Statistical Models & Computing Methods

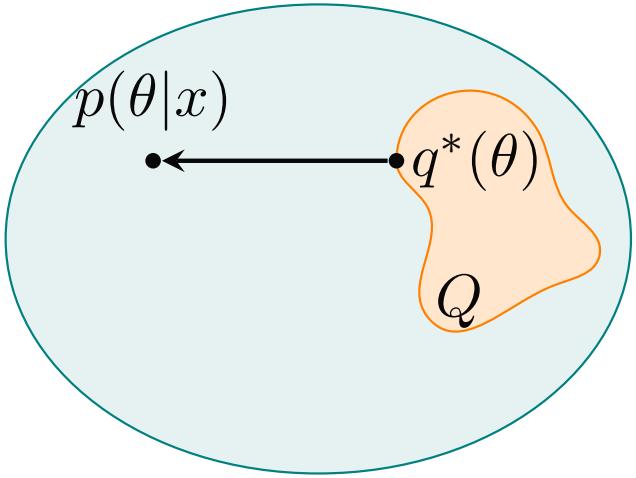
Lecture 13: Variational Inference

Cheng Zhang

School of Mathematical Sciences, Peking University

November 13, 2024

- A Bayesian probabilistic model includes the conditional distribution $p(x|\theta)$ of observed variable x given the model parameter θ, and the prior $p(\theta)$, which gives a joint distribution
$$p(x, \theta) = p(x|\theta)p(\theta)$$
- Inference about the parameter θ is through the **posterior**, the conditional distribution of the parameters given the observations
$$p(\theta|x) = \frac{p(x, \theta)}{p(x)}$$
- For most interesting models, the denominator is not tractable. We appeal to approximate posterior inference.
  - **Markov chain Monte Carlo** – We've introduced
  - **Variational inference** – The topic for this lecture!



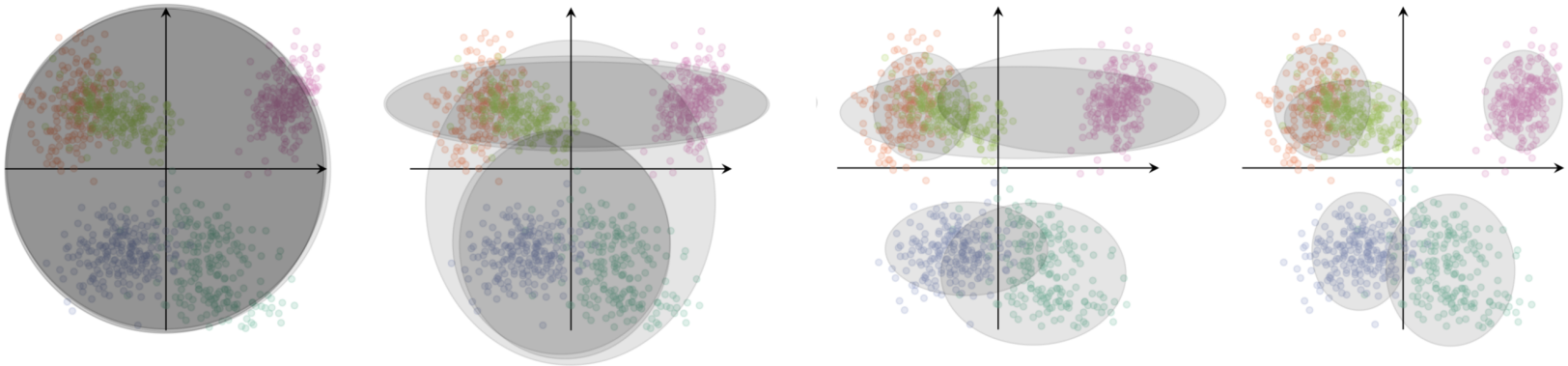
$$q^*(\theta) = \arg\min_{q \in Q} \mathrm{KL}\left(q(\theta) \| p(\theta|x)\right)$$

