General EM (cont'd)

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 10

general EM

complete dataset with latent variables

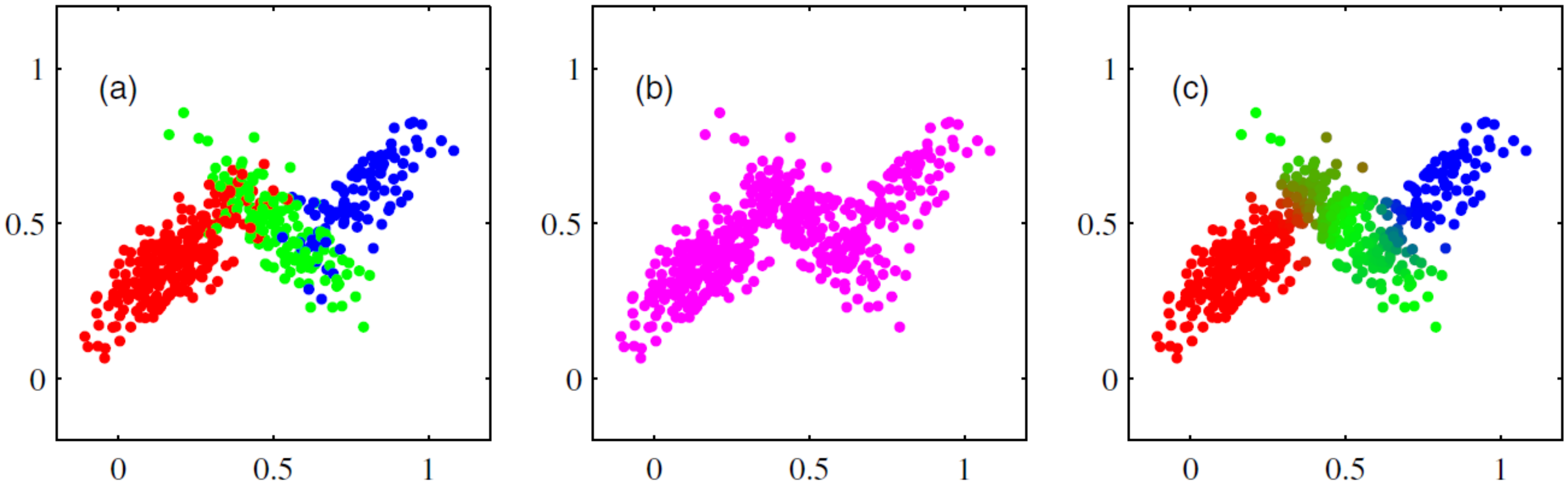
- Suppose $\mathbf{X}$ is the data matrix, and $\mathbf{Z}$ the corresponding latent variables (assumed to be discrete). Then

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

- $\{\mathbf{X}, \mathbf{Z}\}$ is called the complete data set; $\mathbf{X}$ is incomplete
- In practice, we are not given the complete data set; the only way we estimate $\mathbf{Z}$ is by the posterior

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

complete dataset with latent variables



complete data set with both $\{\mathbf{X}, \mathbf{Z}\}$

incomplete data set with only $\mathbf{X}$

$\mathbf{X}$ with the posterior $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$

Image taken from [Bi]

general EM

(E-step): expectation

1. for fixed parameters, find $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
2. calculate the expectation $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

$\mathbb{E}_{p(Z|X,\theta^{old})} \ln p(X, Z|\theta)$

(M-step): maximization

solve for $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$

GMM revisited

- The likelihood of the complete data set is

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{z_{nk}}$$

$p(z|\pi)\,p(x|z,\mu,\Sigma)$

- Taking logarithm yields

$$\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

$$\mathbb{E}\ln p(X,z|\mu,\Sigma,\pi) = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left(\ln\pi_k + \ln\mathcal{N}(x_n|\mu_k,\Sigma_k)\right)$$

$p(z_n = e_k | x_n, \mu, \Sigma, \pi) \propto \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)$

GMM revisited

- By Bayes' theorem, $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}$
- What does this imply for each $\boldsymbol{z}_n$?

$$p(z_n | x_n, \mu, \Sigma, \pi) \propto \prod_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}}$$

- Under the posterior distribution, the conditional expectation of the indicator z_{nk} is given by

$$\mathbb{E}[z_{nk}] = 1 \cdot p(z_{nk} = 1 | x_n, \mu, \Sigma, \pi) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk})$$

$$\| $$

$$p(z_n = e_k | x_n, \mu, \Sigma, \pi)$$

Eq (9.39) in [Bi] is not very straight-forward. Take care when reading it.

