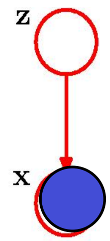
Latent Variable View of EM

Sargur Srihari
srihari@cedar.buffalo.edu

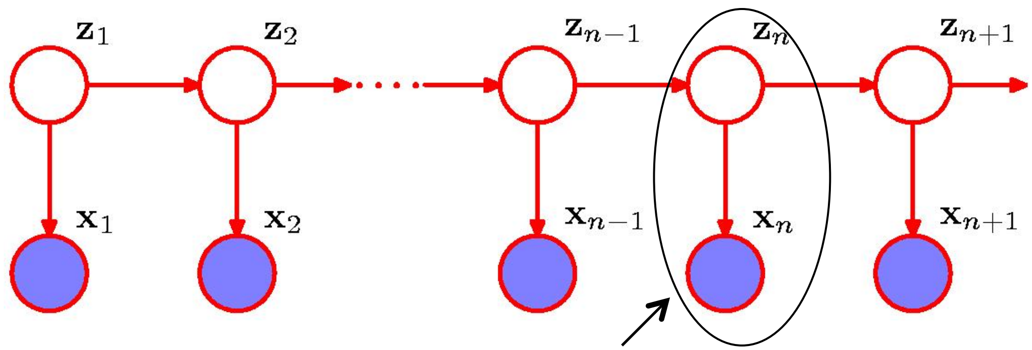
Examples of latent variables

1. Mixture Model



– Joint distribution is $p(\mathrm{x},\mathrm{z})$
  - We don't have values for z

2. Hidden Markov Model



– A single time slice is a mixture with components $p(\mathrm{x}|\mathrm{z})$
– An extension of mixture model
  - Choice of mixture component depends on choice of mixture component for previous distribution
– Latent variables are multinomial variables z_n
  - That describe component responsible for generating x_n

Another example of latent variables

3. Topic Models (Latent Dirichlet Allocation)
   - In NLP unobserved groups explain why some observed data are similar
   - Each document is a mixture of various topics (latent variables)
   - Topics generate words
     - CAT-related: milk, meow, kitten
     - DOG-related: puppy, bark, bone
   - Multinomial distributions over words with Dirichlet priors

Main Idea of EM

- Goal of EM is:
  - find maximum likelihood models for distributions $p(\mathrm{x})$ that have latent (or missing) data
    - E.g., GMMs, HMMs
  - In case of Gaussian mixture models $\boxed{p(\mathrm{x}) = \sum_{k=1}^{K} \pi_k N(\mathrm{x} \mid \mu_k, \Sigma_k)}$
    - We have a complex distribution of observed variables x
    - We wish to estimate its parameters
- Introduce latent variables z so that $\boxed{p(\mathrm{x}) = \sum_{z} p(\mathrm{x},\mathrm{z})}$
  - joint distribution $p(\mathrm{x},\mathrm{z})$ is more tractable (since we know forms of components) $\boxed{p(\mathrm{x} \mid z_k = 1) = N(\mathrm{x} \mid \mu_k, \Sigma_k)}$
  - Complicated form from simpler components
- The original distribution is obtained by marginalizing the joint distribution

Alternative View of EM

- This view recognizes key role of latent variables
- Observed data
  - matrix $X = \begin{bmatrix} x_1 \\ x_2 \\ x_n \\ \end{bmatrix}$

  Latent Variables
  - matrix $Z = \begin{bmatrix} z_1 \\ z_2 \\ z_n \\ \end{bmatrix}$

  - where n^{th} row represents $x_n^T=[x_{n1}\ x_{n2}\ \ x_{nD}]$
  - with corresponding row $z_n^T=[z_{n1}\ z_{n2}\ \ z_{nK}]$

- Goal of EM algorithm is to find maximum likelihood solution for $p(X)$ given some X
- When we do not have Z

Likelihood Function involving Latent Variables

- Joint likelihood function is $p(\mathrm{X},\mathrm{Z}\,|\,\theta)$ where θ is the set of all model parameters
  - E.g., means, covariances, responsibilities
- Marginal likelihood function of observed data
  - From sum rule

