Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 8: MIXTURE MODELS

Reference:
1. "Pattern Recognition and Machine Learning" Chapter 9: Mixture Models and EM
2. Estimating Gaussian Mixture Densities with EM – A Tutorial
3. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and HMMs.

GOAL

We want to identify which data came from which source.

In probabilistic modeling words

"Evaluate the **posterior distribution** *p*(Z /X) of the latent variables **Z** (which source) given the observed (visible) data variables **X**, and the evaluation of expectations computed with respect to this distribution."

Strategy for **parametric models**

Estimate *p*(Z|X) by estimating it's parameter θ

Condition we work on: The data are **independently** generated by **sources** of data (distribution functions) and there are no (or ignorable) dependency between the sources. *Mixture models*

Estimating it's parameter θ by evaluating the **log likelihood** $\mathbf{p(x|\boldsymbol{\theta})}$

A method the solve log likelihood function is using **Expectation Maximization**

MIXTURE OF GAUSSIANS:

- Gaussian Mixture Distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Where $p(z_k = 1) = \pi_k$: prior prob. of $z_k = 1$

$$N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

- Posterior probability of z_k (responsibility) once we observed a point **x**

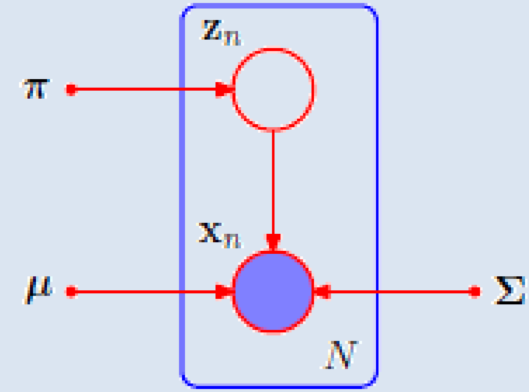
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Where $\theta_{\boldsymbol{k}} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ and $\boldsymbol{\theta}=\{\theta_{\mathbf{1}}, \dots, \theta_K\}$

- Log Likelihood function

$$\ln p\,(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

Mixture of Gaussians Model



$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$

- N number of D dimension data **X**
  $\boldsymbol{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \mathbf{x} = \{x_1, \dots, x_D\}$

- N number of K dim. class variable **Z**
  $\boldsymbol{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_N\}, \boldsymbol{z} = \{z_1, \dots, z_K\}$

ESTIMATING PARAMETER DIRECTLY EVALUATING THE LOG-LIKELIHOOD FUNCTION

$$\ln p(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

Chain rule :
$D\{f(g(x))\} = f'(g(x))g'(x)$

Set derivative of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t means μ_k of the Gaussian components to zero

$$\mu_k = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)\mathbf{x}_n$$

Set derivative of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t Σ_k of the Gaussian components to zero.

$$\Sigma_k = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

Take into account constraint $\sum_k \pi_k = 1$ Set derivative of modified log likelihood w.r.t π_k of the Gaussian components to zero

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda(\sum_k \pi_k - 1)$$

$$\pi_k = \frac{\sum_n \gamma(z_k)}{N}$$

These results are do not constitute a closed-form solution from the parameters of the mixture model because the responsibilities $\gamma(z_k)$ depend on those parameters in a complex way.

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

But can still be used as iterative scheme (i.e. EM) to finding solutions

MIXTURE OF GAUSSIANS: EM FOR GAUSSIAN MIXTURES SUMMARY

1. Initialize $\{\mu_k, \Sigma_k, \pi_k\}$ and evaluate log-likeihood
2. E-Step: Evaluate responsibilities $\gamma(z_k)$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3. M-Step: Re-estimate parameters $\boldsymbol{\theta}$, using current responsibilities $\gamma(z_k)$

$$\mu_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)\mathbf{x}_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k^{\text{new}} = \frac{\sum_n \gamma(z_k)}{N}$$

Maximize log-likelihood

$$\ln p(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

4. Evaluate log-likelihood $\ln p(\boldsymbol{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and check for convergence of either the parameters or the log likelihood.

   If convergence criterion is not satisfied return to step 2.

