Sequential Data Modeling

Tomoki Toda
Graham Neubig
Sakriani Sakti

Augmented Human Communication Laboratory
Graduate School of Information Science

Syllabus

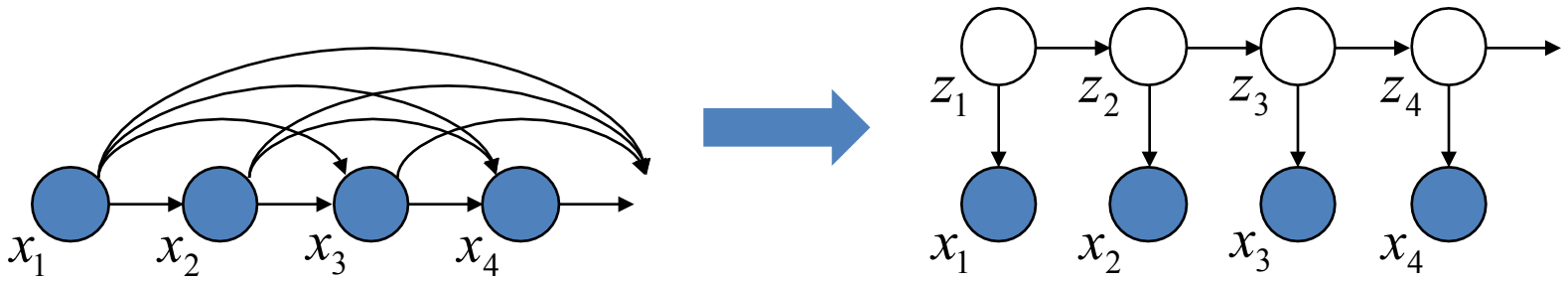
Date	Course description	Lecturer
6/03	Basics of sequential data modeling 1	Graham Neubig
6/10	**Basics of sequential data modeling 2**	**Graham Neubig**
6/17	Discrete latent variable models 1	Tomoki Toda
6/24	Discrete latent variable models 2	Tomoki Toda
7/1	Continuous latent variable models 1	Tomoki Toda
7/15	Discriminative models for sequential labeling 1	Sakriani Sakti
7/29	Continuous latent variable models 2	Tomoki Toda
8/1	Discriminative models for sequential labeling 2	Sakriani Sakti

Questions to: **sdm2016@is.naist.jp**

1st and 2nd Classes (6/3 and 6/10)

- Lecturer: Graham Neubig

- Contents: Basics of sequential data modeling
  - Markov process
  - Latent variables
  - Mixture models
  - Expectation-maximization (EM) algorithm



Sequential Data Modeling

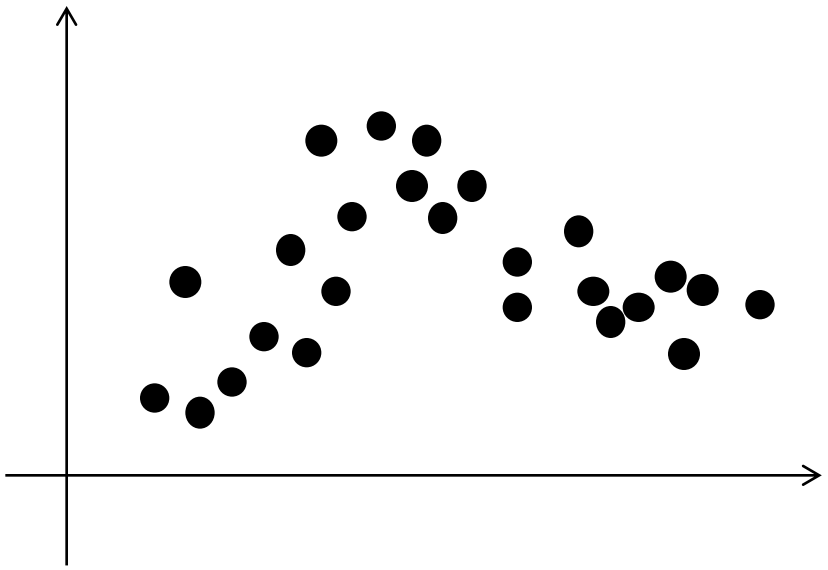
2nd class
"Basics of sequential data modeling 2"

Graham Neubig

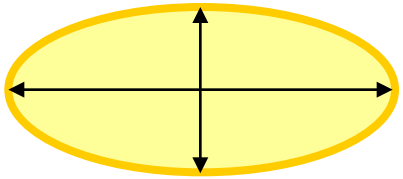
Augmented Human Communication Laboratory
Graduate School of Information Science

Question

2-dimensional data samples are observed as follows:



We can set the following types of distributions to any place.



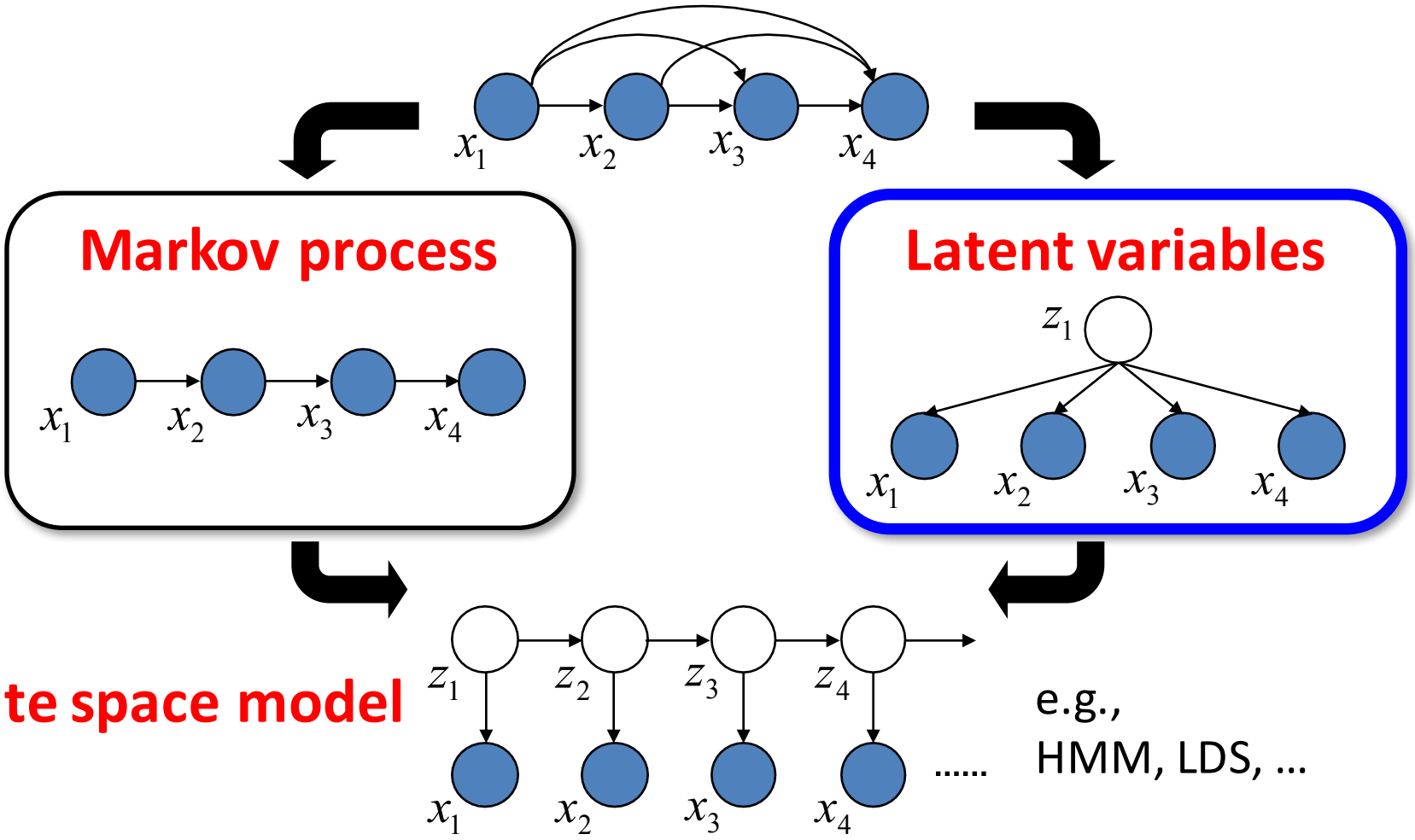
We can control each variance on x-axis and y-axis but we can not rotate this distribution.

Q. How can we model the distribution of these data samples using the above distributions?

After this class, you can answer these questions!

