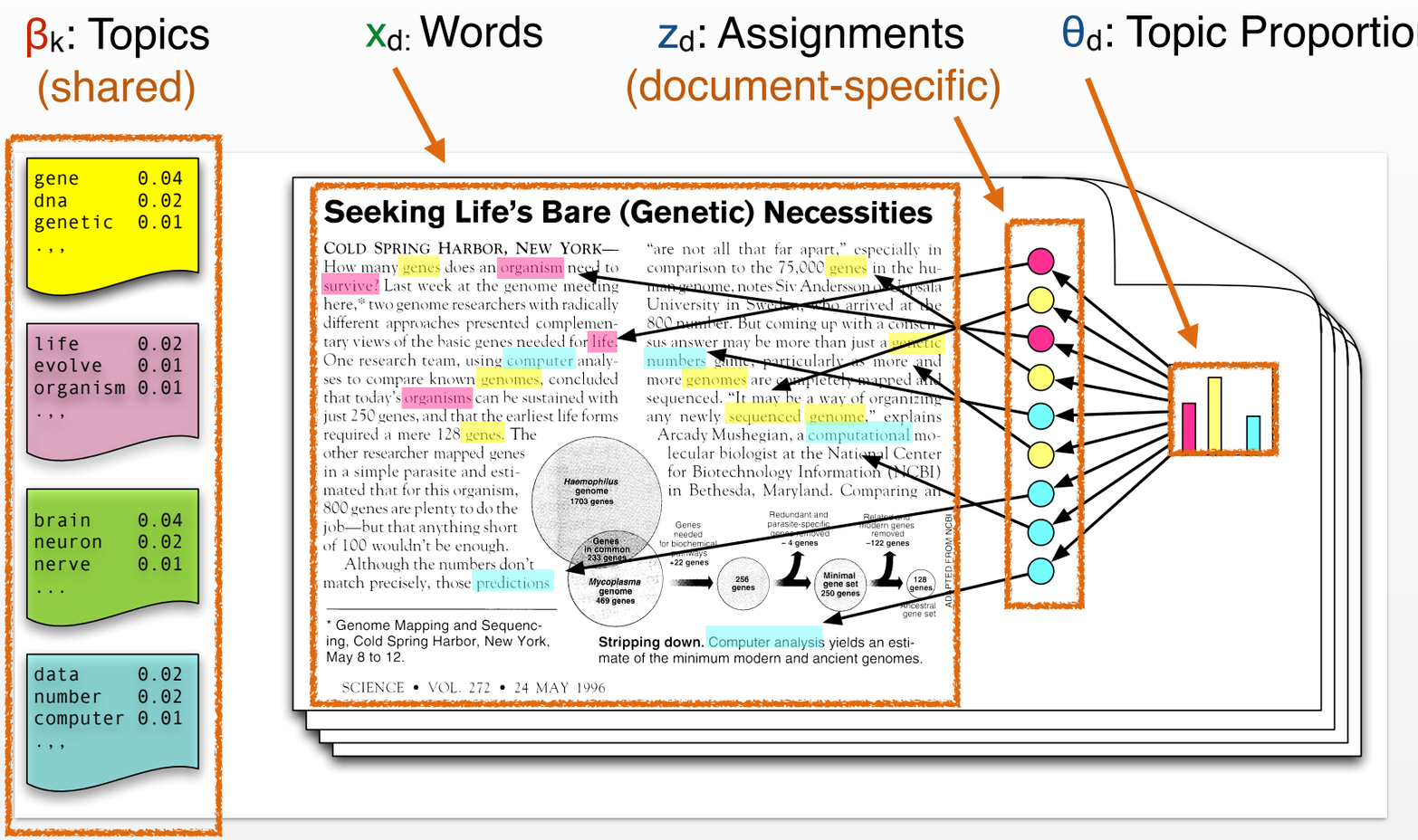
Latent Dirichlet Allocation

Topic Models with Dirichlet Priors

Review: Topic Modeling with PLSA/PLSI



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

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

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

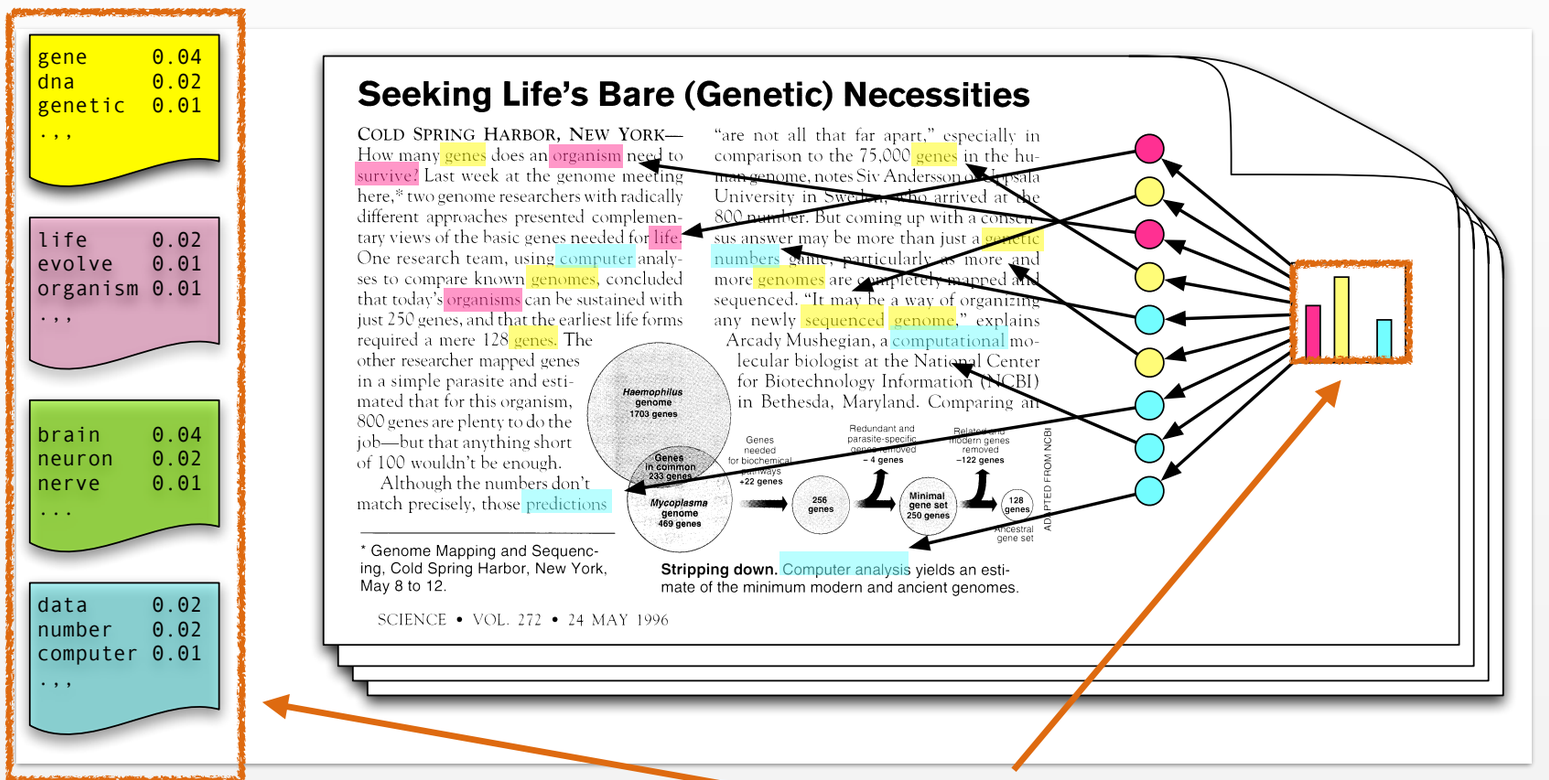
LDA: Add *Dirichlet* Priors

β_k: Topics (shared)

x_d: Words

z_d: Assignments (document-specific)

