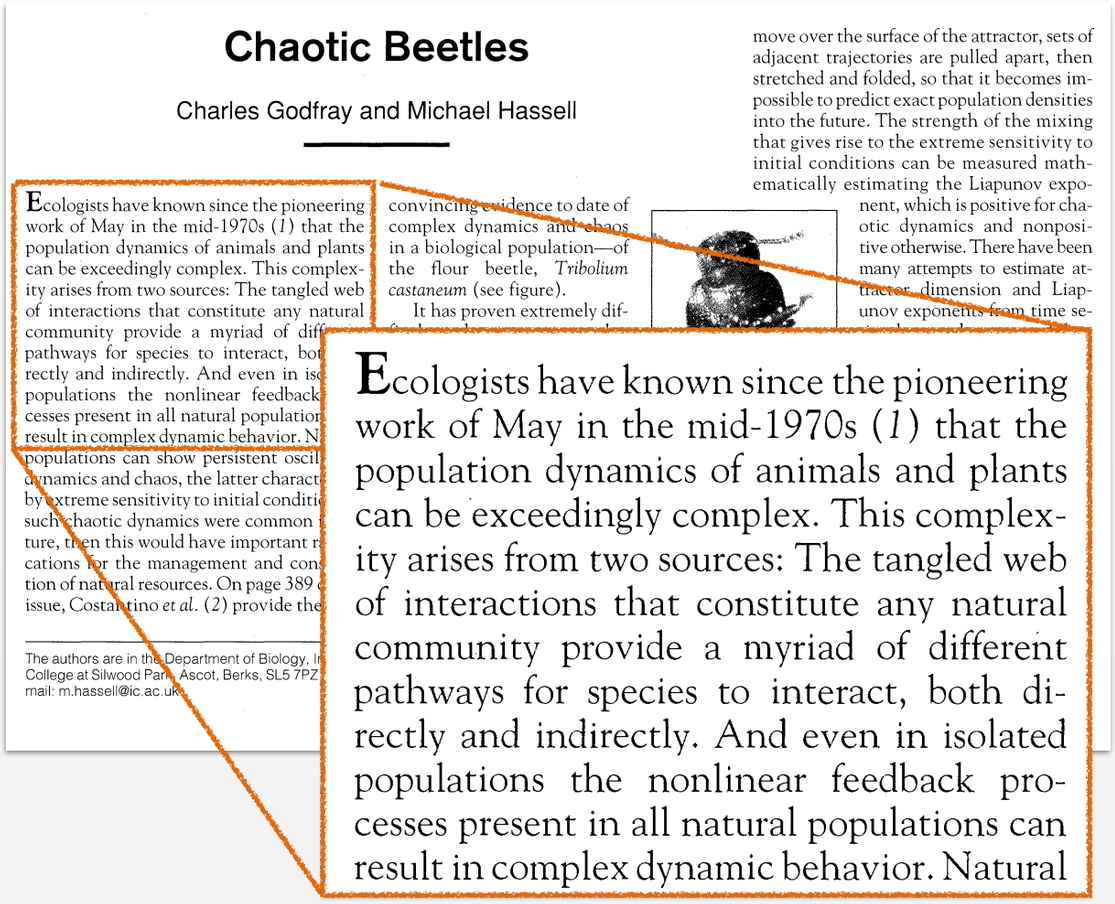
Most Probable Words in Topics

$$x_{d,n} \mid z_{d,n} = k \sim \text{Discrete}(\beta_k)$$

Most frequent (*within topic*)

k=22	k=24	k=61	k=73
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Each Document has Different Topics



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of diff pathways for species to interact, bot rectly and indirectly. And even in isc populations the nonlinear feedback cesses present in all natural populatio result in complex dynamic behavior. N populations can show persistent oscil dynamics and chaos, the latter charact by extreme sensitivity to initial conditi such chaotic dynamics were common ture, then this would have important r cations for the management and con tion of natural resources. On page 389 issue, Costantino *et al.* (*2*) provide the