AN ALTERNATIVE VIEW OF EM: LATENT VARIABLES

- Let X observed data, Z latent variables, parameters.
- Goal: maximize marginal log-likelihood of observed data

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

Summation over the latent variables appears inside the logarithm

Log-sum prevents the logarithm from acting directly on the joint distribution, resulting on complicated expressions for the maximum log likelihood solution.

WHAT IF WE KNOW THE VALUE OF Z?

If we know **complete data** $\{\mathbf{X}, \mathbf{Z}\}$

the complete-data log likelihood function is straight forward.

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

z_{nk} denote $\mathrm{k^{th}}$ component of $\boldsymbol{z}_n$

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

Maximizes trivially in closed form

In practice we don't know $\mathbf{Z}$ but only $\mathbf{X}$

the complete-data log likelihood function is straight forward.

Use expected value under the posterior distribution of the latent variable

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

which we know

LOG-SUM RESOLVED

- State of knowledge about the latent variable $\mathbf{Z}$ known only through Posterior distribution of the latent variable

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

- Expectation of the complete-data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$, using the posterior distributor of the latent variables $p(\mathbf{Z} \,|\mathbf{X}, \boldsymbol{\theta})$

$$\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{z} p(\mathbf{Z} \,|\mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}}) \ln p(\mathbf{X}, \mathbf{Z} \,|\boldsymbol{\theta})$$

Logarithm acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z} \,|\boldsymbol{\theta})$ so maximization is tractable unlike

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

AN ALTERNATIVE VIEW OF EM: GENERAL EM ALGORITHM

Given a joint distribution $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$ over observed variables **X** and latent variables **Z**, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X} \mid \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Initialization: Choose initial set of parameters $\boldsymbol{\theta}^{old}$
2. E-step: use current parameters $\boldsymbol{\theta}^{old}$ to compute.

$$p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}})$$

3. M-step: determine θ^{new} by maximizing $Q(\theta, \theta^{old})$

$$\boldsymbol{\theta}^{\mathbf{new}} = \arg_{\boldsymbol{\theta}} \max Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}).$$

Where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$$

Logarithm acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$ so maximization is tractable

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

4. Check convergence either the log likelihood or the parameter values : stop, or $\boldsymbol{\theta}^{old} \longleftarrow \boldsymbol{\theta}^{new}$ and go to step 2.

Consider the expectation, with respect to the posterior distribution of the latent variables, of the complete-data log likelihood

Posterior distribution :

since $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k^{z_k}$

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

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \propto p(\mathbf{Z}|\boldsymbol{\theta})p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$$
$$= \prod_{n=1}^{N}\prod_{k=1}^{K} [\pi_k N(\mathbf{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k\,]^{z_{nk}}$$

Factorizes over *n*: under posterior distribution $\{z_n\}$ **are independent**

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

Expected value of the indicator variable z_{nk} under this posterior distribution

Expected value of the indicator variable

$$\mathbb{E}[z_{nk}] = \frac{\sum_{\mathbf{z}_n} z_{nk} \prod_{k'} [\pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})]^{z_{nk'}}}{\sum_{\mathbf{z}_n} \prod_{j} [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}}$$

- $\mathrm{E}_{\mathbf{z_n}}[\boldsymbol{z_{nk}}] = \sum_{\mathbf{z_n}} \boldsymbol{z_{nk}}\, p(\mathbf{z_n}|\mathbf{x}_n, \boldsymbol{\theta})$
- $p(\mathbf{z_n}|\mathbf{x}_n, \boldsymbol{\theta}) \propto \prod_{k=1}^{K} [\pi_k N(\mathbf{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k]^{z_{nk}}$

$$= \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \equiv \gamma(z_k)$$

responsibility of component *k* for data point $\mathbf{x}_n$

Expected value of the complete-data log likelihood function

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

Responsibility of component *k* for data point $\mathbf{x_n}$

Note:

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

AN ALTERNATIVE VIEW OF EM: GENERAL EM ALGORITHM FOR GAUSSIAN MIXTURE MODEL

Given a joint distribution $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ over observed variables X and latent variables Z, governed by parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, the goal is to maximize the likelihood function $p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.

1. Initialization: Choose initial set of parameters $\{\boldsymbol{\pi}^{old}, \boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{\mathrm{old}}\}$
2. E-step: use current parameters $\{\boldsymbol{\pi}^{old}, \boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{\mathrm{old}}\}$ to compute.