θ_d: Topic Proportions



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

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

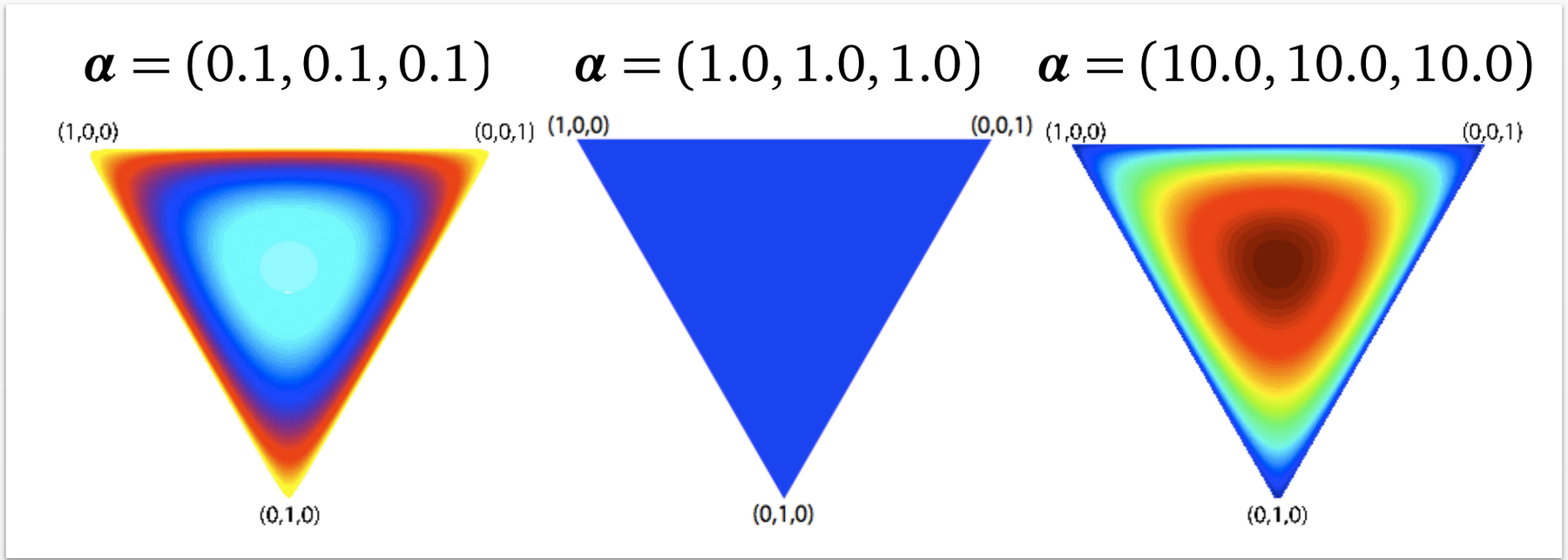
$$\theta_d \sim \text{Dirichlet}(\alpha)$$

