K-Means Clustering and Gaussian Mixture Model

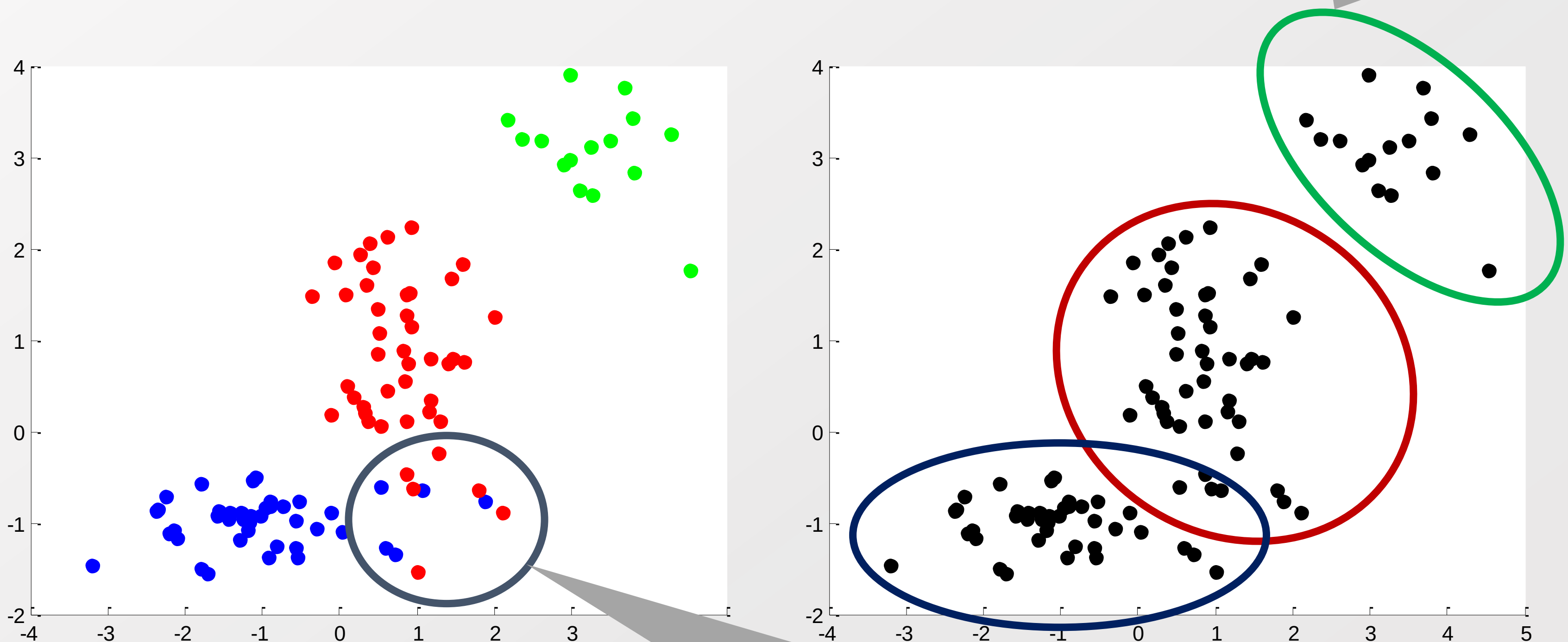
Il-Chul Moon
Department of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

K-MEANS ALGORITHM

Clustering Problem

- How to cluster the unlabeled data points?
  - No concrete knowledge of their classes
  - Latent (hidden) variable of classes
  - Optimal assignment to the latent classes

How to assign data points to classes?
→ Clustering
(here classes == clusters)

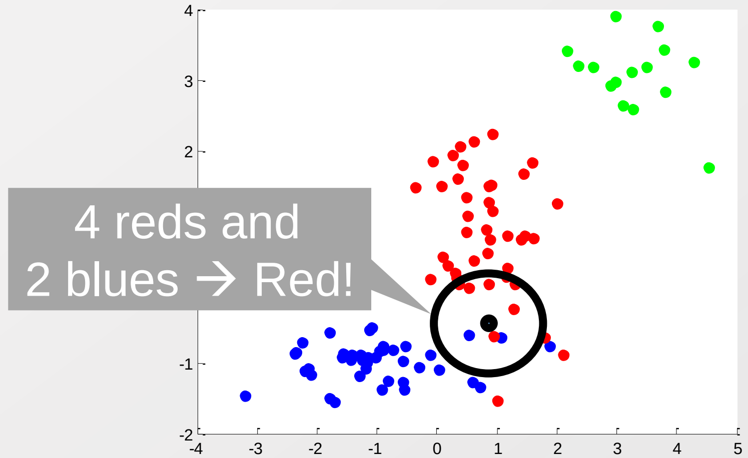


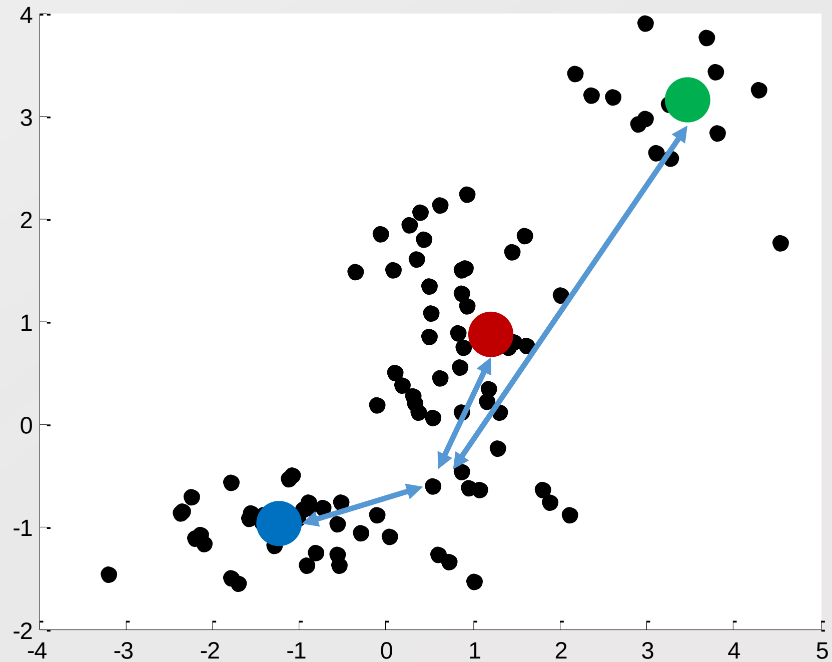
Uncertain area of clustering

K-Means Algorithm

- K-Means algorithm
  - Setup K number of centroids (or prototypes) and cluster data points by the distance from the points to the nearest centroid
- Formally,
  - $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$
  - Minimize J by optimizing
    - r_{nk}: the assignment of data points to clusters
    - μ_k: the location of centroids
  - Iterative optimization
    - Why?
    - Two variables are interacting