Basic Approaches

How to efficiently model joint probability of high-dimensional data (i.e., sequential data)



Towards Understanding Latent Variables

- In this class, we focus on a mixture model using a discrete latent variable.
- We assume that the length of sequential data is constant.

  e.g., if the length of each data is 5,

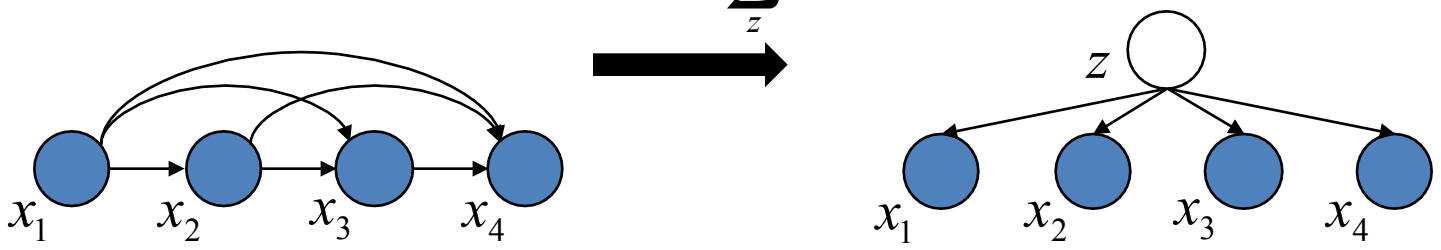
  Data 1: { 0.1, 0.2, -0.3, 1.3, -5.8 }

  Data 2: { 3.2, -4.8, -0.5, 0.6, -1.3 }

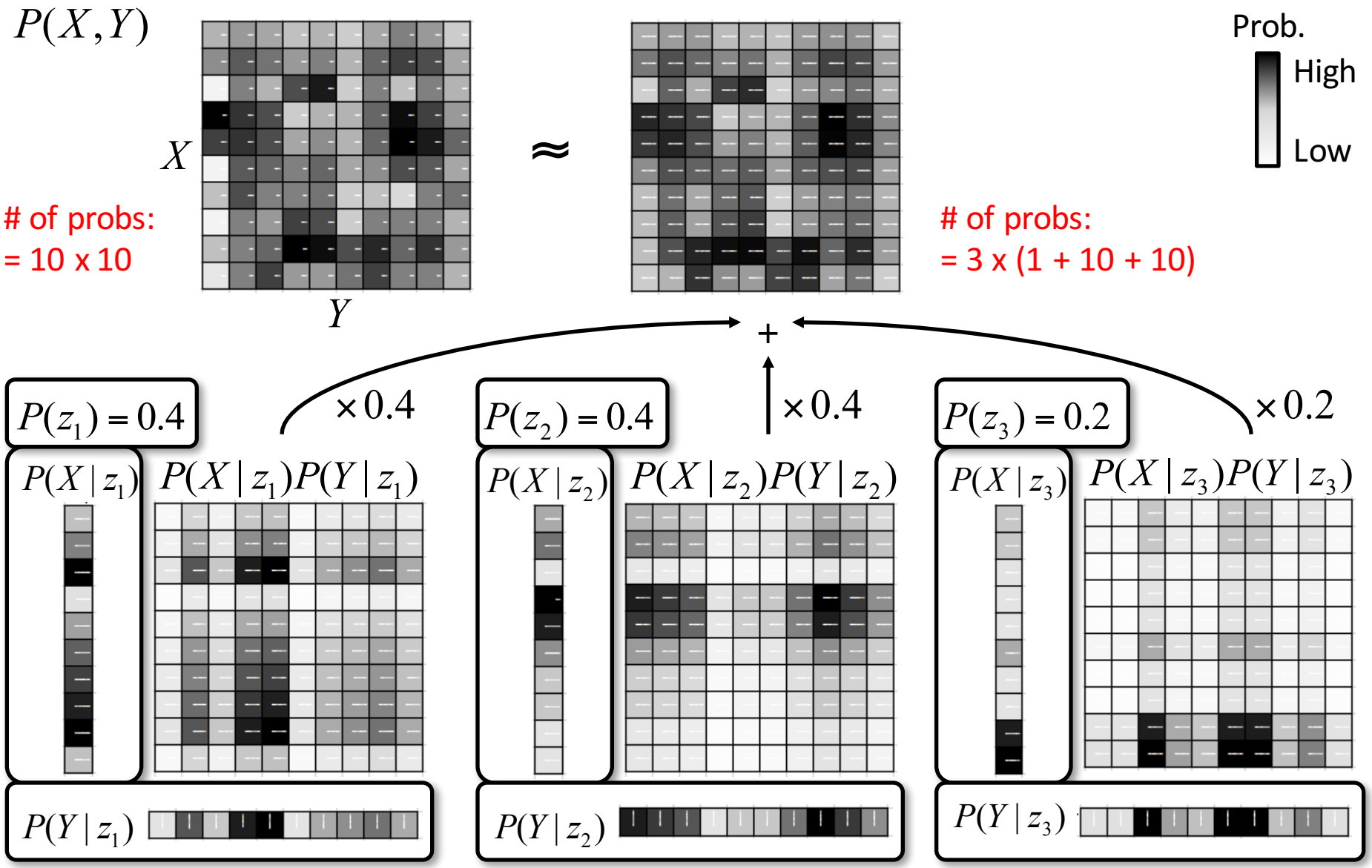
  ⋮

  each data sample is represented as a 5 dimensional vector.
- We learn how to model joint probability density in such a high-dimensional space using the mixture model.

$$P(x_1, x_2, x_3, x_4) = \sum_z P(z)P(x_1 \mid z)P(x_2 \mid z)P(x_3 \mid z)P(x_4 \mid z)$$



Effectiveness of Mixture Model

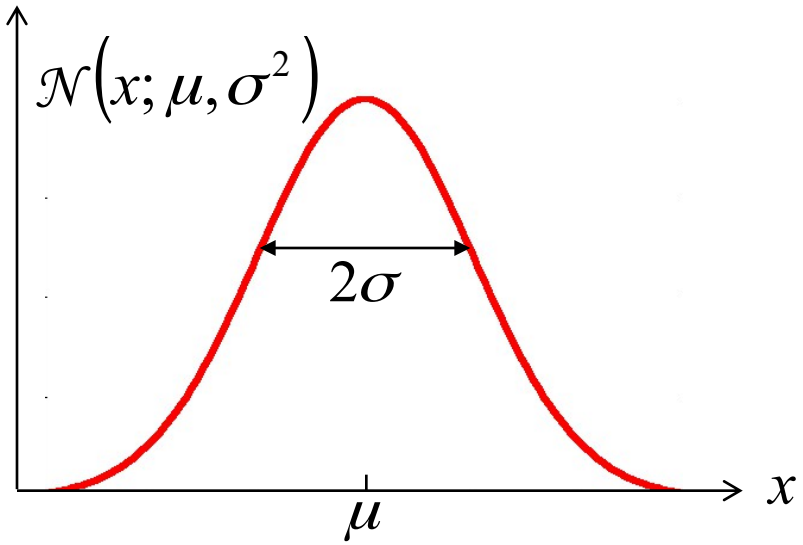


Mixture Model

“Gaussian Mixture Model”

The Gaussian Distribution

- Normal distribution or Gaussian distribution



Parameters { **Mean** : μ ; **Variance** : σ^2

Standard deviation: σ **Precision:** $\beta = 1/\sigma^2$

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Conditions to be satisfied:

$$\mathcal{N}(x;\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x;\mu,\sigma^2)\,dx = 1$$

Multivariate Gaussian Distribution

- Gaussian distribution over a D-dimensional vector $\boldsymbol{x}$ of continuous variables

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

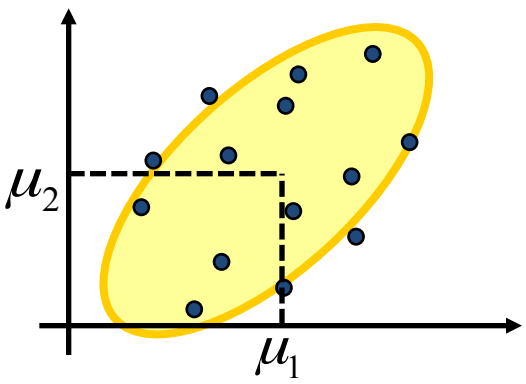
Parameters
- **Mean vector** : $\boldsymbol{\mu}$
- **Covariance matrix** : $\boldsymbol{\Sigma}$

Example of 2-dimensional case:

$$\boldsymbol{\mu} = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix}s_{11} & s_{12}\\ s_{21} & s_{22}\end{bmatrix}$$

