

# EM Algorithm and Variational Inference

Tingyin Wang, Zezhi Wang

School of Management  
University of Science and Technology of China

November 17, 2021

# Plan

- 1 The EM algorithm
  - Maximization-Maximization Aspect
- 2 Mixture Models
  - Collapsing Variance Problem
  - K-means Algorithm
- 3 Variational Inference
  - Evidence Lower Bound
  - Mean-Field Variational Family
  - Communities Detection

# Background

In frequentist framework, we need to find the maximum likelihood estimator (MLE) of parameters. Sometimes it has explicit expression and can directly be solved. But most of the time it does not.

- One approach is to use the gradient-based optimizer, where we often have to enforce constraints and they can be tricky. For example, the parameters of multinomial distribution should be taken within  $[0, 1]$ . Also, the summation of them is restricted to 1.
- We here introduce the Expectation Minimization (EM) algorithm, which is straightforward to extract the maximum under certain circumstance.

# Basic idea

- Among those circumstances, there is a particular type, in which some examples are grouped, censored or truncated data with missing observation and data from mixtures of distribution. .
- To solve this type of problem, we formulate an associated statistical problem with the same parameters with "augmented data" from which it is easier to work out MLE.

# Basic idea

Let  $x_i$  be the visible or observed variables in case  $i$ , and  $z_i$  be the hidden or missing variables. The goal here is to maximize the log likelihood of the observational data:

$$\ell(\theta) = \sum_{i=1}^N \log p(x_i | \theta) = \sum_{i=1}^N \log \left[ \sum_{z_i} p(x_i, z_i | \theta) \right]$$

Unfortunately this is hard to optimize since  $\mathbf{z}_i$  is unknown. Let the complete data be  $y_i = (x_i, z_i)$ , we define the log likelihood w.r.t to the complete data as:

$$\ell_c(\theta | x, z) \triangleq \sum_{i=1}^N \log p(y_i | \theta).$$

# E-step

Furthermore, we define the expected complete data log likelihood as follows (E-step):

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{z \sim p(z|x, \theta^{(t)})} [\ell_c(\theta | z, x)]$$

where  $t$  is the current iteration number.  $Q$  is called the auxiliary function. Noting that

$$\log p(x | \theta) = \log p(y | \theta) - \log p(\mathbf{z} | \mathbf{x}, \theta)$$

Take expectation w.r.t  $z | x, \theta^{(t)}$  and denote  $\mathbb{E}_{z \sim p(z|x, \theta^{(t)})} [\log p(z | x, \theta)]$  as  $H(\theta, \theta^{(t)})$ , we have

$$\log p(x | \theta) = Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)}).$$

## E-step

We show that  $H(\theta, \theta^{(t)})$  is maximized w.r.t  $\theta$  when  $\theta = \theta^{(t)}$ :

$$\begin{aligned} & H(\theta^{(t)} | \theta^{(t)}) - H(\theta | \theta^{(t)}) \\ &= \int -\log \left[ \frac{p(\mathbf{z} | \mathbf{x}, \theta)}{p(\mathbf{z} | \mathbf{x}, \theta^{(t)})} \right] p(\mathbf{z} | \mathbf{x}, \theta^{(t)}) d\mathbf{z} \\ &\geq -\log \int p(\mathbf{z} | \mathbf{x}, \theta) d\mathbf{z} = 0. \end{aligned}$$

# M-step

In the M step, we optimize the  $Q$  function wrt  $\theta$  :

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{(t)}).$$

Therefore,

$$\begin{aligned} \log p(x | \theta^{(t+1)}) &= Q(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) \\ &\geq Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) \\ &= \log p(x | \theta^{(t)}), \end{aligned}$$

i.e., the  $(t + 1)$ th guess  $\theta^{(m+1)}$  through EM will never be less likely than the  $(t)$ th guess  $\theta^{(t)}$ .



# Summary of EM

In summary, the EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

- Expectation step (E step): Define  $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$  as the expected value of the log likelihood function of  $\boldsymbol{\theta}$ , with respect to the current conditional distribution of  $\mathbf{Z}$  given  $\mathbf{X}$  and the current estimates of the parameters  $\boldsymbol{\theta}^{(t)}$  :

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{Z} \sim p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})} [\ell_c(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{X})]$$

- Maximization step (M step): Find the parameters that maximize this quantity:

$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) .$$

# Exponential family

It can be shown that both the E- and M-steps will have particularly simple forms when  $p(y \mid \theta)$  is from an exponential family:

$$p(y \mid \theta) = b(y) \exp \left\{ \mathbf{c}^\top(\theta) \mathbf{t}(y) - a(\theta) \right\},$$

where  $\mathbf{t}(y)$  is the vector of complete-data sufficient statistics. For the exponential families, the E-step can be written as

$$Q(\theta, \theta^{(t)}) \propto \mathbf{c}^\top(\theta) E_{z \sim p(z \mid x, \theta^{(t)})} \mathbf{t}(y) - a(\theta) := \left[ \mathbf{c}^\top(\theta) \mathbf{t}^{(t)} - a(\theta) \right].$$

Hence it is sufficient to compute for M-Step:

$$\theta^{(t+1)} = \arg \max_{\theta} \left[ \mathbf{c}^\top(\theta) \mathbf{t}^{(t)} - a(\theta) \right].$$

For the case that  $\mathbf{c}(\theta) = \theta$ , the M-step is the concave optimization.

