Variational Inference

Evan Glas

Overview

- Motivation
- Mathematical Derivation
- Coordinate Ascent Variational Inference
- Applications

A Familiar Problem

Suppose we are interested in the true value of an unobserved value θ given observations X. What is the posterior distribution of θ given prior beliefs $p(\theta)$?

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

Posterior Computation

Known: Sampling Model (Likelihood)

Known: Prior

Posterior

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

Problem: Evidence

Evidence

$$p(X) = \int_{\theta \in \Theta} p(X|\theta)p(\theta)d\theta$$

- Can become too expensive to compute $p(X)$ under all but the simplest joint distributions $p(X, \theta)$.
  - E.g., $p(X)$ may become intractable if θ is high dimensional or takes on a complicated prior.

Solutions

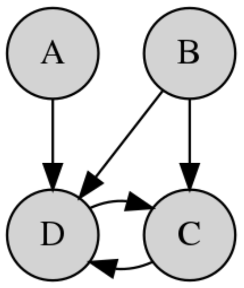
- Expectation Maximization
  - Avoid the problem entirely by simply optimizing the likelihood. Sacrifices computing a posterior distribution around the MLE.
- Sampling
  - Take random samples from the posterior or conditional posterior distributions
    - MCMC
    - Gibbs Sampling
    - Metropolis-Hastings
- Applying conjugate prior-posterior pairs
- Variational Inference

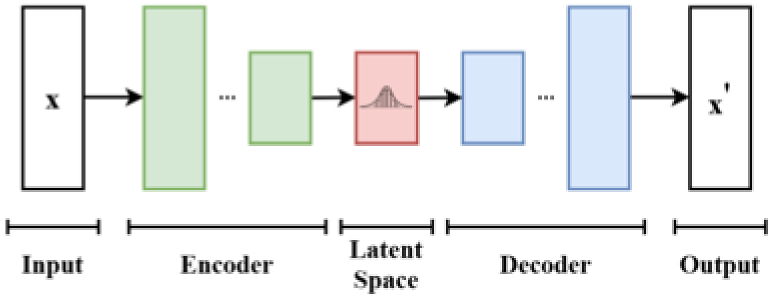
Variational Inference (VI)

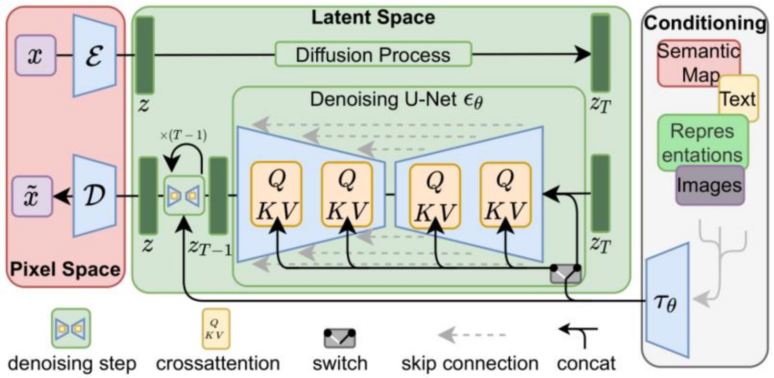
- Variational Inference: a set of techniques for approximating and evaluating a proposal posterior distribution, $q(\theta)$, to match a true posterior $p(\theta|X)$.
  - Frames posterior approximation as an optimization problem.
- Goal: Find $q(\theta)$ as close as possible to $p(\theta|X)$.

VI: Past and Present

- 1976: Rustagi: ***Variational Methods in Statistics***
- 1999: Jordan et. al: ***An Introduction to Variational Methods for Graphical Models***
- 2003: Jordan et. al: ***Latent Dirichlet Allocation***
- 2013: Kingma et. al: ***Auto-Encoding Variational Bayes***
- 2021: Rombach et. al: ***High-Resolution Image Synthesis with Latent Diffusion Models***







What Does “Variational” Mean?

- From ***Calculus of Variations***
  - The study of the optimization of *functionals*
- What is a “functional”?
  - A *functional* is a function-valued mapping.
    - Definite Integral: $\int_a^b f(x)dx$
    - Derivative at a point: $\frac{d}{dx} f(x)|_{\{x=x_0\}}$
  - That is, a functional takes a mapping that takes a function as input.
- How does this relate?
  - We seek to optimize a proposal posterior distribution $q(\theta)$, a function defined over θ.