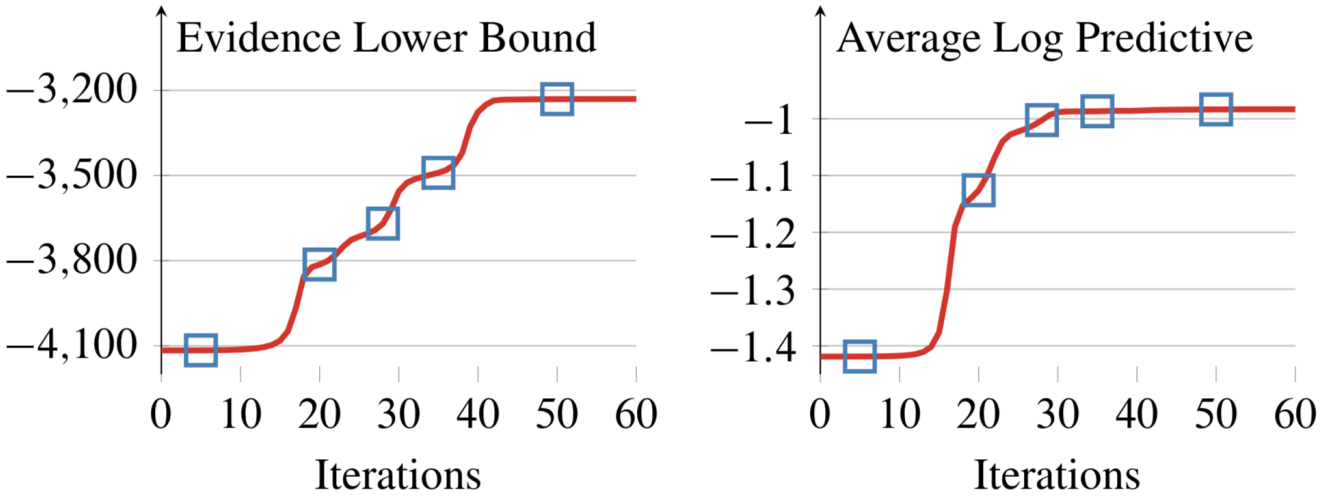
- VI turns inference into optimization
- Specify a variational family of distributions over the model parameters

$$Q = \{q_\phi(\theta); \phi \in \Phi\}$$

- Fit the variational parameters ϕ to minimize the distance (often in terms of KL divergence) to the exact posterior

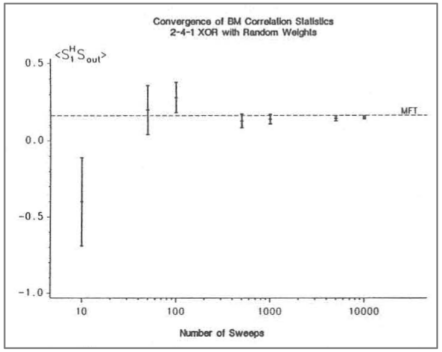
Example: Mixture of Gaussians



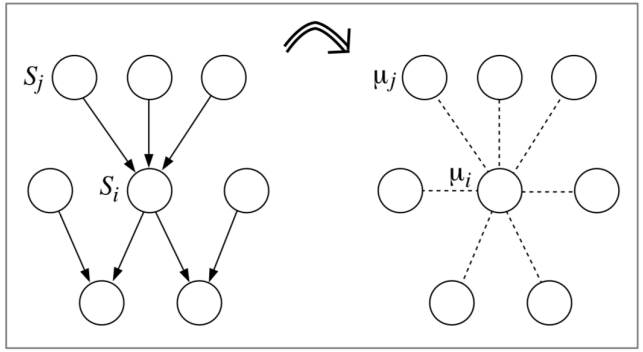


Adapted from David Blei

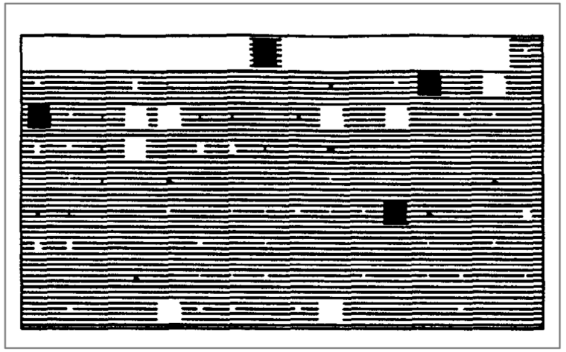
History



[Peterson and Anderson 1987]



[Jordan et al. 1999]

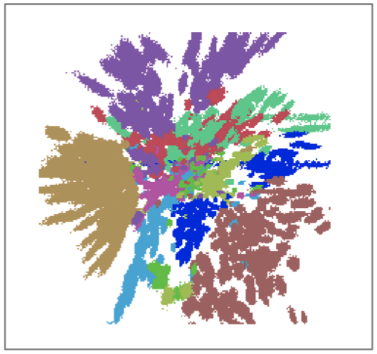


[Hinton and van Camp 1993]