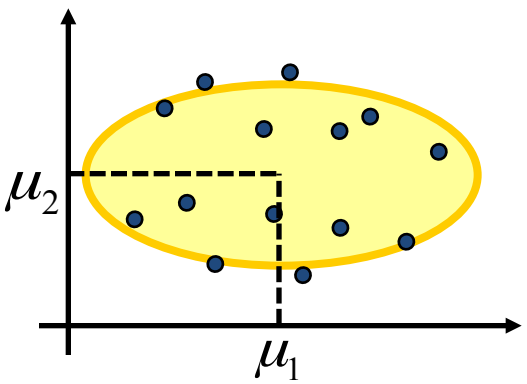
Use of full covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix}s_{11} & s_{12}\\ s_{21} & s_{22}\end{bmatrix}$$



Use of diagonal covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix}s_{11} & 0\\ 0 & s_{22}\end{bmatrix}$$

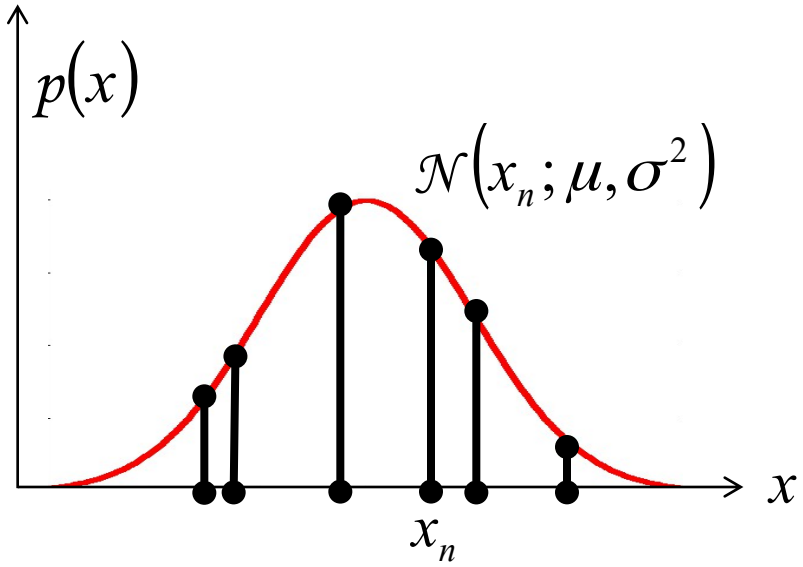


Maximum Likelihood Estimation (MLE)

- Observation data set (N observations of variable $\boldsymbol{x}$) $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$
  - Independent and identically distributed: i.i.d.
  - Data points drawn independently from the same distribution
- Likelihood function

$$p(\boldsymbol{X} \mid \boldsymbol{\lambda}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Because the data set $\boldsymbol{X}$ is i.i.d., $p(\boldsymbol{X} \mid \boldsymbol{\lambda})$ is given by $\prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{\lambda})$.



Determine parameters $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ using an observation data set by maximizing likelihood function!

ML Estimates of Mean and Covariance

- Determine model parameters by maximizing likelihood function

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \ln p(\boldsymbol{X} \mid \boldsymbol{\lambda})$$

ML estimate of mean vector:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\langle \boldsymbol{x} \rangle$$ Sample mean vector

The mean vector is the center of the data points

ML estimate of covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}}$$ Sample covariance

The covariance matrix is the average square of the samples, minus the square of the mean

Sufficient Statistics

Log-scaled likelihood function:

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n \mid \boldsymbol{\lambda})$$

$$= \sum_{n=1}^{N} -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

$$\propto \frac{N}{2}\left(\ln|\boldsymbol{\Sigma}^{-1}| - \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) + \frac{1}{2}\left(\sum_{n=1}^{N}\boldsymbol{x}_n\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\left(\sum_{n=1}^{N}\boldsymbol{x}_n\right) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathsf{T}}\right]$$

$$= \frac{N}{2}\left(\ln|\boldsymbol{\Sigma}^{-1}| - \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) + \frac{1}{2}\langle\boldsymbol{x}\rangle^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{x}\rangle - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\rangle\right]$$

$$\mathrm{tr}[\boldsymbol{ABC}] = \mathrm{tr}[\boldsymbol{BCA}] = \mathrm{tr}[\boldsymbol{CAB}]$$

# of samples : N

Sum of samples : $\langle\boldsymbol{x}\rangle = \sum_{n=1}^{N}\boldsymbol{x}_n \leftarrow \sum_{n=1}^{N}\boldsymbol{x}_n$

Sum of squared samples: $\langle\boldsymbol{x}\,\boldsymbol{x}^{\mathsf{T}}\rangle = \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathsf{T}} \leftarrow \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathsf{T}}$

Deriving ML Estimates

- Determine model parameters by maximizing likelihood function

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \ln p(\boldsymbol{X} \mid \boldsymbol{\lambda})$$

$$\left\{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right\} = \arg\max_{\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}} \frac{N}{2}\left(\ln\left|\boldsymbol{\Sigma}^{-1}\right| - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) + \frac{1}{2}\langle \boldsymbol{x}\rangle^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\langle \boldsymbol{x}\rangle - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\rangle\right]$$

ML estimate of mean vector:

$$\left.\frac{\partial \ln p(\boldsymbol{X} \mid \boldsymbol{\lambda})}{\partial \boldsymbol{\mu}}\right|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = -N\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\Sigma}}^{-1}\langle \boldsymbol{x}\rangle = \boldsymbol{0}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\langle \boldsymbol{x}\rangle \qquad \text{Sample mean vector}$$

ML estimate of covariance matrix:

$$\left.\frac{\partial \ln p(\boldsymbol{X} \mid \boldsymbol{\lambda})}{\partial \boldsymbol{\Sigma}^{-1}}\right|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = \frac{N}{2}\hat{\boldsymbol{\Sigma}} - \frac{N}{2}\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}} + \frac{1}{2}\langle \boldsymbol{x}\rangle\hat{\boldsymbol{\mu}}^{\mathrm{T}} + \frac{1}{2}\hat{\boldsymbol{\mu}}\langle \boldsymbol{x}\rangle^{\mathrm{T}} - \frac{1}{2}\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\rangle = \boldsymbol{0}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\rangle - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}} \qquad \text{Sample covariance}$$

Mixture Model

- Probability density function (*p.d.f.*) given by marginal *p.d.f.* of joint *p.d.f.* over latent variables

$$p(\boldsymbol{x} \mid \boldsymbol{\lambda}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z} \mid \boldsymbol{\lambda}) p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\lambda})$$

Latent variables (1-of-M):

$$\boldsymbol{z} = \{z_1, \cdots, z_M\} \quad z_m \in \{0,1\} \quad \sum_{m=1}^{M} z_m = 1$$

$\boldsymbol{z} = \{1,0,0\}$ use 1st component

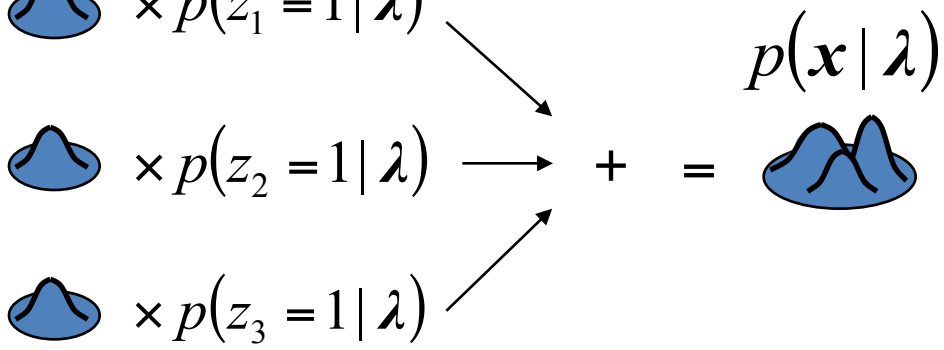
$\boldsymbol{z} = \{0,1,0\}$ use 2nd component

$\boldsymbol{z} = \{0,0,1\}$ use 3rd component

$$p(\boldsymbol{x} \mid z_1 = 1, \boldsymbol{\lambda}) \times p(z_1 = 1 \mid \boldsymbol{\lambda})$$

$$p(\boldsymbol{x} \mid z_2 = 1, \boldsymbol{\lambda}) \times p(z_2 = 1 \mid \boldsymbol{\lambda})$$

$$p(\boldsymbol{x} \mid z_3 = 1, \boldsymbol{\lambda}) \times p(z_3 = 1 \mid \boldsymbol{\lambda})$$



Gaussian Mixture Model (GMM)

- Mixture of multiple Gaussian distributions

$$p(\boldsymbol{x} \mid \boldsymbol{\lambda}) = \sum_{z} p(z \mid \boldsymbol{\lambda}) p(\boldsymbol{x} \mid z, \boldsymbol{\lambda})$$