VI: Optimization Function

- Goal: Find $q(\theta)$ as close as possible to $p(\theta|X)$.
  - What does "close" mean?
- Kullback-Leibler (KL) Divergence
  - A measure of distance between two probability distributions, P and Q

$$D_{KL}(P||Q) = \int_{\theta} p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$

  - We hope to minimize the KL Divergence between $q(\theta)$ and $p(\theta|X)$.

$$\boxed{\min_{q} D_{KL}(Q||P)}$$

Evidence Lower Bound (ELBO)

- The optimization problem "maximize the KL Divergence" is equivalent to a reformulated problem, "minimize the ELBO".

$$D_{KL}(P||Q) = \mathbb{E}_q(\log q(\theta)) - \mathbb{E}_q(\log p(\theta|X))$$
$$= \mathbb{E}_q(\log q(\theta)) - \mathbb{E}_q(\log p(\theta, X)) + \log p(X)$$

- Note: $D_{KL}(P||Q) \geq 0.\ p(\theta)$ is constant with respect to q.

$$\Rightarrow \log p(X) \geq \mathbb{E}_q\left(\log \frac{p(\theta, X)}{q(\theta)}\right) = ELBO$$

- New goal: maximize the lower bound on $\log p(X)$, the ELBO.

Variational Inference: Now What?

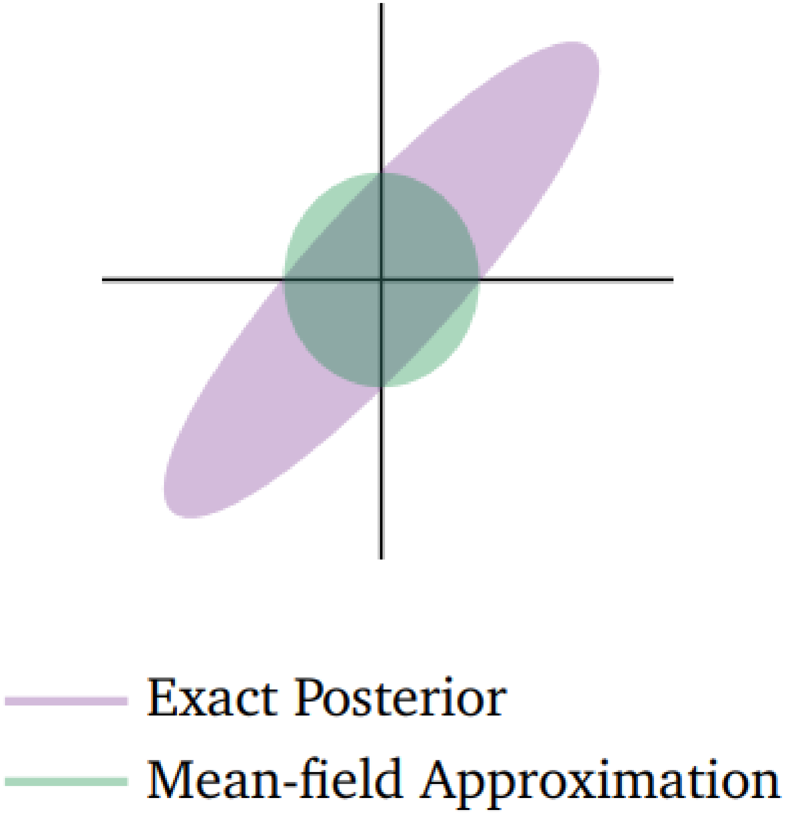
- Variational methods take various approaches from this point.
  - All share the common objective function described previously.
- Choice of $q(\theta)$
  - E.g., is $q(\theta)$ normal, exponential, uniform, etc.?
- Optimization procedure
  - Do we need an exact solution?
  - Do we want a deterministic solution?
  - What computing resources are available?
- Classic Variational Inference: apply mean-field approximation and iterative optimization algorithm

Mean-field Approximation

- We may simplify the optimization problem by constraining $q(\theta)$ to a distribution that factorizes (mean-field approximation):