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

- Log likelihood function is

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

Latent Variables in EM

- Log likelihood function is

$$\ln p(X \mid \theta) = \ln\left\{\sum_{Z} p(\mathrm{X},\mathrm{Z} \mid \theta)\right\}$$

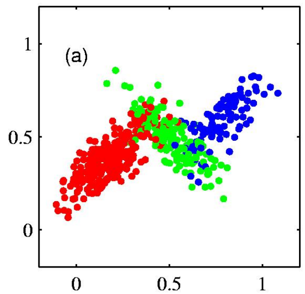
Summation inside brackets
due to marginalization
Not due to log-likelihood

- Key Observation:
  - Summation over latent variables appears inside logarithm
    - Even if joint distribution $p(\mathrm{X},\mathrm{Z} \mid \theta)$ belongs to exponential family the marginal distribution $p(\mathrm{X} \mid \theta)$ does not
      - Taking log of Sum of Gaussians does not give simple quadratic
    - Results in complicated expressions for maximum likelihood solution, i.e., what value of q maximizes the likelihood

Complete and Incomplete Data Sets

Complete Data $\{X,Z\}$

- For each observation in X we know corresponding value of latent variable Z
- Log-likelihood has the form $p(X,Z \mid \theta)$
  - maximization over θ is straightforward

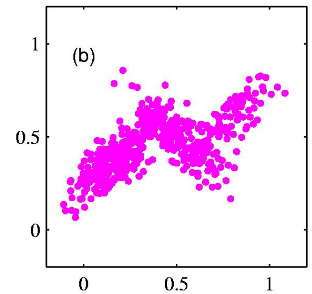


Incomplete Data $\{X\}$

- Actual data set
- Log likelihood function is

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

- Maximization over θ is difficult
  - summations inside logarithm



Expectation of log-likelihood

- Since we don't have the complete data set $\{X,Z\}$ we evaluate the expected log-likelihood, i.e.,

$$E[\ln p(\mathrm{X},\mathrm{Z}|\theta)]$$

- Since we are given X, our state of knowledge of Z is given only by the posterior distribution of the latent variables $p(\mathrm{Z} \mid \mathrm{X}, \theta)$
- Thus expected log-likelihood of complete data is

$$E\,[\,\ln p(X,Z \mid \theta)] = \sum_Z p(Z \mid X,\theta) \underbrace{\ln p(X,Z \mid \theta)}$$

Summation is due to expectation not sum rule!

We maximize this.
Note that the logarithm acts on the joint-- which is tractable

E and M Steps

- *E Step*: Estimate the missing values
  - Use current parameter value θ^{old} to find the posterior distribution of the latent variables given by

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

- *M Step*: Determine revised parameter estimate θ^{new} by *maximizing* $\theta^{\mathrm{new}} = \arg\max_{\theta} Q(\theta, \theta^{\mathrm{old}})$
  - *where*

$$Q\left(\theta,\theta^{old}\right)=\sum_{Z} p\left(Z \mid X,\theta^{old}\right)p\left(X,Z \mid \theta\right)$$

Summation due to expectation

- is the *expectation* of $p(X, Z \mid \theta)$ for some general parameter value θ
  - Evaluate the log-likelihood $\sum_{i=1}^{N} \ln p(Xi, Z \mid \theta)$

General EM Algorithm

- Given joint distribution $p(\mathrm{X},\mathrm{Z}\mid\theta)$ over observed variables X and latent variables Z governed by parameters θ
  goal is to maximize likelihood function $p(\mathrm{X}\mid\theta)$

- Step 1: Choose an initial setting for the parameters θ^{old}
- Step 2: E Step: Evaluate $p(\mathrm{Z}\mid\mathrm{X},\theta^{old})$
- Step 3: M Step: Evaluate θ^{new} given by

$$\theta^{\mathrm{new}} = \arg\max_{\theta} Q(\theta,\theta^{\mathrm{old}})$$

where

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

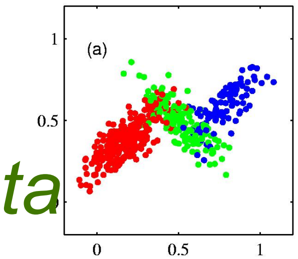
- Check for convergence
  - of either log-likelihood or parameter values
- If not satisfied then let $\theta^{old} \leftarrow \theta^{new}$
- Return to Step 2

