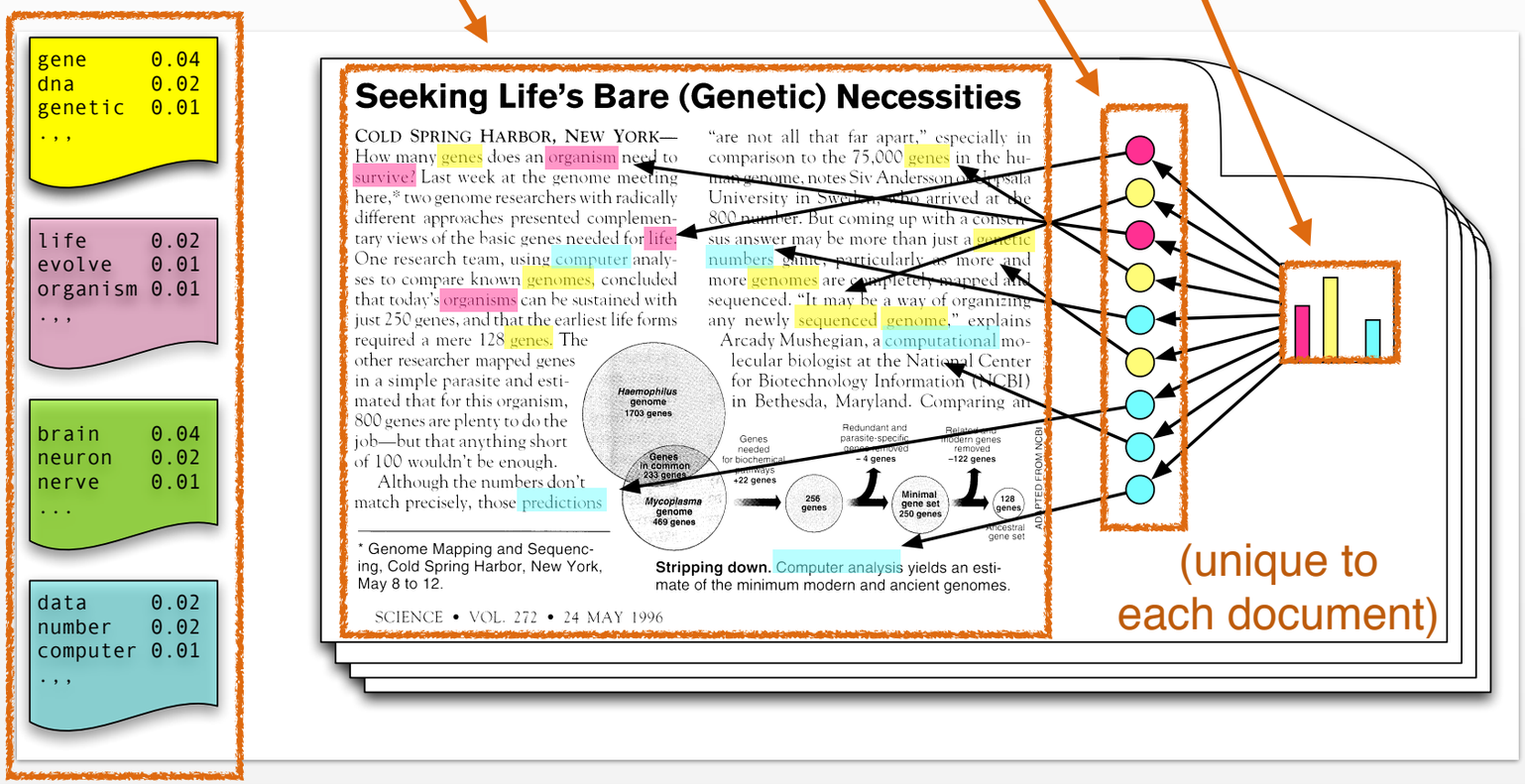
Topic Modeling

β_k: Topics x_d: Words z_d: Assignments θ_d: Topic Proportions



(shared across documents)

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

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

Model Parameters

- $\beta = \{\beta_k\}$: word probabilities, one per topic
- $\theta = \{\theta_d\}$: topic probabilities, one per document

$$p(z_{dn} = k \mid \theta, \beta) = \theta_{dk}$$

$$p(x_{dn} = v \mid z_{dn} = k, \theta, \beta) = \beta_{kv}$$

Interpretation as Matrix Factorization

Log likelihood

$$\log p(x_d \mid \beta, \theta_d) = X_d \log(\theta_d \beta)^T$$

$$\log p(x \mid \beta, \theta) = \sum_{d=1}^{D} X_d \log(\theta_d \beta)^T$$

Bag of word Vector

$X_d = [X_{d1}, X_{d2}, \ \dots \ X_{dV}]$,

$X_{dv} = \sum_{n=1}^{N_d} I[x_{dn} = v]$

Difficult to optimize with gradient ascent.

Need to use EM!

$$\theta = \begin{bmatrix} \theta_{11}, \theta_{12}, & \dots & \theta_{1K} \\ \theta_{21}, \theta_{22}, & \dots & \theta_{2K} \\ & \vdots & \\ \theta_{D1}, \theta_{D2}, & \dots & \theta_{DK} \end{bmatrix} \begin{matrix} \theta_1 \\ \theta_2 \\ \\ \theta_D \end{matrix} \qquad \beta = \begin{bmatrix} \beta_{11}, \beta_{12}, & \dots & \beta_{1V} \\ \beta_{21}, \beta_{22}, & \dots & \beta_{2V} \\ & \vdots & \\ \beta_{K1}, \beta_{K2}, & \dots & \beta_{KV} \end{bmatrix} \begin{matrix} \beta_1 \\ \beta_2 \\ \\ \beta_K \end{matrix}$$

Sketch of EM derivation

Topic models

Observed variable

- $x = \{x_{nd}\}$: all words in the corpus.