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

Review: Dirichlet Distribution

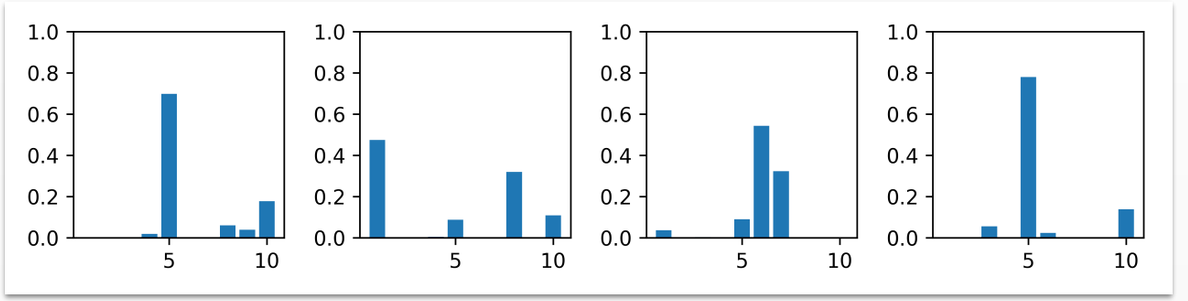
$\theta \sim \text{Dirichlet}(\alpha)$

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

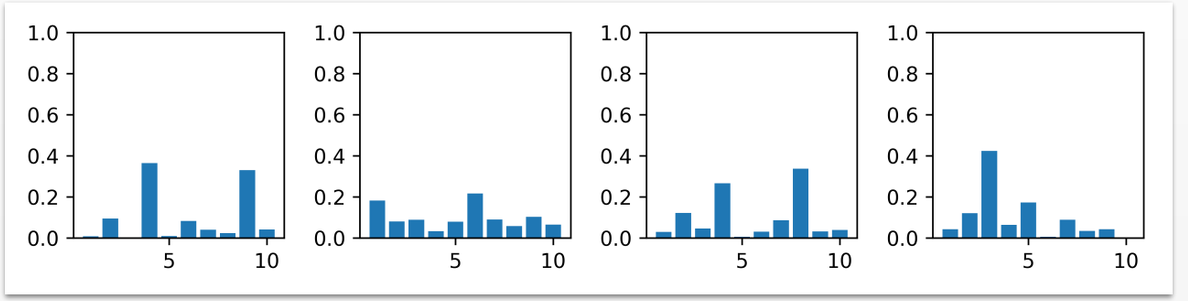


Review: Dirichlet Distribution

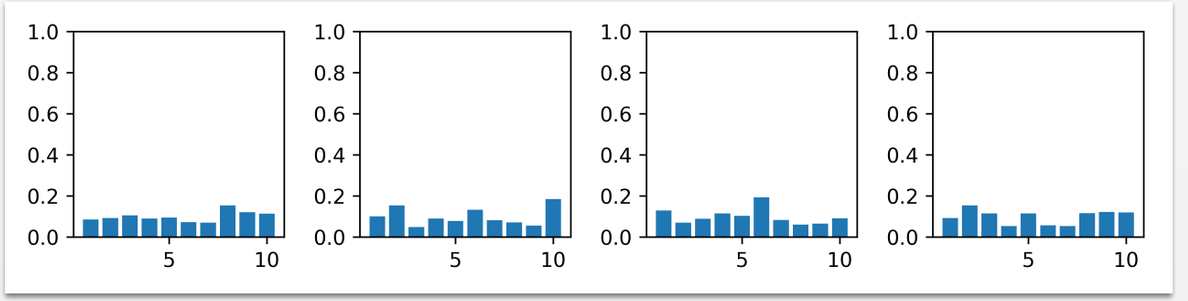
$\alpha_k = 0.1$



$\alpha_k = 1.0$



$\alpha_k = 10.0$



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

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

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

Posterior

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

Not a function of μ

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

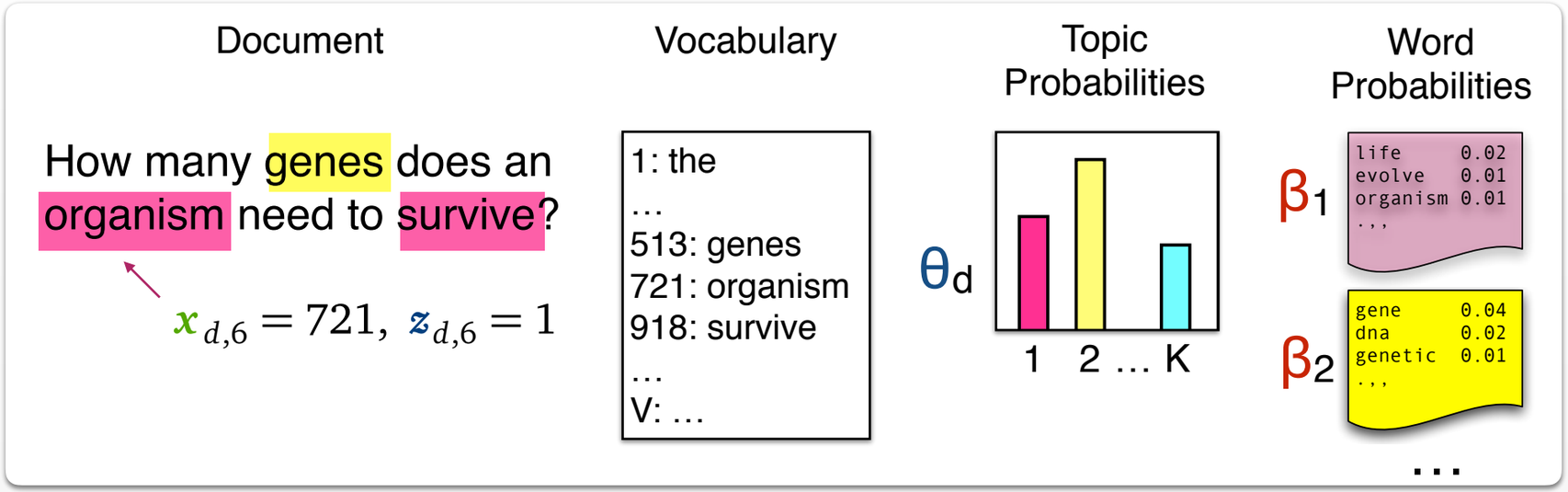
$$\theta_d \sim \text{Dirichlet}(\alpha) \qquad \theta = \{\theta_d\}$$

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

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

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

Estimating the Parameters



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

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

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

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

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

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

MAP estimation for LDA with EM

Generative Model

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

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

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

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

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

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

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

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

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

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

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

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

Estimating the posterior

$$p(Z,\theta,\beta\,|\,X,\alpha,\eta) = \frac{p(Z,\theta,\beta,X\,|\,\alpha,\eta)}{p(X\,|\,\alpha,\eta)}$$

$$= \frac{p(X\,|\,Z,\beta)p(Z\,|\,\theta)p(\theta\,|\,\alpha)p(\beta\,|\,\eta)}{p(X\,|\,\alpha,\eta)}$$

$$= \frac{p(\theta\,|\,\alpha)p(\beta\,|\,\eta)\prod_d\prod_n p(x_{dn}\,|\,z_{dn},\beta)p(z_{dn}\,|\,\theta_d)}{\int_\theta\int_\beta \sum_Z p(Z,\theta,\beta,X\,|\,\alpha,\eta)\; d\theta d\beta}$$

Intractable due to complex dependencies

Observed variables:

X

Unobserved variables:

Z, θ, β

$\theta = \{\theta_d\}$

$\beta = \{\beta_k\}$

Hyper-parameters:

α, η

$\eta = \{\eta_k\}$

Variational Inference approach

Approximate with a simple probability distribution, removing dependencies

Observed variables: X

Unobserved variables: Z, θ, β

$\theta = \{\theta_d\}$

$\beta = \{\beta_k\}$

Hyper-parameters: α, η

$\eta = \{\eta_k\}$

Mean-field assumption

$$q(Z, \theta, \beta \mid X, \alpha, \eta) = q(Z \mid X, \alpha, \beta) \times q(\theta \mid X, \alpha, \beta) \times q(\beta \mid X, \alpha, \eta)$$

$$q(Z, \theta, \beta \mid \phi, \gamma, \lambda) = q(Z \mid \phi(X)) \times q(\theta \mid \gamma(X)) \times q(\beta \mid \lambda(X))$$

Variational Parameters to be learnt from the observed data

$$z_{dn} \sim \text{Discrete}(\phi_{dn1}, \phi_{dn2} \ldots \phi_{dnK})$$

$$\theta_d \sim \text{Dirichlet}(\gamma_{d1}, \gamma_{d2} \ldots \gamma_{dK})$$

$$\beta_k \sim \text{Dirichlet}(\lambda_{k1}, \lambda_{k2} \ldots \lambda_{kV})$$

$$\phi^*, \gamma^*, \lambda^* = \underset{\phi,\gamma,\lambda}{\text{argmin}}\ \text{KL}\left(q(Z, \theta, \beta \mid \phi, \gamma, \lambda) \parallel p(Z, \theta, \beta \mid X, \alpha, \beta)\right)$$

Variational Expectation Maximization (high-level)

Variational E-step: Update ϕ

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

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

$\theta_d \sim \text{Dirichlet}(\gamma_{d1}, \gamma_{d2} \ldots \gamma_{dK})$ $\beta_k \sim \text{Dirichlet}(\lambda_{k1}, \lambda_{k2} \ldots \lambda_{kV})$

Variational M-step: Update γ and λ

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

(analogous to MAP estimation – need to know this)

EM (non-variational)

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

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

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

Take $\log$ on both sides and compare to the variational EM

Variational Expectation Maximization (high-level)

The output of Variational EM is

$$\phi^*, \gamma^*, \lambda^* = \underset{\phi,\gamma,\lambda}{\text{argmin}}\ \text{KL}\Big(q(Z, \theta, \beta \,|\, \phi, \gamma, \lambda) \parallel p(Z, \theta, \beta \,|\, X, \alpha, \beta)\Big)$$

This gives distribution over $z_{dn}, \theta_d, \beta_k$

$$z_{dn} \sim \text{Discrete}(\phi^*_{dn1}, \phi^*_{dn2} \ldots \phi^*_{dnK})$$

$$\theta_d \sim \text{Dirichlet}(\gamma^*_{d1}, \gamma^*_{d2} \ldots \gamma^*_{dK})$$

$$\beta_k \sim \text{Dirichlet}(\lambda^*_{k1}, \lambda^*_{k2} \ldots \lambda^*_{kV})$$

Point estimates can be obtained as

$$\theta^*_{dk} = \mathbf{E}[\theta_{dk} \,|\, \gamma^*_d] = \frac{\gamma_{dk}}{\sum_{k=1}^{K} \gamma_{dk}}$$

$$\beta^*_{kv} = \mathbf{E}[\beta_{kv} \,|\, \lambda^*_k] = \frac{\lambda_{kv}}{\sum_{v=1}^{V} \lambda_{kv}}$$

Point estimates of θ_d, β_k are needed to obtain a representation of a document and a topic, respectively.