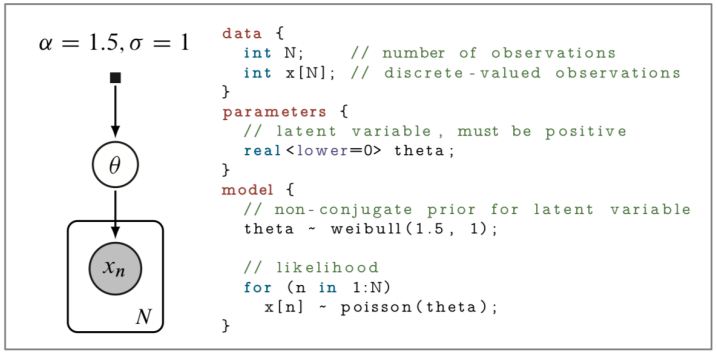
- Idea adapted from statistical physics – mean-field methods to fit a neural network (Peterson and Anderson, 1987).
- Picked up by Jordan's lab in the early 1990s, generalized it to many probabilistic models. (see Jordan et al., 1999 for an overview)
- Contributions from Hinton's group: mean-field for neural networks (Hinton and Van Camp, 1993); connection to the EM algorithm (Neal and Hinton, 1993).

Today

[Kingma and Welling 2013]



[Rezende et al. 2014]



[Kucukelbir et al. 2015]

- More and more work on variational inference now, making it scalable, easier to derive, faster, more accurate, and applying it to more complicated models and applications.
- Modern VI touches many important areas: probabilistic programming, reinforcement learning, neural networks, convex optimization, Bayesian statistics, and myriad applications.

- We use Kullback-Leibler (KL) divergence to measure the distance between two distributions.
- This comes from **information theory**, a field that has deep links to statistics and machine learning.
- The KL divergence for variational inference is

$$\mathrm{KL}(q\|p) = \mathbb{E}_{q(\theta)}\left(\log \frac{q(\theta)}{p(\theta|x)}\right)$$

  - If q is high, p and q should be close
  - if q is low then we don't care
- We choose q that are tractable: easy to take expectations and compute the pdf.
- Reversing the arguments leads to a different kind of variational inference, i.e., expectation propagation, which is in general more computationally expensive.

- We actually can't minimize the KL divergence exactly, but we can find an equivalent formulation that is tractable – the **evidence lower bound** (ELBO)
- By Jensen's inequality

$$\begin{aligned}
\log p(x) &= \log \int p(x,\theta) d\theta \\
&= \log \int q(\theta) \frac{p(x,\theta)}{q(\theta)} d\theta \\
&\geq \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \\
&\geq \mathbb{E}_q(\log p(x,\theta)) - \mathbb{E}_q(\log q(\theta))
\end{aligned}$$

This is the ELBO. Note that this is the free-energy lower bound we derived for EM.

- What does the ELBO have to do with the KL divergence to the posterior?
- Note that $p(\theta|x) = p(x,\theta)/p(x)$, use this in the KL divergence

$$\begin{aligned}
\mathrm{KL}(q(\theta)\|p(\theta|x)) &= \mathbb{E}_q\left(\log\frac{q(\theta)}{p(\theta|x)}\right) \\
&= \mathbb{E}_q\left(\log\frac{q(\theta)}{p(x,\theta)}\right) + \log p(x) \\
&= \log p(x) - \mathbb{E}_q\left(\log\frac{p(x,\theta)}{q(\theta)}\right)
\end{aligned}$$

- Therefore, the KL divergence is just the gap between the ELBO and the model evidence, and minimizing the KL divergence is equivalent to maximizing the ELBO.

More on The ELBO

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_q\left(\log \frac{p(x,\theta)}{q(\theta)}\right) \\ &= \mathbb{E}_q(\log p(x,\theta)) - \mathbb{E}_q(\log q(\theta))\end{aligned}$$

- $\mathcal{L}$ only requires the joint probability $p(x,\theta)$, which is computable.
- The ELBO trades off two terms
  - The first term drives q towards the MAP estimate.
  - The second term encourages q to be diffuse.
- Unfortunately, the ELBO is usually not convex, and VI may end up with local modes.