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

- The mean-field approximation enables us to apply an iterative optimization algorithm
  - Optimize each q_i one at a time without considering the other $q_i(\theta)$



https://arxiv.org/pdf/1601.00670.pdf

VI Algorithm

- From the mean-field approximation, we derive the VI update rule:

$$\mathbb{E}_q\left(\log\frac{p(\theta,X)}{q(\theta)}\right) = ELBO$$

- Differentiate the ELBO with respect to q_i and set to zero.
  - Requires techniques from variational calculus
  - Simplify to get the following result:

$$q_i^*(\theta) = \exp\left(\mathbb{E}_{q_{i\neq j}}(\log p(\theta,X))\right)$$

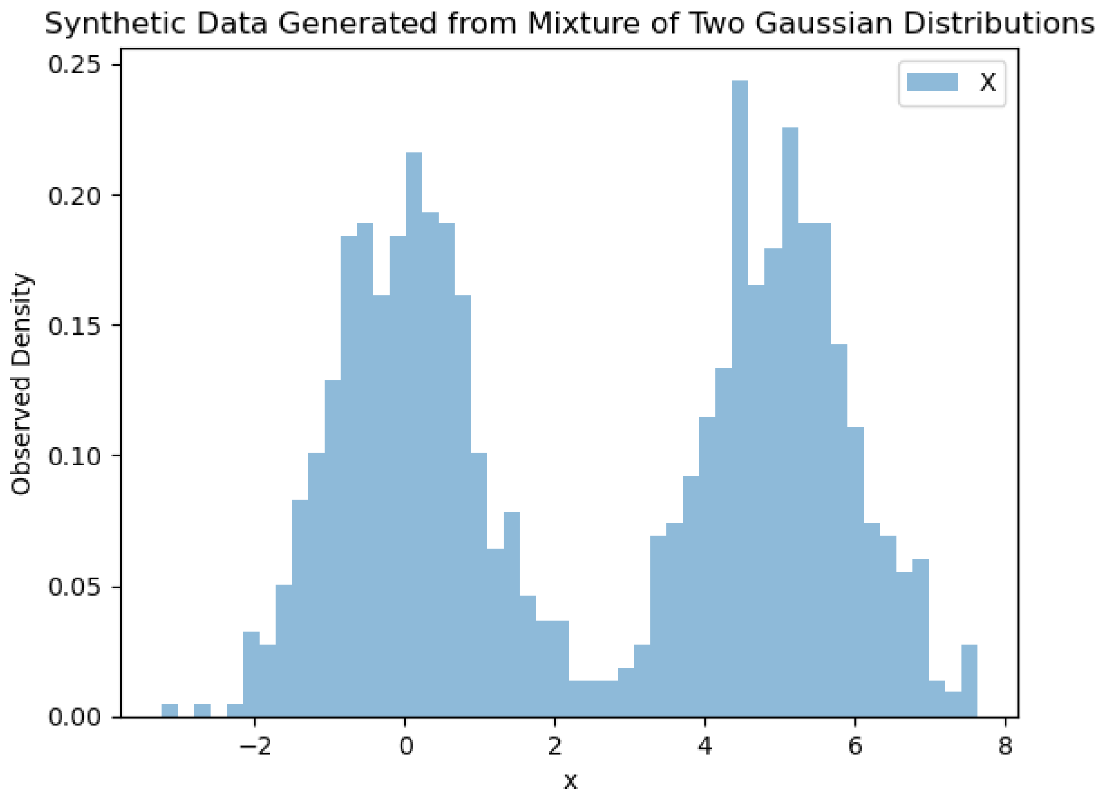
Coordinate Ascent Variational Inference (CAVI)