Unobserved variable

- $z = \{z_{nd}\}$: the topics of all words in the corpus

Model Parameters

- $\beta = \{\beta_k\}$: word probabilities, one per topic
- $\theta = \{\theta_d\}$: topic probabilities, one per document

$\eta = \{\theta, \beta\}$

Log-likelihood

$l(\eta) = \log p(x \mid \eta)$

1) Initialize η as η_0

2) For $t = 0,1\ldots$, repeat until convergence

2a) E-Step:

$\mathcal{Q}^t(\eta) = \mathbf{E}_{q^t(z)}[\log p(x, z \mid \eta)]$,

where $q^t(z) = p(z \mid x, \eta^t)$

2b) M-Step:

$\eta^{t+1} \leftarrow \text{argmax}_\eta \mathcal{Q}^t(\eta)$

Clustering: GMM

Observed variable

- $x = \{x_n\}$: all datapoint

Unobserved variable

- $z = \{z_n\}$: the cluster index of all datapoint

Model Parameters

- $\mu = \{\mu_k\}$: the means of all gaussian clusters.
- $\Sigma = \{\Sigma_k\}$: the covariances of all gaussian clusters
- α: the vector of cluster probabilities

$\eta = \{\alpha, \mu, \Sigma\}$

Log-likelihood

$l(\eta) = \log p(x \mid \eta)$

1) Initialize η as η_0

2) For $t = 0,1\ldots$, repeat until convergence

2a) E-Step:

$\mathcal{Q}^t(\eta) = \mathbf{E}_{q^t(z)}[\log p(x, z \mid \eta)]$,

where $q^t(z) = p(z \mid x, \eta^t)$

2b) M-Step:

$\eta^{t+1} \leftarrow \text{argmax}_\eta \mathcal{Q}^t(\eta)$

PLSI/PLSA*: EM for Topic Models

Generative Model

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

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

Need to evaluate expectation w.r.t. $q^t(z) = p(z \mid x, \theta^t, \beta^t)$ for the E-Step. This is a joint distribution over all the topic assignment variables, one per word in the corpus. It can be decomposed as a product of $\phi_{d,n,k}$.

E-step: Update assignments

Calculate probability that word *n* in document *d* belongs to topic *k*

$$\phi_{d,n,k} = p(z_{d,n}=k \mid x_{d,n}, \beta, \theta_d)$$

M-step: Update parameters

Use assignment probabilities ϕ_d to update topics assignment probabilities θ_d and topic word probabilities β_k

*(Probabilistic Latent Semantic Indexing, a.k.a. Probabilistic Latent Semantic Analysis)

PLSI/PLSA: E-step

$$\phi_{d,n,k} = p(z_{d,n}=k \mid x_{d,n}=v, \beta, \theta_d)$$

$$= \frac{p(x_{d,n}=v,\ z_{d,n}=k \mid \beta, \theta_d)}{p(x_{d,n}=v \mid \beta, \theta_d)}$$

(Apply Bayes' Rule)

$$p(z_{dn}=k \mid \theta, \beta) = \theta_{dk}$$

$$p(x_{dn}=v \mid z_{dn}=k, \theta, \beta) = \beta_{kv}$$

Computationally this form is better →

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

(Substitute results from previous slides)

General Form, with One-hot Indexing Trick

This form is clearly defined →

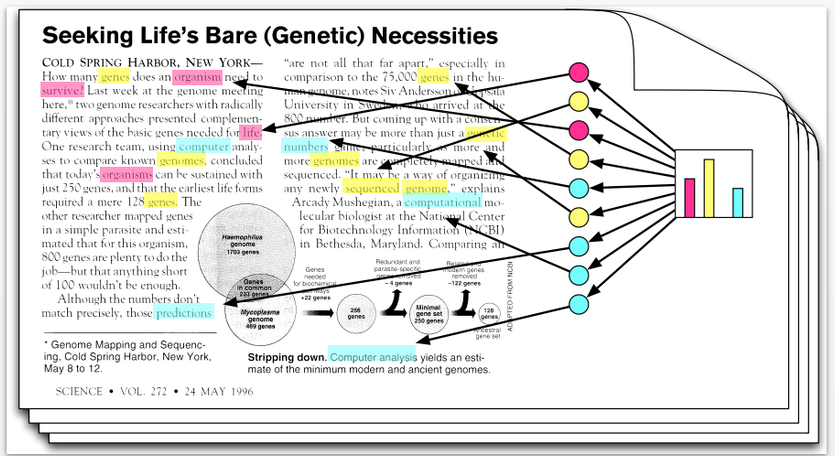
$$\phi_{d,n,k} = \frac{\theta_{d,k}\left(\sum_{v=1}^{V}\beta_{k,v}I[x_{d,n}=v]\right)}{\sum_{l=1}^{K}\theta_{d,l}\left(\sum_{v=1}^{V}\beta_{l,v}I[x_{d,n}=v]\right)}$$

PLSI/PLSA*: EM for Topic Models

Generative Model

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

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



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 completed, 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

E-step: Update assignments

$$\phi_{d,n,k} = p(z_{d,n} = k \mid x_{d,n} = v, \beta, \theta_d)$$

$$= \frac{\theta_{d,k}\left(\sum_{v=1}^{V} \beta_{k,v} I[x_{d,n} = v]\right)}{\sum_{l=1}^{K} \theta_{d,l}\left(\sum_{v=1}^{V} \beta_{l,v} I[x_{d,n} = v]\right)}$$

M-step: Update parameters

Use assignment probabilities ϕ_d to update topics assignment probabilities θ_d and topic word probabilities β_k

*(Probabilistic Latent Semantic Indexing, a.k.a. Probabilistic Latent Semantic Analysis)

PLSI/PLSA: M-Step

Idea: Compute (expected) sufficient statistics

$\phi_{d,n,k}$ Probability that word *n* in document *d* belongs to topic *k*

$\text{TC}_{dk} = \sum_{n=1}^{N_d} \phi_{d,n,k}$ Number of words in document *d* that belong to topic *k*