$$z^*_{dn} = \text{argmax}_k\ \phi^*_{dnk}$$

The point estimated of the topic (if needed) is obtained as the topic having the highest probability

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 $\quad \phi^*, \gamma^*, \lambda^* = \underset{\phi,\gamma,\lambda}{\operatorname{argmin}} \mathrm{KL}\left(q(Z, \theta, \beta \mid \phi, \gamma, \lambda) \parallel p(Z, \theta, \beta \mid X, \alpha, \beta) \right)$

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

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*

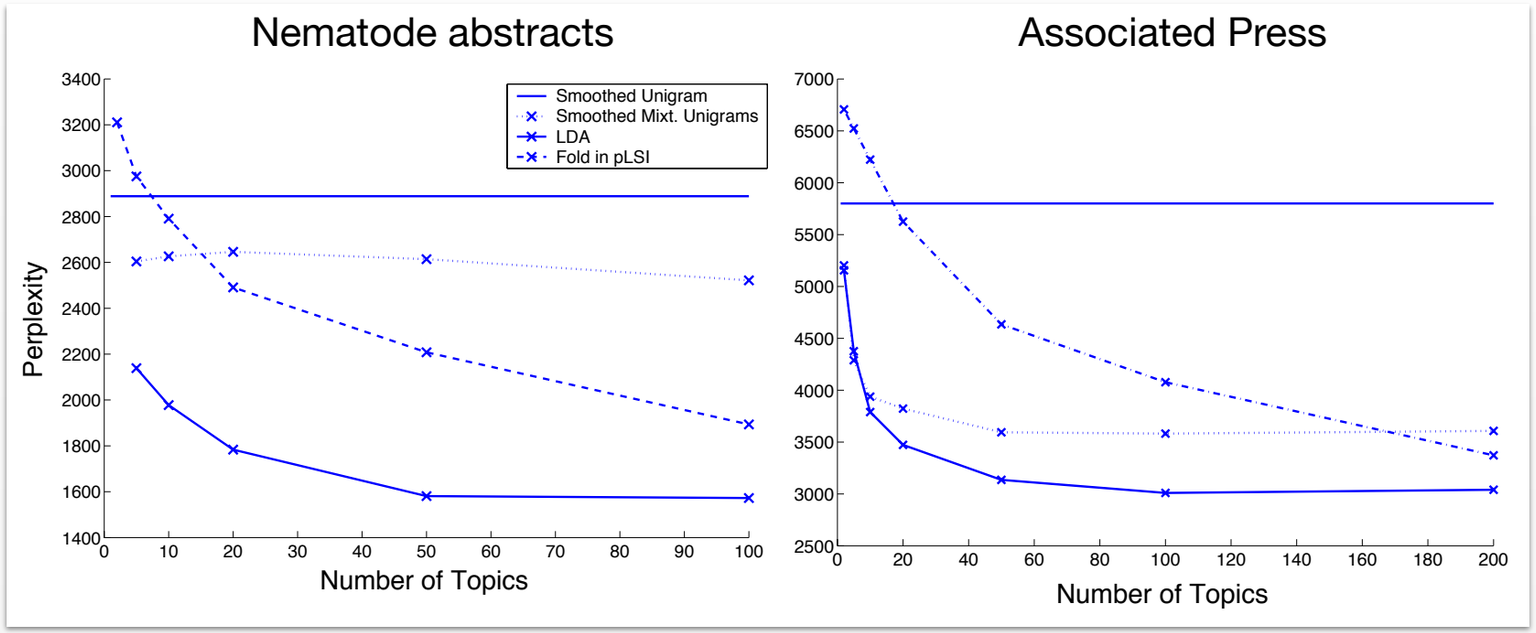
$\beta_{k,v}$ Fraction of words in topic *k* for vocabulary entry *v*

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*

Performance Metric: Perplexity



$$\text{Perplexity} = \exp\left[-\frac{1}{D'}\sum_{d=1}^{D'}\frac{1}{N_d}\log(\boldsymbol{x}'_d \mid \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\eta})\right] \qquad \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{D'}\}$$

Exponent of per-word log predictive probability

Topic Models

Shantanu Jain

Extensions of LDA

Borrowing from:
David Blei
(Columbia)

Extensions of LDA

Latent dirichlet allocation
DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org
Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying ...
Cited by 15971 Related articles All 124 versions Cite Save

Reasons for popularity of LDA:

- Dirichlet prior gives sparser vectors θ_d
- LDA be extended to more sophisticated models

Extensions of LDA

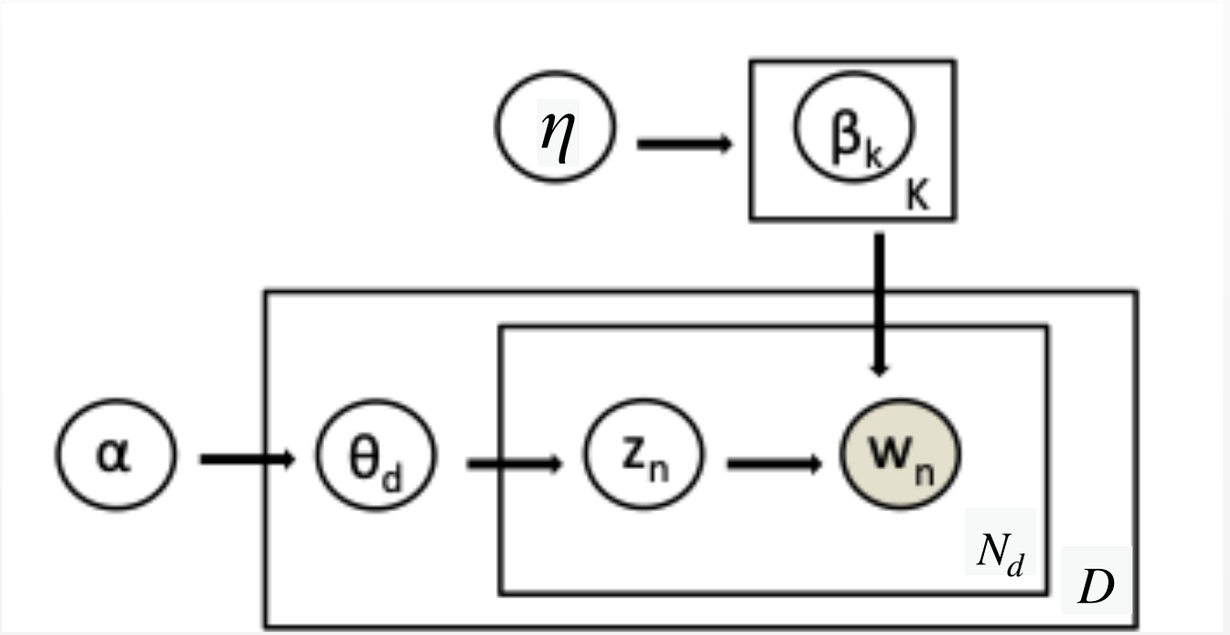
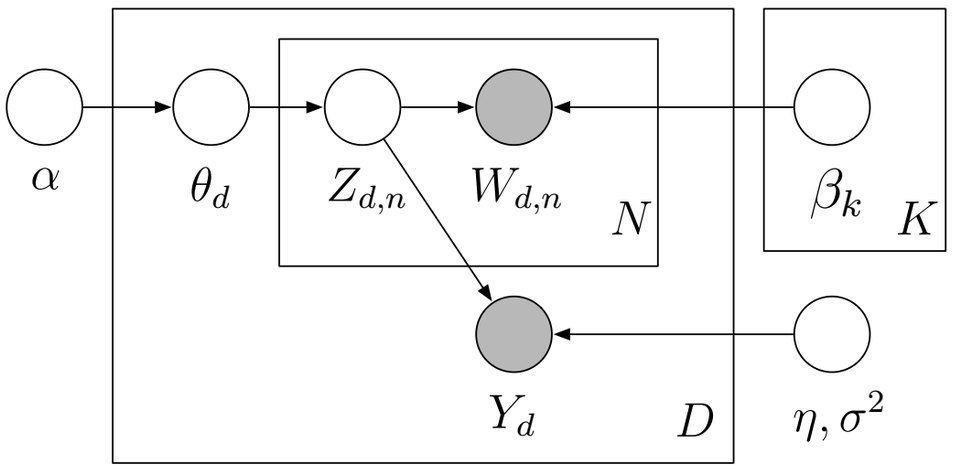