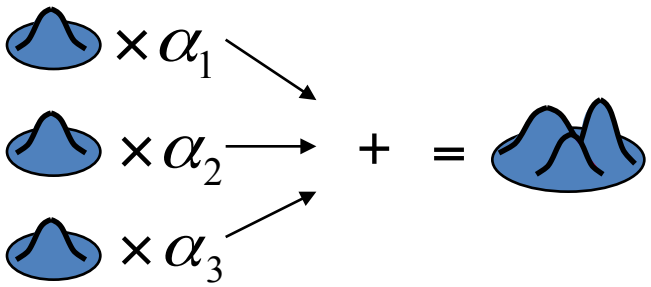
$$= \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

Prior probability of m^{th} mixture component

$$p(z_m = 1 \mid \boldsymbol{\lambda}) = \alpha_m \qquad 0 \le \alpha_m \le 1 \qquad \sum_{m=1}^{M} \alpha_m = 1$$

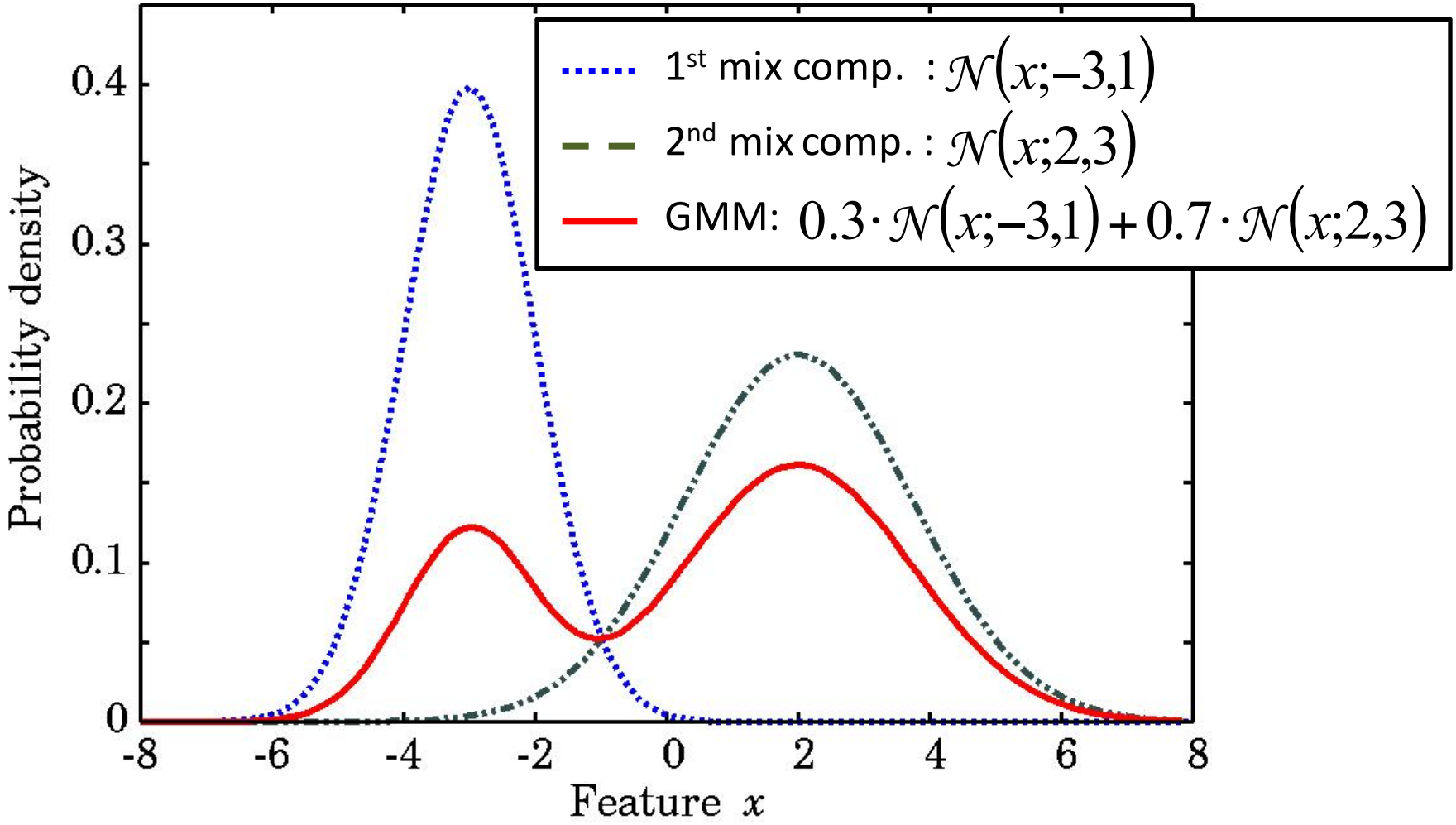
p.d.f. of m^{th} mixture component

$$p(\boldsymbol{x} \mid z_m = 1, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$



Example of $p.d.f.$ of GMM

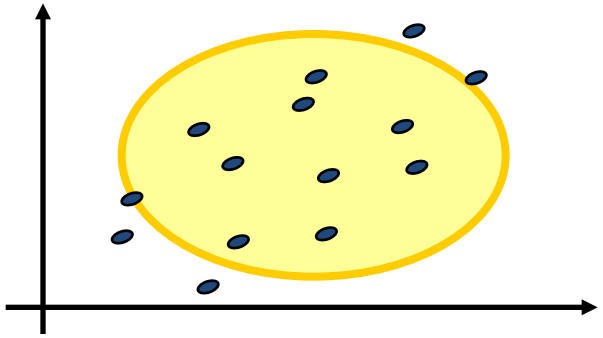
- GMM with 2 mixture components on 1-dimensional space



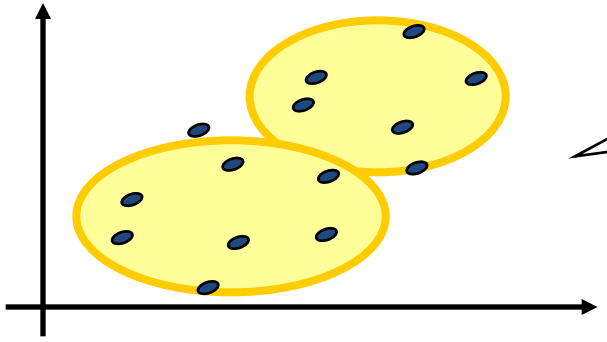
Effectiveness of GMM

- Capable of modeling inter-dimensional correlation even if using diagonal covariance matrices

If using a single Gaussian distribution w/ diagonal covariance matrix…



If using two Gaussian distributions w/ diagonal covariance matrices…



Capable of modeling inter-dimensional correlation!

How to Implement MLE?

Log-scaled likelihood of a mixture model:

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n \mid \boldsymbol{\lambda})$$

$$= \sum_{n=1}^{N} \ln \sum_{z} p(\boldsymbol{z} \mid \boldsymbol{\lambda}) p(\boldsymbol{x}_n \mid \boldsymbol{z}, \boldsymbol{\lambda})$$