- A commonly used variational family is the mean field approximation, a variational family that factorizes

$$q(\theta) = \prod_{i=1}^{d} q_i(\theta_i)$$

  Each variable is independent. We can relax this constraint by using blockwise factorization.
- Note that this family is usually quite limited since the parameters in true posteriors are likely to be dependent.
  - E.g., in the Gaussian mixture model all of the cluster assignments z and the cluster locations μ are dependent on each other given the data x.
  - These dependencies often make the posterior difficult to work with.

- We now turn to optimizing the ELBO for the mean field approximation

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_q(\log p(x,\theta)) - \mathbb{E}_q \sum_{i=1}^{d} \log q_i(\theta_i) \\ &= \mathbb{E}_q(\log p(x,\theta)) - \sum_{i=1}^{d} \mathbb{E}_{q_i} \log q_i(\theta_i)\end{aligned}$$

- For each component $q_i(\theta_i)$

$$\begin{aligned}\mathcal{L} &= \int \prod_{i=1}^{d} q_i(\theta_i) \log p(x,\theta) d\theta - \sum_{i=1}^{d} \mathbb{E}_{q_i} \log q_i(\theta_i) \\ &= \mathbb{E}_{q_i} \mathbb{E}_{-q_i} \left(\log p(x,\theta)\right) - \mathbb{E}_{q_i} \log q_i(\theta_i) + \text{const}\end{aligned}$$

A Coordinate Ascent Algorithm

- Take the derivative w.r.t. $q_i(\theta_i)$

$$\frac{\partial \mathcal{L}}{\partial q_i(\theta_i)} = \mathbb{E}_{-q_i}(\log p(x, \theta)) - \log q_i(\theta_i) - 1 = 0$$

- This leads to a coordinate ascent algorithm

$$q_i^*(\theta_i) \propto \exp\left(\mathbb{E}_{-q_i}\left(\log p(x, \theta)\right)\right)$$

  - The RHS only depends on $q_j(\theta_j), j \neq i$.
  - This determines the form of the optimal $q_i(\theta_i)$. We only specify the factorization before.
  - While the optimal $q_i(\theta_i)$ might not be easy to compute (depending on the form), for many models it is.
- The ELBO converges to a *local minimum*. We use the resulting q as a proxy for the true posterior.

Connection to Gibbs Sampling

There is a strong relationship between Mean-Field VI and Gibbs sampling

- In Gibbs sampling, we sample from the conditional
- In Mean-Field VI (via coordinate ascent), we iteratively set each factor to

$$q_i(\theta_i) \propto \exp\left(\mathbb{E}(\log(\text{conditional}))\right)$$

Example: Multinomial conditionals

- Suppose the conditional is multinomial

$$p(\theta_i|\theta_{-i}, x) \sim \text{Multinomial}(\pi(\theta_{-i}, x))$$

- Then the optimial $q_i(\theta_i)$ is also a multinomial

$$q^*(\theta_i) \propto \exp\left(\mathbb{E}(\log \pi(\theta_{-i}, x))\right)$$

Exponential Family Conditionals

- Suppose each conditional is in the exponential family

$$p(\theta_i|\theta_{-i}, x) = h(\theta_i) \exp\left(\eta(\theta_{-i}, x) \cdot T(\theta_i) - A(\eta(\theta_{-i}, x))\right)$$

  where $\eta(\theta_{-i}, x)$ is the natural parameters and $T(\theta_i)$ is the sufficient statistics.
- This includes a lot of complicated models
  - Bayesian mixture of exponential families with conjugate priors
  - Hierarchical HMMs
  - Mixed-membership models of exponential families
  - Bayesian linear regression

- Compute the log of the conditional

$$\log p(\theta_i|\theta_{-i}, x) = \log h(\theta_i) + \eta(\theta_{-i}, x) \cdot T(\theta_i) - A(\eta(\theta_{-i}, x))$$

- Compute the expectation w.r.t. $q(\theta_{-i})$

$$\mathbb{E}\left(\log p(\theta_i|\theta_{-i}, x)\right) = \log h(\theta_i) + \mathbb{E}\left(\eta(\theta_{-i}, x)\right) \cdot T(\theta_i) - \mathbb{E}\left(A(\eta(\theta_{-i}, x))\right)$$

- Note that the last term does not depend on θ_i, therefore

$$q_i^*(\theta_i) \propto h(\theta_i) \exp\left(\mathbb{E}\left(\eta(\theta_{-i}, x)\right) \cdot T(\theta_i)\right)$$

  and the normalizing constant is $A(\mathbb{E}\left(\eta(\theta_{-i}, x)\right))$
- The optimal $q_i(\theta_i)$ is in the same exponential family as the conditional.