The authors are in the Department of Biology, I College at Silwood Park, Ascot, Berks, SL5 7PZ mail: m.hassell@ic.ac.uk

convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

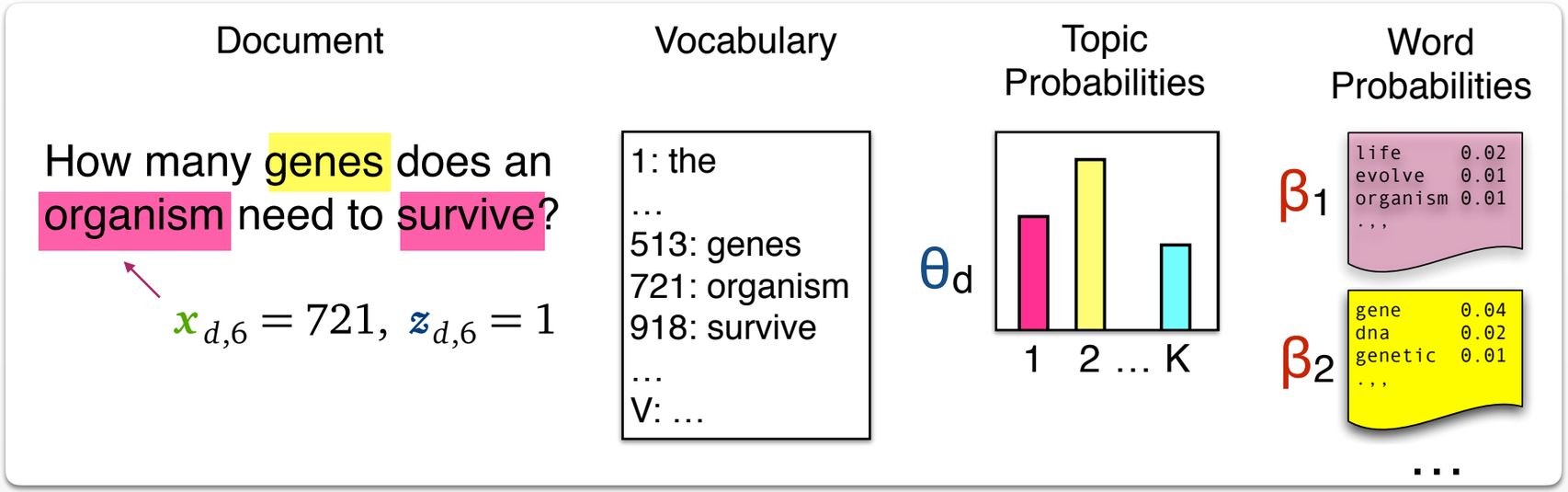
It has proven extremely dif-

move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time se-

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural

problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

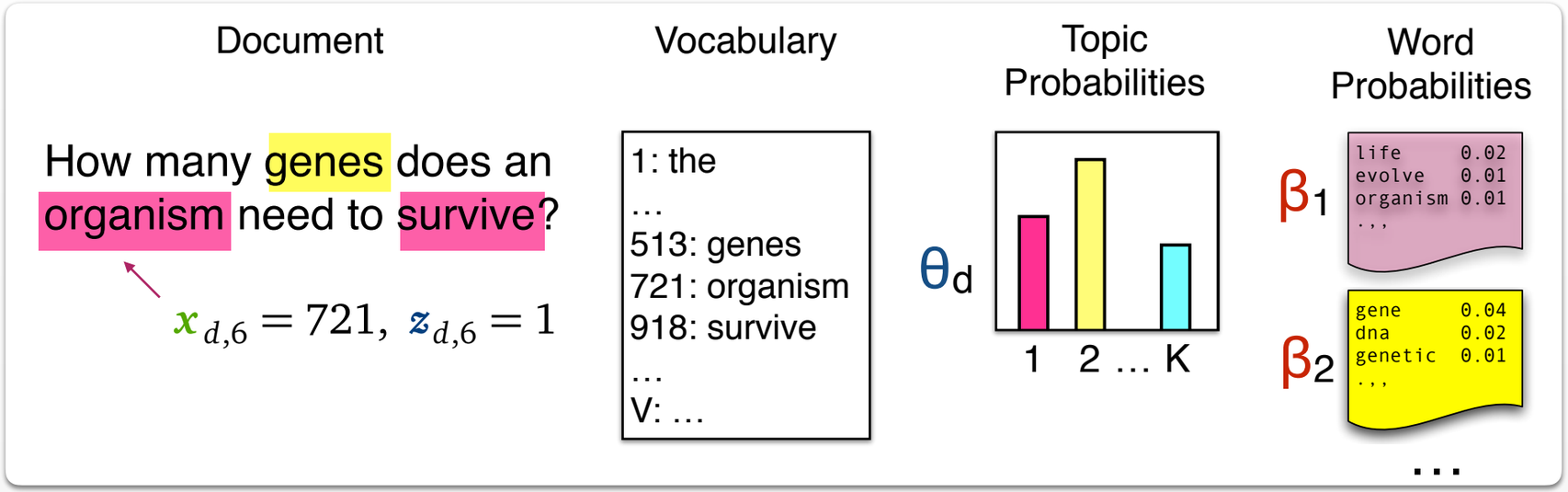
Estimating the Parameters



Maximum Likelihood: $\max_{\theta,\beta} \log p(x \mid \theta, \beta)$

d=1	$[x_{1,1}, x_{1,2}, \ldots, x_{1,N_1}]$	$[\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{1,K}]$	$[\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,V}]$	k=1
d=2	$[x_{2,1}, x_{2,2}, \ldots, x_{2,N_2}]$	$[\theta_{2,1}, \theta_{2,2}, \ldots, \theta_{2,K}]$	$[\beta_{2,1}, \beta_{2,2}, \ldots, \beta_{2,V}]$	k=2
	…	…	…	
d=D	$[x_{D,1}, x_{D,2}, \ldots, x_{D,N_D}]$	$[\theta_{D,1}, \theta_{D,2}, \ldots, \theta_{D,K}]$	$[\beta_{K,1}, \beta_{K,2}, \ldots, \beta_{K,V}]$	k=K
	(not a matrix)	D x K	K x V	