$$= \sum_{n=1}^{N} \ln \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

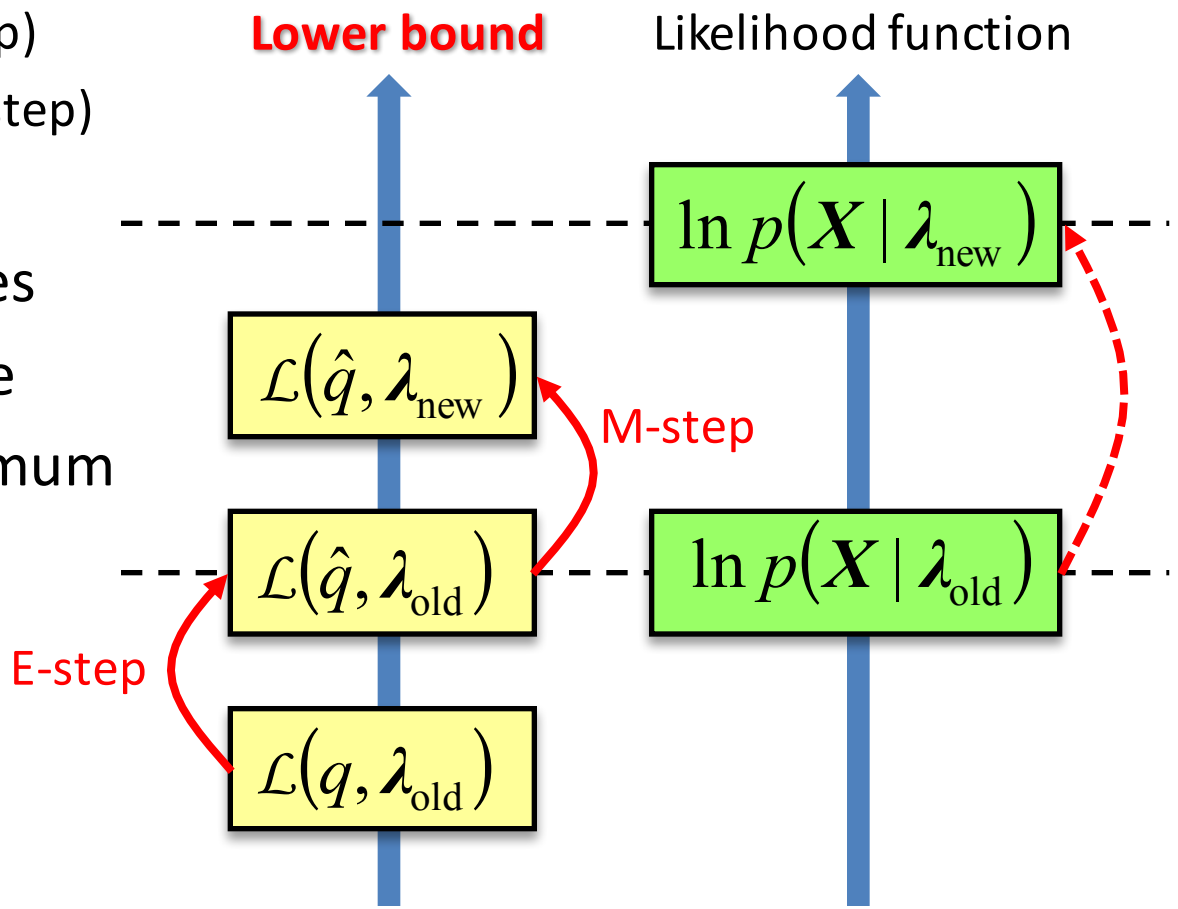
Because of the logarithm of the summation,
its derivatives with respect to model parameters
are not represented as linear equations…
Gradient methods will be necessary…

Can we push the logarithm inside the summation?

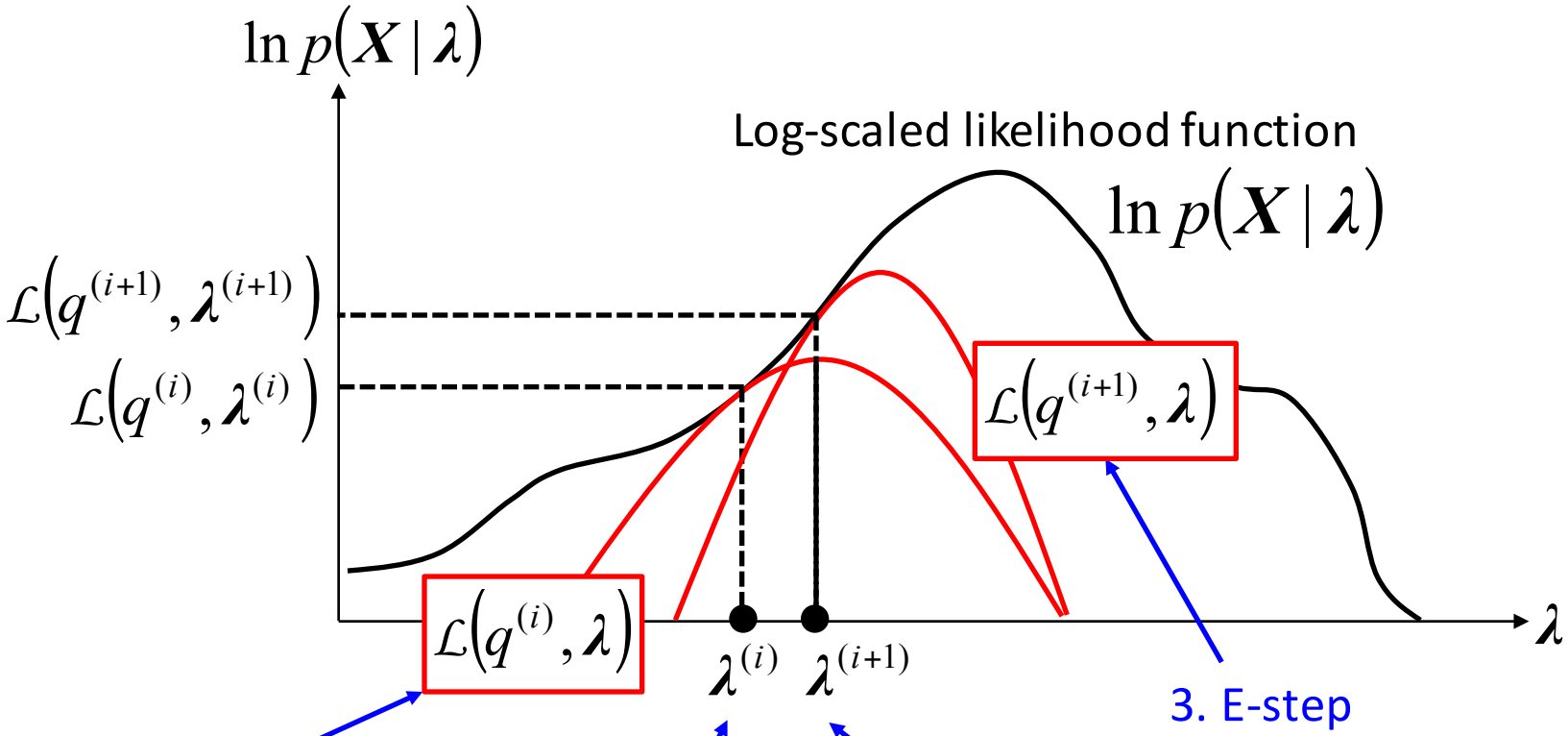
Expectation-Maximization (EM) Algorithm

EM Algorithm

- Iteratively update lower bound of likelihood function through two steps:
  - Expectation step (E-step)
  - Maximization step (M-step)
- Enable closed-form solution of ML estimates
- Guarantee convergence
- Converge to local maximum



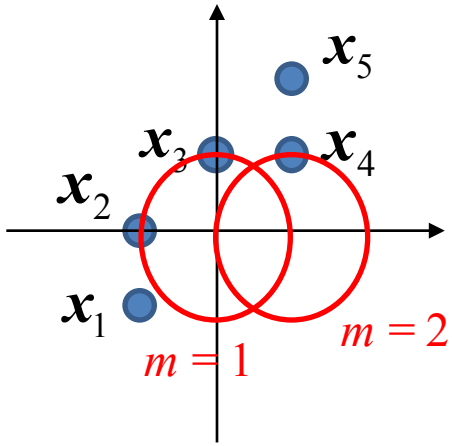
Schematic Image of EM Algorithm



Example in Gaussian Mixture Models

EM Algorithm for GMM

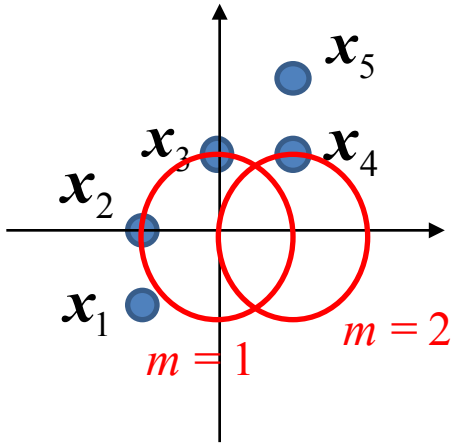
- Initialize: Prepare some training data ***x***
- Iterate:
  - E-step: Guess which distribution each data point in x belongs to
  - M-step: Update the parameters based on these guesses



Example Data

- Five 2-D dimensional samples given as training data

$$\boldsymbol{x}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \quad \boldsymbol{x}_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \boldsymbol{x}_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \boldsymbol{x}_4 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{x}_5 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$



- Initial model

$$p(\boldsymbol{x}_n \mid \boldsymbol{\lambda}) = \sum_{m=1}^{2} \alpha_m \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

1st mixture component w/ a diagonal covariance matrix:

$$\alpha_1 = 0.5 \qquad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

2nd mixture component w/ a diagonal covariance matrix:

$$\alpha_2 = 0.5 \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

E-Step

1. Calculate joint probabilities for individual mixture components

$$p(\boldsymbol{x}_n, z_n = m \mid \boldsymbol{\lambda})$$
$$= \alpha_m \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (\text{as } \boldsymbol{\Sigma}_m = \boldsymbol{I})$$
$$= \alpha_m \frac{1}{\sqrt{(2\pi)^2}} \exp\left(-\frac{1}{2} \sum_{d=1}^{2} (x_{n,d} - \mu_{m,d})^2 \right)$$