$\text{WC}_{kv} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,k} I[x_{d,n} = v]$ Number of times word *v* appears in topic *k* (across ***all*** documents in corpus)

M-Step: Update parameters as

$\theta_{d,k} = \frac{\text{TC}_{dk}}{N_d}$ Fraction of topic *k* in document *d*

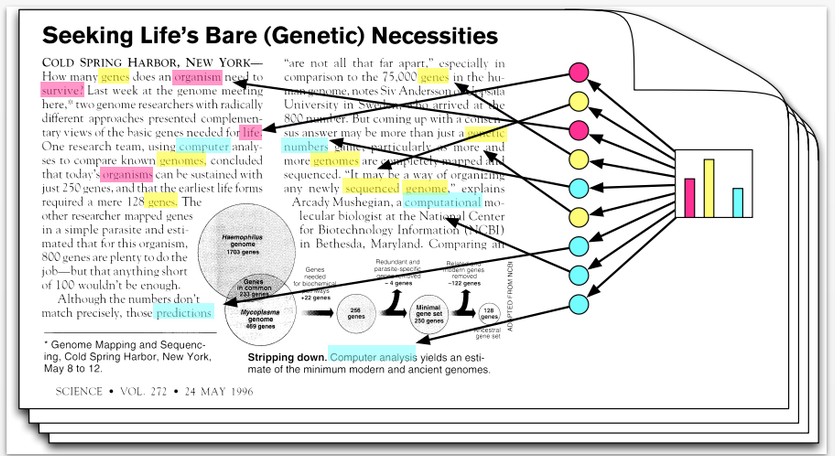
$\beta_{k,v} = \frac{\text{WC}_{kv}}{\sum_{d=1}^{D} \text{TC}_{dk}}$ Fraction of word *v* in topic *k*

PLSI/PLSA*: EM for Topic Models

Generative Model

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

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



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 matter of numbers, since, particularly as more and more genomes are completed, 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

E-step: Update assignments

$$\phi_{d,n,k} = p(z_{d,n} = k \mid x_{d,n} = v, \beta, \theta_d)$$

$$= \frac{\theta_{d,k}\left(\sum_{v=1}^{V} \beta_{k,v} I[x_{d,n} = v]\right)}{\sum_{l=1}^{K} \theta_{d,l}\left(\sum_{v=1}^{V} \beta_{l,v} I[x_{d,n} = v]\right)}$$

M-step: Update parameters

$$\beta_{k,v} = \frac{\text{WC}_{kv}}{\sum_{d=1}^{D} \text{TC}_{dk}} \qquad \text{WC}_{kv} = \sum_{d=1}^{D}\sum_{n=1}^{N_d} \phi_{d,n,k} I[x_{d,n} = v]$$

$$\theta_{d,k} = \frac{\text{TC}_{dk}}{N_d} \qquad \text{TC}_{dk} = \sum_{n=1}^{N_d} \phi_{d,n,k}$$

*(Probabilistic Latent Semantic Indexing, a.k.a. Probabilistic Latent Semantic Analysis)

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

of topics instead of a linear combination of principle components.

PLSI/PLSA:

EM algorithm for maximum likelihood estimation

Topic Models

Shantanu Jain

Latent Dirichlet Allocation

Topic Models with Dirichlet Priors

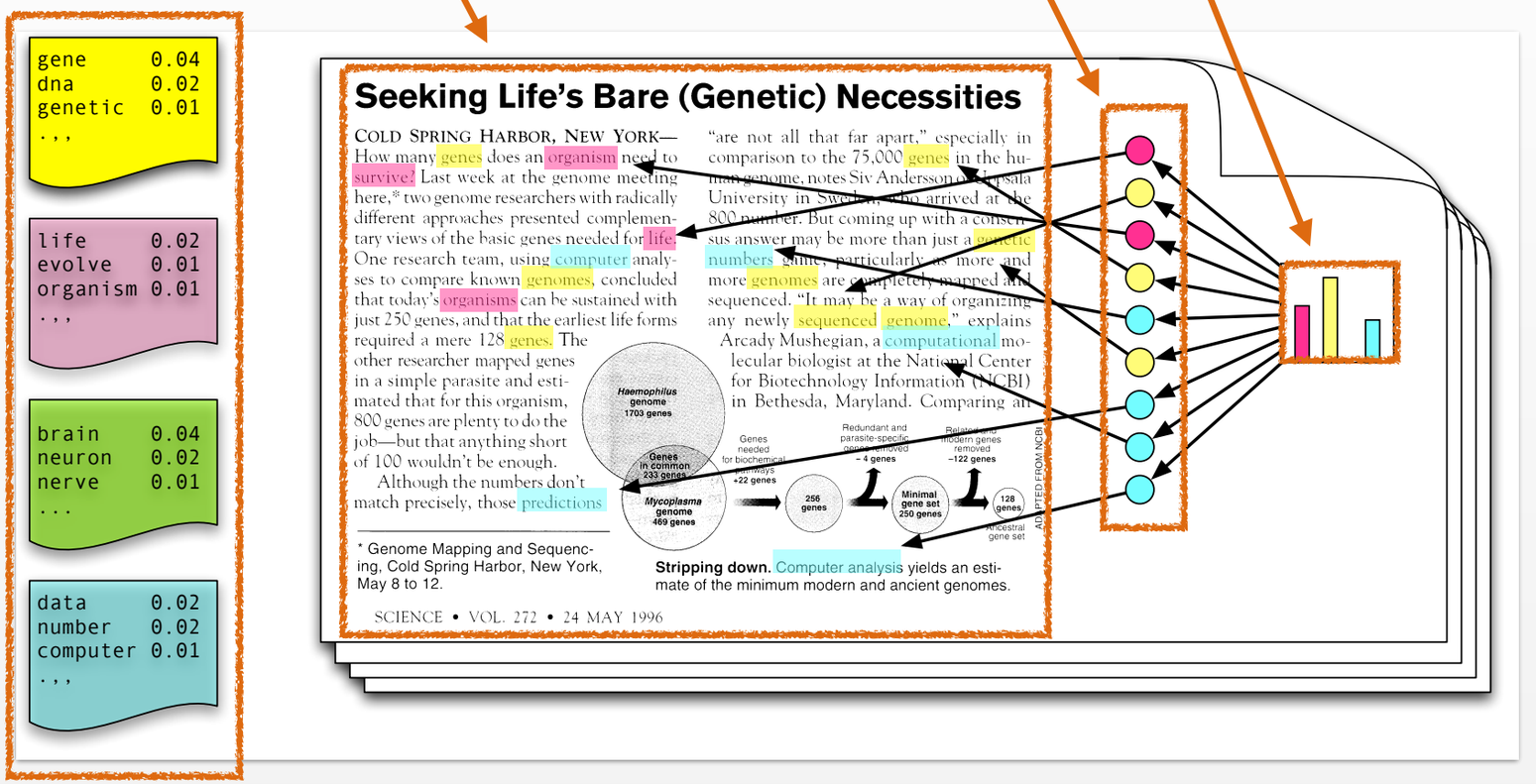
Review: Topic Modeling with PLSA/PLSI

β_k: Topics (shared)

x_d: Words

z_d: Assignments (document-specific)

θ_d: Topic Proportions



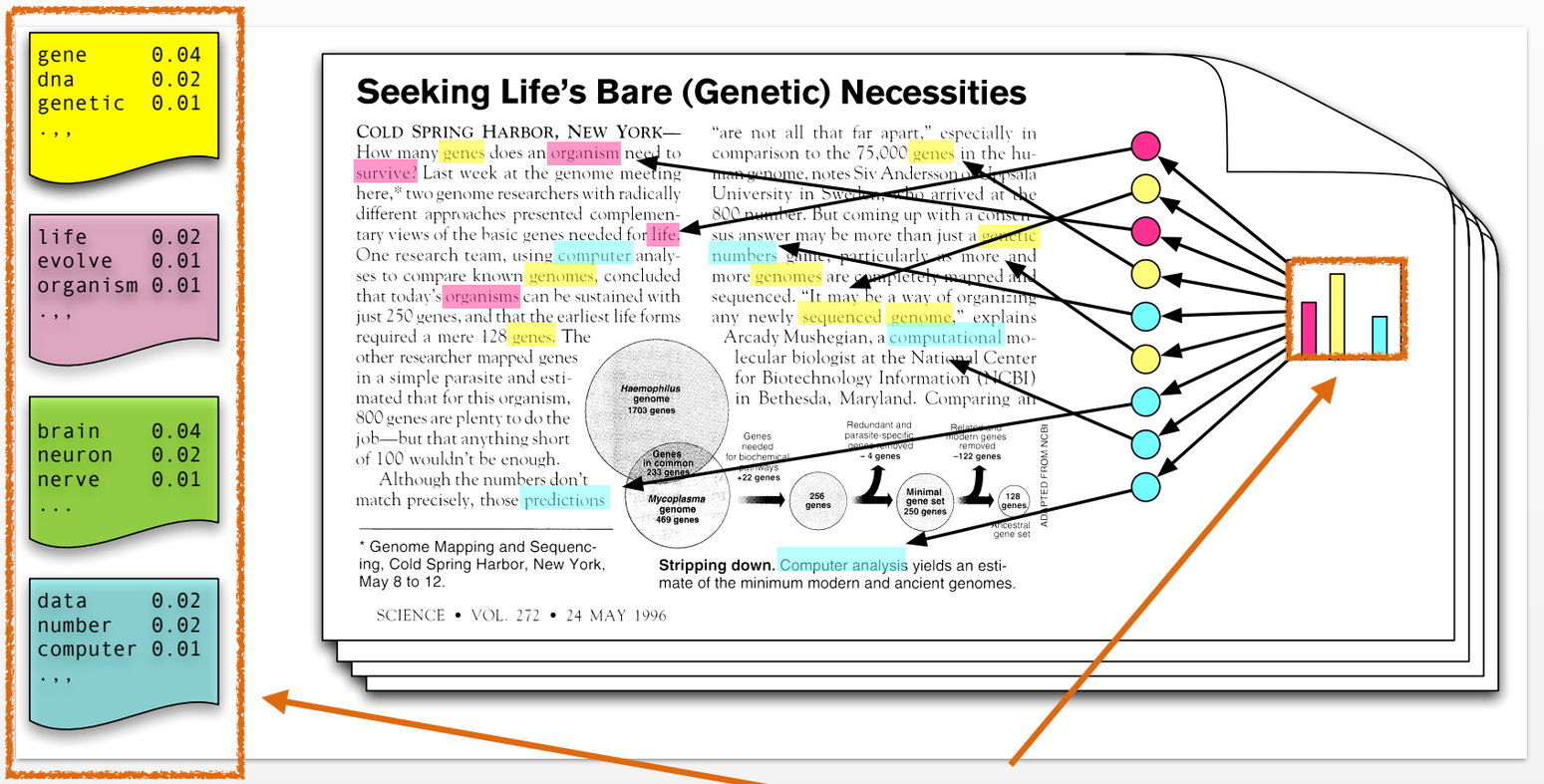
$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

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

- Very flexible model.
- Overfits in practice.
- Doesn't give a sparse fit.