$p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}})$

$$E_z[z_{nk}] = \gamma(z_{nk}) \equiv \frac{\pi_k N(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3. M-step: determine θ^{new} by maximizing $Q(\theta, \theta^{old})$

Closed form solution

$$\{\boldsymbol{\pi}_k{}^{new}, \boldsymbol{\mu}_k{}^{new}, \boldsymbol{\Sigma}_k{}^{new}\} = \arg_{\{\boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}} \max \mathrm{E}_{\mathrm{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)].$$

$$= \arg_{\{\boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}} \max \sum_{n=1}^{N} \gamma(z_{nk}) \{\ln \pi_k + \ln N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

$$\boldsymbol{\theta}^{\mathbf{new}} = \arg_{\boldsymbol{\theta}} \max \sum_z p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}).$$

$$\mu_k^{\mathrm{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k) \mathbf{x}_n$$

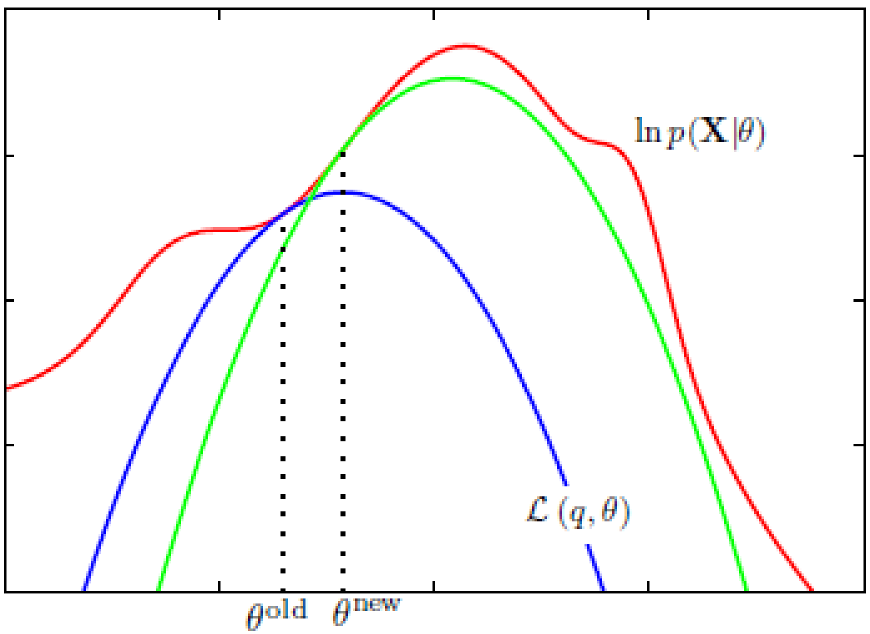
$$\Sigma_k^{\mathrm{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)$$

$$\pi_k^{\mathrm{new}} = \frac{\sum_n \gamma(z_k)}{N}$$

4. Check convergence either the log likelihood or the parameter values : stop, or $\boldsymbol{\theta}^{old} \longleftarrow \boldsymbol{\theta}^{new}$ and go to step 2.

THE EM ALGORITHM IN GENERAL: PICTURE IN PARAMETER SPACE

- E-step resets bound $\mathrm{L}(q|\theta)$ on $\ln p(\boldsymbol{X}|\theta)$ at $\theta = \theta^{old}$, it is
  - tight at $\theta = \theta^{old}$,
  - tangential at $\theta = \theta^{old}$,
  - convex (easy) in θ for exponential family mixture components



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.

THE EM ALGORITHM IN GENERAL: THE EM BOUND

- $\mathrm{L}(\mathrm{q}|\theta)$ is a lower bound on the data log-likelihood
  - $-\mathrm{L}(\mathrm{q}|\theta)$ known as variational free-energy

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

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathrm{L}(\mathrm{q}, \boldsymbol{\theta})$$