$$q_i^*(\theta) = \exp\left(\mathbb{E}_{q_{i\neq j}}(\log p(\theta, X))\right)$$

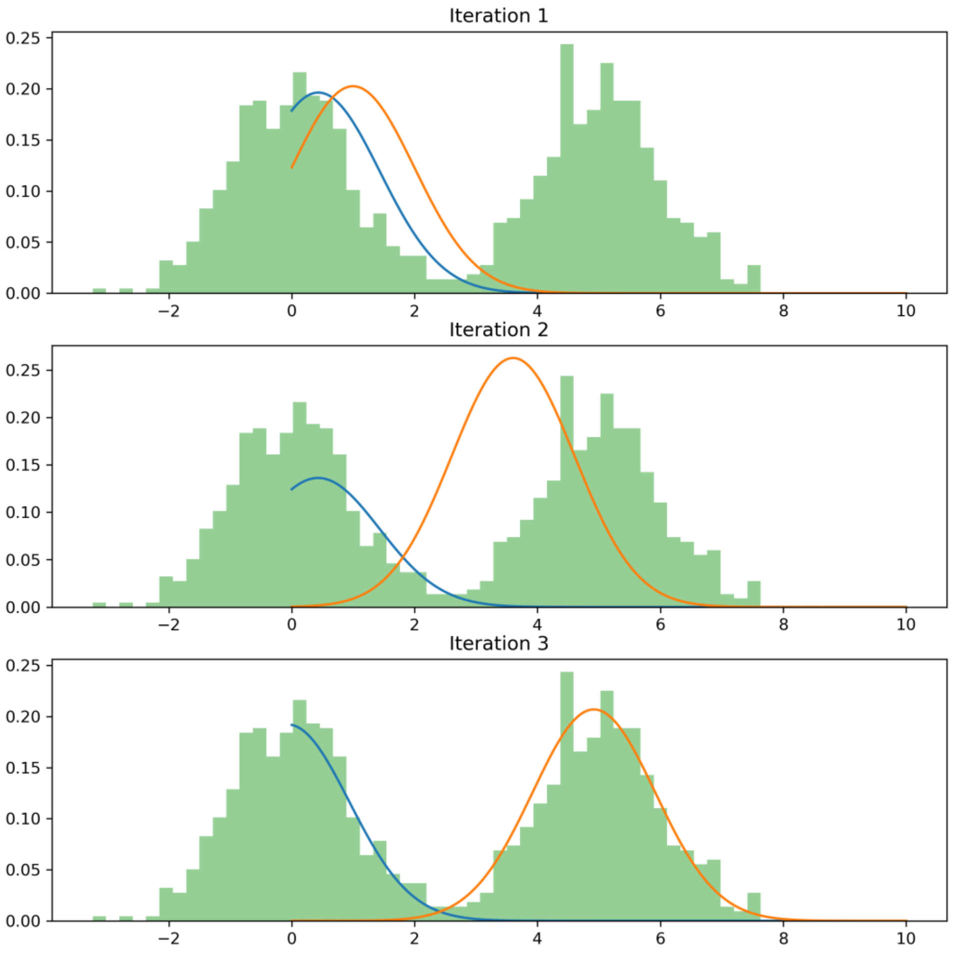
- Like Gibb's sampling, we may apply the above update rule iteratively since $q(\theta)$ factorizes and we may derive the full conditional of θ_j with respect to $\{\theta\}_{i\neq j}$.
- This algorithm is called *Coordinate Ascent Variational Inference (CAVI)*.
  - At each iteration, maximizes the ELBO with respect to q_i while holding the remaining $q_{i\neq j}$ fixed
  - Not guaranteed to reach global optimum (the ELBO is not necessarily concave)