GMM revisited

- The expectation of the complete log likelihood is thus

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}$$

$\sim p(z|x, \mu, \Sigma, \pi)$

- Maximizing the above yields the same parameters $\left\{ \begin{array}{l} \boldsymbol{\mu}_k \\ \boldsymbol{\Sigma}_k \\ \boldsymbol{\pi}_k \end{array}\right.$ as before.

general EM (the general view)

- The expectation-maximization algorithm, or EM algorithm, is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables.
- Consider

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

- Assuming $\mathbf{Z}$ is discrete and dealing with $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is easier than $p(\mathbf{X}|\boldsymbol{\theta})$.
- Introduce a distribution $q(\mathbf{Z})$ over the latent variables.

general EM

KL divergence from q to p

- Claim $$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|\underbrace{p}_{p(z|X,\theta)})$$

where

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$= \mathbb{E}_{q(z)} \ln \left(\frac{q(z)}{p(z|X,\theta)} \right)$$

general EM

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

- To prove the above claim, first note that

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta})$$

← since $p(x, z|\theta) = p(z|x, \theta) \cdot p(x|\theta)$

- Then $\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$ $= ?$

$$= \sum_z q(z) \left(\ln p(z|x, \theta) + \ln p(x|\theta) - \ln q(z) \right)$$

$$= \sum_z q(z) \ln \frac{p(z|x, \theta)}{q(z)} + \left(\sum_z q(z) \right) \ln p(x|\theta)$$

$$= -\mathrm{KL}(q\|p) + \ln p(x|\theta)$$

general EM

- Now we have proved that $\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$
- Fact: $\mathrm{KL}(q \parallel p) \geq 0$
- Therefore, $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound of $\ln p(\mathbf{X}|\boldsymbol{\theta})$

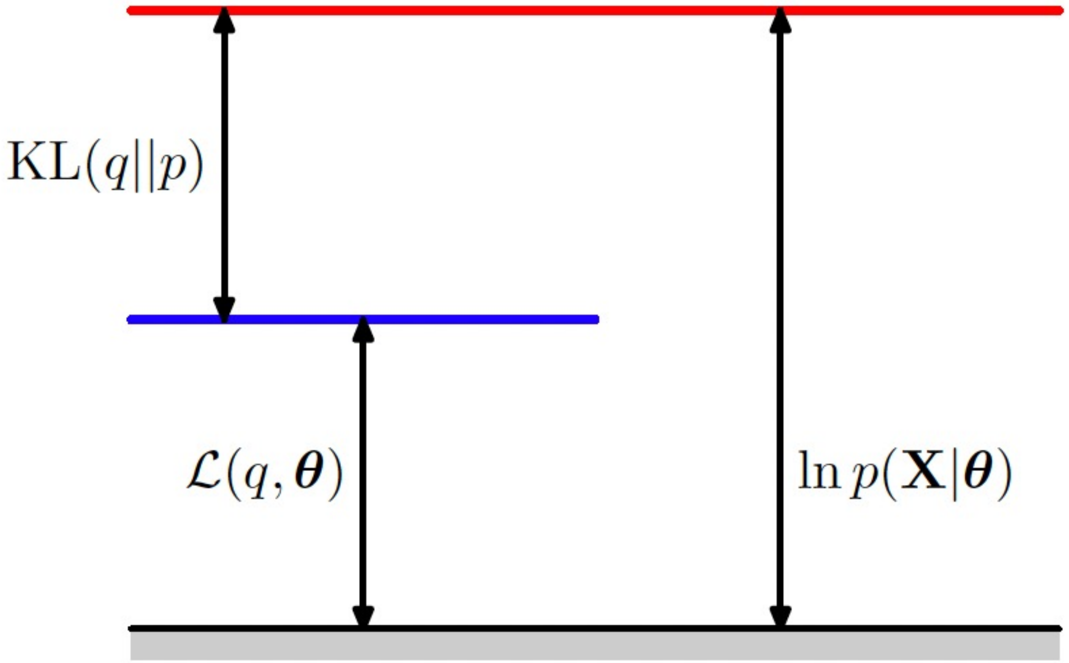
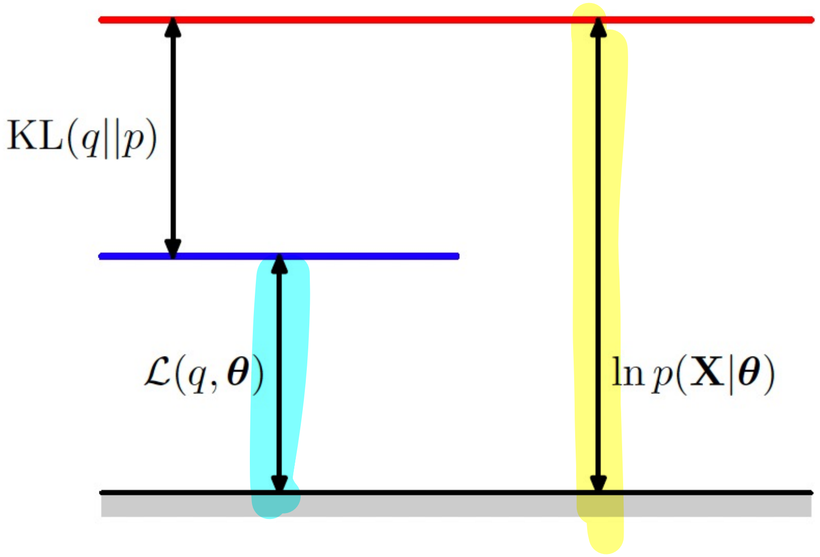