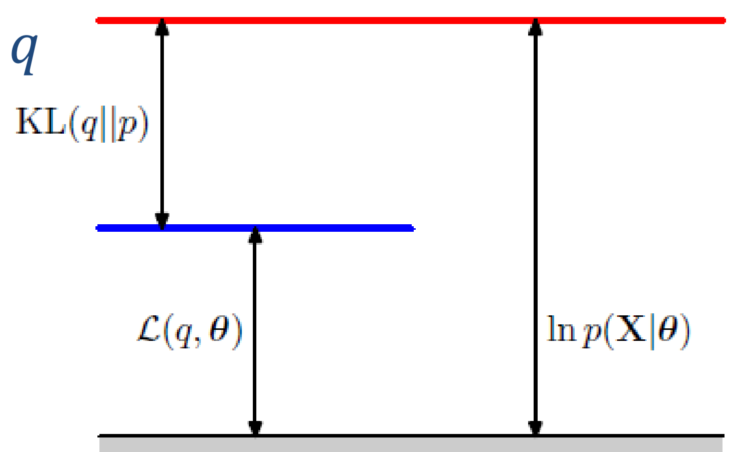
$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \sum_{\boldsymbol{X}} q(\boldsymbol{\theta})(\frac{p(\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})})$$

Jensen's inequality

$$\geq \sum_{\boldsymbol{X}} \ln q(\boldsymbol{\theta}) \left(\frac{p(\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)$$

$$= \mathrm{L}(\mathrm{q}, \boldsymbol{\theta})$$

- The EM algorithm performs coordinate ascent on L
  - E-step maximizes L w.r.t. q for fixed θ
  - M-step maximizes L w.r.t. for fixed q

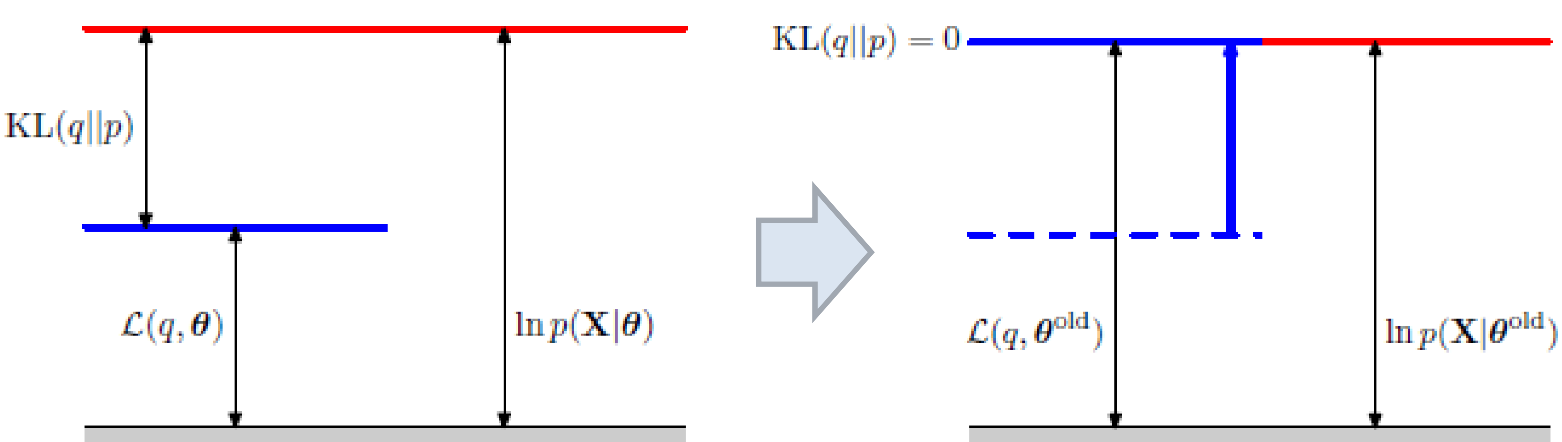


THE EM ALGORITHM IN GENERAL: THE E-STEP

- E-step maximizes $L(q|\theta)$ w.r.t. q for fixed θ

$$\mathcal{L}(q,\theta) \quad = \quad \ln p(\mathbf{X}|\theta) - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\theta))$$

- L maximized for $q(\mathbf{Z}) \leftarrow p(\boldsymbol{Z}|\boldsymbol{X},\theta)$