K-Means! = K-Nearest Neighbor



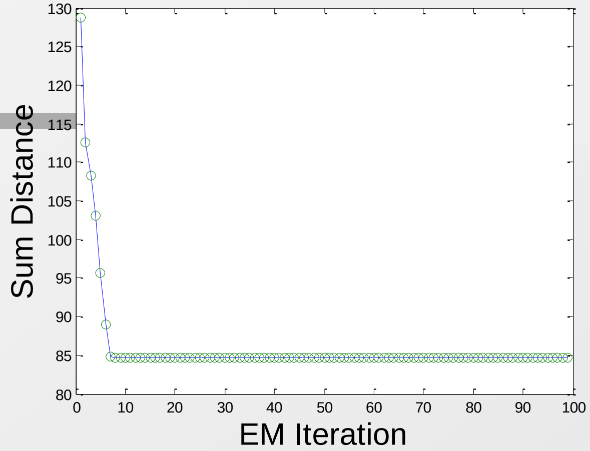


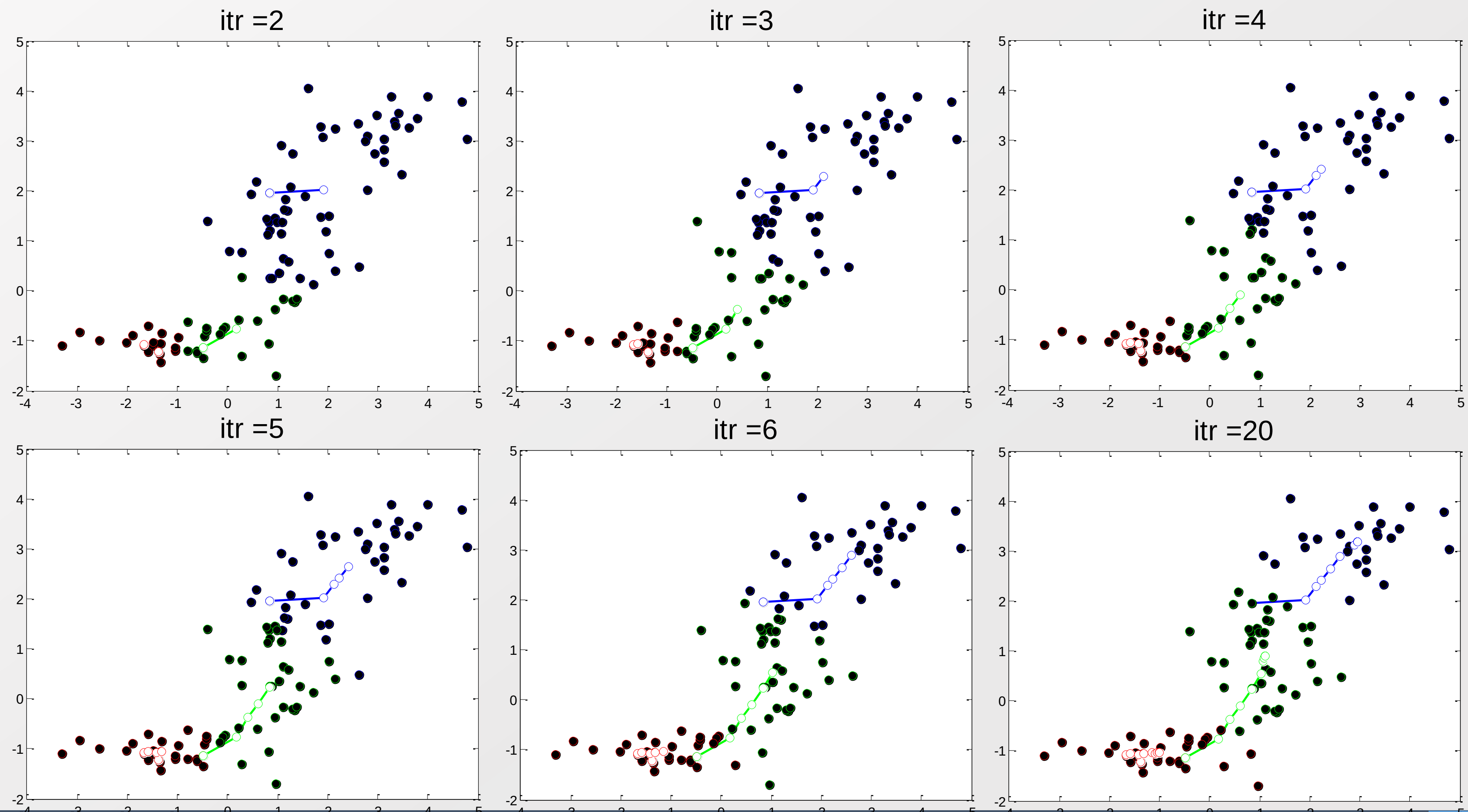
Expectation and Maximization

- $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$
  - Expectation
    - Expectation of the log-likelihood given the parameters
    - Assign the data points to the nearest centroid
  - Maximization
    - Maximization of the parameters with respect to the log-likelihood
    - Update the centroid positions given the assignments
- r_{nk}
  - $r_{nk} = \{0,1\}$
  - Discrete variable
  - Logical choice: the nearest centroid μ_k for a data point of x_n
- μ_k
  - $\frac{dJ}{d\mu_k} = \frac{d}{d\mu_k} \sum_{n=1}^{N} \sum_{l=1}^{K} r_{nl} ||x_n - \mu_l||^2 = \frac{d}{d\mu_k} \sum_{n=1}^{N} r_{nk} ||x_n - \mu_k||^2 = \sum_{n=1}^{N} -2r_{nk}(x_n - \mu_k) = -2(-\sum_{n=1}^{N} r_{nk}\mu_k + \sum_{n=1}^{N} r_{nk}x_n) = 0$
  - $\mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}}$

Progress of K-Means Algorithm

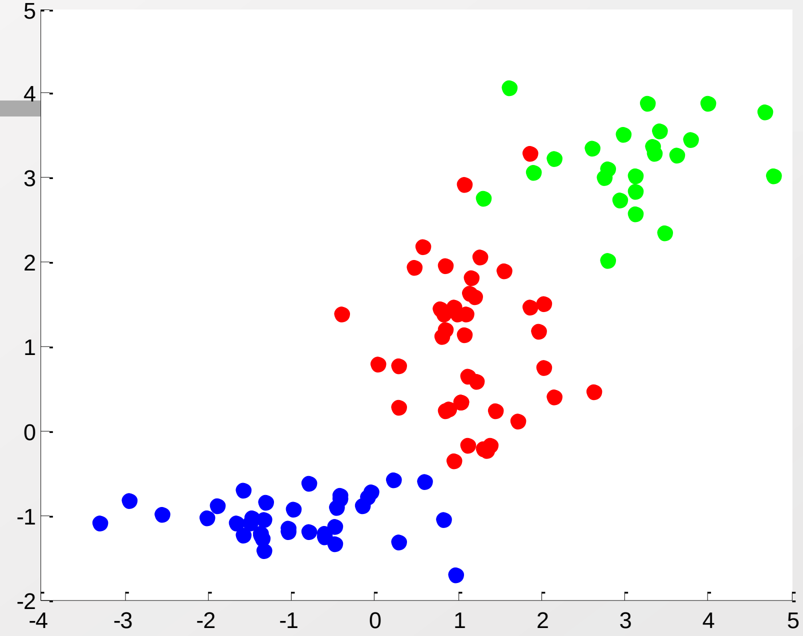
- EM iterations to
  - Optimize the assignments with respect to the sum of distances
  - Optimize the parameters with respect to the sum of distances