# Maximization-Maximization Aspect

Another way to realize the EM algorithm is as a joint maximization procedure. For  $\forall$  density  $q$  of the the missing data  $z$ , noting that

$$\begin{aligned} \text{KL}(q(z) \| p(z | x, \theta)) &= E_q \left[ \log \frac{q(z)}{p(z | x, \theta)} \right] \\ &= E_q[\log q(z)] - E_q[\log p(z | x, \theta)] \\ &= E_q[\log q(z)] - E_q[\log p(y | \theta)] + \log p(x | \theta). \end{aligned}$$

Let  $\text{ELBO}(q, \theta) := E_q[\log p(y | \theta)] - E_q[\log q(z)]$ , then

$$\text{KL}(q(z) \| p(z | x, \theta)) + \text{ELBO}(q, \theta) = \log p(x | \theta),$$

i.e.,  $\text{ELBO}(q, \theta)$  is the evidence lower bound of  $\log p(x | \theta)$ .

# Illustration

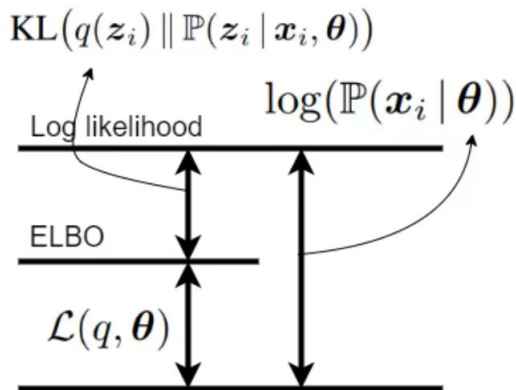


Figure: An illustration of the above equation.

# Maximization-maximization aspect

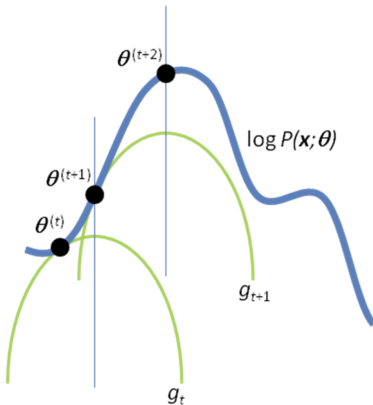
We can iteratively extract  $q$  and  $\theta$  by maximize  $\text{ELBO}(q, \theta)$ .  
When  $\theta^{(t)}$  is fixed,  $q(z) = p(z | x, \theta^{(t)})$  would extinguish the gap between  $\text{ELBO}(q, \theta^{(t)})$  and  $\log p(x | \theta^{(t)})$ . Accordingly,

$$\begin{aligned}\theta^{(t+1)} &= \arg \max_{\theta} \text{ELBO}(p(z | x, \theta^{(t)}), \theta) \\ &= \arg \max_{\theta} E_{z \sim p(z|x, \theta^{(t)})} [\log p(y | \theta)],\end{aligned}$$

therefore, the EM algorithm could be viewed as the coordinate ascend for both  $q$  and  $\theta$ , and

$$\begin{aligned}\log p(x | \theta^{(t+1)}) &\geq \text{ELBO}(p(z | x, \theta^{(t)}), \theta^{(t+1)}) \\ &\geq \text{ELBO}(p(z | x, \theta^{(t)}), \theta^{(t)}) \\ &= \log p(x | \theta^{(t)}).\end{aligned}$$

# Illustration



**Figure:** An illustration, where  $g_t(\theta) = \text{ELBO}(p(z | x, \theta^{(t)}), \theta)$ .

## Further discussion

- In E-step, the calculation of the conditional density and expectation is usually intractable without a closed-form. While some numerical methods (e.g. Laplace approximation, Monte Carlo) could be applied, the computational burden is larger with the increase of the size of missing data.
- In general, we just need to ensure the ascend of Q function in M-step, instead of solving its maximum. The one-step procedure embedded within M-step is more preferable considering the cost of optimization.
- While the M-step is usually the concave optimization. The problem to maximize of the marginal likelihood could be non-convex. Different initial values for EM are necessary to avoid the local maximums.

# Mixture models

- In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. While the information of which sub-population should the observed data belongs to remains unknown.
- Formally, a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population.



# Mixture models

- Suppose  $z_i \in \{1, \dots, K\}$  and  $P(z_i = k) = \pi_k$ .
- For the conditional likelihood, we define

$$p(\mathbf{x}_i | z_i = k, \theta_k) = p_k(\mathbf{x}_i | \theta_k),$$

where  $\theta_k$  is the parameters of  $k$ th sub-distribution.

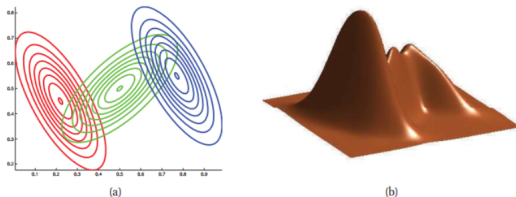
- We are mixing together the  $K$  base distributions as follows:

$$p(\mathbf{x}_i | \Theta) = \sum_{k=1}^K \pi_k p_k(\mathbf{x}_i | \theta_k).$$

# Mixtures of Gaussians

- Each base distribution in the mixture is a multivariate Gaussian with mean  $\mu_k$  and covariance matrix  $\Sigma_k$ .

$$p(\mathbf{x}_i | \Theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k).$$



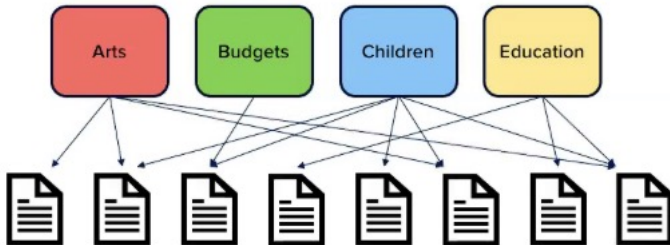
**Figure 11.3** A mixture of 3 Gaussians in 2d. (a) We show the contours of constant probability for each component in the mixture. (b) A surface plot of the overall density. Based on Figure 2.23 of (Bishop 2006a). Figure generated by `mixGaussPlotDemo`.