	$m=1$	$m=2$
$\boldsymbol{x}_1$	0.029	0.007
$\boldsymbol{x}_2$	0.048	0.011
$\boldsymbol{x}_3$	0.048	0.029
$\boldsymbol{x}_4$	0.029	0.048
$\boldsymbol{x}_5$	0.007	0.011

2. Calculate posterior probabilities (i.e., expected # of samples)

$$\gamma_{n,m} = p(z_n = m \mid \boldsymbol{x}_n, \boldsymbol{\lambda})$$
$$= \frac{\alpha_m \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{k=1}^{M} \alpha_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

	$m=1$	$m=2$
$\boldsymbol{x}_1$	0.82	0.18
$\boldsymbol{x}_2$	0.82	0.18
$\boldsymbol{x}_3$	0.62	0.38
$\boldsymbol{x}_4$	0.38	0.62
$\boldsymbol{x}_5$	0.38	0.62

Calculation of Sufficient Statistics

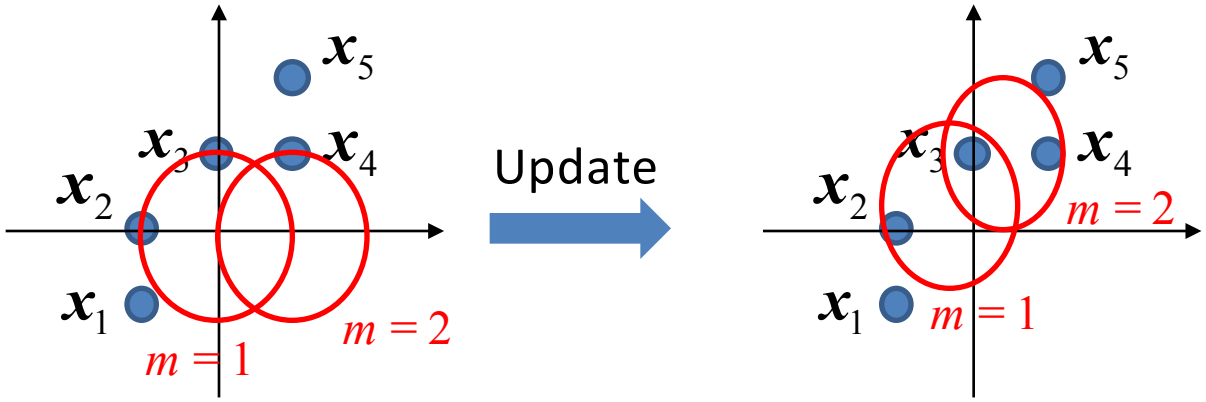
3. Calculate sufficient statistics

		 $m=1$	 $m=2$
# of samples	$\gamma_m = \sum_{n=1}^{N} \gamma_{n,m}$	3.01	1.99
Sum of samples	$\langle \boldsymbol{x} \rangle_m = \sum_{n=1}^{N} \gamma_{n,m} \boldsymbol{x}_n$	$\begin{bmatrix} -0.88 \\ 0.94 \end{bmatrix}$	$\begin{bmatrix} 0.88 \\ 2.06 \end{bmatrix}$
Sum of squared samples	$\langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle_m = \sum_{n=1}^{N} \gamma_{n,m} \boldsymbol{x}_n \boldsymbol{x}_n^{\mathsf{T}}$	$\begin{bmatrix} 2.39 & * \\ * & 3.33 \end{bmatrix}$	$\begin{bmatrix} 1.61 & * \\ * & 3.67 \end{bmatrix}$

M-Step

4. Update model parameters

		$m=1$	$m=2$
Mixture weights	$\hat{\alpha}_m = \frac{\gamma_m}{\sum_{n=1}^{M} \gamma_n}$	0.6	0.4
Mean vectors	$\hat{\boldsymbol{\mu}}_m = \frac{1}{\gamma_m} \langle \boldsymbol{x} \rangle_m$	$\begin{bmatrix} -0.3 \\ 0.3 \end{bmatrix}$	$\begin{bmatrix} 0.4 \\ 1.0 \end{bmatrix}$
Covariance matrices	$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{\gamma_m} \langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle_m - \hat{\boldsymbol{\mu}}_m \hat{\boldsymbol{\mu}}_m^{\mathsf{T}}$	$\begin{bmatrix} 0.7 & 0 \\ 0 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 0.6 & 0 \\ 0 & 0.8 \end{bmatrix}$



Increase of Likelihood

5. Calculate likelihood and compare it before and after update

$$\prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{\lambda}) = \prod_{n=1}^{N} \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

Before update:

	$m=1$	$m=2$	$P(\boldsymbol{x}_n \mid \boldsymbol{\lambda})$
$\boldsymbol{x}_1$	0.029	0.007	0.036
$\boldsymbol{x}_2$	0.048	0.011	0.059
$\boldsymbol{x}_3$	0.048	0.029	0.078
$\boldsymbol{x}_4$	0.029	0.048	0.078
$\boldsymbol{x}_5$	0.007	0.011	0.017

$$\frac{1}{N}\sum_{n=1}^{N} \ln p(\boldsymbol{x}_n \mid \boldsymbol{\lambda}) = -3.07$$

After update:

	$m=1$	$m=2$	$P(\boldsymbol{x}_n \mid \boldsymbol{\lambda})$
$\boldsymbol{x}_1$	0.034	0.001	0.035
$\boldsymbol{x}_2$	0.076	0.008	0.084
$\boldsymbol{x}_3$	0.084	0.078	0.163
$\boldsymbol{x}_4$	0.028	0.071	0.099
$\boldsymbol{x}_5$	0.008	0.039	0.048

$$\frac{1}{N}\sum_{n=1}^{N} \ln p(\boldsymbol{x}_n \mid \boldsymbol{\lambda}) = -2.60$$

Derivation

Lower Bound of Likelihood Function

- Derivation of lower bound of log-scaled likelihood function

Log-scaled likelihood function:

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\lambda})$$

$q(\boldsymbol{z})$ Probability distribution function of latent variables

$$= \sum_{n=1}^{N} \ln \sum_{z_n} q(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$$

Jensen's inequality

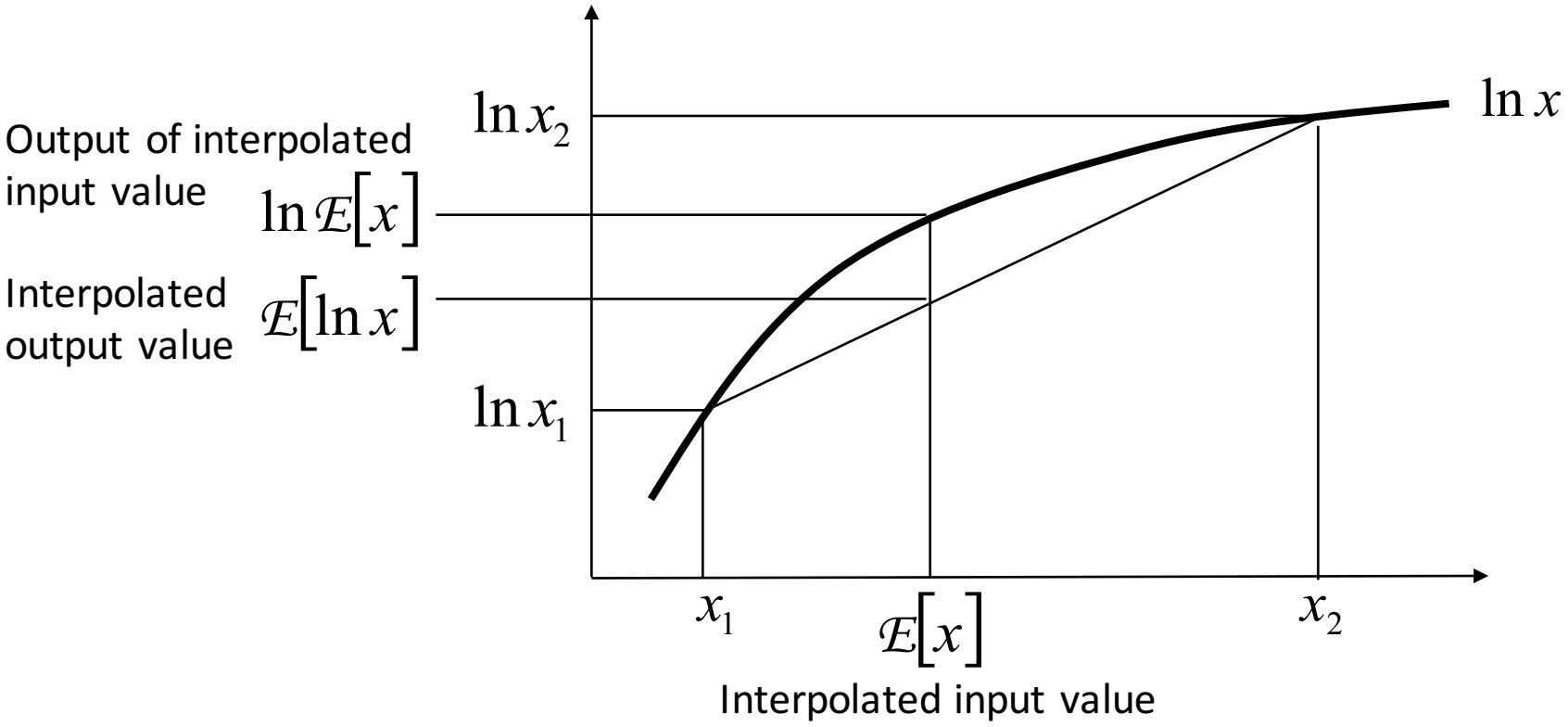
$$\geq \sum_{n=1}^{N} \sum_{z_n} q(\boldsymbol{z}_n) \ln \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$$

$$= \mathcal{L}(q, \boldsymbol{\lambda})$$

Lower bound: $\mathcal{L}(q, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \sum_{z_n} q(\boldsymbol{z}_n) \ln \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$