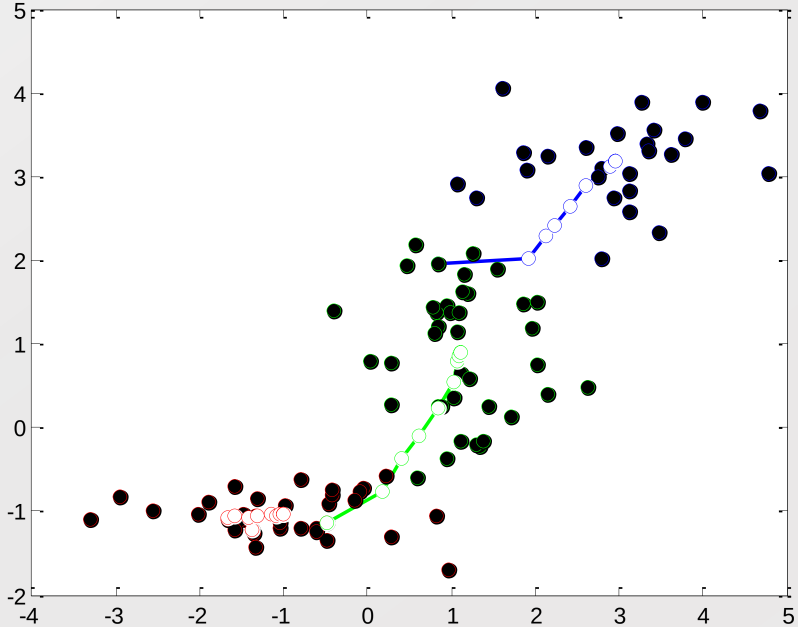


Properties of K-Means Algorithm

- # of clusters is uncertain
- Initial location of centroids
  - Some initial locations might not result in the reasonable results
- Limitation of distance metrics
  - Euclidean distance is very limited knowledge of information
- Hard clustering
  - Hard assignment of data points to clusters
    - $r_{nk} = \{0,1\}$
      - This can be the smoothly distributed probability
    - Any alternatives?
    - Soft clustering



itr =20



GAUSSIAN MIXTURE MODEL

Multinomial Distribution

- Binary variable
  - Selecting 0 or 1 → binomial distribution
- How about K options?
  - $X = (0,0,1,0,0,0)$ when $K = 6$ and selecting the third option
  - $\sum_k x_k = 1, P(X|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k{}^{x_k}\ such\ that\ \mu_k \geq 0, \sum_k \mu_k = 1$
  - A generalization of binomial distribution → Multinomial distribution
- Given a dataset D with N selections, $x_1, \ldots, x_n$
  - $P(X|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k{}^{x_{nk}} = \prod_{k=1}^{K} \mu_k{}^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{K} \mu_k{}^{m_k}$
    - When $m_k = \sum_{n=1}^{N} x_{nk}$
    - Number of selecting k^{th} *option* out of N selections
  - How to determine the maximum likelihood solution of $\boldsymbol{\mu}$?
    - Maximize $P(X|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k{}^{m_k}$
    - Subject to $\mu_k \geq 0, \sum_k \mu_k = 1$

Lagrange Method

Maximize $\boldsymbol{P(X|\mu) = \prod_{k=1}^{K} \mu_k^{m_k}}$
Subject to $\boldsymbol{\mu_k \geq 0, \sum_k \mu_k = 1}$
When $m_k = \sum_{n=1}^{N} x_{nk}$

- Method of finding a local maximum subject to constraints
  - Maximize *f(x,y)*
  - Subject to *g(x,y)=c*
  - Assuming that *f* and *g* have continuous partial derivatives
  - 1) Lagrange function and multiplier (do you recall this?)
    - $L(x,y,\lambda) = f(x,y) + \lambda(g(x,y) - c)$
    - $L(\mu,m,\lambda) = \sum_{k=1}^{K} m_k \ln \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1)$
      - Using the log likelihood
  - 2) Take the partial first-order derivative of variables, and set it to be zero
    - $\frac{d}{d\mu_k} L(\mu,m,\lambda) = \frac{m_k}{\mu_k} + \lambda = 0 \rightarrow \mu_k = -\frac{m_k}{\lambda}$
  - 3) Utilize the constraint to get the optimized value
    - $\sum_k \mu_k = 1 \rightarrow \sum_k -\frac{m_k}{\lambda} = 1 \rightarrow \sum_k m_k = -\lambda \rightarrow \sum_k \sum_{n=1}^{N} x_{nk} = -\lambda \rightarrow N = -\lambda$
    - $\mu_k = \frac{m_k}{N}$: MLE parameter of multinomial distribution