Plate Diagram

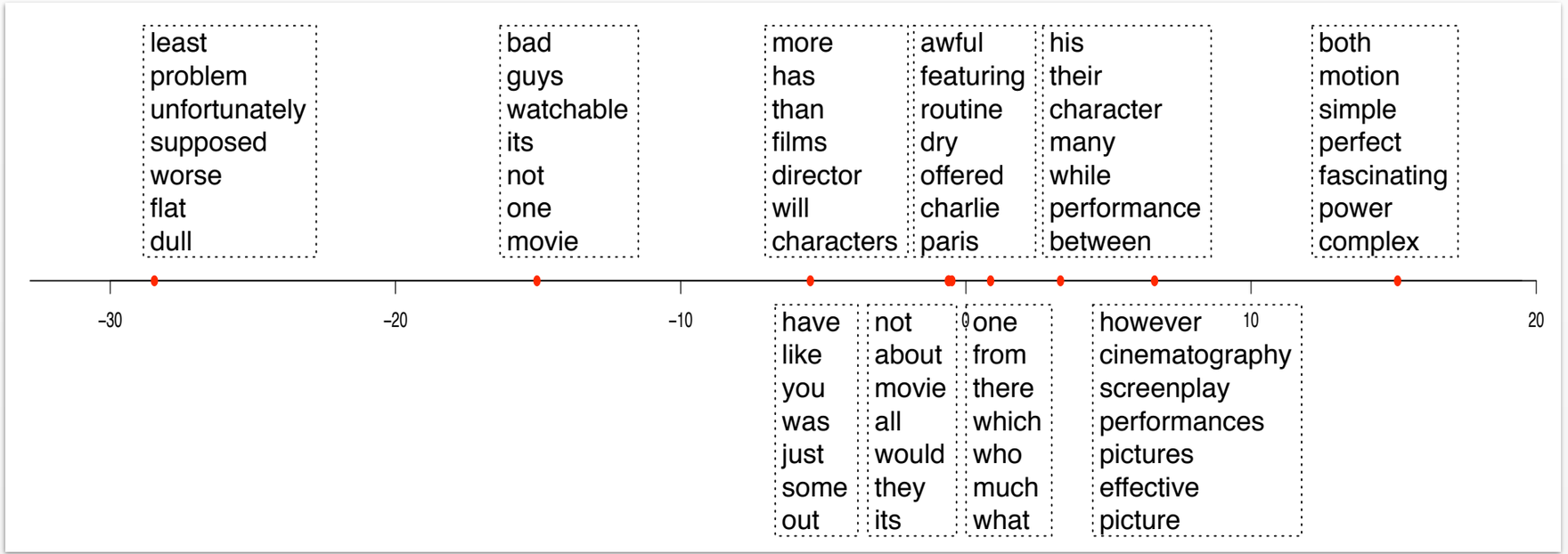
Extensions: Supervised LDA



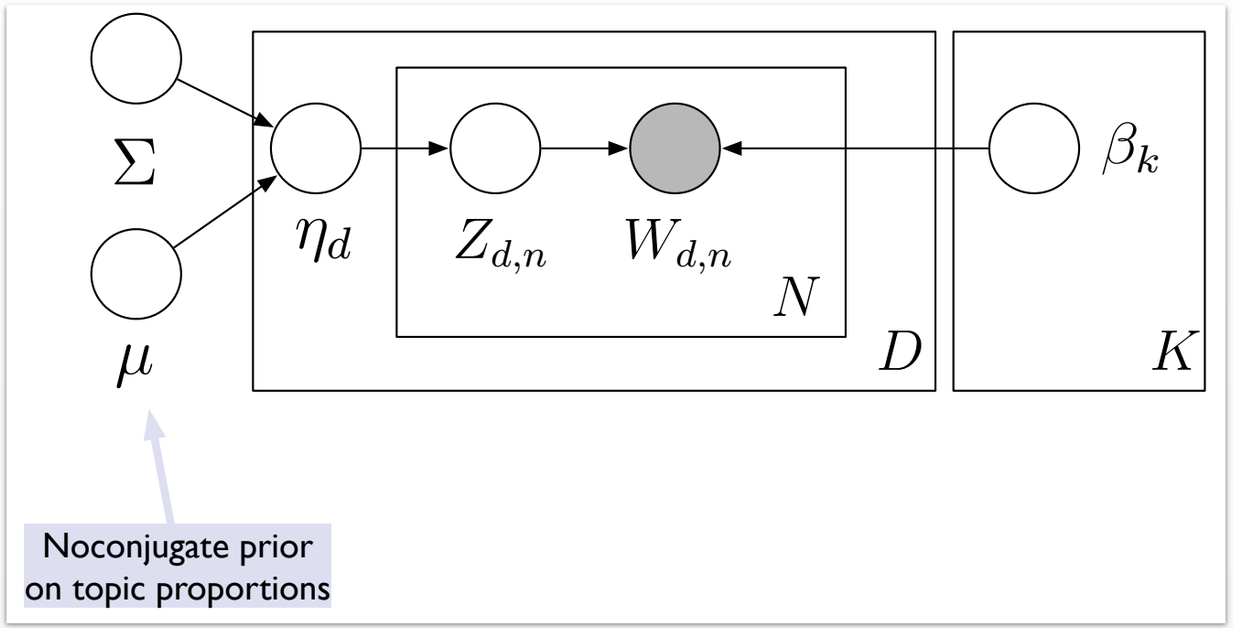
1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Extensions: Supervised LDA



Extensions: Correlated Topic Model



Estimate a covariance matrix Σ that parameterizes correlations between topics in a document