# House prices

- Assume that we observe the prices of  $N$  different houses.
- Different types of houses in different neighborhoods will have vastly different prices, but the price of a particular type of house in a particular neighborhood (e.g., three-bedroom house in moderately upscale neighborhood) will tend to cluster fairly closely around the mean.
- One possible model of such prices would be to assume that the prices are accurately described by a mixture model with  $K$  different components, each distributed as a normal distribution with unknown mean and variance, with each component specifying a particular combination of house type/neighborhood.

# Topics in a document

Assume that a document is composed of  $N$  different words from a total vocabulary of size  $V$ , where each word corresponds to one of  $K$  possible topics. The distribution of such words could be modelled as a mixture of  $K$  different  $V$ -dimensional categorical distributions. A model of this sort is commonly termed a topic model. It's worth noting that  $V \gg N$  usually.



# Using mixture models for clustering

- Generally, the latent variables do not necessarily have to have any meaning, we might simply introduce latent variables to make the model more powerful.
- Use the EM algorithm for clustering, we compute the posterior probability  $p(z_i = k \mid \mathbf{x}_i, \Theta)$

$$r_{ik} \triangleq p(z_i \mid \mathbf{x}_i, \Theta) \propto \pi_k p_k(\mathbf{x}_i \mid \theta_k)$$

- It may be reasonable to compute a hard clustering using the MAP estimate, given by

$$z_i^* = \arg \max_k r_{ik} = \arg \max_k \log p_k(\mathbf{x}_i \mid \theta_k) + \log \pi_k.$$

# EM for GMMs

The expected complete data log likelihood is given by

$$\begin{aligned} Q(\Theta, \Theta^{(t-1)}) &\triangleq \mathbb{E}_{z \sim p(z|x, \Theta^{(t-1)})} \left[ \sum_i \log p(\mathbf{x}_i, z_i | \Theta) \right] \\ &= \sum_i \mathbb{E}_{z \sim p(z|x, \Theta^{(t-1)})} \left[ \log \left[ \prod_{k=1}^K (\pi_k p_k(\mathbf{x}_i | \theta_k))^{I(z_i=k)} \right] \right] \\ &= \sum_i \sum_k p(z_i = k | \mathbf{x}_i, \theta^{(t-1)}) \log [\pi_k p_k(\mathbf{x}_i | \theta_k)] \\ &= \sum_i \sum_k r_{ik}^{(t-1)} \log \pi_k + \sum_i \sum_k r_{ik}^{(t-1)} \log p(\mathbf{x}_i | \theta_k) \end{aligned}$$

where  $r_{ik}^{(t-1)} \triangleq p(z_i = k | \mathbf{x}_i, \theta^{(t-1)})$  is the responsibility that cluster  $k$  takes for data point  $i$ .

# EM for GMMs

- The E step has the following simple form, which is the same for any mixture model:

$$r_{ik}^{(t-1)} = \frac{\pi_k^{(t-1)} p(\mathbf{x}_i | \theta_k^{(t-1)})}{\sum_{k'} \pi_{k'}^{(t-1)} p(\mathbf{x}_i | \theta_{k'}^{(t-1)})}$$

- In the M step, we optimize  $Q$  w.r.t.  $\pi$  and the  $\theta_k$ . For  $\pi$ , we obviously have

$$\pi_k^{(t)} = \frac{1}{N} \sum_i r_{ik}^{(t-1)} = \frac{r_k^{(t-1)}}{N},$$

where  $r_k^{(t-1)} \triangleq \sum_i r_{ik}^{(t-1)}$  is the weighted number of points assigned to cluster  $k$ .

# EM for GMMs

To derive the M step for the  $\mu_k$  and  $\Sigma_k$  terms, we look at the parts of  $Q$  that depend on  $\mu_k$  and  $\Sigma_k$ ,

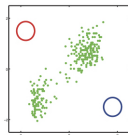
$$\begin{aligned}\ell(\mu_k, \Sigma_k) &= \sum_k \sum_i r_{ik}^{(t-1)} \log p(\mathbf{x}_i | \theta_k) \\ &\propto -\frac{1}{2} \sum_i r_{ik}^{(t-1)} \left[ \log |\Sigma_k| + (\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) \right]\end{aligned}$$

This is just a weighted version of the standard problem of computing the MLEs of an MVN. The new parameter estimates are given by

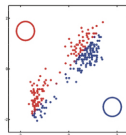
$$\begin{aligned}\mu_k^{(t)} &= \frac{\sum_i r_{ik}^{(t-1)} \mathbf{x}_i}{r_k^{(t-1)}} \\ \Sigma_k^{(t)} &= \frac{\sum_i r_{ik}^{(t-1)} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{r_k^{(t-1)}} = \frac{\sum_i r_{ik}^{(t-1)} \mathbf{x}_i \mathbf{x}_i^T}{r_k^{(t-1)}} - \mu_k^{(t)} \mu_k^{(t)T}.\end{aligned}$$



