

Variational Inference

Michael Thomas Wojnowicz

November 11, 2020

Data Intensive Studies Center, Tufts University

Table of contents

1. Overview
2. Variational Inference and Expectation Maximization
3. Coordinate Ascent Variational Inference (CAVI)
4. Example: Bayesian Mixture Model
5. Tool: CAVI with Exponential Family Complete Conditionals
6. Example: Latent Dirichlet Allocation (LDA)
7. Summary

Some questions

- What is variational inference?
- When is it useful?
- Is it the same as variational bayes?
- Why is it called *variational* inference?
- What is Variational Expectation Maximization (VEM)? Variational Bayes Expectation Maximization (VBEM)?
- How can we apply VI to inference problems?

Overview

Overview

The problem: marginalization

Parametric statistical models

Parametric statistical models

A *parametric statistical model* posits

- x : observed data
- θ : parameters
- z (possibly): latent random variables

Parameters vs. latent variables

Both z and θ are unobserved, but only the dimensionality of z increases with the number of samples in x .

Frequentist vs. Bayesian variants

Frequentists take parameters θ to be fixed (but unknown) constants, whereas the Bayesians take θ to be random variables.

Three statistical modeling paradigms of interest

Let us consider models that present an **intractable marginal**.

Bayesian latent variable models

Examples: Bayesian Mixture Model, Bayesian Hidden Markov Model, Latent Dirichlet Allocation, Bayesian nonparametric versions of the preceding

Bayesian (non-latent variable) models

Examples: Non-conjugate models, Many hierarchical Bayesian models

Frequentist latent variable models

Examples: Hidden Markov Models (although we have handled this case), Variational Autoencoders (the classical kind), Bayesian Generalized Linear Mixed Effects Models

Statistical inference and marginalization

Bayesian latent-variable models

We want the posterior:

$$p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{x})}$$

Note that we need to compute the *evidence*, i.e. the marginal likelihood of the data:

$$p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}, \mathbf{z}) \, d\boldsymbol{\theta} \, d\mathbf{z} \quad (1.1)$$

Statistical inference and marginalization

Bayesian non-latent variable models

We want the posterior:

$$p(\theta | \mathbf{x}) = \frac{p(\mathbf{x} | \theta)p(\theta)}{p(\mathbf{x})}$$

Note that we need to compute the *evidence*, i.e. the marginal likelihood of the data:

$$p(\mathbf{x}) = \int p(\theta, \mathbf{x}) d\theta \tag{1.2}$$

Statistical inference and marginalization

Frequentist latent variable models

We want the maximum likelihood value:

$$\theta_{\text{ML}} := \operatorname{argmax}_{\theta} p(\mathbf{x} \mid \theta) = \operatorname{argmax}_{\theta} \int p(\mathbf{x}, \mathbf{z} \mid \theta) d \mathbf{z} \quad (1.3)$$

In particular, one requires access to the *marginal* likelihood

$$p(\mathbf{x} \mid \theta)_{\text{marginal likelihood}} = \int_{\text{joint (or "complete") likelihood}} p(\mathbf{x}, \mathbf{z} \mid \theta) d \mathbf{z} \quad (1.4)$$

Statistical inference

In general

We must compute the marginal

$$p(\mathbf{x} \mid \mathbf{c}) = \int p(\mathbf{x}, \mathbf{u} \mid \mathbf{c}) \, d\mathbf{u} \quad (1.5)$$

where

- \mathbf{x} : observed data
- \mathbf{u} : unobserved random variables
- \mathbf{c} : constant values

The need for marginalization in statistical inference

Model	Inferential goal	Target marginal
		$p(\mathbf{x} \mathbf{c})$
Bayesian (non-latent)	$p(\boldsymbol{\theta} \mathbf{x})$	$p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta}$
Bayesian latent	$p(\mathbf{z}, \boldsymbol{\theta} \mathbf{x})$	$p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}, \mathbf{z}) d\boldsymbol{\theta} dz$
Frequentist latent	$\text{argmax}_{\boldsymbol{\theta}} p(\mathbf{x} \boldsymbol{\theta})$	$p(\mathbf{x} \boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z} \boldsymbol{\theta}) dz$

Problem: These marginalizations may be intractable

Example: Hidden Markov Model

Define T : the state transition matrix

ϵ_j : the j th emission distribution, $j = 1, \dots, k$

π : the initial latent state distribution

$$\begin{aligned} p(\mathbf{x} | \theta) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z} | \theta) \\ &= \sum_{\mathbf{z}=(z_1, \dots, z_n)} p(\mathbf{x}, \mathbf{z} | \theta) \\ &= \sum_{\mathbf{z}=(z_1, \dots, z_n)} \pi_{z_1} \epsilon_{z_1}(x_1) T_{z_1, z_2} \epsilon_{z_2}(x_2) T_{z_2, z_3}, \dots, T_{z_{n-1}, z_n} \epsilon_{z_n}(x_n) \end{aligned}$$

has $\mathcal{O}(n k^n)$ complexity. 

Consider e.g. that $(k, n) = (5, 100) \rightarrow 10^{72}$ calculations.



Overview

The technique: functional optimization

Towards variational inference

We construct a lower bound on the target marginal.

Variational Lower Bound (VLBO)

Let q be any probability density over \mathbf{u} . Then:

$$\begin{aligned}\ln p(\mathbf{x} \mid \mathbf{c}) &= \ln \int p(\mathbf{u}, \mathbf{x} \mid \mathbf{c}) d\mathbf{u} \\ &= \ln \int q(\mathbf{u}) \frac{p(\mathbf{u}, \mathbf{x} \mid \mathbf{c})}{q(\mathbf{u})} d\mathbf{u} \\ &\stackrel{\text{Jensen's}}{\geq} \int q(\mathbf{u}) \ln \left(\frac{p(\mathbf{u}, \mathbf{x} \mid \mathbf{c})}{q(\mathbf{u})} \right) d\mathbf{u} \\ &:= \text{VLBO}(q)\end{aligned}$$

Variational Inference: Maximizing the VLBO

Variational Inference

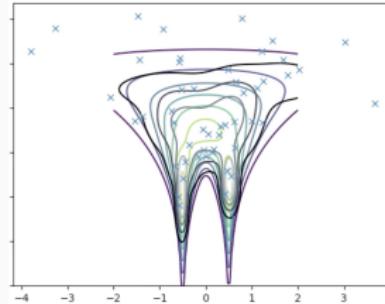
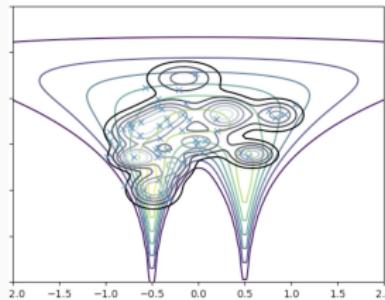
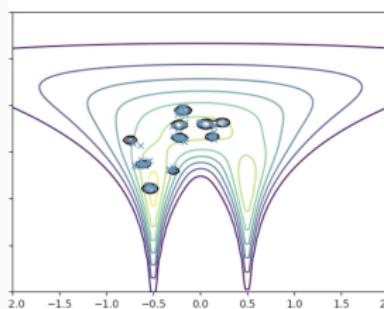
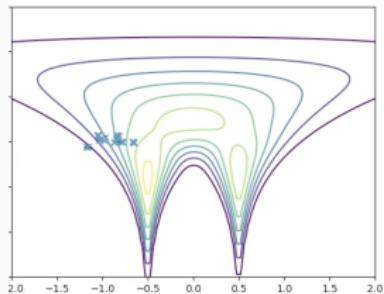
Variational inference (VI) proceeds by finding q^* , the variational density in tractable family \mathcal{Q} which maximizes the VLBO :

$$q^*_{\text{solution}} = \underset{\substack{q \in \mathcal{Q} \\ \text{approximating family}}}{\operatorname{argmax}} \text{VLBO}(q)$$

Rk: Note that we are trying to optimize over a function space (of a particular kind).