LDA: Add *Dirichlet* Priors

β_k: Topics (shared) | x_d: Words | z_d: Assignments (document-specific) | θ_d: Topic Proportions



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

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

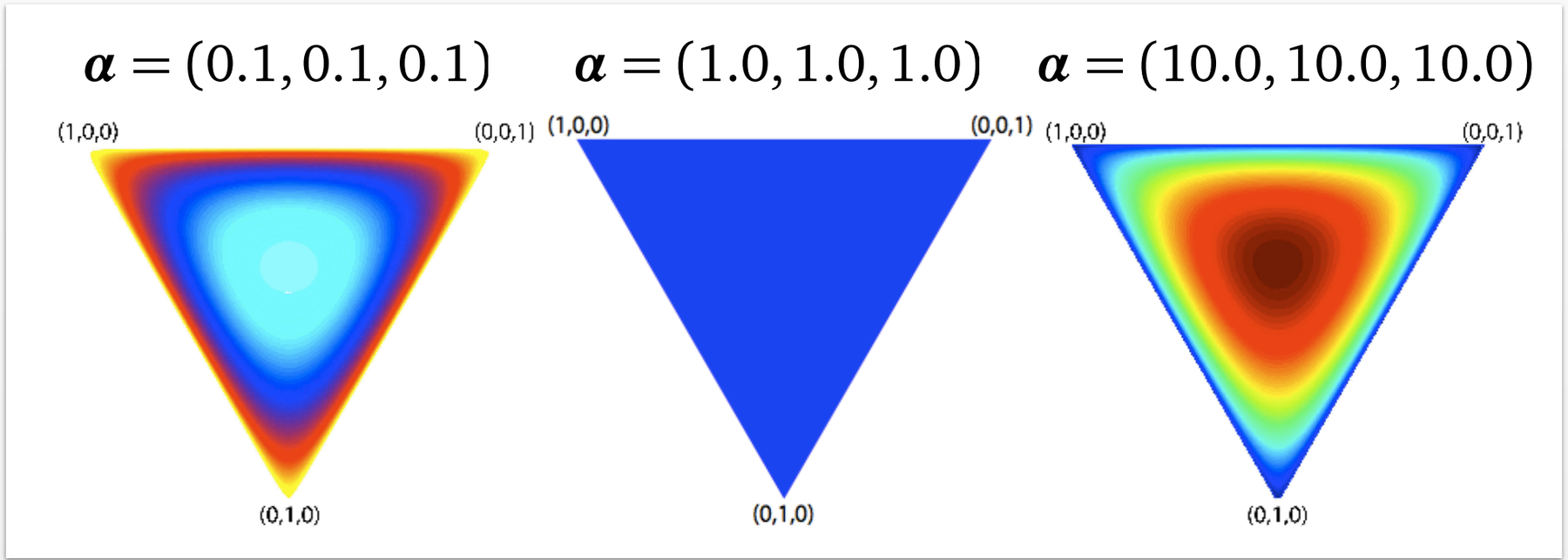
$$\theta_d \sim \text{Dirichlet}(\alpha)$$

$$\beta_k \sim \text{Dirichlet}(\eta_k)$$

Review: Dirichlet Distribution

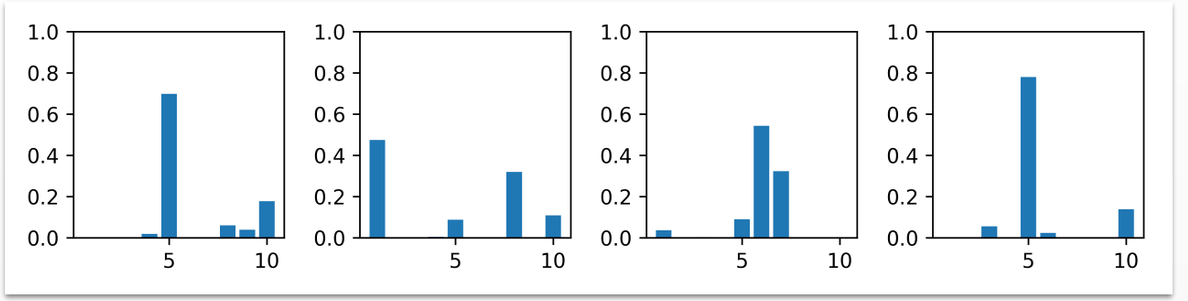
$\theta \sim \text{Dirichlet}(\alpha)$

$$p(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad B(\boldsymbol{\alpha}) := \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

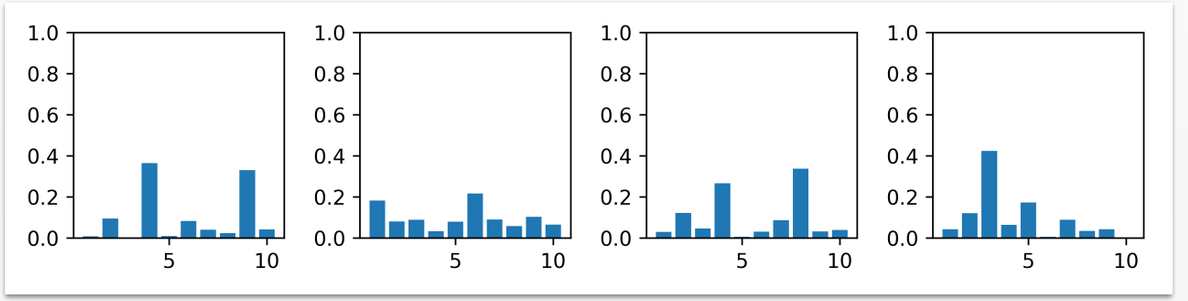


Review: Dirichlet Distribution

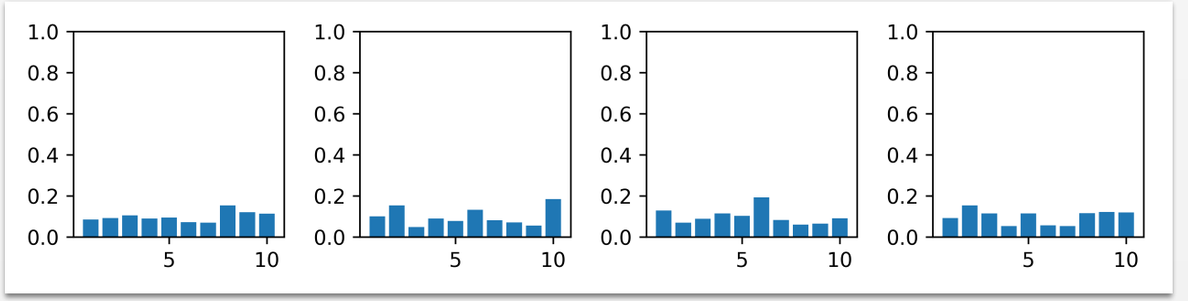
$\alpha_k = 0.1$



$\alpha_k = 1.0$



$\alpha_k = 10.0$



LDA: $\alpha_k = 0.001$ – Enforces Sparsity of Topic Weights θ_d

Maximum likelihood estimation

X: a dataset with N points:

$$X = \{x_n\}_{n=1}^{N}$$

$x_n \in \mathcal{X}$: the space where x_n takes values. Typically $\mathcal{X} = \mathbb{R}^D$

Model Assumption:

$$x_n \sim P_\theta$$

A distribution parametrized by θ (unknown, but not random) defined on the sample space the space $\mathcal{X}$

Maximum likelihood estimate of θ.

$$\theta^{ML} = \text{argmax}_\theta \; p(X|\theta)$$
$$= \text{argmax}_\theta \; \log p(X|\theta)$$

Does not depend on the prior distribution, $p(\theta)$.

Maximum likelihood is a frequents approach. It assumes that the parameters are unknown, but fixed, in the sense that they do not come from a distribution. In the frequentist world the prior distribution does not exist.

Bayesian inference

X: a dataset with N points:

$$X = \{x_n\}_{n=1}^{N}$$

Model Assumption:

$x_n \sim P_\theta$ The data distribution, parametrized by θ

$\theta \sim P_\alpha$ The prior distribution on the parameter θ. In the Bayesian setting, the parameter itself is a random variable having a "prior" distribution, P_α. The prior distribution parameter, α, is called the hyper-parameter. Which is typically assumed to be know or searched via grid-search.

Posterior distribution

$$p(\theta \mid X, \alpha) = \frac{p(X \mid \theta, \alpha)p(\theta \mid \alpha)}{P(X \mid \alpha)}$$

$$= \frac{p(X \mid \theta)p(\theta \mid \alpha)}{P(X \mid \alpha)}$$

Likelihood

Prior

Intractable marginal

The most general goal of Bayesian statistics is to learn the entire posterior distribution, not just an estimate of the parameter. However, this is often very difficult due to the intractable denominator. Have to rely on approximate inference.

As the number of data points increase the contribution of the likelihood term increases and the prior term reduces

$$P(X \mid \theta) = \int_\theta p(X \mid \theta)p(\theta \mid \alpha)d\theta$$

Maximum a posteriori (MAP) estimation

X: a dataset with N points:

$$X = \{x_n\}_{n=1}^{N}$$

MAP Estimate

Posterior

$$\theta^{MAP} = \text{argmax}_\theta \; p(\theta \mid X, \alpha)$$

$$= \text{argmax}_\theta \; \frac{p(X \mid \theta, \alpha) p(\theta \mid \alpha)}{P(X \mid \alpha)}$$

$$= \text{argmax}_\theta \; \frac{p(X \mid \theta) p(\theta \mid \alpha)}{P(X \mid \alpha)}$$

Not a function of θ

$$= \text{argmax}_\theta \; p(X \mid \theta) \times p(\theta \mid \alpha)$$

Likelihood — Prior

As the number of data points increase the contribution of the likelihood term increases and the prior term reduces

Model Assumption:

$x_n \sim P_\theta$ The data distribution, parametrized by θ

$\theta \sim P_\alpha$ The prior distribution on the parameter θ. In the Bayesian setting, the parameter itself is a random variable having a “prior” distribution, P_α. The prior distribution parameter, α, is called the hyper-parameter. Which is typically assumed to be know or searched via grid-search.

$$\theta^{ML} = \text{argmax}_\theta \; p(X \mid \theta)$$

$$= \text{argmax}_\theta \; \log p(X \mid \theta)$$

Estimating Model Parameters

Question: How can we estimate β_k and θ_d?

1. MAP with Expectation Maximization
2. Variational Inference
   (high level)
3. Gibbs Sampling
   (not in this module)

MAP generalization to Discrete distribution

Likelihood

$$p(Z|\theta) = \prod_{n=1}^{N} p(z_n|\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K} \theta_k^{I[x_n=k]}$$

Prior

$$p(\theta|\alpha) = \frac{\prod_{k=1}^{K}\theta_k^{\alpha_k}}{\mathrm{B}(\alpha)}$$

$$p(\theta|Z,\alpha) \propto p(Z|\theta)p(\theta|\alpha)$$

Posterior

$$\propto \left(\prod_{n=1}^{N}\prod_{k=1}^{K}\theta_k^{I[z_n=k]}\right)\frac{\prod_{k=1}^{K}\theta_k^{\alpha_k}}{\mathrm{B}(\alpha)}$$

← Not a function of μ

$$\propto \prod_{k=1}^{K}\theta_k^{\left[\sum_{n=1}^{N} I[z_n=k]\right]}\prod_{k=1}^{K}\theta_k^{\alpha_k}$$

$$\propto \prod_{k=1}^{K}\theta_k^{\mathrm{TC}_k+\alpha_k}$$

$$\mathrm{TC}_k = \sum_{n=1}^{N} I[z_n = k]$$

$$= \mathrm{Dirichlet}(\theta|\mathrm{TC}_1+\alpha_1, \mathrm{TC}_2+\alpha_2, \ldots, \mathrm{TC}_K+\alpha_K)$$

MAP Estimate

$$\theta_k^{MAP} = \frac{\mathrm{TC}_k+\alpha_k-1}{N+\sum_{k=1}^{K}\alpha_k - K}$$