Examples: Bayesian Mixtures of Gaussians

- Consider the clustering of $x = \{x_1, \ldots, x_n\}$ using a finite mixture of Gaussians with generating variance one

$$z_i \sim \text{Discrete}(\pi), \quad x_i | z_i = k \sim \mathcal{N}(\mu_k, 1)$$
$$\mu_k \sim \mathcal{N}(\mu_0, \sigma_0), \quad k = 1, \ldots, K$$

- The joint probability is

$$\begin{aligned}
\log p(x, z, \mu) &= \sum_{i=1}^{n} \log p(x_i, z_i | \mu) + \sum_{k=1}^{K} \log \mathcal{N}(\mu_k | \mu_0, \sigma_0^2) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} 1_{z_i = k} \left(\log \pi_k + \log \mathcal{N}(x_i | \mu_k, 1) \right) \\
&\quad + \sum_{k=1}^{K} \log \mathcal{N}(\mu_k | \mu_0, \sigma_0^2)
\end{aligned}$$

Update $q(z_i)$

- The mean field family is

$$q(\mu, z) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | \tilde{\mu}_k, \tilde{\sigma}_k^2) \prod_{i=1}^{n} q(z_i | \phi_i)$$

- Coordinate ascent update for $q(z_i)$ is

$$q^*(z_i) \propto \exp\left(\mathbb{E}_{-q(z_i)}(\log p(x, z_i, z_{-i}, \mu))\right)$$

- Take expectation and restrict the terms relate to z_i

$$\begin{aligned} q^*(z_i) &\propto \exp(\log \pi_{z_i} + \mathbb{E}(\log \mathcal{N}(x_i | \mu_{z_i}, 1))) \\ &\propto \exp\left(\log \pi_{z_i} + x_i \tilde{\mu}_{z_i} - \frac{\tilde{\mu}_{z_i}^2 + \tilde{\sigma}_{z_i}^2}{2}\right) \end{aligned}$$