Illustration

Here we approximate an probability distribution by finding the best approximation from tractable family $\mathcal{Q} = \{10\text{-component Gaussian mixture models}\}$



Overview

Decompositions: Intuition on the cost function

Decompositions of the VLBO

Energy/Entropy Decomposition of the VLBO

By simply appealing to properties of the logarithm and the definition of expectation, we obtain

$$\begin{aligned} \text{VLBO}(q) &= \int q(\mathbf{u}) \ln p(\mathbf{x}, \mathbf{u} \mid \mathbf{c}) d\mathbf{u} - \int q(\mathbf{u}) \ln q(\mathbf{u}) d\mathbf{u} \\ &= \underset{\text{energy}}{\mathbb{E}_q [\log p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})]} + \underset{\text{entropy}}{\mathbb{H}[q(\mathbf{u})]} \end{aligned}$$

Q What is the effect of the entropy term?

Decompositions of the VLBO

Likelihood/Prior Decomposition of the VLBO

By applying the chain rule to the preceding, and then reapplying the definition of KL divergence, we obtain another nice form

$$\begin{aligned}\text{VLBO}(q) &= \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} | \mathbf{c})] + \mathbb{H}[q(\mathbf{u})] \\ &= \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} | \mathbf{c})] - \mathbb{E}_q[\log q(\mathbf{u})] \\ &= \mathbb{E}_q[\log p(\mathbf{x} | \mathbf{u}, \mathbf{c})] + \mathbb{E}_q[\log p(\mathbf{u} | \mathbf{c})] - \mathbb{E}_q[\log q(\mathbf{u})] \\ &= \mathbb{E}_q[\log p(\mathbf{x} | \mathbf{u}, \mathbf{c})] - \text{KL}(q(\mathbf{u}) || p(\mathbf{u} | \mathbf{c}))\end{aligned}$$

expected log likelihood divergence from prior

Note that the first term grows in magnitude as the number of samples increases; thus, the prior's influence diminishes asymptotically.

Overview

The posterior perspective

Maximizing the VLBO minimizes the KL divergence (to the posterior)

By definition, the KL divergence from the target posterior to the variational density is given by

$$\text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})} \right]$$

By the chain rule, we get

$$\begin{aligned} \text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) &= \underbrace{\mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})]}_{\text{energy/entropy decomposition}} + \log p(\mathbf{x} \mid \mathbf{c}) \\ &= -\text{VLBO}(q) + \text{constant} \end{aligned}$$

Discuss: What is the optimal variational density?

Maximizing the VLBO minimizes the KL divergence (to the posterior)

By definition, the KL divergence from the target posterior to the variational density is given by

$$\text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})} \right]$$

By the chain rule, we get

$$\begin{aligned} \text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) &= \underbrace{\mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})]}_{\text{energy/entropy decomposition}} + \log p(\mathbf{x} \mid \mathbf{c}) \\ &= -\text{VLBO}(q) + \text{constant} \end{aligned}$$

The optimal variational density

The optimal variational density, $q^*(\mathbf{u})$ is the target posterior density $p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})$ when the underlying variational family \mathcal{Q} is unrestricted

Overview

Summary

Summary

- VI is a general tool. It is useful whenever you face intractable marginals.

Model	Inferential goal	Intractable marginal	Variational density	Posterior
General case	infer about θ	$p(x c)$	$q(u)$	$p(u x, c)$
Frequentist latent	$\text{argmax}_{\theta} p(x \theta)$	$p(x \theta) = \int p(x, z \theta) dz$	$q(z)$	$p(z x, \theta)$
Bayesian (non-latent)	$p(\theta x)$	$p(x) = \int p(\theta, x) d\theta$	$q(\theta)$	$p(\theta x)$
Bayesian latent	$p(z, \theta x)$	$p(x) = \int p(\theta, x, z) d\theta dz$	$q(z, \theta)$	$p(z, \theta x)$

How does VI accommodate the goal of statistical inference?

Given selection of variational family \mathcal{Q} , the optimal variational density q^*

...

The marginal perspective

- *For frequentist models:* ... makes the VLBO best approximate the target marginal likelihood, $p(\mathbf{x} | \boldsymbol{\theta})$, which is what we wanted to maximize.
- *For Bayesian models:* ... raises the (approximate) evidence term $p(\mathbf{x})$ (the term used for Bayesian model comparison) as high as possible.

The posterior perspective

- *For Bayesian models:* ... is the family member which is closest to the target posterior $p(\mathbf{u} | \mathbf{x})$.
- *For frequentist models:* ... provides the best substitution $q^*(\mathbf{z}) \approx p(\mathbf{z} | \mathbf{x}, \boldsymbol{\theta}^{\text{curr}})$ into the E-step of the EM algorithm¹

¹See next section for more information.

Overview

Evaluation (in context)

Approximate Bayesian Inference

- The two most prominent strategies for approximating intractable posteriors are VI and Markov Chain Monte Carlo (MCMC).
- MCMC uses **sampling**. We construct a Markov chain over model parameters. The stationary distribution is the posterior. We approximate the posterior with samples.
- VI uses **approximation**. A tractable approximating family is chosen, and parameters are optimized to be close to the posterior.

Variational Inference vs MCMC

Variational Inference scales better to large datasets.

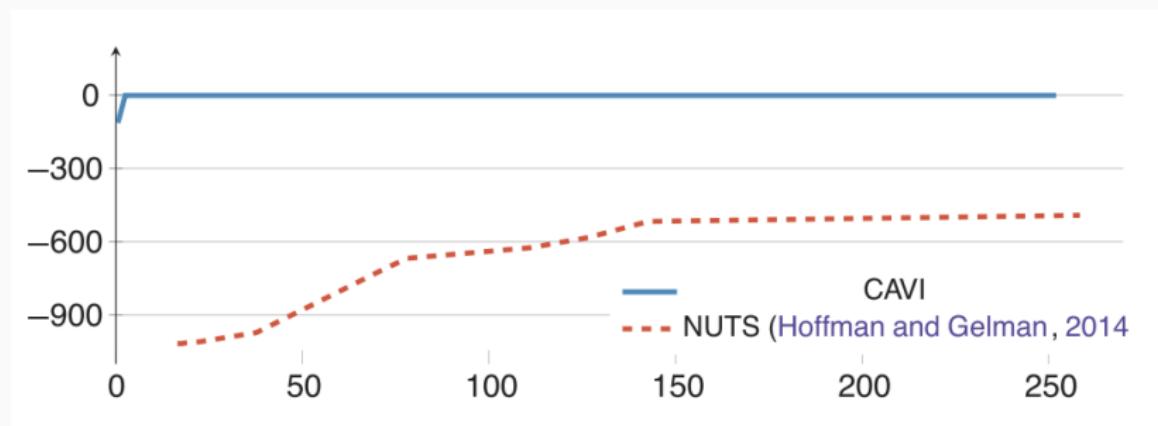


Figure 1: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. The plot shows log predictive test set accuracy by training time (minutes). CAVI fits a Gaussian mixture model to 10,000 images in less than a minute.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Variational Inference vs. Expectation Propagation