Multivariate Gaussian Distribution

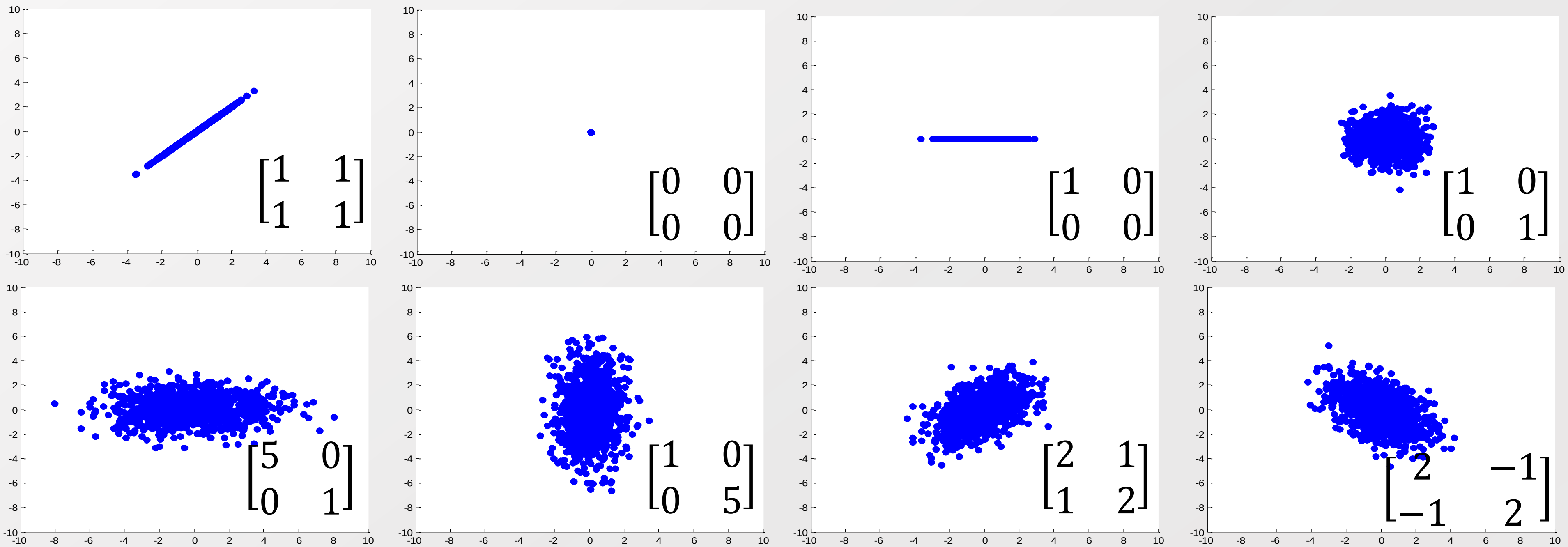
- Probability density function of the Gaussian distribution
  - $N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
  - $\boldsymbol{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$
    - $\mathbf{ln}\boldsymbol{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{C}$
    - $\mathbf{ln}\boldsymbol{N}(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) + \boldsymbol{C}$

      $\propto -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} Tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T]$

      $= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}Tr[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\big((\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T\big)]$
    - $\frac{d}{d\mu}\ln N(X|\mu,\Sigma) = 0 \rightarrow -\frac{1}{2}\times 2\times -1\times\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}}) = 0 \rightarrow \widehat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N}$
    - $\frac{d}{d\boldsymbol{\Sigma}^{-1}}\ln N(X|\mu,\Sigma) = 0 \rightarrow \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})^{\boldsymbol{T}}$
      - Beyond the scope of the course
      - Use "trace trick" and 1) $\frac{d}{dA}\log|A| = A^{-T}$, 2) $\frac{d}{dA}Tr[AB] = \frac{d}{dA}Tr[BA] = B^T$

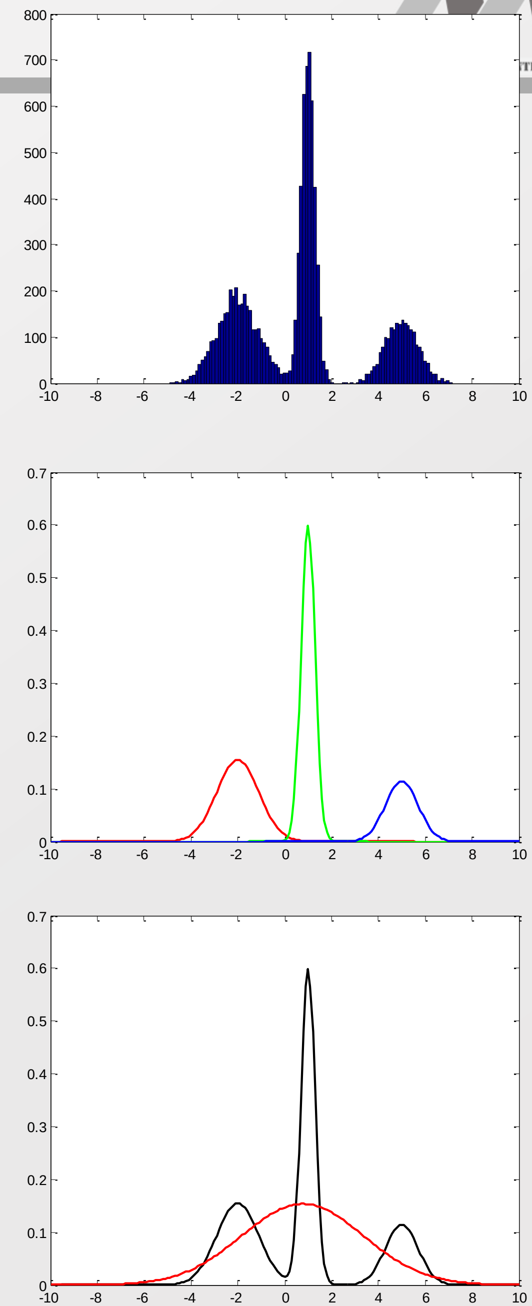
Samples of Multivariate Gaussian Distribution

- Samples of multivariate Gaussian distributions
  - With various covariance matrixes
  - Covariance matrix should a positive-definite matrix
    - $z^T \Sigma z > 0$ for every non-zero column vector z
    - $[a\ b]\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = a^2 + b^2 > 0$ when a, b are non-zero



Mixture Model

- Imagine that the samples are drawn from three different normal distributions
  - Subpopulation
  - The conventional distributions cannot explain the distribution accurately
  - We need to mix the three normal distribution → Create a new distribution adapted to the samples
  - Mixture distribution
- $P(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k)$
  - Mixing coefficients, π_k: A normal distribution is chosen out of K options with probability
    - Works as weighting
    - $\sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$
    - This is a probability (as well as weighting!)
    - Then, which distribution?
    - New variable? Let's say Z!
  - Mixture component, $N(x|\mu_k, \sigma_k)$: A distribution for the subpopulation
- $P(x) = \sum_{k=1}^{K} P(z_k) P(x|z_k)$
  - Why this ordering of variables?

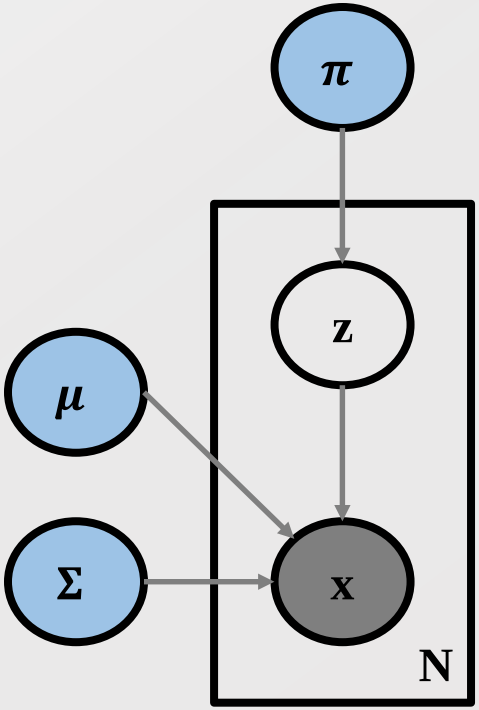


Gaussian Mixture Model

- Let's assume that the data points are drawn from a mixture distribution of multiple multivariate Gaussian distributions
  - $P(x) = \sum_{k=1}^{K} P(z_k)P(x|z) \; = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$
  - How to model such mixture?
    - Mixing coefficient, or Selection variable: z_k
      - The selection is stochastic which follows the multinomial distribution
      - $z_k \in \{0,1\}, \sum_k z_k = 1, P(z_k = 1) = \pi_k, \sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$
      - $P(Z) = \prod_{k=1}^{K} \pi_k{}^{z_k}$
    - Mixture component
      - $P(X|z_k = 1) = N(x|\mu_k, \Sigma_k) \rightarrow P(X|Z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$
  - This is the marginalized probability. How about conditional?
    - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P\left(x\middle|z_j = 1\right)}$

$$= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N\left(x\middle|\mu_j, \Sigma_j\right)}$$