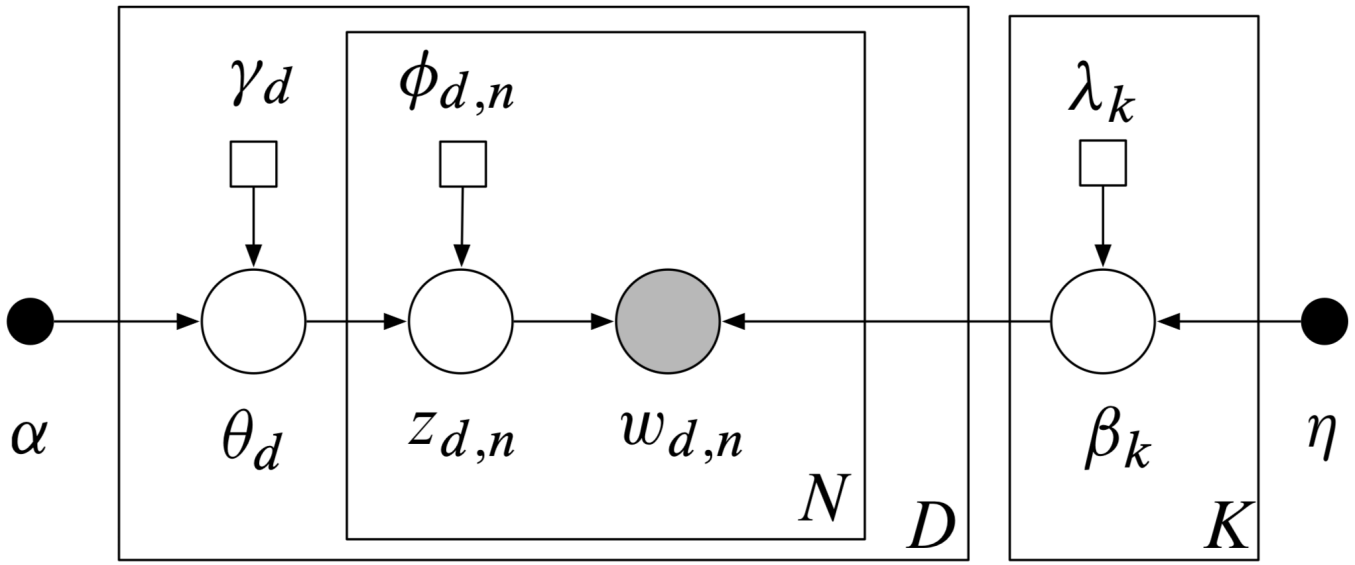
Update $q(\mu_k)$

- Similarly, the coordinate ascent update for $q(\mu_k)$ is

$$
\begin{aligned}
q^*(\mu_k) &\propto \exp\left(\mathbb{E}_{-q(\mu_k)}(\log p(x, z, \mu_k, \mu_{-k}))\right) \\
&\propto \exp\left(\sum_{i=1}^{n} q(z_i = k) \log \mathcal{N}(x_i|\mu_k, 1) + \log \mathcal{N}(\mu_k|\mu_0, \sigma_0^2)\right) \\
&\propto \exp\left(\sum_{i=1}^{n} \phi_{i,k}\left(x_i\mu_k - \frac{1}{2}\mu_k^2\right) + \frac{\mu_0}{\sigma_0^2}\mu_k - \frac{1}{2\sigma_0^2}\mu_k^2\right) \\
&\propto \mathcal{N}\left(\hat{\mu}_k, \hat{\sigma}_k^2\right)
\end{aligned}
$$

where

$$
\hat{\mu}_k = \frac{\dfrac{\mu_0}{\sigma_0^2} + \sum_{i=1}^{n} \phi_{i,k} x_i}{\dfrac{1}{\sigma_0^2} + \sum_{i=1}^{n} \phi_{i,k}}, \qquad \hat{\sigma}_k^2 = \frac{1}{\dfrac{1}{\sigma_0^2} + \sum_{i=1}^{n} \phi_{i,k}}
$$



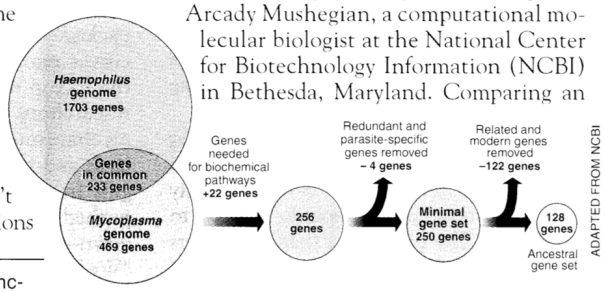
- The local variables are θ_d and z_d.
- The global variables are the topics $\beta_1, \ldots, \beta_K$.
- The variational distribution is

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n})$$

Seeking Life's Bare (Genetic) Necessities

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

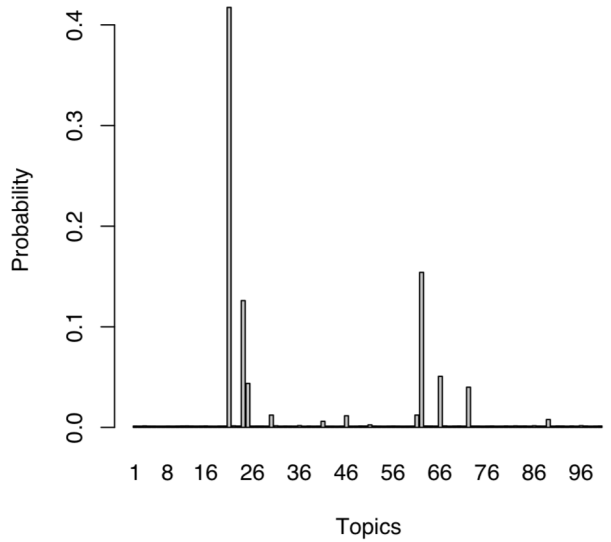
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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



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

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

SCIENCE • VOL. 272 • 24 MAY 1996



Adapted from David Blei

- The complete probability model

$$\theta_d \sim \text{Dirichlet}(\alpha), \quad \beta_k \sim \text{Dirichlet}(\eta)$$

$$z_{d,n}|\theta_d \sim \text{Discrete}(\theta_d), \quad w_{d,n}|z_{d,n}, \beta \sim \text{Discrete}(\beta_{z_{d,n}})$$