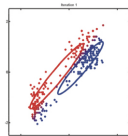
# The convergence of EM



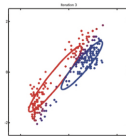
(a)



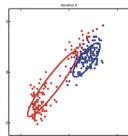
(b)



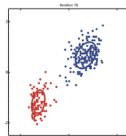
(c)



(d)



(e)



(f)

# Collapsing variance problem

As usual, the EM of GMM may be singular. For example, when  $\Sigma_k = \sigma_k^2 I_D$  and  $K = 2$ . It is possible to get an infinite likelihood by assigning one of the centers, say  $\mu_2$ , to a single data, say  $\mathbf{x}_1$ , i.e.  $\mathbf{x}_1 = \mu_2$ , since

$$\mathcal{N}(\mathbf{x}_1 \mid \mu_2, \Sigma_2) \propto \frac{1}{\sqrt{2\pi\sigma_2^2}},$$

hence  $\sigma_2$  would go to infinite and cause the collapsing variance problem.

# Collapsing variance problem

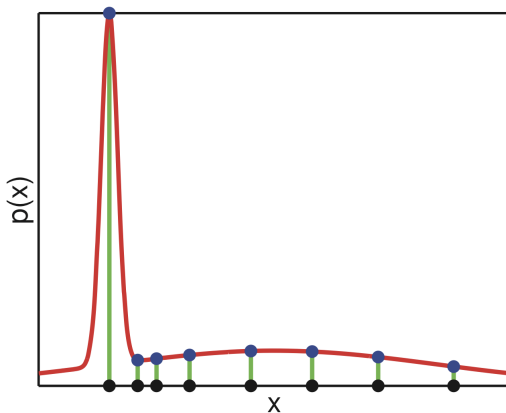


Figure: Illustration of how singularities can arise in EM of GMM.

# Dirichlet prior correction

An easy solution to this is to impose the prior constraint. The new auxiliary function:

$$Q'(\Theta, \Theta^{(t-1)}) = \left[ \sum_i \sum_k r_{ik}^{(t-1)} \log \pi_{ik} + \sum_i \sum_k r_{ik}^{(t-1)} \log p(\mathbf{x}_i | \theta_k) \right] \\ + \log p(\pi) + \sum_k \log p(\theta_k)$$

Note that the E step remains unchanged. It is natural to use a Dirichlet prior:  $\pi \sim \text{Dir}(\alpha)$ , since this is conjugate to the categorical distribution. The estimate of MAP is given by

$$\pi_k = \frac{r_k^{(t-1)} + \alpha_k - 1}{N + \sum_k \alpha_k - K}.$$

Additionally, we assume  $\theta_k \sim G_0$ . We don't specify the form of  $G_0$  here.

# MAP estimation

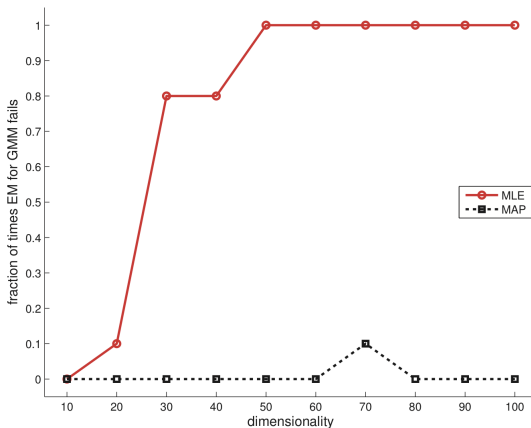


Figure: Comparison of MAP and MLE.

# K-means algorithm

A popular variant of the EM algorithm for GMMs known as the K-means algorithm. Consider a GMM in which we assume  $\Sigma_k = \sigma^2 \mathbf{I}_D$  and  $\pi_k = 1/K$  are fixed, only  $\mu_k \in \mathbb{R}^D$ , have to be estimated.

During the E step:

$$p(z_i = k \mid \mathbf{x}_i, \Theta) \approx \mathbb{I}(k = z_i^*)$$

where  $z_i^* = \operatorname{argmax}_k p(z_i = k \mid \mathbf{x}_i, \Theta)$ . (Hard **EM**). Since we assumed an equal spherical covariance matrix for each cluster,

$$z_i^* = \operatorname{argmin}_k \|\mathbf{x}_i - \mu_k\|_2^2.$$

The M step:

$$\mu_k = \frac{1}{N_k} \sum_{i: z_i=k} \mathbf{x}_i.$$

# K-means algorithm

- EM are somewhat more flexible and with a covariance matrix we can make the boundaries elliptical, as opposed to circular boundaries with K-means.
- EM seem to be more robust.
- EM usually tend to be slower than K-Means because it takes more iterations of the EM algorithm to reach the convergence.
- Both models converge to a local optimal solution and not a global one. One solution is we perform EM with K-Means initializer, in this way we will get a result that is robust and can converge to the optimal.

# Bayesian models

Let  $\mathbf{x} = x_{1:n}$  be a set of observed variables and  $\mathbf{z} = z_{1:m}$  be a set of latent variables or parameter (they are the same thing in Bayesian models), with joint density  $p(\mathbf{z}, \mathbf{x})$ . We omit constants, such as hyperparameters, from the notation.

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{z}, \mathbf{x})}{\int p(\mathbf{z}, \mathbf{x}) d\mathbf{z}}.$$

There is a problem that how to approximate difficult-to-compute probability densities.