Jensen's Inequality

$$\ln \sum_n p_n x_n \geq \sum_n p_n \ln x_n \quad \text{subject to } p_n \geq 0, \sum_n p_n = 1$$



Lower Bound as Functional of q

Lower bound:

$$\mathcal{L}(q,\boldsymbol{\lambda}) = \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln\frac{p(\boldsymbol{x}_n,\boldsymbol{z}_n\mid\boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$$

$$= \frac{p(\boldsymbol{x}_n\mid\boldsymbol{\lambda})p(\boldsymbol{z}_n\mid\boldsymbol{x}_n,\boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$$

$$= \sum_{n=1}^{N}\left\{\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln p(\boldsymbol{x}_n\mid\boldsymbol{\lambda}) + \sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln\frac{p(\boldsymbol{z}_n\mid\boldsymbol{x}_n,\boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}\right\}$$

$$= \sum_{n=1}^{N}\left\{\ln p(\boldsymbol{x}_n\mid\boldsymbol{\lambda}) - \sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln\frac{q(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n\mid\boldsymbol{x}_n,\boldsymbol{\lambda})}\right\}$$

$$= \underbrace{\ln p(\boldsymbol{X}\mid\boldsymbol{\lambda})}_{\text{Log-scaled likelihood}} - \underbrace{\sum_{n=1}^{N}\mathrm{KL}(q(\boldsymbol{z}_n)\|\,p(\boldsymbol{z}_n\mid\boldsymbol{x}_n,\boldsymbol{\lambda}))}_{\text{KL divergence}}$$

"Lower bound" = "log-scaled likelihood" – "KL divergence"

KL Divergence

Discrete variable case:

$$\mathrm{KL}(p \| q) = \sum_{z} p(z) \ln \frac{p(z)}{q(z)}$$

Continuous variable case:

$$\mathrm{KL}(p \| q) = \int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$

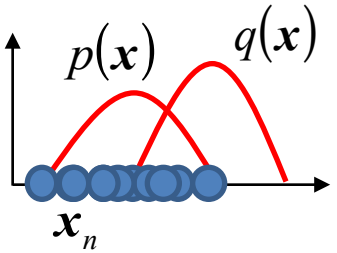
$$\mathrm{KL}(p \| q) = \int p(\boldsymbol{x}) \left(\ln p(\boldsymbol{x}) - \ln q(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x}$$

Approximated using samples $\boldsymbol{x}_n$ randomly generated from $p(\boldsymbol{x})$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n) - \ln q(\boldsymbol{x}_n)$$

Difference of log-scaled likelihoods

Log-scaled likelihood of true distribution ≥ Log-scaled likelihood of another distribution



$$\mathrm{KL}(p \| q) \geq 0$$

If $q = p$, $\mathrm{KL}(p \| q) = 0$

Lower Bound as Function of $\boldsymbol{\lambda}$

Lower bound:

$$\mathcal{L}(q,\boldsymbol{\lambda}) = \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln\frac{p(\boldsymbol{x}_n,\boldsymbol{z}_n\mid\boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$$

$$= \underbrace{\sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln p(\boldsymbol{x}_n,\boldsymbol{z}_n\mid\boldsymbol{\lambda})}_{\text{Auxiliary function}} \underbrace{-\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln q(\boldsymbol{z}_n)}_{\text{Constant}}$$

"Lower bound" $\propto$ "Auxiliary function"

EM: Maximization of Lower Bound

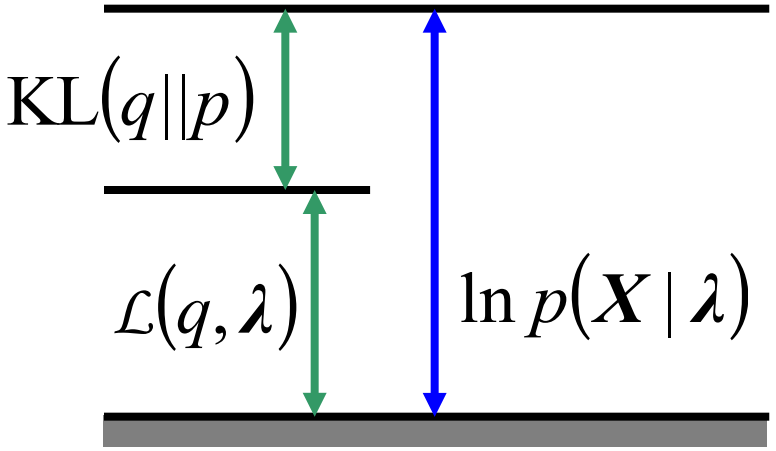
- Lower bound: functional of q and function of $\boldsymbol{\lambda}$

E-step: Maximize lower bound with respect to q while fixing $\boldsymbol{\lambda}$:

$$\mathcal{L}(q,\boldsymbol{\lambda}) = \ln p(\boldsymbol{X}\,|\,\boldsymbol{\lambda}) - \sum_{n=1}^{N} \mathrm{KL}(q(\boldsymbol{z}_n)\,\|\,p(\boldsymbol{z}_n\,|\,\boldsymbol{x}_n,\boldsymbol{\lambda}))$$

M-step: Maximize lower bound
with respect to $\boldsymbol{\lambda}$ while fixing q

$$\mathcal{L}(q,\boldsymbol{\lambda}) \propto \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln p(\boldsymbol{x}_n,\boldsymbol{z}_n\,|\,\boldsymbol{\lambda})$$



* Need to set initial model parameters $\boldsymbol{\lambda}$

E-Step: Update q

- Set KL divergence to 0 under the fixed model parameters $\boldsymbol{\lambda}_{\text{old}}$

$$\sum_{n=1}^{N} \mathrm{KL}\left(q(\boldsymbol{z}_n) \| p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}})\right) = 0 \quad \Rightarrow \quad \hat{q}(\boldsymbol{z}_n) = p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}})$$

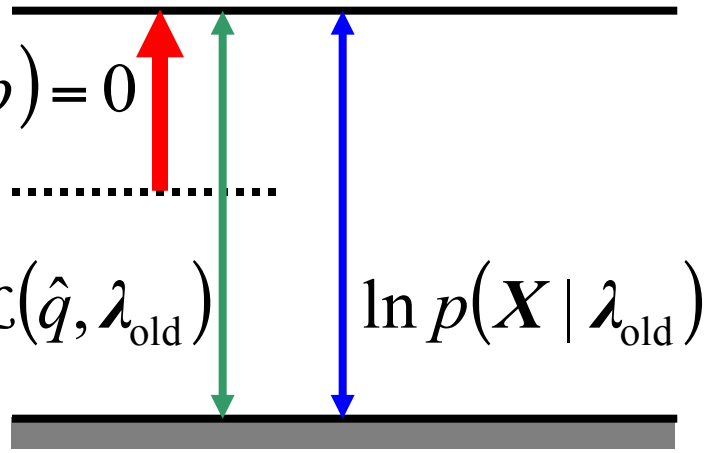
Calculate posterior probabilities of latent variables for each sample

$$p(z_{n,m} = 1 \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}})$$

$$= \frac{p(z_{n,m} = 1 \mid \boldsymbol{\lambda}_{\text{old}}) p(\boldsymbol{x}_n \mid z_{n,m} = 1, \boldsymbol{\lambda}_{\text{old}})}{\sum_{m=1}^{M} p(z_{n,m} = 1 \mid \boldsymbol{\lambda}_{\text{old}}) p(\boldsymbol{x}_n \mid z_{n,m} = 1, \boldsymbol{\lambda}_{\text{old}})}$$

$$= \frac{\alpha_m \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{k=1}^{M} \alpha_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