- The joint probability is

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

- We set $q(\beta_k|\lambda_k), q(\theta_d|\gamma_d), q(z_{d,n}|\phi_{d,n})$ accordingly

$$q(\beta_k|\lambda_k) \sim \text{Dirichlet}(\lambda_k), \quad q(\theta_d|\gamma_d) \sim \text{Dirichlet}(\gamma_d)$$

$$q(z_{d,n}|\phi_{d,n}) \sim \text{Discrete}(\phi_{d,n})$$

- **Update λ**

$$q(\beta_k|\lambda_k^*) \propto \exp\left(\mathbb{E}_{q(\beta_{-k},\theta,z)} \log p(w,z,\theta,\beta)\right)$$

$$\propto \exp\left(\sum_{j=1}^{V}(\eta_j - 1 + \sum_{d=1}^{D}\sum_{n=1}^{N}\phi_{d,n,k}w_{d,n}^j)\log\beta_{k,j}\right)$$

$$\boxed{\Rightarrow \lambda_{k,j}^* = \eta_j + \textstyle\sum_{d=1}^{D}\sum_{n=1}^{N}\phi_{d,n,k}w_{d,n}^j}$$

- **Update γ**

$$q(\theta_d|\gamma_d^*) \propto \exp(\mathbb{E}_{q(\beta,\theta_{-d},z)} \log p(w,z,\theta,\beta))$$

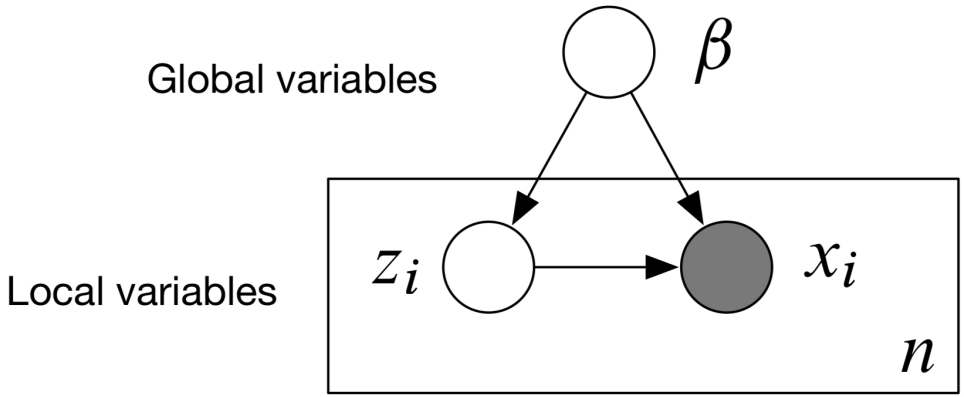
$$\propto \exp\left(\sum_{k=1}^{K}(\alpha_k - 1 + \sum_{n=1}^{N}\phi_{d,n,k})\log\theta_{d,k}\right)$$

$$\boxed{\Rightarrow \gamma_{d,k}^* = \alpha_k + \textstyle\sum_{n=1}^{N}\phi_{d,n,k}}$$

- **Update ϕ**

$$
\begin{aligned}
q(z_{d,n}|\phi^*_{d,n}) &\propto \exp(\mathbb{E}_{q(\beta,\theta,z_{-(d,n)})} \log p(w,z,\theta,\beta)) \\
&\propto \exp(\mathbb{E}_{q(\beta,\theta,z_{-(d,n)})}(\log p(z_{d,n}|\theta_d) \\
&\qquad + \log p(w_{d,n}|z_{d,n},\beta))) \\
&\propto \exp\left(\sum_{k=1}^{K} 1_{z_{d,n}=k}(\mathbb{E}_{\theta_d}(\log \theta_{d,k}) \right. \\
&\qquad \left. + \sum_{j=1}^{V} w^j_{d,n} \mathbb{E}_{\beta_k}(\log \beta_{k,j}))\right)
\end{aligned}
$$

$$
\boxed{\Rightarrow \phi^*_{d,n,k} \propto \exp\left(\mathbb{E}_{\theta_d}(\log \theta_{d,k}) + \textstyle\sum_{j=1}^{V} w^j_{d,n} \mathbb{E}_{\beta_k}(\log \beta_{k,j})\right)}
$$



$$p(\beta, z, x) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i|\beta)$$

- The observations are $x = \{x_1, \ldots, x_n\}$
- The **local** latent variables are $z = \{z_1, \ldots, z_n\}$
- The **global** variables are β
- The i-th data point x_i only depends on z_i and β