# Variational Inference

- In variational inference (VI), we define a flexible family  $\mathcal{Q}$  of distributions over the latent variables. We then find the one that is closest to the probability densities  $p(\mathbf{z} \mid \mathbf{x})$  what we want.
- VI turns to tackle with this problem via optimization, and the reach of the family  $\mathcal{Q}$  manages the complexity of this optimization.
- One of the key ideas behind VI is to choose  $\mathcal{Q}$  to be flexible enough to capture a density close to  $p(\mathbf{z} \mid \mathbf{x})$ , but simple enough for efficient optimization.

# Kullback-Leibler Divergence

We need to define what does "closest" mean, VI use Kullback-Leibler Divergence (KL-div) to describe the distance. Now, we review the KL-div:

$$KL(q\|p) := \int q(x) \ln \frac{q(x)}{p(x)} dx,$$

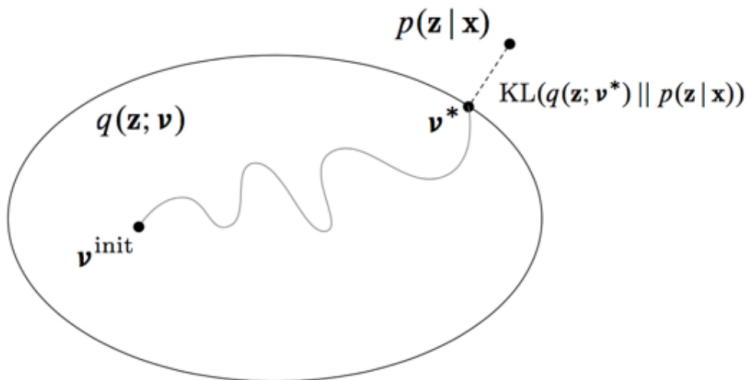
where  $p, q$  are density function. For random variable  $\mathbf{z}$  with density  $q$ , we can see

$$KL(q\|p) = \mathbb{E}_q[\ln q(\mathbf{z})] - \mathbb{E}_q[\ln p(\mathbf{z})].$$

This form is more common here.

# optimization

We see KL-div as distance although it isn't real distance.



# Specific

- We specify a family  $\mathcal{Q}$  of densities over all the latent variables  $\mathbf{z}$  what we want to know.
- Each  $q(\mathbf{z}) \in \mathcal{Q}$  is a candidate approximation and our goal is to find the closest one in KL-div to the exact conditional  $p(\mathbf{z} \mid \mathbf{x})$ .

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \text{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})).$$

- But we don't know  $p(\mathbf{z} \mid \mathbf{x})$ .

# Evidence Lower Bound

We can find that:

$$\begin{aligned}\text{KL}(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x})) \\ &= \mathbb{E}_q[\ln q(\mathbf{z})] - \mathbb{E}_q[\ln p(\mathbf{z} | \mathbf{x})] \\ &= \mathbb{E}_q[\ln q(\mathbf{z})] - \mathbb{E}_q[\ln p(\mathbf{z}, \mathbf{x})] + \ln p(\mathbf{x}),\end{aligned}$$

where all expectations are taken with respect to  $q(\mathbf{z})$ .

Find that  $\ln p(\mathbf{x})$  is a constant with respect to  $q(\mathbf{z})$  and  $p(\mathbf{z}, \mathbf{x})$  is the model what we have known, so we can define the Evidence Lower Bound (ELBO):

$$\text{ELBO}(q) = \mathbb{E}_q[\ln p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_q[\ln q(\mathbf{z})],$$

and maximizing the ELBO is equivalent to minimizing the KL-div.

# Evidence Lower Bound

Notice that the following expression of the evidence:

$$\ln p(\mathbf{x}) = \text{KL}(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x})) + \text{ELBO}(q)$$

we can find that the ELBO is a lower bound of the (ln) evidence,

$$\ln p(\mathbf{x}) \geq \text{ELBO}(q)$$

for any  $q(\mathbf{z})$ , and equal only when  $q(\mathbf{z}) = p(\mathbf{z} | \mathbf{x})$ , this explains the name.

# Review EM Method

Model:

$$p(\mathbf{x}, \mathbf{z}; \theta).$$

Goal:

$$\theta^* = \operatorname{argmax}_{\theta} p(\mathbf{x}; \theta).$$

One of the reasons for using EM algorithm to introduce latent variables is that  $p(\mathbf{x}; \theta)$  is difficult to compute.

# VI and EM

Denote

$$ELBO(q; \theta) = \ln p(\mathbf{x}; \theta) - KL(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x}; \theta)).$$

Assume  $\mathcal{Q}$  is large enough to  $p(\mathbf{z} | \mathbf{x}; \theta) \in \mathcal{Q}$  for any  $\theta$ , then

$$\max_{q \in \mathcal{Q}} ELBO(q; \theta) = ELBO(p(\mathbf{z} | \mathbf{x}; \theta); \theta) = \ln p(\mathbf{x}; \theta).$$

Thus,  $\theta^* = \arg \max_{\theta} \ln p(\mathbf{x}; \theta) = \arg \max_{\theta, q \in \mathcal{Q}} ELBO(q; \theta)$ .



# VI and EM

At k-iteration:

$$q^{(k+1)} = \operatorname{argmax}_{q \in \mathcal{Q}} ELBO(q, \theta^{(k)})$$

$$\theta^{(k+1)} = \operatorname{argmax}_{\theta} ELBO(q^{(k+1)}, \theta)$$

If constraint  $q \in \mathcal{Q}$  is removed, this procedure is completely equivalent to EM algorithm.

We can choose  $\mathcal{Q}$  to trade off between simplicity of calculation and accuracy.