Image taken from [Bi]

general EM



(E-step): expectation

maximize $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ with respect to q

This q will be pushed to $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$!!!

Image taken from [Bi]

general EM

Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values $\boldsymbol{\theta}^{\text{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

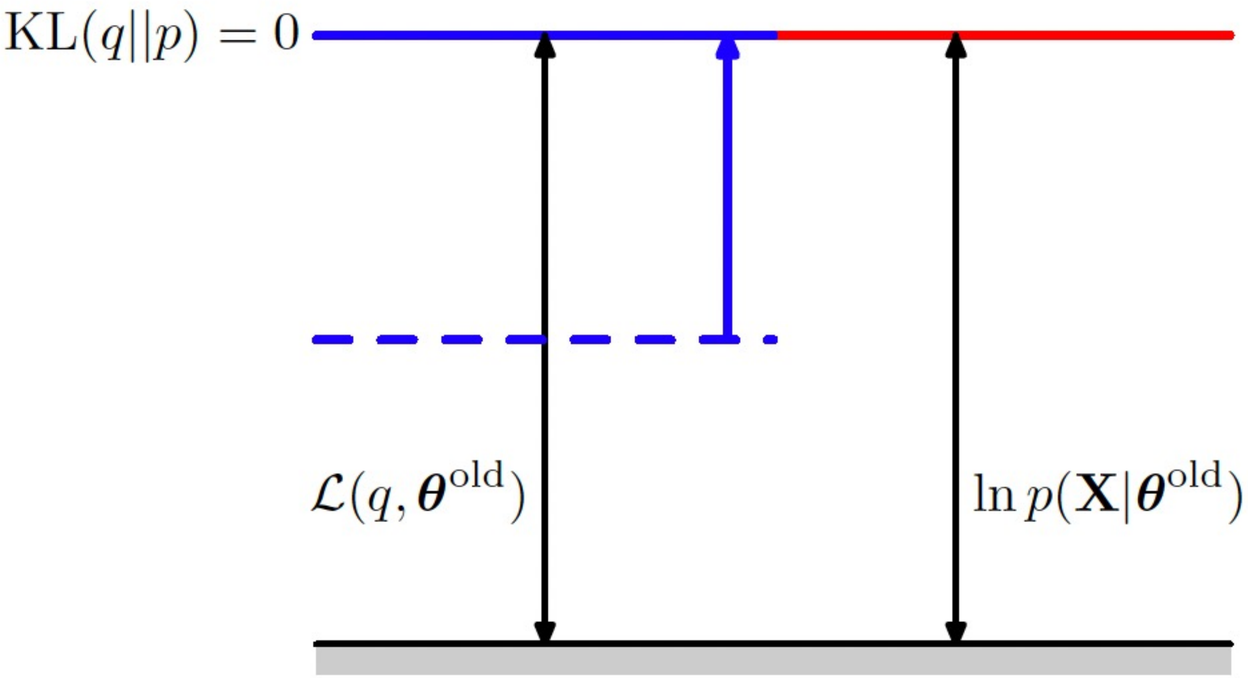


Image taken from [Bi]

general EM

(E-step): expectation

maximize $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ with respect to q