Calculating the Likelihood for each Word



Probability that word *n* is entry *v* in the vocabulary

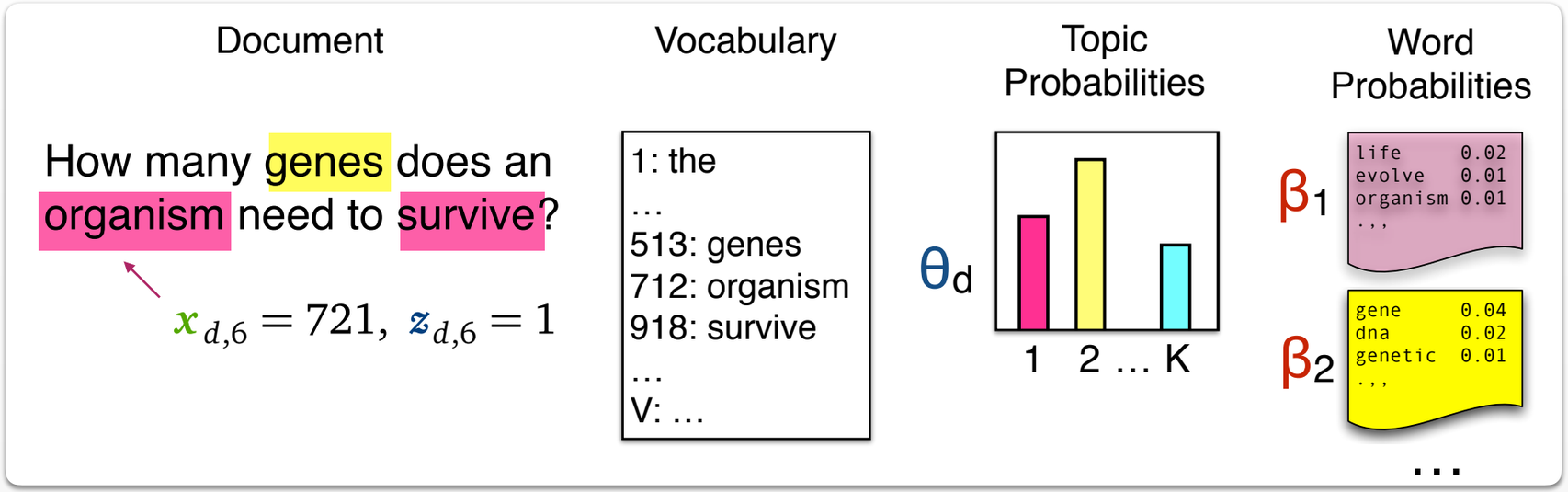
Probability of word *v* given topic *k*

Probability that word belongs to topic *k*

$$p(x_{d,n}=v \mid \beta, \theta_d) = \sum_{k=1}^{K} p(x_{d,n}=v \mid \beta, z_{d,n}=k)\, p(z_{d,n}=k \mid \theta_d)$$

$$= \sum_{k=1}^{K} \beta_{k,v}\, \theta_{d,k}$$

Computing the Likelihood



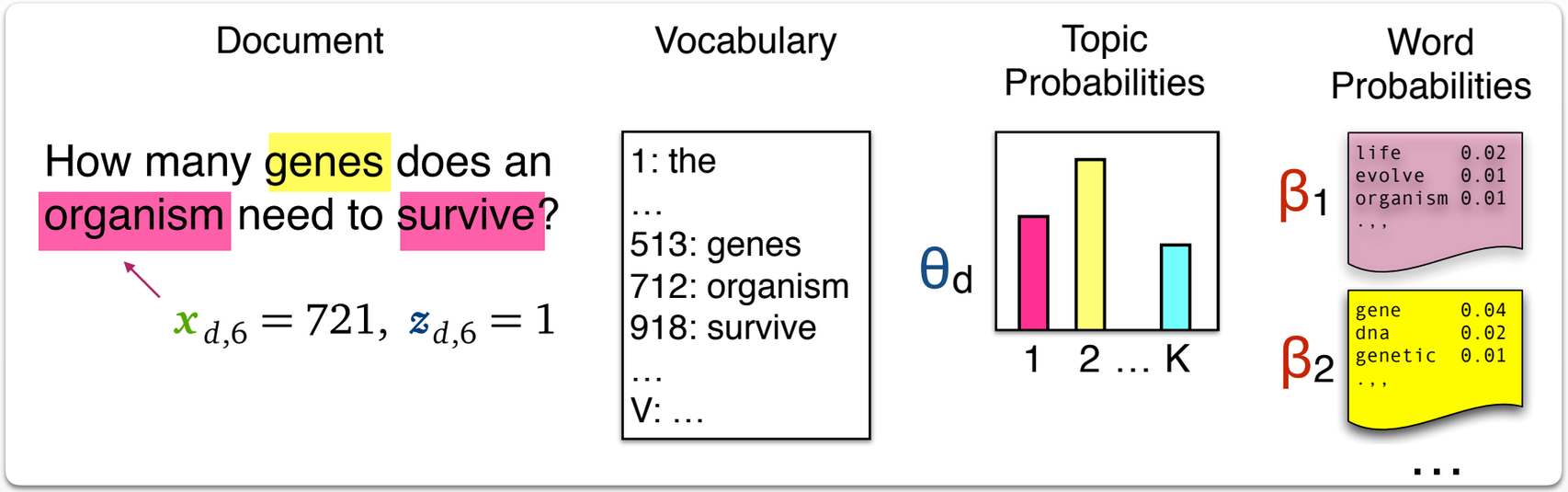
probability of all words $n = 1 \ldots N_d$ in document d (use one-hot trick)