# VI and EM

Denote  $KL^{(k)} = KL(q^{(k)}(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x}; \theta^{(k)}))$ , then

$$\begin{aligned} \ln p(\mathbf{x}; \theta^{(k+1)}) &= ELBO(q^{(k+1)}; \theta^{(k+1)}) + KL^{(k+1)} \\ &\geq ELBO(q^{(k+1)}; \theta^{(k)}) + KL^{(k+1)} \\ &\geq ELBO(q^{(k)}; \theta^{(k)}) + KL^{(k+1)} \\ &= \ln p(\mathbf{x}; \theta^{(k)}) - KL^{(k)} + KL^{(k+1)}. \end{aligned}$$

There is no guarantee that likelihood increases in each iteration, but this is not a completely bad thing. In iterative algorithms, it is not very useful to require accuracy at every step.

# Mean-Field Variational Family

In mean-field variational family  $\mathcal{Q}$ , the latent variables are mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational family is

$$q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j) .$$

# Mean-Field Variational Family

Denote  $\mathbf{z}_{-i} = (z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)$ ,  $q_{-i}(\mathbf{z}_{-i}) = \frac{q(\mathbf{z})}{q_i(z_i)}$ ,

$$\mathbb{E}_i[f(\mathbf{z})] = \int f(\mathbf{z}) q_i(z_i) dz_i,$$

$$\mathbb{E}_{-i}[f(\mathbf{z})] = \int f(\mathbf{z}) q_{-i}(\mathbf{z}_{-i}) d\mathbf{z}_{-i},$$

then,  $\mathbb{E}_q[f(\mathbf{z})] = \mathbb{E}_i \mathbb{E}_{-i}[f(\mathbf{z})]$ .

# Coordinate Ascent Variational Inference

Coordinate Ascent Variational Inference (CAVI) is a method to inference in Mean-Field Variational Family.

Denote density  $\hat{p}(z_j) \propto \exp\{\mathbb{E}_{-j}[\ln p(\mathbf{z}, \mathbf{x})]\}$ . Review

$$\text{ELBO}(q) = \mathbb{E}[\ln p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\ln q(\mathbf{z})].$$

Then, fix  $\mathbf{z}_{-j}$

$$\begin{aligned} & \text{ELBO}(q_j) \\ &= \mathbb{E}_j[\mathbb{E}_{-j}[\ln p(z_j, \mathbf{z}_{-j}, \mathbf{x})]] - \mathbb{E}_j[\mathbb{E}_{-j}[\ln q_j(z_j) + \ln q_{-j}(\mathbf{z}_{-j})]] \\ &= \mathbb{E}_j[\ln \hat{p}(z_j)] - \mathbb{E}_j[\ln q_j(z_j)] + C \\ &= -KL(q_j \| \hat{p}) + C, \end{aligned}$$

$$\Rightarrow q_j^* = \hat{p} \propto \exp\{\mathbb{E}_{-j}[\ln p(\mathbf{z}, \mathbf{x})]\} \propto \exp\{\mathbb{E}_{-j}[\ln p(z_j | \mathbf{z}_{-j}, \mathbf{x})]\}.$$

# Algorithm

---

**Algorithm 1:** Coordinate ascent variational inference (CAVI)

---

**Input:** A model  $p(\mathbf{x}, \mathbf{z})$ , a data set  $\mathbf{x}$

**Output:** A variational density  $q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j)$

**Initialize:** Variational factors  $q_j(z_j)$

**while** *the ELBO has not converged* **do**

**for**  $j \in \{1, \dots, m\}$  **do**

        Set  $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})]\}$

**end**

    Compute  $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$

**end**

**return**  $q(\mathbf{z})$

---

# Gaussian Mixture Model

Data:  $\mathbf{x} = (x_1, \dots, x_n)^T \in R^n$ , latent variable  $\mathbf{z} = (\boldsymbol{\mu}, \mathbf{c})$ ,  
 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T \in R^K$ ,  $\mathbf{c} = (c_1, \dots, c_n)^T \in \{0, 1\}^{n \times K}$ .

$$\mu_k \sim \mathcal{N}(0, \sigma^2), \quad k = 1, \dots, K,$$

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

$$x_i | c_i, \boldsymbol{\mu} \sim \mathcal{N}(c_i^\top \boldsymbol{\mu}, 1), \quad i = 1, \dots, n.$$

Model:  $p(\mathbf{z}, \mathbf{x}) = p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^n p(c_i) p(x_i | c_i, \boldsymbol{\mu})$ .

# VI for Gaussian Mixture Model

Here, the evidence is

$$p(\mathbf{x}) = \int p(\boldsymbol{\mu}) \prod_{i=1}^n \sum_{c_i} p(c_i) p(x_i | c_i, \boldsymbol{\mu}) d\boldsymbol{\mu}.$$

It's a  $K$ -dimensional integral, and the time complexity of numerically evaluation is  $O(K^n)$ , thus we can use  $q(\boldsymbol{\mu}, \mathbf{c})$  to approximate it.

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

We should notice that  $q(\mu_k)$  is the Gaussian distribution and  $q(c_i)$  is the multinomial distribution.