(M-step): maximization

fix q and maximize $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$

general EM

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

- After the E Step, $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$, and therefore the lower bound takes the form

$$\begin{aligned}
\mathcal{L}(q,\boldsymbol{\theta}) &= \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\underline{\boldsymbol{\theta}^{\text{old}}}) \ln p(\mathbf{Z}|\mathbf{X},\underline{\boldsymbol{\theta}^{\text{old}}}) \\
&= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \text{const}
\end{aligned}$$

$\mathbb{E}_{p(z|x,\theta^{\text{old}})} \ln p(x,z|\theta)$

general EM

Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.

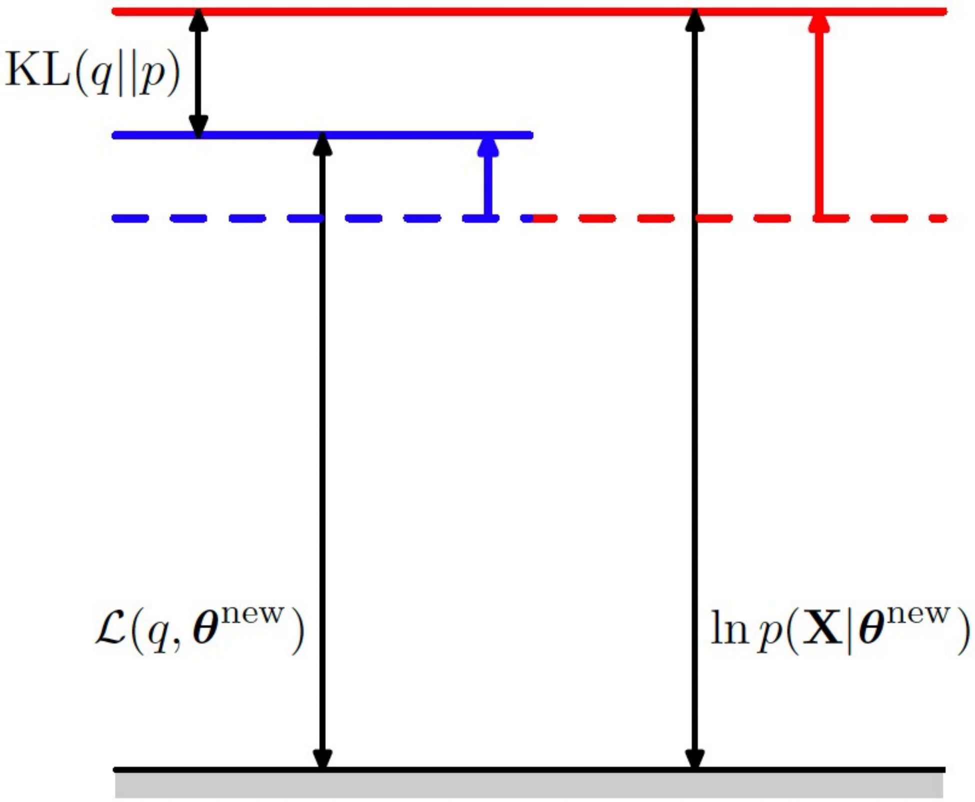
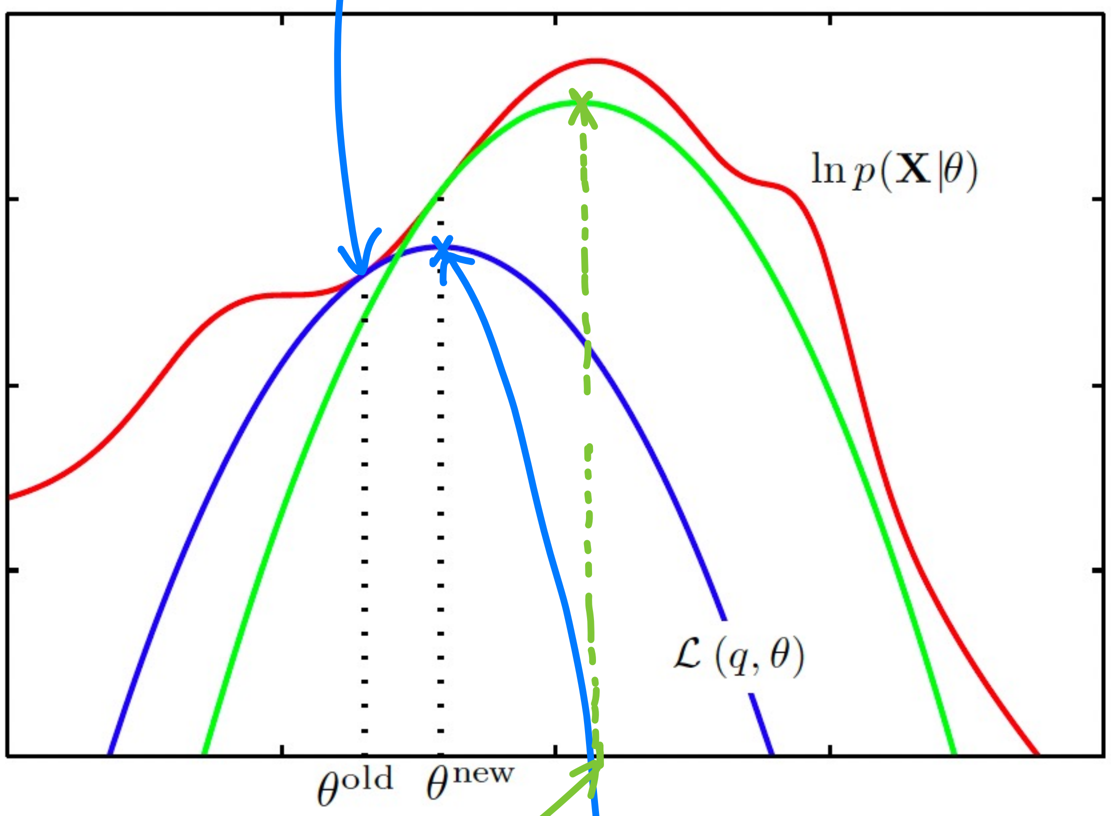