Let us fix a distribution P and consider two optimization strategies

Variational Inference

Minimizing
 $\text{KL}(Q||P)$
 $= \mathbb{E}_Q \left[\log \frac{q(x)}{p(x)} \right]$



Expectation Propagation

Minimizing
 $\text{KL}(P||Q)$
 $= \mathbb{E}_P \left[\log \frac{p(x)}{q(x)} \right]$

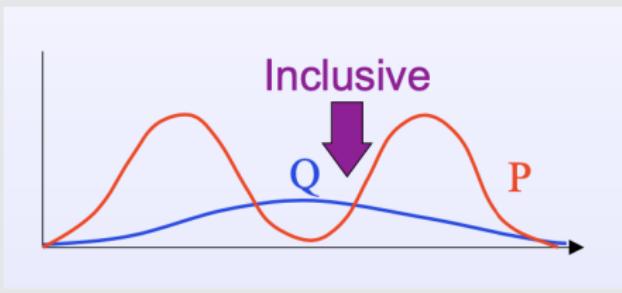
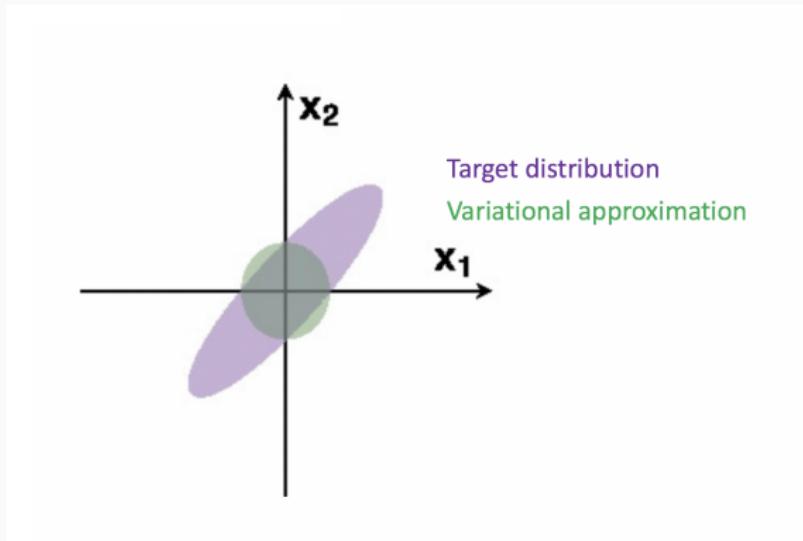


Image Credit: Tushar Tank

Q: What does this say about VI? Which one would you prefer to use?

Shortcoming: VI underestimates variance of the true posterior



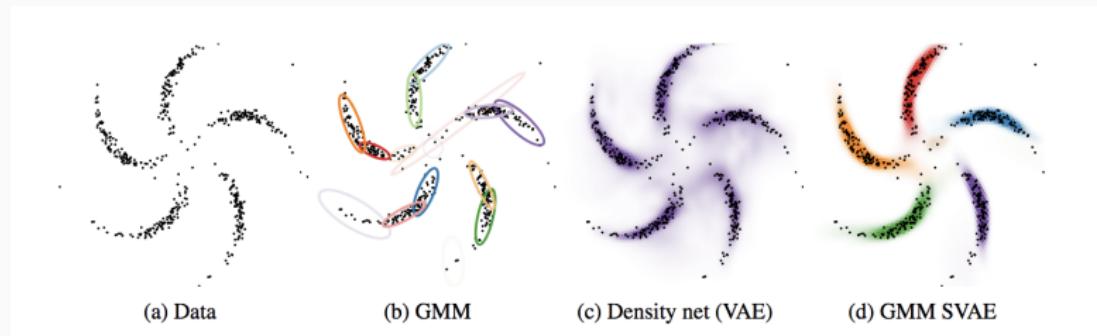
$$\text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})} \right]$$

Intuition

- If $q(\mathbf{u})$ is low, then we don't care (because of the expectation).
- If $q(\mathbf{u})$ is high and $p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})$ is low, then we pay a price

Modern application

We can compose probabilistic graphical models with neural networks to exploit their complementary strengths.



The resulting model is expressive, but also interpretable/decomposable.

Variational Inference and Expectation Maximization

Expectation Maximization (EM)

The EM algorithm refines an initial guess $\theta^{(0)}$ via the recursion

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{p(z | x, \theta^{(t)})} \left[\ln p(x, z | \theta) \right]$$

until convergence to a local optimum.

Example: Exponential Hidden Markov Model

E-step: Compute $p_i := p(z_i | x_i, \theta^{(t)})$ via the forward-backward algorithm.

M-step: Just a computation of weighted empirical reciprocal means:

$$\hat{\theta}_k^{(t)} = \frac{\sum_i (p_i = k)}{\sum_i (p_i = k) x_i}$$

Example in cybersecurity: Kantchelian, A., et al. (2015). *Better malware ground truth: Techniques for weighting anti-virus vendor labels*. In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (pp. 45-56). ACM.

Expectation Maximization (EM)

The EM algorithm refines an initial guess $\theta^{(0)}$ via the recursion

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{p(z|x, \theta^{(t)})} \left[\ln p(x, z | \theta) \right]$$

until convergence to a local optimum.

Example: Gaussian Hidden Markov Model

E-step: Compute $p_i := p(z_i | x_i, \theta^{(t)})$ via the forward-backward algorithm.

M-step: Just a computation of **weighted** empirical means and variances:

$$\hat{\mu}_k^{(t)} = \frac{\sum_i (p_i = k) x_i}{\sum_i (p_i = k)}, \quad \hat{\Sigma}_k^{(t)} = \frac{\sum_i (p_i = k) (x_i - \hat{\mu}^{(t)}) (x_i - \hat{\mu}^{(t)})^T}{\sum_i (p_i = k)}$$

Example in cybersecurity: Kantchelian, A., et al. (2015). *Better malware ground truth: Techniques for weighting anti-virus vendor labels*. In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (pp. 45-56). ACM.

EM from the perspective of VI

For a frequentist latent variable model, the VLBO is

$$\text{VLBO}(q_z, \theta) = \mathbb{E}_q [\log p(x, z | \theta)] + \mathbb{H}[q(z)]$$

Applying coordinate ascent (in the sense of variational calculus), we get the following update equations:

$$\mathbf{q \; update :} \quad q_z^{(t+1)} = \operatorname{argmax}_{q_z} \text{VLBO}(q_z; \theta^{(t)}) \quad (2.1)$$

$$\mathbf{\theta \; update :} \quad \theta^{(t+1)} = \operatorname{argmax}_{\theta} \text{VLBO}(q_z^{(t+1)}; \theta) \quad (2.2)$$