# Variational Density

Update  $c_i$ :

$$\begin{aligned} q^*(c_i) &\propto \exp \{ \mathbb{E}_{-c_i} [\ln p(\mathbf{c}, \boldsymbol{\mu}, \mathbf{x})] \} \\ &\propto \exp \left\{ \mathbb{E}_{-c_i} [\ln p(\boldsymbol{\mu})] + \sum_{j=1}^n \mathbb{E}_{-c_i} [\ln p(c_j)] \right. \\ &\quad \left. + \sum_{j=1}^n \mathbb{E}_{-c_i} [\ln p(x_j | c_j, \boldsymbol{\mu})] \right\} \\ &\propto \exp \{ \ln p(c_i) + \mathbb{E}_{-c_i} [\ln p(x_i | c_i, \boldsymbol{\mu})] \}, \end{aligned}$$

where  $p(c_i) = \frac{1}{K}$ ,  $p(x_i | c_i, \boldsymbol{\mu}) = \prod_{k=1}^K p(x_i | \mu_k)^{c_{ik}}$  and  $p(x_i | \mu_k) \sim N(\mu_k, 1)$ .

# Variational Density

Thus, we can see:

$$\begin{aligned} q^*(c_i) &\propto \exp \left\{ \mathbb{E}_{-c_i} \left[ \sum_{k=1}^K c_{ik} \left( -\frac{1}{2} (x_i - \mu_k)^2 \right) \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \sum_{k=1}^K c_{ik} \mathbb{E}_{\mu_k} \left[ (\mu_k - x_i)^2 \right] \right\} \end{aligned}$$

Assuming  $\mu_k \sim N(m_k, s_k^2) \Rightarrow \mathbb{E}_{\mu_k} [(\mu_k - x_i)^2] = (m_k - x_i)^2 + s_k^2$

$$q^*(c_i) \propto \exp \left\{ -\frac{1}{2} \sum_{k=1}^K c_{ij} \left( m_k^2 + s_k^2 - 2m_k x_i \right) \right\}$$

$$\text{thus, } q^*(c_i) |_{c_{ik=1}} := \varphi_{ik} \propto \exp \left\{ m_j x_i - \frac{m_k^2 + s_k^2}{2} \right\}.$$

# Variational Density

Update  $\mu_k$

$$\begin{aligned} q^*(\mu_k) &\propto \exp \left\{ \ln p(\mu_k) + \sum_{i=1}^n \mathbb{E}_{-\mu_i} \left[ \sum_{k=1}^K c_{ik} \ln p(x_i | \mu_k) \right] \right\} \\ &\propto \exp \left\{ -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^n \mathbb{E}_{-\mu_k} [c_{ik} \ln p(x_i | \mu_k)] \right\} \end{aligned}$$

$$\text{where, } \mathbb{E}_{-\mu_k} [c_{ik} \ln p(x_i | \mu_k)] \propto \varphi_{ik} \cdot \left[ -\frac{1}{2} (x_i - \mu_k)^2 \right]$$

$$\text{then, } q^*(\mu_k) \propto \exp \left\{ -\frac{\mu_k^2}{2\sigma^2} - \frac{1}{2} \left( \sum_{i=1}^n \varphi_{ik} \right) \mu_k^2 + \left( \sum_{i=1}^n \varphi_{ik} x_i \right) \mu_k \right\}$$

$$\text{thus, } s_k^2 = \left( \frac{1}{\sigma^2} + \sum_{i=1}^n \varphi_{ik} \right)^{-1}, \quad m_k^* = s_k^2 \left( \sum_{i=1}^n \varphi_{ik} x_i \right).$$

# CAVI for GMM

---

**Algorithm 2:** CAVI for a Gaussian mixture model

---

**Input:** Data  $x_{1:n}$ , number of components  $K$ , prior variance of component means  $\sigma^2$ **Output:** Variational densities  $q(\mu_k; m_k, s_k^2)$  (Gaussian) and  $q(c_i; \varphi_i)$  ( $K$ -categorical)**Initialize:** Variational parameters  $\mathbf{m} = m_{1:K}$ ,  $\mathbf{s}^2 = s_{1:K}^2$ , and  $\varphi = \varphi_{1:n}$ **while** the ELBO has not converged **do**    **for**  $i \in \{1, \dots, n\}$  **do**        Set  $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2] x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\}$     **end**    **for**  $k \in \{1, \dots, K\}$  **do**

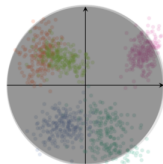
$$\text{Set } m_k \leftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

$$\text{Set } s_k^2 \leftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

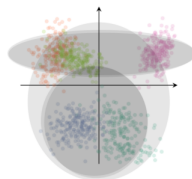
**end**    Compute ELBO( $\mathbf{m}, \mathbf{s}^2, \varphi$ )**end****return**  $q(\mathbf{m}, \mathbf{s}^2, \varphi)$ 

---

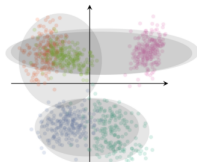
# CAVI for GMM



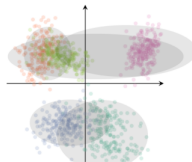
Initialization



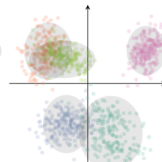
Iteration 20



Iteration 28



Iteration 35



Iteration 50

# Application

One network model groups nodes into modules or communities, with different densities of intra- and inter-connectivity for nodes in the same or different modules.

There is a computationally efficient Bayesian framework for inferring the number of modules, model parameters, and module assignments for such a model, implemented using a variational technique developed only in the past decade.

# Model

- We specify an N-node network by its adjacency matrix  $A$ , where  $A_{ij} = 1$  if there is an edge between nodes  $i$  and  $j$ , and  $A_{ij} = 0$  otherwise.
- Define  $K$  to be the number of modules and  $\sigma_i \in \{1, \dots, K\}$  to be the module which the  $i$ -th node belongs to.
- Assume  $\sigma_i$  correspond to multinomial distribution with probability  $\pi = (\pi_1, \dots, \pi_K)$ , and  $A_{ij}$  corresponds to Bernoulli distribution with probability  $v_c$  when  $\sigma_i = \sigma_j$  or  $v_d$  when  $\sigma_i \neq \sigma_j$ .