- Log likelihood of the entire dataset is
  - $\ln P(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)\}$



Expectation of GMM

- Similar problem of K-means algorithm
  - Two interacting parameters
  - As before, we apply the expectation and the maximization algorithm
    - Expectation: the assignment between the clusters and the data points
    - Maximization: the update of the parameters
- Expectation step
  - Assign a data point to a nearest cluster → the assignment probability
    - Given the parameters and the data point, calculate the likelihood
  - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P\left(x\middle|z_j = 1\right)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N\left(x\middle|\mu_j, \Sigma_j\right)}$
    - Here, x, π, μ, Σ are given, calculate $\gamma(z_{nk})$
    - $\gamma(z_{nk})$ are used to calculate π, μ, Σ
    - The new $\gamma(z_{nk})$ motivates the update of the old parameters

Maximization of GMM

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

$$\ln N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{C}$$

$$\ln N(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) + \boldsymbol{C}$$

$$\frac{d}{d\mu}\ln N(X|\mu,\Sigma) = 0 \rightarrow -\frac{1}{2}\times 2\times -1\times\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}) = 0 \rightarrow \hat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N}$$

$$\frac{d}{d\boldsymbol{\Sigma}^{-1}}\ln N(X|\mu,\Sigma) = 0 \rightarrow \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}})^{\boldsymbol{T}}$$

- Maximization step
  - Update the parameters given $\gamma(z_{nk})$
  - Parameters to update: π, μ, Σ
    - $\ln P(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k$
    - Typical methods
      - Derivative → set the equation to zero when the function is smooth
      - Lagrange method when there is a constraint.
        Which parameter has the constraint?

$$\gamma(z_{nk}) = \frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)}$$

- $\frac{d}{d\mu_k}\ln P(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \widehat{\mu_k}) = 0$

  $\rightarrow \sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n - \widehat{\mu_k})=0 \rightarrow \widehat{\mu_k} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{\sum_{n=1}^{N}\gamma(z_{nk})}$

- $\frac{d}{d\Sigma_k}\ln P(X|\pi,\mu,\Sigma) = 0$

  $\rightarrow \Sigma_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\widehat{\mu_k})(\boldsymbol{x}_n-\widehat{\mu_k})^{\boldsymbol{T}}}{\sum_{n=1}^{N}\gamma(z_{nk})}$

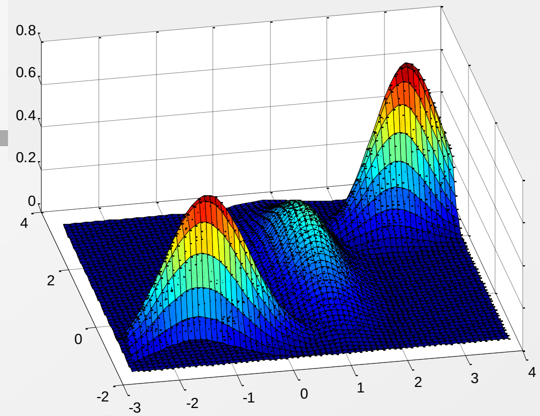
- $\frac{d}{d\pi_k}\ln P(X|\pi,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) = 0$

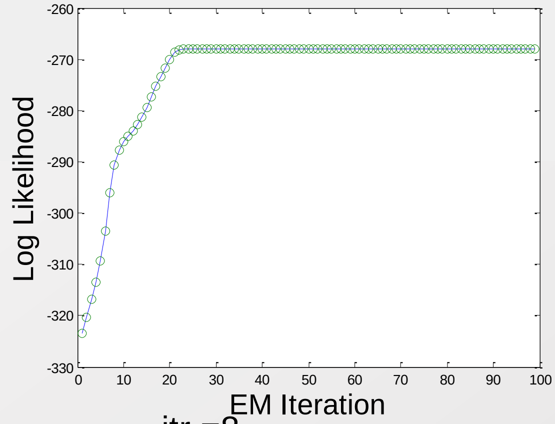
  $\rightarrow \sum_{n=1}^{N}\frac{N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)} + \lambda = 0 \rightarrow \sum_{k=1}^{K}\left\{\sum_{n=1}^{N}\frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)} + \pi_k\lambda\right\} = 0$