Simpler example: single document, observable topics

$$Z = \{z_n\}_{n=1}^{N} \quad z_n \in \{1,2,\ldots,K\}$$

$$z_n \sim \mathrm{Discrete}(\theta) \qquad \sum_{k=1}^{K}\theta_k = 1$$

$$p(z_n|\theta) = \begin{cases} \theta_1 & \text{if } z_n = 1 \\ \theta_2 & \text{if } z_n = 2 \\ \vdots & \\ \theta_K & \text{if } z_n = K \end{cases}$$

$$p(z_n|\theta) = \prod_{k=1}^{K}\theta_k^{I[z_n=k]}$$

$$\theta \sim \mathrm{Dirichlet}(\alpha)$$

$$p(\theta|\alpha) = \frac{\prod_{k=1}^{K}\theta_k^{\alpha_k}}{\mathrm{B}(\alpha)}$$

$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K] \qquad \alpha_k > 0$$

MAP generalization to Discrete distribution

MAP Estimate

$$\theta_{dk}^{MAP} = \frac{\mathrm{TC}_{dk} + \alpha_k - 1}{N_d + \sum_{k=1}^{K} \alpha_k - K}$$

$$\mathrm{TC}_{dk} = \sum_{n=1}^{N_d} I[z_{dn} = k]$$

Count of topic k in the the d^{th} document

Extension to multiple document.

$$Z = \{z_{dn}\} \quad z_{dn} \in \{1,2,\ldots,K\}$$

$$z_{dn} \sim \text{Discrete}(\theta_d) \quad \sum_{k=1}^{K} \theta_{dk} = 1$$

$$p(z_{dn} \mid \theta) = \prod_{k=1}^{K} \theta_{dk}^{I[z_{dn}=k]}$$

$$\theta = \{\theta_d\}$$

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

$$p(\theta_d \mid \alpha) = \frac{\prod_{k=1}^{K} \theta_{dk}^{\alpha_k}}{\mathrm{B}(\alpha)}$$

$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K] \quad \alpha_k > 0$$

MAP generalization to Discrete distribution

MAP Estimate

$$\beta_{kv}^{MAP} = \frac{\text{WC}_{kv} + \eta_{kv} - 1}{\sum_{d=1}^{D} \text{TC}_{dk} + \sum_{v=1}^{V} \eta_{kv} - V}$$

$$\text{WC}_{kv} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} I[z_{dn} = k] I[x_{dn} = v]$$

$$\theta_{dk}^{MAP} = \frac{\text{TC}_{dk} + \alpha_k - 1}{N_d + \sum_{k=1}^{K} \alpha_k - K}$$

$$\text{TC}_{dk} = \sum_{n=1}^{N_d} I[z_{dn} = k]$$

Count of topic k in the the d^{th} document

Estimating word probabilities multiple document.

Simplifying assumption: Topics are observed

$$Z = \{z_{dn}\} \quad z_{dn} \in \{1,2,\ldots,K\}$$

$$X = \{x_{dn}\} \quad x_{dn} \in \{1,2,\ldots,V\}$$

$$z_{dn} \sim \text{Discrete}(\theta_d) \quad \sum_{k=1}^{K} \theta_{dk} = 1$$

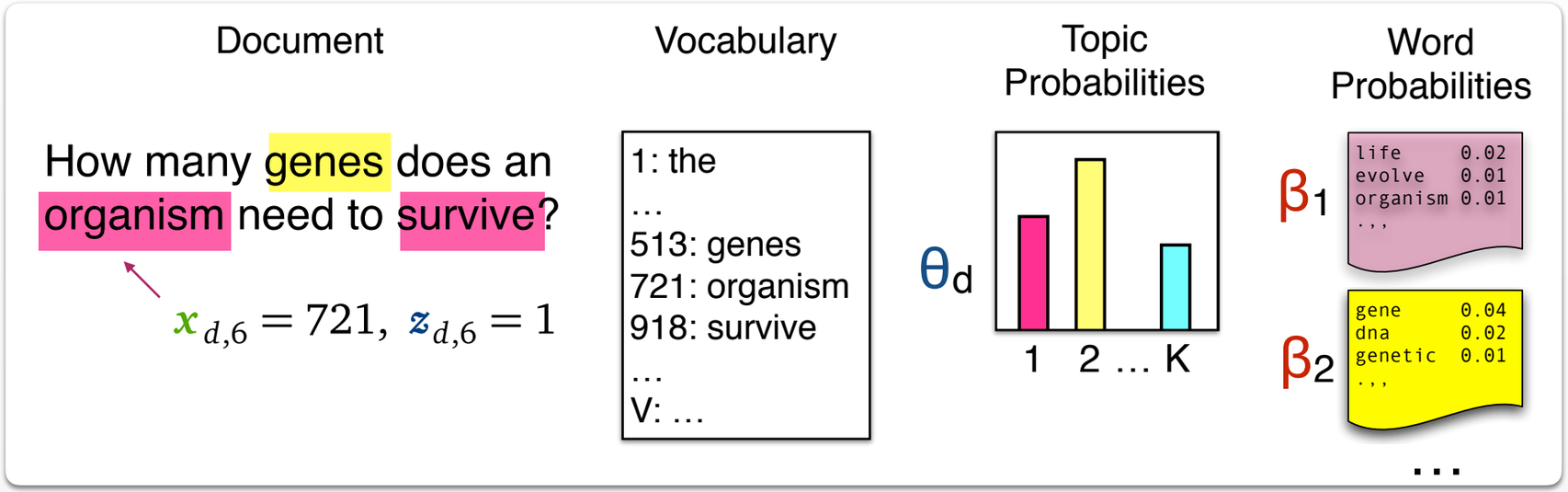
$$\theta_d \sim \text{Dirichlet}(\alpha) \qquad \theta = \{\theta_d\}$$

$$x_{dn} \mid z_{dn} = k \sim \text{Discrete}(\beta_k)$$

$$\beta_k \sim \text{Dirichlet}(\eta_k) \quad \sum_{k=1}^{K} \beta_{kv} = 1$$

$$\beta = \{\beta_k\} \quad \eta = \{\eta_k\}$$

Estimating the Parameters



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

Maximum a Posteriori: $\max_{\theta,\beta} \log p(\theta, \beta \mid x, \alpha, \eta)$ $\eta = \{\eta_k\}$