Image taken from [Bi]

general EM

The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.

E-step: no gap between $\ln p(X|\theta^{old})$ and $\mathcal{L}(q,\theta^{old})$



next θ^{new}

M-step: maximize $\mathcal{L}(q,\theta)$

Image taken from [Bi]

spectral clustering

no need to have exact locations

- For both K-means and EM, we use the location of the data instances $\{\boldsymbol{x}_n\}_{n=1}^{N}$
- However, in a distance-based method, we make decision only based on the relative locations
- Goal of this section: we want to get low dimensional features disregarding the original locations

guide for projection

- Given $\boldsymbol{x}_n \in \mathbb{R}^d$, $n = 1, \cdots, N$.
- Suppose we want to look at low-dimension projection $\boldsymbol{z}_n \in \mathbb{R}^k$, $n = 1, \cdots, N$.

- Suppose further that we are given a similarity score for each pair of instances
    - $w_{nm} = w_{mn}$ is the similarity between x_n and x_m (assume $w_{nn} = 0$)

- The goal is to minimize $\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \| z_n - z_m \|^2 w_{nm}$

case $k = 1$: $z_n \in \mathbb{R}$

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}|z_n - z_m|^2 w_{nm} = ?$$

(Let $d_n := \sum_{m=1}^{N} w_{nm} = \sum_{m=1}^{N} w_{mn}$)

$$= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} z_n^2 w_{nm} + z_m^2 w_{nm} - 2 z_n z_m w_{nm}$$

$$= \frac{1}{2}\left(\sum_{n=1}^{N} z_n^2 \underbrace{\sum_{m=1}^{N} w_{nm}}_{d_n} + \sum_{m=1}^{N} z_m^2 \underbrace{\sum_{n=1}^{N} w_{nm}}_{d_m} - 2\sum_{n=1}^{N}\sum_{m=1}^{N} z_n z_m w_{nm}\right)$$