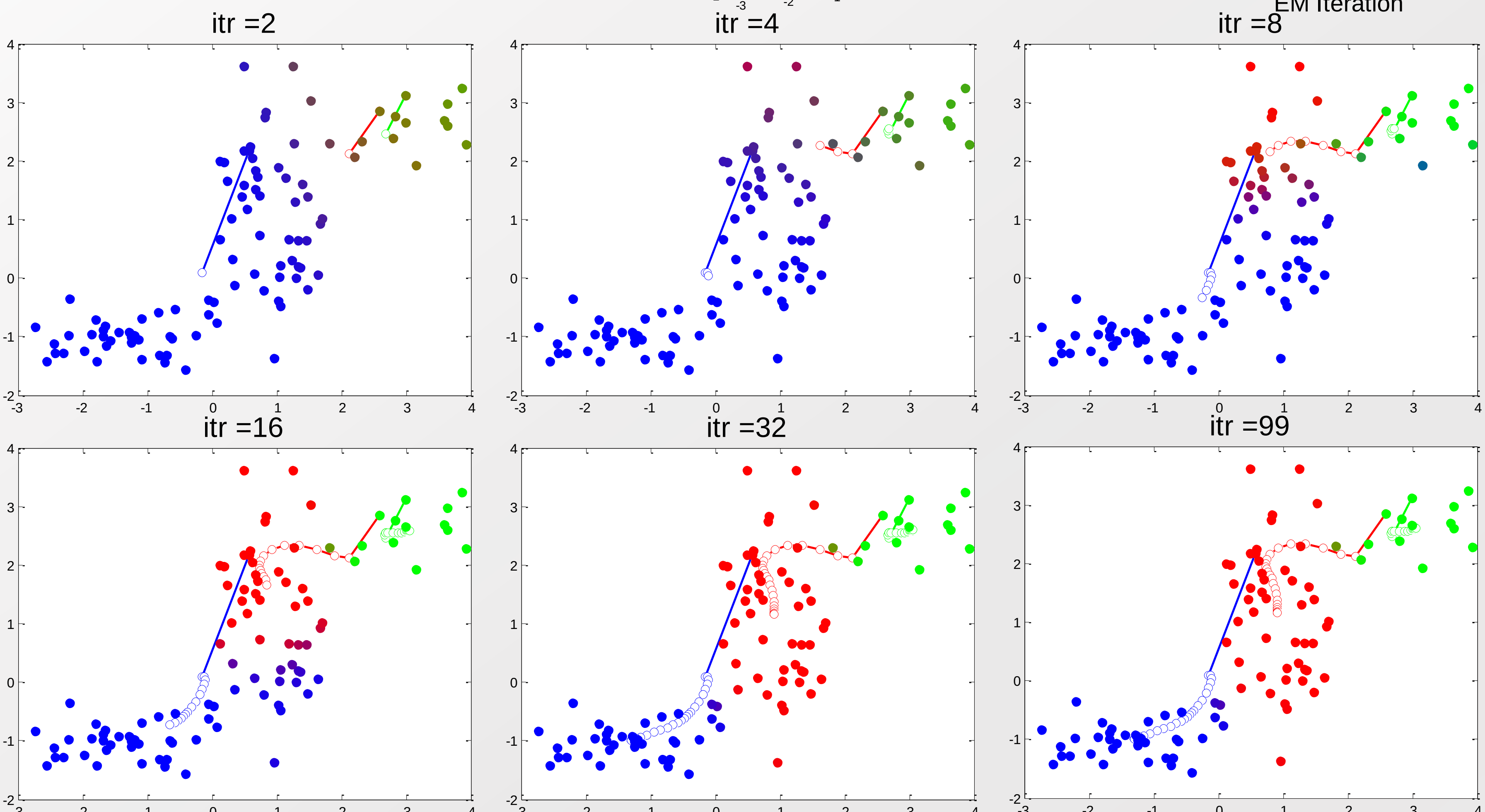
  $\rightarrow \lambda = -N \rightarrow \pi_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N}$

Progress of GMM

- Soft clustering
  - Estimated parameters
  - Soft assignment of data points to clusters

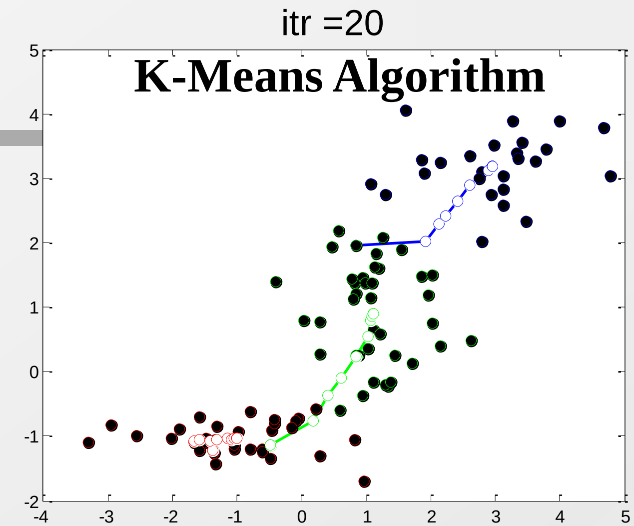


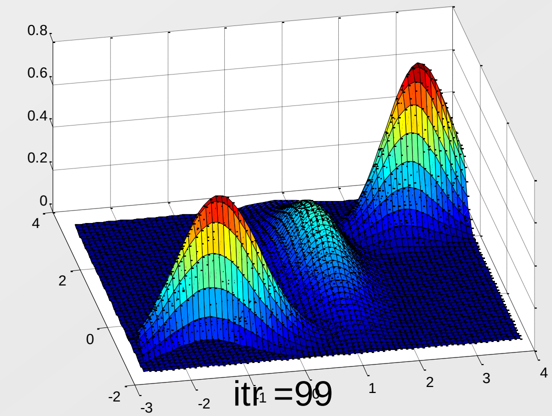


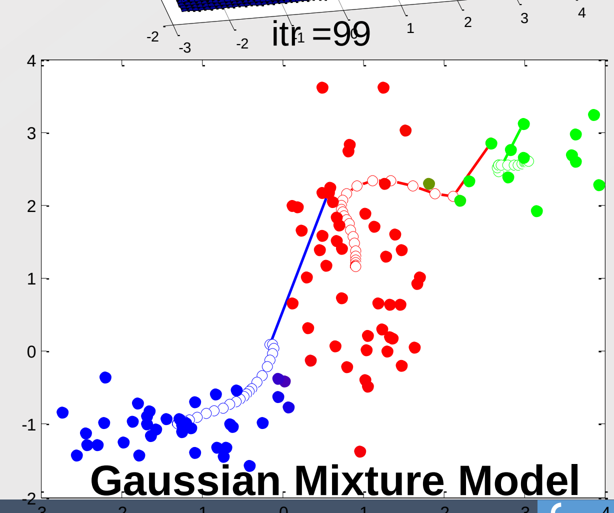


Properties of GMM

- Pros and cons of Gaussian mixture model
  - Pros
    - More information
      - Soft clustering
      - Not a simple and discrete assignment
        - Information loss
    - More and more information
      - Learn the latent distribution
      - Distance is not always the answer of the distribution
  - Cons
    - Long computation time
      - Why?
    - Falling into local maximum
    - Deciding K
- Anyways to mitigate the disadvantage?
  - Fast K-means and slow GMM







Relation between K-Means and GMM

- $N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$
- $P(\boldsymbol{x}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma_k}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k))$
  - Let's say $\Sigma_k = \epsilon \mathrm{I}$
    - Here, I is the identity matrix and ϵ is not updated by the EM process
    - $\mathrm{I} = \mathrm{I}^{-1}$
  - $= \frac{1}{(2\pi)^{D/2}\epsilon^{1/2}}\exp(-\frac{1}{2\epsilon}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}-\boldsymbol{\mu}_k))$
  - $= \frac{1}{(2\pi)^{D/2}\epsilon^{1/2}}\exp(-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2)$

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$$

K-Means Algorithm

- $\gamma(z_{nk}) = \frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N\left(x\middle|\mu_j,\Sigma_j\right)} = \frac{\pi_k \exp(-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2)}{\sum_{j=1}^{K}\pi_j \exp(-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2)}$
  - When $\epsilon \to 0$, the term of smallest $||\boldsymbol{x}-\boldsymbol{\mu}_k||^2$ approaches zero most slowly
  - When all other terms are zero, the term of the smallest $||\boldsymbol{x}-\boldsymbol{\mu}_k||^2$ has a value
  - Now, it becomes the hard assignment
- Still, GMM with $\epsilon\mathrm{I}$ is not K-Means. Why?
  - Soft assignment + Covariance matrix learning