$$= \gamma_{n,m}$$



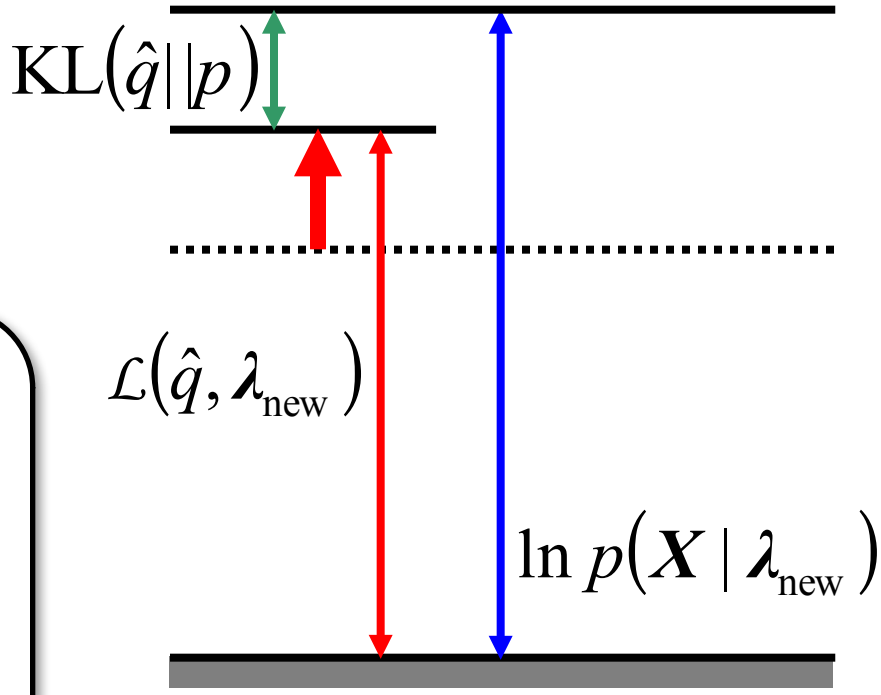
M-Step: Update $\boldsymbol{\lambda}$

- Maximize auxiliary function with respect to model parameters $\boldsymbol{\lambda}_{\text{new}}$

$$\mathcal{L}(\hat{q}, \boldsymbol{\lambda}) \propto \sum_{n=1}^{N} \sum_{z_n} \hat{q}(\boldsymbol{z}_n) \ln p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\lambda})$$

Auxiliary function

$$\begin{aligned} &Q(\boldsymbol{\lambda}_{\text{new}}, \boldsymbol{\lambda}_{\text{old}}) \\ &= \sum_{n=1}^{N} \sum_{z_n} p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}}) \ln p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\lambda}_{\text{new}}) \\ &= \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \left\{ \ln \alpha_m + \ln \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right\} \end{aligned}$$



Sufficient Statistics

Auxiliary function:

$$Q(\boldsymbol{\lambda}_{\text{new}}, \boldsymbol{\lambda}_{\text{old}}) = \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m} \left\{ \ln \alpha_m + \ln \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right\}$$

$$= \sum_{m=1}^{M} \gamma_m \left\{ \ln \alpha_m + \frac{1}{2} \ln \left| \boldsymbol{\Sigma}_m^{-1} \right| - \frac{1}{2} \boldsymbol{\mu}_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m \right\}$$

$$+ \sum_{m=1}^{M} \left\{ \frac{1}{2} \boldsymbol{\mu}_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \langle \boldsymbol{x} \rangle_m + \frac{1}{2} \langle \boldsymbol{x} \rangle_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m - \frac{1}{2} tr\left[\boldsymbol{\Sigma}_m^{-1} \langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle_m \right] \right\}$$

For each mixture component,

# of samples : m $\quad \gamma_m = \sum_{n=1}^{N} \gamma_{n,m} \longleftarrow \sum_{n=1}^{N} m$

Sum of samples : m $\quad \langle \boldsymbol{x} \rangle_m = \sum_{n=1}^{N} \gamma_{n,m} \boldsymbol{x}_n \longleftarrow \sum_{n=1}^{N} m\, n$

Sum of squared samples: m $\quad \langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle_m = \sum_{n=1}^{N} \gamma_{n,m} \boldsymbol{x}_n \boldsymbol{x}_n^{\mathsf{T}} \longleftarrow \sum_{n=1}^{N} m\, n\, n$

ML Estimates

Auxiliary function:

$$Q(\boldsymbol{\lambda}_{\text{new}}, \boldsymbol{\lambda}_{\text{old}}) = \sum_{m=1}^{M} \gamma_m \left\{ \ln \alpha_m + \frac{1}{2} \ln \left| \boldsymbol{\Sigma}_m^{-1} \right| - \frac{1}{2} \boldsymbol{\mu}_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m \right\}$$
$$+ \sum_{m=1}^{M} \left\{ \frac{1}{2} \boldsymbol{\mu}_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \langle \boldsymbol{x} \rangle_m + \frac{1}{2} \langle \boldsymbol{x} \rangle_m^{\mathsf{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m - \frac{1}{2} tr\left[\boldsymbol{\Sigma}_m^{-1} \langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle_m \right] \right\}$$

$$\left. \frac{\partial Q(\boldsymbol{\lambda}_{\text{new}}, \boldsymbol{\lambda}_{\text{old}}) + \beta \left(\sum_{n=1}^{M} \alpha_n - 1 \right)}{\partial \alpha_m} \right|_{\lambda = \hat{\lambda}} = 0$$

$$\left. \frac{\partial Q(\boldsymbol{\lambda}_{\text{new}}, \boldsymbol{\lambda}_{\text{old}})}{\partial \boldsymbol{\mu}_m} \right|_{\lambda = \hat{\lambda}} = \boldsymbol{0}$$

$$\left. \frac{\partial Q(\boldsymbol{\lambda}_{\text{new}}, \boldsymbol{\lambda}_{\text{old}})}{\partial \boldsymbol{\Sigma}_m^{-1}} \right|_{\lambda = \hat{\lambda}} = \boldsymbol{0}$$

For each mixture component,

Mixture weights $\hat{\alpha}_m = \dfrac{\gamma_m}{\sum_{n=1}^{M} \gamma_n}$

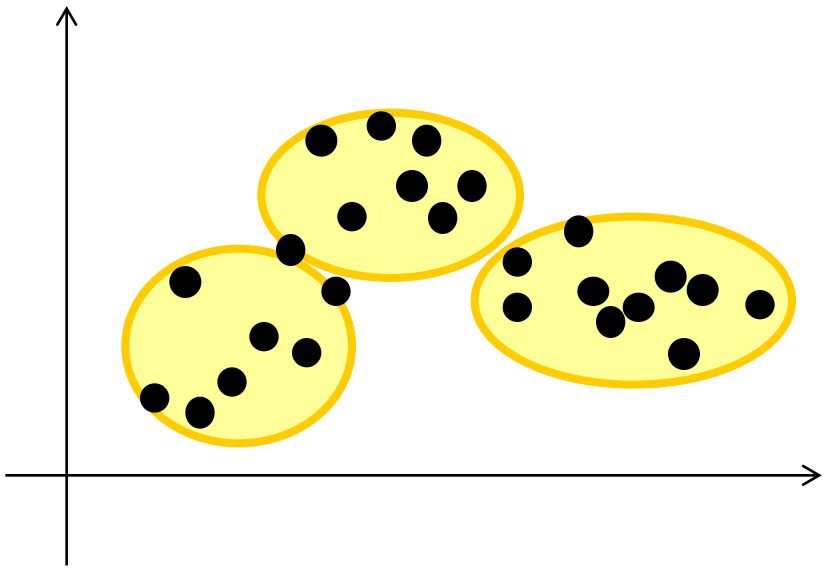
Mean vectors $\hat{\boldsymbol{\mu}}_m = \dfrac{1}{\gamma_m} \langle \boldsymbol{x} \rangle_m$

Covariance matrices $\hat{\boldsymbol{\Sigma}}_m = \dfrac{1}{\gamma_m} \langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \rangle_m - \hat{\boldsymbol{\mu}}_m \hat{\boldsymbol{\mu}}_m^{\mathsf{T}}$

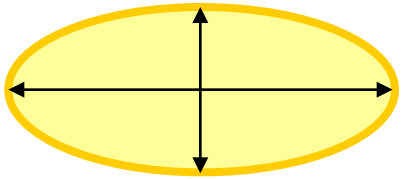
Conclusion

You Can Answer the Question!

2-dimensional data samples are observed as follows:



We can set the following types of distributions to any place.



We can control each variance on x-axis and y-axis but we can not rotate this distribution.

Q. How can we model the distribution of these data samples using the above distributions?

Let's use GMM and optimize it with EM algorithm!