Conclusion: our data is  $A$ , latent variables are  $\sigma$ , parameters are  $\mathbf{v}, \pi, K$ .

# Model

Let's see the joint density function of model.

$$\begin{aligned} & p(A, \sigma, \pi, \mathbf{v}, K) \\ &= p(A \mid \sigma, \pi, \mathbf{v}, K) p(\sigma \mid \pi, \mathbf{v}, K) p(\mathbf{v} \mid \pi, K) p(\pi \mid K) p(K) \\ &= p(A \mid \sigma, \mathbf{v}, K) p(\sigma \mid \pi, K) p(\mathbf{v}) p(\pi \mid K) p(K). \end{aligned}$$

- $p(\mathbf{v}), p(\pi \mid K), p(K)$  are prior of parameters.
- $p(\sigma \mid \pi, K) = \prod_{j=1}^K \pi_j^{n_j}$ .

where  $n_j = \sum_{i=1}^N I_{\{\sigma_i=j\}}$  is the occupation number of the  $j$ -th module.



# Model

$$p(A \mid \sigma, \mathbf{v}, K) = v_c^{c_+} (1 - v_c)^{c_-} v_d^{d_+} (1 - v_d)^{d_-},$$

where

- $c_+ = \sum_{i>j} A_{ij} I_{\{\sigma_i=\sigma_j\}}$   
is the number of edges contained within communities.
- $c_- = \sum_{i>j} (1 - A_{ij}) I_{\{\sigma_i=\sigma_j\}}$   
is the number of non-edges contained within communities.
- $d_+ = \sum_{i>j} A_{ij} I_{\{\sigma_i \neq \sigma_j\}}$   
is the number of edges between different communities.
- $d_- = \sum_{i>j} (1 - A_{ij}) I_{\{\sigma_i \neq \sigma_j\}}$   
is the number of non-edges between different communities.

# Example

Each of the 115 nodes represents an individual team and each of the 613 edges represents a game played between the nodes joined. The algorithm correctly identifies the presence of the 12 conferences which comprise the schedule, where teams tend to play more games within than between conferences, making most modules assortative.

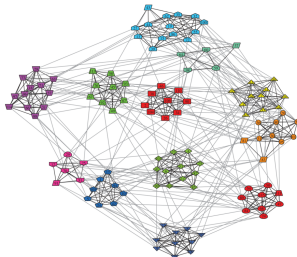


Figure: 2000 NCAA American football.

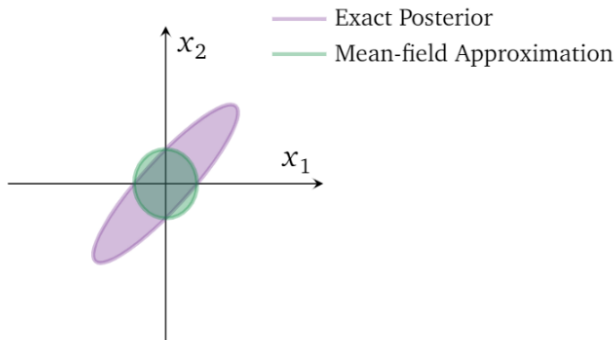
# Discussion

- The idea of variational inference is to construct an analytical approximation to the posterior probability of the set of unobserved variables (parameters and latent variables), given the data. As in other Bayesian methods — but unlike e.g. in expectation maximization (EM) or other maximum likelihood methods — both types of unobserved variables (i.e. parameters and latent variables) are treated the same, i.e. as random variables.
- Estimates for the variables can then be derived in the standard Bayesian ways, e.g. calculating the mean of the distribution to get a single point estimate or deriving a credible interval, highest density region, etc.

# Discussion

- "Analytical approximation" means that a formula can be written down for the posterior distribution. The formula generally consists of a product of well-known probability distributions, each of which factorizes over a set of unobserved variables (i.e. it is conditionally independent of the other variables, given the observed data). This formula is not the true posterior distribution, but an approximation to it;
- However, it cannot capture correlation between them.

# Discussion






## Compared with EM

- VI is often compared with expectation maximization (EM). The actual numerical procedure is quite similar, in that both are alternating iterative procedures that successively converge on optimum parameter values. The initial steps to derive the respective procedures are also vaguely similar, both starting out with formulas for probability densities and both involving significant amounts of mathematical manipulations.

# Differences

- EM computes point estimates of posterior distribution of those random variables that can be categorized as "parameters", but only estimates of the actual posterior distributions of the latent variables. The point estimates computed are the modes of these parameters; no other information is available.
- VI, on the other hand, computes estimates of the actual posterior distribution of all variables, both parameters and latent variables.
- Concomitant with this, the parameters computed in VI do not have the same significance as those in EM.

# References I

-  Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017).  
Variational inference: A review for statisticians.  
*Journal of the American Statistical Association*,  
112(518):859–877.
-  Gupta, M. R. and Chen, Y. (2011).  
*Theory and use of the EM algorithm*.  
Now Publishers Inc.
-  Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W.  
(2013).  
Stochastic variational inference.  
*ArXiv*, abs/1206.7051.



# References II



Hofman, J. M. and Wiggins, C. H. (2008).  
Bayesian approach to network modularity.  
*Phys. Rev. Lett.*, 100:258701.