Extensions: Dynamic Topic Models

1789

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

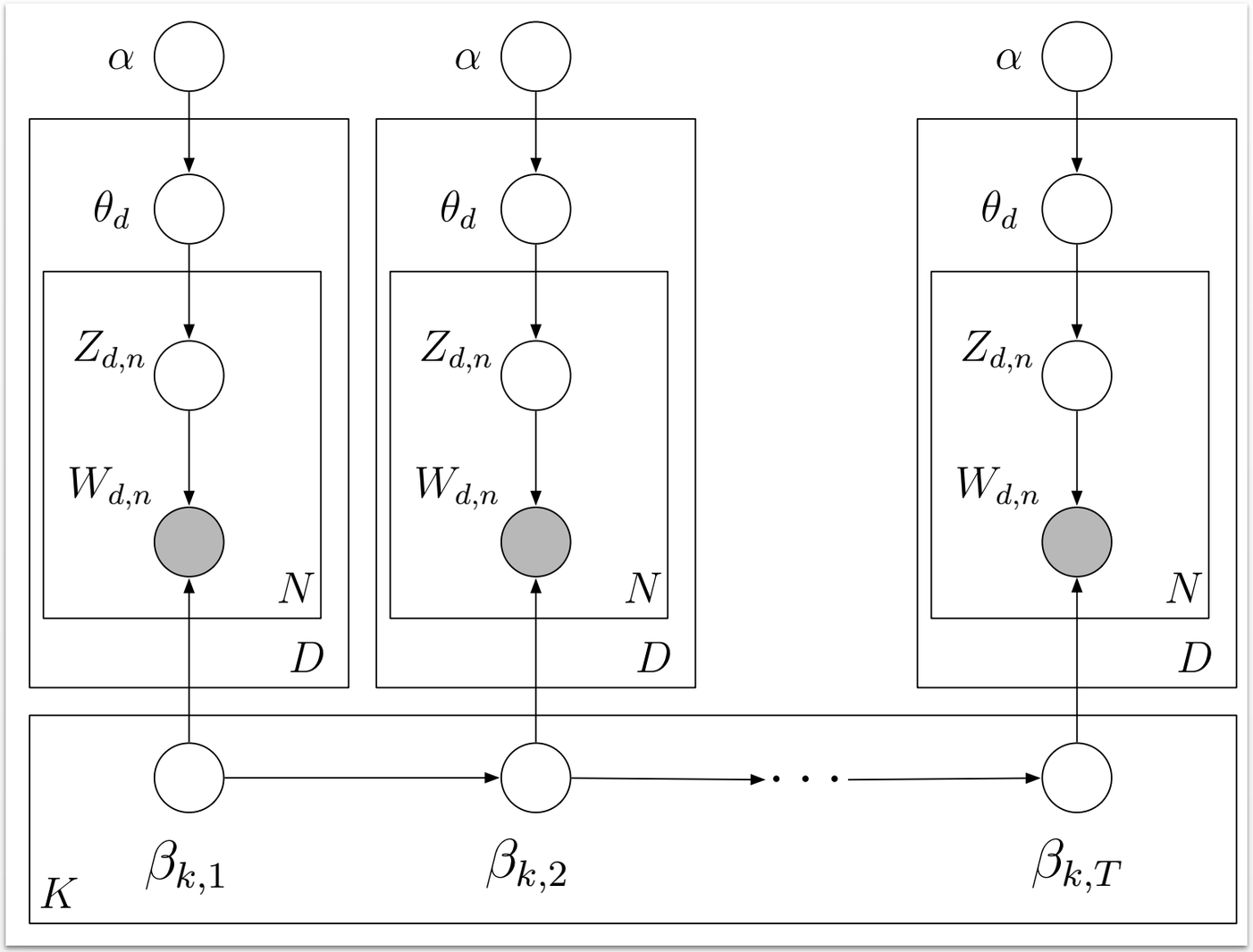
Inaugural addresses

2009

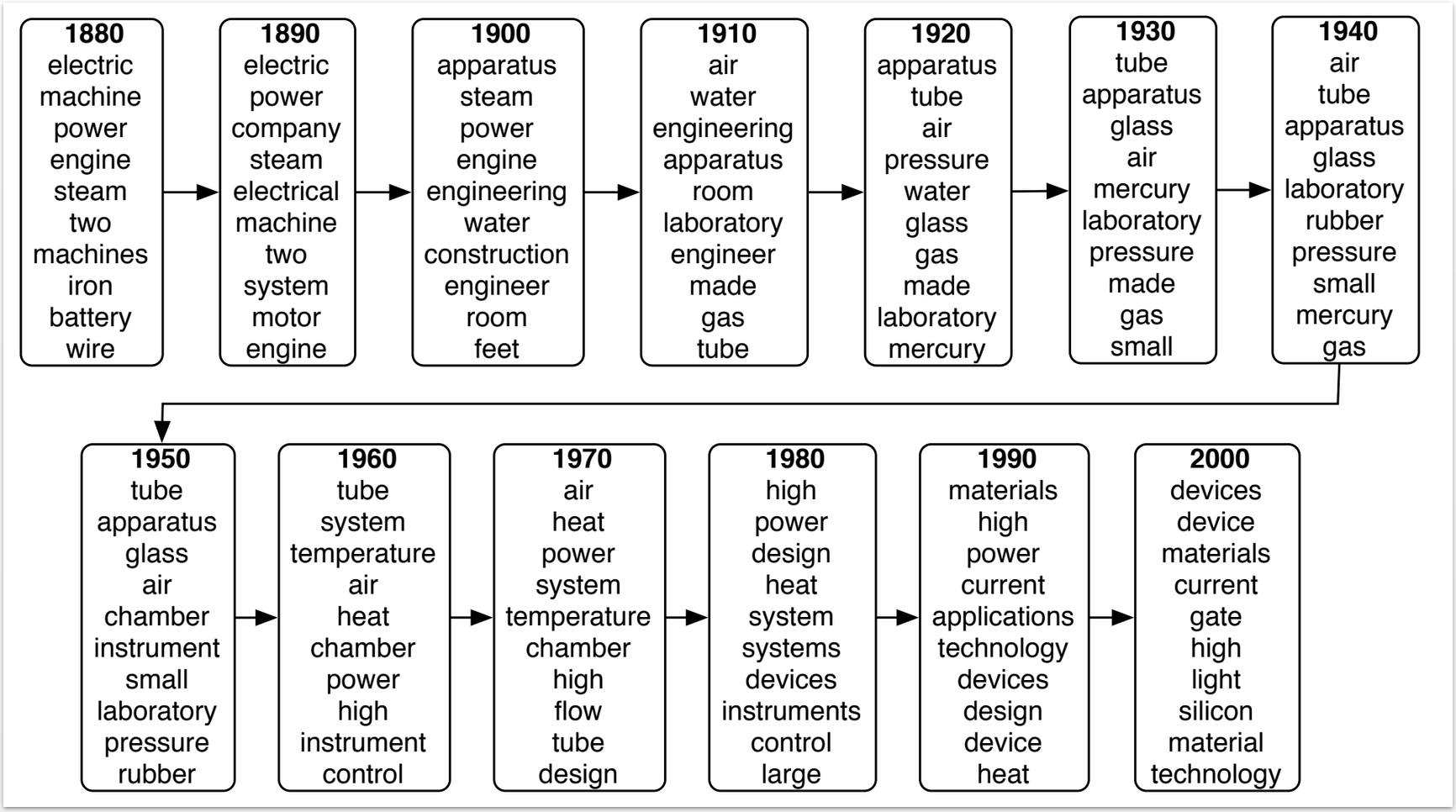
AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

Track changes in word distributions associated with a topic over time.