As argued earlier, we can solve the *q update* exactly by setting

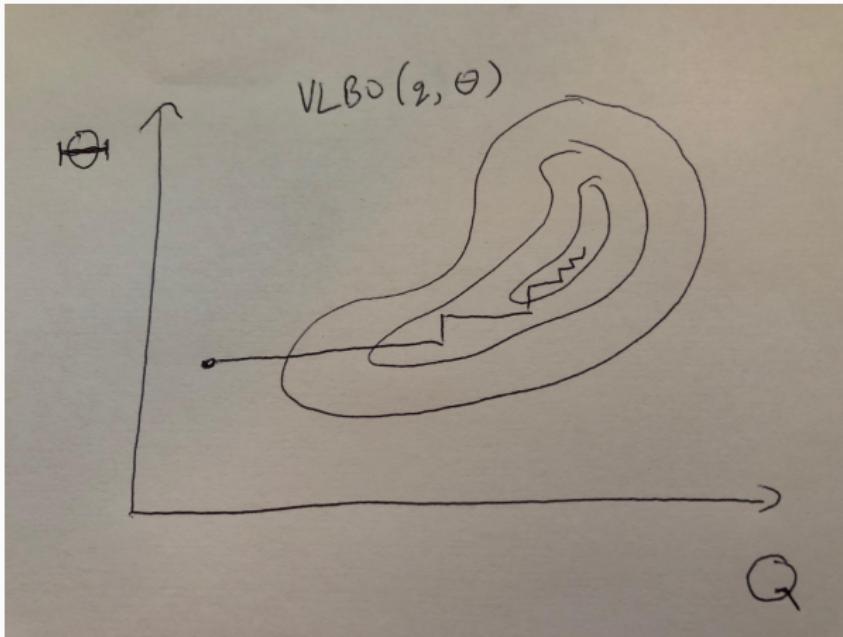
$$q_z^{(t+1)} = p(z | x; \theta^{(t)})$$

in which case the *θ update* becomes

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{p(z | x, \theta^{(t)})} \left[\ln p(x, z | \theta) \right] \quad (2.3)$$

which is precisely the EM algorithm.

EM as coordinate ascent on the VLBO



- If \mathcal{Q} unrestricted, we have EM
- What if we restrict \mathcal{Q} ?

Variational Expectation Maximization (VEM)

Consider a frequentist latent variable model. Since we don't always have access to $p(\mathbf{z} | \mathbf{x}, \boldsymbol{\theta})$, we may restrict our variational family \mathcal{Q} to some convenient form. In this case, coordinate ascent on the VLBO is given by:

$$\begin{aligned} q_{\mathbf{z}}^{(t+1)} &= \operatorname{argmax}_{q_{\mathbf{z}} \in \mathcal{Q}} \text{VLBO}(q_{\mathbf{z}}; \boldsymbol{\theta}^{(t)}) \\ \boldsymbol{\theta}^{(t+1)} &= \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q_{\mathbf{z}}^{(t+1)}} \left[\ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}) \right] \end{aligned}$$

which generalizes the EM algorithm.

Variational Bayes Expectation Maximization (VBEM)

Consider a **Bayesian latent variable model**. So we need to swap $p(x, z, \theta)$ for $p(x, z | \theta)$ in the VLBO.

If we construct the variational density with the factorization

$$q(z, \theta) = q_z(z)q_\theta(\theta)$$

then the VLBO becomes

$$\text{VLBO}(q_z(z), q_\theta(\theta)) := \int \int q_z(z)q_\theta(\theta) \ln \left(\frac{p(z, \theta, x)}{q_z(z)q_\theta(\theta)} \right) d\theta \ dz \quad (2.4)$$

We can perform coordinate ascent on the VLBO with respect to the densities q_z and q_θ :

$$\text{VB-E step : } q_z^{(t+1)} = \operatorname{argmax}_{q_z} \text{VLBO}(q_z; q_\theta^{(t)})$$

$$\text{VB-M step : } q_\theta^{(t+1)} = \operatorname{argmax}_{q_\theta} \text{VLBO}(q_z^{(t+1)}; q_\theta)$$

VBEM: Derivation

See notes.

VBEM: Update Equations

The coordinate ascent equations have the form

$$\textbf{VB-E step : } q_z^{(t+1)} \propto \exp \left(\mathbb{E}_{q_\theta^{(t)}} [\ln p(x, z | \theta)] \right) \quad (2.5)$$

$$\textbf{VB-M step : } q_\theta^{(t+1)} \propto p(\theta) \exp \left(\mathbb{E}_{q_z^{(t)}} [\ln p(x, z | \theta)] \right) \quad (2.6)$$

Prior-likelihood decomposition

Bayes' rule

$$p(\theta | x) \propto \underset{\textit{posterior}}{p(\theta)} \underset{\textit{prior}}{p(x | \theta)} \underset{\textit{likelihood}}{}$$

VB-M update

$$\underset{\textit{variational posterior}}{q_\theta^{(t+1)}} \propto \underset{\textit{prior}}{p(\theta)} \underset{\textit{expected likelihood under variational distribution}}{\exp \left(\mathbb{E}_{q_z^{(t)}} [\ln p(x, z | \theta)] \right)}$$

VI and EM: Summary

Variational inference can be considered as a generalization of the expectation maximization algorithm (which is generally used by frequentists). It

- relaxes the need for tractable computation of the posterior distribution $p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta})$.
- relaxes the assumption that $\boldsymbol{\theta}$ is a deterministic variable; variational calculus lets us do coordinate ascent on the *distribution* governing $\boldsymbol{\theta}$.

Coordinate Ascent Variational Inference (CAVI)

Coordinate Ascent Variational Inference (CAVI)

Coordinate ascent variational inference (CAVI) is a general approach to fitting models using VI.

This approach generalizes VBEM.

Mean Field Coordinate Ascent Variational Inference (MF-CAVI)

Mean field variational families

A variational family \mathcal{Q} is mean field if it factorizes

$$q(u_1, \dots, u_K) = \prod_{k=1}^K q_k(u_k) \quad (3.1)$$

Mean field coordinate ascent variational inference (MF-CAVI) is CAVI performed under the mean field assumption (3.1).

Update equations for MF-CAVI

To perform coordinate ascent on the VLBO under the mean field assumption (3.1), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp \left\{ \mathbb{E}_{q_{-k}} \left[\log p(u_k \mid \mathbf{u}_{-k}, \mathbf{x}, \mathbf{c}) \right] \right\} \quad (3.2)$$

The derivation uses variational calculus, and is nearly syntactically identical to the derivation of the VBEM updates.

Update equations for MF-CAVI

To perform coordinate ascent on the VLBO under the mean field assumption (3.1), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp \left\{ \mathbb{E}_{q_{-k}} \left[\log p(u_k \mid \mathbf{u}_{-k}, \mathbf{x}, \mathbf{c}) \right] \right\} \quad (3.2)$$

The derivation uses variational calculus, and is nearly syntactically identical to the derivation of the VBEM updates.

Rk: Note the connection to Gibbs sampling. In the MCMC literature, the distributions $p(u_k \mid \text{rest})$ are known as *full conditionals* or *complete conditionals*. Gibbs sampling involves successive draws from the full conditionals. In mean-field variational inference, we take expectations of the same distributions, in order to update our posterior approximations.

Example: Bayesian Mixture Model

Example: Bayesian Gaussian Mixture Model

To see the mean field CAVI algorithm (3.2) in a concrete context, consider a version of the Bayesian Gaussian Mixture Model.

$$\mu_k \sim \text{Normal}(M_k = 0, V_k = \sigma^2) \quad k = 1, \dots, K$$

$$c_i \sim \text{Categorical}(\pi_1, \dots, \pi_K) \quad i = 1, \dots, n$$

$$x_i \mid c_i, \mu \sim \text{Normal}(\mu_{c_i}, 1) \quad i = 1, \dots, n$$

(The model is simple in that it assumes univariate observations and that each mixture component has unit variance.)