Missing Variables

- EM has been described for maximum likelihood function when there are discrete latent variables
- It can also be applied when there are unobserved variables corresponding to missing values in data set
  - Take the joint distribution of all variables and then marginalize over missing ones
  - EM is then used to maximize corresponding likelihood function
- Method is valid when data is missing at random
  - Not if missing value depends on unobserved values
  - E.g., if quantity exceeds some threshold

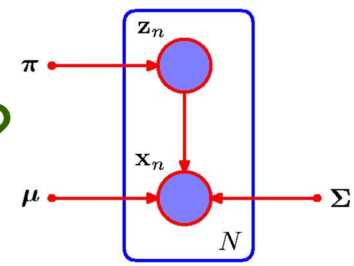
Gaussian Mixtures Revisited

- Apply EM (latent variable view) to GMM
- In the E-step we compute
  - Expectation of *log-likelihood of complete data* $\{X,Z\}$ wrt posterior of latent Variables Z



$$\boxed{Q(\theta,\theta^{old}) = \sum_Z p(Z \mid X,\, \theta^{old}) \ln p(X,Z \mid \theta)}$$

  - What is the form of the two product terms?



- In the M-step we maximize $Q(\theta,\theta^{old})$ wrt θ
  - Will show that this leads to the same m.l estimates for GMM parameters π, μ, Σ as before

Likelihood for Complete Data

- Likelihood function for the complete data set is

$$p(X,Z|\pi,\mu,\Sigma)=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}N\left(x_n|\mu_k,\Sigma_k\right)^{z_{nk}}$$

- Log-likelihood is

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

- Much simpler than log-likelihood for incomplete data:

$$\ln p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N(\mathrm{x}_n|\mu_k,\Sigma_k)\right\}$$

- Maximum likelihood solution for complete data can be obtained in closed form
- Since we don't have values for latent variables, we obtain its expectation wrt the posterior distribution of latent variables

Posterior Distribution of Latent Variables

- From $p(\mathrm{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(\mathrm{x} \mid \mathrm{z}) = \prod_{k=1}^{K} N(x \mid \mu_k, \Sigma_k)^{z_k}$ we have

$$\boxed{p(Z \mid X, \mu, \Sigma) \; \alpha \; \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k N\left(x_n \mid \mu_k, \Sigma_k\right)\right)^{z_{nk}}}$$

- From which we can get the expected value for the indicator variable as

$$\boxed{E[z_{nk}] = \frac{\pi_k N\left(x_n \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j N\left(x_n \mid \mu_j, \Sigma_j\right)} = \gamma\left(z_{nk}\right)}$$

- Substituting into complete log-likelihood:

$$\boxed{E_Z\left[\ln p(X, Z \mid \pi, \mu, \Sigma)\right] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{nk}\right)\left\{\ln \pi_k + \ln N\left(x_n \mid \mu_k, \Sigma_k\right)\right\}}$$

- Final procedure: choose initial values for $\pi^{old}, \mu^{old}, \Sigma^{old}$
  - Evaluate the responsibilities (E-step)
  - Keep responsibilities fixed and use closed-form solutions for $\pi^{new}, \mu^{new}, \Sigma^{new}$

$$\boxed{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n} \qquad \boxed{\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathrm{x}_n - \mu_k)(\mathrm{x}_n - \mu_k)^T} \qquad \boxed{\pi_k = \frac{N_k}{N}}$$

Relation to K-means

- EM for Gaussian mixtures has close similarity to K-means
- K-means performs a hard assignment of data points to clusters
  - Each data point is associated uniquely with one cluster
- EM makes a soft assignment based on posterior probabilities
- K-means does not estimate the covariances of the clusters but only the cluster means