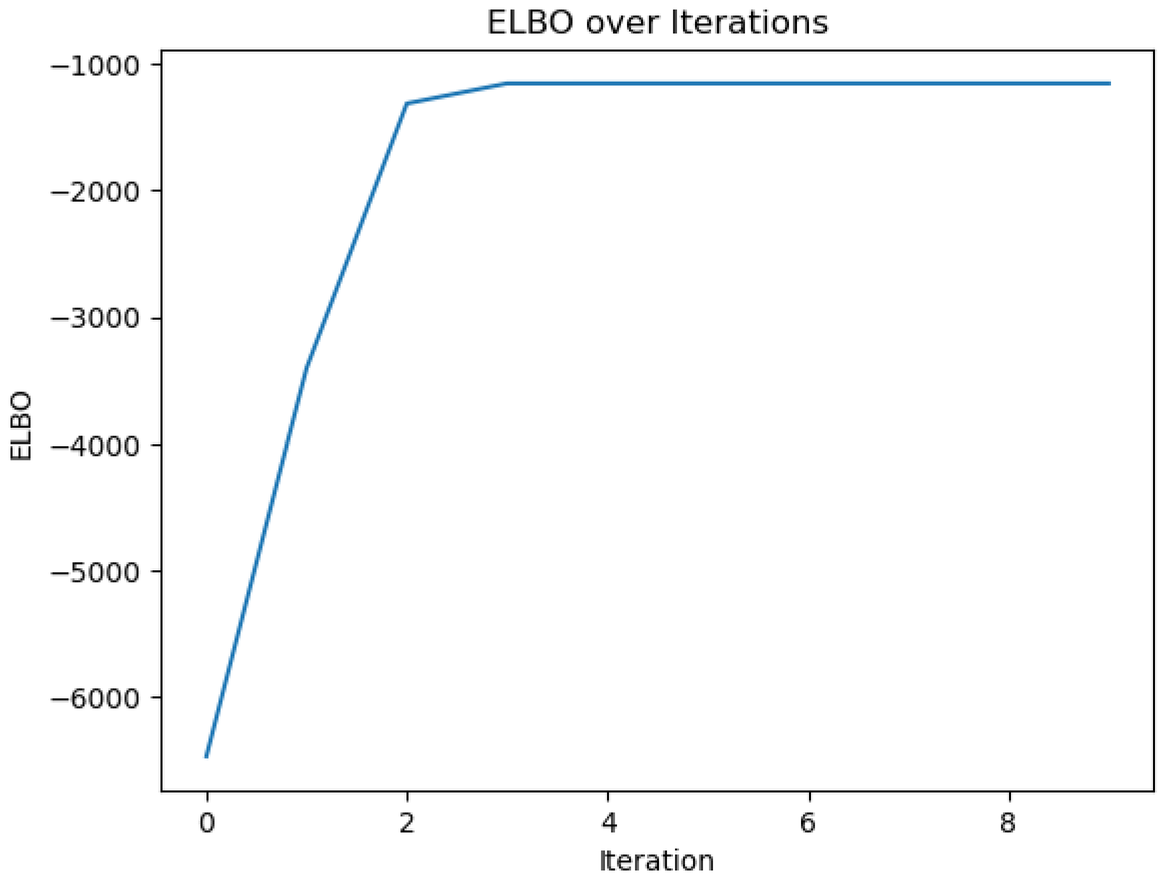
CAVI Example: Gaussian Mixture Model

- We may apply CAVI towards the problem of fitting a Gaussian Mixture model.
- Consider the following observed data →
- We make the following assumptions:
  - $x_i \sim \begin{cases} N(\mu_1, \sigma = 1),\ prob. = \phi_i \\ N(\mu_2, \sigma = 1),\ prob. = 1 - \phi_i \end{cases}$
  - $(\mu_1, \mu_2, \{\phi\}_i) \sim q$



CAVI Fitting: Visualization



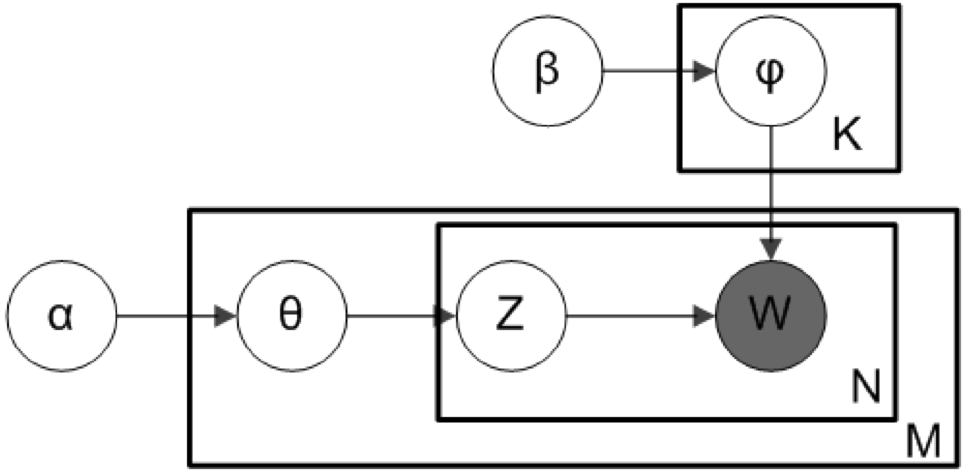


Beyond CAVI

- CAVI Limitations
  - Requires one to compute the update rule explicitly, which may become difficult in high dimensions or complex $q(\theta)$.
  - Necessary to loop over the entire dataset at each iteration to compute the expectation $\mathbb{E}_{q_{i \neq j}}(\log p(\theta, X))$.
- Stochastic Gradient Descent
  - Can instead apply SGD, using only a portion of the data each step.
- SGD offers the computational efficiency to apply VI to large datasets and potentially complex $q(\theta)$.