The joint density, by chain rule, is

$$p(x, c, \mu) = p(\mu) \prod_{i=1}^n p(c_i) p(x_i \mid c_i, \mu)$$

And a mean-field variational family is given by

$$q(c, \mu) = \prod_{k=1}^K q(\mu_k) \prod_{i=1}^n q(c_i)$$

Example: Bayesian Gaussian Mixture Model

We apply (3.2) to determine the coordinate updates for q_{c_i} , the variational factors governing cluster assignments.

$$\begin{aligned} q(c_{ik}) &\propto \exp \left\{ \mathbb{E}_{q_{\mu_k}} \left[\log p(c_i = k) + \log p(x_i | c_i = k, \mu) \right] \right\} \\ &\propto \exp \left\{ \mathbb{E}_{q_{\mu_k}} \left[\log \pi_k + x_i \mu_k - \frac{1}{2} \mu_k^2 \right] \right\} \\ &\propto \pi_k \exp \left\{ x_i \mathbb{E}_{q_{\mu_k}} [\mu_k] - \frac{1}{2} \mathbb{E}_{q_{\mu_k}} [\mu_k^2] \right\} \end{aligned}$$

The coordinate updates for q_{μ_k} are derived similarly. They reveal that q_{μ_k} are Gaussian, and hence the above expectations are easy to compute.

Note: We abuse notation, and write $q(c_{ik})$ as shorthand for $q(c_i = k)$

Bayesian Gaussian Mixture Model: Updates to mixture component means

Using the same strategy as when updating cluster assignments c_i , we obtain

$$\begin{aligned} q(\mu_k) &\propto \exp \left\{ \mathbb{E}_{-q_{\mu_k}} \left[\log p(\mu_k) + \sum_{i=1}^n \log p(x_i \mid c_i = k, \mu) \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \mu_k^2 + \sum_{i=1}^n \mathbb{E}_{q_c} \left[\mathbf{1}_{c_i=k} \left(x_i \mu_k - \frac{1}{2} \mu_k^2 \right) \right] \right\} \\ &\propto \exp \left\{ \left(\sum_{i=1}^n q(c_{ik}) x_i \right) \mu_k + -\frac{1}{2} \left(\frac{1}{\sigma^2} + \sum_{i=1}^n q(c_{ik}) \right) \mu_k^2 \right\} \end{aligned}$$

which is an exponential family distribution with sufficient statistics (μ_k, μ_k^2) and base measure $\propto 1$; hence it is Gaussian.

Bayesian Gaussian Mixture Model: Updates to mixture component means

It is easy to show² that for a Gaussian with mean M and variance V , the natural parameters are given by

$$\eta_1 = \frac{M}{V}, \quad \eta_2 = -\frac{1}{2V}$$

From the last slide, the variational density $q(\mu_k)$ has natural parameters

$$\eta_1 = \left(\sum_{i=1}^n q(c_{ik}) x_i \right), \quad \eta_2 = -\frac{1}{2} \left(\frac{1}{\sigma^2} + \sum_{i=1}^n q(c_{ik}) \right)$$

Using this, we can backsolve to determine the updates to the mean and variance of the Gaussian variational density governing the k th cluster mean:

$$M_k = \frac{\sum_{i=1}^n q(c_{ik}) x_i}{1/\sigma^2 + \sum_{i=1}^n q(c_{ik})}, \quad V_k = \frac{1}{1/\sigma^2 + \sum_{i=1}^n q(c_{ik})}$$

²Indeed, in this course we have seen

Example: Bayesian Mixture Model

Why *Bayesian* mixture models?

VB Predictive vs. ML Solution

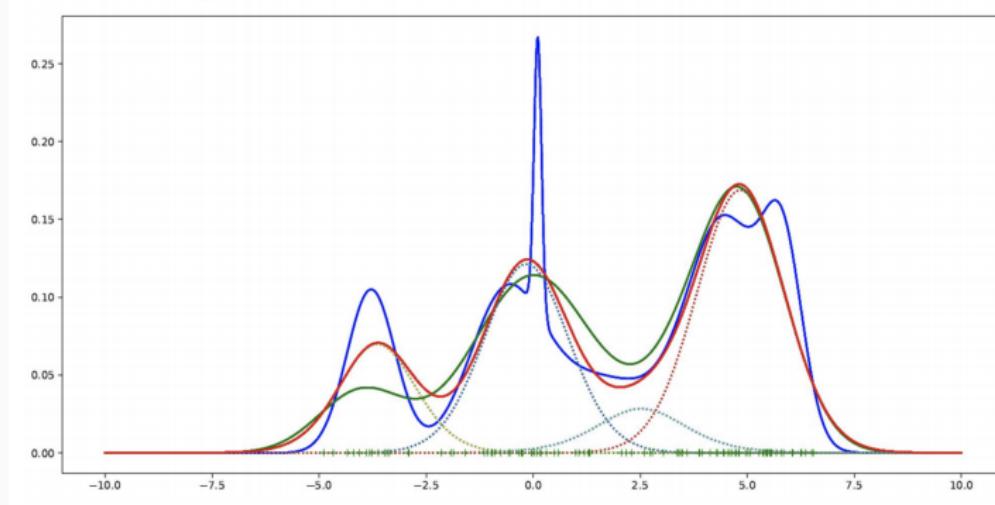


Image Credit: Lukas Burget

- VB was initialized from the ML solution
- VB recovers from ML overfitting and is closer to the true distribution for generating the training data

Extensions of Bayesian Mixture Models

The Bayesian framework can be used to endow mixture models with many nice properties.

Example: Dirichlet Process Mixture Models (DPMMs)

DPMM's have an unbounded number of mixture components.

The model automatically adapts its number of components.

- Click [here](#) for Demo 1.
- Click [here](#) for Demo 2.

Tool: CAVI with Exponential Family Complete Conditionals

Motivation

The form of (3.2) suggests that MF-CAVI can be simplified when the complete conditional has known form. We consider the case where complete conditionals are in the exponential family

This situation describes a lot of models:

Models with exponential family complete conditionals

- Bayesian mixture models (where the mixture components are exponential families and conjugate priors are used)
- Bayesian Hidden Markov Models (HMMs)
- Hierarchical HMMs
- Switching Kalman Filters
- Certain hierarchical regression models (Linear regression, Poisson regression, probit regression)
- Matrix factorization models
- [...]

The Exponential Family

We define an *exponential family* of probability distributions as those distributions whose density has the following form

$$p(x | \eta) = h(x) \exp\{\eta^T s(x) - a(\eta)\} \quad (5.1)$$

where we refer to h as the base measure, η as the natural parameter, s as the sufficient statistics, and a as the log normalizer.

Note: $x \perp\!\!\!\perp \theta | t(x)$

MF-CAVI updates on random variables with exponential family complete conditionals

Claim

Consider a model with unobserved random variables (u_1, \dots, u_k) . Let

1. Variational density q have mean field factorization $q = q_k(u_k)q_{-k}(u_{-k})$.
2. $p(u_k | u_{-k}, x)$ be in exponential family \mathcal{E} with natural parameter $\eta_k(u_{-k}, x)$.

Then optimal mean field CAVI update (3.2) puts

$$q_k \in \mathcal{E} \tag{5.2}$$

with natural parameter

$$\nu_k = \mathbb{E}_{q_{-k}}[\eta_k(u_{-k}, x)] \tag{5.3}$$

Take Home

- Variational factor is in same family as complete conditional.
- Its natural parameter is the expectation (with respect to the other variational factors) of the natural parameter of the complete conditional.

Proof

If the k th complete conditional is in the exponential family, then we have

$$p(u_k \mid u_{-k}, x) = h(u_k) \exp\{\eta_k^T s(u_k) - a(\eta_k)\} \quad (5.4)$$

Note that our notation suppresses that the natural parameter η_k depends on the conditioning variables (u_{-k}, x) .