E step of the EM

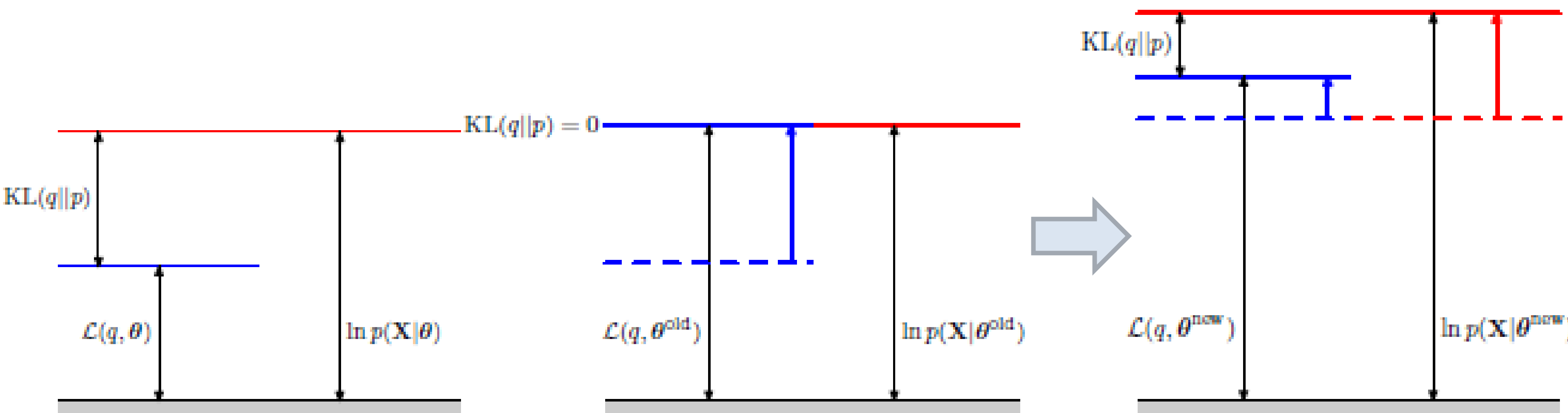
THE EM ALGORITHM IN GENERAL: THE M-STEP

- M-step maximizes $\mathrm{L}(q|\theta)$ w.r.t. for fixed q

$$\mathcal{L}(q,\theta) \;=\; \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z})$$

- L maximized for

$$\theta = \arg\max_{\theta} \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X},\mathbf{Z}|\theta)$$



M step of the EM

A BAYESIAN METHOD OF MODEL SELECTION CRITERIA FOR GAUSSIAN MIXTURE MODELS

- Finding the number of components
  - Modeling the parameters

1. Use the **lower bound** as the model selection score.

$$P(D, \mu, T, s|\pi) = P(D|\mu, T, s)P(s|\pi)P(\mu)P(T).$$

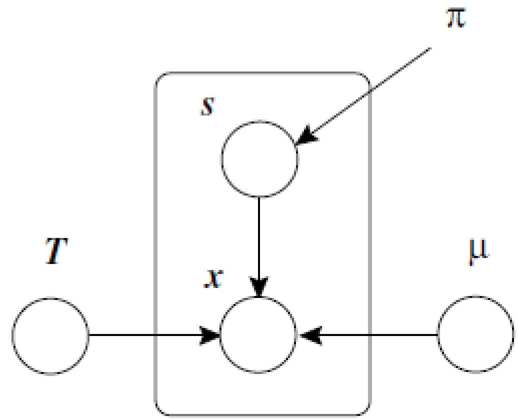
$$P(D|\mu, T, s) = \prod_{n=1}^{N}\prod_{i=1}^{M}\mathcal{N}(x_n|\mu_i, T_i)^{s_{in}}.$$