$$p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \prod_{n=1}^{N_d} \prod_{v=1}^{V} p(\boldsymbol{x}_{d,n} = v \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)^{I[\boldsymbol{x}_{d,n}=v]}$$

take log probability, substitute result from previous slide

$$\log p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \sum_{n=1}^{N_d} \sum_{v=1}^{V} I[\boldsymbol{x}_{d,n} = v] \log \left(\sum_{k=1}^{K} \boldsymbol{\beta}_{k,v} \boldsymbol{\theta}_{d,k} \right)$$

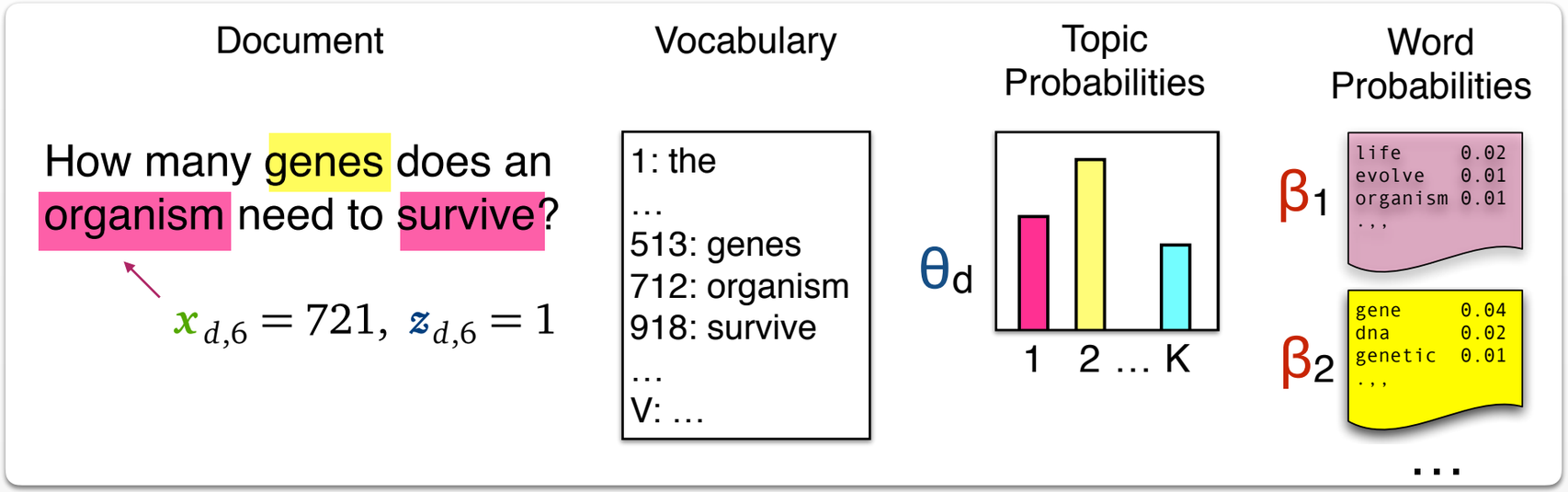
Calculating the Likelihood for all Words



log probability of all words in document *d*

$$\log p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \sum_{n=1}^{N_d} \sum_{v=1}^{V} I[\boldsymbol{x}_{d,n} = v] \log \left(\sum_{k=1}^{K} \boldsymbol{\beta}_{k,v} \boldsymbol{\theta}_{d,k} \right)$$