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

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

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

$$\beta_k \sim \text{Dirichlet}(\eta_k)$$

MAP estimation for LDA with EM

Generative Model

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

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\eta}_k)$$

$$z_{d,n} \sim \text{Discrete}(\boldsymbol{\theta}_d)$$

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

E-step: Update assignments

$$\phi_{d,n,k} = p(z_{d,n}=k \mid x_{d,n}=v, \boldsymbol{\beta}, \boldsymbol{\theta}_d)$$

$$= \frac{\theta_{d,k}\left(\sum_{v=1}^{V} \beta_{k,v} I[x_{d,n}=v]\right)}{\sum_{l=1}^{K} \theta_{d,l}\left(\sum_{v=1}^{V} \beta_{l,v} I[x_{d,n}=v]\right)}$$

M-step: Update parameters

$$\text{TC}_{dk} = \sum_{n=1}^{N_d} I[z_{dn}=k]$$

$$\text{TC}_{dk} = \sum_{n=1}^{N_d} \phi_{dnk}$$

$$\text{WC}_{kv} = \sum_{d=1}^{D}\sum_{n=1}^{N_d} I[z_{dn}=k] I[x_{dn}=v]$$

$$\text{WC}_{kv} = \sum_{d=1}^{D}\sum_{n=1}^{N_d} \phi_{dnk} I[x_{dn}=v]$$

$$\beta_{kv}^{MAP} = \frac{\text{WC}_{kv}+\eta_{kv}-1}{\sum_{d=1}^{D}\text{TC}_{dk}+\sum_{v=1}^{V}\eta_{kv}-V}$$

$$\theta_{dk}^{MAP} = \frac{\text{TC}_{dk}+\alpha_k-1}{N_d+\sum_{k=1}^{K}\alpha_k-K}$$

(not used in practice; requires $\alpha_k > 1$ and $\eta_{kv} > 1$)

Variational Expectation Maximization (high-level)

Idea: Approximate $p(z, \theta, \beta \mid x)$ with $q(z, \theta, \beta)$

$$\phi, \gamma, \lambda = \underset{\phi,\gamma,\lambda}{\operatorname{argmin}} \ \mathrm{KL}(q(z, \theta, \beta) \,||\, p(z, \theta, \beta \mid x))$$

$$q(z, \theta, \beta) = q(z; \phi)\, q(\theta; \gamma)\, q(\beta; \lambda)$$

Discrete Dirichlet Dirichlet

Variational E-step: Update ϕ

$$\phi_{d,n,k} = \exp\left(\mathbb{E}_q \left[\log \theta_{d,k} + \sum_{v=1}^{V} I[x_{d,n} = v] \log \beta_{k,v} \right] \right)$$

(won't derive this – but can be computed in closed form)

Variational M-step: Update γ and λ

$$\gamma_{d,k} = \alpha_k + \mathrm{TC}_{dk} \qquad \lambda_{k,v} = \eta_{k,v} + \mathrm{WC}_{kv}$$

(analogous to MAP estimation – need to know this)

EM vs. Variational EM

EM $\quad \boldsymbol{\theta}, \boldsymbol{\beta} = \underset{\boldsymbol{\theta},\boldsymbol{\beta}}{\operatorname{argmax}} \ \log p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\beta})$

E-step: $\quad \boldsymbol{\phi}_{d,n,k} \propto \boldsymbol{\theta}_{d,k} \left(\sum_{v=1}^{V} \boldsymbol{\beta}_{k,v} I[\boldsymbol{x}_{d,n} = v] \right)$

M-step: $\quad \boldsymbol{\theta}_{d,k} = \frac{\mathrm{TC}_{dk}}{N_d} \qquad \boldsymbol{\beta}_{k,v} = \frac{\mathrm{WC}_{kv}}{\sum_{d=1}^{D} \mathrm{TC}_{dk}}$

Variational EM $\quad \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda} = \underset{\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\lambda}}{\operatorname{argmin}} \ \mathrm{KL}(q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) \,||\, p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{x}))$

E-step: $\quad \boldsymbol{\phi}_{d,n,k} = \exp\left(\mathbb{E}_q \left[\log \boldsymbol{\theta}_{d,k} + \sum_{v=1}^{V} I[\boldsymbol{x}_{d,n} = v] \log \boldsymbol{\beta}_{k,v} \right] \right)$

M-step: $\quad \boldsymbol{\gamma}_{d,k} = \boldsymbol{\alpha}_k + \mathrm{TC}_{dk} \qquad \boldsymbol{\lambda}_{k,v} = \boldsymbol{\eta}_{k,v} + \mathrm{WC}_{kv}$

EM vs. Variational EM

Commonalities: Both compute sufficient statistics

$\phi_{d,n,k}$ Probability that word *n* in document *d* belongs to topic *k*

TC_{dk} Number of words in document *d* that belong to topic *k*

WC_{kv} Number of times word *v* appears in topic *k* (across ***all*** documents in corpus)

Differences: Point estimates vs Distributions

EM: Computes most likely values for parameters

$\theta_{d,k}$ Fraction of words in document *d* for topic *k*

$\beta_{k,v}$ Fraction of words in topic *k* for vocabulary entry *v*

Variational EM: Estimate *Posterior* over Parameters

$q(\theta_d; \gamma_d)$ Approximation of topic distribution for document *d*

$q(\beta_k; \lambda_k)$ Approximation of word distribution for topic *k*