- **Goal**: compute $p(\beta, z|x)$
- Exponential family and conditional conjugacy

$$p(x_i, z_i|\beta) = h(x_i, z_i) \exp\left(\beta \cdot T(x_i, z_i) - A_\ell(\beta)\right)$$

$$\begin{aligned} p(\beta) &= h(\beta) \exp\left(\alpha \cdot T(\beta) - A_g(\alpha)\right) \\ &= h(\beta) \exp\left(\alpha_1 \cdot \beta - \alpha_2 A_\ell(\beta) - A_g(\alpha)\right) \end{aligned}$$

- Complete conditionals

$$p(\beta|x, z) = h(\beta) \exp\left(\eta_g(x, z) \cdot T(\beta) - A_g(\eta_g(x, z))\right)$$

$$p(z_i|x_i, \beta) = h(z_i) \exp(\eta_\ell(x_i, \beta) \cdot T(z_i) - A_\ell(\eta_l(x_i, \beta)))$$

where $\eta_g(x, z) = (\alpha_1 + \sum_{i=1}^n T(x_i, z_i), \alpha_2 + n)$

- The mean-field variational family

$$q(\beta, z) = q(\beta|\lambda) \prod_{i=1}^{n} q(z_i|\phi_i)$$

  - The global parameters λ govern the global variables
  - The local parameters ϕ_i govern the local variables
- Moreover, we set $q(\beta|\lambda), q(z_i|\phi_i)$ to be in the same exponential family

$$q(\beta|\lambda) = h(\beta) \exp(\lambda \cdot T(\beta) - A_g(\lambda))$$

$$q(z_i|\phi_i) = h(z_i) \exp(\phi_i \cdot T(z_i) - A_\ell(\phi_i))$$

- Update λ

$$
\begin{aligned}
q(\beta|\lambda^*) &\propto \exp(\mathbb{E}_{q(z)}(\log p(x,z,\beta)) \\
&\propto \exp\left(\mathbb{E}_{q(z)}(\log p(\beta) + \sum_{i=1}^{n} \log p(x_i, z_i|\beta))\right) \\
&\propto h(\beta)\exp(\mathbb{E}_{q(z)}(\eta_g(x,z)) \cdot T(\beta))
\end{aligned}
$$

Therefore

$$\boxed{\lambda^* = \mathbb{E}_{q(z)}(\eta_g(x,z))}$$

- Update ϕ_i

$$\begin{aligned} q(z_i|\phi_i^*) &\propto \exp(\mathbb{E}_{q(\beta,z_{-i})}(\log p(x,z,\beta))) \\ &\propto \exp(\mathbb{E}_{q(\beta)}(\log p(z_i|x_i,\beta))) \\ &\propto \exp(\mathbb{E}_{q(\beta)}(\log h(z_i) + \eta_\ell(x_i,\beta)\cdot T(z_i))) \\ &\propto h(z_i)\exp(\mathbb{E}_{q(\beta)}(\eta_\ell(x_i,\beta))\cdot T(z_i)) \end{aligned}$$

Therefore

$$\boxed{\phi_i^* = \mathbb{E}_{q(\beta)}(\eta_\ell(x_i,\beta))}$$

- We then iteratively update each parameter, holding others fixed.

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.

Initialize λ randomly.

repeat

  for *each data point* i **do**

    Set local parameter $\phi_i \leftarrow \mathbb{E}_\lambda[\eta_\ell(\beta, x_i)]$.

  end

  Set global parameter

$$\lambda \leftarrow \alpha + \sum_{i=1}^{n} \mathbb{E}_{\phi_i}[t(Z_i, x_i)].$$

until *the ELBO has converged*

- We introduced **variational inference** (VI), an alternative method to MCMC for approximate Bayesian inference.
- For models with conditional conjugacy, a mean-field approximation can be learned via coordinate ascent.
- This strategy is applicable to a generic class of models, including Bayesian mixture models, time series models (e.g., HMM), factorial models, multilevel regression, and mixed-membership models (e.g., LDA), etc.

Pros and Cons for Mean-field VI

- can be fast to train (compared to MCMC).
- may provide poor approximation, depending on the complexity of the posterior.

References

- M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graph- ical models. Machine Learning, 37:183–233, 1999.
- Ghahramani and Beal, Propagation Algorithms for Variational Bayesian Learning, 2001
- M. J. Beal and Z. Ghahramani, "The variational bayesian em algorithm for in- complete data: With application to scoring graphical model structures", Bayesian statistics, vol. 7, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D Heckerman, A. Smith, M West, et al., Eds., pp. 453–464, 2003.

References

- R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models, 89:355–368, 1998.
- D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.
- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518):859–877, 2017.