Calculating the Likelihood for all Words



log probability of all words in document *d*

$$\log p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \sum_{v=1}^{V} \sum_{n=1}^{N_d} I[\boldsymbol{x}_{d,n} = v] \log \left(\sum_{k=1}^{K} \boldsymbol{\beta}_{k,v} \boldsymbol{\theta}_{d,k} \right)$$

$$= \boldsymbol{X}_d \log (\boldsymbol{\theta}_d \boldsymbol{\beta})^\top$$

bag-of-word vector

$$\boldsymbol{X}_{d,v} = \sum_{n=1}^{N_d} I[\boldsymbol{x}_{d,n} = v]$$

inner product between bag of word vector,
and log weighted average over topics

Interpretation as Matrix Factorization

Log likelihood

$$\log(p(\boldsymbol{X}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)) = \boldsymbol{X}_d \log(\boldsymbol{\theta}_d \boldsymbol{\beta})^\top$$

Bag of Word Vector

$$\boldsymbol{X}_{d,v} = \sum_{n=1}^{N_d} I[\boldsymbol{x}_{d,n} = v]$$

Word Counts | Topic Counts | Topic Word Probabilities

$$\mathbb{E}\left[\begin{bmatrix} \boldsymbol{X}_{1,1} & \cdots & \boldsymbol{X}_{1,V} \\ \cdots & & \\ \boldsymbol{X}_{D,1} & \cdots & \boldsymbol{X}_{D,V} \end{bmatrix}\right] = \begin{bmatrix} N_1\boldsymbol{\theta}_{1,1} & \cdots & N_1\boldsymbol{\theta}_{1,K} \\ \cdots & & \\ N_D\boldsymbol{\theta}_{D,1} & \cdots & N_D\boldsymbol{\theta}_{D,K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{1,1} & \cdots & \boldsymbol{\beta}_{1,V} \\ \cdots & & \\ \boldsymbol{\beta}_{K,1} & \cdots & \boldsymbol{\beta}_{K,V} \end{bmatrix}$$

(D x V) (D x K) (K x V)

stocks	chairman	the		wins	game
2	4	8	…	0	1
…					
0	1	7	…	2	3

$$\simeq$$

finance	Sports		stocks	game
$112 \cdot 0.91$	…	$112 \cdot 0.01$	0.0081	0.0002
…			…	
$234 \cdot 0.02$	…	$234 \cdot 0.86$	0.0001	0.0072

Relationship to Latent Semantic Analysis

LSA: Factorize word counts (using PCA)

$\mathbf{X}$ (V x D) $\approx$ $\mathbf{U}$ (V x K) $\mathbf{Z}$ (K x D)

$$\begin{pmatrix} \text{stocks: } 2 & \cdots & 0 \\ \text{chairman: } 4 & \cdots & 1 \\ \text{the: } 8 & \cdots & 7 \\ \cdots & \vdots & \vdots \\ \text{wins: } 0 & \cdots & 2 \\ \text{game: } 1 & \cdots & 3 \end{pmatrix} \approx \begin{pmatrix} 0.4 & \cdot\cdot & -0.001 \\ 0.8 & \cdot\cdot & 0.03 \\ 0.01 & \cdot\cdot & 0.04 \\ \vdots & \cdot\cdot & \vdots \\ 0.002 & \cdot\cdot & 2.3 \\ 0.003 & \cdot\cdot & 1.9 \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

Topic Models: Factorize word counts (using mixture model)

$\mathbb{E}[\boldsymbol{X}^\top]$ (V x D) $=$ $\boldsymbol{\beta}^\top$ (V x K) $\boldsymbol{\theta}^\top$ (K x D) $\boldsymbol{N}\,\boldsymbol{I}$ (D x D)

$\mathbb{E}[\boldsymbol{X}]$ (D x V) $=$ $\boldsymbol{N}\,\boldsymbol{I}$ (D x D) $\boldsymbol{\theta}$ (D x K) $\boldsymbol{\beta}$ (K x V)

Topic Models: *Summary so far*

Core Idea:

Model documents as ***mixtures*** over topics

Model Parameters:

θ_d Topic probabilities for each document

(K-dimensional vector)

β_k Word probabilities for each topic

(V-dimensional vector)

Relationship to Dimensionality Reduction:

Similar to LSA, but assumes Discrete mixture

instead of Gaussian distribution on word counts