Application: Topic Modeling

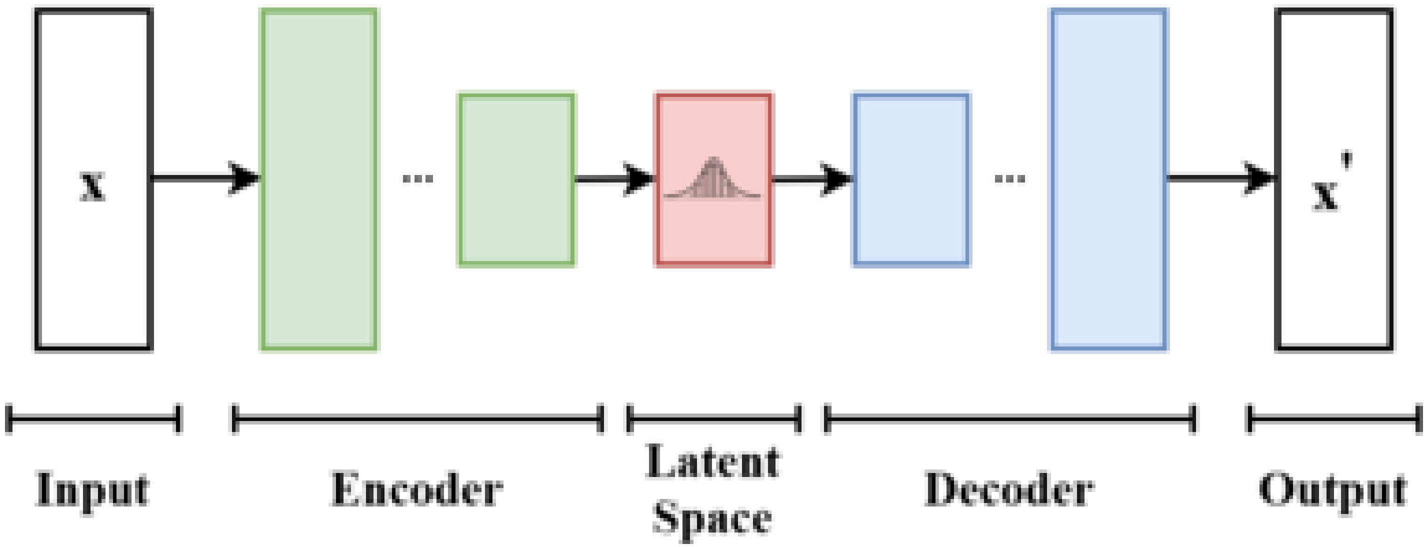
- Goal: cluster a set of documents given a bag of word assumption over potential topics.
  - Applies a Bayesian network to model the distribution of topics and words
- Latent Dirichlet Allocation
  - In Jordan et. al proposed modelling the posterior distribution of topics across documents and words across documents as Dirichlet priors
  - Optimize the posterior via VI

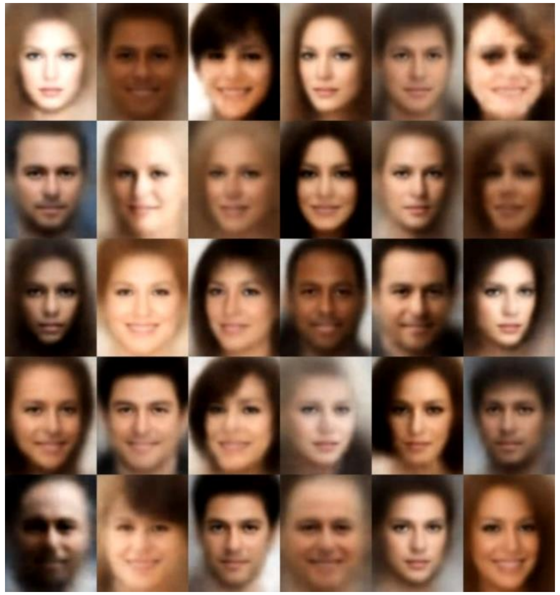


- M documents
- N_i words per document
- Z: topic of word j
- θ: distribution of topics for document i
- ϕ: distribution of words for topic K
- α, β: Dirichlet distribution parameters

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation#cite_note-blei2003-3

Application: Variational Autoencoder: VAE





- A machine learning model to approximate arbitrary distributions $q(\theta)$
  - Represents $q(\theta)$ via a neural network
  - Jointly estimates $p(\theta, X)$
- Uses the same optimization function (maximize the ELBO)

Why VI?

- VI has several advantages over alternative posterior density estimation tools.
  - Faster convergence: stochastic sampling methods may take extremely long to converge.
  - Deterministic: not (necessarily) stochastic
  - Parallelizable: can apply posterior updates in parallel (assuming mean-field approximation)
- VI retains a broad range of applications within fields including statistics, physics, and machine learning.
  - E.g., stable diffusion