By substituting (5.4) into (3.2) and discarding factors that do not depend on u_k , we obtain

$$\begin{aligned} q_k(u_k) &\propto \exp \left\{ \mathbb{E}_{q_{-k}} \left[\eta_k^T s(u_k) + \log h(u_k) \right] \right\} \\ &\propto h(u_k) \exp \left\{ \mathbb{E}_{q_{-k}} [\eta_k]^T s(u_k) \right\} \end{aligned}$$

Since h and s are identical to those of (5.4), the claim holds.

Example: Latent Dirichlet Allocation (LDA)

Acknowledgements

This section, especially the intro, borrows heavily from David Blei's 2012 ICML tutorial.

Overview

LDA is a generative probabilistic model of a corpus of documents of text.

LDA assumes:

- There is a set of topics that describe the corpus
- Each document exhibits these topics to varying degrees (each word in a document was generated by one of these topics.).

So:

- The topics and how they relate to the documents are hidden structure
- The main computational problem is to infer this hidden structure

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

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.

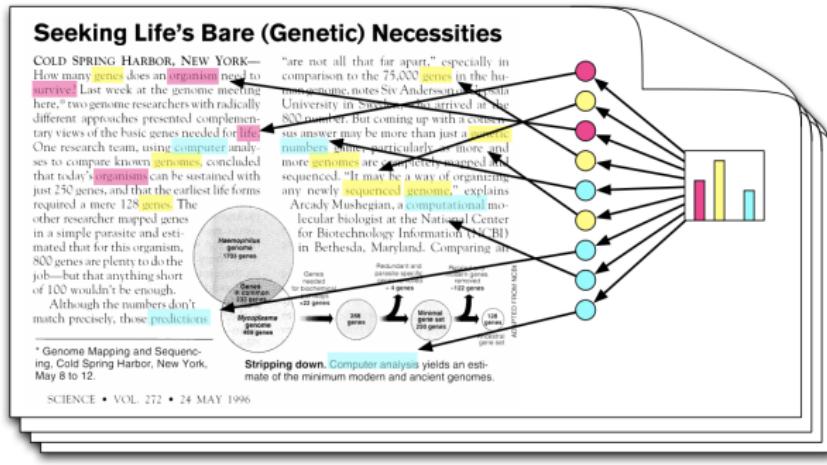
Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Stéphane Anderson of the University in Sweden who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game; particularly as more and more genomes are being collected and sequenced. "It may be a way of organizing and easily sequencing genomes," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



- Each **topic** is a distribution over words.

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

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

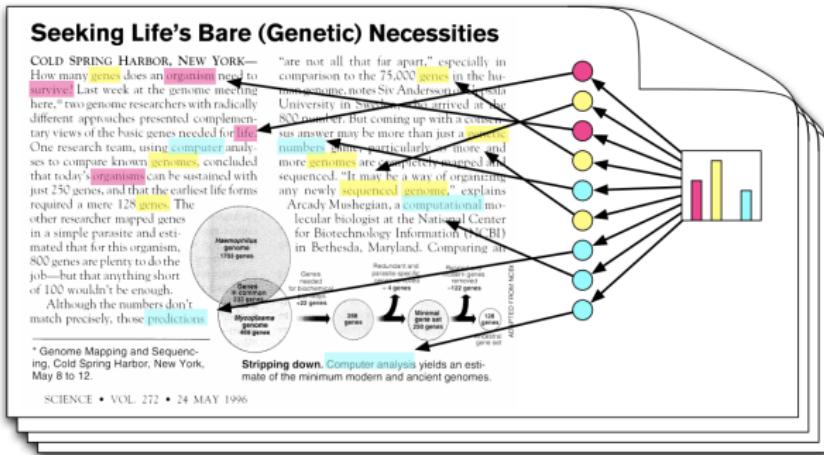
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Stéphane Anderson of the University in Sweden who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game; particularly as more and more genomes are being collected and sequenced. "It may be a way of organizing and easily sequencing genomes," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

the genomes, he says, "we can find the minimum set of genes that are common to all." The diagram illustrates the process of stripping down a genome to its bare necessities.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics.

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

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

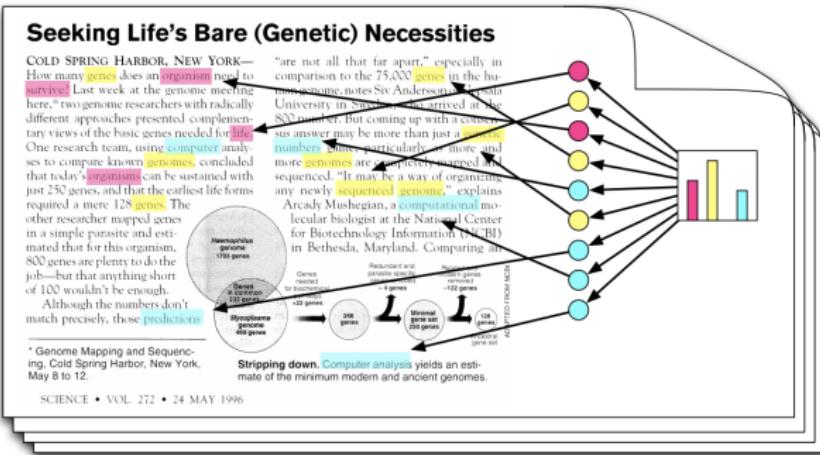
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Stéphane Andréasson, a geneticist at the University in Sweden who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game; particularly, as more and more genomes are collected, compared and sequenced. "It may be a way of organizing and easily sequencing genomes," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

the genomes, he says, will help researchers identify the minimum set of genes needed for life.

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

SCIENCE • VOL. 272 • 24 MAY 1996

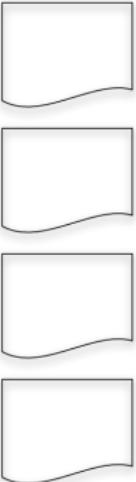
Topic proportions and assignments



- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics.
- Each **word** is drawn from one of those topics.

Source: David Blei, 2012 ICML Tutorial

Topics



Documents

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,¹ 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

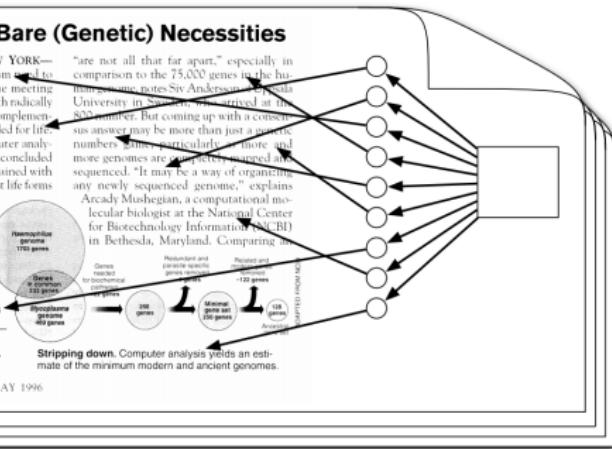
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Umeå University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a matter of numbers. Some predictability: more and more genomes are being pieced together and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Aradhye Mudigian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

¹ Genome Mapping and Sequencing. Cold Spring Harbor, New York. May 8 to 12.