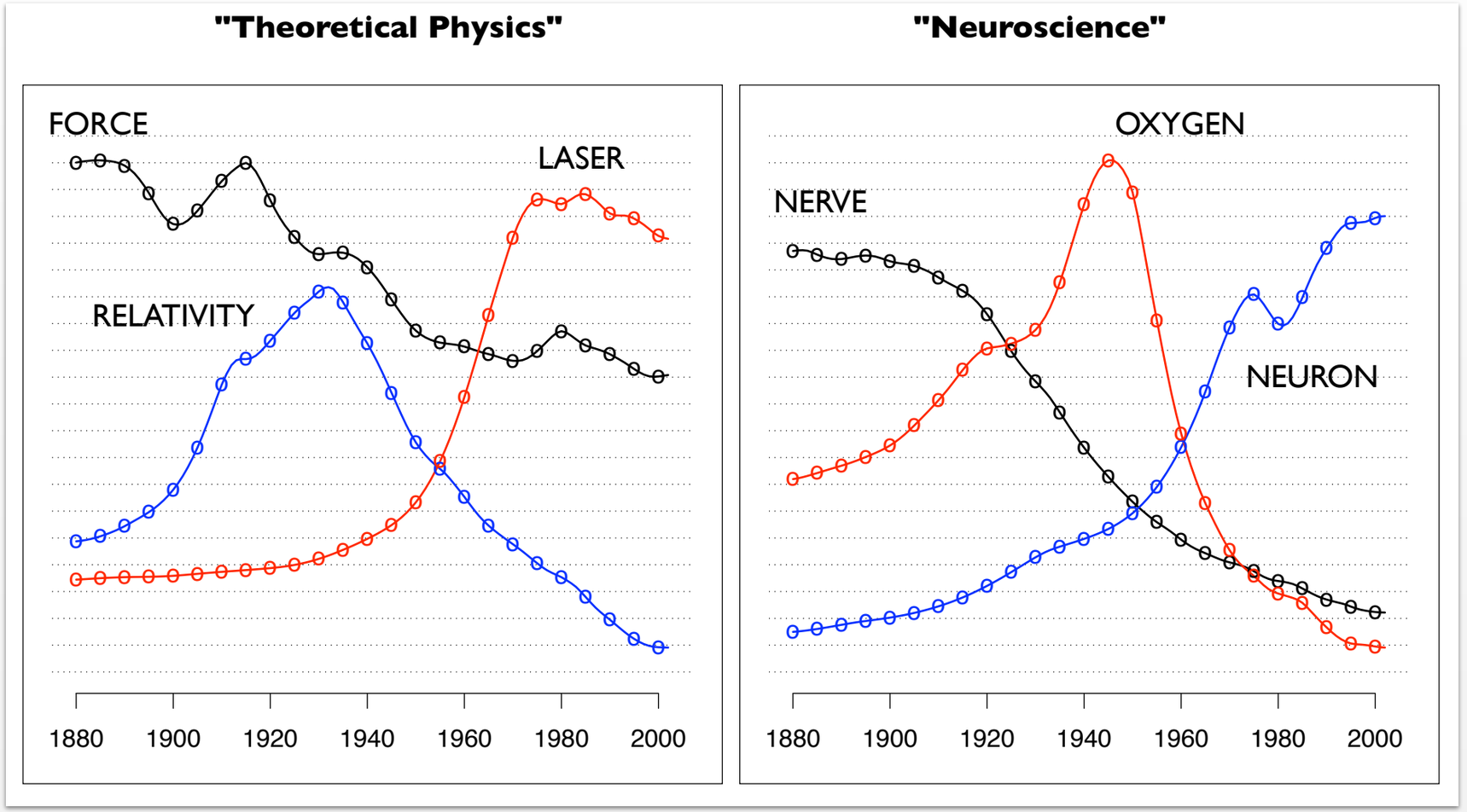
Extensions: Dynamic Topic Models



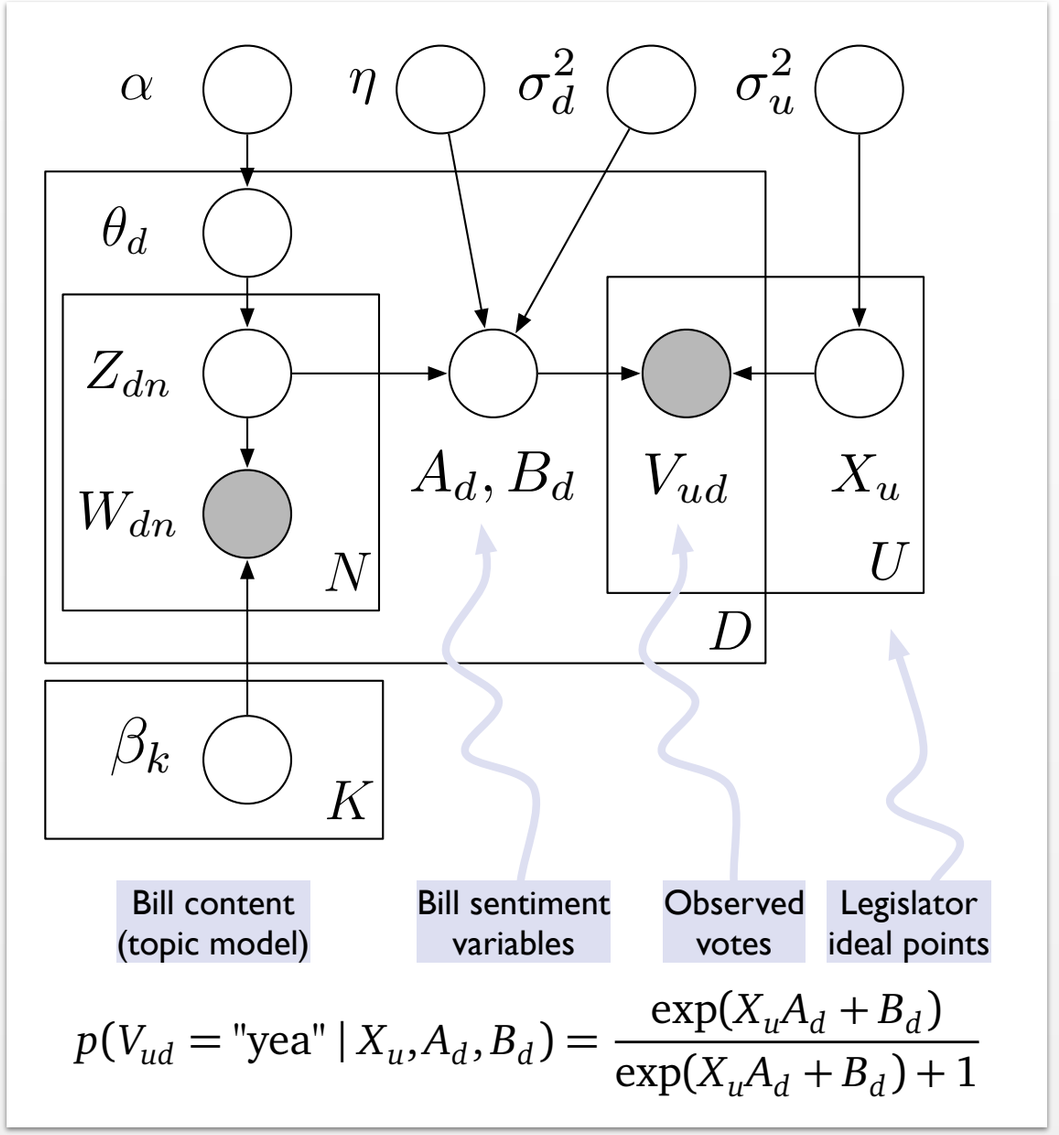
Extensions: Dynamic Topic Models



Extensions: Dynamic Topic Models

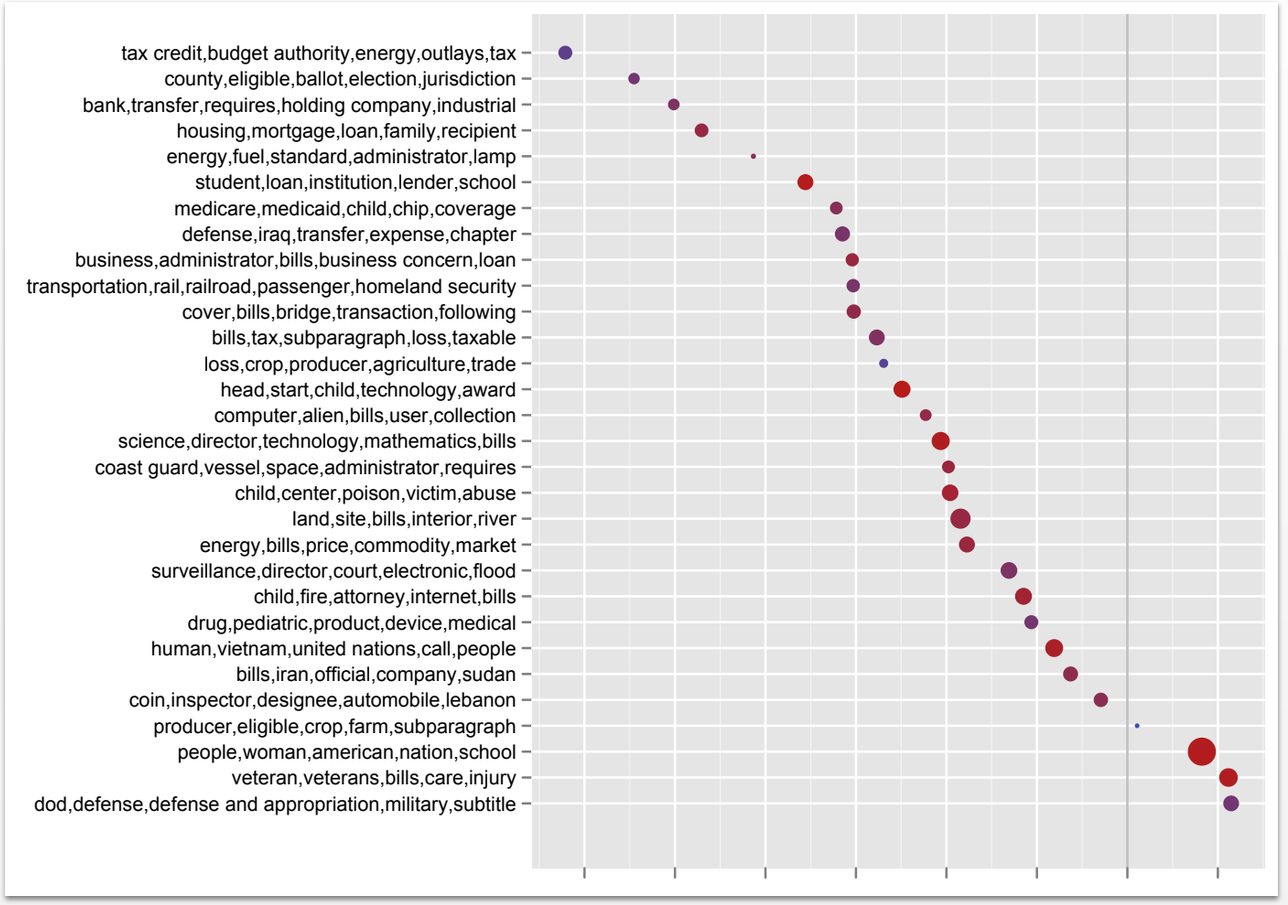


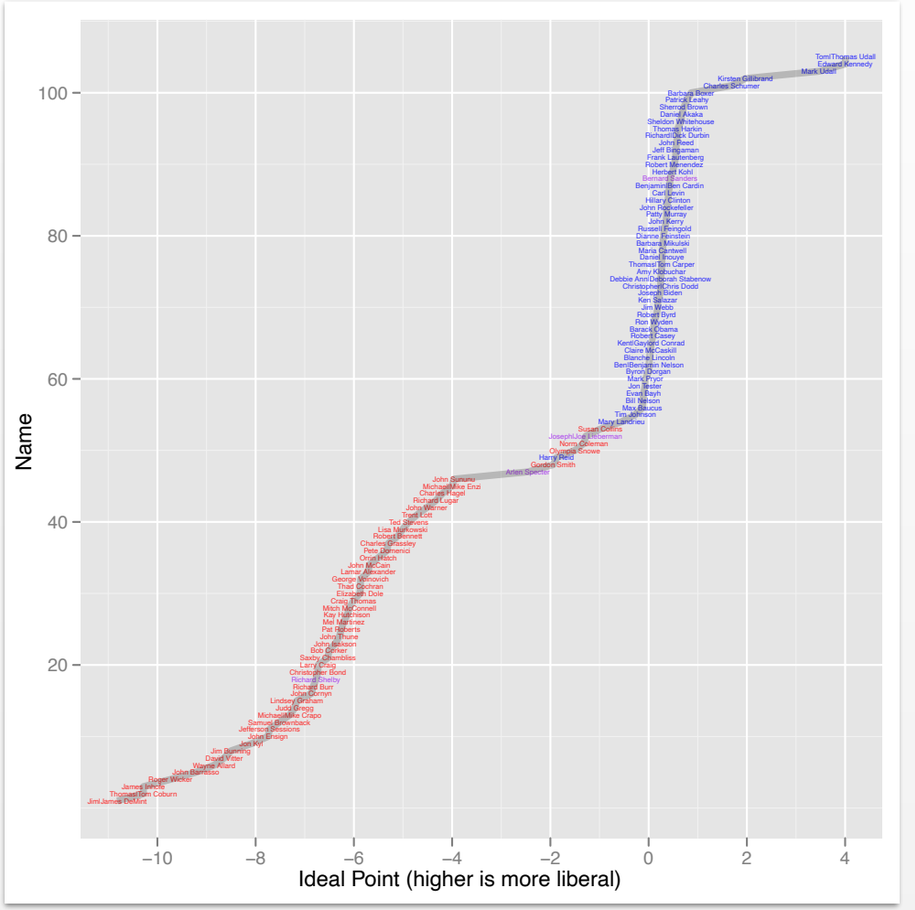
Extensions: Ideal Point Topic Models



$$p(V_{ud} = \text{"yea"} \mid X_u, A_d, B_d) = \frac{\exp(X_u A_d + B_d)}{\exp(X_u A_d + B_d) + 1}$$

Extensions: Ideal Point Topic Models





$$p(V_{ud} = \text{"yea"} \mid X_u, A_d, B_d) = \frac{\exp(X_u A_d + B_d)}{\exp(X_u A_d + B_d) + 1}$$

LDA: *Summary*

- **Idea:** Model documents as ***mixtures*** over topics
- **Model parameters (estimate with VBEM)**
  θ_d Topic probabilities for each document
  (K-dimensional vector for each document)
  β_k Word probabilities for each topic
  (V-dimensional vector for each topic)
- **Dirichlet Priors:** Enforce sparsity, associate a small number of topics which each document
- **Extensions:** Can design graphical models that build on LDA for a variety of modeling tasks