EM ALGORITHM

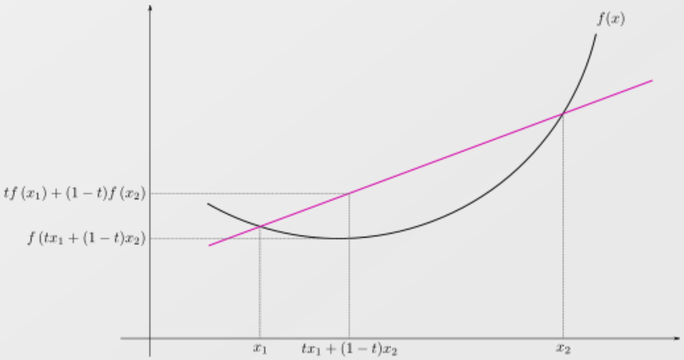
Inference with Latent Variables

- Difference between classification and clustering
- Let's say
  - $\{X, Z\}$: complete set of variables
  - X: observed variables
  - Z: hidden (latent) variables
  - θ: parameters for distributions
  - $P(X|\theta) = \sum_Z P(X, Z|\theta) \rightarrow \ln P(X|\theta) = \ln\{\sum_Z P(X, Z|\theta)\}$
    - Any problem here?
    - The locations of summation and log make this complicated
    - Eventually, we want to exchange the locations of the two operators
- What we want to know is
  - The values of Z and θ
    - Optimizing $P(X|\theta) = \sum_Z P(X, Z|\theta)$
  - The interacting terms for the optimization

Probability Decomposition

- $l(\theta) = \ln P(X|\theta) = \ln\{\sum_Z P(X,Z|\theta)\} = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\}$
  - Use the Jensen's inequality
  - $\ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)}$
- $= \sum_Z q(Z)\ln P(X,Z|\theta) - q(Z)\ln q(Z)$
  - Recall the second term?
  - $H(X) = -\sum_X P(X=x)log_b P(X=x)$
- $= E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
  - $Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
  - This hold for any distribution of q
  - This is only the lower bound of $l(\theta)$
    - Need to make it tight!
    - How to?

Jensen's Inequality



When φ(x) is concave

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_j}\right) \geq \frac{\sum a_i \varphi(x_i)}{\sum a_j}$$

When φ(x) is convex

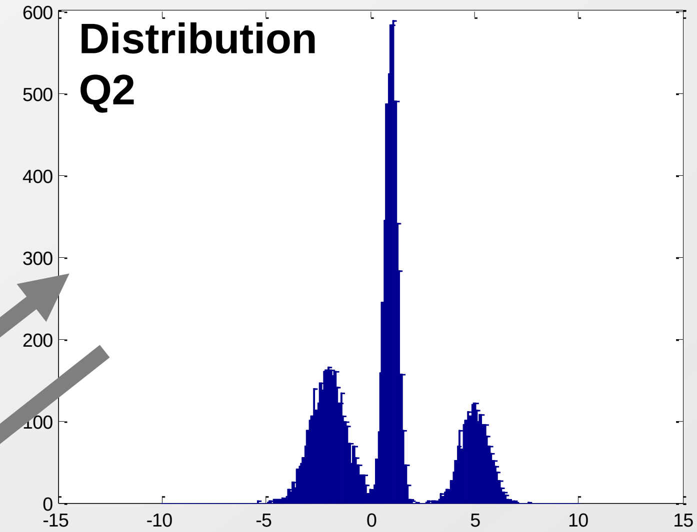
$$\varphi\left(\frac{\sum a_i x_i}{\sum a_j}\right) \leq \frac{\sum a_i \varphi(x_i)}{\sum a_j}$$

Maximizing the Lower Bound (1)

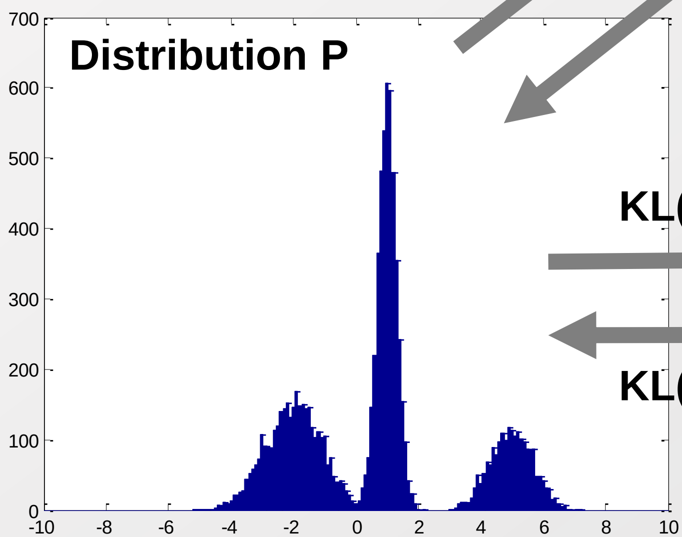
- $l(\theta) = \ln P(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = Q(\theta,q)$
  - $Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
- The other storyline is
  - $l(\theta) \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = \sum_Z q(Z)\ln\frac{P(Z|X,\theta)P(X|\theta)}{q(Z)}$
    $= \sum_Z\left\{q(Z)\ln\frac{P(Z|X,\theta)}{q(Z)} + q(Z)\ln P(X|\theta)\right\} = \ln P(X|\theta) + \sum_Z\left\{q(Z)\ln\frac{P(Z|X,\theta)}{q(Z)}\right\}$
  - $L(\theta,q) = \ln P(X|\theta) - \sum_Z\left\{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\right\}$
- Here, the second term is a very special term
  - $KL\big(q(Z)\|P(Z|X,\theta)\big) = \sum_Z\left\{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\right\}$
  - Kullback-Leiber divergence, or KL divergence: $KL(P\|Q) = \sum_i P(i)\ln\left(\frac{P(i)}{Q(i)}\right)$
  - Non-symmetric measure of the difference between two probability distributions, or $KL(P\|Q)$
  - Measures the difference
    - $KL(P\|Q) \geq 0$
    - When there is no difference between P and Q, $KL(P\|Q) = 0$

KL Divergence

- Kullback-Leiber divergence, or KL divergence: $KL(P||Q) = \sum_i P(i) \ln\left(\frac{P(i)}{Q(i)}\right)$
  - Measures the matching performance of P and Q
  - Consider Gaussian distribution and Gaussian mixture distribution