$$P(s|\pi) = \prod_{i=1}^{M}\prod_{n=1}^{N}\pi_i^{s_{in}}.$$

$$P(\mu) = \prod_{i=1}^{M}\mathcal{N}(\mu_i|0, \beta I)$$

$$P(T) = \prod_{i=1}^{M}\mathcal{W}(T_i|\nu, V)$$

W : Wishart distribution



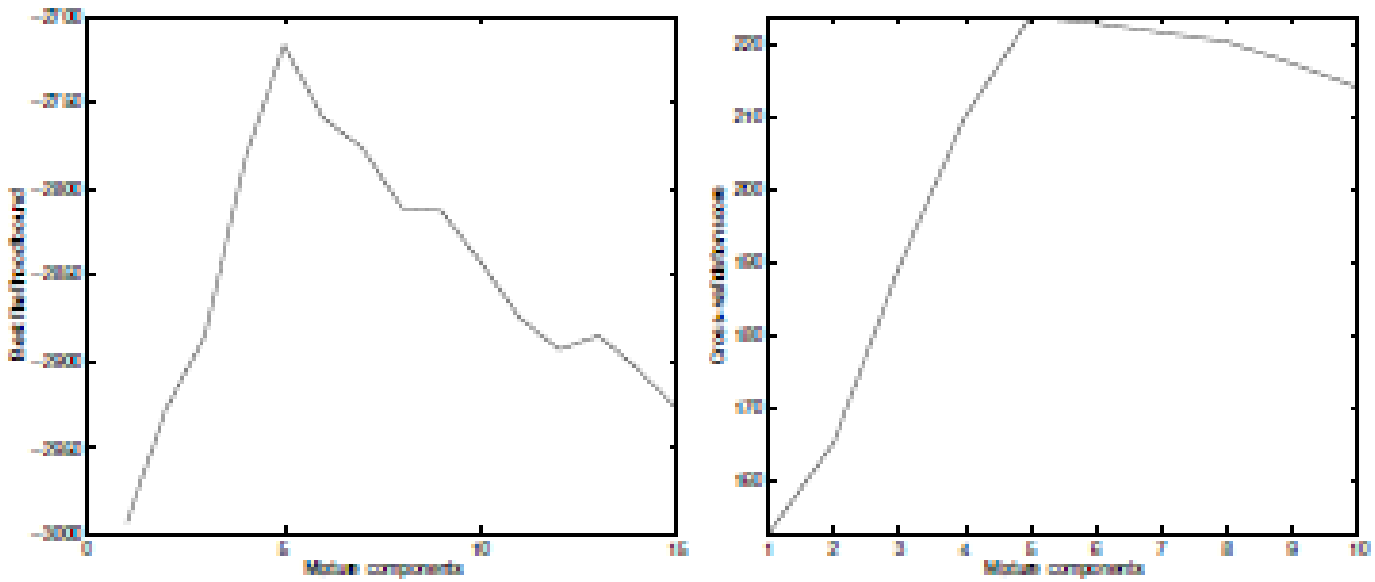
$$\pi_i = \frac{1}{N}\sum_{n=1}^{N} p_{in}.$$

$$p_{in} = \frac{\widetilde{p}_{in}}{\sum_{j=1}^{M}\widetilde{p}_{jn}}$$

$$\begin{aligned}\widetilde{p}_{in} &= \exp(\langle \ln|T_i|\rangle/2 + \ln \pi_i \\ &-\frac{1}{2}\mathrm{Tr}\{\langle \mathrm{T_i}\rangle(\mathrm{x_n x_n^T} - \langle \mu_\mathrm{i}\rangle \mathrm{x_n^T} \\ &-x_n\langle\mu_\mathrm{i}\rangle^\mathrm{T} + \langle \mu_\mathrm{i}\mu_\mathrm{i}^\mathrm{T}\rangle)\})\end{aligned}$$

2. remove components with very small **mixing coefficients**

A. Corduneanu and C. Bishop. Variational Bayesian Model Selection for Mixture Distributions 2001


0
5
10
15
Mixture components
220
210
200
190
180
170
160
1
2
3
4
5
6
7
8
9
10
Mixture components

THE EM ALGORITHM IN GENERAL: FINAL THOUGHTS

- Local maxima of $\mathrm{L}(\mathrm{q}|\theta)$ correspond to those of $\ln p(\boldsymbol{X}|\theta)$
- EM converges to local maximum of likelihood
  - Coordinate ascent on $\mathrm{L}(\mathrm{q}|\theta)$ and $\mathrm{L}(\mathrm{q}|\theta) = \ln p(\boldsymbol{X}|\theta)$ after E-step
- Alternative schemes to optimize the bound
  - Generalized EM: relax M-step from maximizing to increasing L
  - Expectation Conditional Maximization: M-step maximizes w.r.t. groups of parameters in turn
  - Incremental EM: E-step per data point, incremental M-step
  - Variational EM: relax E-step from maximizing to increasing L
    - no longer L = $\ln p(\boldsymbol{X}|\theta)$ after E-step