$$= \sum_{n=1}^{N} z_n d_n z_n - \sum_{n=1}^{N}\sum_{m=1}^{N} z_n w_{nm} z_m$$

$$= z^T D z - z^T W z$$

where $D = \mathrm{diag}(d_1, \cdots, d_N)$, $W = (w_{nm})_{n,m=1}^{N}$

case $k = 1$: $z_n \in \mathbb{R}$

- Now that $\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} |z_n - z_m|^2 w_{nm} = z^T(\boldsymbol{D} - \boldsymbol{W})z$ where
  - $\boldsymbol{W}$ is the symmetric matrix whose (n, m)-th entry is w_{nm} (let's define $w_{nn} = 0$ for each n)
  - $\boldsymbol{D}$ is the diagonal matrix with $D_{nn} = \sum_{m \neq n} w_{nm}$
  - Let $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, we call $\boldsymbol{L}$ the graph Laplacian of the similarity graph of our data. (just a remark)

eigenvector of Laplacian

- Now we need to minimize $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{z}$.
- $\boldsymbol{z} = (z_1, \cdots, z_N)^{\mathrm{T}}$ should be an eigenvector of $\boldsymbol{L}$.
  - Note that there is a trivial eigenvector $\boldsymbol{c} = \frac{1}{\sqrt{N}}(1,1,\cdots,1)^{\mathrm{T}}$
    - $\boldsymbol{L}\boldsymbol{c} = \boldsymbol{D}\boldsymbol{c} - \boldsymbol{W}\boldsymbol{c}$, whose n-th entry is $\frac{1}{\sqrt{N}}\left[D_{nn} - \sum_{m \neq n} w_{nm}\right] = 0$
    - but this is not interesting because this means z_n is the same for different data point $\boldsymbol{x}_n$.
  - We should have some $\boldsymbol{z}$ that is orthogonal to the above trivial choice.
  - Therefore, we take $\boldsymbol{z}$ to be the eigenvector corresponding to the second smallest eigenvalue.

Laplacian eigenmap

- The map that maps $\{\boldsymbol{x}_n\}_{n=1}^N$ to $\{z_n\}_{n=1}^N$ is called the Laplacian eigenmap where $\boldsymbol{z} = (z_1, \cdots, z_N)^{\mathrm{T}}$ is the eigenvector corresponding to the second smallest eigenvalue of L.
- In case $k > 1$ (more than one feature in $\boldsymbol{z}$), we need to take the next features in $\boldsymbol{z}$ to be the eigenvectors corresponding to the next smallest eigenvalues.
- Specifically, we will get a feature matrix $\boldsymbol{Z}$, whose size is N-by-k, and whose k columns are given by the eigenvectors corresponding to the 2nd to the $(k+1)$-th smallest eigenvalues of $\boldsymbol{L}$.

spectral clustering

- Once we have the features $\{\boldsymbol{z}_n\}_{n=1}^N$, we can do clustering on the features instead of on the original data points $\{\boldsymbol{x}_n\}_{n=1}^N$.

- For instance, we can apply K-means on $\{\boldsymbol{z}_n\}_{n=1}^N$.

normalized Laplacian

- In practice, instead of working with the Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, we may want to work with a normalized version of Laplacian, e.g.
  - $\boldsymbol{L}_{\mathrm{rw}} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}$
  - $\boldsymbol{L}_{\mathrm{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$

choice of similarity score

- We only consider similarities locally (in a neighborhood):
  - either set $w_{nm} = 0$ for $\| \boldsymbol{x}_n - \boldsymbol{x}_m \| > \epsilon$ for some preset threshold ϵ;
  - or using k-NN: $w_{nm} \neq 0$ if only if $\boldsymbol{x}_m$ is among the k nearest neighbors of $\boldsymbol{x}_m$ or vice versa.
- For $w_{nm} \neq 0$, one possible similarity score is to set

$$w_{nm} = \exp\left(-\frac{\| \boldsymbol{x}_n - \boldsymbol{x}_m \|^2}{2\sigma^2}\right)$$