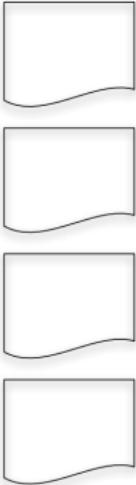
SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

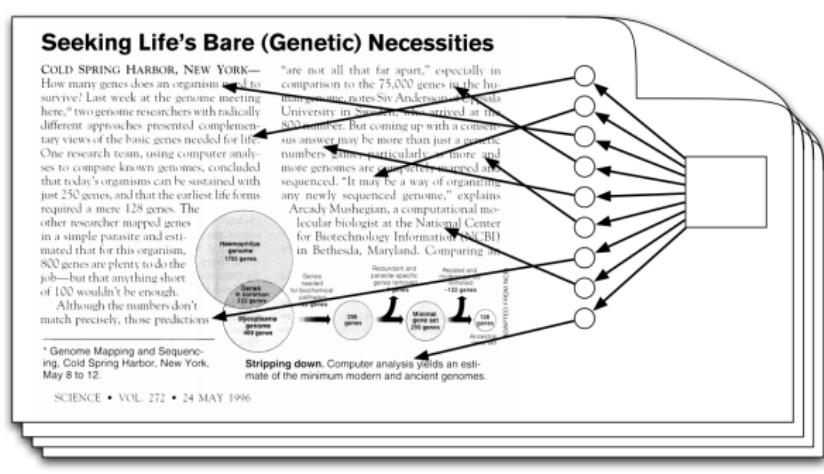


- In reality, we only observe the documents.
- The other structure is **hidden variables**.

Topics



Documents



- Our goal is to **infer** the hidden variables.
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

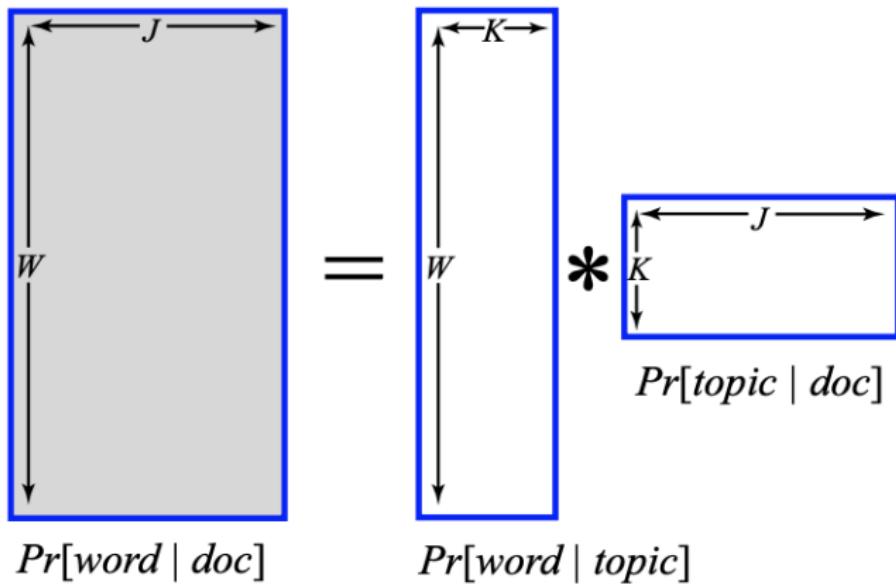
LDA: Generative Process

- Fix a vocabulary of V words, and set the number of topics, K .
- Set hyperparameters $\eta \in \mathbb{R}^V, \alpha \in \mathbb{R}^K$.
- For k in $(1, K)$:
 - Define the *topic* (a distribution over words), $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$.³
- For d in $(1, D)$:
 - Choose (per-document) *topic proportions* $\theta_d \in \mathbb{R}^K \sim \text{Dir}(\alpha)$.⁴
 - For n in $(1, N_d)$:
 - Choose the *topic assignment* $z_{d,n} \sim \text{Categorical}_K(\theta_d)$
 - Choose *word* $w_{d,n} \sim \text{Categorical}_V(\beta_{z_{d,n}})$

³Interpretation of η : psuedocount of vocabulary words observed across prior topics.

⁴Interpretation of α : psuedocount of topics observed across prior documents.

LDA as Probabilistic Matrix Factorization



LDA can be seen as a probabilistically constrained factorization of the matrix describing the bag of words composing each group, or document.

The number K of latent topics determines the factorization's rank.

Aside: Dirichlet Distribution

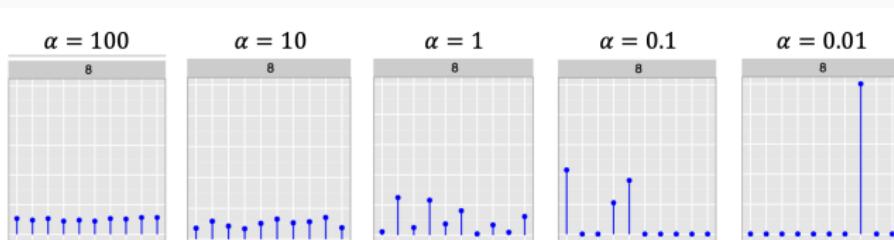
- Dirichlet distribution is *conjugate prior* of Multinomial

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$

- The parameter α controls the shape and sparsity of the θ_d 's.

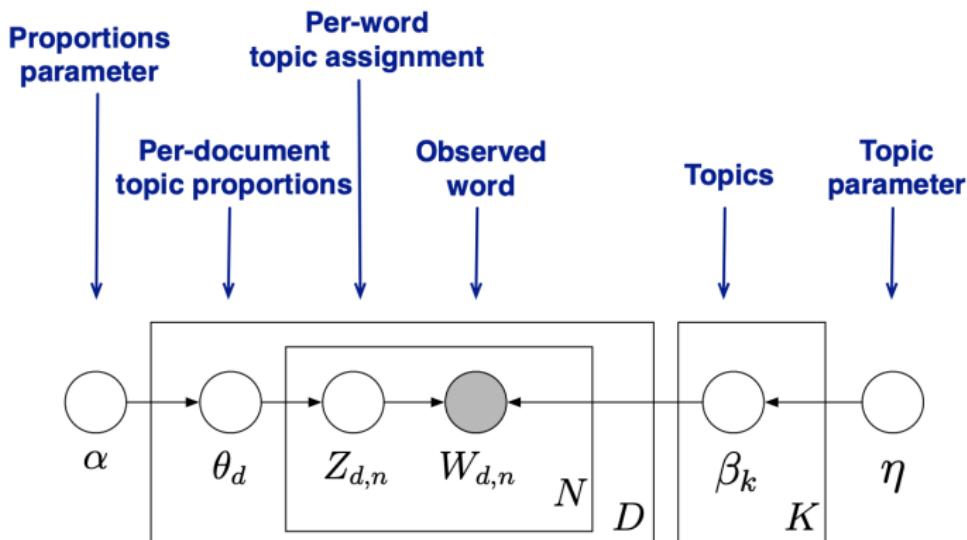
(per-document topic proportions)

- high α : typical θ_d (from the prior) will be uniform
- small α : a typical θ_d (from the prior) will be sparse



- Likewise, η controls the shape and sparsity of the β_k 's (the topics – distribution over words)

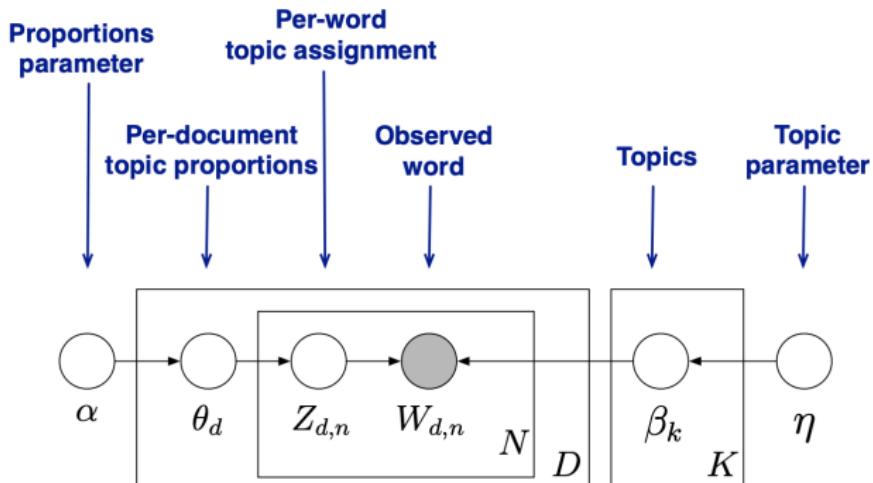
LDA as a graphical model



Recall:

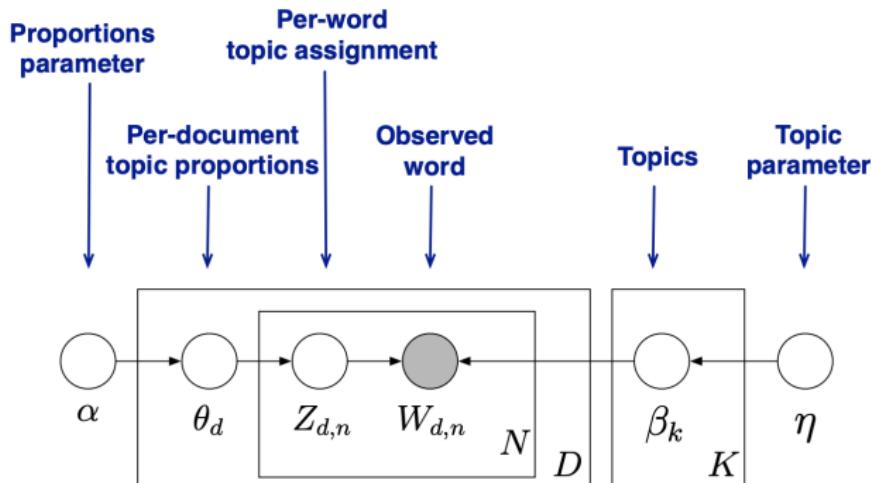
- Nodes are random variables.
- Shaded nodes are observed.
- Plates indicate replicated variables.
- Each node is conditionally independent from its non-descendants given its parents.

Joint distribution



$$p(z, \theta, w, \beta | \alpha, \eta) = \prod_{k=1}^K p(\beta_k | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta_k) \quad (6.1)$$

Joint distribution



$$p(z, \theta, w, \beta | \alpha, \eta) = \prod_{k=1}^K p(\beta_k | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta_k) \quad (6.1)$$

Remark. Note that LDA is a "bag-of-words" model; i.e. the probability of a word (or document) is invariant to word order.

Variational distribution

We approximate the posterior $p(\theta | z, w)$ using mean field variational inference (3.1). In particular, we assume that the variational family \mathcal{Q} has a density which factorizes as

$$\begin{aligned} q &= q_\delta(\theta) q_\tau(z) \\ &= \prod_{d=1}^D \underbrace{q_\delta(\theta_d)}_{\text{Dirichlet}} \prod_{n=1}^{N_d} \underbrace{q_{\tau_n}(z_{d,n})}_{\text{Categorical}} \end{aligned} \tag{6.2}$$

Note We are treating the topics β_k as a constant for simplicity. For a fuller treatment, see Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Update equations

LDA coordinate ascent update equations

$$\delta_{d,k} = \frac{\alpha_k}{\text{prior counts}} + \sum_{n=1}^{N_d} \frac{\tau_{d,n,k}}{\text{var. expected assignment}} \quad (6.3)$$

$$\begin{aligned} \tau_{d,n,k} &\propto \exp \left\{ \mathbb{E}_{q_\delta(\theta)} \left[\log \theta_{d,k} \right] \right\} \frac{\beta_{k,[w_{d,n}]}}{\text{likelihood}} \\ &= \left(\Psi(\delta_k) - \Psi \left(\sum_j \delta_j \right) \right) \beta_{k,[w_{d,n}]} \end{aligned} \quad (6.4)$$

where $\Psi(\cdot)$ is the first derivative of the log Γ function.

- Derivable via VBEM (see notes).
- Characteristic form: latent variable update depends on the data, global parameter update depends on the latent variable

Note We are treating β as a constant for brevity. We could fit also β to the data via VEM. (VEM does "empirical Bayes" for you.) More generally, we could treat β with VBEM. For a fuller treatment, see Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

The role of analytical computations

The variational categorical update crucially hinges on facts about the exponential family.

In particular, the meat of the proof of the variational categorical update depends crucially on the fact that the Dirichlet of a single probability component is given by

$$\mathbb{E}_{q_\delta(\theta)} \left[\log \theta_i \right] = \Psi(\delta_i) - \Psi\left(\sum_k \delta_k\right) \quad (6.5)$$

where $\Psi(\cdot)$ is the first derivative of the $\log \Gamma$ function.

This fact is justified via facts about the exponential family (such as that the derivative of the log normalization factor with respect to the natural parameter is equal to the sufficient statistic).

LDA Example Inference

- Data: 17K Science documents from 1990-2000 (11M words, 20K unique terms)
- Model: 100-topic LDA model, fit using variational inference

1 dna gene sequence genes sequences human genome genetic analysis two	2 protein cell cells proteins receptor fig binding activity activation kinase	3 water climate atmospheric temperature global surface ocean carbon atmosphere changes	4 says researchers new university just science like work first years	5 mantle high earth pressure seismic crust temperature earths lower earthquakes
6 end article start science readers service news card circle letters	7 time data two model to system number different with on	8 materials surface high structure temperature molecules chemical molecular to university	9 dna rna transcription protein site binding sequence proteins specific sequences	10 disease cancer patients human gene medical studies drug normal drugs
11 years million ago age university north early fig evidence record	12 species evolution population evolutionary university populations natural studies genetic today	13 protein structure proteins two amino binding acid residues molecular structural	14 cells cell virus hiv infection immune human antigen infected viral	15 space solar observations earth stars university mass sun astronomers telescope
16 fax manager science aaas advertising sales member recruitment associate washington	17 cells cell gene genes expression development mutant mice fig biology	18 energy electron state light quantum physics electrons high laser magnetic	19 research science national scientific scientists new states university united health	20 neurons brain cells activity fig channels university cortex neuronal visual

Application: Anomaly Detection in Network Traffic Traces

LDA can be applied to *documents* that can be just about anything!

IP Addresses

An **ip address**, like 72.194.113.177, is (roughly) an address assigned to each device connected to the Internet. My laptop has one, my iphone has one, every website (Google, Apple, etc.) has one, etc.

One application

(Newton, B. D. (2012). Anomaly Detection in Network Traffic Traces Using Latent Dirichlet Allocation.)

- “Documents” = the session of a specific IP address
- “words” = the full external IP address and port number combinations.
- The “words” in each “document” are counted and then this data set is processed by LDA to yield a compact model of the data.

Q: Ok, but how to perform anomaly detection? Q: What assumptions are being made by LDA?

I analyzed each of the anomalies detected in the last half hour of the trace [...]. The second anomaly with nearly 300 thousand messages exchanged with an SMTP server, is a bit more troubling. It is possible that this was actually a malicious client participating in a Mailbomb attack. According to the DARPA Intrusion Detection Attacks Database [8] a Mailbomb attack “is one in which the attacker sends many messages to a server, overflowing that server’s mail queue and possibly causing a system failure.”

Questions?