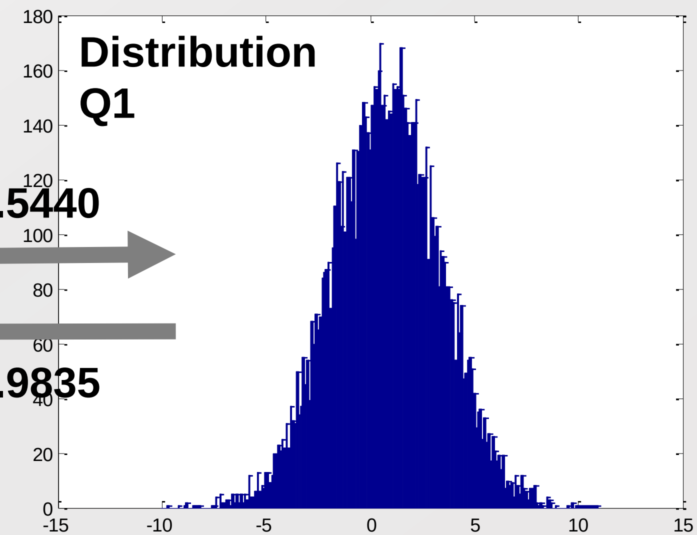


KL(P||Q2)=0.0104 KL(Q2||P)=0.0101



KL(P||Q1)=0.5440

KL(Q1||P)=0.9835



Maximizing the Lower Bound (2)

- $l(\theta) = \ln P(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = Q(\theta,q)$
  - $Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
  - $L(\theta,q) = \ln P(X|\theta) - \sum_Z\left\{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\right\}$
- Why do we compute $L(\theta,q)$?
  - We do not know how to optimize $Q(\theta,q)$ without further knowledge of $q(Z)$
  - The second term of $L(\theta,q)$ tells how to set $q(Z)$
    - The first term is fixed when θ is fixed ***at time t***
    - The second term can be minimized to maximize $L(\theta,q)$
      - $KL(q(Z)||P(Z|X,\theta)) = 0 \rightarrow q^t(Z) = P(Z|X,\theta^t)$
  - Now, the lower bound with optimized *q* is
    - $Q(\theta,q^t) = E_{q^t(Z)}\ln P(X,Z|\theta^t) + H(q^t)$
- Then, optimizing θ to retrieve the tight lower bound is
  - $\theta^{t+1} = argmax_\theta Q(\theta,q^t) = argmax_\theta E_{q^t(Z)}\ln P(X,Z|\theta)$
    - $q^t(Z) \rightarrow$ Distribution parameters for latent variable is at time t
    - $\ln P(X,Z|\theta) \rightarrow$ optimized log likelihood parameters is at time $t+1$

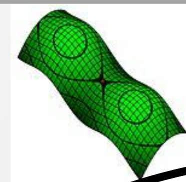
Tells how to setup Z by setting $q^t(Z) = P(Z|X,\theta^t)$

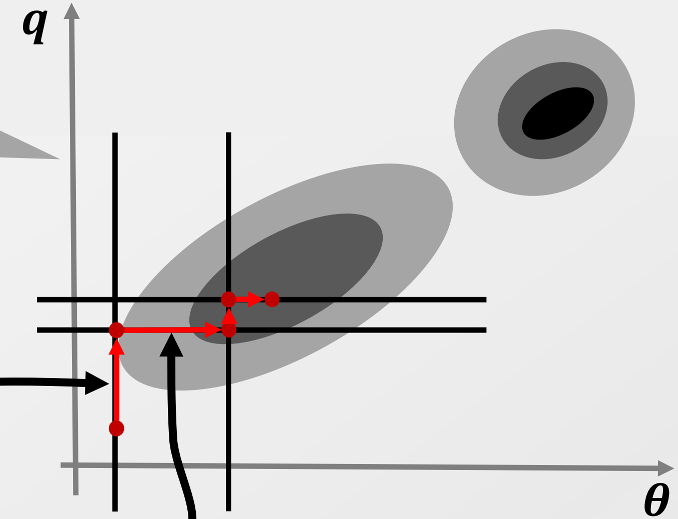
Relax the KL divergence by updating θ^t to θ^{t+1}

Graphical Interpretation of Lower Bound Maximization

- $l(\theta) = \ln P(X|\theta) \geq L(\theta, q)$
  $= \ln P(X|\theta) - \sum_Z \left\{ q(Z) \ln \frac{q(Z)}{P(Z|X,\theta)} \right\}$
  - $\ln P(X|\theta) = L(\theta, q) + \sum_Z \left\{ q(Z) \ln \frac{q(Z)}{P(Z|X,\theta)} \right\}$
    $= L(\theta, q) + KL(q||p)$

Fall into a local maxima or ???





$KL(q||p)$

$\ln P(X|\theta)$

$L(\theta, q)$

Optimize q

$KL(q||p) = 0$

$L(\theta^t, q)$

$\ln P(X|\theta^t)$

Optimize θ

$KL(q||p)$

$L(\theta^{t+1}, q)$

$\ln P\left(X|\theta^{t+1}\right)$

Setting
$q^t(Z) = P(Z|X, \theta^t)$

Setting
$\theta^{t+1} = argmax_\theta E_{q^t(Z)} \ln P(X, Z|\theta)$

EM Algorithm

$$l(\theta) = \ln P(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = Q(\theta,q)$$

$$Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$$

$$L(\theta,q) = \ln P(X|\theta) - \sum_Z \{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\}$$

- EM algorithm
  - Finds the maximum likelihood solutions for models with latent variables
  - $P(X|\theta) = \sum_Z P(X,Z|\theta) \rightarrow \ln P(X|\theta) = \ln\{\sum_Z P(X,Z|\theta)\}$
- EM algorithm
  - Initialize θ^0 to an arbitrary point
  - Loop until the likelihood converges
    - Expectation step
      - $q^{t+1}(z) = argmax_q Q(\theta^t,q) = argmax_q L(\theta^t,q) = argmin_q KL(q||P(Z|X,\theta^t))$
      - → $q^t(z) = P(Z|X,\theta)$→Assign Z by $P(Z|X,\theta)$
    - Maximization step
      - $\theta^{t+1} = argmax_\theta Q(\theta,q^{t+1}) = argmax_\theta L(\theta,q^{t+1})$
      - → fixed Z means that there is no unobserved variables
      - → Same optimization of ordinary MLE

Rethinking GMM Learning Process

- GMM, K-Means
  - We used EM algorithm to find the assignment of latent variables and the related distribution parameters
- EM algorithm
  - Initialize θ^0 to an arbitrary point
  - Loop until the likelihood converges
    - Expectation step
      - Assign Z by $P(Z|X,\theta)$
      - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P\left(x\middle|z_j = 1\right)} = \frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N\left(x\middle|\mu_j,\Sigma_j\right)}$
    - Maximization step
      - Same optimization of ordinary MLE
      - $\frac{d}{d\mu_k}\ln P(X|\pi,\mu,\Sigma) = 0, \frac{d}{d\Sigma_k}\ln P(X|\pi,\mu,\Sigma) = 0, \frac{d}{d\pi_k}\ln P(X|\pi,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) = 0$
      - $\widehat{\mu_k} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}, \Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \widehat{\mu_k})(\boldsymbol{x}_n - \widehat{\mu_k})^{\boldsymbol{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})}, \pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$

Further Readings

- Bishop Chapter 2 and 9
- Murphy Chapter 11