example: three concentric clusters

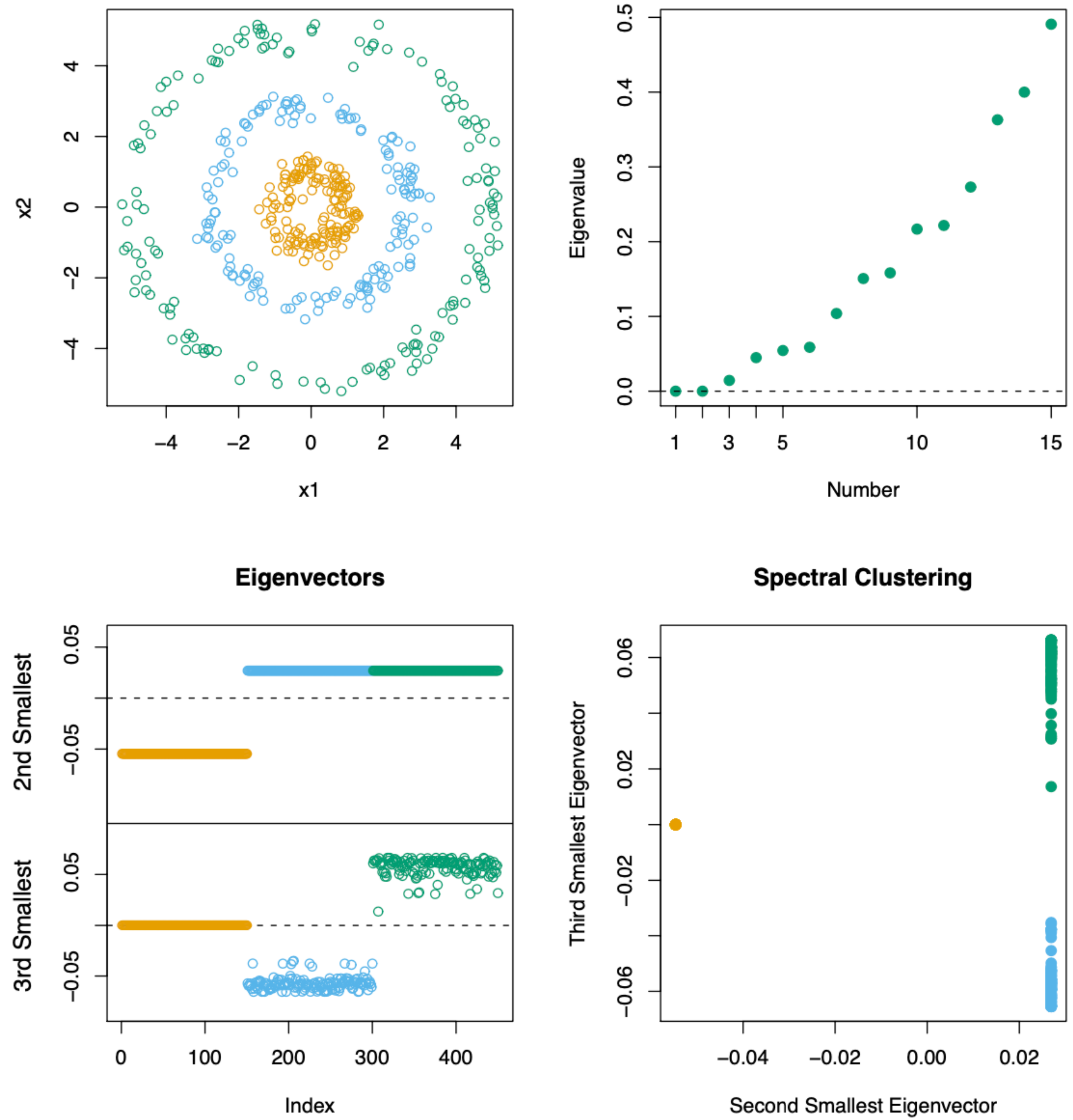


Image taken from [HaTF]

Questions?

Reference

- *K-means:*
  - *[Al] Ch.7.3*
  - *[HaTF] Ch.13.2.1*
  - *[Bi] Ch.9.1*
- *EM:*
  - *[Al] Ch.7.2, 7.4*
  - *[HaTF] Ch.13.2.3*
  - *[Bi] Ch.9.2-9.4*
- *Spectral clustering:*
  - *[Al] Ch.6.12 7.7*
  - *[HaTF] Ch.14.5.3*